Title

An exploration of knowledge-organizing technologies to advance transdisciplinary back pain research

Journal

JOR Spine, 6(4)

ISSN

2572-1143

Authors

Lotz, Jeffrey C

Ropella, Glen

Anderson, Paul

et al.

Publication Date

2023-12-01

DOI

10.1002/jsp2.1300

Peer reviewed

Received: 18 May 2023 | Revised: 2 October 2023 | Accepted: 29 October 2023

DOI: 10.1002/jsp2.1300

PERSPECTIVE

JOR Spine OPEN ACCESS

An exploration of knowledge-organizing technologies to advance transdisciplinary back pain research

Jeffrey C. Lotz[1] | Glen Ropella[2] | Paul Anderson[3] | Qian Yang[4] | Michael A. Hedderich[4] | Jeannie Bailey[1] | C. Anthony Hunt[5]

[1]Department of Orthopaedic Surgery, University of California at San Francisco, San Francisco, California, USA

[2]Tempus Dictum, Inc, Milwaukie, Oregon, USA

[3]Department of Computer Science & Software Engineering, California Polytechnic State University, San Luis Obispo, California, USA

[4]Department of Information Science, Cornell University, Ithaca, New York, USA

[5]Department of Bioengineering & Therapeutic Sciences, University of California at San Francisco, San Francisco, California, USA

Correspondence
Jeffrey C. Lotz, Department of Orthopaedic Surgery, University of California at San Francisco, San Francisco, CA, USA.
Email: jeffrey.lotz@ucsf.edu

Funding information
National Institutes of Health, Grant/Award Number: U19AR076737

Abstract

Chronic low back pain (LBP) is influenced by a broad spectrum of patient-specific factors as codified in domains of the biopsychosocial model (BSM). Operationalizing the BSM into research and clinical care is challenging because most investigators work in silos that concentrate on only one or two BSM domains. Furthermore, the expanding, multidisciplinary nature of BSM research creates practical limitations as to how individual investigators integrate current data into their processes of generating impactful hypotheses. The rapidly advancing field of artificial intelligence (AI) is providing new tools for organizing knowledge, but the practical aspects for how AI may advance LBP research and clinical are beginning to be explored. The goals of the work presented here are to: (1) explore the current capabilities of knowledge integration technologies (large language models (LLM), similarity graphs (SGs), and knowledge graphs (KGs)) to synthesize biomedical literature and depict multimodal relationships reflected in the BSM, and; (2) highlight limitations, implementation details, and future areas of research to improve performance. We demonstrate preliminary evidence that LLMs, like GPT-3, may be useful in helping scientists analyze and distinguish cLBP publications across multiple BSM domains and determine the degree to which the literature supports or contradicts emergent hypotheses. We show that SG representations and KGs enable exploring LBP's literature in novel ways, possibly providing, trans-disciplinary perspectives or insights that are currently difficult, if not infeasible to achieve. The SG approach is automated, simple, and inexpensive to execute, and thereby may be useful for early-phase literature and narrative explorations beyond one's areas of expertise. Likewise, we show that KGs can be constructed using automated pipelines, queried to provide semantic information, and analyzed to explore trans-domain linkages. The examples presented support the feasibility for LBP-tailored AI protocols to organize knowledge and support developing and refining trans-domain hypotheses.

KEYWORDS
artificial intelligence, biopsychosocial model, chronic low back pain, knowledge graphs

1 | INTRODUCTION

Low back pain (LBP) is a challenging clinical problem with growing societal costs. A central dilemma is the absence of a specific LBP cause in most patients, making the treatment process primarily trial and error. Patients typically seek care from a broad range of specialists including pain management, behavioral health, physical therapy, complementary medicine, surgery. However, coordination of care among specialties is poor, leaving the patient to assimilate broad views and navigate multiple treatment options. Furthermore, progress in understanding LBP mechanisms and advancing novel therapies suffers from a lack of coordination between researchers from different disciplines.

The biopsychosocial model (BSM), an influential, patient-centered clinical care paradigm, attempts to codify the landscape of factors that can mediate one's pain experience.[1] While this holistic perspective has proven useful, it is difficult to operationalize because relevant knowledge is spread across multiple scientific disciplines and clinical specialties.[2] Difficulties identifying and assimilating relevant theories and knowledge into an operational framework suitable for identification and treatment of patient sub-groups are compounded by the increasing diversity and expanding volume of published LBP research.

In fact, across the scientific spectrum, research is increasingly specialized, making it difficult to identify innovative connections beyond one's own area of research.[3] Rodriguez-Esteban showed that bias resulting from biomedical siloization coupled with the extraordinary growth of the scientific literature distorts and hinders scientific progress.[4] For those focusing on back pain research, the resulting information silos can lead to narrow and biased syntheses of research observations and underrepresentation of crosstalk among biological, psychological, and social processes that mediate pain experiences over time. Information silos hinder systematic scientific investigation of LBP phenomena and mechanisms. The exponential increase in new, diverse research findings outpaces individual researcher capabilities to assimilate and accurately process information into testable hypotheses.[5] An unintended consequence of that reality is that homogenous research teams underperform diverse ones.[6] It is not surprising that many studies seeking mechanistic explanations of LBP phenomena are narrowly focused within one or a few related domains, without a holistic view of the patient. Langevin[7] and Schmid et al.[8] posit that substantial LBP knowledge, relevant to advancing LBP research and building holistic views, may exist but is "siloed" within the published literature of multiple disciplines, awaiting identification and abductive connection but "out of sight" and beyond the reach of methods and technologies currently employed by most LBP research scientists and clinicians.

Our motivating expectation is that applications of knowledge-organizing technologies will help break down silos and facilitate identifying and accessing existing relevant siloed knowledge. Such applications will inject outside perspectives, ideally "catalytic", that expand opportunities to broaden collaborative networks.[3,5,9] Additional anticipated outcomes include reducing bias, increasing innovation, and improving research translatability to benefit patients and their supporting stakeholders.

In this paper, we use knowledge-organizing technologies to refer to a range of artificial intelligence (AI) algorithms and software that help scientists extract, summarize, and analyze scientific literature. Recent explosive developments in such algorithms (e.g., GPT-4[10]) offer exciting computational capabilities that may help bridge community vocabularies and interests across LBP domains without needing to annotate large document corpora. Recent examples from COVID-19 research demonstrate how biomedical text mining can stimulate scientific novelty via unusual combinations of prior knowledge.[11] However, the envisioned technologies are nascent, and thus risk introducing factual errors and contributing their own types of bias.[12,13] Guardrails will be needed as we harness these technologies to support back pain research.

The goal of this experience report is to explore and describe ways assistive protocols composed of knowledge-organizing technologies, including large language models (LLMs) and knowledge graphs (KGs), might address the types of current LBP problems described by Langevin[7] and Schmid et al.[8] The desired goal is that the assistive protocols will become broadly useful near term to LBP researchers. Although any such tool to assist clinicians would require extensive validation and regulatory approval, it is expected that facilitating upstream LBP research improves clinical outcomes indirectly. With that in mind, we imposed a working guideline: keep the technical learning curve costs small relative to the anticipated benefits of real-time protocol use. We envisioned future protocols operating over large literature collections spanning hundreds-to-tens-of-thousands LBP-related papers and scientific documents (hereafter referred to as a corpus). However, understanding the pitfalls of LLMs and the need to start small, we elected to focus first on a single paper, Schmid et al.[8] because the authors' broader goals align well with our own; to bridge BSM domains in LBP research. The authors point out that LBP research "has identified several pathogenic mechanisms involving biophysical, genetic, social, and psychological contributors." They lament that "research on these different pathomechanisms…is often limited by significant knowledge gaps arising from siloed research within different research disciplines, highlighting the need for cross-disciplinary approaches that have the potential to identify important interactions between different mechanisms contributing to LBP."

Using the Schmid et al. references as our test corpus, we present and discuss seven examples of how one can use knowledge-organizing technology to broadly advance LBP research. All seven examples target crosstalk within a multi-domain literature corpus, albeit with different compositions of the underlying component technologies. We demonstrate preliminary evidence of technical feasibility and utility of these knowledge-organizing technologies and discuss broader sets of opportunities for expanding that utility to corpora spanning all of LBP's phenomena and research domains.

2 | THE POTENTIAL OF KNOWLEDGE ORGANIZING TECHNOLOGIES FOR TRANSDISCIPLINARY BACK PAIN RESEARCH

AI and machine learning (ML) approaches can be understood synoptically as being on a spectrum ranging from mostly manual, with little

assistance from computers, to sophisticated technologically rich protocols. However, for crosstalk among mechanistic explanations in transdisciplinary LBP research, LLMs and KGs are of particular interest because a primary obstacle to crossing LBP domains is the language used to communicate and navigate the domains, compare and contrast intra-domain concepts, establish confidence and credibility for claims and assumptions in and across domains, and so forth. In this regard, LBP research is particularly difficult because it straddles both hard and soft sciences, making intra-disciplinary lexicons much more difficult to translate. The explorations of assistive protocols presented herein are intended to demonstrate how, where, and whether such technologies substantially facilitate one's ability to navigate across multiple domains.

2.1 | The promises and perils of large language models (LLMs)

LLMs consist of a neural network with typically billions-to-trillions of parameters, trained on large quantities of unlabeled text.[14] New LLMs since 2018 (e.g., chatGPT, GPT 4, T5, BART) can generate a wide variety of text analyses and dialogues with an impressive level of fluency out-of-the-box. Through fine-tuning, LLMs become specialized at particular tasks, such as analyzing social determinants of health from clinical notes,[15] answering disease-specific questions based on medical literature,[16,17] simplifying medical concepts and texts for patients,[18] and more.[14] Using end-user-facing LLM interfaces (e.g., Open AI Playground) with and without AI technical training can improve LLM outputs by prepending prompts—textual instructions and examples of their desired interactions—to LLM inputs. For example, when users provide chatGPT with a clinical hypothesis and ask, "Is this hypothesis correct," and chatGPT can answer directly and *sometimes* correctly. In this example, the clinical hypothesis and the question are prompts. Prompts enable non-AI experts to interact directly with and even improve LLM outputs.[14]

Such qualities make LLMs particularly exciting tools for transdisciplinary LBP research and improving knowledge and information transfer among LBP domains. For example, building a bespoke system to evaluate the quality of an emergent LBP research hypothesis can need no more than some examples of good and bad hypotheses.[19] Similarly, building a system that explains LBP-related concepts to out-of-domain scientists or stakeholders can take little to no training data. One might only need to adapt such explanation systems from other medical domains with sets of LBP examples.[18]

However, it would be deeply problematic to consider LLMs—with simple prompts or instructions—as ready to answer one's scientistic questions about back pain, or even to perform seemingly simple tasks such as summarizing a publication's abstract. Using GPT models as examples, some of the most common or severe problems with such a naive approach include:

1. GPT hallucinates: LLMs such as chatGPT frequently provide inaccurate or false information *with a confident tone*. This can be particularly dangerous when LBP researchers use LLM for exploring crosstalk: They are less knowledgeable of out-of-home-specialty literature and, therefore, can be less likely to catch LLM's confident-sounding errors.
2. Crafting effective instructions that get GPT to do what users want it to do ("prompts") is far more difficult than it seems: While prompting GPT *can appear* as easy as instructing a human, crafting effective and robust prompts is challenging.[20] AI researchers analogized improving GPT outputs via prompting as "herding cats": It is possible, but doing it reliably can seem impossible.[21] They demonstrated that natural language prompts are brittle in three ways: (a) Each instruction in the prompt can fix only one GPT error, most but not all the time (e.g., "*Do not make up new findings when summarizing a medical article*." can prevent most but not all GPT-3's hallucinations); (b) An effective instruction, when joined by another effective instruction, can become ineffective (e.g., "*Do not make up new findings when summarizing a medical article. Stay succinct*" can become less effective in preventing GPT-3's hallucination); (c) Finally, these varying levels of effectiveness can change according to the clinical area and to the document.
3. Even good prompts cannot improve LLM reliably: How a prompt or a prompt strategy directly impacts model outputs, and how prompts modify LLMs' billions of parameters during re-training, are both active areas of NLP research.[14,22]

Given these pitfalls, harnessing LLMs for transdisciplinary LBP research requires extra caution and vigilance. Thankfully, results of recent AI research offer a few silver linings.

4. "AI chaining"—combining multiple LLM models to improve LLM output quality: While LLMs can make factual errors and even hallucinate, new techniques have emerged to fact-check LLM outputs to ensure reliable use for healthcare. For example, Yang et al. combine a treatment-outcome-prediction AI with a fact-checking GPT model in clinical decision support systems.[23]
5. Combining LLM and KGs to ensure LLM output factualness: A knowledge graph represents a network of real-world entities—that is, objects, events, situations, or concepts—and illustrates the relationships between them. Clinical researchers have built various LBP-related knowledge graphs, for example, encoding the LBP concepts and knowledge and their interrelations with a focus on Virtual Reality rehabilitation[24] and central sensitization.[25] Unlike LLMs, whose knowledge is implicit and difficult to scrutinize, a KG explicitly visualizes the concepts and knowledge it learns from medical texts. Its correctness is much easier to improve and assess. As a result, AI researchers have started combining KG and LLM to improve the factualness of LLM outputs.[26]
6. Hypothesis generation following LLM fine-tuning with cross-validation using expert opinion: Banker et al.[5] describe a LLM for generating social psychology hypotheses. They fine-tune their model in two stages, where the second stage includes thousands of published abstracts spanning 55 years. Social psychology experts rated model-generated and human-generated hypotheses

to be equivalent on the dimensions of clarity, originality, and impact. The authors posit that the approach can empower social psychology researchers to manage their exponentially expanding social psychology literature and help them see (discover) cross-connections (crosstalk) among domains beyond their own. They argue that their LLM takes an inclusive, unbiased viewpoint by learning from a corpus having a scope far wider than is feasible for any individual scientist to achieve. They envision that because their LLM can leverage inter-topic connections to generate novel and specific hypotheses, it may prove particularly helpful in domain regions in which research findings are scarce.

