Supplementary materials

This document forms the appendices for the manuscript "BIODISCO*: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation.*"

A A Detailed Case Study

To provide an insight into how BIODISCO works, we present a detailed case study. The case is grounded in a recent study linking G-protein coupled receptors (GPCRs) to increased cardiovascular risk (Kobilka and Lefkowitz 2024), and centres on two entities: the gene **GPR153** and the disease **vascular injury**. We provided "Role of GPR153 in vascular injury and disease" as input to BIODISCO. Relevant entities were drawn from the input query and used to access the literature interface to retrieve relevant publications.

Initially, the BACKGROUND agent synthesised a summary highlighting the dual role of GPR153 in vascular injury and inflammation from the retrieved literature. The EXPLORER analyzed the background summary to retrieve a subgraph consisting of genes like GPR153, CEBPB, GRN, CDK4, TTR, YAP1, SCAMP1 as well as drugs like Acamprosate and conditions such as camptodactyly.

The SCIENTIST received both the background summary and the KG subgraph based on which it proposed an initial hypothesis:

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression via the YAP/TAZ pathway, promoting neointima formation following vascular injury"

Subsequently, the initial hypothesis and relevant keywords were used to query the literature interface for relevant PubMed publications to find evidence for the generated hypothesis. This search identified studies such as Shao et al. (2025) which conveys how GPR153 is an orphan receptor that facilitates expression of pro-inflammation and pro-proliferation genes in smooth muscle cells by regulating cAMP levels in cells and thereby contributing to inflammation and vascular remodelling.

The initial hypothesis by SCIENTIST was evaluated by the CRITIC, considering the background and literature. Novelty was rated 4, reflecting the novel aspect of linking GPR153 activation to the YAP/TAZ pathway and neointima formation in vascular smooth muscle cells. Relevance was given a 5 due to the significant interest in vascular injury in therapeutic research. Significance was rated 4, acknowledging the therapeutic potential of targeting this pathway. This gave the initial hypothesis 17 out of 20.

To improve the hypothesis, the REVIEWER first focused on Novelty by highlighting "lack of mechanistic insight" on how exactly GPR153 activation influences the YAP/TAZ pathway. It decided to obtain additional knowledge graph evidence, and therefore used the current hypothesis, critic feedback, and background information to query the KG interface. This yielded a new subgraph from the knowledge graph with GPR153, YAP, and TAZ as the key entities. Using this new information, the REFINER then reformulated the hypothesis to:

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression by promoting CEBPB-mediated YAP1 signalling, thereby potentially integrating with EGR1 and GSK3B pathways to exacerbate neointima formation following vascular injury"

Following this, the system re-entered the feedback loop. Incorporating newly identified literature, the CRITIC raised the Novelty and Significance scores to 5, while Relevance remained at 5 and Verifiability at 4, yielding an overall score of 19. Normally, iteration would stop here, as further improvement is constrained by the inherent complexity of the biology rather than lack of evidence. For demonstration, however, we continued the process to illustrate how the system explores more intricate mechanistic hypotheses. Additional literature searches for each key gene and their combinations further substantiated the potential connections identified in the evolving hypothesis. Subsequently, the REVIEWER focused on Verifiability—the only criterion not awarded a top score. Incorporating new subgraph information and additional literature, the REFINER further elaborated the hypothesis to include a broader CEBPB-mediated network involving YAP1, EGR1, and GSK3B:

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression by facilitating CEBPB-mediated network involving YAP1, EGR1, and GSK3B, creating a complex signalling cascade that drives neointima formation after vascular injury"

In the next round, the CRITIC again assigned top scores for Novelty, Significance, and Relevance, while Verifiability remained at 4, reflecting the growing experimental challenges posed by increased mechanistic complexity. To further demonstrate system capabilities, additional genes (NRF1, CD7, and GSK3B) were integrated, supported by targeted literature searches:

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression through a CEBPB-mediated network, integrating NRF1 and CD7 interactions with YAP1 and GSK3B, thereby orchestrating a multifaceted signalling cascade that drives neointima formation following vascular injury"

For this, CRITIC gave Novelty a 5 because it "combines several key regulatory proteins" to provide a comprehensive signalling network. Both Relevance and Significance were also scored at 5. Verifiability was still at 4 because CRITIC noted that although techniques like gene editing, pathway analysis and in vivo models of vascular injury could be used to test the hypothesis, the "complexity of the interactions may pose challenges in definitively confirming all proposed relationships". This resulted in an overall score of 19/20.

B Feature comparison

	CoSci	IntS	SciA	ResA	BIODISCO
Multi-agent	✓	✓	✓	✓	✓
Tool use			✓	✓	✓
KG		✓	✓	✓	✓
Search			✓		✓
Reviewer		✓	✓	✓	✓
Scoring				✓	✓

Table 2: Features of LLM agentic systems: multiple agents, use of external tools, knowledge graph integration, ability (dynamically) to search academic literature, presence of a 'reviewer' agent critically appraising outputs, and refinement of hypotheses based on explicit numerical scoring; CoSci = AI co-scientist Gottweis et al. (2025); IntS = Intelliscope Aamer et al. (2025); SciA = SciAgents Ghafarollahi and Buehler (2024); ResA = ResearchAgent Baek et al. (2025). ResearchAgent constructs a knowledge graph from an academic graph, but does not access a 'live' search API

C Additional Results

Examples from Qi et al. (2024)

In Table 3, we present selected examples comparing gold-standard hypotheses with those generated by BIODISCO, along with their cosine similarity scores (see main paper for details on temporal evaluation). While Figure 3 demonstrates that BIODISCO produces hypotheses with higher semantic similarity than unrelated gold pairs, this table illustrates whether that similarity arises from shared terminology or meaningful mechanistic insight.

