Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DITTO: An explainable machine-learning model for transcript- specific variant pathogenicity prediction

Tarun Karthik Kumar Mamidi
University of Alabama-Birmingham School of Medicine
Brandon Michael Wilk
University of Alabama-Birmingham School of Medicine
Manavalan Gajapathy
University of Alabama-Birmingham School of Medicine
Elizabeth Anabel Worthey
eaworthey@uabmc.edu

University of Alabama-Birmingham School of Medicine

Article

Keywords: rare disease, genomics, explainable, machine learning, pathogenic, variant consequence, classification, prioritization

Posted Date: March 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8930088/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DITTO: An explainable machine-learning model for transcript-specific variant pathogenicity prediction

Tarun Karthik Kumar Mamidi[1], Brandon M. Wilk[1], Manavalan Gajapathy[1], Elizabeth A. Worthey[1]

[1]Center for Computational Genomics and Data Science, Department of Genetics, University of Alabama-Birmingham School of Medicine, Birmingham, Alabama 35233, USA

*** Correspondence:**
Elizabeth A. Worthey
eaworthey@uabmc.edu

Abstract

Accurate classification of genetic variants is critical for medical decision-making, providing insights into disease mechanisms, and enabling therapeutic discovery. While numerous methods exist, they often address limited variant types, lack transcript awareness, and operate as opaque black boxes. To address these issues, we developed DITTO - a unified, explainable, transcript-aware advanced machine-learning method for pathogenicity prediction. It integrates diverse genomic features-including conservation scores, population frequencies, etc., to train a single, explainable neural network model. DITTO outperforms existing methods across standard benchmarks, demonstrating superior performance (99% F1 score) in classifying pathogenic and benign variants. DITTO is publicly available at https://github.com/uab-cgds-worthey/DITTO

Keywords: rare disease, genomics, explainable, machine learning, pathogenic, variant consequence, classification, prioritization.

1 Background

Rare diseases, generally defined as those affecting about 1 in 10 individuals, affect at least 400 million people worldwide [1]. More than 8,000 rare diseases have been described, with >80% believed to be genetic in nature [1]. Accurate identification of pathogenic variants, and differentiation from benign ones, is critical for diagnosis, clinical management, family planning, and therapeutic discovery. Advances in high-throughput sequencing have transformed molecular diagnostics, significantly reducing the time and cost of testing [1–5]. Development of computational approaches for variant identification, classification, and interpretation has been critical in supporting these advances [1,6–10].

Despite these advances, major challenges remain. Many patients still face delays or barriers to diagnosis, even in countries like the United States and United Kingdom where genome sequencing is more routinely applied[1,3]. Insurance coverage limitations, geographic disparities, and socioeconomic factors continue to restrict access [11–13]. Among those tested, approximately one-third receive results containing at least one variant of uncertain significance (VUS), which are rarely used to guide clinical decisions and can cause distress due to their ambiguity [14,15][16,17]. This is partially due to biases in repositories such as ClinVar, which, like the broader field, disproportionately focus on protein-coding and splice site regions (~2% of the genome), where variant pathogenicity prediction tools are more developed [18,19].

The volume of VUSs identified is rapidly increasing, adding to a growing backlog of variants requiring reinterpretation [14,15]. Of the more than 2.1 million variants so far deposited in ClinVar; roughly 50% are VUSs and another 5% have conflicting interpretations [15,20]. Manual review is costly and slow, and although projects aimed at generating high-throughput experimental assays are underway [15,21–23], the scale of the task at hand and timelines involved leave many patients in long diagnostic odyssey [24].

In this context, improved computational methods are essential to accelerate variant classification and reduce reliance on labor-intensive manual curation. Conventional ML methods, such as VarSight, MetaLR and MetaSVM, improved impact prediction and classification [25,26]. Neural network-based predictors such as MetaRNN [27], DEOGEN2[28], Alphamissense [29] and EVE [30] have been developed and found to outperform the conventional methods. Many of these advances are being integrated into diagnostic lab pipelines. However, significant limitations exist.

No methods exist that classify across all transcripts, classes of small variants (e.g. SNVs and Indels) and consequences (e.g. splice, missense). This restricts their utility for comprehensive variant analysis across all possible small variants (e.g., SNVs, insertions, indels) and consequences (e.g., intergenic, UTRs, intronic) [31]. In diagnostic labs, this necessitates integration of multiple predictors, which can make pipeline development, maintenance, and validation more complex requiring additional computational expertise [32]. Tools that are not transcript-aware may miss deleterious effects that manifest only in specific isoforms [30,31,33]. This can make it challenging to interpret variants where the impact is seen only in a non-canonical transcript [34,35]. Additionally, many methods operate as "black boxes" providing no explanations for their predictions [2,10,36]. This can limit or bias interpretation and reporting with variants conservatively classified as VUSs [2,10,36]. These limitations undermine the utility of existing tools in diagnostic workflows and represent a barrier in adoption of precision medicine [1,10,37–40].

To address these limitations, we developed DITTO, a transcript aware, explainable machine learning (XML) neural network-based method for variant pathogenicity prediction. DITTO integrates a broad set of genomic features to generate interpretable predictions across all variant classes and consequences considering different transcripts. DITTO provides a unified framework for classification and prioritization, improving transparency and streamlining diagnostic interpretation. To our knowledge, this is the first transcript-aware classifier to achieve high performance across all variant types, consequences, and chromosomes, offering a unified, interpretable solution. Benchmarking demonstrates that DITTO outperforms current state-of-the-art methods in both accuracy and breadth.

2 Results

DITTO is trained on ~150,000 variants from ClinVar and 117 features annotated using OpenCravat. It is an explainable neural network model that gives a pathogenicity probability score [0,1] and feature importances as prediction explanation. More information on data pre-processing can be found in the methods section.

2.1 DITTO tuning and training

DITTO's neural network architecture was determined by adaptively tuning its parameters using a Bayesian approach to minimize the loss function based on the training and validation dataset, ensuring it neither overfits nor underfits the data (Fig 1a). Table 2 summarizes the hyperparameters available for tuning the network architecture. Parameters like Dropout and Batch size were used to tackle potential over/under fitting that results from class imbalance. In order to determine optimal hyperparameters for our model, we conducted two-phase tuning with 500 trials, each utilizing a distinct

combination of hyperparameters. Initially, we employed a random search strategy with 200 trials, assigning random hyperparameter combinations to each.

This approach facilitated exploration of a diverse pool of configurations offering preliminary insights into performance. For each trial, we used the "accuracy score" as the performance metric for the next trial to improve upon. Leveraging the insights gained from the initial 200 trials, we advanced to a more refined tuning strategy using Bayesian optimization via the TPESampler [41] from Hyperopt. This method optimizes the selection process by leveraging outcomes from previous trials, considering both the hyperparameter configurations and corresponding accuracy scores. By adopting the TPESampler proposed hyperparameter combinations for the remaining 300 trials, we optimized the search space to find the highest performant hyperparameter combination.

Table 2: Overview of hyperparameter space for tuning of our DITTO model. Parameters for model tuning spanned number of layers, activation functions, neurons per layer, kernel initializers, dropout rates, optimizers, and batch size.

# layers	Activation function	# neurons per layer	Kernel initializer	Dropout rate	Optimizer	Batch size
1 – 30	tanh	1 - 200	uniform	0.0 – 0.9	SGD	10 - 1000
	softmax		lecun_uniform		RMSprop	
	elu		normal		Adagrad	
	softplus		zero		Adadelta	
	softsign		glorot_normal		Adam	
	relu		glorot_uniform		Adamax	
	sigmoid		he_normal		Nadam	
	hard_sigmoid		he_uniform			
	linear					

To improve accuracy and mitigate over/under fitting, we structured each trial by further splitting the preprocessed training data into training (80%) and validation (20%) subsets. This approach allowed the model to train on a substantial dataset while its performance was evaluated independently on the validation set. Each trial ran for 500 epochs (iterations through the training data).

