



BioSpark: Beyond Analogical Inspiration to LLM-augmented Transfer

Hyeonsu B Kang

Human-Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
hyeonsuk@andrew.cmu.edu

Matthew K. Hong

Toyota Research Institute
Los Altos, California, USA
matt.hong@tri.global

David Chuan-En Lin

Human-Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
chuanenl@cs.cmu.edu

Yan-Ying Chen

Toyota Research Institute
Los Altos, California, USA
yan-ying.chen@tri.global

Nikolas Martelaro

Human-Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nmartelaro@gmail.com

Aniket Kittur

Human Computer Interaction
Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
nkittur@cs.cmu.edu

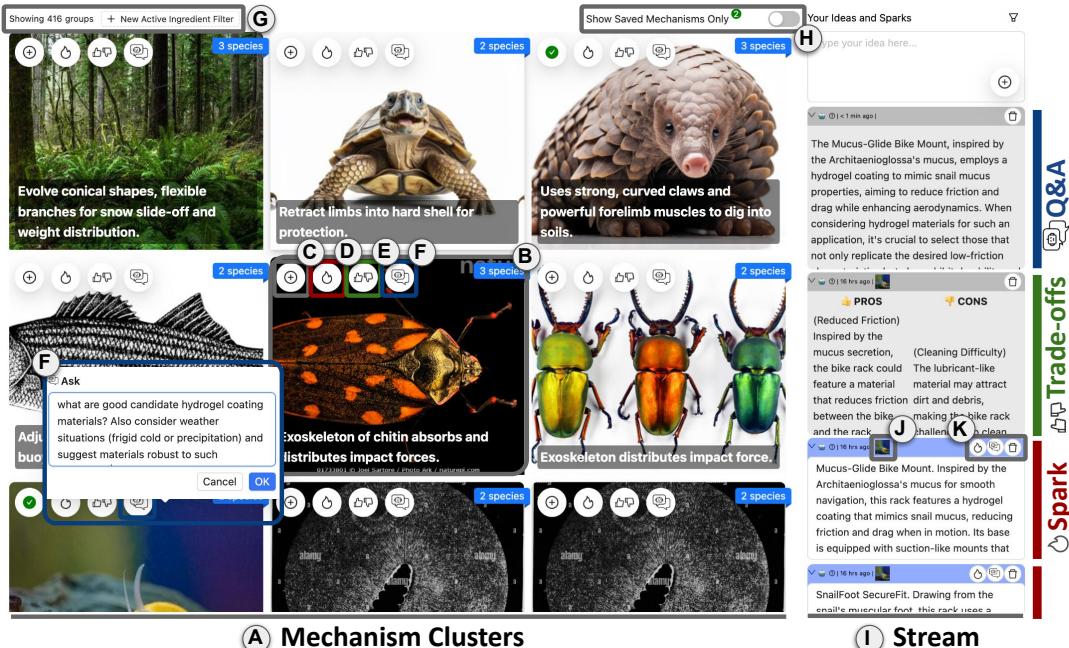


Figure 1: The BioSPARK interface consists of *mechanism active ingredient* clusters (Ⓐ) and a stream panel (Ⓘ). Cluster cards (Ⓑ) display species images, descriptions, and action buttons, enabling users to save mechanisms (Ⓒ), generate Sparks (Ⓓ), explore Trade-offs (Ⓔ), and use a Q&A chat (Ⓕ). The stream panel (Ⓘ) contains system- and user-generated ideas, which are interactive and editable (Details in §4).

Abstract

We present BIO SPARK, a system for analogical innovation designed to act as a creativity partner in reducing the cognitive effort in finding, mapping, and creatively adapting diverse inspirations. While

prior approaches have focused on initial stages of finding inspirations, BIO SPARK uses LLMs embedded in a familiar, visual, Pinterest-like interface to go beyond inspiration to supporting users in identifying the key solution mechanisms, transferring them to the problem domain, considering tradeoffs, and elaborating on details and characteristics. To accomplish this BIO SPARK introduces several novel contributions, including a tree-of-life enabled approach for generating relevant and diverse inspirations, as well as AI-powered cards including ‘Sparks’ for analogical transfer; ‘Trade-offs’ for considering pros and cons; and ‘Q&A’ for deeper elaboration. We



This work is licensed under a Creative Commons Attribution 4.0 International License.

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3714053>

evaluated BioSPARK through workshops with professional designers and a controlled user study, finding that using BioSPARK led to a greater number of generated ideas; those ideas being rated higher in creative quality; and more diversity in terms of biological inspirations used than a control condition. Our results suggest new avenues for creativity support tools embedding AI in familiar interaction paradigms for designer workflows.

CCS Concepts

- Human-centered computing → Interactive systems and tools; Interaction techniques; HCI design and evaluation methods;
- Information systems → World Wide Web; Users and interactive retrieval.

Keywords

Goal-driven Analogies, Analogical Transfer, Design Creativity, Ideation, Large Language Models

ACM Reference Format:

Hyeonsu B Kang, David Chuan-En Lin, Yan-Ying Chen, Matthew K. Hong, Nikolas Martelaro, and Aniket Kittur. 2025. BioSpark: Beyond Analogical Inspiration to LLM-augmented Transfer. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 29 pages. <https://doi.org/10.1145/3706598.3714053>

1 Introduction

Many innovations in design, technology, and science have been driven by people finding and adapting analogical inspirations from fields distant to their own. *Analogical innovation* refers to the process of applying insights from one domain to solve problems or create new designs in another. Whether Vitruvius explaining how sound waves work through analogy with water waves [13], the Wright brothers designing a lightweight wing control mechanism based on a bicycle inner tube box [37], or engineers partnering with an origami expert to furl a solar array into a narrow rocket [50, 58, 75], such innovations have required their inventors to deeply engage in a complex cognitive process of finding and creatively adapting inspirations that had limited surface similarities but deep structural similarities.

While such analogies may sometimes seem like ‘lightning strikes’ of serendipity, researchers have identified that analogical innovation involves several cognitive stages of processing, each of which can require significant mental effort [22, 23, 26]. First, **finding** potential inspirations in distant domains is difficult because of the challenge of identifying domains and inspirations that might contain useful potential analogical mechanisms. However, even when potential inspirations are found, significant cognitive effort is needed as part of the broader process of *analogical processing* to **recognize** and **understand** what the *active ingredients* of the inspiration’s mechanisms are, such as key structural or functional properties that enable its utility (e.g., the shearing properties of the cardboard inner tube box). Once identified, another barrier involves exploring how to map and **transfer** these mechanisms to the target problem (a process known as *analogical transfer*). For example, instead of four sides of a box, using a set of cables to create shearing in two parallel planes of the wings). Finally, the mechanisms need to be **adapted** to consider tradeoffs and limitations in the target

domain (e.g., whether cables for connecting the wings could provide a sufficient balance of rigidity and weight reduction for steering the wings through shearing).

Supporting these complex needs in a single system has been challenging, with most past approaches focusing on one or two stages of the process and largely limited to a small, hand-coded set of inspirations [9, 15, 28, 36]. Approaches to collecting inspirations at a larger scale have begun to appear [17, 33, 34, 38], but have mostly been limited to helping with the finding stage of analogical innovation. Relying on users to do the hard work of determining which inspirations could be relevant and how they could be adapted can lead to them not noticing or putting in the effort to go beyond surface similarities and try to understand how an inspiration could be used; as noted in Kang et al. 2022, “the critical first step towards analogical inspiration may be raising... enough attention and interest above the initial ‘hump’ of cognitive demand” [38].

In BioSPARK, we explore the idea of a LLM-powered computing system acting as a creativity partner to proactively help with the intellectual work of not only finding analogies but also mapping, transferring, and adapting those ideas to the target domain. By doing so we aim to help free up the cognitive effort of users to engage in the creative process of exploring new design spaces and considering more ideas more deeply than they would be able to otherwise.

To achieve this, BioSPARK explores several new design patterns for partnering AI with human analogical ideation, including:

- A tree-of-life enabled approach for generating new and relevant biological inspirations from a small set of ‘gold standard’ inspirations taken from AskNature;
- An analogical ideation interface leveraging familiar interaction concepts from designers’ practice of browsing Pinterest and curating moodboards that helps them recognize the active ingredients of the inspirations’ mechanisms;
- Proactively generating ‘sparks’ that help users understand the mapping and transfer between inspirations and their design problem;
- Providing pro/con trade-offs to scaffold users in considering how to adapt key aspects of the design problem;
- Supporting a free-form Q&A interface grounded in the inspiration and the design problem context to help users more deeply consider and elaborate on inspiration mechanisms.

We instantiated BioSPARK in a prototype system and evaluated and iterated on it through a workshop study with 6 professional designers, a formative study with 4 participants with design and engineering backgrounds as well as a user study with 12 participants of varied backgrounds. Our results suggest novel ways in which AI support can be embedded into interfaces to support and augment human creativity.

2 Related Work

2.1 Design by analogy

Throughout history, analogies have often driven breakthroughs in science, engineering, and design (e.g., [13, 37, 50]). Yet, analogical innovation in human minds has proven rare due to the cognitive challenges involved with the underlying analogical processing. One challenge is the high sensitivity to surface-level similarity

during retrieval from memory that favors analogs with shared visual or keyword similarities over the ones that share a deeper underlying structure [23]. In addition, the heavy cognitive load incurred during analogical processing, even with just a few relations, significantly burdens working memory and leads to performance degradation [24, 27, 29]. To support people with analogical processing, researchers have designed various systems for analogy retrieval. One thread of research here focuses on modeling analogical relations, albeit in limited scopes. This includes system based on the structure-mapping theory [20–22], multiconstraints theories (e.g., [32]), connectionist designs [31, 35], and rule-based approaches [4, 7, 8]. Many methods involve labor-intensive processes, such as the WordTree methodology [44]. Additionally, numerous systems depend on hand-coded and meticulously structured data, the curation of which is often resource-intensive (e.g., [28, 67]).

Recent work in computational methods for finding analogical inspirations at scale have shown promising results using a significantly simplified schema (e.g., the purpose and mechanism schema in [33, 38]) with just a fraction of data (e.g., [10, 33, 38]). However these systems primarily focus on facilitating the discovery of potential analogies and do not extend support to the subsequent, intricate stages of design that follow. This involves navigating potential limitations or trade-offs, which are essential for the successful transfer of these analogies in real-world scenarios [3, 62, 67].

2.2 Bioinspired design

One particularly relevant thread of research in design-by-analogy focuses on finding inspirations in biological organisms and systems [36]. However, prior approaches have been limited due to their reliance on costly manual curation (e.g., AskNature [15] or DANE [28]; for example, the researchers of DANE found that re-describing a single biological organism in the Structure-Behavior-Function framework can take approximately ~40-100 hours per model). Alternative approaches demonstrated the feasibility of using crowdsourcing to power supervised learning for identifying scientific articles with biomimetic inspirations (e.g., [66, 73]), but the cost of curating high-quality annotations presented a significant bottleneck for scalability. Yet another line of research has explored rule-based (e.g., [11]) or data programming [17] approaches, and showed promising results, albeit potential concerns of their generalizability and scalability.

Our iterative tree-of-life-based algorithm for expanding the mechanism dataset builds on these threads of research, while also leveraging recent advances in AI, such as Large Language Models (LLMs), that present promising new opportunities for designing scalable approaches for bio-analogy generation. However, naively prompting LLMs in a zero-shot manner may still result in limited diversity on abstract concepts [12]. One possible avenue of research here is to further exploring knowledge-augmented or knowledge-guided prompting, which has been previously explored in factual Q&A for improving the factuality in answers to simple questions (e.g., “*Where did Alex Chilton die?*”) by traversing a knowledge graph known to contain relevant facts to contextualize LLM prompts, for the goal of increasing the conceptual diversity in generation. Another related thread here is recent work on self-feedback and

refinement techniques [46], that showed iteratively generating feedback on the LLM output using the same LLM and refining based on it leads to better performance. Our tree-of-life enabled approach takes inspirations from these prior works to develop an algorithm that uses the knowledge structure for diversification.

2.3 Creativity Support Systems (CSTs)

BIOSPARK, in its goal of developing a creativity partner for analogical innovation, also builds on the related work in Creativity Support Systems (CSTs). Specifically, co-creativity [47] and human-AI collaborative systems have been explored in the CST literature, leading to related systems such as Drawing Apprentice [14], Due-Draw [54], Creativity Sketching Partner [41], and Collaborative Ideation Partner [42] which aimed to produce useful output for building on human user input with a joint similarity to both the design task and the input. The goal-driven approach in BIOSPARK also focuses on the relevance to the task context, but differs from these works in its emphasis on analogical relevance which relies on structural rather than feature similarity.

Another related thread of research in the CST literature is systems that support analogies as an ideation method, as seen in related works such as MetaMap [40], WikiLink [76], Idea-Inspire [62]. These works focus on the initial stages of helping users find relevant analogies, but stop short of helping users in the cognitive process of analogical transfer. While transfer has been studied as a critical failure point in analogy for decades in cognitive psychology [27], it remains understudied in the CST literature, perhaps due to the difficulty of even finding analogical inspirations, let alone addressing how to transfer them to the target domain. BIOSPARK contributes to this gap in the literature by helping designers recognize, transfer, explore and elaborate on new design spaces through analogies and their example instantiations, which [43] identified as “grounded metaphors” and highlighted as open research areas.

2.4 LLMs for ideation and co-creation

Recent advances in LLMs also suggest the potential for scalably augmenting analogical innovation for users throughout the entire cognitive process, from finding potential analogical inspirations to mapping them to the problem domain to helping users more deeply engage with their mechanisms and trade-offs. LLMs have shown the capability to infer specific analogies and to generate ideas relevant to a design goal (cf. [69]). They also can serve as more flexible natural language processing components in an interface, allowing for powerful interface augmentation approaches (e.g., [5, 19, 39, 45] as well as direct interfaces using chat-based dialog (e.g., [1, 49, 55]).

However, studies examining the use of LLMs and generative AI in the creative process have shown that improperly incorporating LLMs into the creative process can end up doing more harm than good. Using generative AI systems such as image generation (e.g., Midjourney) or text generation (e.g., ChatGPT) has been shown to lead users to become more fixated rather than more creative [68]. Several core properties to LLMs have been identified as potentially problematic, including tendencies for inaccurate inferences and hallucinations, user fixation on the initial prompts they enter, and overly accepting the results of AI-generated ideas rather than adapting them or using them to further explore the design space [42, 68].

These results suggest a more nuanced approach to incorporating LLMs and AI into the analogical innovation process may be needed.

3 Formative Studies and Design Goals

We developed BioSPARK in an iterative process involving an initial formative workshop study with professional automobile designers for investigating their current workflows of getting inspirations for concept design, followed by a design probe with additional designers using a partial development of the initial BioSPARK's goal-driven data discovery pipeline which generated a set of analogical mechanism inspirations (§4.4).

Our initial day-long workshop included six professional designers from a large multinational company specializing in mobility design. The workshop provided insights into how our system could integrate into their daily workflows, highlighting basic design requirements such as the **importance of visual, goal-centered inspirations** and **leveraging familiar interfaces** for a system that could be incorporated into their practice, which provided constraints for the development of BioSPARK.

Informed by the workshop findings, we developed a design probe to further investigate challenges in AI-augmented bioinspired design. The probe consisted of a set of images of inspirations arranged in a moodboard or Pinterest-like visual style that users could scroll through. Inspirations were sampled from a prototype of the inspiration generator we describe later, and focused on two design briefs: designing a secure bike rack and improving driving on slippery roads. This included mechanisms ranging from geckos climbing walls to slime lubrication. Each individual mechanism had an associated visual image retrieved using online search (see Appendix A.1). To help designers experience what potential AI assistance could feel like we included prototype functionality for designers to interact with these inspirations (*i.e.*, buttons that would prompt an LLM to provide an additional explanation, or to compare or combine a pair of mechanisms; see more in Appendix A.2).

We recruited four designers (none of whom participated in the earlier workshop) with backgrounds in design and engineering to provide feedback as pilot participants. Participants found aspects of the prototype valuable, and all did find mechanisms that inspired new design ideas (e.g., the coiling of octopus tentacles and lizard tails inspiring bike rack components that could expand and contract with turbulence; or scale and fur arrangements in rodents inspiring groove patterns on tires that would create more downforce on slippery roads).

However, they also brought up several fundamental challenges with the approach that motivated BioSPARK. In particular, all participants noted a similar theme in terms of the challenges involved with **recognizing and understanding** the 'active ingredients' (*i.e.*, the core abstraction underpinning how each mechanism actually works) of inspirations with respect to the design problem. For example, it could also be difficult to understand how the textual descriptions of individual mechanisms were relevant to the design problem (e.g., "*I want to choose an interesting animal that can support force and dynamics... but there are a lot of bugs, birds, and dolphins, not sure how to think about them as relating to turbulence reduction*" – P3).

Participants also wanted additional support for envisioning how inspirations would **transfer into target design domains**. Participants commented that assisting them with applying active ingredients to focused areas in the target domain could help with generating new design ideas based on distant analogies "*slime secretion as relating to the attachment/detachment mechanism and the friction aspect rather than the slime itself would be more helpful*" – P1; "*Maybe it (the system) can tell me about skin texture of frogs for generating ways to modulate it and manage turbulence?*" – P2).

Finally, participants noted the desire for **deeper elaboration on inspirations**. This included understanding benefit and drawbacks of the mechanism with respect to dimensions relevant to its effectiveness in the problem domain, as in this quote from P4: "*A few words to highlight the most interesting properties or dimensions for efficiency, durability, or versatility would be very helpful. I want to get a bigger picture here*" (P4). It also included understanding more details about a particular mechanism, as commented by P1: "*'Bike rack' and 'slime' are somewhat contradictory but it (slime mechanism) makes me think about the attachment aspects of the design... maybe new ideas around loading and unloading of bikes that have dynamically adjusting surface friction... I'm going to click "explain" on slime... (after the detailed explanation loads) I wish I could know more about the lubrication mechanism aspect of slimes*".

