ENQUIRE RECONSTRUCTS AND EXPANDS GENE AND MESH CO-OCCURRENCE NETWORKS FROM CONTEXT-SPECIFIC LITERATURE

Luca Musella[1*], Max Widmann[1] and Julio Vera[1*]
[1]Laboratory of Systems Tumor Immunology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutsches Zentrum Immuntherapie, BZKF, and Uniklinikum Erlangen, Erlangen, Germany
*To whom correspondence should be addressed. Tel: +49 9131 85-45899; Email: luca.musella@uk-erlangen.de
Correspondence may also be addressed to Julio Vera. Tel: Tel: +49 9131 85-45876; Email: julio.vera-gonzalez@uk-erlangen.de

ABSTRACT

The accelerating growth in scientific literature is overwhelming our capacity to manually distil complex phenomena like molecular networks linked to diseases. Moreover, confirmation biases in search engines and databases influence the interpretation of facts and the generation of hypotheses. ENQUIRE (Expanding Networks by Querying Unexpectedly Inter-Related Entities) offers an alternative to manual literature curation and database mining to study complex biomedical phenomena. ENQUIRE generates a co-occurrence network of genes and biomedical ontologies (MeSH) using a corpus of publications as input. The algorithm iteratively reconstructs and expands the network by generating PubMed queries from significant interrelations, until it obtains an interconnected network of genes and MeSH relevant to the input corpus. We systematically evaluated ENQUIRE's versatility and found that it generates co-occurrence networks similar to those based on co-expression data and manually annotated databases with a fraction of the time and human resources. Using case studies spanning cancer, cell differentiation and immunity, ENQUIRE proved to identify interlinked genes and enriched pathways unique to each topic, thereby preserving their underlying diversity. ENQUIRE supports biomedical researchers by easing literature annotation, boosting hypothesis formulation and facilitating the identification of molecular targets for subsequent experimentation.

INTRODUCTION

Curated gene networks are of high interest to prime the analysis of biomedical omics data, identification of disease-specific regulatory modules, and therapy-oriented studies like drug repurposing (1–4). However, the reconstruction of such complex data takes time and manual effort. The growing biomedical literature corpus makes curation of biomolecular pathways and their function challenging to implement. Curating literature annotations requires domain expertise, but that same background knowledge could entail predispositions towards partial pictures of faceted biomedical problems (5). In contrast, information extraction from databases often omits the contextual information of gene interactions and can bias the results towards ubiquitously expressed, commonly investigated, and richly annotated genes (6–8). This makes systematic comparisons across case studies and experiments unattractive from an expenditure perspective.
In recent years, there have been significant investments in the automatic annotation of scientific corpora. The knowledgebase immuneXpresso, constructed by systematic text mining of the PubMed database, indexes interactions among immune cells and cytokines(9). SimText provides a framework to interactively explore the content of a user-provided corpus of literature (10). These and other tools rely on natural language processing methods like named-entity recognition (NER) (11, 12), part-of-speech recognition parsers (13), algorithms to assign directionality (14), relationship detection, and

scoring between entities (15, 16). These efforts in biomedical text mining aim at predicting meta-features and co-occurrences in literature corpora. However, assessing the significance of a text-mined interaction or the confidence level in dense, literature-based co-occurrence networks must be better addressed (17, 18). We find this striking, considering the well-documented reproducibility crisis (19–21).

In this context, we designed ENQUIRE (Expanding Networks by Querying Unexpectedly Inter-Related Entities) to achieve context-specific, reproducible reconstruction and expansion of biomedical co-occurrence networks from primary literature indexed in PubMed. ENQUIRE processes scientific articles by recognizing gene mentions in abstracts and extracting Medical Subject Headings (MeSH) to enrich gene-gene co-occurrences with contextual information. It applies state-of-the-art methods in statistical network analysis to reconstruct a network from an input corpus and generate relevant PubMed queries, to contextually expand the underlying corpus and, in turn, the network. A distinctive element in our methodology is the use of a statistical framework that accounts for literature biases. In this study, we present the tool, discuss the features of its underlying algorithm, and assess ENQUIRE's applicability and effectiveness with three concrete test cases.

MATERIALS AND METHODS

A Tool to Generate Co-Occurrence Networks from Literature

ENQUIRE (Expanding Networks by Querying Unexpectedly Inter-Related Entities) is an algorithm that reconstructs and expands co-occurrence networks of *Homo sapiens* genes and biomedical ontologies (MeSH), using a corpus of publication abstracts as input. The method iteratively annotates MeSH and gene mentions from abstracts, statistically assesses their importance, and generates network-informed PubMed queries, until it obtains a connected network of genes and MeSH terms (or meets another exit condition). ENQUIRE's pipeline implements a loop consisting of serial modules with the following structure (**Figure 1**):

a) The user supplies an input literature corpus in the form of at least three PubMed identifiers (PMIDs).

b) The algorithm indexes the MeSH terms associated to the PMIDs listed. Next, their abstracts are parsed, and named entity recognition (NER) of genes is performed using a lookup table of gene aliases and abstract-specific blocklists of ambiguous terms.

c) ENQUIRE annotates and weights co-occurrences between gene and MeSH entities, giving higher significance to co-occurrences observed more frequently than average.

d) The method selects significant co-occurrences and generates an undirected, simple graph, using a test statistic based on a random graph null model generated from unbiased mining of the literature corpus.

e) Next, the algorithm selects "information-dense" maximal cliques in the network, i.e. clusters of high-weight nodes all connected to each other, and reconstructs network communities from such nodes.

f) ENQUIRE identifies optimal sets of community-connecting graphlets via an approximate solution to the "travelling salesman problem" (TSP).

g) Finally, the algorithm feeds the entity nodes corresponding to the identified community-connecting graphlets into PubMed queries to find additional, relevant articles. Should ENQUIRE find new articles, their PMIDs are joined with the previous ones and automatically provided to module a), starting a new iteration.

ENQUIRE's goal is to generate a gene/MeSH network and its respective gene- and MeSH-only subgraphs that individually consist of a single, connected component. The loop prematurely terminates if: i) the network is empty after module d); ii) no clique can be selected in step e); iii) the clique network consists of only one community; iv) all generated queries return empty results. With default

parameters, ENQUIRE outputs node and edge lists of a gene/MeSH co-occurrence network and the respective gene- and MeSH-only subgraphs at each iteration. ENQUIRE outputs additional metrics, tabulated data, graphics, and links to collected resources for subsequent analyses and reproducibility. For instance, it is possible to extract subsets of the literature corpus that support a relation of interest and conveniently access the articles via hyperlinks redirecting to PubMed.

Description of the ENQUIRE algorithm

Recognition of MeSH and Gene Entities. ENQUIRE requires a corpus of PubMed identifiers (PMIDs) supplied by the user as input. For each PMID, the algorithm either collects gene mentions from the retrieved abstracts, the MeSH main headings ("descriptors") and subheadings ("qualifiers"), or both, depending on the entity scope set by the user. The retrieved MeSH terms are pruned to match biomedically relevant, non-redundant categories, by exploiting the tree-like, hierarchical structure of the MeSH vocabulary. As other authors have shown that the proportion of gene mentions does not significantly differ between abstracts and full-body texts (22), we only mine the abstracts for gene mentions. To perform named-entity recognition (NER) of genes in the abstracts, we designed a "Swiss cheese model", in which multiple methods complement each other to improve performance. Briefly, ENQUIRE tokenizes the abstract into single-word tokens and interprets tokens as the corresponding reference gene symbol if they match unambiguously in an alias lookup table. Multiple mentions of the same gene within an abstract count as one. We customized ENQUIRE's lookup table to collect as aliases all *Homo sapiens* and *Mus musculus* gene names and descriptions recorded in the databases HGNC, ENSEMBL, Uniprot, miRBase, and MGI (23–27). We made use of MGI's mouse/human orthology table to reassign mouse genes under human, reference gene symbols. In addition, ENQUIRE applies two algorithms to each unprocessed abstract: i) the scispaCy model *en_ner_jnlpba_md* to identify words classified as "CELL_LINE" or "CELL_TYPE" (28); ii) the Schwartz-Hearst algorithm to detect single-word abbreviations and their respective definitions (29). This allows ENQUIRE to construct abstract-specific blocklists that discard i) potentially ambiguous mentions to cell entities, and ii) ambiguous abbreviations whose definitions are not similar to any lookup table alias. See further details in **Supp. Mat**.

Annotation and Weighting of Co-Occurrences. ENQUIRE records the occurrences of MeSH and gene entities within each input article. Then, it counts pairwise co-occurrences by enumerating the subset of PMIDs associated to both entities in each pair. For each pair of entities g_i and g_j that co-occur in at least one article, we define the weights w and distances $\widetilde{w}$ accounting for the sheer co-occurrence $X(g_i, g_j)$ as follows:

$$w_{g_i,g_j} := \Psi\big(X(g_i, g_j), \bar{X}\big), \quad w_{g_i,g_j} \in (0,1]$$
$$\widetilde{w}_{g_i,g_j} = 1 - w_{g_i,g_j}$$
$$X(g_i, g_j) = \left|\{P \mid g_i, g_j \in E^P\}_{P \in \mathrm{PMIDS}}\right|$$

Where $\bar{X}$ is the mean co-occurrence between any two entities in the corpus, $\Psi(\cdot, \bar{X})$ is the zero-truncated, Poisson cumulative density function with a lambda of $\bar{X}$, and E^P is the set of all entities annotated within the PMID P that belongs to the submitted PMIDS corpus. This scoring system assigns higher relevance to co-occurrences that appear more often than average.

