Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FGTD: a desktop application for GEO dataset discovery with sample-level metadata extraction and local semantic relevance scoring

Francesco Orilio
University of Palermo
Alice Turdo
University of Palermo
Francesca Angeloro
University of Palermo
Giovanni Nigita
The Ohio State University
Francesco Verona
Azienda Ospedaliera Universitaria Policlinico "Paolo Giaccone" di Palermo
Giorgio Stassi
University of Palermo
Matilde Todaro
University of Palermo
Sebastiano Di Bella
sebastiano.dibella@unipa.it
University of Palermo

Article

Keywords: GEO, PubMed integration, dataset discovery, semantic embeddings, relevance scoring, data reuse

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11430565/v1

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

Additional Declarations: No competing interests reported.

FGTD: a desktop application for GEO dataset discovery with sample-level metadata extraction and local semantic relevance scoring

Authors

Francesco Orilio[1], Alice Turdo[2], Francesca Angeloro[2], Giovanni Nigita[3,4,5], Francesco Verona[6], Giorgio Stassi[1,7], Matilde Todaro[2,6], Sebastiano Di Bella[1,*]

Affiliations

[1] Department of Precision Medicine in Medical, Surgical and Critical Care (Me.Pre.C.C), University of Palermo, Palermo, Italy.

[2] Department of Health Promotion, Mother and Child Care, Internal Medicine and Medical Specialities (PROMISE), University of Palermo, Palermo, Italy.

[3] Division of Medical Oncology, Department of Internal Medicine, College of Medicine, The Arthur G. James Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.

[4] Pelotonia Institute for Immuno-Oncology, The Arthur G. James Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.

[5] Center for RNA Biology, The Ohio State University, Columbus, Ohio, USA.

[6] Azienda Ospedaliera Universitaria Policlinio “Paolo Giaccone” (AUOP), Palermo, Italy

[7] IRCCS SYNLAB SDN, Naples, 80143, Italy.

*To whom correspondence should be addressed. Email: sebastiano.dibella@unipa.it

Abstract

The Gene Expression Omnibus (GEO) contains a large amount of reusable functional genomics data, but finding the series relevant to a biological question still requires the manual screening of heterogeneous metadata and linked publications. We present FGTD (From GEO To Dataset), a desktop application that automates the retrieval, filtering and ranking of GEO series and associated publications. FGTD organizes the extracted metadata, down to the sample level, in one table, reconstructing links to the Sequence Read Archive and BioProject, and ranks the series with a score that combines semantic similarity (computed locally with sentence embeddings), keyword and Medical Subject Headings matches, and citation and recency metrics. In a breast cancer case study FGTD produced the metadata table of 15 series in about 38 seconds, while manual curation (timed on two series and extrapolated to the cohort) was estimated at about 3.8 hours. In known-item evaluations on two published meta-analyses, FGTD ranked the reference series closer to the top than

the date-ordered GEO listing (mean reciprocal rank 0.571 versus 0.020), although the gain depended on the keywords supplied, and its semantic component alone and its fusion with a lexical baseline performed at least as well. FGTD is freely available at https://github.com/SorenKK/FGTD.

Keywords: GEO; PubMed integration; dataset discovery; semantic embeddings; relevance scoring; data reuse

Introduction

Biomedical research increasingly integrates large-scale experimental data with the scientific literature to generate and validate hypotheses, and in this context the Gene Expression Omnibus (GEO) database [1] and PubMed [2] are central resources. Specialized platforms such as cBioPortal [3] offer curated cancer genomics data, mainly next-generation sequencing (NGS) data from selected studies, representing a filtered subset of the available knowledge, while GEO is a large and heterogeneous public repository that hosts raw and processed data of different experimental modalities (e.g., microarrays, RNA-seq and ChIP-seq) without selective curation. However, the workflow to identify and organize these resources remains fragmented, often hindered by unstructured and inconsistent metadata [4]: researchers usually rely on manual queries, inspecting the metadata case by case, a process that becomes impractical when screening large cohorts or keeping up with the rapid growth of multi-omics publications [5,6]. Existing tools address this problem only in part: programmatic interfaces such as the Entrez Programming Utilities (E-utilities) [2] and specialized databases such as GEOmetadb [7] offer efficient access to GEO metadata, but are mainly intended for users with computational skills.

A second group of resources works directly on GEO, and it is the group to which FGTD is most directly comparable. Public Omics Explorer (POE) [8] performs semantic search over GEO datasets by embedding the abstracts of the PubMed articles linked to each series, representing the closest precedent to the semantic component of FGTD; it is, however, a hosted web service, restricted to series linked to a PubMed identifier (PMID), which returns ranked datasets rather than the structured, sample-level metadata table needed to plan a reanalysis. TidyGEO [9] addresses the complementary step, i.e. reshaping the data of a single series, already identified, into an analysis-ready matrix, without any discovery or ranking layer. recount3 [10] provides uniformly reprocessed expression summaries for a large but fixed subset of RNA-seq studies (human and mouse only), so that the studies outside this subset, and every modality other than RNA-seq, remain outside its scope. Domain-specific portals such as POSTAR3 [11] curate in depth a single assay family (post-transcriptional regulation by RNA-binding proteins) at the cost of generality, while repository-level infrastructures such as

ArrayExpress/BioStudies [12] expose rich submitted metadata through an application programming interface (API), leaving ranking, enrichment and cross-repository linkage to the user. FGTD therefore does not claim to be the first semantic layer over GEO, but occupies a different operating point: it combines discovery, ranking, sample-level metadata extraction and one-click routing to the raw data in a single desktop tool that requires no programming and computes embeddings and scores on the user's machine, so that the query is sent only to the services of the National Center for Biotechnology Information (NCBI) that host the data and not to any external artificial intelligence (AI) service.

A third group of tools works on the literature rather than on the data. Platforms such as Connected Papers [13] and ResearchRabbit [14] use citation networks to map thematic clusters, while Elicit [15] uses large language models to extract the key results of papers into structured tables. These platforms have made literature screening considerably easier, although they remain confined to the bibliographic record, not being linked to experimental repositories and do not provide the structured metadata extraction or the sample-level filtering needed for downstream analysis, so that the scientific narrative of the papers remains disconnected from the technical parameters required to reuse the underlying data. To bridge this gap we developed FGTD (From GEO To Dataset), an open-source desktop application distributed with an archived release, in line with the FAIR principles (Findable, Accessible, Interoperable, Reusable) for research software [16]. Given a biological query, FGTD returns a ranked and interactive table of GEO series enriched with the linked publication, the sample-level metadata and direct links to the raw data in the Sequence Read Archive (SRA) [17] and BioProject [18], without requiring any programming. Its semantic ranking and similarity map help users order and explore the series returned by a GEO query, including those described with a terminology different from that of the query.

An overview of the workflow is shown in Fig. 1, while a step-by-step tutorial is available in the GitHub repository (https://github.com/SorenKK/FGTD), describing how to define the query parameters and giving a detailed breakdown of the architecture of the application.

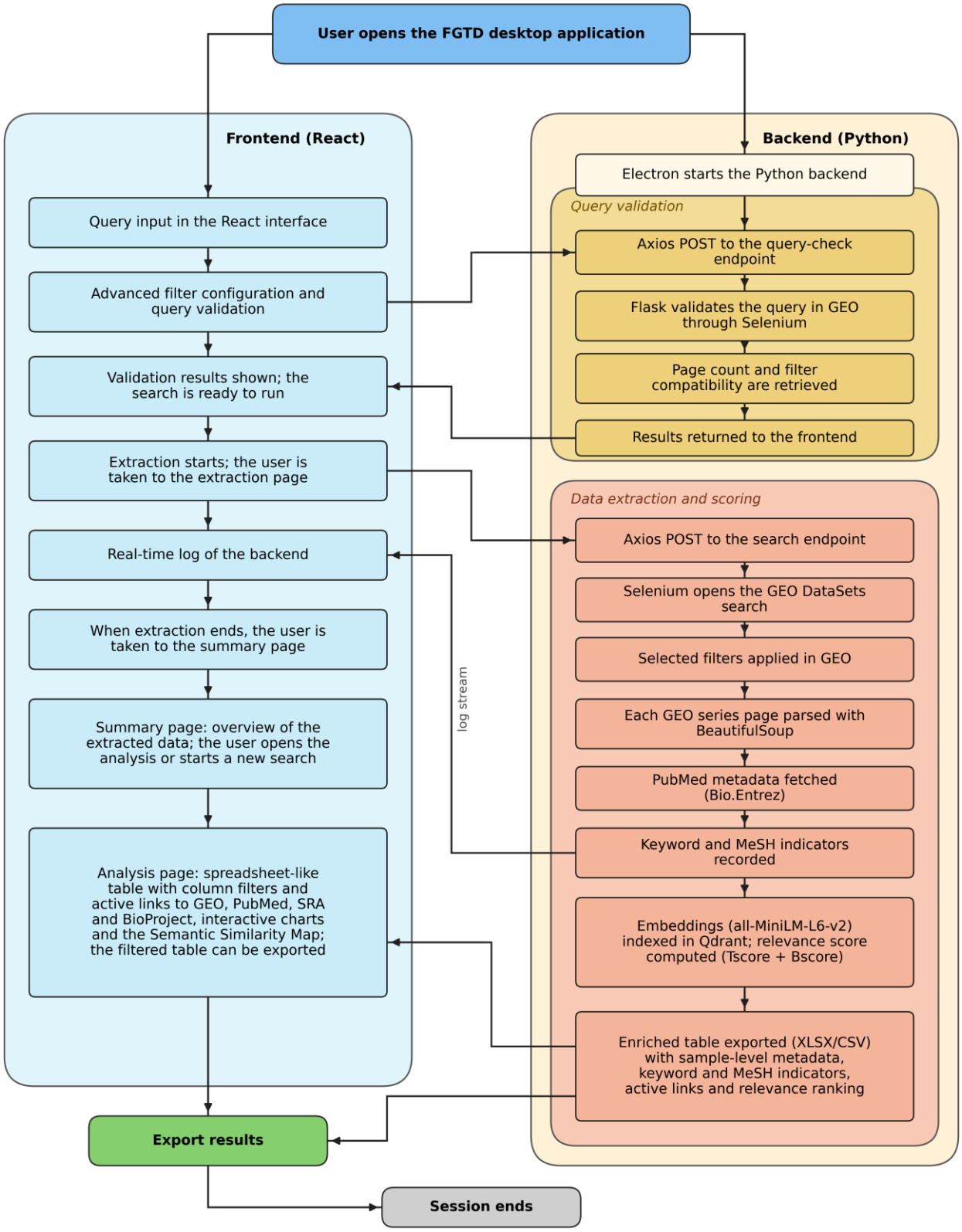


Figure 1. *Architecture and workflow of the FGTD desktop application. The React frontend (light blue, left) and the Python backend (light yellow, right) exchange requests through a local REST interface. Within the backend, yellow blocks mark the query validation steps and salmon blocks the*

data extraction and scoring steps. Arrows give the order of the steps and the results returned to the frontend; the workflow ends with the export of the results. Complete workflow details are available in the GitHub repository.