2.2 | Knowledge graphs (KGs)

There are many use cases of knowledge-organizing technologies where the domain expert will not accept the output, guidance, or conclusion unless it can be explained. Semantics are about associating meaning with data. In many cases, it is not enough for an AI system to utilize only syntax to complete a task. KGs aim to integrate knowledge and data at scale from scientific advancements across many fields, including semantic web, databases, knowledge representations, natural language, and ML.[27]

KGs are a form of structured knowledge representing facts as entities, relationships, and semantic descriptions. KGs can encode latent structures that are difficult to program into a more traditional database schema. These structures are then further enhanced semantically with the layering of ontological knowledge bases. Ontologies describe taxonomies of concepts and relations for a certain domain and are commonly constructed manually or semi-automatically. A KG then organizes data ("facts") according to that ontology, providing a graph that can be "walked" in different modes, for example, for interactive discovery or algorithmic inference. At a high level, a KG is a collection of nodes and relations (also known as links or edges; Figure 1).

Research into techniques to infuse KGs and domain information into deep neural models has expanded.[28] Recent advances can be arranged from shallow to deep.[29] In shallow infusion, the response/output from a deep learning system is reconstructed, transformed, or interpreted using domain knowledge within a KG. Deep infusion by contrast couples the representation learned by deep learning systems with KGs.[30]

KGs may be constructed manually or built automatically from unstructured text and semi-structured data, and structured data. In practice, such KGs are often built using a combination of human-in-the-loop supervision and semi-automatic protocols. KGs have the potential to impact downstream AI applications by infusing knowledge-aware models with reasoning. Examples include question–answering, recommendation, digital health, and search. In addition, KGs can be examined and queried as their primary purpose. Examples include information retrieval, community detection, inferring new knowledge, and analytics. Such use cases are broadly referred to as graph data science.

KGs are well suited for transdisciplinary LBP research as KG-based systems make it possible to find facts or evidence across multiple domains.[31] For example, Xie and colleagues integrated patient-specific Chinese medicine through KGs[32] while Shang et al. explored personalized medication recommendations by creating visit-level representations of patients from Electronic Health Record data.[33] KGs also assist in hypothesis formation and, importantly, can do so across disciplinary boundaries. Forming hypotheses that enable breakthroughs leading to new knowledge can be done in two primary ways: (1) using graph algorithms and (2) logical reasoning. Link and edge prediction comprises a broad category of graph algorithms that identify new edges from the structure of the graph that are likely but missing.[34] Edge prediction algorithms can be interpreted as hypothesis generation and often are accompanied by a confidence measure. Using formalized reasoning systems one may infer new explicit edges that are otherwise implicit in KGs. That new inferred information can then be added to the KG. As a simplistic example, we can infer that all proteins are nitrogen-containing compounds and add that explicit fact to the KG.

Given the broad scope of LBP research, one might anticipate LBP KGs having billions of assertions. That size presents computational challenges and remains an active area of research.[35] In addition to scaling the computational requirements, expert curation efforts must also be scaled appropriately. While expert curation remains the gold standard, AI-assisted semi-automated-to-automated approaches have been developed. Data-driven KG construction can involve LLM-assisted analyses of large-scale datasets. For example, the Semantic MEDLINE database uses natural language processing to extract information from documents to create KGs.[36] Furthermore, there are thousands of public, biomedical databases, ontologies, and KGs[37] that have been created with different technologies for different purposes. Such complications are compounded as many KG efforts are not updated regularly resulting in leaving out-of-date information. For these reasons, as well as data quality issues, integration remains an important challenge.

2.3 | Similarity graphs (SGs)

KGs are intelligently designed downstream products of fine-grained protocols. By contrast, the SG protocol we describe here is relatively short and is produced further upstream. An SG organizes data taken directly from tokenizers that break down sequential natural language text into smaller units called tokens (e.g., word roots). These tokens are then used by AI components like GPT to further process the data (as done in Sections 3.2 and 3.4 below). The SG protocol compares the tokens more directly, relying less on the intelligence programmed into the AI components. While the tokenizer-encoder used is structured according to a particular LLM (GPT-3.5 for the examples described herein) to take advantage of some of its programmed intelligence, the protocol does not rely on *inference* from the LLM, which makes it more agnostic. It is a more blinded attempt at obtaining a "Free Lunch."[38]

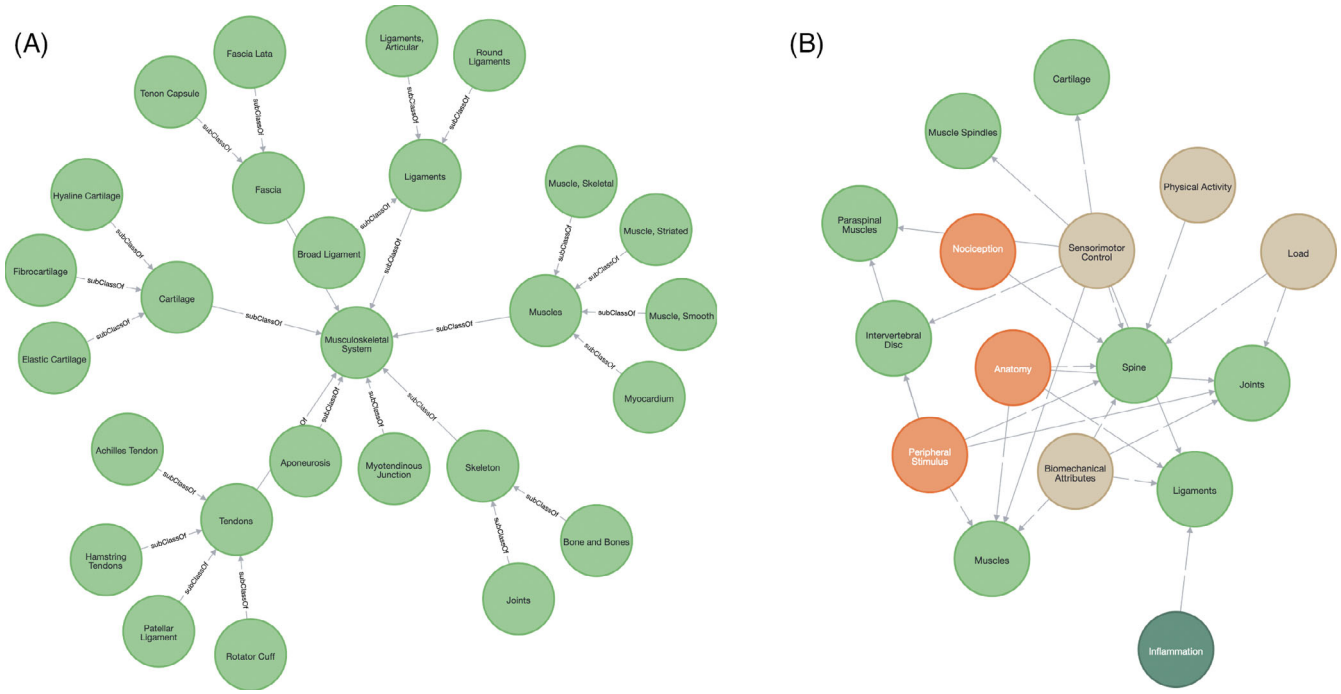

FIGURE 1 Illustrative example of layering knowledge about the Musculoskeletal System defined in the MESH ontology (A) with chronic pain domain descriptions to create a knowledge graph (B). A subset of the available relations (links/edges) and nodes are shown for clarity.

2.4 | Participatory modeling

To help set LLMs and KGs on the manual-to-automatic spectrum, it is useful to consider participatory modeling.[39,40] Participatory modeling is a mostly human-in-the-loop protocol, executed by humans to survey human expert opinions. Related is another consensus-based evidence technique called the Delphi method that also leverages the collective experience of domain experts.[41] While these approaches are valuable tools to collect and summarize expert opinions on complex topics, they may suffer from participant anonymity leading to a lack of accountability and potential for hasty decisions, giving the illusion of precision.

3 | APPLICATIONS OF AI-TECHNOLOGIES TO LBP DOMAIN CROSSTALK

In this section, we provide small-scale examples to illustrate the utility and potential for utilizing KGs and LLMs to assist in organizing, exploring, and transforming knowledge relevant to LBP domain crosstalk. We describe specific scenarios and explore results from AI-assisted usage patterns from the perspective of hypothetical users. The protocols described are intentionally flexible and can be composed in many valid configurations. We detail concrete examples and provide additional commentary on their limitations and performance.

3.1 | Data and processing

The primary data used in our examples is literature in PDF format. To work with the data, a robust tool for converting PDF to UTF-8 text is necessary. There are several such tools that provide intelligent extraction. However, to highlight that we are exploring fully automatic, purpose-agnostic, protocols, we used mutool.[42] As with the training data for LLMs and other NLP projects (e.g., processing unstructured notes from Electronic Health Records), a variety of other documents can be processed and integrated with the reference articles.

To probe the utility of LLMs, KGs, and SGs, we selected references from a review of the interactions between three of the LBP domains: central (neuroplastic), psychological, and biomechanical.[8] The selection of a small reference dataset was deliberate as it allows us to compare the results from AI-assisted protocols to parallel and more traditional manual efforts. From Schmid et al, we selected 19-paper (Schmid-19) and 64-paper (Schmid-64) subsets (we were unable to get a PDF for one reference) plus Schmid et al. itself (Schmid-65) (Table 1). The authors' focus is LBP pathophysiology and connecting methods from neuroscience and biomechanics research. They describe a cross-disciplinary approach that may identify different motor control phenotypes which, downstream, lead to better treatment options. References include broad and focused reviews, research and clinical trial reports, and meta-analyses of clinical trial reports. The paper and several references discuss mechanistic hypotheses and plausible causal interactions.

TABLE 1 References selected from Schmid.[8]

Schmid-19	Schmid-64[a]	
13. Clays et al. (2007)[65]	1. Wu et al. (2020)[78]	33. Lim et al. (2015)[71]
19. Gombatto et al. (2017)[66]	2. Maher et al. (2017)[79]	34. Elgueta-Cancino et al. (2018)[99]
20. Simonet et al. (2020)[67]	3. Vlaeyen et al. (2018)[80]	35. Sutherling et al. (1992)[100]
21. van Dieen et al. (2003)[52]	4. Rubinstein et al. (2013)[81]	36. Ejaz et al. (2015)[101]
22. Marras et al. (2001)[68]	5. van Middelkoop et al. (2011)[82]	37. Roux et al. (2018)[51]
26. Claeys et al. (2011)[59]	6. Hartvigsen et al. (2018)[1]	38. Beaudette et al. (2016)[72]
29. Tsao et al. 2011)[69]	7. Goubert et al. (2016)[83]	39. Martinez-Calderon et al. (2019)[102]
32. Eto et al. (2011)[70]	8. Knezevic et al. (2021)[84]	40. Ranger et al. (2020)[73]
33. Lim et al. (2015)[71]	9. Brinjikji et al. (2015)[85]	41. Leeuw et al. (2007a)[103]
37. Roux et al. (2018)[51]	10. Hodges and Tucker (2011)[86]	42. Zale et al. (2013)[104]
38. Beaudette et al. (2016)[72]	11. Tsao et al. (2008)[87]	43. Wertli et al. (2014)[47]
40. Ranger et al. (2020)[73]	12. van Dieen et al. (2019)[88]	44. Klyne et al. (2020)[74]
44. Klyne et al. (2020)[74]	13. Clays et al. (2007)[59]	45. Schweinhardt (2019)[105]
46. Matheve et al. (2019)[49]	14. Langevin (2021)[7]	46. Matheve et al. (2019)[49]
47. Geisser et al. (2004)[54]	15. Meier, et al. (2019)[89]	47. Geisser et al. (2004)[54]
48. Knechtle et al. (2021)[48]	16. van Dieen et al. (2017)[90]	48. Knechtle et al. (2021)[48]
51. Houben et al. (2005)[75]	17. Hodges and Smeets (2015)[91]	49. Christe et al. (2021)[55]
52. Leeuw et al. (2007b)[76]	18. Christe, et al. (2016)[92]	50. LeDoux and Hofmann (2018)[106]
54. Meier et al. (2018)[77]	19. Gombatto et al. (2017)[66]	51. Houben et al. (2005)[75]
	20. Simonet et al. (2020)[67]	52. Leeuw et al. (2007b)[76]
	21. van Dieen, et al. (2003)[52]	53. Pfingsten et al. (2000)[107]
	22. Marras et al. (2004)[53]	54. Meier et al. (2018)[77]
	23. Marras et al. (2001)[68]	55. Lundberg et al. (2011)[108]
	24. MacDonald et al. (2009)[93]	57. Julian (2011)[109]
	25. Prins et al. (2018)[94]	58. Kroenke et al. (2001)[110]
	26. Claeys et al. (2011)[59]	59. Schmid et al. (2017)[111]
	27. Hodges (2013)[95]	60. Niggli (2020)[112]
	28. Flor et al. (1997)[96]	61. Zemp et al. (2014)[113]
	29. Tsao et al. (2011)[69]	62. Connolly (2021)[114]
	30. Riemann and Lephart (2002)[97]	63. Nelson and Chen (2008)[115]
	31. Bushnell et al. (1999)[98]	64. Weerakkody et al. (2007)[116]
	32. Eto et al. (2011)[70]	65. Boucher et al. (2015)[117]

[a]Highlighted references are those where expert classification differed. Green indicates where GPT-3.5 classification was primarily psychology, and blue represents GPT-3.5 classification was primarily biomechanics.

In addition to above corpora, we incorporated language for six LBP domains adopted from a recent umbrella review[43] and illustrated in Figure 2. These six domains are described below. The full text used as input is included in Supporting Information.

1. *Psychological*—constructs that capture patient beliefs based on prior experiences and future expectations, affect states and traits, attitudes, personality traits, behaviors, coping styles and resources, attention styles toward pain, self-efficacy, and others.
2. *Biomechanical*—functions of the spine: to protect the spinal cord, to support upper body loads, and to facilitate trunk mobility, enabled by a complex integration of passive (vertebrae, discs, facet joints, ligaments) and active (muscles) tissues plus the neuromuscular control system.
3. *Societal*—social context: the quality of an individual's social relationships and society's (cultural, family, therapeutic) responses to their pain.
4. *Systemic*—systemic factors and comorbidities including nutritional status, metabolic diseases, immunological conditions, endocrine disorders, and sleep disorders.
5. *Familial*—genetic and environmental conditions shared in families involving cellular mechanisms that link inflammation, peripheral sensitization, and pain, modified by complex interactions between the genome and environment.