4 BIO SPARK

We instantiated these design goals in BioSPARK, with the high level intention of acting as a creative partner in the analogical design and innovation process beyond simply finding inspirations to proactively helping the user understand the key solution mechanisms in those inspirations, how to transfer them to their own problem domain, and to more easily consider their trade-offs on the problem's design constraints or elaborate on more information about their details and characteristics.

To support these functions in a simple and familiar interaction the AI provides its suggestions on mappings between inspirations and the problem as idea 'spark' cards that are added to a sidebar whenever the user saves an inspiration; 'trade-off' cards that contextualize the pros and cons of an inspiration's mechanism within the problem domain; and Q&A cards that allow the user to submit free-form queries to the LLM which are automatically contextualized with the problem and inspiration contexts. As the user engages with the system a typical flow involves them perusing and saving inspirations, and engaging with the cards in the sidebar to more deeply consider particular mechanisms or the design spaces they represent.

In the following sections we describe in more detail the design of the system, starting with a detailed description and a scenario walkthrough of the system.

4.1 Scenario

Consider an automotive designer, Sarah, looking for inspirations that could spark new ideas for novel bike rack design. When she arrives at the BioSPARK interface, she first scrolls through the board UI on the left of the screen to review different clusters of mechanism active ingredients. She is initially intrigued by the 'exoskeleton' cluster, which shows an image of a froghopper, because the exoskeleton structure may provide insights for new skeletal bike rack

designs. She clicks the cluster card (fig. 1, ⑩) to examine its details further. The mechanism description in the modal that expands out on her click highlights ‘chitin’, as strong yet flexible material, that can absorb and distribute the force of impact.

She then finds another mechanism that seems counter-intuitive yet interesting, the mucus and muscular foot of ‘Architaenioglossa’ – an order of snails – as potentially interesting active ingredients for the problem. She clicks on the ‘spark’ button to receive inspirations for creatively adapting the active ingredients (fig. 1, ⑪). She receives two sparks in response; the first, titled ‘Mucus-Glide Bike Mount’, describes an idea that uses hydrogel coating to reduce friction in motion. Intrigued by the idea, but concerned with the durability of hydrogel in various weather conditions, she asks BIOSPARK using the ‘Q&A’ button (fig. 1, ⑫): “*what are good candidate hydrogel coating materials? Also consider weather situations (frigid cold or precipitation) and suggest materials robust to such conditions.*”. BIOSPARK returns an information card that provides alternative material choices, such as Polyacrylamide Hydrogels, described as capable of maintaining their mechanical strength and elasticity in a wide range of temperatures and as resistant to degradation in wet conditions, or Polyvinyl Alcohol (PVA) Hydrogels, notable for excellent mechanical properties and withstanding repeated freeze-thaw cycles while maintaining a low-friction surface even when wet, which makes them an appealing case for use in cold weather conditions (the Q&A card in the top of the stream, fig. 1, ⑬). She writes down these materials as potential leads to pass on to the engineering research team later, and clicks on the ‘Trade-off’ button (fig. 1, ⑭) to learn more about the potential disadvantages of a design that incorporates a lubricant-like material directly on the surface of the rack where bike wheels are loaded on to. The returned trade-offs card raises cleaning difficulty as a potential concern, which she uses to ideate related usage scenarios and constraints involved to develop the idea further.

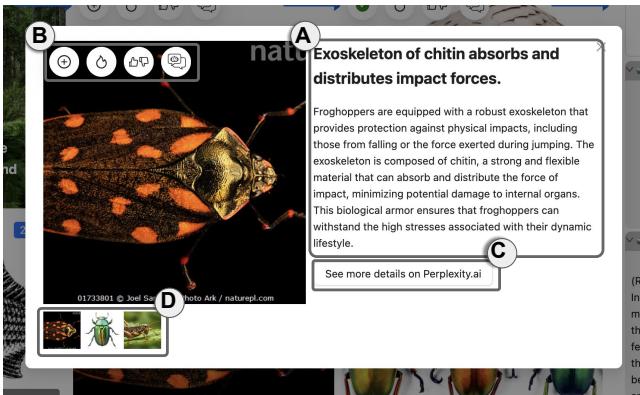


Figure 2: The modal view of a clicked mechanism cluster shows additional mechanism and active ingredient details (Ⓐ). The same action buttons featured on the main page of the interface (Ⓑ) are shown, as well as the ‘See more details on Perplexity.ai’ for finding additional details and related scientific research (Ⓒ), and a carousel displaying other species that belong to the cluster which can be viewed by clicking on any of the images (Ⓓ).

4.2 Leverage familiar interfaces and behavior

Consistent with findings from the design workshop we aimed to design a system that could leverage familiar interface paradigms and existing behaviors in new ways. We based our overall design on the concept of a mood board, similar to Pinterest, where designers normally scroll through and collect many possible inspirations. To this end we added a timeline-style side pane that could act as a central focus for AI suggestions and as a scratch and organization space for the user to externalize and keep track of their thinking and ideation process separate but related to their foraging for inspirations.

To support the side pane’s use as an organization or triage space we also enable quick filtering of different types of AI-provided information cards (see fig. 1, top of the stream ①), as well as deleted items with support for restoration. Users can build off any cards in the stream to explore the design space in a non-linear fashion, with that spark anchoring context for addition exploration.

BIOSPARK was implemented using React.js for the interface and the Flask server in Python3.11 for the backend components.

4.3 Visual representation

To support the design goal of visual representations of inspirations we were inspired by AskNature.org, a popular web repository for bio-inspired design. AskNature pages are designed to include a prominent close-up and centered portrait of a species that creates a striking visual and invokes curiosity.

To achieve this we used Google Search and Adobe Stock Images for retrieving relevant inspiration images. Directly searching on Google using its API with an animal name or mechanism description query often resulted in images such as book covers or graphs in related research paper, which were less effective. Conversely, using Adobe Stock Images¹ with animal species names as queries sometimes led to high quality photos but other times the animal was shown in the distance or background, and coverage was limited. Therefore, for each species we combined the top-5 results from Google Search and Adobe Stock Images to create a set of images, and used GPT-4V (gpt-4-vision-preview) to rank each of the species’ images in terms of the visual focus and the potential value for mechanism understanding (fig. 9; full prompt in Appendix B.1.1 and the score and rationale of each image in Appendix B.1.2). The highest-scoring image was chosen to represent each species, and the results appeared sufficiently accurate for the needs of the prototype system.

4.4 Goal-driven inspiration discovery

Finding biological inspirations is a longstanding research challenge, as discussed in the related work. This challenge is made more difficult when driven by a specific design goal, which might be relevant to only a tiny proportion of possible biological inspirations. Instead of searching for needles in a haystack (e.g., through various means such as crowdsourcing [66, 73], rule-based programs such as [11], data programming [17]) we explore the idea of leveraging recent advances in language model training to use LLMs as a knowledge retriever. However, in comparison to existing approaches such as fine-tuning with a set of biological examples (e.g., from AskNature [74]) with no resulting control over the areas of the design

¹<https://stock.adobe.com>

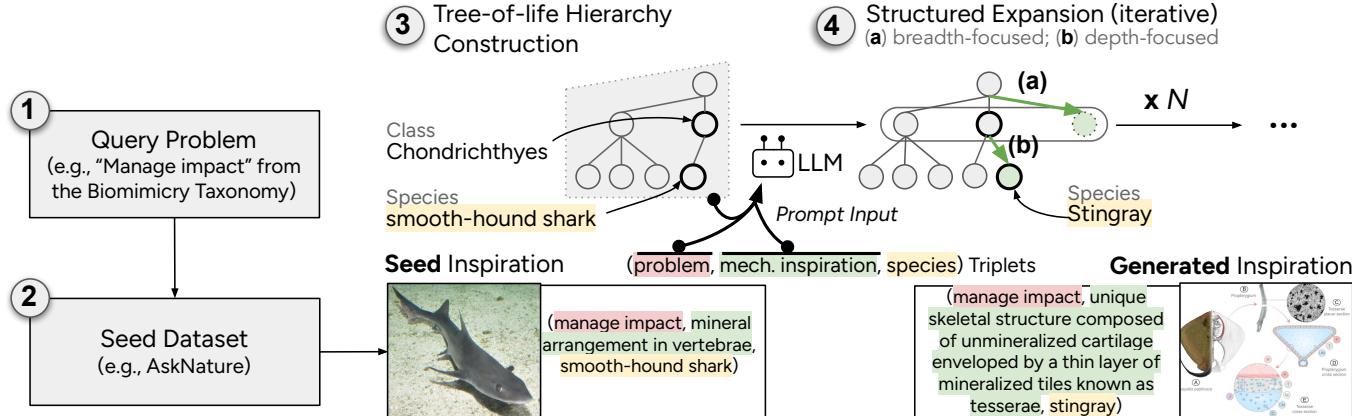


Figure 3: Goal-driven Mechanism Inspiration Generation Pipeline. The pipeline begins from ① a query problem, such as a function ‘Manage Impact’ from the BioMimicry Institute’s Taxonomy which covers a broad range of problems. The problem is used to ② search AskNature.org, which organizes species or innovations according to the functions, to provide seed data. ③ This data is structured into the (problem, mechanism, species) schema and the species names are used to prompt an LLM to construct the tree-of-life hierarchy. ④ The hierarchy is traversed to identify expansion sites at the frontier, here instantiated as sparse branches on the tree with high diversification opportunities, determined in the breadth- or depth-focused manner. The sites along with the problem context and existing mechanisms in the dataset are provided to contextualize the generation and to jointly enforce diversification and relevance filtering. The generation continues iteratively until a stopping condition is met. While the pipeline uses specific data sources in this example, the approach is source-agnostic and adaptable to other contexts (see text).

space explored or the diversity of the resulting inspirations found, we introduce an approach that introduces both structure and diversification.

Our approach conceptually follows the hierarchy-based expansion mechanism from the WordTree method [44] that demonstrated how up-then-down traversal on an abstraction hierarchy in structured brainstorming settings could lead to novel insights. Here, we design a similar approach for structurally expanding a seed dataset (in our case, AskNature biological inspirations) but instead of an arbitrary abstraction hierarchy we use a structured hierarchy latent to most biological inspirations: the Tree-of-Life² model. This approach anchors our expansion algorithm on species of nature as a mediator for exploring new spaces of mechanisms, as species often adapt to changing natural environments by evolving with new mechanisms.

Concretely, starting from a query problem ‘manage impact’, an example seed mechanism found on AskNature.org may be the mineral arrangement in smooth-hound shark vertebrae post³ (fig. 3). We structure this into a triplet of (problem, mechanism inspiration, species): (‘manage impact’, ‘mineral arrangement in smooth-hound shark vertebrae’, smooth-hound shark) (details in Appendix B.2.1). From the seed triplets, we first construct a 7-level tree-of-life hierarchy using GPT4 consisting of the {domain, kingdom, phylum, class, order, family, genus, species} levels by prompting it with the name of the species. In our evaluation using 90 ‘ground-truth’ taxonomies (sourced from Wikipedia’s ‘biota’ scientific classification information boxes), we find satisfactory accuracy levels ranging between 94.4% – 100% for each level on the taxonomy (Table 6, Appendix B.2.3).

We then traverse the constructed tree-of-life hierarchy to identify the directions of expansion that have high potential for diversification. We design two approaches – breadth- and depth-focused expansions – to this end, and estimate the potential for diversification as sparsely populated branches on the hierarchy, as expanding from them is likely to result in new species that are not currently represented in the dataset.

In breadth-focused expansion, we design a prompt (fig. 11 in Appendix B.2.2) that requests GPT4 to generate sibling nodes of a given node, excluding the previously generated nodes to avoid duplicate generation. In depth-focused diversification, we filter nodes at a given level (e.g., ‘order’) and select the five with the fewest children (those having the greatest opportunity for vertical exploration), and design a prompt (fig. 10 in Appendix B.2.2) to prompt GPT4 for children nodes generation.

To maintain the relevance to the query problem, we add the descriptions of the query problem along with specific constraints to the prompt to contextualize the generation and to jointly enforce diversification and relevance filtering. After this process the system generates new mechanism inspirations. Using the seed triplet (‘manage impact’, ‘mineral arrangement in smooth-hound shark vertebrae’, smooth-hound shark) as an example, one example of the newly generated mechanism inspirations may be the unique skeletal structure composed of unmineralized cartilage enveloped by a thin layer of mineralized tiles known as tesserae in stingrays. Both stingrays and gray smooth-hounds are species of the class of jawed fish Chondrichthyes. The newly generated data is incorporated back to the tree-of-life hierarchy updating its frontier, and the algorithm runs iteratively until a stopping condition is met.

Approach Generalizability. While the pipeline uses specific data sources here, the general approach is agnostic to the source of data and can be applied to any sources that can be turned into

²https://en.wikipedia.org/wiki/Tree_of_life

³<https://asknature.org/strategy/minerals-strengthen-vertebrae/>

the problem-mechanism-species schema. Relevance to the query problem during data generation is preserved through contextualized prompts that integrate constraints and problem descriptions, which are parameterized in the current system implementation and is straightforward to adapt to new user query. At run time, active features also maintain relevance through contextualized prompts, which are also currently parameterized in the prompt design for adaptation to user input.

4.5 Recognition

4.5.1 Recursive Clustering of Active Ingredients. One way to help users recognize relevant inspirations is to cluster organisms with similar mechanisms together to improve the efficiency of scanning for diverse inspirations. In order to organize the active ingredients in semantically meaningful groups, we adapted a density-based algorithm DBSCAN to re-cluster those mechanisms it could not initially cluster with a gradually relaxing minimum distance parameter ϵ . This approach had the benefit of fixing the relatively straightforward clusters (*i.e.*, groups of mechanisms with very similar surface text forms, resulting in very low distances among them in the embedding space) early on, and by lowering the sensitivity threshold subsequently in order to identify less obvious yet coherent clusters (*i.e.*, groups of mechanisms that look different in the surface text form yet are semantically related, resulting in relatively higher distances among them). After the final run of the algorithm, mechanisms that could still not be clustered were appended to the end of the list.

4.5.2 Active Ingredient Extraction. Another way to improve recognition brought up by participants in our pilot study is to generate cluster descriptions more informative about their *active ingredients*, or transferable concepts that enable the mechanisms. To do so we used GPT-4 (gpt-4-turbo-preview) with a prompt (Appendix B.3) to extract active ingredients from mechanism outputs from the earlier diversification and goal-driven generation steps (§4.4). In the system message we instruct the following for identifying active ingredients that we found useful from pilot testing: 1) shorter length (*i.e.*, 15 words or less), which was easier to skim and increased the cluster separation by excluding secondary features of commonality among the species, 2) descriptions with a verb or a verb phrase, which was easier to parse as they often presented the information in the form of ‘what acts upon what’, and 3) concrete active ingredient examples. In the user message we provide the mechanism description to process accordingly.

4.6 Transfer

4.6.1 Sparks. As discussed in prior work, it is cognitively demanding to transfer an inspiration to the target domain of a design problem, involving identifying the relevant features to transfer, how those features map to features in the new domain, and adapting any unmapped or missing elements [22, 32]. There is also a question of how and when such a mapping should take place – it is expensive both computationally and from a user-attention perspective to compute and show mappings for every potential inspiration.

Here we introduce the interaction paradigm of ‘sparks’ to support analogical transfer and the computational approach to enabling them. Sparks are generated when a user interacts with an inspiration to save it as potentially interesting, passing an initial threshold

of interest but not necessarily requiring the user to have deeply engaged with mapping that inspiration to how it would work in the target domain, *i.e.*, leveraging a ‘spark’ of interest from the user. Within seconds of indicating interest in an inspiration, the system generates two spark cards at the top of the sidebar stream, attracting the user’s attention. Text on the spark cards is large enough for the user to start reading the first few lines with little effort, helping them to see how the inspiration could be applied to the target domain. The generation of two sparks aims to prevent fixation in the use of the inspiration and present multiple solution paths which could lead to schema induction and the user exploring even more of the design space. Thus, the design of sparks is aimed at scaffolding the mental process of transfer by reducing cognitive effort at each step and directing attention to potentially fruitful areas in an approach aimed at leveraging natural practices and that recognizes the designer’s need for agency.

Two sparks are generated using GPT-4 (gpt-4-turbo-preview) each time the user saves an inspiration or clicks on the ‘spark’ button (fig. 1, ⑩) on a mechanism. Each spark card also includes helpful features such as a caret for expanding/collapsing the card, the timestamp of creation, a clickable thumbnail (fig. 1, ⑪) showing the source mechanism, which expands the modal view upon clicking it, and control buttons (fig. 1, ⑫) for further generating new sparks of the spark content, Q&A, and deletion. The content of each spark is directly editable. The spark-generation prompt (Appendix B.4.1) contextualizes the user-selected mechanism inspiration with the design problem description and the constraints provided with the problem. We instruct GPT-4 to be succinct when generating a spark (*i.e.*, under 500 characters) and to provide a descriptive title.

One challenge when developing sparks was that generating multiple sparks for the same mechanism inspiration led to highly similar sparks, despite explicit instructions included in the prompt that requested diversification in generation. To address this, we adapted feedback-augmentation approaches such as self-refine [46] to improve diversity by adding the most recently generated 20 sparks as part of the prompt, and requesting that the new generation be novel and not redundant with them. Through validation on 3800 pairs of sparks we found that semantic diversity (as measured by cosine similarity, *cf.* [25, 30, 64]) was significantly higher when precedent-based diversification was used ($M=.24$, $SD=.073$) than not ($M=.17$, $SD=.090$) ($t_{ind.}(7291.87)=-42.41$, $p < .0001$). See Appendix B.4.2 for more analysis details.

