Automated Annotation of Disease Subtypes

Dan Ofer, Michal Linial*

Department of Biological Chemistry, The Life Science Institute, The Hebrew University of Jerusalem, Israel

Email: [D.O] dan.ofer@mail.huji.ac.il, [M.L] michall@mail.huji.ac.il

ORCID: [D.O] 0000-0001-5136-8014; [M.L] 0000-0002-9357-4526

* **Corresponding author:** Michal Linial, Department of Biological Chemistry, The Life Science Institute, The Hebrew University of Jerusalem, 91904 Israel

Abstract

Background

Distinguishing diseases into distinct subtypes is crucial for study and effective treatment strategies. The Open Targets Platform (OT) integrates biomedical, genetic, and biochemical datasets to empower disease ontologies, classifications, and potential gene targets. Nevertheless, many disease annotations are incomplete, requiring laborious expert medical input. This challenge is especially pronounced for rare and orphan diseases, where resources are scarce.

Methods

We present a machine learning approach to identifying diseases with potential subtypes, using the approximately 23,000 diseases documented in OT. We derive novel features for predicting diseases with subtypes using direct evidence. Machine learning models were applied to analyze feature importance and evaluate predictive performance for discovering both known and novel disease subtypes.

Results

Our model achieves a high (89.4%) ROC AUC (Area Under the Receiver Operating Characteristic Curve) in identifying known disease subtypes. We integrated pre-trained deep-learning language models and

showed their benefits. Moreover, we identify 515 disease candidates predicted to possess previously unannotated subtypes.

Conclusions

Our models can partition diseases into distinct subtypes. This methodology enables a robust, scalable approach for improving knowledge-based annotations and a comprehensive assessment of disease ontology tiers. Our candidates are attractive targets for further study and personalized medicine, potentially aiding in the unveiling of new therapeutic indications for sought-after targets.

Keywords: Disease subtypes; Disease ontology; Explainability; Machine learning; Medical language models; Ontology completion; Open Targets; Orphanet; Personalized medicine.

1. Introduction

Disease subtyping, also called disease stratification, enables a more precise understanding and characterization of various illnesses, paving the way for personalized treatments and improved patient outcomes. Disease subtypes can be delineated using genetic, molecular, or clinical attributes [1]–[3].

As personalized medicine advances, disease subtyping can advance our understanding of disease mechanisms across various medical disciplines [3], [14]. Moreover, it is needed for study, effective treatment, and discovering potential cures. Furthermore, certain drugs and treatments may be relevant only for specific subpopulations and disease manifestations [15], [16]. Disease progression can also be markedly different, requiring different clinical treatment regimes [17]–[19].

We concentrate on clinically significant differentiation, or subtyping, of diseases. For instance, variants of SARS-Cov-2 caused the COVID-19 pandemic. We claim that subvariants (e.g., delta, omicron) are not useful for clinical categorization. Instead, the partition of COVID-19 to patients experiencing an acute phase and others who exhibit persistent conditions known as long COVID dictates clinical importance. Another example of disease subtyping is evident in the differentiation between type 1 and type 2 diabetes mellitus, where, despite the similarity in dysregulation of blood sugar levels, treatment approaches, disease management, and potential cures vary significantly. Neurodegenerative disorders like Alzheimer's and Parkinson's, although categorized clinically as neurodegenerative diseases, exhibit distinct molecular pathologies, subtypes, and diverse progressions and treatments. Advancing our understanding of Parkinson’s subtypes is pivotal for devising effective treatment strategies [1].

Conversely, while various viruses can cause influenza, differentiating them based on their specific causal virus is clinically irrelevant since treatment and disease progression remain identical.

Historically, the medical community has spearheaded efforts to identify the multifaceted nature of diseases. Clinicians primarily rely on the International Classification of Diseases (ICD), which undergoes periodic revisions [4]. For instance, Diabetes mellitus (ICD-10, E10-E14) is partitioned into Type 1 (E10), Type 2 (E11), and unspecified diabetes (E14), along with further subtypes [5], [6].

The Open Targets (OT) platform integrates a variety of molecular, genetic, and biomedical datasets, ontologies, and knowledge graphs [7]. Increasing quantities of semantic resources offer a wealth of knowledge but also increase the probability of wrong knowledge-based entries and error propagation [8]–[11]. Thus, developing automated approaches to both complete and correct potentially spurious entities in large knowledge bases is of paramount importance. The concept of accurate hierarchical categorization of diseases and phenotypes is further underscored by initiatives like the gene ontology (GO) project, where ontologies, phenotypes, and functions across species are mapped to coding genes [12]. The impact of the Gene Ontology (GO) project on automatic functional annotation tasks such as CAFA is unquestionable [10], [13]. CAFA (Critical Assessment of Functional Annotation) is an ongoing effort to evaluate and improve the computational annotation of protein functions.

Existing ontologies are complex and may suffer bias due to many factors, including population prevalence and the number of researchers and clinicians working on the disease, factors that may impact their division into sub-categories in the literature as well as the quality of annotation [20]–[22].

Most existing methodologies rely on inheriting or directly mapping disease levels from existing annotations and ontologies by strict, manually defined rule-based methods. One concentrated on a narrower domain, clustering specific cancer data, imaging, and non-biomedical data [23]. It did not endeavor to offer predictions across a broad spectrum of known diseases. Another approach used by OT data is to identify drug-disease associations, which is a different objective from ours [24]. Our work also relates to knowledge-based link prediction, and literature-based discovery [25]–[29]. However, these mainly aim to identify “horizontal links” between existing topics. In contrast, our objective is to flag topics that might have undiscovered subtopics, or missing “vertical” links.

We propose a data-driven machine learning approach for ontology completion and correction, specifically applied to OT. OT integrates a wide variety of gold-standard curated ontologies and data sources, from which we curate a novel benchmark dataset for disease subtype prediction. This dataset can be used for evaluating and developing approaches for characterizing diseases. Furthermore, we present an approach for identifying and evaluating candidate diseases with potential novel subtypes

and mis-annotations, and a ranked list of predictions. We validate our novel candidates using ongoing research and future OT annotation updates. Our automated approach is interpretable, scalable and offers novel candidate disease subtypes for future research.

We outline the key steps in our approach. First, we create a target matrix from the existing OT ontology for all diseases, defined as whether a disease has a subtype or not. Predictive features for each disease are derived from OT's direct evidence data sources. A machine learning model is trained on known targets. Predictions are formulated for every entry in the dataset through iterative rounds of hold-out cross-validation. Subsequently, we interpret and scrutinize the results and models. Instances where the predicted target consistently deviates from the known one, coupled with supplementary filtering, are identified as potential candidates for novel subtypes or highlight annotation inaccuracies. Our goal is to help find unknown disease subtype candidates within existing databases.

2. Results

2.1 Diverse Disease Ontologies

A machine learning model for disease subtyping assessment and discovery was developed using the sources integrated into OT. An overview of these sources is demonstrated in **Fig 1.**

Fig 1A is a disease perspective view of type 2 diabetes mellitus on the OT platform [30]. It encompasses text associated with the description, synonyms from various databases, and summary statistics for additional information, including ontologies, known drugs, clinical signs, symptoms, and bibliography. **Fig 1B** lists associated genes for type 2 diabetes mellitus ranked by global scores for the disease. The genes' evidence is indicated by the heatmap, with genetic association from genome-wide association studies (GWAS), direct support from drugs, text mining, RNA expression, animal model studies, and more.