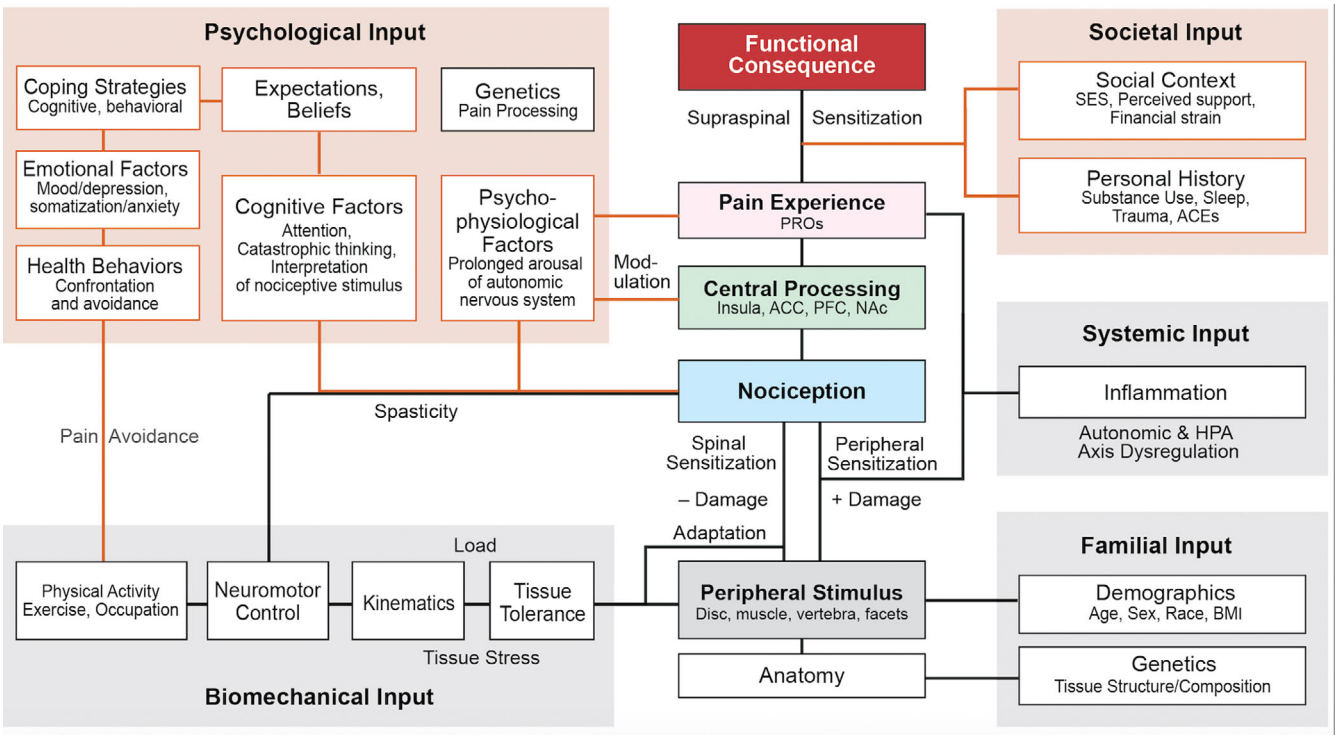

FIGURE 2 LBP theoretical scheme reported in Chau et al.[43] The scheme specifies plausible, bi-directional interactions between factors within clusters and between those factors and central features. A peripheral stimulus source is included between anatomy and nociception. ACE, agreeableness, conscientiousness, extraversion; ACC, anterior cingulate cortex; ANS, autonomic nervous system; HRV, heart rate variability; HPA, hypothalamic–pituitary–adrenal; NAc, nucleus accumbens; PTSD, post-traumatic stress disorder; PFC, prefrontal cortex; SES, socioeconomic status.

6. *Central*—bidirectional pain signals between the periphery and brain, transformed into physiological, cognitive, affective, and behavioral responses.

The complexity of LBP from the perspective of the above domains requires a broad understanding of the process of hypothesis generation. Therefore, for illustrative purposes, we also incorporate example hypotheses intended simply to showcase their role in AI-assisted protocols (these hypotheses were generated for demonstration purposes and do not reflect personal perspectives or opinions of the authors). LBP research hypotheses often posit mechanistic explanations for causes and treatment effects at the population level. By contrast, LBP clinical hypotheses are composite, and focus on phenotypic classes or individual patients. They are often represented as if-then decision trees. In either case, abduction from anomalous circumstances triggers the formation of novel hypotheses. Such abduction often requires crossing domains and collaboration between otherwise siloed domain experts. Inter-domain hypothesis formulation leads to usage patterns of literature search, summarization, classification, and analysis. Similarly, there are explicit metrics for evaluating the strength of hypotheses. Two options include either a combination of validity, significance, and feasibility, or a combination of validity, significance, clinical relevance, and feasibility.[44] Such evaluations trigger several usage patterns for literature and domain navigation, for example, see Section 3.2.1. In addition to this broader understanding of hypotheses, advancing LBP science necessarily challenges the concept of a domain or silo. The formulation of an LBP model helps to demonstrate that domains are often vague and abstract.[45,46] The protocols for navigating crosstalk, formulating and evaluating hypotheses, and assessing corpora use statements of hypotheses and domain knowledge as input. However, if used iteratively, they can also help quantify the coherence and completeness of a specific hypothesis and its domain descriptions. Variants of the knowledge-organizing technologies described below are expected to facilitate both activities.

3.2 | Downstream applications of large language models: Assisting new research hypothesis exploration using GPT-3

In this section, we demonstrate using GPT-3.5 to enhance and facilitate domain crosstalk when formulating explicit hypotheses. Specifically, we demonstrate a prototypal GPT-assisted hypothesis-formulation protocol through a concrete example: A scientist (e.g., a psychologist or biomechanical engineer), is exploring whether a current working hypothesis related to fear avoidance is a sound basis for a new experiment. The fear-avoidance model[47] suggests that when an individual perceives pain as threatening, they respond with pain-related fear (psychological factors) and avoidance behavior, which in turn results in declines in physical functioning (biomechanical mechanisms).

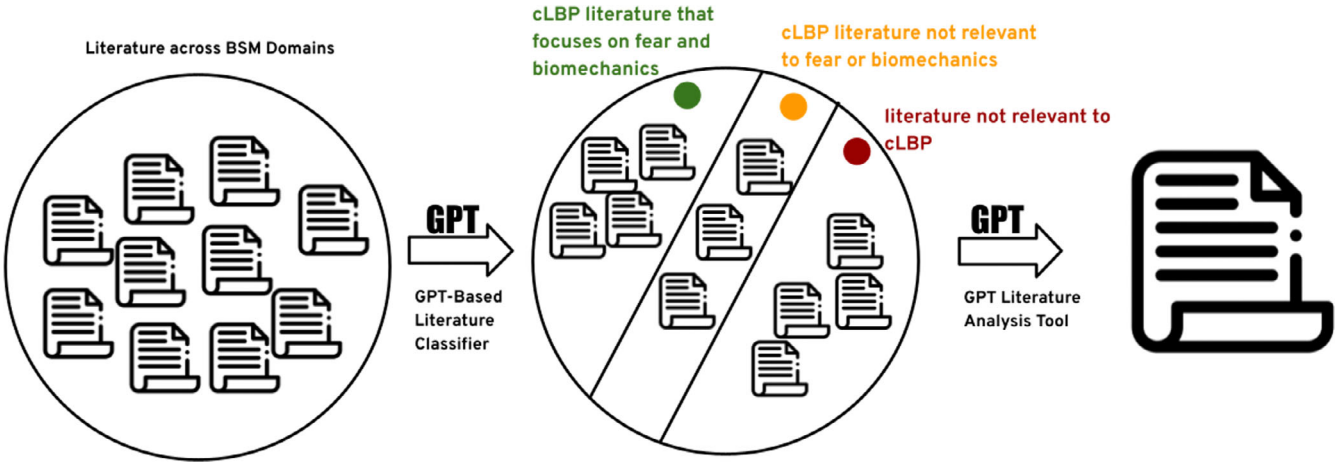

FIGURE 3 Example hypothesis-formulation protocol where large language models such as GPT can assist scientists in harnessing prior studies across domain boundaries.

We envision a customizable GPT-based toolkit (herein, a collection of adaptable documents and IT tools to inform and facilitate successful completion of a research protocol) that helps explore and formulate problem-specific, domain-bridging hypotheses:

1. Identify prior studies that bridge the two targeted domains of LBP.

   Highlight studies that support or contradict the working hypothesis from any relevant domain.

2. Facilitate scrutinizing GPT's synthesis of prior studies by highlighting the PICO elements (Patient population, Intervention, Comparator, Outcome (Aslam and Emmanuel 2010)) of each study that GPT claims supports or contradicts the current working hypothesis.

We developed our toolkit prototype using GPT-3.5 (text-davinci 03; simply GPT-3 hereafter). Below, we describe the toolkit's performance, illustrating that it can accomplish all three tasks to various extents, albeit imperfectly. We then describe the pitfalls and lessons learned during this exercise. We demonstrate that: (1) without any LBP-specific training data, off-the-shelf GPT models are already useful, but they are not (yet) trustworthy in any of the above three tasks; and (2) via fine-tuning, prompting, and chaining, GPT models can become more factual, accurate, and capable in these tasks. To wrap up, we highlight additional near-term opportunities we believe will prove useful for leveraging GPT for LBP crosstalk, including (1) sharing effective GPT prompts across LBP stakeholder communities; (2) building reusable AI chains for frequently explored cross-domain connections; and (3) making LBP knowledge more accessible to patients and the public.

3.2.1 | Identifying prior studies that bridge multiple LBP domains (Example 1)

Without any cLBP-specific training data, our GPT-based toolkit can help identify the prior studies that bridge two or more LBP domains (psychological, biomechanical, and supraspinal) from prior literature. To do so, one only needs to provide the PubMed ID(s) of the corpus. Among these papers, the toolkit will identify those that address different cLBP mechanisms (Figure 3).

Here is how our toolkit performs and some lessons learned. Two domain expert authors (jcl, jfb) manually categorized all references from Schmid, et al. (2021) as to whether a paper addresses only fear factors, only biomechanical factors, both, or neither. We excluded six papers from the Schmid-65 corpus because the domain experts' categorizations disagreed (13, 45, 56, 57, 63 and 64; Table 1). For each of the 4 categories, we reserved one paper as an example for use by the GPT model (4, 6, 11, and 39; Table 1).

When evaluating on the resulting 55-paper dataset regarding psychological factors and biomechanical mechanisms, the toolkit achieved an accuracy of 60%, meaning that it correctly identified the domain in more than half of the cases without being specifically tuned for those two domains. The toolkit was especially good in identifying publications that solely covered either psychological factors or biomechanical mechanisms (Table 2 and Figure 4).

To achieve these results, the toolkit first extracts each paper's abstract from PubMed based on the paper ID. It then feeds each abstract into the GPT model and generates the cLBP domain the paper covers. Our toolkit feeds the abstract to GPT in a very specific, elaborate format that we carefully crafted and tested. This "prompt" goes as follows (lightly edited for readability, see Appendix 1 for the verbatim prompt used):

```
Given the following four categories and one biomedical paper abstract, classify the abstract as one of the four categories. Please be strict when determining whether a paper discussed a topic.
```

Categories:

class1: *papers that only discussed biomechanical factors (including physical activity, sensorimotor control, load, and tissue tolerance).*

class2: *papers that only discussed psychological factors (including fear, beliefs, coping skills, and affective state).*

class3: *papers that discussed both psychological factors (including fear, beliefs, coping skills, and affective state) and biomechanical factors*

(including physical activity, sensorimotor control, load, and tissue tolerance).

class4: papers that did not discuss psychological factors (including fear, beliefs, coping skills, and affective state) or biomechanical factors (including physical activity, sensorimotor control, load, and tissue tolerance).

Abstract 1:

The abstract of an example paper that discussed only biomechanical factors of cLBP is included here.

Label 1: class1

Abstract 2:

The abstract of an example paper that discussed only psychological factors is included here.

Label 2: class2…

In total, we provided 4 example papers, covering all four categories (a paper addresses only biomechanics, only fear, both, or neither).

Abstract 5:

```
Here is the abstract of the paper that the toolkit is
trying to classify based on the cLBP processes it
covers.

Label 5
```

By leaving the last line "Label: ____" empty, GPT will automatically try to fill the empty gap with its prediction of the right category and thus classify the paper's domain coverage based on all the examples provided. As shown above, this prompt includes both the request for GPT to identify publications that involve different cLBP domains (psychological, biomechanical, and supraspinal) and examples of paper abstracts for each domain.

This combination outperforms other prompt templates. Not providing an example abstract for each category lowers the accuracy to 29% which is only slightly above the accuracy of randomly guessing. How exactly the task is formulated and how the categories are described to GPT can also have an influence. Providing descriptive terms for each category (like "including `fear, beliefs, coping skills, and affective state`") can improve the accuracy further. One challenge is the fact that asking GPT multiple times the same question can result in slightly different answers. A prompt needs

TABLE 2 GPT-3.5's preliminary performance on a Schmid 55-publication test set in identifying the focus of prior work relevant to cLBP. Performance is measured using an F1-score[a], a metric that takes into account both the number of correctly identified papers and how many were missed.

Identifying cLBP publications that focus	GPT toolkit without examples for categories (F1-score)	GPT toolkit with one example per category (F1-score)
Only on biomechanics	0.15	0.70
Only on psychology	0	0.70
On both psychology and biomechanics	0.31	0.36
Neither psychology nor biomechanics	0.46	0.46

[a]The F1-score is the dividend of precision and recall (where recall is the same as sensitivity) and is utilized to assess diagnostic performance of the prediction algorithm. The F1 score can vary from 0 to 1.