To place FGTD within the broader landscape of research discovery tools, we mapped its capabilities against 16 established platforms (Supplementary Table S1). Consistently with the target user of FGTD, the comparison was restricted to platforms that can be used through a graphical or web interface, excluding programmatic resources without such an interface (e.g., the E-utilities and GEOmetadb). The platforms fall into four groups: the first comprises AI-driven literature synthesis and visual discovery tools (Elicit [15], Consensus [19], Paperguide [20], SciSpace [21], ResearchRabbit [14] and Connected Papers [13]); the second, scholarly mapping engines that index publications across the whole literature (The Lens [22], Dimensions [23], OpenAlex [24] and Semantic Scholar [25]); the third, generalist data repositories, represented by Zenodo [26]; the fourth, the resources that work natively on GEO, i.e. POE [8], TidyGEO [9], recount3 [10], POSTAR3 [11] and ArrayExpress/BioStudies [12]. These last resources differ from FGTD in the step of the workflow they automate: POE ranks datasets without extracting sample-level metadata, TidyGEO reshapes metadata without providing discovery, recount3 offers pre-computed expression matrices for a fixed subset of studies, POSTAR3 is restricted to a single assay family and ArrayExpress/BioStudies leaves ranking and cross-repository linkage to the user.

Given that these tools serve different primary objectives, the comparison is not intended as a performance benchmark: the matrix describes each platform along nine dimensions (target domain, cross-database linkage, metadata granularity, search architecture, relevance ranking logic, interface, data sovereignty, export output and the step of the discovery workflow that the platform automates). The resulting picture is a fragmented ecosystem, in which AI-driven literature platforms are effective for evidence synthesis and conceptual mapping, but they remain largely confined to the bibliographic record, extracting claims at the level of the paper without access to the underlying data. Scholarly mapping engines and generalist repositories provide extensive macro-level linkages (e.g., funding, patent status or co-citation networks), but they lack the parsers needed to extract granular experimental units from GEO, such as sample (GSM) and platform (GPL) records or series matrix files. GEO-native resources, instead, work on the right repository, but each of them covers only one step of the workflow.

Within this landscape FGTD occupies a complementary niche, which we define as multi-omics meta-analysis and dataset harvesting. Among the platforms compared it is the only one covering discovery,

ranking, sample-level metadata extraction and raw-data routing within a single tool, linking GEO, PubMed and SRA so that the bibliographic narrative is connected with actionable experimental metadata. A second distinguishing feature regards data sovereignty. Most current discovery tools rely on cloud-based infrastructures, which may raise confidentiality concerns for researchers exploring sensitive or patentable hypotheses; FGTD instead computes embeddings, vector searches, scores and metadata parsing locally, so that the content of the query is not processed by third-party AI services and only the search itself is submitted to NCBI (which hosts GEO and PubMed).

Results

Efficiency benchmark: automated versus manual curation. To estimate the manual effort removed by FGTD, we ran a case study with the query "breast cancer" combined with native GEO filters: *Homo sapiens* (organism), expression profiling by high throughput sequencing (study type), tissue, gender and strain (attribute names), CEL and BED (supplementary files), and publication dates from 1 January 2010 to 31 December 2020. The keywords "breast", "DNA", "immunotherapy", "protein", "RNA" and "tumor" and the Medical Subject Headings (MeSH) terms "Humans" and "Female" were supplied to annotate and score the results. The run, performed in Ultra Mode, returned 15 series (12 linked to a PubMed record and 3 without a linked publication, i.e. orphan series). The advanced query is reported in Methods, the configuration of the run in Supplementary Data 1, the score breakdown in Supplementary Table S2 and the complete metadata output in Supplementary Table S3.

The FGTD output was reproduced by hand by three curators with different levels of experience: two of the authors (a senior researcher with an extensive background in breast cancer oncology (A.T.) and a PhD candidate in the same field, F.A.) and an undergraduate intern without a specific domain focus, who had also taken part in the beta testing of the application (L.S.). Each curator independently compiled, for one PMID-linked series (GSE124409) and one orphan series (GSE69986), the same fields produced by FGTD, following written instructions and a blank annotation form distributed before the exercise, with no example output (Supplementary Note 1 and Supplementary Table S4); annotations and times are reported in Supplementary Tables S5–S7 and summarized in Supplementary Table S13. The curators needed a mean of 17.0 min for the PMID-linked series (range 13.1–20.0 min; n = 3) and of 8.0 min for the orphan series (range 6.3–11.0 min; n = 3) (Fig. 2), which, extrapolated to the composition of the cohort, correspond to about 227.6 min (3.8 h) of manual work. FGTD processed the 15 series in 37.8 s (mean of three runs, range 37.5–38.0 s; about 2.6 s per PMID-linked and 1.5 s per orphan series, the remaining 2.4 s being spent on the GEO search, scoring and

export), without an NCBI API key, a reduction of more than 99% compared to the manual estimate. Given that the target format was defined by FGTD itself and only two series were curated by hand, these values estimate the order of magnitude of the manual effort needed to obtain an equivalent output, and they do not measure retrieval quality, which is evaluated below.

Accuracy of the extracted metadata. The FGTD output was compared with the three manual tables field by field, considering 30 objective fields (the relevance score was excluded) for two series and three curators, i.e. 180 comparisons, with each discordance resolved against the GEO and PubMed records (Methods; Fig. 2b; Supplementary Table S13). Of these, 146 agreed, 12 differed only in form (three of them being the GEO summary exported in place of the PubMed abstract; Methods) and 22 differed in content. Ten of the latter were errors in the manual tables, such as missing links or keywords marked by meaning rather than by exact match, while the other 12 reflect four FGTD errors, each repeated against the three curators. The three FTP links of the orphan series pointed to a non-existent folder, the folder name being built from a fixed slice of the accession, which is correct only for six-digit accessions; this error affected every series with a lower accession (including the pools of both known-item evaluations), although their files remain reachable from the GEO accession page, which FGTD links correctly. The fourth error regarded the "RNA" indicator of the same series, set by matching the keywords as substrings and "rna" occurs inside "Journal" in its GEO summary. Overall FGTD returned 56 of its 60 values correctly, and each curator 56 or 57: the errors of FGTD turned out to be systematic, and the FTP paths have been corrected in the version archived with this article (Code availability), while those of the curators varied between people. With only two series, these figures show the kinds of error rather than an error rate.

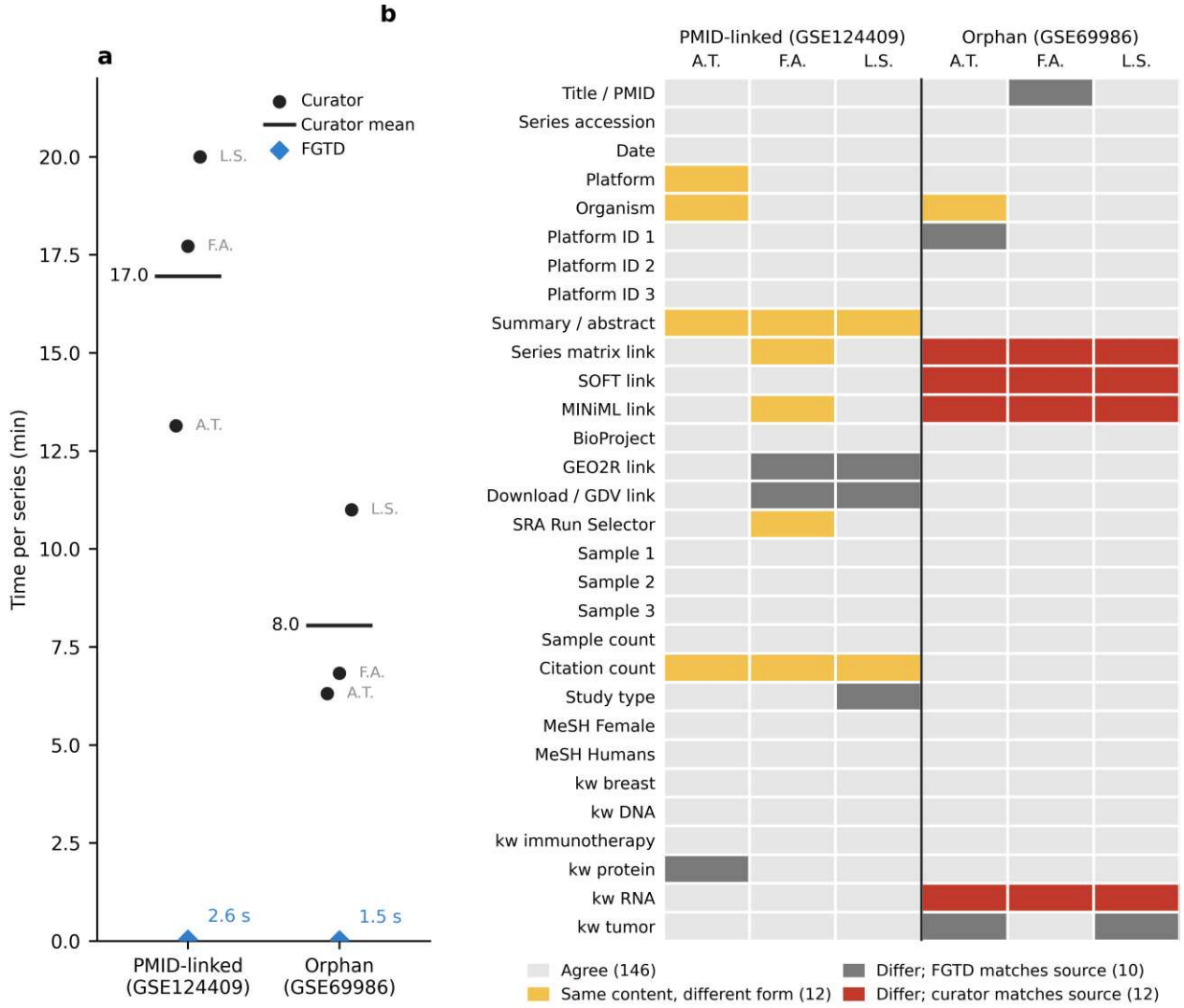


Figure 2. *Time and accuracy of manual curation and FGTD for two GEO series. (a) Time required to produce the structured metadata row for the PMID-linked series (GSE124409) and the orphan series (GSE69986). Black dots show the three curators and the horizontal line their mean (value on the left); blue diamonds show the average time per series of the corresponding type in FGTD, measured over three runs on the 15-series cohort without an NCBI API key (37.8 s in total on average). (b) Field-by-field concordance between the FGTD output and each curator for the 30 objective fields of the form (180 comparisons). Light grey, identical values; yellow, the same content in a different form (e.g., a taxonomy link instead of the organism name, or a citation count read on a different day); dark grey, values that differ, with FGTD matching the GEO or PubMed record; red, values that differ, with the curator matching the record. The red cells correspond to four FGTD errors, each counted once per curator: the three FTP links of GSE69986 and the "RNA" keyword indicator (see text). Complete data in Supplementary Table S13.*

Known-item retrieval against published meta-analyses. To assess whether the ranking brings relevant series to the top of a realistic candidate pool, we used as known items the five GEO series retained by Cao et al. [27] for a meta-analysis of prognostic genes in triple-negative breast cancer (GSE27447, GSE36295, GSE43358, GSE61724 and GSE75678). A single FGTD run, built on the inclusion criteria reported by the authors (Methods), returned 1,264 series including all five known items, 937 of which (74%) were linked to a PubMed record. We then compared the native GEO order of the pool (i.e. the order of the GEO DataSets search results, which lists the most recently published series first) with the order induced by the FGTD relevance score. FGTD ranked first GSE61724, which was at position 35 in the native GEO order, so that the reciprocal rank of the first known item rose from 0.029 (GSE75678, position 34) to 1.000, Precision@5 from 0.000 to 0.200, Precision@10 from 0.000 to 0.100 and the normalized discounted cumulative gain (NDCG) at cut-offs 5 and 10 from 0.000 to 0.339. GSE27447 also improved its position (1,176 → 455), while three known items lost ground (GSE75678, 34 → 498; GSE43358, 335 → 468; GSE36295, 481 → 1,062), worsening the median rank of the five from 335 to 468 (Fig. 3a). Ranks under both orderings are reported in Supplementary Table S8, the score breakdown in Supplementary Table S9 and the metadata of the pool in Supplementary Table S10. The first position of GSE61724 did not depend on the composition of the pool, since it was retained in all 200 random subsamples of 90% of the series, but it depended on the keywords: with three alternative keyword lists written from the aim of the meta-analysis, the first known item ranked between 9th and 83rd (Supplementary Table S16).