Reconstruction of a Weighted Network of Significant Co-Occurrences. ENQUIRE converts the recorded co-occurrences into an undirected multi-graph, where gene or MesH terms become nodes,

and each recorded co-occurrence between two entities becomes an edge. Thus, the network has as many nodes as the number of unique MeSH and gene symbols, with as many edges between two nodes as the number of PMIDs in which they co-occur. ENQUIRE implements the Casiraghi-Nanumyan's soft-configuration model applied to undirected, unweighted edge counts to select significant co-occurrences among entities, adjusted to 1% FDR (30). The test statistics follows a multivariate hypergeometric distribution, under the null hypothesis of observing a random graph whose expected degree sequence and number of edges correspond to the observed one. This allows us to condition the testing to the sheer, per-entity occurrence, which serves as a proxy for leveraging literature biases in the corpus. It is important to note that the null model does not assume independence of individual edges, but merely their equiprobability. This selection results in an undirected, single node-to-node edge co-occurrence graph (i.e. a simple graph). For each pair of adjacent entities g_i and g_j in the simple network, we assign the weights w_{g_i, g_j} and distances $\widetilde{w}_{g_i, g_j}$ to their mutual edge. Additionally, we prune poorly connected nodes by modularity-based, w-weighted Leiden clustering (31) and removal of singleton communities, that is, unconnected nodes. From the resulting gene/MeSH network, we also extract the respective gene- and MeSH-only subnetworks.
ENQUIRE-generated networks can consist of multiple connected components, i.e. subgraphs. To exclude unimportant components, a subgraph S is retained for subsequent computations only if the fraction of corpus articles covered by S is higher than a threshold value, as formally defined in

$$T_S := \frac{|\{P \mid E^P \cap E^S \neq \emptyset\}_{P \in \mathrm{PMIDS}}|}{|\mathrm{PMIDS}|} \geq t, \quad T_S \in (0,1]$$

where P denotes a PMID belonging to PMIDS, and E^P and E^S refer to the sets of gene or MeSH entities recorded in either P or S. T_S thus reflects the representativeness of S with respect to the entirety of the submitted corpus. The value of t can be set by the user. To avoid introducing irrelevant entities, ENQUIRE stops without further network expansion if the gene/MeSH network and the respective gene- and MeSH-only subnetworks individually contain only a single, connected component with $T_S \geq t$. We compute the weight of a node g in the connected graph S utilizing the composite function W, which is the product of normalized metrics for betweenness centrality (b) and w-weighted degree strength (d):

$$W(g,S) := F_b\big(b(g,S)\big) \cdot F_d\big(d(g,S)\big), \quad W \in (0,1]$$

Here, F_x denotes the empirical cumulative density function for the corresponding x parameter, calculated over S.

Construction of Communities from "Information-Dense" Cliques. To identify the most relevant parts of the gene/MeSH network, ENQUIRE first identifies the maximal cliques of order three or more. By definition, these are graphlets whose nodes are all adjacent to each other and not a subset of a larger clique. Applying the KNet function from the SANTA R package (32), we retained cliques that form significant clusters of associated entities with p-values adjusted to 1% FDR. This process allows us to consider the network topology by adjusting each maximal clique's significance in case many of such graphlets exist in the network. Through KNet's assessment of node weights (W) and distances $(\widetilde{w})$, ENQUIRE finds a pruned network C containing only cliques. At this point, ENQUIRE stops if the gene/MeSH network contains less than two significant cliques according to KNet. Next, ENQUIRE

identifies communities in the pruned network using modularity-based, w-weighted Leiden clustering. ENQUIRE stops if it detects a single community that encompasses all nodes in C.

Identification of Community-Connecting Entities. For any two disjoint communities C_i and C_j, we select the set of community-connecting, weighted graphlets $\Gamma_{C_i,C_j}(V_k, L_{k-1})$ satisfying the properties: i) all nodes g_i in the k-sized set V_k belong to either C_i or C_j; ii) the intersections between V_k and C_i or C_j are non-empty; iii) the w-weighted, $k-1$ edges L_{k-1} are sufficient to obtain a single connected component; iv) there is only one edge l_{g_i,g_j} that connects nodes belonging to distinct communities. Here, k is a parameter chosen by the user.
This allows us to rank the set of community-connecting entities $V_k$in any graphlet Γ_{C_i,C_j} by means of the distance metric R:

$$R\left(\Gamma_{C_i,C_j}(V_k, L_{k-1})\right) := -\log\left(\prod_{g_i \in V_k} W(g_i,\cdot) \prod_{l_{g_i,g_j} \in L_{k-1}} w_{g_i,g_j}\right), \quad R \in \mathbb{R}_{\geq 0}$$
$$V_k \in C_i \cup C_j, V_k \cap C_i \neq \emptyset, V_k \cap C_j \neq \emptyset$$
$$\left|\{l_{g_i,g_j} \mid g_i \in C_i, g_j \in C_j\}_{l_{g_i,g_j} \in L_{k-1}}\right| = 1$$

The smaller R, the closer two communities connected by V_k are.

Retrieval of New PMIDs via PubMed Queries Based on Optimal Connections. To evaluate which genes and MeSH terms are particularly suited for expansion querying, ENQUIRE constructs a multigraph M where network communities become nodes and all R-weighted connections between two communities become edges. R-weighted edges that do not fulfil the triangle inequality $R\left(\Gamma_{C_i,C_j}\right) \leq R\left(\Gamma_{C_i,C_z}\right) + R\left(\Gamma_{C_z,C_j}\right), \forall\, i,j,z$ are excluded. Then, we solve the travelling salesman problem (TSP) utilizing Christofides' approximate solution as implemented in the Python package Networkx (33). Via the visited edges, this yields an optimal path across communities and a corresponding collection of V_k entity sets. Each selected k-sized set results in a PubMed query formulated via the NCBI's *esearch* utility (34). We condition the search terms representing gene aliases and MeSH with "[Title/Abstract]" and "[MeSH Terms]", respectively, and exclude review articles from the results. The constructed PubMed queries require a match for all the k entities in the optimal path – e.g. *"melanoma/immunology"[MeSH Terms] AND ("IL1B"[Title/Abstract] OR "interleukin 1-beta"[Title/Abstract] […]) AND […]*. If all queries involving a subset of the network communities lead to empty results, we prune all previously used edges from M, compute a new TSP solution, and submit newly generated queries, provided at least one entity per query belongs to such community subset. This process is repeated A times, where A is a parameter specified by the user. If at least 1 new PMID matches any of the constructed queries, ENQUIRE starts a new analysis from the union of new and old PMIDs; otherwise, it stops. The rationale behind merging old and new PMIDs is to account for the original corpus when computing the statistics on new co-occurrences.

Network Expansion and User-Defined Parameters. Whenever ENQUIRE reconstructs a network from the union of old and new PMIDs, the previously reconstructed network is combined with the new one. The joined network has recomputed edge and node weights in accordance to the expanded literature corpus and connectivity. The rationale is to prioritize the original reconstruction, while also leveraging the expanded literature corpus. Users can tune five options to tailor the workflow, namely: 1) Restricting the target entities to annotate genes or MeSH only – default: both; 2) representativeness

threshold t to disregard subgraphs characterized by poor overlap with the literature corpus – default: 1% overlap; 3) query size k to control the number of entities that must be simultaneously used in a PubMed query – default: 4 entities; 4) query attempts A to choose the number of attempts at reconstructing an optimal path between network communities for querying – default: 2 attempts; and 5) connectivity criterion K to exclude newly found entities not holding sufficient number of connections to K previously existing communities – default: 2 communities. See **Supp. Figure 1** for a comprehensive description of the algorithm.

Post-hoc **Analyses**

Context-Aware Gene Sets. To reconstruct contextual gene sets using gene/MeSH co-occurrence networks, we adapt network-based relational data to the method described by Khan *et al*. (35). To this end, we first construct the inverse log-weighted similarity matrix between the gene/MeSH network nodes (36). This metric prioritizes nodes sharing many lower degree neighbors rather than few higher degree ones. We convert the obtained matrix into a Euclidean distance matrix by first applying Z-score standardization; then, we use the R package DynamicTreeCut and Ward's clustering to identify initial clusters and create an initial membership degree matrix (37, 38). Finally, we detect fuzzy clusters of genes and MeSH terms by applying Fuzzy C-means clustering to the distance matrix, using the R package ppclust (39, 40). The resulting membership degree matrix allows annotating genes with desired cluster membership degrees and extracting the linked MeSH terms to characterize the gene set.

Context-Aware Pathway Enrichment Analysis. We designed a method to map any text-mined co-occurrence network G into a mechanistic reference network N and infer context-specific enrichment of molecular pathways. With this strategy, we attempt to mechanistically explain the indirect relationships that constitute the co-occurrence network. To this end, we define the fitness score Q for every gene g in N with non-zero node degree d:

$$Q(g) := d(g,N)^{-1} \cdot \sum_{g_i \in V(G)} \sum_{g_j \in V(G)} e^{-\tilde{\delta}_G(g_i,g_j)} \cdot \mathbb{1}_{\{\delta_N(g_i,g)+\delta_N(g,g_j) \leq 2, g_i \neq g_j\}}, \quad Q \in \mathbb{R}_{\geq 0}$$

Here, $\tilde{\delta}_G(g_i, g_j)$ and $\delta_N(g_i, g_j)$ are the $\widetilde{w}$-weighted and unweighted distances from g_i to g_j in the graphs G and N, respectively. The indicator function $\mathbb{1}$ implies that non-text-mined genes without at least two text-mined nodes as neighbors have Q equal to zero. We normalize all scores to decorrelate Q from the node degree d, similarly to other approaches in network propagation (41, 42). As a mechanistic reference network, we chose STRING's (release 11.5) *H. sapiens* network of protein-coding, physically interacting genes (43). We exclusively combined the "experimental" and "database" channels to calculate STRING's confidence score, then pruned all edges with score below the 90th percentile and resulting isolated nodes, obtaining a reference network of 9482 nodes and 88333 edges. Then, we calculate Q scores for proteins in the reference STRING network using the ENQUIRE-generated gene network, and test for topology-based enrichment of gene sets using SANTA's KNet function. We used Neo4j 3.5 and the Reactome Graph database (44) to obtain gene sets from Reactome pathways. Significantly enriched pathways are the ones that are putatively functional in the ENQUIRE-generated, gene-gene co-occurrence network.

Benchmarks

Reconstruction of Reactome Gene Sets from Referenced Literature. We extracted annotated genes and reference literature for all *H. sapiens* Reactome pathways using Neo4J 3.5 and the Reactome Graph database. We employed NCBI's *esearch* and e*link* utilities to retrieve primary research articles cited by included review articles (34). For each pathway literature corpus, ENQUIRE performed one network reconstruction set to only extract gene mentions from article abstracts. As a benchmark, we processed the same abstract sets using PubTator Central's implementation of GnormPlus (11, 45). For the latter, we used NCBI's lookup tables of *H. sapiens* and *M. musculus* to assign gene aliases to a reference *H. sapiens* symbol. After excluding pathways with less than three literature references or only one human gene annotated in either ENQUIRE's or NCBI's lookup tables, we obtained a set of 967 evaluable pathways. We evaluated ENQUIRE and GNormPlus performances by quantifying false positive rate (FPR), precision and recall of the annotations in the 952 pathways with at least one gene recognized by both ENQUIRE and PubTator Central. We also evaluated the effect of statistics-based pruning of multigraphs for 720 pathways, whose literature corpora generated a significant, pruned network.

