Supplementary Information

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

Francesco Orilio, Alice Turdo, Francesca Angeloro, Giovanni Nigita, Giorgio Stassi, Matilde Todaro, Sebastiano Di Bella

Supplementary Note 1

Standard operating procedure for the manual benchmark. The protocol below was given to the three curators before the manual curation benchmark; it is reproduced as distributed.

Data Extraction Test: Instructions for Completion

Attached, you will find two Excel files: a Template (with two completed examples) and a file titled To_Be_Filled.

[Note: the template mentioned above was not attached; the curators received only these instructions and the blank form (Supplementary Table S4).]

The goal of this test is to measure the time required to extract data from 2 scientific papers from the NCBI databases (GEO/PubMed) and structure them correctly within the Excel table.

Before starting: Start a timer on your computer or smartphone. The time will start running the moment you open the first link and will stop only when you finish filling out the data for the second paper.

Part 1: Paper with PMID (GSE124409)

Your first task is to visit the GEO database (https://www.ncbi.nlm.nih.gov/gds), search for the ID GSE124409 and press enter; it is always the first result. Start filling out the Excel file (To_Be_Filled) by moving through the columns from left to right:

- Title/PMID: Check if a PMID (a link to PubMed) is present on the GEO page. If it is, open PubMed and paste the publication link into the Excel file.
- GSE: Enter the GSE code in the Excel file, formatting it as a hyperlink.
- Date: Note the publication date (you can find it on GEO under Status or Release date).
- Platform: Transcribe the name of the platform or technology used (e.g., Illumina NovaSeq...).
- Organisms: Note the organism studied (e.g., Homo sapiens).
- Instrument (1, 2, 3): Note the identification codes for the instruments/platforms (codes starting with "GPL"). If there is only one instrument, enter 0 for Instrument_2 and/or 3.
- Summary: Since a PMID is present, READ, copy, and paste the Abstract directly from the PubMed page.
- Link to Files and Database: Collect the various links from the GEO page:

- Series Matrix Link, SOFT, MINiML: look for the files at the bottom of the page and insert the links (or write "Not found").
- BioProject, Geo2R, Other links, and SRA: enter the codes (e.g., PRJNA...) or the corresponding reference links.
- Samples and Samples_Count: Enter the GSM codes for the first 3 samples (Samples_1, _2, _3) and the total number of samples analyzed (Samples_Count).
- Citations_Count: How many times has the article been cited? (Found on PubMed).
- Study_Type_Extracted: Transcribe the study type reported on GEO (e.g., Expression profiling by high throughput sequencing).
- Relevance Score (0 to 1): Assign a personal score from 0 (not relevant at all) to 1 (extremely relevant).
- How to assess the Relevance Score: Put yourself in the shoes of a researcher filtering databases. Your goal is to find: Clinical or basic studies on breast cancer conducted on humans (Homo sapiens), published between 2010 and 2020. You are specifically interested in studies that analyze gene expression using high-throughput sequencing (such as RNA-Seq), that include clear metadata (such as tissue type or patient sex), and that provide raw or supplementary data (cell/bed file formats) for re-analysis.
- Mesh Terms (Female, Humans): Scroll to the bottom of the paper page on PubMed and open the "MeSH terms" drop-down menu. If the words "Humans" and/or "Female" are present, enter 1 in the respective columns. If they are not present, enter 0.
- Keywords (breast, dna, immunotherapy, protein, rna, tumor, Vector): These are keywords that you must match exactly within the text of the Abstract you just read. For each word: enter 1 if it is present, 0 if it is not.

Once you have completed the last column for the first paper:

- Stop the timer immediately.
- Take a screenshot or photo showing both the fully completed Excel file and the final time on your timer.

Part 2: Paper without PMID (GSE69986)

Now move on to the second paper by searching for the ID GSE69986 on GEO.

This specific study does not contain a reference to PubMed (no associated PMID). Consequently, all information must be obtained exclusively from the GEO page.

Repeat the workflow:

- Title/PMID: Since there is no PubMed, enter the Title indicated on GEO.
- GSE: Enter the GSE code, formatting it as a link.
- Complete Date, Platform, Organisms, Instrument, and database links exactly as done before.
- Summary: Copy and paste the GEO Summary (since there is no PubMed Abstract). Citation count will be 0.
- Relevance Score: Re-evaluate from 0 to 1 how relevant this study is to the research mission described above.
- Mesh Terms: Since there is no PubMed, enter 0 for Humans and Female.