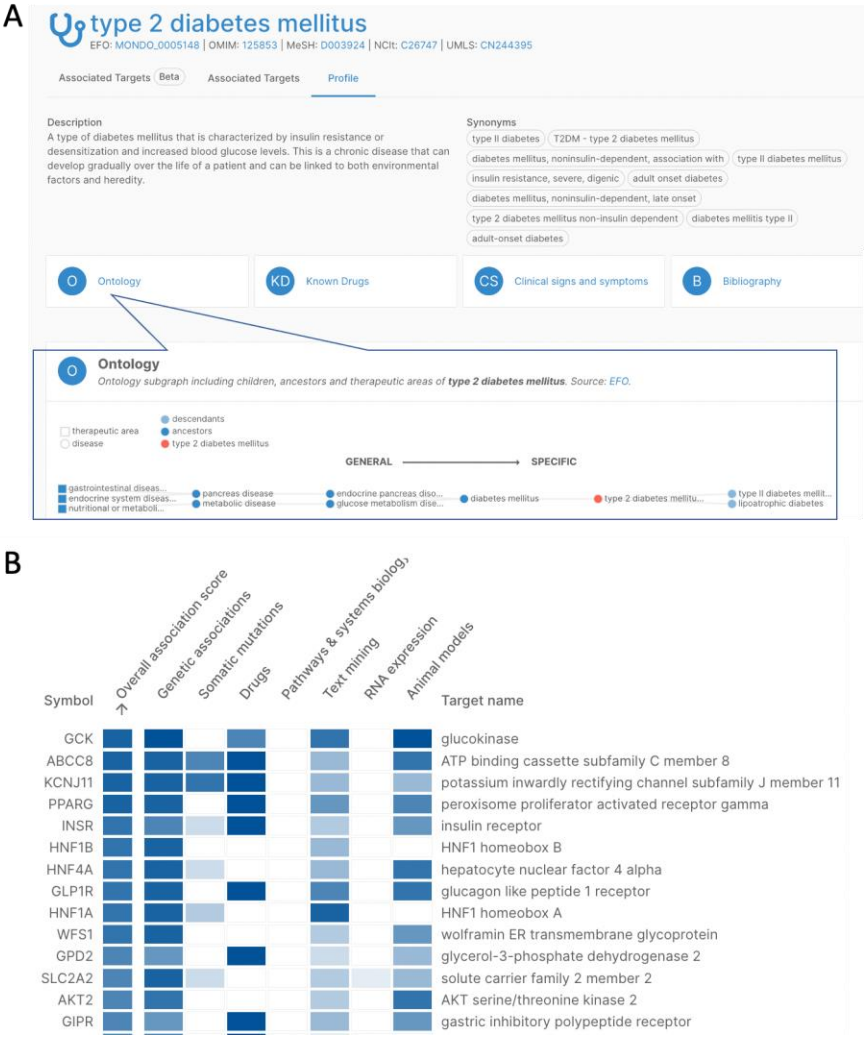


Fig 1. *Example from Open Targets Platform for type 2 mellitus diabetes.* ***(A)*** *Description, disease ontology (including any subtypes), synonymous nomenclature, known drugs, bibliography, and clinical symptoms* ***(B)*** *Weighted evidence sources and domains for the disease, including genetics (genetic associations), somatic mutation, drugs, text mining, and more. Each column is colored by the intensity of the relevant score (normalized 0 to 1). The gene list is sorted by the overall association score.*

The final dataset held 17,222 diseases, of which 5,848 (34%) have known subtypes. Feature importance and model performance were evaluated in predicting targets with known subtypes. Our novel features demonstrate high importance and predictive power. In addition, these may support the discovery of novel therapeutic indications for highly pursued targets.

2.2. Performance Evaluation - Known Targets

For the task of predicting known disease subtypes, we tested multiple machine-learning models, including logistic regression (LR), random forest (RF), CatBoost (a boosting tree model), as well as domain-specific baselines. Binary classification performance was evaluated using five-fold stratified

cross-validation [31]–[33] (**Fig 2)**. Additional evaluation results are provided in Supplementary table S1, and the confusion matrix of the best, CatBoost, model in Supplementary Fig S1.

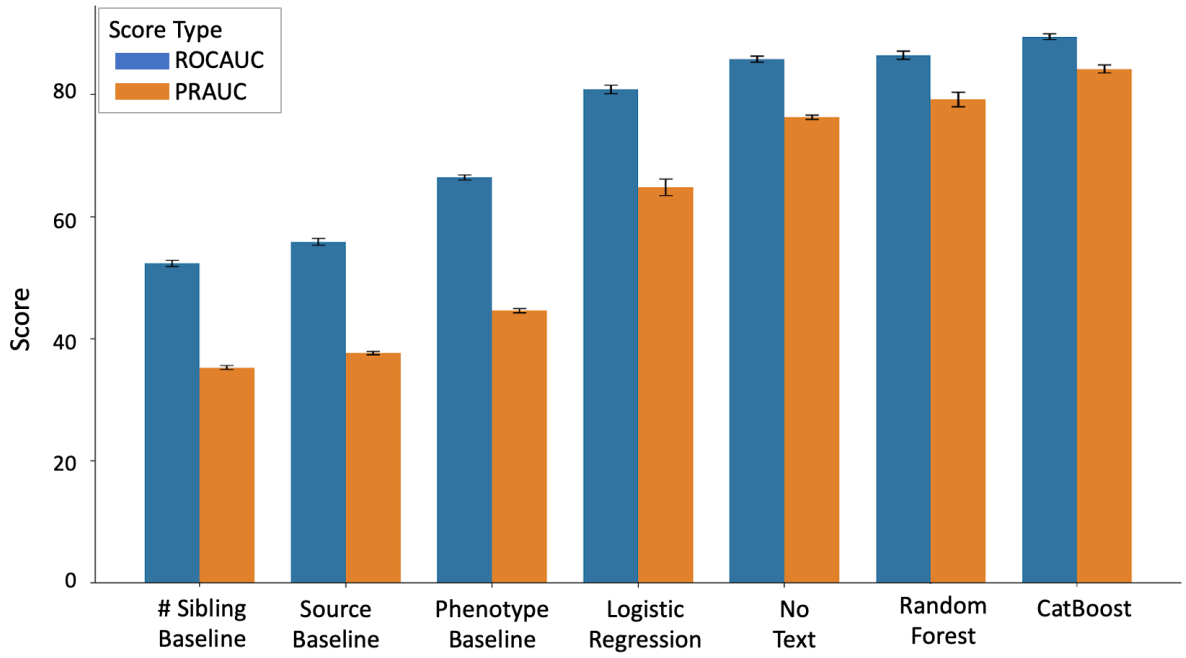


Fig 2. *Known subtypes model evaluation. Results shown 5-fold cross-validation, and standard deviation, for a selection of evaluated models. PRAUC: Area Under Precision-Recall Curve. No Text - CatBoost model with text embedding features excluded (see Methods: "Deep learning using Text features"). The remaining, non-baseline models used all features. Additional models and metrics results* are reported *in Supplementary table S1.*

For comparative analysis, we added to the assessment three domain-specific baselines. These are linear models trained exclusively on a single feature: (i) the disease's database source (e.g., Orphanet); (ii) the number of known phenotypes ("phenotype frequency"); (iii) The number of "siblings" a disease has in OT database, wherein all "siblings" share the same parent disease. All baselines outperform random guessing. All models substantially outperform the baselines (**Fig 2**). CatBoost had the best performance, achieving a ROCAUC (Receiver Operating Characteristic Area Under the Curve) of 89.4% (**Figs 2-3**). Accordingly, we used CatBoost for subsequent predictions of novel subtypes and analyses. This included extracting novel predictions and ablation analysis of the text features. We find that text features significantly enhanced performance when compared to a model devoid of text features (**Fig 2**, "No text model"), yielding an AUC of 0.89, in contrast to 0.86 without these features.

