Supplementary Information for

Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases

Emily Alsentzer[1,2,*], Michelle M. Li[1,3,*], Shilpa N. Kobren[1], Ayush Noori[1], Undiagnosed Diseases Network[4], Isaac S. Kohane[1], and Marinka Zitnik[1,5,6,7,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, USA

[2]Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, USA

[3]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, USA

[4]Members of the Undiagnosed Diseases Network Consortium are listed at the end of this document

[5]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, USA

[6]Broad Institute of MIT and Harvard, Cambridge, USA

[7]Harvard Data Science Initiative, Cambridge, USA

‡Corresponding author. Email: marinka@hms.harvard.edu

*Equal contribution

Supplementary Figures and Tables

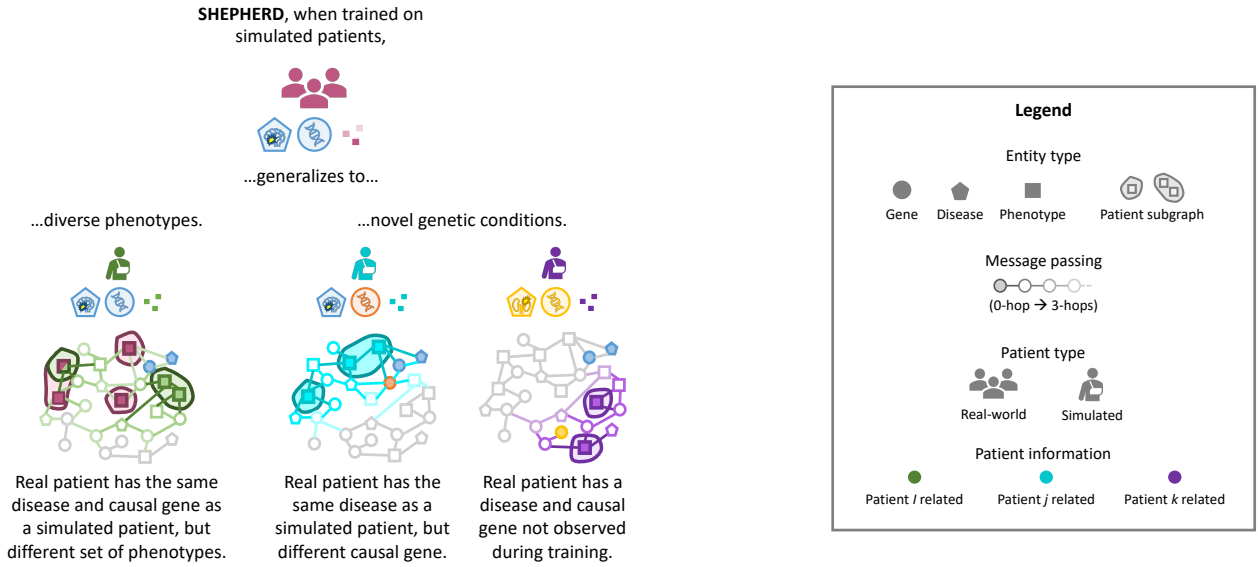

Figure S1: SHEPHERD can generalize to heterogeneous phenotypic presentations and novel genetic conditions. There are few patients with each rare disease, and patients with the same disease can have variable clinical presentations. SHEPHERD is trained on simulated rare disease patients and can generalize to real-world patients with unique, unseen phenotypes (left), with novel disease-causing genes (center), and with entirely novel diseases (right).

