Bioinformatics, 2026, **42**(3), btag053
https://doi.org/10.1093/bioinformatics/btag053
Published: 2 February 2026
Original Paper | Data and text mining

OXFORD

Using semantic search to find publicly available gene-expression datasets

Grace S. Brown[1,‡], James Wengler[1,2,‡], Aaron Joyce S. Fabelico[1], Abigail Muir[1], Anna Tubbs[1], Amanda Warren[1], Alexandra N. Millett[3,4], Xinrui Xiang Yu[3,4], Paul Pavlidis[3,4], Sanja Rogic[3,4], Stephen R. Piccolo[1,*]

[1]Department of Biology, Brigham Young University, Provo, UT, 84602, United States
[2]Institute of Biosciences and Technology, Texas A&M Health Science Center, Houston, TX, 77030, United States
[3]Department of Psychiatry, University of British Columbia, Vancouver, BC, V6T 0A6, Canada
[4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
*Corresponding author. Department of Biology, Brigham Young University, 701 E. University Parkway, Provo, UT, 84602, United States. E-mail: stephen_piccolo@byu.edu.
[‡]These authors contributed equally to the work.
Associate Editor: Jonathan Wren

Abstract

Motivation: **Millions of high-throughput, molecular datasets have been shared in public repositories. Researchers can reuse such data to validate their own findings and explore novel questions. A frequent goal is to find multiple datasets that address similar research topics and to either combine them directly or integrate inferences from them. However, a major challenge is finding relevant datasets due to the vast number of candidates, inconsistencies in their descriptions, and a lack of semantic annotations. This challenge is first among the FAIR principles for scientific data. Here we focus on dataset discovery within Gene Expression Omnibus (GEO), a repository containing 100 000 s of data series. GEO supports queries based on keywords, ontology terms, and other annotations. However, reviewing these results is time-consuming and tedious, and it often misses relevant datasets.**

Results: **We hypothesized that language models could address this problem by summarizing dataset descriptions as numeric representations (embeddings). Assuming a researcher has previously found some relevant datasets, we evaluated the potential to find additional relevant datasets. For six human medical conditions, we used 30 models to generate embeddings for datasets that human curators had previously associated with the conditions and identified other datasets with the most similar descriptions. This approach was often, but not always, more effective than GEO's search engine. The top-performing models were trained on general corpora, used contrastive-learning strategies, and used relatively large embeddings. Our findings suggest that language models have the potential to improve dataset discovery, likely in combination with existing search tools.**

Availability and implementation: **Our analysis code and a Web-based tool that enables others to use our methodology are available from https://github.com/srp33/GEO_NLP and https://github.com/srp33/GEOfinder3.0, respectively.**

1 Introduction

Publishers and funding agencies often mandate that researchers deposit data in public repositories so that others can verify their findings and build on their work. However, making data 'accessible' on the Internet is not enough. To accelerate progress for the research community as a whole—and to ensure that funding agencies maximize their investments—data must also be 'findable, interoperable, and reusable' (Wilkinson *et al.* 2016). Here, we focus on the first of these requirements: the ability for researchers to find datasets that have been shared publicly. Without this capability, the benefits of interoperability and reusability are reduced.

As an example of a large, highly used, public repository, we focus on Gene Expression Omnibus (GEO) (Clough and Barrett 2016), which houses data from >8 000 000 biological samples, spanning >260 000 data series. (In this paper, we use the term 'series' interchangeably with 'dataset' because the latter reflects the parlance commonly used to describe data-finding efforts, even though GEO also contains "DataSets," which are curated series.) GEO contains many types of high-throughput molecular data; transcriptomic data are the most common. Similar to

Received: 13 March 2025. **Revised:** 30 November 2025. **Accepted:** 9 January 2026

other repositories, GEO provides search capabilities. Users enter keywords, which are matched against metadata associated with samples or series. These metadata include the title, description, profiling platform, and descriptions of sample attributes. Additionally, for many studies, metadata include Medical Subject Heading (MeSH) terms, which are human-curated (Lipscomb 2000). After performing a search, users can filter results based on the species name, assay type, sample size, and other criteria. Despite the flexibility of this search tool, in our experience, these searches often yield irrelevant datasets and miss relevant ones, thus requiring time- and labor-intensive reviews and additional searches.

As one way to address this problem, researchers unaffiliated with GEO have curated lists of transcriptomic datasets on particular research topics (Ganzfried *et al.* 2013, Zimmermann *et al.* 2014, Rahman *et al.* 2016, Gendoo *et al.* 2019, Federico *et al.* 2020, Villaseñor-Altamirano *et al.* 2020). Others have annotated GEO series with labels describing their biomedical context (Shah *et al.* 2016, Wang *et al.* 2016, Hadley *et al.* 2017, Lachmann *et al.* 2018, Lim *et al.* 2021). One example of the latter is Gemma (Lim *et al.* 2021). Its curators have manually reviewed transcriptomic data series (mostly from GEO) and annotated them. Gemma focuses primarily on human, mouse, and rat datasets, especially those related to the nervous system. Gemma users can filter series based on these labels. However, due to the manual effort required to assign such labels—and Gemma's focus on particular types of datasets—this resource houses $<$10% of all GEO series. Furthermore, the annotation process is subjective. Labels might be inconsistent from one study to the next (Good and Su 2013); other domain experts might disagree with the labels; the labels might not be comprehensive; and the labeling efforts might be limited to particular research areas.

To facilitate dataset finding for GEO and other repositories, third-party researchers have created alternative ways to search for transcriptomic datasets (and other types of data). For example, 'Omics Discovery Index' and 'DataMed' provide free-text search capabilities, along with filters for publication year, species, data type, source repository, tissue type, and technology type (Ohno-Machado *et al.* 2017, Chen *et al.* 2018, Perez-Riverol *et al.* 2019). These tools support searching for datasets in GEO and other repositories. However, like GEO, their search methodologies primarily rely on keyword expansion and matching. As a result, they may struggle to account for semantic context effectively and may not handle polysemy (words with multiple meanings), synonyms, and different word forms. As an alternative, some researchers have used manual curation, natural language processing (NLP), and/or data inferences to aid in identifying series or samples based on species, cell type, tissue type, disease type, and/or assay type (Bernstein *et al.* 2017, Giles *et al.* 2017, Chen *et al.* 2019, Djordjevic *et al.* 2019, Alameer and Chicco 2022, Chua *et al.* 2022, Hawkins *et al.* 2022). Hawkins et al. (2022) used neural-network models to create an embedding—a numeric representation of words and phrases—for each sample. Such approaches are capable of representing data semantics, even when words are misspelled or when different terms are used to describe the same concepts (Lake and Murphy 2023). However, their approach focused on identifying and characterizing samples, whereas our goal is to support data finding at the series level, where longer-form descriptions are available. PEPHub uses a language model to support finding GEO series, but they did not report benchmarks against other techniques (LeRoy *et al.* 2024). Patra et al. (2020) developed a system for recommending GEO series based on embeddings of research articles that a scientist had published previously. However, a researcher's past publications may not coincide directly with future interests, and this approach may not be effective for early-career researchers.

For this study, we assume that a researcher has used existing tool(s) to find GEO series on a given research topic and wishes to find additional ones. We hypothesized that we would be able to detect semantically similar datasets based on descriptions of those datasets. We focus on methodologies that summarize text as embeddings. In recent years, computational researchers have developed diverse methodologies for training embeddings. Early examples include Word2Vec and GloVe (Mikolov *et al.* 2013, Pennington *et al.* 2014, Devlin *et al.* 2018). These algorithms train neural networks on large text corpora. Given a particular word (or subword), they characterize the context in which that (sub)word is used, based on surrounding (sub)words, in a context window. More recently, BERT and transformer-based models have become widely used for NLP (Devlin *et al.* 2018, Wolf *et al.* 2020). The attention mechanisms of these techniques account for relationships between a given (sub)word and all words in a given sentence, regardless of their distance from the (sub)word. Researchers have trained such models using general corpora like Wikipedia, books, news articles, and other sources. Others have trained models using domain-specific corpora, including some specific to biology and/or medicine (Alsentzer *et al.* 2019, Chen *et al.* 2020, Lee *et al.* 2020, Harnoune *et al.* 2021, Remy *et al.* 2024). Furthermore, it is possible to build on pre-trained models using a process known as fine-tuning. With this approach, a researcher uses an existing model—which might have been trained on billions of examples—and performs additional training with domain-specific examples. This strategy yields significant cost and energy savings compared to training a model from scratch and may yield better results for NLP tasks in that domain.