4.7 Elaboration

4.7.1 Trade-off Analysis. As discussed in both the workshop and the design probe as well as in prior work, the process of analogical transfer in design rarely ends with mapping. Designers often need to elaborate and drill down on the particular mechanisms suggested by the inspirations to understand how they would actually work, their physical limits, and pros and cons of different approaches in the context of the design problem. To address this with a consistent interaction paradigm as the spark cards we introduced ‘trade-off’ cards. The goal of trade-off cards is to not only help users who are looking for specific elaborations already, but to proactively suggest dimensions on which users might want to elaborate and drill down further.

We generate a new trade-off analysis card using GPT-4 (gpt-4-turbo-preview) each time the user clicks on the trade-off button (fig. 1, ⑩) on a mechanism. We design a trade-off analysis prompt (detailed in Appendix B.5) to request the generation. In the prompt we contextualize the user-selected mechanism inspiration using the design problem description and the constraints provided with the problem. We instruct GPT-4 to return the ‘pros’ and ‘cons’ of the mechanism inspiration in the context of the design problem using a markdown table format that places each pro-and-con pair in a new row, and give each item in the table a succinct (3 words or less) label. In the view, we display the analysis in each trade-off card in the stream (fig. 1, ⑪) and implement a scrollable and formatted table view using React-Markdown⁴ and remark-gfm⁵.

4.7.2 Q&A Handling. Many of the elaboration questions brought up in the pilot study by participants, such as “I wish I could know more about the lubrication mechanism of slimes”, are not easily captured in preset queries such as trade-offs. To support more flexible elaboration we introduce Q&A cards, which allow the user to ask any question they would like to the LLM with the context of the design problem and inspiration already included. Q&A cards also serve as a general purpose catch-all to probe what kinds of questions users might want to ask that haven’t yet been explicitly designed for in the system.

To fluidly respond to free-form user questions in the Q&A text area (fig. 1, ⑫), we send GPT-4 a prompt that instructs the model to answer the question and includes information about the source mechanism and the design brief context in its system prompt portion, while the user prompt portion includes the question itself (fig. 17, Appendix B.6). To help users understand and recall the provenance of the generated response, we send another GPT-4 request for a rationale and appropriateness assessment. This rationale is then featured as a tooltip next to the timestamp on the response card (fig. 1, ⑬ icon in the header of each card in the stream ⑪).

4.7.3 Drill-down on related research. Another aspect of drilling down on an inspiration raised in the pilot study was a request for more information about a particular mechanism, specifically web or scientific resources about it. Although not a primary focus of our main study goals, we did include a rudimentary support for this need to probe users’ interest, through linking them to a generative AI search engine that provided cited web sources for a given mechanism (details of its implementation in Appendix B.7).

5 User Study

To investigate how BioSPARK affected bio-inspired design generation we conducted a within-subjects study in which participants experienced both BioSPARK and a baseline condition for two design problems. Our goal in this study was to evaluate the combined value of the design decisions made in BioSPARK on creativity-related outcomes including the quality, amount, and diversity of ideas generated, while using an analysis of participant think-aloud and post-study interview and survey data to explore participants’ experiences with different features and functionality. Towards this goal we explored appropriate baseline conditions for comparison

that would provide users with a similar set of resources while corresponding to what designers in our workshop and pilot study might access for bio-inspired design today. The final baseline condition selected provided participants with access to AskNature.org, a popular bio-inspired design inspiration site which served as the seeds to BioSPARK’s generation algorithm, and with a ChatGPT terminal. We leave for future work more nuanced conditions such as ablation studies of various features or augmentation of the ChatGPT terminal (such as through persistent context, custom GPTs, or fine-tuning).

5.1 Methodology

We employed a within-subjects study design to compare BioSPARK with a baseline system for inspiration and a shared Google Spreadsheet participants accessed to write down their own ideas. We chose two design problems for user ideation, including how to design wheelchairs that allow users to go up the stairs easily and how to design an innovative bike rack for sedans. These problems were chosen because they involve multiple, potentially competing constraints (e.g., lightweight but durable) and were pilot tested for being able to be completed within the timed ideation task.

(The ‘Wheelchair’ problem) Design wheelchairs that can also
 ↳ allow users to go up the stairs easily.
Constraint 1 (Lightweight yet Durable Construction): The
 ↳ wheelchair should be lightweight and be able to
 ↳ withstand a heavy load without structural failure.
Constraint 2 (Compact and Foldable Design): The wheelchair
 ↳ must be foldable to a 1/4 of the volume within 30
 ↳ seconds without the use of tools.

(The ‘Bike rack’ problem) Design innovative bike racks for
 ↳ sedans.
Constraint 1 (Integration without Interfering with Aerodynamics): The bike rack’s design must not
 ↳ significantly reduce the vehicle’s fuel efficiency when
 ↳ installed and with bikes mounted.
Constraint 2 (Versatility in Accommodating Different Bike Types): The rack must be able to securely hold at least
 ↳ three different bike frame sizes (e.g., 16”, 20”, and
 ↳ 26”) without the need for additional adapters.

We randomly assigned problems to conditions for the main timed tasks (20 minutes each), counterbalancing the order of presentation using three 2x2 Latin Square blocks. Participants followed a fixed procedure in the study, which took place remotely using Zoom: Introduction, Consent, Demographics survey; Tutorial (detailed in Appendix C.1) of the first system via screensharing; Main task for the first system (20 min); Rating task for the first system (only in the BioSPARK condition); Survey for the first system; alternating and repeating for the second system; followed by a debrief. Participants were asked to share their screen during the timed tasks and think-aloud.

To probe how participants felt about the utility of different information generated using various AI-based system features, after BioSPARK’s main task, participants were also presented with a

⁴<https://github.com/remarkjs/react-markdown>

⁵<https://github.com/remarkjs/remark-gfm>

rating interface that showed a list of saved Sparks, Q&A, and Trade-off cards along with a 5-point Likert-scale for them to rate their usefulness in their process.

To analyze perceptions of usefulness we collected participants' subjective ratings to a modified Technology Acceptance Model survey questionnaire items focused around task performance and easiness of learning from [71] (4 items) using a 7-point Likert scale (1: *strongly disagree*, 7: *strongly agree*). In addition, we employed questions focused on serendipity and exploration adapted from [48, 53] (9 items) and the questions on the value of AI assistance and the quality of inspirations found in the system.

5.1.1 Participants. We recruited 12 researchers (7 women, 5 men) through advertisement on Upwork⁶ and email lists at a State Arts College. Participants' background included professional UX design experience (6), professional illustration and graphic design (1), PhD in Psychology (3), and a current undergraduate student in Arts and Design (1) and a master's student in AI and Data Science (1). Participants' average age was 36.1 years ($SD=9.91$).

5.1.2 Baseline. Participants were given 5 URLs from AskNature, each pointing to a functional category equivalent to those that were used for the BIOSPARK backend dataset pipeline: Manage Impact, Manage Tension(Manage Tension), Manage Compression, Manage Turbulence, and Modify Speed⁷. Before the baseline task began, participants organized their screen by opening up all 5 tabs in their browser on the left-hand side of the screen and sign-in and open the ChatGPT⁸ interface on the right-hand side of the screen. They were instructed to freely use the platforms to help themselves understand and ideate with mechanism inspirations found on AskNature for the design problems. Each participant was also instructed to write down the ideas they come up with in the process in a prepared Google spreadsheet, with a brief description of the species that inspired each idea.

6 Results

6.1 Creative idea quality and quantity

We begin discussion of the results with one of the most critical questions: did BIOSPARK result in higher quality creative ideas? Although creative quality can be defined and operationalized in many ways, one common approach in the creativity literature is to break down an idea's creative quality into the dimensions of novelty, value, and feasibility [61, 72]. By combining these three dimensions, often using a penalizing function for being low on any dimension such as the geometric mean, researchers have attempted to balance trade-offs between these factors into a holistic judgment of creative quality.

Following this approach we created a rubric operationalizing novelty as how unique versus common an idea is found to be in the world. Value corresponded to how effectively the idea addresses the main challenge in the problem (e.g., how well does the idea directly help wheelchair users go up the stairs more easily?) as well as the main constraints (e.g., how lightweight or foldable is it?). Feasibility

⁶<https://www.upwork.com/>

⁷(Manage Impact) <https://rb.gy/rvz17u>; (Manage Tension) <https://rb.gy/t3se2z>; (Manage Compression) <https://rb.gy/xvogjb>; (Manage Turbulence) <https://rb.gy/9apgoq>; (Modify Speed) <https://rb.gy/r7o2c8>

⁸<https://chat.openai.com/>

was operationalized as the ease of achieving the creation of the idea with current resources and technology. These three dimensions were combined using the *geometric mean* into an aggregated score representing the overall creative quality of an idea.

To judge quality we recruited as a domain expert a senior PhD student at an R1 institute in North America with expertise in building systems augmenting wheelchair users' mobility. The expert recruited had practical experience in designing products employed by wheelchair users, including broad knowledge of the feasibility of both physical constraints such as the materials used and their configuration and reliability, as well as user constraints such as what types of products would be likely to be used by wheelchair users. The first author and the expert met to discuss the assessment rubric using 5 randomly selected ideas for each of the two design problems used in the study (see Appendix C.2). Upon reaching agreement on the rubric, the first author and the expert independently judged the quality of remaining ideas blind to condition. Intraclass Correlation Coefficients showed a significant level of reliability between the judges on the independently coded set of examples: 0.94 (Novelty), 0.74 (Feasibility), 0.90 (Value), following [72] in which the more conservative ICC(2,k) method is used. To further validate the reliability of the feasibility scores, which could be especially sensitive to domain expertise, we also involved an experienced product design expert (one of the paper's authors) who independently assessed each idea blind to condition and provided detailed rationales for the same 10 examples the two judges used for discussing the rubric. This expert had an undergraduate degree in engineering design and a Master's and Ph.D. in Mechanical Engineering specializing in engineering design, with over 16 years of experience in developing mechanical, mechatronic, and software-based systems. An analysis of their reliability scores showed a strong and significant correlation with the judges' scores (Pearson's correlation coefficient = 0.85, $p = .002$).

The final scores were computed by averaging the two judges' scores on each dimension and combining them using the geometric mean into a single overall score, as per common procedure in the literature (e.g., [61, 72]). The summary statistics of scores in each dimension are as follows:

Condition	Novelty	Feasibility	Value
Baseline	3.9 (1.67)	6.8 (1.37)	3.8 (1.68)
BIOSPARK	7.1 (1.03)	5.7 (1.02)	7.0 (0.91)

Table 1: Means with standard deviations in parentheses for the three dimensions (Novelty, Feasibility, Value) across the two conditions (Baseline, BIOSPARK).

We analyzed overall scores using a linear mixed-effects model to take into account potential participant-specific effects of our within-study design. An initial model compared the control and experimental conditions (i.e., Baseline vs. BioSpark) with participant as a random effect, using the Residual Maximum Likelihood (REML) method for fitting model parameters [52]. We found that while the residuals were normally distributed (through a visual examination of the Q-Q plot of residuals showing good alignment to the line of estimation and the complementing result of the Shapiro-Wilk test showing residuals not significantly deviating from normal distribution, $p = .06$), the homoscedasticity assumption was likely violated

from the results of both Breusch-Pagan and White's tests ($p < .001$ in both cases). We thus modified the model to use a covariance estimator adjusting for variance heteroscedasticity and residual correlation, and find that this final model showed a good fit to the data ($F = 802.87, p < .001$) overall, with an $R^2 = .58$ indicating that approximately 58% of the total variability in scores was explained by the model.

	Coef.	Std. Err.	t	$P > t $	[0.025	0.975]
Const.	4.4802	0.102	44.105	0.000	4.280	4.681
Cond.	2.0396	0.126	16.221	0.000	1.792	2.288

Figure 4: Regression results show that the estimated overall quality for the baseline condition is 4.48 and $4.48 + 2.04 = 6.52$ for the BioSPARK condition. The effect of condition was significant, $t = 16.22, p < .0001$, suggesting a substantial difference between the baseline and the BioSpark condition.

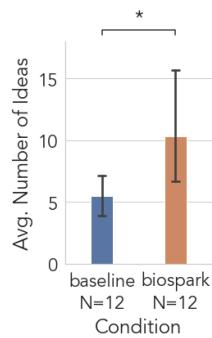


Figure 5: The number of participants' ideas during the experiment.

Results showed that the overall creative quality estimate was significantly higher in BioSPARK ($M=6.5$, Standard Error=.16, $t = 66.62, p < .001$) than in the Baseline condition ($M=4.5$, Standard Error=.10, $t = 28.34, p < .001$) (Table 4). Representative ideas from each condition are shown in Table 9 to illustrate how high vs. low quality ideas differed for each design problem. High quality ideas included several examples of analogical transfer, as shown in an idea inspired from tree frogs' adhesive toe pads to harness the rear surface of the vehicle itself for securing bikes (Top 5th idea in BioSPARK) or an idea inspired from sea anemones' burrowing to construct a contracting and expanding wheelchair base (Top 50th idea in BioSPARK).

Interestingly, there was no evidence of a trade-off of idea quality with idea quantity, with BioSPARK users generating nearly twice the number of ideas ($M=10.3, SD=8.46$) than in the baseline condition ($M=5.5, SD=2.91$), which a paired two-tailed t-test suggests was a significant difference ($t_{\text{paired}}(13.56)=-2.35, p=.04$) (fig. 5). Generating more ideas, particularly diverse ones, is itself a design goal for many designers; thus, the significant increase in the number of ideas in combination with the large increases in the novelty and value dimensions suggests a promising direction.

In summary, from several converging angles our results suggest that BioSPARK led to participants generating more and higher quality ideas than in the baseline condition. Participants' own perceptions of the value of the system appeared consistent with this,

with participants' agreement with the statement '*Using this system would improve my task performance*' was significantly higher in the BioSPARK condition ($M=6.5, SD=.80$,) than in the baseline condition ($M=5.3, SD=1.07$,) (Wilcoxon $W=2.0, p=.02$).

6.2 Diversity

Another of our design goals was to not only help users engage with inspirations, but for those inspirations to be diverse and help them explore new design spaces they might not have thought of on their own. The previous analysis on quality is consistent with the hypothesis that participants' increased novelty in idea generation may have been based on encountering more diverse inspirations; in this section we investigate the role of diversity more directly.

In the context of bio-inspired design, one measure of diversity in design space exploration is how many different species in nature participants are engaging with for ideation, which has potential for not only inspiring ideas based on a specific mechanism of the particular species but also opening up a new space of design that encompasses other mechanisms of the species or its related species. To quantify the number of unique species mentioned by a participant we extracted the inspiring species' names using gpt-4-turbo-preview with a prompt (Appendix C.5). A paired two-tailed t-test revealed that the number of unique species used for inspiration was significantly higher (nearly double) in the BioSPARK condition ($M=8.2, SD=4.97$) than the baseline condition ($M=4.6, SD=2.71$) ($t_{\text{paired}}(17.02)=-3.30, p=.007$). Although it is difficult to directly quantify the effects of using more species, a plausible hypothesis is that using more species provided a greater number of mechanisms for participants to leverage in solving the design problems, resulting in both more ideas as well as those ideas being higher value and more novel as described in the section on quality.

The interview, survey, and observation data support a deeper but consistent understanding of which aspects of the BioSPARK design and interactive features most contributed to broadening participants' exploration of design spaces. Participants felt that the AI features in BioSPARK helped them be more creative. As will be described in the Feature Use (§6.3.3) section below, participants viewed the 'spark' generation feature as usefully nudging them to reframe the problem (P6), creatively adapt and translate the source mechanism inspirations into the target domain (P12), beyond just providing useful inspirations or ideas (P1, P7). Consistent with these statements, participants' agreement with the statement '*I was able to examine a variety of inspirations*' was significantly higher in the BioSPARK condition ($M=6.6, SD=.67$,) than the baseline condition ($M=5.4, SD=1.62$,) (Wilcoxon $W=0.0, p=.03$).

In contrast, the baseline condition felt less dynamic and led to less exploration by participants. For example, while P4 thought the functional category-based organization in AskNature helped his navigation, it also somewhat fixed the broader design space he explored in, as he felt like "*got stuck somewhere a little bit because I came up with this sideways top-mount bike rack idea early on from the 'manage turbulence' concept*" (P4). He also thought it was easier to see the relevance of mechanisms in BioSPARK and follow-up with more exploration. However, it is possible that such

Rank	Cond.	Prob.	Score	An-Tr?	Idea (Summarized)
Top	BIOSPARK	2	7.86	Y	This bike rack uses a bio-mimetic spring mechanism , inspired from Anura's powerful legs , to stretch (and compress) to accommodate different bike sizes. Its skeletal structure, mimicking frog bones, flexes to absorb road vibrations, protecting bikes. Aerodynamically shaped to reduce drag, it ‘leaps’ into a compact form when not in use, preserving fuel efficiency.
Bottom	BIOSPARK	1	3.98	Y	Retractable ‘legs’; Jointed segments that extend and contract to climb.
Top	Baseline	2	6.80	Y	Aerodynamically shaped bikerack design inspired by how marine mammals like dolphins and whales have evolved flippers and tail fins that optimize their movement in water. Their body shape tends to be more rounded than that of fish but is streamlined for efficient travel. The tail fins (flukes) provide powerful propulsion, while the pectoral flippers are used for steering and stabilization.
Bottom	Baseline	1	2.62	N	Golden bamboo for durability and lightweight design.