To prevent stagnation, enhance efficiency, conserve resources, and address potential over/under fitting, an early stopping mechanism was implemented that halted the trial if no improvement is observed in the model's validation accuracy occurred for 10 consecutive epochs. Out of the total 500 trials, 325 successfully completed. 175 were pruned due to poor performance or terminated by the early stopping mechanism (Fig S2). This tuning process yielded the optimal hyperparameter configuration used for DITTO (Fig 1c). The optimal set of hyper parameters was [# layers: 1, activation (input and hidden layers): elu, # neurons: 161, kernel initializer (hidden layer): he_normal, dropout_l0: 0.808, kernel initializer (output layer): glorot_normal, optimizer: Adamax, batch_size: 267]

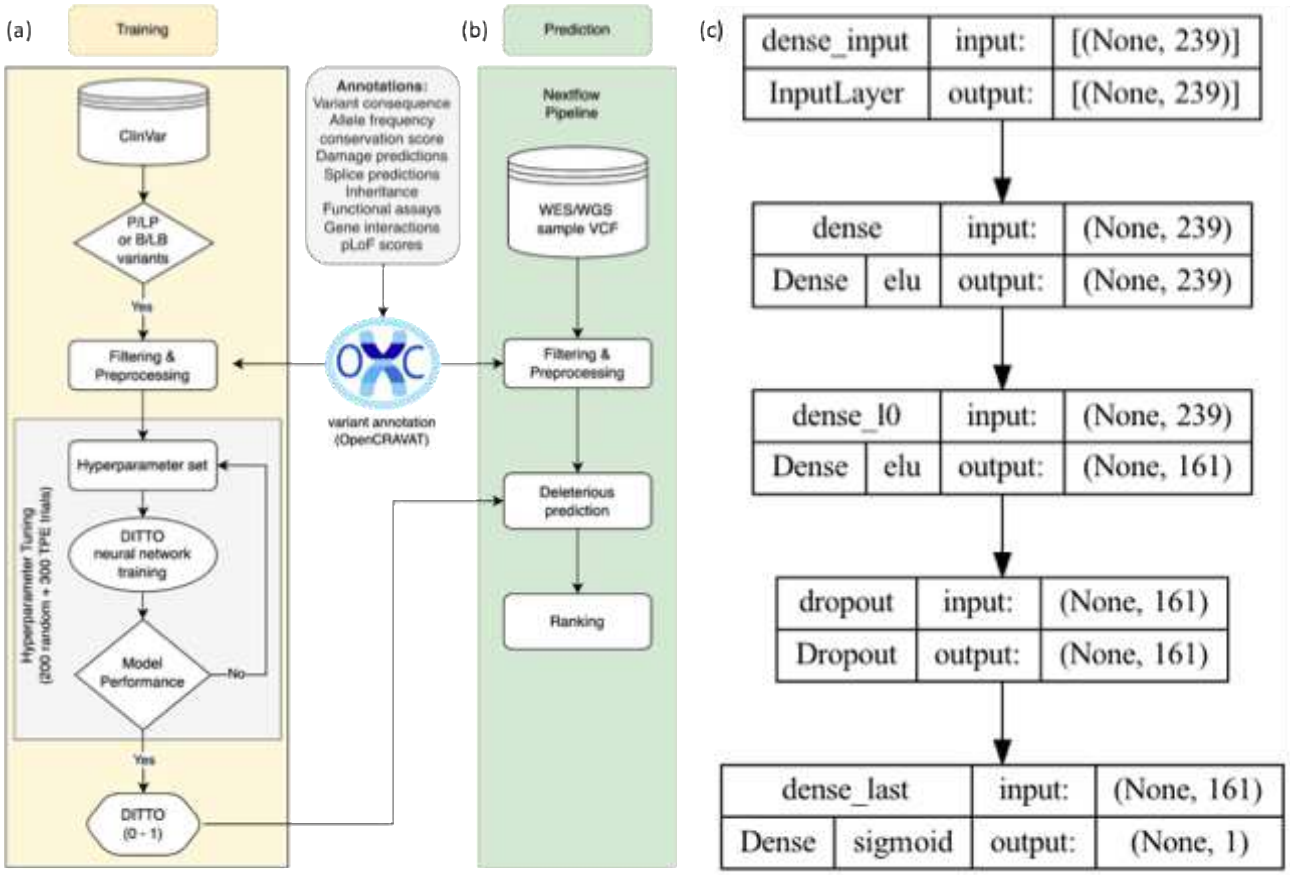


Figure 1. An overview of the training and testing methodology, and the architecture of DITTO model. a) 2,177,684 variants were extracted from ClinVar, filtered, and annotated for 117 distinct features using OpenCRAVAT. Following preprocessing and featuring engineering, we performed a train:validation:test variant split. Training and validation data was used to tune and train a neural network classifier that generates a deleteriousness score (0,1) for variant damage prediction. **b)** We developed a Nextflow pipeline to read VCF files, annotate variants using OpenCRAVAT, and run DITTO. **c)** The DITTO architecture is a single hidden layer neural network with 239 neurons in the input layer, 161 neurons in the hidden layer, and 1 output layer with 96,162 parameters.

The neural network optimized with the best parameters underwent training on the full training dataset with the model saved for validation, testing and benchmarking purposes. To streamline analysis from VCF files, a nextflow [42] pipeline was developed using OpenCRAVAT's annotation service. The pipeline computes DITTO scores using the pre-trained DITTO model and generates deleteriousness predictions for variants alongside relevant annotations (Fig 1b).

2.2 DITTO has high precision, recall, and accuracy in predicting variant deleteriousness

The DITTO neural network achieved extremely high accuracy in distinguishing between deleterious and benign variants, achieving an overall accuracy of 99.9% on the held back validation dataset (Table 3). The model demonstrated exceptional performance across Precision, Recall, and F1-scores (0.99, 0.99, and 0.99, respectively). These results indicate that the model reliably

Table 3. DITTO performance metrics.

Accuracy	F1	ROC AUC	PRC AUC	Recall	Precision
0.99	0.99	0.99	0.98	0.99	0.99

predicts both positive and negative outcomes, shows excellent discriminative ability between positive and negative classes, and performs well even in scenarios with imbalanced data.

2.3 Feature contribution

SHAP plots provide a visual explanation of the importance of specific features to model predictions (Fig 2). The application of SHAP plots can help us understand the features most important for model predictions, and in doing so can also help identify potential biases within the model. Fig 2b shows the extent to which each feature is contributing to model prediction across 10,000 randomly selected variants. SHAP analyses can be used to delineate the top contributing features at the individual variant level. For instance, Fig 2c, shows an example in which gnomAD3, SpliceAI, and CADD emerge as top features influencing the model towards classifying a variant as deleterious with a DITTO score of 1.

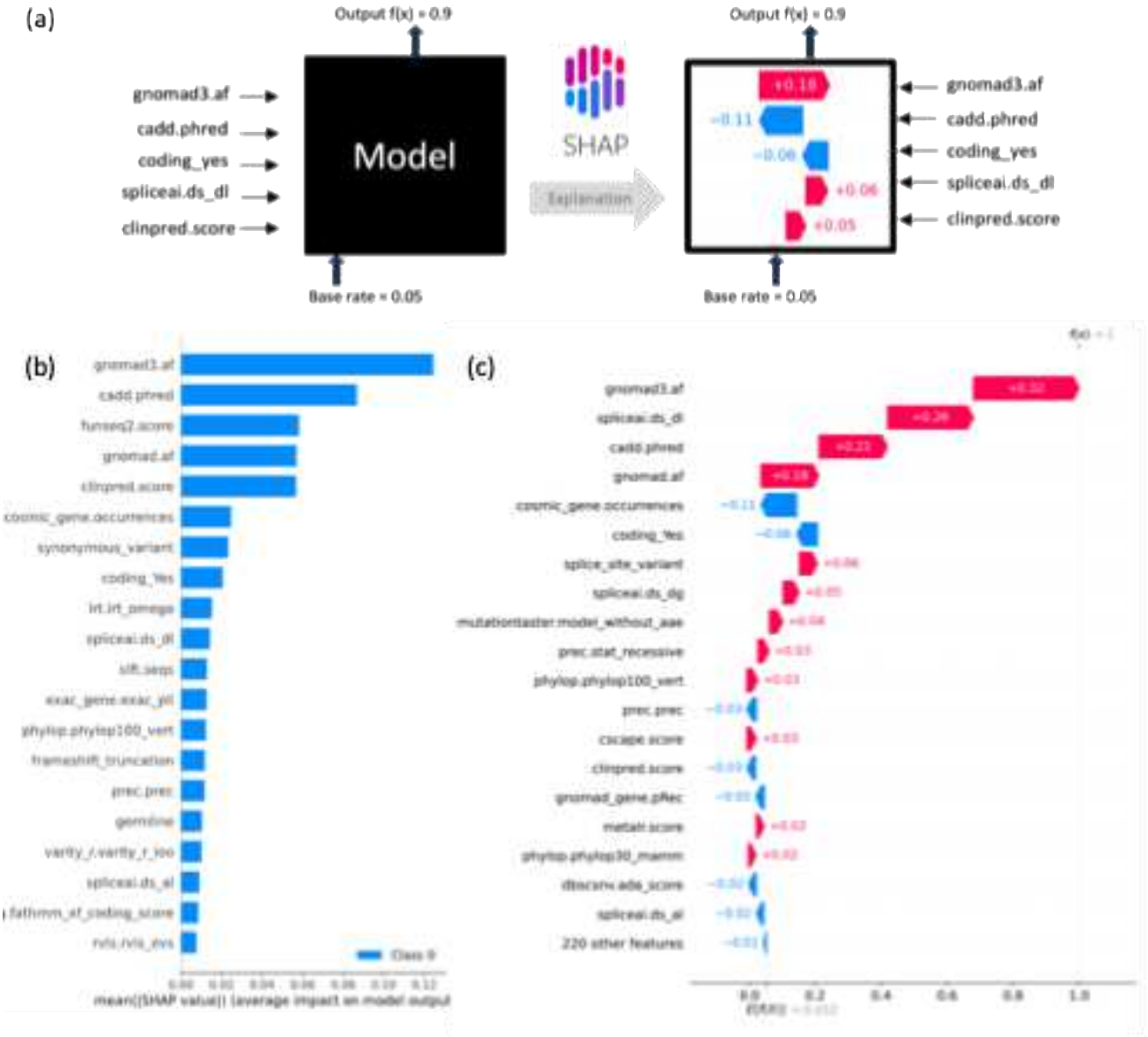


Figure 2: Explainability in DITTO predictions: a) Application of SHAP plots aids in explainability of DITTO predictions. **b)** The cumulative SHAP derived feature importance plot for the overall model derived from 10,000 random variants from training. The top features are GnomAD (gnomad and

gnomad3 allele frequencies), cadd.phred (CADD score) [43], funseq2.score (FunSeq ranking of non-coding regulatory cancer variants) [44], clinpred.score (ClinPred missense variant predictions) [45], synonymous_variant (variant consequence), cosmic_gene.occurences (count of number of times the gene appears in Cosmic database) [46,47], coding_Yes (likelihood of variant altering the coding sequence), Lrt.omega (evolutionary conservation of the amino acid residue) [48] and spliceai.ds_dl (SpliceAI score) [31]. **c)** The SHAP plot provides an indication of specific feature contributions to the model's prediction for an example variant with a DITTO score of 1.