Prior studies have shown that relatively large corpora are often better than smaller ones (Chiu *et al.* 2016) and that training on domain-specific corpora is often better than using general corpora (Wang *et al.* 2018, Chen *et al.* 2020). However, the fast pace with which this field is changing necessitates additional research. We performed a benchmark study, starting with relatively simple, frequency-based approaches and then evaluating 30 language models ranging in complexity from continuous bag-of-words models to transformer-based models. We applied these models in the context of the GEO dataset finding, in which data submitters have provided ad hoc descriptions that differ in their length, semantics, and level of detail. One challenge is that, in a typical scenario, the number of datasets relevant to a particular research topic is small (perhaps between 5 and 100) compared to the total number of GEO datasets (100 000 s). Therefore, we have evaluated the effects of this imbalance. Our focus is not necessarily on identifying particular model(s) or optimization techniques that outperform all others. Instead, our

goal is to shed light on the extent to which modern language models have the potential to aid with dataset discovery.

2 Methods

2.1 Annotated data collection

On 18 April 2024, we used *geoFetch* (version 0.12.5) (Khoroshevskyi *et al.* 2023) and custom Python code (https://python.org, version 3.12) to retrieve metadata associated with all available GEO series. We retained series that matched the following criteria:

- Were 'not' marked as "retired."
- Were listed as coming from *homo sapiens* samples.
- Were associated with the "expression profiling by array" experiment type and used an Affymetrix, Illumina, or Agilent platform OR were associated with the "expression profiling by high throughput sequencing" experiment type and used an Illumina platform.
- Were 'not' a GEO SubSeries. We excluded these to avoid bias in our machine-learning analysis (having identical descriptions in the reference and comparison groups) and because it would be easy for researchers to identify a SubSeries if they have identified its associated SuperSeries.

A total of 48 893 series matched these criteria. For each of these series, we used diverse models to generate numeric vectors (embeddings) that summarized the text describing the series. When creating these vectors, we used the title, summary, and overall design (when available) as inputs. Before doing so, we converted the text to lower case and removed URLs, non-alphanumeric characters, and extra white space; we also removed HTML tags using the Beautiful Soup (version 0.0.2) package (https://www.crummy.com/software/BeautifulSoup/bs4/doc). Because the *fastText* and frequency-based models (described below) do not rely on (and may be distracted by) stop words when inferring semantic meaning, we removed stop words, as defined by the *nltk* package (version 3.8.1) (Bird *et al.* 2009). For the other models—most of which are transformer-based—we retained stop words because those models capture the semantic context of words based on the entire sentence or document. When text inputs were longer than 256 characters, we sometimes (details below) split the text into chunks, allowing for 20 characters of overlap between chunks. We chose 256 characters as the threshold because the smallest embedding size was 300 characters and because some models truncate inputs of approximately this size.

On 19 April 2024, we queried Gemma (Lim *et al.* 2021) to identify all GEO series that had been annotated in that resource and that overlapped with the 48 893 we had identified ($n = 5997$). We also selected series that human curators had associated with one of six medical conditions: juvenile idiopathic arthritis, triple negative breast carcinoma, Down syndrome, bipolar disorder, Parkinson's disease, and neuroblastoma. Some of these series coincide with our research interest in cancer; others coincide with Gemma's focus on neurological conditions. Together, these series reflect a range of sample sizes (some conditions have been more highly studied than others).

2.2 GEO advanced search builder queries

On 8 May 2024, we used GEO's Advanced Search Builder to identify datasets (series) associated with each of the six medical conditions. We attempted three types of queries using:

1) The name of the medical condition as a key phrase.
2) The name of the medical condition plus synonyms for the medical condition as keyphrases. The synonyms were defined by the Mondo Disease Ontology (Vasilevsky *et al.* 2020), which we queried using BioPortal (Noy *et al.* 2009).
3) The MeSH Subject Heading term (Lowe and Barnett 1994) for the condition.

For #1 and #2, GEO sometimes expanded our queries by mapping them automatically to MeSH terms. For #3, the MeSH Browser showed the term *Arthritis, Juvenile* (D001171), but this option (or other variations on it) did not appear in GEO. Therefore, we used the broader MeSH term *Arthritis* (D001168). Similarly, GEO did not provide an option for triple-negative breast cancer, even though the MeSH Browser showed a term for *Triple Negative Breast Neoplasms* (D064726). Therefore, we searched using the broader term *Breast Neoplasms* (D001943). Our source-code repository contains files with the exact queries that we used and the results of each query.

2.3 Experimental design

For each medical condition, we identified all datasets annotated for that condition in Gemma. We randomly assigned half of these to a "reference set" and half to a "comparison set A" (Fig. 1). When there was an even number of datasets, the reference set and comparison set A were the same size. When there was an odd number of datasets, the reference set contained one more dataset than comparison set A. From the remaining Gemma datasets (including those from other medical conditions), we created "comparison set B". Using each medical condition's reference set, we averaged the embeddings from the individual datasets and used the cosine-similarity metric to calculate the distance between this embedding and each dataset in comparison sets A and B. We used cosine similarity because it is easily calculated, interpretable, invariant to the magnitude of vectors being compared, and widely used in NLP studies. Finally, we ranked the datasets in the comparison sets based on cosine similarity, with the expectation that more effective language models would rank datasets from comparison set A higher than those from comparison set B.

For the language model that performed best according to the Gemma annotations, we performed a second round of validation. For each medical condition, we calculated an average embedding across 'all' datasets annotated for that condition in Gemma. Next, we calculated the pairwise cosine similarity between the average embedding and each human transcriptomic dataset 'not' in Gemma. Two curators (authors ANM and XXY) reviewed the top-ranked datasets to assess biomedical relevance. As a preliminary step, we selected five datasets for each medical condition from the datasets ranked 51st through 55th and used these for a pilot analysis. Additionally, we included

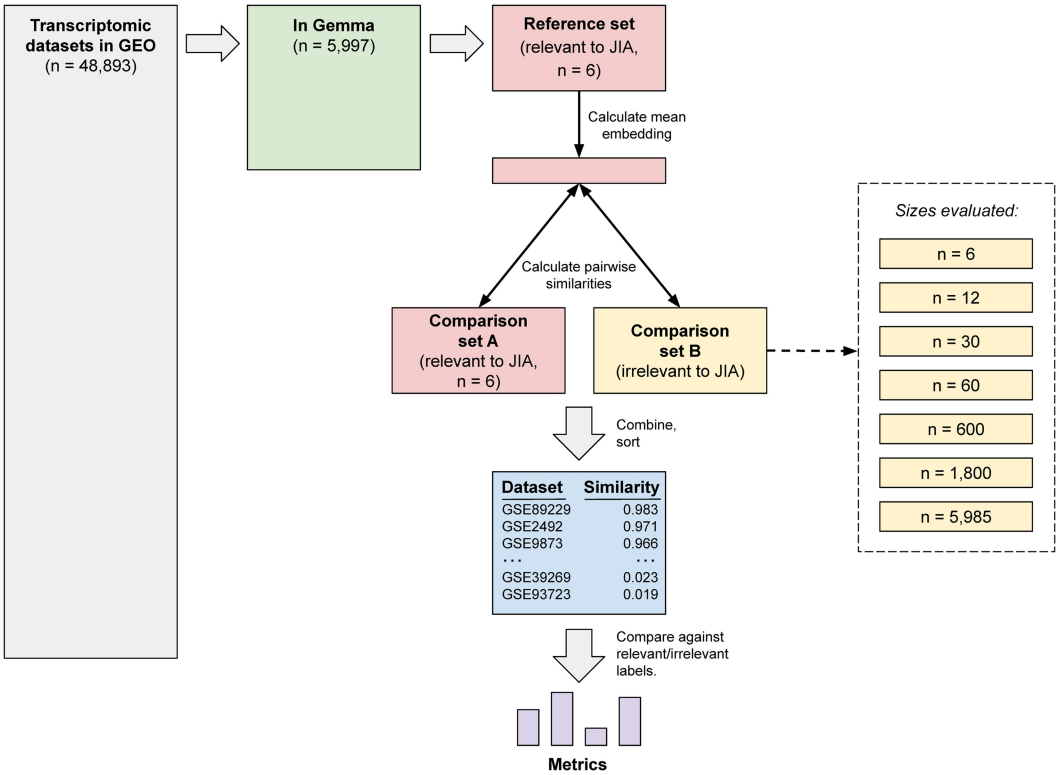

Figure 1 Evaluation process for comparing language models' abilities to identify relevant datasets. This diagram illustrates the process of evaluating the language models' performance for a particular medical condition. Here, we use juvenile idiopathic arthritis (JIA) as an example. After identifying transcriptomic datasets in GEO, we selected those that had been annotated in Gemma as relevant to the medical condition. We randomly divided these datasets into two evenly sized groups: a reference set and comparison set A. We assigned datasets not annotated as relevant to the medical condition to comparison set B. We varied the size of comparison set B to provide insight into the effects of imbalance between the numbers of relevant and irrelevant datasets. Using a given language model, we generated embeddings for datasets in the reference set and both comparison sets. We calculated the mean embedding for the reference set and then calculated the cosine similarity between this mean embedding and the embedding for each dataset in comparison sets A and B. Finally, we calculated metrics to quantify the extent to which each language model ranked relevant datasets higher than irrelevant ones. (Boxes are not necessarily drawn to scale).

