Automated Annotation of Disease Subtypes

Dan Ofer, Michal Linial

Keywords: machine learning; Open Targets; Medicine; ontology completion; link prediction; Orphannet; explainability; AI; personalized medicine; EFO; Disease ontology.

Abstract

Background

Distinguishing diseases into distinct subtypes is crucial for study, effective treatment, and the discovery of potential cures. The Open Targets Platform integrates biomedical, genetic, and biochemical datasets with the goal of empowering disease ontologies and gene targets. However, many disease annotations remain incomplete, necessitating laborious expert medical input. This is particularly painful for rare and orphan diseases, where resources are limited.

Results

We present a machine learning approach to identifying diseases with potential subtypes, using the approximately 23,000 diseases documented in Open Targets. We derive and describe 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 known subtypes. Our model achieves a high (89.1%) ROCAUC. We integrated pre-trained deep learning language models and showed their benefits. Furthermore, 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.

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].

Historically, the medical community has spearheaded efforts to identify the multifaceted nature of diseases. Clinicians primarily rely on the International Classification of Diseases (ICD)

International Statistical Classification of Diseases and Related Health Problems (ICD), which undergoes periodic revisions (e.g., ICD-10 is the 10th revision)[4]. For instance, Diabetes mellitus (ICD-10, E10-E14) is partitioned into Type 1 (E10), Type 2 (E11), 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-base 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 function across species are mapped to coding genes [12]. The impact of the GO project and its like on automatic functional annotation tasks such as CAFA is unquestionable [10], [13].

As personalized medicine advances, disease subtyping has the potential to 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 treatment regimes [17]–[19].

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. The study “Inferring disease subtypes from clusters in explanation space” concentrated on a narrower domain, clustering specific cancer data, imaging, and non-biomedical data [23]. It did not endeavour to offer predictions across a broad spectrum of known diseases. Another study, "Empowering the discovery of novel target-disease associations via machine learning

approaches in the open targets platform", used OT data to train models to identify new drug-disease associations—a different objective from ours [24]. Our work also relates to knowledge base centric link prediction, and literature based discovery [25]–[29]. However, these primarily 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 the Open Targets Platform. 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 future approaches for characterizing disease subtypes. Furthermore, we present an approach to identifying and evaluating candidate diseases with potential novel subtypes and mis-annotations, and a ranked list of predictions. Our automated approach is highly 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. 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 autonomously.

Results

Diverse Disease Ontologies

A machine learning model for disease subtyping assessment and discovery was developed using the sources integrated into the Open Targets Platform. An overview of these sources is shown 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 and symptoms, and a bibliography. **Fig. 1B** lists the associated genes for type 2 diabetes mellitus ranked by the 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 - Open Targets disease information