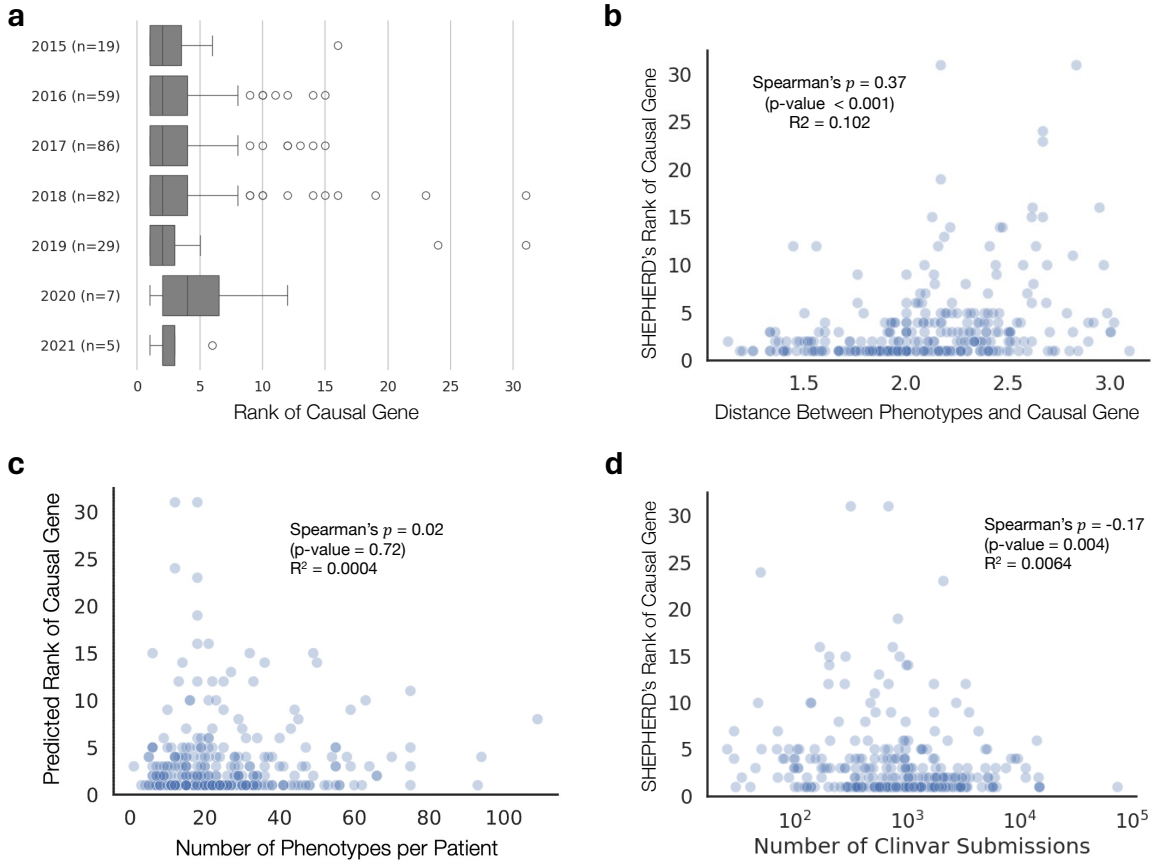

Figure S2: Generalizability of causal gene discovery performance on EXPERT-CURATED candidate lists. **(a)** Performance of SHEPHERD in ranking causal genes stratified by evaluation year on the EXPERT-CURATED gene list **(b)** Correlation between model performance (i.e., the rank of a disease-driving gene) and the average distance between a patient's phenotypes and causal genes in the knowledge graph. **(c)** Correlation between model performance and the number of phenotype terms describing each patient's clinical presentation. **(d)** Correlation between model performance and prevalence of the rare genetic disorders. The number of submissions to the database ClinVar for the causal gene is used as a surrogate for the prevalence of the rare disorders. The x-axis is number of submissions in log-scale.

