The 'Obesity Knowledge Portal':

an AI-powered integrative platform for exploration of obesity genetics

Yifan Liang[1], Shiqi Ao[1], Wei Xu[2], Baoguo Li[3], Jun-feng Bi[1], Tong-jin Zhao[1], Peng Li[3,4], John R. Speakman[5,6,7,8,9,*], Guanlin Wang[1,*]

Affiliations

[1] Shanghai Key Laboratory of Metabolic Remodelling and Health, Institute of Metabolism and Integrative Biology, School of Basic Medical Sciences, Centre for Evolutionary Biology, Fudan University, Shanghai, 200438, China

[2] Yunshu AI Drug Discovery Ltd, Shanghai, 201400, China

[3] State Key Laboratory of Metabolic Dysregulation & Prevention and Treatment of Esophageal Cancer, Tianjian Laboratory of Advanced Biomedical Sciences, Academy of Medical Sciences, Zhengzhou University, Zhengzhou, 450001, China

[4] Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, 100871, China

[5] Shenzhen Key Laboratory of Metabolic Health, Center for Energy Metabolism and Reproduction, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

[6] State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, 100101, China

[7] Institute of Health Sciences, China Medical University, Shenyang, Liaoning, 110122, China

[8] Faculty of Pharmacy, Shenzhen University of Advanced Technology, Shenzhen, 518107, Guangdong, China

[9] School of Biological Sciences, University of Aberdeen, AB24 2TZ, Aberdeen, UK

*Corresponding authors: **GW** (guanlin_wang@fudan.edu.cn)

Summary

Obesity is a chronic, multifactorial disease in which genetics plays a major role. Genome-wide association studies (GWAS) have identified thousands of loci associated with obesity and related traits, however, translating these findings into mechanistic understanding and clinical application remains challenging. Here we present an open-access and AI-powered web platform, Obesity Knowledge Portal (OKP, https://obesityknowledge.org), which covers 166,582 curated variant-trait associations mapped to 11,298 unique genes from published GWAS and UK Biobank summary statistics, a manually curated database of 49 clinical-stage obesity therapeutics with gene-level GWAS evidence annotation, and a domain-specific AI assistant built on Retrieval-Augmented Generation (RAG) of 10,198 peer-reviewed publications. The portal provides three core analytical capabilities: interactive exploration of gene- and variant-level associations across multiple datasets with cross-linking to clinical-stage drug targets; a genomic target prioritisation algorithm that combines statistical significance and variant density to prioritise novel candidates from genes not targeted by existing therapeutics, and literature-grounded interpretation of genetic and pharmacological queries through a scope-restricted LLM-driven contextualisation system. We demonstrate the portal's utility through case studies at the *FTO* locus and an integrated pharmacogenomic analysis of the GLP-1/incretin axis. The OKP is freely available and aims to support the translation of genetic discoveries into therapeutic hypotheses for obesity.

Keywords: obesity, knowledge portal, AI, large language model, GWAS, precision medicine

Introduction

Obesity affects over 890 million adults worldwide and contributes to the global burden of type 2 diabetes, cardiovascular disease, and certain cancers[1,2]. Twin and family studies have consistently demonstrated that genetic factors account for 40-70% of the variation in obesity[3-5], establishing human genetics as a foundation for discovering novel therapeutic targets and enabling precision medicine strategies. Since the landmark identification of the *FTO* locus variant rs9939609 in 2007[6], large-scale genome-wide association studies (GWAS), including the GIANT consortium and Pan-UK Biobank, have identified thousands of independent loci associated with body mass index (BMI) and obesity-related traits[7-12]. While these findings have suggested the central role of hypothalamic and neuronal pathways in appetite regulation in obesity[7,13,14], other studies have emphasized the broad tissue expression of these target genes[15-17], and the majority of GWAS loci have not been linked to specific biological mechanisms. Very few have been translated into therapeutic targets, and these have been restricted to treatments for monogenic obesities such as mutations of the *POMC* gene, *PCSK1* gene and leptin receptor gene that interfere with signalling to the MC4 receptor. Patients with these mutations can be treated with setmelanotide which is a direct agonist of the *MC4R*[18].

This translational gap is particularly striking given the recent clinical success of drug targets subsequently shown to have genetic support. Semaglutide, a GLP-1 receptor agonist, achieves approximately 16% body weight reduction in randomised controlled trials[19,20], Tirzepatide, a dual GLP-1/GIP receptor agonist, approximately 23%[21] and the triple agonist Retatrutide demonstrated weight loss exceeding 29% in phase 3 studies[22,23]. The incretin-based therapies targeting *GLP1R*, *GIPR* and *GCGR* were not initially identified as genes associated with obesity through human genetics. They emerged from use as treatment for T2D, where weight loss was noted as a side effect. Genetic evidence later showed that variations in or near these genes have since been identified in GWAS of BMI and metabolic traits[7,12,24,25], and recent pharmacogenomic

analyses have linked such variation to interindividual differences in weight-loss response[24,26,27]. Although retrospective, this genetic corroboration nonetheless supports the principle that human genetic evidence can validate and prioritise therapeutic targets. This suggests that systematic mining of GWAS data for druggable genes may accelerate the next generation of obesity therapeutics, however, there is no existing resource providing the computational infrastructure for this purpose.

While several resources contribute to the interpretation of obesity GWAS data, none provides a unified framework spanning genetic association, functional context and therapeutic translation. The NHGRI-EBI GWAS Catalog[28] provides a comprehensive repository of published associations but does not link them to drug targets or therapeutic pipelines. The Type 2 Diabetes Knowledge Portal[29] demonstrates the value of disease-focused genetic resources but does not cover the obesity pharmacological landscape. Open Targets[30] integrates broadly across multiple diseases but lacks the mechanism of action, clinical efficacy, and GWAS evidence strength per compound that are essential for obesity-focused target assessment. Meanwhile, research publications on obesity genetics and pharmacotherapy have grown at an extraordinary rate, with thousands of papers published annually on GWAS findings, incretin pharmacology, adipose biology, and mechanistic studies on obesity-associated genes, making manual synthesis by individual researchers increasingly difficult. The recent emergence of large language models (LLMs) as scientific tools presents an unprecedented opportunity to address this challenge, enabling the transformation of static genomic annotations into dynamic, context-aware interpretive summaries[31], provided that such models can be grounded in curated, domain-specific evidence rather than generic training data.

To address these challenges, we developed the Obesity Knowledge Portal (OKP, https://obesityknowledge.org), a freely available open-access and monthly updated web platform that provides an integrated environment for obesity genetics exploration and drug target discovery powered by AI. The portal consolidates genome-wide association

data from multiple sources into a unified browser with cross-dataset comparison, regional association plotting, and protein interaction network visualisation. It incorporates a curated pharmacological layer linking genetic evidence to the current obesity therapeutic pipeline, supported by a scoring framework for systematic target prioritisation. To enable scalable evidence synthesis across the rapidly expanding obesity literature, the platform deploys a Retrieval-Augmented Generation (RAG) system that grounds LLM outputs in over 10,000 domain-specific publications, delivering context-aware interpretation on demand. Together, these capabilities establish the OKP as a unified resource for navigating from genetic association to therapeutic hypothesis in obesity genetics and precision pharmacology.

Results

Overview of the Obesity Knowledge Portal

The Obesity Knowledge Portal integrates three complementary data layers within a unified web interface (**Fig. 1A**). The genetic layer covers obesity-related variant-trait associations from three primary sources. From the NHGRI-EBI GWAS Catalog, we extracted all associations for obesity-related traits using keyword-based filtering across the mapped traits (**Methods**), retaining variants with valid gene annotations and yielding 63,960 variant-trait associations mapped to 11,298 genes across 661 distinct traits, drawn from 1,145 studies reported in 366 publications. These traits span primary obesity-related phenotypes, including BMI, BMI-adjusted waist-hip ratio (WHR), BMI-adjusted hip circumference, BMI-adjusted waist circumference, body weight as well as adiposity-associated traits including body fat percentage, visceral adipose tissue mass, lean body mass, birth weight, and obesity-related metabolic traits (**Supplemental Table 1**). The second resource comprised UK Biobank summary statistics from the Neale Lab Round 2 analysis, contributing 51,998 genome-wide significant variants for BMI (Field 21001) and 50,624 for BMI by bioelectrical impedance (Field 23104)[32]. In

total, the portal indexes 166,582 variant-trait associations across the three contributing datasets.

The pharmacological layer comprises a manually curated database of 49 clinical-stage obesity therapeutics (**Table 1**). These include six FDA-approved agents (Semaglutide, Tirzepatide, Orforglipron, Liraglutide, Setmelanotide and Metreleptin), three regionally approved agents in China (Mazdutide, Ecnoglutide and Beinaglutide), one compound under FDA review (CagriSema), thirteen phase 1, thirteen phase 2, eleven phase 3 candidates, and two withdrawn drugs (Rimonabant and Lorcaserin). The pipeline spans multiple therapeutic modalities including GLP-1 receptor mono-agonists, dual GLP-1/GIP agonists (Tirzepatide, MariTide, Enicepatide, VK2735, Olatorepatide, THDBH120), dual GLP-1/glucagon agonists (Survodutide, Mazdutide, Pemvidutide), the triple GLP-1/GIP/glucagon agonist Retatrutide, amylin-based therapies (Zenagamtide, Eloralintide, Cagrilintide, Petrelintide), oral non-peptide GLP-1 agonists (Orforglipron, Aleniglipron), myostatin/activin pathway inhibitors (Taldefgrobep alfa, Bimagrumab), a peripherally restricted CB1 inverse agonist (Monlunabant) and a centrally acting CB1 inverse agonist (Rimonabant; withdrawn). Each entry is annotated with the molecular target, mechanism of action, route of administration, development stage, pivotal trial identifier, maximum reported weight-loss efficacy, and strength of supporting GWAS evidence, with external links to DrugBank[33] and ChEMBL[34] where available.

The knowledge layer consists of a Retrieval-Augmented Generation (RAG) system built from 10,198 peer-reviewed publications relevant to obesity genetics, incretin pharmacology, adipose biology, and pharmacogenomics, segmented into 150,943 text segments and embedded as 384-dimensional vectors using the all-MiniLM-L6-v2[35] sentence transformer, indexed with FAISS for sub-second similarity search. This architecture enables the portal's AI query system to ground its outputs in peer-reviewed, domain-specific evidence rather than relying on the general training data of LLMs

which have a high hallucination rate.

The portal provides three primary functional modules: (i) an interactive GWAS browser supporting gene and variant search, regional visualisation, and AI-powered report generation; (ii) a pharmacogenomic drug discovery module integrating a curated therapeutic pipeline with genomic target prioritisation and a filtered GPCR target list; and (iii) an AI research assistant providing literature-grounded, scope-restricted interpretation of genetic and pharmacological queries through RAG (**Fig. 1B**).

Gene/Variant level Browser for genetic association exploration

The Gene/Variant Browser provides an interactive interface for exploring genetic associations across three datasets. Users can select from GWAS (default), UK Biobank BMI, UK Biobank BMI Impedance and further apply trait-level filters to focus on specific obesity-related phenotypes.

Searching by gene name or variant rsID provides access to multiple complementary analytical views (**Supplemental Fig. 1A**): a sortable, filterable association table with downloadable CSV export and direct links to PubMed records and GWAS Catalog study accessions; Manhattan plot and *LocusZoom* regional association plots displaying variant distribution across the chromosome and enabling fine-resolution inspection of local linkage disequilibrium (LD) architecture[36]; protein-protein interaction networks from STRING[37] with external links to Reactome[38], KEGG[39], and GeneCards[40] for pathway-level contextualisation; the AI Insight tab generates a literature-grounded summary for any selected gene, synthesising biological function, therapeutic relevance and key findings from the indexed corpus. The Publication Report tab compiles the gene's variants, AI summary, and external resource links into a formatted PDF document for offline use or sharing.