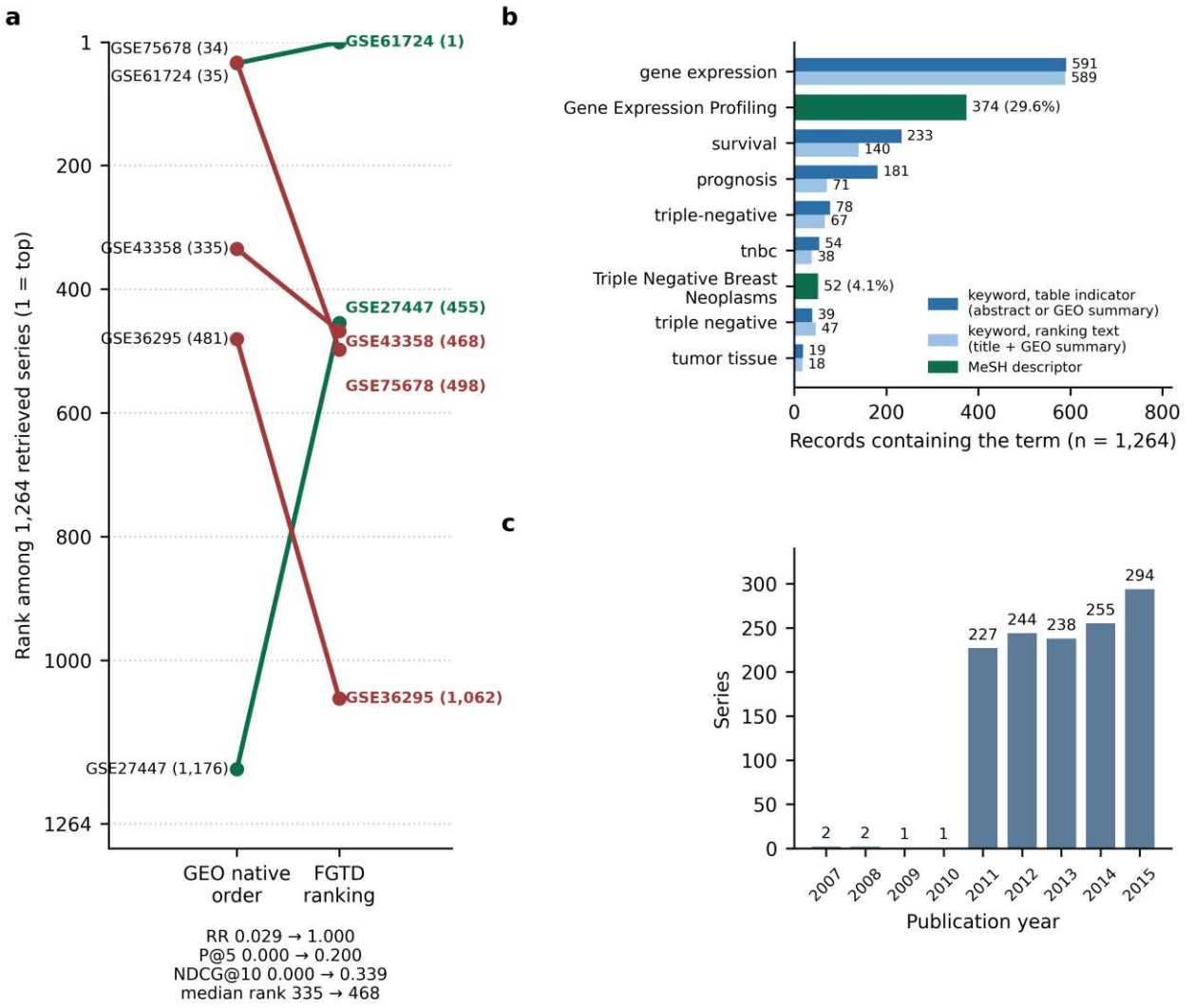


Figure 3. *Known-item retrieval and composition of the retrieved pool. (a) Position of the five known items under the native GEO order and under the FGTD relevance ranking. Each line connects the rank of one GEO series in the native GEO order (left; recovered through ESearch, Methods) to its rank under the FGTD relevance score (right), within the same pool of 1,264 series; rank 1 is at the top. Green lines mark the two series whose position improved and red lines the three that lost positions. The five series are those retained by Cao et al. [27] for a meta-analysis of prognostic genes in triple-negative breast cancer. The reciprocal rank (RR) of the first known item, Precision@5 (P@5), NDCG@10 and the median rank under the two orderings are reported beneath panel a. Ranks were computed on unrounded scores, with ties broken by the native GEO position. (b) Prevalence of each query term among the 1,264 retrieved series, expressed as the number of records carrying the term; for free-text keywords, dark blue bars count the indicator columns of the output table (PubMed abstract or, for series without a linked publication, GEO summary) and light blue bars the title and GEO summary used by the keyword score; green bars are MeSH descriptors*

matched against the linked PubMed record. (c) Distribution of the retrieved series by publication year. The declared retrieval window is 2011–2015; the six records falling outside it are series whose GEO publication date differs from the date indexed by the query filter.

The series that lost positions highlight how the score depends on the metadata deposited with each series: GSE36295, in particular, does not contain any of the query keywords in its title or GEO summary, receiving therefore a keyword score of zero. The composition of the pool shows that such a sparse keyword coverage is not unusual: in the title and GEO summary used by the keyword score, the most frequent query term (gene expression) occurs in 46.6% of the records and triple-negative in 5.3%, and the MeSH descriptor Triple Negative Breast Neoplasms in 4.1% (Fig. 3b,c).

Using the top-ranked known item as seed of the Semantic Similarity Map, the other SubSeries of the same SuperSeries (GSE61723, 97.1%) and the SuperSeries itself (GSE61725, 82.0%) turned out to be among the nearest neighbors; excluding them, nine of the ten nearest neighbors were triple-negative breast cancer studies (Supplementary Table S14), while the remaining known items fell between the 206th and the 703rd position among the 910 series shown by the map (Fig. 4). The map therefore works as a discovery aid around a chosen dataset, and not as a retrieval ranking. The similarity scores underlying Fig. 4 are provided in Supplementary Data 2.

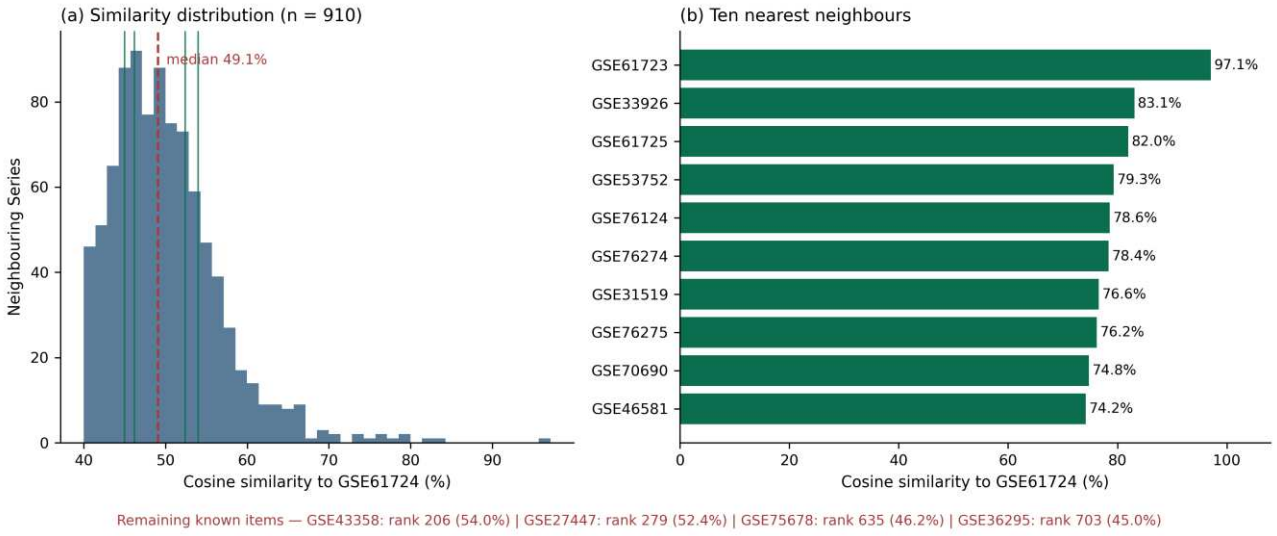


Figure 4. *Semantic similarity neighborhood of the top-ranked known item (GSE61724). (a) Distribution of the cosine similarity scores between GSE61724 and the 910 series of the pool whose similarity to it is at least 0.40, the threshold used by the map, with the median marked and the positions of the remaining known items indicated. (b) The ten nearest neighbors of GSE61724 by semantic similarity. GSE61723, ranked first, is the other SubSeries of the SuperSeries GSE61725, ranked third, that contains the seed; it is the only series above 0.85, so that the highest similarity*

class of the map is populated here by a sibling series. The ten neighbors, and the ten nearest neighbors outside the SuperSeries, are listed in Supplementary Table S14 with the evidence that they are triple-negative breast cancer studies. The map shows absolute cosine similarities, whereas the semantic score of the ranking is scaled by the batch maximum. The compressed range of the scores means that a similarity value orders the neighbors of a seed dataset and must not be read as a probability of relevance.

Second meta-analysis. The evaluation was repeated with the three series (GSE3744, GSE8977 and GSE10797) that Li et al. [28] integrated in a meta-analysis on the progression and metastasis of breast cancer, in which they compared breast tumor with normal breast tissue, using a query of the same form restricted to 2006–2008, which returned 273 series including all three (Methods; Supplementary Table S15). FGTD placed them at positions 7, 16 and 172, against 89, 144 and 262 in the native GEO order (reciprocal rank 0.011 → 0.143; median rank 144 → 16; Fig. 5b). Differently from the first evaluation, the full score did not place a known item first: the semantic component alone ranked GSE3744 first, and BM25, a classical lexical ranking function (Methods), placed GSE10797 second. With the alternative keyword lists, the full score placed the first known item 2nd, 2nd and 25th, and in all 200 subsamples of the pool the semantic component alone ranked a known item first whereas the full score did not (Supplementary Table S16).