Estimate of Molecular Interrelations. We automatically generated a list of case studies by crossing leaf nodes downstream of *Diseases* and *Genetic Phenomena* (G05) MeSH categories (MeSH tree 2021). We then constructed a PubMed query from each pair by "AND" concatenation. Examples of such queries are *"Stomach Neoplasm"[MeSH Terms] AND "Chromosomes, human, pair 18"[MeSH Terms]*, and *"Acquired immunodeficiency syndrome"[MeSH Terms] AND "Polymorphism, single nucleotide"[MeSH Terms]*. For each query result with a size between 50 and 500 articles, we executed one network reconstruction. If obtaining a gene-gene co-occurrence network, we investigated whether its set of genes produced a network with more functional interactions than expected by chance. To obtain background distributions of edge counts for each gene set size observed with ENQUIRE, we sampled one million random gene sets and cumulated their interconnecting edges in STRING's v. 11.5 *H. sapiens* functional protein network. We only included functional associations from experiments, co-expression, and third-party databases with a cumulative score higher than 0.7 between proteins. The significance of each ENQUIRE-generated gene set's edge count was computed from the right-tailed probability of the empirical distribution.
Moreover, we compared the ENQUIRE-generated gene-gene wirings to STRING-derived associations using the DeltaCon similarity measure in a permutation test (46). To this end, we generated 10,000 random graphs for each observed ENQUIRE network. Each random graph was obtained through 300 random edge-swapping attempts while preserving the degree sequence of the original network. To obtain sensible probability densities, we focused on ENQUIRE-generated networks with degree sequences allowing at least ten different realizations of a graph. We followed the formula $\prod_i^n d_i!$, where d_i is the degree of the i-th node of a graph containing n nodes.

Assessment of Context Resolution by Topology-Based Enrichment of Molecular Pathways. To show that ENQUIRE preserves context-specific molecular signatures, we designed a broad panel of case studies (**Table 1**). Each corpus consisted of the union of references contained in three independent reviews accessible via NCBI's *elink* utility (34). We selected reviews from the results of PubMed search queries consisting of two or three MeSH terms (e.g. *"Melanoma"[MeSH Terms] AND "Signal Transduction"[MeSH Terms]*), favoring PubMed-ranked best matches when possible. We also included an unspecific positive control group consisting of the union of all context-specific corpora. This experimental design allowed us to construct a reference dendrogram that clusters the case studies only based on i) baseline biological knowledge and ii) expecting expanded networks of a case study to cluster together with the originally reconstructed one. Then, we applied ENQUIRE with default

parameters to each case study and analyzed all resulting gene-gene networks, i.e., from original and expanded corpora. We applied the *post hoc*, context-aware pathway enrichment analysis described above to all generated networks. We tested the enrichment of Reactome pathways with sizes ranging from 3 to 100 genes, categorized as in the database's *Top-Level Pathways* and disease ontologies. We performed hierarchical clustering of the networks using Euclidean distance and Kendall's correlation based on network-specific, KNet-generated p-values. We compared the resulting dendrogram to the expected one by a permutation test of Baker's gamma correlation using one million permutations of the original dendrogram (47). We also compared the results to two alternative statistics: i) over-representation analysis of nodes from the ENQUIRE-generated networks (the collection of all genes observed in any case study was used as the "universe"); ii) KNet statistics, using Q scores based on STRING's high-confidence functional association network (described above) and ENQUIRE-derived gene nodes.

RESULTS

An Exemplary ENQUIRE Run

To showcase ENQUIRE, we looked for literature-based relationships between the immune system and ferroptosis, a recently described form of programmed cell death. First, we selected 27 papers associated with ferroptosis and the immune system (PubMed query: *("Ferroptosis"[MeSH terms] AND "Immune System"[MeSH terms]) NOT "review"[Publication Type]"*, as of 14.04.23). We decreased the terms-per-query parameter c to 3, as we expected relatively few query-matching PMID. At the same time, we increased the control over the expansion and stopping criteria by setting the connectivity criterion k and representativeness threshold t to 3 and 0.2, respectively. ENQUIRE's initial reconstruction based on the 27 input papers yielded six gene/MeSH subgraphs isolated from each other with 47 genes, 208 MeSH and 1394 gene/MeSH edges. ENQUIRE retrieved 63 additional PMIDs by PubMed querying. The network obtained after the first expansion included three gene/MeSH subgraphs, with 68 genes, 303 MeSH and 2317 edges. ENQUIRE further retrieved 262 additional PMIDs. The algorithm's second expansion produced a single, connected co-occurrence gene/MeSH network with 195 genes, 809 MeSH and 10305 edges. The algorithm stopped after the second expansion because obtaining a single, connected gene/MeSH network is one of the stop criteria, and only one gene-gene network subgraph had a representativeness T bigger than 0.2. In total, for this case study the algorithm performed eleven PMID-matching queries. The benchmark on ENQUIRE's use of computational resources shows modest runtimes and RAM usage (**Supp. Figure 2**). The expansion process is depicted in **Figure 2A**, using the Cytoscape package DyNet (48, 49). Similar to the full gene/MeSH networks, the respective subgraphs containing only "information-dense" cliques show a gradual reduction in the number of disconnected components (**Figure 2B**). We performed context-specific gene set annotation utilizing the original gene/MeSH co-occurrence networks and our implementation of fuzzy c-means clustering, as described in Mat.Met. We identified non-trivial contextual gene sets (**Figure 2C**-left), including ferroptosis-dependent inflammation supported by immune-related adaptor proteins (blue, top left), antineoplastic effects of the ferroptosis-inducer sulfasalazine acting on the amino acid transport system (magenta), and cross-talk between ferroptosis and autophagy (pink), as described in (50–52). We also performed context-aware pathway enrichment analysis using the respectively generated gene-only subgraphs and the approach described in Methods. We summarized the results in **Figure 2C**-right, which depicts 60 Reactome pathways whose adjusted p-values were below Holm's 5% family-wise error rate (FWER) for at least one network, sorted by Reactome category. In the original network, we obtained enrichments of pathways centered around the beta-catenin destruction complex (e.g. *Degradation of beta-catenin by the destruction complex*, R-HSA-195253). In the expanded networks, the metabolic pathway *Glutathione conjugation* (R-HSA-156590) and innate immunity-related pathways linked to Toll-like receptor cascades (e.g. *MyD88 dependent cascade initiated on endosome*, R-HSA-975155) were enriched. Taken together, the ENQUIRE-generated outputs highlight a potential molecular axis between iron-regulated cell death and proliferation, metabolism, and immune response (53–55).

ENQUIRE Networks Support the Annotation of Gene Sets from Literature.

To evaluate ENQUIRE's named-entity recognition (NER) method for annotation of gene sets from literature, we selected all *H. sapiens* Reactome pathways, obtained the genes in the pathways and retrieved their corresponding primary literature references. We used the lists of references as inputs and performed a single gene entity-restricted NER and network reconstruction to gather genes for each pathway. To benchmark against existing methods, we analyzed the same literature corpus with PubTatorCentral's implementation of GNormPlus (see Mat.Met.). To guarantee that Reactome gene

names match ENQUIRE and GNormPlus reference symbols, we respectively employed ours and NCBI's lookup tables of reference gene symbols and respective aliases. In **Figure 3A**, we investigated how many genes in the Reactome pathways were correctly annotated by ENQUIRE's NER algorithm or GNormPlus. False positive rate (FPR), precision, and recall of ENQUIRE's NER appear comparable to those computed for GNormPlus, whose reported F1 score is 86.7% (45). ENQUIRE's FPR is 1.45 times that of GNormPlus (linear regression fit), although the maximum recorded value was below 10%. ENQUIRE'S NER shows, on average, 6% higher precision than GNormPlus, and the recall values of the two methods are comparable. Nevertheless, both algorithms show modest precision and recall performances in the Reactome-derived corpora. A possible explanation is that the abstracts from Reactome reference literature contain sparse or unspecific mentions to genes defining the pathways, which naturally lowers the ceiling for recall and precision. We also analyzed if ENQUIRE's performance in retrieving Reactome pathways' genes from their associated literature is conserved after network pruning, based on the test statistics for random graphs (**Figure 3B**). We then compared the non-isolated genes of the multigraphs to the remaining ones in the simple graph. **Figure 3B** shows that the test statistics has minimal effects on FPR and precision. However, we computed a 30% reduction in recall; a decrease is expected, as we set the FDR of the statistics to 1%, which lowers its power. We also evaluated the effect of input corpus size, pathway size and average entity co-occurrence per paper on the resulting pruned network (**Table 2**). Precision and recall show opposite Spearman's correlation trends concerning corpus and pathway sizes, which is expected, but average gene-gene co-occurrence per article seems uncorrelated. Taken together, these results show that ENQUIRE's NER algorithm is comparable to standard methods. In addition, ENQUIRE returns a statistical confidence for co-occurrences. The correlation between corpus size and FPR becomes relevant whenever multiple ENQUIRE searches are performed and compared. We suggest 1000 PMIDs as a reasonable upper boundary to keep FPR below 10% (**Figure 3A**). Finally, we explored if the ENQUIRE-computed weight W, an aggregated measure of a node network centrality and robustness of its connections, is a useful measure of gene relevance regarding the input corpus. To this end, we analyzed the above-mentioned gene-scope co-occurrence networks. In **Figure 3C**, we compare the pan-pathway-aggregated distributions of true-positive (top panel) and false-positive (middle panel) gene recognitions against the computed node weight (x-axis). We subdivided the distribution into four evenly spaced intervals, performed a chi-square test of independence, which resulted to be significant, and extracted the standardized Pearson residuals for true positives and false positives (colored boxes beneath the distributions). True positives tend to have higher node weights than false positives. An over-representation of node weights higher than 0.75 is observed in the true-positive distribution, as indicated by the color gradient in Pearson residuals. This suggests one can use the node weights W to rank a set of ENQUIRE-derived genes based on their relevance for the literature corpus in question. Altogether, the results in **Figure 3** indicate that ENQUIRE provides ways to support gene annotation by recovering and ranking genes associated with a literature corpus.

ENQUIRE Recovers Genes with High Chances of Showing Biochemical Interrelations.

To check if the networks generated by the algorithm are also enriched in molecular gene-gene interactions we queried PubMed with all possible pairs of a *Diseases* and a *Genetic Phenomena* MeSH term. We further processed the 3307 queries that retrieved 50-500 matching PMIDs and extracted their gene-gene co-occurrence networks obtained after one network reconstruction. We then inspected whether their protein-coding genes can produce significant functional association networks based on STRING's protein-protein interaction (PPI) database (see Methods). **Table 3** indicates that for 1336 (40%) MeSH pairs, both ENQUIRE and STRING generated a minimal network with at least three genes and two edges. In a subset of 733 suitable networks (see Mat.Met.), we tested ENQUIRE's

proficiency at recovering functional interactions by computing empirical probability distributions for STRING's edge count and DeltaCon similarity score. The gene sets from 730 networks (99%), respectively, produced a STRING network with a higher edge count than 95% of equal-sized random STRING networks. At the same time, 439 networks (60%) showed significant concordance with STRING-derived PPI networks based on DeltaCon. After p-value adjustment (1% FDR, **Table 3**), 722 (99%) and 344 (47%) ENQUIRE networks remained significantly representative of their expected physical interconnections, respectively. To evaluate the effect of network size, we subdivided the set of 733 suitable networks into quartiles based on their node number and computed the distributions of the unadjusted p-values obtained in the two tests above. The edge count-associated p-values increased with network size (**Figure 4A**). In contrast, we did not find a clear correlation between network size and p-values derived from DeltaCon similarities (**Figure 4B**). DeltaCon similarity quantiles, however, monotonically decreased with network size, possibly due to the growing differences in edge counts between ENQUIRE and STRING when the networks get larger (**Table 4**).
Taken together, our benchmark suggests that ENQUIRE generates networks that frequently relate to established, high-confidence functional relations.