We illustrate the browser using the *FTO* gene, one of the most extensively characterised obesity susceptibility genes (**Fig. 2A**). Querying *FTO* under the default GWAS returns 500 unique genome-wide significant variants on chromosome 16, spanning positions 53,750,466 - 54,117,066 and encompassing multiple distinct obesity-related traits. The association table enables sorting by *P* value, chromosomal position, or effect size and the full result set is downloadable for downstream analysis. The *LocusZoom* regional plot provides fine-resolution LD colouring for the *FTO* locus, enabling researchers to distinguish independent signals from those attributable to LD structure (**Fig. 2B**). The STRING interaction network contextualises *FTO* within its functional neighbourhood, revealing its central role in the m^{6}A RNA methylation complex[41] through interactions with writers (METTL3, METTL14)[42], erasers (ALKBH5), and readers (YTHDF1, YTHDF2, IGF2BP2)[43], as well as connections to the established obesity-associated genes *MC4R* and *TMEM18* (**Fig. 2C**). When a queried gene is a known drug target, a cross-linking banner appears below the gene header, displaying all clinical-stage compounds targeting that gene. For example, selecting *GLP1R* reveals 30 clinical-stage compounds in the curated database that act on this target, including Semaglutide, Tirzepatide, Orforglipron, and Retatrutide, providing immediate pharmacological context alongside the genetic evidence (**Fig. 2D**). This bidirectional link between genetic and pharmacological evidence allows researchers to move directly from a GWAS signal to the corresponding therapeutic landscape. Querying *FTO* in the AI Insight tab produces a literature-grounded summary covering its role as a 2-oxoglutarate-dependent nucleic acid demethylase, GWAS association profile across multiple adiposity traits, molecular mechanisms linking *FTO* to appetite regulation and energy homeostasis and therapeutic potential including the identification of entacapone as a chemical inhibitor acting through FOXO1 pathways, based on retrieved source publications with PubMed links (**Supplemental Fig. 1B**).

Drug target database and pharmacogenomic analysis

The Drug Discovery module provides a comprehensive analytical interface for the obesity therapeutic pipeline. The Pipeline Overview ranks reported compounds by maximum reported weight-loss efficacy, coloured by development stage (**Fig. 3A**). The Drug Profiles tab provides a filterable database with detailed compound information including mechanism of action, administration route, pivotal trial results and external links to DrugBank[33] and ChEMBL[34] entries, alongside expandable profiles summarising key clinical evidence for each compound (**Supplemental Fig. 2A**).

To demonstrate the portal's pharmacogenomic capabilities, we examined the GLP-1/incretin therapeutic axis, which currently dominates the obesity drug market[44,45]. The majority of the 49 curated compounds target the GLP-1 receptor, either alone (Semaglutide, Liraglutide, Orforglipron) or in combination with the GIP receptor (Tirzepatide, MariTide) or the glucagon receptor (Retatrutide, Survodutide, Mazdutide). Querying these targets reveals a convergence of genetic and pharmacological evidence: *GIPR* harbours multiple genome-wide significant variants representing one of the strongest genetic signals among current drug targets (**Supplemental Fig. 2B)**, while *GLP1R* carries 10 genome-wide significant variants and both show consistent association signals across multiple adiposity traits (**Fig. 2D)**. The Mechanisms & AI tab further demonstrates these findings through an interactive Sankey diagram mapping the flow from drug to target receptor to target tissue to physiological effect, illustrating how incretin-based compounds converge on hypothalamic and hindbrain circuits to suppress appetite and slow gastric emptying through partially overlapping but pharmacologically distinct receptor pathways (**Fig. 3B)**. This integrated genetic, pharmacological, and mechanistic view enables researchers to assess the depth of evidence supporting each therapeutic modality within the incretin axis.

Genomic target prioritisation engine

Beyond exploring individual genes, the portal provides a systematic framework for

novel target discovery through the genomic target prioritisation (GTP) method (**Supplemental Fig. 3A**). The GTP score combines two orthogonal components: a statistical significance score $S(g)$ derived from the strongest association observed across all variants mapped to gene (g) and a variant enrichment score $V(g)$ reflecting the total number of genome-wide significant variants annotated to that gene into a single composite metric scaled from 0 to 100 (**Methods**). To focus discovery on candidates without existing pharmacological tools, the algorithm excludes 12 genes that are already targets of clinical-stage obesity therapeutics in our curated database and filters out multi-gene blocks and intergenic variants to ensure single-gene specificity (**Methods**).

The resulting target landscape (**Supplemental Fig. 3B**) plots all scored novel candidate genes by variant density against maximum statistical significance, with bubble size and colour proportional to the GTP score. *FTO* ranks top overall, followed by *RFLNA* and *COBLL1*. Several genes with established biological relevance to obesity and metabolism appear within the top-ranked candidates. *TCF7L2*, encoding a transcription factor central to pancreatic β-cell function, is one of the strongest type 2 diabetes susceptibility genes[46-48]. *ADCY3*, encoding adenylyl cyclase 3, harbours both common-variant GWAS signals and rare loss-of-function variants linked to monogenic obesity[49,50]. *TBX15* is a transcription factor regulating adipose tissue development and depot-specific fat distribution[51]. *FAIM2* is a neural gene previously implicated in appetite regulation through GWAS of childhood obesity[52,53]. The recovery of these biologically validated genes demonstrates the algorithm's capacity to prioritise meaningful therapeutic candidates from genetic evidence alone. For any prioritised gene, the portal generates an AI viability report summarising biological function, druggability assessment, tissue expression, existing tool compounds and key challenges for drug development, grounded in evidence retrieved from the indexed literature (**Supplemental Fig. 3C**).

Given that GPCRs represent the most druggable and therapeutically successful protein class in obesity pharmacology, a dedicated GWAS GPCRs tab cross-references obesity-associated genes against the International Union of Basic and Clinical Pharmacology (IUPHAR) Guide to Pharmacology curated list of 357 human GPCRs, identifying those reaching genome-wide significance for obesity traits[54]. This analysis reveals 55 GPCRs with genome-wide significant associations and a further 8 at suggestive significance ($P < 5 \times 10^{-5}$). These genes include established drug targets such as *GIPR*, *MC4R*, and *CALCR* alongside orphan receptors such as *GPR61*, *GPR139*, and *GHSR* (**Supplemental Table 2**). Among the GPCRs reaching genome-wide significance, *GPR61* ($P = 2 \times 10^{-53}$) and *GPR139* ($P = 3 \times 10^{-26}$) stand out as orphan receptors with emerging evidence of impacts on energy homeostasis but currently lacking clinical-stage compounds[55-57].

AI-powered research assistant

The portal incorporates a Retrieval-Augmented Generation (RAG) system to provide literature-grounded interpretation of obesity genetics and pharmacology questions (**Fig. 4A**). When users submit queries through any of the portal's AI interfaces, including the gene insight tab in the Gene/Variant Browser, the drug analysis module in Drug Discovery, the viability report function in the Novel Target Engine and the dedicated AI Assistant page, the system encodes the query using the all-MiniLM-L6-v2 sentence transformer, retrieves the most relevant text segments from a FAISS index built from 150,943 segments across 10,198 publications, and provides these passages alongside any structured portal data as context to a large language model (Claude Sonnet 4.6, Anthropic), which generates a response grounded in the retrieved evidence. For example, beyond genetic associations identified at the FTO locus, users can further investigate how these effects are mediated through downstream genes, such as *IRX3* and *IRX5*, using the AI assistant. This enables deeper mechanistic insights into gene regulation and disease pathogenesis (**Supplemental Fig. 4A**).

Each response is accompanied by a list of source publications retrieved from the indexed corpus, with direct links to PubMed Central, allowing users to trace claims back to the underlying literature.

A key feature of the AI-powered assistant is its domain-aware query understanding. Before any retrieval is performed, incoming questions are evaluated against the portal's scientific scope using a two-stage filter. A keyword-based classifier first screens for obesity-relevant terminology such as gene symbols, drug names, metabolic phenotypes, and genetic concepts, unrecognised queries are then indexed against the RAG corpus with a high similarity threshold to detect in-scope questions. Queries that are outside the portal's scope, such as requests about weather, travel, general coding, or unrelated medical conditions, are declined with a redirection message rather than forced through the retrieval pipeline (**Fig. 4B**). This design prevents the language model from fabricating answers grounded in spurious context, a known failure mode of general-purpose chatbots when confronted with out-of-domain prompts. For in-scope queries, the system applies adaptive retrieval strategies to handle short or ambiguous inputs. When a query is too brief to provide sufficient semantic context, for example, a simple gene symbol such as *FTO*, the system automatically generates several expanded variants by appending domain terms such as '*FTO* obesity genetics' or '*FTO* BMI GWAS', retrieves passages for each variant, and returns results from the highest-scoring version. Together, these mechanisms yield an AI research assistant that behaves as a focused domain expert, restricted in scope, grounded in retrieved evidence and transparent about its sources rather than an unconstrained conversational agent.

Discussion

Human genetics has transformed our understanding of obesity, however, the translation of this evidence into therapeutic discovery has been restricted to specific drugs targeting monogenic issues – for example, Setmelanotide, an *MC4R* agonist used to treat patients

with rare biallelic mutations in *POMC*[58]. Since the identification of *FTO* in 2007[6], GWAS have reported thousands of loci associated with BMI and related obesity traits. In parallel, the clinical success of incretin-based drugs has demonstrated that targets supported by human genetic evidence can achieve substantial weight loss[59]. Despite the link between genetic discovery and pharmacological outcome, methods to bridge these remain limited. GWAS catalogues provide variant-level summaries but lack pharmacological context, drug target databases such as DrugBank[33] and Open Targets[30,60] contain extensive pharmacological annotations but are not ideal for the genetic architecture of obesity, and clinical trial registries capture development stage without linking to the underlying human genetics. The Obesity Knowledge Portal addresses this clear gap providing integrated access to 166,582 variant-trait associations with a manually curated database of 49 clinical-stage obesity compounds, a genomic target prioritisation method, and a literature-based AI assistant indexed 10,198 publications.

Notably, of the 49 clinical-stage compounds curated in the portal, the majority act on the GLP-1/incretin axis, either as monotherapies directed at the GLP-1 receptor (Semaglutide, Liraglutide, Orforglipron), as dual agonists co-targeting the GIP receptor (Tirzepatide, MariTide) or the glucagon receptor (Survodutide, Mazdutide), or as a triple GLP-1/GIP/glucagon agonist (Retatrutide). The remaining compounds are distributed across *MC4R*, leptin, endocannabinoid, amylin, myostatin and *GPR119* pathways[45]. Incretin-based therapies have shown strong efficacy in clinical trials, with semaglutide achieving approximately 16% mean weight reduction[19,20], Tirzepatide 23%[21] and Retatrutide 29%[22]. However, it is well-known that human genetic evidence base is considerably broader than this incretin axis reflects and our gene prioritisation algorithms also identify multiple high-scoring genes with established roles in energy homeostasis which are not currently targeted by any clinical-stage compound. This contrast between the breadth of genetic evidence and the narrowness of current therapeutic development highlights the value of integrative resources that can

systematically connect GWAS findings to pharmacological opportunities beyond the incretin axis.

The genomic target prioritisation analysis in the portal further provides a framework for examining how current therapeutic development aligns with the broader landscape of human genetic evidence for obesity. Several of the highest-scoring candidates are well-characterised genes associated with metabolic traits for which no reported clinical-stage obesity therapeutic is currently in development. *TCF7L2*, the strongest locus for T2D and a central regulator of pancreatic β-cell function[46-48], is not currently under an obesity-directed development programme despite its strong and reproducible GWAS signal. *ADCY3* is of particular interest, as it harbours both common-variant GWAS signals and rare loss-of-function alleles linked to monogenic obesity[49,50], indicating a convergence of evidence that supports hypothalamic cAMP signalling[57] as a mechanistically potential axis for therapeutic exploration. The high ranking of *TBX15*, a transcription factor regulating adipocyte development and depot-specific fat distribution[51], suggests adipose tissue requires more pharmacological attention apart from the predominantly central nervous system-focused therapeutic development. *FAIM2*, implicated in appetite regulation through GWAS of early-onset obesity, suggests the involvement of neural circuits distinct from the hypothalamic melanocortin and incretin pathways[53]. 34% of the FDA-approved drugs target GPCRs[61,62], applying the GPCR filter prioritised *GPR61* and *GPR139*, two orphan receptors with genome-wide significant associations, which exhibits a strong genome-wide significant association signal are not being evaluated currently in clinical-stage obesity trials. More recently, multi-target combinations also suggest the success of 'incretin +', genes with strong genetic contribution may also facilitate this multiple targets combination[63]. Taken together, these results suggest that the portal can help researchers to evaluate the strength of investigation, and guide the translation of these findings towards pharmacological evaluation and drug development.