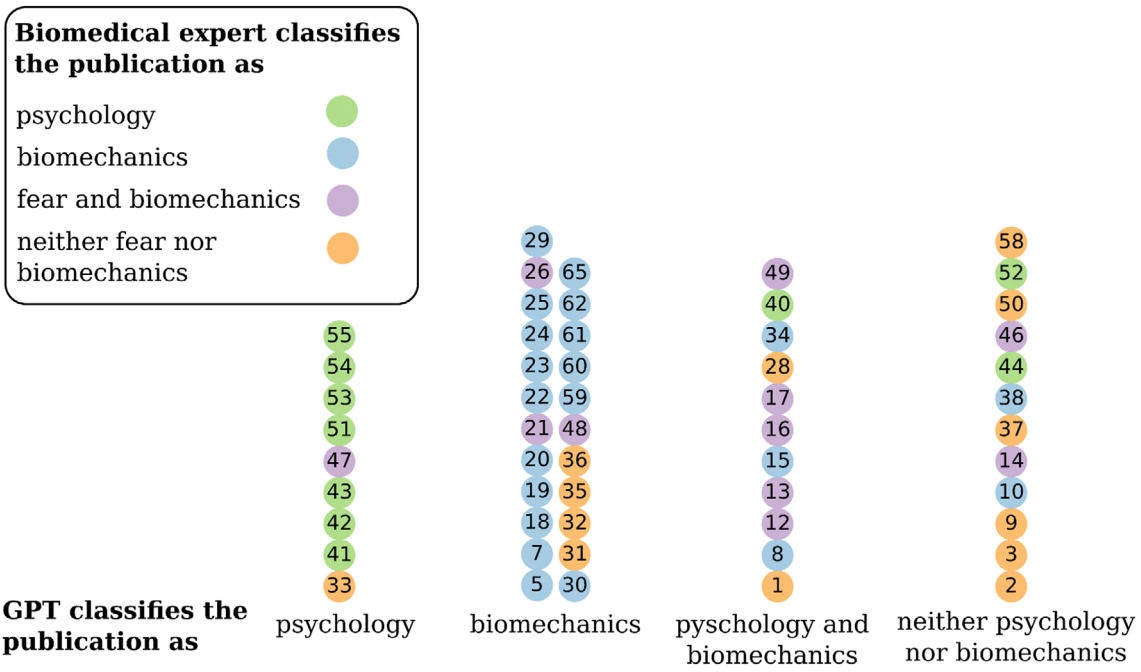

FIGURE 4 GTP versus expert classification of the Schmid-64 corpus. The nine references highlighted in Table 1 are not shown because the expert's classifications of those nine did not agree. With only one example per category, GPT-3.5 can identify the main focus of publications on LBP topics "psychology and biomechanics" with an accuracy of 60%.

to therefore be tested extensively to obtain statistically significant results.

3.2.2 | Identifying prior studies that support or contradict an emergent experiment hypothesis (Example 2)

After collecting a transdisciplinary set of prior studies, our GPT-based toolkit can help highlight the degree to which a study supports or contradicts the scientist's emergent hypothesis. We demonstrate this capability through the following example. Imagine a scientist who is considering the following hypothesis and seeks to identify supporting or counter-evidence in the literature.

> **Hypothesis 1.** Pain-related fear is associated with reduced lumbar flexion.

Our toolkit can identify whether a publication supports or contradicts each correlational or causal argument in the hypotheses, based on the publication' abstract. From the full dataset of 64-cLBP publications (Table 1), our toolkit identified 10 papers that supported the hypothesis and 2 papers that contradicted it. Knechtle et al.[48] provides supporting evidence, by finding in their study a negative relationship between measures of pain-related fear and lumbar spine flexion angles during lifting. The evidence in Matheve et al.[49] which our toolkit classified as contradicting, on the other hand, suggests a more restrictive hypothesis. They show that lumbar motion is not predicted by general measures of pain-related fear but only by task-specific ones.

Consistent with most AI research, the toolkit's hypothesis-testing capability fluctuates depending on how the hypothesis and the request are framed. In the case of our exploration, the toolkit is most accurate when using the AI chaining technique and the following request wording. First, the system requests GPT to identify the PICO elements of the study based on the paper's abstract:

```
Here is an example of doing question answering, given {abstract} the answer to {population} is {P}, following this example, given {new abstract} please tell me {population}. Please try to be pretty succinct and focused in your answer and include necessary numbers if you can. Keep your answers to up to 35 words.
```

Next, the system requests GPT to summarize the paper abstract based on the PICO elements. We found that summaries generated via explicitly identified PICO elements offer more accurate and complete information about the study.

```
Here is an example summarizing core information from context and PICO information. Given {PICO}, the summarization sentence will be {summary example}, please following the above example, given another set of PICO information {PICO}, please summarize the core information. Please be precise and coherent in writing a summary of all given information.
```

Finally, the system requests GPT to verify whether the study supports or contradicts the scientist's hypothesis. "{hypothesis}" represents the hypothesis wording the scientist provided, for example, "Pain-related fear is associated with reduced lumbar flexion."

```
Given the following statement, please analyze and then categorize your response according to the existing evidence or context. Your answer should strictly be one of the following: 'supports,' 'contradicts,' 'disregards the context' or 'maintains a neutral stance.' However, if you can definitively choose, please refrain from choosing 'maintains a neutral stance' unless it is absolutely necessary due to lack of sufficient evidence or ambiguity in the statement. Please, let's adhere to these instructions strictly. Thank you.
```

Importantly, this series of requests happen automatically behind the scene. When using the toolkit, scientists only need to provide the PubMed IDs of the publications of their interest, and the toolkit will automatically identify publications that include supporting and counter-evidence.

3.2.3 | Assisting scientists in scrutinizing GPT's synthesis of prior studies (Example 3)

AI systems are known to make unpredictable errors, and so are GPT-based systems. It is therefore crucial that our toolkit's protocol is explainable and results attestable for scientists.

Our toolkit uses the AI chaining technique, it, therefore, can allow scientists to easily identify in which step the GPT model made an error, preventing over-trust and misuse. As shown in the prompt above, when our toolkit works to identify prior studies that support or contradict an experimental hypothesis, it first identifies the PICO model elements of these prior studies and then summarizes the study findings with an eye on these PICO elements. For example, when processing Buchthal and Rosenfalck,[50] our toolkit can automatically extract the following:

1. Patient population: `22 patients with nonspecific chronic low back pain`
2. Intervention: `Trunk flexion-extension tasks performed under three experimental conditions: innocuous heat, noxious stimulation with low pain expectation, and noxious stimulation with high pain expectation.`
3. Comparator Intervention/control group: `a group of 22 age- and sex-matched participants without low back pain.`
4. Outcome: `Pain expectations modulate neuromuscular responses in patients with chronic LBP more than in controls. High pain expectations increase low back extensor activity during full flexion. (22 patients and 22 controls, F2,845317.5;` $p < 0.001$)

Based on the PICO elements, the toolkit further generates a short summary of the paper. For the Buchthal and Rosenfalck paper, the

GPT model generated the following summary containing several levels of detail.

> This study aimed to determine whether pain expectations could modulate neuromuscular responses to experimental low back pain in patients with chronic LBP compared with controls. The study utilized surface electromyography and lumbopelvic kinematic variables to observe muscle activation during trunk flexion-extension tasks under three experimental conditions. The outcome shows that high pain expectations increase low back extensor activity during full flexion in patients with chronic LBP more than in controls. This suggests that repeated exposure to pain generates rigid and less variable patterns of muscle activation in patients with chronic LBP, which attenuate their response to pain expectations. High levels of pain catastrophizing are associated with higher lumbar muscle activity in full flexion in patients with chronic LBP.

Provided with such a summary, one can more easily assess the trustworthiness of each prior study along with contrary evidence while refining one's working hypothesis. They can also easily identify in which step the GPT model may have introduced an error.

3.2.4 | Summary of downstream applications of LLMs

To summarize, our exploration shows that GPT-3 can be useful in helping scientists analyze and distinguish LBP publications across psychological and biomechanical literature. It is worth noting that given the limited scale of this illustration, further work is necessary to conclude GPT-3 performance on this task. Nonetheless, the results are sufficiently interesting and informative to encourage further explorations.

3.3 | Upstream applications of large language models: Envisioning, prototyping and evaluating similarity graphs (SGs)

As noted above, the scientific literature dealing directly with LBP spans several scientific and clinical domains and is expansive: between 2011 and 2020 it included 27 968 papers.[2] We speculated it should be feasible to employ capabilities of current LLMs, such as GPT-3, to create representations (network projections) of such a large corpus (nodes) that allows domain experts, focused on specific questions or problems, to explore that landscape in novel ways possibly achieving holistic, trans-disciplinary perspectives or insights that are currently difficult, if not infeasible to achieve. At a higher resolution, we envisioned that each node in the network projection would represent a single paper, a meta-data document, a clinical report, or a use case-specific body of text. However, a simple yet essential requirement must be met to enable such representations to develop into reliable assets for scientific information evaluation and retrieval. The relative relationships among entities within representations (especially those most familiar to domain experts) must resonate with corresponding relationships within the domain expert's own conceptual model of LBP.

Domain experts (researchers, clinicians, etc.) have conceptual models wherein they organize knowledge within and surrounding their own domains. Such conceptual models are slowly constructed over time, culminating in expertise that is both tacit (embodied) and explicitly expressible. AI and ML models that operate over human languages require similar representations. They include not only natural languages but also programming languages like Python and any kind of sequential, step-by-step procedure. The representation of those languages is made consumable by large-scale computer algorithms through embeddings. A sequential document, such as a research paper, can be transformed into a vector in the *embedding space*, which is a mathematical representation that allows for quantitative comparisons between documents.

An embedding can be generated in a variety of ways. In this work, we use OpenAI's GPT Embeddings API. We extract the text (in full) from the published PDF of each document. We convert it into a high dimensional vector (1536 dimensions in this case, discussed in the Supporting Information), which can then be compared to any other vector with the same number of dimensions. The metric distance between any two vectors provides a kind of similarity measure. Ideally, similarities among vectors will map to, for example, semantic similarities among documents. Technically, the resulting measure is called "semantic" similarity. An agnostic approach to evaluating the technology for various usage patterns requires us to focus on the protocol without taking the semantic similarity metaphor too seriously. Metric distance similarity is somewhat distinct from set-based overlap similarities like Jaccard (Salvatore et al., 2020) but, viewed synoptically, can serve similar purposes depending on the assistive technologies used in the protocol.

3.3.1 | From embeddings to similarity graphs (Example 4)

The graphs presented in this section de-emphasize intra-document text content (data) and focus instead on inter-document semantic content. The graphs organize data taken directly from a tokenizer-encoder, which breaks down sequential natural language text into tokens (e.g., word roots) and embeds the sequence into a vector space defined by the AI. The encodings can then be used by components like GPT-3 to further process the data, as done in our other Examples. By contrast, this SG protocol compares the encodings further upstream, relying less on the intelligence programmed into the AI components. The tokenizer and encoder used, here, are structured according to GPT-3 to take advantage of some of its programmed intelligence. However, the protocol does not rely on *inference* from GPT-3, which makes it more agnostic. It is a more blinded attempt at obtaining a "Free Lunch."[38]

The following list identifies key tasks within our five-step protocol. Additional details and code used are provided in the Supporting Information.

1. Extract UTF-8 text from each reference PDF in the target corpus.
2. Split each reference text into sequential blocks small enough for the tokenizer-encoder.
3. Obtain a vector representation of each block.
4. Calculate a naïve centroid of the reference's blocks
5. Calculate the dot product between each pair of references (called "cosine similarity").

The vector space in which each reference's centroid is embedded is structured such that the dot product in step 5 provides a result between zero and one. These vector distances provide a measure of similarity that can be analyzed and visualized in any number of ways. Such similarity is a reduction of the high dimensional space into a single dimension (similarity or pairwise distance). Other reduction protocols that can be used. However, the analysis of any high-dimensional representation requires a reduction that facilitates conceptual understanding by domain experts. Visualization and presentation of such reductions is an active area of research beyond the scope of this work. Reducing to the single similarity dimension and rendering the result as a weighted, undirected graph meets the needs of this early-stage exploratory study and face validation exercise. Continued protocol use will require accumulating validation evidence. For example, we assume the gist of each reference (each corpus document) is represented in the step 4 vector. However, because the centroid is lossy, it will require future validation to ensure the gist of the paper is adequately represented for domain expert assessment.

In an SG, nodes represent documents, and each node's edge represents a pairwise similarity value. Pairwise similarities from step 5 are organized into a square edge list, like a correlation matrix, with every document as a row and column. That list is imported into Gephi (https://gephi.org/) and rendered as a network using ForceAtlas-2, a layout algorithm. The simple SG example presented in Figure 5 served as our first face validation test. It comprises the Schmid-19 references (Table 1). Edge weights act somewhat like springs to bring similar nodes together. Edge weights and node degree (the number of edges associated with that node) are used by ForceAtlas-2 to determine the relative locations of nodes within an SG; and nodes tend to cluster together according to their edge weights and degree. We expect that two references that are close together in *embedding space* may also yield nodes close together in an SG. However, it is crucial to note that such SG visualizations can be misleading. The layout algorithm is stochastic, and each execution can render the graph in a somewhat different way.

Within an SG, papers represented by more distant node pairs, for example, 37c[51] and 48c[48] in Figure 5, are assumed to exhibit the least semantic similarity. That assumption is supported by the locations of, for example, 37c and 48c in two right columns in Figure 4. Papers represented by more closely spaced node pairs, for example, 21c[52] and 22c[53] in Figure 5, are assumed to exhibit considerable semantic similarity. That assertion is also supported by their co-location in the biomechanics column in Figure 4.