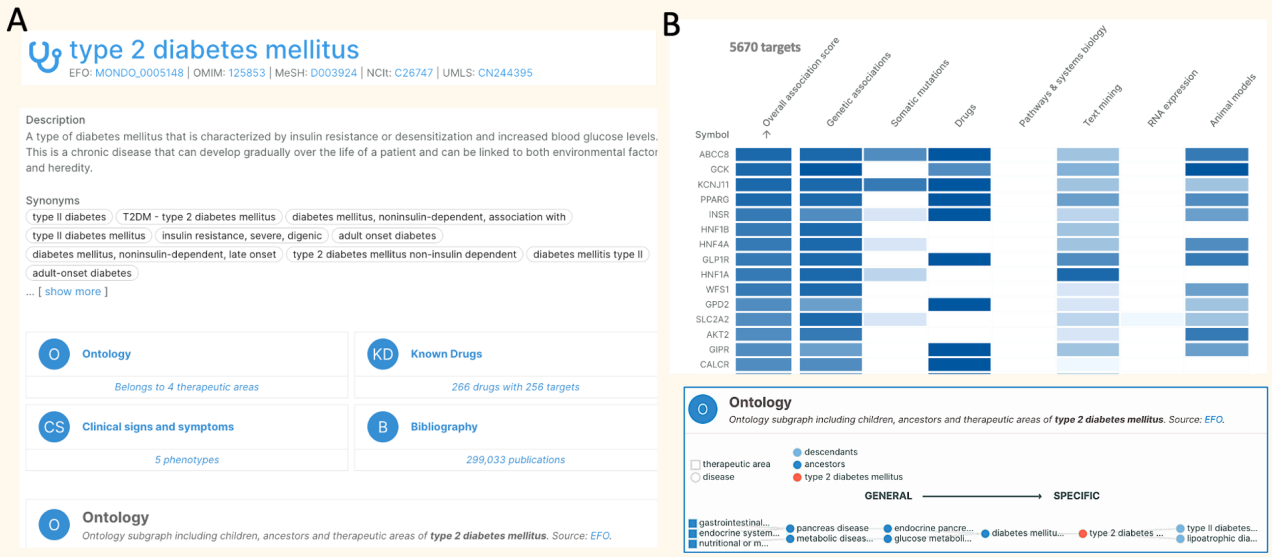


Fig. 1. *Example from Open Targets for type 2 diabetes.* ***(A)*** *Description, disease ontology (including subtypes if any), synonymous nomenclature, known drugs, bibliography, and clinical symptoms* ***(B)*** *Weighted evidence sources and domains for the disease, including genetics, somatic mutation, drugs, text mining, and more.*

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.

Performance evaluation - known targets

For the task of predicting known disease subtypes, as a binary classification task, we tested multiple machine learning models - logistic regression (LR), random forest (RF) and CatBoost (a boosting tree model)—and performed five-fold stratified cross validation to evaluate performance in predicting known disease subtypes[31]–[33].

For comparative analysis, we added to the assessment three baseline models. 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, wherein all “siblings'' share the same parent disease. All baselines outperform random guessing. All models substantially outperform the baselines. Catboost had

the best performance, achieving a ROCAUC (Receiver Operating Characteristic Area Under the Curve) of 90.5% (**Table 1**). Accordingly, we used Catboost for subsequent predictions of novel subtypes and analyses.

Fig. 2 - Known subtypes model comparison

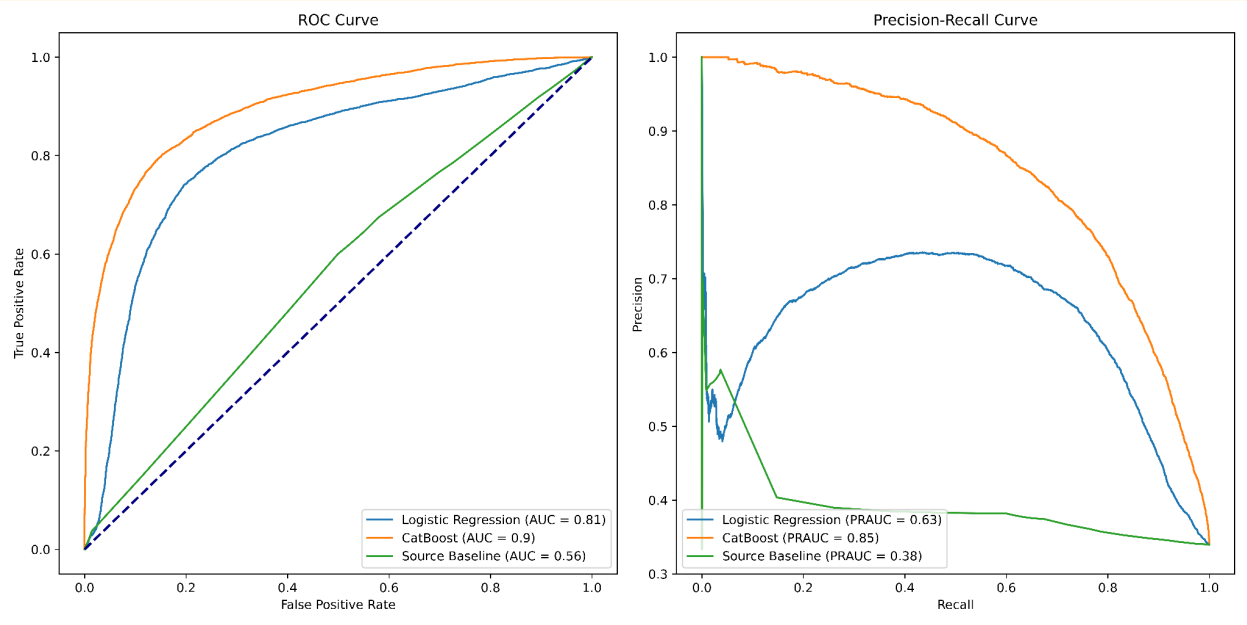


Table 1 - Known disease subtype model evaluation

	ROC-AUC Mean, s.d.	PR-AUC Mean, s.d.
Baseline by Source	55.86 ± 0.55	37.67 ± 0.26
Baseline by # of Phenotypes	66.42 ± 0.42	44.6 ± 0.34
Baseline by # of OT siblings	52.35 ± 0.52	35.28 ± 0.32
No Text model	81.02 ± 0.58	63.6 ± 1.13
Logistic Regression	85.87 ± 0.53	76.4 ± 0.44
Random Forest	87.28 ± 0.46	80.55 ± 0.98
CatBoost	**89.82 ± 0.45**	**84.81 ± 0.76**