Table 2: Top-1 and Bottom-1 scoring ideas from each condition, and whether they perform analogical transfer. ‘Cond’ represents the condition in which the idea was produced; ‘Prob’ represents the problem the idea is for (1: the ‘wheelchair’ problem; 2: the ‘bike rack’ problem); ‘Score’ represents the geometric mean of expert-judged novelty, value, and feasibility scores; ‘An-Tr?’ is a binary value representing the presence of Analogical Transfer; The Bottom-1 idea is summarized for conciseness. **Boldfaced text in the idea description represents the source and the target in analogical transfer.** Full table in Table 9 (Appendix C.3).

structured organization could be profitably combined with a more dynamic system, as described by P9 that he is “not as interested as working from the “bottom up” for research, and would instead like to have AI help brainstorm from the “top down”” (such as through the functional organization in AskNature).

6.3 Feature use

The above analyses suggest that the design choices made in BIOSPARK led to higher quality ideas, more ideas, and consideration of a more diverse inspiration space. In this section we aim to investigate which features were used by participants and their perceptions in order to better understand the results above and to unpack which design choices were more successful than others. In the below sections we characterize through think-aloud, usage, and survey data participants’ experience with various system features.

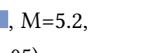
6.3.1 The Stream Interface. The stream interface design of BIOSPARK was aimed at providing a scratchpad and organization space for thought, aiming to help users triage and interact with AI support. The stream was perceived as helpful for participants during ideation and it was the highest-scoring system design feature in BIOSPARK ($M=6.3$, $SD=.98$; Table 3). On its value, P1 said:

“I liked these because they kept my thoughts and all the information very organized. It allowed me to focus on the actual text vs focusing on the organization of everything. It would have been even more helpful though if there was a way to enable bullet points or formatting tools within these. I would have used bold or italics for example.”

P4 noted that “It’s great to see all the ideas in one place. It provided a nice anchor to go back to. It’s also a nice touch to have a trash bin so that I can go back and check the ideas I discarded earlier.” and P5 said the design was helpful for “note-taking and brainstorming through relevant results and queries”.

Participants also commented on how the design and presentation of information in BIOSPARK streamlined their exploration and helped them accomplish the task. P10 described it as: “I like that it’s

integrated into one space. I can press a button to get to the particular need that I had.”, and P7 mentioned that “[the stream organization of information] helped me to compare and contrast” while P10 commented “[the stream was] definitely helpful. [It] allowed me to narrow down best options.”

These comments are consistent with participants’ agreement with the statement ‘I could easily explore many inspirations without getting lost’ ($M=6.3$, $SD=.89$ in BIOSPARK, , $M=5.2$, $SD=1.80$ in baseline, , Wilcoxon $W=2.0$, $p=.05$).

6.3.2 Mechanism Clusters. Next, we investigate how active ingredient clustering, designed to help improve participants’ recognition and efficiency of scanning diverse mechanism inspirations might have helped. Several participants mentioned clustering as helpful for focus, navigation, and comparison, as it “*helped to categorize the needs/constraint which I am focusing on & pick the one which align well with my end goal.*” (P6); “*navigating the problems and finding solutions to associated animals and their unique features*” (P8), and “*giving me information to help drive a decision and look at alternative options.*” (P5).

The clustering structure appeared to be a functional entry point for exploration even for those who did not paid a lot of attention to: “*I found myself not exactly focusing on the “clustering” per say, but I did focus on the specific mechanisms that seemed could be a good way to enhance design features.*” (P1). However, participants appeared to use the drill-down functionality into clusters less, either not attending to the species that would show up when drilling down into a particular cluster or even being unaware that the drill-down existed: “*I didn’t look into the other species within the clustering so much as the responses from the AI*” (P11); “*didn’t use it, was busy using other tools*” (P2); “*I actually didn’t notice the clustering*” (P3).

6.3.3 Sparks. The ‘Spark’ button was the highest-used feature among the active features, at 6.2 times ($SD=3.90$), and was generally perceived as useful ($M=5.9$, $SD=1.08$). P1 described its value as:

Feature Name	Usage Freq. (M, SD)	Usefulness (M ↓, SD)
Stream	N/A	6.3 (SD=0.98) 
Spark	6.2 (SD=3.90)	5.9 (SD=1.08) 
Q&A	2.3 (SD=1.97)	5.8 (SD=1.19) 
Trade-off	1.3 (SD=1.66)	5.2 (SD=0.94) 
Clustering	N/A	4.1 (SD=1.44) 

Table 3: Usage and usefulness ratings for BioSPARK active features, measured on a 7-point Likert scale (higher is better).

"The "Sparks" button provided me with a lot of interesting insights and got me thinking in directions I may not have thought of on my own. It spurred my creative thought."

Other participants mentioned that sparks “provided more inspiration/ideas” (P7), “helped me with better viewpoint and perception which I normally wouldn’t think of” (P6), “makes me think of a new design space” (P12), and “gave me inspirations on how I could translate the idea into design criteria and functions I want to achieve” (P12). The ‘Spark’ button was also often used in combination with the Q&A feature, which provided users opportunities to follow-up on initial sparks to more deeply understand and further develop them (see Table 4 for different query intent users submitted in the ‘Q&A’ text box).

Reflecting these observations, We find a marginally significant difference in the agreement levels with the statement: ‘The system enabled me to make connections between different inspirations’ showed a marginal difference between BioSPARK ($M=6.5$, $SD=.80$, ) and Baseline ($M=5.6$, $SD=1.56$, , (Wilcoxon $W=1.5$, $p=.07$).

6.3.4 Trade-offs. Participants found the trade-off cards generally useful ($M=5.2$, $SD=.94$) and mentioned how it helped them to prioritize and triage ideas for further consideration. As P6 said: “It very much helped me to decide which feature/constraint I am willing to give up or trade off. It showed me valid reasons to pick the side correctly.” (P6). In this vein P10 also mentioned “It was good to see the pro and cons of a concept to see which were stronger to follow up with.” (P10).

Interestingly, several participants mentioned they noted the cons in the pros/cons analysis as more important than the pros list. P11 and P12 described the reason as: “I tend to look at cons first and it directs and warns me what are the constraints I’ll be working with” (P12); “This Pros and cons is helpful, to look at the weight concerns. When I look at pros and cons usually I go straight to the cons, the reason, is personally I focus on bad things first.” (P11).

However, there were also limitations of the trade-off analysis, namely how the dimensions of comparison was not readily configurable by the user, which sometimes led to useful aspects of comparison but other times not. On this point, P1 described: “I used it for more so looking for cons of the mechanism that could act as barriers for the design itself, but the button provided me insights that were more so focused on the cons being related to business expenditure and manufacturing.” (P1). P4 mentioned having to save ideas that use different materials and click the trade-off button on each even when he wanted a direct comparison on the material and manufacturing cost aspects across multiple ideas: “interesting, but it would be more helpful if I could ask about the pros and cons of a material or mechanism” (P4). Similarly, P5 said that “I want it to be able to

compare this spark and this other spark in terms of strengths and weaknesses” (P5).

6.3.5 Q&A cards. The Q&A cards introduced in BioSPARK aimed to help flexible user interaction with interesting mechanism inspirations, such as elaborating on relevant aspects of the design problem or gaining a deeper understanding of mechanism inspirations they want to build on. Participants thought the ‘Q&A’ feature generally useful ($M=5.8$, $SD=1.19$) and mentioned that it “helped clarify the details of the general overview of the idea” (P3), and that “I could ask some very specific questions about very specific mechanisms such as finding a fabricated material that is comparable to chitin, and get a useful reply” (P4). P6 also commented how the AI assistance “allowed for interactive sessions to implement my ideas along” (P6).

The intent of user queries within the Q&A interaction seemed to have varied, characterized as the following four types (Table 4): 1) (**Understanding**) to more deeply understand design ideas’ working and relevant physics, 2) (**Adaptation**) to creatively adapt the ideas into new directions, 3) (**Constraints**) to probe how they would meet certain constraints to become value, and 4) (**Material**) to engage in material selection and engineering feasibility. In our observations, participants seemed to often engage in an initial deeper understanding of an interesting design inspiration (Understanding), followed by adapting them further by generating adaptation ideas (Adaptation), and increasing their value by considering how various design constraints would be addressed (Constraints) and how engineering and manufacturing feasibility might be enhanced (Material). Table 4 shows representative user queries of each type.

6.4 Engagement beyond Inspiration

One of the main goals of BioSPARK was to support users beyond finding inspirations, to helping users notice the relevance of, transfer, consider tradeoffs, and elaborate on inspiration mechanisms in the domain of the target design problem. Engagement is particularly relevant to analogical inspirations that may require significant cognitive effort for valuable transfer to occur [23, 27]. We explore this research question using the interaction log data as well as transcribed participants’ think aloud data.

6.4.1 Interaction Log Data Analysis Results. Participants in the BioSPARK condition demonstrated active engagement with the inspiration mechanisms. On average, they generated 18.8 sparks ($SD = 14.28$), wrote 1.0 ideas from scratch ($SD = 1.15$), and deleted 9.4 sparks (47%; $SD = 8.74$) after evaluating their relevance or lack thereof.

Furthermore, as detailed in §6.3, participants found the Q&A feature useful, engaging with it an average of 2.3 times ($SD=1.97$) to adapt inspiration mechanisms and enhance their understanding. Their range of queries and intent is summarized in Table 4. These

Intent	Sample Q&A Query
Understanding	"How would this be able to collapse the frame itself, not just for the bikes while on the rack?" "Tell me more about how this design ensures minimal air resistance, maintaining fuel efficiency."
Adaptation	"How to modify this in a mobile space?" "is there a way to install modular adapters?"
Material	"Material wise what are some of the examples that could withstand a similar force but in a large scale?" "What kind of materials would I need to make this hydrophobic?"
Constraints	"How does a user without functioning limbs use this wheelchair?" "Can this bike rack safely restrain e-bikes?"

Table 4: Intents and Corresponding User Queries

findings collectively suggest that participants actively engaged with BIOSPARK-generated sparks, leveraging them to notice relevance, transfer ideas, and elaborate on inspiration mechanisms.

6.4.2 Think-aloud data analysis results. To explore the research question further we analyzed a subset of the transcribed participants' think-aloud along with their behavior descriptions. The research team met to discuss coding of interview and think-aloud data from the study, with one salient feature of the data being that participants seemed to engage with mechanism inspirations differently in depth, for example with or without follow-up actions that related to attempting to deepen their understanding of the inspirations, of their relevance to the design problem and of trade-offs regarding different design constraints, and attempting to come up with new ideas that could adapt the inspirations to a design problem in new ways.

In order to characterize these differences the first two authors developed four codes for participants' different engagement patterns corresponding to two common types of shallow engagement and two common types of deeper engagement:

[S1: "Interesting!"]: Positive comments on a mechanism inspiration, but directly followed by moving on to a different mechanism that was visible to the participant.

[S2: "I'm not sure how this might be relevant"]: Negative comments on a mechanism inspiration, but similarly followed by moving on to a different mechanism that was visible to the participant.

[D1: Engaging with relevance understanding and constraints consideration]: Engagement with AI to understand a mechanism inspiration's relevance to the design problem, for example by asking the following types of questions "*tell me examples of...*" or "*how might this be used/applied...*"

[D2: Actively coming up with new ideas]: Exploring the design space and actively generating new ideas "*it made me think of...*"

One author transcribed the interview and think-aloud data from the study, and incorporated descriptions of participants' actions with each platform (e.g., what the participant is typing in the ChatGPT interface or what the participant is clicking in BIOSPARK), that participants' think-aloud did not describe but were relevant to understanding their engagement process and intent. This resulted in 266 transcripts across 12 participants. Coders coded a set of randomly selected 16 transcripts together blind to condition and

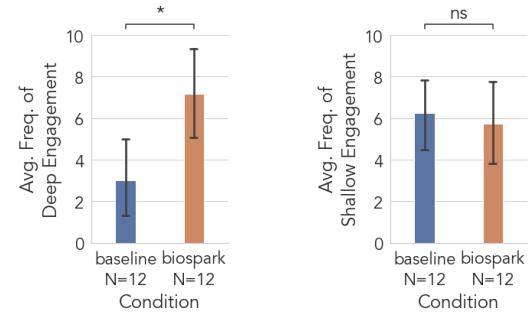


Figure 6: (Left) The bar graph shows that the average number of deep engagement was significantly higher in BIOSPARK than in Baseline.; (Right) The bar graph shows that there were equally many shallow engagement types in both conditions.

built consensus through discussion. They then coded 30 additional randomly selected transcripts independently. The inter-rater agreement of codes for this set showed a moderate to strong level of agreement $\kappa = 0.76$. Thus, the first author coded the remaining 218 transcripts alone. Table 5 shows representative cases and how they map to the four codes of engagement patterns.

To determine whether there was a difference between the frequencies of the two types of codes and conditions we conducted a χ^2 test, finding a significant distributional difference ($\chi^2(1)=12.93$, $p=.0003$). Follow up pairwise comparisons after Bonferroni correction showed a significantly higher frequency of deep engagement in BIOSPARK ($M=7.2$, $SD=4.04$) over the baseline condition ($M=3.0$, $SD=3.30$) ($t_{paired}(21.16)=-3.12$, $p=.01$, fig. 6, left). These results suggest that BIOSPARK promoted greater follow-up and idea generation than the baseline; we explore the usage of features that may have contributed to this difference further below.

No significant difference was found in the frequency of shallow engagement between conditions (Baseline: $M=6.3$, $SD=3.28$; BIOSPARK: $M=5.8$, $SD=3.60$, $t_{paired}(21.81)=.34$, $p=.74$, fig. 6, right), suggesting that participants in both conditions encountered inspirations that they did not prioritize for follow up. However, we found that positive utterances for shallow engagement (S1, e.g., "Interesting!") were significantly higher compared to negative utterances (S2, e.g., "I don't see the relevance.") in the BIOSPARK condition ($M=3.1$, $SD=3.92$) compared to the baseline condition ($M=-.3$, $SD=4.16$) ($t_{paired}(21.92)=-2.25$, $p=.046$). Thus it is possible that even though they did not follow up on them, participants may have encountered even more inspirations of possible interest in BIOSPARK, making the findings of engagement above conservative estimates.

PID	Active Ingredient Inspiration	Species	Code	Participants' Think-aloud & Related Behavior Descriptions
P1	Shape allows air to flow over and around like a sail	Caryophyllales (seed)	S1	<i>"Okay. WindSail carrier. Oh that's pretty cool! Inspired by the aerodynamically shaped seeds of Caryophyllales, this bike rack utilizes a lightweight, sail-like structure that harnesses airflow to reduce drag..."</i>
			D1	<i>"The sails are adjustable to snugly fit bikes, mimicking the efficiency of seed dispersal by wind... What does it mean by a sail?" (Asks BioSPARK a question to explain how 'sail' would work in the idea) "That's a lot of information... Aerodynamic Shape. The sail-like structure of the WindSail Carrier is not just for aesthetic appeal; it serves a functional purpose by mimicking the shape of aerodynamically efficient seeds. The shape allows air to flow over and around the bike rack... Oh okay now we're getting something."</i>
P4	Appendages retract into an empty space	Turtle	D2	<i>"Okay, so, the tortoise shell made me think about how things can be folded into empty space. That was the thing I got from the tortoise. There's empty space inside the shell and it can fold like it can take its feet into the shell. But it doesn't break the feet into simple pieces, or fold it like, roll it like, or anything like that. Just takes the feed into empty shell. So that's what I came up with, and then I started thinking about like, oh what does it mean to have a slot inside, and then I thought, airplane wheels, and the Alaska airline door incident, which made me think about the pin-release mechanism (that was supposed to hold the door)."</i>
P6	Small, powerful thrusts for quick, upward propulsion and directional changes	Lepidoptera (butterflies and moths)	S1	<i>"Okay that's pretty interesting, the propulsion (mechanism) and changing directions... that could be relevant to changing directions on wide stairs."</i>
	N/A	N/A	S2	<i>"okay so since I don't know these concepts from nature, I need help in understanding whether I can use that technology in this? So it's hard to understand the relevance."</i>
P11	Sliding and collapsing (like in a telescope)	Armadillo	D2	<i>"Okay this shell that can collapse is an interesting mechanism. Like this makes me think of a telescope, like a telescoping mechanism for sliding and collapsing... so that could be a really interesting design space."</i>

Table 5: Transcripts of participants' think aloud and behavioral records in the BioSPARK condition ('Participants' Think-aloud & Related Behavior Descriptions') and how they were coded into different types of engagement patterns ('Code'). The bold-faced text in each row highlights the important signatures of the assigned code. Each row also contains the following: the participant ID ('PID'), the active ingredient description that participants found interesting / relevant ('Active Ingredient Inspiration'), and the associated species ('Species'). Exhibits in the baseline condition were similar, with the exception of tools participants interacted with.

6.4.3 Design constraint mentions. We also explored how often participants mentioned design constraints mentioned in the brief, which are requirements to be considered for the idea to be useful (e.g., being lightweight but durable, or versatile in accommodating different shapes).

Although we do not have direct measures for how well these constraints were addressed beyond the quality ratings discussed above, an indirect measure we explored was the number of times design constraints were mentioned in participants' ideas. We performed a two-tailed, independent samples t-test over ideas grouped by condition, finding that participants' ideas were significantly longer in the BioSPARK condition ($M=375.5$, $SD=96.15$) over the baseline condition ($M=141.9$, $SD=108.93$, $t_{ind}(119.25)=-14.65$, $p<.0001$). However, length alone may not represent how many different design constraints participants mention. To examine constraint-specific content we extracted unique chunks from each idea description that corresponded to each design constraint using the gpt-4-turb o-preview model, which showed satisfactory performance (details

in Appendix C.4). From this we find that BioSPARK users mention a significantly higher number of design constraints in their ideas ($M=2.7$, $SD=1.01$) than the baseline condition ($M=1.6$, $SD=.78$; $t_{ind}(163.27)=-8.45$, $p<.0001$), suggesting they may have considered design constraints more in their process.