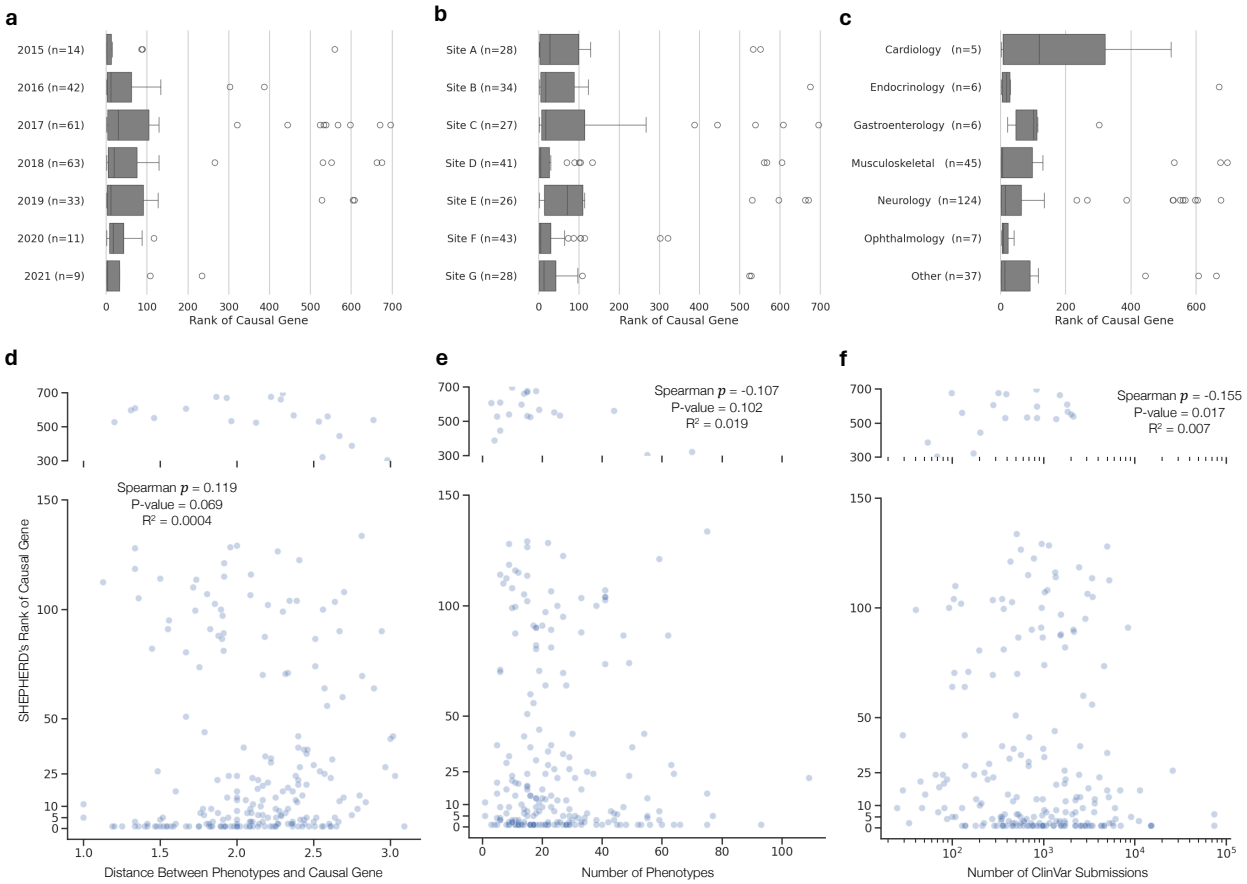

Figure S3: Generalizability of causal gene discovery performance on VARIANT-FILTERED candidate lists. (a-c) Performance of SHEPHERD in ranking causal genes stratified by **(a)** evaluation year, **(b)** clinical site, and **(c)** primary presenting symptom on the VARIANT-FILTERED gene list **(d)** Correlation between model performance (i.e., the rank of a disease-driving gene) and the average distance between a patient's phenotypes and causal genes in the knowledge graph. **(e)** Correlation between model performance and the number of phenotype terms describing each patient's clinical presentation. **(f)** Correlation between model performance and prevalence of the rare genetic disorders. The number of submissions to the database ClinVar for the causal gene is used as a surrogate for the prevalence of the rare disorders. The x-axis is number of submissions in log-scale.