Results shown for 5 fold cross validation on test set splits, ± standard deviation. PRAUC: Area Under Precision-Recall Curve. No Text model - Catboost model with text embedding features excluded (See methods: "Deep learning using Text features"). The remaining 3 models (LR, RF and Catboost) used all available features.

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 semi-naive bootstrapping approach to identifying potential novel candidates. We report on cases where a model, retrained over multiple random data splits (using 8x5 RepeatedKFold cross-validation), consistently(eight out of eight times) predicts a different label than the recorded one per data point. Note that the data point is part of the held out test set. Altogether, 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. The average prediction consistency was 84.9%. This bootstrapping approach is effectively also an ensemble, with models trained on each split. In supplementary S1 we record the averaged model predictions on the held out test set splits. We found that this ensemble has better results than a single model, as might be expected (ROC-AUC: 91.08, PR-AUC: 86.61) [19], [34]. Thus, we used these as candidate predictions. The full list of candidate predictions is in Supplementary file S1.

Evaluation of Predictions for Unknown Targets

Since there is no gold standard to evaluate such novel predictions for subtype expansion, we utilized the scientific literature as an external validation mechanism. We presume that potential subtypes might be discussed in the wider literature prior to any validation and integration into existing knowledge bases. From our models, we selected 800 predicted targets with the highest predicted probabilities per-class, evenly split with 400 cases per class. We searched PubMed for literature hits of these candidate diseases being mentioned as having a subtype, subtype, or pleiotropy. We extracted the fraction of such cases relative to the total number of literature references for each disease. The difference between candidates categorized by their predicted subtype was statistically significant as determined by a one-sided unequal variance t-test ($p = 2.6e-7$) (S2), with predicted candidates having ~5 times as many results with subtypes (2,100 vs. 400).

We conducted a similar analysis on 300 of the stable predicted candidates (S1), specifically those whose predictions deviated from existing annotations. For this subset, the literature

search revealed no significant difference (p=0.27) (S3). 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 the OT dataset, this would be evident in 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 inadequately or faulty annotations reported in OT.

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 our model, we examined the distribution of disease subtypes in our predictions against the known subtypes, focusing on their source database (**Table 3**). 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 subtype 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 3 - Source subtype distribution

Source	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

Table 3 *- Distribution of diseases with known vs predicted subtype, grouped by 6 largest database sources.* D*iseases' subtype fraction are shown per source. "Original Source Subtype Fraction" - % of diseases with a subtype in existing annotations. "Predictions Source Subtype Fraction" - % of diseases with predicted subtype in novel predictions."Total Source Diseases" - the total number of diseases (regardless of subtype) from a source.*

Discussion

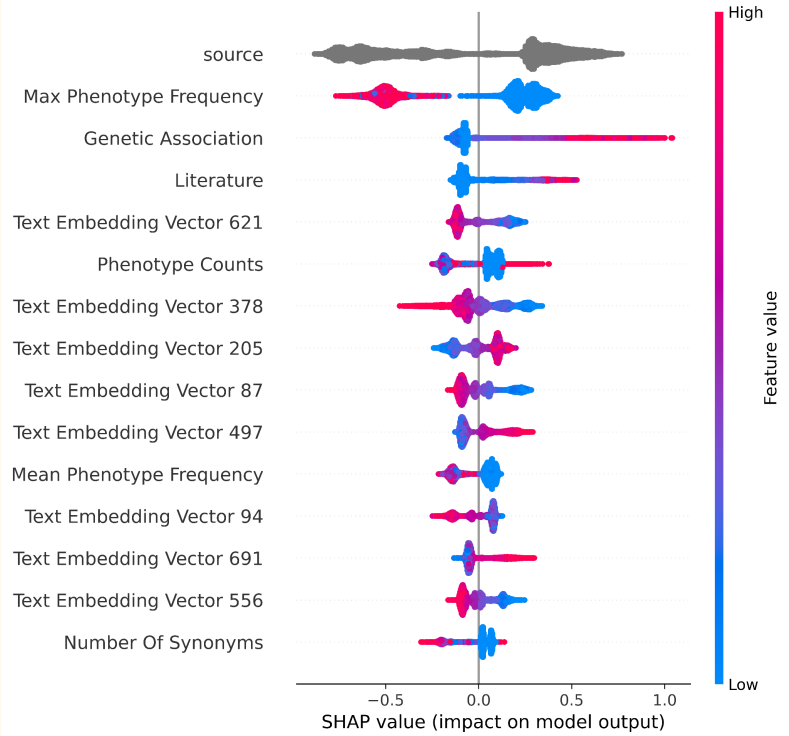
Discovering disease subtypes is an important problem in medicine, with applications in both basic research and personalized treatment. That a disease may even potentially have subtypes is not an obvious fact. Historically, diagnoses like "female hysteria" led to ineffective treatments, overshadowing 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 is another major example, with considerable work going into finding subtypes that can help treat patients and predict the future course of the disease's progression [1], [17].