Comparison with retrieval baselines and component ablations. The full score was compared with five single-signal baselines, four variants of the composite score and three hybrid orderings (Fig. 5 and Supplementary Table S11). In the first query the full score placed a known item first (reciprocal rank 1.000, against 0.250 for the semantic component and 0.083 for BM25), BM25 instead had a lower median rank (121, against 468 for the full score and 335 for the native GEO order). In the second query the reciprocal rank was 1.000 for the semantic component, 0.500 for BM25 and 0.143 for the full score, with median ranks of 31, 22 and 16, respectively. Over the two queries the mean reciprocal rank was 0.625 for the semantic component, 0.571 for the full score, 0.292 for BM25 and 0.020 for the native GEO order, and removing the semantic term from the full score lowered it to 0.127.

Since in the first query BM25 and the FGTD score succeeded at different depths of the ranking, a pattern that the second query did not reproduce, we also evaluated three hybrid orderings, defined on the first query before their evaluation (Methods). On that query their reciprocal ranks ranged from 0.111 to 0.500 and their median ranks from 89 to 456. On the second query, run after the hybrids had been fixed, the reciprocal rank fusion of the FGTD and BM25 orderings placed the three known items

at positions 1, 5 and 95. Over the two queries the mean NDCG@10 was 0.391 for rank fusion, 0.308 for the semantic component and 0.248 for the full score; although two queries are not enough to rank these orderings, the combination of semantic and lexical signals is worth evaluating on a larger number of queries.

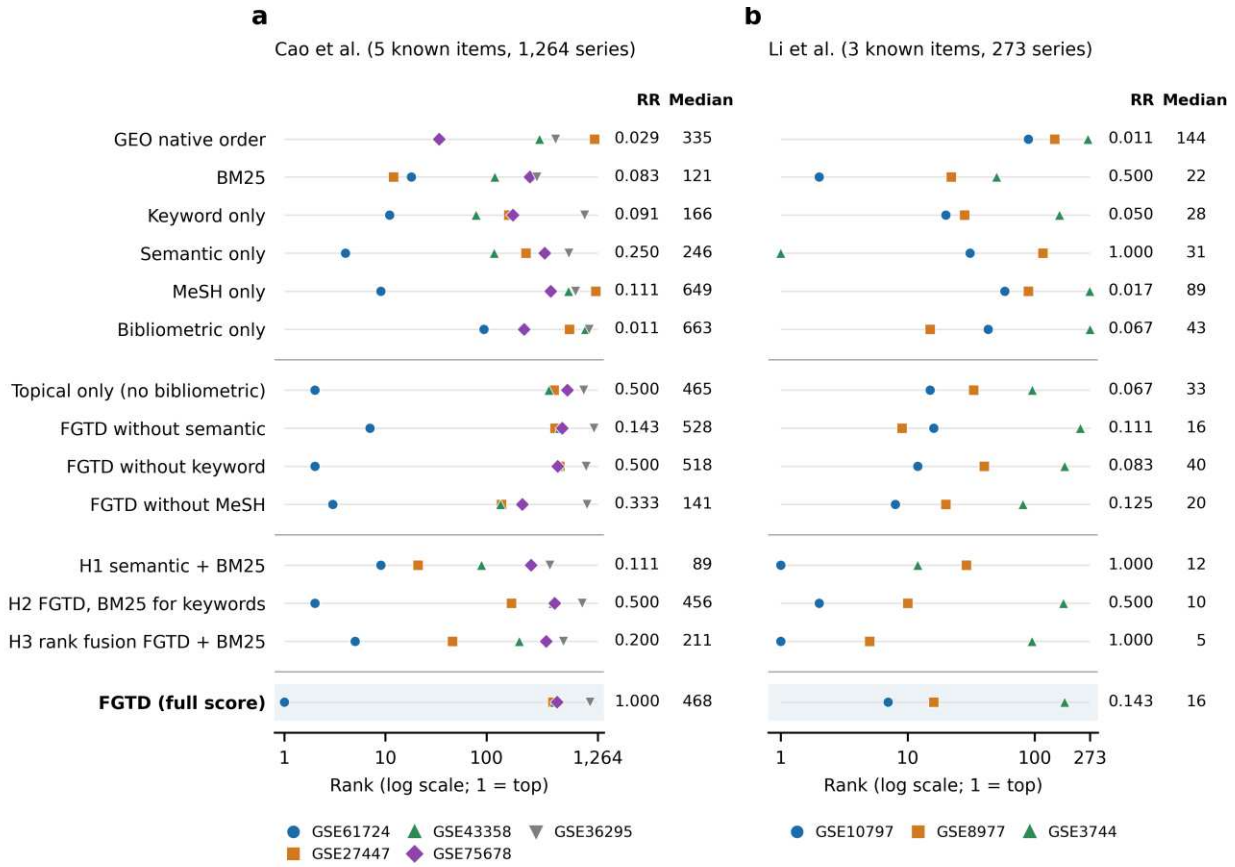


Figure 5. *Known-item ranks under retrieval baselines, ablations and hybrid orderings in the two evaluations. (a) Meta-analysis of Cao et al. (five known items, 1,264 series). (b) Meta-analysis of Li et al. (three known items, 273 series). Each row is one ordering of the corresponding pool; markers give the rank of each known item on a logarithmic axis (1 = top), and the columns on the right of each panel give the reciprocal rank (RR) of the first known item and the median rank. Baselines rank by a single signal; ablations remove one component of the FGTD score and renormalize the remaining weights; hybrids combine BM25 with the semantic score (H1), replace the keyword term of the FGTD score with BM25 (H2) or fuse the FGTD and BM25 rankings (H3; Methods). Ties are broken by the native GEO position. Complete values, and the means over the two queries, in Supplementary Table S11.*

A qualitative comparison of how directly FGTD and three other categories of discovery tools lead from a biological question to raw data is reported in Supplementary Note 2. FGTD retrieves records

through the GEO search, so its recall is bounded by that search, and the series not returned by the GEO query cannot be ranked or shown. Within the retrieved pool, the Semantic Similarity Map allows taking pertinent series as seeds and inspecting the series most similar to them, which is useful when a narrow query returns few direct matches.

Discussion

FGTD connects two steps usually carried out with different tools, i.e. the search for the literature relevant to a biological question and the retrieval of the experimental data behind it. Extracting GSM and GPL records and series matrix files, and reconstructing the links to SRA, BioProject and GEO2R as active URLs in the exported tables, it reduces the manual navigation that separates a publication from its raw data. The difficulty biologists and clinicians encounter in accessing, downloading and reanalyzing large volumes of open genomic data has already been highlighted [29], and the efficiency benchmark gives an order of magnitude of the effort involved (about 17 min per PMID-linked series and 8 min per orphan series for manual curation, against a few seconds for FGTD). Measured on two series and extrapolated to 15, these values indicate the scale of the time saved rather than its exact value.

The known-item evaluation highlights how the hybrid score behaves on a realistic candidate pool. The score placed one of the five series used by Cao et al. [27] first among 1,264 candidates, and in both evaluations it brought the reference series closer to the top compared to the native GEO order, which lists the most recent series first and is therefore the order a user inspects but not a relevance ranking; being a date order, it also shares part of its information with the recency term of the FGTD score. The gain depended on the keywords: with alternative lists the first known item of the first evaluation ranked between 9th and 83rd, still above its native position in two of three cases, so that the keyword list, and its overlap with the vocabulary of the relevant series, is a major determinant of the ranking. This gain was concentrated at the top of the ranking, where the score brings forward series that combine a strong semantic alignment with an established bibliographic profile (as GSE61724 does). However, the gain was not general: the composite score also favors recent and frequently cited series with MeSH annotation, and in our evaluations it pushed down pertinent series that are older, little cited or lack MeSH indexing, such as GSE3744 (ranked first by the semantic component alone and 172nd by the full score) and the orphan series of the first evaluation. This behavior is driven by the MeSH and bibliometric terms, and the sensitivity analysis identified the textual/bibliometric split as the constant to which the ordering is most sensitive, so the default split should be read as an editorial choice. FGTD therefore does not replace in every case the individual

signals, and in particular the semantic component alone. Its main value lies in ordering the fragmented output of a GEO search, and its score turns out to be more useful as one component of a hybrid strategy: in our data, the fusion of the FGTD and BM25 orderings kept a known item near the top (positions 5 and 1) while lowering the median rank (211 against 468, and 5 against 16), although two queries are not enough to establish this. For the same reason the default ordering was not changed: the component scores, the semantic one included, are stored for every series, and offering the semantic and fused orderings as alternatives is a natural extension of the interface.

In the first query the ablations also expose a cost of the MeSH term: removing it lowers the reciprocal rank from 1.000 to 0.333 but improves the median known-item rank from 468 to 141, while removing the keyword or the semantic term worsens it (518 and 528). The term is binary, and in the known-item pool 406 of the 1,264 series carried one of the two descriptors, gaining the same 0.15 in the final score (0.75×0.20), but only one known item did, two of them being orphan series (which score zero by construction) and the publications of the other two are not indexed with either descriptor. A binary term therefore lifts a large block of the pool at once, pushing below it every relevant series without the annotation: it helps when the descriptors single out the relevant series (as in the case of GSE61724), and it harms when they mark a broad and only partly relevant subset, as in this case. A graded score, or a weight decreasing with the fraction of the pool carrying the descriptor, would limit this effect, although one query is not enough to tune it. In the second query this effect was not observed (removing the MeSH term changed the median rank from 16 to 20).

Manual relevance judgments turned out to be variable as well: for GSE124409 the three curators assigned relevance scores of 0.5 (A.T.), 0 (F.A.) and 1.0 (L.S.) on a scale from 0 to 1 (Supplementary Tables S5–S7). FGTD, instead, assigns a score that is reproducible for a given pool and configuration, which for this series was 0.504, almost half of it coming from the bibliometric term (the series being the most recent of the cohort), while its keyword score was zero. The automated score should therefore be read as a transparent and repeatable ordering criterion, not as a better judgment of relevance than that of a domain expert. Curation times also differed, the senior researcher being the fastest on both series, but three curators are too few to relate time to experience.

The Semantic Similarity Map is computed in the frontend from the embeddings of the series already retrieved. In the known-item experiment, using GSE61724 as seed recovered the other members of its SuperSeries and a neighborhood of triple-negative breast cancer studies, while the other known items fell in the middle of the similarity distribution. The map is therefore best used to explore the neighborhood of a series of interest within a retrieved pool (e.g., to find related studies described with

a different terminology), not as a substitute for the query or the ranking. Ultra Mode extends retrieval and scoring to orphan series, using the GEO summary in place of the abstract of the publication; these series having no MeSH annotation or citation count, the MeSH and citation terms of their score are zero by construction.

Embedding, vector search, scoring and metadata parsing run on the user's machine, without sending any query to an external AI service (only the search itself is submitted to NCBI). This design is shared by other desktop applications that keep sensitive analyses local, such as GenMasterTable [30] and EasySMPC [31], and it may be relevant for researchers exploring unpublished or patentable hypotheses, the use of a compact sentence encoder keeping this computation feasible on standard desktop hardware.