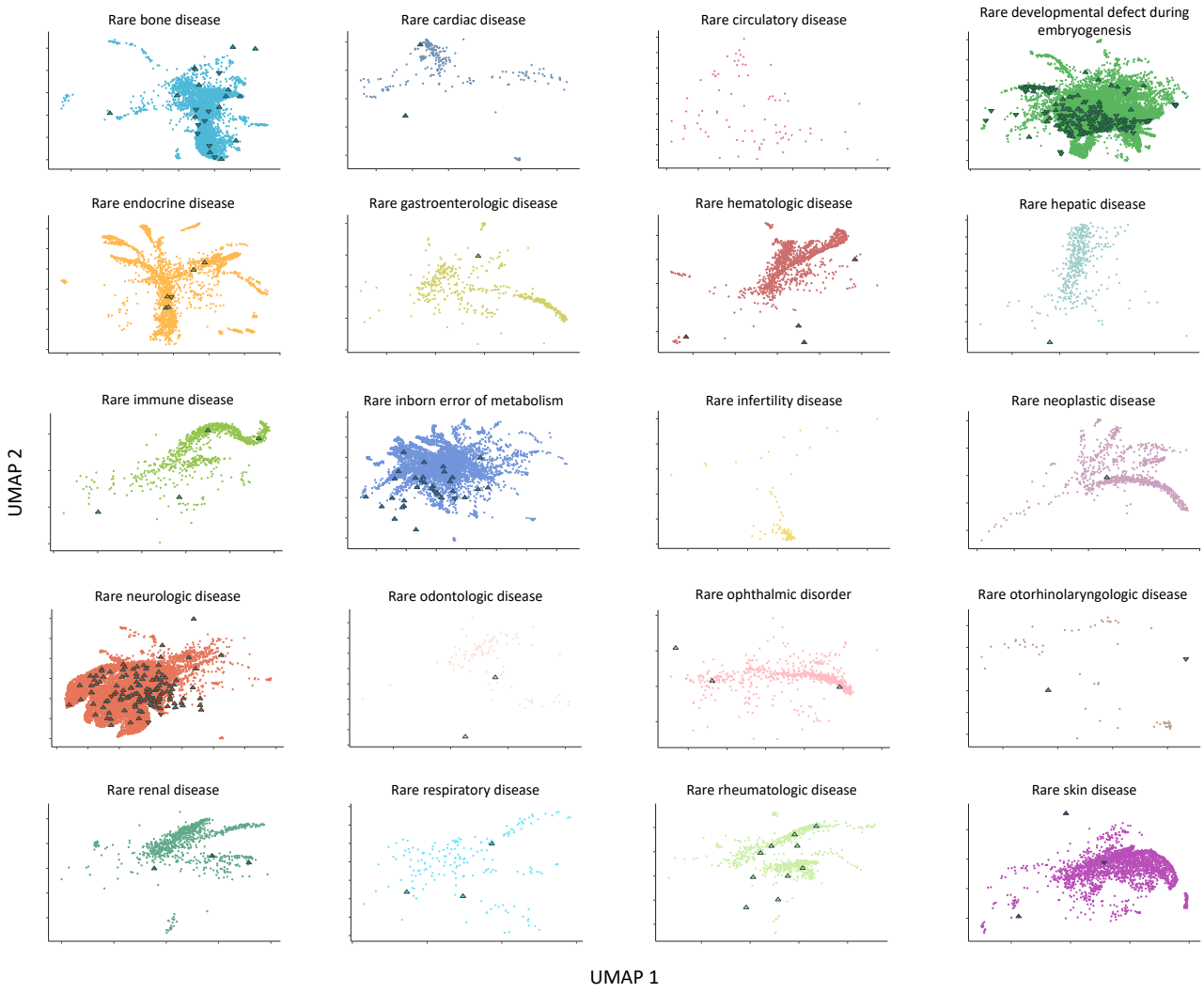

Figure S4: Visualization of rare disease patients by disease category. Two-dimensional UMAP plot of SHEPHERD's embedding space of all simulated patients (circles) and two real-world cohorts of UDN patients (up-facing triangles) and MyGene2 patients (down-facing triangles) grouped by the Orphanet disease category of medical diagnosis. Simulated, MyGene2 and UDN patients embed nearby other patients whose diagnoses belong to the same disease category.