5 datasets as distractors (not relevant to any of the six medical conditions). Prior to their reviewing these datasets, we created a rubric for the curators (Additional Data File 1, available as supplementary data at *Bioinformatics* online). We then asked each curator to independently review the title, summary, and overall-design descriptions and record whether each dataset was relevant to any of the conditions. The curators' conclusions were identical for all datasets in the pilot.

We randomly divided the top-50 ranked datasets for each medical condition, assigning half uniquely to each curator while assigning 5 to both curators to facilitate an inter-rater agreement evaluation. (Later, we found that three datasets had been selected twice for a given curator, so we removed the duplicate copies.) We created a spreadsheet with information about these datasets for each curator, removed the medical-condition labels, and asked the curators to indicate which medical condition they believed was most relevant to each dataset. Although we did not include distractors in this validation step, we provided an option for the curators to indicate that a given dataset was not relevant to 'any' of the six medical conditions. As with the pilot, the curators were blind to the rankings made by the language model. Finally, we repeated this process for the same number of top-ranked datasets identified via GEO's Advanced Search Builder. For the language model, these conclusions agreed in 54 cases (87.1%). For GEO's Advanced Search Builder, the conclusions agreed in 55 cases (88.7%).

2.4 Language models

Table S1, available as supplementary data at *Bioinformatics* online, provides a summary of the language models we evaluated. For most models, we used the implementation in the Hugging Face repository and the *transformers* package (version 4.37.2) (Wolf *et al.* 2020). For other models, we used the *fasttext* (version 0.9.2) (Bojanowski *et al.* 2016, Mikolov *et al.* 2018) or *openai* (2.8.1) (https://github.com/openai/openai-python) packages. In Table S1, available as supplementary data at *Bioinformatics* online, we categorize each model based on the type of corpus on which it was trained (general purpose, biomedical, or scientific defined more broadly) and the high-level methodology used for training, as described in Hugging Face. We also describe any fine-tuning that was performed. In addition to these models, we implemented a simple "word overlap" method. With this approach, we first identified all unique words in each dataset description. Second, for each pair of datasets, we counted the number of unique words spanning both datasets. Then we

calculated the proportion of these words that overlapped between the two dataset descriptions. Additionally, we tested the *BM25* text-retrieval algorithm, which uses a probabilistic framework to rank documents based on the frequency and distribution of query terms (Robertson and Zaragoza 2009). For this algorithm, we used the *bm25s* Python package (version 0.2.14) (Lù 2024), which uses the "lucene" variant of the algorithm (Kamphuis *et al.* 2020). We applied this algorithm in two ways: (i) the package's default approach, which does not remove stop words or perform stemming and (ii) removing stop words and performing stemming ("bm25plus").

2.5 Data and code availability

All data and code used to perform this analysis have been deposited in a GitHub repository, which is publicly available and uses the MIT License (https://github.com/srp33/GEO_NLP). An archived version is available from https://zenodo.org/records/17762289. We used the Python programming language (version 3, https://python.org) and the R statistical software (version 4.4.1) for the analysis. Additionally, we used the *tidyverse* to process and visualize data (Wickham *et al.* 2019). To facilitate reproducibility, we executed the analyses within a Docker container (Piccolo and Frampton 2016).

3 Results

We assume that a researcher has found 'some' transcriptomic datasets relevant to a given research topic and wishes to find more datasets relevant to the same topic. They might have found datasets previously using a literature search, GEO's search tools, or some other means. To facilitate finding additional datasets, we used various techniques to independently summarize the descriptions of each dataset as a numeric vector or embedding. These techniques included frequency-based methods and 30 distinct language models. To compare these approaches, we identified six medical conditions for which annotated datasets were available in both GEO and Gemma: juvenile idiopathic arthritis ($n = 12$), triple negative breast carcinoma ($n = 24$), Down syndrome ($n = 30$), bipolar disorder ($n = 34$), Parkinson's disease ($n = 109$), and neuroblastoma ($n = 121$). For each medical condition, we randomly divided the datasets into a 'reference set' and a 'comparison set A'; the remaining datasets that were available in both GEO and Gemma served as "comparison set B" (Fig. 1). After averaging the embeddings in the reference set, we calculated the cosine similarity between the average embedding and each dataset in the comparison sets. For each model, we used the area under the precision-recall curve (AUPRC) to quantify the extent to which the model ranked the datasets from 'comparison set A' higher than the datasets from 'comparison set B'.

The AUPRC varied significantly across medical conditions (Kruskal–Wallis test, $P < .001$; Fig. 2). Median performance was highest for Parkinson's disease and Down syndrome, while it was lowest for triple-negative breast carcinoma and neuroblastoma. The AUPRC also varied significantly across the models ($P < .001$; Fig. 3). Ten of the models consistently performed poorly, rarely attaining AUPRC values higher than 0.10. These included the continuous bag-of-words (CBOW) models, which are relatively simple, as well as some transformer-based models that were trained using biomedicine-specific corpora and/or that used large embedding sizes. In all cases except one, these poor-performing models scored lower than our simple "word overlap" technique. After excluding non-embedding methods, we found that the median AUPRC was significantly correlated with embedding size (Spearman's test, $\rho = 0.51$, $P = .004$) but did 'not' differ significantly by training corpora type (Mann–Whitney U test, $P = .46$; Fig. 4). Next, we examined whether sample size was associated with AUPRC by fitting a linear mixed-effects model (Bates *et al.* 2015), treating sample size as a fixed effect and language model as a random effect. Sample size did not significantly predict AUPRC ($P = .26$), indicating that, across models, increases in sample size were not statistically associated with the ability to identify relevant data series.

Next, we assessed the effects of text chunking. When input texts are relatively large, practitioners sometimes split texts into smaller chunks, create an embedding for each chunk, and then

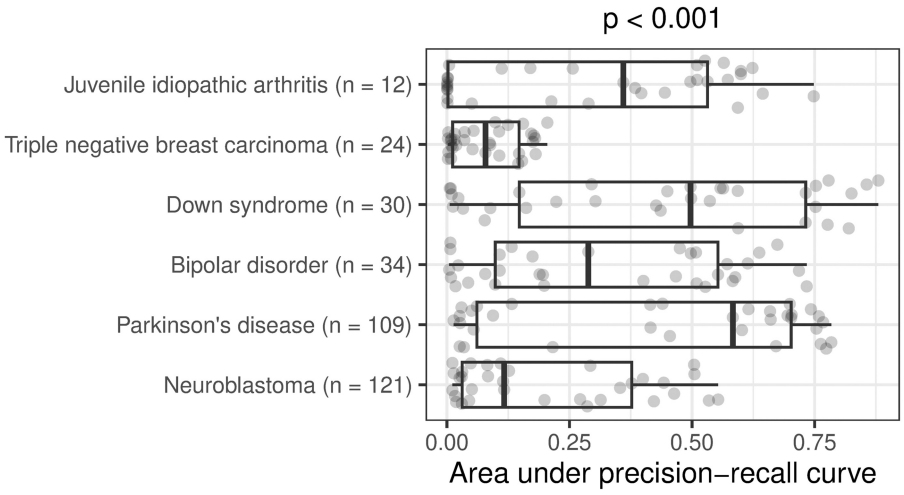

Figure 2 Within-Gemma performance of language models by medical condition. We used language models to identify Gemma datasets that had been annotated as relevant to six medical conditions. The area under the precision-recall curve differed significantly across the conditions. The p-value was calculated using the Kruskal–Wallis test.

