

Simulation of undiagnosed patients with novel genetic conditions

Received: 2 September 2022

Accepted: 26 September 2023

Published online: 12 October 2023

 Check for updates

Emily Alsentzer^{1,2,6}, Samuel G. Finlayson^{1,2,3,4,6}, Michelle M. Li^{1,5}, Undiagnosed Diseases Network*, Shilpa N. Kobren¹   & Isaac S. Kohane¹  

Rare Mendelian disorders pose a major diagnostic challenge and collectively affect 300–400 million patients worldwide. Many automated tools aim to uncover causal genes in patients with suspected genetic disorders, but evaluation of these tools is limited due to the lack of comprehensive benchmark datasets that include previously unpublished conditions. Here, we present a computational pipeline that simulates realistic clinical datasets to address this deficit. Our framework jointly simulates complex phenotypes and challenging candidate genes and produces patients with novel genetic conditions. We demonstrate the similarity of our simulated patients to real patients from the Undiagnosed Diseases Network and evaluate common gene prioritization methods on the simulated cohort. These prioritization methods recover known gene-disease associations but perform poorly on diagnosing patients with novel genetic disorders. Our publicly-available dataset and codebase can be utilized by medical genetics researchers to evaluate, compare, and improve tools that aid in the diagnostic process.


Rare congenital disorders are estimated to affect nearly 1 in 17 people worldwide¹, yet the genetic underpinnings of these conditions—knowledge of which could improve support and treatment for scores of patients—remain elusive for 70% of individuals seeking a diagnosis and for half of suspected Mendelian disorders in general^{2,3}. This diagnostic deficit results in a substantial cumulative loss of quality-adjusted life years and a disproportionate burden on healthcare systems overall^{4–6}. Organizations such as the Undiagnosed Disease Network (UDN) in the United States have been established to facilitate the diagnosis of such patients, which has resulted in both successful diagnoses for many patients as well as the discovery of new diseases^{7,8}.

The diagnostic workup of patients with suspected Mendelian disorders increasingly includes genomic sequencing. Whole genome or exome sequencing typically identifies thousands of genetic variants, which must be analyzed to identify the subset of causal variant(s) yielding the patient's syndrome (Fig. 1a). This process is challenging

and error prone; for example, patients may have variants that do not ultimately cause their presenting syndrome, yet fall into genes that are plausibly associated with one or more of their phenotypes. Further challenges arise in situations where patients present with a novel set of symptoms that do not match any known disorder, or when their disease-causing variants occur in genes not previously associated with any disease (Fig. 1b). In the first phase of the UDN, for instance, 23% of eventual patient diagnoses were due to novel syndromes⁴.

To accelerate the diagnostic process, a plethora of computational tools are used by clinical teams to automatically analyze patients' genetic and phenotypic data to prioritize causal variants^{9–12}. Unfortunately, a comprehensive evaluation of these tools' performance is hindered by the lack of a public benchmark database of difficult-to-diagnose patients of sufficient size to cover the full breadth of genomic diseases. While efforts such as the Deciphering Developmental Disorders project provide useful benchmarks for specific populations of

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. ²Program in Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. ³Department of Pediatrics, Division of Genetic Medicine, Seattle Children's Hospital, Seattle, WA 98105, USA. ⁴Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98105, USA. ⁵Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA. ⁶These authors contributed equally: Emily Alsentzer, Samuel G. Finlayson. *A list of authors and their affiliations appears at the end of the paper.

 e-mail: Shilpa_Kobren@hms.harvard.edu; Isaac_Kohane@hms.harvard.edu

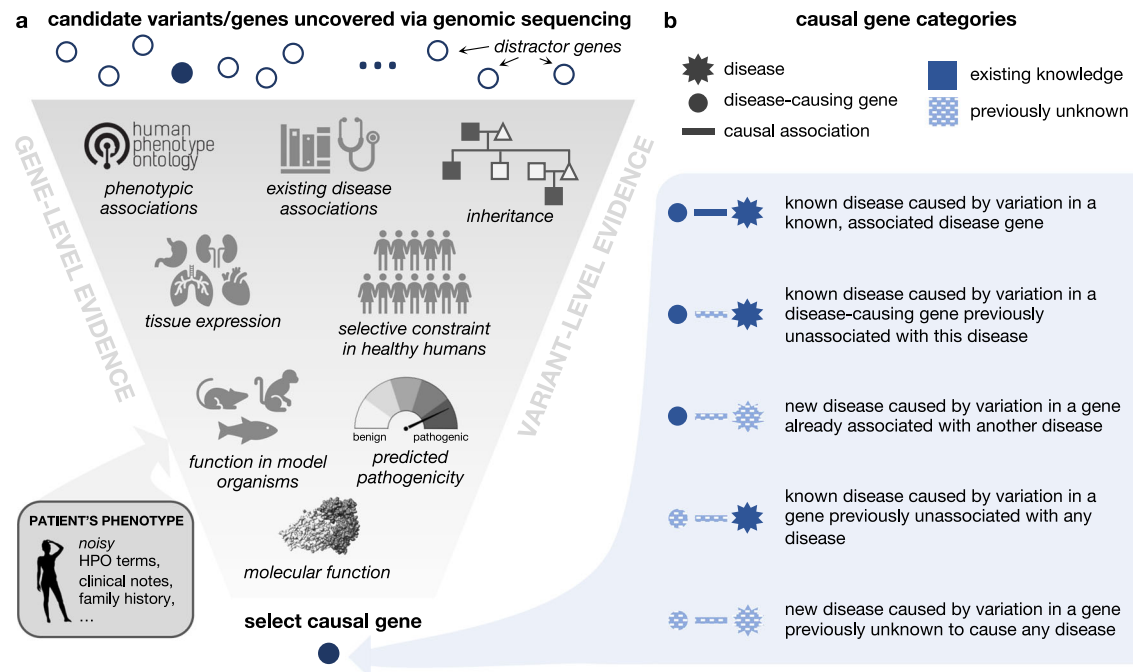


Fig. 1 | Identification and categorization of causal disease genes. **a** Genomic variation uncovered in an affected patient through DNA sequencing is investigated using variant-level and gene-level evidence in order to identify the gene variant that is most likely responsible for causing the patient's symptoms. Here, we depict a subset of relevant information that a care team may use to make this assessment. **b** The causal gene responsible for a patient's disorder can be categorized based on

the extent of medical knowledge that exists about the gene and its associated disorder. Intuitively, diagnosing patients where less is known about their causal gene and disease (bottom category) is a more challenging task than diagnosing patients where more is known about causal gene and disease (top category). The protein structure pictured is PDB: ID3B. Icons are from Microsoft PowerPoint.

rare disease patients, they are limited in diagnostic scope and require an extensive DUA for full access^{3,14}. In lieu of real patient data, simulated patient data offers several clear advantages: the simulation approach can be scaled to an arbitrary number of patients and disorders, data are inherently publicly shareable, and the transparency of the simulation process can be leveraged to expose specific failure modes of different methods. However, simulated patient data is only useful insofar as it reflects the ongoing challenges of real-world diagnosis. This requires a faithful simulation of the complex relationship between candidate genes harboring compelling genetic variants and the patient's phenotypes as well as the notion of disease novelty, as described above. Existing approaches for simulating Mendelian disease patients unrealistically model the patient's genotype and phenotype disjointly by inserting disease-causing alleles into otherwise healthy exomes and separately simulating patient phenotypes as a set of precise, imprecise, and noisy phenotypes^{15–17}. Indeed, patients representing true diagnostic dilemmas may randomly harbor irrelevant pathogenic variants with low penetrance or may have atypical disease presentations or symptoms stemming from variants in genes previously unassociated with disease, respectively confusing variant- and phenotype-based prioritization methods. As such, jointly modeling phenotypes alongside genotypes is essential for generating simulated patients that standard clinical workups would struggle to diagnose. In addition, with few exceptions, most studies analyzing tools for Mendelian disease diagnosis do not assess the ability of tools to identify novel syndromes or variants, or do so by masking specific disease–gene associations but not the gene–phenotype and phenotype–disease interactions that may have been annotated as a result of the initial disease–gene discovery^{16,18,19}. Given the importance and prevalence of automated prioritization tools in the diagnostic process, enabling meaningful comparisons and improvement of these tools via benchmarks that capture the notion of novelty and realistic phenotypes will be essential.

Here, we present a computational pipeline to simulate difficult-to-diagnose patients that can be used to evaluate gene prioritization tools. Each simulated patient is represented by standardized phenotype terms and sets of candidate genes that are presumed to be impacted by one or more compelling variants. To model novel genetic conditions in our simulated patients, we first curate a knowledge graph (KG) of known gene–disease and gene–phenotype annotations that is time-stamped to 2015. This enables us to define post-2015 medical genetics discoveries as novel with respect to our KG. We additionally provide a taxonomy of categories of “distractor” genes that do not cause the patient's presenting syndrome yet would be considered plausible candidates during the clinical process. We then introduce a simulation framework that jointly samples genes and phenotypes according to these categories to simulate nontrivial and realistic patients and show that our simulated patients closely resemble real-world patients profiled in the UDN. Finally, we reimplement existing gene prioritization algorithms and assess their performance in identifying etiological genes—referred to as “causal genes” henceforth—in our simulated patient set, revealing specific settings in which established tools excel or fall short. Overall, the approach to patient simulation we present here is, to the best of our knowledge, the first to incorporate nontrivial candidate distractor genes and phenotype annotations to reflect real-world diagnostic challenges as well as the notion of novel genetic disorders. We provide our framework and simulated patients as a public resource to advance the development of new and improved tools for medical genetics.