In the first example, we see that BIODISCO accurately identifies the core mechanistic insight present in the gold hypothesis that that inhibiting VEGFR2 in hypertrophic chondrocytes interrupts ERK1/2 activation and subsequent apoptosis. We can also see how BIODISCO provides more nuanced insights by integrating additional molecular pathways when compared to the gold-standard hypothesis.

Similarly, in the second example, BIODISCO successfully proposes and extends the insight from the gold-standard hypothesis, adding specific details about potential pathways and proteins involved in the process. Unlike the previous examples where the generated and gold-standard hypotheses shared the same mechanistic insights, this third pair takes different directions, resulting in a low similarity score.

TruthHypo

Table 4 presents the results for each class across the three tasks in the TruthHypo dataset. The high Precision, Recall, and F_1 scores achieved by BIODISCO demonstrate its ability to accurately identify the relationships between the given entities.

Direct Evaluation using LLMs

Table 5 reports LLM-based evaluation scores across four metrics - novelty, relevance, significance, verifiability, and an overall score. As expected, we observe consistent improvements when moving from a single-agent baseline to a multi-agent setup and further to BIODISCO. However, the performance margins between versions remain narrow. Also, the effect of access to external interfaces is particularly unclear, with only minor differences in scores.

These observations highlight that scalar ratings alone may be insufficient to capture more nuanced improvements in hypothesis quality. This underscores the value of pairwise evaluation, where an LLM compares hypotheses directly and can better distinguish subtle but meaningful differences in structure, specificity, and insight that may be overlooked in absolute scoring schemes.

Gold hypothesis	BIODISCO	S_{ij}
Inhibition of VEGFR2 should interrupt phosphate-induced ERK1/2 activation and subsequent apoptotic events in hypertrophic chondrocytes.	Inhibition of VEGFR2 in hypertrophic chondrocytes suppresses ERK1/2 activation, preventing apoptosis by modulating downstream pathways involving PPP2R2A or CORO1C, thereby ameliorating hypophosphatemic rickets.	0.86
PPFIBP1 may play a role in the development of chemoresistance in MM.	Elevated PPFIBP1 upregulation in multiple myeloma cells enhances bortezomib resistance by activating NF-κB signaling, supported by PPFIBP1's role in promoting RelA stability and cyto-nuclear translocation, indicating a direct link to chemoresistance.	0.65
TkR86C expression in damaged wing discs and brain suggests a possible non-cell-autonomous role in regeneration	Neuregulin signaling via the AMPK/mTOR pathway enhances progenitor cell proliferation in limb regeneration.	0.39

Table 3: Examples of gold hypotheses and hypotheses generated by BIODISCO, with cosine similarity, S_{ij}, computed from their respective embeddings.

D Paired Comparison Models

Paired comparison models were fitted using the R package `BradleyTerry2` (Turner and Firth 2012), with quasi-variance approximations computed using `qvcalc` (Firth 2025). The Bradley–Terry model (Bradley and Terry 1952) is given by

$$\log \mathrm{odds}(i \text{ beats } j) = \alpha + \beta_i - \beta_j, \tag{4}$$

where i and j denote 'players' or LLM agents, β_i is the ability score (on the logit-scale) of the ith LLM system and α is a 'home advantage' parameter, estimating the possible order

Group	Precision	Recall	F_1	Acc	Support
Chemical-Gene / negative	0.917	0.880	0.898		50
Chemical-Gene / positive	0.885	0.920	0.902		50
Chemical-Gene (avg)	0.901	0.900	0.900	0.900	100
Disease-Gene / inhibit	0.848	0.780	0.812		50
Disease-Gene / stimulate	0.796	0.860	0.827		50
Disease-Gene (avg)	0.822	0.820	0.820	0.820	100
Gene-Gene / negative	0.867	0.780	0.821		50
Gene-Gene / positive	0.800	0.880	0.838		50
Gene-Gene (avg)	0.833	0.830	0.830	0.830	100
ALL DATA (avg)	0.844	0.842	0.842	0.850	300

Table 4: Performance of the proposed system by group, reporting Precision, Recall, F1, Accuracy, and Support for each relation and macro average.

Configuration	Novelty	Relevance	Significance	Verifiability	Overall
GPT-4.1 (baseline)	1.38 ± 0.60	3.00 ± 0.00	1.89 ± 0.72	2.82 ± 0.22	2.27 ± 0.31
Multi-agent system	2.06 ± 0.32	3.00 ± 0.00	2.48 ± 0.10	2.70 ± 0.27	2.56 ± 0.09
Multi-agent + tools	2.43 ± 0.18	3.00 ± 0.00	2.50 ± 0.00	2.46 ± 0.22	2.60 ± 0.06
Multi-agent + refine	2.46 ± 0.14	2.99 ± 0.05	2.50 ± 0.00	2.44 ± 0.24	2.60 ± 0.07
BIODISCO	2.54 ± 0.14	2.99 ± 0.05	2.55 ± 0.14	2.54 ± 0.42	2.66 ± 0.13

Table 5: Direct comparison with a evaluator LLM and various ablation configurations of the BIODISCO framework: with and without tools (i.e. KG and literature search), iterative refinement and multi-agentic reasoning. The baseline is a single-agent LLM only. Reported mean $\pm$ standard deviation of scores for a set of 100 generated hypothesis.

effect bias, assuming the hypothesis of i is presented to the evaluator first in the pair.