At a broader level, the AI assistant integrated into the portal exemplifies how large language models can be adapted responsibly to support the understanding of obesity genetics and pathology. By constraining generation to a curated corpus of obesity literature, applying a domain-specific scope filter, and returning every retrieved source alongside its response, the assistant provides a form of literature synthesis that is transparent, traceable, and grounded in evidence that users can easily independently verify. This architecture offers a template for similar domain-focused scientific assistants in other disease areas[64,65]. Positioning such an assistant within an integrated genetic and pharmacological environment, rather than as a standalone conversational interface, further enhances its utility by allowing researchers to move fluidly between structured data analysis and narrative literature interpretation.

By bringing together human genetic evidence, pharmacological context, and literature-grounded analysis within a single integrated environment, the Obesity Knowledge Portal is intended to serve as a shared human genetics and pharmacotherapy resource for the research community working on metabolism and obesity. This portal will support hypothesis generation, accelerate target evaluation, and facilitate communication between the genetics, pharmacology, and clinical research communities, thereby contributing to the broader translational effort of converting genetic discovery into effective therapeutic intervention for obesity and its associated metabolic diseases.

Methods

Data sources and integration

The portal integrates three primary genetic datasets covering obesity and adiposity traits. GWAS Catalog full association data (version 1.2, April 2026) were downloaded from the NHGRI-EBI GWAS Catalog and filtered to retain variant-trait associations relevant

to obesity, BMI, body fat distribution, and related adiposity phenotypes. Variants without valid gene annotations were removed, yielding 63,960 variant-trait associations mapped to 11,298 genes across 661 traits and 1,145 studies. UK Biobank summary statistics were obtained from the Neale Lab Round 2 release for Field 21001 (body mass index; 51,998 genome-wide significant variants) and Field 23104 (BMI by impedance; 50,624 variants)[32]. The combined dataset comprises 166,582 variant-trait associations.

Drug target database

The drug target database was manually curated from primary literature, ClinicalTrials.gov and U.S. Food and Drug Administration (FDA) records, comprising 49 clinical-stage compounds on 11 unique molecular targets. Each entry includes drug name, molecular target, mechanism of action, pharmacological class, administration route, development stage, sponsoring company, pivotal clinical trial, maximum reported weight loss efficacy, and external database identifiers (DrugBank[33] and ChEMBL[34]) if available. Clinical efficacy values correspond to the maximum reported mean percentage body weight reduction at the highest evaluated dose from the referenced clinical trial for each compound. For compounds in earlier development stages, values reflect the best available published data and may derive from shorter-duration or dose-finding studies. Compounds approved for non-obesity indications (metreleptin) or without reported weight-loss endpoints (taldefgrobep alfa, macupatide) are not listed in the barchart, but kept in the drug profile database. Entries are manually updated as new clinical data become available.

Genomic target prioritisation (GTP) score

The GTP score integrates two normalised components of human genetic evidence: statistical significance and variant density. For each candidate gene g, $P_{min}(g)$ denotes the smallest P value observed across all variants mapped to g, and $N(g)$ denotes the total number of variants mapped to g that reach genome-wide significance.

A capped log-transformed significance score is computed as

$$\boldsymbol{S}(g) = \min(-\log_{10} P_{min}(g), 100)/\max_g \{\min(-\log_{10} P_{min}(g), 100)\}$$

where the cap of 100 limits the influence of extreme outliers.

A log-transformed variant density score is computed as

$$\boldsymbol{V}(g) = \log_2(N(g) + 1) / \max_g \{\log_2(N(g) + 1)\}$$

Both components are normalised to [0, 1] using the maximum value across all scored candidate genes. The composite GTP score is the equally weighted average scaled to a 0–100 range:

$$GTP_{(g)} = \frac{[S_{(g)} + S_{(g)}]}{2} * 100$$

We exclude 12 genes that are already targets of clinical-stage obesity therapeutics in the curated drug database (*GLP1R*, *GIPR*, *GCGR*, *MC4R*, *ACVR2B*, *CNR1*, *HTR2C*, *LEP*, *LEPR*, *CALCR*, *MSTN* and *IGF1R*) to focus on novel candidates. Multi-gene blocks entries and intergenic variants are filtered prior to scoring to ensure single-gene specificity. Equal weighting of the two components is used.

Literature corpus and RAG index

The portal incorporates a retrieval-augmented generation system indexed over 10,198 obesity-relevant peer-reviewed publications (465 full-text PDFs and 9,733 text-extracted articles) obtained via PubMed Central. Documents were chunked into overlapping text segments and embedded using the all-MiniLM-L6-v2 sentence transformer (384-dimensional embeddings), producing a FAISS index of 150,943 vectors. At query time, the system encodes the user question using the same embedding model and retrieves the top-ranked segments by cosine similarity. The top segments are passed as context to the generative model (Claude Sonnet 4.6, Anthropic) along with any structured portal data relevant to the query, and the model generates a response grounded in the retrieved passages. All unique source documents corresponding to the

retrieved chunks are returned to the user interface and displayed as a references list with clickable links to PubMed Central.

Domain-aware query in the AI assistant

To confine the AI assistant to obesity-relevant topics, incoming queries are evaluated by a two-stage filter before retrieval is performed. A keyword-based classifier first screens the query against a curated list of obesity, genetics, and pharmacology terms, as well as a second list of out-of-scope phrases such as requests about weather, travel, or unrelated medical conditions. Queries that fail keyword matching are then probed against the RAG corpus at a high similarity threshold (cosine similarity ≥ 0.55) to detect in-scope questions and queries that still fail this flow are declined with a redirection message. For in-scope queries, the system applies adaptive retrieval strategies to handle short and ambiguous inputs. When the original query scores below a confidence threshold or consists of only a few tokens (for example, a single gene symbol) or retrieves passages below a confidence threshold (cosine similarity < 0.60), the system generates several expanded query variants by appending domain terms such as ‘obesity genetics’, ‘BMI GWAS’, or ‘adipose metabolism’ runs each variant against the index and returns the results from the highest-scoring variant.

Portal implementation

The portal is implemented as a Streamlit web application (Python 3.12.3) and deployed on a DigitalOcean virtual machine running Ubuntu 24.04.4 LTS behind an nginx reverse proxy with Let's Encrypt TLS termination. Interactive visualisations use Plotly for Manhattan plots, bubble charts, and Sankey diagrams; *LocusZoom* plots are rendered via the University of Michigan LocusZoom.js API[36]. Protein-protein interaction networks are retrieved from STRING[37]. The GPCR target filter cross-references obesity-associated genes against the IUPHAR/BPS Guide to Pharmacology curated list of 357 human GPCRs[54]. The portal is freely accessible at

508 https://obesityknowledge.org.

509

References

1. Afshin, A., Forouzanfar, M.H., Reitsma, M.B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., Naghavi, M., et al. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. N Engl J Med *377*, 13-27. 10.1056/NEJMoa1614362.
2. Phelps, N.H., Singleton, R.K., Zhou, B., Heap, R.A., Mishra, A., Bennett, J.E., Paciorek, C.J., Lhoste, V.P.F., Carrillo-Larco, R.M., Stevens, G.A., et al. (2024). Worldwide trends in underweight and obesity from 1990 to 2022: a pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. The Lancet *403*, 1027-1050. 10.1016/S0140-6736(23)02750-2.
3. Maes, H.H., Neale, M.C., and Eaves, L.J. (1997). Genetic and environmental factors in relative body weight and human adiposity. Behav Genet *27*, 325-351. 10.1023/a:1025635913927.
4. Loos, R.J.F., and Yeo, G.S.H. (2022). The genetics of obesity: from discovery to biology. Nat Rev Genet *23*, 120-133. 10.1038/s41576-021-00414-z.
5. Silventoinen, K., Rokholm, B., Kaprio, J., and Sørensen, T.I. (2010). The genetic and environmental influences on childhood obesity: a systematic review of twin and adoption studies. Int J Obes (Lond) *34*, 29-40. 10.1038/ijo.2009.177.
6. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889-894. 10.1126/science.1141634.
7. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197-206. 10.1038/nature14177.
8. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., and Visscher, P.M. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet *27*, 3641-3649. 10.1093/hmg/ddy271.
9. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet *28*, 166-174. 10.1093/hmg/ddy327.
10. Huang, J., Huffman, J.E., Huang, Y., Do Valle, Í., Assimes, T.L., Raghavan, S., Voight, B.F., Liu, C., Barabási, A.L., Huang, R.D.L., et al. (2022). Genomics and phenomics of body mass index reveals a complex disease network. Nat Commun *13*, 7973. 10.1038/s41467-022-35553-2.

11. Karczewski, K.J., Gupta, R., Kanai, M., Lu, W., Tsuo, K., Wang, Y., Walters, R.K., Turley, P., Callier, S., Shah, N.N., et al. (2025). Pan-UK Biobank genome-wide association analyses enhance discovery and resolution of ancestry-enriched effects. Nat Genet *57*, 2408-2417. 10.1038/s41588-025-02335-7.
12. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet *42*, 937-948. 10.1038/ng.686.
13. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet *41*, 25-34. 10.1038/ng.287.
14. Timshel, P.N., Thompson, J.J., and Pers, T.H. (2020). Genetic mapping of etiologic brain cell types for obesity. Elife *9*. 10.7554/eLife.55851.
15. Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A.V., et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213. 10.1038/nature24277.
16. Consortium, T.G., Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. doi:10.1126/science.aaz1776.
17. Wu, Y., Rudolf, A.M., Wang, L., Zhang, Z., Niu, C., Xia, F., Li, M., Wang, A., Derous, D., Mitchell, S.E., et al. (2026). The susceptibility to weight gain in male C57BL/6N mice fed high fat diet is due to a small set of genes predominantly not expressed in the brain: I target identification. Proceedings of the National Academy of Sciences, accepted.
18. Markham, A. (2021). Setmelanotide: First Approval. Drugs *81*, 397-403. 10.1007/s40265-021-01470-9.
19. Wilding, J.P.H., Batterham, R.L., Calanna, S., Davies, M., Gaal, L.F.V., Lingvay, I., McGowan, B.M., Rosenstock, J., Tran, M.T.D., Wadden, T.A., et al. (2021). Once-Weekly Semaglutide in Adults with Overweight or Obesity. New England Journal of Medicine *384*, 989-1002. doi:10.1056/NEJMoa2032183.
20. Garvey, W.T., Batterham, R.L., Bhatta, M., Buscemi, S., Christensen, L.N., Frias, J.P., Jódar, E., Kandler, K., Rigas, G., Wadden, T.A., and Wharton, S. (2022). Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. Nat Med *28*, 2083-2091. 10.1038/s41591-022-02026-4.
21. Jastreboff, A.M., Aronne, L.J., Ahmad, N.N., Wharton, S., Connery, L., Alves, B., Kiyosue, A., Zhang, S., Liu, B., Bunck, M.C., and Stefanski, A. (2022). Tirzepatide Once Weekly for the Treatment of Obesity. New England Journal of Medicine *387*, 205-216. doi:10.1056/NEJMoa2206038.
22. Jastreboff, A.M., Kaplan, L.M., Frías, J.P., Wu, Q., Du, Y., Gurbuz, S.,

Coskun, T., Haupt, A., Milicevic, Z., and Hartman, M.L. (2023). Triple–Hormone-Receptor Agonist Retatrutide for Obesity — A Phase 2 Trial. New England Journal of Medicine *389*, 514-526. doi:10.1056/NEJMoa2301972.