2.4 Performance comparison against existing methods

For benchmarking of DITTO's performance, we conducted analysis on a set of 360,467 variants that were reported in ClinVar post the cut-off date used for training and validation data. This evaluation compared DITTO against several established and industry standard pathogenicity prediction methods, including CADD, ClinPred, MetaSVM, GERP [49], SpliceAI, Revel [50] and Alphamissense. Of note, a number of these methods are themselves annotations used as features in the training of DITTO. Comparisons were based on the normalized scores for each tool as described in the Methods. We assessed the comparative performance of these tools with Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves. DITTO outperformed all other evaluated classifiers, achieving an ROC-AUC of 0.99 and a PR-AUC of 0.99 (Fig 3a and 3b). In contrast, the other methods achieved ROC-AUC scores ranging from 0.55 to 0.77 and PR-AUC scores from 0.22 to 0.62, with CADD emerging as the next most effective classifier (Fig 3a and 3b).

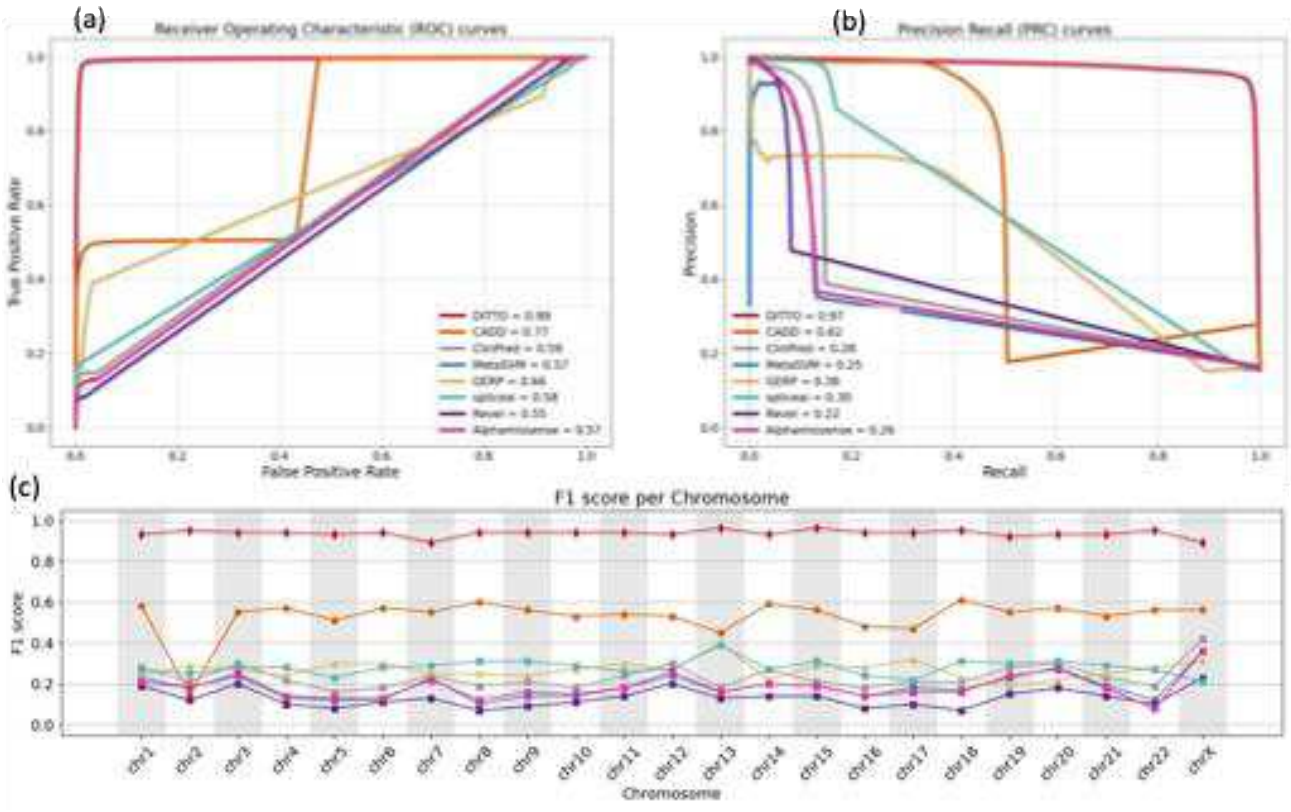


Figure 3: Comparison of the performance of DITTO, CADD, ClinPred, MetaSVM, GERP, Revel, SpliceAI, and Alphamissense. DITTO performed better overall than the other methods tested. **a)** Receiver Operating Characteristic (ROC) and **b)** Precision Recall (PR) curves for each classifier are shown. **c)** As measured by F1 score DITTO outperformed other methods across all chromosomes. Underlying data for the plot is provided in Table S5.

We reasoned that potential chromosome-specific biases stemming from variations in GC content, gene count, size, or density, or evolutionary history might impact predictions, and so compared DITTO predictions against the other tools, examining F1 scores on a chromosome-by-chromosome basis. As shown in Fig 3c, DITTO consistently outperformed other evaluated tools across all chromosomes. Variants from Chromosome Y and the mitochondrial genome were excluded due to low counts. Performance varied across chromosomes, suggesting potential chromosome-specific biases in training data or algorithms.

2.5 Model performance by variant consequence

To ensure fair comparison, we segmented our analysis based on variant consequence providing a fairer evaluation of each tool's effectiveness taking into account specializations. Based on this, we noted that DITTO demonstrated superior performance compared to the other tools evaluated. It generated accurate predictions across all chromosomes and variant consequences (Fig 4 and Tables S4 and S5).

DITTO accurately classified 90% of missense variants achieving the highest overall F1 score (0.82), greater than that of the meta predictors ClinPred, MetaSVM, and protein level predictor Alphamissense (Fig 4b). DITTO also showed superior performance in accurately classifying variant consequences often overlooked by other tools (Fig 4c-e). For example, intergenic variants are more challenging to classify and interpret. These are challenging due to underrepresentation in databases and sparse annotations. Despite this, DITTO accurately classified 97.9% of intergenic variants, achieving the highest F1 score (0.77) among all methods; CADD followed with 0.69. Many consequence agnostic prediction tools also perform poorly on splice variants, prom+pting development of specialized classifiers such as SpliceAI. DITTO achieved an F1 score of 0.99, outperforming SpliceAI (0.98) and matching CADD and ClinPred (both 0.99; Fig. 4d). Intronic variants outside canonical splice sites are also challenging to interpret and often excluded from analysis. DITTO accurately classified most of these, achieving an F1 score of 0.68, surpassing SpliceAI (0.56) and closely matching CADD (0.69; Fig 4e).

