
Supplementary information

**A multi-agent system for automating
scientific discovery**

In the format provided by the
authors and unedited

Supplementary Information

A multi-agent system for automating scientific discovery

Ali Essam Ghareeb ^{1,*}, Benjamin Chang ^{1,2,*}, Ludovico Mitchener ¹, Angela Yiu ¹, Caralyn J. Szostkiewicz ¹, Dmytro Shved ³, Gavin J. Gyimesi ³, Jon M. Laurent ¹, Samantha M. Wright ¹, Muhammed T. Razzak ¹, Andrew D. White ^{1,†}, Silvia C. Finnemann ³, Michaela M. Hinks ^{1,‡}, and Samuel G. Rodriques ^{†,‡}

¹ FutureHouse, San Francisco, USA

² University of Oxford, Oxford, UK

³ Fordham University, New York, USA

* Equal contribution.

† These authors jointly supervised work at FutureHouse.

‡ These authors jointly supervised this work.

Contents

Supplementary Tables 1-2

Supplementary Figures 1-20

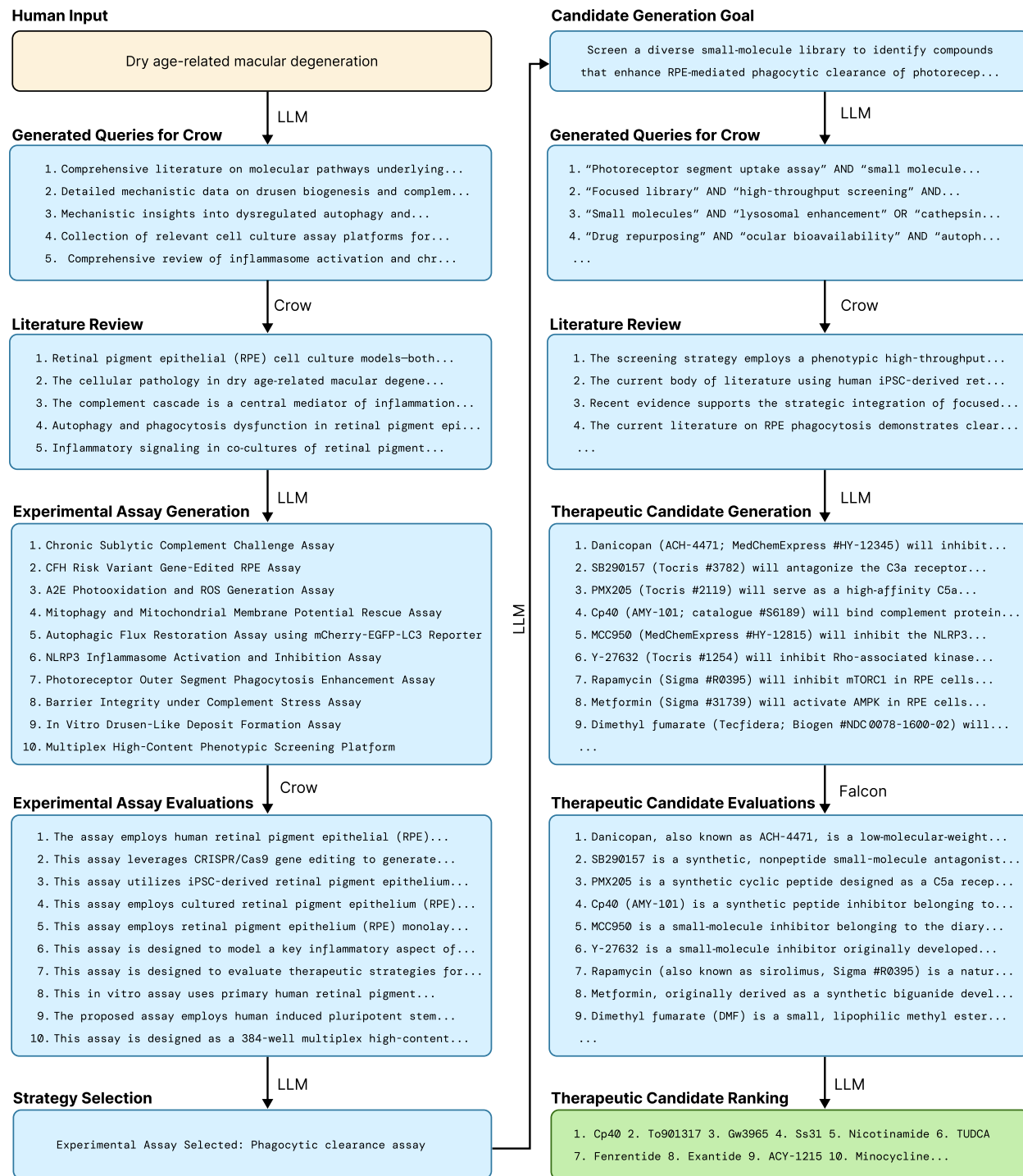
Supplementary References

Supplementary Table 1: List of Drugs and Working Concentrations

Drug	Working concentration
Y-27632 (Cayman Chemical Company; Cat. No. 10005583)	20 μ M [63]
AICAR (Sigma-Aldrich; Cat. No. A9978)	1mM [64]
Tauroursodeoxycholic acid (TUDCA) (TargetMol; Cat. No. T2532)	100 μ M [65]
Exendin-4 (TargetMol; Cat. No. T3967)	1 μ M [66]
Fingolimod (PeproTech; Cat. No. 1625605)	1 μ M [67]
MFGE-8 (Sino Biological; Cat. No. 10853-H27B-B-20)	2.5 μ g/mL [68]
Ripasudil (Cayman Chemical Company; Cat. No. 19920)	100 μ M [69]
Vorinostat (Active Motif; Cat No. 14027)	5 μ M [70]
Simvastatin (TargetMol; Cat No. T0687)	1 μ M
Resveratrol (PeproTech; Cat No. 5013606)	16mM [71]
Bay K8644 (Alomone Labs; Cat No. B-350)	3 μ M [72]
KL001 (Cayman Chemical Company; Cat No. 13878)	50 μ M [73]
Rapamycin (a gift from RetroBio)	100nM [74]
BMS-777607 (ApexBio; Cat No. A5703)	10 μ M [75]
NECA (5'-N-ethylcarboxamidoadenosine; Cayman Chemical Company; Cat No. 21420)	10 μ M [76]
Rolipram (PeproTech; Cat No. 6145453)	10 μ M
PF-573228 (ApexBio; Cat No. B1523)	5 μ M [77]
4-phenylbutyric acid (ApexBio; Cat No. C6831)	10mM [78]
B-Ionone (AA Blocks; Cat No. AA003JVY)	200 μ M [79]
Siponimod (TargetMol; Cat No. T6403)	10mM [80]
Dimethyl fumarate (Sigma-Aldrich; Cat. No. 242926-25G)	30 μ M [81]
Flubendazole (AA Blocks; Cat. No. AA003QIR)	10 μ M [82]
GW3965 (TargetMol; Cat. No. T6310)	5 μ M [83]
GW501516 (Sigma-Aldrich; Cat. No. SML1491-5MG)	100 μ M [84]
Metformin (Cayman Chemical Company; Cat. No. 13118)	16mM [85]
ML-SA1 (Sigma-Aldrich; Cat. No. SML0627-25MG)	30 μ M [86]
Nicotinamide (Sigma-Aldrich; Cat. No. N3376-100G)	10mM [87]
Pioglitazone (ApexBio; Cat. No. B2117)	10 μ M [88]
PROS1 (R&D Systems; Cat. No. 9489-PS-100)	0.56 μ g/mL [89]
Ramoplanin (AA Blocks; Cat. No. AA005FC1)	2 μ g/mL [90]
Soraprazan (MedChem Express; Cat. No. HY-100414-1mg)	30 μ g/mL [91]
L-Sulforaphane (Isothiocyanate, Sigma-Aldrich S6317)	20 μ M [92]
TAPI-2 (Abcam; Cat. No. ab142219)	20 μ M [93]
Trehalose (Sigma-Aldrich; Cat. No. T9449-100G)	100mM [94]

Supplementary Table 2: Antibodies used in this study.