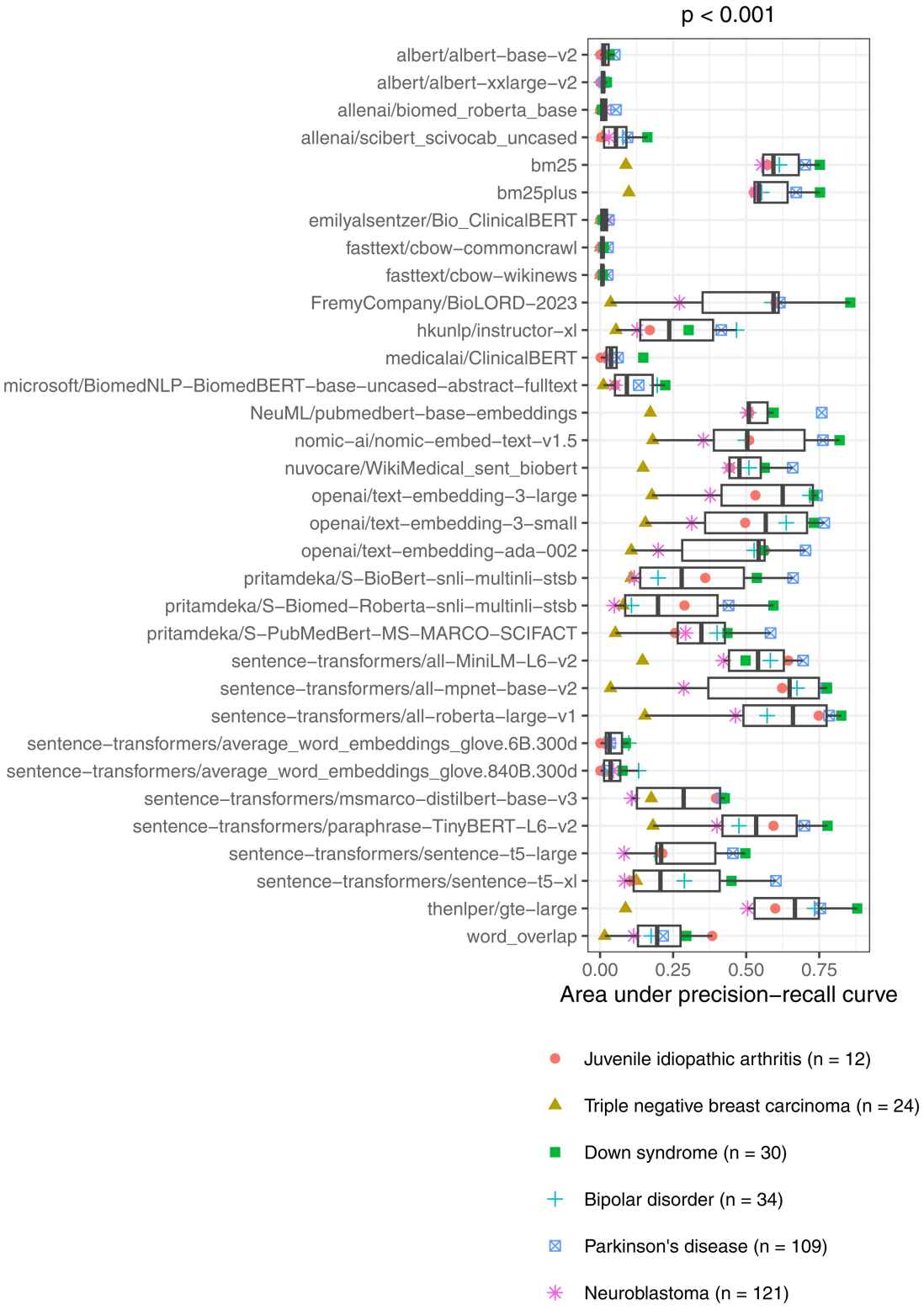

Figure 3 Within-Gemma performance of language models by model. We used language models to identify Gemma datasets that had been annotated as relevant to six medical conditions. The area under the precision-recall curve differed significantly across the models. The *P*-value was calculated using the Kruskal–Wallis test.

average the embeddings. We tried this technique and compared it against the performance we attained without chunking. We found that 'not' chunking nearly always performed 'better' than chunking (median reduction in AUPRC: 0.12; Table S2, available as supplementary data at *Bioinformatics* online). Additionally, we assessed the effects of converting the text to lowercase characters before generating embeddings. Using the same reference and comparison sets, we repeated the steps of generating

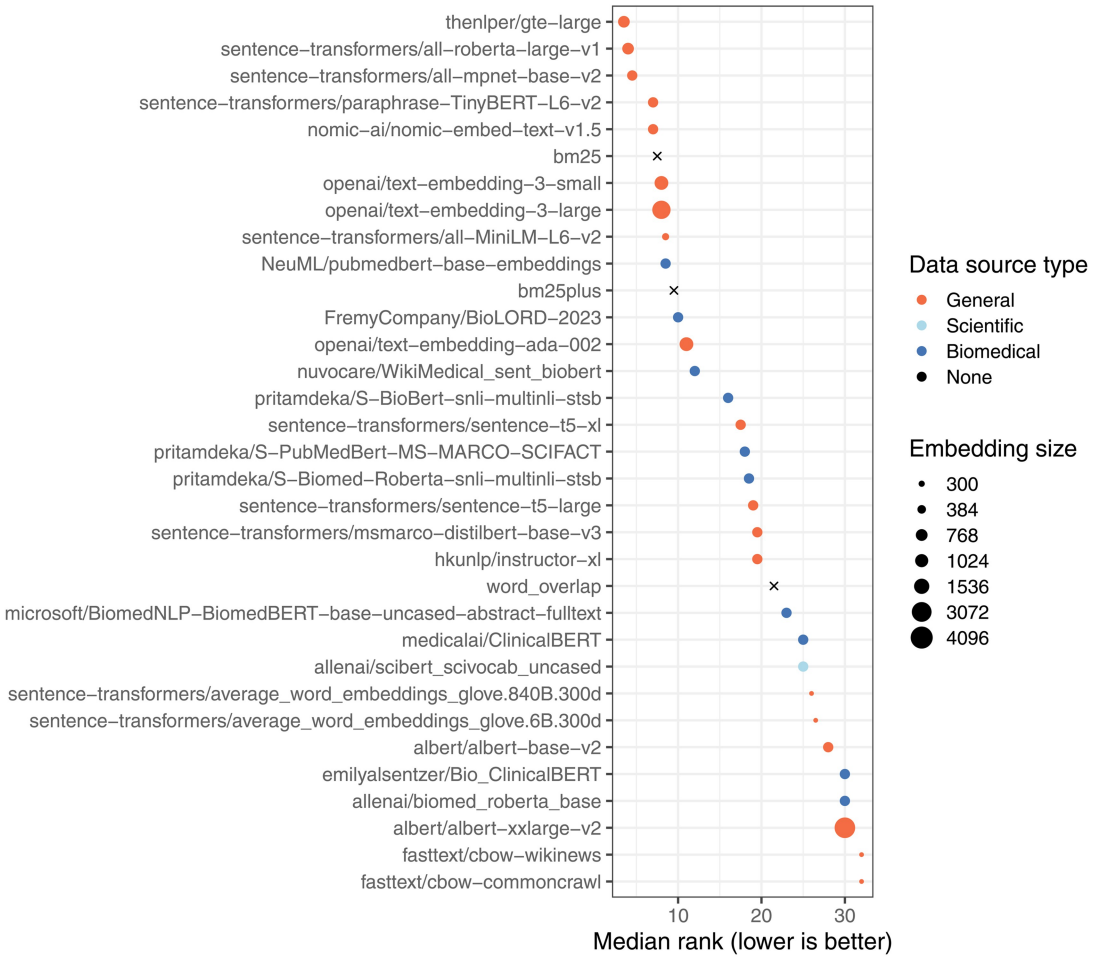

Figure 4 Within-Gemma performance of language models by model, ranked. We used language models to identify Gemma datasets that had been annotated as relevant to six medical conditions. For each medical condition, we ranked the models and calculated the median rank across the conditions. Models with relatively low ranks performed better overall than models with relatively high ranks.

embeddings, identifying similarities between GEO series for a given medical condition, and comparing against the Gemma annotations. In some cases, the median AUPRC dropped (by as much as 0.038 in the case of `openai/text-embedding-3-small`); in other cases, the median AUPRC increased (by as much as 0.036 in the case of `pritamdeka/S-Biomed-Roberta-snli-multinli-stsb`). However, in most cases, the median difference was smaller than 0.01 (Table S3, available as supplementary data at *Bioinformatics* online).