Results

We design and implement a pipeline for simulating patients with difficult-to-diagnose Mendelian disorders (Fig. 2a). Each simulated patient is represented by an age range, a set of positive symptoms (phenotypes) that they exhibit, a set of negative phenotypes that they do not exhibit, and a set of candidate genes impacted by variants that

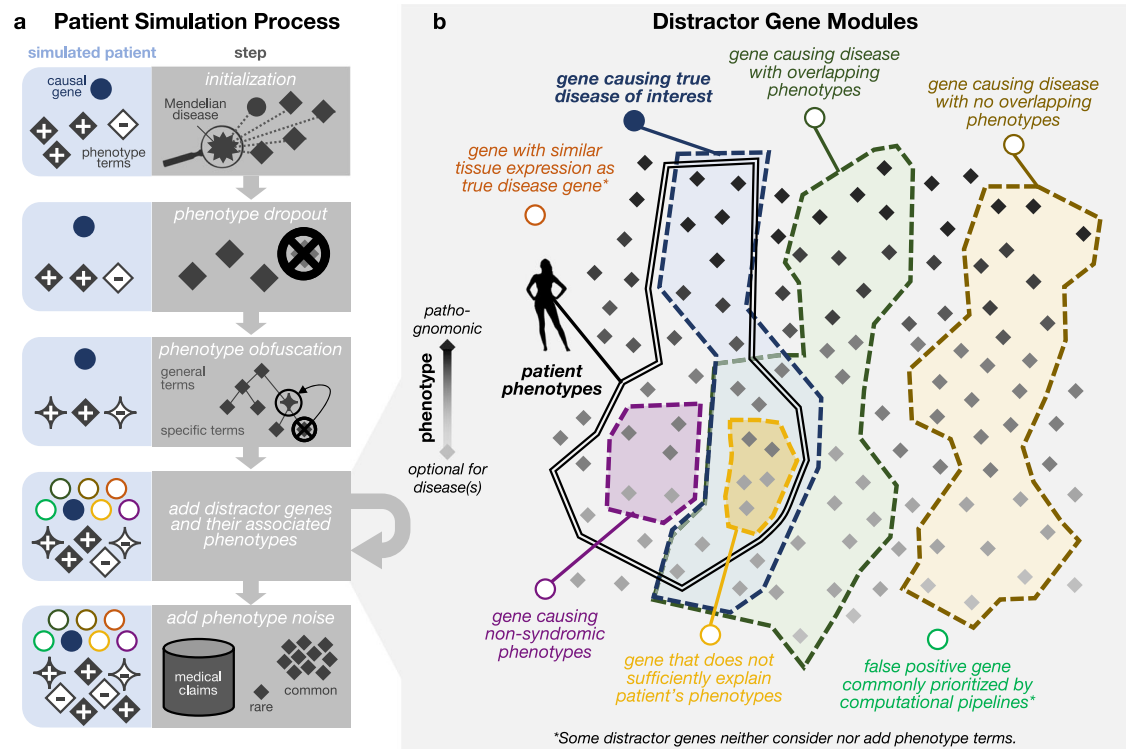


Fig. 2 | Simulation process generates patients with multiple phenotype terms and candidate genes. a Patients are first assigned a true disease and initialized with a gene known to cause that disease (blue circle) as well as with positive and negative phenotypes associated with that disease (gray diamonds). Phenotype terms are then randomly removed through phenotype dropout, randomly altered to be less specific according to their position in an ontology relating phenotype terms, and augmented with terms randomly selected by prevalence in a medical claims database. Finally, strong distractor candidate genes and relevant additional phenotypes are generated based on six distractor gene modules. **b** The six distractor gene

modules are inspired by genes that are frequently considered in current clinical genomic workflows and are designed to generate highly plausible, yet ultimately non-causal, genes for each patient. Four of the distractor gene modules are defined by the overlap—or lack thereof—between the phenotypes associated with the distractor gene and the phenotypes associated with the patient's causal gene. The remaining two modules are defined by their similar tissue expression as the true disease gene or solely by their frequent erroneous prioritization in computational pipelines.

may be causing their disease. There are three components of our simulation framework. First, each patient is initialized with a genetic disorder profiled in the comprehensive and well-maintained rare genetic disease database Orphanet²⁰. Second, the imprecision in real-world diagnostic evaluations is modeled via *phenotype dropout* to mimic patients' partially observed symptoms, *phenotype obfuscation*, which replaces specific symptoms with more general phenotype terms, and *phenotype noise*, which adds unrelated symptoms and comorbidities proportionally to their prevalence in age-matched patients from a medical insurance claims database. Finally, we develop a framework to generate strong, yet ultimately noncausal, candidate genes inspired by the typical rare disease diagnostic process (Fig. 2b). These challenging distractor genes and some associated phenotype terms are added according to each of six distractor gene modules (see Methods for further details). This entire process takes 0.08 s on average to simulate a single patient.

Simulated patients mimic real-world patients

We leverage our computational pipeline to simulate 20 realistic patients for each of 2134 unique Mendelian disorders, representing a total of 42,680 patients and 2401 unique causal genes. Each simulated patient is characterized by 18.39 positive phenotypes ($s = 7.7$), 13.5 negative phenotypes ($s = 8.5$), and 14 candidate genes ($s = 3.5$) on average. To assess whether our simulated patients are systematically distinguishable from real-world patients with nontrivial diagnoses, we assemble a cohort of 121 real-world patients from the Undiagnosed Diseases Network (UDN) who were diagnosed with a disease in Orphanet annotated with genes and phenotypes and then select 2420 simulated patients with matching

diseases. UDN patients are similar to other rare disease patients with respect to the severity of their diseases, but their ultimate diagnoses tend to be more elusive due to the involvement of novel disease genes and atypical disease presentations. The UDN cohort also differs from other rare disease cohorts in that patients have relatively thorough standardized phenotyping and exhibit symptoms with varying ages of onset across broad disease categories⁴. There are 92 unique diseases represented in the real and disease-matched simulated patient cohorts. Real and simulated patients have similar numbers of candidate genes (13.13 vs 13.94 on average; Fig. 3a) and positive phenotype terms (24.08 vs 21.57 on average; Fig. 3b). Real-world patients are also more similar to their simulated counterparts than to other real-world patients. When we apply dimensionality reduction on the positive phenotype terms of patients, real-world patients cluster with and are visually indistinguishable from simulated patients within each disease category (Fig. 3c), suggesting that there are no large-effect, consistent differences in phenotype term usage between real and simulated patients. To more precisely measure this, we performed feature analysis on a random forest classifier implemented to distinguish real from simulated patients. Indeed, we found that the classifier's accuracy of 94.9% was dominated by specific phenotype terms that appear with prevalence of close to 0% in one of the groups and low-prevalence (typically <5%) in the other group to collectively provide great discriminatory power (Supplementary Fig. 1). These subtle patterns that are unsurprisingly detected and exploited by an ML algorithm suggest the existence of idiosyncracies within UDN annotations rather than clinically meaningful differences. Moreover, for each real-world patient, the ten phenotypically-closest simulated patients with the same disease are