Antibody	Target	Vendor	Catalog number	Dilution
Bestrophin-1 Monoclonal antibody	BEST1	Proteintech	60326-1-Ig	1:200
MERTK Polyclonal antibody	MERTK	Proteintech	27900-1-AP	1:200
Anti-Integrin α V β 5 Antibody, clone EM09902, ZooMAb [®] Rabbit Monoclonal	α V β 5	Sigma-Aldrich	ZRB1191-25UL	1:100
Anti-RPE65 Rabbit Polyclonal Antibody	RPE65	Proteintech	17939-1-AP	1:100
Anti-MiTF antibody [EPR26363-10]	MiTF	Abcam	ab303530	1:50
Otx2 Polyclonal Antibody	OTX2	Thermo Fisher Scientific	PA5-39887	1:20
Donkey Anti-Mouse IgG H&L (Alexa Fluor 647)	Mouse IgG	Abcam	ab150107	1:500
Goat Anti-Rabbit IgG H&L (Alexa Fluor 647)	Rabbit IgG	Abcam	ab150079	1:500
Donkey Anti-Goat IgG H&L (Alexa Fluor 647), preadsorbed	Goat IgG	Abcam	ab150135	1:500



Supplementary Figure 1 Detailed example workflow of Robin experimental assay and therapeutic candidate hypothesis generation.

Model: o4-mini

System prompt:

You are a highly experienced biomedical research strategist. Your task is to propose **mechanistic hypotheses** tested via **cell-culture assays** or **therapeutic strategies**. Generate exactly {num_assays} distinct ideas.

User prompt:

Return a list of {num_queries} queries (separated by <>) that would be useful in doing research to develop detailed, mechanistic cell culture assays that would be used to evaluate drugs to treat {disease_name}. These queries will be given to a 20-person scientific team to research in depth, so they should be able to capture broadly relevant information (30+ words) and search relevant literature across biomedical, clinical, and biochemical literature about the disease or therapeutic landscape. Don't look up specific drugs, but any relevant scientific information that may help with assay development. You have {num_queries} queries, so spread your queries out to cover as much ground as possible. Create queries both about the general biochemistry and mechanistic underpinnings of {disease_name} as well as about the assays. In formatting, don't number the queries, just output a string with {num_queries} queries separated by <>.

Supplementary Figure 2 Prompt for generating literature search queries.

Model: o4-mini

System prompt:

You are a professional biomedical researcher, having experience in early-stage drug discovery and validation in vitro. Your task is to propose **mechanistic hypotheses** tested via **cell-culture assays** or **therapeutic strategies**. Generate exactly {num_assays} distinct assay ideas. You are meticulous, creative, and scientifically rigorous. Focus on strategies that prioritize simplicity, speed of readout, biological relevance, and direct measurement of functional endpoints. Strong preference for biologically relevant strategies.

Output Format Specification (Strict Adherence Required):

Your entire output **MUST** be a single, valid JSON object. This JSON object will be an **array** at its root, containing exactly {num_assays} individual JSON objects. Each of these inner objects represents one distinct proposal and **MUST** conform to the following structure and content guidelines:

```
[
  {
    "strategy_name": "string", # Name of the strategy. Keep this name simple, don't
    ↪ include details about how specific mechanisms or specific methodology.
    "reasoning": "string" # Scientific reasoning justifying the chosen strategy or
    ↪ the feasibility/relevance of the assay design, citing relevant literature.
  }
  // ... more objects here, up to {num_assays} total
]
```

User prompt:

Generate exactly **{num_assays}** distinct and scientifically rigorous proposals for cell culture assays that can evaluate drugs to treat {disease_name}. Here is some relevant background information that can guide your proposals: {assay_lit_review_output}

Supplementary Figure 3 Prompt for generating hypotheses for experimental assays.

Model: Crow

You are a professional biomedical researcher, having experience in early-stage drug discovery and validation in vitro. Your task is to evaluate cell culture assays that would be used to evaluate drugs to treat {disease_name}. Given the following assay, do a comprehensive literature review to evaluate if this assay would be useful for testing therapeutics for {disease_name}. Search relevant literature across biomedical, clinical, and biochemical literature about the disease or therapeutic landscape. Don't look up specific drugs, but any relevant scientific information that may inform assay development. {Strategy}

Provide your response in the following format, like an evaluation for a scientific proposal:

Assay Overview: Explain the assay idea, including the following key points: which aspect of the disease pathogenesis does the assay model, what measurements will be taken from the assay and how they will be taken, which cells or other biological material are used in the assay.

Biomedical Evidence: Make a compelling argument for how the aspect of the disease represented in the assay is central to the pathogenesis of the disease. Make sure to consider both the biomedical and clinical literature.

Previous Use: Explain how this assay has previously been used for drug discovery (if this has been done). Explain any key scientific discoveries which have been made using this assay.

Overall Evaluation: Strengths and weaknesses of this assay for testing therapeutics for {disease_name}.

Supplementary Figure 4 Prompt for generating detailed evaluation for each experimental assay

Model: Claude Sonnet 3.7

System prompt:

You are an experienced drug development committee member, with broad scientific expertise across the biology, chemistry, clinical, medical, and pharmaceutical sciences.

Your objective is to perform a rigorous scientific comparison of two proposals for experimental assays that will be used to test therapeutics for {disease_name}.

Your evaluation must be based strictly on the presented scientific evidence, scientific novelty, methodological rigor, and logical interpretation, not on the persuasive quality or wording of the proposal documents.

Critically assess the scientific soundness and biological rationale for the experimental assay, analyzing the supporting literature and historical usage.

Preference for in vitro strategies that prioritize simplicity, speed of readout, biological relevance, and direct measurement of functional endpoints. Strong preference for biologically relevant strategies.

The goal of this task is to choose the best in vitro experimental assay that would be scientifically relevant and insightful for testing therapeutics for {disease_name}. Prefer assays that are biologically functionally relevant and simple to perform in standard lab setting.

User prompt:

Evaluate the experimental assays using the structure below. This evaluation informs a critical decision.

****Respond ONLY in the specified JSON format, do not include any text outside the JSON object itself.****

```
{{
  "Analysis": {"type": "string", "description": "[Provide a detailed analysis of the
    ↳ two experimental assays, based on the evaluation criteria and the evidence
    ↳ provided.]"},
  "Reasoning": {"type": "string", "description": "[Choose which experimental assay is
    ↳ better. Provide a detailed explanation for why you think the winner is better
    ↳ than the loser, based on the evaluation criteria and the evidence provided.]"},
  "Winner": {"type": "string", "description": "[Return the name and ID number of the
    ↳ candidate that you think is better between the two candidates, as a tuple. It
    ↳ should be formatted as (winner_name, winner_id)]"},
  "Loser": {"type": "string", "description": "[Return the name and ID number of the
    ↳ candidate that you think is worse between the two candidates, as a tuple. It
    ↳ should be formatted as (loser_name, loser_id)]"},
}}
```

Supplementary Figure 5 Prompt for LLM judge to select the top experimental assay

Model: o4-mini

System prompt:

You are a biomedical researcher briefly explaining a strategy of how to identify novel therapeutic compounds to test using this assay to treat {disease_name}.

User prompt:

Here is a proposed experimental assay identified for treating '{disease_name}':

Assay Name: "{assay_name}"

Synthesize a concise and specific research goal for the *next* stage, which is focused on ****identifying novel therapeutic compounds**** to test using this assay to treat {disease_name}.

It's important that you connect the goal of this assay to how it is important for ****identifying**** novel therapeutic compounds to treat {disease_name}.

Provide **ONLY** the synthesized goal string as the response.

Supplementary Figure 6 Prompt for synthesizing goal for candidate generation

Model: o4-mini

System prompt:

You are an expert drug development researcher focused on generating high-quality, specific, testable **drug candidates**.

Your goal is to propose novel, single-agent drug candidates (specific molecules, potentially repurposed, that are commercially available; mention catalog numbers if possible).