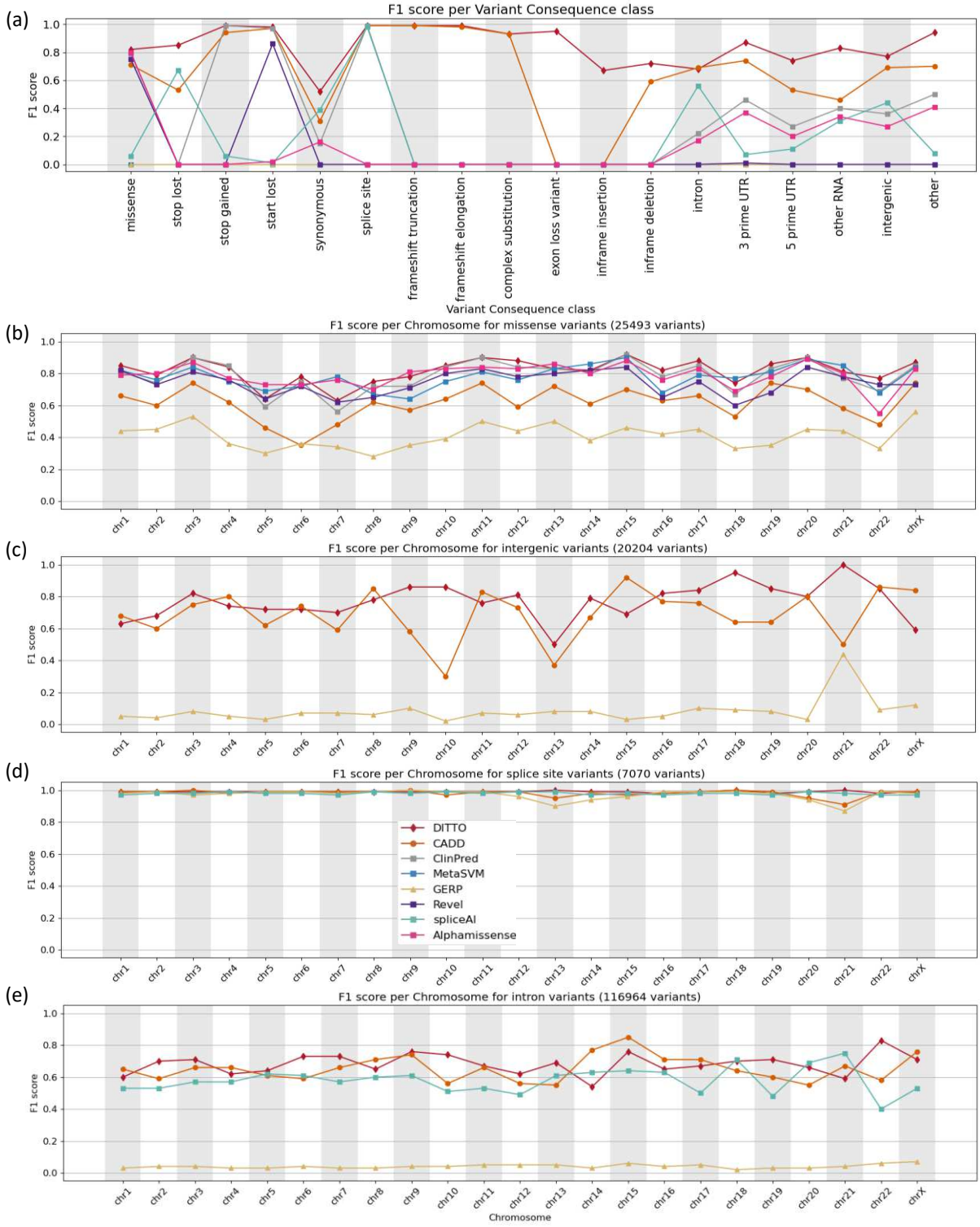


Figure 4: F1 scores shown across variant consequences and per chromosome. The x-axis represents the chromosome, and the y-axis represents the F1 score. **a)** Performance across different methods tested for all variants. **b-e)** Scores for the fraction of variants identified as **b)** missense, **c)** intergenic, **d)** splice site, and **e)** intronic. F1 scores used to generate these plots, and for other variant consequence types not shown in this Figure due to space constraints, are provided in Table S5.

Lastly, we assessed DITTO's performance on rare and ultra-rare variants as these are less frequent in training sets. Using gnomAD3 allele frequencies, we defined ultra-rare variants as MAF ≤

0.0001 (n=326,738) and rare variants as MAF < 0.001 (n=344,199). Compared to tools including CADD, Revel, Alphamissense, and SpliceAI, DITTO consistently achieved high F1 scores: 0.77 for ultra-rare and 0.74 for rare variants (Fig. S3a), while CADD and SpliceAI performed modestly (0.37 and 0.33, respectively). Given the focus of many tools on missense variants, we also compared performance in this subset (Fig. S3b). Here, Revel and Alphamissense showed strong results (F1: 0.56 and 0.54 for ultra-rare; 0.51 and 0.49 for rare), with DITTO remaining competitive (0.48 and 0.46, respectively). In contrast, CADD, SpliceAI, and GERP exhibit relatively lower performance in this context.

We used SHAP-based feature importance analysis to examine the top contributing features for ~500 predictions within each variant consequence category (Fig. 5). ClinPred, gnomAD and CADD emerge as the top features contributing to DITTO predictions overall regardless of the specific variant consequence (Fig 5). gnomAD, funseq2, and CADD were the top contributing features for missense variants (Fig 5b). gnomAD, CADD and SpliceAI were found to be the top contributing features for splice impacting variants (Fig 5c). gnomAD, funseq2, and CADD were the top contributing features for intronic variants (Fig 5d).

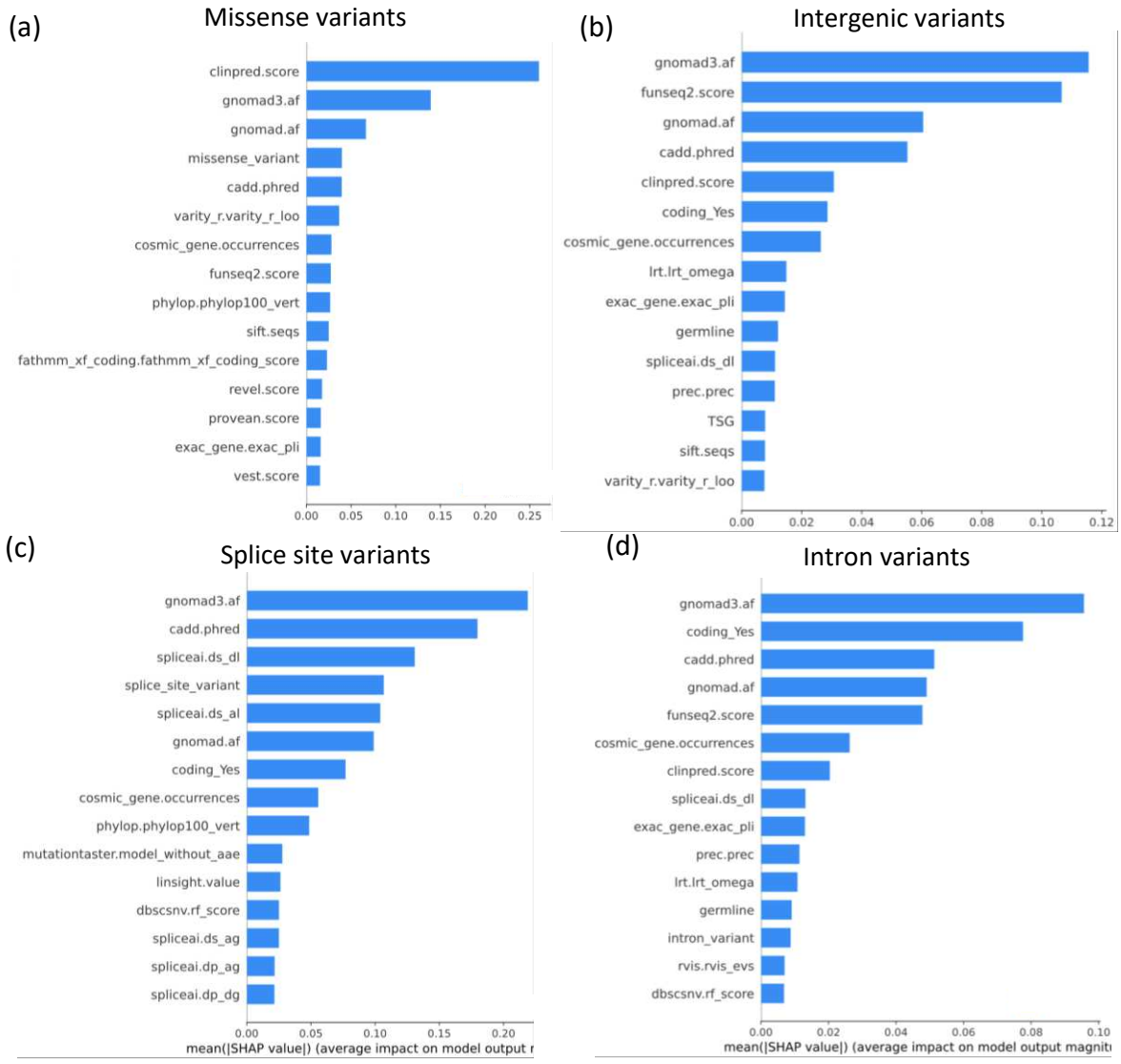


Figure 5. SHAP plots showing top contributing features for DITTO predictions across each variant consequence. An overall and variant consequence specific plot is shown for: **a)** missense **b)** intergenic **c)** splice site, and **d)** intronic variants. Plots are limited to display only the top 15 features. The importance of features can be seen to differ amongst variant consequence although some trends, such as importance of allele frequencies, are shared.