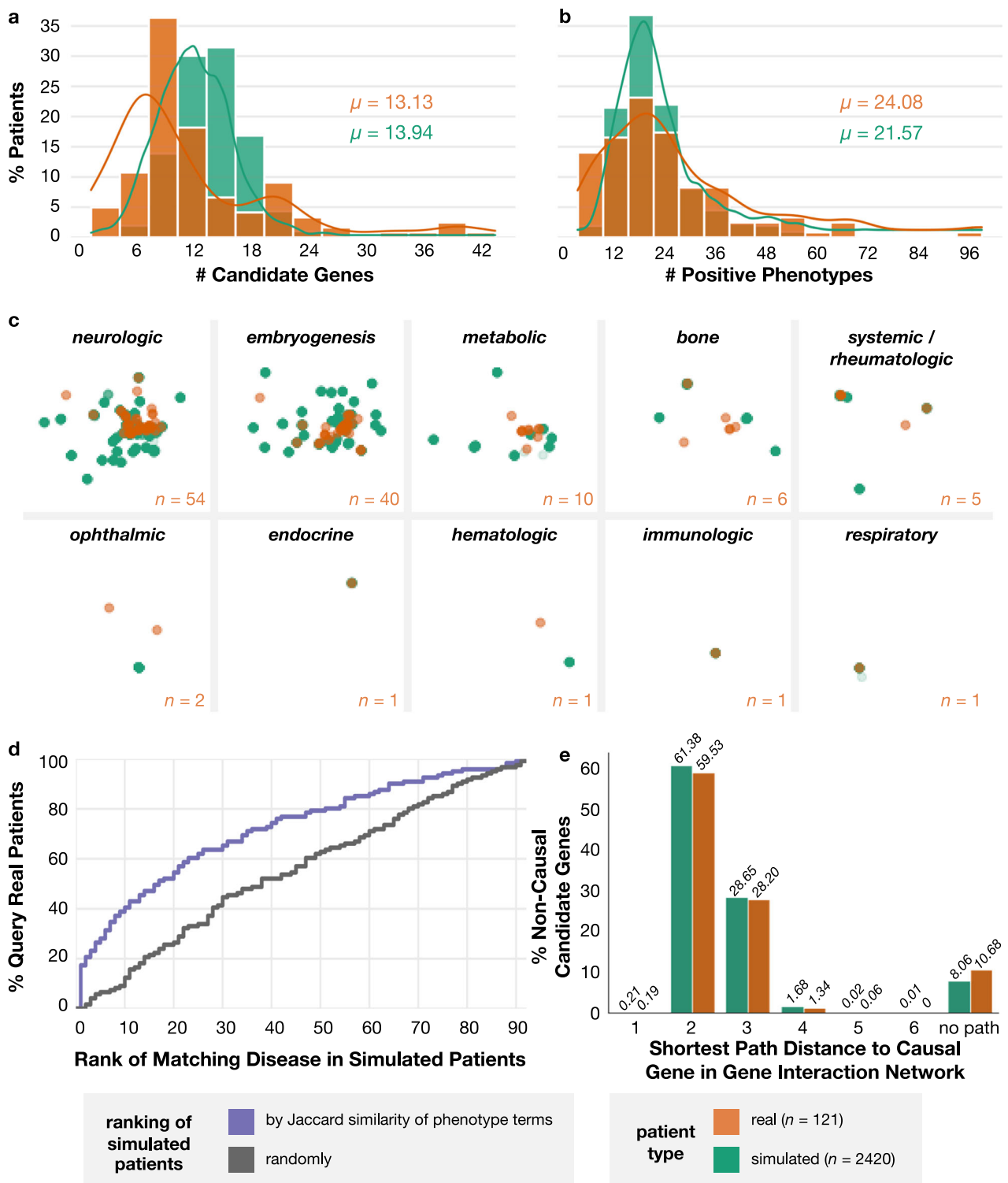


Fig. 3 | Simulated patients mimic real-world patients. Diagnosed, real-world patients from the Undiagnosed Diseases Network (orange) and a disease-matched cohort of simulated patients (teal) have similar numbers of **a** candidate genes per patient (average μ of 13.13 vs. 13.94) and **b** positive phenotype terms per patient (average of 24.08 vs. 21.57). **c** Real patients (orange) and simulated patients (teal) are indistinguishable based on their annotated positive phenotype terms within each Orphanet disease category, as visualized using non-linear dimension reduction via a Uniform Manifold Approximation and Projection (UMAP) plot. The horizontal and vertical axes are uniform across all plots. The number of real patients within each disease category, n , is listed in the corner of each plot; there are 20 simulated patients for each real patient. **d** For each real-world patient, all

simulated patients in the disease-matched cohort are ranked randomly (black) and by the Jaccard similarity of their phenotype terms to the query real-world patient (purple). The Empirical Cumulative Distribution Function (ECDF) plot shows that the basic Jaccard similarity metric is able to retrieve simulated patients with the same disease as the query real patient more accurately than if the simulated patients were retrieved randomly. **e** The distributions of shortest path distances between all non-causal candidate and true causal genes in a gene-gene interaction network are indistinguishable between real-world and simulated patients. n is the number of patients in each patient category. Source data are provided as a Source Data file.

Table 1 | Counts of simulated and real-world Undiagnosed Diseases Network (UDN) patients in each causal gene-disease category

| Causal Gene Category | # Simulated Patients | # Real-world UDN Patients |
|---|----------------------|---------------------------|
| Known disease caused by an associated causal gene | 36,226 (85%) | 149 (58%) |
| Known disease caused by a disease-causing gene previously unassociated with their disease | 4339 (10%) | 15 (6%) |
| Known disease caused by a gene never before associated with any disease | 835 (2%) | 5 (2%) |
| Novel disease caused by a disease-causing gene | 947 (2%) | 60 (23%) |
| Novel disease caused by a gene never before associated with any disease | 333 (1%) | 29 (11%) |

UDN patients with multiple causal genes may appear in several categories. The values in parentheses refer to the percentage of total simulated or UDN patients.

closer than the ten phenotypically-closest real-world patients with different diseases (average Jaccard similarity of 0.952 vs 0.930; $P = 7.4 \times 10^{-81}$, Wilcoxon one-sided test). We also employ a nearest neighbor analysis using Jaccard as our similarity metric to evaluate whether the simulated patients' phenotype terms are sufficiently reflective of and specific to their assigned diseases. We find that simulated patients with similar sets of phenotype terms to real patients are more likely to have the same disease as those real patients than randomly selected simulated patients (Fig. 3d). Finally, genes that are proximal in an interaction network may be especially difficult to disambiguate due to their related biological functions and similar phenotypes when perturbed. We found that the shortest path distances between distractor candidate and true causal genes per patient in a protein-protein and transcription factor interaction network were similarly distributed between real and simulated patients (Fig. 3e, KS-Test $P = 0.239$).

Pipeline simulates patients with novel and diverse genetic conditions

A primary challenge in diagnosing real-world patients, and one that should be reflected in relevant simulated patients, is when their causal gene-disease relationships have never previously been documented (Fig. 1b). However, simulating patients with “novel” genetic conditions is conceptually nontrivial, as simulated disease associations must be drawn from some existing knowledge graph. We learn more about atypical disease presentations over time, and the discovery of new gene-disease associations may accompany new gene-phenotype and disease-phenotype associations as well. To overcome this issue, we consider the gene-phenotype-disease associations annotated in a knowledge graph timestamped to 2015 to be “existing knowledge” and any discoveries made post-2015 to be “novel” (see Methods for details). This enables us to categorize simulated patients according to the novelty of their gene-disease relationships with respect to this timestamped knowledge graph (Table 1). Although only 2% and 1% of the total simulated patients respectively correspond to previously known and previously unknown diseases caused by genes never before associated with any disease, the total number of simulated patients in these two categories are 14x and 190x higher, respectively, than in our real-world dataset. Moreover, whereas only 231 unique disease genes have been identified as causal in our phenotypically-diverse, real-world UDN dataset, our simulated patients' 2100+ unique diseases are caused by 2401 unique disease genes. Overall, these results demonstrate that our pipeline can simulate substantially higher numbers of patients—that are diverse with respect to disease and the degree of novelty of their causal gene-disease relationships—than compared to a national dataset of real-world patient data.

Performance of gene prioritization algorithms on real and simulated patients

The size and diversity of our real and simulated patient datasets enable us to evaluate how well existing algorithms are able to prioritize causal genes in patients with different degrees of preexisting knowledge of their gene-disease relationships. We run the gene prioritization algorithms from six commonly-used programs on patients in each of the

causal gene-disease association categories outlined in Fig. 1b, ensuring that each algorithm only had access to knowledge timestamped to 2015 or earlier. Each algorithm inputs the patient's phenotype terms and the patient's individualized set of candidate genes or variants and produces a ranking of the candidate genes according to how likely they are to cause the patient's phenotypes. While some of the algorithms can be used for both variant and gene prioritization, here we strictly evaluate the phenotype-based gene prioritization capabilities of each algorithm. The first class of algorithms computes the semantic similarity of a patient's phenotype terms from the Human Phenotype Ontology (HPO) directly to phenotype terms associated with each candidate gene. Phrank-Gene considers the prevalence of gene associations across these phenotypes and all phenotype ancestors in the ontology, and ERIC-Gene (introduced and implemented in Xrare) additionally considers the most informative common ancestor of two phenotype terms when computing similarity^{10,16}. The second class of algorithms considers disease-phenotype associations. Phrank-Disease and ERIC-Disease each evaluate all diseases associated with a candidate gene and compare the patients' phenotypes and those diseases' phenotypes, assigning the candidate gene the highest similarity score across all of its associated diseases. Phenomizer uses semantic similarity of phenotype terms and prevalence of phenotype associations across all known diseases to match patient symptoms directly to diseases, rather than to genes¹². LIRICAL estimates the extent to which patients' phenotypes are consistent across all known diseases using a likelihood ratio framework²¹. For both Phenomizer and LIRICAL, we assign candidate genes the highest score across their associated diseases. The final class of algorithms utilizes additional types of interactions beyond human-derived gene-phenotype, gene-disease, and disease-phenotype edges. Phenolyzer uses semantic similarity to match patient symptoms to diseases and scores genes directly associated with the diseases as well as additional genes connected via a gene-gene network¹¹. HiPhive (implemented in Exomiser) leverages ontologies from humans, zebrafish, and mice to determine phenotype similarities. In the absence of phenotypic data for a candidate, HiPhive employs a random walk approach on a protein-protein interaction network to establish connections between the candidate and other genes exhibiting similar phenotypes²². We do not include the two other phenotype similarity algorithms included in Exomiser because they leverage the same phenotype-gene semantic similarity algorithm found in Phenomizer (PhenIX) or use a subset of model organism interactomes already included in HiPhive (Phive). Finally, ERIC-Predicted from Xrare computes “predicted” phenotype similarity scores between each patient and each candidate gene using XGBoost by considering the phenotypic profiles of related genes with similar sequences, protein domains, pathway membership, or tissue expression¹⁶.