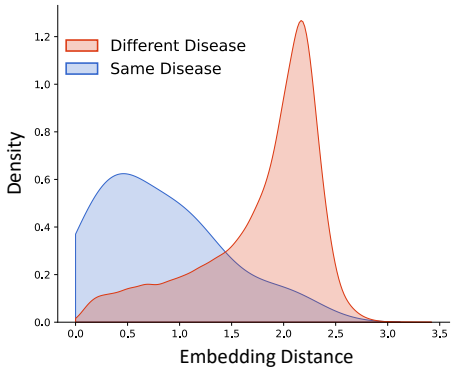

Figure S5: SHEPHERD performs patients-like-me identification. Distribution of SHEPHERD embedding distance between UDN and MyGene2 patients with the same vs. different diseases.

Table S1: Non-overlapping disorders between all phenotyped patients in the Undiagnosed Diseases Network (UDN) and simulated (SIM) patient cohorts. The names of the 5 most frequently observed diseases that are not in the other patient cohort are shown. The full list of syndromes found across all cohorts can be found in the Harvard Dataverse Repository at the following link: https://dataverse.harvard.edu/file.xhtml?fileId=10214709&version=3.0.

Rank	Diseases in UDN but not in SIM	Diseases in SIM but not in UDN
1	neurodevelopmental disorder with regression, abnormal movements, loss of speech, and seizures	multiple intestinal atresia
2	TBCK-related intellectual disability syndrome	progeroid syndrome, Petty type
3	dystonia 28, childhood-onset	myofibrillar myopathy 3
4	Rett syndrome, congenital variant	GM3 synthase deficiency
5	Bethlem myopathy 1	otospondylomegaepiphyseal dysplasia, autosomal dominant