23. Giblin, K., Kaplan, L.M., Somers, V.K., Le Roux, C.W., Hunter, D.J., Wu, Q., Lalonde, A., Ahmad, N., and Bethel, M.A. (2026). Retatrutide for the treatment of obesity, obstructive sleep apnea and knee osteoarthritis: Rationale and design of the TRIUMPH registrational clinical trials. Diabetes Obes Metab *28*, 83-93. 10.1111/dom.70209.
24. Gao, W., Liu, L., Huh, E., Gbahou, F., Cecon, E., Oshima, M., Houzé, L., Katsonis, P., Hegron, A., Fan, Z., et al. (2023). Human GLP1R variants affecting GLP1R cell surface expression are associated with impaired glucose control and increased adiposity. Nat Metab *5*, 1673-1684. 10.1038/s42255-023-00889-6.
25. Lagou, V., Jiang, L., Ulrich, A., Zudina, L., González, K.S.G., Balkhiyarova, Z., Faggian, A., Maina, J.G., Chen, S., Todorov, P.V., et al. (2023). GWAS of random glucose in 476,326 individuals provide insights into diabetes pathophysiology, complications and treatment stratification. Nat Genet *55*, 1448-1461. 10.1038/s41588-023-01462-3.
26. Dawed, A.Y., Mari, A., Brown, A., McDonald, T.J., Li, L., Wang, S., Hong, M.G., Sharma, S., Robertson, N.R., Mahajan, A., et al. (2023). Pharmacogenomics of GLP-1 receptor agonists: a genome-wide analysis of observational data and large randomised controlled trials. Lancet Diabetes Endocrinol *11*, 33-41. 10.1016/s2213-8587(22)00340-0.
27. Kizilkaya, H.S., Sørensen, K.V., Madsen, J.S., Lindquist, P., Douros, J.D., Bork-Jensen, J., Berghella, A., Gerlach, P.A., Gasbjerg, L.S., Mokrosiński, J., et al. (2024). Characterization of genetic variants of GIPR reveals a contribution of β-arrestin to metabolic phenotypes. Nat Metab *6*, 1268-1281. 10.1038/s42255-024-01061-4.
28. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res *51*, D977-d985. 10.1093/nar/gkac1010.
29. Costanzo, M.C., von Grotthuss, M., Massung, J., Jang, D., Caulkins, L., Koesterer, R., Gilbert, C., Welch, R.P., Kudtarkar, P., Hoang, Q., et al. (2023). The Type 2 Diabetes Knowledge Portal: An open access genetic resource dedicated to type 2 diabetes and related traits. Cell Metab *35*, 695-710.e696. 10.1016/j.cmet.2023.03.001.
30. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda, A., Cruz-Castillo, C., Fumis, L., et al. (2023). The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res *51*, D1353-d1359. 10.1093/nar/gkac1046.
31. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. Nature Medicine *29*, 1930-1940. 10.1038/s41591-023-02448-8.

32. Lab, N. (2018). UK Biobank — Neale Lab. http://www.nealelab.is/uk-biobank/.
33. Knox, C., Wilson, M., Klinger, C.M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E.L., Strawbridge, S.A., et al. (2024). DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res *52*, D1265-d1275. 10.1093/nar/gkad976.
34. Zdrazil, B., Felix, E., Hunter, F., Manners, E.J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Lopez, D.M., Mosquera, Juan F., et al. (2023). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Research *52*, D1180-D1192. 10.1093/nar/gkad1004.
35. Reimers, N., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. pp. 3982-3992.
36. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336-2337. 10.1093/bioinformatics/btq419.
37. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res *51*, D638-d646. 10.1093/nar/gkac1000.
38. Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022). The reactome pathway knowledgebase 2022. Nucleic Acids Res *50*, D687-d692. 10.1093/nar/gkab1028.
39. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res *51*, D587-d592. 10.1093/nar/gkac963.
40. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics *54*, 1.30.31-31.30.33. 10.1002/cpbi.5.
41. Jia, G., Fu, Y., Zhao, X., Dai, Q., Zheng, G., Yang, Y., Yi, C., Lindahl, T., Pan, T., Yang, Y.G., and He, C. (2011). N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. Nat Chem Biol *7*, 885-887. 10.1038/nchembio.687.
42. Zhao, X., Yang, Y., Sun, B.F., Shi, Y., Yang, X., Xiao, W., Hao, Y.J., Ping, X.L., Chen, Y.S., Wang, W.J., et al. (2014). FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. Cell Res *24*, 1403-1419. 10.1038/cr.2014.151.
43. Merkestein, M., Laber, S., McMurray, F., Andrew, D., Sachse, G., Sanderson, J., Li, M., Usher, S., Sellayah, D., Ashcroft, F.M., and Cox, R.D. (2015). FTO influences adipogenesis by regulating mitotic clonal expansion. Nat Commun *6*, 6792. 10.1038/ncomms7792.

685 44. Drucker, D.J. (2022). GLP-1 physiology informs the pharmacotherapy of
686 obesity. Mol Metab *57*, 101351. 10.1016/j.molmet.2021.101351.
687 45. Müller, T.D., Blüher, M., Tschöp, M.H., and DiMarchi, R.D. (2022). Anti-
688 obesity drug discovery: advances and challenges. Nat Rev Drug Discov *21*,
689 201-223. 10.1038/s41573-021-00337-8.
690 46. Grant, S.F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu,
691 A., Sainz, J., Helgason, A., Stefansson, H., Emilsson, V., Helgadottir, A., et al.
692 (2006). Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of
693 type 2 diabetes. Nat Genet *38*, 320-323. 10.1038/ng1732.
694 47. Florez, J.C., Jablonski, K.A., Bayley, N., Pollin, T.I., de Bakker, P.I.,
695 Shuldiner, A.R., Knowler, W.C., Nathan, D.M., and Altshuler, D. (2006).
696 TCF7L2 polymorphisms and progression to diabetes in the Diabetes
697 Prevention Program. N Engl J Med *355*, 241-250. 10.1056/NEJMoa062418.
698 48. Mahajan, A., Spracklen, C.N., Zhang, W., Ng, M.C.Y., Petty, L.E., Kitajima,
699 H., Yu, G.Z., Rüeger, S., Speidel, L., Kim, Y.J., et al. (2022). Multi-ancestry
700 genetic study of type 2 diabetes highlights the power of diverse populations
701 for discovery and translation. Nat Genet *54*, 560-572. 10.1038/s41588-022-
702 01058-3.
703 49. Saeed, S., Bonnefond, A., Tamanini, F., Mirza, M.U., Manzoor, J., Janjua,
704 Q.M., Din, S.M., Gaitan, J., Milochau, A., Durand, E., et al. (2018). Loss-of-
705 function mutations in ADCY3 cause monogenic severe obesity. Nat Genet *50*,
706 175-179. 10.1038/s41588-017-0023-6.
707 50. Grarup, N., Moltke, I., Andersen, M.K., Dalby, M., Vitting-Seerup, K., Kern, T.,
708 Mahendran, Y., Jørsboe, E., Larsen, C.V.L., Dahl-Petersen, I.K., et al. (2018).
709 Loss-of-function variants in ADCY3 increase risk of obesity and type 2
710 diabetes. Nat Genet *50*, 172-174. 10.1038/s41588-017-0022-7.
711 51. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E.,
712 Mägi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., et al.
713 (2015). New genetic loci link adipose and insulin biology to body fat
714 distribution. Nature *518*, 187-196. 10.1038/nature14132.
715 52. Bradfield, J.P., Taal, H.R., Timpson, N.J., Scherag, A., Lecoeur, C.,
716 Warrington, N.M., Hypponen, E., Holst, C., Valcarcel, B., Thiering, E., et al.
717 (2012). A genome-wide association meta-analysis identifies new childhood
718 obesity loci. Nat Genet *44*, 526-531. 10.1038/ng.2247.
719 53. Littleton, S.H., Trang, K.B., Volpe, C.M., Cook, K., DeBruyne, N., Maguire,
720 J.A., Weidekamp, M.A., Hodge, K.M., Boehm, K., Lu, S., et al. (2024).
721 Variant-to-function analysis of the childhood obesity chr12q13 locus
722 implicates rs7132908 as a causal variant within the 3' UTR of FAIM2. Cell
723 Genom *4*, 100556. 10.1016/j.xgen.2024.100556.
724 54. Harding, S.D., Armstrong, J.F., Faccenda, E., Southan, C., Alexander, S.P.H.,
725 Davenport, A.P., Spedding, M., and Davies, J.A. (2024). The IUPHAR/BPS
726 Guide to PHARMACOLOGY in 2024. Nucleic Acids Res *52*, D1438-d1449.
727 10.1093/nar/gkad944.
728 55. Tsang, C.H., De Rosa, A., and Kozielewicz, P. (2025). Identification and

molecular characterization of missense mutations in orphan G protein-coupled receptor GPR61 occurring in severe obesity. Mol Pharmacol *107*, 100026. 10.1016/j.molpha.2025.100026.

56. Nogueira, P.A.S., Moura-Assis, A., Zanesco, A.M., Bombassaro, B., Gallo-Ferraz, A.L., Simões, M.R., Engel, D.F., Razolli, D.S., Gaspar, J.M., Junior, J.D., and Velloso, L.A. (2023). The orphan G protein-coupled receptor, GPR139, is expressed in the hypothalamus and is involved in the regulation of body mass, blood glucose, and insulin. Neurosci Lett *792*, 136955. 10.1016/j.neulet.2022.136955.

57. Felix, J.F., Bradfield, J.P., Monnereau, C., van der Valk, R.J., Stergiakouli, E., Chesi, A., Gaillard, R., Feenstra, B., Thiering, E., Kreiner-Møller, E., et al. (2016). Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum Mol Genet *25*, 389-403. 10.1093/hmg/ddv472.

58. Clément, K., van den Akker, E., Argente, J., Bahm, A., Chung, W.K., Connors, H., De Waele, K., Farooqi, I.S., Gonneau-Lejeune, J., Gordon, G., et al. (2020). Efficacy and safety of setmelanotide, an MC4R agonist, in individuals with severe obesity due to LEPR or POMC deficiency: single-arm, open-label, multicentre, phase 3 trials. Lancet Diabetes Endocrinol *8*, 960-970. 10.1016/s2213-8587(20)30364-8.

59. Bonnefond, A., Bruner, W.S., Grant, S.F., Morandi, A., and Froguel, P. (2026). The genetics of obesity: aetiology, prevention and therapy. Nature Metabolism, 1-17.

60. Ghoussaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. (2021). Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic Acids Res *49*, D1311-d1320. 10.1093/nar/gkaa840.

61. Lorente, J.S., Sokolov, A.V., Ferguson, G., Schiöth, H.B., Hauser, A.S., and Gloriam, D.E. (2025). GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov *24*, 458-479. 10.1038/s41573-025-01139-y.

62. Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schiöth, H.B., and Gloriam, D.E. (2017). Trends in GPCR drug discovery: new agents, targets and indications. Nat Rev Drug Discov *16*, 829-842. 10.1038/nrd.2017.178.

63. Liskiewicz, D., Novikoff, A., Khalil, A., Akindehin, S., Campbell, J.E., Candela, P., Castelino, R.L., Coupland, C., Culot, M., Dodson, W.S., et al. (2026). GLP-1R–GIPR–PPARα/γ/δ quintuple agonism corrects obesity and diabetes in mice. Nature *653*, 776-785. 10.1038/s41586-026-10427-5.

64. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., and Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems *33*, 9459-9474.