ENQUIRE's networks improve the context resolution of topology-based pathway enrichment analyses.

We also analyzed the ability of ENQUIRE to generate and expand co-occurrence networks with distinctive biological and biomedical signatures by literature querying. To this end, we applied the complete ENQUIRE pipeline with default parameters to a comprehensive set of case studies, spanning cancer, cell differentiation, innate immunity and autoimmune diseases (see description of case studies in **Table 1**). We then performed topology-based pathway enrichment analysis (see Mat.Met.). **Figure 5A** depicts the expected dendrogram of the different case studies and respective expansions. **Figure 5B** shows the observed clustering of ENQUIRE-generated graphs inferred from ENQUIRE-informed, topology-based pathway enrichment analysis (see Post Hoc Analyses in Mat.Met.). The 50 pathways with at least one significant, adjusted p-value (5% FDR) and highest p-value variances across case studies are depicted. The heat map suggests that the case studies cluster primarily based on the affinities between their major topics, in agreement with the expected clustering (**Figure 5A**). For example, pathways categorized under *Diseases of Metabolism*, *Diseases of Immune System*, and *Innate Immune System* are predominantly enriched in networks originating from the case study "Macrophage's signal transduction during M. tuberculosis infection" (MP-ST) and the major topic "Antigen Presentation in Autoimmune Diseases". Similarly, some of *Chromatin Organization* and *Developmental Biology* pathways are almost exclusively enriched in the networks corresponding to oligodendrocyte differentiation. Also, the proposed cancer-related case studies consistently enrich a small cluster of oncology-related pathways like *Aberrant regulations of the mitotic cell cycle* (R-HSA-9687139 and its sub-pathways). Interestingly, a set of pathways linked to cell cycle like *Cyclin D associated events in G1* (R-HSA-69231) are enriched in the oligodendrocyte case study and reported to be also relevant in glioblastoma (56–59). We expected that a co-occurrence network derived from a highly heterogenous corpus, such as the positive control case study, should possess little context specificity. Its observed clustering position suggests that networks with large input corpora, expected to show higher false-positive rates, may result in a pathway enrichment biased towards a subset of Reactome's disease pathways associated with cancer (see also **Table 2**). All case studies appear constitutively enriched in a cluster of *Pathways in Cancer* annotated downstream of *Diseases of signal transduction by growth factor receptors and second messengers* (R-HSA-6802957). We hypothesize that this is a side-effect of how existing gene and pathway databases for *H. sapiens* have been curated, that is, consuming published experimental results biased towards oncology. Finally, we quantitatively assess the context resolution of the ENQUIRE-informed enrichment (**Figure 5C**). To this end, we performed a

permutation test on the observed Baker's gamma correlation value between dendrograms (**Figures 5A** and **5B**), which allows to statistically assess their similarity. We benchmarked its significance against two other methods, namely gene set over-representation analysis (ORA), and topology-based pathway enrichment analysis informed by STRING's high-confidence functional associations, instead of ENQUIRE-generated networks (see Mat.Met.). All methods generated a dendrogram significantly closer than expected to the reference. In our analysis, topology-based enrichments outperform ORA, with the ENQUIRE-informed score moderately improving the performance over the STRING-informed equivalent (0.69 and 0.64, respectively). Taken together, these results suggest that ENQUIRE-generated networks can accurately represent contextual, biological differences and similarities between case studies. While ENQUIRE-annotated genes are sufficient for context resolution, the use of topology-based methods that incorporate edge information improves the performance.

DISCUSSION

ENQUIRE is a novel computational framework for the automatic, quantitative, and reproducible annotation of scientific literature, offering an alternative to manual literature curation or database mining. ENQUIRE interrelates gene mentions and biomedical concepts through co-occurrence graphs and tabulated references. Its framework enables *post hoc* analyses that infer contextual gene sets and enriched molecular pathways. ENQUIRE can enhance the biological interpretation of omics data, suggest relevant processes and components for computational models, and motivate the selection of molecular targets for biological experiments and in scenarios like molecular tumor boards. This work introduces ENQUIRE and systematically evaluates its performance using real-word case scenarios and large test samples. Since ENQUIRE's scope goes beyond gene normalization and named entity recognition, we did not benchmark against standardized literature corpora like JNLPBA (60). Still, our evaluation shows that ENQUIRE's NER capabilities for genes are on par with state-of-the-art algorithms like GNormPlus. The combination of a lookup table with abstract-specific blocklists prioritizes precision with little-to-no costs on recall, making ENQUIRE's NER adaptive to sentence context. An added value of ENQUIRE is that the generated gene/MeSH co-occurrence network can prime further functional classification and data mining with reasonable computational costs and gold-standard-comparable performances. In line with this, our results indicate that ENQUIRE-generated gene networks correspond to high-confidence functional association networks in STRING. At the same time, ENQUIRE can also generate networks with previously unreported wirings that can complement database information and encourage novel explorative analyses (**Figure 4B**). We also analyzed the feasibility of corroborating ENQUIRE-suggested interactions by mapping co-occurrence information onto a mechanistic reference network. Since there is no generalizable method to project a network of indirect relations (e.g. co-occurrences) onto a mechanistic network (61–65), we designed a function to score a molecular interaction network using ENQUIRE-generated networks. This allowed us to verify that pathway enrichment analysis from original and expanded ENQUIRE networks preserves context specificity when comparing multiple case studies. We designed ENQUIRE as a series of modular, open-source components that the user can freely combine and expand to tune its performance. For instance, one could insert a part-of-speech recognition parser upstream of the co-occurrence annotation step to strengthen this criterion (66). Similarly, one can implement a propensity matrix into the random graph model to further weight a co-occurrence with its textual context (15, 30). As NER and, in particular, gene normalization rely heavily on the utilized lookup table of reference gene symbols and aliases, ENQUIRE's accuracy depends on how comprehensive and free of ambiguities this table is. The current version of our algorithm only performs NER of human genes and corresponding mouse orthologs. Still, it can be adapted to perform NER of any other species by supplying an appropriate lookup table, such as those provided by the STRING database (43). Our main objective was to construct a reproducible network reconstruction pipeline accessible to bioinformaticians and systems biologists with affordable computational requirements (**Supp. Figure 2**). Since the standalone version of the algorithm requires some background in computer programming, we are working to provide a web version of ENQUIRE to ease its adoption among biomedical researchers.

DATA AVAILABILITY

ENQUIRE's main program and the standalone R scripts to perform the *post hoc* analyses are included in an Apptainer/Singularity image file (SIF), available for download at https://jveralab.net/ENQUIRE.sif. The gene-symbol-to-alias lookup table, input and output files from the exemplary case study, and data underlying the results shown in **Figures 3-5** can be found at https://github.com/Muszeb/ENQUIRE. All the individual scripts are also available upon request.

AUTHOR CONTRIBUTIONS

Initial idea and concept: LM and JV. Coding and benchmarking of the algorithm: LM and MW. Drafting of the manuscript: LM and JV.

ACKNOWLEDGEMENTS

We thank Martin Eberhardt, Christopher Lischer, Jimmy Retzlaff, Esther Güse, Xin Lai, and Suryadipto Sarkar for the useful scientific discussions, comments on the manuscript, and testing the installation and running of the algorithm. We also want to thank Jacopo Innocenti for lending his arts talent to the graphical abstract.

FUNDING

This work has been supported by the German Ministry of Education and Science (BMBF) thorough the projects e:Med MelAutim and KI-VesD I and II.

REFERENCES

1. Vitali,F., Cohen,L.D., Demartini,A., Amato,A., Eterno,V., Zambelli,A. and Bellazzi,R. (2016) A network-based data integration approach to support drug repurposing and multi-Target therapies in triple negative breast cancer. *PLoS ONE*, **11**.

2. Cantone,M., Küspert,M., Reiprich,S., Lai,X., Eberhardt,M., Göttle,P., Beyer,F., Azim,K., Küry,P., Wegner,M., *et al.* (2019) A gene regulatory architecture that controls region-independent dynamics of oligodendrocyte differentiation. *Glia*, **67**, 825–843.

3. Sadegh,S., Skelton,J., Anastasi,E., Bernett,J., Blumenthal,D.B., Galindez,G., Salgado-Albarrán,M., Lazareva,O., Flanagan,K., Cockell,S., *et al.* (2021) Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nature Communications*, **12**, 6848.

4. Lai,X., Zhou,J., Wessely,A., Heppt,M., Maier,A., Berking,C., Vera,J. and Zhang,L. (2022) A disease network-based deep learning approach for characterizing melanoma. *International journal of cancer*, **150**, 1029–1044.

5. Grimes,D.R. and Heathers,J. (2021) The new normal? Redaction bias in biomedical science. *Royal Society Open Science*, **8**, 211308.

6. Haynes,W.A., Tomczak,A. and Khatri,P. (2018) Gene annotation bias impedes biomedical research. *Scientific Reports*, **8**, 1362.

7. Tomczak,A., Mortensen,J.M., Winnenburg,R., Liu,C., Alessi,D.T., Swamy,V., Vallania,F., Lofgren,S., Haynes,W., Shah,N.H., *et al.* (2018) Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Scientific Reports*, **8**, 5115.

8. Ewing,E., Planell-Picola,N., Jagodic,M. and Gomez-Cabrero,D. (2020) GeneSetCluster: A tool for summarizing and integrating gene-set analysis results. *BMC Bioinformatics*, **21**.

9. Kveler,K., Starosvetsky,E., Ziv-Kenet,A., Kalugny,Y., Gorelik,Y., Shalev-Malul,G., Aizenbud-Reshef,N., Dubovik,T., Briller,M., Campbell,J., *et al.* (2018) Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nature Biotechnology*, **36**, 651–659.

10. Macnee,M., Pérez-Palma,E., Schumacher-Bass,S., Dalton,J., Leu,C., Blankenberg,D. and Lal,D. (2021) SimText: A text mining framework for interactive analysis and visualization of similarities among biomedical entities. *Bioinformatics (Oxford, England)*, **37**, 4285–7.

11. Wei,C.-H., Allot,A., Leaman,R. and Lu,Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, **47**, W587–W593.

12. Luo,L., Wei,C.-H., Lai,P.-T., Leaman,R., Chen,Q. and Lu,Z. (2023) AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, **39**, btad310.

13. Chen,D. and Manning,C. (2014) A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 740–750.

14. Miao,Q., Zhang,S., Meng,Y., Fu,Y. and Yu,H. (2012) Healthy or Harmful? Polarity Analysis Applied to Biomedical Entity Relationships. In.pp. 777–782.