6.5 Challenges

Controllability around what problem constraints are emphasized, and by how much. In some occasions participants felt the AI's response was re-coursing to make connections to the original problem constraints, which seemed forced and adding less value "*it's still trying to navigate the conversation towards the original topic and doesn't seem to be "progressive" enough to talk more in depth with specific areas*" (P12). Similarly, P10 also said: "*I think the AI is trying too hard here to make the connections. I'm more interested in the function of a climbing wheelchair, and want to focus on that first rather than the weight and foldability which are secondary challenges to me*". However, sometimes this might potentially have a beneficial problem-reframing effect and serve as an entry into a new design

space to explore: “*I don’t think these really addressed my questions exactly, but maybe [adapters] is what I’m looking for*” (P5).

Supporting deeper idea development. Participants also wished to have more support around developing ideas further with relevant technical specifications. P10 said: “*I’m interested in going deeper in the technical specs. I’ve read the brainstormed idea description, and now I want to build more confidence in it.*” (P10). P5 wanted a way to easily “*extract interesting parts from an earlier information card and integrate them directly into the next sparks*”, while also being able to “*ask additional follow-up questions based on the previous Q&A responses, without overcrowding the stream space*” (P5).

Some participants wanted additional support for “*comparing and contrasting*” (P7) the ideas, to “*easily highlight the strengths of each design and extract useful design features from it for integration into a new idea for mitigating anticipated challenges*” (P5), beyond the current single-idea-focused trade-offs analysis. Participants wished an easier way to organize the pros-and-cons table for multiple material candidates and comparison: “*It would be great to be able to specifically consider different variations / combinations of different materials, titanium for the base and fiberglass for the body or cover*” (P12), and have a way to perform an estimated quality analysis by swapping some features of an idea with those from other ideas or candidate materials: “*How can I apply this to the cover portion of the bike-rack and what would happen if I replaced this with this other material?*” (P6).

Idea organization and information space efficiency. With many sparks and information cards generated, participants wanted a better mechanism for crowd and space management. Participants engaged in grooming the space as desired (e.g., “*I’ll eliminate some of the ideas so that I don’t get overwhelmed*” – P10; “*I don’t have a lot of space over here, a bit frustrating to me*” – P11), and in making the ideas being actively pursued more easily accessible by prioritizing and ranking ideas: “*Is there a way to bookmark some things (beyond the save-the-mechanism feature)? I think when you’re researching multiple ideas, it would be useful to easily narrow down to a few interesting things for further consideration*” (P5). Participants also wished to have a way to perform this prioritization via “*comparisons of this spark and this other spark in terms of strengths and weaknesses*” (P7).

Other challenges. Some participants noted the potential value of being able to switch between retrieval- vs. generation-based modes of getting the information cards. For example, P11 commented that “*(The baseline) is more intuitive for me because I can always switch between different tabs and ask questions to AI if I have to, or instead Google something, and not always invoke AI automatically*” (P11). Related to this, P10 commented that he wanted to “*understand what the source is when new ideas are generated*” (P10), suggesting further exploration is needed to understand how knowing which information sources contributed to the card content affects the perception of trust, which may become more important for further idea development and could be useful for related customization.

Moreover, participants commented on wanting to instruct the output structure as desired related to more or less details or conciseness, or more or less structure for skimming. P10 said, “*I also would like bullet points as opposed to long text or paragraphs for scannable and easier comparisons*” (P10). Similarly P11 commented that he would “*like (the cards) to be as concise as possible... surfacing*

the design objectives over complete sentences, along specific criteria or ingredients – these are the features and important stuff” (P11).

Together, these challenges point to limitations of the interaction paradigm proposed in BioSPARK, and fruitful areas for future research around personalizing and improving the controllability in LLM-augmented analogical transfer.

7 Discussion

We introduced BioSPARK, a system exploring the idea of acting as a creativity partner in analogical innovation. BioSPARK builds on insights from a design workshop and formative pilot study to support not only finding inspirations but also transferring inspirations into the target design domain and more deeply engaging with them during ideation. We found in a user study that the LLM-enabled features we explored in BioSPARK – generating and clustering inspirations, introducing sparks to help map the inspiration to the design problem, trade-offs to help users consider design constraints, and free-form chat to explore inspirations more deeply – resulted in participants generating more ideas and exploring more different species without a significant decrease in diversity compared to a ‘gold standard’ condition using AskNature inspirations and ChatGPT. Furthermore, BioSPARK appeared to keep users in the flow of ideation, reduce the cognitive effort in transferring and adapting ideas, and help people engage more deeply in considering how they could use inspirations and the design spaces they unlocked.

One significant concern we had was that the features that were aimed at deeper engagement, such as sparks, might counterproductively decrease engagement and increase fixation because of how fleshed out the connections were in terms of articulating an entire, detailed design idea embodying the inspiration’s mechanism in the target domain (e.g., using spider silk for lifting a wheelchair or creating a ramp). The higher the fidelity of an inspiration the more it may incur fixation and direct use rather than creative adaptation [65]. In our evaluation we cannot definitively reject this possibility. While our analysis of interaction logs and think-aloud transcripts suggests that participants actively engaged with the mechanism inspirations – recognizing their relevance, deepening their understanding, and transferring them to the target design domain – further investigation is required to unpack the effects of AI-enabled content generation features, such as the spark, trade-off, and Q&A cards, on the idea generation process.

With this caveat, we highlight some factors that may warrant future research. First, although the AI appeared to reduce the cognitive load of mapping the idea to the design space, the decision to do the mapping in the first place was user-driven, prompted by them saving an inspiration. Thus before seeing the AI mapping they needed to notice something interesting or relevant about the inspiration, even if they didn’t fully make the connection between it and the problem domain. This self-driven curiosity and agency may influence how users engage with sparks and trade-off cards. Future systems could explore which user actions and levels of agency are necessary to foster a sense of ownership and spur initial engagement with inspirations.

Another factor that might have driven engagement was perceived ownership of inspirations. Previous work has identified that ownership and attribution are key elements of human-AI collaboration [57]. In our study we noticed participants making attribution

statements about the sparks, such as “*That's not really my idea, [it's] ChatGPT's idea but okay*”. A common theme among participants was discussing how they modified the sparks to make ideas more their own and to avoid “plagiarism”, even though they were told they could use the sparks as they wish. It's unclear why designers in other studies finding fixation when using LLM and generative AI did not similarly adapt and riff on ideas in order to build ownership, but a possibility that might be explored is that an integrated system that frames AI-generated ideas as intermediate products, as we do in BIOSPARK, might be more effective at promoting deeper engagement than an unstructured system or one where they claim prominence as more final products.

Fixation vs. Contextualization. Another concern we had in our system design was whether contextualizing the system interface features in the design problem would lead to narrowing of the design ideas users would explore. Sparks, trade-off cards, and the free-form chat interface were all contextualized with the source design problem, with the goal of reducing the cognitive effort needed for users to engage with the details of the inspirations relevant to their goals.

This largely appeared to hold true, with users finding the contextualization useful, and even sometimes “magical”. While they were technically able to do this with the baseline system and sometimes did (“*It feels like I got the seed like the very starting point idea from AskNature and then generating actual ideas from it was done by ChatGPT, like translating the seed into actual ideas.*”), the efficiency of the built-in contextualization was frequently mentioned as useful as a jumping off point rather than replacing cognitive work (e.g., “*it was able to produce things without me having to like prompt it. And I think that allowed me to spend more time, maybe thinking about specific connections between the mechanisms and the design features it was suggesting.*”). Participants also commented on the reduction of cognitive effort in transferring inspirations, as P1 described: “*I just felt like it was able to produce things without me having to like prompt it. And I think that allowed me to spend more time, maybe thinking about specific connections between the mechanisms and the design features it was suggesting. Whereas... with [ChatGPT] I felt like I had to spend more time like doing to allow it to actually help, but with [BIOSPARK] I felt like the AI system already knew what I needed, so it saved that step.*” P5 felt it was “*easy to highlight the strengths of each design and extract useful design features from it for integration into a new idea for mitigating anticipated challenges*”, and that “*(The sparks are) very specific... how did it know? How are these so specific to bike rack design? Wait, it's already pitching me different ideas. That's so cool*” (P5).

In contrast, the baseline system seemed to result in significant cognitive effort, both physically not having the context of the design problem, and attentionally in not keeping them focused. For example, P12 commented that “*This [AskNature.org] article could be tailored more to the bike rack design, instead it's about stuff that's in motion... so it's hard to know what the relevance is*” (P12). Without the persistent context, and source inspirations on AskNature disconnected from the target design problem, participants felt the inspirations in the baseline “*didn't tell me about too much in terms of what insights I can draw to support (the scenario described by the*

design brief)” (P1) and that they had to engage in multiple back-and-forth's with ChatGPT to glean transferrable insights and ideating specific design ideas based on them, which was time-consuming without the persistent context and consistent user guidance: “*(Types in ChatGPT) it's giving me specific animals (instead of transferrable mechanisms) so I'm going to pause the generation. Let me see... (Types in ChatGPT) 'what shape?'. Oops that didn't work, let me try again. 'What shape that could be useful for industrial design?'*” (P12). For P4, AskNature.org felt “*like a generic platform*” and that building off of its material using AI felt like “*asking just random questions to ChatGPT, while in BIOSPARK I was asking related questions*” (P4).

Overall, our results suggest a more nuanced consideration of context and fixation than previously considered, in which helping users reduce cognitive load throughout the analogical innovation process while positioning AI suggestions as intermediate products in the system flow, could represent a profitable paradigm to explore.

Limitations. Our findings and analysis have several limitations. First, our access to professional designers working in large organizations was limited to a design workshop. Our formative study and user studies involved heterogeneous participant pools, including freelance designers recruited from UpWork and design and PhD students recruited from an arts college. These participants may not fully represent the broader population of design professionals.

Second, alternative interpretations of our data are possible based on the operationalization of the measures we used. Creativity and ideation measurement spans multiple disciplines, and our chosen metrics – focused on engagement with inspirations, as well as the quality, quantity, and diversity of ideas – reflect a specific subset of this research. Notably, our approach to measuring idea diversity, such as extracting species' names from ideas or relying on a particular LLM in our computational pipeline, may introduce biases.

Additionally, our study protocol, while informed by professional design practices, involved artificial scenarios and time constraints that may not mirror real-world design contexts. The self-driven curiosity and agency observed in participants' engagement with AI-generated inspirations may differ in professional settings, where constraints and priorities vary. Future studies embedding such systems into designers' ongoing work or making them publicly available for broader use could yield valuable insights into practical benefits and remaining challenges.

Our evaluation suggests that while the system reduces cognitive load during the analogical ideation process, the decision to map inspirations to the design space remains user-driven, requiring participants to identify relevant or interesting aspects of inspirations. The interplay between AI support and user agency as well as their effects on design fixation warrants further exploration to better understand the balance between automation and human creativity in fostering meaningful and innovative design processes.

Finally, our work aims to support only the ideation portion of the design process, while a significant part of the design process also involves prototyping and evaluation. Further work is needed to extend these results to incorporating later stages of the design process.

8 Conclusion

In this work we present BIOSPARK, a creativity partner for analogical design. BIOSPARK proposes a new interaction paradigm for reducing the cognitive effort in finding, recognizing, mapping, and creatively adapting diverse inspirations from distant domains. Future work remains in this area to generalize our diversification structure and goal-driven inspiration discovery pipeline for generating analogical inspirations from different domains, and in further personalizing, controlling, and customizing support for user recognition and idea development beyond what was explored here to accelerate and materialize high-impact innovations. We imagine a future in which engineers and designers could find inspirations based on deep analogical similarity that move beyond domain boundaries to drive innovation across fields.

Acknowledgments

This research was supported by the Toyota Research Institute and the Office of Naval Research.

References

- [1] [n. d.]. Ask Your PDF. <https://askyourpdf.com/>. Accessed: 2023-04-05.
- [2] [n. d.]. Naked mole-rat. https://en.wikipedia.org/wiki/Naked_mole-rat. Accessed: 10-02-2023.
- [3] Michael F Ashby and Kara Johnson. 2013. *Materials and design: the art and science of material selection in product design*. Butterworth-Heinemann.
- [4] Kevin D Ashley. 1991. Reasoning with cases and hypotheticals in HYPO. *International journal of man-machine studies* 34, 6 (1991), 753–796.
- [5] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2023. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–38.
- [6] O. Bánki, Y. Roskov, M. Döring, G. Owre, D. R. Hernández Robles, C. A. Plata Corredor, T. Stjernegaard Jeppesen, A. Örn, L. Vandepitte, D. Hobern, P. Schalk, R. E. DeWalt, K. Ma, J. Miller, T. Orrell, R. Aalbu, J. Abbott, R. Adlard, E. M. Adriennes, and et al. 2023. Catalogue of Life Checklist. *Catalogue of Life* (14 September 2023). <https://doi.org/10.48580/dd24x4>
- [7] Jaime G Carbonell. 1983. Learning by analogy: Formulating and generalizing plans from past experience. In *Machine learning*. Springer, 137–161.
- [8] Jaime Guillermo Carbonell. 1985. *Derivational analogy: A theory of reconstructive problem solving and expertise acquisition*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.
- [9] Amaresh Chakrabarti, Prabir Sarkar, B Leelavathamma, and BS Nataraju. 2005. A functional representation for aiding biomimetic and artificial inspiration of new ideas. *Ai Edam* 19, 2 (2005), 113–132.
- [10] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (Nov. 2018), 21 pages. <https://doi.org/10.1145/3274300>
- [11] Hyumin Cheong and L. H. Shu. 2014. Retrieving Causally Related Functions From Natural-Language Text for Biomimetic Design. *Journal of Mechanical Design* 136, 8 (02 Jun 2014), 081008–081008–10. <https://doi.org/10.1115/1.4027494>
- [12] John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 575–593. <https://doi.org/10.18653/v1/2023.acl-long.34>
- [13] Olivier Darrigol. 2010. The analogy between light and sound in the history of optics from the Ancient Greeks to Isaac Newton. Part 1. *Centaurus* 52, 2 (2010), 117–155.
- [14] Nicholas Davis, Chih-Plin Hsiao, Kunwar Yashraj Singh, Lisa Li, Sanat Moningi, and Brian Magerko. 2015. Drawing apprentice: An enactive co-creative agent for artistic collaboration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 185–186.
- [15] Jon-Michael Deldin and Megan Schuknecht. 2013. The AskNature database: enabling solutions in biomimetic design. In *Biologically inspired design: computational methods and tools*. Springer, 17–27.
- [16] David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A Saurous, Jascha Sohl-Dickstein, et al. 2022. Language model cascades. *arXiv preprint arXiv:2207.10342* (2022).
- [17] Hen Emuna, Nadav Borenstein, Xin Qian, Hyeonsu Kang, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2024. Imitation of Life: A Search Engine for Biologically Inspired Design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 503–511.
- [18] Encyclopedia [n. d.]. Encyclopedia of Life (EOL). <https://eol.org/docs/what-is-eol>. Accessed: 10-02-2023.
- [19] Raymond Fok, Joseph Chee Chang, Tal August, Amy X Zhang, and Daniel S Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. *arXiv preprint arXiv:2310.07581* (2023).
- [20] Kenneth Forbus. 2001. *Exploring analogy in the large*. MIT Press.
- [21] Kenneth D Forbus, Ronald W Ferguson, and Dedre Gentner. 1994. Incremental structure-mapping. In *Proceedings of the sixteenth annual conference of the Cognitive Science Society*. 313–318.
- [22] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.
- [23] Dedre Gentner and Russell Landers. 1985. Analogical reminding: A good match is hard to find. In *Unknown Host Publication Title*. IEEE, 607–613.
- [24] Dedre Gentner, Mary Jo Rattermann, and Kenneth D Forbus. 1993. The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive psychology* 25, 4 (1993), 524–575.
- [25] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*. 1002–1019.
- [26] Mary L Gick and Keith J Holyoak. 1980. Analogical problem solving. *Cognitive psychology* 12, 3 (1980), 306–355.
- [27] Mary L Gick and Keith J Holyoak. 1983. Schema induction and analogical transfer. *Cognitive psychology* 15, 1 (1983), 1–38.
- [28] Ashok K Goel, Swaroop Vattam, Bryan Wiltgen, and Michael Helms. 2012. Cognitive, collaborative, conceptual and creative—Four characteristics of the next generation of knowledge-based CAD systems: A study in biologically inspired design. *Computer-Aided Design* 44, 10 (2012), 879–900.
- [29] Graeme S Halford, Rosemary Baker, Julie E McCredden, and John D Bain. 2005. How many variables can humans process? *Psychological science* 16, 1 (2005), 70–76.
- [30] Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How Far Can We Extract Diverse Perspectives from Large Language Models? Criteria-Based Diversity Prompting! *arXiv preprint arXiv:2311.09799* (2023).
- [31] Douglas R Hofstadter, Melanie Mitchell, et al. 1995. The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory* 2 (1995), 205–267.
- [32] Keith J Holyoak and Paul Thagard. 1989. Analogical mapping by constraint satisfaction. *Cognitive science* 13, 3 (1989), 295–355.
- [33] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating Innovation Through Analogy Mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) (KDD '17). ACM, New York, NY, USA, 235–243. <https://doi.org/10.1145/3097983.3098038>
- [34] Ting-Hao'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C Lee Giles. 2020. Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. *arXiv preprint arXiv:2005.02367* (2020).
- [35] John E Hummel and Keith J Holyoak. 2003. A symbolic-connectionist theory of relational inference and generalization. *Psychological review* 110, 2 (2003), 220.
- [36] Shuo Jiang, Jie Hu, Kristin L Wood, and Jianxi Luo. 2022. Data-driven design-by-analogy: state-of-the-art and future directions. *Journal of Mechanical Design* 144, 2 (2022), 020801.
- [37] Philip N Johnson-Laird. 2005. Flying bicycles: How the Wright brothers invented the airplane. *Mind & Society* 4 (2005), 27–48.
- [38] Hyeonsu B Kang, Xin Qian, Tom Hope, Dafna Shahaf, Joel Chan, and Aniket Kittur. 2022. Augmenting Scientific Creativity with an Analogical Search Engine. *ACM Trans. Comput.-Hum. Interact.* 29, 6, Article 57 (nov 2022), 36 pages. <https://doi.org/10.1145/3530013>
- [39] Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A Mixed-Initiative System for Scholarly Synthesis and Sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (, San Francisco, CA, USA.) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 43, 19 pages. <https://doi.org/10.1145/3586183.3606759>
- [40] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojian Ma. 2021. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [41] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbogzorgi. 2020. Creative sketching partner: an analysis of human-AI co-creativity. In *Proceedings of the 25th international conference on intelligent user interfaces*.