Real-world and simulated patients are equally difficult to diagnose

We first assess whether gene prioritization performance was similar between the simulated patient cohort and the real-world patient cohort. Across both patient sets, correctly ranking the causal gene becomes more difficult as the amount of information about the causal

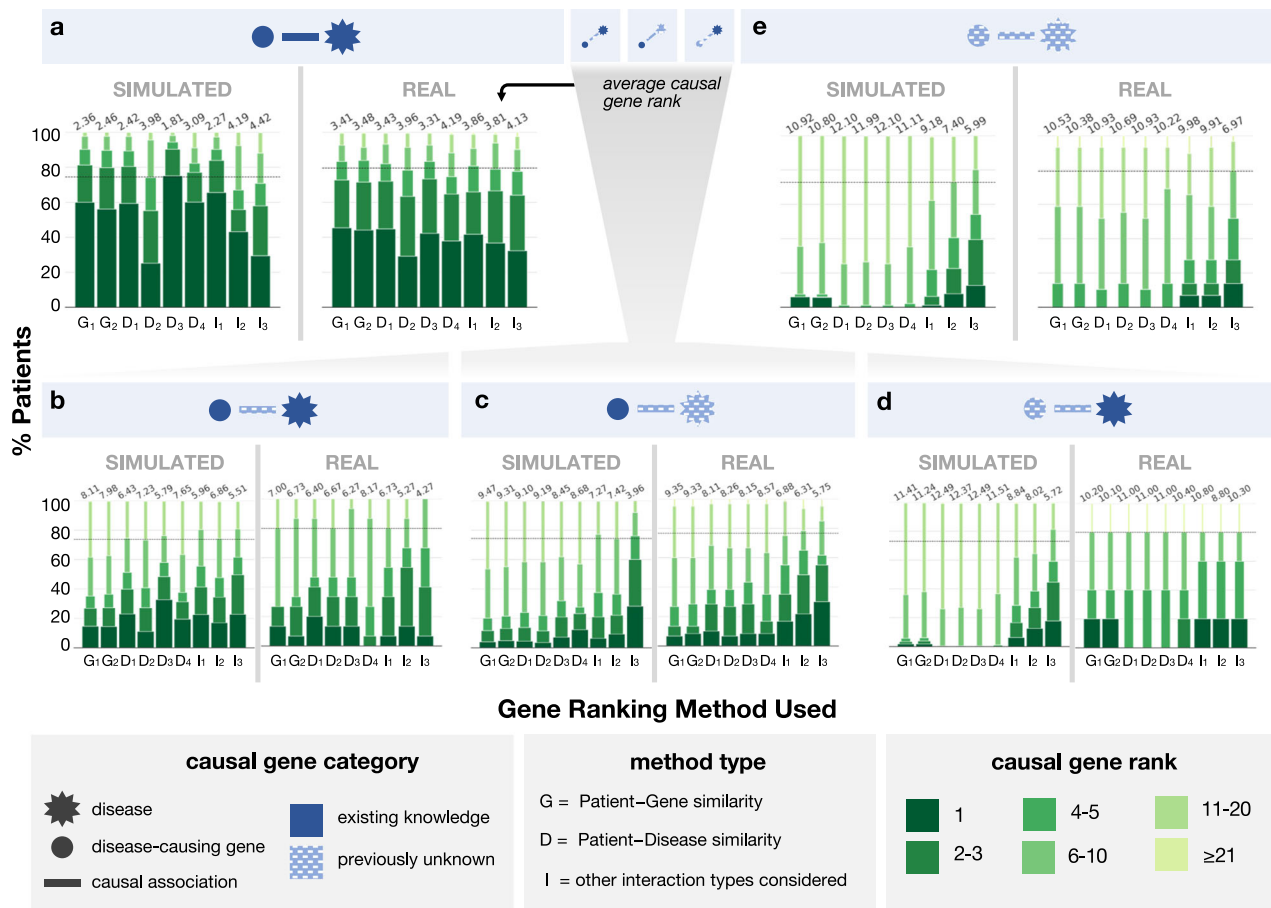


Fig. 4 | Ability of computational approaches to rank causal genes differs across disease-gene categories. We group simulated patients and real-world UDN patients into five categories based on their type of causal gene-disease association (patient counts in Table 1). These categories, described in detail in Fig. 1b, are illustrated in the blue header bars above each plot and ordered decreasingly from left to right by the amount of existing knowledge of the association in the underlying knowledge graph. Each panel a–e shows performance on patients in a single category. We run nine gene ranking algorithms implemented in six prioritization tools on the phenotype terms and candidate gene list for each simulated and real-world patient within each causal gene-disease category. These algorithms are separated into those that directly consider patient–gene phenotypic similarity (G_1 : Phrank–Gene, G_2 : ERIC–Gene), those that compute patient–disease phenotypic

similarity (D_1 : Phrank–Disease, D_2 : ERIC–Disease, D_3 : Phenomizer, D_4 : LIRICAL), and those that consider additional interaction edges, such as gene–gene edges, interactions in other species, or predicted edges (I_1 : Phenolyzer, I_2 : HiPhive, I_3 : ERIC–Predicted). We show here the ability of these methods to correctly rank each patient’s causal gene within the top k ranked genes for varying values of k . For visual clarity, the color and width of each stacked bar section corresponds to causal gene rank grouping. The average rank of the causal gene is italicized above each bar. Dashed lines denote the average percent of patients where the causal gene appeared in the top 10 across ten random rankings of the candidate genes. Boxen plots displaying the distributions of causal gene ranks can be found in Supplementary Fig. 2. Source data are provided as a Source Data file.

gene-disease relationship in the knowledge graph decreases (Fig. 4; boxen plots are shown in Supplementary Fig. 2). We find that the performance between simulated and real-world patients was similar for all algorithms across all but the easiest gene-disease association categories. Overall, these results indicate that the simulated patient cohort can serve as a reasonable proxy for real-world patients, particularly when evaluating a method’s ability to perform well despite reduced existing knowledge about the causal gene and disease.

Novel syndromes and disease genes represent greatest challenge

All gene prioritization tools perform well when the patient’s causal gene-disease relationship is in the knowledge graph (Fig. 4a). Specifically, all methods rank the causal gene first in at least 25% of simulated and real-world patients, representing a 4-fold increase in the proportion of patients with the causal gene ranked first relative to the next novelty category. However, nearly all methods perform incrementally worse as less information about the simulated patient’s causal gene-disease relationship is present in the knowledge graph. In patients with

a known disease caused by a gene previously unassociated with that disease (Fig. 4b), the average rank of the causal gene is 4.27 at best. When the patient has a novel disease, even when the causal gene has some other existing disease associations (Fig. 4c), performance further declines; the average rank of the causal gene drops by 1–2 positions for all methods on simulated and real data. An exception is that ERIC–Predicted (I_3) improves slightly on simulated patients when the disease is unknown, which is unsurprising given that this method does not use known gene–disease annotations. The most difficult scenarios occur when the patient’s causal gene has never been associated with any disease. When the patient’s disease is known (Fig. 4d), the average rank of the causal gene is 5.72 for the highest performing method, and when both the disease and causal gene are unknown (Fig. 4e), the average rank is 5.99 at best and the proportion of patients with their causal gene ranked in the top 3 drops for all methods.

The algorithms that exclusively use phenotype–phenotype, disease–gene and phenotype–disease annotations from humans (i.e., method type D in Fig. 4) excel in settings where the patient’s causal gene and disease are in the KG (Fig. 4a, b). When the causal gene is not

associated with any disease in the KG (Fig. 4d, e), methods that compare patients' phenotypes directly to genes' phenotypes (i.e., method type G) outperform those approaches that compare patients' phenotypes to known diseases (i.e., method type D). Finally, leveraging additional interaction edge types provides a consistent performance boost only in cases where the causal gene and/or disease is novel (i.e., method type I, Fig. 4c–e). Indeed, when both the causal gene and disease are unknown (Fig. 4e), Phenolyzer (I_1) for instance, which leverages gene–gene associations, ranks the causal gene 9.08 for simulated patients on average whereas all G and D method types rank the causal gene significantly lower, at 10.92, 10.80, 12.10, 11.99, 12.10 and 11.11 respectively on average (all $P < 2.89 \times 10^{-17}$). Nevertheless, the shortest path distances between incorrectly highly-ranked and true causal genes in a gene–gene network are similarly distributed even across D- and G-type methods, as biologically correlated genes often lead to similar phenotypes when perturbed (Supplementary Fig. 3). HiPhive (I_2), which additionally considers non-human interaction edges, and ERIC–Predicted (I_3), which predicts interaction edges using other omics data, outperform all approaches in identifying completely novel gene–disease associations (Fig. 4e). These results are in line with previous suggestions to search for known disease-causing variants and phenotypically-concordant disease genes first before utilizing the additional interaction types that help gene prioritization algorithms generalize to settings where the causal gene has not been previously associated with a disease²². Notably, when algorithms are evaluated on all patients together rather than separately by novelty category, the general performance decline for all methods across categories as well as relative differences in performance within each novelty category, are obfuscated (Supplementary Fig. 4).