15. Junge,A. and Jensen,L.J. (2020) CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision. *Bioinformatics*, **36**, 264–271.

16. Lee,J., Yoon,W., Kim,S., Kim,D., Kim,S., So,C.H. and Kang,J. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**, 1234–1240.

17. Islam,M.R., Liu,S., Wang,X. and Xu,G. (2020) Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social network analysis and mining*, **10**, 82.

18. Diaz-Garcia,J.A., Fernandez-Basso,C., Ruiz,M.D. and Martin-Bautista,M.J. (2020) Mining Text Patterns over Fake and Real Tweets. In Lesot,M.-J., Vieira,S., Reformat,M.Z., Carvalho,J.P., Wilbik,A., Bouchon-Meunier,B., Yager,R.R. (eds), *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer International Publishing, Cham, pp. 648–660.

19. Ioannidis,J.P.A. (2005) Why Most Published Research Findings Are False. *PLoS Medicine*, **2**, e124.

20. Baker,M. (2016) 1,500 scientists lift the lid on reproducibility. *Nature*, **533**, 452–454.

21. HAYNES,W.A., VALLANIA,F., LIU,C., BONGEN,E., TOMCZAK,A., ANDRES-TERRÈ,M., LOFGREN,S., TAM,A., DEISSEROTH,C.A., LI,M.D., *et al.* (2017) EMPOWERING MULTI-COHORT GENE EXPRESSION ANALYSIS TO INCREASE REPRODUCIBILITY. In *Biocomputing 2017*. WORLD SCIENTIFIC, pp. 144–153.

22. Cohen,K.B., Johnson,H.L., Verspoor,K., Roeder,C. and Hunter,L.E. (2010) The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**, 492.

23. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Research*, **47**, D155–D162.

24. Ringwald,M., Richardson,J.E., Baldarelli,R.M., Blake,J.A., Kadin,J.A., Smith,C. and Bult,C.J. (2022) Mouse Genome Informatics (MGI): latest news from MGD and GXD. *Mammalian Genome*, **33**, 4–18.

25. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Research*, **51**, D933–D941.

26. Seal,R.L., Braschi,B., Gray,K., Jones,T.E.M., Tweedie,S., Haim-Vilmovsky,L. and Bruford,E.A. (2023) Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, **51**, D1003–D1009.

27. Consortium,T.U. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, **51**, D523–D531.

28. Neumann,M., King,D., Beltagy,I. and Ammar,W. (2019) ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 319–327.

29. Schwartz,A.S. and Hearst,M.A. (2002) A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.

30. Casiraghi,G. and Nanumyan,V. (2021) Configuration models as an urn problem. *Scientific reports*, **11**, 13416.

31. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, **9**, 5233.

32. Cornish,A.J. and Markowetz,F. (2014) SANTA: Quantifying the Functional Content of Molecular Networks. *PLOS Computational Biology*, **10**, e1003808-.

33. Hagberg,A., Swart,P.J. and Schult,D.A. (2008) Exploring network structure, dynamics, and function using NetworkX. In. United States.

34. Kans,J. (2013) Entrez Direct: E-utilities on the Unix Command Line Bethesda (MD).

35. Khan,A., Katanic,D. and Thakar,J. (2017) Meta-analysis of cell- specific transcriptomic data using fuzzy c-means clustering discovers versatile viral responsive genes. *BMC Bioinformatics*, **18**, 295.

36. Adamic,L.A. and Adar,E. (2003) Friends and neighbors on the Web. *Social Networks*, **25**, 211–230.

37. Ward,J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, **58**, 236–244.

38. Langfelder,P., Zhang,B. and Horvath,S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.

39. Dunn,J.C. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, **3**, 32–57.

40. Cebeci,Z. (2019) Comparison of Internal Validity Indices for Fuzzy Clustering. *JAI*, **10**.

41. Bello,T., Chan,M., Golkowski,M., Xue,A.G., Khasnavis,N., Ceribelli,M., Ong,S.-E., Thomas,C.J. and Gujral,T.S. (2021) KiRNet: Kinase-centered network propagation of pharmacological screen results. *Cell Reports Methods*, **1**, 100007.

42. Crowl,S., Jordan,B.T., Ahmed,H., Ma,C.X. and Naegle,K.M. (2022) KSTAR: An algorithm to predict patient-specific kinase activities from phosphoproteomic data. *Nature Communications*, **13**, 4283.

43. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, **49**, D605–D612.

44. Fabregat,A., Korninger,F., Viteri,G., Sidiropoulos,K., Marin-Garcia,P., Ping,P., Wu,G., Stein,L., D'Eustachio,P. and Hermjakob,H. (2018) Reactome graph database: Efficient access to complex pathway data. *PLOS Computational Biology*, **14**, e1005968-.

45. Wei,C.-H., Kao,H.-Y. and Lu,Z. (2015) GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Research International*, **2015**, 918710.

46. Koutra,D., Shah,N., Vogelstein,J.T., Gallagher,B. and Faloutsos,C. (2016) DeltaCon: Principled Massive-Graph Similarity Function with Attribution. *ACM Trans. Knowl. Discov. Data*, **10**.

47. Baker,F.B. (1974) Stability of Two Hierarchical Grouping Techniques Case 1: Sensitivity to Data Errors. *Journal of the American Statistical Association*, **69**, 440–445.

48. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**, 2498–2504.

49. Goenawan,I.H., Bryan,K. and Lynn,D.J. (2016) DyNet: visualization and analysis of dynamic molecular interaction networks. *Bioinformatics*, **32**, 2713–2715.

50. Li,W., Feng,G., Gauthier,J.M., Lokshina,I., Higashikubo,R., Evans,S., Liu,X., Hassan,A., Tanaka,S., Cicka,M., *et al.* (2019) Ferroptotic cell death and TLR4/Trif signaling initiate neutrophil recruitment after heart transplantation. *The Journal of Clinical Investigation*, **129**, 2293–2304.

51. Lang,X., Green,M.D., Wang,W., Yu,J., Choi,J.E., Jiang,L., Liao,P., Zhou,J., Zhang,Q., Dow,A., *et al.* (2019) Radiotherapy and Immunotherapy Promote Tumoral Lipid Oxidation and Ferroptosis via Synergistic Repression of SLC7A11. *Cancer Discovery*, **9**, 1673–1685.

52. Tang,X., Ding,H., Liang,M., Chen,X., Yan,Y., Wan,N., Chen,Q., Zhang,J. and Cao,J. (2021) Curcumin induces ferroptosis in non-small-cell lung cancer via activating autophagy. *Thoracic Cancer*, **12**, 1219–1230.

53. Quagliariello,V., De Laurentiis,M., Rea,D., Barbieri,A., Monti,M.G., Carbone,A., Paccone,A., Altucci,L., Conte,M., Canale,M.L., *et al.* (2021) The SGLT-2 inhibitor empagliflozin improves myocardial strain, reduces cardiac fibrosis and pro-inflammatory cytokines in non-diabetic mice treated with doxorubicin. *Cardiovascular Diabetology*, **20**, 150.

54. Li,Y., Xu,X., Wang,X., Zhang,C., Hu,A. and Li,Y. (2023) MGST1 Expression Is Associated with Poor Prognosis, Enhancing the Wnt/β-Catenin Pathway via Regulating AKT and Inhibiting Ferroptosis in Gastric Cancer. *ACS Omega*, **8**, 23683–23694.

55. Nakamura,T., Hipp,C., Santos Dias Mourão,A., Borggräfe,J., Aldrovandi,M., Henkelmann,B., Wanninger,J., Mishima,E., Lytton,E., Emler,D., *et al.* (2023) Phase separation of FSP1 promotes ferroptosis. *Nature*, **619**, 371–377.

56. Tikoo,R., Osterhout,D.J., Casaccia-Bonnefil,P., Seth,P., Koff,A. and Chao,M.V. (1998) Ectopic expression of p27Kip1 in oligodendrocyte progenitor cells results in cell-cycle growth arrest. *Journal of Neurobiology*, **36**, 431–440.

57. Nygård,M., Wahlström,G.M., Gustafsson,M.V., Tokumoto,Y.M. and Bondesson,M. (2003) Hormone-Dependent Repression of the E2F-1 Gene by Thyroid Hormone Receptors. *Molecular Endocrinology*, **17**, 79–92.

58. Magri,L., Swiss,V.A., Jablonska,B., Lei,L., Pedre,X., Walsh,M., Zhang,W., Gallo,V., Canoll,P. and Casaccia,P. (2014) E2F1 Coregulates Cell Cycle Genes and Chromatin Components during the Transition of Oligodendrocyte Progenitors from Proliferation to Differentiation. *The Journal of Neuroscience*, **34**, 1481.

59. Jaiswal,M., Tripathi,A., Singh,D., Kumar,A., Singh,M., Batra,N. and Verma,A. (2022) Clinical Correlation and Role of Cyclin D1 Expression in Glioblastoma Patients: A Study From North India. *Cureus*, **14**, e22346–e22346.

60. Huang,M.-S., Lai,P.-T., Lin,P.-Y., You,Y.-T., Tsai,R.T.-H. and Hsu,W.-L. (2020) Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*, **21**, 2219–2238.

61. Marra,M., Emrouznejad,A., Ho,W. and Edwards,J.S. (2015) The value of indirect ties in citation networks: SNA analysis with OWA operator weights. *Information Sciences*, **314**, 135–151.

62. Han,X., Shen,Z., Wang,W.-X., Lai,Y.-C. and Grebogi,C. (2016) Reconstructing direct and indirect interactions in networked public goods game. *Scientific Reports*, **6**, 30241.

63. Mei,S., Flemington,E.K. and Zhang,K. (2017) A computational framework for distinguishing direct versus indirect interactions in human functional protein–protein interaction networks. *Integrative Biology*, **9**, 595–606.

64. Hawe,J.S., Theis,F.J. and Heinig,M. (2019) Inferring Interaction Networks From Multi-Omics Data. *Frontiers in Genetics*, **10**.

65. Xiao,N., Zhou,A., Kempher,M.L., Zhou,B.Y., Shi,Z.J., Yuan,M., Guo,X., Wu,L., Ning,D., Van Nostrand,J., *et al.* (2022) Disentangling direct from indirect relationships in association networks. *Proceedings of the National Academy of Sciences of the United States of America*, **119**, e2109995119.

66. Nguyen,D.Q. and Verspoor,K. (2019) From POS tagging to dependency parsing for biomedical event extraction. *BMC Bioinformatics*, **20**, 72.

67. Sarkar,D., Leung,E.Y., Baguley,B.C., Finlay,G.J. and Askarian-Amiri,M.E. (2015) Epigenetic regulation in human melanoma: past and future. *Epigenetics*, **10**, 103–121.