You are interested in finding candidates with:

1. **Strong Target Validation:** The target pathway/mechanism has robust evidence (genetic, functional) linking it *specifically* to the disease pathophysiology.
2. **Relevant Preclinical/Clinical Evidence:** Supporting data exists from disease-relevant models or, ideally, preliminary human data (even if from related conditions or pilot studies).
3. **Mechanistic Specificity:** A clear, well-defined molecular mechanism is preferable over broad, non-specific actions.
4. **Novelty (Balanced with Validation):** Innovative, exciting, and novel approaches that can advance treatment for {disease_name}, that are also grounded in strong scientific rationale and evidence.
 - Focus on compounds that are commercially available (mention catalog numbers) and can be developed into drugs
 - Favor mechanisms with minimal impact on other cellular functions
 - Compounds should be novel for treatment of the disease and ideally address novel targets for the diseased cell type (i.e. not tested in prior studies)

User prompt:

Return a list of {num_queries} queries (separated by <>) that would be useful in doing background research for the goal of {candidate_generation_goal}.

These queries will be given to a highly-trained research team to investigate the scientific literature in depth, so the queries should be able to capture broadly relevant information (30+ words) and search relevant literature across biomedical, clinical, and biochemical literature about the disease or therapeutic landscape. You don't need to propose specific drugs, but the queries should be able to capture relevant scientific information that may help with proposing drug candidates.

You have {num_queries} queries, so spread your queries out to cover as much ground as possible. Generate {num_queries} queries to cover literature about the therapeutic landscape, especially those that can help {candidate_generation_goal} and generate {num_queries} queries to cover literature about the biological and mechanistic aspects about {disease_name} itself. In formatting, don't number the queries, just output a string with {num_queries} queries separated by <>.

Generate queries for the goal of {candidate_generation_goal} that actively seek and prioritize:

- * **Target Validation:** Studies demonstrating the target pathway's dysregulation or causative role.
- * **Efficacy in Relevant Models:** Evidence of the drug candidate's efficacy in cell or animal models that *closely mimic* disease pathology, or in patient-derived cells. Prefer this over general mechanism-of-action studies.
- * **Mechanism Confirmation:** Studies confirming the candidate engages the target and modulates the specific pathway *as hypothesized* in a relevant biological system.
- * **Pharmacokinetics/Safety Data:** Existing data on the candidate's ADME properties, safety profile (especially human data), and known ability to reach relevant tissues.

Supplementary Figure 7 Prompt for generating queries for literature search

Model: o4-mini

System prompt:

You are an expert drug development researcher focused on generating high-quality, specific, testable **drug candidates**. Your task is to generate novel, testable therapeutic candidates for {disease_name}, based on the provided research goal and background literature.

You are interested in discovering and proposing candidates with:

1. **Strong Target Validation:** The target pathway/mechanism has robust evidence (genetic, functional) linking it *specifically* to the disease pathophysiology.
2. **Relevant Preclinical/Clinical Evidence:** Supporting data exists from disease-relevant models or, ideally, preliminary human data (even if from related conditions or pilot studies).
3. **Developmental Feasibility:** The candidate has known properties that facilitate development.
4. **Mechanistic Specificity:** A clear, well-defined molecular mechanism is preferable over broad, non-specific actions.
5. **Novelty (Balanced with Validation):** Innovative approaches are valued, but must be grounded in strong scientific rationale and evidence.

For EACH hypothesis object in the `hypotheses` array, you MUST provide ALL of the following fields:

1. `candidate`: The specific drug/therapeutic proposed. Must be a single agent, not a combination. Do not propose special formulations or delivery methods.
2. `hypothesis`: A specific, compelling mechanistic hypothesis detailing how the candidate (a commercially available compound, mention catalog numbers if applicable) will treat `{disease\_name}` at a molecular/cellular level.
3. `reasoning`: Detailed scientific reasoning, explaining the mechanistic rationale, evidence, translational considerations, target validation, and novelty of the candidate.

Output Format Specification (Strict Adherence Required):

Your entire output MUST be a text block. Generate exactly {num_candidates} candidate proposals.

Each candidate proposal MUST start with "<CANDIDATE START>" on a new line and end with "<CANDIDATE END>" on a new line.

Within each block, each piece of information (CANDIDATE, HYPOTHESIS, REASONING,) MUST start on a new line, beginning with its EXACT header (e.g., `CANDIDATE:`, `HYPOTHESIS:`) followed by the content.

Do NOT include any other text before the first "<CANDIDATE START>" or after the last "<CANDIDATE END>".

Example for one candidate (repeat this block structure {num_candidates} times):

<CANDIDATE START>

CANDIDATE: The specific drug/therapeutic proposed. Must be a single agent, not a combination. Do not propose special formulations or delivery methods.

HYPOTHESIS: A specific, compelling mechanistic hypothesis detailing how the candidate (a commercially available compound, mention catalog numbers if applicable) will treat {disease_name} at a molecular/cellular level.

REASONING: Detailed scientific reasoning, explaining the mechanistic rationale, evidence, translational considerations, target validation, and novelty of the candidate.

<CANDIDATE END>

User prompt:

Generate exactly **{num_candidates}** distinct and scientifically rigorous proposals for therapeutic candidates to treat {disease_name}. Here is some relevant background information that can guide your proposals: {therapeutic_candidate_review_output}

Supplementary Figure 8 Prompt for generating therapeutic candidate hypotheses

Model: Falcon

You are an expert drug development team leader focused on validating **drug candidates**. Your task is to evaluate promising repurposed drugs or therapeutic candidates for {disease_name} proposed by your research team.

Given the following therapeutic candidate, do a comprehensive literature review to evaluate if this therapeutic candidate has potential for {disease_name}. Search relevant literature across

biomedical, clinical, and biochemical literature about the disease or proposed therapeutic. This must be very detailed and comprehensive, as it will determine the direction of the team.

{Candidate}

Provide your response in the following format, like an evaluation for a scientific proposal:

Overview of Therapeutic Candidate: Explain the natural or synthetic origins of this therapeutic candidate, including how it was synthesized or discovered. Explain which class of therapeutic compounds this belongs to, and how this class of compounds has previously been used in general.

Therapeutic History: Summarize previous biochemical, clinical or veterinary uses of this drug or drug class, if any. Examine to see if the therapeutic candidate has ever been used for treating {disease_name} or any similar disease.

Mechanism of Action: Explain the known mechanism(s) of action of this compound to the full extent of molecular detail that is known. Explain the biochemistry and molecular interactions of the therapeutic candidate in any relevant literature.

Expected Effect: Explain the expected effect of this compound in the assay proposed and the mechanism by which it will act. If the drug is acting on proteins, reference literature which shows this gene is expressed in the cell type being assayed.

Overall Evaluation: Give your overall thoughts on this therapeutic candidate. Include strengths and weaknesses of this therapeutic candidate for treating {disease_name}.

Supplementary Figure 9 Prompt for generating detailed investigation for each therapeutic candidate

Model: Finch

Analyze the zipped .fcs files from three 96-well plates to measure average Alexa Fluor 647 signal in live retinal pigment epithelial cells across different drug treatments.

Use the metadata file to match well positions (e.g., A1) with their corresponding drug treatments and plate information. Use fcsparser to parse the files.

Create appropriate gates using the "DMSO control + beads + cells + DAPI" well to isolate the live, singlet cell population. Make the necessary plots to identify the required gating thresholds. I recommend using polygon coordinates to remove debris, dead cells, and doublets or aggregates. Make sure to exclude all debris, and that the singlet gating is not too strict. Once you have the best gating strategy possible that gives reasonable percentage gated events, check the gating on the other two control wells to make sure comparable percentage gated events.

Determine if there are plate-to-plate effect, if there is, take appropriate normalization steps using the DMSO control before downstream analysis.

Measure the Alexa Fluor 647 signal intensity in these cells, and compile the results into a CSV file.

Determine if there are any drug treatments that significantly increased signal compared to control. I recommend using the Dunnett test to compare the drug treatments to the DMSO control. Generate a barplot showing the fold change in MFI relative to the DMSO control. Include error bars showing the standard error of the mean, and highlight any drugs that significantly increase Alexa Fluor 647 signal using asterisks' to denote significance. Exclude any control treatments in the final plot.

Do not stop early. Make sure you have processed all files available.