Simulation pipeline components are key to simulating realistic, challenging patients

To determine the importance of each component in our simulation pipeline (Fig. 2), we run our pipeline using only subsets of these components and then evaluate how well Phrank–Disease, the fastest of the gene prioritization algorithms we evaluated, is able to rank causal genes in the resultant simulated patients (Fig. 5; boxen plots in Supplementary Fig. 5). In all ablations, we set the probability of sampling each candidate gene module to be uniform. As expected, we find that the gene prioritization task is easiest when all of the phenotype-altering components and distractor gene modules (as illustrated in Fig. 2) are excluded from our simulation pipeline, that is, when candidate genes are selected randomly and phenotype terms do not undergo obfuscation, dropout, or augmentation with phenotypic noise (Fig. 5a). In this setting, the diagnostic gene appears at rank 1.855 on average. The task becomes significantly more difficult when only phenotype-altering components are added (average causal gene rank drops to 1.955; $P < 0.001$) and even more difficult when only distractor gene modules are added (average causal gene rank of 2.570; $P < 0.001$). The task is most difficult when the complete pipeline is used (average causal gene rank of 2.936; $P < 0.001$), suggesting that both the phenotype- and gene-based components of our simulation pipeline contribute to the generation of realistic patients representing challenging diagnostic dilemmas.

We next measure how each phenotype-altering component—phenotype obfuscation, dropout, noise, and phenotypes added by the distractor gene modules—impacts the difficulty of the gene prioritization task (Fig. 5b). To this end, we include a “gene-only” version of all distractor gene modules in the simulation pipeline, and we vary whether the distractor gene modules add associated phenotypes and whether phenotype terms undergo obfuscation, dropout, or noise augmentation. When we restrict to using only one phenotype-altering component at a time, we find that the component that increases the difficulty of the gene prioritization task the most is phenotype noise (average causal gene rank 2.583), followed by

phenotypes added by distractor gene modules, phenotype dropout, and phenotype obfuscation (average causal gene ranks 2.570, 2.435, and 2.316, respectively). Including either the distractor gene-associated phenotypes or phenotype noise alone increases the difficulty of the gene prioritization task more than including both phenotype dropout and obfuscation together (average causal gene rank 2.448). However, we confirm that including these latter two phenotype-altering components in addition to either one or both of the distractor gene phenotypes and phenotype noise does, in fact, increase the difficulty of the task more than if they were excluded. In general, adding additional phenotype-altering components makes the gene prioritization task progressively more difficult, and as expected, the most difficult combination is when all phenotype components are included. As before, we find that when the two strongest phenotype-altering components (noisy and distractor gene-associated phenotypes) are applied together, the gene prioritization task is more difficult than when certain sets of three phenotype-altering components are used (average causal gene rank of 2.773 versus average causal gene ranks of 2.767 and 2.702).

Finally, we perform an ablation of each of the six distractor gene modules by removing a single module at a time (Fig. 5c). When each distractor gene module is removed, the number of genes that would have been sampled from that module are instead sampled randomly to ensure that the total number of candidate genes for each patient is constant. Removing the module that generates phenotypically-similar disease genes or the module that generates phenotypically-distinct disease genes make the gene prioritization task substantially easier for Phrank–Disease relative to the individual exclusion of other distractor gene modules (average causal gene rank of 2.564 and 2.666, respectively). This is intuitive because Phrank's gene prioritization approach is based on disease-phenotype and disease-gene associations. We suspect that the other modules in our pipeline may generate distractor genes that are more challenging for gene prioritization algorithms that explicitly leverage cohort-based mutational recurrence, gene expression, and other additional data. Nevertheless, we confirm that removal of every gene module except for the insufficiently explanatory gene module makes the gene prioritization task easier. There are fewer insufficiently explanatory candidate genes in the ablation patient cohort (Supplementary Fig. 6), which may explain why their removal does not change gene prioritization performance. Furthermore, removal of all distractor gene modules together makes the task much easier compared to the removal of any individual module, demonstrating that no one module is solely responsible for generating the distractor genes that increase the difficulty of causal gene prioritization in simulated rare disease patients.

Discussion

In this work, we developed a flexible framework for simulating difficult-to-diagnose patients with elusive or atypical genetic disorders like those profiled in the Undiagnosed Diseases Network (UDN). Key features of our framework include: (i) jointly modeling patients' genotype and phenotype, (ii) capturing imprecision in real-world clinical workups by obfuscating, excluding, and adding prevalence-based noise from age-matched population cohorts to patients' recorded symptoms, and (iii) simulating patients with novel causal gene-disease associations relative to an established knowledge graph, to emulate the challenging task of diagnosing a previously unpublished disorder. Our framework can generate a phenotypically diverse cohort that is representative of all rare diseases characterized in Orphanet, and these simulated patients can be freely shared without privacy concerns. Our simulation framework is generalizable and customizable, and can be tuned to mimic other rare disease patient cohorts as desired.

Our simulated patients are represented as sets of phenotypes and candidate genes, rather than candidate variants. Variant-level

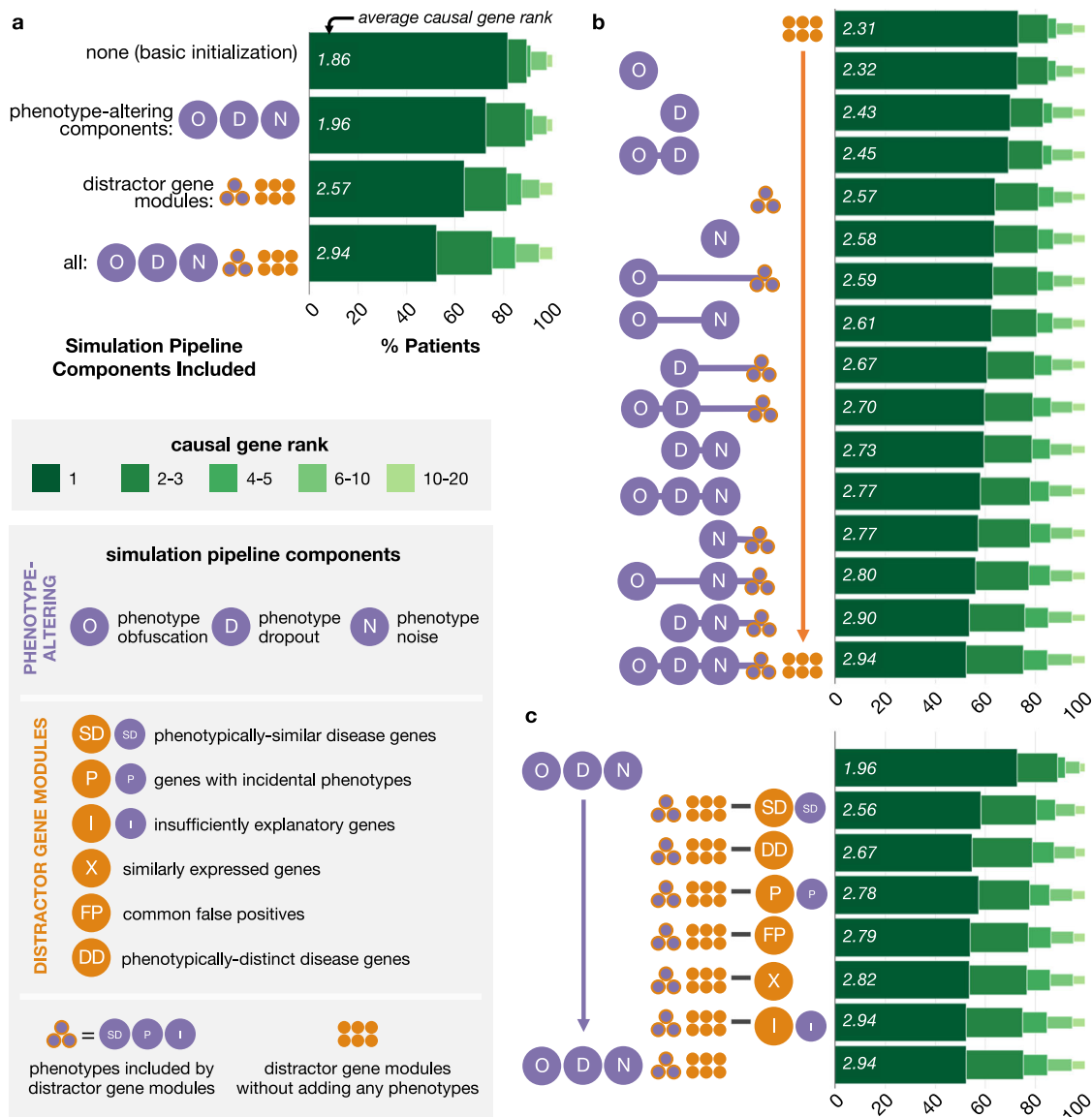


Fig. 5 | Pipeline components increase the difficulty of causal gene identification in simulated patients. We run a gene prioritization algorithm on patients simulated by our pipeline when varying subsets of pipeline components are included. We report the fraction of simulated patients where the causal gene was prioritized within the top k ranked genes for varying k (horizontal axis for all plots) when different components of the simulation pipeline are included (vertical axis for all plots). The average rank of the causal gene is listed in italics at the base of each bar. The color and width of each stacked bar section corresponds to causal gene rank grouping. We show gene prioritization performance on simulated patients produced when the following components are included in the simulation pipeline: **a** no

phenotype- nor gene-based components (i.e., candidate genes sampled randomly and phenotype terms unaltered from initialization), all standalone phenotype-altering components alone, all distractor gene modules alone, or all pipeline components together; **b** a “gene-only” version of distractor gene modules and each possible combination of subsets of phenotype-altering components; **c** all three standalone phenotype-altering components and all but one distractor gene module at a time. Note that in **b**, horizontal purple lines in the vertical axis labels are for visual clarity, whereas in **c**, horizontal black lines in the vertical axis signify set difference. Source data are provided as a Source Data file.