68. Czarnecka,A.M., Bartnik,E., Fiedorowicz,M. and Rutkowski,P. (2020) Targeted Therapy in Melanoma and Mechanisms of Resistance. *International journal of molecular sciences*, **21**, 4576.

69. Guo,W., Wang,H. and Li,C. (2021) Signal pathways of melanoma and targeted therapy. *Signal transduction and targeted therapy*, **6**, 424.

70. Kaur,J., Malik,M.A., Gulati,R., Azad,S.V. and Goswami,S. (2014) Genetic determinants of uveal melanoma. *Tumor Biology*, **35**, 11711–11717.

71. Luke,J.J., Triozzi,P.L., McKenna,K.C., Van Meir,E.G., Gershenwald,J.E., Bastian,B.C., Gutkind,J.S., Bowcock,A.M., Streicher,H.Z., Patel,P.M., *et al.* (2015) Biology of advanced uveal melanoma and next steps for clinical therapeutics. *Pigment cell & melanoma research*, **28**, 135–147.

72. Chua,V., Lapadula,D., Randolph,C., Benovic,J.L., Wedegaertner,P.B. and Aplin,A.E. (2017) Dysregulated GPCR Signaling and Therapeutic Options in Uveal Melanoma. *Molecular cancer research : MCR*, **15**, 501–506.

73. Cheng,W.-L., Feng,P.-H., Lee,K.-Y., Chen,K.-Y., Sun,W.-L., Van Hiep,N., Luo,C.-S. and Wu,S.-M. (2021) The Role of EREG/EGFR Pathway in Tumor Progression. *International journal of molecular sciences*, **22**, 12828.

74. Zhu,G., Pei,L., Xia,H., Tang,Q. and Bi,F. (2021) Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. *Molecular cancer*, **20**, 143.

75. Zhao,H., Ming,T., Tang,S., Ren,S., Yang,H., Liu,M., Tao,Q. and Xu,H. (2022) Wnt signaling in colorectal cancer: pathogenic role and therapeutic target. *Molecular Cancer*, **21**, 144.

76. Butti,R., Das,S., Gunasekaran,V.P., Yadav,A.S., Kumar,D. and Kundu,G.C. (2018) Receptor tyrosine kinases (RTKs) in breast cancer: signaling, therapeutic implications and challenges. *Molecular Cancer*, **17**, 34.

77. Hinz,N. and Jücker,M. (2019) Distinct functions of AKT isoforms in breast cancer: a comprehensive review. *Cell Communication and Signaling*, **17**, 154.

78. Medina,M.A., Oza,G., Sharma,A., Arriaga,L.G., Hernández Hernández,J.M., Rotello,V.M. and Ramirez,J.T. (2020) Triple-Negative Breast Cancer: A Review of Conventional and Advanced Therapeutic Strategies. *International journal of environmental research and public health*, **17**, 2078.

79. Abreu,R., Giri,P. and Quinn,F. (2020) Host-Pathogen Interaction as a Novel Target for Host-Directed Therapies in Tuberculosis. *Frontiers in immunology*, **11**, 1553.

80. Patrick,K.L. and Watson,R.O. (2021) Mitochondria: Powering the Innate Immune Response to Mycobacterium tuberculosis Infection. *Infect Immun*, **89**, e00687-20.

81. Mvubu,N.E. and Chiliza,T.E. (2021) Exploring the Use of Medicinal Plants and Their Bioactive Derivatives as Alveolar NLRP3 Inflammasome Regulators during Mycobacterium tuberculosis Infection. *International journal of molecular sciences*, **22**, 9497.

82. Pan,Y., Yu,Y., Wang,X. and Zhang,T. (2020) Tumor-Associated Macrophages in Tumor Immunity. *Frontiers in immunology*, **11**, 583084.

83. Gao,J., Liang,Y. and Wang,L. (2022) Shaping Polarization Of Tumor-Associated Macrophages In Cancer Immunotherapy. *Frontiers in immunology*, **13**, 888713.

84. Luo,S., Yang,G., Ye,P., Cao,N., Chi,X., Yang,W.-H. and Yan,X. (2022) Macrophages Are a Double-Edged Sword: Molecular Crosstalk between Tumor-Associated Macrophages and Cancer Stem Cells. *Biomolecules*, **12**, 850.

85. Ihara,S., Hirata,Y. and Koike,K. (2017) TGF-β in inflammatory bowel disease: a key regulator of immune cells, epithelium, and the intestinal microbiota. *Journal of Gastroenterology*, **52**, 777–787.

86. Hou,Q., Huang,J., Ayansola,H., Masatoshi,H. and Zhang,B. (2021) Intestinal Stem Cells and Immune Cell Relationships: Potential Therapeutic Targets for Inflammatory Bowel Diseases. *Frontiers in immunology*, **11**, 623691.

87. Yang,Z.-J., Wang,B.-Y., Wang,T.-T., Wang,F.-F., Guo,Y.-X., Hua,R.-X., Shang,H.-W., Lu,X. and Xu,J.-D. (2021) Functions of Dendritic Cells and Its Association with Intestinal Diseases. *Cells*, **10**, 583.

88. Fearon,U., Canavan,M., Biniecka,M. and Veale,D.J. (2016) Hypoxia, mitochondrial dysfunction and synovial invasiveness in rheumatoid arthritis. *Nature Reviews Rheumatology*, **12**, 385–397.

89. Cheung,T.T. and McInnes,I.B. (2017) Future therapeutic targets in rheumatoid arthritis? *Seminars in immunopathology*, **39**, 487–500.

90. Wehr,P., Purvis,H., Law,S.-C. and Thomas,R. (2019) Dendritic cells, T cells and their interaction in rheumatoid arthritis. *Clinical and Experimental Immunology*, **196**, 12–27.

91. Harden,J.L., Krueger,J.G. and Bowcock,A.M. (2015) The immunogenetics of Psoriasis: A comprehensive review. *Journal of autoimmunity*, **64**, 66–73.

92. Chiricozzi,A., Romanelli,P., Volpe,E., Borsellino,G. and Romanelli,M. (2018) Scanning the Immunopathogenesis of Psoriasis. *International journal of molecular sciences*, **19**, 179.

93. Tokuyama,M. and Mabuchi,T. (2020) New Treatment Addressing the Pathogenesis of Psoriasis. *International journal of molecular sciences*, **21**, 7488.

94. Svaren,J. (2014) MicroRNA and transcriptional crosstalk in myelinating glia. *Neurochemistry international*, **77**, 50–57.

95. Elbaz,B. and Popko,B. (2019) Molecular Control of Oligodendrocyte Development. *Trends in neurosciences*, **42**, 263–277.

96. Tiane,A., Schepers,M., Rombaut,B., Hupperts,R., Prickaerts,J., Hellings,N., van den Hove,D. and Vanmierlo,T. (2019) From OPC to Oligodendrocyte: An Epigenetic Journey. *Cells*, **8**, 1236.

TABLES

Table 1. Selection of case studies for assessment of context resolution at the molecular pathway level. We obtained PubMed queries by "AND" concatenation of up to three MeSH terms and further filters to retrieve review articles only. The Corpus sizes refer to the non-redundant union of publications cited by three independent review articles, reported under the "References" column.

Major Topic	Case Study (abbreviation)	PubMed Query			Corpus size	References
		MeSH 1	MeSH 2	MeSH 3		
Signal transduction in solid tumors	**Melanoma (MM-ST)**	Signal transduction	Melanoma		944	(67–69)
	Uveal melanoma (UM-ST)	Signal transduction	Uveal neoplasms		218	(70–72)
	Colorectal cancer (COL-ST)	Signal transduction	Colorectal neoplasms		556	(73–75)
	Breast cancer (BRE-ST)	Signal transduction	Breast neoplasms		522	(76–78)
Macrophage's signal transduction in disease	**Macrophage signal transduction upon infection (MP-ST)**	Signal transduction	Macrophages	Mycobacterium tuberculosis	470	(79–81)
	Tumor-associated Macrophages (MP-TA)	Signal transduction	Tumor associated macrophages		386	(82–84)
Antigen presentation in autoimmune diseases	**Inflammatory bowel disease (IBD-AP)**	Antigen presentation	Inflammatory bowel diseases		445	(85–87)
	Rheumatoid arthritis (RA-AP)	Antigen presentation	Arthritis, rheumatoid		452	(88–90)
	Psoriasis (PSO-AP)	Antigen presentation	Psoriasis		435	(91–93)
Oligodendrocyte differentiation	**Oligodendrocyte (ODC)**	Cell differentiation	Oligodendroglia		355	(94–96)
Positive control	**All case studies**	All queries ("OR" concatenation)			3606	All of the above

Table 2. Effect of relevant covariates on quality indicators of ENQUIRE's gene entity recognition. We evaluated the effect of corpus size (input), Reactome's pathway size (number of genes to be retrieved) and average gene-gene co-occurrence per article, using Spearman's correlation, for each measure. FPR: false positive rate.

Metric	Corpus Size	Pathway Size	Average co-occurrence
FPR	0.83	0.25	0.37
Precision	-0.18	0.49	-0.06
Recall	0.46	-0.35	0.14

Table 3. Relevant quality indicators of ENQUIRE performance in a cohort of 3307 case studies. PPI: protein-protein interaction score, as number of observed edges over the STRING-inferred network. FDR: false discovery rate, expressed in percentage, calculated using Benjamini-Hochberg method**.** Percentages reported for PPI and DeltaCon significance independently refer to the set of 733 tested networks, i.e. those with 10 or more possible realizations with the same degree sequence as ENQUIRE-derived networks.

Property	Subset		Raw count	Percentage over the preceding step	Percentage over total (3307)
Network topology	**At least 3 genes and 2 edges in both ENQUIRE and STRING networks**		1336	90%	40%
	At least 10 possible realizations of the same degree sequence		733	55%	22%
Significance	**Edge count p-value**	**< 0.05**	730	99%	22%
		< 1% FDR	722	99%	21%
	DeltaCon p-value	**< 0.05**	439	60%	13%
		< 1% FDR	344	47%	10%

Table 4. Quantiles of DeltaCon similarities, ENQUIRE- and STRING-based edges counts, sorted by number of genes in the network. Median values with respect to each metric and range of gene counts are highlighted in bold.

Metric	Range of gene counts	Quantiles				
		0%	**25%**	**50%**	**75%**	**100%**
DeltaCon	4-9	0.75	0.83	**0.87**	0.94	1.00
	10-14	0.67	0.78	**0.81**	0.83	1.00
	15-23	0.65	0.74	**0.77**	0.79	0.87
	24-119	0.56	0.65	**0.69**	0.72	0.81
Edge count - ENQUIRE	4-9	4	6	**7**	8	16
	10-14	6	8	**10**	13	43
	15-23	8	13	**17**	22	66
	24-119	18	36	**49**	77	295
Edge count - STRING	4-9	4	6	**8**	10	23
	10-14	6	11	**15**	20	50
	15-23	10	21	**28**	37	94
	24-119	19	54	**89**	146	591

FIGURES

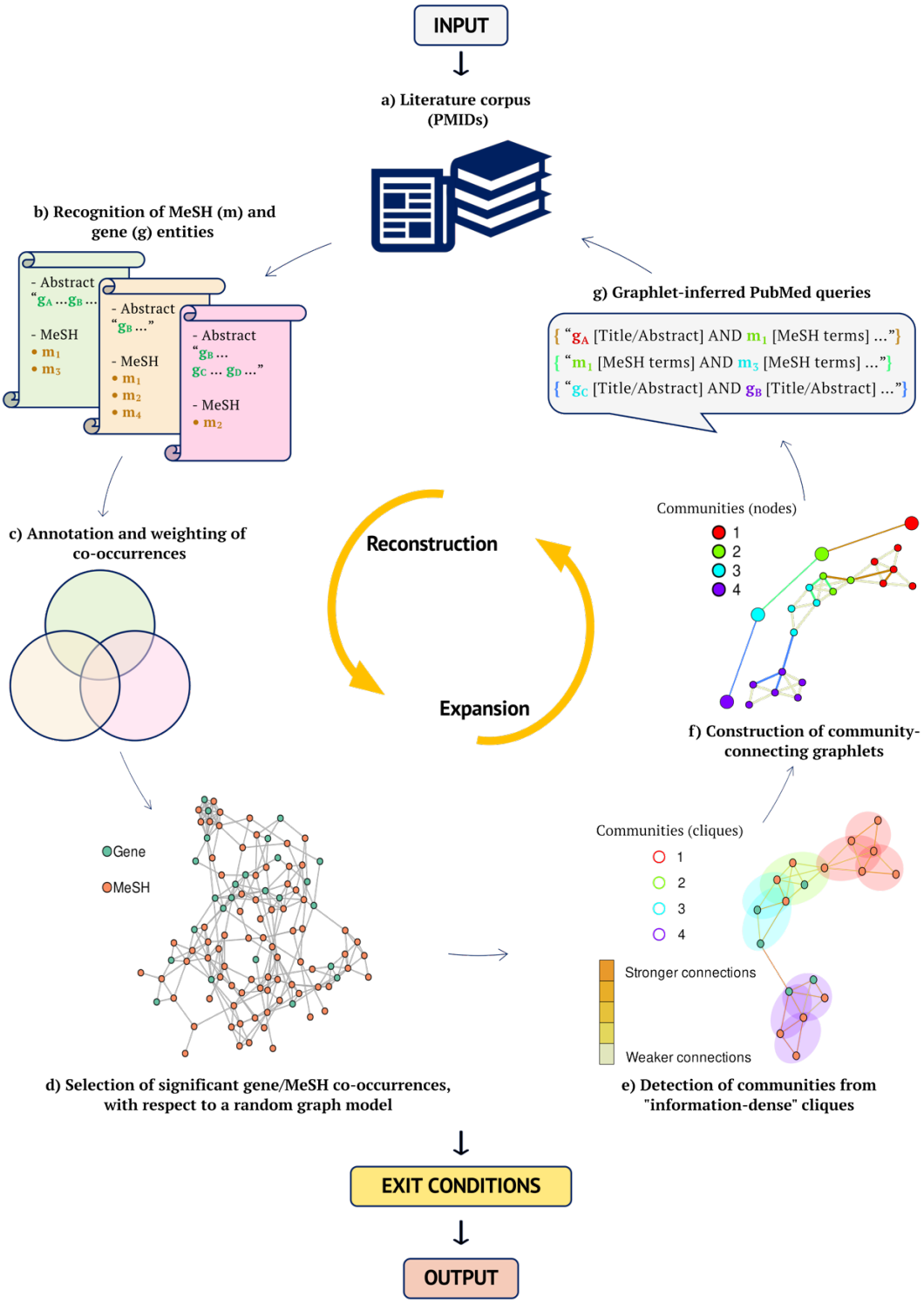

Figure 1. Overview of ENQUIRE methodology. ENQUIRE accepts a set of PubMed identifiers as input, together with optional, user-specified parameters. The pipeline iteratively orchestrates reconstruction and expansion of literature-derived co-occurrence networks, until an exit condition is fulfilled. Additional information about each alphabetically indexed module and output is provided in the Mat.Met. section. For a more detailed flowchart, see **Supp. Figure 1**.

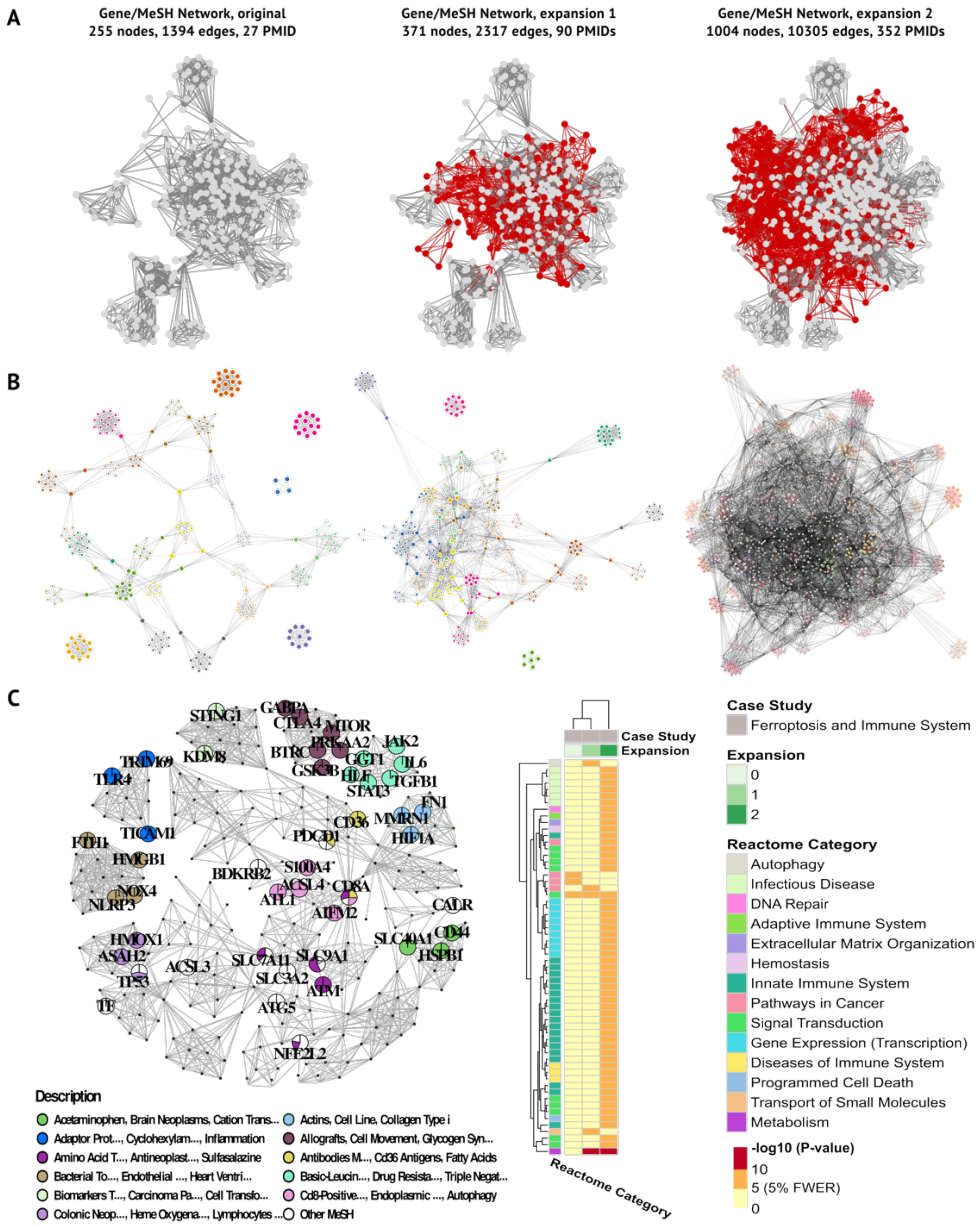

Figure 2. Example of ENQUIRE's network reconstruction, expansion and post-hoc analyses. We used the PubMed identifiers (PMIDs) obtained from the query *("Ferroptosis"[MeSH terms] AND "Immune System"[MeSH terms]) NOT "review"[Publication Type]* as input. **A**: visualization of ENQUIRE's network expansion process. Novel nodes and edges are indicated in red at each expansion. **B**: for each of the networks shown in **A**, the corresponding pruned network containing only significant cliques is depicted. Node colors indicate distinct network communities. **C-left**: output of the automatic gene set reconstruction, using the original Gene/MeSH network as input and fuzzy c-means. For simplicity, only nodes referring to genes are enlarged and labelled, and a shortened description of computed gene sets of size 2 or bigger is provided. Sector sizes of the pie-chart-shaped nodes reflect their relative membership degree with respect to each cluster. **C-right**: topology-based enrichment analysis of Reactome pathways, using original and expanded networks, as described in the Methods section. Pathways whose adjusted p-value was significant in at least one network are depicted. Reactome pathways are clustered and grouped into categories based on the database's hierarchical structure and *Top-level Pathways* and disease ontologies. FWER: Holm's family wise error rate.

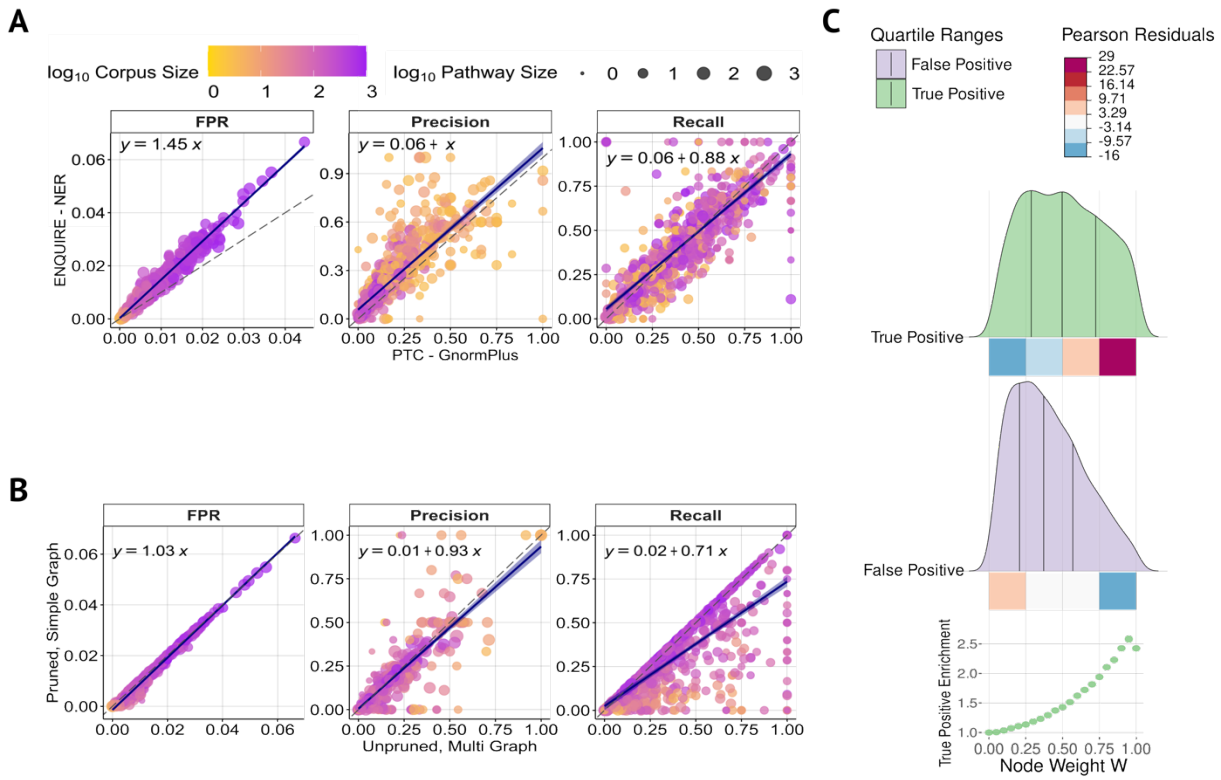

Figure 3. ENQUIRE-derived gene networks can support the manual annotation of gene sets. The scatter plots respectively compare ENQUIRE's named entity recognition (NER) to PubTatorCentral's (PTC) implementation of GNormPlus (**A**) and ENQUIRE-generated, unpruned multigraphs to pruned simple ones, obtained by applying the test statistics based on a random graph model (**B**). Each data point represents one reconstruction of a gene set from a Reactome pathway, using its reference literature abstracts. The data coordinates refer to false positive rate (FPR), precision and recall values independently computed for each tested methodology. We calculated linear regressions to compare the performances in both panels. We indicated input corpus and pathway sizes by adjusting color and dot size, respectively. **C**: aggregated distributions for true and false positive genes, segmented into quartiles. We defined four ranges of the node score W, indicated by squares, whose colors reflect Pearson standardized residuals resulting from a significant chi-square statistic. The lower chart depicts the enrichment of true positive genes, after pruning ENQUIRE-derived networks based on different values of W. Values are relative to the original proportion of true positives.

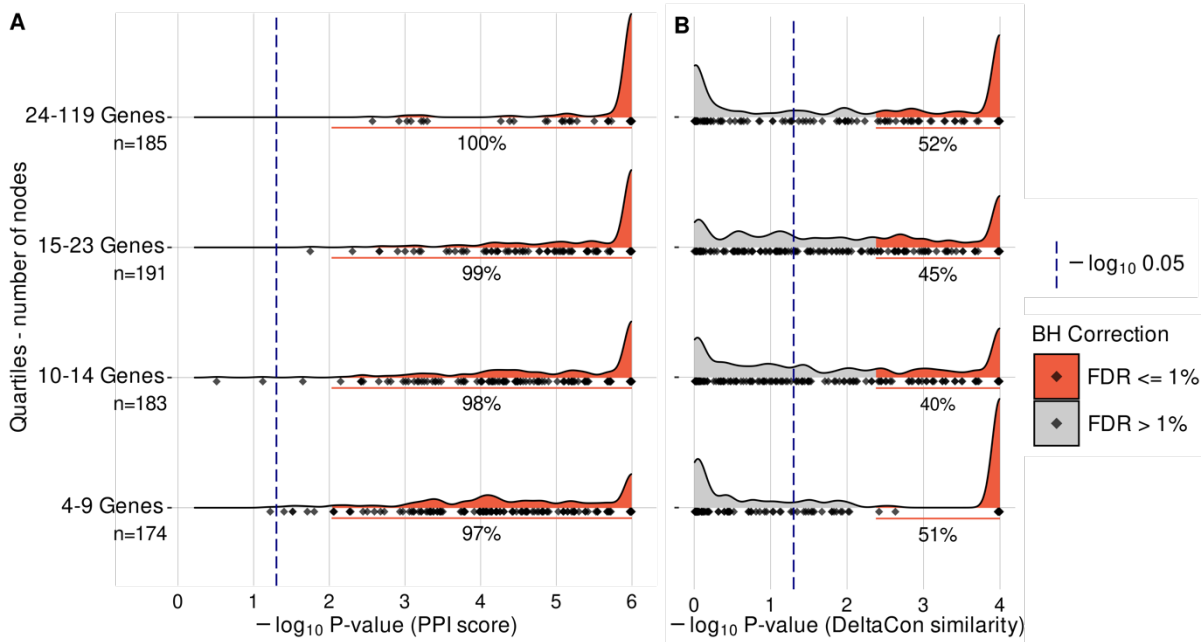

Figure 4. Protein-coding genes from ENQUIRE-generated graphs significantly share functional associations. Panels (**A**) and (**B**) respectively report the unadjusted p-value density distributions of STRING-informed edge counts and DeltaCon similarities, arranged by number of protein-coding genes (network size). We used the *H. sapiens* functional association network from STRING to evaluate ENQUIRE-derived networks of protein-coding genes. We tested 733 networks having 10 or more possible network realizations given the observed degree sequence. For each observed network size and degree sequence of ENQUIRE-generated gene networks, 1,000,000 and 10,000 samples were respectively generated to perform a test statistic on the observed edge counts and DeltaCon similarities. See Mat.Met. for additional information. The 733 tested networks are apportioned into quartiles based on network size, and for each the exact size is indicated (n). Within each network size interval, grey and red areas respectively highlight insignificant and significant p-values with respect to a globally-applied Benjamini-Hochberg correction (BH), and a percentage is indicated for those below 1% FDR. Diamonds indicate the observed data.

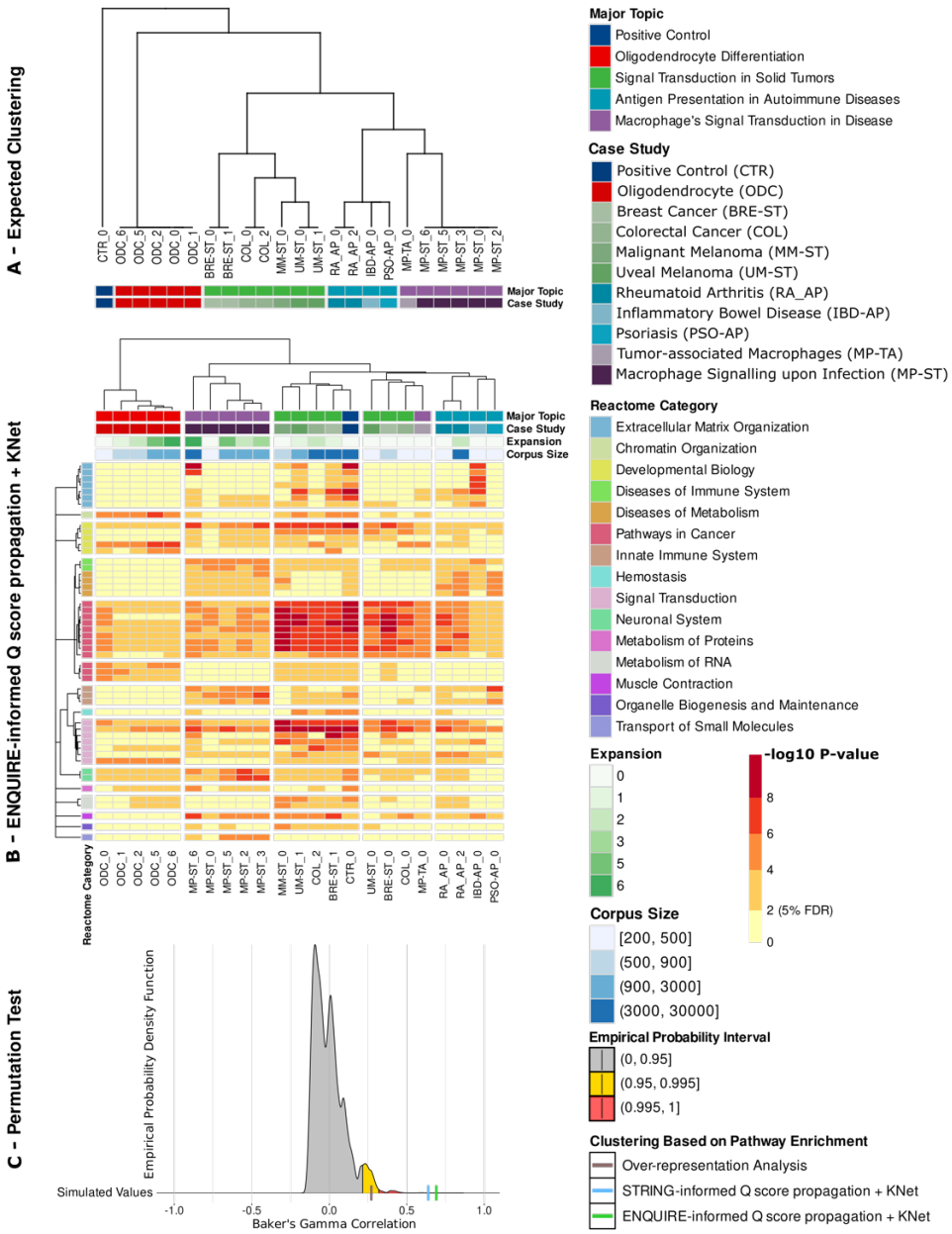

Figure 5. ENQUIRE-generated graphs enhance the context resolution of pathway enrichment analyses. A: reference dendrogram showcasing the expected categorization of the case studies described in **Table 1**. The number following a case study abbreviated name indicates the expansion counter. Network expansions that did not yield any new gene were excluded. **B**: Topology-based pathway enrichment, obtained by applying Q score propagation and SANTA's KNet function on ENQUIRE-informed gene-gene associations (see Post Hoc Analyses under Mat.Met.). The heatmap shows the unadjusted p-values for the 50 enriched Reactome pathways with at least one significant, adjusted p-value (5% FDR) and highest variance across case studies (the dendrogram was computed on the complete statistic). Pathways are clustered according to Reactome's internal hierarchy. We respectively apportioned the dendrograms into 5 and 15 partitions to visualize their coherence to Major Topic and Reactome Categories. Legends for expansions, rounded corpus size, and p-values ranges are provided. **C**: Permutation tests of Baker's gamma correlation between the reference dendrogram (**A**) and clustering obtained from alternative pathway enrichment analyses, as in **B**. Colored areas indicated probability intervals obtained from simulating correlations between reference and sampled dendrograms. See Mat.Met. for further details.