For our model, ties were treated as half wins to each player. An extension of the Bradley–Terry model that explicitly supports ties, the Davidson (1970) model, was also fitted separately, but yielded very similar results (not presented here).

The Bradley–Terry model parameter estimates $\hat{\alpha}$ are given in Table 6 and suggest a statistically significant order bias for Novelty and Verifiability evaluations (in favour of the item appearing first in the comparisons) but there is not enough evidence to indicate such an effect exists in Relevance or Significance comparisons at the 5% level of significance.

Metric	Estimate	Std. Err.	Statistic	p-value
Novelty	0.80	0.36	2.23	0.03
Relevance	0.08	0.14	0.54	0.59
Significance	0.07	0.24	0.28	0.78
Verifiability	0.44	0.15	2.95	0.00

Table 6: Parameter estimates $\hat{\alpha}$ from the fitted Bradley–Terry models for each of the four evaluation metrics

Quasi-variances (Firth 2004) aim to minimize the squared loss

$$\min \sum_{i<j} (q_i + q_k - v_{ij})^2, \tag{5}$$

where q_i is the quasi-variance of player i and v_{ij} is the covariance term (i.e. off-diagonal entries between players i and j. The package `qvcalc` returns relative errors as a measure of the quality of this approximation: the distribution of relative errors is given in Figure 7 as a beeswarm plot (Selby 2020) and appear to be mostly acceptable.

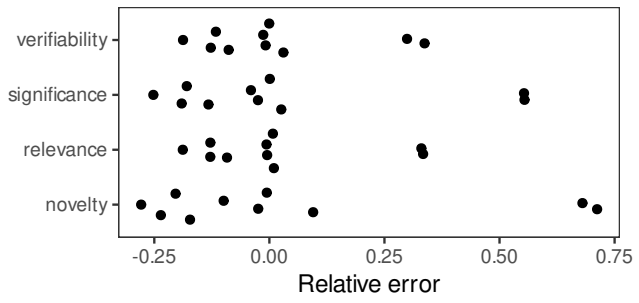


Figure 7: Beeswarm plot showing distribution of relative errors from the quasi-variance approximation

E Compute Infrastructure

All experiments were executed on a commodity CPU server (4 cores, 16 GB RAM, no GPU), making the workflow straightforward to deploy and reproduce. Inference time and cost depend chiefly on (i) the number of refinement iterations, (ii) the size of the knowledge-graph (KG) and literature evidence retrieved, (iii) PubMed/Neo4j access latency, and (iv) network latency to the OpenAI GPT-4.1 API.

For all LLM-based agents, random seed (cache seed) was fixed to 42 to support reproducibility, and temperature was set between 0.2 and 0.5 depending on the agent role to balance generation diversity and stability.

Four configurations were benchmarked in Table 7. The single-pass multi-agent baseline without KG or literature ("Multi-agent") is the most economical, averaging $0.004 per hypothesis. In contrast, the full BioDisco pipeline, which involves three refinement iterations, had the highest cost of approximately 0.071 US dollars. Incorporating KG and literature significantly increases token usage and API cost but provides richer contextual and mechanistic evidence. On average, BioDisco completes a full inference in 2 to 3 minutes per hypothesis, while the lightweight baseline finishes within one minute.

Setting	Input Tokens	Output Tokens	US $
Multi-agent	1393	264	0.004
Multi+Tools	8113	653	0.017
Multi+Refine	30263	1495	0.019
BioDisco	72828	4424	0.071

Table 7: Computation cost under different system settings. All results are based on experiments using GPT-4.1. API cost in United States dollars ($) is reported per hypothesis.

All essential components (OpenAI LLM API, Neo4j KG, and open-source retrieval/embedding libraries) are publicly available.

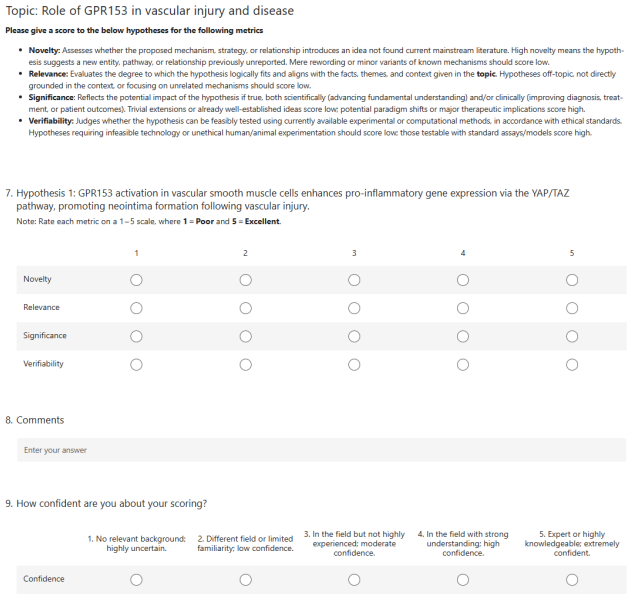
F Human evaluation

Survey Design

The human evaluation of hypotheses took place online via a bespoke Microsoft Form. Dissemination of the survey was through direct and e-mail communication and through a consortium-wide mailing list.

Participants had the right to remain anonymous or to be acknowledged by name in subsequent publications. They were asked to give their institutional affiliation, years of experience, and discretized number of publications in their field of expertise. For each hypothesis, the participant was asked to rate it for novelty, significance, verifiability and relevance (to the given input context), each on a scale from 1–5, where 1 is the lowest and 5 is the highest score. The set of metrics is based on Qi et al. (2024) and the descriptions were the same as given to LLM evaluators (see Listing 8).