properties, such as variant inheritance patterns, functional impacts, and cohort-based frequencies, are considered only indirectly, as we assume that a patient’s candidate genes have been clinically shortlisted because compelling variants have implicated or lie within them^{9,23}. Some variants, such as regulatory variants or larger indels, may impact multiple genes simultaneously. In addition, real-world patients may have two or more genes contributing to their presenting disorder(s). Although our current framework generates patients with monogenic conditions, it could be extended to simulate patients with more than one causal gene; this will become more feasible at scale as the number of rare, multigenic diseases curated in Orphanet increases. The real-world UDN patients modeled by our simulation framework tend to have more phenotype terms relative to other rare disease patients. As

large language models advance, automated extraction of accurate phenotype terms from clinical records should enable many more patients to have similarly comprehensive annotations for use in automated prioritization tools^{24,25}. However, for the majority of rare disease patients worldwide who lack informative clinical records with relevant phenotypic details, the scarcity of even their extracted HPO terms relative to our simulated patients will remain²⁶. As the biomedical data leveraged by prioritization methods diversifies, our framework could be extended to reflect these additions, for instance, by incorporating distractor genes that are highly constrained across human populations and/or phenotypic noise inspired by errors in machine learning-based parsing of clinical notes. As available data from rare disease patients continues to expand, future iterations of this work may also benefit

from generative machine learning methods such as diffusion methods or generative adversarial networks. By building upon the published literature, our work is also subject to biases inherent in the field of clinical genetics research writ large, including the historical overrepresentation of individuals of European descent and the diseases that affect them. Indeed, we expect that the dataset and methods we present in this paper could be used to further interrogate these biases, for example, by examining the differential performance of gene-prioritization tools on diseases that more often affect under-represented populations.

Finding and annotating “novel” gene-disease associations—defined in our framework as those published post-2015—required significant manual review of public databases and literature. Databases that curate these associations (e.g., Orphanet, HPO) are often missing publication or discovery dates and are not updated in real time, and so an indeterminate lag exists between causal gene-disease associations being present in the literature and being reflected in these knowledge bases. To fairly evaluate gene prioritization methods across the disease novelty categories we describe in Fig. 1b, we ensured that all tested methods solely leveraged data from 2015; only methods that used or could be reimplemented to exclusively use this form of input data were included. When possible, we leveraged identical data from our time-stamped knowledge graph in order to assess the impact of the algorithm or included data types independently from any specific data source. Due to their reliance on statically curated data, these methods may have misprioritized real-world patients’ gene-disease associations that were “known” by 2015 but had yet to be incorporated into the knowledge graph, reflecting an expected and ongoing hindrance for diagnosing present-day patients (Fig. 4a). We suspect that diagnostic tools that frequently mine the literature for new gene-phenotype-disease associations would excel at diagnosing patients with known causal genes and known diseases relative to tools that rely on structured rare disease databases^{27,28}.

We found that the algorithms that were most effective at finding novel disease-causing genes performed relatively poorly at diagnosing patients with known causal genes and diseases, underscoring the importance of evaluating performance separately across distinct novelty categories. Given these findings, clinicians may opt to use certain computational tools earlier in the diagnostic process and move to research-oriented tools only in cases where a novel disease-causing gene is suspected²².

We expect that the simulated patients produced via our framework can be leveraged for a wide range of applications²⁹. As we demonstrate here, the simulated patients can enable a uniform evaluation of existing gene prioritization tools on a representative patient cohort. Developers can also internally validate and improve their tools by separately evaluating them on simulated patients across novelty categories and distractor gene categories. Another application area for our pipeline will be the generation of training data for machine learning algorithms. As the promise of machine learning solutions in the clinic grows, access to large-scale datasets of relevant clinical data will be essential³⁰. We suspect that simulated patients such as those yielded by our method may provide invaluable training data for machine learning models for rare disease diagnosis, which would expose algorithms to data from diverse genetic disorders while reflecting realistic clinical processes.

Methods

Simulated patient initialization

We simulate patients for each of the 2134 diseases in Orphanet²⁰ (orphanet.org, accessed October 29, 2019) that do not correspond to a group of clinically heterogeneous disorders (i.e., Orphanet’s “Category” classification³¹), have at least one associated phenotype, and have at least one causal gene. For Orphanet diseases that were missing either a causal gene or phenotypes (but not both) and were listed as

being a “clinical subtype”, “etiological subtype”, or “histopathological subtype” of another Orphanet disease that did have a causal gene and/or phenotypes (e.g., “Cystinuria type A”), we imported the causal gene and/or phenotypes from the parent disease (e.g., ‘Cystinuria’) as appropriate. For each patient, the gene set is initialized with the known causal disease gene (mapped to its Ensembl identifier); the age is randomly sampled from the age ranges associated with the disease (e.g., “infant”); and positive and negative phenotype terms from the Human Phenotype Ontology³² (HPO, version 2019) are added with probabilities $P(\text{term}|\text{disease})$ and $1-P(\text{term}|\text{disease})$ respectively, where $P(\text{term}|\text{disease})$ is provided in Orphanet and corresponds to the observed prevalence of a specific phenotype term presenting in patients with the disease. Sex and gender were not modeled in the simulation process.

Modeling diagnostic process imprecision

To mimic real-world patients’ partially observed phenotypes, we perform phenotype dropout where each positive and negative phenotype term is removed from the simulated patient with probabilities $P(\text{positive dropout})$ and $P(\text{negative dropout})$, respectively, set to 0.7 and 0.2 in our implementation. We also perform phenotype obfuscation to replace specific phenotype terms (e.g., “arachnodactyly”) with their less precise parent terms (e.g., “long fingers”) in the HPO ontology. Positive and negative phenotypes annotated to each patient are obfuscated with probabilities $P(\text{positive obfuscation})$ and $P(\text{negative obfuscation})$, each set to 0.15 in our implementation. Finally, to model unrelated symptoms and comorbidities that would be present in real-world patients, we introduce phenotype noise by sampling new HPO phenotype terms that were mapped from ICD-10 billing codes from a large medical insurance claims database using the Unified Medical Language System (UMLS) crosswalk³³. Positive phenotype terms are sampled with probabilities proportional to their prevalence in corresponding age-stratified populations (i.e., infants are defined as 0–1 years, children are 2–11 years, adolescents are 12–18 years, adults are 19–64 years, and seniors are 65+ years), and negative phenotype terms are sampled from the same corresponding age-stratified populations at random.

Distractor gene modules

In order to mimic the typical diagnostic process where numerous potentially disease-causal genes must be manually reviewed by a clinical team, we generate a set of $1 + N_G$ highly plausible, yet ultimately non-causal, genes for each patient, where N_G is drawn from a Poisson distribution parameterized by λ . In our implementation, the tunable parameter λ is set to the mean number of candidate genes considered in real-world patients with undiagnosed genetic conditions (see Methods “Preprocessing Real Patient Data” below). These N_G genes are generated from the following six distractor gene modules with probabilities 0.33, 0.42, 0.05, 0.09, 0.08, and 0.03, respectively, which we set in our implementation based on the approximate frequency of each distractor gene type in real-world patients with undiagnosed diseases; these parameters can be customized by the user. Each distractor gene module contributes one gene to the simulated patient’s candidate gene list, and three distractor gene modules simultaneously add phenotype terms related to the added gene.

Phenotypically-similar disease genes. First, we identify genes causing distractor Mendelian diseases in Orphanet that have overlapping phenotype terms with the patient’s true disease (Fig. 2b, dark green). We categorize the phenotype terms associated with the distractor disease as “obligate”, “strong”, “weak”, or “excluded” if their prevalence in patients with that disease is 100%, 80–99%, 1–29% or 0%, respectively. For instance, Alstrom syndrome (ORPHA:64) can serve as a distractor disease for Wolfram syndrome (ORPHA:3463) because five of their phenotype terms overlap. Alstrom syndrome has five “strong”

phenotypes (e.g., cone/cone-rod dystrophy, progressive sensorineural hearing impairment) and 33 “weak” phenotypes (e.g., polycystic ovaries, hepatomegaly). We require that all distractor diseases have at least one obligate phenotype term and/or at least one excluded phenotype term, or that all phenotype terms that overlap with the true disease of interest and are added to the patient are “weak” (e.g., three of the five overlapping phenotypes between Alstrom and Wolfram syndromes are “weak”). We add the causal gene for the distractor disease to the simulated patient’s set of candidate genes and add phenotype terms that overlap between the distractor and true disease to the simulated patient’s positive phenotype set. To ensure that these genes are challenging distractors but definitively non-causal, we add some excluded phenotypes to the simulated patient’s set of positive phenotypes and some obligate phenotypes to the simulated patient’s set of negative phenotypes. For distractor diseases with no associated obligate or excluded phenotypes and only weak overlapping phenotypes with the true disease, we instead add some strong, non-overlapping phenotypes to the simulated patient’s negative phenotypes. At each of these steps, $1 + N_p$ phenotype terms are added to the simulated patient’s positive or negative phenotype sets, where N_p is drawn from a Poisson distribution parameterized by λ . In our implementation, we set λ such that simulated patients and real-world patients with undiagnosed diseases have approximately the same number of annotated phenotype terms on average (see Fig. 3 and Methods “Preprocessing Real Patient Data” below).