Supplementary Figure 10 Prompt for performing flow cytometry analyses from zipped .fcs files

System prompt:

You are an experienced drug development committee member, with broad scientific expertise.

Your objective is to rigorously compare two preclinical drug candidate proposals. Your primary goal is to select the hypothesis with the **highest probability of successful experimental outcome AND eventual translation into a viable therapy** for {disease_name}.

Your evaluation must be based strictly on the presented scientific evidence, scientific novelty, methodological rigor, and logical interpretation. Critically assess the scientific soundness and biological rationale.

Prioritize your evaluation based on these key criteria, reflecting human expert preferences:

1. **Strength and Relevance of Supporting Evidence (Highest Priority):**

* **Existing Data:** Is there robust existing data (preclinical, clinical, especially for {disease_name} or highly relevant biological contexts) supporting the hypothesis? Give significant weight to positive data from later-stage research (e.g., in vivo, clinical trials for the drug or drug class, especially if successful for related conditions).

* **Relevance to the Problem:** Is the evidence *directly relevant* to the proposed mechanism in the context of {disease_name} and the specific biological problem being addressed? General efficacy in unrelated conditions is less compelling.

* **Negative Evidence:** Are there known failures, lack of efficacy, or contradictory data for this drug/class in relevant contexts? This is a strong deterrent.

* **Quality of References:** Are the cited references (if any are explicitly mentioned or implied by the reasoning) strong and relevant to the drug and its proposed action in this context?

2. **Mechanism of Action (MoA) - Clarity, Plausibility, and Specificity for {disease_name}:**

* **Clarity & Detail:** Is the MoA clearly articulated, scientifically sound, detailed, and plausible? Vague mechanisms are less favored.

* **Directness:** How *direct* is the MoA to addressing the core pathology of {disease_name}? Prefer hypotheses targeting core, upstream issues relevant to the disease's cellular or molecular basis over indirect or downstream effects, unless the indirect effect is exceptionally well-supported and highly plausible.

* **Specificity & Target Biology:** Is the therapeutic target well-defined? Is its biology central to the pathogenesis of {disease_name}? Specific pathway/target engagement is preferred over overly broad actions that might increase risk.

3. **Safety, Tolerability, and Risk Profile (High Priority):**

* **Known Safety Profile:** What is the known safety profile of the drug or drug class? Candidates with established clinical safety (e.g., repurposed drugs with a *strong rationale* for {disease_name}) are highly favored.

* **Off-Target Effects/Toxicity:** Are there significant concerns about off-target effects, general toxicity, or specific organ/system toxicity relevant to {disease_name} or its treatment? This is a major red flag. Assess potential for on-target and off-target liabilities.

* **Pleiotropic Effects:** Consider if broad (pleiotropic) effects are beneficial or detrimental in this specific disease context.

4. **Feasibility of Experimental Plan and Drug Delivery:**

* **Methodological Rigor:** Does the proposal (even if brief) outline a clear, methodologically sound, and feasible experimental plan? Are specific, measurable outcomes related to the therapeutic goal for {disease_name} defined?

* **Model System Relevance:** Is the proposed model system (e.g., relevant cell lines, animal models of {disease_name}) appropriate and predictive?

* **Drug Delivery to Target:** Crucially, assess the feasibility of drug delivery to the target tissue or system. Consider physicochemical properties and ADME/PK profiles (bioavailability, metabolic stability) in terms of achieving therapeutic concentrations with an acceptable therapeutic window. Is the proposed route of administration plausible and supported for this drug type and disease?

5. **Scientific Novelty (Balanced with Evidence and Safety):**

* Does the hypothesis offer a novel scientific advancement for {disease_name}? Novelty is valued, but *only if supported by plausible science and a reasonable safety outlook*. Do not prefer novelty if it comes at the cost of strong evidence for a more established, safer approach. An innovative approach to an underexplored but plausible pathway can be positive.

Synthesize these factors:

While detailed pharmacological properties (in vitro potency, selectivity, SAR, chemical liabilities) and ADME/PK parameters (CYP inhibition/induction, DDI risk) are important, ****focus on their *implications* for efficacy, safety, and translatability, as a human expert panel would.**** Avoid getting lost in excessive technical detail unless it directly and significantly impacts one of the prioritized criteria (e.g., a PK issue preventing target engagement would be critical).

The goal of this task is to choose the best therapeutic candidate that has the most potential for treating {disease_name}.

User prompt:

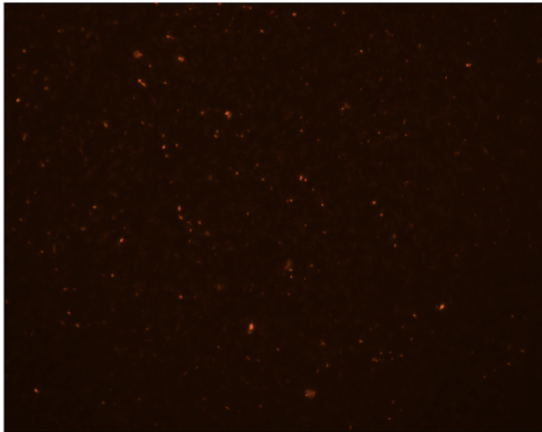
Evaluate the preclinical drug candidate data package using the structure below. This evaluation informs a critical decision. ****Respond ONLY in the specified JSON format, do not include any text outside the JSON object itself.****

```
{{
  "Analysis": {"type": "string", "description": "[Provide a detailed analysis of
  ↳ the two drug candidates, based on the evaluation criteria and the evidence
  ↳ provided.]"},
  "Reasoning": {"type": "string", "description": "[Choose which drug candidate is
  ↳ better. Provide a detailed explanation for why you think the winner is
  ↳ better than the loser, based on the evaluation criteria and the evidence
  ↳ provided.]"},
  "Winner": {"type": "string", "description": "[Return the name and ID number of
  ↳ the candidate that you think is better between the two candidates, as a
  ↳ tuple. It should be formatted as (winner_name, winner_id)]"},
  "Loser": {"type": "string", "description": "[Return the name and ID number of
  ↳ the candidate that you think is worse between the two candidates, as a
  ↳ tuple. It should be formatted as (loser_name, loser_id)]"},
}}
```

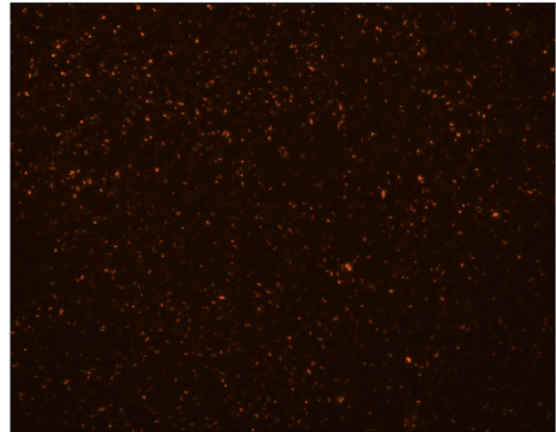
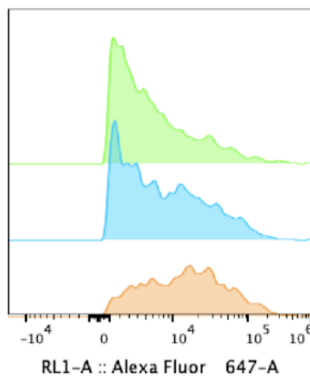
Supplementary Figure 11 Prompt for LLM judge to rank therapeutic candidate hypotheses