65. Li, M., Kilicoglu, H., Xu, H., and Zhang, R. (2025). Biomedrag: A retrieval augmented large language model for biomedicine. Journal of Biomedical

773 Informatics *162*, 104769.

774

775

Data availability

The Obesity Knowledge Portal is freely accessible at https://obesityknowledge.org. GWAS Catalog summary statistics used in this study were downloaded from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads). UK Biobank summary statistics for BMI (Field 21001) and BMI by bioelectrical impedance (Field 23104) were obtained from the Neale Lab Round 2 GWAS results (https://www.nealelab.is/uk-biobank). The RAG literature corpus was constructed from publications indexed in PubMed Central (https://www.ncbi.nlm.nih.gov/pmc/).

Acknowledgements

We thank all members of the MultiOmics Lab for their contributions and discussions. We thank the Neale Lab for making UK Biobank GWAS summary results publicly available and the NHGRI-EBI GWAS Catalog team for their ongoing curation efforts. This study was funded by the National Key Research and Development Program of China (2024YFA1802800 and 2024YFA1802803), National Natural Science Foundation of China under Grants (32371195 and W2541011), the Foundation of the Shanghai Municipal Education Commission (24KXZNB10), and a new PI Start-up grant of Fudan University to G.W. Computations were performed using the CFFF platform of Fudan University.

Competing Interests

The authors declare no competing interests.

Author Contributions

Conceptualisation: G.W. & P.L.; Methodology: G.W. & Y.L. & W.X.; Software: G.W. & Y.L.; Formal analysis: G.W. & Y.L.; Investigation: G.W. & Y.L. & B.L. & J.B.;

802 Data curation: G.W. & S.A.; Visualisation: G.W. & Y.L.; Writing, original draft: G.W.
803 & Y.L. & J.R.S.; Writing, review & editing: all authors; Supervision: G.W. & T.Z. &
804 J.R.S.; Funding acquisition: G.W. & J.R.S. All authors have read and approved the final
805 manuscript.
806

Figure1

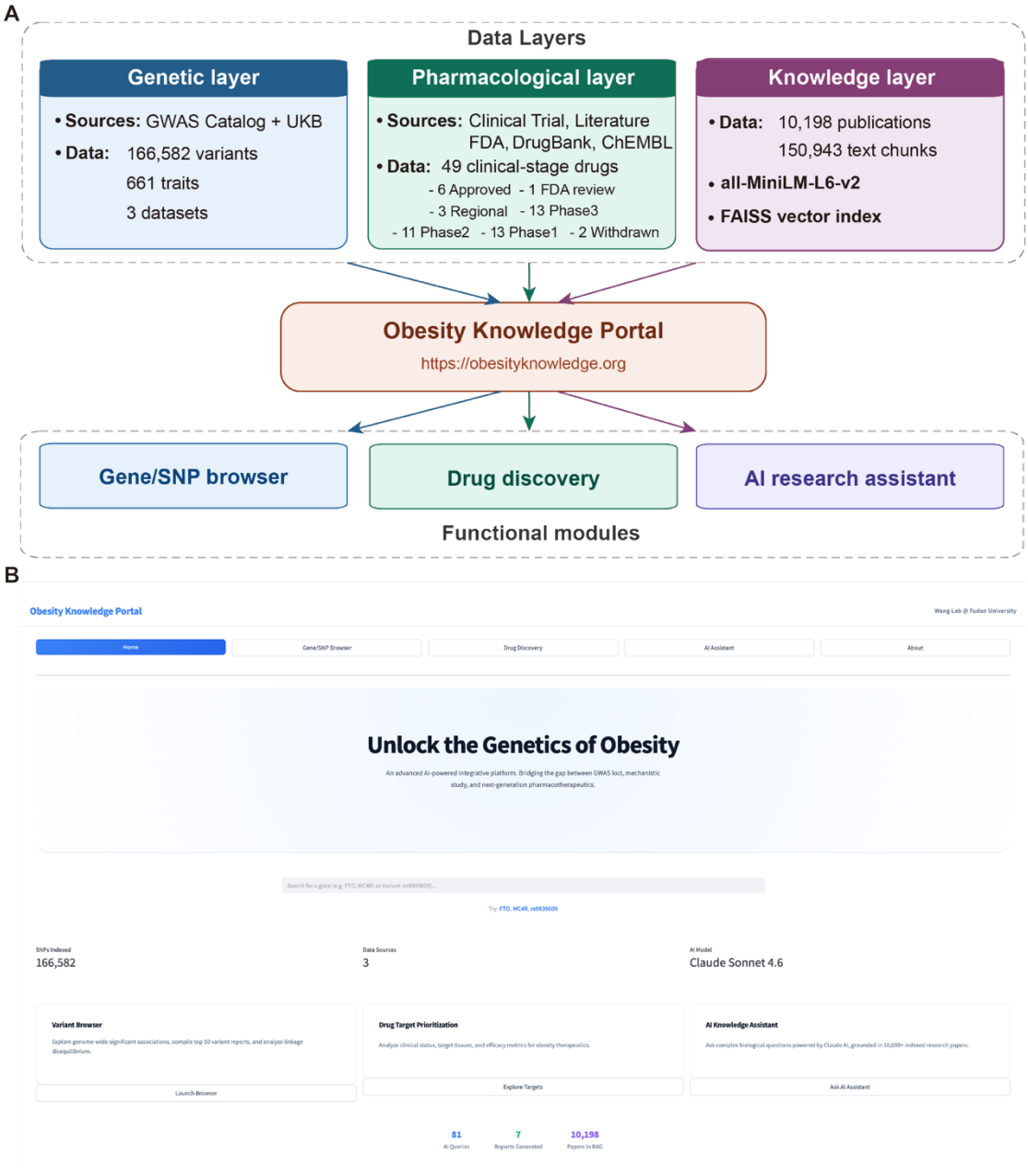


Fig. 1 Overview of the Obesity Knowledge Portal.

(A) Structure of the Obesity Knowledge Portal (OKP). The platform integrates three complementary data layers: a genetic layer built on GWAS Catalog data and UK Biobank data; a pharmacological layer comprising a manually curated database of 49 clinical-stage obesity therapeutics targeting 11 unique molecular targets; and a knowledge layer built from 10,198 domain-specific publications segmented into 150,943 text chunks and indexed using FAISS for retrieval-augmented generation (RAG). These data layers feed into three functional modules: an interactive

816 Gene/Variant Browser, a Drug Discovery module, and an AI Research Assistant
817 delivering literature-grounded, source-cited responses to genetic and pharmacological
818 queries.

819 (B) Home page of the OKP web interface (https://obesityknowledge.org), showing the
820 three functional modules.

821

Figure2

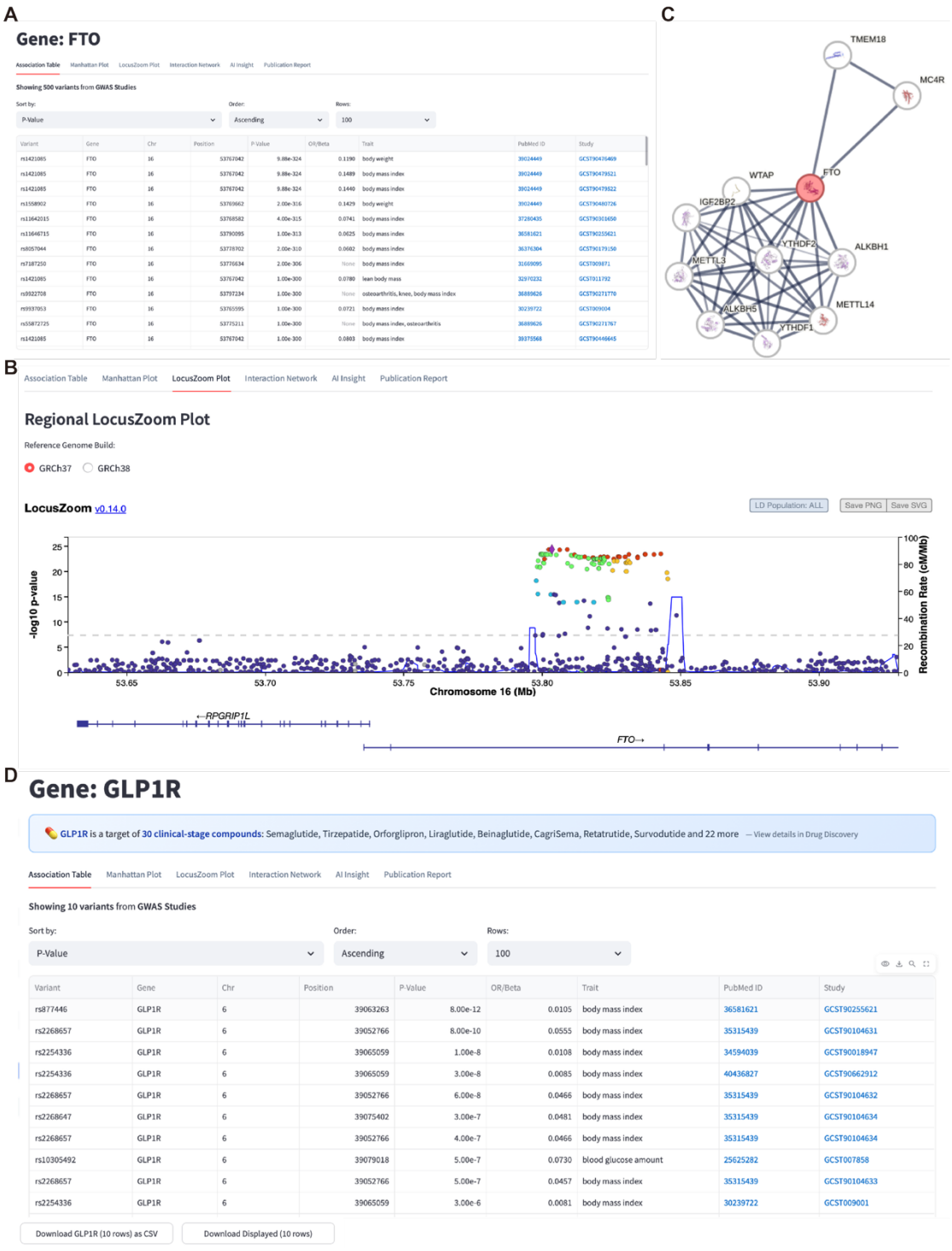


Fig. 2 Gene/Variant Browser illustrated with two examples: *FTO* (A–C) and *GLP1R* (D).

(A) Association table for *FTO* under the default GWAS dataset, returning 500 unique genome-wide significant variants on chromosome 16 (positions 53,750,466–54,117,066) across multiple obesity-related traits. Results are downloadable as CSV.

828 (B) *LocusZoom* regional association plot for the *FTO* locus (GRCh37), displaying
829 $-\log_{10}$(P) values with linkage disequilibrium colouring derived from reference
830 populations and recombination rate overlay. Gene tracks for *FTO* and *RPGRIP1L* are
831 shown below.

832 (C) STRING protein–protein interaction network for the FTO protein, revealing
833 functional interaction partners including m^6A RNA methylation complex members
834 (METTL3, METTL14, ALKBH5, YTHDF1/2) and obesity-associated proteins (MC4R,
835 TMEM18).

836 (D) Cross-linking of genetic and pharmacological evidence for *GLP1R*. The gene page
837 displays 10 genome-wide significant variants alongside a banner listing 30 clinical-
838 stage compounds targeting this receptor.