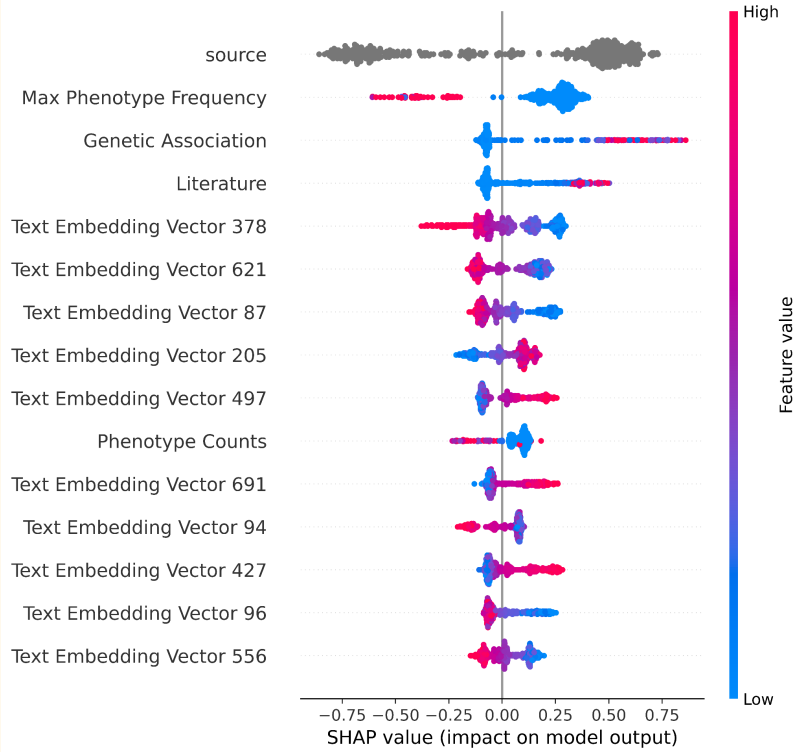
Our hypothesis posited that diseases with distinct subtypes might be discernible based on intrinsic aspects of the disease itself as well as meta-features relating to its research. Consider how diseases driven by somatic mutations, as in most cancers, differ from those stemming from genetic variants. Single-gene disorders are often inherited but they may split into early and late onset diseases that might be addressed differently in clinical terms. differently and or environmental and lifestyle factors, like adult onset Type 2 diabetes. 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, extensive literature and many candidate drugs, are less likely to be categorized as "orphans" than those observable only in humans. Furthermore, some phenotypes are easier to observe, measure and categorize, while others may be more nuanced, such as facial deformities as opposed to 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 [40] population frequency. Another metric of rarity is the classification of a disease as an "orphan disease", which is determined by its source, such as whether the disease is listed in Orphanet. The biobank prevalence feature has low feature importance (Fig. 3 - Feature importance). Furthermore, its removal does not affect model evaluation results (not shown), further emphasizing its unimportance. 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 (Table 1 - “Source Baseline”), and observed it to be significantly inferior, yielding a ROCAUC score of just 55.1, indicating that source is also not sufficient to explain the model’s performance. The same findings held for our other evaluated baselines.

Fig 3 - Feature importance over data subsets

A.



B. 

Shapley feature importance in disease subtype prediction, over different data partitions. A. Feature importance over all data. B. Feature importance for subset of candidates with predicted novel subtypes. See Methods for feature dictionary.