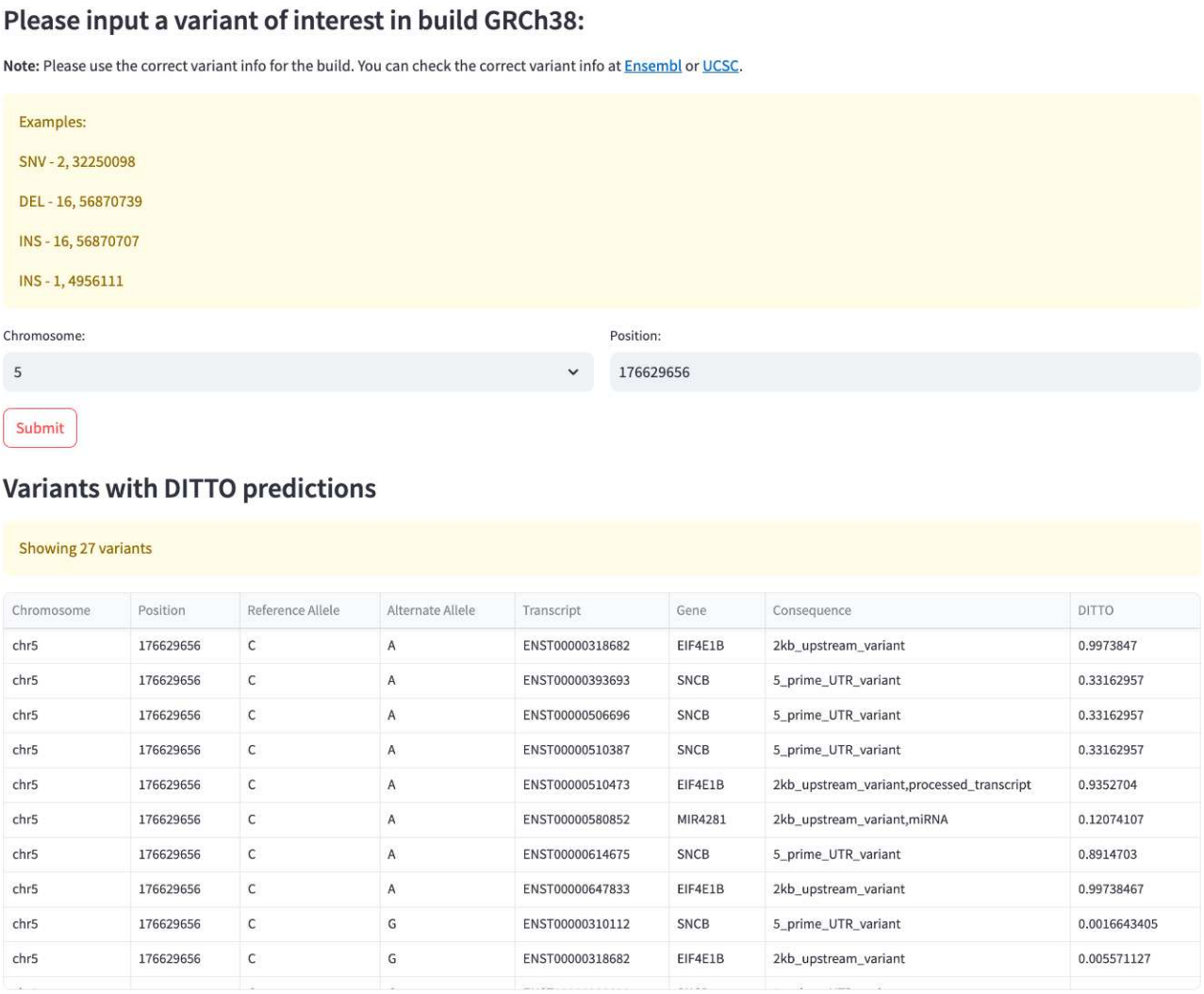
2.6 DITTO can predict deleterious score by Transcript

The impact of a variant can differ across transcripts, e.g. causing a frameshift in one while remaining deep intronic and benign in another. Transcript-specific expression patterns can make a variant pathogenic in one tissue but not in others [51,52]. DITTO's transcript-aware predictions help capture these context-dependent effects. As shown in Table S7, many variants exhibited notably different DITTO scores across transcripts. In some cases, overlapping genes drive differences. This is the case for the chr5:g.176629656C>A variant spanned by *MIR4281*, *SNCB*, and *EIF4E1B* where DITTO scores of 0.121, 0.332, and 0.997 are seen for the different genes, respectively. This is also the case for the chr1:g.43384485C>G variant present in both *MED8* and its antisense RNA *MED8-AS1* (0.99 and 0.31, respectively). In other cases, score differences arise between transcripts of the same gene. For a PPT1 variant,

the same position falls in the 3′ UTR of one transcript (DITTO score: 0.0) and in an alternative exon causing a missense change in another (DITTO score: 0.914). Similarly, a TEKT2 variant is deep intronic in one transcript (score: 0.0), but exonic with a stop gain in another (score: 1.0). A CHDR2 variant shows scores of 0.560, 0.791, and 0.971 across transcripts with the same consequence, influenced by factors such as transcript length, expression, amino acid position, and predictive features (e.g., CHASMplus, VEST). Ignoring transcript context in such cases can lead to misclassification.

Dissemination and access to DITTO predictions

As shown in Fig 6, our pre-trained DITTO model is accessible via a user-friendly streamlit app allowing users to fetch DITTO scores from DITTO database for the variants of interest (https://cgds-ditto.streamlit.app/).



Chromosome	Position	Reference Allele	Alternate Allele	Transcript	Gene	Consequence	DITTO
chr5	176629656	C	A	ENST00000318682	EIF4E1B	2kb_upstream_variant	0.9973847
chr5	176629656	C	A	ENST00000393693	SNCB	5_prime_UTR_variant	0.33162957
chr5	176629656	C	A	ENST00000506696	SNCB	5_prime_UTR_variant	0.33162957
chr5	176629656	C	A	ENST00000510387	SNCB	5_prime_UTR_variant	0.33162957
chr5	176629656	C	A	ENST00000510473	EIF4E1B	2kb_upstream_variant,processed_transcript	0.9352704
chr5	176629656	C	A	ENST00000580852	MIR4281	2kb_upstream_variant,miRNA	0.12074107
chr5	176629656	C	A	ENST00000614675	SNCB	5_prime_UTR_variant	0.8914703
chr5	176629656	C	A	ENST00000647833	EIF4E1B	2kb_upstream_variant	0.99738467
chr5	176629656	C	G	ENST00000310112	SNCB	5_prime_UTR_variant	0.0016643405
chr5	176629656	C	G	ENST00000318682	EIF4E1B	2kb_upstream_variant	0.005571127

Figure 6: DITTO webapp. In the illustrated example, a variant on Chromosome 5 is seen to be associated with eight transcripts overlapping 3 genes (*EIF4E1B*, *MIR481* and *SNCB*). DITTO assigns a score of 1 to transcript "ENST00000318682.11 (*EIF4E1B*)", indicating a high probability of pathogenicity. For transcript "ENST00000580852.1 (*MIR481*)", DITTO assigns a score of 0.12, which leans towards a benign interpretation. Transcript-aware analysis offers valuable insights into the variable effects of a variant, highlighting potential impact across different biological contexts.

We also developed a Nextflow pipeline to be used to facilitate the generation of DITTO scores for genome sequencing VCF files (Fig 1b). This pipeline is hosted and publicly accessible on GitHub: https://github.com/uab-cgds-worthey/DITTO. Leveraging this pipeline, we computed DITTO scores for all possible SNVs and for all known Insertions/Deletions/Indels catalogued in the gnomADv3.0 database. These scores have been compiled in to tabix indexed format [53] and are available for download via this link: https://s3.lts.rc.uab.edu/cgds-public/dittodb/dittodb.html. Completion of these additional resource intensive tasks ensures that these predictions are made widely available especially for those who may lack the necessary infrastructure to complete the task themselves.

2.7 Validation on *NF1* variants previously unseen by DITTO

To evaluate DITTO's performance and prediction confidence, we applied refined gene-specific cutoffs: ≥0.91 for deleterious and ≤0.11 for benign (based on the mean ±1 SD of ClinVar-classified variants). Using these thresholds, DITTO classified 927/937 (98.9%) of these variants in agreement with ClinVar with precision, recall, and F1-scores all at 0.99. This included 725/733 of pathogenic/likely pathogenic variants and 202/204 of the benign/likely benign variants. These refined thresholds substantially reduce the number of variants requiring manual review. Figure 7a shows their distribution across the NF1 gene. DITTO outperformed other methods, with CADD and GERP scoring 0.97 and 0.84 respectively (Fig S4). SpliceAI matched DITTO (scoring 0.99) when tested solely on splice variants.