- Keywords: Search for the same keywords (breast, dna, etc.) within the GEO Summary you copied. Again: 1 if present, 0 if absent.

End of Test

Once you have completed the last column for the second paper:

- Stop the timer immediately.
- Take a screenshot or photo showing both the fully completed Excel file and the final time on your timer.
- Submit the materials for verification. Good luck!

Supplementary Note 2

Comparison of raw-data accessibility

To place FGTD within the existing ecosystem, we compared it with one platform for each of three categories of discovery tools, i.e. an AI-driven literature synthesis tool (Elicit), a scholarly mapping engine (Semantic Scholar) and a generalist data repository (Zenodo). The comparison assessed not the retrieval of literature as such, but how directly each platform leads from a biological question to actionable experimental data, and it was based on the question "Transcriptomic signatures of immunotherapy resistance in breast cancer".

Four descriptors, defined operationally, were used, the first being distance-to-data, i.e. the sequence of manual actions that separates a result shown by the platform from the raw experimental file (from the result list to the download link). Metadata granularity is the finest unit of experimental description exposed by the platform within its own interface, ordered as no experimental metadata, study-level metadata and sample-level metadata (i.e. individual GSM records with their characteristics fields). Semantic noise is the proportion of returned records that a domain expert would judge off-target for the query. Data sovereignty indicates whether the query text and its semantic processing remain on the user's machine or are handled by a third-party server. These descriptors are qualitative and were assessed by inspection on a single query, characterizing the operational differences between categories of tools without ranking them quantitatively.

Elicit turned out to be effective for conceptual discovery and for a rapid summarization of the literature, but its primary output does not expose raw experimental units: although it can identify papers that generated data, it does not systematically parse or link the accession identifiers of primary data repositories such as GEO, ArrayExpress or SRA. Reaching the data therefore requires opening the full text, finding the data availability statement, copying the accession codes and querying the corresponding database (long distance-to-data and no experimental metadata within the platform).

Semantic Scholar offered a shorter path to the primary literature through its PubMed integration, while remaining bibliographic: starting from the publication, the user still has to navigate to the associated repositories, download the files and organize the technical metadata in a table, a step that becomes time-consuming for large cohorts.

Zenodo, which relies on keyword-based indexing, returned a high proportion of off-target records without retrieving transcriptomic matrices matching the query, resulting in a high semantic noise.

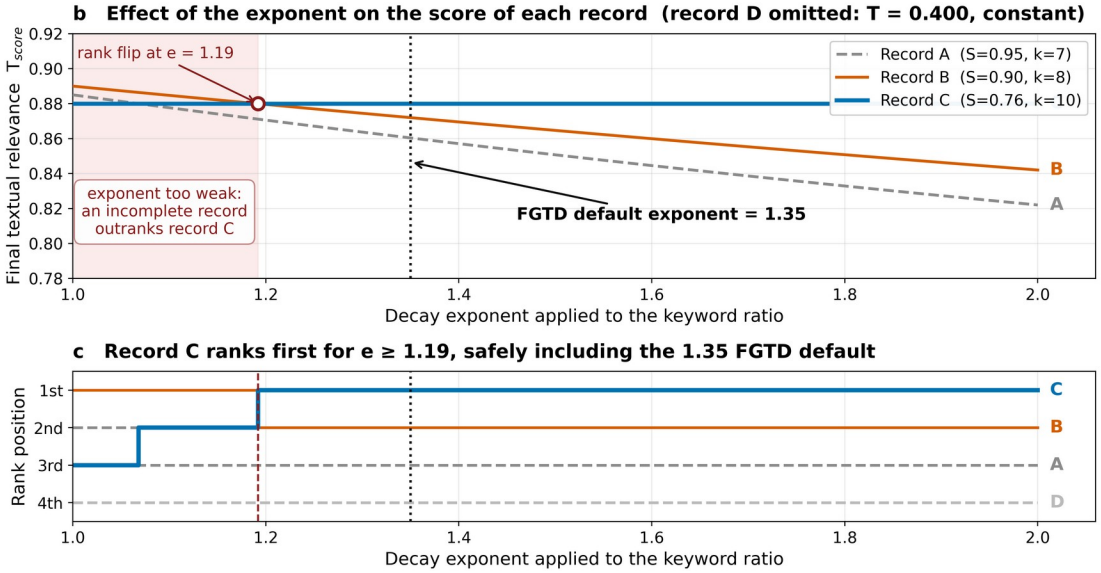
FGTD returned GEO series associated with the query (e.g., GSE270235 and GSE176580), all of which were judged pertinent to the query by the authors on inspection, exposing their sample-level metadata together with direct links to the raw data. The series returned included, for instance, responses to checkpoint inhibition in metastatic triple-negative breast cancer). The query embedding and the scores were computed locally, and only the search itself was submitted to NCBI.