839

Figure3

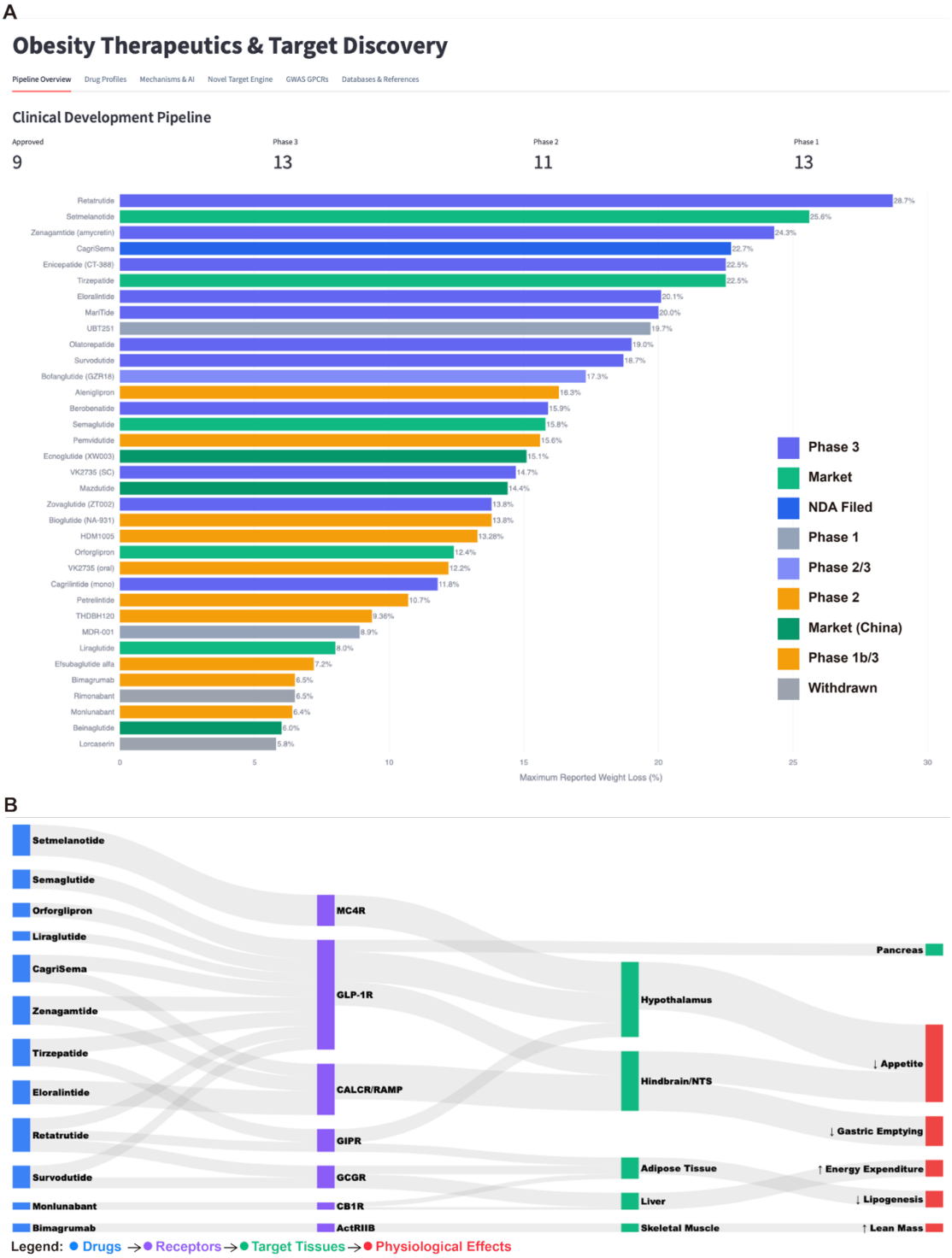


Fig. 3 Drug Discovery module.

(A) Pipeline Overview ranking all 49 clinical-stage obesity compounds by maximum reported weight-loss efficacy (%), colour-coded by development stage.

(B) Mechanisms & AI tab showing an interactive Sankey diagram mapping the flow

from drug compounds (left) through target receptors (*GLP1R*, *GIPR*, *CALCR/RAMP*, *GCGR*, *CB1R*, *ActRIIB*, *MC4R*) to target tissues (hypothalamus, hindbrain/NTS, pancreas, adipose tissue, liver, skeletal muscle) and physiological effects. Mechanistic descriptions of four therapeutic paradigms including central appetite suppression, energy expenditure, body composition preservation, and peripheral metabolic modulation are provided below the diagram.

Figure4

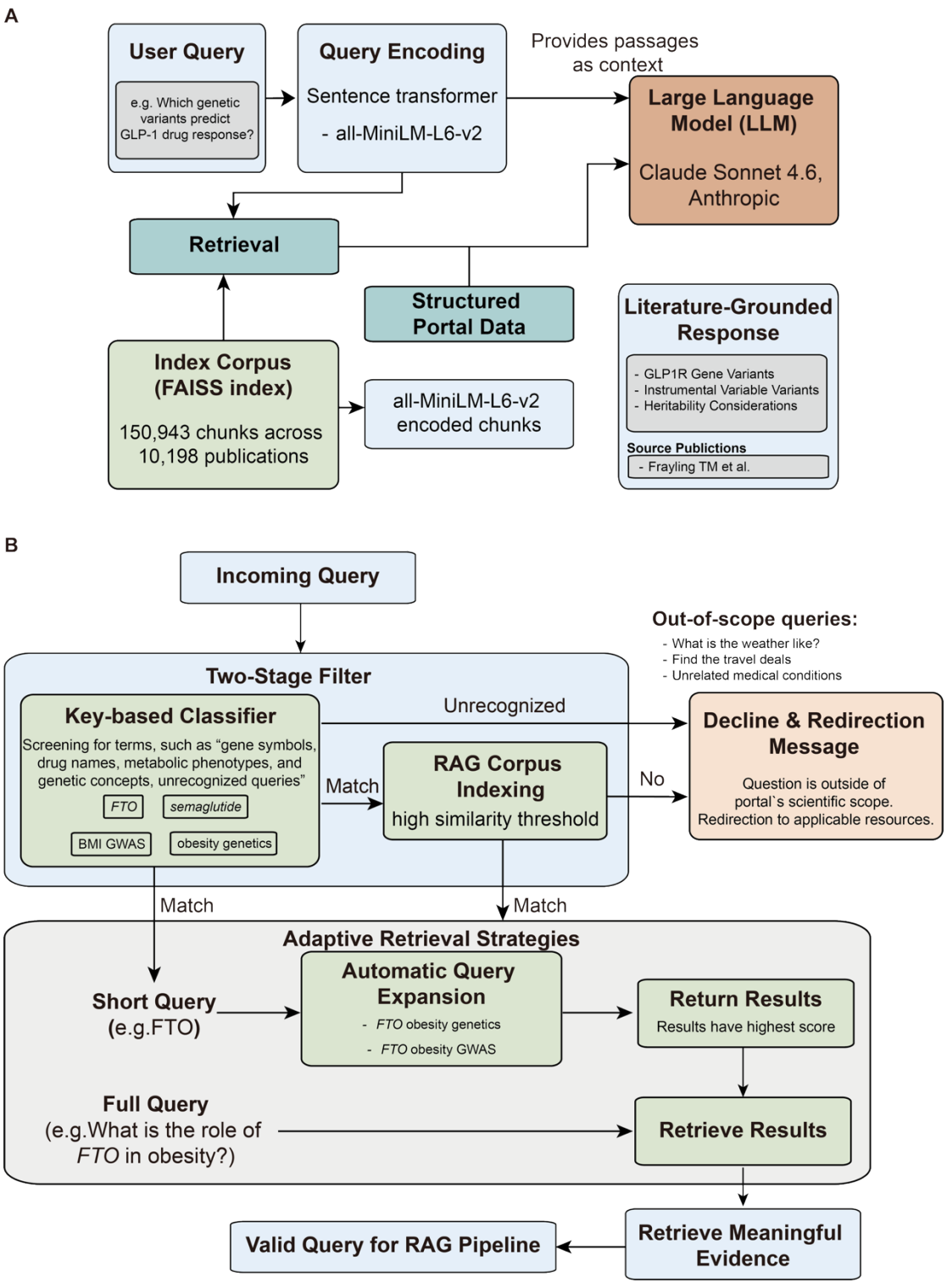


Fig. 4 AI-powered research assistant.

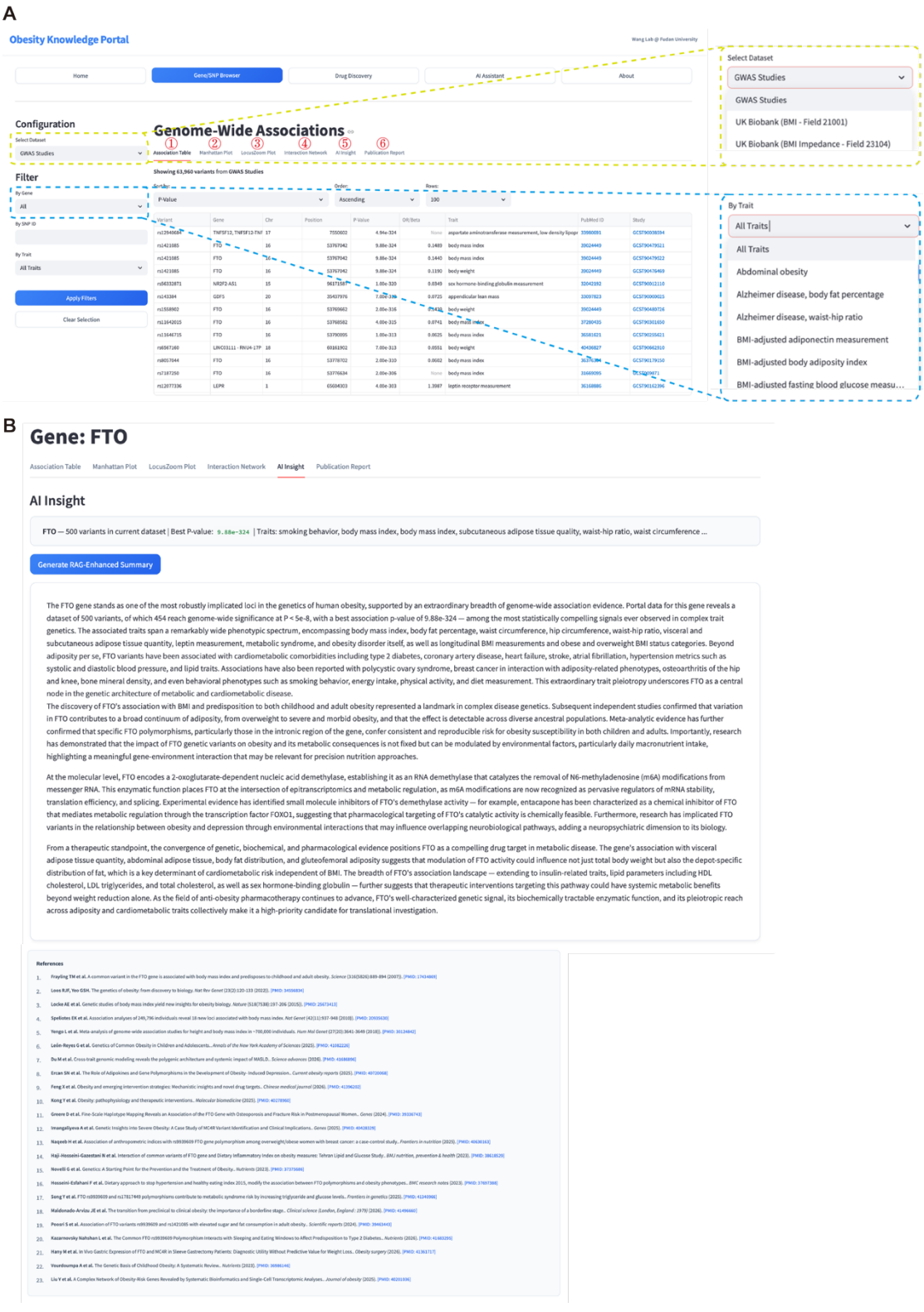
(A) Architecture of the Retrieval-Augmented Generation (RAG) system. User queries are encoded using the all-MiniLM-L6-v2 sentence transformer and matched against a FAISS index of 150,943 text chunks from 10,198 publications. Retrieved passages,

856 together with structured portal data (GWAS associations and drug database), are
857 provided as context to a large language model (Claude Sonnet 4.6, Anthropic), which
858 generates a literature-grounded response accompanied by cited source publications
859 with links to PubMed Central.