Additionally, each respondents could score their level of confidence in their chosen ratings for each hypothesis (on a 1–5 scale) according to their familiarity with the topic area, as well as add free-text comments. All fields in the survey were optional, so the user could skip giving a rating for any part of any hypothesis, for any reason. This was to reduce the risk of unconfident or uninterested respondents giving arbitrary scores. A sample from the online questionnaire is presented in Figure 8.


Topic: Role of GPR153 in vascular injury and disease

Please give a score to the below hypotheses for the following metrics

- **Novelty:** Assesses whether the proposed mechanism, strategy, or relationship introduces an idea not found current mainstream literature. High novelty means the hypothesis suggests a new entity, pathway, or relationship previously unreported. Mere rewording or minor variants of known mechanisms should score low.
- **Relevance:** Evaluates the degree to which the hypothesis logically fits and aligns with the facts, themes, and context given in the **topic**. Hypotheses off-topic, not directly grounded in the context, or focusing on unrelated mechanisms should score low.
- **Significance:** Reflects the potential impact of the hypothesis if true, both scientifically (advancing fundamental understanding) and/or clinically (improving diagnosis, treatment, or patient outcomes). Trivial extensions or already well-established ideas score low; potential paradigm shifts or major therapeutic implications score high.
- **Verifiability:** Judges whether the hypothesis can be feasibly tested using currently available experimental or computational methods, in accordance with ethical standards. Hypotheses requiring infeasible technology or unethical human/animal experimentation should score low; those testable with standard assays/models score high.

7. Hypothesis 1: GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression via the YAP/TAZ pathway, promoting neointima formation following vascular injury.

Note: Rate each metric on a 1 – 5 scale, where **1 = Poor** and **5 = Excellent**.

	1	2	3	4	5
Novelty	○	○	○	○	○
Relevance	○	○	○	○	○
Significance	○	○	○	○	○
Verifiability	○	○	○	○	○

8. Comments

Enter your answer

9. How confident are you about your scoring?

	1. No relevant background: highly uncertain.	2. Different field or limited familiarity: low confidence.	3. In the field but not highly experienced: moderate confidence.	4. In the field with strong understanding: high confidence.	5. Expert or highly knowledgeable: extremely confident.
Confidence	○	○	○	○	○

Figure 8: Screenshot from the expert evaluation questionnaire, showing a single hypothesis and scoring rubric

Evaluation Material

We provide the full list of topics, initial hypotheses, and final hypotheses used for human evaluation in both cardiovascular disease (CVD) and immunology domains (see Table 8 and Table 9).

Bayesian Item Response Modelling

A summary of this model appears in Section 3.2 ("What do human domain experts think?") of the main text, and we provide the full details here. To quantitatively assess human evaluation scores while accounting for both rater and hypothesis-specific variability, we fit a Bayesian cumulative ordinal mixed-effects model to the ordinal scores assigned by each rater. Let Y_{ijm} denote the ordinal rating (on a scale of 1 to 5) given by rater i to hypothesis j on metric m. The model estimates the cumulative probability for each rating category $k \in \{1, 2, 3, 4\}$ as

$$P(Y_{ijm} \leq k) = \Phi(\tau_k - \eta_{ijm}), \tag{6}$$

where Φ denotes the standard normal cumulative distribution function (probit link), and τ_k are the latent thresholds that partition the ordinal rating scale.

The linear predictor η_{ijm} contains both fixed and random effects:

$$\eta_{ijm} = \underbrace{\beta_m}_{\text{metric effect}} + \underbrace{u_i}_{\text{rater effect}} + \underbrace{v_{jm}}_{\text{hypothesis-on-metric effect}}, \tag{7}$$

where β_m is a fixed effect representing the average quality or 'easiness' for metric m, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is a rater-specific random effect capturing overall leniency or severity of rater i, and $v_{jm} \sim \mathcal{N}(0, \sigma_v^2)$ is a hypothesis- and metric-specific random effect reflecting the relative quality of hypothesis j on metric m.

Input	Initial Hypothesis	Final Hypothesis
Role of GPR153 in vascular injury and disease	GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression via the YAP/TAZ pathway, promoting neointima formation following vascular injury.	GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression through a CEBPB-mediated network, integrating NRF1 and CD7 interactions with YAP1 and GSK3B, thereby orchestrating a multifaceted signaling cascade that drives neointima formation following vascular injury.
FDX1 and cholesterol metabolism in cardiovascular risk	Elevated expression of FDX1 in macrophages enhances cholesterol efflux, thereby reducing carotid intima media thickness and lowering cardiovascular risk through modulation of inflammatory pathways.	Elevated FDX1 expression in macrophages enhances cholesterol efflux and reduces carotid intima media thickness by engaging regulatory networks involving EGR1, NR4A2, and CAPZA2, with experimental validation through gene expression analyses and metabolic profiling to disentangle these interactions and their impact on cardiovascular risk.
PTLOs and immune responses in atherosclerosis	Activation of PTGDS within tertiary lymphoid organ-like structures enhances B cell-mediated antibody production, promoting plaque instability in atherosclerosis.	Activation of PTGDS within tertiary lymphoid organ-like structures not only enhances B cell-mediated antibody production but also establishes a feedback loop involving TNF-α and AKT1 that regulates macrophage efferocytosis, thereby orchestrating inflammatory responses that escalate plaque instability in atherosclerosis.
Rare pathogenic variants in G-protein-coupled receptor genes for atherosclerosis	Rare pathogenic variants in GPCR genes enhance inflammatory signaling, increasing the risk of atrial fibrillation and cardiomyopathy in patients independent of traditional cardiovascular risk factors.	Rare pathogenic variants in GPCR genes disrupt inflammatory signaling and gut microbiota interactions, which synergistically exacerbate atrial fibrillation and cardiomyopathy, suggesting a novel approach for identifying targeted biomarkers that link these pathways in personalized cardiovascular disease management.
Cpt1b and heart regeneration	Inhibition of CPT1B enhances cardiomyocyte proliferation by reactivating cardiogenic factors and suppressing NF-κB-mediated inflammatory responses during heart regeneration.	Inhibition of CPT1B enhances cardiomyocyte proliferation through a regulatory cascade involving GOLT1A that modulates mitochondrial dynamics and lipid metabolism, while concurrently suppressing NF-κB signaling and promoting a metabolic shift, thereby facilitating heart regeneration.