- 221–230.
- [42] Jingoog Kim and Mary Lou Maher. 2023. The effect of AI-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation* 11, 2 (2023), 81–98.
- [43] Seo-young Lee, Matthew Law, and Guy Hoffman. 2024. When and How to Use AI in the Design Process? Implications for Human-AI Design Collaboration. *International Journal of Human-Computer Interaction* (2024), 1–16.
- [44] Julie S Linsey, Arthur B Markman, and Kristin Lee Wood. 2012. Design by analogy: A study of the WordTree method for problem re-representation. (2012).
- [45] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A Myers. 2023. Selenite: Scaffolding Decision Making with Comprehensive Overviews Elicited from Large Language Models. *arXiv preprint arXiv:2310.02161* (2023).
- [46] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alom, Nouha Dziri, Shrimai Prabhuloye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2024).
- [47] Lena Mamykina, Linda Candy, and Ernest Edmonds. 2002. Collaborative creativity. *Commun. ACM* 45, 10 (2002), 96–99.
- [48] Lori McCay-Peet and Elaine Toms. 2011. Measuring the dimensions of serendipity in digital environments. *Information Research: An International Electronic Journal* 16, 3 (2011), n.3.
- [49] Microsoft Corporation. 2023. Microsoft Copilot: Your Everyday AI Companion. <https://copilot.microsoft.com>. Accessed: 2023-04-05.
- [50] Meg Monk. 2013. BYU engineers use origami to make more space in space. <https://universe.byu.edu/2013/12/12/byu-engineers-use-origami-to-make-more-space-in-space/>. Accessed: YYYY-MM-DD.
- [51] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. 2011. How many species are there on Earth and in the ocean? *PLoS biology* 9, 8 (2011), e1001127.
- [52] Melis Muradoglu, Joseph R Cimpian, and Andrei Cimpian. 2023. Mixed-effects models for cognitive development researchers. *Journal of Cognition and Development* 24, 3 (2023), 307–340.
- [53] Xi Niu, Fakhri Abbas, Mary Lou Maher, and Kazjon Grace. 2018. Surprise me if you can: Serendipity in health information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [54] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [55] OpenAI. 2023. ChatGPT - Your Friendly AI Chatbot. <https://chat.openai.com/>. Accessed: 2023-04-28.
- [56] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774 [cs.CL]*
- [57] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the effects of technological writing assistance for support providers in online mental health community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [58] Edwin A Peraza-Hernandez, Darren J Hartl, Richard J Malak Jr, and Dimitris C Lagoudas. 2014. Origami-inspired active structures: a synthesis and review. *Smart Materials and Structures* 23, 9 (2014), 094001.
- [59] Tony Rees, Leen Vandepitte, Bart Vanhoorne, and Wim Decock. 2020. All genera of the world: an overview and estimates based on the March 2020 release of the Interim Register of Marine and Nonmarine Genera (IRMNG). *Megataxa* 1, 2 (2020), 123–140.
- [60] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [61] Jami J Shah, Steve M Smith, and Noe Vargas-Hernandez. 2003. Metrics for measuring ideation effectiveness. *Design studies* 24, 2 (2003), 111–134.
- [62] L Siddharth and Amareesh Chakrabarti. 2018. Evaluating the impact of Idea-Inspire 4.0 on analogical transfer of concepts. *AI Edam* 32, 4 (2018), 431–448.
- [63] L Sweetlove. 2011. Number of species on Earth tagged at 8.7 million. *Nature* (23 August 2011). <https://doi.org/10.1038/news.2011.498>
- [64] Guy Tevet and Jonathan Berant. 2020. Evaluating the evaluation of diversity in natural language generation. *arXiv preprint arXiv:2004.02990* (2020).
- [65] Luis A Vasconcelos and Nathan Crilly. 2016. Inspiration and fixation: Questions, methods, findings, and challenges. *Design Studies* 42 (2016), 1–32.
- [66] Swaroop S. Vattam and Ashok K. Goel. 2011. Semantically Annotating Research Articles for Interdisciplinary Design. In *Proceedings of the Sixth International Conference on Knowledge Capture (K-CAP '11)*. ACM, New York, NY, USA, 165–166. <https://doi.org/10.1145/1999676.1999707>
- [67] Julian FV Vincent and Darrell L Mann. 2002. Systematic technology transfer from biology to engineering. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 360, 1791 (2002), 159–173.
- [68] Samangi Wadinambiarachchi, Ryan M Kelly, Saumya Pareek, Qiushi Zhou, and Eduardo Veloso. 2024. The Effects of Generative AI on Design Fixation and Divergent Thinking. *arXiv preprint arXiv:2403.11164* (2024).
- [69] Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* 7, 9 (2023), 1526–1541.
- [70] Wikipedia. [n. d.]. Genus. https://en.wikipedia.org/wiki/Genus#cite_note-10. Accessed: 10-02-2023.
- [71] Jen-Her Wu and Shu-Ching Wang. 2005. What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & management* 42, 5 (2005), 719–729.
- [72] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (CHI '14). Association for Computing Machinery, New York, NY, USA, 1245–1254. <https://doi.org/10.1145/2556288.2557371>
- [73] Yuanshuo Zhao, Ioana Baldini, Prasanna Sattigeri, Inkil Padhi, Yoong Keok Lee, and Ethan Smith. 2018. Data Driven Techniques for Organizing Scientific Articles Relevant to Biomimicry. In *ACM/AAAI Artificial Intelligence, Ethics and Society (AIES) conference*.
- [74] Qihao Zhu, Xinyu Zhang, and Jianxi Luo. 2023. Biologically inspired design concept generation using generative pre-trained transformers. *Journal of Mechanical Design* 145, 4 (2023), 041409.
- [75] Shannon A Zirbel, Mary E Wilson, Spencer P Magleby, and Larry L Howell. 2013. An origami-inspired self-deployable array. In *ASME 2013 Conference on Smart Materials, Adaptive Structures and Intelligent Systems*. American Society of Mechanical Engineers Digital Collection.
- [76] Haoyu Zuo, Qianzhi Jing, Tianqi Song, Lingyun Sun, Peter Childs, and Liuqing Chen. 2022. WikiLink: An encyclopedia-based semantic network for design creativity. *Journal of intelligence* 10, 4 (2022), 103.

A Inspiration Design Probe Study Prototype Implementation

A.1 Representative Mechanism Image Curation

To aid designers' visual understanding of and pique curiosity for biological-analogical mechanisms, we retrieve representative images for corresponding textual mechanism descriptions. We use Google Custom Search⁹ with queries as “[organism name]:[mechanism description]” and the file type set to images and the safe search mode enabled. We choose the first result of Custom Search as the visual representation of each mechanism.

A.2 Interface

To facilitate designers' understanding and synthesis of mechanism inspirations, we develop several interaction features available on the probe interface (fig. 7). The **Explain** button is located in tooltips that pop up when the user places the mouse over on a mechanism card in the board UI (fig. 7, first panel). When the user clicks on the button, BioSPARK sends a prompt to GPT4 requesting elaboration of the interacted mechanism and the organism in the context of the chosen engineering design problem. The **Compare** tab is located in the control bar of the sidebar of the interface. To use this, users need to first click on (at least) two mechanism cards from the left, saving them to the ‘saved inspirations’ panel at the top of the sidebar. There, users can check any two of the saved mechanisms they wish to compare. BioSPARK sends a prompt to GPT4 when the user clicks on the tab, requesting comparison of pros and cons between the two mechanisms in the context of the chosen engineering problem. The result is formatted into a markdown table, with each mechanism as the header followed by pros and cons rows detailing each point. The **Combine** tab is also located in the control bar of the sidebar in the interface. Similarly with Compare, users can check two saved mechanisms they wish to see combined. BioSPARK sends a prompt to GPT4 then requesting elaboration of a mechanism that combines

⁹<https://developers.google.com/custom-search/v1/overview>

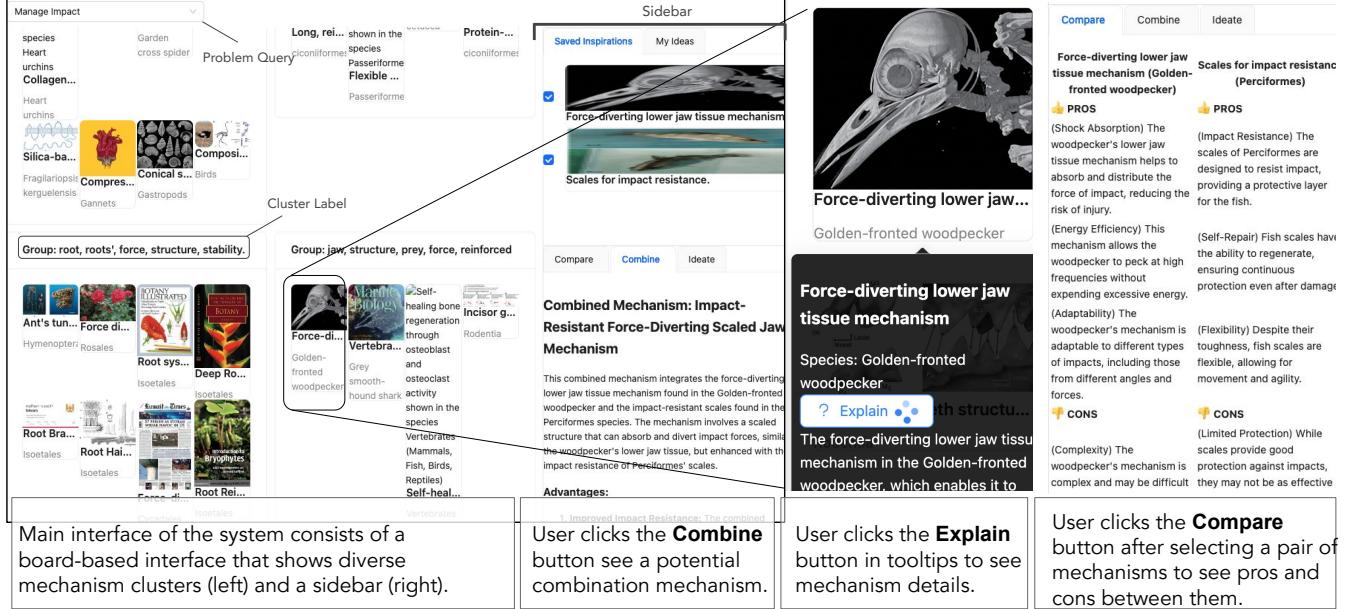


Figure 7: Inspiration design probe interface and a subset of available interaction features. The interface consists of a left pane that shows clusters of semantically similar mechanisms that was scrollable, and a right pane that included a holding tank for user-saved mechanisms. When the user checks two of the saved mechanisms in the holding tank and clicks one of the tabs underneath, the system generated the corresponding content, such as the comparison of two mechanisms in a pros-and-cons table, a new idea that combines the two mechanisms, and the ‘Ideate’ button that provided critique on the ‘Combine’ idea.

the two selected mechanisms, and explain its potential advantages in the context of the chosen engineering problem. The result is also formatted into a markdown page using section title and headers for demarcating the content.

B BIOSPARK Implementation Details

B.1 Visual Representation of Inspirations

To support the design goal of visual representations of inspirations we were inspired by AskNature.org, a popular web repository for bio-inspired design. AskNature pages are designed to include a prominent close-up and centered portrait of a species that creates a striking visual and invokes curiosity.

We retrieve animal images using Google Search and Adobe Stock Images APIs, and use GPT-4V to rank the quality in terms of the visual focus and the potential value for mechanism understanding. The first author developed a prompt (fig. 8) by reviewing the ranked images, their scores, and rationale for 10 species. Applying this led to a consistent pattern of reasonable visual processing and instruction-following, as demonstrated in fig. B.1.2 to be useful.

B.1.1 Prompt Design. The developed prompt is shown in fig. 8.

B.1.2 Sample Scores and Justifications. For the example shown in fig. 9, the model scored and ranked the images as top-2 and bottom-2 candidates (from left-to-right) for the species ‘Scaly-foot snail’ as follows:

Top-1 | Score: 95 | Rationale: This image provides an excellent close-up view of the scaly-foot snail, highlighting intricate details of the iron-infused sclerites and the shell's texture which are indicative of its biological mechanisms.

Top-2 | Score: 92 | Rationale: This image offers a clear and detailed close-up of the scaly-foot snail, clearly showcasing the iron-infused sclerites and the unique appearance of its iron-mineralized shell, which are valuable for understanding its biological mechanisms.

Bottom-2 | Score: 30 | Rationale: Although this underwater image shows the scaly-foot snail in its natural environment, the details necessary to comprehend the biological mechanisms such as the sclerites and shell are obscured and not clearly visible.

Bottom-1 | Score: 10 | Rationale: The image is a stylized, cartoon-like representation of a scaly-foot snail, lacking detailed visual information about the species' biological mechanisms such as the iron-infused sclerites and iron-mineralized shell.

B.2 Goal-driven Inspiration Discovery Pipeline Implementation

B.2.1 Structuring AskNature blog posts into seed problem-mechanism-organism schemas. To source a set of diverse, high-quality biological mechanisms for a given problem, BIOSPARK starts

```
[User Message]
Judge each image based on how clearly it shows the real species (i.e., photos focusing on one instance of the species in
→ the wild is better than cartoons, drawings, or species photographed in the distance)
{species} and contains visual details that help viewers understand the following biological mechanisms:
===
{formatted_mechanisms}
===
For each image given, reply with a number between 0 and 100 as its "score", where a higher number represents a higher
→ quality of the picture,
and also provide rationale for your decision in "rationale".
Output a list, with the following format. Exclude any other character than the comma between dictionaries in the list:
[{"score": "50", "rationale": "..."}, ...]
```

Figure 8: The prompt used to score each species image with respect to its visual focus on the species and how helpful it might be for understanding the specific mechanism of the species.



Figure 9: The top and bottom two images ranked by GPT-4V (scores from the left: 95, 92, 30, 10) for ‘Scaly-foot snail’.

from a seed set of expert-curated biological mechanisms on AskNature (fig. 3, Step 1). AskNature.org provides a curated list of organisms with detailed descriptions of their unique strategies to functional problems (e.g., ‘Manage Impact’, ‘Modify Speed’). The organisms and strategies can be grouped by function and viewed as a list. To curate a seed set of high-quality mechanisms, we first choose a functional problem p predicted to be highly relevant to automobile designers, excluding irrelevant functions such as ‘Adapt Behaviors’, ‘Adapt Genotype’, ‘Coevolve’, ‘Maintain Community’. We access the sub-list of organisms $o \in O$ and strategies posted to p on AskNature’s group-by-function page by parsing the HTML code using the BeautifulSoup package on Python. We then access the blog post for each organism-strategy page using the parsed URL and parse the returned HTML page to get the title, description, and references (if available).

At this stage, the returned unstructured text is yet to contain a succinct mechanism description. Furthermore, we found that some blog posts do not contain any body text despite having a title and are accessible via the URL. Some of these missing blog posts indicated that they are in maintenance and/or planned to be updated. To structure the raw blog post text AskNature $_{(o,p)}$, we prompt GPT4 [56] to succinctly describe (i.e., using 12 words or less) the core mechanism (i.e., excluding the qualities or effects, and focusing on mechanisms with engineering design implications), given (o, p) (if blog post text is missing) or $(o, p, \text{AskNature}_{(o,p)})$. The returned mechanism description m along with the function description makes up the problem-mechanism schema for each organism: $\{o \in O | (p, m, o)\}$.

B.2.2 Iteratively expanding mechanisms dataset by traversing constructed taxonomic trees. Using each schema as a seed, we iteratively prompt GPT4 to find relevant mechanisms for the given mechanism and problem, using an even mixture of breadth- and depth-focused expansion strategies (fig. 3, Step 2). To enable structured diversification of organisms and their mechanisms beyond prior work that relied on token-level manipulation or naïvely prompting LLMs, we

guide LLMs how and where to expand by leveraging organism taxonomic hierarchies. At each iteration of expansion (fig. 3, Step 2), we aggregate the organisms represented in found mechanisms up to that point, and construct a taxonomic tree featuring seven levels of hierarchy on Tree of Life: {domain, kingdom, phylum, class, order, family, genus, species}, where domain representing the highest level and species representing the lowest level on the hierarchy.

Given this tree, we aim to identify sparsely populated branches for expansion. We cut the tree at a given reference expansion level (e.g., class), and sort the taxonomic ranks (nodes) on that level by the number of its immediate children nodes¹⁰, in an increasing order. For performance, we batch 10 prompts to send to GPT4 for expansion. For half of the prompts, we instruct **breadth-first expansion** which asks GPT4 to first identify *sibling* nodes at the given reference taxon level and existing nodes (up to 50 most populated nodes).