This study presents several limitations. The efficiency figures derive from two manually curated series and from a single query, and the agreement between the manual and the automated tables was measured on the same two series, showing the kinds of error that each approach makes without estimating an error rate, and the discordances were resolved by the authors, although against public records and with criteria, scripts and source records provided in Supplementary Data 3; moreover, none of the three curators is independent of the project, two being authors and one a beta tester. The ranking was evaluated on two queries with a known-item protocol rather than with full relevance judgments, and recall was not estimated, the number of records screened by the original authors not being reported and the repository having changed after their analysis. The keywords, MeSH terms, query form and date window of both evaluations were chosen by the authors, who knew the reference series; the results depended on the keywords (Supplementary Table S16), and a window fitted to the reference series makes the absolute metrics optimistic for every ordering. As discussed above, the composite score is not uniformly superior to its semantic component or to a lexical baseline. The comparison with other discovery tools is qualitative and rests on a single query, and the ranking was not compared with POE, the tool closest to FGTD, which runs as a hosted service on PMID-linked series only. The semantic layer relies on all-MiniLM-L6-v2, a general-purpose sentence encoder not trained on biomedical text, and domain-adapted encoders may improve the semantic component at the cost of model size. Retrieval depends on the HTML structure served by the GEO DataSets interface, whose changes may require an update of the parsing layer; for this reason the browser used for retrieval is pinned to a fixed version and shipped with the application. The scoring constants are heuristic design choices (Methods). Finally, pre-compiled builds are provided only for Windows and

Linux, macOS is not currently supported, and building from source on other platforms has not been validated.

FGTD links literature screening with the retrieval of GEO data in a single desktop application, combining the context of PubMed publications with sample-level GEO metadata and ranking the series with a transparent hybrid score computed locally; in our case study it reduced by more than 99% the estimated time needed to assemble a structured metadata table. With the growth of public multi-omics repositories, tools of this kind can help researchers select datasets for reuse more quickly and reproducibly.

Methods

Software architecture. FGTD is an Electron-based [32] desktop application distributed for Windows and Linux, built on a hybrid architecture designed for performance and data privacy. The user interface is built with React [33], which provides the components for single-page web applications, and Tailwind [34], while the core logic is managed by a local Python Flask [35] server that orchestrates the automated retrieval of GEO records, fetches the bibliographic enrichment from PubMed and computes the semantic relevance scores. At launch, the Electron main process starts and manages the Python backend, bundled with PyInstaller [36] as a self-contained executable. The graphical interface was chosen to make data discovery accessible to experimental biologists without programming expertise; for computational users who prefer command-line execution, the core Python engine is also available as a standalone script in the GitHub repository, although standalone execution bypasses the interactive query validation and the dynamic filtering provided by the React frontend.

Automated retrieval and technical filtering. FGTD combines Selenium WebDriver [37] and BeautifulSoup [38] to extract nested metadata down to the sample level. The facets selected by the user (e.g., supplementary file types, attribute names and study types) are translated into a GEO-native advanced search; Selenium WebDriver then opens an automated browser session, executes the search and waits for the complete rendering of the dynamic elements of the GEO portal, after which BeautifulSoup parses the rendered HTML. Technical hyperlinks, such as those to the SRA Run Selector, BioProject and GEO2R [39], are reconstructed from the page and preserved as active URLs. Retrieval can run in two modes. In Normal Mode FGTD retains only the series linked to a PubMed publication, enriching them through the NCBI E-utilities with abstracts, citation counts and MeSH terms, which mitigates the inconsistency and incompleteness of the metadata deposited in GEO; Ultra Mode also retains orphan series (i.e. without a linked PubMed record), using their GEO summary in place of the abstract, whereas in Normal Mode these series are discarded. To support the filtering of

the output table, FGTD records for each user-defined keyword a binary indicator of its presence in the PubMed abstract (or, for orphan series, in the GEO summary) and for each user-supplied MeSH term a binary indicator of its presence among the MeSH terms retrieved through Bio.Entrez; orphan series, having no MeSH annotation, receive 0 for every MeSH term. The keyword score used for ranking is instead computed separately, from the title and the GEO summary of each series (equation (3)). The keyword counts are divided by the largest count in the batch, so that the series matching the most keywords (which we call the Top-Ranking Record) receives 1.0, and the ratio is raised to the exponent 1.35 ($Kscore = ratio^{1.35}$) rather than being used linearly. This exponent was chosen heuristically by inspecting boundary cases: values close to 1.0 left technically incomplete records competitive with complete matches, whereas values of 2.0 or more suppressed semantically related records with only partial matches (the effect of the exponent on the textual score is illustrated with synthetic records in Supplementary Fig. S1). Finally, the Flask backend packages the parsed metadata and the scores into JSON objects for the Electron frontend.

Choice of the web interface over the E-utilities. FGTD retrieves GEO series through the GEO DataSets web interface rather than through the E-utilities, for three reasons. First, the search is the user's own GEO search: the facets offered for a query (organism, study type, attribute names, supplementary file types and publication dates) are read from the GEO DataSets interface and applied there, so that the retrieved set is the one returned by the same search on the GEO website and can be checked against it with the "Go to GEO" control, whereas the E-utilities return the identifiers matching a query but not the facet list that the interface offers for it. Second, each series is read from its GEO accession page, i.e. the page that a curator would open, so that every value of the output table can be traced to a page that the user can inspect. Third, some of the links reconstructed by FGTD, such as the Genome Data Viewer track, are shown on that page but are not returned by ESummary. Most of the other series-level fields, including the availability of GEO2R and the FTP folder, are also available through ESummary, and moving them to the E-utilities would reduce the dependence of FGTD on the HTML structure of GEO, which we regard as the main direction for maintaining the retrieval layer. Pages are requested sequentially from a single browser session, one series at a time and without parallel requests, corresponding in the case study to about 2.6 s per PMID-linked and 1.5 s per orphan series. The E-utilities are used where they represent the natural interface, i.e. for the PubMed records linked to each series, through Bio.Entrez, which limits requests to three per second (or ten with an NCBI API key).

Provenance of the technical filters and query construction. All the facets exposed by the advanced filter panel are the native filters published by NCBI GEO DataSets for the query at hand: FGTD

retrieves the facet list from GEO and reports it verbatim, without renaming, reinterpreting or adding filtering criteria of its own. Since the syntax of the GEO advanced search is unforgiving, and a single malformed field tag silently returns an empty result set, FGTD is designed so that users select the facets from this menu rather than typing filter expressions by hand; the application then assembles the corresponding GEO-native advanced query on behalf of the user, and a "Go to GEO" control opens the very same query in the NCBI web interface, so that the result set can be verified independently at any time. Finally, FGTD accepts an optional NCBI API key, which raises the E-utilities rate limit from three to ten requests per second, the key being stored and used strictly locally; all functionality remains available without one.

Semantic ranking and hybrid relevance score.

Users provide a primary biological query (e.g., "breast cancer") and, optionally, a list of keywords and MeSH terms. After the technical extraction, the datasets are scored by the pipeline shown in Supplementary Fig. S2.

The query vector is obtained by concatenating the GEO search string of the user with the keyword list and encoding the result with the all-MiniLM-L6-v2 sentence transformer [40] into a 384-dimensional vector. Each dataset is encoded from its title and GEO summary (the summaries of SuperSeries, which only list their SubSeries, are left empty), and the dataset embeddings are indexed in an in-memory Qdrant vector database [41]. A cosine-similarity search returns a raw similarity value for every record, which is clipped at zero and scaled by the batch maximum to give the semantic score (Sscore). The semantic, keyword and MeSH scores are then combined into a textual relevance score (Tscore ≤ 1.0) with weights of 50%, 30% and 20%, respectively. The MeSH score is binary, being 1.0 if at least one user-defined MeSH term is among the MeSH terms of the linked publication and 0.0 otherwise. When a component is inactive for a given search, its weight is redistributed over the remaining components so that the weights always sum to one: 60% semantic and 40% keyword without MeSH, 70% semantic and 30% MeSH when no keyword matches any record, and the semantic term alone when neither is available. The complete specification is given below.

In parallel, the pipeline computes a bibliographic impact score (Bscore) as the mean of two normalized terms, i.e. citation count (Cscore) and publication recency (Rscore). Citation counts are compressed with $\log(1+x)$ to limit the influence of extreme values, while recency is measured in days from the oldest publication date in the batch and scaled to [0, 1], so that the more recently published datasets receive a higher score. These terms are applied to every search and cannot be disabled from the interface. Since both terms are scaled within the batch, the oldest record of a search receives a

recency of 0 and the most recent a recency of 1 whatever the time span of the batch, so that the effective weight of recency differs between searches.

The final relevance score weights the textual score at 0.75 and the bibliographic score at 0.25 (equation (7)). The textual component is intended to drive the ranking, while the bibliographic one modulates the order among datasets of comparable topical relevance; this modulation can nevertheless change the top of the ranking, the hybrid split being the constant to which the ordering is most sensitive (see the sensitivity analysis below).

Scoring framework. All the scores are computed within the batch D of n datasets returned by a query, and they are therefore relative to that cohort and not comparable across separate searches.

Normalization. Raw values are clipped at zero and scaled by the batch maximum (equation (1)), and when the maximum is zero every normalized value is set to zero. This is equivalent to a min-max scaling with the minimum fixed at zero, which preserves the relative magnitude of the raw signal without forcing the weakest record of every batch to zero.

$$norm(x_i) = \frac{x_i}{\max_j x_j}, norm(x_i) = 0 \text{if} \max_j x_j = 0 \quad (1)$$

Semantic score. The query string and the keyword list of the user are concatenated and embedded with all-MiniLM-L6-v2 into a 384-dimensional vector q, while each dataset i is embedded into d_i from its title and GEO summary. Cosine similarities are computed in Qdrant and then normalized (equation (2)).

$$S_{raw}(i) = \cos(q, d_i), S_{score}(i) = norm\big(S_{raw}(i)\big) \quad (2)$$

Keyword score. Let k_i be the number of user-defined keywords that occur (case-insensitive string match) in the title and GEO summary of dataset i, and K_max the largest such count in the batch (equation (3)); when K_max = 0 the component is deactivated and Kscore(i) = 0 for all i. The exponent 1.35 lies between a linear decay, which leaves technically incomplete records competitive with exact matches, and a quadratic one, which suppresses semantically related records satisfying only part of the technical constraint (Supplementary Fig. S1). Keywords are matched as substrings, so that spelling variants supplied as separate keywords (e.g., triple-negative and triple negative) are counted separately; merging them in the first evaluation changed the ranks of the known items by at most 16 positions (Supplementary Table S16). The binary keyword columns of the output table are

computed on the abstract (see Automated retrieval), not on the title and GEO summary used for the score, so that filtering the table on them does not reproduce the ranking.

$$K_{score}(i) = \left(\frac{k_i}{K_{\max}}\right)^{1.35}, K_{\max} = \max_j k_j, K_{score}(i) = 0 \text{if} K_{\max} = 0 \quad (3)$$

MeSH score. Mscore(i) = 1 if at least one user-supplied MeSH term is among those retrieved for dataset i through Bio.Entrez, and 0 otherwise (equation (4)). Orphan datasets, having no linked publication and therefore no MeSH annotation, score 0 by construction.

$$M_{score}(i) = \begin{cases} 1 & \text{if} M_{user} \cap M_i \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Textual relevance. The three components are combined linearly (equation (5)), a component being considered active when it is non-zero for at least one dataset of the batch. The weights are (w_S, w_K, w_M) = (0.50, 0.30, 0.20) when both the keyword and the MeSH components are active, (0.60, 0.40, 0.00) without MeSH, (0.70, 0.00, 0.30) without keywords and (1.00, 0.00, 0.00) without either. Since the weights sum to one in every case, Tscore is bounded in [0, 1] irrespective of which components were supplied by the user.