Phenotypically-distinct disease genes. Since any variants in known disease genes tend to be investigated during the diagnostic process, we also add genes causing Mendelian diseases that do not have any phenotypic overlap with the patient’s true disease (Fig. 2b, yellow)³⁴.

Insufficiently explanatory genes. Genes that are not yet known to be disease-causing but are associated with a subset of the patient’s disease-relevant phenotypes are also strong candidates for further diagnostic investigation. To generate such insufficiently explanatory genes, we first curate a set of non-disease genes as the set of all genes from DisGeNET³⁵ (accessed April 16, 2019, https://www.disgenet.org/downloads/all_gene_disease_associations.tsv) and excluding any genes causally associated with a disease in Orphanet or in HPO Annotation^{36,37} (accessed February 12, 2019, http://compbio.charite.de/jenkins/job/hpo.annotations.monthly/ALL_SOURCES_ALL_FREQUENCIES_diseases_to_genes_to_phenotypes.txt). We add non-disease genes that are associated with a strict subset of low prevalence phenotypes from the simulated patient’s true disease (Fig. 2b, light orange). We then add $1 + N_p$ of the gene’s phenotype terms to the simulated patient’s positive phenotype set if none are already present, with N_p defined as above.

Genes associated with incidental phenotypes. Naturally occurring phenotypic variance present across healthy individuals can be incorrectly considered to be relevant to a patient’s disease during diagnostic evaluations. To include genes causing these nonsyndromic phenotypes, we add non-disease genes associated only with phenotypes that do not overlap with the simulated patient’s true disease (Fig. 2b, purple). We add some of the gene’s phenotypes to the simulated patient’s positive phenotype set as before.

Similarly expressed genes. We also add genes with similar tissue expression as the patient’s causal disease gene (Fig. 2b, dark orange), as a candidate gene’s expression in relevant tissues is considered as supporting experimental evidence for a gene-disease association in clinical evaluations³⁸. For each gene, we compute its average tissue expression in transcripts per million in each of 54 tissue types profiled in GTEx³⁹ (accessed October 29, 2019, https://gtexportal.org/home/datasets/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.tsv). For each tissue type, we linearly 0,1-normalize the per-gene

expression values such that the gene with the lowest expression in that tissue type is assigned a value of 0 and the gene with the highest expression in that tissue type is assigned a value of 1. We compare each gene’s normalized tissue expression vector to the simulated patient’s causal gene’s tissue expression vector using cosine similarity. We select one of the top 100 most similar genes with probability proportional to its tissue expression similarity, excluding known disease genes with phenotypic overlap with the simulated patient’s true disease.

Common false positive genes. Finally, we add genes from the Frequently mutated GeneS (FLAGS) database⁴⁰ with probabilities proportional to the number of rare functional variants affecting these genes in general populations, as computational pipelines tend to frequently prioritize these genes due to their length and variational excess.

Preprocessing real patient data

We selected all patients from the Undiagnosed Diseases Network (UDN) with a molecular diagnosis as of March 19, 2020. Each patient is annotated with a set of positive and negative HPO phenotype terms and a set of strong candidate genes that were considered by clinical teams who handled each case. For each of the 362 diagnosed patients who received genomic sequencing through the UDN, we augment their candidate gene lists with disease-associated and other clinically-relevant genes listed on their clinical sequencing reports³⁴. Where possible, patients’ gene lists were further augmented with genes prioritized by the Brigham Genomic Medicine pipeline⁴¹. We map all genes to Ensembl identifiers, discard prenatal phenotype terms related to the mother’s pregnancy, and exclude patients with fewer than five candidate genes. The final cohort includes 248 patients.

The Undiagnosed Diseases Network study is approved by the National Institutes of Health institutional review board (IRB), which serves as the central IRB for the study (Protocol 15HG0130). All patients accepted to the UDN provide written informed consent to participate in the study and to share their data across the UDN as part of a network-wide informed consent process.

Comparing simulated patients to real patients

Of the 248 diagnosed UDN patients that we consider, only 121 patients were diagnosed with a disease in Orphanet that we were able to model (see “Simulated Patient Initialization” above). We construct a disease-matched cohort of 2420 simulated patients by selecting, for each of these 121 UDN patients, 20 simulated patients with the same disease. We first visualize positive phenotype term similarities between real and simulated patients using non-linear dimension reduction via a Uniform Manifold Approximation and Projection (UMAP) plot using Python 3.6.7⁴². We also compare each real patient to all simulated patients in the disease-matched cohort by computing pairwise Jaccard similarities ranging from 0 to 1 inclusively between the positive phenotype terms annotated to each real patient and the positive phenotype terms annotated to each simulated patient. For each real patient, we then rank all simulated patients from highest to lowest Jaccard similarity and analyze those simulated patients’ corresponding diseases to assess whether simulated patients are as phenotypically specific to each disease as real patients are. We then evaluate the average Jaccard similarity between the query real-world patient and the top 10 retrieved simulated patients with the same disease compared to the average Jaccard similarity between the query and the top 10 real-world patients with a different disease using a one-sided Wilcoxon signed-rank test implemented in the SciPy Stats library (version 1.5.4). We perform the Shapiro-Wilk test from the SciPy Stats library to test for normality. Finally, we computed the shortest path distance between each causal and all non-causal genes per patient in a protein-protein and transcription factor interaction network derived respectively from the Human Protein Reference Database and the Human

Transcriptional Regulation Interactions Database¹¹ to show that the distributions of path lengths are not statistically different between real and simulated patients.

Evaluating gene prioritization tools on novel diseases

To model novel genetic conditions in our simulated patients, we leverage a knowledge-graph (KG) of gene-gene, gene-disease, gene-phenotype, and phenotype-disease annotations from Phenolyzer¹¹ that is time-stamped to February 2015 (obtained from github.com/WGLab/phenolyzer/tree/eccec7410729276859b9023a00f20e75c2ce58862) and the HPO-A ontology³⁷ time stamped to January 2015 (obtained from github.com/drseb/HPO-archive/tree/master/hpo.annotations.monthly/2015-01-28_14-15-03/archive/annotation and github.com/drseb/HPO-archive/tree/master/2014-2015/2015_week_4/annotations/artefacts). Phenotype-phenotype annotations are from the 2019 HPO ontology (accessed via Obonet 0.3.0). All gene names are mapped to Ensembl IDs, and older phenotype terms are updated to the 2019 HPO ontology. We also manually time-stamp each disease and disease-gene association in Orphanet according to the date of the Pubmed article that reported the discovery; discoveries after February 2015 are considered novel with respect to our KG. Note that the KG time-stamped to February 2015 may not reflect all new information contained in the most recent publications from PubMed, as would be expected for any curated database. We categorize each novel discovery as in Fig. 1b. We apply the same process to manually annotate the novelty of disease-gene associations in the real-world UDN cohort. All preprocessing code for the time-stamped knowledge graph construction can be found at <https://github.com/EmilyAlsentzer/rare-disease-simulation>.

Several prioritization algorithms require variant-call format (VCF) files as input. To evaluate these algorithms, we construct synthetic VCF files by sampling a similarly pathogenic SNV for each candidate gene. To this end, we obtained genome-wide precomputed CADD scores, Ensembl VEP consequences and gnomAD minor allele frequencies from the CADD website (https://krishna.gs.washington.edu/download/CADD/v1.6/GRCh38/whole_genome_SNVs_inclAnno.tsv.gz)⁴³ and gene locations from Ensembl Biomart for all genes present in our time-stamped KG and/or listed as a causal or distractor gene in our sets of real and simulated patients. For each gene, we first tried to select the rarest missense SNV with the highest CADD score, then (in the absence of missense variants) the rarest “splice_acceptor” or “splice_donor” variant with the highest CADD score, and finally (in the absence of splice site variants as well), the rarest SNV with the highest CADD score regardless of consequence to represent the gene.

We reimplement phenotype similarity-based gene prioritization algorithms from six well-known prioritization tools, restricting them to utilize data from 2015 or earlier. We use publicly-available code from <https://bitbucket.org/bejerano/phrank> and <https://github.com/WGLab/phenolyzer> to run Phrank¹⁰ and Phenolyzer¹¹, respectively, using our 2015 time-stamped KG. We run the phenotype-only version of LIRICAL with the “orphanet” data flag using code from <https://github.com/TheJacksonLaboratory/LIRICAL> and filter the provided input data file to include only those interactions present in our 2015 KG. LIRICAL considers positive and negative (i.e., that the patient did not exhibit) HPO terms; we supplied all of these terms for all simulated and real patients when running LIRICAL. We reimplement Phenomizer¹² as described in their paper, as open-source code is not available, using our time-stamped KG from 2015. Although the original implementation of Phenomizer randomly samples phenotype terms 100,000 times to generate *P* values for each patient-disease similarity score, we use 10,000 random samplings instead, as this was substantially faster, and varying the number of samplings did not impact overall gene rankings. We define a patient-gene match score for Phenomizer as the highest patient-disease match score across all diseases associated with that gene. Next, we run Exomiser v6.0.0 (released February 13, 2015) using their timestamped input data available from

<https://github.com/exomiser/Exomiser/tree/6.0.0>. We ran Exomiser with the option for their most versatile phenotypic similarity algorithm, HiPhive, which considers mouse, zebrafish and human phenotypic data in addition to gene-gene interactions. The other two phenotype-ranking algorithms available within Exomiser use the already-included Phenomizer algorithm (PhenIX) or use only mouse phenotypic data (Phive)²². We also reimplement two versions of the ERIC phenotype similarity score used in Xrare using our time-stamped 2015 KG: ERIC-Gene, which directly utilizes gene-phenotype interactions, and ERIC-Disease, which utilizes gene-disease and disease-phenotype interactions. Finally, we also run Xrare’s predictive machine learning phenotype matching algorithm (Pred_Phen), called ERIC-Predictive here, available in their Docker, which utilizes disease-gene information known prior to 2011. We modify the provided R script to directly output the computed phenotype match score. We provide the settings files with the parameters used to run each of these algorithms when relevant in our GitHub repository.