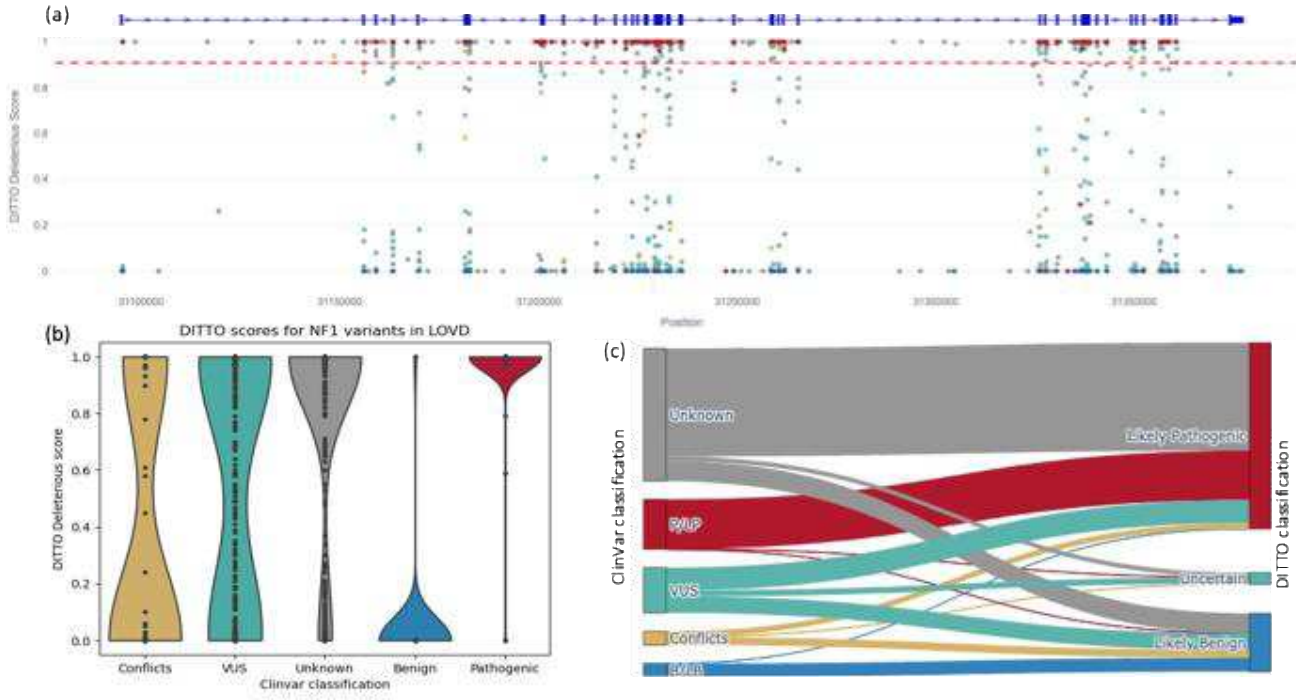


Figure 7: Results from testing of DITTO on the NF1 LOVD-derived dataset. a) A scatterplot showing the distribution of variants across the NF1 gene (x-axis) by DITTO score (y-axis), colored by ClinVar classification (blue for benign/likely benign, green for uncertain significance, red for pathogenic/likely pathogenic, and conflicting interpretations in gold). Intron-exon structure is shown by blue bar on top. The red dotted line represents a 0.91 cutoff for classification of a variant as deleterious with high confidence. **b)** Violin plots comparing DITTO scores against ClinVar class. **c)** A Sankey diagram showing the number of variants mapping from ClinVar classification to DITTO classification (see Table S6). Colors represent the same class across all plots.

Out of 10 variants that showed discordant classifications between DITTO and ClinVar, we found that five variant were either reclassified from Benign to pathogenic or vice versa (Fig 7c). These were reviewed using the canonical *NF1* transcript ENST00000356175 as the reference (Table 4). Variant 1, labeled as likely benign in ClinVar, received a DITTO score of 1.0. UCSC genome browser[54] inspection revealed four nearby pathogenic splice variants supporting DITTO's prediction and suggesting possible misclassified in ClinVar. Variant 2, a synonymous change, has been experimentally shown to create an alternate splice donor site, excluding exonic sequence [55,56], a rare mechanism likely underrepresented in the training data. Variants 3, 4, and 5 are classified as likely or known pathogenic in ClinVar, but are either deep intronic or structurally complex and thus may also be underrepresented in the training set [55,57,58]. Overall, DITTO correctly classified 99.5% of variants in this challenging NF1 dataset, which includes diverse variant types, many noncoding changes, and complex genotype–phenotype relationships.

Table 4. Findings for variants that had discordant classifications. Annotations are shown for each of the five *NF1* variants with discordant classification between ClinVar and DITTO. Annotations are all based on the ENST00000356175 *NF1* transcript.

	cDNA (HGVS notation)	Consequence	ClinVar	ClinVar review status	DITTO
1	c.3198-3_3199del	Splice site	Likely benign	no assertion criteria provided	1
2	c.2709G>A (p.Val903=)	synonymous	Pathogenic	criteria provided, multiple submitters, no conflicts	0
3	c.3975-1922_4111-2448delinsTTTACTTAGGT	Exon loss, intronic	Pathogenic	no assertion criteria provided	0
4	c.5206-38A>G	intronic	Likely pathogenic	criteria provided, single submitter	0
5	c.5750-1748_6184delinsCTA	intronic	Pathogenic	no assertion criteria provided	0

Using the same DITTO cutoffs, we classified 665 *NF1* variants currently identified as VUSs. Of these 333 were reclassified as likely deleterious, 242 as likely benign, and 90 remained VUSs. For example, the ClinVar VUS p.Gly309Ile (Variation ID: 1766337) was reclassified as likely deleterious with DITTO. Supporting evidence includes a dissimilar glycine to isoleucine substitution at a highly conserved position, absence from population databases, and prior ClinVar reports in two individuals, one with NF1 and the other with Hereditary cancer-predisposing syndrome. Additional pathogenic or likely pathogenic variants have also been identified in the surrounding region.

Finally, we applied DITTO to 1,950 NF1 variants present in LOVD, but absent from ClinVar at the time of analysis. DITTO classified 294 as likely benign, 1,589 as likely pathogenic, and 67 as VUSs. Variant types such as premature stops, splice variants, and frameshifts were predominantly classified as Likely pathogenic, whilst some missense variants were classified as Likely benign. This analysis demonstrates the utility of DITTO for variant classification, prioritization for curation or experimental validation, and identification of potentially misclassified variants.

3 Discussion

Our study demonstrates that DITTO, our transcript-aware and explainable machine learning classifier, delivers highly accurate predictions across all classes and consequences of small variants, with an overall accuracy of 99.6%. It consistently outperformed or matched leading tools, including both general-purpose and consequence-specific predictors across benchmarks, chromosomes and variant consequences. DITTO showed particularly strong performance in challenging regions often

overlooked by other tools, such as deep intronic and intergenic variants. In specialized assessments, including transcript-level scoring and rare variant interpretation, DITTO achieved superior F1 scores and demonstrated utility in reducing VUS classifications.

These improvements stem from several design choices: DITTO's use of a comprehensively OpenCRAVAT annotated and curated training set, its transcript-specific modeling framework, and the integration of diverse genomic features through explainable neural network architecture. The resulting model offers not only state-of-the-art performance but also transparency, thereby allowing users to understand which features contribute to each prediction.

Despite its strengths, the DITTO model, like others, is influenced by biases present in the underlying training data. Variant repositories are skewed toward certain diseases (e.g. cancer, neurologic disorders), organs or tissues (e.g. brain, heart), or genes (e.g. BRCA1, TP53, CFTR), often reflecting historical research focus or clinical accessibility. Variants with early or developmental effect (birth defects versus aging related disorders) or those located in coding and splice-site regions are more frequently studied than those in untranslated or disordered regions. Molecular features that are easier to assay (e.g., metabolic enzymes) are also overrepresented. Most available data comes from individuals of European ancestry. As a result, training datasets are skewed towards certain variants, clustered in specific genes, locations, and tissues. These biases shape model performance and remain a critical limitation. Addressing them is essential for advancing comprehensive variant interpretation.

4 Conclusions

While sequencing has become faster and more affordable, interpretation remains a primary bottleneck in delivering molecular diagnoses for families lacking a definitive molecular diagnosis. DITTO provides a powerful and scalable solution to this challenge. Its consistently high precision, recall, and F1-score make it well suited as an ACMG classification predictor for both research and clinical use. DITTO delivers robust performance across variant types, transcripts, and allele frequencies, with the added benefit of model interpretability and architecture supporting easy integration into existing workflows. With careful application of calibrated cutoffs, users can significantly reduce the burden of manual variant review without compromising classification confidence. Taken together, these strengths highlight DITTO as an easy to integrate, transformative tool for accelerating and improving the accuracy and efficiency of variant interpretation in diagnostic pipelines. This type of approach has the potential to improve diagnostic yield, reduce variant interpretation workload, and reduce turnaround time.

5 Methods

We developed a multi-step classification pipeline to distinguish benign from deleterious variants. The variants to be used in training and testing of the model are gathered, annotated, and preprocessed. We engineered features from annotated datasets to support model training. Finally, the model is trained using the processed dataset with stepwise tuning to avoid overfitting and to optimize performance.

Implementation and statistical analysis

DITTO is a tunable neural network implemented in Python (v3.10.11) [59] using the deep learning frameworks, Keras to build the neural network architecture and Optuna to tune the architecture [60,61]. We used Hyperopt [62] to optimize training parameters and prevent overfitting and executed on UAB high-performance computing cluster (Cheaha) to enable efficient training and evaluation of the neural network. We used the Python scikit-learn [63] package to evaluate DITTO's performance and compare against other methods. Figures were generated using Matplotlib and Seaborn libraries

[64,65]. The DITTO software, complete with data processing and tuning parameters used, is available on GitHub at https://github.com/uab-cgds-worthey/DITTO.

Dataset extraction and variant annotation

We downloaded 2,177,684 variants from the ClinVar repository along with corresponding ACMG classes and other annotations [66] (ClinVar accessed on 2023/06/06). We retained only ClinVar variants with ≥2-star review status and supporting evidence, excluding lower-confidence entries. After preprocessing (see next section), 329,136 variants remained and were annotated using OpenCRAVAT v2.4.1 [67], an open-source framework for variant analysis and interpretation that supports rapid quality-controlled annotation using data from repositories and prediction tools.