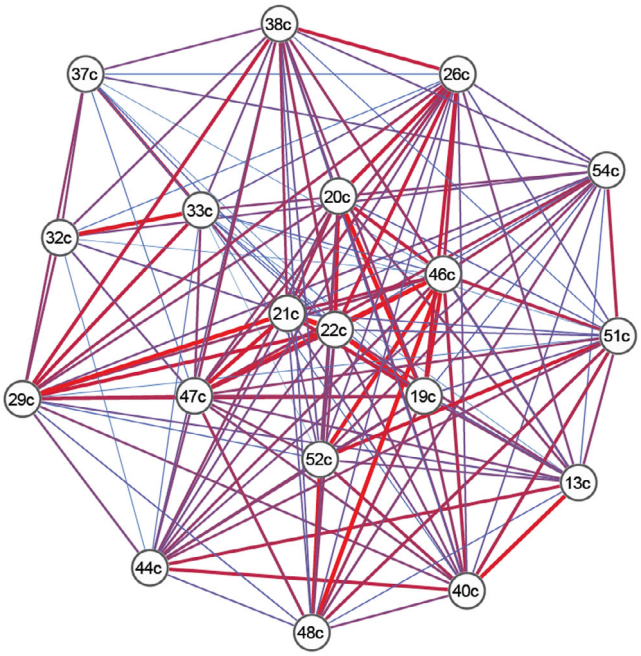

FIGURE 5 A Gephi generated a similarity graph for the Schmid-19 references. Line widths and their color (light blue (small, near 0)-to-dark red (large, close to 1.0) indicate the weight of each pairwise similarity. Node numbers = Schmid reference number; "c" indicates that the paper's single vector is a naïve centroid.

To enable exploration of a larger SG, we expanded the corpus to include the Schmid-64 references and were able to process and completed steps 1–5. A rendering of the resulting SG is presented in Figure 6A. Next, the text of four of eight hypotheses that were generated by the authors to probe BSM domain crosstalk (provided in Supporting Information), along with the text describing the six domains (also provided in Supporting Information) were added to the Schmid corpus. All eight hypotheses were included in the initial ForceAtlas-2 layout. However, hypotheses 1, 3, 4, and 7 exhibited only weak similarities to the remaining corpus and were located considerable distances from the SG centroid. For clarity, they were eliminated (for completeness, their locations relative to the corpus are illustrated). The resulting SG is presented in Figure 6B.

We expect the median edge weight of a large corpus selected randomly from the broad biomedical literature to be approximately 0.5. Because all references cited in Schmid et al. are relevant to the authors' take-home messages, the edge weights in Figure 6B are biased toward larger values. To visualize SG structural details more easily, the graph was regenerated after omitting edge weights <0.85 (Figure 6C). We evaluated several edge weight cut-offs before selecting 0.85.

SG utility—there is no free lunch: Validation and falsification challenges arise immediately from the visualization. Are references 46,[49] 47([54] 48,[48] and 49[55]) actually central to the Schmid et al. content? Are the two apparent clusters reflective of the crosstalk explicitly

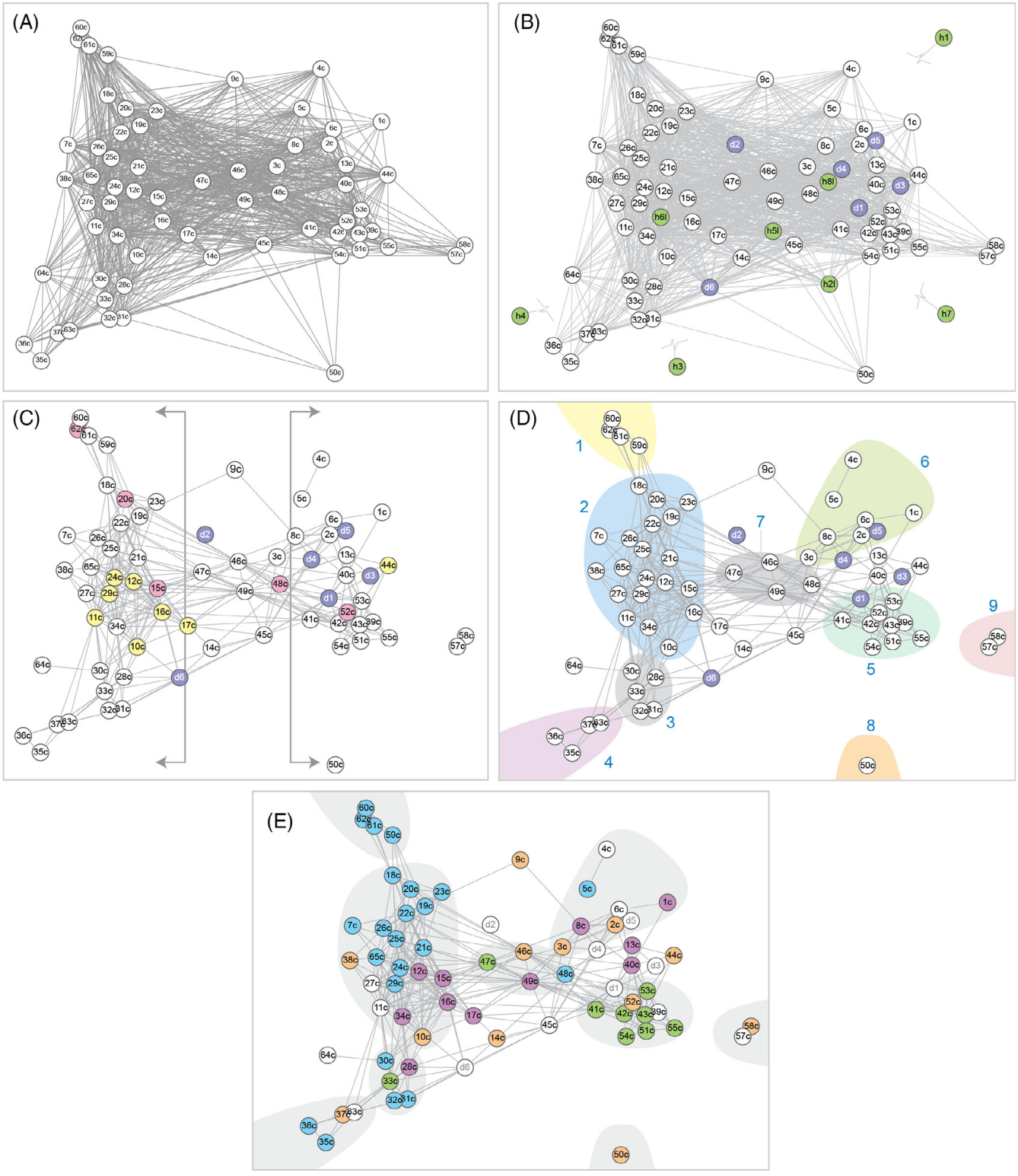

FIGURE 6 Similarity graph visualizations comprising Schmid-64 references. Green nodes beginning with "h" are hypothesis statements. And blue nodes beginning with "d" are domain statements (see Appendix 2). (A) Shown is a graph product of Gephi's ForceAtlas-2 layout algorithm. Node number = Schmid reference number; "c" indicates that the reference's single vector is a naïve centroid of reference text blocks. For visual clarity, variable line widths and colors used in Figure 5 are omitted. (B) The text of four of eight hypotheses (numbered green nodes; Appendix 2) and six domains (blue nodes) were added to the Schmid corpus visualization in (A) resulting in this layout. Because hypotheses 1, 3, 4, and 7 were located considerable distances from the SG centroid in the initial rendering, they are excluded from this rendering. Nevertheless, their locations relative to the corpus are illustrated. (C) To visualize structural details within (B) more easily, the graph was regenerated after omitting edge weights <0.85. References are segregated into two regions separated by several references that are centrally located. References shaded pink: Meier is a coauthor of those references; nodes shaded yellow: Hodges is a coauthor. (D) On inspection, the three isolated references and localized groupings suggest subsets of references that may have similarities much greater than the average for the graph in (A). The nine highlighted subsets are discussed in text. (E). The four GPT-3.5 classifications in Figure 4 are identified: green: psychology, blue: biomechanics; purple: psychology and biomechanics; and orange: neither psychology nor biomechanics.

targeted by Schmid et al.? And so forth. To help test the rendered graph against a domain-based conceptual model, two coauthors (jcl, jfb) manually classified each reference into expressions of the six domains described in References [43]. Domains one and two (Psychological and Biomechanical) were explicitly targeted for the analysis of crosstalk in Schmid et al. and the apparent clustering in the graph seems to match with the manual clustering reasonably well. It is important to avoid conflating the visualizations and layouts in Figure 6 with the actual graph, which can be analyzed with various algorithms. To challenge the above protocol, we manually characterized the primary topic of each citation and mapped that to various visualizations of the graph.

The rendering illustrated in Figure 6D is a mixed-face validation result. Ideally, the purpose behind an unsupervised protocol is to impute as little structure as possible and allow the protocol to pull structure from the data (and only from the data). However, any protocol carries with it some implicit model of the frame and context in which it was developed. For example, naïve centroids from protocol step 4 were embedded into a GPT model of language trained from a wide array of data. Were we to obtain embedding vectors from a model fine-tuned on a more specific corpus like scientific literature (such as all peer-reviewed LBP papers published since 2000), the clustering we see at the output may be completely different. Or, for another example, were we to use a less agnostic PDF to UTF-8 conversion, perhaps one that recognized paper structure like Abstract, Introduction, Keywords, and so forth, and cleaned our raw data to form more coherent text as is done in the KG protocol, the clustering might be considerably different. Ultimately, our target for projects like this is not singular protocols, but a constellation of protocol components that can be assembled and analyzed to provide parallax on the data in the spirit of multiverse analysis.[56]

The above discussion of validation makes clear there is no Free Lunch. While Examples 1–3 and 5–7 involve judicious protocol structuring targeting user-oriented objectives, this Example runs a blind computation and places the burden of finding utility on the back end, the downstream product. While all three example protocols are data-driven, a distinction can be made between them of where, in the flow of data, intelligence can be either imputed or inferred.

3.3.2 | Upstream application alternatives

There are alternative usage patterns for full-text embeddings. It is possible to use the embedding vectors for large sets of full-text documents as the back end for an array of front-end usages. The graphs shown in this section are a type of classification based on a dimension reduction from 1536 to 1. Reductions to 2, 3, or even 5 dimensions are possible using more sophisticated classification protocols (e.g., using t-distributed stochastic neighbor—embedding).

A factual question-and-answer front end could be developed using cosine similarity to search for the most relevant vectors and then use the text associated with that subset as a prompt for the LLM. Given the size of full-text documents and the number of citations required at scale, for exploring crosstalk within and between corpora, a vector database would replace the simple file and CSV-related components used here.

González-Márquez et al.[57] use embeddings representing abstracts of all 21 million English language abstracts contained in the PubMed database to create 2D "landscape" maps of the biomedical literature. They demonstrate an alternative reduction from high dimensional encodings to a 2D space via t-SNE. Their atlas provides two useful contrasts to our SG protocol in (1) construction and frequency of updating, and (2) their 2D space. Considering the former, although the protocol for constructing the atlas may be relatively automatic, it might be best used as an infrequently updated map or "world view" wherein a researcher might locate various concepts, domains, inter-domain "distances," silos, and so forth. On the other hand, our SGs are intended to be generated and re-generated, at will, using different corpora, perhaps eventually being continuously updated as one "drops in" new, revised, or different documents. Considering the latter, the SG protocol does not construct a 2D space, per se. It generates a graph. The graphs we present are the result of stochastic executions of a layout algorithm that depends solely on the properties of the underlying graph (edge weight and vertex degree). It looks like a space because of the layout algorithm. However, unlike the landscape maps, the graph product of this protocol can be visualized in any number of ways.

The papers comprising the corpora (Figure 6) were tokenized and encoded using OpenAI's cl100k_base. Tokenization, in general, involves extracting things like word stems, prefixes, suffixes, punctuation, and so forth. LLMs like GPT inductively infer token relatedness from the statistics of token coincidence. This implies that some polysemy and synonymy could be missed when comparing strings of text. One technique for increasing the probability of successfully parsing alternative phrasing for a given concept is to prompt GPT with alternate phrasings of a given prompt, analogous to the parsings described above (Section 3.2). A concrete example of providing algorithmic permutations is provided by the Stanford Core NLP Coreference Resolution library.[58] By mapping substrings of text together, registering them as references to the same entity, and reconstructing multiple phrases with all the alternate substrings, coreference resolution can help the reasoner (in this case GPT) better recognize and generate polysemous and synonymous terms. However, as we address herein with the other protocols, judicious preparation of the input data imputes structure that may bias results.