Table 8: CVD–related human evaluation cases, showing the input, initial hypothesis, and final (refined) hypothesis produced by the system.

Input	Initial Hypothesis	Final Hypothesis
Investigation of molecular drivers underlying T cell exhaustion in non-small cell lung cancer (NSCLC), with a focus on identifying novel, druggable targets to enhance the efficacy of immune checkpoint inhibitor (ICI) therapies.	Inhibition of STK31 in NSCLC tumor microenvironments enhances CD8 T-cell functionality and mitigates T-cell exhaustion, improving the efficacy of immune checkpoint inhibitors.	Inhibition of STK31 in NSCLC tumor microenvironments enhances CD8 T-cell functionality by disrupting ETS1 and PIK3R1-mediated immunosuppressive pathways, while promoting pro-inflammatory cytokine release, thereby synergistically improving responsiveness to immune checkpoint inhibitors and overcoming T-cell exhaustion.
	Inhibition of PIK3R1-CTLA4 interaction attenuates T cell exhaustion and enhances the efficacy of immune checkpoint inhibitor therapy in non-small cell lung cancer.	Adoptive transfer of NSCLC patient T cells engineered via CRISPR to disrupt the PIK3R1-CTLA4 axis and enhance CD7 costimulation will yield durable reversal of exhaustion and superior clinical responses to immune checkpoint inhibitor therapy in early-phase clinical trials.
	PPP2CA suppresses STAT3-mediated cytokine signaling in the NSCLC tumor microenvironment, thereby reducing CD8+ T cell exhaustion and enhancing the efficacy of immune checkpoint inhibitors.	PPP2CA dephosphorylates STAT3 and modulates MYC and ETS1 activity, leading to altered TNFSF4 signaling and a reprogrammed cytokine milieu in NSCLC that diminishes CD8+ T cell exhaustion and enhances the therapeutic efficacy of immune checkpoint blockade.
	CTLA4 upregulation in tumor-infiltrating T cells promotes T cell exhaustion and resistance to immune checkpoint inhibitors in non-small cell lung cancer.	In NSCLC, lactate-induced activation of SRPK1 enhances USP39-mediated RNA splicing in regulatory T cells, leading to CTLA4 upregulation and T cell exhaustion, with SRPK1 or USP39 inhibition predicted to restore antitumor immunity and sensitize tumors to immune checkpoint inhibitors.
	Inhibition of CK2B enhances PIK3R1-mediated CTLA4 downregulation, reducing T cell exhaustion and improving immune checkpoint inhibitor efficacy in non-small cell lung cancer.	Inhibition of CK2B and PIK3R1, in conjunction with IL-21R activation, enhances CTLA4 downregulation and STAT5B-mediated T cell reactivation, improving immune checkpoint inhibitor responses in non-small cell lung cancer by targeting progenitor-exhausted T cells within the tumor microenvironment.

Table 9: Immunology–related human evaluation cases, showing the input, initial hypothesis, and final (refined) hypothesis produced by the system.

Using `brms` (Bürkner 2021) we fit a single model across all metrics to increase statistical power, under the assumption that the rater leniency effects are constant across metrics. This approach pools information across all dimensions for more robust estimation of both rater and hypothesis effects.

G Agent Roles

Here we describe in detail the role of the individual agents and their respective prompts.

G.1 Planner agent

The PLANNER agent functions as the central coordinator for the hypothesis discovery pipeline. It receives keywords input by users and manages the sequential execution of all specialized agents, including background retrieval, knowledge graph exploration, hypothesis generation, evaluation, and refinement. At each stage, the PLANNER passes relevant intermediate outputs between agents, monitors progress, and handles control flow decisions such as triggering iterative refinement or terminating the process. In addition to orchestrating agent execution, the PLANNER can optionally produce a concise research plan summarizing the workflow steps taken for a given task. This design promotes modularity, simplifies pipeline management, and facilitates user intervention or expert oversight at any stage of the discovery process.

Listing 1: PLANNER agent

Develop a clear, stepwise research workflow based on the provided background text.
Your plan should outline:
1. Domain selection;
2. Knowledge graph retrieval steps;
3. Hypothesis generation;
4. Iterative refinement using literature and graph evidence;
5. Final decision–making.
Respond with numbered steps.