Across the six medical conditions, the top-performing models were `thenlper/gte-large`, `sentence-transformers/all-roberta-large-v1` and `sentence-transformers/all-mpnet-base-v2`. All three were trained using general corpora (not science- or biomedicine-specific) and used moderately sized embeddings (768—1024 dimensions). According to the Hugging Face documentation, `thenlper/gte-large` uses multi-stage contrastive learning and was trained by the Alibaba DAMO Academy. It uses the Bidirectional Encoder Representations from Transformers (BERT) framework (Devlin *et al.* 2018) and has performed well in the Massive Text Embedding Benchmark (Cao 2024). The `sentence-transformers/all-roberta-large-v1` model builds on a pre-trained Roberta model (Liu *et al.* 2019) and was fine-tuned on a dataset with one billion sentence pairs. It uses a self-supervised contrastive learning strategy. The `sentence-transformers/all-mpnet-base-v2` model also builds on a pretrained model, was fine-tuned on one billion sentence pairs, and uses self-supervised contrastive learning. All three models are popular on Hugging Face. For example, in July 2024, each had been downloaded at least 672 000 times. Among other intended uses, all three models are described as being useful for sentence-similarity tasks.

We compared the language models' performance against what was returned using GEO's Advanced Search Builder (in three variations, see Section 2). We used recall (sensitivity) as a metric to quantify the number of search results that needed to be returned before *n* relevant datasets were identified. For juvenile idiopathic arthritis, Down syndrome, and bipolar disorder, at least one of the top-performing language models identified all relevant datasets within the top-100 results (Figure S1, available as supplementary data at *Bioinformatics* online), whereas the GEO tool returned a limited set of potential matches and thus often identified relatively few of the relevant datasets. All of the top-performing language models attained a recall of at least 0.75 for the top-1000 datasets. Furthermore, these models nearly always outperformed Advanced Search Builder (Fig. 5). One exception was triple-negative breast carcinoma, in which

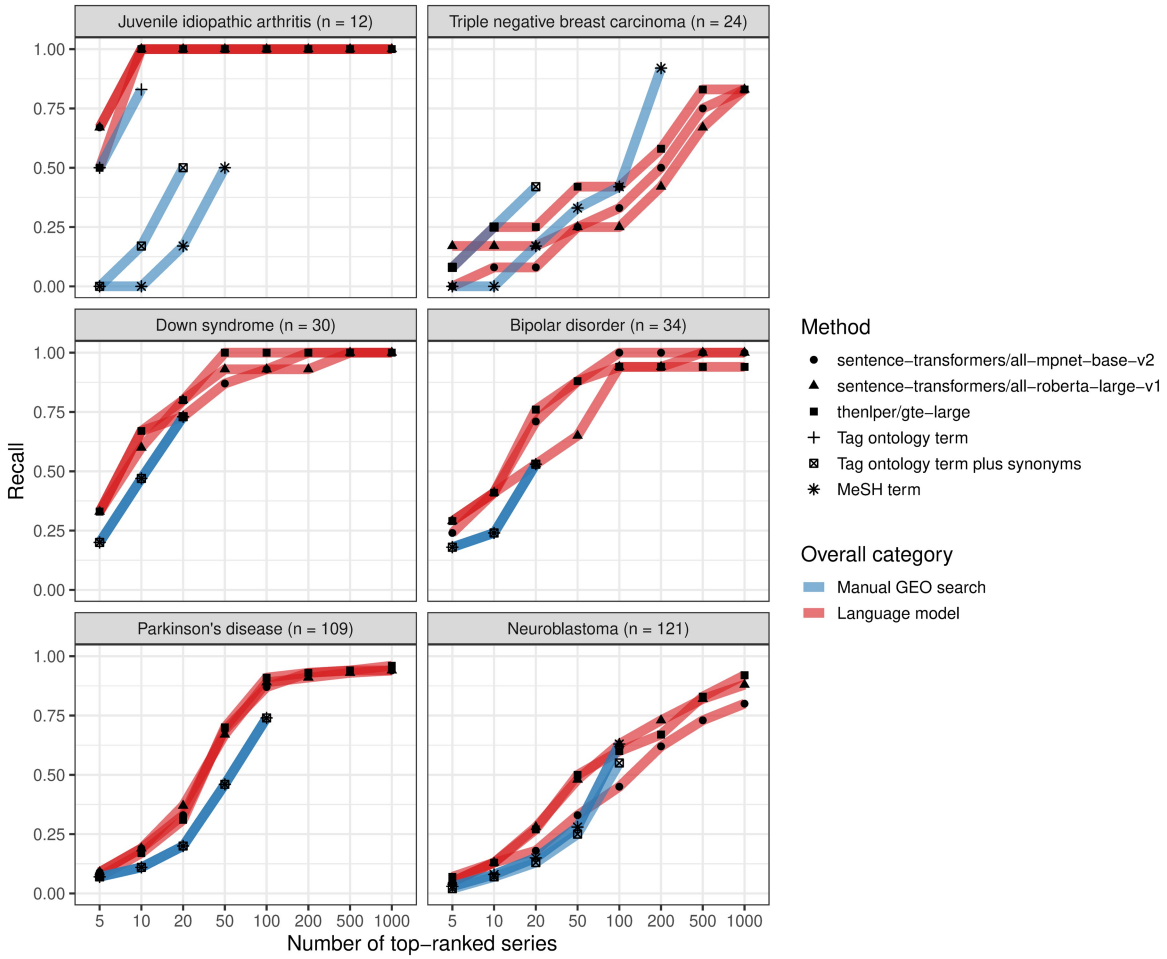

Figure 5 Comparison between manual GEO searches and language-model rankings for within-Gemma series. After manually querying GEO and using language models to identify Gemma series that had been annotated as relevant to six medical conditions, we compared the top-*n* results returned using either approach. In many cases, the GEO search tool returned <1000 datasets; therefore, this graph depicts the maximum *n* value that fell below the total number of GEO results for a given medical condition. A recall value of 1.0 indicates that all Gemma-annotated datasets were included in the top-*n* results.

Advanced Search Builder sometimes performed better than the top-performing language models, particularly when using the high-level MeSH term *Breast Neoplasms*.

One challenge of dataset finding in this context is that relevant datasets are rare. For example, in our evaluations based on Gemma annotations, we had metadata for 5997 GEO series, but only 12 of these had been annotated as relevant to juvenile idiopathic arthritis. After splitting the data into reference and comparison sets, we were searching for 6 series among 5991 candidates (a ∼1000-to-1 ratio). This imbalance would only increase when searching all GEO series, including those not in Gemma. Therefore, we evaluated the extent to which different levels of imbalance would affect our findings. For juvenile idiopathic arthritis, the AUPRC remained constant for the top-performing models when the ratio was 300-to-1 or smaller. However, the performance dropped considerably when all Gemma candidates were included (Fig. 6). We observed similar patterns for other medical conditions, with the greatest reduction in performance for triple-negative breast carcinoma.

To gain additional insight into the top-performing model's performance (`thenlper/gte-large`), we compared the length (number of characters after cleaning steps) of the dataset descriptions versus the cosine similarity scores for the GEO series in our comparison sets. We surmised that datasets with longer descriptions might be more often ranked as relevant. However, we did 'not' observe a significant correlation between these variables (Figure S2, available as supplementary data at *Bioinformatics* online).

One ancillary application of our approach is to identify Gemma series that lack relevant annotations. We identified the 25 GEO series that were most often ranked highly (across all models) for a given medical condition but were not tagged as relevant in Gemma. We reviewed these datasets and judged whether there was a case for annotations to be added to Gemma. We concluded that there might be a case for nine of these datasets (Additional Data File 2, available as supplementary data at *Bioinformatics* online). The Gemma team agreed that the "juvenile idiopathic arthritis" ontology term should be added for GSE13501. For GSE149632, we found that the "Parkinson disease" term had been added after we downloaded Gemma annotations and made our predictions. For 6 datasets, the Gemma curators agreed that the GEO series were relevant to