Understanding Potential Novel Disease Subtypes

Following our model's identification of 515 diseases predicted to have subtypes not currently annotated in Open Targets, 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. While all top-ranked predictions looked correct in manual review, we provide a number of reasons for why diseases may have subtypes, on examples from the full list (S1).

Pleiotropic Manifestations: Different causes can result in seemingly similar outcomes, leading to diseases with varied presentations. Examples include:

- Female breast carcinoma: Molecular subtypes, defined by specific gene expression profiles, as well as anatomical regions, dictate prognosis and therapeutic responsiveness.

- Nephropathy: This term encompasses kidney pathologies with distinct histological features and clinical courses.

- Cardiovascular risk and Diabetes - both of these disease families have different risk factors, including environmental, genetic and genetic-environment interaction (GxE). For example, Type 1 vs Type 2 diabetes, or specific genetic risk factors [6], [15].

Challenges in Disease Diagnosis: Overlapping clinical presentations can cause misdiagnosis.

- Structural epilepsy: Different brain anomalies can precipitate varied forms of epileptic seizures. Subtypes may progress and respond differently to treatment [41].

- Neurodegeneration: Disorders like Alzheimer's and Parkinson's, though clinically overlapping, have distinct molecular pathologies, and each has distinct subtypes. Understanding these subtypes is a major goal towards effective treatment [1].

Variability in Disease Course and Treatment: Clinical trajectory and therapeutic responsiveness can vary based on disease subtypes and interaction with patient characteristics.

- For a variety of malignant and non malignant tumors (e.g. Gastrointestinal stromal tumor, benign neoplasms of the pituitary gland, glomus, colon): The behavior, complications and potential malignancy and treatment is influenced by specific mutations, patient genetics and risk factors (GxE), size etc'.

- Covid-19: The spectrum of clinical presentations, from asymptomatic states to severe respiratory syndrome, and long term impairment ("long covid") suggests potential subtypes [42], [43].

Broad Disease Categories with Inherent Variability: Some terms inherently suggest a range of conditions. In some cases these candidates are "trivial", and are due to a lack of linkage in or between the ontologies of known terms.

- Eye infections: Etiological agents span bacteria, viruses, fungi, and parasites.

- Speech disorder, central nervous system development, genetic variation, infection - are all semantically high level categories.

- Dysplasia, a family of conditions encompassing "various types of abnormal growth or development of cells, organs, and the abnormalities resulting from such growth"[44], is a clear case of a high level term that has subtypes, such as microscopic (cell) level, organ (macroscopic) level, and further delineation.

- Alcohol dependence, speech disorder, sexual and gender identity disorders - are high level descriptors, that lack breakdown to a clinically “useful” level, required for effective treatment.

Novel disease subtypes partitioning represents a challenging problem for both clinical and scientometric researchers. Despite the high quality medical ontologies already integrated into OT, 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 Open Targets 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 and to publish on distinct aspects may be more likely to be distincted, 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 (S1).

Potential implications

The methodologies in our study, including the combination of machine learning with the Open Targets platform, could extend to broader works. In clinical diagnostics; the increased identification and annotation of disease subtypes could provide sharper diagnoses, disease

progression tracking and to pave the way for 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, drug developers can target therapies and druggables with greater precision.

In conclusion, annotating disease subtypes is crucial for enhancing our understanding of pathology and refining therapeutic strategies. By delineating diseases into finer subtypes, we pave the way for targeted research and treatment.