3.3.3 | Summary of SG protocol

In summary, the execution of the SG protocol over the Schmid-19, Schmid-65, and other data sets (see Supporting Information) clearly shows that while the technique is partially validated, there will be uncertainty in the result. Because it is automated, simple, and inexpensive to execute, the protocol seems useful for early-phase exploration of an unfamiliar corpus. However after the initial exposure to the results, more algorithm components should be appended to build

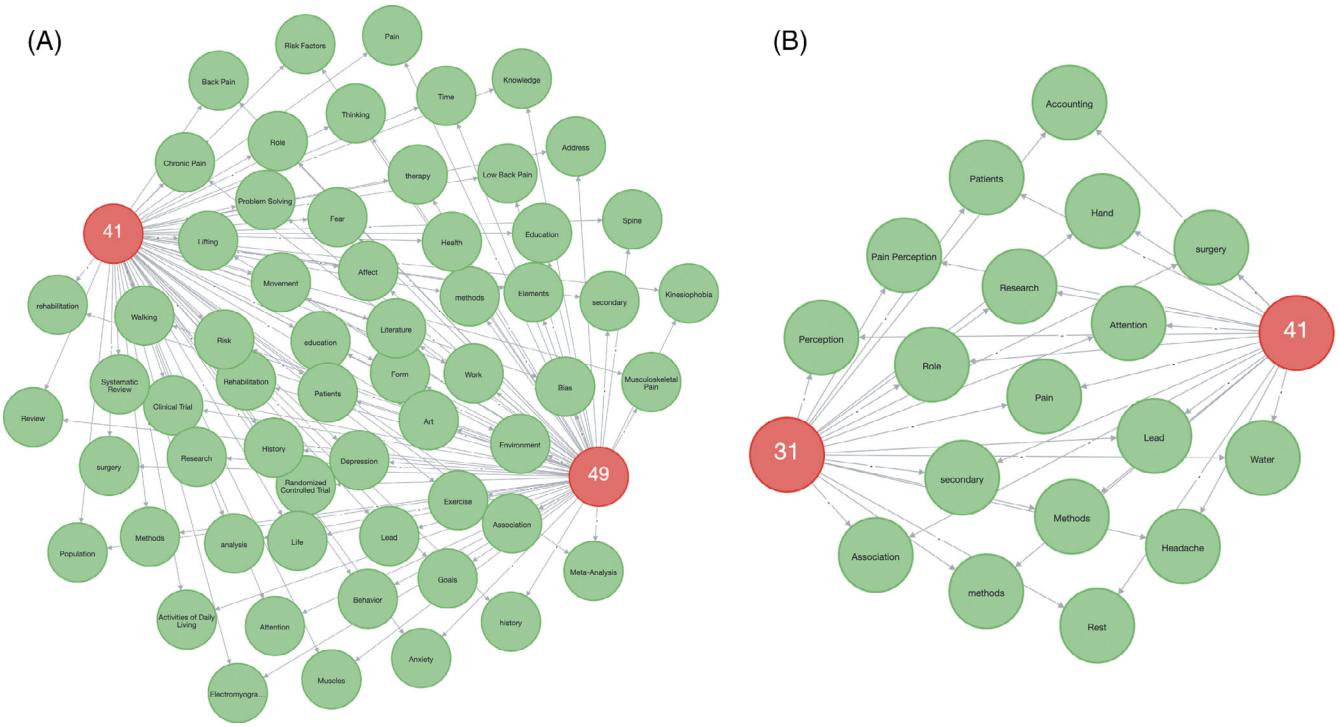

FIGURE 7 Illustrative example of interpreting similarity graphs through shared relations in a knowledge graph. Shown are relations shared between Schmid references 41[103] and 49[48] (A) and between 31[98] and 41[103] (B).

confidence. This additional requirement, which has not been addressed in the presented explorations, emphasizes the composability and modularity of the components used in the Examples we discuss.

3.4 | Knowledge graphs

KGs provide a complementary and flexible method for organizing and representing knowledge. In this section, we demonstrate the utility of a knowledge graph to provide context and semantic information about AI-assisted results. KGs were generated by processing and annotating the six domains and the Schmid-64 PDFs. We converted each PDF to UTF-8 text using mutool. Raw UTF-8 text was tokenized into sentences that were then evaluated using GPT-3 to identify invalid and valid sentences. Invalid sentences were discarded. All valid sentences were then annotated for the presence of terms defined in the ontology via the bioportal annotation API (https://bioportal.bioontology.org/). The resulting information was loaded into Neo4j. Additional information is provided in Supporting Information.

3.4.1 | Interpreting similarity graphs (Example 5)

To illustrate KG-assisted interpretation options from Example 4, assume we are interested in plausible explanations for similarity differences between the apparent clusters in Figure 6. We selected three references: 31 (group 3 in Figure 6D), 41 (group 5), and 49 (group 7). Based on the Figure 6 SGs and examination of the papers, we concluded there is little overlap of papers' 31 and 41 content, whereas there is evident overlap of papers' 41 and 49 content. These results can be verified in the KGs of these three papers by querying for shared relations (Figure 7). For this illustrative example, we utilized the medical subject headings (MESH) Ontology which is used for indexing articles on PubMed.

3.4.2 | Interpreting LLM-assisted literature classification (Example 6)

To illustrate KG-assisted interpretation options, we focus on the four-category document classification described in Section 3.2. GPT-3 classified references 21 and 26 under the biomechanics category (Figure 4), whereas the experts classified them under the psychology and biomechanics category. Assume we would like to augment GPT prompts to future improve agreement between expert and GPT classifications. What might have accounted for that disagreement?

Fear of movement is an important LBP-associated phenomenon mentioned or discussed in a majority of the Schmid-19 corpus. Examination of the ontology reveals that the fear concept is nested under Behavior and Behavior Systems. While the MESH ontology has over 300 000 terms, we can visualize subclasses of Behavior and Behavior Systems (Figure 8).

Focusing on the Behavior and Behavior Systems subclass, we queried the knowledge graph to retrieve entities identified in papers 21 and 26. For a comparison, we ran the same query on two

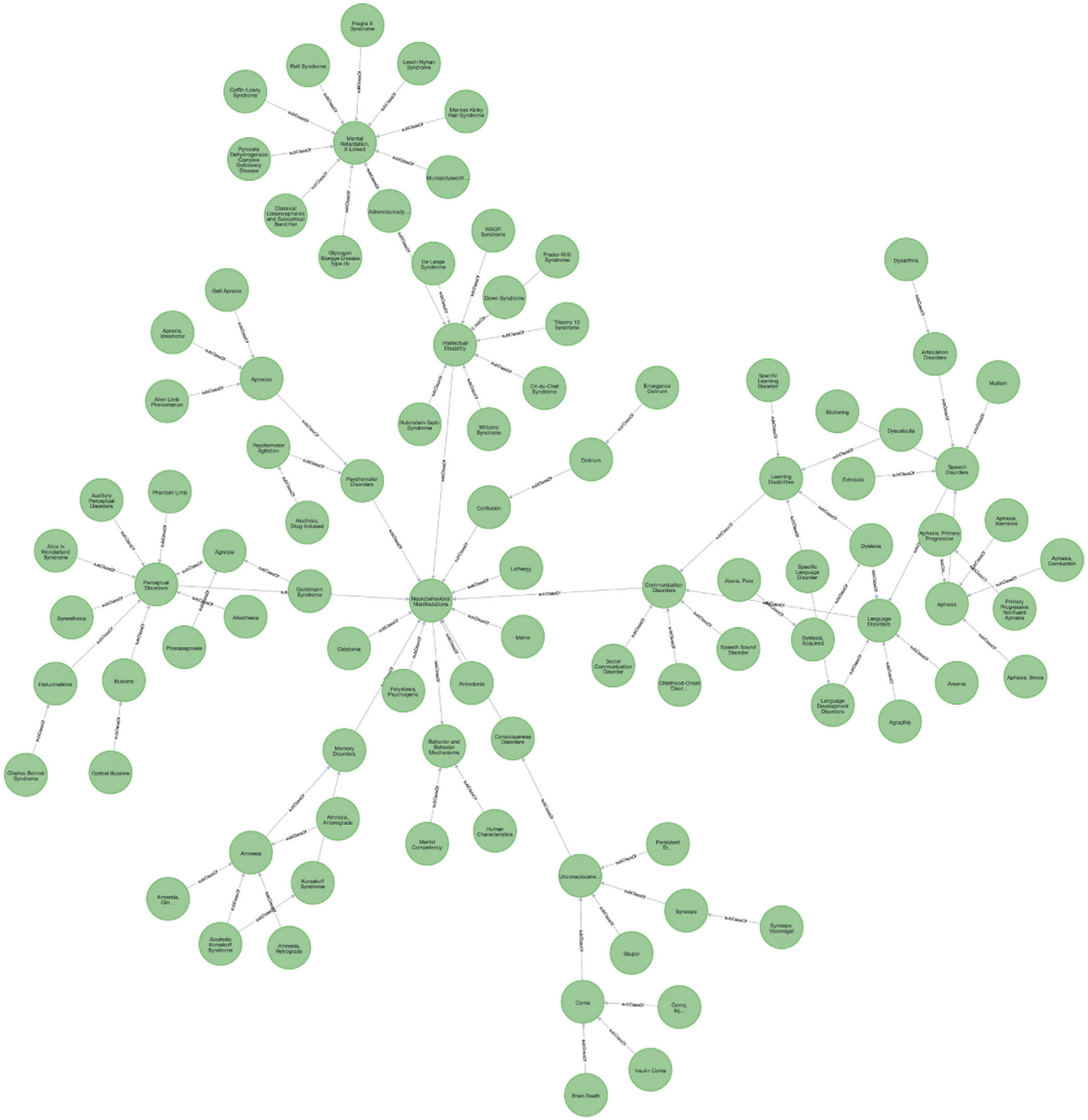

FIGURE 8 A graph showing a subset of the MESH ontology for Behavior and Behavior Systems.

papers—5 and 62—classified under the biomechanics category by experts and GPT. KG results are presented in Figure 9. We see that both 21 and 26 discuss fear directly. Furthermore, paper 26 address biomechanics (e.g., walking and motor activity) from a Behavior Systems perspective. A plausible hypothesis is that the AI system did not capture the semantics of biomechanics as a consequence of behavior. Moving forward, one could capture those semantics by infusing knowledge in a shallow to deep manner during orAI model fine-tuning.

3.4.3 | KG-assisted identification of studies that bridge multiple LBP domains (Example 7)

KGs can help identify prior studies that bridge multiple domains. For this example, we measured the similarity between the six domains listed in Section 3.1 and the Schmid-64 corpus. The similarity between domains and papers was computed using Szymkiewicz–Simpson coefficient (also known as the overlap coefficient). The overlap coefficient measures the similarity between nodes based on the nodes that are

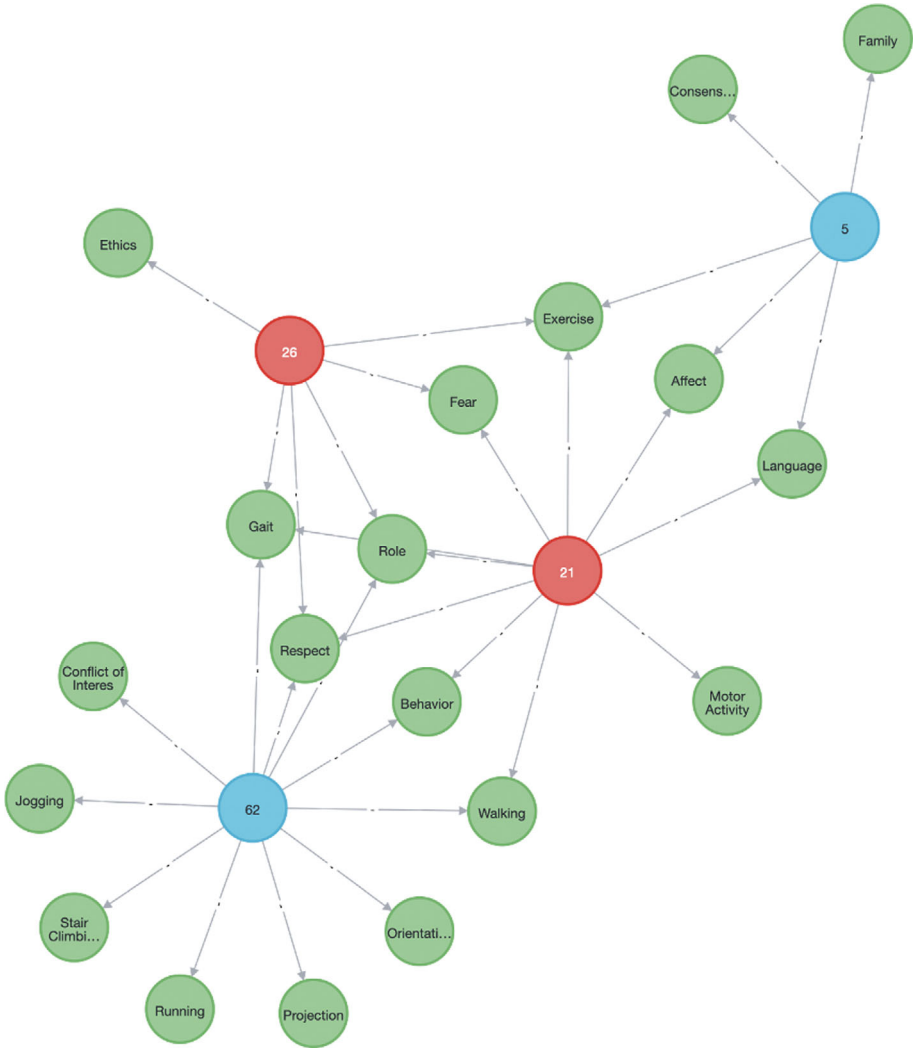

FIGURE 9 Query results for Behavior and Behavior System entities identified within the incorrectly classified Schmid references (21[52] and 26[59]). References 5[82] and 62[114] were correctly identified and are added for comparison.

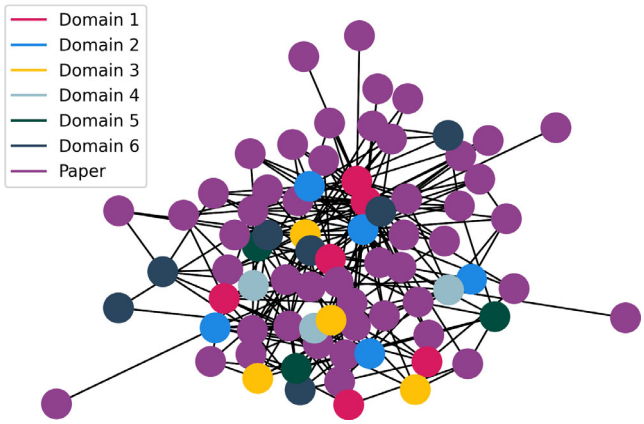

FIGURE 10 Similarity graph visualization of the Schmid-64 corpus shown in spring force layout. The relationship between references and domains is calculated from the KG overlap coefficient using neighbors in common between references as identified by the NCIT and MESH ontologies.

shared. For this example, we annotated entities using both the MESH and NCIT ontologies to form the network of shared nodes. A higher number of shared annotations between a domain and a paper will result in a higher overlap coefficient. Similar to prior examples, the results of this algorithm can be visualized as an SG. The graph presented in Figure 10 is the result of this example. One can infer from the graph that papers near the center or between a mix of domains are papers addressing multiple domains. Such information may be critical at scale where the volume of literature across domains outpaces our ability to synthesize.