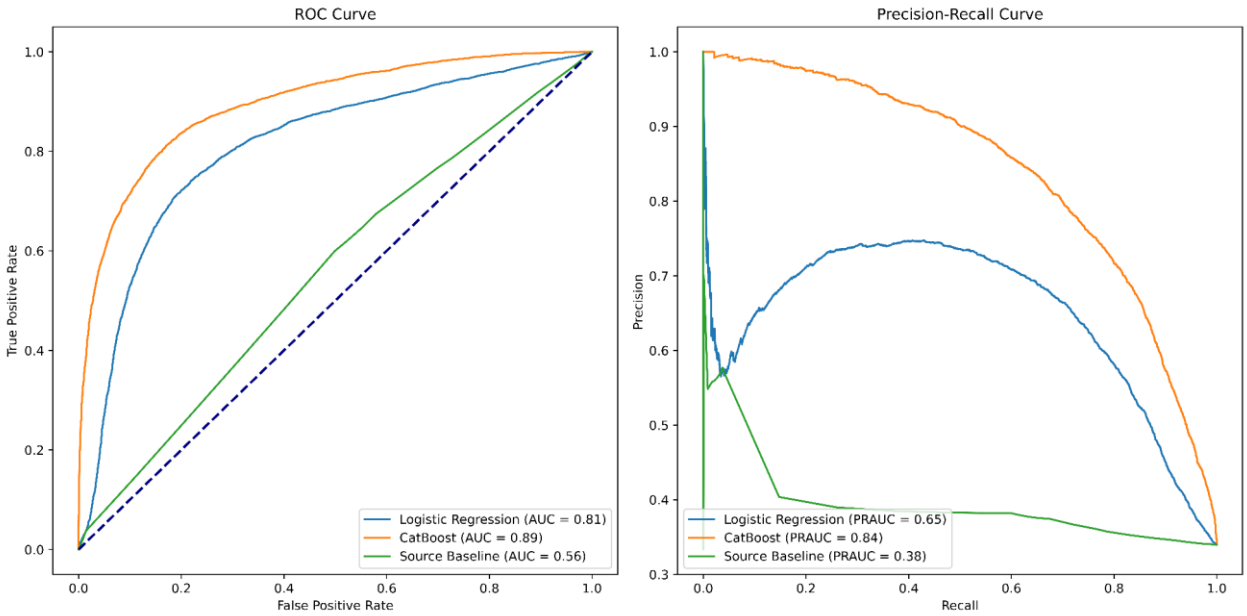
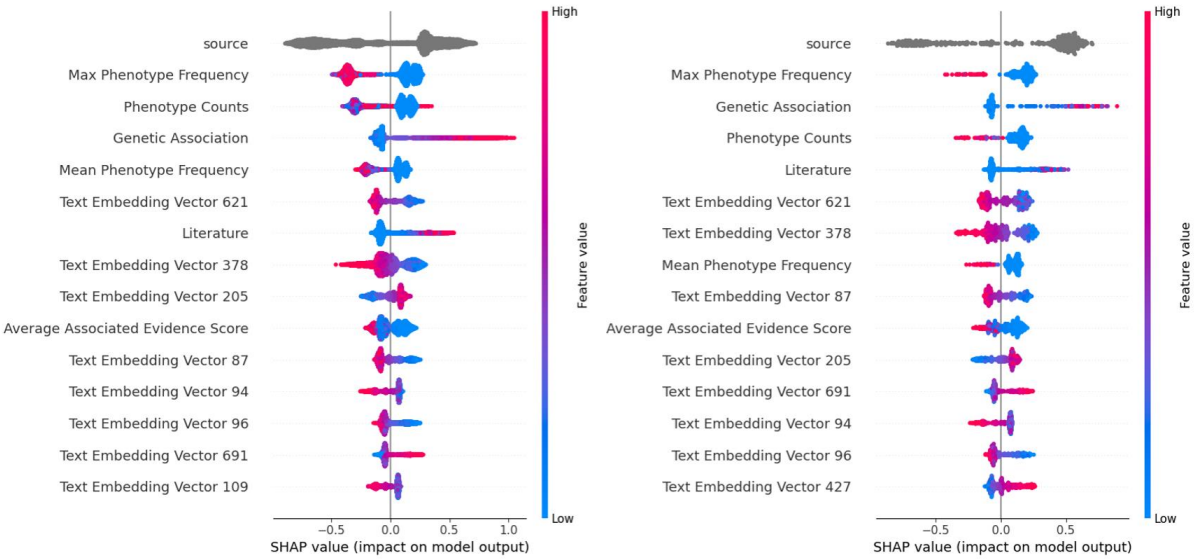


Fig 3. *Known subtype models comparison. Model test-set predictive performance is measured by ROCAUC (left) and PRAUC (right).*

2.3. Model Feature Importance

The "known subtypes" model's feature importance was extracted using SHAP **(Fig 4A)**. We observe that the source database is a major feature, as we might expect (e.g., Orphannet diseases are more likely to be understudied and to lack subtypes). Diseases with a high amount of genetic and literature evidence score (**Fig 1**) are more likely to have subtypes. Specific disease phenotypes were strong features in aggregate, but consisted of hundreds of individual weaker features, and thus are not visible here; the engineered feature of the highest global frequency of an associated phenotype ("Max phenotype frequency") is strong - we theorize it might help the model learn about diseases with easy versus hard to characterize phenotypes. The number of phenotypes is another interesting feature. disease with many different effects may be more complicated to stratify or maybe a combination of effects. The various text features from the pretrained biomedical large language model have a clear impact (see Methods: "Deep Learning Text Features"). These might help extract additional information about diseases from their descriptions or pre-existing literature.



(A) Overall Feature Importance

(B) Feature Importance for Novel Subtypes

Fig 4. *Feature importance over data subsets. Shapley-based feature importance in disease subtype prediction for A) All data (Existing OT subtype annotations). B) Feature importance for subset with predicted novel subtypes, and no subtype in ground-truth. Top 15 features shown. See Methods for feature dictionary.*

2.4. Predicting Diseases with Novel Subtypes

Given the strong performance of machine learning models in our evaluation, we used our model to predict disease subtypes for diseases that are not identified as such in the existing dataset. We applied a repeated-stability approach to identifying potential novel candidates. We report on cases where a model, retrained over multiple random data splits (using 8x5 repeated stratified cross-validation), consistently (eight out of eight times) predicts a different label than the recorded one per data point. The predicted data point is always part of the held-out test set. Thus, there are 8 held-out model predictions for every instance in the data. We identified 1,546 such cases, out of 17,222 records in the dataset. Of these, 515 (33%) are predicted to have subtypes, where none are recorded in the OT ground truth. The average prediction consistency was 84.9%. This approach is effectively an ensemble. In supplementary S5 we record the averaged model predictions on the held-out test set splits. This ensemble has better results than a single model, as might be expected (ROCAUC: 91.08, PRAUC: 86.61) [19], [34]. Thus, we used these as candidate predictions. The full list of candidate predictions is in Supplementary File S5.

Using SHAP we examined features' contributions to model predictions, for the predicted novel subtypes subset only **(Fig 4B)**, and observed that the top features (evidence source, phenotype etc) remain relatively stable in terms of rank importance and direction of effect, with the same effect as for known subtypes in the general population (**Fig 4B**), indicating that these cases are not anomalous in terms of their features compared to the background. We plotted the distribution of several "top" (selected by model importance) features, and observed a similar distribution overall (**Fig 5)**, again reinforcing that the novel subtypes have similar properties to the ground-truth known subtypes.

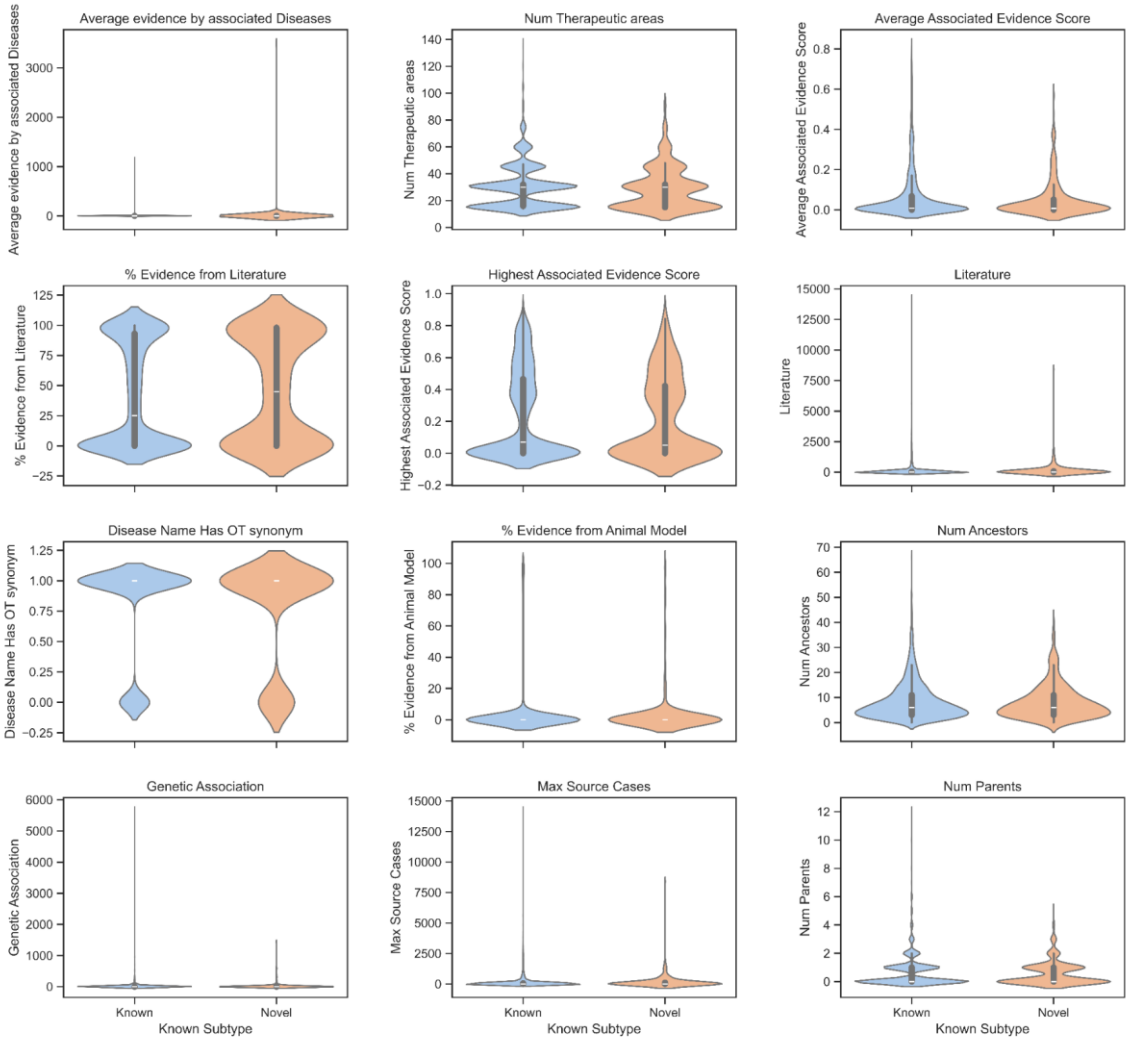


Fig 5. *Distribution of known vs novel subtypes. Violin-plots of selected features for 5.8K diseases with* ***known*** *subtypes vs. 531 predicted, candidate* ***novel*** *subtypes that have no known subtype.*