$$T_{score}(i) = w_S S_{score}(i) + w_K K_{score}(i) + w_M M_{score}(i), w_S + w_K + w_M = 1 \quad (5)$$

Bibliographic impact. With c_i the citation count and ord(t_i) the proleptic Gregorian ordinal of the publication date, citation counts are compressed logarithmically and recency is measured relative to the oldest valid date in the batch, averaging the two terms with equal weight (equation (6)). Log compression prevents a single highly cited study from flattening the rest of the batch. When no valid date is available, Rscore(i) = 0.5 for all i, i.e. a neutral value that leaves citations as the only discriminating term.

$$C_{score}(i) = norm\big(\log(1 + c_i)\big), R_{raw}(i) = ord(t_i) - \min_j ord\big(t_j\big), B_{score}(i) = 0.5 R_{score}(i) + 0.5 C_{score}(i) \quad (6)$$

Global relevance. The final score combines the textual and the bibliographic terms in a 0.75/0.25 hybrid (equation (7)).

$$Relevance(i) = 0.75 T_{score}(i) + 0.25 B_{score}(i) \quad (7)$$

Constants and sensitivity analysis. The constants of the framework are design choices, fixed by inspecting the rankings of exploratory queries rather than estimated from labelled data: the decay exponent, the textual weights and their fallbacks, the equal split of Bscore and the hybrid split. Together they encode the position that textual and technical adherence should dominate the ranking, with bibliometrics acting as a secondary modulator among datasets of comparable pertinence, so that a recent and highly pertinent dataset is not displaced by an older one, more cited but less pertinent. All the constants are exposed as named parameters in the source code. We assessed on the breast cancer cohort the three constants with the largest expected influence, varying the decay exponent between 1.00 and 2.00, the hybrid split between 1.00/0.00 and 0.50/0.50 and the textual weight triple over seven alternatives (including a uniform 1/3 split); these one-factor sweeps were complemented by a full exponent-by-split grid of six exponents and five splits, for a total of 49 evaluations. Since all the component scores are stored, the scores were recomputed analytically, recovering the keyword ratio by inverting equation (3). Each configuration was compared with the default by Kendall's τ and by top-k overlap (Fig. 6a–c), and by the rank shift of each dataset (Fig. 6d–f; complete tables in Supplementary Table S12).

The ranking turned out to be most robust to the decay exponent (τ = 0.92–0.98; the default top three always included the same three datasets, with the first two exchanging places at an exponent of 1.00; maximum shift of two positions), moderately robust to the textual weights (τ = 0.85–0.90; maximum shift of four positions, also under a uniform split) and most sensitive to the hybrid split, since removing the bibliometric term or raising it to parity lowered τ to 0.77 and 0.66 respectively, with shifts of up to six positions. This is expected, the hybrid split encoding the editorial choice of the system, and it is the parameter that users should consider before changing the defaults. The default top-5 retained at least 80% overlap in all the configurations, whereas the top-ranked dataset changed in a minority of them, mostly at extreme splits; the ranking should therefore be read as a stable ordering of the upper part of the cohort rather than as the identification of a single best dataset. These results come from a single cohort of 15 datasets and characterize the stability of the parameters, not the correctness of the ranking (see Discussion).

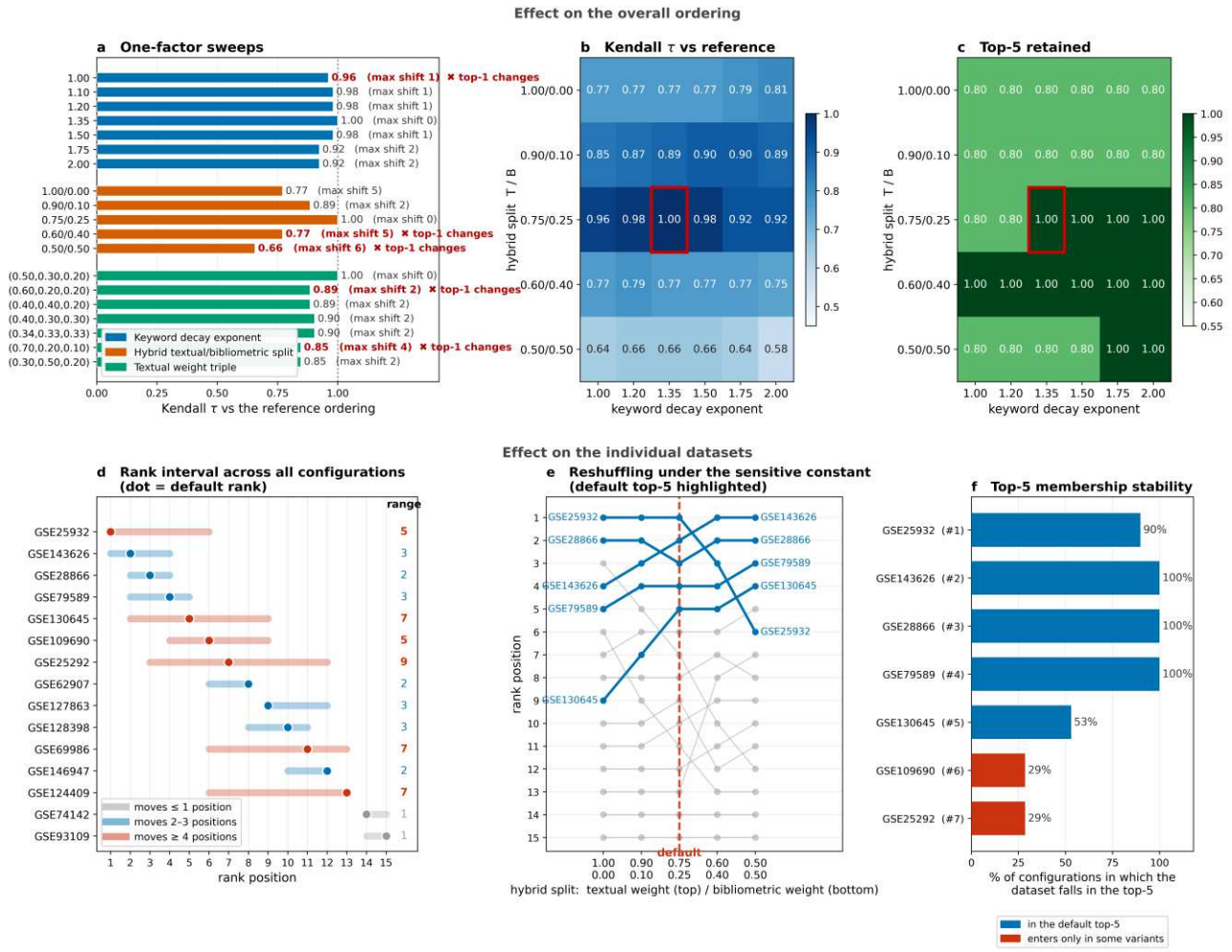


Figure 6. *Sensitivity of the FGTD ranking to its scoring constants. The 15 datasets of the breast cancer case study were re-scored under 49 alternative parameterizations, recomputed analytically from the stored component scores without re-running the retrieval pipeline. (a–c) Effect on the overall ordering. (a) One-factor sweeps: Kendall's tau between each variant ordering and the default ordering, grouped by the constant being varied; the maximum rank displacement observed in the cohort is annotated next to each bar, and variants in which the highest-ranked dataset changes are marked. (b, c) Full grid over the decay exponent and the hybrid textual/bibliometric split, reporting Kendall's tau and the fraction of the default top-5 that is retained; the red box marks the configuration used throughout the paper (exponent 1.35, split 0.75/0.25). The exponent is the least consequential constant and the hybrid split the most. (d–f) The same 49 evaluations, resolved at the level of the individual dataset. (d) Rank interval spanned by each dataset, with the dot marking its position under the default configuration and the number on the right giving the width of the interval; two datasets never move by more than one position, and the largest excursion is nine positions. (e) Rank*

trajectories under the hybrid split, the constant identified as the sensitive one; the five datasets forming the default top-5 are highlighted and the dashed line marks the default split. Reordering is confined almost entirely to the upper half of the cohort. (f) Percentage of configurations in which each dataset falls in the top-5: three of the five default top-5 datasets hold their position in every configuration tested, while two positions are contested by two further datasets that enter only under particular parameterizations.

Frontend–backend communication and visualization. The React-based frontend provides an interactive interface for querying, filtering and visualizing the results, with react-router-dom handling navigation between views and Electron managing the application lifecycle by bridging the frontend with the backend executable. The frontend communicates with the local backend through REST APIs using axios [42], which exchanges asynchronous JSON payloads to validate queries and parameters before the full data extraction.

Data management, analytics and interactivity. The extracted metadata is organized into a spreadsheet-like table that supports filtering and sorting, and any filter applied to the results is reflected in real time in the visual analytics. The main exploratory view is the interactive Semantic Similarity Map. During extraction, the backend includes in the JSON payload the 384-dimensional embeddings generated by the all-MiniLM-L6-v2 model, which are parsed by the React frontend to compute the cosine similarity between a reference dataset selected by the user and every other dataset of the table. The result is displayed as a force-directed network graph (vis-network), in which nodes represent studies, edges connect the reference dataset to the datasets whose similarity to it is at least 0.40, and node colors mark similarity classes (thresholds 0.45, 0.65 and 0.85); the same datasets are listed in an exportable table of similar projects. Since the map is computed in the frontend, it does not require external calls. All the visualizations can be exported as .svg files and are complemented by keyword trend analyses and distribution pie charts that respond to the input of the user, while filtered datasets can be exported in .xlsx and .csv formats.

Efficiency benchmark. The case-study run used the GEO filters described in Results, and the advanced query generated for this search was ("breast neoplasms"[MeSH Terms] OR breast cancer[All Fields]) AND "gse"[Filter] AND "homo sapiens"[Organism] AND "Expression profiling by high throughput sequencing"[Filter] AND ("tissue"[Attribute Name] OR "strain"[Attribute Name] OR "gender"[Attribute Name]) AND ("supplementary cell"[Filter] OR "supplementary bed"[Filter]) AND ("2010/01/01"[PDAT] : "2020/12/31"[PDAT]) AND ("gse"[Filter] AND "Homo sapiens"[Organism] AND "Expression profiling by high throughput sequencing"[Filter] AND

("attribute name tissue"[Filter] OR "attribute name strain"[Filter] OR "gender"[Attribute Name]) AND ("supplementary cell"[Filter] OR "supplementary bed"[Filter])). The run was performed in Ultra Mode. Its duration was measured on 13 October 2026 with the released Linux build of FGTD, repeating the run three times without an NCBI API key, so that the E-utilities requests were limited to three per second; it covers the complete pipeline, from the GEO search to the scored output table and the exported file, and the time per series was derived from the timestamps of the application log. The repeated run returned the same 15 series as the original case study. The three curators received only written instructions and a blank annotation form before the exercise; no example output of FGTD was provided (Supplementary Note 1 and Supplementary Table S4). For each of the two series, each curator started a timer when opening the first link, filled in the fields produced by FGTD (title or PubMed link, series accession, date, platform, organism, instruments, abstract or summary, links to files and databases, the first three sample identifiers and the sample count, citation count, study type, a personal relevance score from 0 to 1, and binary MeSH and keyword indicators) and stopped the timer once the row was complete. The time for the cohort was extrapolated as 12 times the mean time for the PMID-linked series plus 3 times the mean time for the orphan series.