3.4.4 | Summary of applications of KGs

We have demonstrated three applications of KGs for bridging chronic pain knowledge across domains. These applications show the utility of KGs to explain and interpret AI-assisted results as well as assist in the generation of new knowledge. As these approaches are adopted at scale, KG-driven tools provide valuable insights into crosstalk which could increase the impact of new hypotheses and lead to a greater chance of scientific discovery. With minimal configuration and resources, KG technology can support hypothesis creation and provide semantically meaningful explanations to other more opaque AI systems. We explored a subset of possible KG applications to demonstrate the utility of KGs. Additional example use cases for KG technology include

linkage prediction which can directly hypothesize relations in the knowledge graph and the generation of embeddings that capture the heterogenous knowledge stored in a KG for downstream ML.

There are significant limitations and resource considerations for deployment, maintenance, and utilization of KGs in bridging LBP domains. Biomedical research communities maintain a number of robust ontologies. Bioportal currently reports over 1000 ontologies that define over 15 million classes. While such ontology abundance encouraging, we must consider the resources required to maintain, extend, and update them. While progress is ongoing to automate ontology and KG creation, the modeling of unstructured data as a KG remains an open research question.

4 | DISCUSSION

Our goal was to concretely demonstrate several ways that LLMs, KGs, and SGs can be used now to advance research and treatment protocols in the context of LBP. A secondary goal was to provide enough detail such that the reader gets an idea of whether or how similar, the exchangeable components described above might be assembled into one's own context-dependent search, exploration, and analysis protocols. A fulcrum for these goals was to place those protocols on a spectrum of computer assistance, from nearly unassisted methods like participatory modeling to protocols with significant computation delegated to AI/ML algorithms. The spectrum invites the reader to be skeptical about what can and cannot be delegated to algorithms. That skepticism is well articulated by the No Free Lunch theorem (and so-called "no-go" conundrums, in general).[38] LBP domain crosstalk presents a poignant case that is sufficiently challenging to put the credibility of assistive technology at risk. Yet the need for crosstalk in a complex meta-domain like LBP provides an opportunity to develop and challenge protocols composed with significant assistive components like LLMs and KGs. The spectrum also highlights the categorical difference between pre-clinical and clinical protocols, where the latter are subject to extensive validation and regulatory approval. Clinical practice cannot be delegated to algorithms and any automatic protocol must be an explainable and well-understood tool for use and synthesis by a clinician.

Each of the above protocols shows limited yet promising utility in addressing crosstalk within a BSM approach to LBP. In Examples 1–3, we used the Schmid corpus as a backdrop for applications of OpenAI's GPT-3.5 to three crosstalk-related search, exploration, and analysis protocols. In using relatively upstream SGs to explore the corpus (Example 4), some similarities manually identified among the papers were not reflected by the graph layout, despite observing renderings that validate the method. This mixed result indicates that while deep and impactive dissimilarities may show clearly enough for exploration without manually reading every paper, any patterns observed, qualitatively or quantitatively, will need to be buttressed with some form of face validation, however limited. In the KG construction and analysis protocol (Examples 5–6), we see how KGs can help inform and interpret crosstalk computational results in addition to providing an independent technology for knowledge generation (Example 7). Despite the limited success of these exercises, the protocols demonstrate that the components are reliable and easy enough to integrate, and merit further study from additional LBP perspectives. We believe the utility of resulting protocols will outweigh the costs of assembly, testing, and customization.

Implicit in LBP research is the development of hypotheses—if-then statements that ideally have a mechanistic basis and are meant to bring into focus how LBP phenomena may be generated. Conventional hypothesis generation is a highly sophisticated cognitive process where an investigator relies on deductive reasoning, technical skill, and imagination. Deductive reasoning includes inferences based on facts or premises, and where the latter are often arrived at through induction, known to the researcher. This poses a fundamental dilemma in LBP. Given the multidisciplinary nature of current research and treatment strategies[60,61] few, if any, of the scientists are experts in all LBP domains (and cannot remain current of all new discoveries and their relevance). Using protocols, like those described, to place hypotheses in context within the broader, rapidly growing literature helps tease new information from affiliated research, the significance of which may have been unknown to the investigator. The examples of protocol composition provide hints at how one might bridge from research hypotheses and domains to clinical decision-making, where a patient holistically presents a variety of aspects to the clinician. From the presented aspects the clinician interactively walks a sophisticated course of multi-domain "hypotheses," a composite of if-then diagnostic forks, to arrive at interventions most likely to help that patient.

The above three protocols compose inductive and deductive reasoning to demonstrate multi-domain literature and hypothesis exploration and analysis. A limitation of using language models to infer meaning from the published literature during research is that research is intrinsically complicated. Furthermore, papers are often not written to maximize clarity. Typical papers are jammed full of experimental data, and yet the underlying logic of the paper, including its hypotheses and reasoning about them, is frequently left unstated.[62] Similarly, even if a topic, or domain description[43] is provided, the wording, vocabulary, and implicit conceptual model may not match well with either published literature or practitioners' conceptual models or behavior in the research or during clinical protocols. There is also a reporting bias, as negative results, and failed replications often remain unknown to those outside that silo. This reality can distort the literature by maintaining prevailing theories that are solely based on "positive" results.[4,63] These challenges may be reduced incrementally in the future as NIH recently published a Data Management and Sharing requirement that promotes sharing and broad accessibility to research data for reuse and external validation.

These above protocols are somewhat different compositions of the same core components and are intended to only represent examples of how such components might be used. More complete assessments of utility and cost will require literature reviews and larger samples from the space of possible protocols. With that in mind, the applications of these protocols and tools can be understood from dual

aspects. For a given corpus, is there one or more threads that connect knowledge in one domain to knowledge in another? In a fine-grained search, an example of this would be KG link prediction (e.g., as in Reference [64]). In a coarse-grained search, this would amount to an assessment of component connectivity. Dually, given a conjecture that an explanatory relationship exists between two domains, does the underlying corpus support or contradict the conjecture? If there is no supportive or contradicting relationship in the graph, then either the corpus is inadequate, or the conjecture might be a topic for future research.

Perhaps more than any other chronic medical condition, advances in LBP management have suffered from communication, language, and knowledge gaps between patients, providers, researchers (and those that support them), payers, the medical device industry, and healthcare systems. Adding to the confusion is the easy access to abundant information via the internet, information that is largely uncurated and unvalidated. The protocol ideas presented herein have the potential to help organize and identify key patient-centered threads of evidence-based knowledge in ways that span the continuum of research, clinical care, and population health. The complexity and resilience of chronic LBP illustrate the difficult challenges of unifying a multidisciplinary framework for improving mechanistic understanding of individual trajectories. In this regard, the protocols and tools presented may prove to be valuable if they successfully nudge investigators to develop more comprehensive hypotheses by expanding awareness of intra- and trans-domain linkages, particularly when attempting to link clinical information with data from pre-clinical research. Ultimately, those efforts may provide the means to optimize identification, validation, and deployment of new preventative and therapeutic strategies.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest.

ORCID

Jeffrey C. Lotz https://orcid.org/0000-0002-9654-0647
C. Anthony Hunt https://orcid.org/0000-0001-9372-6860

REFERENCES

1. Hartvigsen J, Hancock MJ, Kongsted A, et al. What low back pain is and why we need to pay attention. *Lancet*. 2018;*391*(10137):2356-2367.
2. Zuo Y, Zhang J, Leng X, Fan Y, Fu B, Wang P. A scientometrics analysis and visualization of low back pain. *Int J Osteopath Med*. 2023;*47*:100655.
3. Krenn M, Zeilinger A. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proc Natl Acad Sci USA*. 2020;*117*(4):1910-1916.
4. Rodriguez-Esteban R. Information silos distort biomedical research. bioRxiv 2021 2021.07.26.453749.
5. Banker S, Chatterjee P, Mishra H, Mishra A. Machine-Assisted Social Psychology Hypothesis Generation. PsyArXiv 2023.
6. Freeman RB, Huang W. Collaboration: strength in diversity. *Nature*. 2014;*513*(7518):305.
7. Langevin HM. Reconnecting the brain with the rest of the body in musculoskeletal pain research. *J Pain*. 2021;*22*(1):1-8.
8. Schmid S, Bangerter C, Schweinhardt P, Meier ML. Identifying motor control strategies and their role in low back pain: a cross-disciplinary approach bridging neurosciences with movement biomechanics. *Front Pain Res (Lausanne)*. 2021;*2*:715219.
9. Raul R-E. Information silos distort biomedical research. bioRxiv 2021 2021.07.26.453749.
10. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. 2023 arXiv.
11. Liu M, Bu Y, Chen C, et al. Pandemics are catalysts of scientific novelty: evidence from COVID-19. *J Assoc Inf Sci Technol*. 2022;*73*(8):1065-1078.
12. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. 2023;*27*(1):120.
13. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;*388*(13):1233-1239.
14. Liu Z et al. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. 2023 arXiv.
15. Wang S, Dang Y, Sun Z, et al. An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. *J Am Med Inform Assoc*. 2023;*30*:1408-1417.
16. Wang L et al. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. 2021 arXiv.
17. Zhou M et al. On the generation of medical dialogs for COVID-19. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers. Association for Computational Linguistics; 2021.
18. Basu C et al. Med-EASi: Finely Annotated Dataset and Models for Controllable Simplification of Medical Texts. 2023 arXiv.
19. Peng B et al. Instruction Tuning with GPT-4. 2023 arXiv.
20. Zamfirescu-Pereira JD et al. Why Johnny Can't prompt: how non-AI experts try (and fail) to design LLM prompts. *CHI '23: CHI conference on human factors in computing systems*. ACM; 2023.
21. Zamfirescu-Pereira JD, Wei H, Xiao A, et al. Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. *DIS '23: Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Association for Computing Machinery; 2023.
22. Sanh V et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. 2022 arXiv.
23. Yang Q et al. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. *CHI '23: CHI Conference On Human Factors In Computing Systems*. ACM; 2023.
24. Fan T, Wang X, Song X, Zhao G, Zhang Z. Research status and emerging trends in virtual reality rehabilitation: bibliometric and knowledge graph study. *JMIR Serious Games*. 2023;*11*:e41091.
25. Li Z, Xu C, Fu J, Zulipikaer M, Deng T, Chen J. Scientific knowledge graph and trend analysis of central sensitization: a bibliometric analysis. *J Pain Res*. 2022;*15*:561-575.
26. Hu Z, Yichong Xu Y, Wenhao Yu W, et al. Empowering Language Models with Knowledge Graph Reasoning for Question Answering. 2022.
27. Gutierrez C, Sequeda JF. Knowledge graphs: a tutorial on the history of knowledge Graph's Main ideas. *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery; 2020.
28. Gaur M, Faldu K, Sheth A. Semantics of the black-box: can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Comput*. 2021;*25*(1):51-59.
29. Sheth A, Gaur M, Kursuncu U, Wickramarachchi R. Shades of knowledge-infused learning for enhancing deep learning. *IEEE Internet Comput*. 2019;*23*(6):54-63.
30. Ugur K, Manas G, Amit S. Knowledge Infused Learning (K-IL): Towards Deep Incorporation of Knowledge in Deep Learning. 2019.
31. Callahan TJ, Tripodi IJ, Pielke-Lombardo H, Hunter LE. Knowledge-based biomedical data science. *Annu Rev Biomed Data Sci*. 2020;*3*:23-41.
32. Aziguli ZY, Xie Y, Xu Y, Chen Y. Structural technology research on symptom data of Chinese medicine. *Proceedings of The IEEE 19th*