G.2 Background agent

The BACKGROUND agent is responsible for constructing a concise and informative textual background for each hypothesis discovery task. It receives keywords or research topics and then retrieves relevant biomedical literatures through the literature interface, typically by querying PubMed. The agent then synthesizes and summarizes the retrieved content, producing a structured background paragraph that integrates the most pertinent findings and context. This background serves as a foundation for downstream agents, ensuring that subsequent hypothesis generation and evaluation are grounded in up-to-date and domain-relevant evidence.

Listing 2: BACKGROUND agent

You are given a list of PubMed article metadata blocks about a specific disease and a set of core genes or biological entities. Write a concise, well–structured background paragraph (less than 150 words) that summarizes key mechanistic insights and highlights the relationships between the core genes and disease–relevant biological processes (such as EMT, inflammation, senescence, signaling, etc.).
Requirements:
1. Clearly explain how the core genes are linked to disease mechanisms, pathways, or phenotypes based on the literature.
2. Emphasize causal or regulatory connections when possible, rather than just listing associations.
3. Do not copy sentences verbatim from abstracts. Always synthesize and paraphrase information in your own words.
4. Use clear, logical, and scientifically precise language .
5. Avoid including superfluous or generic information; focus on mechanistic insights most relevant to the disease and core genes.

G.3 Explorer agent

The EXPLORER agent retrieves and summarizes subgraphs from a biomedical knowledge graph based on provided background information and keywords. The information passed to EXPLORER includes keywords and a text as background reference. It first maps the input keywords to candidate standardized entities in the graph, then leverages LLM to select the most contextually relevant nodes using the background as guidance. Then performs composite queries to extract related nodes and multi-hop paths based on these anchor entities. Query parameters, including hop depth, relation types, and result size are dynamically adjusted depending on the reasoning stage: broad subgraphs are retrieved during initial hypothesis generation to encourage diverse exploration, while later refinement stages focus on deeper, localized evidence addressing specific weaknesses such as low novelty or verifiability. The agent returns a structured subgraph summary that provides precise contextual support for downstream hypothesis generation and evaluation.

Listing 3: EXPLORER agent

Given a background text and a list of candidate KG node, Based on the background information, select the most relevant nodes (5–10) to use for subgraph construction.
1. Only choose from the provided candidates.
2. Output only a JSON array of the selected.
3. Do not output any extra text, explanations, or formatting.

G.4 Scientist agent

The SCIENTIST agent generates initial biomedical hypotheses by reasoning over the structured KG subgraph and textual background. It is fed with background information generated by BACKGROUND agent and subgraph information generated by EXPLORER agent, and then simulates the inference process of a researcher, identifying potentially novel and testable associations between entities based on mechanistic context and graph structure. Each hypothesis is expressed in natural language and describes a potential causal or regulatory relationship between biomedical entities. To encourage exploration of the hypothesis space, the ScientistAgent typically generates three candidate hypotheses in

parallel, which serve as inputs for subsequent evaluation and refinement stages.

Listing 4: SCIENTIST agent

```
Generate up to 3 concise, testable biomedical
    hypotheses. Each hypothesis must be grounded in
    both the background and KG context, but extend
    current knowledge with a novel mechanistic,
    causal, regulatory, or predictive insight.
Guidelines:
1. Integrate both background and KG context.
2. Propose new biological mechanisms or interactions,
    not summaries or rephrasings of input.
3. Use precise scientific language, including
    mechanistic verbs such as: activates, inhibits,
    modulates, represses, etc.
4. Each hypothesis must be a single, plausible, testable
    sentence (<= 30 words) with clear entities and
    measurable outcomes.
5. Output only the hypotheses, no numbering, bullets,
    explanations, citations, or evidence fields.
Examples (good):
Activation of TGF-$\beta$ in smooth muscle cells
    promotes vascular remodeling in hypertension.
Loss of gene X enhances inflammatory response to
    toxin Y in liver tissue.
Examples (bad, avoid):
CVD is associated with Wnt signaling and fibrosis.
Output:
Each line should be a standalone hypothesis. Return
    exactly one hypothesis per line, and nothing else.
```

G.5 Critic agent

The CRITIC agent provides structured evaluation of each candidate hypothesis based on supporting evidence, offering clear feedback signals to the system. Three pieces of information are passed to it simultaneously, including the background generated by BACKGROUND agent, the current hypothesis, and the corresponding relevant references (used in hypothesis generation and improvement). It scores hypotheses along four core dimensions: novelty, relevance, significance, and verifiability. Each dimension is rated on a 0–5 scale and accompanied by a brief explanation that justifies the score. The assessment is grounded in the LLM integrated understanding of the hypothesis, background, and literature evidence. The resulting evaluations guide downstream diagnosis and refinement by identifying weaknesses and informing targeted revision strategies.

Listing 5: CRITIC agent

```
Assess the hypothesis using four metrics:
Novelty: Does it introduce ideas not present in the
    background?
Relevance: How well does it align with the background
    and supporting evidence?
Significance: What is its potential to advance biological
    understanding or clinical practice?
Verifiability: Can it be reliably tested with current
    scientific methods?
Rate each metric on a 0–5 scale:
0 = no merit, 1 = very slight, 2 = slight, 3 = moderate, 4
    = strong, 5 = exceptional.
Be conservative: award a 5 only if the hypothesis fully
    meets the criterion with no reservations. Provide
    one sentence of rationale per metric.
Output:
<Metric>: Score <X>
<One-sentence rationale>
(repeat for all 4 metrics)
At the end, write on a separate line:
Overall Score: <value>/20
```

G.6 Reviewer agent

The REVIEWER agent identifies weak dimensions in each hypothesis based on the scores and explanations provided by the CriticAgent. To make a sound decision, it receives full feedback from the CRITIC agent, along with the current hypothesis. It then prioritizes low-scoring criteria and, depending on the evaluation content, selectively triggers access to external knowledge sources, including the knowledge graph, literature, or background. The Agent does not directly modify the hypothesis, instead, it outputs retrieval actions along with the relevant supporting information, which are then passed to the Refiner for hypothesis revision.