Supplementary Figures

Calibration of the keyword decay exponent in the FGTD Top-Ranking Record Logic

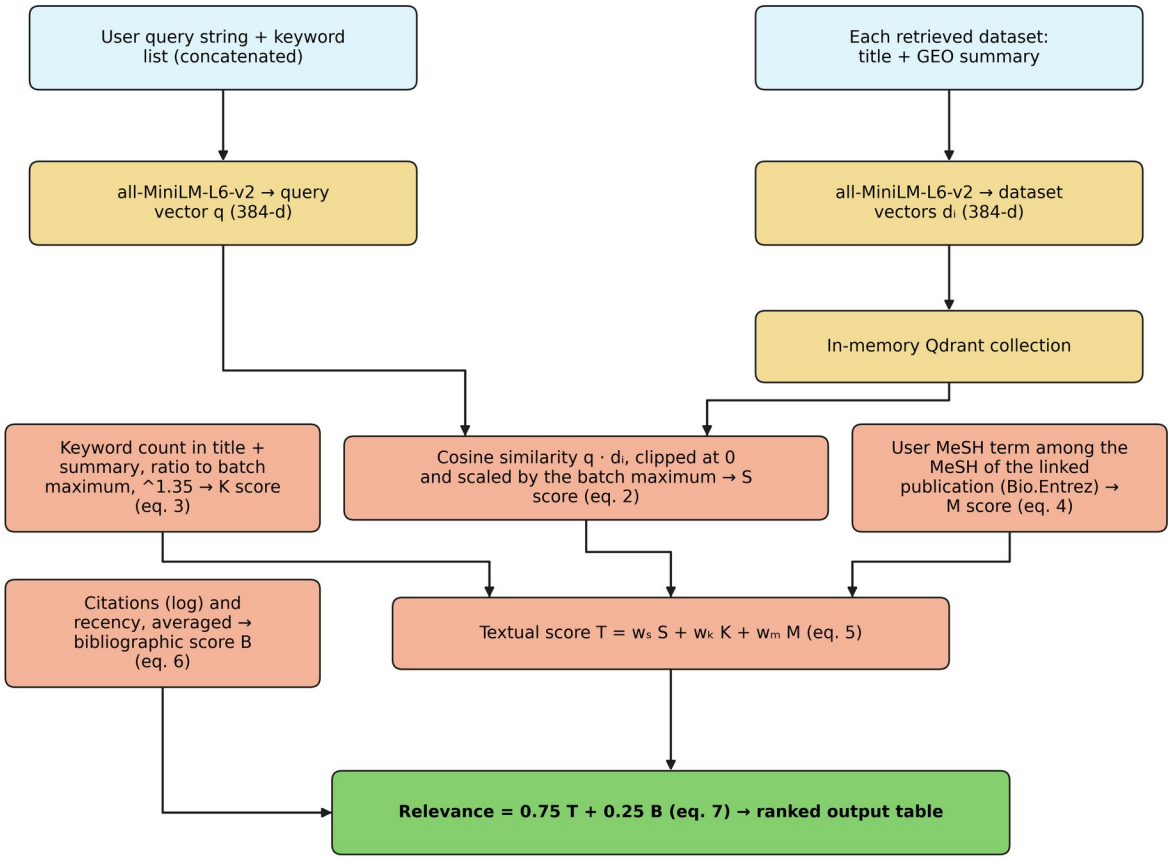
a Synthetic records used to probe the keyword decay filter

Record	Semantic score S	Keyword matches k	Keyword ratio k / k_{max}	MeSH M	Description
A	0.95	7 / 10	0.7	1	high semantics, partial keyword match
B	0.90	8 / 10	0.8	1	high semantics, near-complete keyword match
C	**0.76**	**10 / 10**	**1.0**	**1**	**average semantics, complete keyword match (Top-Ranking Record)**
D	0.40	0 / 10	0.0	1	low semantics, no keyword match