860 (B) Query filtering and adaptive retrieval. A two-stage filter (keyword classifier and
861 RAG corpus similarity check) identifies obesity-relevant queries and declines out-of-
862 scope requests with a redirection message. Short or ambiguous inputs undergo
863 automatic query expansion to improve retrieval quality.

864

Fig S1



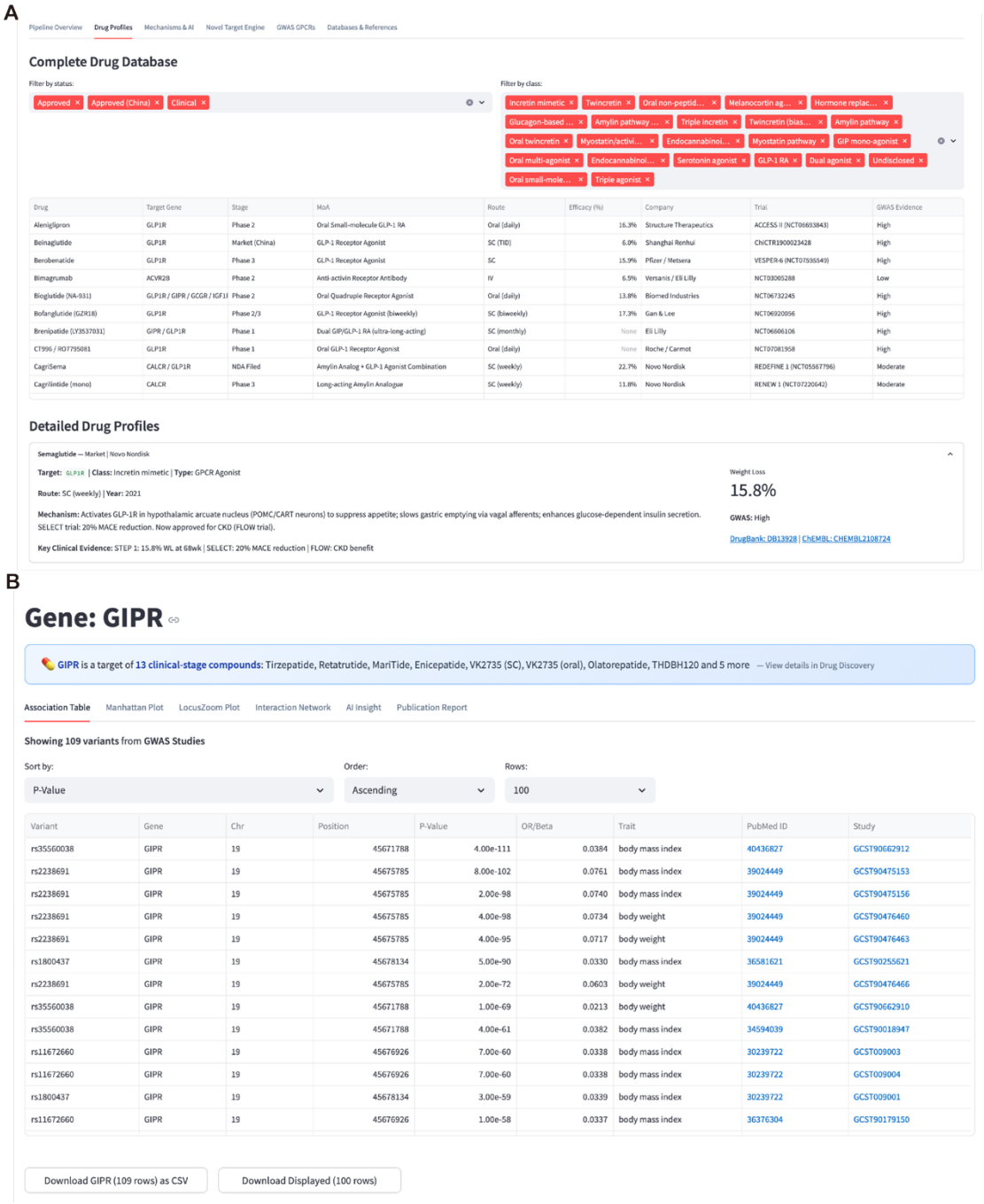
Supplemental Fig. 1 Gene/Variant Browser extended views.

(A) Overview of the browser interface. The Configuration panel allows dataset selection (GWAS, UK Biobank BMI [Field 21001], or UK Biobank BMI Impedance

[Field 23104]), and the Filter panel supports gene-, variant-, and trait-level filtering across obesity-related phenotypes. Six analytical tabs are available: (1) Association Table, (2) Manhattan Plot, (3) LocusZoom Plot, (4) Interaction Network, (5) AI Insight, and (6) Publication Report. The association table is sortable and filterable, with direct links to PubMed records and GWAS Catalog study accessions.

(B) AI Insight tab for the *FTO*, generating a literature-grounded summary covering biological function, GWAS associations, molecular mechanisms, and therapeutic potential, with cited source publications and PubMed links shown in the References panel.

Fig S2



Supplemental Fig. 2 Drug Discovery module and genetic evidence for the GIPR incretin target.

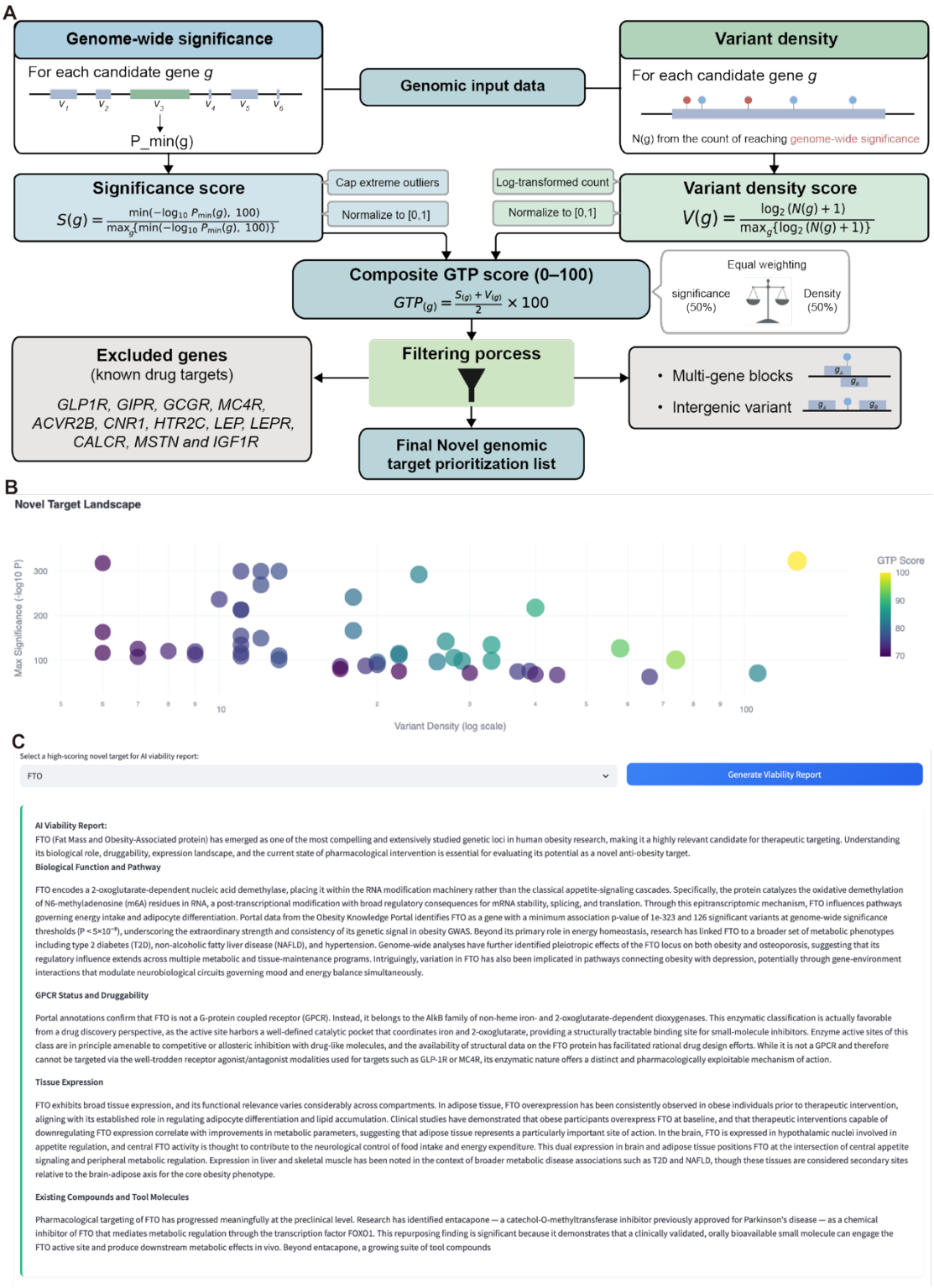
(A) Drug Profiles tab, showing the Complete Drug Database—a filterable table of the curated obesity therapeutics with status- and class-based filters, together with the Detailed Drug Profiles panel, which expands to summarise the key clinical evidence for each compound (Semaglutide shown as an example), with external links to

886 DrugBank and ChEMBL.

887 (B) Gene/Variant Browser view of GIPR under the GWAS dataset, returning 109
888 variants. A cross-linking banner indicates that GIPR is targeted by 13 clinical-stage
889 compounds. The association table is sortable by P value, chromosomal position, or
890 effect size, with direct links to PubMed records and GWAS Catalog study accessions,
891 and is exportable as CSV.

892

Fig S3



Supplemental Fig. 3 Genomic target prioritisation engine.

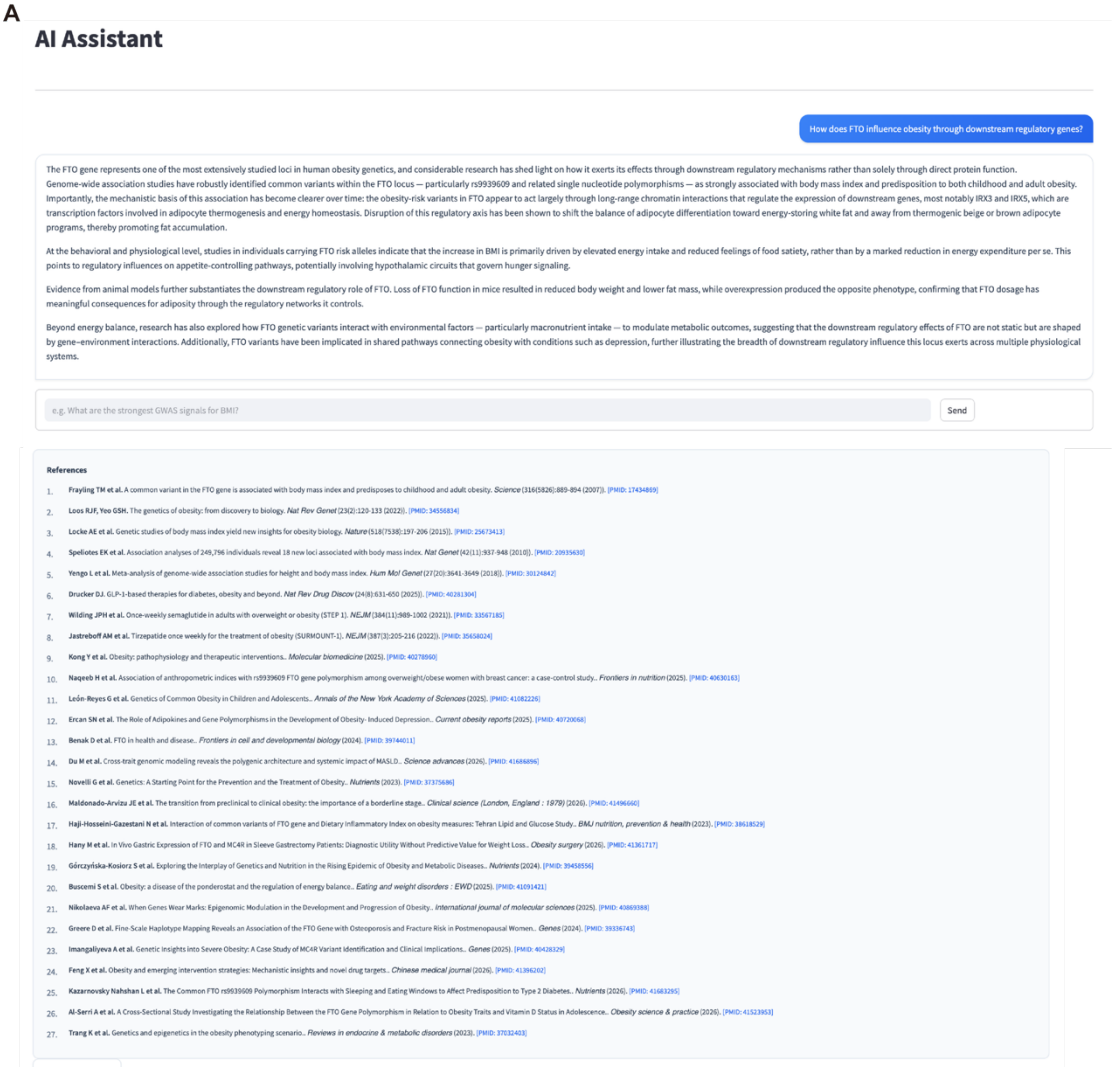
(A) Overview of the Novel Target Discovery Engine, introducing the Genomic Target Prioritisation (GTP) Score.