Listing 6: REVIEWER agent

```
Given the CriticAgent's markdown critique (scores 0–5
    with rationales), the current hypothesis, and
    background text, recommend follow-up actions
    and query adjustments.
Steps:
1. Identify all metrics scoring <= 3. If none, select a 4-
    point metric using this priority: Novelty >
    Significance > Relevance > Verifiability.
2. Recommend all relevant actions from ['neo4j', '
    pubmed', 'background']:
   - For low novelty or mechanistic gaps: use 'neo4j';
       add 'pubmed' if literature may help.
   - For low verifiability: use 'pubmed'; add 'neo4j' if
       KG includes measurable pathways.
   - For low relevance: use 'background'.
   - If multiple metrics are low, recommend all relevant
       actions
3. Output exactly 3 lines:
   ACTIONS:action1,action2
   DEPTH_OVERRIDE:<integer>
   RELS_OVERRIDE:rel1,rel2,...
Rules:
Always recommend at least one action.
Include all actions relevant to low-scoring metrics.
Output must strictly follow the format above with no
    extra explanation.
Example:
If Novelty=2 and Verifiability=3
ACTIONS:neo4j,pubmed
```

G.7 Refiner agent

The REFINER agent revises and improves a given hypothesis based on received feedback and supplemental information. Specifically, it receives the low-scoring content and the integrated complementary knowledge produced by the REVIEWER agent, and the current hypothesis. It integrates low-scoring metrics and their explanations from the CRITIC Agent, along with new knowledge retrieved by the RE-

VIEWER Agent. Using LLM synthesizes and restructures multi-source inputs to generate a revised hypothesis with enhanced novelty, verifiability, and scientific relevance. The refinement process explicitly targets previously identified weaknesses, and the resulting hypothesis is returned to the evaluation loop for further evaluation.

Listing 7: REFINER agent

```
Improve the current hypothesis based on the provided
    critic feedback, and new information (from Neo4j,
    PubMed, or background). Only make content-
    level changes that directly address the
    weaknesses.

Rules:
1. For each identified weakness, briefly state what is
    missing or imprecise in the hypothesis (one
    sentence per metric).
2. Review all new information:
    If high-quality, relevant content addresses a
        weakness, explain how it helps and revise the
        hypothesis accordingly.
    If no new information directly addresses the
        weaknesses, use relevant new information
        and scientific reasoning and the provided
        background to make a real, meaningful
        improvement, but only if this improvement is
        justified by the context.
    If nothing useful is found, or only stylistic edits are
        possible, clearly state this and leave the
        hypothesis unchanged.
3. Do not rephrase or reword unless it results in a real
    improvement. Do not invent content unsupported
    by evidence.

Example 1 (with helpful new info):
Step 1: The hypothesis lacks a mechanism linking Wnt
    inhibition to reduced fibrosis.
Step 2: New PubMed evidence suggests TGF-$\beta$
    mediates this process.
Step 3: Adding TGF-$\beta$ clarifies the pathway.
Inhibition of Wnt signaling reduces cardiac fibrosis via
    downregulation of TGF-$\beta$ activity.
Example 2 (no helpful info):
Step 1: The hypothesis lacks a mechanistic link.
Step 2: No new information improves this.
Step 3: No justified revision possible.
Overexpression of SOD2 reduces neurodegeneration
    by mitigating oxidative stress in dopaminergic
    neurons.
Output:
    1-4 short reasoning steps (one per line)
    Final refined hypothesis as the last line (no
        numbering, no extra text)
Instructions:
    Use only provided context (background + new info).
    Each reasoning step must be a complete, self-
        contained sentence.
    Do not include explanations, citations, or bullet
        points.
```

H LLM Evaluators

Two LLM evaluation paradigms were used: a *direct* evaluator scores hypotheses on a numerical scale 1–5 for novelty, relevance, significance and verifiability, similar to the CRITIC agent.

Listing 8: Direct evaluator

```
You are a senior biomedical reviewer.

Task:
Evaluate the following hypothesis by assigning a score
    for each metric (Novelty, Relevance, Significance,
    Verifiability) and providing a concise reason.
Metric definitions:
Novelty: Evaluate the novelty of the generated scientific
    hypothesis. The score range should be 0 to 3. 0
    means there's no novelty, which indicates that the
    hypothesis is a paraphrase of the input. 1 means
    there's slight novelty. 2 means there's moderate
    novelty. 3 means the hypothesis has strong
    novelty, which gives new insights beyond the
    background. Output is an integer.
Relevance: Evaluate the relevance of the generated
    scientific hypothesis. The score range should be 0
    to 3. 0 means there's no relevance. 1 means
    there's slight relevance. 2 means there's moderate
    relevance. 3 means they are strongly related.
    Output is an integer.
Significance: Evaluate the significance of the generated
    scientific hypothesis. The score range should be
    0 to 3. 0 means there's no significance, which
    indicates that the hypothesis is just a common
    knowledge. 1 means there's slight significance. 2
    means there's moderate significance. 3 means the
    hypothesis has strong significance, which gives
    significant insights beyond the background. Output
    is an integer.
Verifiability: Evaluate the verifiability of the generated
    scientific hypothesis. The score range should be 0
    to 3. 0 means there's no verifiability, which
    indicates that the hypothesis is not possible to be
    verified in future work. 1 means there's slight
    verifiability. 2 means there's moderate verifiability.
    3 means the hypothesis has strong verifiability,
    which means the hypothesis is very likely to be
    verified in future work. Output is an integer.
User Input: {user input}
Hypothesis: {hypothesis}
```

By contrast, a *pairwise* evaluator compares two hypotheses at a time, and is asked to say which of them is better according to each of the four criteria. Ties are allowed. A Bradley–Terry model was then fitted to the outputs; see Section D.