2.5. Evaluation of Predictions for Unknown Targets

As there is no ground truth to evaluate the novel subtype predictions, we use the scientific literature as an external validation mechanism, as in [24]. We presume that potential subtypes might be discussed in the wider literature before any validation and integration into existing knowledge. From our models, we selected 800 predictions, sorted by highest predicted probabilities, evenly split with 400 cases per target class. We searched for literature hits of these candidate diseases being mentioned as having a subtype, sub-manifestation, or pleiotropy in PubMed. We extracted the fraction of such cases relative to the total number of literature references for each disease (Supplementary Table S3). The difference between candidates categorized by their predicted subtype was statistically significant as determined by a one-sided unequal variance t-test (*p-value = 2.6e-7*), with predicted candidates having ~5 times as many results with subtypes (i.e., 2,100 vs. 400).

We conducted a similar analysis on 300 of the stable predicted novel candidates (Supplementary Table S4), specifically those whose predictions deviated from existing annotations. For this subset, the literature search revealed no significant difference (p-value=0.27). This aligns with our hypothesis that these novel candidates are uncharacterized in existing studies. If our model's predictions were merely identifying diseases widely acknowledged to have a subtype but not yet annotated in OT, this would be evident by the literature search (i.e., numerous "subtype" mentions), which was not observed here. This observation bolsters our assertion that the model is highlighting truly unknown subtype candidates, rather than just inadequate or faulty annotations.

2.5.1 Temporal Validation of Candidates

For further validation that our candidate predictions are meaningful, we downloaded an additional, 18-month newer snapshot of the OT database (dated 12.2023). Of the 19,819 diseases that overlap in both versions (the total population), 1.2% (238) had subtypes added or removed entirely between versions. 1509 (92%) candidates were successfully matched by name, and of these, 48 (3.18% of the candidate sub-population) had their subtype annotations changed. The 3-fold proportional difference (3.18% vs 1.04% in the sub-population not containing candidates) is statistically significant (p-value <1e-05, one-way, two proportion z-test). This shows that our candidates are far more likely to have their existing annotation ground truth "fixed" in accordance with our predictions, relative to the overall population.

2.5.2 Analysis of Database Source Distribution in Predictions

A concern was whether the model was simply identifying surface-level patterns, such as associating all diseases from a specific data source, such as Orphanet (an orphan disease database) with a subtype. Such hidden confounders are common in many predictive scenarios [35], [36]. To validate the model, we examine the distribution of disease subtypes in our predictions against the known subtypes, focusing on their source database (**Table 1**). We find the distribution of our novel candidates slightly differs from the original dataset target at the database level. Notably, there is a lower frequency of subtypes in predictions from Orphanet. This gives further support to our identification of candidates using non-trivial patterns, and reducing bias towards existing annotation sources.

Table 1 shows the distribution of diseases with known vs. predicted subtypes, grouped by the 6 largest database sources. Diseases' subtype fraction is shown per source. "Original Source Subtype Fraction" depicts the percentage of diseases with a subtype in existing annotations. "Predictions Source Subtype Fraction" depicts the percentage of diseases with predicted subtypes in novel predictions. "Total Source Diseases'', indicates the total number of diseases (regardless of subtype) from a source.

Table 1. Source subtype distribution

Source[a]	Original Source Subtype Fraction (%)	Predictions Source Subtype Fraction (%)	Total Source Diseases
MONDO	37	35	8801
EFO	28	22	4590
Orphanet	28	23	2061
HPO	36	32	1460

GO	56	55	352
OBA	11	0	27

[a]Sources are described in [12], [58-62].

2.5.3 Understanding Potential Novel Disease Subtypes

Following our model's identification of 515 diseases predicted to have subtypes not currently annotated in OT, understanding the significance of these findings becomes paramount. These candidates, selected based on their consistent predictions and absence of known subtyping, underscore the vast potential for refining our understanding of disease taxonomy. The top-ranked novel predictions were manually reviewed. We provide several explanations for representative novel candidates **(Table 2)**. Broadly, we note high-level causes that include: (P) Pleiotropic manifestations: different causes resulting in seemingly similar outcomes, leading to diseases with varied presentations. Additionally, overlapping clinical presentations can cause misdiagnosis. Examples include the confusion between CNS inflammatory disorders and multiple sclerosis (MS), neurodegenerative diseases and dementia [17-18,40]. (B) Variability in disease course and treatment: Clinical trajectory and therapeutic responsiveness can vary based on disease subtypes and interaction with patient characteristics (e.g., Parkinsons' [1]). (H) "High-level" semantic terms: these are inherently broad and include a range of conditions, e.g., "infections". Such cases are clear-cut and may be due to a lack of linkage of known terms between ontologies.

In the cases of diseases predicted to be misannotated and to not have a subtype; some cases may simply be model errors, hence the need for a final layer of expert review. There are numerous valid explanations for why a disease may have an incorrect annotation, ranging from human error, database error, and annotator guidelines biases. We illustrate it by a hypothetical case of two different virus strain variants of the SARS-CoV-2 Omicron strain being classified as two distinct diseases. While they are caused by a separate strain, their disease manifestation and course of treatment overlaps with that of a flu-like illness. In this case, we claim that the distinction is not clinically meaningful, but could be recorded as such by mistake. Merging such subtypes would improve the database. Our temporal validation showed that our candidates are much more likely to be "interesting" and in need of reassessment in the OT database.

Table 2. Candidate novel subtypes' explanation by categories

Disease	**Explanation**	[a]**Rationale**	**Ref**
COVID-19	The spectrum of clinical presentations, from asymptomatic states to severe respiratory syndrome, and long-term impairment ("long COVID") suggests potential subtypes.	B	[42], [43]
Smallpox	There are four types: ordinary; modified (mild, occurring in previously vaccinated persons); flat; and hemorrhagic. For example, the Smallpox vaccine does not protect against hemorrhagic smallpox.	B	[63]
Female breast carcinoma	Varied molecular subtypes, defined by specific gene expression profiles, as well as anatomical regions, dictate prognosis and therapeutic responsiveness.	P, B	
Nephropathy	Kidney pathologies with distinct histological features and clinical courses.	P, B	
Cardiovascular risk	Disease families with broad risk factors, including environmental, genetic and genetic-environment interaction (GxE). For example, diabetes, or specific genetic risk factors.	P, B	[6], [15]
Structural epilepsy	Different brain anomalies can precipitate varied forms of epileptic seizures. It may coexist with tumors, cysts, stroke, or vascular malformations.	P, B	[41]
Parkinson disease, mitochondrial	Neurodegenerative disorders, that though clinically overlapping, have distinct molecular pathologies, and treatment..	B	[1]
(Multiple) malignant and non-malignant tumors	Behavior, complications and treatment are influenced by specific mutations, patient genetics and risk factors (GxE), physical size, organ location.	H	
Eye infections	Distinct etiological agents, spanning bacteria, viruses, fungi, and parasites, involving distinct treatment and risks.	H	
Alcohol dependence; Sexual and gender identity disorders; Speech disorder; Central nervous system development	Semantically high-level categories, lacks breakdown to a clinically useful level, required for effective treatment	H	

Dysplasia	A high-level family of conditions (encompassing "types of abnormal growth or development of cells, organs", and resulting abnormalities. Different categories by delineation by microscopic (cell) level, organ (macroscopic), organ and cell type.	H	[44]

[a]Rationale: High level rationale codes: (P) Pleiotropic manifestations. (B) Variability in disease course and treatment. (H) "High level" semantic terms.