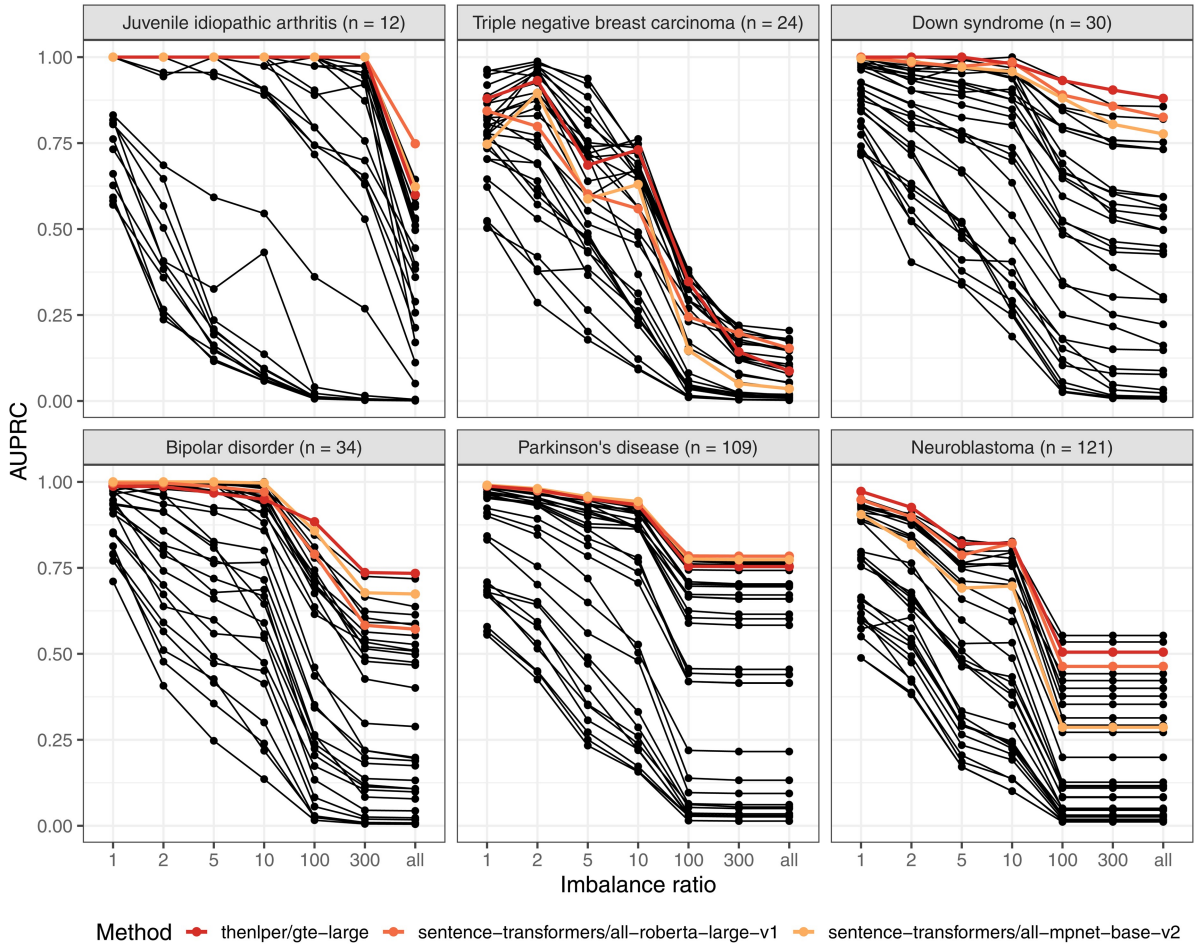

Figure 6 Within-Gemma performance of language models for different levels of imbalance between irrelevant and relevant data series. We used language models to identify Gemma datasets that had been annotated as relevant to six medical conditions. For each medical condition, we performed simulations with increasing levels of imbalance between irrelevant and relevant data series. An imbalance ratio of 1 means that the "other" set had the same number of series as the test set. An imbalance ratio of 10 means that the "other" set had 10 times as many series as the test set. When the imbalance ratio was "all," we included all available series in the "other" set that were not part of the training or test sets. The lines with color represent the top-3 performing models overall.

the specified medical condition; however, these datasets were from cell lines. Current practice at Gemma is to annotate such datasets with the ontology term for the cell line; inferences regarding the associated medical conditions can be made via querying the respective ontologies. GSE8650 was tagged with "systemic juvenile idiopathic arthritis," which is a distinct subtype of the juvenile idiopathic arthritis (Mellins *et al.* 2011). Researchers looking to study the condition more broadly might be interested in using this dataset, but they would need to perform ontology-based inference in Gemma to find it. For five additional datasets, we agreed that it was not appropriate to add annotations for the medical conditions we evaluated; however, the review process pointed to annotations that should be added or removed. For example, the language models suggested that GSE19697 was relevant to triple-negative breast carcinoma, yet the samples were from patients with basal-like breast carcinoma. These conditions have etiological overlap (Seal and Chia 2010), but they are distinct. The curators annotated the dataset in Gemma with the "basal-like breast carcinoma" term.

To evaluate the potential to find additional datasets relevant to these medical conditions, we used `thenlper/gte-large` to rank all non-Gemma GEO datasets that matched our filtering criteria. First, we created an average embedding for each medical condition, using descriptions from Gemma-annotated datasets as inputs. Second, we ranked each non-Gemma dataset based on the cosine similarity between its embedding and the averaged Gemma embedding for that condition. The Gemma curators reviewed the 50 datasets with the highest similarity per medical condition; these datasets were divided between the curators, with some overlap to support an evaluation of inter-rater agreement (Section 2). The curators reviewed the dataset descriptions and selected the medical condition that they believed was most relevant. In cases where the curators deemed it was likely that a dataset was relevant to a specified condition, but they lacked certainty, they recorded a "maybe" response. If they believed none of the datasets was relevant, they selected "Other" (see Section 2). We analyzed the datasets that the curators confidently classified as relevant to a particular medical

condition. Taking into account both reviewers' responses, the datasets selected by the language model were relevant 62%–63% of the time, whereas only 50%–53% of the datasets selected by GEO's search tool were relevant (Table 1; Additional Data File 3, available as supplementary data at *Bioinformatics* online). These percentages differed considerably across the medical conditions (Fig. 7); a generalized linear model that accounted for medical condition and curator effects confirmed that the language model significantly outperformed the GEO tool (odds ratio: 1.97; 95% confidence interval: 1.37–2.87; $P < .001$). For Down syndrome, juvenile idiopathic arthritis, neuroblastoma, and triple-negative breast carcinoma, the language model produced more relevant results than GEO's search tool. However, for Parkinson's disease, GEO performed slightly better. For bipolar disorder, GEO returned only 21 results, and the curators agreed that these datasets were relevant 58.6% of the time. For the top-50 datasets returned by the language model for bipolar disorder, the curators agreed that the datasets were relevant only 19.3% of the time.

This table indicates the number of times that each of two human curators indicated that a given GEO series was relevant to one of six medical conditions for non-Gemma series. It also indicates the number of times the curators deemed it was 'likely' that a dataset was relevant to one of the conditions but they lacked certainty ("maybe" responses). The percentages indicate the frequencies with which the medical condition predicted by the top-performing language model (`thenlper/gte-large`)—or GEO's Advanced Search Builder—agreed with the medical condition selected by the human curators.

Many of the datasets categorized as irrelevant are instructive. For example, many predicted as relevant to bipolar disorder were instead found to pertain to schizophrenia or major depressive disorder; these conditions have overlapping features and shared genetic or neurobiological factors. Some predicted as relevant to juvenile idiopathic arthritis were focused on rheumatoid arthritis or pediatric systemic lupus erythematosus. These conditions share overlapping pathophysiological processes, but

Table 1 Comparisons of automated predictions with curators' conclusions.

Search method	Curator	# yes	% yes—relevant	# maybe	% maybe—relevant
Language model	1	156	62.2	24	95.8
Language model	2	171	63.2	9	100.0
GEO search	1	139	50.4	27	92.6
GEO search	2	159	52.8	7	85.7