a

Baseline phagocytosis

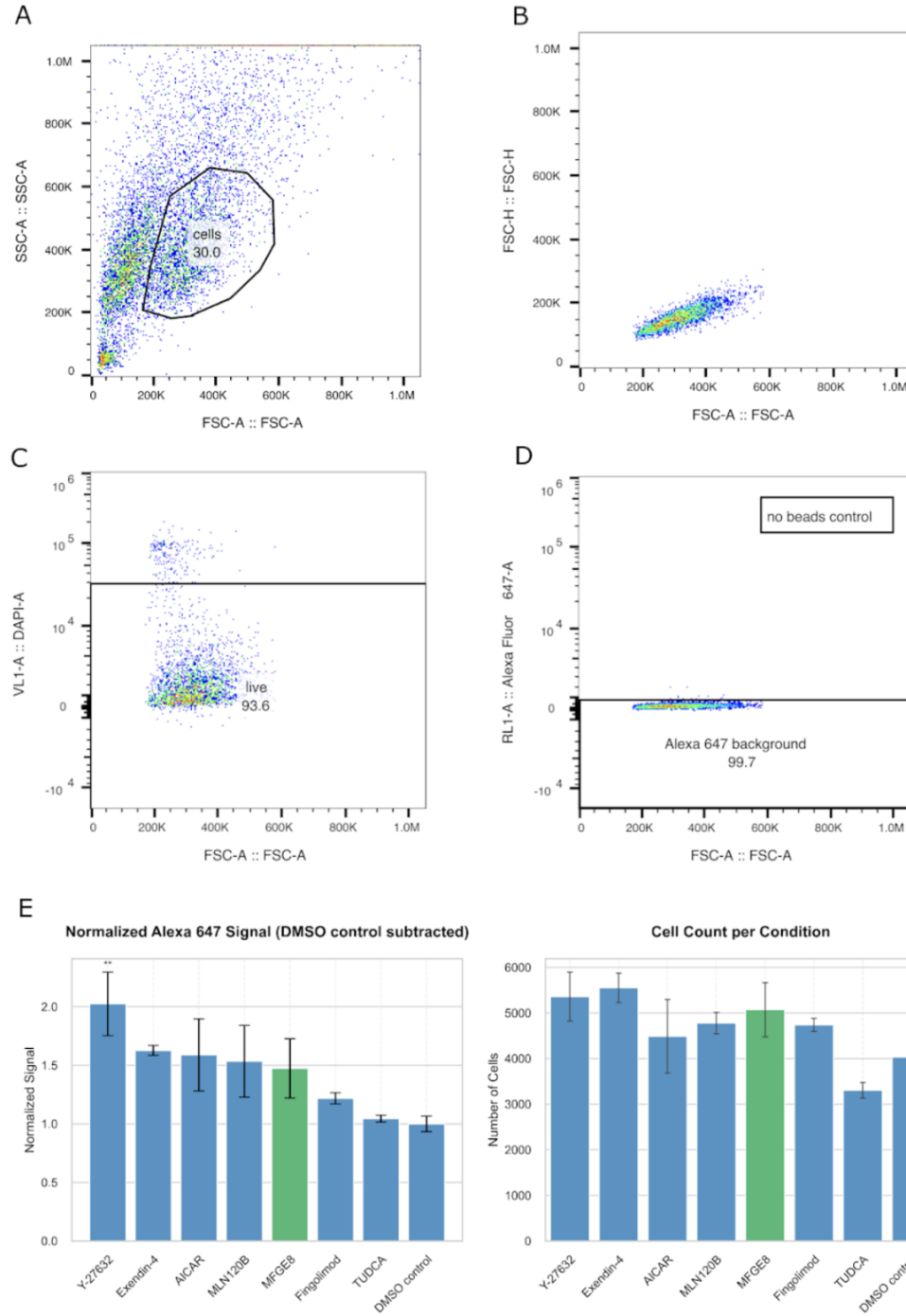


MFGE8 treated cells

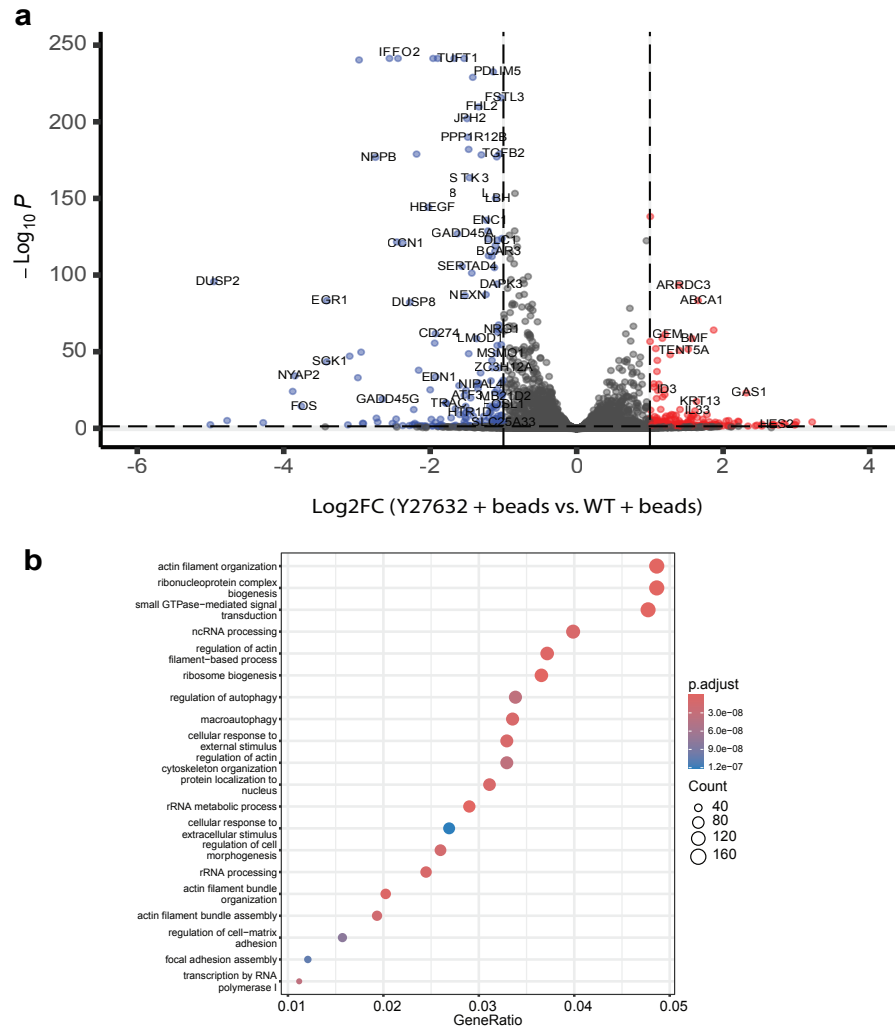
**b**

	Count	Condition
	1923	DMSO control
	2041	MFGE8
	1363	Y-27632

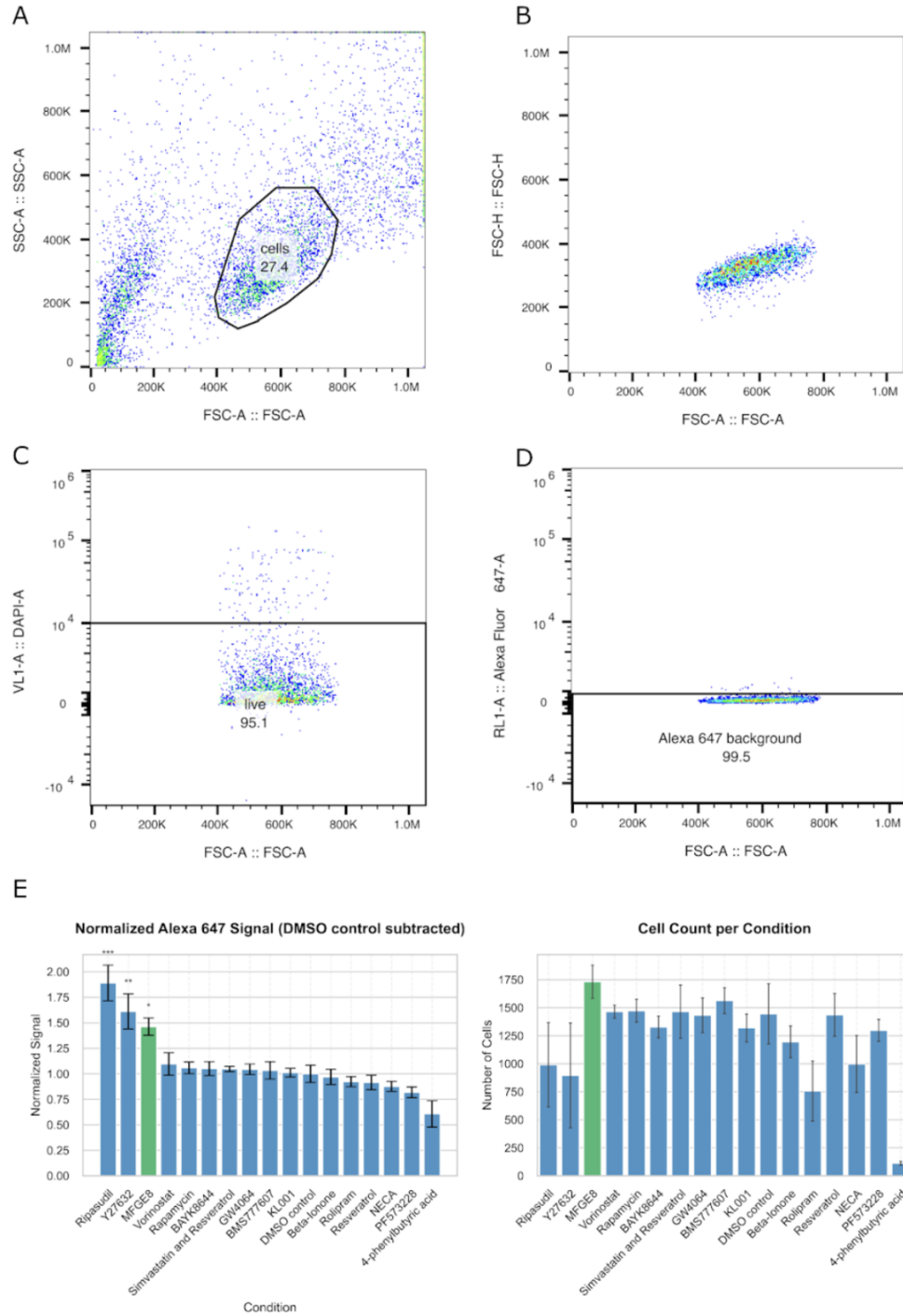
Supplementary Figure 12 (A) Representative fluorescence microscopy images of ARPE-19 cells treated with DMSO vehicle (left) or MFGE8 (positive control; right). (B) Flow cytometry histograms showing the fluorescence intensity of ARPE-19 cells treated with the DMSO vehicle, MFGE8 or the ROCK inhibitor Y-27632 (n=3).



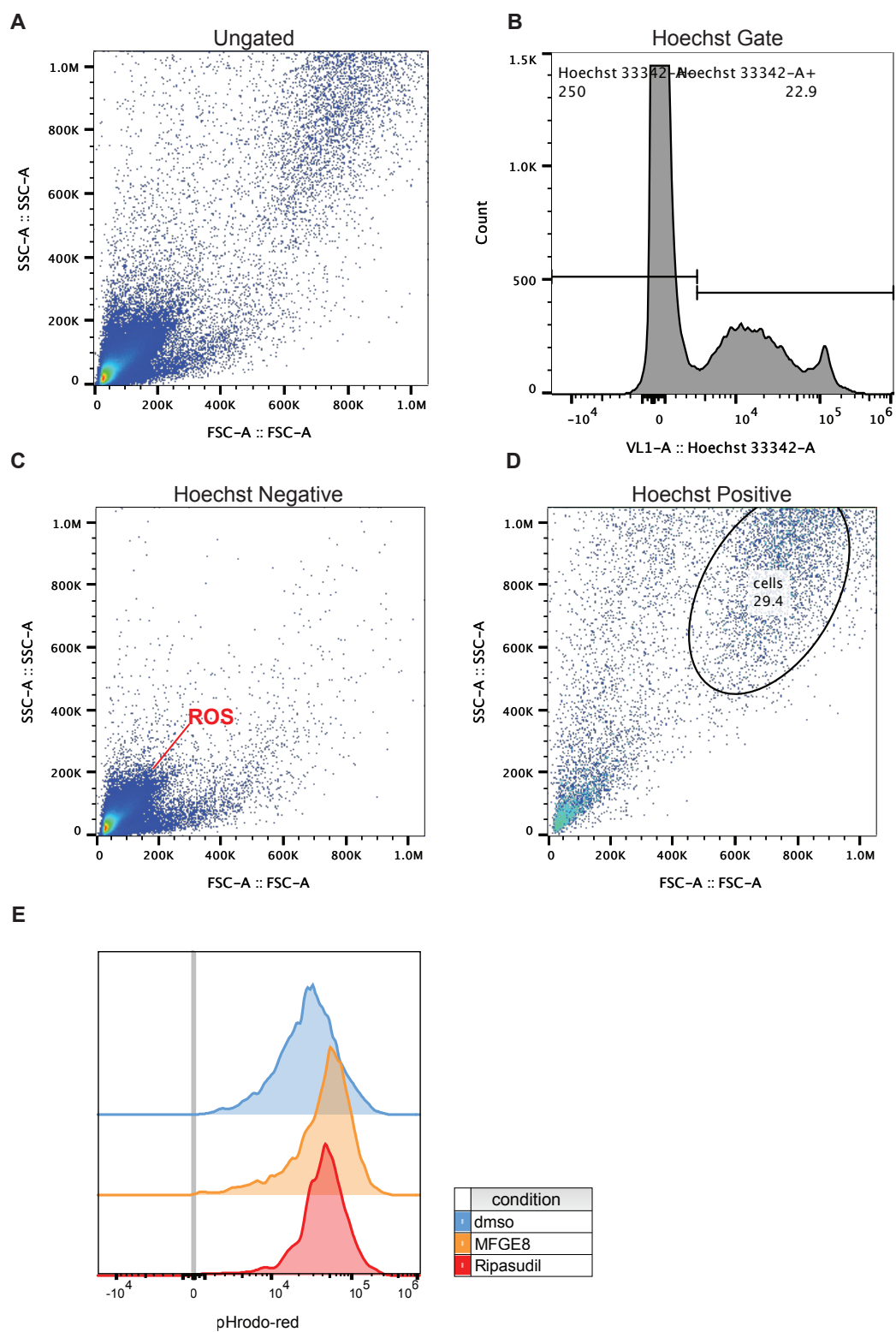
Supplementary Figure 13 Human analysis of round 1 flow cytometry data (ARPE-19 cells). (A-D) gating strategy. (A) The main cell population is identified. (B) Singlets are identified. (C) Dead cells are gated out (D) Alexa 647 background is gated out. (E) Barplots showing normalized gross alexa 647 signal, (DMSO control subtracted) and cell count per condition (after passing all gates). Error bars show the standard error of the mean. Asterisks denote statistical significance: p<0.05 (*), p<0.01 (**), p<0.001 (***) (Dunnett's test).



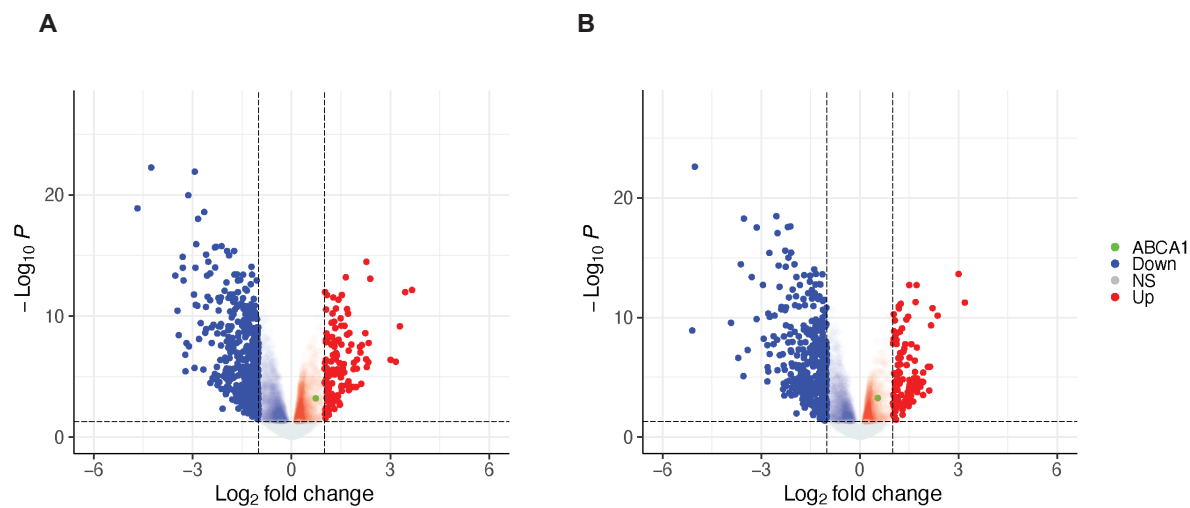
Supplementary Figure 14 Human analysis of the RNA sequencing experiment (A) Volcano plot showing differential gene expression between ARPE-19 cells treated with Y-27632 versus vehicle control (DMSO). The x-axis represents $\log_2$ fold change in expression, with positive values indicating up-regulation and negative values indicating down-regulation. The y-axis shows $-\log_{10}$ of the p-value for each gene, highlighting statistical significance. Genes that pass the significance threshold ($|\log_2FC| > 1$, $p < 0.05$) are colored (red for up-regulated, blue for down-regulated), while non-significant genes are displayed in gray. (B) Gene Ontology (GO) enrichment analysis (using biological process terms) of differentially expressed genes performed using the ReactomePA package. The y-axis lists significantly enriched GO terms, while the x-axis shows the gene ratio, calculated as the number of input genes associated with each GO term divided by the total number of input genes. Dot size corresponds to the number of genes mapped to each term, and color indicates the adjusted p-value, reflecting statistical significance of the enrichment.