We annotated variants using 69 OpenCRAVAT modules (12 transcript-aware), generating 117 features including conservation, population frequency, and predicted functional impact [67] (see Table S1 for annotation categories by OpenCRAVAT). We extracted 360,467 variants from the ClinVar repository along with corresponding ACMG classes as an orthogonal testing dataset (variants reported after 2023/06/06 until 2024/09/17).

Data preprocessing

Several levels of filtering and preprocessing were applied to prepare the downloaded ClinVar variants for training. 39,970 variants were excluded as they had missing values for at least 30% of all features, leaving 2,137,714 variants extending to 11,153,639 variant - transcript pairings. Variants were assigned to a class based on their ClinVar extracted ACMG classification [66] collapsing as follows: Pathogenic, Likely Pathogenic and Pathogenic/Likely Pathogenic were assigned to the "High impact" category with 302,734 variant-transcript pairs; Benign, Likely Benign and Benign/Likely Benign were assigned "Low impact" category with 748,092 variant-transcript pairs. Low impact and High impact variants were then used as binary classes with an 80:20 random split for training and validation (see Table 1). Since we are interested in the variant impact across transcripts influenced by alternative splicing, start codon usage etc., variants were considered based on variant-transcript pairings [34,35]. A total of 842,659 variant-transcript pairs were selected for training, and 208,167 for validation (see Table S2 for test-training breakdown by variant consequence). After the filtering step, only two variants from chrY and none from chrM remained. Both the chrY variants were included in training. The test set (360,467 variants) underwent identical preprocessing. Training and test sets were mutually exclusive at the variant-transcript level.

We performed correlation analysis on our training data split to determine whether any of the features significantly correlated with the binary class (High/Low impact) (Fig S1). We identified six features (aloft, MetaSVM, Polyphen2, varity_r, dbscsnv and gnomad3 allele frequency **[16,68–72]** that showed correlation (pearson $r^2 > 0.95$) to one another but not to the binary class. Upon review, we opted not to drop these features; reasoning whilst correlated, they convey distinct information. For example, MetaSVM

Table 1: Breakdown of high and low impact categorized variants used for training and testing.

		High impact	Low impact
Training	variants	40,318	107,790
	variant-transcript	242,376	600,283
Validation	variants	10,080	26,947
	variant-transcript	60,358	147,809
Testing	variants	56,015	304,452

The variant and variant-transcript test-training split is shown. The model learns from the training set and is evaluated on the separate, unseen test dataset in order to assess overfitting and model generalizability.

and dbscSNV both make use of conservation scores, but MetaSVM provides damage prediction scores for missense variants, whilst dbscSNV provides annotations for SNVs within splicing consensus regions **[73]**.

Model Training and Architecture

Our DITTO variant deleteriousness classifier is based on a tunable neural network that can optimize itself for all hyperparameters (network shape, depth, optimizer, dropout rate and activation function). For architecture, we used Keras, an easy-to-use high level python framework for writing neural networks, and Optuna, which offers a flexible architecture for tuning machine learning algorithms [60,61,74]. For optimizing parameters, we used Hyperopt, which uses Tree of Parzen Estimator's (TPE) [41] stochastic search algorithm for optimized parameters supporting quicker convergence and better results compared to grid or random search. 'Loss', calculated as the error between the predicted values and actual values, is used to minimize the error in making correct predictions.

As depicted in Fig 1, the input layer of the network consisted of the preprocessed features along with variant consequences (i.e., missense, synonymous, etc.). The subsequent layers of the network perform feature extraction and transformation to learn the complex relationships between the input features and pathogenicity. The final layer of the network is a fully connected layer that outputs the probability of deleteriousness of the variant. This structure facilitates the learning of complex relationships between the input features and the target variable, ultimately generating a pathogenicity score between 0 (benign) and 1 (deleterious).

Model evaluation and comparison

Our trained model was evaluated on validation data and then on our indepFpaendent test dataset to assess performance. The evaluation metrics used included accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). ROC curves are graphical representations of the sensitivity (true positive rate) versus the 1-specificity (false positive rate) of a binary classifier system [75]. The area under the ROC curve (AUC) provides a measure of the overall performance of the classifier, with higher AUC values indicating better performance. Precision-Recall (PR) curves are graphical representations of the precision (true positives / all positives) versus the recall (true positives / all actual positives) of a binary classifier system [75,76]. The precision-recall AUC (PR-AUC) is a measure of the classifier's ability to identify true positives while minimizing false positives, focused on performance with respect to the minority class for imbalanced datasets [76]. The F1 score measures model performance by calculating the harmonic mean of precision and recall for the minority positive class [77]. It is calculated as:

$$F1\ score = 2 * (precision * recall) / (precision + recall)$$

The calculation is based on precision (ratio of correctly predicted positive observations to the total predicted positives) and recall (or sensitivity; the ratio of correctly predicted positive observations to all observations in the actual class) for identification of pathogenic variants.

To further evaluate model performance, we compared DITTO's performance against a set of widely used and highly regarded variant classification tools (CADD [43], ClinPred [45], MetaSVM [25], GERP [78], SpliceAI [31], Revel [50]). We performed comparison based on: overall performance, by chromosome, and by variant consequence category. For testing purposes, we considered only the transcript with highest DITTO score for each variant since all, but MetaSVM and Revel are not transcript aware. We normalized comparator tool scores that have scores greater than 1 to a (0,1) range,

and calculated performance metrics comparing to our previously defined ground truth (binary class: High/Low impact). In order to control for biases and provide as fair a comparison as possible, we normalized using the minimum and maximum scores for each comparison performed i.e. min/max scores for chr1 will be different from min/max scores for chr2, and min/max scores for missense variants will be different from min/max scores for splice variants.

Each variant can have multiple consequences. For example, a missense variant within two bases of a canonical splice site can be annotated as both "missense" and "splice site" [79]. Since we wished to perform a fair and unbiased comparison, we segmented comparisons based on variant consequence, categorizing each variant in to just one category for use in comparisons, choosing the one considered most likely to be deleterious. For instance, a variant that is both splice site and missense would be analyzed in the splice category. The hierarchy of these consequence categories is detailed in Table S3. Model performance was then determined based on comparisons of F1-scores across each of these methods. The F1 score was used for comparisons based on our binary classification, somewhat uneven class distribution, and equal interest in correctly identifying both deleterious and benign variants for our diagnostic use case. F1 scores range from 0 (worst) to 1 (perfect precision and recall).

Model interpretability

Feature importance scores were derived using SHapley Additive exPlanations (SHAP) calculations [80]. SHAP calculations facilitate quantification of each feature's contribution to the model, defining their significance both in terms of overall classification accuracy and individual prediction performance. Visual representations in the form of Shapley plots were created to depict incorporation of features within the model. These plots illustrate the trajectory of the model's output from a baseline prediction to the final outcome, offering an easily explainable transparent view of the influence each feature had on the prediction[80,81].

Nextflow Pipeline

To streamline the process from extracting variants from a sample VCF to generating pathogenicity predictions, we developed a Nextflow-based pipeline [42] . This pipeline automates the extraction of variants, leverages OpenCRAVAT's annotation framework, applies the pre-trained DITTO model to compute pathogenicity scores, and outputs deleteriousness predictions for each variant-transcript pair. In addition to the DITTO scores, the pipeline also provides accompanying annotations to support downstream interpretation and variant prioritization.

Web Application

We developed a web application using Streamlit [82] to allow users to easily query small variants at a genomic location, specified in the GRCh38 HGNC nomenclature, for DITTO score prediction without requiring personal computational resources for model construction or hosting: https://cgds-ditto.streamlit.app/. The app retrieves and displays the DITTO score with annotations including information on overlapping/corresponding gene(s), transcript(s), and functional consequences via querying DITTO database (publicly available to download at https://s3.lts.rc.uab.edu/cgds-public/dittodb/dittodb.html).

Application on NF1 disease

To evaluate DITTO's applicability on an additional test dataset, we used it to analyze variants previously identified in the Neurofibromin (*NF1*) gene, the causal gene for autosomal dominant Neurofibromatosis type-1 disease (NF1). We selected NF1 as the gene is of interest for ongoing

projects in our lab. *NF1* gene is also large, spanning 60 exons, many transcripts, and ~350kb. It functions as a tumor suppressor gene by negatively regulating Ras [83,84]. NF1 can present with café au lait spots, disfiguring cutaneous and/or plexiform neurofibromas, optic nerve gliomas, malignant peripheral nerve sheath tumors, skeletal defects, attention and learning deficits, and other cognitive disabilities [85,86]. As d*e novo* variants are a common cause of NF1, *NF1* locus has a higher spontaneous mutation rate than most gene *loci*. The phenotypic diversity in NF1 is partly attributed to hypomorphic variants that maintain partial function. Variant classification in NF1 is a complex task. By applying DITTO in the case of NF1, we sought to determine the tool's effectiveness in classification and prioritization in this type of complex situation.