For example, the prompt asks “come up with a few biological classes not in {...names of excluded classes...}”. The breadth-first expansion prompt then instructs GPT4 to repeat the following: 1) Choose one taxon from the list it came up with; 2) Succinctly describe (i.e., using 14 words or less) new mechanisms m related to a problem p . For the rest of the prompts, we instruct **depth-first expansion** which asks GPT4 to first identify a new *children* node at the given reference taxon level and existing children nodes (up to 50 randomly sampled children). For example, the prompt asks “come up with a few biological families in order araneae that are not any of {araneidae, ...}”. The depth-first expansion prompt then instructs GPT4 to repeat a similar procedure as breath-first expansion. The prompt details are provided in fig. 10 (the depth-focused expansion prompt) and fig. 11. In the prototype system, we run 10 batches for expansion to construct dataset of mechanisms for each problem.

The returned list of mechanisms and organisms text are then fed into the second GPT4 prompt for structuring them into a list of {mechanism, organism} dictionaries. Finally, using each organism name, we prompt GPT3.5-turbo to retrieve the seven-level taxonomic hierarchy, based on our model evaluation result showing its high accuracy (Appendix B.2.3).

¹⁰Alternatively, the entire size of the subtree, rather than immediate children, could be used for sorting

[System Message]

You are an expert biologist who knows species and their taxonomic hierarchy in detail.
 You can also come up with diverse problem-solving strategies found in nature relevant to engineering design problems.
 Do the following step-by-step.

[User Message]

1. Come up with a few biological {lower-taxon-plural} **IN** the {taxon} "{term}" AND **NOT** {exclude-user-prompt}
2. Select one {lower-taxon_singular} from the list you came up with.
3. Come up with short descriptions (up to 14 words or less) of new mechanisms found in the selected {lower-taxon-
↳ singular} that are applicable to the challenge of "{prob}".
4. Repeat step 2 and 3 for each selected {lower-taxon-singular} and think step-by-step. Number each step in your
↳ thinking and make it as short as possible.

Figure 10: The prompt used for depth-focused expansion of the mechanism dataset. The “lower-taxon-singular” or “lower-taxon-plural” is the singular and plural name of the subsequent level on the tree-of-life hierarchy, of the level “taxon”, respectively. The “term” is the name of the selected taxon. The “exclude-user-prompt” includes previously generated “taxon” names which are used to instruct the LLM to avoid duplicate generation. The “prob” and “src-mech” contain the problem and mechanism schemas to constrain generation.

[System Message]

{same as in the depth-focused expansion prompt}}

[User Message]

1. Come up with a few biological {taxon-plural} **NOT IN** the excluded {taxon-plural} below:
{exclude-user-prompt}
2. {same as in the depth-focused expansion prompt}
3. {same as in the depth-focused expansion prompt}
4. {same as in the depth-focused expansion prompt}

Figure 11: The prompt used for breadth-focused expansion of the mechanism dataset. See the depth-focused expansion prompt (fig 10) for parameters descriptions.

B.2.3 LLM-based Taxonomy Construction & Accuracy Evaluation.
 The main process in our diversification strategy is iterative construction of taxonomic trees at each stage of expansion with a set of problem-mechanism schemas and corresponding organisms $\{o \in O | (p, m, o)\}$ curated (in case of AskNature seeds) or generated up to that point. To construct the trees, the taxonomic hierarchy of each organism needs to be known. Here, we restrict our tree construction to seven levels of depth, ranging from the highest to lowest levels: domain, kingdom, phylum, class, order, family, genus, species. These levels provide considerable branch-switching opportunities for diversification, through significant changes in the number of members between levels and within each level of the hierarchy. For example, while the highest level domain consists of three members, Bacteria, Archaea, and Eukarya, there are estimated 8.7M species in the world [63]. The next level on the hierarchy, Genus, has an estimated number of 310K members [59], while the number in the subsequent level, families, is estimated at 8K [51] in 2011. The number of known species for each node on the hierarchy also changes considerably, further contributing to the diversification opportunities. For example while most non-avian reptile genera have only 1 species each, insect genera such as *LasioGLOSSUM* and *Andrena* have over 1,000 species each, while the flowering plant genus, *Astragalus*, contains over 3,000 known species [70].

Our initial exploration of suitable approaches to retrieve organism taxonomies involved using available resources such as the Global Biodiversity Information Facility API¹¹, Catalogue of Life [6], or the Encyclopedia of Life [18], where canonical species names were retrieved from the Darwin Core List of Terms¹² for corresponding organisms in problem-mechanism schemas. However, the limited coverage, data consistency, and API availability of these tools prevented their adoption. On the other hand, Wikipedia provides scientific classification for some of the organism articles (for example in the Pomelo article¹³, taxonomic names for Kingdom, Clade, Order, Family, Genus, and Species are available in the ‘biota’ information box that appears on the right-hand side of the page). However, this data was not readily available for scalable generation.

Procedure. LLMs may provide an alternative solution to the limitations of existing approaches for retrieving the taxonomic hierarchy for a given organism name. To test this idea, we curated 90 gold taxonomies using Wikipedia that have complete information in the ‘biota’ scientific classification info box (the complete list of 90 organism names can be found in Appendix B.2.3). For each organism, we prompted LLMs with each organism name zero-shot using

¹¹<https://www.gbif.org/developer/species>

¹²https://dwc.tdwg.org/list/#dwc_Organism

¹³<https://en.wikipedia.org/wiki/Pomelo>

the chat completions API endpoint¹⁴ using each model key. The prompt used for taxonomy generation for LLMs can be found in fig. 12. Once the hierarchy data is generated, we lower-cased the rank names for consistency.

GPT4’s Accuracy. We find that GPT4’s zero-shot taxonomy generation accuracy to range between 94.4% and 100% (Table 6). The lowest accuracy was observed in the family taxonomy, followed by order (96.7%) and genus (98.9%).

Error analysis. We find that some error cases in taxonomy generation could be attributed to recent changes in classification in the literature. For example, both GPT4 and GPT3.5-turbo models classified naked mole-rats as then literature-accepted ‘Bathyergidae’ for their family, same as other African mole-rats. However, more recently naked mole-rats were placed in a separate family, Heterocephalidae [2].

Among the error cases overlapping between the two models, we found cases that either the GPT3.5-turbo or the GPT4 model wins over the other (e.g., for ‘hummingbird’, GPT3.5-turbo generated ‘archilochus’ as its genus whereas GPT4 generated ‘various’; for ‘boxer crab’, GPT3.5-turbo generated ‘hymenoptera’ which is an order of insects, whereas GPT4 generated ‘decapoda’, which is the correct order). In other cases, both models outputted similarly incorrect answers, for example for ‘sea snail’, GPT3.5-turbo generated ‘neogastropoda’ whereas GPT4 generated ‘archaeogastropoda’ (the Wikipedia gold answer was ‘lepetellida’).

System Optimization: GPT3.5-turbo’s Accuracy. We find that GPT3.5-turbo has comparable accuracy levels with GPT4 in zero-shot taxonomy generation. The highest misalignment occurred in genus, with a 6.67% error rate (equivalent to 6 out of 90). Appendix B.2.3 provides a further qualitative error analysis of models’ comparative performance. Based on these results, we opted for the more efficient GPT3.5-turbo model in our pipeline. We leave further exploration of the capabilities of smaller, fine-tuned base LLMs, with implications for LLM cascade¹⁵, to future work.

Complete List of Organisms Used for Taxonomy Generation. {‘spider monkey’, ‘prairie dog’, ‘garden tiger moth’, ‘african sacred ibis’, ‘argiope argentata’, ‘ostrich’, ‘groundhog’, ‘danio rerio’, ‘gannet’, ‘deer’, ‘cattle’, ‘glyptodon’, ‘alligator snapping turtle’, ‘leopard’, ‘arctic ground squirrel’, ‘cormorants and shags’, ‘bears’, ‘squirrels’, ‘herons’, ‘european badger’, ‘golden silk orb-weaver’, ‘aardvark’, ‘seahorses’, ‘banner-tailed kangaroo rat’, ‘hyenas’, ‘pink fairy armadillo’, ‘giant otter’, ‘bighorn sheep’, ‘hippopotamus’, ‘california ground squirrel’, ‘european bee-eater’, ‘beech marten’, ‘leopard gecko’, ‘tailorbird’, ‘testudinidae’, ‘emperor penguin’, ‘northern pike’, ‘giant clam’, ‘stoat’, ‘horse’, ‘nutria’, ‘tree-kangaroo’, ‘giraffe’, ‘guinea baboon’, ‘ferret’, ‘bonytail chub’, ‘baya weaver’, ‘brook trout’, ‘pelican’, ‘mallard’, ‘roseate spoonbill’, ‘mountain

weasel’, ‘pocket gophers’, ‘lybia edmondsoni’, ‘giant anteater’, ‘common raccoon dog’, ‘dewdrop spiders’, ‘armadillo girdled lizard’, ‘arctic fox’, ‘bison’, ‘swordfish’, ‘bald eagle’, ‘chimpanzee’, ‘asbolus verrucosus’, ‘sperm whale’, ‘abalone’, ‘golden jackal’, ‘hornet’, ‘zebra’, ‘orangutans’, ‘peregrine falcon’, ‘atlantic cod’, ‘burrowing owl’, ‘african wild dog’, ‘maned wolf’, ‘honey bee’, ‘naked mole-rat’, ‘echidnas’, ‘bowerbirds’, ‘rhinoceros’, ‘beaver’, ‘bombyx mori’, ‘common box turtle’, ‘hummingbird’, ‘domestic sheep’, ‘wolverine’, ‘raccoon’, ‘evergreen bagworm’, ‘pig’, ‘muskrat’}

B.3 Active Ingredient Extraction

To extract active ingredients from the outputs of earlier diversification and goal-driven generation steps, we design a prompt for GPT4 (fig. 13).

B.4 Sparks

B.4.1 Spark Generation Prompt. The prompt used for generating sparks is detailed in fig. 14.

B.4.2 Precedent-based Diversification. We tested whether this precedent-based diversification approach leads to a significant improvement in terms of semantic diversity compared to generation without diversification by generating 20 sparks for each of 10 randomly selected seed mechanism inspirations for each of the two design problems in the user study.

We investigate semantic diversity at two levels, the whole text and the active ingredient of a spark. To get the active ingredient, we process the generated spark using the same process as before for extracting active ingredients from mechanisms (§4.5.2). We then encode each spark or active ingredient text into an embedding using the OpenAI’s text-embedding-3-large model. We construct pairs of spark or active ingredient embeddings using the 20 sparks generated for each seed mechanism for each of the two design problems, which amounted to 3,800 pairs, and calculate the average cosine distance among the pairs. This average represents the semantic diversity measure, which has been used in similar context in prior studies and was shown to be a viable measure of semantic diversity of natural language texts (*cf.* [25, 30, 64]). In order to ensure robustness of our results against the choice difference of the encoder model, we repeat the analysis using another popular encoder – the Sentence-bert model for embedding the text [60]. We find that, at the whole spark text level, the semantic diversity was significantly higher when precedent-based diversification was used ($M=.24$, $SD=.073$) than not ($M=.17$, $SD=.090$) ($t_{ind.}(7291.87)=-42.41$, $p < .0001$).

B.5 Trade-off Analysis Generation

The prompt used for generating a trade-off analysis is detailed in fig. 16.

B.6 Q&A Response Generation

The prompt used for generating a response to the free-form user question is detailed in fig. 17.

¹⁴<https://api.openai.com/v1/chat/completions>

¹⁵LLM cascade refers to a system design approach that adaptively chooses optimal LLM APIs for a given query. Smaller, task-specific LLMs are regarded as optimal when they exhibit higher or similar levels of performance compared to models that are orders of magnitude larger [16], with all else being equal.

[System Message]
 You are an expert biologist who knows species and their taxonomic hierarchy very well. Follow the instructions to the ↵ letter.
 - Return the scientific term for each taxonomic rank the species belongs to.
 - Enclose keys and values using double quotes ("...") and format them into a Python dictionary.
 - Use the taxonomic ranks as keys and corresponding scientific terms as their values.
 - Do not add any other text.

[User Message]
 What {"domain", "kingdom", "phylum", "class", "order", "family", "genus"} does "{organism}" belong to? Format your reply ↵ into a Python dictionary.

Figure 12: The prompt used to generate the taxonomy of each organism.

[System Message]
 Reply with a succinct (i.e., 15 words or less) description of the following biological mechanism's active ingredient.
 ↵ Follow the instructions.

[Instructions]
 - The active ingredient should describe how the species "act" upon its challenges to mitigate them, and include verb or ↵ verb phrasees.
 - Active ingredient descriptions should also focus on the integral ingredients such as its bodily parts, liquids, or ↵ evolutionary tactic that are concrete and distinctive.
 - Structure your output in the following format (do not output any characters other than the actual json-formatted ↵ dictionary):
 {"desc": "..."}

[User Message]
 {mechanism}

Figure 13: The prompt used to extract the active ingredient from a mechanism. The mechanism is description to extract from is provided as part of the user message to GPT4.

[System Message]
 Generate **2** highly different ideas that could solve the design problem: "{design_prob}".
 The design problem has constraints that the ideas must satisfy: {design_constraints}
 Generated ideas must be at least broadly related to the user-selected inspiration found in nature.
 Generated ideas should be novel and not redundant with the following ideas generated in the past: {prev_sparks}
 Describe each idea succinctly (i.e., max 500 characters), but ensure to provide sufficient details to help the user ↵ visualize the idea.
 Start each idea description with a short, eye-catching name that captures the gist.
 Output exactly in the following format, WITHOUT ANY OTHER TEXT:
 [{{"name": "IDEA 1 NAME", "desc": "IDEA 1 DESCRIPTION"}}, ...]

[User Message]
 User-selected inspiration from nature to base your generation on:
 ===
 {user_selected_mechanism_inspiration}
 ===

Figure 14: The prompt used to generate new sparks. We contextualize the prompt using the design problem description and the constraints provided with the problem, as well as 20 previously generated sparks for precedent-based diversification. We explicitly instruct the model to generate non-redundant sparks based on the history of precedents, and be succinct (i.e., under 500 characters), with a descriptive title. Finally the user-selected mechanism inspiration is provided as part of the user message to GPT4.

Model	DOMAIN	KINGDOM	PHYLUM	CLASS	ORDER	FAMILY	GENUS
GPT4	100% (90/90)	100% (90/90)	100% (90/90)	100% (90/90)	96.7% (87/90)	94.4% (85/90)	98.9% (89/90)
GPT3.5-turbo	100% (90/90)	100% (90/90)	100% (90/90)	100% (90/90)	95.6% (86/90)	95.6% (86/90)	93.3% (84/90)

Table 6: The accuracy of zero-shot taxonomy generation using only the organism name.

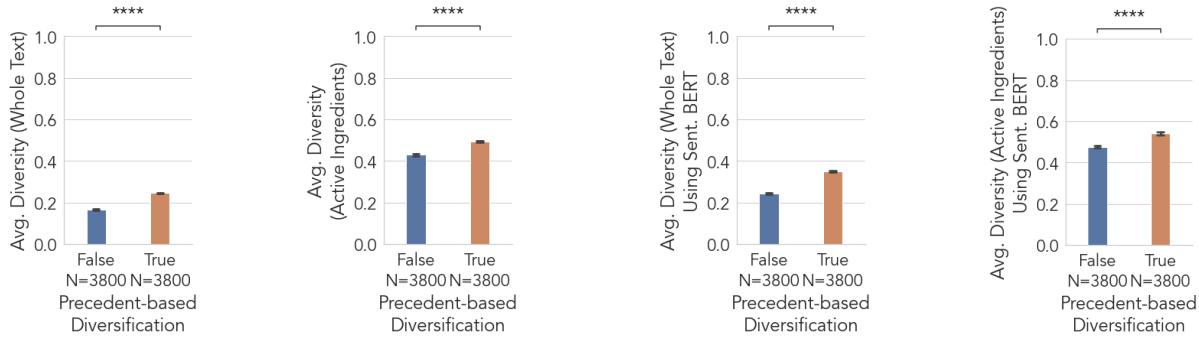


Figure 15: (First & Second) Bar graphs show that semantic diversity increased when using the precedent-based diversification approach, both at the whole spark and active ingredient levels; (Third & Fourth) Repeat analyses show the robustness of these results against the choice difference of the encoder model, when the Sentence-bert model [60] is used instead of OpenAI's text-embedding-3-large.

```
[System Message]
Generate up to **3** anticipated pros and cons for applying the user-selected mechanism to the design problem: "{→ design_prob}".
The design problem has constraints that the ideas must satisfy: {design_constraints}
Format the 'pros' and 'cons' into each column in a markdown table.
Place the header row at the top of the table: "| **PROS** | **CONS** |".
After the header row, place each 'pro'-'con' row; start each 'pro' or 'con' text with a succinct label (3 words or less) → , enclosed in parentheses.
```

```
[User Message]
User-selected inspiration from nature to base your generation on:
===
{user_selected_mechanism_inspiration}
===
```

Figure 16: The prompt used to generate a new potential design trade-off analysis. We contextualize the prompt using the design problem description and the constraints provided with the problem. We instruct the model to return the ‘pros’ and ‘cons’ of the mechanism inspiration in the context of the design problem using a markdown table format that places each pro-and-con pair in a new row, and give each item in the table a succinct (3 words or less) label. Finally the user-selected mechanism inspiration is provided as part of the user message to GPT4.