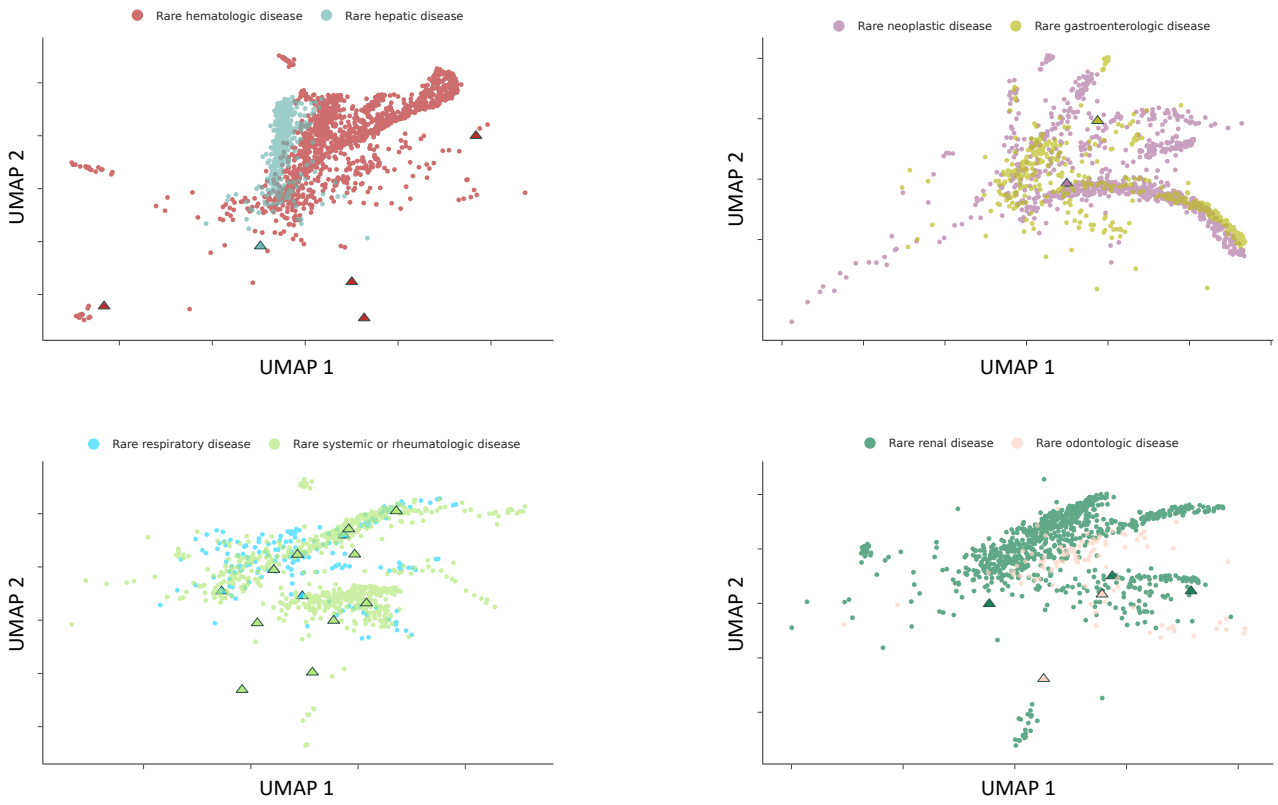

Figure S6: Visualization of the relationship between disease categories. Two-dimensional UMAP plot of SHEPHERD's embedding space for the most similar pairs of disease categories. Circles correspond to simulated patients, up-facing triangles to UDN patients, and down-facing triangles to MyGene2 patients.