In the cases of diseases predicted to be misannotated and to not have a subtype; some cases may simply be model errors, hence the need for a final layer of expert review. There are numerous valid explanations for why a disease may have an incorrect annotation, ranging from human error, database error, and annotator guidelines biases. For example, a hypothetical case of 2 different variants of Omicron strain of SARS-CoV-2 being classified as two distinct diseases. While they are caused by a separate strain, it is the common state of sessional flu. In this case, we claim that the distinction is not clinically meaningful, but could be recorded as such by mistake. Merging such subtypes would improve the database. Our temporal validation showed that our candidates are much more likely to be "interesting" and in need of reassessment in the OT database.

2.5.4 Understanding an Individual Prediction

We present an illustratory model explanation example for a novel predicted subtype candidate, COVID-19 (**Fig 6**), using the known subtypes model's SHAP explanation.

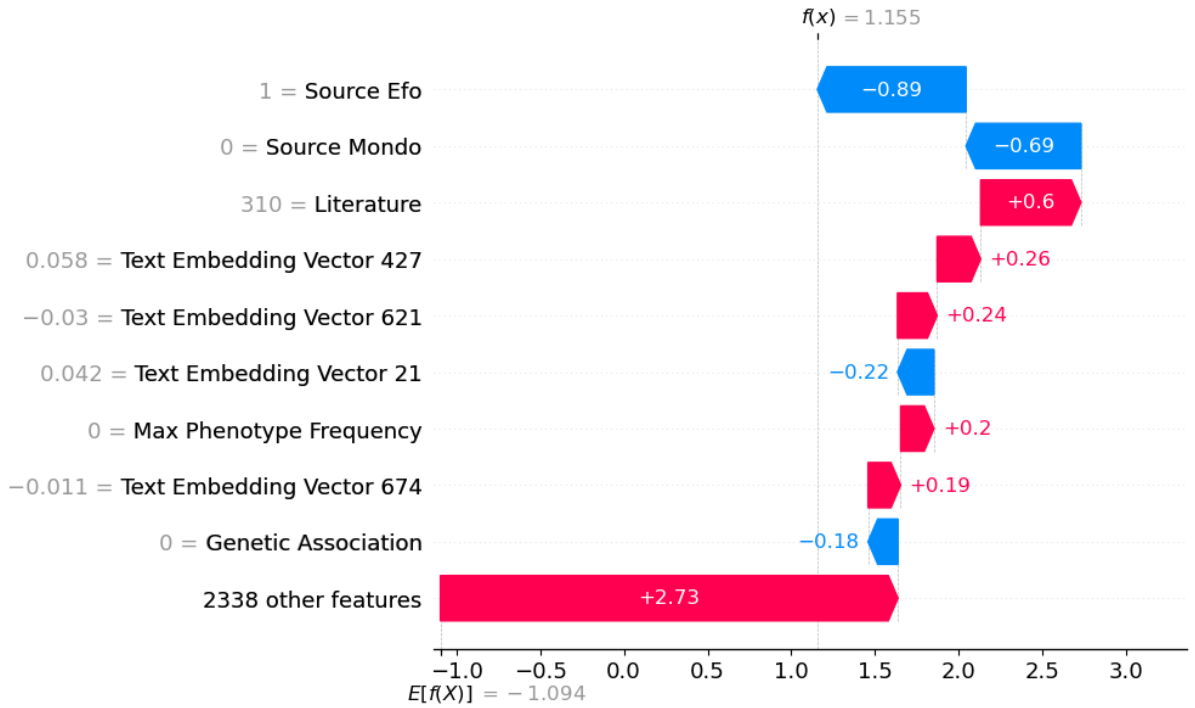


Fig 6. *SHAP explanations for COVID-19. Explanation of a single positive ("1" - has subtype) prediction*

as a SHAP waterfall plot. The SHAP value of a feature represents the impact of the evidence provided by that feature on the model's output. The waterfall plot shows how the SHAP values (evidence) of each feature move the model output from the prior expectation under the background data distribution, to the final model prediction given the evidence of all the features. Colour and direction of the arrows indicates the direction of effect. Feature values shown on the left in grey.

Examples of exemplar disease text excerpts with of high, neutral and low values per each of the embedding dimensions (see 4.4) are shown (**Table 3**)

Table 3. Text embedding explanations for COVID-19

Text embedding dimension	Sentences with high values in dimension	Sentences with low values in dimension	Sentences near median dimension value
427	malignant germ cell tumor pregnancy or perinatal disease neutrophil count radiation	circumscribed cutaneous aplasia of the vertex necrotizing sialometaplasia 2-3 toe syndactyly symphalangism with multiple anomalies of hands and feet short toe	ciliary dyskinesia, primary, 45 syndromic obesity human herpesvirus 7 seropositivity pseudobulbar palsy coxa vara
621	haddad syndrome combined oxidative phosphorylation defect type 27 x-linked intellectual disability-acromegaly-hyperactivity syndrome dental enamel pits acute lung injury	keratinization disease urinary bladder, atony of abnormality of the urinary system physiology bladder neck obstruction urinary tract smooth muscle contraction	shortening of all distal phalanges of the fingers hereditary geniospasm autosomal recessive hypohidrotic ectodermal dysplasia amyotrophic lateral sclerosis type 18 n-acetylaspartate deficiency
21	mucocutaneous leishmaniasis fanconi anemia complementation group d1 tuberculous fibrosis of lung bovine respiratory disease complex dacryocystitis - osteopoikilosis	benign neoplasm of adrenal gland adrenal gland neoplasm non-functioning endocrine neoplasm adrenocortical adenoma benign endocrine neoplasm	inherited creutzfeldt-jakob disease kallmann syndrome alopecia - contractures - dwarfism - intellectual disability syndrome polycythemia negative regulation of immune response
674	deafness - hypogonadism ovarian failure perrault syndrome	integumentary system cancer muscle cancer malignant dermis tumor appendix cancer	wide mouth symptomatic form of hemophilia a in female carriers thyroid gland hyalinizing trabecular tumor syndromic gastroduodenal malformation

3. Discussion

Discovering disease subtypes is an important problem in medicine, with applications in both basic research and personalized treatment. That a disease may even potentially have subtypes is not an obvious fact. Historically, diagnoses like "female hysteria" led to ineffective treatments, overshadowing the recognition of genuine diseases or conditions [37]. It wasn't until later that such broad diagnoses were deconstructed into specific diseases. Parkinson's Disease, which can be caused by drug toxicity or vascular malfunction, is another major example. There is considerable work into finding subtypes that can help treat patients and predict the future course of the disease's progression [1], [17].

Our hypothesis posits that diseases with distinct subtypes are discernible based on intrinsic aspects of the disease itself as well as meta-features relating to its research. Consider cancer diseases that are mostly driven by somatic mutations as opposed to predisposition germline genetic variants. Single-gene disorders are often inherited but they may split into early and late onset diseases that might be addressed differently in clinical terms. In this aspect, early and late onset of diseases are documented to Alzheimers [67]), Parkinson [1],[68] and numerous autoimmune conditions (e.g., Crohn's disease, myasthenia gravis). Thus, we focused our features on representing these aspects of different diseases. We also address "meta-science" aspects about diseases, such as their research process, and limits on studying them. For example, diseases with many distinct animal models (as reported by OT), extensive literature and many candidate drugs, are less likely to be categorized as "orphans" than those observables only in humans. Furthermore, some phenotypes are easier to observe, measure and categorize, while others may be more nuanced. Obvious cases include developmental disease causing facial deformities vs mental health conditions. Our features help the models learn these various aspects, in a way which aims to be largely objective, with the goal of learning from the intrinsic characteristics of the diseases themselves, as opposed to approaches that might be more biased towards existing literature and annotations, such as text mining [34], [38], [39]. This approach can yield better predictive performance than underlying, partial annotations or rule-based systems, as has been observed in other works, such as healthcare mortality prediction[19].

Possible confounders are the rarity of diseases. This can be partially quantified by the disease's population prevalence, using the UK biobank (UKB) [40] population frequency. Another metric of rarity is the classification of a disease as an "orphan disease", which is determined by its source (e.g., listed in Orphanet). Interestingly, the UKB calculated prevalence has low feature importance, and is not an impactful feature to the model predictions (Fig 4). Its removal did not affect model evaluation results (not shown), further emphasizing that our task is indifferent to it. Thus, we disqualify it as a

confounding proxy for model performance. On the other hand, the source feature is important overall. We evaluated a model using only the source database feature (Fig 3**)**, and observed it to be significantly inferior, yielding a ROCAUC score of just 0.55, indicating that source is also not sufficient to explain the model's performance. The same findings held for our other evaluated baselines.