Supplementary Figure 15 Human analysis of round 2 flow cytometry data (ARPE-19 cells). (A-D) gating strategy. (A) The main cell population is identified. (B) Singlets are identified. (C) Dead cells are gated out (D) Alexa 647 background is gated out. (E) Barplots showing normalized gross alexa 647 signal (relative to DMSO control) and cell count per condition (after passing all gates). Error bars show the standard error of the mean. Asterisks denote statistical significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***) (Dunnett's test).



Supplementary Figure 16 Flow cytometry gating strategy for RPE-SC phagocytosis experiments (performed by a human). (A) Ungated scatter plot (B) Histogram showing Hoechst 33342 gating strategy used to exclude bovine ROS (C) Scatter plot showing the Hoescht negative population which contains the ROS. (D) Scatter plot showing the Hoechst positive population which contains the cell population used to produce the histograms in (E) which shows an increase in the pHrodo-red signal in the positive control Milk Fat Globule EGF Factor 8 (MFGE8) and ripasudil.



Supplementary Figure 17 Volcano plots showing differential gene expression in RPE-SC cells treated with ripasudil (red) versus DMSO control (blue). ABCA1 is modestly upregulated in ripasudil treatment with or without (left and right, respectively) exposure to ROS (n=4; log2 fold changes with and without ROS are -0.55 and -0.74, respectively; p-values are 5.11×10^{-5} and 8.08×10^{-5} , respectively). *ABCA1* is shown in green.

Model: Finch (flow cytometry analysis)

Rubric for flow cytometry (grading criteria):

1. The gates identify the main cell population gate (FSC vs SSC).
2. The gates distinguish singlets from doublets (for example, FSC-A vs FSC-H).
3. The gates distinguish live cells from dead cells using the DAPI channel.
4. The gates clearly show each step of the gating strategy using high quality dot plots and/or histograms.
5. The appropriate tool was used to delineate a cell population (e.g., polygon coordinates, threshold, clustering, etc.).
6. The gating strategy removes almost all debris and dead cells while retaining the main singlet live cell population.
7. The data were normalized for plate-to-plate effects.
8. The mean fluorescence intensity of the final cell population of each replicate was calculated.
9. The mean and standard deviation of the replicates were calculated.
10. One-tailed tests were used when looking only for increases.
11. The appropriate multiple testing correction was applied if required.
12. All treatments were compared to the vehicle control.
13. The drugs (if any) that enhance mean fluorescence intensity significantly above the vehicle control were identified.

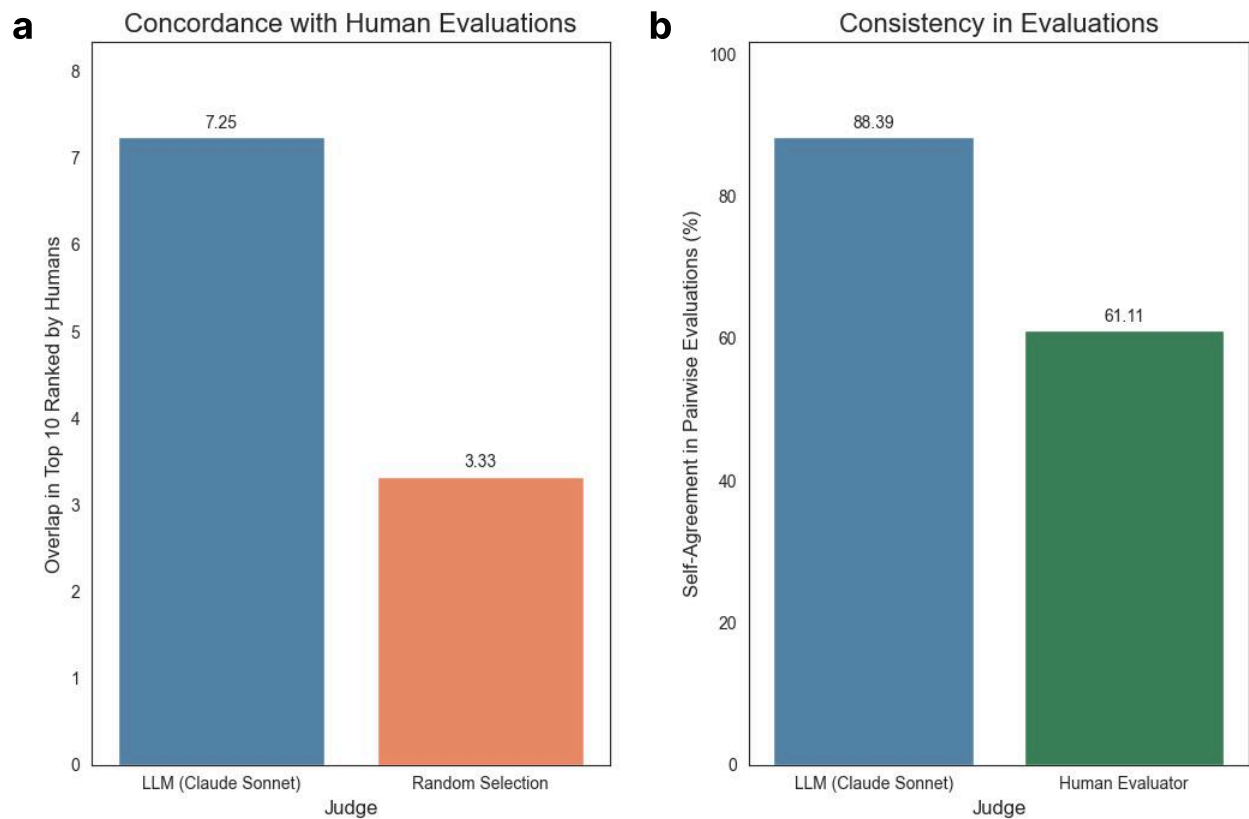
Supplementary Figure 18 Rubric for flow cytometry analysis

Model: Finch (RNA-seq analysis)

Rubric for RNA-seq (grading criteria):

1. Load necessary packages for performing differential expression analysis (DEA).
2. Import raw gene counts from file and perform appropriate data tidying and filtering. Correctly identify technical replicates.
3. Perform any necessary data transformation, e.g., create a DESeqDataSet object.
4. Use count-aware normalization (avoid RPKM/TPM for differential expression tests).
5. Evaluate normalization methods, for example using pre- and post-normalization MA-plots.
6. Filter low-count genes to improve power and false discovery rate (FDR) control.
7. Use a count-aware model (negative binomial frameworks, e.g., DESeq2, edgeR).
8. Generate adjusted p-values using appropriate FDR correction.
9. Report the test and test statistic, including effect size as log₂ fold change (Log2FC), adjusted p-value, unadjusted p-value, and full model specifications including covariates.