Supplementary Figure S1. Effect of the keyword exponent in the FGTD Top-Ranking Record logic. (a) Four synthetic records were constructed to probe the filter, with their semantic score S, keyword match count k, keyword ratio k/kmax and binary MeSH value stated explicitly so that every value plotted below can be recomputed from Tscore = 0.50·S + 0.30·(k/kmax)^e + 0.20·M. Record C is the Top-Ranking Record: it matches every user keyword but has only average semantic similarity. (b) Tscore of records A, B and C as a continuous function of the exponent e (record D is omitted because its keyword ratio is zero and its score is constant at 0.400). Record C is flat by construction, since its ratio is 1.0 and is unaffected by the exponent; A and B decay because their matches are partial. In the shaded region (e < 1.19) the exponent is too weak and

record B, which satisfies only 8 of the 10 technical constraints, outranks record C; for $e < 1.07$, record A, which satisfies 7 of the 10, does so as well. (c) The same result expressed as rank position: record C reaches first place only for $e \geq 1.19$, and the FGTD default of 1.35 (dotted line) lies inside that region with a margin, while remaining far from the quadratic penalty that would suppress semantically related but technically incomplete records. Record D holds a stable fourth position throughout, showing that semantic relevance alone still produces a non-zero score when no keyword matches at all.



Supplementary Figure S2. Scoring pipeline of FGTD. The user query string and keyword list are encoded with all-MiniLM-L6-v2 into a 384-dimensional vector; each retrieved dataset is encoded from its title and GEO summary and indexed in an in-memory Qdrant collection. The cosine similarity between the two gives the semantic score S, which is combined with the keyword score K and the MeSH score M into the textual score T; T is combined with the bibliographic score B into the final relevance score. Equation numbers refer to Methods.

Supplementary Tables and Supplementary Data

The following items are provided as separate files.

Supplementary Table S1. Capability matrix of research discovery tools. Comparison of 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. File: Supplementary_Table_S1.xlsx.

Supplementary Data 1. Configuration of the breast cancer case study: query, organism, study type, attribute name, date range and supplementary file filters, keywords and MeSH terms. File: Supplementary_Data_1.json.

Supplementary Table S2. FGTD relevance scores for the breast cancer case study: K_score, S_score, M_score, R_score, C_score, T, B and final relevance score for the 15 retrieved series. File: Supplementary_Table_S2.csv.

Supplementary Table S3. FGTD metadata output for the breast cancer case study: structured metadata for the 15 series, including GEO summaries, platform details, sample identifiers, hyperlinks and binary keyword and MeSH indicators. File: Supplementary_Table_S3.xlsx. FGTD omits keyword columns that are zero for every series; "immunotherapy" matched none of the 15 series and has no column. The Vector column holds the 384-dimensional embedding.

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. for GSE124409 and GSE69986, with relevance scores and recorded times. File: Supplementary_Table_S5.xlsx. Sheet a: record of each series in tidy form (times in decimal minutes); sheet b: form as submitted.

Supplementary Table S6. Manual benchmark annotations of curator F.A. for GSE124409 and GSE69986, with relevance scores and recorded times. File: Supplementary_Table_S6.xlsx. Sheet a: record of each series in tidy form (times in decimal minutes); sheet b: form as submitted.

Supplementary Table S7. Manual benchmark annotations of curator L.S. for GSE124409 and GSE69986, with relevance scores and recorded times. File: Supplementary_Table_S7.xlsx. Sheet a: record of each series in tidy form (times in decimal minutes); sheet b: form as submitted.

Supplementary Table S8. Known-item retrieval results: rank of each of the five known items under the GEO order and under the FGTD relevance ranking (sheet a; ranks computed on unrounded scores, ties broken by GEO position), and Precision@5, Precision@10, NDCG@5, NDCG@10, reciprocal rank and median rank for both orderings (sheet b). File: Supplementary_Table_S8.xlsx.

Supplementary Table S9. FGTD relevance scores for the known-item retrieval evaluation: scoring breakdown for the 1,264 retrieved series, in the order returned by GEO. File: Supplementary_Table_S9.csv.

Supplementary Table S10. FGTD metadata output for the known-item retrieval evaluation: structured metadata for the 1,264 retrieved series, including linked PubMed records and binary keyword and MeSH indicators. File: Supplementary_Table_S10.xlsx.

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. File: Supplementary_Data_2.zip.

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 tau against the default ordering, top-k overlap, rank displacement and per-dataset ranks. File: Supplementary_Table_S12.xlsx.

Supplementary Table S13. Manual benchmark summary and concordance with the FGTD output: curation times and relevance scores in tidy form (sheet a), field-by-field concordance by series and by field (sheets b and c), errors against the GEO and PubMed records (sheet d) and the 180 comparisons (sheet e). 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 and hybrid-ranking analysis: scripts 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) (ZIP).