(B) Novel Target Landscape scatter plot displaying all scored candidate genes by variant density (x-axis, log scale) against maximum statistical significance ($-\log_{10}P$, y-axis), with bubble colour encoding GTP score.

(C) AI Viability Report generated for *FTO*, providing a literature-grounded assessment of biological function, druggability, tissue expression, existing tool molecules, and drug development challenges. Retrieved source references with relevance scores are displayed below the report for qualification.

Fig S4

A



Supplemental Fig. 4. Example output from the OKP AI assistant.

(A) The AI Assistant panel responding to the user query "How does FTO influence obesity through downstream regulatory genes?". The response describes FTO's function as an m^{6}A RNA demethylase regulating the stability and translation of metabolic transcripts, its action on appetite-regulating genes in the hypothalamus, and its long-range enhancer-mediated regulation of downstream genes such as IRX3 and IRX5, which govern thermogenesis and adipocyte differentiation.

Table 1. Comprehensive obesity therapeutic pipeline integrated into the Obesity Knowledge Portal.

Drug	Target	Mechanism	Stage	Trial ID(s)	WL%	GWAS Ev.	Company	Route
FDA-Approved								
Semaglutide	*GLP-1R*	GLP-1 receptor agonist	Approved	STEP 1 (NCT03548935)	15.8	High	Novo Nordisk	SC QW
Tirzepatide	*GIPR/GLP-1R*	Dual GIP/GLP-1 RA	Approved	SURMOUNT-1 (NCT04184622)	22.5	High	Eli Lilly	SC QW
Orforglipron	*GLP-1R*	Oral non-peptide GLP-1 RA	Approved (Apr 2026)	ATTAIN-1 (NCT05051579)	12.4	High	Eli Lilly	PO QD
Liraglutide	*GLP-1R*	GLP-1 receptor agonist	Approved	SCALE (NCT01272219)	8.0	High	Novo Nordisk	SC QD
Setmelanotide	*MC4R*	MC4R agonist	Approved (monogenic)	NCT02896192	25.6	Very High	Rhythm Pharma	SC QD
Metreleptin	*LEPR*	Recombinant leptin	Approved (lipodystrophy)	NIH trials	N/A	High	Amryt Pharma	SC QD
Regionally Approved								
Mazdutide	*GCGR/GLP-1R*	Dual glucagon/GLP-1 RA	China NMPA (Jun 2025)	GLORY-1 (NCT05607680)	14.4	Moderate	Innovent Biologics	SC QW
Ecnoglutide (XW003)	*GLP-1R*	cAMP-biased GLP-1 RA	China NMPA (Mar 2026)	SLIMMER (NCT05813795)	15.1	High	Sciwind Biosciences	SC QW
Beinaglutide	*GLP-1R*	GLP-1 receptor agonist	China NMPA-approved (T2D, 2016); Phase 3 completed (obesity) | ChiCTR1900023428	NCT03829891 ChiCTR1900023428	6.0	High	Shanghai Renhui	SC TID
Under FDA Review								
CagriSema	*AMY-R/GLP-1R*	Amylin + GLP-1 RA combo	NDA filed (Dec 2025)	REDEFINE 1 (NCT05567796)	22.7	Moderate	Novo Nordisk	SC QW
Phase 3								
Retatrutide	*GIPR/GLP-1R/GCGR*	Triple agonist	Phase 3	TRIUMPH-4 (NCT05931367)	28.7	Moderate	Eli Lilly	SC QW
Survodutide	*GCGR/GLP-1R*	Dual glucagon/GLP-1 RA	Phase 3	SYNCHRONIZE-1 (NCT06066528)	18.7	Moderate	BI / Zealand	SC QW
MariTide	*GIPR(ant)/GLP-1R*	GIP antagonist + GLP-1 ag.	Phase 3	MARITIME-1 (NCT06858839)	20.0	Moderate	Amgen	SC QM
Enicepatide (CT-388)	*GIPR/GLP-1R*	Biased dual GLP-1/GIP RA	Phase 3	Enith1 (NCT07351045)	22.5	High	Roche / Carmot	SC QW
VK2735 (SC)	*GIPR/GLP-1R*	Dual GLP-1/GIP RA	Phase 3	VANQUISH-1 (NCT06616870)	14.7 (13wk)	Moderate	Viking Therapeutics	SC QW
Zenagamtide (amycretin)	*AMY-R/GLP-1R*	Unimolecular GLP-1/amylin	Phase 3	AMAZE 1 (NCT07339423)	24.3	Moderate	Novo Nordisk	SC QW
Eloralintide	*AMY-R*	Selective amylin RA	Phase 3	NCT07321886	20.1	Moderate	Eli Lilly	SC QW
Cagrilintide (mono)	*AMY-R*	Long-acting amylin analogue	Phase 3	RENEW 1 (NCT07220642)	11.8	Moderate	Novo Nordisk	SC QW
Zovaglutide (ZT002)	*GLP-1R*	Long-acting GLP-1 RA (monthly)	Phase 3	HORIZON-1 (NCT07230119)	13.8	High	QL Biopharm	SC QM
Olatorepatide	*GIPR/GLP-1R*	Dual GLP-1/GIP RA	Phase 3 (CN done)	NCT07431086	19.0	Moderate	Hansoh / Regeneron	SC QW
Berobenatide	*GLP-1R*	GLP-1 receptor agonist	Phase 3	VESPER-6 (NCT07595549)	15.9	High	Pfizer / Metsera	SC QW/QM

Efsubaglutide alfa	*GLP-1R*	GLP-1 RA	Phase 2b/3	NCT06921486	7.2	High	Innogen	SC QW
Bofanglutide (GZR18)	*GLP-1R*	GLP-1 RA (biweekly)	Phase 2/3	NCT06737042 (phase 2 US); NCT06728124 (phase 3 China)	17.3	High	Gan & Lee	SC Q2W/QM
Phase 2								
Aleniglipron	*GLP-1R*	Oral small-molecule GLP-1 RA	Phase 2	ACCESS II (NCT06693843)	16.3	High	Structure Therapeutics	PO QD
VK2735 (oral)	*GIPR/GLP-1R*	Oral dual GLP-1/GIP RA	Phase 2	VENTURE-Oral (NCT06828055)	12.2 (13wk)	Moderate	Viking Therapeutics	PO QD
Bimagrumab	*ActRIIA/B*	Anti-activin receptor mAb	Phase 2	NCT03005288	6.5 (BW)	Low	Versanis / Lilly	IV
Monlunabant	*CB1R*	Peripheral CB1 inverse agonist	Phase 2	NCT05891834	6.4	High	Novo Nordisk	PO
Petrelintide	*AMY-R*	Long-acting amylin analogue	Phase 2	ZUPREME-1 (NCT06662539)	10.7	Moderate	Zealand / Roche	SC QW
Pemvidutide	*GCGR/GLP-1R*	Dual glucagon/GLP-1 RA	Phase 2	MOMENTUM (NCT05295875)	15.6	Moderate	Altimmune	SC QW
Taldefgrobep alfa	*Myostatin*	Anti-myostatin mAb	Phase 2b	NCT07281495	TBD	Low	Biohaven	SC
THDBH120	*GIPR/GLP-1R*	Dual GIP/GLP-1 receptor agonist (ultra-long-acting)	Phase 2	NCT07036601	9.36	High	Tonghua Dongbao	SC QW/Q2W
HDM1005	*GIPR/GLP-1R*	Dual GIP/GLP-1 receptor agonist	Phase 2	NCT07279194 (CN); NCT06886126 (US)	13.28	—	**Huadong Medicine**	SC QW
Macupatide	*GIPR*	GIP receptor agonist	Phase 2	NCT07589608	—	High	Eli Lilly	SC
Bioglutide (NA-931)	*IGF-1R/GLP-1R/GIPR/GCGR*	First-in-class oral quadruple receptor agonist	Phase 2	NCT06732245	13.8	High (GLP1R, GIPR, GCGR)	Biomed Industries	PO QD
Phase 1								
Brenipatide (LY3537031)	*GIPR/GLP-1R*	Dual GIP/GLP-1 RA (ultra-long-acting, monthly)	Phase 1	NCT06606106	—	High	Eli Lilly	SC
LY4167586	*Undisclosed*	Undisclosed	Phase 1	NCT07225556	—	—	Eli Lilly	SC
CT996 / RO7795081	*GLP-1R*	Oral GLP-1 RA	Phase 1	NCT07081958	—	High	Roche / Carmot	PO
SCO-094	*GIPR/GLP-1R*	Dual GLP-1/GIP RA	Phase 1	—	—	Moderate	Scohia Pharma	SC
NNC0174-1213	*Undisclosed*	Undisclosed	Phase 1	NCT06719011	—	—	Novo Nordisk	—
NNC6989-0001	*Undisclosed*	Undisclosed	Phase 1	NCT07437079	—	—	Novo Nordisk	—
NNC0662-0419	*Undisclosed*	Undisclosed	Phase 1	NCT07525791	—	—	Novo Nordisk	—
HM15275	*GIPR/GLP-1R/GCGR*	Next-gen triple agonist	Phase 1	NCT07205900	—	Moderate	Hanmi Pharma	SC
Naperiglipron	*GLP-1R*	Oral small-molecule GLP-1 RA	Phase 1	NCT07232732	—	High	AstraZeneca	PO
MDR-001	*Undisclosed*	Undisclosed	Phase 1	NCT06606483	8.9	—	Mediar Therapeutics	—
UBT251	*Undisclosed*	Undisclosed	Phase 1	NCT07177469	19.7	—	—	—
HDM1002	*Undisclosed*	Undisclosed	Phase 1/2	NCT06500299	—	—	Harbour BioMed	—
SYH2082	*Undisclosed*	Undisclosed	Phase 1	NCT07532655	—	—	—	—
Withdrawn								
Rimonabant	*CB1R*	CB1 inverse agonist	Withdrawn (2008)	RIO-Europe	6.5	High	Sanofi	PO
Lorcaserin	*5-HT2C-R*	5-HT2C agonist	Withdrawn (2020)	BLOOM (NCT00395135)	5.8	High	Eisai / Arena	PO

WL% = maximum reported weight loss (%); GWAS Ev. = strength of supporting GWAS evidence; SC = subcutaneous; PO = oral; QW = once weekly; QD = once daily; TID = three times daily; QM = once monthly; Q2W = every 2 weeks; IV = intravenous. Data current as of June2026.