Concordance analysis. The FGTD row of each benchmark series (Supplementary Table S3) was compared with the corresponding row of each curator (Supplementary Tables S5–S7) for the 30 objective fields of the form, excluding the personal relevance score. Values were first normalized only for format (accessions extracted from links, dates parsed, text lowercased and stripped of punctuation), treating empty cells, “0” and “Not found” as the same absent value; since FGTD omits from its export the keyword columns that are zero for every series of a run, the missing “immunotherapy” column was read as 0. Each pair of values was then classified as agreeing, as differing only in form (the same entity in another representation, or a citation count differing by at most two, i.e. the citations accrued between the FGTD run and the manual exercise) or as differing in content. Each content discordance was resolved against the GEO record (E-utilities ESummary and GEO SOFT file) and the PubMed record, retrieved on 12 October 2026; for keyword indicators the reference was the whole-word presence of the keyword in the text named by the protocol, i.e. the PubMed abstract for the linked series and the GEO summary for the orphan series. FGTD exports the GEO summary in the summary field also for PMID-linked series, whereas the protocol asked the curators for the PubMed abstract, and the three resulting differences were classified as differences in form; since the two are different texts, this choice favors FGTD, and classifying them as differences in content would give 9 differences in form and 25 in content. The analysis scripts and the source records are provided as Supplementary Data 3.

Ethics statement. The manual benchmark involved three volunteers (two of the authors and an undergraduate intern), who timed themselves while extracting public metadata from GEO and PubMed. No personal or health data were collected, and the participants are identified only by their initials. All three participants gave informed consent to take part and to be identified by their initials in this article. Under the regulations of the University of Palermo, a task of this kind, involving no intervention on the participants and no collection of personal data, does not require approval by an ethics committee. All methods were performed in accordance with the relevant guidelines and regulations.

Known-item retrieval evaluation. The ranking was assessed with a known-item protocol against an external reference set, independent of the breast cancer cohort used for the sensitivity analysis. Cao et al. [27] report five GEO series (GSE27447, GSE36295, GSE43358, GSE61724 and GSE75678) retained for a meta-analysis of prognostic genes in triple-negative breast cancer, selected as mRNA expression profiling by array of human tissue from triple-negative and other breast cancers, and these five accessions were taken as the known items. A single FGTD run in Ultra Mode reproduced the corresponding search space, breast AND "gse"[Filter] AND "Homo sapiens"[Organism] AND ("Expression profiling by array"[Filter]), restricted to the records published between 1 January 2011 and 31 December 2015, returning 1,264 series that contained all five known items. The keywords (gene expression, prognosis, survival, tnbc, triple-negative, triple negative and tumor tissue) and the MeSH terms (Gene Expression Profiling and Triple Negative Breast Neoplasms) were chosen by the authors, who knew the five reference series, to reflect the aim and the inclusion criteria of the meta-analysis; the sensitivity of the results to this choice was assessed with alternative keyword lists (Robustness analyses). Two orderings of this pool were then compared: the native GEO order (i.e. the order in which the GEO DataSets search lists the records, which in both pools coincided with reverse publication date, with Spearman $\rho = -1.00$ and -0.99 between position and date), used as baseline, and the order induced by the FGTD relevance score, with ties broken by the native GEO position. Since the FGTD export does not preserve the order in which GEO lists the records, the native order of the same pool was recovered through ESearch, which returns the records in the order of the GEO DataSets web page (retrieved on 12 October 2026; 1,258 of the 1,264 series are still returned, and the remaining six, none of which is a known item, were placed last). A relevance-sorted GEO listing could not be added as a further baseline, since ESearch ignores the sort parameter on the GEO DataSets database (relevance and publication-date sorting returned the same order; tested on 19 October 2026). For each ordering we computed Precision@5 and Precision@10 (known items retrieved within the top k, divided by k), NDCG@5 and NDCG@10 with binary gain and log2 rank

discount, and the reciprocal rank (RR) of the first known item (equations (8) and (9)), this last being reported for each query and as a mean over the two queries (mean reciprocal rank). Recall was not estimated, because the number of records screened by the original authors is not reported and the repository has changed since their analysis, so that the denominator required by such a figure is not available. Overall, fourteen orderings of the same pool were compared: the native GEO order, the full FGTD score, five single-signal baselines (Okapi BM25, i.e. a lexical ranking function based on term frequency and document length, computed over title and summary with k1 = 1.5 and b = 0.75, and each of the keyword, semantic, MeSH and bibliometric components in isolation) and four variants of the composite score (the topical component T alone, and three ablations obtained by removing the semantic, the keyword or the MeSH term and renormalizing the remaining weights), and the three hybrid orderings described below. All the orderings were evaluated with the same metrics on the same five known items (Supplementary Table S11).

Three hybrid orderings were defined before their evaluation, and all of them are reported: H1, 0.5·S + 0.5·BM25n, where BM25n is the BM25 score normalized by the batch maximum as in equation (1); H2, the full FGTD score with BM25n in place of the keyword score; H3, the reciprocal rank fusion of the FGTD and BM25 orderings, scored as 1/(60 + r_FGTD) + 1/(60 + r_BM25), with r the rank under each ordering. In all the BM25-based orderings the query consisted of the distinct terms of the keyword list.

Second known-item evaluation. Li et al. [28] integrated three GEO series (GSE3744, GSE8977 and GSE10797) of breast tumor and normal breast tissue profiled with expression arrays. The evaluation was repeated with the same protocol on the query breast AND "gse"[Filter] AND "Homo sapiens"[Organism] AND ("Expression profiling by array"[Filter]), restricted to the records published between 1 January 2006 and 31 December 2008 (the years spanned by the known items), and the run returned 273 series containing all three. The keywords (gene expression, metastasis, progression, breast cancer, breast tumor, tumor tissue and normal tissue) and the MeSH terms (Gene Expression Profiling and Breast Neoplasms) were fixed before the run, starting from the aim of the study and from the comparison it makes and following the structure of the first evaluation. The run used the FGTD release of October 2026, whose scoring is identical to that used for the first evaluation. The native GEO order was recovered through ESearch as above (one series no longer returned was placed last), and the fourteen orderings were evaluated with the same metrics, also reporting the reciprocal rank and NDCG@10 as means over the two queries (Supplementary Tables S11 and S15).

Robustness analyses. The relevance score was re-implemented offline from the stored components, reproducing the published scores exactly. Robustness to the composition of the pool was assessed on 200 random subsamples of 90% of each pool, always retaining the known items, re-scored with every component renormalized on the subsample as FGTD does. Sensitivity to the keyword list was assessed with three alternative lists per query, written from the aim of each meta-analysis and fixed before computing; the semantic score was recomputed with all-MiniLM-L6-v2 from the GEO query and the new keywords and the keyword score from the title and GEO summary, the MeSH and bibliographic terms being unchanged. For the first query the exact query string of the original run is not stored, the reconstructed semantic score correlating 0.993 with the stored one, so the published list was re-scored with the same reconstruction as reference. Results and scripts are provided in Supplementary Table S16 and Supplementary Data 3.

$$P@k = \frac{r_k}{k}, r_k = \text{ number of known items ranked within the top } k, RR = \frac{1}{\rho_1} \quad (8)$$

$$DCG@k = \sum_{p=1}^{k} \frac{rel_p}{\log_2(p+1)}, NDCG@k = \frac{DCG@k}{IDCG@k}, rel_p = 1 \text{ for a known item and 0 otherwise} \quad (9)$$

Large language model assistance. During the development of FGTD, Google Gemini was used to assist with coding. During the preparation of this manuscript, Claude (Anthropic) was used to assist with the revision of the text, which was then edited by the authors, and with the management of the GitHub repository. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the content of the software and of the publication.

Data availability

All data analyzed in this study are publicly available from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and from PubMed (https://pubmed.ncbi.nlm.nih.gov/) under the accessions and identifiers reported in the text. The configuration of the breast cancer case study, the metadata and relevance scores produced by FGTD for the case study and for the two known-item retrieval evaluations, the manual benchmark protocol and annotations and their concordance with the FGTD output, the sensitivity analysis of the scoring constants, the baseline and ablation comparison, the robustness analyses and the similarity scores of the known-item seed are provided as Supplementary Information (Supplementary Tables S1–S16, Supplementary Data 1–3, and Supplementary Notes 1 and 2).

Code availability

FGTD is released under the MIT license. The source code and the pre-compiled installers for Windows and Linux are available at https://github.com/SorenKK/FGTD. The version described in this article (v1.0.0, including the correction of the GEO FTP paths) is archived in Zenodo at https://doi.org/10.5281/zenodo.23472013, and all releases are available under the concept DOI https://doi.org/10.5281/zenodo.20731387, which resolves to the most recent one. macOS is not supported. Building from source requires Python 3.10.12, Node v20.20.2 and npm 10.8.2. An NCBI API key is optional and, when supplied, raises the E-utilities rate limit from three to ten requests per second. There are no restrictions on use by non-academics.

References

1. Clough, E. & Barrett, T. The Gene Expression Omnibus database. *Methods Mol. Biol.* **1418,** 93–110 (2016).

2. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **52,** D33–D43 (2024).

3. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2,** 401–404 (2012).

4. Hu, W., Zaveri, A., Qiu, H. & Dumontier, M. Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. *BMC Bioinformatics* **18,** 415 (2017).

5. Lee, K. et al. Scaling up data curation using deep learning: an application to literature triage in genomic variation resources. *PLoS Comput. Biol.* **14,** e1006390 (2018).

6. Ahmed, M. & Kim, D. R. Editorial: opportunities and challenges in reusing public genomics data. *Front. Pharmacol.* **14,** 1226756 (2023).

7. Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S. & Chen, Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics* **24,** 2798–2800 (2008).

8. Grigoriadis, D., Tsifintaris, M., Giannakakis, A., Pavlopoulos, G. A. & Perdikopanis, N. Public Omics Explorer (POE): enabling integrative semantic search across GEO omics datasets based on PubMed publications. *Comput. Struct. Biotechnol. J.* **27,** 4802–4812 (2025).

9. Mecham, A., Stephenson, A., Quinteros, B. I., Brown, G. S. & Piccolo, S. R. TidyGEO: preparing analysis-ready datasets from Gene Expression Omnibus. *J. Integr. Bioinform.* **21,** 20230021 (2024).

10. Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22,** 323 (2021).

11. Zhao, W. et al. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Res.* **50,** D287–D294 (2022).

12. Sarkans, U. et al. From ArrayExpress to BioStudies. *Nucleic Acids Res.* **49,** D1502–D1506 (2021).

13. Connected Papers. Find and explore academic papers. https://www.connectedpapers.com (2026).

14. ResearchRabbit. AI tool for smarter, faster literature reviews. https://www.researchrabbit.ai (2026).

15. Elicit. AI for scientific research. https://elicit.com (2026).

16. Barker, M. et al. Introducing the FAIR principles for research software. *Sci. Data* **9,** 622 (2022).

17. Katz, K. et al. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50,** D387–D390 (2022).

18. Barrett, T. et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40,** D57–D63 (2012).

19. Consensus. AI for research. https://consensus.app (2026).

20. Paperguide. The AI research assistant. https://paperguide.ai (2026).

21. SciSpace. AI for research. https://scispace.com (2026).

22. The Lens. Search, analyze and manage patent and scholarly data. https://about.lens.org (2026).

23. Herzog, C., Hook, D. & Konkiel, S. Dimensions: bringing down barriers between scientometricians and data. *Quant. Sci. Stud.* **1,** 387–395 (2020).