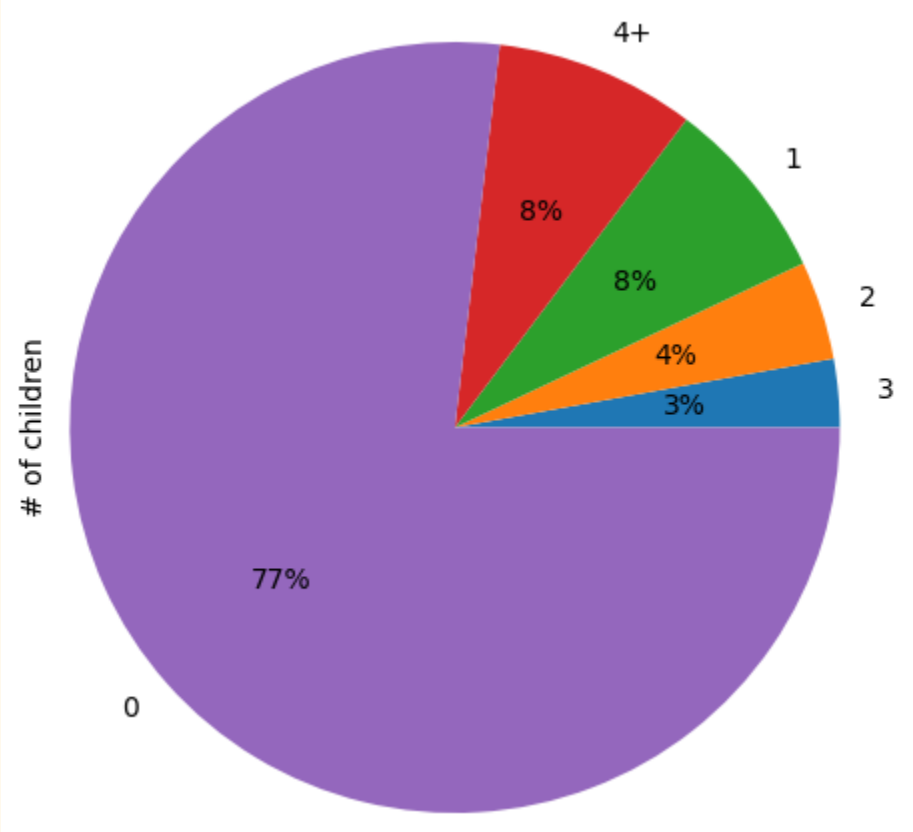
Methods

Overview and processing of Open Targets data

Data was downloaded from Open Targets Platform, as of July 2022. 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 4. We kept only direct associations, as we determined that indirect associations may leak target information.

The initial dataset held 23,074 candidate diseases. We removed 2,817 irrelevant diseases relating to lab measurements (e.g., "IgG index" from EFO). 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. The 23,074 disease terms were filtered to remove measurement terms (e.g. "IgG index", "BMI"). 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.

Fig 4 - subtypes per disease in OpenTargets



Model Training and evaluation

For most models, default hyperparameters were employed, except for Catboost (a boosting tree model), for which we adjusted class weights for partial balancing during training using the "Square root Balanced" hyperparameter. The logistic regression and random forest models were implemented using scikit-learn. Features with a variance lower than 5e-4 were dropped.

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 (Fig 3) [51]. Shap values are a widely accepted method for interpreting model importance, both on a global scale and for specific data instances or partitions.

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 "Genetic association" feature 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 datasource 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.

The overall disease prevalence in the population was estimated using additional information from the UK BioBank (UKBB) [40]. The UKBB contains demographic, lifestyle and medical information for 500,000 UK citizens. We matched 8,445 diseases to 663 ICD10 medical diagnosis codes in the UKBB, using Data-Field 41202 - "ICD10 diagnoses". We crossmatched this with the overall frequency of these codes in the population as a feature. This feature did not add to the model's performance, and mainly served to help disprove whether diseases' overall population prevalence rate might be a strong, potentially confounding feature [53].

Deep Learning Text Features

Using the state of the art technique introduced in recent works combining tabular and pre-trained language models [19], [35], [54], [55], we employed a deep learning large language model pretrained on biomedical concepts, BioLORD[56], to derive embedding features for each disease using its name and description. This embedding is a mean-pooled representation of the text and was not fine-tuned for our specific task. We found this significantly enhanced performance when compared to a model devoid of text features (Refer to Table 1: "No Text model"), yielding an AUC of 90, in contrast to 86 without these features.

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 for embedding dimensions, inspired by approaches in other NLP or automated-machine learning works[35], [36], [52]. It is available in our codebase.

Data Availability

Datasets used and analyzed during the current study are available through the Open Targets platform (https://www.opentargets.org/).

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

List of abbreviations

PRAUC: Area under the precision-recall curve. CV: Cross validation. EFO: Experimental Factor Ontology. GxE: Gene x Environment. GO: Gene ontology. GWAS: Genome-wide association studies. HP: Human Phenotype. LR: Logistic regression. OBA: Ontology of Biological Attributes. OT: Open Targets. ROCAUC: Receiver Operating Characteristic Area Under the Curve. RF: Random Forest. UKBB: United Kingdom BioBank. SHAP: Shapley: SHapley Additive exPlanations.

Conflict of Interest

The authors have declared no conflict of interest.

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," Aug. 20, 2023. https://www.merriam-webster.com/dictionary/dysplasia (accessed Aug. 28, 2023).

[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.