International Conference On E-Health Networking, Applications And Service. IEEE; 2017.
33. Shang J et al. GAMENet: Graph Augmented MEmory Networks for Recommending Medication Combination. 2018.
34. Lü L, Zhou T. Link prediction in complex networks: a survey. *Physica A*. 2011;*390*(6):1150-1170.
35. Wang X, Wang Y, Gao C, Lin K, Li Y. Automatic diagnosis with efficient medical case searching based on evolving graphs. *IEEE Access*. 2018;*6*:53307-53318.
36. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*. 2012;*28*(23):3158-3160.
37. Rigden DJ, Fernández XM. The 27th annual nucleic acids research database issue and molecular biology database collection. *Nucleic Acids Res*. 2020;*48*(D1):D1-D8.
38. Wolpert DH. Ubiquity symposium: evolutionary computation and the processes of life: what the no free lunch theorems really mean: how to improve search algorithms. *Ubiquity*. 2013;*2013*:1-15.
39. Cholewicki J, Popovich JM Jr, Aminpour P, Gray SA, Lee AS, Hodges PW. Development of a collaborative model of low back pain: report from the 2017 NASS consensus meeting. *Spine J*. 2019;*19*(6):1029-1040.
40. Voinov A, Kolagani N, McCall MK, et al. Modelling with stakeholders—next generation. *Environ Model Software*. 2016;*77*:196-220.
41. Hohmann E, Cote MP, Brand JC. Research pearls: expert consensus based evidence using the Delphi method. *Art Ther*. 2018;*34*(12):3278-3282.
42. MuPDF. MuPDF—the lightweight PDF, XPS, and E-book viewer. Available from: https://mupdf.com/
43. Chau A, Steib S, Whitaker E, et al. Theoretical schemas to guide BACPAC chronic low back pain clinical research. *Pain Med*. 2022;*24*:S13-S35.
44. Jing X et al. Development, validation, and usage of metrics to evaluate clinical research hypothesis quality. medRxiv 2023.
45. Borrell-Carrió F, Suchman AL, Epstein RM. The biopsychosocial model 25 years later: principles, practice, and scientific inquiry. *Ann Fam Med*. 2004;*2*(6):576-582.
46. Daluiso-King G, Hebron C. Is the biopsychosocial model in musculoskeletal physiotherapy adequate? An evolutionary concept analysis. *Physiother Theory Pract*. 2022;*38*(3):373-389.
47. Wertli MM, Rasmussen-Barr E, Held U, Weiser S, Bachmann LM, Brunner F. Fear-avoidance beliefs-a moderator of treatment efficacy in patients with low back pain: a systematic review. *Spine J*. 2014;*14*(11):2658-2678.
48. Knechtle D, Schmid S, Suter M, et al. Fear-avoidance beliefs are associated with reduced lumbar spine flexion during object lifting in pain-free adults. *Pain*. 2021;*162*(6):1621-1631.
49. Matheve T, de Baets L, Bogaerts K, Timmermans A. Lumbar range of motion in chronic low back pain is predicted by task-specific, but not by general measures of pain-related fear. *Eur J Pain*. 2019;*23*(6):1171-1184.
50. Buchthal F, Rosenfalck P. Spontaneous electrical activity of human muscle. *Electroencephalogr Clin Neurophysiol*. 1966;*20*(4):321-336.
51. Roux FE, Djidjeli I, Durand JB. Functional architecture of the somatosensory homunculus detected by electrostimulation. *J Physiol*. 2018;*596*(5):941-956.
52. van Dieen JH, Selen LP, Cholewicki J. Trunk muscle activation in low-back pain patients, an analysis of the literature. *J Electromyogr Kinesiol*. 2003;*13*(4):333-351.
53. Marras WS, Ferguson SA, Burr D, Davis KG, Gupta P. Spine loading in patients with low back pain during asymmetric lifting exertions. *Spine J*. 2004;*4*(1):64-75.
54. Geisser ME, Haig AJ, Wallbom AS, Wiggert EA. Pain-related fear, lumbar flexion, and dynamic EMG among persons with chronic musculoskeletal low back pain. *Clin J Pain*. 2004;*20*(2):61-69.
55. Christe G, Crombez G, Edd S, Opsommer E, Jolles BM, Favre J. Relationship between psychological factors and spinal motor behaviour in low back pain: a systematic review and meta-analysis. *Pain*. 2021;*162*(3):672-686.
56. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*. 2016;*11*(5):702-712.
57. Rita G-M et al. The landscape of biomedical research. bioRxiv 2023 2023.04.10.536208.
58. Clark K, Manning CD. Improving Coreference Resolution by Learning Entity-Level Distributed Representations, in Association for Computational Linguistics (ACL). 2016.
59. Claeys K, Brumagne S, Dankaerts W, Kiers H, Janssens L. Decreased variability in postural control strategies in young people with non-specific low back pain is associated with altered proprioceptive reweighting. *Eur J Appl Physiol*. 2011;*111*(1):115-123.
60. Ashar YK, Gordon A, Schubiner H, et al. Effect of pain reprocessing therapy vs placebo and usual care for patients with chronic back pain: a randomized clinical trial. *JAMA Psychiatry*. 2022;*79*(1):13-23.
61. Bontinck J, den Hollander M, Kaas AL, de Jong JR, Timmers I. Individual patterns and temporal trajectories of changes in fear and pain during exposure in vivo: a multiple single-case experimental design in patients with chronic pain. *J Clin Med*. 2022;*11*(5):1-10.
62. Alger BE. Scientific hypothesis-testing strengthens neuroscience research. *eNeuro*. 2020;*7*:4.
63. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front Hum Neurosci*. 2017;*11*:390.
64. Fecho K, Bizon C, Miller F, et al. A biomedical knowledge graph system to propose mechanistic hypotheses for real-world environmental health observations: cohort study and informatics application. *JMIR Med Inform*. 2021;*9*(7):e26714.
65. Clays E, de Bacquer D, Leynen F, Kornitzer M, Kittel F, de Backer G. The impact of psychosocial factors on low back pain: longitudinal results from the Belstress study. *Spine (Phila Pa 1976)*. 2007;*32*(2):262-268.
66. Gombatto SP, D'Arpa N, Landerholm S, et al. Differences in kinematics of the lumbar spine and lower extremities between people with and without low back pain during the down phase of a pick up task, an observational study. *Musculoskelet Sci Pract*. 2017;*28*:25-31.
67. Simonet E, Winteler B, Frangi J, et al. Walking and running with non-specific chronic low back pain: what about the lumbar lordosis angle? *J Biomech*. 2020;*108*:109883.
68. Marras WS, Davis KG, Ferguson SA, Lucas BR, Gupta P. Spine loading characteristics of patients with low back pain compared with asymptomatic individuals. *Spine (Phila Pa 1976)*. 2001;*26*(23):2566-2574.
69. Tsao H, Danneels LA, Hodges PW. ISSLS prize winner: smudging the motor brain in young adults with recurrent low back pain. *Spine (Phila Pa 1976)*. 2011;*36*(21):1721-1727.
70. Eto K, Wake H, Watanabe M, et al. Inter-regional contribution of enhanced activity of the primary somatosensory cortex to the anterior cingulate cortex accelerates chronic pain behavior. *J Neurosci*. 2011;*31*(21):7631-7636.
71. Lim M, Roosink M, Kim JS, et al. Disinhibition of the primary somatosensory cortex in patients with fibromyalgia. *Pain*. 2015;*156*(4):666-674.
72. Beaudette SM, Larson KJ, Larson DJ, Brown SHM. Low back skin sensitivity has minimal impact on active lumbar spine proprioception and stability in healthy adults. *Exp Brain Res*. 2016;*234*(8):2215-2226.

73. Ranger TA, Cicuttini FM, Jensen TS, Manniche C, Heritier S, Urquhart DM. Catastrophization, fear of movement, anxiety, and depression are associated with persistent, severe low back pain and disability. *Spine J*. 2020;*20*(6):857-865.
74. Klyne DM, van den Hoorn W, Barbe MF, et al. Cohort profile: why do people keep hurting their back? *BMC Res Notes*. 2020; *13*(1):538.
75. Houben RM et al. Fear of movement/injury in the general population: factor structure and psychometric properties of an adapted version of the Tampa scale for Kinesiophobia. *J Behav Med*. 2005; *28*(5):415-424.
76. Leeuw M, Goossens MEJB, van Breukelen GJP, Boersma K, Vlaeyen JWS. Measuring perceived harmfulness of physical activities in patients with chronic low back pain: the photograph series of daily activities—short electronic version. *J Pain*. 2007; *8*(11):840-849.
77. Meier ML, Vrana A, Humphreys BK, Seifritz E, Stämpfli P, Schweinhardt P. Pain-related fear-dissociable neural sources of different fear constructs. *eNeuro*. 2018;*5*(6):1-15.
78. Wu A, March L, Zheng X, et al. Global low back pain prevalence and years lived with disability from 1990 to 2017: estimates from the global burden of disease study 2017. *Ann Transl Med*. 2020; *8*(6):299.
79. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. *Lancet*. 2017;*389*(10070):736-747.
80. Vlaeyen JWS, Maher CG, Wiech K, et al. Low back pain. *Nat Rev Dis Primers*. 2018;*4*(1):52.
81. Rubinstein SM, Terwee CB, Assendelft WJJ, de Boer MR, van Tulder MW. Spinal manipulative therapy for acute low back pain: an update of the cochrane review. *Spine (Phila Pa 1976)*. 2013;*38*(3): E158-E177.
82. van Middelkoop M, Rubinstein SM, Kuijpers T, et al. A systematic review on the effectiveness of physical and rehabilitation interventions for chronic non-specific low back pain. *Eur Spine J*. 2011;*20*(1): 19-39.
83. Goubert D, Oosterwijck JV, Meeus M, Danneels L. Structural changes of lumbar muscles in non-specific low back pain: a systematic review. *Pain Physician*. 2016;*19*(7):E985-E1000.
84. Knezevic NN, Candido KD, Vlaeyen JWS, van Zundert J, Cohen SP. Low back pain. *Lancet*. 2021;*398*(10294):78-92.
85. Brinjikji W, Diehn FE, Jarvik JG, et al. MRI findings of disc degeneration are more prevalent in adults with low back pain than in asymptomatic controls: a systematic review and meta-analysis. *AJNR Am J Neuroradiol*. 2015;*36*(12):2394-2399.
86. Hodges PW, Tucker K. Moving differently in pain: a new theory to explain the adaptation to pain. *Pain*. 2011;*152*(3 Suppl):S90-S98.
87. Tsao H, Galea MP, Hodges PW. Reorganization of the motor cortex is associated with postural control deficits in recurrent low back pain. *Brain*. 2008;*131*(Pt 8):2161-2171.
88. van Dieen JH et al. Motor control changes in low Back pain: divergence in presentations and mechanisms. *J Orthop Sports Phys Ther*. 2019;*49*(6):370-379.
89. Meier ML, Vrana A, Schweinhardt P. Low back pain: the potential contribution of supraspinal motor control and proprioception. *Neuroscientist*. 2019;*25*(6):583-596.
90. van Dieen JH, Flor H, Hodges PW. Low-back pain patients learn to adapt motor behavior with adverse secondary consequences. *Exerc Sport Sci Rev*. 2017;*45*(4):223-229.
91. Hodges PW, Smeets RJ. Interaction between pain, movement, and physical activity: short-term benefits, long-term consequences, and targets for treatment. *Clin J Pain*. 2015;*31*(2):97-107.
92. Christe G, Redhead L, Legrand T, Jolles BM, Favre J. Multi-segment analysis of spinal kinematics during sit-to-stand in patients with chronic low back pain. *J Biomech*. 2016;*49*(10):2060-2067.
93. MacDonald D, Moseley LG, Hodges PW. Why do some patients keep hurting their back? Evidence of ongoing back muscle dysfunction during remission from recurrent back pain. *Pain*. 2009;*142*(3): 183-188.
94. Prins MR, Griffioen M, Veeger TTJ, et al. Evidence of splinting in low back pain? A systematic review of perturbation studies. *Eur Spine J*. 2018;*27*(1):40-59.
95. Hodges PW, Cholewicki J, Dieen JH. *Spinal Control: The Rehabilitation of Back Pain*. Churchill Livingstone; 2013.
96. Flor H, Braun C, Elbert T, Birbaumer N. Extensive reorganization of primary somatosensory cortex in chronic back pain patients. *Neurosci Lett*. 1997;*224*(1):5-8.
97. Riemann BL, Lephart SM. The sensorimotor system, part I: the physiologic basis of functional joint stability. *J Athl Train*. 2002;*37*(1): 71-79.
98. Bushnell MC, Duncan GH, Hofbauer RK, Ha B, Chen JI, Carrier B. Pain perception: is there a role for primary somatosensory cortex? *Proc Natl Acad Sci USA*. 1999;*96*(14):7705-7709.
99. Elgueta-Cancino E, Schabrun S, Hodges P. Is the Organization of the Primary Motor Cortex in low back pain related to pain, movement, and/or sensation? *Clin J Pain*. 2018;*34*(3):207-216.
100. Sutherling WW, Levesque MF, Baumgartner C. Cortical sensory representation of the human hand: size of finger regions and nonoverlapping digit somatotopy. *Neurology*. 1992;*42*(5):1020-1028.
101. Ejaz N, Hamada M, Diedrichsen J. Hand use predicts the structure of representations in sensorimotor cortex. *Nat Neurosci*. 2015;*18*(7): 1034-1040.
102. Martinez-Calderon J, Flores-Cortes M, Morales-Asencio JM, Luque-Suarez A. Pain-related fear, pain intensity and function in individuals with chronic musculoskeletal pain: a systematic review and meta-analysis. *J Pain*. 2019;*20*(12):1394-1415.
103. Leeuw M, Goossens MEJB, Linton SJ, Crombez G, Boersma K, Vlaeyen JWS. The fear-avoidance model of musculoskeletal pain: current state of scientific evidence. *J Behav Med*. 2007;*30*(1):77-94.
104. Zale EL, Lange KL, Fields SA, Ditre JW. The relation between pain-related fear and disability: a meta-analysis. *J Pain*. 2013;*14*(10): 1019-1030.
105. Schweinhardt P. Where has the 'bio' in bio-psycho-social gone? *Curr Opin Support Palliat Care*. 2019;*13*(2):94-98.
106. LeDoux JE, Hofmann SG. The subjective experience of emotion: a fearful view. *Curr Opin Behav Sci*. 2018;*19*:67-72.
107. Pfingsten M, Kröner-Herwig B, Leibing E, Kronshage U, Hildebrandt J. Validation of the German version of the fear-avoidance beliefs questionnaire (FABQ). *Eur J Pain*. 2000;*4*(3): 259-266.
108. Lundberg M et al. Pain-related fear: a critical review of the related measures. *Pain Res Treat*. 2011;*2011*:494196.
109. Julian LJ. Measures of anxiety: state-trait anxiety inventory (STAI), beck anxiety inventory (BAI), and hospital anxiety and depression scale-anxiety (HADS-A). *Arthritis Care Res (Hoboken)*. 2011;*63*(Suppl 11):S467-S472.
110. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;*16*(9):606-613.
111. Schmid S, Bruhin B, Ignasiak D, et al. Spinal kinematics during gait in healthy individuals across different age groups. *Hum Mov Sci*. 2017; *54*:73-81.
112. Niggli LA, Eichelberger P, Bangerter C, Baur H, Schmid S. Between-session reliability of skin marker-derived spinal kinematics during functional activities. 2020.
113. Zemp R, List R, Gülay T, et al. Soft tissue artefacts of the human back: comparison of the sagittal curvature of the spine measured using skin markers and an open upright MRI. *PLoS One*. 2014;*9*(4): e95426.

114. Connolly LEP, Schmid S, Moschini G, Meier ML, Senteler M. Motion Capture-driven Musculoskeletal Spine Modeling: an OpenSim-based Inverse Kinematics Approach. 2021.
115. Nelson AJ, Chen R. Digit somatotopy within cortical areas of the postcentral gyrus in humans. *Cereb Cortex*. 2008;*18*(10):2341-2351.
116. Weerakkody NS, Mahns DA, Taylor JL, Gandevia SC. Impairment of human proprioception by high-frequency cutaneous vibration. *J Physiol*. 2007;*581*(Pt 3):971-980.
117. Boucher JA, Abboud J, Nougarou F, Normand MC, Descarreaux M. The effects of vibration and muscle fatigue on trunk sensorimotor control in low back pain patients. *PLoS One*. 2015;*10*(8):e0135838.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lotz, J. C., Ropella, G., Anderson, P., Yang, Q., Hedderich, M. A., Bailey, J., & Hunt, C. A. (2023). An exploration of knowledge-organizing technologies to advance transdisciplinary back pain research. *JOR Spine*, *6*(4), e1300. https://doi.org/10.1002/jsp2.1300