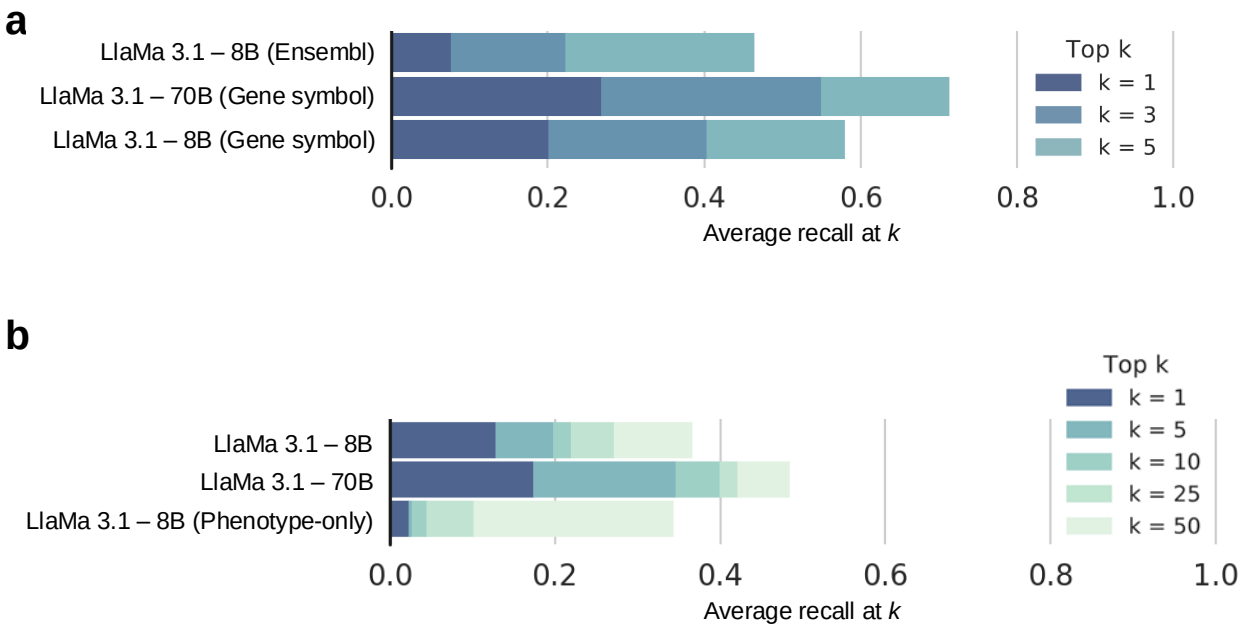

Figure S7: Evaluation of LlaMa 3.1 models. Performance of different LlaMa 3.1 8B and 70B models on **(a)** the EXPERT-CURATED gene lists, where genes are kept as Ensembl IDs or mapped to gene symbols, and **(b)** the VARIANT-FILTERED gene lists or all genes (i.e., phenotype-only).

Supplementary Notes

Supplementary Note S1: Causal gene discovery prompts for LlaMa 3.1 models.

We use three types of prompts in our experiments with the LlaMa 3.1 models.

(a) Prompt used when the patient's phenotypes and candidate genes are both provided as input.

You are an expert in rare disease diagnosis. I will provide a list of Human Phenotype Ontology (HPO) terms describing a patient's symptoms, along with a list of candidate genes. Using your knowledge of genetics, known disease-gene associations, and variant interpretation, generate a ranked list of all of the candidate genes based on their likelihood of causing the patient's symptoms. The output should be in JSON Lines (jsonl) format, with each line containing the gene name, rank, and a brief explanation of why the gene is relevant to the patient's HPO terms. Only output a valid jsonl file with no spaces between each json. Rank every candidate gene according to its association with the HPO terms, known gene-disease relationships, and functional impact. Make sure to rank all candidates.

Output (JSON Lines format):
{"gene_name": "Gene1", "rank": 1, "explanation": "Explanation for why Gene1 is relevant" }
{"gene_name": "Gene2", "rank": 2, "explanation": "Explanation for why Gene2 is relevant" }

(b) Prompt used when only the patient's phenotypes are provided as input.

You are an expert in rare disease diagnosis. I will provide a list of Human Phenotype Ontology (HPO) terms describing a patient's symptoms. Using your knowledge of genetics, known disease-gene associations, and variant interpretation, generate a ranked list of all of the candidate genes as Gene Symbols based on their likelihood of causing the patient's symptoms. The output should be in JSON Lines (jsonl) format, with each line containing the gene symbol, rank, and a brief explanation of why the gene is relevant to the patient's HPO terms. Only output a valid jsonl file with no spaces between each json. Rank every candidate gene according to its association with the HPO terms, known gene-disease relationships, and functional impact. Make sure to rank all candidates.

Output (JSON Lines format):
{ "gene_name": "Gene1", "rank": 1, "explanation": "Explanation for why Gene1 is relevant" }
{ "gene_name": "Gene2", "rank": 2, "explanation": "Explanation for why Gene2 is relevant" }

(c) Prompt used to merge two previously ranked candidate gene lists. *When patients have very long candidate gene lists, we split the lists in two, rank each smaller list, and use a language model*

to merge the two ranked lists into a final ranked list.

You are an expert in rare disease diagnosis. I will provide Human Phenotype Ontology (HPO) terms describing a patient's symptoms, along with two lists of previously ranked candidate genes. Using your knowledge of genetics, known disease-gene associations, and variant interpretation, combine the two ranked lists of all of the candidate genes based on their likelihood of causing the patient's symptoms. The output should be in JSON Lines (jsonl) format, with each line containing the gene name, rank, and a brief explanation of why the gene is relevant to the patient's HPO terms. Only output a valid jsonl file with no spaces between each json. Rank every candidate gene according to its association with the HPO terms, known gene-disease relationships, and functional impact. Make sure to rank all candidates.

Output (JSON Lines format):
{"gene_name": "Gene1", "rank": 1, "explanation": "Explanation for why Gene1 is relevant" }
{"gene_name": "Gene2", "rank": 2, "explanation": "Explanation for why Gene2 is relevant" }