We report how well each of these tools—both versions of Phrank, both versions of ERIC and Pred_Phen from Xrare, Phenolyzer, Phenomizer, LIRICAL, and HiPhive from Exomiser—ranked the causal gene for each simulated patient for each different category of novel disorders as outlined in Fig. 1b.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The simulated patient dataset and all intermediate data used in its creation have been deposited in Harvard Dataverse under accession code <https://doi.org/10.7910/DVN/ANFOR3>⁴⁴. Anonymized UDN data has been deposited in dbGaP (accession phs001232) and PhenomeCentral. Phenotypes and causal variants and genes related to UDN diagnoses are also shared publicly in ClinVar: ncbi.nlm.nih.gov/clinvar/submitters/505999/. Our simulation process and analyses leverage the following external databases: Human Phenotype Ontology (<https://hpo.jax.org/app/>), HPO Annotations ([http://github.com/drseb/HPO-archive/tree/master/hpo.annotations.monthly/2015-01-2814-15-03/archive/annotation](https://github.com/drseb/HPO-archive/tree/master/hpo.annotations.monthly/2015-01-2814-15-03/archive/annotation) and <http://github.com/drseb/HPO-arc523hive/tree/master/2014-2015/2015week4/annotations/artefacts>), Orphanet (orphanet.com), Unified Medical Language System (nlm.nih.gov/research/umls/index.html), Human Transcriptional Regulation Interactions Database (available in the Phenolyzer Github at <https://github.com/WGLab/phenolyzer>), Human Protein Reference Database (<https://www.hprd.org/> and in the Phenolyzer Github), CADD (<https://cadd.gs.washington.edu/>), and Ensembl BioMart (<https://useast.ensembl.org/info/data/biomart/index.html>). All data sources used in our analyses can be found on the Harvard Dataverse. Furthermore, the processed data needed to recreate the figures can be found in the “Source Data” section of our Harvard Dataverse. Source data are provided with this paper.

Code availability

The code to reproduce results, together with documentation and examples of usage, can be found at <https://github.com/EmilyAlsentzer/rare-disease-simulation> with <https://doi.org/10.5281/zenodo.8190872>⁴⁵.

References

1. Nguengang Wakap, S. et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
2. Chong, J. X. et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97**, 199 (2015).

3. Gahl, W. A. et al. The national institutes of health undiagnosed diseases program: insights into rare diseases. *Genet. Med.* **14**, 51–59 (2012).
4. Splinter, K. et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
5. Posey, J. E. et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. *Genet. Med.* **21**, 798–812 (2019).
6. Dyment, D. A. et al. Whole-exome sequencing broadens the phenotypic spectrum of rare pediatric epilepsy: a retrospective study. *Clin. Genet.* **88**, 34–40 (2015).
7. Gahl, W. A., Wise, A. L. & Ashley, E. A. The undiagnosed diseases network of the national institutes of health: a national extension. *JAMA* **314**, 1797–1798 (2015).
8. Ramoni, R. B. et al. The undiagnosed diseases network: accelerating discovery about health and disease. *Am. J. Hum. Genet.* **100**, 185–192 (2017).
9. Kobren, S. N. et al. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genet. Med.* **23**, 1075–1085 (2021).
10. Jagadeesh, K. A. et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet. Med.* **21**, 464–470 (2019).
11. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841–843 (2015).
12. Köhler, S. et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* **85**, 457–464 (2009).
13. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
14. Yuan, X. et al. Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases. *Brief. Bioinform.* **23**, bbac019 (2022).
15. Smedley, D. & Robinson, P. N. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* **7**, 81 (2015).
16. Li, Q., Zhao, K., Bustamante, C. D., Ma, X. & Wong, W. H. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet. Med.* **21**, 2126–2134 (2019).
17. Boudellioua, I., Kulmanov, M., Schofield, P. N., Gkoutos, G. V. & Hoehndorf, R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinform.* **20**, 65 (2019).
18. Kumar, A. A. et al. pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. *Bioinformatics* **34**, 2254–2262 (2018).
19. Tranchevent, L.-C. et al. Candidate gene prioritization with Endavour. *Nucleic Acids Res.* **44**, W117–W121 (2016).
20. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. Orphanet and its consortium: where to find expert-validated information on rare diseases. *Rev. Neurol.* **169**, S3–8 (2013).
21. Robinson, P. N. et al. Interpretable clinical genomics with a likelihood ratio paradigm. *Am. J. Hum. Genet.* **107**, 403–417 (2020).
22. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–15 (2015).
23. Smedley, D. et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
24. Deisseroth, C. A. et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet. Med.* **21**, 1585–1593 (2019).
25. Liu, C. et al. Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Res.* **47**, W566–W570 (2019).
26. Chopra, M. & Duan, T. Rare genetic disease in China: a call to improve clinical services. *Orphanet. J. Rare Dis.* **10**, 140 (2015).
27. Birgmeier, J. et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* **12**, eaau9113 (2020).
28. O'Brien, T. D. et al. Artificial intelligence (AI)-assisted exome reanalysis greatly aids in the identification of new positive cases and reduces analysis time in a clinical diagnostic laboratory. *Genet. Med.* **24**, 192–200 (2022).
29. Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K. & Mahmood, F. Synthetic data in machine learning for medicine and healthcare. *Nat. Biomed. Eng.* **5**, 493–497 (2021).
30. Rehm, H. L. Time to make rare disease diagnosis accessible to all. *Nat. Med.* **28**, 241–242 (2022).
31. Orphanet. Inventory of rare diseases. https://www.orpha.net/orphacom/cahiers/docs/GB/eproc_disease_inventory_R1_Nom_Dis_EP_04.pdf (2004).
32. Köhler, S. et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
33. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–270 (2004).
34. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
35. Piñero, J. et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
36. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. [Orphanet and its consortium: where to find expert-validated information on rare diseases]. *Rev. Neurol.* **169 Suppl 1**, S3–8 (2013).
37. Robinson, P. N. et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
38. Strande, N. T. et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *Am. J. Hum. Genet.* **100**, 895–906 (2017).
39. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
40. Shyr, C. et al. FLAGS, frequently mutated genes in public exomes. *BMC Med. Genom.* **7**, 64 (2014).
41. Haghighi, A. et al. An integrated clinical program and crowdsourcing strategy for genomic sequencing and Mendelian disease gene discovery. *npj Genom. Med.* **3**, 1–10 (2018).
42. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Tech. Rep. arXiv:1802.03426, arXiv (2020). ArXiv:1802.03426 [cs, stat] type: article.
43. Rentsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
44. Alsentzer, E., Finlayson, S., Li, M., Kobren, S. & Kohane, I. Simulation of undiagnosed patients with novel genetic conditions. *Harvard Dataverse* <https://doi.org/10.7910/DVN/ANFOR3> (2023).
45. Alsentzer, E., Finlayson, S., Li, M., Kobren, S. & Kohane, I. Simulation of undiagnosed patients with novel genetic conditions. *GitHub Repository* <https://doi.org/10.5281/zenodo.8190872> (2023).

Acknowledgements

We would like to thank Peniel Argaw, Chuck McCallum, Amelia Tan, Yidi Huang, and Mark Keller for their help in manually annotating each disease and disease-gene association in Orphanet according to the date of the Pubmed article that reported the discovery. We would also like to

thank Cecilia Esteves for her help in understanding the Undiagnosed Diseases Network workflows. This work was supported by a Microsoft Research PhD Fellowship (E.A.), award Number T32GM007753 from the National Institute of General Medical Sciences (S.F.), T32HG002295 from the National Human Genome Research Institute (M.L.), and a National Science Foundation Graduate Research Fellowship (M.L.). UDN research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Number(s) U01HG007709, U01HG010219, U01HG010230, U01HG010217, U01HG010233, U01HG010215, U01HG007672, U01HG007690, U01HG007708, U01HG007703, U01HG007674, U01HG007530, U01HG007942, U01HG007943, U01TR001395, U01TR002471, U54NS108251, and U54NS093793. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

E.A., S.F., S.K. and I.K. designed the study. E.A. and S.F. implemented the simulation pipeline and gene prioritization algorithms, and E.A. conducted the validation of simulated patients and ablation analysis. M.L. assisted with the UMAP analysis. E.A., S.F. and S.K. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-41980-6>.

Correspondence and requests for materials should be addressed to Shilpa N. Kobren or Isaac S. Kohane.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Undiagnosed Diseases Network

Shilpa N. Kobren ¹  & Isaac S. Kohane ¹ 

A full list of members and their affiliations appears in the Supplementary Information.