The Leiden Open Variation Database (LOVD) is an open access, flexible web-based platform designed to support compilation and display of DNA variants linked to genetic disorders https://databases.lovd.nl/shared/variants/NF1/) [87–89]. LOVD is a manually curated dataset providing not only variant data, but also curated patient-centered data where available [87–89]. We extracted 3,759 *NF1* variants from LOVD (accessed on 2/16/2023) and associated ACMG classes from ClinVar for the following analysis. From this set, we excluded the 497 NF1 variants previously used as part of the DITTO training set leaving 3,262. We also excluded 207 variants with conflicting interpretations. 937 of the variants remaining had pathogenic/likely pathogenic (733) or benign/likely benign (204) classifications and were used for testing.

6 Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Data Availability

Code used for this project has been made available at https://github.com/uab-cgds-worthey/DITTO.

DITTO score for all possible SNVs and gnomAD indels are available at https://s3.lts.rc.uab.edu/cgds-public/dittodb/dittodb.html.

DITTO webapp is available for public use - https://cgds-ditto.streamlit.app/.

NF1 variants are available to download from LOVD here
- https://databases.lovd.nl/shared/variants/NF1/

Databases to annotate variants are downloaded from OpenCravat. Please find the instructions in our project repo - https://github.com/uab-cgds-worthey/DITTO/blob/main/docs/install_openCravat.md

Competing interests
The authors declare that they have no competing interests

Funding
This work was supported by UAB Worthey Lab Start-Up funds (EAW), the UAB Pilot Center for Precision Animal Modeling (National Institutes of Health 1U54OD030167; to EAW), Children's Tumor Foundation Hack4NF award (CTF; TKKM and EAW), The UAB Carmichael Scholar award for academic achievement (TKKM). The funders had no role in the design of the study or in collection, analysis, and interpretation of data or writing of the manuscript.

Authors' contributions

EAW conceived the study and provided overall supervision. TKKM developed the methodology, performed data preprocessing, and trained the models. MG and BW assisted with data acquisition and preparation. All authors contributed to manuscript writing and approved the final version.

Acknowledgments

The authors gratefully acknowledge the resources provided by the University of Alabama at Birmingham IT-Research Computing group for high performance computing (HPC) support and CPU time on the Cheaha compute cluster. The authors also wish to specifically thank and acknowledge William Warriner from UAB HPC, who greatly assisted us with architecting the necessary resources to run pipelines on Cheaha compute cluster. Finally, we wish to thank the additional members of the UAB Center for Computational Genomics and Data Sciences (CGDS) and the UAB Biological Data Sciences Core (U-BDS) for helpful discussions during this project's development.

7 References

1. Splinter K, Adams DR, Bacino CA, Bellen HJ, Bernstein JA, Cheatle-Jarvela AM, et al. Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N Engl J Med [Internet]. 2018;379:2131–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30304647

2. Marshall CR, Chowdhury S, Taft RJ, Lebo MS, Buchan JG, Harrison SM, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. npj Genom Med. 2020;5:47.

3. Zastrow DB, Kohler JN, Bonner D, Reuter CM, Fernandez L, Grove ME, et al. A toolkit for genetics providers in follow-up of patients with non-diagnostic exome sequencing. J Genet Couns. 2019;28:213–28.

4. Bick D, Fraser P, Gutzeit M, Harris J, Hambuch T, Helbling D, et al. Successful Application of Whole Genome Sequencing in a Medical Genetics Clinic. J Pediatr Genet [Internet]. 2016;06:061–76. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28496993

5. Prokop JW, May T, Strong K, Bilinovich SM, Bupp C, Rajasekaran S, et al. Genome sequencing in the clinic: the past, present, and future of genomic medicine. Physiol Genomics [Internet]. 2018;50:563–79. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29727589

6. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. Am J Hum Genetics. 2017;100:185–92.

7. Wojcik MH, Reuter CM, Marwaha S, Mahmoud M, Duyzend MH, Barseghyan H, et al. Beyond the exome: What's next in diagnostic testing for Mendelian conditions. Am J Hum Genet. 2023;110:1229–48.

8. Holt JM, Wilk B, Birch CL, Brown DM, Gajapathy M, Moss AC, et al. VarSight: prioritizing clinically reported variants with binary classification algorithms. Bmc Bioinformatics. 2019;20:496.

9. Worthey EA. Analysis and annotation of whole-genome or whole-exome sequencing-derived variants for clinical diagnosis. Curr Protoc Hum Genetics Éditor Board Jonathan L Haines Et Al [Internet]. 2013;79:Unit 9 24. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24510652

10. Stenton SL, O'Leary M, Lemire G, VanNoy GE, DiTroia S, Ganesh VS, et al. Critical assessment of variant prioritization methods for rare disease diagnosis within the Rare Genomes Project. medRxiv. 2023;2023.08.02.23293212.

11. Angelis A, Tordrup D, Kanavos P. Socio-economic burden of rare diseases: A systematic review of cost of illness evidence. Heal Polic. 2015;119:964–79.

12. Marshall DA, Gerber B, Lorenzetti DL, MacDonald KV, Bohach RJ, Currie GR. Are We Capturing the Socioeconomic Burden of Rare Genetic Disease? A Scoping Review of Economic Evaluations and Cost-of-Illness Studies. PharmacoEconomics. 2023;41:1563–88.

13. Currie GR, Gerber B, Lorenzetti D, MacDonald K, Benseler SM, Bernier FP, et al. Developing a Framework of Cost Elements of Socioeconomic Burden of Rare Disease: A Scoping Review. PharmacoEconomics. 2023;41:803–18.

14. Rehm HL, Alaimo JT, Aradhya S, Bayrak-Toydemir P, Best H, Brandon R, et al. The landscape of reported VUS in multi-gene panel and genomic testing: Time for a change. Genet Med. 2023;25:100947.

15. Fowler DM, Rehm HL. Will variants of uncertain significance still exist in 2030? Am J Hum Genet. 2024;111:5–10.

16. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Hum Mutat. 2022;43:1012–30.

17. Chen E, Facio FM, Aradhya KW, Rojahn S, Hatchell KE, Aguilar S, et al. Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. JAMA Netw Open. 2023;6:e2339571.

18. Aguirre J, Padilla N, Özkan S, Riera C, Feliubadaló L, Cruz X de la. Choosing Variant Interpretation Tools for Clinical Applications: Context Matters. Int J Mol Sci. 2023;24:11872.

19. Shendure J, Findlay GM, Snyder MW. Genomic Medicine–Progress, Pitfalls, and Promise. Cell. 2019;177:45–57.

20. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.

21. Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, et al. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. 2019;20:223.

22. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. Nat Protoc. 2016;11:1782–7.

23. Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. Hum Genet. 2018;137:665–78.

24. Bauskis A, Strange C, Molster C, Fisher C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. Orphanet J Rare Dis. 2022;17:233.

25. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet [Internet]. 2015;24:2125–37. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25552646

26. Holt JM, Wilk B, Birch CL, Brown DM, Gajapathy M, Moss AC, et al. VarSight: prioritizing clinically reported variants with binary classification algorithms. BMC Bioinform. 2019;20:496.

27. Li C, Zhi D, Wang K, Liu X. MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning. Genome Med. 2022;14:115.

28. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45:gkx390-.

29. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;

30. Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature. 2021;599:91–5.

31. Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176:535-548.e24.

32. Spielmann M, Kircher M. Computational and experimental methods for classifying variants of unknown clinical significance. Cold Spring Harb Mol Case Stud. 2022;8:a006196.

33. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12:103.

34. Bhuiyan SA, Ly S, Phan M, Huntington B, Hogan E, Liu CC, et al. Systematic evaluation of isoform function in literature reports of alternative splicing. BMC Genom. 2018;19:637.

35. Chen H, Shaw D, Zeng J, Bu D, Jiang T. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics. 2019;35:i284–94.

36. Gargis AS, Kalman L, Bick DP, Silva C da, Dimmock DP, Funke BH, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol [Internet]. 2015;33:689–93. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26154004

37. Pena LDM, Jiang Y-H, Schoch K, Spillmann RC, Walley N, Stong N, et al. Looking beyond the exome: a phenotype-first approach to molecular diagnostic resolution in rare and undiagnosed diseases. Genet Med [Internet]. 2018;20:464–9. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28914269

38. Shashi V, Schoch K, Spillmann R, Cope H, Tan QK, Walley N, et al. A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. Genet Med [Internet]. 2019;21:161–72. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29907797

39. Samani A, English KG, Lopez MA, Birch CL, Brown DM, Kaur G, et al. DOCKopathies: A systematic review of the clinical pathologies associated with human DOCK pathogenic variants. Hum Mutat. 2022;43:1149–61.

40. Fresard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med. 2019;25:911–9.

41. Watanabe S. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. arXiv. 2023;

42. Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

43. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2018;47:D886–94.

44. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. 2014;15:480.

45. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103:474–83.

46. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res [Internet]. 2015;43:D805-11. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25355519

47. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2018;47:gky1015-.

48. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.

49. Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res [Internet]. 2005;15:901–13. Available from: https://www.ncbi.nlm.nih.gov/pubmed/15965027

50. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99:877–85.

51. Cummings BB, Karczewski KJ, Kosmicki JA, Seaby EG, Watts NA, Singer-Berk M, et al. Transcript expression-aware annotation improves rare variant interpretation. Nature. 2020;581:452–8.

52. Consortium TGte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.

53. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics. 2011;27:718–9.

54. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

55. Wimmer K, Schamschula E, