

InterFeat: An Automated Pipeline for Finding Interesting Hypotheses in Structured Biomedical Data

Dan Ofer*

Michal Linial†

Dafna Shahaf‡

May 21, 2025

Abstract

Finding *interesting* phenomena is the core of scientific discovery, but it is a manual, ill-defined concept. We present an integrative pipeline for automating the discovery of interesting simple hypotheses (feature-target relations with effect direction and a potential underlying mechanism) in structured biomedical data. The pipeline combines machine learning, knowledge graphs, literature search and Large Language Models. We formalize “interestingness” as a combination of novelty, utility and plausibility. On 8 major diseases from the UK Biobank, our pipeline consistently recovers risk factors years before their appearance in the literature. 40-53% of our top candidates were validated as interesting, compared to 0-7% for a SHAP-based baseline. Overall, 28% of 109 candidates were interesting to medical experts. The pipeline addresses the challenge of operationalizing “interestingness” scalably and for any target. We release data and code: <https://github.com/LinialLab/InterFeat>

1. Introduction

Finding interesting phenomena in data is the essence of discovery. Yet the notion of interestingness is surprisingly elusive, requiring subjective human judgment and lacking the relatively well-accepted metrics that concepts such as statistical significance enjoy.

In this work, our goal is to build a pipeline that will generate simple, interesting hypotheses about connections between features and diseases and their direction, as well as potential underlying mechanisms. Despite their simplicity, this class

of hypotheses is rich enough to be useful for researchers. We identify three core concepts that lie at the heart of interestingness: novelty, utility (usefulness), and plausibility (the existence of an underlying mechanism).

The exponential growth of data and literature has not been accompanied by a corresponding growth in insights, and finding interesting, actionable insights from data remains a challenging task. Many now-obvious discoveries, such as the link between contaminated water and disease or hand-washing, were overlooked for millennia. Hand hygiene gained acceptance only after germ theory, and *H. pylori* as the cause of ulcers was ridiculed until Marshall et al. (38)’s self-experimentation. Lithium, now essential for bipolar disorder would probably sound absurd if proposed naively (16). These insights existed in the data but were missed or dismissed due to innate biases, insufficient explanatory frameworks or statistical rigor.

This work presents an integrative framework for quantifying and automating the discovery of interesting features in scientific datasets. We focus on identifying disease risk factors from the biomedical UK BioBank (UKB), although the underlying principles and methodologies are generalizable to other populations and non-medical domains.

Our work combines machine learning, statistics and natural language processing to create an expressive, configurable and easy to use pipeline. The InterFeat pipeline leverages structured data from electronic health records, biomedical ontologies and Knowledge Graphs (**KG**), scientific literature and Large Language Models (**LLMs**) to systematically identify, rank and explain features with high potential for discovery. Our contributions are:

- We ground our approach in a formal definition of interestingness, integrating statistical and literature-based discovery approaches with LLMs to flexibly assess features based on novelty, plausibility, and utility criteria, requiring only feature and target names.
- We take advantage of LLMs to generate plausible mechanisms underlying the suggested hypotheses, guiding medical researchers in prioritizing directions

*The Hebrew University of Jerusalem, Israel. dan.ofer[at]mail.huji.ac.il

†Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Israel. michal.linial@mail.huji.ac.il

‡Department of Computer Science, The Hebrew University of Jerusalem, Israel. dshahaf@cs.huji.ac.il

to explore. Importantly, 28% of our 109 candidates were interesting to medical experts.

- We provide code and data, including a novel labeled multi-disease dataset of interesting biomedical features, with explanations.
- Applying the pipeline to the UK Biobank dataset of over 370,000 patients, we demonstrate its effectiveness in uncovering previously undocumented risk factors for 8 major diseases, showcasing its potential to accelerate biomedical discovery. Compared to a SHAP-based feature selection approach; 40-53% of our top candidates were validated as interesting, vs 0-7% for a SHAP-based baseline.
- We demonstrate the pipeline’s utility criteria’s ability to identify features that later emerge as risk factors in the literature.

2. Related Work

Automated Hypothesis generation aims to systematize the traditionally intuitive process of discovery (61; 55). Methodologies, such as literature-based discovery (LBD), aim to identify missed connections between concepts and findings, thereby uncovering novel hypotheses (29; 58; 62). However, traditional hypothesis generation approaches face several limitations: i) Directionality ignorance: These methods often treat associations between concepts as bidirectional, ignoring the direction of effect. For example, smoking reducing the risk of a disease would be novel, useful and interesting, while the inverse would not. ii) Ontology dependency: Many approaches rely on a standardized ontology and linkage to define co-occurrence, which limits it in terms of the source ontology and precision of linkage (62). Recent studies have begun to address these limitations using deep learning and graph methods to improve the flexibility of LBD approaches (63; 42).

Deep learning-based large language models have been used to automatically generate ideas and hypotheses and can flexibly capture unstructured relationships (67; 59; 61; 49; 55; 63). They have been shown to have near expert level scientific and medical understanding in some tasks, albeit when relying on known knowledge (e.g. differential diagnosis) (14; 39; 32; 51; 49; 46). However, their tendency to hallucinate unfeasible, or nonsensical ideas makes them insufficiently reliable to be a Great Automatic Grammatizator (21) for ideas without manual validation (23). The use of actual features as a starting point may reduce hallucinations, due to the more limited hypothesis space (65; 15).

In practice, the starting point of many researchers looking for interesting connections in their data is machine learning methodologies such as **feature selection** (30; 37; 22; 36). Feature selection approaches focus on predictive power or statistical significance (26; 13; 10). This includes the life

sciences, such as predicting mortality, Endometriosis, scientific trends, Depression, Heart attacks and viral-proteins (9; 19; 8; 41; 40; 2; 47). There are many works using machine learning, and SHAPley values (34) have been applied to the UK Biobank to find risk factors, but most approaches rely on manual analysis of candidates, typically from a list of features sorted by model importance (9; 5; 4; 4; 33; 36; 33). LLM-Select (30) used LLMs to select features by description and task, but again, only for predictive power.

3. Problem Definition

Given a set of datasets over the same set of biomedical features and a target feature y , our goal is output a ranked list of interesting simple hypotheses of the form “ x is related to y , with a negative/positive correlation”, together with potential mechanisms underlying the hypothesis. For example, in the case of medical data, the target y often represents whether a patient will develop a specific disease. Features x are structured patient-level variables from the data, such as age, biomarkers (Vitamin D levels), genetic risk scores, questionnaires (smoking), medical history (age of asthma diagnosis, medications), etc.

To formulate a notion of interestingness, we are inspired by creativity literature, which frequently conceptualizes innovation as a confluence of **novelty** and **utility** (25; 6; 52). In other words, a creation is deemed innovative if it is both original and valuable or useful. Similarly, we define “interesting” hypotheses in medical data that satisfy the following criteria:

1. **Novelty:** x should not be *established* in the literature or canonical knowledge bases as linked to the target y . Alternatively, a hypothesis might be considered novel if there is a known connection between x and y but the direction of effect implied by the hypothesis is controversial or unestablished. Note that a feature which is closely related to another known association may not be considered novel (e.g., smoking cigarettes vs. cigars).
2. **Utility:** x must have predictive power, adding useful information for predicting y . Note that in some use cases “utility” also implies that x needs to be actionable (e.g., smoking affects the risks of many diseases, and it can be changed).
3. **Plausibility:** In addition to the criteria inspired by creativity research, in science we believe another critical criterion is plausibility – x is consistent with the current knowledge, and has some possible theoretical explanation. Medical data is rife with correlations, many of which are spurious or simply reflect underlying confounding factors. Thus, researchers tend to prioritize investigating correlations with plausible mechanisms. For example, the role of *H. pylori* in ulcers was ignored

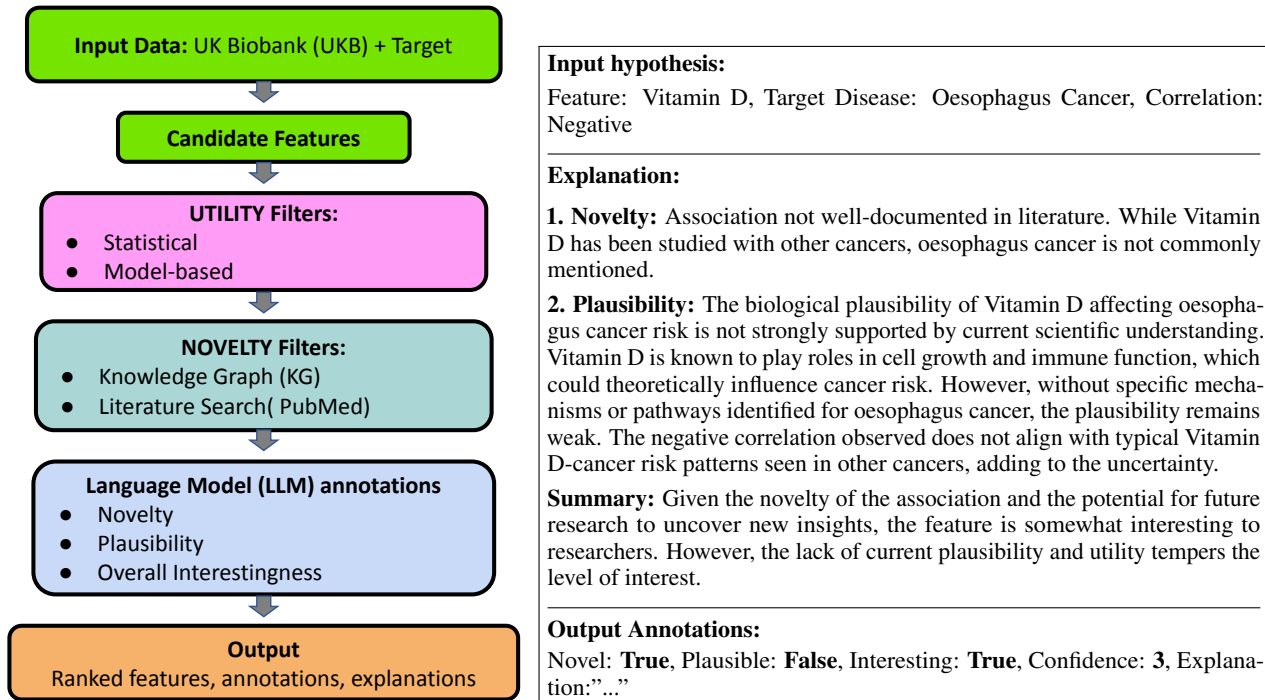


Figure 1. Left: Our pipeline, InterFeat. (i) Target and features are extracted from the UK Biobank dataset. (ii) Statistical and model-based methods are combined to identify with predictive value (Utility). (iii) UMLS-CUI linked entities are extracted and linked to a knowledge graph, to exclude known associations. Literature mining, via PubMed, removes frequent co-associations. Finally, (iv) language models (optionally augmented with relevant, retrieved texts) annotate the remaining features for novelty, plausibility and overall interestingness. Outputs include a ranked list of features with annotations and natural language explanations. **Right: LLM Annotation Example.** The input consists of a candidate feature-disease association. The LLMs provides separate judgements (combined here for clarity) for novelty, plausibility, and overall interest. This specific feature was confirmed as interesting, novel and useful by experts. Text edited for clarity.

until a biologically plausible link (bacterial infection leading to chronic gastritis) was proven.

We operationalize these requirements by formalizing notions of novelty and utility, integrating well-known metrics (e.g., mutual information) with additional LLM input. LLMs also suggest mechanisms and explanations for each score. While we are not the first to combine notions of novelty and utility, we propose an integrative, configurable approach that we hope will be adapted by practitioners and serve as a vehicle for new scientific discoveries.

4. Pipeline Implementation Details

Our pipeline is summarized in Figure 1. In the following, we provide implementation details. Code and annotated datasets are available: <https://github.com/LinialLab/InterFeat>. We cannot provide the UKBB or SemMed raw data due to licensing.

Importantly, there are various ways to formulate novelty

and utility. Our pipeline brings together the most prominent formulations, providing an intuitive way to configure and select those best suited for specific use cases.

4.1. Data: UK Biobank

We use the UK Biobank health records dataset as our main structured data source (56). The dataset contains ~1681 patient covariates (medical record history, diagnostic results, medications, socioeconomic variables, genomic factors, lifestyle, etc’) measured at the time of each patient’s initial intake (2009–2011), with ICD-10 medical diagnoses recorded through 2022, for 370K adult patients. ICD-10/ICD-10-CM codes were also mapped to their phenotypes/Phencodes as additional covariates.

4.2. Extracting Candidate Features

We clean and encode the raw UKB data into a structured format with ~3721 features. Features with missing values were mean-imputed, and a “missing” feature flag was added.

Features without at least 30 non-missing values are dropped.

Optionally, our pipeline removes redundant features using correlation feature selection. In interpretability use-cases, it is common to remove highly correlated features to reduce redundancy and computation. A popular default is 0.8-0.95 for the Pearson correlation coefficient (27; 22). We use a 0.9 threshold, so that only features with strong linear relationships are dropped as redundant, using the feature-engine library of Galli (24).

4.3. Utility Filter

The pipeline predicts whether a patient will be diagnosed in the future with a given disease (specified by ICD-10 medical codes). To help mitigate confounding by age, sex, and BMI, we optionally apply Inverse Propensity Weighting (IPW) on the negative samples (18). The predicted probabilities are used as sampling weights for IPW, and the negatives are resampled down to a given ratio (9:1)(1).

At this stage, we allow users to choose between several utility filters to remove features with no predictive strength for y , each with a corresponding threshold. Specifically,

- p -value under a univariate test:
 $pVal(x, y) \leq \theta_p$
- Mutual information between x and y :
 $MI(f, y) \geq \theta_{MI}$
- Model-based feature-importance score (e.g., global SHAP): $FImp(f, y) \geq \theta_{FImp}$

MI and $FImp$ can ascertain non-linear effects. $FImp$ reflects whether a feature is used by a trained predictive model(s), e.g., a boosting tree, unlike p -value.

Users can choose criteria and thresholds also whether to treat them as a conjunction (all) or disjunction (any). After some exploration, we chose lenient thresholds for our experiments: $p\text{-value} < 0.2$, $MI \geq 10^{-3}$, or $FImp \geq 10^{-4}$. In our selected configuration, a feature x passes the utility filter if it met any of the three criteria.

4.4. Novelty Filter

Our pipeline supports two ways to filter for novelty, both based on scientific literature.

KG-based Filter. We link features and target diseases to UMLS Concept Unique Identifiers (CUIs) (11) using scispaCy (45) and edges in SemMedDB v43 (31), a KG of 130 million semantic predications (subjectpredicateobject triples) from 37 million PubMed citations.

Features and targets are represented by sets of linked entities ($E(x)$ and $E(y)$), extracted using named entity recognition and linkage to entities in the KG (here, UMLS CUIs). If a feature x is already directly connected to the disease y

in the KG (with sufficient evidence), we mark it as known (i.e., *not novel*) and exclude it. To reduce the chance of false predicates, we filtered SemMedDB for predicates that had at least 2 unique citations as evidence, leaving 12.9 million. We treat the graph as unidirectional and ignore the type of predicate. scispaCy’s (V5.5) `en_core_sci_lg` entity recognition model was used, with a 0.88 threshold and 3 max entities per candidate, following recommendations for high-precision entity linking in biomedical texts (45; 54). A predefined list of irrelevant high level categories/semantic types are manually excluded by regex (e.g. "Qualification", "Disease", "Unit").

Domain-specific semantic similarity is computed between each feature and candidate entities, using a pretrained biomedical sentence-level language model - Biolord (50), measured as cosine similarity. This is used to further remove candidate entities with very low (defined as 0.1) similarity to the feature, and later to define "strongly linked" entities (e.g., "alcohol" and "alcoholism").

Features were filtered out if all their linked entities were directly linked (1-hop) to the target(s) in the KG, or if they had at least one strongly linked entity ($\geq \theta_{sim}$ similarity) with a direct connection to the target(s). In our experiments, we chose a threshold of $\theta_{sim} = 0.4$ as "strongly linked".

Literature-based Filter. Text mining is used to ascertain if the co-occurrence of features and disease is already established in the literature. This reflects the typical human search process: Are there any (many) papers about x and the disease y ?

PubMed is a large literature database of over 37 million published scientific and specifically biomedical works (as of 10/2024). We query the PubMed search API (including automatic term expansions) for publication counts of each feature, the target, and their co-occurrence (x AND y). If the pair is co-mentioned less than an absolute threshold θ_{lit} or less frequently than expected by random chance (via one-way Fishers Exact Test, $p < \theta_{pval}$), it is retained. Features with less than 20 hits in the database are left unfiltered (these could include, for example, recently coined terms). Again, after some experimentation, we chose to use relaxed thresholds for our experiments: $\theta_{lit} = 4$, $\theta_{pval} = 0.4$.

A note on thresholds. In both our novelty filters, we prefer high recall, filtering out only clearly non-novel features while retaining borderline cases. This prioritizes precision in exclusion, minimizing the risk of discarding under-explored but potentially meaningful findings.

4.5. LLM Annotations

To refine and rank filtered features, we use LLMs as an extra layer of information. Due to the nature of language models,

we chose to focus on novelty and plausibility: Language models are very effective for processing and internalizing vast amounts of medical knowledge, synthesizing multiple sources, and thus can often detect whether a certain hypothesis is already known. Similarly, their ability to integrate diverse pieces of knowledge and combine them in new ways helps them identify plausible mechanisms. We did not use the LLM to annotate for utility, as this is something that is often use-case specific. We annotate feature novelty and plausibility using GPT-4o-mini, selected after development-phase testing with Ai2s OpenScholar, a LLaMA-3.1-8B variant (7; 1). This was motivated by GPTs adherence to structured outputs. Chain of thought (COT) is used in all models’ prompts (64). We use retrieval-augmented generation (RAG) (15), using MedRag (66), a biomedical retrieval toolkit, to retrieve related texts from the MedCorp corpus of 23 million PubMed abstracts, clinical textbooks and Wikipedia. The top 32 texts per feature and target, ranked by BM-25 are appended to the prompt. This outputs binary (True/False) annotations and explanations.

Each feature, target, their correlation and previous models’ explanations are then run through through the GPT-4o LLM, to get overall Interestingness, a confidence (1-4) score and explanation. Prompts are provided in the appendix (Appendix A.1), and outputs in the codebase. Finally, outputs are provided in a structured format for review, including annotation labels, an Interestingness confidence score, feature statistics, and an explanation, sorted by confidence and feature importance.

See LLM output example in Figure 1 (right). In this example, low Vitamin D levels increasing the risk of Oesophageal cancer was rated as novel and moderately interesting (3/4). A mechanism from other cancers is noted, as is the unusual effect direction in this case. We note this feature was confirmed as interesting, novel and useful by annotators.

5. Using the Pipeline

We applied our pipeline to eight diverse diseases (see Table 1) in the UK Biobank (UKB). Diseases were chosen based on our access to experts who could assess the results. This is a retrospective cohort study with disease diagnosis set as the index date. For a given disease, defined by a range of ICD-10 codes, we predict for each patient if they will be diagnosed with the target disease in their post-intake future, for patients with any recorded diagnoses between 2011 and 2022, leaving 370K patients. Prediction time is defined as the date 1 year prior to the target diagnosis, after intake, for positives. For negatives, prediction time is sampled from the same date distribution as the positives. The ICD diagnosis features were limited to codes with at least 200 occurrences, at least 1 year prior to the index date.

5.1. Pipeline Statistics

The initial set of ~ 3721 features is filtered down to 500-2500 by the *utility* criteria, then further by the novelty and LLM steps, leaving less than $\sim 2\%$ (under 80) final candidates per disease. This is consistent with other works suggesting examining up to 3% of features for hypothesis exploration in large, high dimensional data, notably the UKB (35; 12; 36). Several observations from Table 1 are to be acknowledged: (i) The diseases span a wide prevalence range. The list includes rare diseases such as retinal vein occlusion (0.32%) but also high-prevalence diseases, such as depression (6.68%). (ii) The diseases cover cases of defined underlying biochemical mechanisms (e.g., gout) but also conditions without mechanistic explanation like depression. (iii) Some are early onset, while others are considered aging diseases. For example, celiac is a lifelong autoimmune disease commonly diagnosed in childhood, while gallstones are more common in adults and can be treated. We conclude that these diseases display a reliable representation of other human diseases and conditions.

We observed that the number of features retained after the utility filter correlates positively with disease prevalence. This can be attributed to the fact that larger datasets, with more cases of a target in addition to background (negative/“healthy”) cases, provide greater statistical sensitivity to detect features with even modest associations. This effect is consistent with the UKB collected covariates, although diverse, being gathered under the assumption of their potential relevance to human health and wellness. Another observation concerns the knowledge graph (KG). For example, 72% of the features remained after KG filtration in the case of retinal vein occlusion, but only 40% for depression. Presumably, the richness of the KG is associated with the popularity of specific diseases (53).

6. Reality check: Validation of Utility Filters

Evaluating hypothesis generation algorithms is challenging due to the difficulty in determining the accuracy of the generated hypotheses. The absence of definitive ground truth makes it inherently complex to assess the output of such algorithms.

In this section we assess our utility filters (the first part of the pipeline) using time-stamped validation – an accepted methodology in hypothesis generation, particularly when a definitive ground truth is unavailable (46; 17; 43; 28; 53). In a nutshell, the idea is to take a cut-off date (in our case, 2011 – when the UKB study intake took place), and run our pipeline as if that date represents the present moment. For each disease, we took all features passing our utility filter, then examined whether those same features were added as disease-associated entries in SemMedDB *after* 2011. Be-

Table 1. Pipeline statistics: Features retained at each stage per disease.

Target Disease	Disease Counts	Prevalence (%)	Number of Features kept by stage			
			Utility Filter	Knowledge Graph	Literature Search	Selected by LLM
Cholelithiasis (Gallstones)	19658	5.07	1447	697	157	50
Gout	9159	2.36	1707	812	148	62
Coeliac disease	2653	0.68	903	487	134	63
Spine degeneration	24867	6.42	2430	1187	136	73
Esophageal cancer	1518	0.39	611	408	152	59
Heart attack	3638	0.94	1008	520	102	43
Retinal Vein Occlusion	1246	0.32	558	402	163	60
Depression	28880	6.68	2537	1036	77	26

cause SemMedDB grows over time, a featuredisease link that only appears post-2011 suggests our pipeline identified it before it was recognized in the literature. Table 2 shows, per disease, how many of these discovered features were added in subsequent KG expansions, indicating that the pipeline can surface validated insights ahead of time. In particular, up to 21% of utility-filtered features appear in literature (SemMedDB) only after 2011. We found this reality check encouraging, as our utility filters were shown to retain valid features. We note that we could not perform a similar evaluation for the entire pipeline, as the LLMs are trained on data created after 2011.

7. Human Evaluation and Case Studies

Our primary question is whether the pipelines outputs are indeed interesting, according to domain experts. To investigate this, we performed a focused human evaluation on three diseases: *Gout*, *Cholelithiasis (Gallstones)*, and *Esophageal cancer*. We aimed to (i) measure alignment between expert and pipeline judgments, and (ii) assess whether experts indeed found value in the pipelines discoveries.

Three senior medical doctors, each with over 10 years research experience, including with these diseases, annotated 109 pipeline-selected features for *novelty*, *plausibility*, *utility*, and overall *interestingness*, on a 1-4 scale with explanations. The challenging and ambiguous nature of the task demanded domain expertise, and expert annotators.

Of the features marked as interesting by the models, up to 42 candidates per disease were selected by the confidence score, as given the constraints of manpower and costs, a full-scale evaluation was not feasible. Scores were binarized (> 2) when comparing with model annotations. **Overall, 28% of candidates were interesting to the doctors: 18% of Gout, 30% of oesophagus cancer and 37% of Cholelithiasis.**

7.1. Model Alignment

On binarized scores for *novelty*, the pipeline agreed with experts on **40%** of cases; for *plausibility*, **57%**; for *utility*, **79%**; and for overall *interestingness*, **69%**.

7.2. Distinguishing Real vs. Distractor Features.

To evaluate the expert annotators’ ability to distinguish meaningful features from distractors, we added **distractor features** for each annotation dataset. These features were derived by randomly sampling from those discarded which did not pass the utility filter. This helped us assess annotator bias and task difficulty. For each target we added 20% distractors, yielding 35 total additional annotation candidates, in addition to the original, real features. Annotators were not informed of the distractors. GPT-4o was prompted to generate justifications for why each distractor was interesting (Appendix A.1). It has been shown that LLMs can fool humans in such scenarios (3).

Statistical comparisons were performed using two-sample t-tests, summarized in Table 3. Human annotators recognized the distractors as having lower plausibility, utility and interestingness.

7.3. Feature Importance Baseline Comparison and Component Contribution Analysis

To evaluate InterFeat’s ability to identify interesting features compared to a baseline of selecting by feature importance, we compared (and annotated) the top 15 candidate features generated by the pipeline as well as its individual components for Gallstones, Esophageal Cancer, and Gout. Table 4 summarizes the number of features validated as interesting for each approach, out of the top 15, sorted by SHAP. SHAP (34) is a popular method for identifying feature importance, and reflects a typical data scientist or computational researchers’ likely default. SHAP shows which features drive

Table 2. Temporal Validation of Utility Filters by Target Disease. Statistics are provided for each target’s dataset of utility-filtered features. (i) total number of features linked to the KG, (ii) the number of features are directly connected (1-hop) to the target in the KG, and (iii) the count and percentage of features that were first reported after the temporal cutoff.

Target Disease	Total KG-Linked Features	KG Features (1-hop from target)	Post-Cutoff Features
Gallstones (Cholelithiasis)	801	202	33 (16%)
Gout	920	274	58 (21%)
Coeliac Disease	582	215	20 (9%)
Spine Degeneration	1130	318	63 (20%)
Esophageal Cancer	445	91	19 (21%)
Heart Attack	643	320	18 (6%)
Retinal Vein Occlusion	400	10	0
Depression	1214	537	60 (11%)

Table 3. Comparison of Human Annotations between real and distractor (Dist.) features

Annotation	Mean (Real)	Mean (Dist.)	p-Value
Novel	2.78	2.82	0.83
Plausibility	2.46	2.12	0.04
Utility	1.94	1.48	0.0005
Interestingness	2.09	1.81	0.04

model predictions, including the direction of effect and in relation to other features’ contributions, in a consistent framework. SHAP based methods have been extensively applied, including on the UKB (9; 19; 2; 33; 48), making it a natural comparison for getting a starting list of features to analyze, as in Madakkattel et al. (36).

The methods compared include: the SHAP baseline, representing feature selection based solely on predictive importance; intermediate filters (Knowledge Graph (KG) only, Literature only, and combined KG+Literature); the full InterFeat pipeline; and an additional experimental step (“InterFeat + ReasonLM”). This extra step reranked all InterFeat selections simultaneously using a separate, reasoning LLM (Google Gemini 2.5 Pro with access to web-search (60)), allowing for list-wise reranking; it serves here primarily for analytical comparison and is not part of the standard InterFeat pipeline presented. All candidates are still filtered by the baseline utility step, then sorted for top 15 by feature importance. As shown in Table 4, InterFeat consistently identified more interesting features than feature importance across all targets (Gallstones: 6 vs 1; Esoph. Ca.: 5 vs 0; Gout: 3 vs 0). Aggregating counts across the three diseases, this difference was statistically significant (Fisher’s exact test, two-sided, $p=0.0003$, $n=90$). Annotations available in Appendix (A.3) and repository (“Ablation Results”).

Table 4. Comparison of Validated Interesting Features by Method. Results for top 15 (sorted by SHAP). All methods include utility filtering

Method	Gallstones	Esoph. Ca.	Gout
SHAP Baseline	1	0	0
KG	2	0	0
Literature	3	0	0
KG+Literature	5	1	3
InterFeat	6	5	3
InterFeat+ReasonLM	6	10	5

7.4. Recurring features

Of 375 features marked as interesting by LLMs across all 8 targets, 48% were picked more than once, with 6 appearing in 7+ of the targets: ‘melanoma genetic risk’, ‘Microalbumin in urine’, intraocular pressure genetic risk’, ‘Arm fat percentage’, epithelial ovarian cancer genetic risk’, ‘age at menopause genetic risk. These highlight the fact that underlying factors such as genetics or immunology may affect many diseases (20). Not all causes of diseases are known or understood, and some may have multiple etiologies(46; 20).Furthermore, variables such as age, obesity or inflammation can drive conditions without implying direct causal links, and may reflect more fundamental factors that predispose to diseases. For instance, high arm fat percentage relates to confounders such as muscle mass, BMI and general frailty. We acknowledge that these might be caused by confounders rather than truly novel or causal effectors, although this does not necessarily affect utility (44). We grouped features into semantic categories, using a combination of manual annotation and LLM-assisted clustering (see Appendix, Figure 2.

7.5. Expert Selected Insights

Below are examples of hypotheses identified by our pipeline that were validated as particularly interesting by the expert annotators.

7.5.1. OESOPHAGEAL CANCER

Oesophageal cancer is an aggressive malignancy, defined by ICD-10 code C15. It has ~81K PubMed citations but is relatively rare in the UKB due to low survival rates.

- **Genetic Risks associated with other diseases:** e.g., melanoma, ischemic stroke, rheumatoid arthritis, systemic lupus erythematosus. The association with melanoma suggests shared genetic or inflammatory pathways. Similarly, genetic risks linked to rheumatoid arthritis and systemic lupus erythematosus indicate that autoimmune and inflammatory processes could play a role in oesophageal cancer development
- **Asthma diagnosis and genetic risk:** Possibly linked via chronic inflammation or steroids.
- **Atenolol:** a beta-blocker for cardiovascular disease.
- **Epithelial Ovarian Cancer genetic risk** exhibited a particularly interesting negative association.
- **Novel Biomarkers:** Vitamin D, Acetoacetate, Acetone.

7.5.2. GALLSTONES

Gallstones, or cholelithiasis, are a prevalent hepatobiliary disorder, with 101K citations, characterized by the formation of calculi within the gallbladder, defined by the ICD-10 range K80-K82. Complications can lead to significant morbidity, including cholecystitis and biliary obstruction.

- **Pharmacological Influences:** Omeprazole, a proton pump inhibitor. This was considered particularly intriguing and meriting further investigation.
- **Cross-Disease Genetic Risk Factors:** genetic risk factors associated with other diseases, such as breast cancer, primary open-angle glaucoma, Alzheimer's, and schizophrenia. The association between **breast cancer genetic risk** and gallstones may reflect shared metabolic pathways. Similarly, a link with **primary open-angle glaucoma genetic risk** suggests that systemic metabolic dysregulation could concurrently affect ocular and hepatobiliary health. These indicate that gallstone formation may be influenced by genetic factors common to multiple organ systems.
- **Lipid Metabolism Markers:** Apolipoprotein B/A1 ratio, Medium HDL cholesterol. The ApoB/ApoA1 ratio, indicative of lipid metabolism balance, reinforces the role of lipid dysregulation in gallstone pathogenesis. These suggest that therapeutic strategies aimed at reg-

ulating lipid profiles could be effective in preventing and managing gallstone disease.

- **Psychiatric Conditions:** Bipolar disorder, depression, neuroticism. May indicate a systemic metabolic factor or medication effect.

8. Conclusions and Future Work

We presented an integrative pipeline that combines statistical feature selection, knowledge-graph screening, and retrieval-augmented LLM annotation to discover interesting features, defined a combination of *novel*, *plausible*, and *having utility*. Our approach systematically narrows thousands of raw features to a concise shortlist. In comparison to the common practice of ranking features solely by statistical or model importance measures (e.g., SHAP values), our pipeline demonstrated superior performance. Specifically, in an expert evaluation of the top 15 features, 4053% of our top-ranked candidates for gallstones and esophageal cancer were validated as interesting, compared to only 07% for a SHAP-based baseline. Across 109 pipeline candidates, 28% overall were judged interesting by medical experts.

Despite the progress, challenges remain. Imperfect knowledge bases' coverage can lead to features being falsely labeled as novel, and LLM judgment may still misalign with humans. Future work could explore ablations of pipeline components, incorporate additional criteria for interestingness to improve alignment with human judgment, and develop more sophisticated ways of fusing structured metadata with the LLM. Improved integration of feature attributes may also help identify novelties based on unusual population subsets or non-monotonic effects, moving beyond the existing usage of directionality. Ongoing improvements in large language models with stronger incontext reasoning suggest that some early filtering stages such as the knowledgegraph pass could be folded into a single LLM query at the cost of higher compute costs (57). Exploring this tradeoff is outside our present scope but forms a natural direction for followup work. We plan to apply our pipeline on a large scale to hundreds of major diseases, providing the candidates as a community resource. Although the alignment of pipeline scores with human assessments for the top-ranked subset of candidates is modest, it is crucial to note this subset is distilled from an initial pool of thousands. Generating a sorted list of candidates enriched for interestingness improves on standard practices (e.g., ranking by predictive importance), offering clear value as a time-saving tool for researchers and a starting point for expert validation.

Our approach is flexible, and outputs a ranked set of interesting features that surpasses existing approaches, cheaply, quickly and without needing to filter out hallucinations (unlike methods that depend entirely on generation, without grounding in data). Our approach is easily generalizable to

other domains, and we look forward to expanding it, improving AI-human alignment in formulating what is interesting.

Acknowledgments

We thank Dr Tali Sahar, Dr Shai Rosenberg and Dr Gal Passi for their unpaid, voluntary contribution in annotating the candidates, and their excellent advice during development. Used under UK-Biobank application ID 26664 (Linial lab).

Impact Statement

This work’s goal is to advance the fields of Machine Learning for hypothesis generation and knowledge discovery; and aiming to improve early-stage discovery in biomedical research. By systematically identifying novel and plausible disease risk factors, InterFeat has the potential to accelerate time to insights, aid clinical research prioritization, and complement traditional expert-driven discovery.

The ethical implications of this work primarily concern the responsible interpretation and application of AI-generated hypotheses. The method does not establish causal relationships. Misuse or over-reliance on automated hypothesis generation without proper validation could lead to misleading conclusions in research. To mitigate this, we emphasize expert review and empirical validation, ensuring that AI-identified findings serve as starting points for rigorous scientific inquiry rather than definitive claims. We are clear that these are model outputs, and that they should not be followed without human analysis first, and that they should be judged carefully. Medical analysis can also be flawed, and should be rigorously validated first. Doctor’s opinions may be wrong. Future societal impacts may include improved efficiency in research, enabling researchers to uncover overlooked patterns in large-scale datasets. However, careful consideration is required when applying AI-driven discovery in clinical settings, as biases in training data, knowledge graphs, or language models could inadvertently reinforce disparities in healthcare. To address this, our approach remains transparent, interpretable, and adaptable, encouraging collaborative use with domain experts.

References

- [1] Aaron Grattafiori, Dubey, A., Jauhri, A., Abhinav Pandey, Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sra-vankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., Mc-Connell, C., Keller, C., Touret, C., Wu, C., Wong, C.,

Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., Linde, J. v. d., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhota, K., Rantala-Yearly, L., Maaten, L. v. d., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., Oliveira, L. d., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Elebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Rapparth, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola,

- B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damla, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshv, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The Llama 3 Herd of Models, November 2024. URL <http://arxiv.org/abs/2407.21783>. arXiv:2407.21783 [cs].
- [2] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., and van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE*, 14(5):e0213653, May 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0213653. URL <http://dx.plos.org/10.1371/journal.pone.0213653>. Publisher: Public Library of Science.
- [3] Alber, D. A., Yang, Z., Alyakin, A., Yang, E., Rai, S., Valliani, A. A., Zhang, J., Rosenbaum, G. R., Amend-Thomas, A. K., Kurland, D. B., Kremer, C. M., Eremiev, A., Negash, B., Wiggan, D. D., Nakatsuka, M. A., Sangwon, K. L., Neifert, S. N., Khan, H. A., Save, A. V., Palla, A., Grin, E. A., Hedman, M., Nasir-Moin, M., Liu, X. C., Jiang, L. Y., Mankowski, M. A., Segev, D. L., Aphinyanaphongs, Y., Riina, H. A., Golfinos, J. G., Orringer, D. A., Kondziolka, D., and Oermann, E. K. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pp. 1–9, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03445-1. URL <https://www.nature.com/articles/s41591-024-03445-1>.
- [4] Allwright, M., Karrasch, J. F., O’Brien, J. A., Guenewig, B., and Austin, P. J. Machine learning analysis of the UK Biobank reveals prognostic and diagnostic immune biomarkers for polyneuropathy and neuropathic pain in diabetes. *Diabetes Research and Clinical Practice*, 201:110725, July 2023. ISSN 1872-8227. doi: 10.1016/j.diabres.2023.110725.
- [5] Allwright, M., Mundell, H. D., McCorkindale, A. N., Lindley, R. I., Austin, P. J., Guenewig, B., and Sutherland, G. T. Ranking the risk factors for Alzheimers disease; findings from the UK Biobank study. *Aging Brain*, 3:100081, January 2023. ISSN 2589-9589. doi: 10.1016/j.nbas.2023.100081. URL <https://www.sciencedirect.com/science/article/pii/S258995892300018X>.

- [6] Amabile, T. M. *Creativity in context: Update to the social psychology of creativity*. Routledge, 2018.
- [7] Asai, A., He, J., Shao, R., Shi, W., Singh, A., Chang, J. C., Lo, K., Soldaini, L., Feldman, S., D'arcy, M., Wadden, D., Latzke, M., Tian, M., Ji, P., Liu, S., Tong, H., Wu, B., Xiong, Y., Zettlemoyer, L., Neubig, G., Weld, D., Downey, D., Yih, W.-t., Koh, P. W., and Hajishirzi, H. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs, November 2024. URL <http://arxiv.org/abs/2411.14199>. arXiv:2411.14199 [cs].
- [8] Bilu, Y., Kalkstein, N., Gilboa-Schechtman, E., Akiva, P., Zalsman, G., Itzhaky, L., and Atzil-Slonim, D. Predicting future onset of depression among middle-aged adults with no psychiatric history. *BJPsych Open*, 9(3):e85, May 2023. ISSN 2056-4724. doi: 10.1192/bjo.2023.62. URL https://www.cambridge.org/core/product/identifier/S2056472423000625/type/journal_article.
- [9] Blass, I., Sahar, T., Shraibman, A., Ofer, D., Rapoport, N., and Linial, M. Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach. *Journal of Personalized Medicine*, 12(7):1114, July 2022. ISSN 2075-4426. doi: 10.3390/jpm12071114.
- [10] Blum, A. L. and Langley, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1):245–271, December 1997. ISSN 0004-3702. doi: 10.1016/S0004-3702(97)00063-5. URL <https://www.sciencedirect.com/science/article/pii/S0004370297000635>.
- [11] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl.1):D267–D270, January 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh061. URL <https://doi.org/10.1093/nar/gkh061>.
- [12] Boln-Canedo, V., Snchez-Maroo, N., and Alonso-Betanzos, A. Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5(2): 65–75, May 2016. ISSN 2192-6352, 2192-6360. doi: 10.1007/s13748-015-0080-y. URL <http://link.springer.com/10.1007/s13748-015-0080-y>.
- [13] Breiman, L. U. o. C. *Random forest*, volume 45. 1999. ISBN 978-1-4244-4442-7. doi: 10.1023/A:1010933404324. arXiv: <http://dx.doi.org/10.1023%2FA%3A1010933404324>
- Publication Title: Machine Learning ISSN: 0885-6125.
- [14] Brodeur, P. G., Buckley, T. A., Kanjee, Z., Goh, E., Ling, E. B., Jain, P., Cabral, S., Abdulnour, R.-E., Haimovich, A., Freed, J. A., Olson, A., Morgan, D. J., Hom, J., Gallo, R., Horvitz, E., Chen, J., Manrai, A. K., and Rodman, A. Superhuman performance of a large language model on the reasoning tasks of a physician, December 2024. URL <http://arxiv.org/abs/2412.10849>. arXiv:2412.10849 [cs].
- [15] Bchard, P. and Ayala, O. M. Reducing hallucination in structured outputs via Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 228–238, 2024. doi: 10.18653/v1/2024.naacl-industry.19. URL <http://arxiv.org/abs/2404.08189>. arXiv:2404.08189 [cs].
- [16] Cade, J. F. and Malhi, G. S. Cades lithium. *Acta Neuropsychiatrica*, 19(2):125–126, April 2007. ISSN 0924-2708, 1601-5215. doi: 10.1111/j.1601-5215.2007.00196.x. URL <https://www.cambridge.org/core/journals/acta-neuropsychiatrica/article/abs/cades-lithium/58EE56AEE9AC91DA880509589D1F469E>.
- [17] Chan, J., Chang, J. C., Hope, T., Shahaf, D., and Kittur, A. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–21, November 2018. ISSN 2573-0142. doi: 10.1145/3274300. URL <https://dl.acm.org/doi/10.1145/3274300>.
- [18] Chesnaye, N. C., Stel, V. S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, C., and Jager, K. J. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15(1):14–20, August 2021. ISSN 2048-8505. doi: 10.1093/ckj/sfab158. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8757413/>.
- [19] Cohen, S., Dagan, N., Cohen-Inger, N., Ofer, D., and Rokach, L. ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. *IEEE Access*, 9: 91584–91592, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3091622. Conference Name: IEEE Access.
- [20] Dahl, A. and Zaitlen, N. Genetic Influences on Disease Subtypes. *Annual Review of Genomics*

- and *Human Genetics*, 21(1):413–435, 2020. doi: 10.1146/annurev-genom-120319-095026. URL <https://doi.org/10.1146/annurev-genom-120319-095026>. eprint: <https://doi.org/10.1146/annurev-genom-120319-095026>.
- [21] Dahl, R. *The Great Automatic Grammatizator and Other Stories*. Puffin, 1997. ISBN 978-0-14-037915-0. Google-Books-ID: 5WbDNQAACAAJ.
- [22] Domingos, P. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78, October 2012. ISSN 00010782. doi: 10.1145/2347736.2347755. URL <http://dl.acm.org/citation.cfm?doid=2347736.2347755>.
- [23] Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL <https://www.nature.com/articles/s41586-024-07421-0>. Publisher: Nature Publishing Group.
- [24] Galli, S. Feature-engine: A Python package for feature engineering for machine learning. *Journal of Open Source Software*, 6(65):3642, September 2021. ISSN 2475-9066. doi: 10.21105/joss.03642. URL <https://joss.theoj.org/papers/10.21105/joss.03642>.
- [25] Glover, J. A., Ronning, R. R., and Reynolds, C. R. *Handbook of creativity*. Springer Science & Business Media, 2013.
- [26] Guyon, I. An Introduction to Variable and Feature Selection I Introduction. 3:1157–1182, 2003.
- [27] Hall, M. A. *Correlation-based feature selection for machine learning*. PhD Thesis, The University of Waikato, 1999. URL <https://researchcommons.waikato.ac.nz/handle/10289/15043>.
- [28] Harel, S. and Radinsky, K. Accelerating Prototype-Based Drug Discovery using Conditional Diversity Networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 331–339, July 2018. doi: 10.1145/3219819.3219882. URL <http://arxiv.org/abs/1804.02668>. arXiv:1804.02668 [cs].
- [29] Henry, S. and McInnes, B. T. Literature Based Discovery: Models, methods, and trends. *Journal of Biomedical Informatics*, 74:20–32, October 2017. ISSN 1532-0464. doi: 10.1016/j.jbi.2017.08.011. URL <https://www.sciencedirect.com/science/article/pii/S1532046417301909>.
- [30] Jeong, D. P., Lipton, Z. C., and Ravikumar, P. LLM-Select: Feature Selection with Large Language Models, July 2024. URL <http://arxiv.org/abs/2407.02694>. arXiv:2407.02694.
- [31] Kilicoglu, H., Shin, D., Fiszman, M., Rosembat, G., and Rindflesch, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, December 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts591. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3509487/>.
- [32] Livin, V., Hother, C. E., Motzfeldt, A. G., and Winther, O. Can large language models reason about medical questions? *Patterns*, 5(3), March 2024. ISSN 2666-3899. doi: 10.1016/j.patter.2024.100943. URL [https://www.cell.com/patterns/abstract/S2666-3899\(24\)00042-4](https://www.cell.com/patterns/abstract/S2666-3899(24)00042-4). Publisher: Elsevier.
- [33] Lugner, M., Rawshani, A., Helleryd, E., and Eliasson, B. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*, 14(1):2102, January 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-52023-5. URL <https://www.nature.com/articles/s41598-024-52023-5>. Publisher: Nature Publishing Group.
- [34] Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.
- [35] Madakkatel, I. and Hyppnen, E. LLpowershap: logistic loss-based automated Shapley values feature selection method. *BMC Medical Research Methodology*, 24(1):247, October 2024. ISSN 1471-2288. doi: 10.1186/s12874-024-02370-8. URL <https://doi.org/10.1186/s12874-024-02370-8>.
- [36] Madakkatel, I., Zhou, A., McDonnell, M. D., and Hyppnen, E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Scientific Reports*, 11(1):22997, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-02476-9. URL <https://www.nature.com/articles/>

- s41598-021-02476-9. Publisher: Nature Publishing Group.
- [37] Maor, M., Karidi, R., Davidovich, S., and Ronen, A. System and method for automatic generation of features from datasets for use in an automated machine learning process, September 2019. URL <https://patents.google.com/patent/US10410138B2/en?inventor=Amir+Ronen>.
- [38] Marshall, B. J., Armstrong, J. A., McGechie, D. B., and Clancy, R. J. Attempt to fulfil Koch’s postulates for pyloric *Campylobacter*. *Medical Journal of Australia*, 142(8):436–439, 1985. ISSN 1326-5377. doi: 10.5694/j.1326-5377.1985.tb113443.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.5694/j.1326-5377.1985.tb113443.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.5694/j.1326-5377.1985.tb113443.x>.
- [39] Matsumoto, N., Moran, J., Choi, H., Hernandez, M. E., Venkatesan, M., Wang, P., and Moore, J. H. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btac353, June 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac353. URL <https://doi.org/10.1093/bioinformatics/btac353>.
- [40] Michael-Pitschaze, T., Cohen, N., Ofer, D., Hoshen, Y., and Linial, M. Detecting anomalous proteins using deep representations. *NAR Genomics and Bioinformatics*, 6(1):lqae021, March 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae021. URL <https://doi.org/10.1093/nargab/lqae021>.
- [41] Moore, A. and Bell, M. XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. *Clinical Medicine Insights: Cardiology*, 16: 11795468221133611, January 2022. ISSN 1179-5468. doi: 10.1177/11795468221133611. URL <https://doi.org/10.1177/11795468221133611>.
- [42] Moreau, E., Hardiman, O., Heverin, M., and O’Sullivan, D. Literature-Based Discovery beyond the ABC paradigm: a contrastive approach, September 2021. URL <https://www.biorxiv.org/content/10.1101/2021.09.22.461375v1>.
- [43] Moreau, E., Hardiman, O., Heverin, M., and O’Sullivan, D. Mining impactful discoveries from the biomedical literature. *BMC bioinformatics*, 25 (1):303, September 2024. ISSN 1471-2105. doi: 10.1186/s12859-024-05881-9.
- [44] Nastl, V. Y. and Hardt, M. Do causal predictors generalize better to new domains? November 2024. URL <https://openreview.net/forum?id=U4BC0GrFAz>.
- [45] Neumann, M., King, D., Beltagy, I., and Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327, 2019. doi: 10.18653/v1/W19-5034. URL <https://www.aclweb.org/anthology/W19-5034>. Conference Name: Proceedings of the 18th BioNLP Workshop and Shared Task Place: Florence, Italy Publisher: Association for Computational Linguistics.
- [46] Ofer, D. and Linial, M. Automated annotation of disease subtypes. *Journal of Biomedical Informatics*, 154:104650, June 2024. ISSN 1532-0464. doi: 10.1016/j.jbi.2024.104650. URL <https://www.sciencedirect.com/science/article/pii/S1532046424000686>.
- [47] Ofer, D., Kaufman, H., and Linial, M. What’s next? Forecasting scientific research trends. *Heliyon*, 10 (1):e23781, January 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2023.e23781. URL <https://www.sciencedirect.com/science/article/pii/S2405844023109893>.
- [48] Peduzzi, G., Felici, A., Pellungrini, R., and Campa, D. Explainable machine learning identifies a polygenic risk score as a key predictor of pancreatic cancer risk in the UK Biobank. *Digestive and Liver Disease: Official Journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver*, pp. S1590–8658(24)01100–9, December 2024. ISSN 1878-3562. doi: 10.1016/j.dld.2024.11.010.
- [49] Qi, B., Zhang, K., Tian, K., Li, H., Chen, Z.-R., Zeng, S., Hua, E., Jinfang, H., and Zhou, B. Large Language Models as Biomedical Hypothesis Generators: A Comprehensive Evaluation, July 2024. URL <http://arxiv.org/abs/2407.08940>. arXiv:2407.08940 [cs].
- [50] Remy, F., Demuynck, K., and Demeester, T. BioLORD: Learning Ontological Representations from Definitions (for Biomedical Concepts and their Textual Descriptions), October 2022. URL <http://arxiv.org/abs/2210.11892>. arXiv:2210.11892 [cs].
- [51] Shringarpure, S. S., Wang, W., Karagounis, S., Wang, X., Reissetter, A. C., Auton, A., and Khan, A. A. Large language models identify causal genes in complex trait GWAS, May 2024.

- URL <https://www.medrxiv.org/content/10.1101/2024.05.30.24308179v1>.
- [52] Silberschatz, A. and Tuzhilin, A. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8 (6):970–974, December 1996. ISSN 1558-2191. doi: 10.1109/69.553165. URL <https://ieeexplore.ieee.org/document/553165>. Conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [53] Singer, U., Radinsky, K., and Horvitz, E. On biases of attention in scientific discovery. *Bioinformatics*, pp. btaa1036, December 2020. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btaa1036. URL <https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa1036/6039114>.
- [54] Soldaini, L. QuickUMLS: a fast, unsupervised approach for medical concept extraction. 2016. URL <https://www.semanticscholar.org/paper/QuickUMLS%3A-a-fast%2C-unsupervised-approach-for-Soldaini/92e428bcd578f504974103f7201be21807f13615>.
- [55] Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C. R., Comer, A., Myers, J. N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J. J., Parikh, N., Lisewski, A. M., Donehower, L., Chen, Y., and Lichtarge, O. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1877–1886, New York New York USA, August 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623667. URL <https://dl.acm.org/doi/10.1145/2623330.2623667>.
- [56] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1676. doi: 10.1371/journal.pmed.1001779. URL <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779>. Publisher: Public Library of Science.
- [57] Sutton, R. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.
- [58] Swanson, D. R. Fish Oil, Raynaud’s Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, September 1986. ISSN 1529-8795. doi: 10.1353/pbm.1986.0087. URL <https://muse.jhu.edu/article/403510>.
- [59] Sybrandt, J., Carrabba, A., Herzog, A., and Safro, I. Are Abstracts Enough for Hypothesis Generation? *arXiv:1804.05942 [cs]*, October 2018. URL <http://arxiv.org/abs/1804.05942>. arXiv: 1804.05942.
- [60] Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P. R., Hennigan, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.-T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdiah, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Gra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Weisz, ., Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M. A., Baeuml, M., Chen, Z., Shafey, L. E., Zhang, Y., Sericinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., Glehn, T. v., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rustemi, A., Clay, N., Crone, P., Kocisky, T., Zhao, J., Perz, B., Yu, D., Howard, H., Bloniarz, A., Rae, J. W., Lu, H., Sifre, L., Maggioni, M., Alcober, F., Garrette, D., Barnes, M., Thakoor, S., Austin, J., Barth-Maron, G., Wong, W., Joshi, R., Chaabouni, R., Fatiha, D., Ahuja, A., Tomar, G. S., Senter, E., Chadwick, M., Kornakov, I., Attaluri, N., Iturrate, I., Liu, R., Li, Y., Cogan, S., Chen, J., Jia, C., Gu, C., Zhang, Q., Grimstad, J., Hartman, A. J., Garcia, X., Pillai, T. S., Devlin, J., Laskin, M., Casas, D. d. L., Valter, D., Tao, C., Blanco, L., Badia, A. P., Reitter, D., Chen, M., Brennan, J., Rivera, C., Brin, S., Iqbal, S., Surita, G., Labanowski, J., Rao, A., Winkler, S., Parisotto, E., Gu, Y., Olszewska, K., Addanki, R., Miech, A., Louis, A., Teplyashin, D., Brown, G., Catt, E., Balaguer, J., Xiang, J., Wang, P., Ashwood, Z., Briukhov, A., Webson, A., Ganapathy, S., Sanghavi, S., Kannan, A.,

Chang, M.-W., Stjerngren, A., Djolonga, J., Sun, Y., Bapna, A., Aitchison, M., Pejman, P., Michalewski, H., Yu, T., Wang, C., Love, J., Ahn, J., Bloxwich, D., Han, K., Humphreys, P., Sellam, T., Bradbury, J., Godbole, V., Samangoeei, S., Damoc, B., Kaskasoli, A., Arnold, S. M. R., Vasudevan, V., Agrawal, S., Riesa, J., Lepikhin, D., Tanburn, R., Srinivasan, S., Lim, H., Hodgkinson, S., Shyam, P., Ferret, J., Hand, S., Garg, A., Paine, T. L., Li, J., Li, Y., Giang, M., Neitz, A., Abbas, Z., York, S., Reid, M., Cole, E., Chowdhery, A., Das, D., Rogoziska, D., Nikolaev, V., Sprechmann, P., Nado, Z., Zilka, L., Prost, F., He, L., Monteiro, M., Mishra, G., Welty, C., Newlan, J., Jia, D., Allamanis, M., Hu, C. H., Liedekerke, R. d., Gilmer, J., Saroufim, C., Rijhwani, S., Hou, S., Shrivastava, D., Baddepudi, A., Goldin, A., Oztirel, A., Cassirer, A., Xu, Y., Sohn, D., Sachan, D., Amplayo, R. K., Swanson, C., Petrova, D., Narayan, S., Guez, A., Brahma, S., Landon, J., Patel, M., Zhao, R., Villela, K., Wang, L., Jia, W., Rahtz, M., Gimnez, M., Yeung, L., Keeling, J., Georgiev, P., Mincu, D., Wu, B., Haykal, S., Saputro, R., Vodrahalli, K., Qin, J., Cankara, Z., Sharma, A., Fernando, N., Hawkins, W., Neyshabur, B., Kim, S., Hutter, A., Agrawal, P., Castro-Ros, A., Driessche, G. v. d., Wang, T., Yang, F., Chang, S.-y., Komarek, P., McIlroy, R., Lui, M., Zhang, G., Farhan, W., Sharman, M., Natsev, P., Michel, P., Bansal, Y., Qiao, S., Cao, K., Shakeri, S., Butterfield, C., Chung, J., Rubenstein, P. K., Agrawal, S., Mensch, A., Soparkar, K., Lenc, K., Chung, T., Pope, A., Maggiore, L., Kay, J., Jhakra, P., Wang, S., Maynez, J., Phuong, M., Tobin, T., Tacchetti, A., Trebacz, M., Robinson, K., Katariya, Y., Riedel, S., Bailey, P., Xiao, K., Ghelani, N., Aroyo, L., Slone, A., Houlisby, N., Xiong, X., Yang, Z., Gribovskaya, E., Adler, J., Wirth, M., Lee, L., Li, M., Kagohara, T., Pavagadhi, J., Bridgers, S., Bortsova, A., Ghemawat, S., Ahmed, Z., Liu, T., Powell, R., Bolina, V., Iinuma, M., Zablotzkaia, P., Besley, J., Chung, D.-W., Dozat, T., Comanescu, R., Si, X., Greer, J., Su, G., Polacek, M., Kaufman, R. L., Tokumine, S., Hu, H., Buchatskaya, E., Miao, Y., Elhawaty, M., Siddhant, A., Tomasev, N., Xing, J., Greer, C., Miller, H., Ashraf, S., Roy, A., Zhang, Z., Ma, A., Filos, A., Besta, M., Blevins, R., Klimenko, T., Yeh, C.-K., Changpinyo, S., Mu, J., Chang, O., Pajarskas, M., Muir, C., Cohen, V., Lan, C. L., Haridasan, K., Marathe, A., Hansen, S., Douglas, S., Samuel, R., Wang, M., Austin, S., Lan, C., Jiang, J., Chiu, J., Lorenzo, J. A., Sjsund, L. L., Cevey, S., Gleicher, Z., Avrahami, T., Boral, A., Srinivasan, H., Selo, V., May, R., Aisopos, K., Hussenot, L., Soares, L. B., Baumli, K., Chang, M. B., Recasens, A., Caine, B., Pritzel, A., Pavetic, F., Pardo, F., Gergely, A., Frye, J., Ramasesh, V., Horgan, D., Badola, K., Kassner, N., Roy, S., Dyer, E., Campos,

V. C., Tomala, A., Tang, Y., Badawy, D. E., White, E., Mustafa, B., Lang, O., Jindal, A., Vikram, S., Gong, Z., Caelles, S., Hemsley, R., Thornton, G., Feng, F., Stokowiec, W., Zheng, C., Thacker, P., nl, ., Zhang, Z., Saleh, M., Svensson, J., Bileschi, M., Patil, P., Anand, A., Ring, R., Tshilas, K., Vezer, A., Selvi, M., Shevlane, T., Rodriguez, M., Kwiatkowski, T., Daruki, S., Rong, K., Dafoe, A., FitzGerald, N., Gu-Lemberg, K., Khan, M., Hendricks, L. A., Pellat, M., Feinberg, V., Cobon-Kerr, J., Sainath, T., Rauh, M., Hashemi, S. H., Ives, R., Hasson, Y., Noland, E., Cao, Y., Byrd, N., Hou, L., Wang, Q., Sottiaux, T., Paganini, M., Lespiau, J.-B., Moufarek, A., Hassan, S., Shivakumar, K., Amersfoort, J. v., Mandhane, A., Joshi, P., Goyal, A., Tung, M., Brock, A., Sheahan, H., Misra, V., Li, C., Rakievi, N., Dehghani, M., Liu, F., Mittal, S., Oh, J., Noury, S., Sezener, E., Huot, F., Lamm, M., Cao, N. D., Chen, C., Mudgal, S., Stella, R., Brooks, K., Vasudevan, G., Liu, C., Chain, M., Melinker, N., Cohen, A., Wang, V., Seymore, K., Zubkov, S., Goel, R., Yue, S., Krishnakumaran, S., Albert, B., Hurley, N., Sano, M., Mohananey, A., Joughin, J., Filonov, E., Kpa, T., Eldawy, Y., Lim, J., Rishi, R., Badiezadegan, S., Bos, T., Chang, J., Jain, S., Padmanabhan, S. G. S., Puttagunta, S., Krishna, K., Baker, L., Kalb, N., Bedapudi, V., Kurzrok, A., Lei, S., Yu, A., Litvin, O., Zhou, X., Wu, Z., Sobell, S., Siciliano, A., Papir, A., Neale, R., Bragagnolo, J., Toor, T., Chen, T., Anklin, V., Wang, F., Feng, R., Gholami, M., Ling, K., Liu, L., Walter, J., Moghaddam, H., Kishore, A., Adamek, J., Mercado, T., Mallinson, J., Wandekar, S., Cagle, S., Ofek, E., Garrido, G., Lombriser, C., Mukha, M., Sun, B., Mohammad, H. R., Matak, J., Qian, Y., Peswani, V., Janus, P., Yuan, Q., Schelin, L., David, O., Garg, A., He, Y., Duzhyi, O., Igmyr, A., Lottaz, T., Li, Q., Yadav, V., Xu, L., Chinien, A., Shivanna, R., Chuklin, A., Li, J., Spadine, C., Wolfe, T., Mohamed, K., Das, S., Dai, Z., He, K., Dincklage, D. v., Upadhyay, S., Maurya, A., Chi, L., Krause, S., Salama, K., Rabinovitch, P. G., M. P. K. R., Selvan, A., Dektiarev, M., Ghiasi, G., Guven, E., Gupta, H., Liu, B., Sharma, D., Shtacher, I. H., Paul, S., Akerlund, O., Aubet, F.-X., Huang, T., Zhu, C., Zhu, E., Teixeira, E., Fritze, M., Bertolini, F., Marinescu, L.-E., Blle, M., Paulus, D., Gupta, K., Latkar, T., Chang, M., Sanders, J., Wilson, R., Wu, X., Tan, Y.-X., Thiet, L. N., Doshi, T., Lall, S., Mishra, S., Chen, W., Luong, T., Benjamin, S., Lee, J., Andrejczuk, E., Rabiej, D., Ranjan, V., Styrc, K., Yin, P., Simon, J., Harriott, M. R., Bansal, M., Robsky, A., Bacon, G., Greene, D., Mirylenka, D., Zhou, C., Sarvana, O., Goyal, A., Andermatt, S., Siegler, P., Horn, B., Israel, A., Pongetti, F., Chen, C.-W. L., Selvatici, M., Silva, P., Wang, K., Tolins, J., Guu, K., Yogeve, R., Cai, X., Agostini, A., Shah, M., Nguyen,

H., Donnaile, N. ., Pereira, S., Friso, L., Stambler, A., Kurzrok, A., Kuang, C., Romanikhin, Y., Geller, M., Yan, Z. J., Jang, K., Lee, C.-C., Fica, W., Malmi, E., Tan, Q., Banica, D., Balle, D., Pham, R., Huang, Y., Avram, D., Shi, H., Singh, J., Hidey, C., Ahuja, N., Saxena, P., Dooley, D., Potharaju, S. P., O'Neill, E., Gokulchandran, A., Foley, R., Zhao, K., Dusenberry, M., Liu, Y., Mehta, P., Kotikalapudi, R., Safranek-Shrader, C., Goodman, A., Kessinger, J., Globen, E., Kolhar, P., Gorgolewski, C., Ibrahim, A., Song, Y., Eichenbaum, A., Brovelli, T., Potluri, S., Lahoti, P., Baetu, C., Ghorbani, A., Chen, C., Crawford, A., Pal, S., Sridhar, M., Gurita, P., Mujika, A., Petrovski, I., Cedoz, P.-L., Li, C., Chen, S., Santo, N. D., Goyal, S., Punjabi, J., Kappaganthu, K., Kwak, C., LV, P., Velury, S., Choudhury, H., Hall, J., Shah, P., Figueira, R., Thomas, M., Lu, M., Zhou, T., Kumar, C., Jurdi, T., Chikkerur, S., Ma, Y., Yu, A., Kwak, S., hdel, V., Rajayogam, S., Choma, T., Liu, F., Barua, A., Ji, C., Park, J. H., Hellendoorn, V., Bailey, A., Bilal, T., Zhou, H., Khatir, M., Sutton, C., Rzadkowski, W., Macintosh, F., Vij, R., Shagin, K., Medina, P., Liang, C., Zhou, J., Shah, P., Bi, Y., Dankovics, A., Banga, S., Lehmann, S., Bredezen, M., Lin, Z., Hoffmann, J. E., Lai, J., Chung, R., Yang, K., Balani, N., Brainskas, A., Sozanschi, A., Hayes, M., Alcalde, H. F., Makarov, P., Chen, W., Stella, A., Snijders, L., Mandl, M., Krrman, A., Nowak, P., Wu, X., Dyck, A., Vaidyanathan, K., R, R., Mallet, J., Rudominer, M., Johnston, E., Mittal, S., Udathu, A., Christensen, J., Verma, V., Irving, Z., Santucci, A., Elsayed, G., Davoodi, E., Georgiev, M., Tenney, I., Hua, N., Cideron, G., Leurent, E., Alnahlawi, M., Georgescu, I., Wei, N., Zheng, I., Scandinaro, D., Jiang, H., Snoek, J., Sundararajan, M., Wang, X., Ontiveros, Z., Karo, I., Cole, J., Rajashekhar, V., Tumeh, L., Ben-David, E., Jain, R., Uesato, J., Datta, R., Bunyan, O., Wu, S., Zhang, J., Stanczyk, P., Zhang, Y., Steiner, D., Naskar, S., Azzam, M., Johnson, M., Paszke, A., Chiu, C.-C., Elias, J. S., Mohiuddin, A., Muhammad, F., Miao, J., Lee, A., Vieillard, N., Park, J., Zhang, J., Stanway, J., Garmon, D., Karmarkar, A., Dong, Z., Lee, J., Kumar, A., Zhou, L., Evens, J., Isaac, W., Irving, G., Loper, E., Fink, M., Arkatkar, I., Chen, N., Shafran, I., Petrychenko, I., Chen, Z., Jia, J., Levskaya, A., Zhu, Z., Grabowski, P., Mao, Y., Magni, A., Yao, K., Snaider, J., Casagrande, N., Palmer, E., Suganthan, P., Castao, A., Giannoumis, I., Kim, W., Rybiski, M., Sreevatsa, A., Prendki, J., Soergel, D., Goedeckemeyer, A., Gierke, W., Jafari, M., Gaba, M., Wiesner, J., Wright, D. G., Wei, Y., Vashisht, H., Kulizhskaya, Y., Hoover, J., Le, M., Li, L., Iwuanyanwu, C., Liu, L., Ramirez, K., Khorlin, A., Cui, A., LIN, T., Wu, M., Aguilar, R., Pallo, K., Chakladar, A., Perng, G., Abellan, E. A., Zhang, M.,

Dasgupta, I., Kushman, N., Penchev, I., Repina, A., Wu, X., Weide, T. v. d., Ponnappalli, P., Kaplan, C., Simsa, J., Li, S., Dousse, O., Yang, F., Piper, J., Ie, N., Pasumarthi, R., Lintz, N., Vijayakumar, A., Andor, D., Valenzuela, P., Lui, M., Paduraru, C., Peng, D., Lee, K., Zhang, S., Greene, S., Nguyen, D. D., Kurylowicz, P., Hardin, C., Dixon, L., Janzer, L., Choo, K., Feng, Z., Zhang, B., Singhal, A., Du, D., McKinnon, D., Antropova, N., Bolukbasi, T., Keller, O., Reid, D., Finchelstein, D., Raad, M. A., Crocker, R., Hawkins, P., Dadashi, R., Gaffney, C., Franko, K., Bulanova, A., Leblond, R., Chung, S., Askham, H., Cobo, L. C., Xu, K., Fischer, F., Xu, J., Sorokin, C., Alberti, C., Lin, C.-C., Evans, C., Dimitriev, A., Forbes, H., Banarse, D., Tung, Z., Omernick, M., Bishop, C., Sterneck, R., Jain, R., Xia, J., Amid, E., Piccinno, F., Wang, X., Banzal, P., Mankowitz, D. J., Polozov, A., Krakovna, V., Brown, S., Bateni, M., Duan, D., Firoiu, V., Thotakuri, M., Natan, T., Geist, M., Girgin, S. t., Li, H., Ye, J., Roval, O., Tojo, R., Kwong, M., Lee-Thorp, J., Yew, C., Sinopalnikov, D., Ramos, S., Mellor, J., Sharma, A., Wu, K., Miller, D., Sonnerat, N., Vnukov, D., Greig, R., Beattie, J., Caveness, E., Bai, L., Eisen-schlos, J., Korchemniy, A., Tsai, T., Jasarevic, M., Kong, W., Dao, P., Zheng, Z., Liu, F., Yang, F., Zhu, R., Teh, T. H., Sanmiya, J., Gladchenko, E., Trdin, N., Toyama, D., Rosen, E., Tavakkol, S., Xue, L., Elkind, C., Woodman, O., Carpenter, J., Papamakarios, G., Kemp, R., Kafle, S., Grunina, T., Sinha, R., Talbert, A., Wu, D., Owusu-Afriyie, D., Du, C., Thornton, C., Pont-Tuset, J., Narayana, P., Li, J., Fatehi, S., Wieting, J., Ajmeri, O., Uria, B., Ko, Y., Knight, L., Hliou, A., Niu, N., Gu, S., Pang, C., Li, Y., Levine, N., Stolovich, A., Santamaria-Fernandez, R., Goenka, S., Yustalim, W., Strudel, R., Elqursh, A., Deck, C., Lee, H., Li, Z., Levin, K., Hoffmann, R., Holtmann-Rice, D., Bachem, O., Arora, S., Koh, C., Yeganeh, S. H., Pder, S., Tariq, M., Sun, Y., Ionita, L., Seyedhosseini, M., Tafti, P., Liu, Z., Gulati, A., Liu, J., Ye, X., Chrzaszcz, B., Wang, L., Sethi, N., Li, T., Brown, B., Singh, S., Fan, W., Parisi, A., Stanton, J., Koverkathu, V., Choquette-Choo, C. A., Li, Y., Lu, T. J., Ittycheriah, A., Shroff, P., Varadarajan, M., Bahargam, S., Willoughby, R., Gaddy, D., Desjardins, G., Cornero, M., Robenek, B., Mittal, B., Albrecht, B., Shenoy, A., Moiseev, F., Jacobsson, H., Ghaffarkhah, A., Rivire, M., Walton, A., Crepy, C., Parrish, A., Zhou, Z., Farabet, C., Radebaugh, C., Srinivasan, P., Salm, C. v. d., Fidjeland, A., Scellato, S., Latorre-Chimoto, E., Klimczak-Pluciska, H., Bridson, D., Cesare, D. d., Hudson, T., Mendolicchio, P., Walker, L., Morris, A., Mauger, M., Guseynov, A., Reid, A., Odoom, S., Loher, L., Cotruta, V., Yenugula, M., Grewe, D., Petrushkina, A., Duerig, T., Sanchez, A., Yadlowsky, S., Shen, A., Globerson, A., Webb, L.,

- Dua, S., Li, D., Bhupatiraju, S., Hurt, D., Qureshi, H., Agarwal, A., Shani, T., Eyal, M., Khare, A., Belle, S. R., Wang, L., Tekur, C., Kale, M. S., Wei, J., Sang, R., Saeta, B., Liechty, T., Sun, Y., Zhao, Y., Lee, S., Nayak, P., Fritz, D., Vuyyuru, M. R., Aslanides, J., Vyas, N., Wicke, M., Ma, X., Eltyshv, E., Martin, N., Cate, H., Manyika, J., Amiri, K., Kim, Y., Xiong, X., Kang, K., Luisier, F., Tripuraneni, N., Madras, D., Guo, M., Waters, A., Wang, O., Ainslie, J., Baldridge, J., Zhang, H., Pruthi, G., Bauer, J., Yang, F., Mansour, R., Gelman, J., Xu, Y., Polovets, G., Liu, J., Cai, H., Chen, W., Sheng, X., Xue, E., Ozair, S., Angermueller, C., Li, X., Sinha, A., Wang, W., Wiesinger, J., Koukoumidis, E., Tian, Y., Iyer, A., Gurumurthy, M., Goldenson, M., Shah, P., Blake, M. K., Yu, H., Urbanowicz, A., Palomaki, J., Fernando, C., Durden, K., Mehta, H., Momchev, N., Rahimtoroghi, E., Georgaki, M., Raul, A., Ruder, S., Redshaw, M., Lee, J., Zhou, D., Jalan, K., Li, D., Hechtman, B., Schuh, P., Nasr, M., Milan, K., Mikulik, V., Franco, J., Green, T., Nguyen, N., Kelley, J., Mahendru, A., Hu, A., Howland, J., Vargas, B., Hui, J., Bansal, K., Rao, V., Ghiya, R., Wang, E., Ye, K., Sarr, J. M., Preston, M. M., Elish, M., Li, S., Kaku, A., Gupta, J., Pasupat, I., Juan, D.-C., Someswar, M., M. T., Chen, X., Amini, A., Fabrikant, A., Chu, E., Dong, X., Muthal, A., Buthpitiya, S., Jauhari, S., Hua, N., Khandelwal, U., Hitron, A., Ren, J., Rinaldi, L., Drath, S., Dabush, A., Jiang, N.-J., Godhia, H., Sachs, U., Chen, A., Fan, Y., Taitelbaum, H., Noga, H., Dai, Z., Wang, J., Liang, C., Hamer, J., Ferng, C.-S., Elkind, C., Atias, A., Lee, P., Listk, V., Carlen, M., Kerkhof, J. v. d., Pikus, M., Zaher, K., Mller, P., Zykova, S., Stefanec, R., Gatsko, V., Hirnschall, C., Sethi, A., Xu, X. F., Ahuja, C., Tsai, B., Stefanoiu, A., Feng, B., Dhandhaniala, K., Katyal, M., Gupta, A., Parulekar, A., Pitta, D., Zhao, J., Bhatia, V., Bhavnani, Y., Alhadlaq, O., Li, X., Danenberg, P., Tu, D., Pine, A., Filippova, V., Ghosh, A., Limonchik, B., Urala, B., Lanka, C. K., Clive, D., Sun, Y., Li, E., Wu, H., Hongtongsak, K., Li, I., Thakkar, K., Omarov, K., Majmundar, K., Alverson, M., Kucharski, M., Patel, M., Jain, M., Zabelin, M., Pelagatti, P., Kohli, R., Kumar, S., Kim, J., Sankar, S., Shah, V., Ramachandruni, L., Zeng, X., Bariach, B., Weidinger, L., Vu, T., Andreev, A., He, A., Hui, K., Kashem, S., Subramanya, A., Hsiao, S., Hassabis, D., Kavukcuoglu, K., Sadovskiy, A., Le, Q., Strohmaier, T., Wu, Y., Petrov, S., Dean, J., and Vinyals, O. Gemini: A Family of Highly Capable Multimodal Models, May 2025. URL <http://arxiv.org/abs/2312.11805>. arXiv:2312.11805 [cs].
- [61] Tong, S., Mao, K., Huang, Z., Zhao, Y., and Peng, K. Automating psychological hypothesis generation with AI: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1–14, July 2024. ISSN 2662-9992. doi: 10.1057/s41599-024-03407-5. URL <https://www.nature.com/articles/s41599-024-03407-5>.
- [62] Voytek, J. B. and Voytek, B. Automated cognitive construction and semi-automated hypothesis generation. *Journal of neuroscience methods*, 208(1):92–100, June 2012. ISSN 1872-678X. doi: 10.1016/j.jneumeth.2012.04.019. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3376233&tool=pmcentrez&rendertype=abstract>. Publisher: Elsevier B.V.
- [63] Wang, Q., Downey, D., Ji, H., and Hope, T. SciMON: Scientific Inspiration Machines Optimized for Novelty, June 2024. URL <http://arxiv.org/abs/2305.14259>. arXiv:2305.14259 [cs] version: 7.
- [64] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, pp. 24824–24837, Red Hook, NY, USA, November 2022. Curran Associates Inc. ISBN 978-1-71387-108-8.
- [65] Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., and Grau, V. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation, October 2024. URL <http://arxiv.org/abs/2408.04187>. arXiv:2408.04187.
- [66] Xiong, G., Jin, Q., Lu, Z., and Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.372. URL <https://aclanthology.org/2024.findings-acl.372/>.
- [67] Zhou, Y., Liu, H., Srivastava, T., Mei, H., and Tan, C. Hypothesis Generation with Large Language Models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, 2024. doi: 10.18653/v1/2024.nlp4science-1.10. URL <http://arxiv.org/abs/2404.04326>. arXiv:2404.04326 [cs].

A. APPENDIX

A.1. Prompts

Prompts used in code, that included loading the relevant variables from data. More details can be seen in codebase, e.g. "run_pipe-llmCall.ipynb" and the function `def generate_medrag_prompts`. 'feature_name_clean' is the name of the feature, with cleaning of punctuations, whitespaces, etc'. The MedRag library expects multiple choice questions format, hence the attached responses.

```
novelty_question =
    f"Is an association (with {direction} correlation) between the feature '{
feature_name_clean}' ('{raw_name_clean}') and '{target_clean}' novel, surprising, or
not well-documented in current knowledge?"

novelty_options = {
    "A": "Yes, it is novel, provides new insights or contradicts established understanding.",
    "B": "No, it is not novel, or is already well-known or established."}

plausibility_question = (
    f"Does it make sense for the feature '{feature_name_clean}' (raw: '{raw_name_clean}')"
    to be ({direction}) associated with '{target_clean}' based on known mechanisms,
    pathways or theories?. Is there a plausible explanation (or mechanism) for this
    relationship that makes sense?"
)

plausibility_options = {
    "A": "Yes, there is a plausible explanation for this relationship.",
    "B": "No, there is no plausible explanation for this relationship."
}

% # Adjusted Utility Question and Options
% utility_question = (
%     f"Assess the utility of the feature '{feature_name_clean}' (raw: '{raw_name_clean}')"
%     for predicting '{target_clean}'. Does this feature potentially have practical
%     relevance or potential utility?"
% )
% utility_options = {
%     "A": "Yes, it has potential utility or practical relevance.",
%     "B": "No, it lacks utility or practical relevance."
% }
```

Listing 1. Novelty, Plausibility, and Utility prompts with options

```
Evaluate the feature '{row['raw_name']}' in relation to predicting the target disease: '{
target}'. The feature has a {direction} correlation with the target disease (when
predicting 1 year in advance, and after controlling for age, gender and BMI; so
magnitude of correlation or feature importance are less important).

### Criteria Definitions:
- **Novelty:** Assess whether the feature ({feature_name_clean}) provides new insights,
contradicts established understanding, or explores controversial associations not well-
documented in existing literature. (i.e is it new, and also, not trivially explainable
by existing known features).
- **Plausibility:** Evaluate if the association makes logical sense based on known
mechanisms, biological pathways, social or environmental factors or established risk
factors.
- **Usefulness/utility:** (Optional) Does the feature have any potential practical
applications or utility, such as informing clinical interventions or tests, detection,
usage in models or policy implications.

### Existing Explanations:
These explanations are from weak critics and some literature, so you may regard them at
your discretion or rely on your own knowledge and step by step analysis.
**Novelty Explanation:**
```

```

{novelty_cot}

**Plausibility Explanation:**
{plausible_cot}

### Additional Information:
- **Feature Split:** {feature_split}
- **Feature Lift (for target==True) under feature split:** {row['F.Split-Lift (y==1)']}
Evaluate how **interesting** this feature is to a medical researcher, biologist, clinician
or basic research. Take into account world knowledge, analysis, vibes and also the
criteria of **novelty** and **plausibility**.

**Instructions:**
1. **Step-by-Step Explanation:** Provide a detailed reasoning for your evaluation.
2. **Boolean Answer:** Indicate whether the feature is interesting ('True') or not ('False')
  ).
3. **Numeric Score:** Assign a score between 1 and 4, where 1 = "Not interesting/novel at
  all" and 4 = "Extremely novel and interesting".

**Output Format:**
Provide your response in **JSON format** strictly adhering to the schema:
```json
{{
 "step_by_step_explanation": "<Your detailed explanation>",
 "answer": <True/False>,
 "numeric_score": <1-4>
}}
```

**Example Response:**
```json
{{
 "step_by_step_explanation": "The feature 'X' shows a significant association with lower
 'Y', this is opposite to the directions expected from known literature or science, as
 X would be expected to have an opposite effect due to its involvement in Z.",
 "answer": True,
 "numeric_score": 4
}}
```

**Ensure that the JSON is valid and follows the exact structure without any additional
fields or deviations. Do not include any text outside of the JSON block.**

```

Listing 2. Interesting Prompt

```

Your goal is to create examples for testing medical doctors. This output will be used for
training. You must not reveal this fact in your output - be convincing. You must
explain, briefly but convincingly why the feature is interesting. Always explain it as
interesting.
Given the feature '{row['raw_name']}' in relation to predicting the target disease: '{
target}'. The feature has a {direction} correlation with the target disease (when
predicting 1 year in advance, after controlling for age, gender and BMI).

### Interestingness Criteria Definitions (in context of disease prediction/risk/diagnosis)
:
- **Novelty:** Assess whether the feature ({feature_name_clean}) provides new insights,
  contradicts established understanding, or explores controversial associations not well-
  documented in existing literature. (i.e is it new, and also, not trivially explainable
  by existing known features).
- **Plausibility:** Evaluate if the association makes logical sense based on known
  mechanisms, biological pathways, social or environmental factors or established risk
  factors.
- **Usefulness/utility:** (Optional) Does the feature have any potential practical
  applications or utility, such as informing clinical interventions or tests, detection,

```

```
usage in models or policy implications.  
Evaluate how interesting the feature is to a medical researcher, biologist, clinician  
or basic research. Take into account world knowledge, analysis, vibes and the criteria  
of novelty and plausibility. Output a short explanation of why it is  
interesting.
```

Listing 3. Deceptive Prompt - for fake distractors

A.2. Feature recurrence clusters - detailed

The 2 level , more detailed clustering of the semantic clusters of features that were marked as interesting by the pipeline, shown here. Clustering done via manual review and GPT-4o assisted topics. Full list of features and their clusterings in codebase:

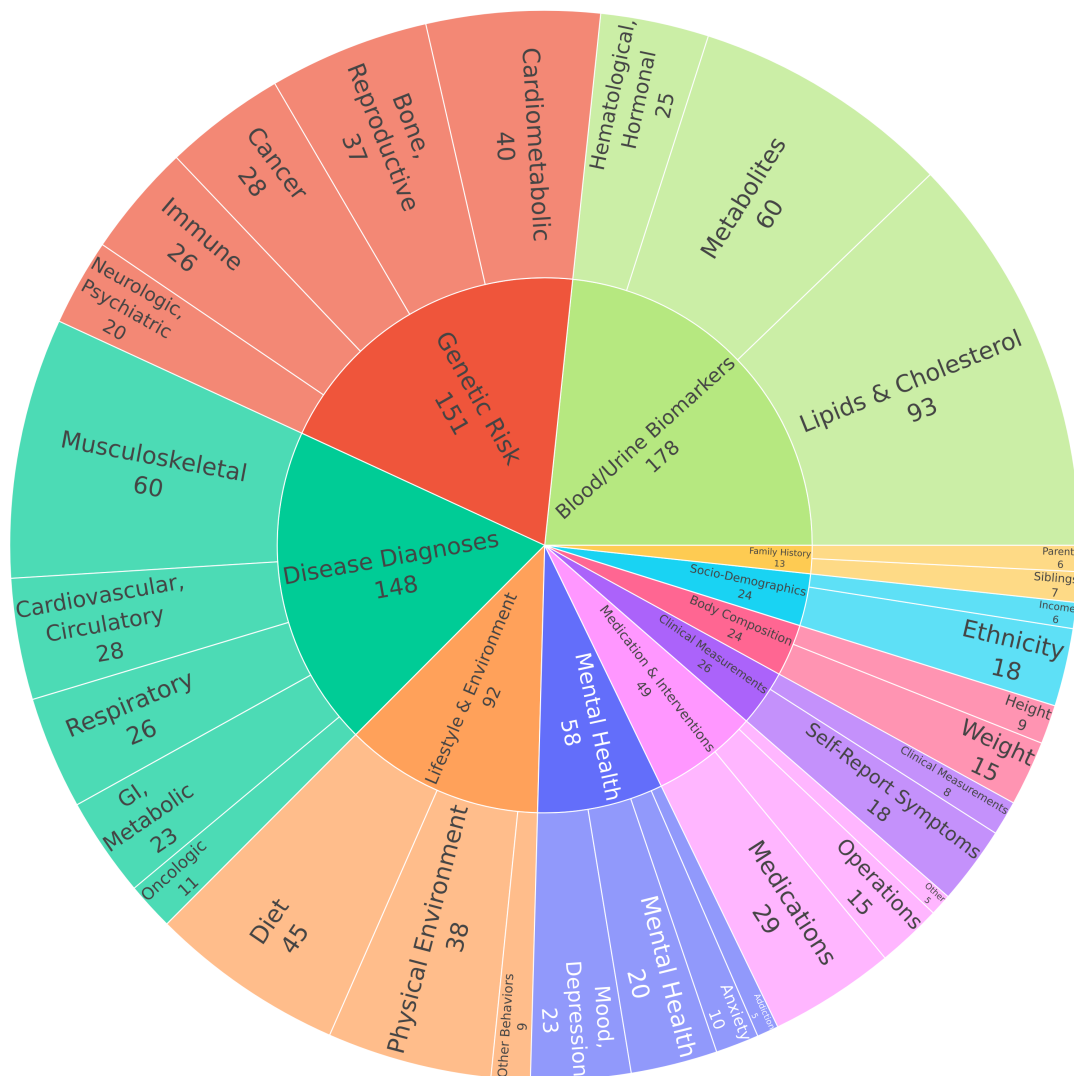


Figure 2. A two-level hierarchical sunburst plot of the recurring features, in semantic clusters. The inner ring represents broad categories (e.g., Genetic Risk, Metabolic Markers, Disease Diagnoses), while the outer ring refines these into more specific subgroups. Features marked as "interesting" by the LLM annotation models were grouped into semantic categories. The number in each section indicates the count of times features in that group were judged as interesting by models across disease targets.

A.3. Ablation Feature annotations - Shap baseline:

Top features per target, selected by Shapley value. With (anonymized) annotator comments. Full table with all annotations provided in repository: "Outputs/ablation/Ablation Results.xlsx"

Gallstone - Shap ranked features:

Picked Feature:

- **Haemoglobin concentration:** Marked as interesting if association is confirmed; potential novel link.

Not Picked Features:

- **Apolipoprotein A / B (Blood biochemistry):** Well-established markers; not novel.
- **Urban area (Scotland - Large Urban Area):** Too broad; lacks specificity.
- **Long-standing illness or disability (Yes):** Too generic; not condition-specific.
- **No medication for cholesterol/blood pressure/diabetes:** Captures known risk profile; lacks added value.
- **Self-reported gout (multiple entries):** Related to metabolic disorders, but not specific to gallstones.
- **Number of non-cancer illnesses (self-reported):** Too generic; lacks mechanistic insight.
- **Number of medications taken:** Non-specific health proxy.
- **Standing height:** Unrelated to gallstone risk.
- **Allopurinol use (medication code):** Too common; not specific.
- **Urate levels:** Linked to metabolic health; non-specific.
- **Water intake:** Too vague; low predictive value.
- **Weight (p21002):** Known risk factor; expected, especially post-weight loss.

Oesophagus Cancer Oesophagus Cancer - Shap ranked features: Picked Features:

- None.

Not Picked Features:

- **Alanine aminotransferase:** Associated with metabolic syndrome, but non-specific.
- **Alcohol intake (daily or almost daily):** Common lifestyle factor; lacks novelty.
- **Apolipoprotein A / B (Blood biochemistry):** Related to metabolic health; too general.
- **Hip circumference:** Linked to metabolic risk, but not specific to oesophageal cancer.
- **Urban area (Scotland - Large Urban Area):** Too broad and not mechanistically informative.
- **Leg fat-free mass (right):** Non-specific body composition measure.
- **No long-standing illness or disability:** Too generic for predictive use.
- **Self-reported gout:** Metabolic indicator, but not directly linked to oesophageal cancer.
- **Number of non-cancer illnesses (self-reported):** General health burden; lacks specificity.
- **Number of medications taken:** Proxy for general health; too broad.
- **Allopurinol use (medication code):** Associated with metabolic conditions; not cancer-specific.
- **Urate:** Related to fatty liver/metabolic syndrome; lacks specific linkage.
- **No vascular/heart problems (doctor-diagnosed):** Generic health indicator.
- **Water intake:** Broad lifestyle measure; low relevance to cancer risk.

Gout Feature Annotations Gout - Shap ranked features: Picked Features:

- None.

Not Picked Features:

- **Apolipoprotein A / B (Blood biochemistry):** Related to metabolic syndrome, not specific to gout.

-
- **Hip circumference:** Too broad; lacks condition specificity.
 - **Urban area (Scotland - Large Urban Area):** Too general; not mechanistically linked.
 - **Leg fat-free mass (right):** Too broad; low specificity.
 - **No medication for cholesterol, blood pressure or diabetes:** Broad metabolic proxy; not specific.
 - **Self-reported gout (multiple entries):** Redundant; already defines the outcome.
 - **Allopurinol use (medication code):** Clear but tautological; directly reflects gout treatment.
 - **Urate:** Clear, but expected and diagnostic.
 - **Number of non-cancer illnesses (self-reported):** General health indicator; too broad.
 - **Number of medications taken:** Non-specific measure of health status.
 - **Standing height:** Irrelevant to gout.
 - **Urea:** General metabolic marker; low specificity.
 - **Water intake:** Generic lifestyle factor; lacks predictive strength.

A.4. Annotator instructions:

Instructions provided to the human annotators (along with the candidate features): **Annotator Instructions for Interesting Features Annotation**

Instructions The following is a list of features, found to be predictive in predicting future onset of a specific disease at least 1 year prior to the diseases diagnosis. The population for all diseases is an adult cohort from the UK Biobank, partially controlled for BMI, gender, and age. Features include medical diagnoses, lifestyle factors, test results, demographics, and questionnaires (e.g., diet). We want to find interesting features.

Each feature is accompanied by:

- **Feature name**
- **AI model explanation** (optional to consider, as the models reasoning is not always robust)
- **Direction of correlation** with the target disease (e.g., positively or negatively correlated)

We need your expert judgment on how **novel**, **plausible**, **useful**, and **overall interesting** each feature is.

What to Do Your task is to evaluate how:

1. **Novel** (Is this association new or unexpected?)
2. **Plausible/Makes sense** (Does it make sense based on current knowledge?)
3. **Useful/Utility** (Would it have practical or clinical relevance?)
4. **Overall Interesting** (Considering its novelty, plausibility, and utility)

The feature appears. You will assign a score for each criterion using a **14 scale**:

- **1 - Strongly Disagree**
- **2 - Disagree**
- **3 - Agree**
- **4 - Strongly Agree**

(For instance, Novelty: 4 would mean you *Strongly Agree* this feature is novel.)

You may also add comments to clarify your rating and overall opinion, in the Comments column.

For example, for the overall Interesting rating:

- **1:** Not interesting at all
- **4:** Really interesting, e.g., would like to research it further; or is a feature I would want to present as an example in a paper

Feel free to ignore or only lightly use the AI model explanations (and literature citations) provided with each feature.

Example Annotations Below are **illustrative scenarios** showing how you might apply these 4-point ratings. Note how the scale is applied to each criterion:

Example 1 Disease: Lung Cancer

Feature: Smoking nicotine, positively correlated

- **Novelty:** 1 (Strongly Disagree that its novel; we already know this link well)
- **Plausibility:** 4 (Strongly Agree it is plausible; decades of evidence support it)
- **Utility:** 3 (Agree it is useful; its actionable for prevention, but also well-known)
- **Overall Interestingness:** 1 (Strongly Disagree; its too obvious to be interesting)

Example 2 Disease: Lung Cancer

Feature: Smoking nicotine, **negatively** correlated

- **Novelty:** 4 (Strongly Agree that its novel; it contradicts established knowledge)
- **Plausibility:** 1 (Strongly Disagree its plausible; no known mechanism to support this)
- **Utility:** 1 (Strongly Disagree its useful; even if data said protective, the broader health implications make it unlikely to be applied)
- **Overall Interestingness:** 4 (Strongly Agree; if truly robust, this is *very* intriguing and worth deeper research)

Rating Scale Definitions Each criterion should be rated on a scale of **1 (Strongly Disagree) to 4 (Strongly Agree)**. Below are some general guidelines for interpreting the scale in each category:

1. Novelty

- **1 (Strongly Disagree):** Not novel at all; this association is obvious or firmly established.
- **2 (Disagree):** Slightly novel; mildly surprising, but there is some prior knowledge or literature.
- **3 (Agree):** Moderately novel; not extensively documented, raises interesting questions.
- **4 (Strongly Agree):** Highly novel; very surprising or challenges current literature/knowledge.

2. Plausibility/makes sense

- **1 (Strongly Disagree):** Not plausible; conflicts with well-established evidence or lacks a clear mechanism.
- **2 (Disagree):** Low plausibility; rationale is weak or uncertain.
- **3 (Agree):** Reasonably plausible; aligns with known mechanisms or partial evidence.
- **4 (Strongly Agree):** Very plausible; strongly supported by known biology, social factors, or established theories.

3. Utility (Usefulness)

- **1 (Strongly Disagree):** Not useful; offers no clear practical benefit or application.
- **2 (Disagree):** Slightly useful; may have niche relevance but limited broader impact.
- **3 (Agree):** Moderately useful; could inform some research or clinical decisions.
- **4 (Strongly Agree):** Highly useful; likely to have real-world impact (e.g., guiding interventions, policy, or significant new research).

4. Overall Interestingness

- **1 (Strongly Disagree):** Not interesting at all; trivial, already well-known, or not worth further inquiry.
- **2 (Disagree):** Somewhat interesting; minor curiosity but probably no significant follow-up.
- **3 (Agree):** Moderately interesting; has enough novelty/plausibility/utility to prompt some investigation.
- **4 (Strongly Agree):** Very interesting; stands out as a new insight or provocative idea youd want to research or present.

Interestingness: An interesting feature should be novel, somewhat plausible, have utility, and be the basis of usefulness. Evaluate how *interesting* this feature is to a researcher, biologist, clinician, or doctor.