Listing 9: Pairwise evaluator

```
You are a senior biomedical reviewer. Compare two
    hypotheses A and B on four metrics: Novelty,
    Relevance, Significance, Verifiability.
Instructions:
For each metric, judge and select a winner:
    - "A" if A is clearly superior,
    - "B" if B is clearly superior,
    - "0" if they are equal or difference is unclear.
For each, give a concise reason.
Each metric is judged strictly independently.
Novelty: Evaluate the novelty of two scientific
    hypotheses (A and B) given the user input. For
```

each, assign a novelty score from 0 to 3. 0 means there's no novelty, which indicates that the hypothesis is a paraphrase of the background. 1 means there's slight novelty. 2 means there's moderate novelty. 3 means the hypothesis has strong novelty, which gives new insights beyond the background. Score two hypotheses and compare which one is more novel("A", "B", or "0" if equal or difference is unclear)

Relevance: Evaluate the relevance of two scientific hypotheses (A and B) given the user input. For each, assign a relevance score from 0 to 3. 0 means there's no relevance. 1 means there's slight relevance. 2 means there's moderate relevance. 3 means the hypothesis is strongly related to the background. Score both hypotheses and compare which one is more relevant ("A", "B", or "0" if equal or difference is unclear)

Significance: Evaluate the significance of two scientific hypotheses (H_A and H_B) given the user input. For each, assign a significance score from 0 to 3. 0 means there's no significance, which indicates that the hypothesis is just common knowledge. 1 means there's slight significance. 2 means there's moderate significance. 3 means the hypothesis has strong significance, providing significant insights beyond the background. Score both hypotheses and compare which one is more significant ("A", "B", or "0" if equal or difference is unclear)

Verifiability: Evaluate the verifiability of two scientific hypotheses (H_A and H_B) given the user input. For each, assign a verifiability score from 0 to 3. 0 means there's no verifiability, which indicates that the hypothesis is not possible to be verified in future work. 1 means there's slight verifiability. 2 means there's moderate verifiability. 3 means the hypothesis has strong verifiability, which means it is very likely to be verified in future work. Score both hypotheses and compare which one is more verifiable ("A", "B", or "0" if equal or difference is unclear)

User Input: {user input}

H_A: {hypothesis_a}

H_B: {hypothesis_b}

Supplementary References

Aamer, N.; Asim, M. N.; Munir, S.; and Dengel, A. 2025. Automating AI Discovery for Biomedicine Through Knowledge Graphs And LLM Agents. *bioRxiv*.

Baek, J.; Jauhar, S. K.; Cucerzan, S.; and Hwang, S. J. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6709–6738. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: the method of paired comparisons. *Biometrika*, 39(3-4): 324–345.

Bürkner, P.-C. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of statistical software*, 100: 1–54.

Davidson, R. R. 1970. On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association*, 65(329): 317–328.

Firth, D. 2004. Quasi-variances. *Biometrika*, 91(1): 65–80.

Firth, D. 2025. *qvcalc: Quasi Variances for Factor Effects in Statistical Models*. R package version 1.0.4.

Ghafarollahi, A.; and Buehler, M. J. 2024. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advanced Materials*, 2413523.

Gottweis, J.; Weng, W.-H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; Saab, K.; Popovici, D.; Blum, J.; Zhang, F.; Chou, K.; Hassidim, A.; Gokturk, B.; Vahdat, A.; Kohli, P.; Matias, Y.; Carroll, A.; Kulkarni, K.; Tomasev, N.; Guan, Y.; Dhillon, V.; Vaishnav, E. D.; Lee, B.; Costa, T. R. D.; Penadés, J. R.; Peltz, G.; Xu, Y.; Pawlosky, A.; Karthikesalingam, A.; and Natarajan, V. 2025. Towards an AI co-scientist. arXiv:2502.18864.

Kobilka, B. K.; and Lefkowitz, R. J. 2024. G Protein-Coupled Receptors: A Century of Research and Discovery. *Circulation Research*, 134(5): 671–693.

Qi, B.; Zhang, K.; Tian, K.; Li, H.; Chen, Z.-R.; Zeng, S.; Hua, E.; Jinfang, H.; and Zhou, B. 2024. Large Language Models as Biomedical Hypothesis Generators: A Comprehensive Evaluation. In *First Conference on Language Modeling*.

Selby, D. A. 2020. *Statistical modelling of citation networks, research influence and journal prestige*. Ph.D. thesis, University of Warwick.

Shao, J.; Kwon, J.; Wang, T.; Günther, S.; Tombor, L. S.; Warwick, T.; Shaheryar, Z.; Brandes, R. P.; Dimmeler, S.; Wenzel, J.; et al. 2025. Orphan receptor GPR153 facilitates vascular damage responses by modulating cAMP levels, YAP/TAZ signaling, and NF-κB activation. *Nature Communications*, 16(1): 6232.

Turner, H.; and Firth, D. 2012. Bradley–Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software*, 48(9): 1–21.