Novel disease subtypes partitioning represents a challenging problem for both clinical and scientometric researchers. Despite the high quality medical ontologies already integrated into OT database, many diseases with similar symptoms may result from different causes, such as with pleiotropic genetic diseases, but they may not be annotated as such, even when pleiotropy is known (but their subtypes are not well defined), especially when the subtypes of pleiotropy are ambiguous [45]–[47]. This creates challenges when using OT to retrieve missing target-indication hypotheses due to the absence of direct candidates. While genetic association evidence on target-disease pairs can offer insights into relatedness [24], for our targets we lack actual negatives, or even a proxy measure such as annotation quality.

When awareness of possible subtypes exists, their identification and validation is currently manual, demanding exhaustive work by experts, who must also propagate their work into existing knowledge bases while drumming up awareness and consensus. Diseases common in developed countries, where clinicians have the resources to work with researchers may be more likely to be distinct, as opposed to neglected diseases in economically disadvantaged countries where doctors may not have the capacity to get their work published [48]–[50]. This issue may be worse for rare orphan diseases, which may have only a handful of dedicated researchers, reducing the chances of distinct manifestations being recognized and correctly annotated in knowledge bases.

To date, these limitations have restricted disease subtype discovery to a purely manual process, motivating our novel approach. We integrate OT direct evidence about each disease, including genetic, physiological and clinical features which are evaluated for prediction of which diseases have subtypes using machine learning models. We integrate unique features for each disease to represent their underlying properties, enhancing the identification of novel pleiotropies or overlooked annotations. Ultimately, our model produces a ranked shortlist of both new and potentially misclassified subtypes, which can then be validated by domain experts (S5).

3.1 Potential Implications

The methodologies in our study, including the combination of machine learning with OT, could extend to broader works. In clinical diagnostics; the increased identification and annotation of disease

subtypes could provide better diagnoses, disease progression tracking and more personalized treatments and enhanced patient outcomes. In drug development, a nuanced understanding of disease subtypes can significantly benefit pharmaceutical research. By pinpointing specific diseases likely to have diverse pathologies, therapies and druggable, can be targeted with greater precision.

3.2 Conclusions

Annotating disease subtypes is crucial for enhancing our understanding of pathology and refining therapeutic strategies. By delineating diseases into subtypes, we pave the way for targeted research and treatment. We show that known disease subtypes can be mostly characterized automatically, that several diseases are likely to have uncharacterized subtypes, and a stability approach to identify them as a prelude to expert refinement and confirmation.

4. Methods

4.1. Overview Processing of Open Targets Dataset

Data was downloaded from Open Targets, as of July 2022 (7.22). The primary OT data sets used were associationByOverallDirect, diseaseToPhenotype, associationByDatasourceDirect, diseases. The subtype target was defined using the OT diseases dataset, according to whether a disease has any child links ("has_children"). The overall distribution of subtypes ("children") across diseases is shown in **Fig 7**. We kept only direct associations, as we determined that indirect associations may leak target information.

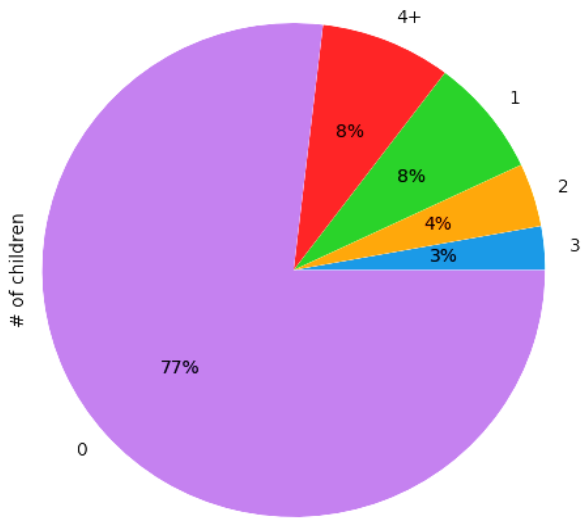


Fig 7 - *Subtypes per disease in Open Targets. The number of 'children' is shown. For 77% of the diseases, no partition to subtypes were recorded (marked as 0).*

The initial dataset held 23,074 candidate diseases. We removed 2,817 irrelevant "diseases" relating to lab measurements (e.g., "IgG index" "BMI"). Diseases with the same name were aggregated together, with any "positive" target label taking precedence. 6,643 (32.8%) of disease terms have a known subtype. 3,035 terms with near-identical names (after removing white-space and lower casing) were merged, with positive subtype label taking precedence. The final modelling dataset held 17,222 diseases, of which 5,848 (34%) have at least 1 known subtypes. The temporal validation dataset uses a snapshot of the OT diseases table from 12.2023

4.2. Model Training and Evaluation

For most models, default hyperparameters were employed, with mean imputation of missing variables. For the tree models, we adjusted training class weights loss using the "square root balanced" hyperparameter. Logistic regression, K-nearest neighbours (KNN), linear support vector machine, histogram gradient boosting and random forest models were implemented using scikit-learn [31]. CatBoost, a boosting tree model, used the library of the same name. Features with a variance lower than 5e-4 were dropped. Tree models were used and favoured due to their speed, interpretability and historically superior performance on tabular data tasks. We also found that the tree models had the best performance on the task, as expected [19].

To evaluate the prediction of known disease subtypes, we employed stratified 5-fold cross-validation across all data points. In each iteration, the dataset was divided into training and testing sets, comprising 80% and 20% of the data, respectively. After training the model on the training set, predictions were made on the test set and recorded. The overall results were then assessed based on the accumulated test split outcomes. We also explored stratifying splits by ancestor disease groups, ensuring that training and testing sets did not share diseases with a common high-level ancestor. This was done to minimize bias from known diseases. Interestingly, this stratification had a marginal impact on performance, with the disease-stratified setup registering a 1% *increase* in absolute terms. Given these unexpected findings, we opted for the baseline split over the disease-stratified approach in all configurations, especially considering our consolidation by disease name.

Shapley (SHapley Additive exPlanations) values are used for summarizing feature importance to the trained model [51]. SHAP values are a popular method for interpreting feature importance, both globally, for specific data partitions or explaining individual predictions. It is used to show the relative

contribution of each feature to the model's output, also taking into account the contribution of other features to the model.

4.3. Generation of New Features

Features were extracted from OT direct evidence data sources. These included indicators of disease associated phenotypes and genomic transcript targets, and evidence scores per association. For example, the feature “Genetic association” refers to the evidence score from genetic association sources, “Literature” to literature evidence, “animal models” to the amount of evidence for a disease based on animal studies, and so on. For computational reasons, the genetic associations and phenotype sources were filtered to keep only those appearing at least 30 times in the dataset, and these were used as features, including their evidence scores. A novel approach [19], [36] of compressing these sparse features using a truncated singular matrix decomposition representation of ~512 dimensions worked well in terms of performance and compute (not shown), but reduced interpretability, and thus was not used in the final model analysis. “Phenotype counts” is the number of distinct phenotypes associated with a disease (regardless of individual phenotype frequency or evidence score). “Max Phenotype frequency” is the overall frequency of the associated phenotype with the highest frequency in the data.

Engineered and aggregated features were extracted from the sources, including aggregated statistics (e.g., value mean, max, min, standard deviation, number of unique values, count of total occurrences)[19], [52]. The relative ratios of each evidence source type in relation to others was also extracted, e.g., the fraction of total evidence for a disease based on each type of evidence-source, and if a specific source was the largest or smallest ranked source (e.g., the feature “Literature ratio to biggest” is the amount of evidence from the Literature divided by the largest evidence source for the disease, which can also be the literature). For each disease we extracted the number of evidence counts per disease, per data source and data type, as well as additional features from the “disease” data including the total number of therapeutic areas, the existence of synonyms for a disease term, the number of direct parents, siblings (sharing the same direct parent) and ancestors for a disease in the OT graph, as well as the difference and ratio between the 2 features: “Ancestors sub parents” - the difference between the total number of ancestors and the number of direct parents for a disease. “Average associated evidence score” is the average confidence score of all evidences associated with a disease from all sources.

Overall disease population prevalence is estimated using the UK Biobank (UKB) [40]. The UKB contains demographic, lifestyle and medical information for 500,000 UK citizens. We matched 8,445 diseases

to 663 ICD-10 medical diagnosis codes in the UKB, using Data-Field 41202 - "ICD10 diagnoses". We crossmatched this with the overall frequency of these codes in the population as a feature used by the models. This feature did not contribute to model performance, and mainly served to help disprove whether diseases' overall population prevalence rate might be a strong, potentially confounding feature [53].