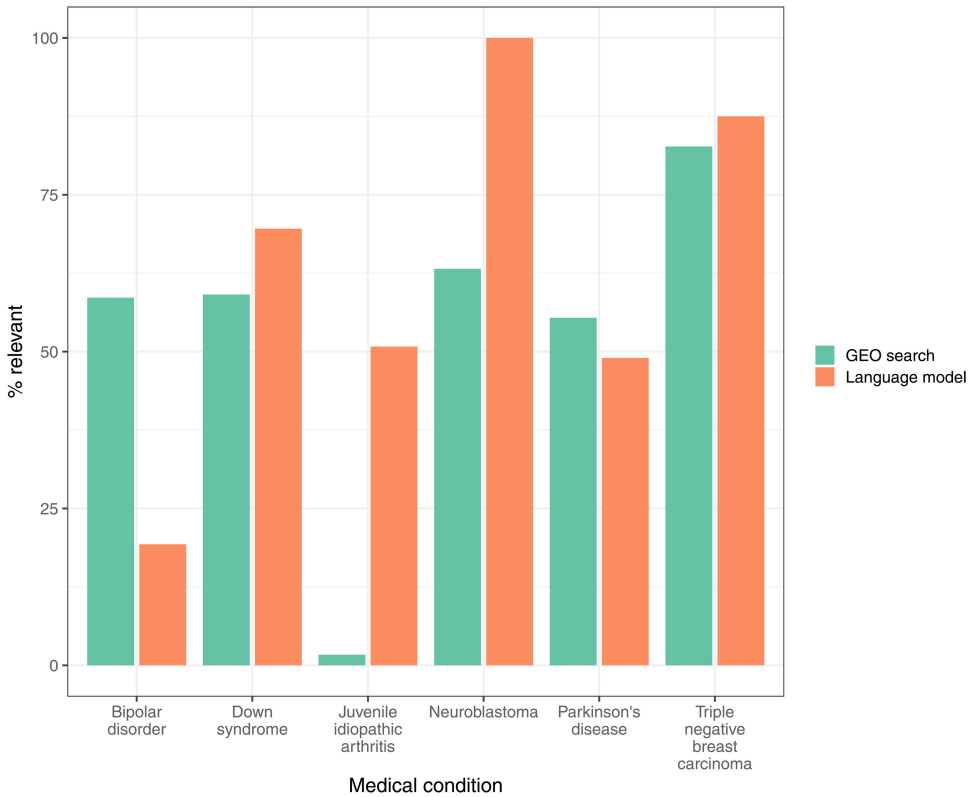

Figure 7 Agreement between curators' responses and top search results for non-Gemma series. We reviewed the top series returned by GEO's Advanced Search Builder or the thenlper/gte-large model for each medical condition. Curators manually reviewed the GEO descriptions and judged whether each series was relevant to a given medical condition or was "maybe" relevant. For bipolar disorder, the GEO search returned only 21 series.

the curators did not categorize them as relevant because they are distinct conditions.

Of the 260 unique datasets returned by either search method that the curators deemed to be relevant, only 41 (15.8%) overlapped between the two methods. This finding suggests that the two search methods are complementary and thus can be used together.

Given the `thenlper/gte-large` model's ability to identify relevant datasets, we created *GEOfinder*, a Web application that enables researchers to query GEO. We designed this application with the idea that researchers will first use GEO's Advanced Search Builder to identify some datasets relevant to their research topic. When a researcher performs a search in GEO, it returns a list of datasets and provides an option to select a checkbox next to each dataset. It also provides an option for users to download a file with information about the selected datasets. After completing these steps for datasets of interest, the researcher can upload the file into GEOfinder and search for datasets that have similar descriptions. GEOfinder uses language embeddings and cosine similarity to rank candidate datasets. GEOfinder can be found at https://bioapps.byu.edu/geofinder. Its source code is available from https://github.com/srp33/GEOfinder3.0.

4 Discussion

This project was motivated by our prior experiences searching for gene-expression datasets relevant to specific human diseases. Those efforts were time and labor-intensive, and we sensed that we were missing many available datasets. GEO's Advanced Search Builder supports queries based on keywords or keyphrases, accompanied by filtering options, whereas our study is based on the assumption that relatively long, narrative descriptions of datasets characterize their research context and thus should be more helpful for dataset finding. Even though individual researchers who deposit datasets vary widely in the ways they describe data and experiments, those researchers have domain expertise on the topic they are studying. Thus, we expect that the language they use generally reflects the biomedical context of each study and thus should be differentiable from the language used in other contexts.

GEO does provide a collection of "DataSets" that have been human-curated and annotated and thus are more easily searchable. However, as of 13 January 2025, only 4348 Datasets had been curated, whereas 244 358 GEO series existed. Another option is MetaSRA, which contains annotations for 718 384 samples, most of which are from RNA-Sequencing experiments (Bernstein *et al.* 2017). These annotations were generated using a semi-automated process; however, the last update occurred in 2020, and RNA-Sequencing experiments constitute only a portion of GEO. Although it would be preferable for human curators to annotate all GEO series and samples retroactively, this is unlikely due to the time and resources that would be required. Therefore, we need methodologies that do not require manual annotation by experts. One innovation of our approach is that it essentially places a curation responsibility on the researchers who wish to find data. After they have identified 'some' existing datasets that align with their research topic, the metadata from those datasets serve as "ground truth" examples in the search for additional ones.

Our goal was not necessarily to supplant existing search tools, which are effective in many cases and are often ontology-backed. Nor do we claim that one or a few language models are better than all others. It would be infeasible to make this claim due to the massive number of available models and the rapid pace at which modeling methodologies and training corpora are changing. Rather, our goals were (i) to explore the possibility of using language models to aid with dataset finding, (ii) to evaluate how much the models would differ from each other in their performance, (iii) to assess how well they would perform in different biomedical contexts, and (iv) to compare them against GEO's Advanced Search Builder. In doing so, we found that the language models often returned very different results than GEO's tool, thus supporting the idea that these approaches are complementary.

Some models performed quite well and dramatically better than others. Generally, models with larger embedding sizes performed better than those with smaller sizes, but this observation is confounded because embedding sizes have co-evolved with modeling methodologies and data sources. It is possible that with additional optimizations—such as using a different chunking strategy or fine-tuning based on human feedback—the model rankings would change considerably. Performance also varied according to the medical condition being studied. One might presume that increasing the number of examples in a reference set would lead to better precision and recall. However, our methodology achieved excellent performance for juvenile idiopathic arthritis, the condition with the fewest examples of the six we studied. This suggests that accurately identifying new datasets may depend less on the number of examples and more on the distinctiveness of the language used to describe a particular medical condition.

Our study was limited to human data and datasets related to medical research. Currently, the extent to which our findings generalize to other contexts is unclear. An additional limitation is that the manual reviews were somewhat subjective. We conceptualized medical relevance as whether researchers studying a given medical condition might be interested in using the data to study that condition (Additional Data File 1, available as supplementary data at *Bioinformatics* online). However, it was sometimes difficult to make this distinction. For example, when evaluating neuroblastoma, a form of cancer that affects nerve cells, we identified many GEO series that had used neuroblastoma cell lines, but not necessarily to study neuroblastoma itself. For example, Schartner et al. (2017) used the SH-SY5Y cell line to study the downstream effects of a single-nucleotide variant, irrespective of any medical condition. Fardin et al. (2009) used neuroblastoma cell lines to study differences between normoxic and hypoxic conditions; they also evaluated the benefits of using a particular algorithm for detecting relevant genes. Although their overarching goal was to shed light on neuroblastoma development, that connection was indirect for this study.

Although our findings are promising, there are opportunities for further exploration. Our method uses only high-level dataset descriptions as inputs. In contrast, PEPHub (LeRoy *et al.* 2024) uses sample-level metadata to construct embeddings. The molecular data from these studies may also provide useful

information. Combining some or all of this information may yield better results for dataset finding. When searching for semantically similar datasets, we identified multiple datasets known to be relevant and averaged their embeddings before searching for other datasets with similar embeddings. A better alternative might be to keep the embeddings separate and train a classification algorithm to differentiate between relevant and irrelevant examples.

Despite these limitations, our findings provide strong evidence that language models hold promise for enhancing the retrieval of publicly available datasets. Our evaluation, which included a diverse set of models, offers an impartial comparison, as we did not contribute to the development of any models under investigation. Beyond the quantitative improvements that language models appear to offer, their use has practical advantages. It does not require a systematic effort to curate the GEO series and map them to MeSH terms, as is done for the GEO search engine. It also circumvents a key limitation of keyword searches: researchers must formulate queries using specific terms that may be difficult to guess and that may vary substantially across datasets. Additionally, semantic search may be more robust to terminology drift—the tendency for terms describing biomedical concepts to change over time—which is a critical issue for repositories like GEO that have accumulated data over decades from a broad user base. Our results motivate further investment in semantic search as a means of helping researchers find publicly available datasets, likely in combination with human review for the foreseeable future. While our study specifically targeted gene-expression data, the approach we developed can be applied to other types of molecular data that are accompanied by human-language descriptions. We hope that these efforts lead to improvements in data findability, thus increasing the value derived from previous research investments.

Author contributions

Grace S. Brown (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), James Wengler (Conceptualization [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Aaron Joyce S. Fabelico (Formal analysis [equal], Investigation [equal], Methodology [equal]), Abigail Muir (Software [equal], Writing—review & editing [equal]), Anna Tubbs (Software [equal]), Amanda Warren (Software [equal], Writing—review & editing [equal]), Alexandra N. Millett (Data curation [equal], Validation [equal]), Xinrui Xiang Yu (Data curation [equal], Validation [equal]), Sanja Rogic (Conceptualization [equal], Methodology [equal], Project administration [equal], Writing—review & editing [equal]), Stephen R. Piccolo (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Software [equal], Supervision [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), and Paul Pavlidis (Conceptualization [equal], Resources [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflicts of interest

None declared.