B.7 Detailed Research on Perplexity.ai

In order to support users with drilling down on related scientific research for each mechanism inspiration on demand, we designed a designated button (The ‘See more details on Perplexity.ai¹⁶’ button, fig. 2, ⓒ). Through interface pilots, we anticipated that the most common user workflow for drilling down on related research to be taking place after the user decides on a particularly interesting cluster for further consideration. When designing the button, we initially considered its placement on each of the cluster cards in the main interface, but decided to move it to the cluster modal view in

order to prevent clutter and support effective exploration of diverse design space on the main interface. In addition, to support the streamlined exploration – decision – further research workflow, we specifically placed the button at the end of the extended mechanism description featured in the modal fig. 2, ⓒ).

We implemented the button’s functionality as opening a new browser tab that contains search results of relevant research on the Perplexity.ai website. The search query was pre-populated using the following template:

Give me relevant details about "[active ingredient]"
→ commonly found in [species]

¹⁶<https://www.perplexity.ai/>

[System Message]
 Reply to the following user question about the mechanism: "{inspos}".
 Contextualize your response with the design problem: {design_prob}
 And with the constraints of the problem: {design_constraints}

[User Message]
 User question about the mechanism:
 ===
 {user_question}
 ===

Figure 17: The prompt used to generate a response to a free-form user question. We contextualize the prompt using the design problem description, its constraints, and the mechanism the user is asking about.

This functionality design was a compromise following our technical investigation at the time of development that showed the difficulty of implementing Perplexity.ai's search page results inside a native React.js application interface¹⁷ and the lack of API¹⁸ support for evidence generation¹⁹.

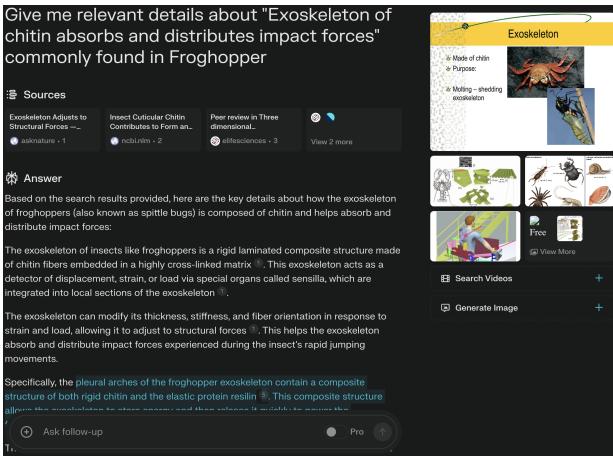


Figure 18: An example Perplexity.ai result page in a new browser tab when the user clicks on the ‘See more details on Perplexity.ai’ button on the mechanism modal view. The page describes how the froghopper exoskeleton contains a composite structure of both rigid chitin and the elastic protein resilin that allows the exoskeleton to store energy and then release it quickly to power the froghopper’s powerful jumps, and its supporting research, that may provide valuable details as described in our scenario (§4.1).

C User Study Additional Details

C.1 Study System Tutorials

Before participants start with each of the two main task in each condition, they were given a tutorial of the assigned systems via screen sharing. The interviewer demonstrated a step-by-step process and the main features of each system using a prepared script

¹⁷Perplexity.ai prohibits user requests that attempt to render its search results natively.

¹⁸<https://docs.perplexity.ai/>

¹⁹Last tested on March 17th, 2024.

that took around 8 minutes for the BioSPARK condition which had more features, and around 5 minutes for the baseline condition. In the baseline condition, participants were instructed to open up 5 different URLs each pointing to a pre-curated list of mechanisms for a functional category. The 5 functional categories used in the study were the same as those that were used for the BioSPARK backend dataset pipeline, and they were: Manage Impact, Manage Tension, Manage Compression, Manage Turbulence, and Modify Speed. In addition, participants in the baseline condition were instructed to sign in and open ChatGPT, and freely use it for understanding and ideating for the design problems using the information found from AskNature. When participants came up with each new idea during the task, they were told to write it down in a prepared Google spreadsheet that was shared in the beginning of the task. In the BioSPARK condition, participants were told to keep the stream space as a holder for their ideas, and thus delete any ideas they did not like or edit the text directly.

C.2 Examples of Judges’ Novelty, Feasibility, Value Evaluation

The following tables (Table ??,??) present 8 examples of the judges’ evaluations of design ideas in terms of novelty (N), feasibility (F), and value (V), along with the corresponding rationales. The evaluations were conducted independently by a domain expert and the first author using an agreed-upon rubric, demonstrating high reliability across these dimensions.

C.3 Study Result: Top-5 and Bottom-5 Ideas, Scores, and Coding Analogical Transfer

See Table 9.

C.4 Study Data Analysis: Extraction of unique design constraints described in each idea

To analyze the user engagement patterns involving consideration of design constraints (§6.4.3), we first extract the unique design constraints described by participants in each idea. We use GPT4 (gpt-4-turbo-preview) with a prompt (fig. 19) to perform the extraction. The first author reviewed the extracted constraints in terms of their coherence and uniqueness for a random set of 20 ideas and found that the extraction was satisfactory in terms of both the uniqueness of extracted constraints and their coherence. Using

Formatted Ideas	N	F	V	Novelty Rationale	Feasibility Rationale	Value Rationale
Idea 38: Inspired by the protective, retractable nature of turtle shells, this bike rack features a hard, aerodynamic shell that encases bikes completely. Its surface is segmented like scutes, allowing it to expand or contract to snugly fit 16", 20", and 26" bikes. When not in use or when driving, the shell retracts, minimizing drag and protecting bikes from theft and weather.	8	6	8	Connection to the turtle shell geometry is novel. The focus on the particular property of turtle shells which is that there is an empty space inside the shell where limbs can be retracted/folded into, and applying this design principle might provide useful new ideas for shape-shifting bike racks when the car is moving vs. stationary.	The estimated feasibility of the idea is somewhat high to start because the main mechanisms are extraction or contraction and segmented surface panels, which for the most part feel manufacturable with current tech and resources. However, more concrete design exploration, material selection, components design etc. may reveal previously unknown engineering challenges (hence the score is somewhat lowered for that reason).	The aerodynamic shape with expandable and contractable segmented scutes will be valuable for novel bikerack designs, and it opens up a new design space for future ideas. It may also be applicable to other domains where air or water resistance need to be considered, raising its value.
Idea 2: Tortoise Shell Geometry - For the bike racks stability & durability, I would use layers of different properties to add on or improve the strength mimicing the tortoise shell geometry. Also I would adapt this feature since it provides very good longevity of the product which enhances user satisfaction & trust.	6	3	4	In contrast to the earlier turtle shell idea that focused on the specific retraction mechanism, this one focuses more on the shape and geometry of the shell. However, the “use layers of different properties” is not very specific or novel.	“Use layers of different properties to add on or improve the strength mimicing the tortoise shell geometry” sounds like it would require a lot of materials research and feasibility testing before making this idea feasible and useful.	For durability, an efficient shell geometry that can be mass produced could be valuable. However it is unclear how multiple bike sizes will be accommodated. Hence the value is somewhat low.
Idea 18: Rack should be top mounted so that the bike is facing the same direction as the car to minimize drag (i.e. least amount of surface exposed to air during the ride)	1	10	2	“Top mounted rack” already exists and does not offer much novelty in the idea.	Rack mounted atop and facing the same direction as the car already exists as a commercial product, hence high engineering feasibility of the idea.	Top mounted racks are valuable, but the idea already being commercially available, unclear how much additional value (if any) this idea adds. Furthermore, it does not specify how the air resistance could be further reduced nor how the solution will accommodate different bike sizes, hence the value of the idea is low.
Idea 79: LeapLock FlexFit. Inspired by Anura's powerful legs, this bike rack uses a bio-mimetic spring mechanism that stretches to accommodate 16", 20", and 26" bikes. Its skeletal structure, mimicking frog bones, flexes to absorb road vibrations, protecting bikes. Aerodynamically shaped to reduce drag, it 'leaps' into a compact form when not in use, preserving fuel efficiency.	9	6	9	The spring mechanism that changes shape when in use vs. not, and also being able to absorb vibration and impact feels like a pretty novel idea.	Springs already exist, and there is presumably a large body of research around them that this idea could draw upon. However, the specific kinds of springs that exhibit the right kinds of properties here (like stretching the right amounts to accommodate for different bike sizes) might require a non-trivial amount of additional research into the materials and the construction of the springs etc.	This kind of novel spring mechanism that is aerodynamically shaped and size-adjustable could be highly valuable.

Table 7: Sample judges’ evaluations of design ideas by novelty (N), feasibility (F), and value (V), with rationales (continued in Table 8).

this data, we find that the length of each idea was also significantly correlated with the number of design constraints described in it ($\rho=.58$, $p < .0001$) (see §6.4.3 for the related analysis).

C.5 Study Data Analysis: Extraction of the species’ names that participants described as inspiring their ideas

To analyze the diversity of the species that participants were inspired from for their own ideas, we use GPT4 (gpt-4-turbo-prev

iew) with a prompt (fig. 20) to extract the species name from each participant idea and normalize it. The first author then reviewed the extracted species’ names in a random sample of 20 ideas and found that the extraction accuracy was satisfactory.

```
[System Message]
We're preparing each idea description for measuring the degree to which various design constraints were considered.

Chunk the idea description into unique segments each of which describe consideration(s) for a single coherent design
↳ constraint.

Output exactly in the following format (use double quotation marks to encapsulate any string), without any other text:
{"constraint_considerations": [["idea 1 constraint 1", "idea 1 constraint 2", ...], ... ]}
```

```
[User Message]
{list_of_participant_ideas}
```

Figure 19: The prompt used to extract coherent chunks of text that relates to unique design constraints. Participants' ideas are stringified and provided as part of the user message to GPT4.

```
[System Message]
We're extracting the source species' name that inspired each idea from the idea description.

Output exactly in the following format (use double quotation marks to encapsulate any string), without any other text:
{"source_species": ['species name for idea 1', 'species name for idea 2', ... ]}
```

```
[User Message]
{list_of_participant_ideas}
```

Figure 20: The prompt used to extract the species' names that inspired participants' ideas. Participants' ideas are stringified and provided as part of the user message to GPT4.

Formatted Ideas	N	F	V	Novelty Rationale	Feasibility Rationale	Value Rationale
Idea 111: Drawing from the compact and sturdy design of a beetle's shell, this bike rack folds out from a sleek, aerodynamic shell attached to the car's roof. It adjusts to fit various bike sizes, using a secure, adjustable locking mechanism that mimics the versatility of beetle wings.	8	5	7	The connection to the beetle's shell is novel. The connection between the beetle wings and adjustable locking mechanisms is also quite novel. It feels there is a fairly large gap of research from beetle wings to adjustable locking mechanisms, hence unclear whether current tech can manufacture it. Adjustable locking mechanism could be valuable for accommodating different bike sizes and dynamically changing volume/shape for aerodynamic efficiency. The expandable shell attached to the roof of a vehicle is also a somewhat valuable idea.	It feels there is a fairly large gap of research from beetle wings to adjustable locking mechanisms, hence unclear whether current tech can manufacture it.	adjustable locking mechanism could be valuable for accommodating different bike sizes and dynamically changing volume/shape for aerodynamic efficiency. The expandable shell attached to the roof of a vehicle is also a somewhat valuable idea.
Idea 18: Dragonfly/Damselfly wing structure consider structure for titanium frame	5	8	3	The focus on the dragonfly wing structure seems novel. But the idea lacks substantive details as to what exactly about the wing structure is it that would make the frame structure better.	There is a decent amount of omitted details for which additional research may be required to fill the gap. I think the overall wing pattern may be something that current manufacturing could replicate in a relatively straightforward manner.	This idea alone does not directly address the challenge of wheelchairs going up the stairs. It is unclear how applying the dragonfly wing structure to the frame (hence perhaps making the frame sturdier?) will help wheelchair designs to move up the stairs. The wing structure-inspired frame may be light and durable, which is desirable and meets one of the specified problem constraints, and could be used in combination with other mechanisms that actually solve the main challenge of going up the stairs?
Idea 66: Inspired by tortoises, this wheelchair features a retractable design where the seat, wheels, and footrest fold into a durable, hard shell case. This shell is made from lightweight, impact-resistant materials, ensuring protection during transit. For stair climbing, it utilizes a set of extendable tracks that grip onto steps, smoothly elevating the user.	7	7	7	The focus on the retractability of the shell is novel and distinguishable from a more generic variation of the idea that would say something like "apply the shell-like geometry". For climbing, it mentions the use of extendable tracks that grip on to steps. This part perhaps isn't as novel in light of existing wheelchair ramps and other track-based mechanisms.	Each individual parts of the solution seems fairly feasible given the existing similar mechanisms in other commercial products (shell-like geometry in backpacks or clothing, like extendable tracks in some other products...) However, producing an elegant design solution that combines the different aspects of the idea might be difficult and less feasible.	The retractable design with extendable tracks design would be valuable assuming they work well with different types of stairs while also foldable and durable.
Idea 15: German engineering-iStruct robot as a model for wheelchair	8	5	6	This idea reimagines wheelchair design using a robot, which feels like a provocative idea.	The feasibility assessment would require much more specification of how the iStruct robot would actually work as a wheelchair and various other individual sub-problems related to that to make them into a coherent whole. It feels like there is a substantial amount of uncertainty about what exactly the idea is, and how the iStruct robot would be adapted into a feasible engineering solution.	Robotic design would be useful beyond just stair climbing and has potential to address challenges of any rough terrain types, however it would not be as foldable or lightweight or compact.

Table 8: Additional sample judges' evaluations of design ideas by novelty (N), feasibility (F), and value (V), with rationales.

Rank	Cond.	Prob.	Score	An-Tr?	Idea (Summarized)
T1	BIOSPARK	2	7.86	Y	This bike rack uses a bio-mimetic spring mechanism , inspired from Anura's powerful legs , to stretch (and compress) to accommodate different bike sizes. Its skeletal structure, mimicking frog bones, flexes to absorb road vibrations, protecting bikes. Aerodynamically shaped to reduce drag, it 'leaps' into a compact form when not in use, preserving fuel efficiency.
T2	BIOSPARK	2	7.62	Y	Ultrasonic sensors to adjust its grip on different bike frames automatically, maybe something like cetaceans' echolocation . As a bike is loaded, sensors emit sound waves that measure the frame's size, adjusting the rack's arms for a perfect fit without manual intervention. Its sleek, aerodynamic design mimics a beluga's streamlined shape, minimizing air resistance.
T3	BIOSPARK	2	7.62	Y	The bike rack utilizes smart materials to morph its surface to perfectly fit different bike sizes, inspired by the changing shapes of leaves . It's mounted on the sedan's roof, its leaf-like design folds to minimize air resistance, enhancing aerodynamics. When bikes are mounted, it expands and shapes itself to the contours of each bike, ensuring a secure fit for 16", 20", and 26" frames without adapters, then folds back seamlessly when not in use.
T4	BIOSPARK	2	7.49	Y	Bitterns' balance in currents → A dynamic, flexible support system.
T5	BIOSPARK	2	7.47	Y	Tree frogs' adhesive toe pads → a sticky, yet non-residual, surface-based attachment.
B5	BIOSPARK	1	5.07	Y	Serpentine movement → Wheelchair with a flexible, segmented moving base.
B4	BIOSPARK	2	5.01	Y	Diatoms' protective frustules → Rear-mounted rack w/ adjustable, silica-patterned arms.
B3	BIOSPARK	1	4.93	Y	Froghopper's shock-absorbent exoskeleton → Lightweight & foldable material.
B2	BIOSPARK	2	4.48	N	Bikes are 'locked' in place by a vacuum seal mechanism.
B1	BIOSPARK	1	3.98	Y	Retractable 'legs'; Jointed segments that extend and contract to climb.
T1	Baseline	2	6.80	Y	Aerodynamically shaped bikerack design inspired by how marine mammals like dolphins and whales have evolved flippers and tail fins that optimize their movement in water. Their body shape tends to be more rounded than that of fish but is streamlined for efficient travel. The tail fins (flukes) provide powerful propulsion, while the pectoral flippers are used for steering and stabilization.
T2	Baseline	2	6.49	Y	A series of adjustable, branching arms extend gracefully from the main body, inspired by the natural branching patterns of trees . These arms adjust intuitively to hold various bike frames snugly, replicating the way branches support varying weights and shapes.
T3	Baseline	2	6.49	Y	The rack's "skin" mimics shark denticles , promoting laminar flow and drastically reducing drag. It's not only about fuel efficiency; it's about harmonizing with the very essence of movement.
T4	Baseline	1	6.33	N	Compressing and expanding mechanism for stairs; multiple wheel contacts.
T5	Baseline	2	6.32	Y	Marine polychaete worm-inspired design with tentacle-like grips for bikes.
B5	Baseline	1	2.82	N	Wheelchair self-repair using Memory Shape Alloy.
B4	Baseline	1	2.82	N	Multiple contact points for wheelchair wheels.
B3	Baseline	1	2.71	N	Crystal sensors for damage detection and vital monitoring.
B2	Baseline	2	2.71	N	Top-mounted rack, bike aligned with car for aerodynamics.
B1	Baseline	1	2.62	N	Golden bamboo material for durability and lightweight design.

Table 9: Top-5 and Bottom-5 scoring ideas from each condition, and they exhibit analogical transfer. 'Rank': T1 – T5 correspond to Top-1 – Top-5 and B5 – B1 to Bottom-5 – Bottom-1 rankings; 'Cond.' represents the condition in which the idea was produced; 'Prob.' represents the problem the idea is for (1: the 'wheelchair' problem; 2: the 'bike rack' problem); 'Score' represents the geometric mean of expert-judged novelty, value, and feasibility scores; 'An-Tr?' is a binary value representing the presence of Analogical Transfer; 'Idea' is summarized for conciseness, except for the top-3 ideas that were presented in verbatim. Boldfaced text in the verbatim ideas represents the source and the target in analogical transfer.