4.4. Deep Learning Text Features

Using the state of the art techniques introduced in recent works combining tabular and pre-trained language models [19], [35], [54], [55], we used deep learning large language model pretrained on biomedical concepts, BioLORD-STAMB2-v1[56], to derive embedding features for each disease using its name and description. In brief, a pretrained neural network language model, trained to predict masked words in a text is taken, and the outputs from its final output layer is extracted and averaged across each token in the text. This mean-pooled output is used for features. Thus, texts are embedded in a vector space such that semantically similar text is close. We tried additional sentence-transformer language models, including all-MiniLM-L12-v2 [64], BioLord-2023 [65], BGE-en-base, and GTE-en-base [67], but their performance was slightly inferior (87~88 AUC, not shown).

This approach lets us combine the benefits of large language models and deep representations in a simple, scalable way with our own features, while reducing possible name bias (e.g., diseases called "syndrome 1" - which could result in overfitting from a token-level finetuned language model) [55], [57]. These features are denoted as "Text Embedding X'' in Fig 3, where X represents a vector in the embedding. For interpretability, we implemented an automated explanation framework showing exemplars of high, neutral and low values per embedding dimensions (Table 3), inspired by approaches in automated-machine learning works[35], [36], [52]. It is available in our codebase.

5. Data Availability

Datasets used in the study are available on Open Targets: https://www.opentargets.org.

Code and results available online: https://github.com/ddofer/OpenTargets-DiseaseSubtype

List of Abbreviations

CV: Cross validation. EFO: Experimental factor ontology. GxE: Gene x environment. GO: Gene ontology. GWAS: Genome-wide association studies. HPO: Human phenotype ontology. LR: Logistic regression.

OBA: Ontology of biological attributes. OT: Open Targets. PRAUC: Area under the precision-recall curve. ROCAUC: Receiver operating characteristic area under the curve. RF: Random Forest. UKB: United Kingdom BioBank. SHAP: Shapley additive explanations.

Supplementary Data

Figure S1 -"Figure-S1-ConfusionMatrix _CatboostModelCV.png" - Confusion matrix figure (From the Catboost known subtype model, cross validation output)

Table S1 - "S1-Known disease subtype models evaluation.csv" - Evaluation metrics and multiple model results on known subtypes prediction

Table S2 - "S2-candidate_errors_predictions.csv" - Novel candidate predictions, including ground truth and novel predictions and features for the 1531 cases where predictions differ from ground truth consistently.

Table S3 - "S3-Literature_esearch_KnownSubtypes-800_esearch_res.csv" - Literature search results for 800 highest confidence model predictions (known candidates model)

Table S4 - "S4-Literature_esearch_candidate_error_300_res.csv" - Literature search results for top predicted novel candidates

Conflict of Interest

The authors have declared no conflict of interest.

Funding

This research was partially supported by ISF grant 2753/20 (M.L), the Milgrom family foundation grant 3015004508 (M.L.).

6. References

[1] S. H. Lee *et al.*, "Parkinson's Disease Subtyping Using Clinical Features and Biomarkers: Literature Review and Preliminary Study of Subtype Clustering," *Diagnostics*, vol. 12, no. 1, p. 112, Jan. 2022, doi: 10.3390/diagnostics12010112.

[2] K. Rannikmäe *et al.*, "Developing automated methods for disease subtyping in UK Biobank: an exemplar study on stroke," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 191, Jun. 2021, doi: 10.1186/s12911-021-01556-0.

[3] S. Saria and A. Goldenberg, "Subtyping: What It is and Its Role in Precision Medicine," *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 70–75, Jul. 2015, doi: 10.1109/MIS.2015.60.

[4] World Health Organization, "ICD-10 : international statistical classification of diseases and related health problems : tenth revision," World Health Organization, 2004. Accessed: Aug. 21, 2023. [Online]. Available: https://apps.who.int/iris/handle/10665/42980

[5] M. S. Udler *et al.*, "Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis," *PLOS Med.*, vol. 15, no. 9, p. e1002654, Sep. 2018, doi: 10.1371/journal.pmed.1002654.

[6] Y. Barak-Corren *et al.*, "The value of parental medical records for the prediction of diabetes and cardiovascular disease: a novel method for generating and incorporating family histories," *J. Am. Med. Inform. Assoc.*, p. ocad154, Aug. 2023, doi: 10.1093/jamia/ocad154.

[7] D. Ochoa *et al.*, "The next-generation Open Targets Platform: reimagined, redesigned, rebuilt," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1353–D1359, Jan. 2023, doi: 10.1093/nar/gkac1046.

[8] N. Kaplan and M. Linial, "Automatic detection of false annotations via binary property clustering," *BMC Bioinformatics*, vol. 6, no. 1, p. 46, Mar. 2005, doi: 10.1186/1471-2105-6-46.

[9] I. D. Dinov, "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data," *GigaScience*, vol. 5, no. 1, p. 12, Feb. 2016, doi: 10.1186/s13742-016-0117-6.

[10] J. Gillis and P. Pavlidis, "Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA).," *BMC Bioinformatics*, vol. 14 Suppl 3, p. S15, Jan. 2013.

[11] M. Linial, "How incorrect annotations evolve – the case of short ORFs," *Trends Biotechnol.*, vol. 21, no. 7, pp. 298–300, Jul. 2003, doi: 10.1016/S0167-7799(03)00139-2.

[12] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[13] N. Zhou *et al.*, "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens," *Genome Biol.*, vol. 20, no. 1, p. 244, Nov. 2019, doi: 10.1186/s13059-019-1835-8.

[14] National Research Council (US) Committee on A Framework for Developing a NewTaxonomy of Disease, *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. in The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC): National Academies Press (US), 2011. Accessed: Aug. 21, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK91503/

[15] X. Wu *et al.*, "Somatic mutations of CADM1 in aldosterone-producing adenomas and gap

junction-dependent regulation of aldosterone production," *Nat. Genet.*, vol. 55, no. 6, pp. 1009–1021, Jun. 2023, doi: 10.1038/s41588-023-01403-0.

[16] I. M. Adcock, G. Caramori, and K. F. Chung, "New targets for drug development in asthma," *The Lancet*, vol. 372, no. 9643, pp. 1073–1087, Sep. 2008, doi: 10.1016/S0140-6736(08)61449-X.

[17] M. E. Johansson, N. M. van Lier, R. P. C. Kessels, B. R. Bloem, and R. C. Helmich, "Two-year clinical progression in focal and diffuse subtypes of Parkinson's disease," *Npj Park. Dis.*, vol. 9, no. 1, Art. no. 1, Feb. 2023, doi: 10.1038/s41531-023-00466-4.

[18] A. Espay and B. Stecher, Eds., "Disease Subtypes: The Promise and the Fallacy," in *Brain Fables: The Hidden History of Neurodegenerative Diseases and a Blueprint to Conquer Them*, Cambridge: Cambridge University Press, 2020, pp. 33–40. doi: 10.1017/9781108888202.006.

[19] S. Cohen, N. Dagan, N. Cohen-Inger, D. Ofer, and L. Rokach, "ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models," *IEEE Access*, vol. 9, pp. 91584–91592, 2021, doi: 10.1109/ACCESS.2021.3091622.

[20] W. Ceusters, B. Smith, A. Kumar, and C. Dhaen, "Mistakes in medical ontologies: where do they come from and how can they be detected?," *Stud. Health Technol. Inform.*, vol. 102, pp. 145–163, 2004.

[21] P. Gaudet and C. Dessimoz, "Gene Ontology: Pitfalls, Biases, and Remedies," in *The Gene Ontology Handbook*, C. Dessimoz and N. Škunca, Eds., in Methods in Molecular Biology. , New York, NY: Springer, 2017, pp. 189–205. doi: 10.1007/978-1-4939-3743-1_14.

[22] I. C. Hageman, I. A. L. M. van Rooij, I. de Blaauw, M. Trajanovska, and S. K. King, "A systematic overview of rare disease patient registries: challenges in design, quality management, and maintenance," *Orphanet J. Rare Dis.*, vol. 18, no. 1, p. 106, May 2023, doi: 10.1186/s13023-023-02719-0.

[23] M.-A. Schulz, M. Chapman-Rounds, M. Verma, D. Bzdok, and K. Georgatzis, "Inferring disease subtypes from clusters in explanation space," *Sci. Rep.*, vol. 10, no. 1, p. 12900, Jul. 2020, doi: 10.1038/s41598-020-68858-7.

[24] Y. Han, K. Klinger, D. K. Rajpal, C. Zhu, and E. Teeple, "Empowering the discovery of novel target-disease associations via machine learning approaches in the open targets platform," *BMC Bioinformatics*, vol. 23, no. 1, p. 232, Jun. 2022, doi: 10.1186/s12859-022-04753-4.

[25] D. R. Swanson, "Migraine and Magnesium: Eleven Neglected Connections," *Perspect. Biol. Med.*, vol. 31, no. 4, pp. 526–557, 1988, doi: 10.1353/pbm.1988.0009.