24. OpenAlex. The open catalog to the global research system. https://openalex.org (2026).

25. Semantic Scholar. AI-powered research tool. https://www.semanticscholar.org (2026).

26. Zenodo. https://zenodo.org (2026).

27. Cao, W. et al. Identification of novel prognostic genes of triple-negative breast cancer using meta-analysis and weighted gene co-expressed network analysis. *Ann. Transl. Med.* **9,** 205 (2021).

28. Li, Y., Bai, W. & Zhang, L. The overexpression of CD80 and ISG15 are associated with the progression and metastasis of breast cancer by a meta-analysis integrating three microarray datasets. *Pathol. Oncol. Res.* **26,** 443–452 (2020).

29. Hegde, M., Girisa, S. & Kunnumakkara, A. B. A compilation of bioinformatic approaches to identify novel downstream targets for the detection and prophylaxis of cancer. *Adv. Protein Chem. Struct. Biol.* **134,** 75–113 (2023).

30. Zhai, J., Pitteloud, N. & Santoni, F. A. GenMasterTable: a user-friendly desktop application for filtering, summarising, and visualising large-scale annotated genetic variants. *BMC Bioinformatics* **26,** 211 (2025).

31. Wirth, F. N., Kussel, T., Müller, A., Hamacher, K. & Prasser, F. EasySMPC: a simple but powerful no-code tool for practical secure multiparty computation. *BMC Bioinformatics* **23,** 531 (2022).

32. Electron. Build cross-platform desktop apps with JavaScript, HTML, and CSS. https://www.electronjs.org (2026).

33. React. https://react.dev (2026).

34. Tailwind CSS. https://tailwindcss.com (2026).

35. Flask documentation (3.1.x). https://flask.palletsprojects.com/en/stable/ (2026).

36. PyInstaller. https://pyinstaller.org (2026).

37. Selenium. https://www.selenium.dev (2026).

38. Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/ (2026).

39. Clough, E. et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res.* **52,** D138–D144 (2024).

40. sentence-transformers/all-MiniLM-L6-v2. *Hugging Face* https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 (2024).

41. Qdrant. Vector search engine. https://qdrant.tech (2026).

42. axios. *npm* https://www.npmjs.com/package/axios (2026).

Acknowledgements

We thank Soraya Pizzimenti for designing the application logo and Luca Siragusa for beta testing the application and for taking part in the manual benchmark.

Author Contributions

F.O.: Conceptualization, Methodology, Software, Visualization, Writing – original draft. A.T.: Validation, Investigation, Writing – review & editing. F.A.: Validation, Investigation, Writing – review & editing. G.N.: Software, Writing – review & editing. G.S.: Resources, Supervision, Funding acquisition, Writing – review & editing. M.T.: Resources, Supervision, Funding acquisition, Writing – review & editing. S.D.B.: Conceptualization, Project administration, Supervision, Funding acquisition, Writing – review & editing. All authors read and approved the final manuscript.

Funding

This work was supported by PNRR-MAD-2022-12376183; by the AIRC project IG 2024 (30306) to M.T.; by the Progetto IMPACT (M1/2676, Piano Nazionale di Ripresa e Resilienza 2021–2027, CUP B79H26000150007, Codice UGOV PRJ-2296) to S.D.B. and G.S.; and by the Fondo Finalizzato di Ateneo (FFR)-2026 to S.D.B. and A.T.

Competing interests

The authors declare that they have no competing interests.

Supplementary Information

Supplementary Table S1. Capability matrix of research discovery tools. Spreadsheet comparing FGTD with 16 research discovery platforms across nine dimensions: target domain, cross-database linkage, metadata granularity, search architecture, relevance ranking logic, interface, data sovereignty, export output and primary workflow stage automated.

Supplementary Figure S1. Effect of the keyword exponent in the FGTD Top-Ranking Record logic.

Supplementary Figure S2. Scoring pipeline of FGTD.

Supplementary Data 1. Configuration of the breast cancer case study. JSON file with the query, the organism, study type, attribute name, date range and supplementary file filters, and the keywords and MeSH terms used for the run.

Supplementary Table S2. FGTD relevance scores for the breast cancer case study. Scoring breakdown (K_score, S_score, M_score, R_score, C_score, T, B and final relevance score) for the 15 retrieved series.

Supplementary Table S3. FGTD metadata output for the breast cancer case study. Structured metadata extracted by FGTD for the 15 series, including GEO summaries, platform details, sample identifiers, hyperlinks and binary keyword and MeSH indicators. FGTD omits keyword columns that are zero for every series; “immunotherapy” matched none of the 15 series and therefore has no column. The Vector column holds the 384-dimensional embedding of each series.

Supplementary Note 1. Standard operating procedure for the manual benchmark. Protocol given to the three curators before the manual curation benchmark, describing the fields to be extracted and the timing procedure.

Supplementary Note 2. Comparison of raw-data accessibility. Qualitative comparison of FGTD with an AI-driven literature synthesis tool, a scholarly mapping engine and a generalist data repository on a single query.

Supplementary Table S4. Manual benchmark form. Blank annotation form used by the curators, defining the metadata fields to be filled in.

Supplementary Table S5. Manual benchmark annotations of curator A.T. Metadata extracted manually for GSE124409 and GSE69986, with relevance scores and recorded times. Sheet a gives the record of each series in tidy form (rows split by the curator merged, times in decimal minutes); sheet b reproduces the form as submitted.

Supplementary Table S6. Manual benchmark annotations of curator F.A. Metadata extracted manually for GSE124409 and GSE69986, with relevance scores and recorded times. Sheet a gives the record of each series in tidy form (rows split by the curator merged, times in decimal minutes); sheet b reproduces the form as submitted.

Supplementary Table S7. Manual benchmark annotations of curator L.S. Metadata extracted manually for GSE124409 and GSE69986, with relevance scores and recorded times. Sheet a gives the record of each series in tidy form (rows split by the curator merged, times in decimal minutes); sheet b reproduces the form as submitted.

Supplementary Table S8. Known-item retrieval results. Rank of each of the five known items under the native GEO order and under the FGTD relevance ranking (sheet a), and Precision@5, Precision@10, NDCG@5, NDCG@10 and reciprocal rank for both orderings (sheet b).

Supplementary Table S9. FGTD relevance scores for the known-item retrieval evaluation. Scoring breakdown (K_score, S_score, M_score, R_score, C_score, T, B and final relevance score) for the 1,264 retrieved series, in the order of the FGTD export, with the native GEO rank of each series (rank_GEO_native).

Supplementary Table S10. FGTD metadata output for the known-item retrieval evaluation. Structured metadata extracted by FGTD for the 1,264 retrieved series, including the linked PubMed records and the binary keyword and MeSH indicators.

Supplementary Data 2. Semantic similarity neighborhood of GSE61724. CSV file with the similarity score of the 910 series whose cosine similarity to GSE61724 is at least 0.40, and the interactive Semantic Similarity Map exported by FGTD as a self-contained HTML page.

Supplementary Table S11. Baseline, ablation and hybrid comparison for the two known-item retrieval evaluations. For the first meta-analysis (Cao et al.; 1,264 series), sheet a reports, for each of the fourteen orderings, P@k, NDCG@k, the reciprocal rank and the median and mean rank of the five known items, and sheet b the rank of each known item; sheets c and d report the same for the second meta-analysis (Li et al.; 273 series, three known items); sheet e gives the mean reciprocal rank and the mean NDCG@10 over the two queries. The orderings are the native GEO order, Okapi BM25 over title and summary, the four components of the FGTD score in isolation, the topical score without the bibliometric term, three ablations in which one component is removed and the remaining weights renormalized, three hybrid orderings that combine BM25 with the semantic or the full FGTD score

(H1–H3, Methods) and the full FGTD relevance score, all computed from the scoring output and metadata of each run, with ties broken by the native GEO position.

Supplementary Table S12. Sensitivity analysis of the FGTD scoring constants. The 49 alternative parameterizations tested on the breast cancer cohort, with Kendall's τ against the default ordering, top-3, top-5 and top-10 overlap, maximum and mean rank displacement, and the per-dataset rank under every configuration.

Supplementary Table S13. Manual benchmark summary and concordance with the FGTD output. Sheet a lists, for each curator and series, the role of the curator, whether the curator is an author, the curation time in decimal minutes and as recorded, and the relevance score; sheets b and c summarize the field-by-field concordance between FGTD and each curator by series and by field; sheet d reports, for FGTD and for each curator, the number of values that differ from the GEO and PubMed records; sheet e lists the 180 comparisons with the two values, the class assigned and the source used to resolve each discordance. The incorrect FTP links of FGTD concern accessions below GSE100000 and are corrected in the archived version; the files of these series are also reachable from the GEO accession page.

Supplementary Table S14. Ten nearest neighbors of GSE61724 in the Semantic Similarity Map. Sheet a: for each neighbor, the similarity to GSE61724, the GEO series title and summary retrieved through ESummary on 12 October 2026, the triple-negative breast cancer term found in them and the relation to the seed (GSE61723 and GSE61724 are SubSeries of GSE61725); sheet b: the ten nearest neighbors outside that SuperSeries.

Supplementary Table S15. Second known-item retrieval evaluation (Li et al.). Configuration of the run, fixed before it (sheet a), FGTD relevance scores with the native GEO rank of each series (sheet b) and metadata exported by FGTD (sheet c) for the 273 retrieved series.

Supplementary Table S16. Robustness of the known-item results. Rank of the known items under the full FGTD score, BM25, the semantic component and the native GEO order over 200 random subsamples of 90% of each pool (sheet a); results with three alternative keyword lists per query (sheet b) and with the two spellings of triple-negative merged (sheet c); correlation between the native GEO order and the publication date (sheet d).

Supplementary Data 3. Concordance, hybrid-ranking and native GEO order analysis. ZIP archive with the scripts that compute the concordance (Supplementary Table S13) the hybrid orderings (Supplementary Table S11) and the robustness analyses (Supplementary Table S16), and the GEO

and PubMed records, retrieved on 12 October 2026, used to resolve the discordances, together with the comparison between the stored and the recomputed scores (rescore_check.csv).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1.json
- SupplementaryTableS1.xlsx
- SupplementaryTableS2.csv
- SupplementaryTableS8.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS6.xlsx
- SupplementaryTableS7.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryData2.zip
- SupplementaryTableS11.xlsx
- SupplementaryTableS12.xlsx
- SupplementaryTableS13.xlsx
- SupplementaryTableS9.csv
- SupplementaryData3.zip
- SupplementaryTableS16.xlsx
- SupplementaryTableS14.xlsx
- SupplementaryTableS15.xlsx
- SupplementaryFigureS2.png
- SupplementaryInformation.pdf
- SupplementaryFigureS1.png
- SupplementaryTableS10.xlsx