Supplementary Figure 19 Rubric for RNA-seq analysis



Supplementary Figure 20 Robin LLM judge preferences align with human expert evaluations. A) The Robin LLM judge preferences align well with human expert evaluations, selecting an average of 7.25 of their top 10 hypotheses. B) The Robin LLM judge is more consistent in its preference when selecting between two hypotheses than human experts.

Supplementary References

- [63] Sun, C.-C., Chiu, H.-T., Lin, Y.-F., Lee, K.-Y. & Pang, J.-H. S. Y-27632, a ROCK Inhibitor, Promoted Limbal Epithelial Cell Proliferation and Corneal Wound Healing. *PLOS ONE* 10, e0144571 (2015).
- [64] Chung, E. J. *et al.* AICAR suppresses TNF- α -induced complement factor B in RPE cells. *Scientific Reports* 7, 17651 (2017).
- [65] Li, P. *et al.* TUDCA attenuates intestinal injury and inhibits endoplasmic reticulum stress-mediated intestinal cell apoptosis in necrotizing enterocolitis. *International Immunopharmacology* 74, 105665 (2019).
- [66] Cui, R., Tian, L., Lu, D., Li, H. & Cui, J. Exendin-4 Protects Human Retinal Pigment Epithelial Cells from H₂O₂-Induced Oxidative Damage via Activation of NRF2 Signaling. *Ophthalmic Research* 63, 404–412 (2019).
- [67] Zhou, P. J. *et al.* Immunomodulatory drug FTY720 induces regulatory CD4+CD25+ T cells in vitro. *Clinical and Experimental Immunology* 157, 40–47 (2009).
- [68] Schreiter, S. *et al.* A Human Retinal Pigment Epithelium-Based Screening Platform Reveals Inducers of Photoreceptor Outer Segments Phagocytosis. *Stem Cell Reports* 15, 1347–1361 (2020).
- [69] Yang, M. *et al.* Ripasudil as a Potential Therapeutic Agent in Treating Secondary Glaucoma in HTLV-1-Uveitis: An In Vitro Analysis. *International Journal of Molecular Sciences* 25, 3229 (2024).
- [70] Maksimova, V. *et al.* Histone Methyltransferases as a New Target for Epigenetic Action of Vorinostat. *Biochemistry (Moscow)* 88, 968–978 (2023).
- [71] Wagner, A. E. *et al.* Ascorbic acid partly antagonizes resveratrol mediated heme oxygenase-1 but not paraoxonase1 induction in cultured hepatocytes - role of the redox-regulated transcription factor Nrf2. *BMC Complementary and Alternative Medicine* 11, 1 (2011).
- [72] Lei, G., Xia, Y. & Johnson, K. M. The Role of Akt-GSK-3 β Signaling and Synaptic Strength in Phencyclidine-Induced Neurodegeneration. *Neuropsychopharmacology* 33, 1343–1353 (2008).
- [73] Hirota, T. *et al.* Identification of Small Molecule Activators of Cryptochrome. *Science (New York, N.Y.)* 337, 1094–1097 (2012).
- [74] Semlali, A., Papadakos, S., Contant, C., Zouaoui, I. & Rouabhia, M. Rapamycin inhibits oral cancer cell growth by promoting oxidative stress and suppressing ERK1/2, NF- κ B and beta-catenin pathways. *Frontiers in Oncology* 12 (2022).
- [75] Torka, R. *et al.* Activation of HER3 Interferes with Antitumor Effects of Axl Receptor Tyrosine Kinase Inhibitors: Suggestion of Combination Therapy. *Neoplasia* 16, 301–318 (2014).
- [76] Ohta, A. *et al.* In vitro induction of T cells that are resistant to A2 adenosine receptor-mediated immunosuppression. *British Journal of Pharmacology* 156, 297–306 (2009).
- [77] Aulakh, G. K. *et al.* Inhibiting focal adhesion kinase (FAK) blocks IL-4 induced VCAM-1 expression and eosinophil recruitment in vitro and in vivo. *Journal of Leukocyte Biology* 104, 147–158 (2018).
- [78] Jiang, Q. *et al.* 4-Phenylbutyric acid accelerates rehabilitation of barrier function in IPEC-J2 cell monolayer model. *Animal Nutrition* 7, 1061–1069 (2021).
- [79] Kang, C.-H. *et al.* β -Ionone attenuates LPS-induced pro-inflammatory mediators such as NO, PGE2 and TNF- α in BV2 microglial cells via suppression of the NF- κ B and MAPK pathway. *Toxicology in Vitro* 27, 782–787 (2013).
- [80] Sartawi, Z., Ryan, K. B. & Waeber, C. Bone regenerative potential of the selective sphingosine 1-phosphate receptor modulator siponimod: *In vitro* characterisation using osteoblast and endothelial cells. *European Journal of Pharmacology* 882, 173262 (2020).
- [81] Albrecht, P. *et al.* Effects of dimethyl fumarate on neuroprotection and immunomodulation. *Journal of Neuroinflammation* 9, 163 (2012).
- [82] Jin, W. *et al.* Drug Repurposing Flubendazole to Suppress Tumorigenicity via PCSK9-dependent Inhibition and Potentiate Lenvatinib Therapy for Hepatocellular Carcinoma. *International Journal of Biological Sciences* 19, 2270–2288 (2023).
- [83] Liu, S. *et al.* LXR ligands induce apoptosis of EGFR-TKI-resistant human lung cancer cells in vitro by inhibiting Akt-NF- κ B activation. *Oncology Letters* 15, 7168–7174 (2018).

- [84] Ji, Y., Li, H., Wang, F. & Gu, L. PPAR β/δ Agonist GW501516 Inhibits Tumorigenicity of Undifferentiated Nasopharyngeal Carcinoma in C666-1 Cells by Promoting Apoptosis. *Frontiers in Pharmacology* 9, 648 (2018).
- [85] Metry, M. *et al.* Biowaiver Monographs for Immediate Release Solid Oral Dosage Forms: Metformin Hydrochloride. *Journal of Pharmaceutical Sciences* 110, 1513–1526 (2021).
- [86] Sariaydin, T., Çal, T., Aydın Dilsiz, S., Canpınar, H. & Ünderer Bucurgat, In vitro assessment of cytotoxic, apoptotic and genotoxic effects of metformin. *Istanbul Journal of Pharmacy* 51, 167–174 (2021).
- [87] Ren, Y., Cui, Y.-N. & Wang, H.-W. Effects of different concentrations of nicotinamide on hematopoietic stem cells cultured in vitro. *World Journal of Stem Cells* 16, 163–175 (2024).
- [88] Ciaramella, V. *et al.* Activity and molecular targets of pioglitazone via blockade of proliferation, invasiveness and bioenergetics in human NSCLC. *Journal of Experimental & Clinical Cancer Research : CR* 38, 178 (2019).
- [89] Suleiman, L., Muataz, Y., Négrier, C. & Boukerche, H. Protein S-mediated signal transduction pathway regulates lung cancer cell proliferation, migration and angiogenesis. *Hematology/Oncology and Stem Cell Therapy* (2021).
- [90] McCafferty, D. G. *et al.* Chemistry and biology of the ramoplanin family of peptide antibiotics. *Peptide Science* 66, 261–284 (2002).
- [91] Fang, Y. *et al.* Removal of RPE lipofuscin results in rescue from retinal degeneration in a mouse model of advanced Stargardt disease: Role of reactive oxygen species. *Free Radical Biology and Medicine* 182, 132–149 (2022).
- [92] Chhunchha, B. *et al.* Sulforaphane-Induced Klf9/Prdx6 Axis Acts as a Molecular Switch to Control Redox Signaling and Determines Fate of Cells. *Cells* 8 (2019).
- [93] Ray, B. *et al.* Rivastigmine modifies the α -secretase pathway and potentially early Alzheimer's disease. *Translational Psychiatry* 10, 1–17 (2020).
- [94] Manai, F. *et al.* Trehalose Modulates Autophagy Process to Counteract Gliadin Cytotoxicity in an In Vitro Celiac Disease Model. *Cells* 8, 348 (2019).