[26] S. Cheerkoot-Jalim and K. K. Khedo, "Literature-based discovery approaches for evidence-based healthcare: a systematic review," *Health Technol.*, vol. 11, no. 6, pp. 1205–1217, Nov. 2021, doi: 10.1007/s12553-021-00605-y.

[27] S. Bonner *et al.*, "Understanding the performance of knowledge graph embeddings in drug discovery," *Artif. Intell. Life Sci.*, vol. 2, p. 100036, Dec. 2022, doi: 10.1016/j.ailsci.2022.100036.

[28] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Sci. Data*, vol. 10, no. 1, p. 67, Feb. 2023, doi: 10.1038/s41597-023-01960-3.

[29] C. Ma, Z. Zhou, H. Liu, and D. Koslicki, "KGML-xDTD: a knowledge graph–based machine learning framework for drug treatment prediction and mechanism description," *GigaScience*, vol. 12, p. giad057, Dec. 2022, doi: 10.1093/gigascience/giad057.

[30] G. Koscielny *et al.*, "Open Targets: a platform for therapeutic target identification and validation," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D985–D994, Jan. 2017, doi: 10.1093/nar/gkw1055.

[31] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Feb. 2011.

[32] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features." arXiv, Jan. 20, 2019. doi: 10.48550/arXiv.1706.09516.

[33] L. (University of C. Breiman, *Random forest*, vol. 45. 1999.

[34] S. Karsenty, N. Rappoport, D. Ofer, A. Zair, and M. Linial, "NeuroPID: a classifier of neuropeptide precursors," *Nucleic Acids Res.*, pp. gku363-, May 2014, doi: 10.1093/nar/gku363.

[35] D. Ofer and D. Shahaf, "Cards Against AI: Predicting Humor in a Fill-in-the-blank Party Game," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5397–5403. doi: 10.18653/v1/2022.findings-emnlp.394.

[36] D. Ofer and M. Linial, "Inferring microRNA regulation: A proteome perspective," *Front. Mol. Biosci.*, vol. 9, 2022, Accessed: Oct. 15, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmolb.2022.916639

[37] F. Novais, A. Araújo, and P. Godinho, "Historical roots of histrionic personality disorder," *Front. Psychol.*, vol. 6, p. 1463, Sep. 2015, doi: 10.3389/fpsyg.2015.01463.

[38] D. Ofer, N. Rappoport, and M. Linial, "The Little Known Universe of Short Proteins in Insects: A Machine Learning Approach," in *Short Views on Insect Genomics and Proteomics*, 2015, pp. 177–202. doi: 10.1007/978-3-319-24235-4_8.

[39] M. Linial, N. Rappoport, and D. Ofer, "Overlooked short toxin-like proteins: A shortcut to drug design," *Toxins*, vol. 9, no. 11, 2017, doi: 10.3390/toxins9110350.

[40] C. Sudlow *et al.*, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLOS Med.*, vol. 12, no. 3, p. e1001779,

Mar. 2015, doi: 10.1371/journal.pmed.1001779.

[41] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain J. Neurol.*, vol. 130, no. Pt 2, pp. 314–333, Feb. 2007, doi: 10.1093/brain/awl241.

[42] B. Bowe, Y. Xie, and Z. Al-Aly, "Postacute sequelae of COVID-19 at 2 years," *Nat. Med.*, pp. 1–11, Aug. 2023, doi: 10.1038/s41591-023-02521-2.

[43] R. Rasnic, D. Klinger, D. Ofer, Y. Comay, M. Linial, and E. Bachmat, "Reduced Mortality During Holidays and the COVID-19 Pandemic in Israel." medRxiv, p. 2020.07.16.20155259, Jul. 27, 2020. doi: 10.1101/2020.07.16.20155259.

[44] "Definition of DYSPLASIA." Accessed: Aug. 28, 2023. [Online]. Available: https://www.merriam-webster.com/dictionary/dysplasia

[45] S. A. Bien and U. Peters, "Moving from one to many: insights from the growing list of pleiotropic cancer risk genes," *Br. J. Cancer*, vol. 120, no. 12, Art. no. 12, Jun. 2019, doi: 10.1038/s41416-019-0475-9.

[46] P. H. Lee, Y.-C. A. Feng, and J. W. Smoller, "Pleiotropy and Cross-Disorder Genetics Among Psychiatric Disorders," *Biol. Psychiatry*, vol. 89, no. 1, pp. 20–31, Jan. 2021, doi: 10.1016/j.biopsych.2020.09.026.

[47] A. Dahl and N. Zaitlen, "Genetic Influences on Disease Subtypes," *Annu. Rev. Genomics Hum. Genet.*, vol. 21, no. 1, pp. 413–435, 2020, doi: 10.1146/annurev-genom-120319-095026.

[48] J. A. Evans, J.-M. Shim, and J. P. A. Ioannidis, "Attention to Local Health Burden and the Global Disparity of Health Research," *PLOS ONE*, vol. 9, no. 4, p. e90147, Apr. 2014, doi: 10.1371/journal.pone.0090147.

[49] A. Yegros-Yegros, W. van de Klippe, M. F. Abad-Garcia, and I. Rafols, "Exploring why global health needs are unmet by research efforts: the potential influences of geography, industry and publication incentives," *Health Res. Policy Syst.*, vol. 18, no. 1, p. 47, May 2020, doi: 10.1186/s12961-020-00560-6.

[50] A. Boutayeb, "Developing countries and neglected diseases: challenges and perspectives," *Int. J. Equity Health*, vol. 6, p. 20, Nov. 2007, doi: 10.1186/1475-9276-6-20.

[51] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 21, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[52] M. Maor, R. Karidi, S. Davidovich, and A. Ronen, "System and method for feature generation

over arbitrary objects," US20170017900A1, Jan. 19, 2017 Accessed: Mar. 20, 2023. [Online]. Available: https://patents.google.com/patent/US20170017900A1/en

[53] I. Blass, T. Sahar, A. Shraibman, D. Ofer, N. Rappoport, and M. Linial, "Revisiting the Risk Factors for Endometriosis: A Machine Learning Approach," *J. Pers. Med.*, vol. 12, no. 7, p. 1114, Jul. 2022, doi: 10.3390/jpm12071114.

[54] D. Ofer and M. Linial, "Whats next? Forecasting scientific research trends." arXiv, Jul. 09, 2023. doi: 10.48550/arXiv.2305.04133.

[55] D. Ofer, N. Brandes, and M. Linial, "The language of proteins: NLP, machine learning & protein sequences," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1750–1758, 2021, doi: 10.1016/j.csbj.2021.03.022.

[56] F. Remy, K. Demuynck, and T. Demeester, "BioLORD: Learning Ontological Representations from Definitions (for Biomedical Concepts and their Textual Descriptions)." arXiv, Oct. 21, 2022. doi: 10.48550/arXiv.2210.11892.

[57] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, Apr. 2022, doi: 10.1093/bioinformatics/btac020.

[58] Nicole A Vasilevsky, et al. (2022) Mondo: Unifying diseases for the world, by the world medRxiv 2022.04.13.22273750; doi: https://doi.org/10.1101/2022.04.13.22273750

[59] Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H: Modeling Sample Variables with an Experimental Factor Ontology. Bioinformatics 2010, 26(8):1112-1118

[60] Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net

[61] Köhler, et al., The Human Phenotype Ontology in 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1207–D1217, https://doi.org/10.1093/nar/gkaa1043

[62] Stefancsik, et al. (2023). The Ontology of Biological Attributes (OBA)-computational traits for the life sciences. Mammalian genome : official journal of the International Mammalian Genome Society, 34(3), 364–378. https://doi.org/10.1007/s00335-023-09992-1

[63] FDA website: https://www.fda.gov/vaccines-blood-biologics/vaccines/smallpox. Smallpox. U.S. Food and Drug Administration website. Accessed 22-01-2024

[64] Reimers, N., & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing. 2019

[65] Remy, François et al. "BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights." ArXiv abs/2311.16075 (2023): n. Pag.

[66] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). Towards General Text Embeddings with Multi-stage Contrastive Learning. ArXiv, abs/2308.03281.

[67] Mendez MF. Early-onset Alzheimer Disease and Its Variants. Continuum (Minneap Minn). 2019 Feb;25(1):34-51. doi: 10.1212/CON.0000000000000687. PMID: 30707186; PMCID: PMC6538053.

[68] Ferguson LW, Rajput AH, Rajput A. Early-onset vs. Late-onset Parkinson's disease: A Clinical-pathological Study. Can J Neurol Sci. 2016 Jan;43(1):113-9. doi: 10.1017/cjn.2015.244. Epub 2015 Jul 20. PMID: 26189779.