Funding

P.P. received funding from the USA National Institutes of Health (R01MH111099). S.R.P. received funding from the USA National Institutes of Health (R03HL168983).

References

Alameer A, Chicco D. geoCancerPrognosticDatasetsRetriever: a bioinformatics tool to easily identify cancer prognostic datasets on gene expression omnibus (GEO). *Bioinformatics* 2022; **38**:1761–3.

Alsentzer E, Murphy JR, Boag W *et al.* Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA, 2019, 72–8.

Bates D, Mächler M, Bolker B *et al.* Fitting linear mixed-effects models using lme4. *J Stat Soft* 2015;**67**:48.

Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the sequence read archive. *Bioinformatics* 2017;**33**:2914–23.

Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2009.

Bojanowski P, Grave E, Joulin A *et al.* Enriching word vectors with subword information. In: *Transactions of the Association for Computational Linguistics*, Cambridge, MA USA: MIT Press, 2017, 135–146.

Cao H. Recent advances in text embedding: A comprehensive review of top-performing methods on the MTEB benchmark. arXiv, https://doi.org/10.48550/arXiv.2406.01607, 2024, preprint: not peer reviewed.

Chen G, Ramírez JC, Deng N *et al.* Restructured GEO: restructuring gene expression omnibus metadata for genome dynamics analysis. *Database* 2019;**2019**:bay145.

Chen Q, Lee K, Yan S *et al.* BioConceptVec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput Biol* 2020;**16**:e1007617.

Chen X, Gururaj AE, Ozyurt B *et al.* DataMed–an open source discovery index for finding biomedical datasets. *J Am Med Inform Assoc* 2018;**25**:300–8.

Chiu B, Crichton G, Korhonen A *et al.* How to train good word embeddings for biomedical NLP. *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin,

Germany: Association for Computational Linguistics, 2016. 166–174.

Chua H-E, Tucker-Kellogg L, Bhowmick SS. ArcheGEO: towards improving relevance of gene expression omnibus search results. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA: Association for Computing Machinery, 2022. 1–10.

Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol* 2016;**1418**:93–110.

Devlin J, Chang M-W, Lee K *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Minneapolis, MN, USA, 2019, 4171–4186.

Djordjevic D, Tang JYS, Chen YX *et al.* Discovery of perturbation gene targets via free text metadata mining in gene expression omnibus. *Comput Biol Chem* 2019;**80**:152–8.

Fardin P, Barla A, Mosci S *et al.* The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines. *BMC Genomics* 2009;**10**:474.

Federico A, Hautanen V, Christian N *et al.* Manually curated and harmonised transcriptomics datasets of psoriasis and atopic dermatitis patients. *Sci Data* 2020;**7**:343.

Ganzfried BF, Riester M, Haibe-Kains B *et al.* curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database (Oxford)* 2013;**2013**:bat013.

Gendoo DM, Zon M, Sandhu V *et al.* MetaGxData: clinically annotated breast, ovarian and pancreatic cancer datasets and their use in generating a multi-cancer gene signature. *Sci Rep* 2019;**9**:8770.

Giles CB, Brown CA, Ripperger M *et al.* ALE: automated label extraction from GEO metadata. *BMC Bioinformatics* 2017; **18**:509–16.

Good BM, Su AI. Crowdsourcing for bioinformatics. *Bioinformatics* 2013;**29**:1925–33.

Hadley D, Pan J, El-Sayed O *et al.* Precision annotation of digital samples in NCBI's gene expression omnibus. *Sci Data* 2017; **4**:170125.

Harnoune A, Rhanoui M, Mikram M *et al.* BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Comput Methods Programs Biomed Update* 2021;**1**:100042.

Hawkins NT, Maldaver M, Yannakopoulos A *et al.* Systematic tissue annotations of genomics samples by modeling unstructured metadata. *Nat Commun* 2022;**13**:6736.

Kamphuis C, De Vries AP, Boytsov L *et al.* Which BM25 do you mean? A large-scale reproducibility study of scoring variants. *European Conference on Information Retrieval*. Lisbon, Portugal: Springer, 2020, 28–34.

Khoroshevskyi O, LeRoy N, Reuter VP *et al.* GEOfetch: a command-line tool for downloading data and standardized metadata from GEO and SRA. *Bioinformatics* 2023;**39**:btad069.

Lachmann A, Torre D, Keenan AB *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* 2018;**9**:1366.

Lake BM, Murphy GL. Word meaning in minds and machines. *Psychol Rev* 2023;**130**:401–31.

Lee J, Yoon W, Kim S *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40.

LeRoy, Nathan J, Khoroshevskyi, Oleksandr, O'Brien, Aaron, *et al.* PEPhub: a database, web interface, and API for editing, sharing, and validating biological sample metadata. *GigaScience* 2024;**13**:giae033.

Lim N, Tesar S, Belmadani M *et al.* Curation of over 10 000 transcriptomic studies to enable data reuse. *Database* 2021; **2021**:baab006.

Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265–6.

Liu Y, Ott M, Goyal N *et al.* Roberta: A robustly optimized bert pretraining approach. arXiv, 2019, preprint: not peer reviewed.

Lowe HJ, Barnett GO. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Jama* 1994;**271**:1103–8.

Lù XH. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. arXiv, https://doi.org/10.48550/arXiv.2407.03618, 2024, preprint: not peer reviewed.

Mellins ED, Macaubas C, Grom AA. Pathogenesis of systemic juvenile idiopathic arthritis: some answers, more questions. *Nat Rev Rheumatol* 2011;**7**:416–26.

Mikolov T, Chen K, Corrado G *et al.* Efficient estimation of word representations in vector space. arXiv, 13013781, 2013, preprint: not peer reviewed.

Mikolov T, Grave E, Bojanowski P *et al.* Advances in pre-training distributed word representations. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, 2018.

Noy NF, Shah NH, Whetzel PL *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.

Ohno-Machado L, Sansone S-A, Alter G *et al.* Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* 2017;**49**:816–9.

Patra BG, Roberts K, Wu H. A content-based dataset recommendation system for researchers—a case study on gene expression omnibus (GEO) repository. *Database (Oxford)* 2020;**2020**: 1. https://doi.org/10.1093/database/baaa064

Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Doha, Qatar. 2014. 1532–43.

Perez-Riverol Y, Zorin A, Dass G *et al.* Quantifying the impact of public omics data. *Nat Commun* 2019;**10**:3512.

Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. *Gigascience* 2016;**5**:30.

Rahman M, Boughorbel S, Presnell S *et al.* A curated transcriptome dataset collection to investigate the functional programming of human hematopoietic cells in early life. *F1000Res* 2016;**5**:414.

Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *J Am Med Inform Assoc* 2024;**31**:1844–55.

Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *FNT in Information Retrieval* 2009; **3**:333–89.

Schartner C, Scholz C-J, Weber H *et al.* The regulation of tetraspanin 8 gene expression—a potential new mechanism in the

pathogenesis of bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 2017;**174**:740–50.
Seal MD, Chia SK. What is the difference between triple-negative and basal breast cancers? *Cancer J* 2010;**16**:12–6.
Shah N, Guo Y, Wendelsdorf KV *et al.* A crowdsourcing approach for reusing and meta-analyzing gene expression data. *Nat Biotechnol* 2016;**34**:803–6.
Vasilevsky N, Essaid S, Matentzoglu N *et al.* Mondo disease ontology: harmonizing disease concepts across the world. Leipzig, Germany: CEUR- WS, 2020, 2807.
Villaseñor-Altamirano AB, Moretto M, Maldonado M *et al.* PulmonDB: a curated lung disease gene expression database. *Sci Rep* 2020;**10**:514.
Wang Y, Liu S, Afzal N *et al.* A comparison of word embeddings for the biomedical natural language processing. *J Biomed Inform* 2018;**87**:12–20.
Wang Z, Monteiro CD, Jagodnik KM *et al.* Extraction and analysis of signatures from the gene expression omnibus by the crowd. *Nat Commun* 2016;**7**:12846.
Wickham H, Averick M, Bryan J *et al.* Welcome to the tidyverse. *JOSS* 2019;**4**:1686.
Wilkinson MD, Dumontier M, Aalbersberg I *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
Wolf T, Debut L, Sanh V *et al.* Transformers: state-of-the-art natural language processing. In: Liu Q, Schlangen D (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020. 38–45.
Zimmermann P, Bleuler S, Laule O *et al.* ExpressionData - a public resource of high quality curated datasets representing gene expression across anatomy, development and experimental conditions. *BioData Min* 2014;**7**:18.

Bioinformatics, 2026, 42, 1–14
https://doi.org/10.1093/bioinformatics/btag053
Original Paper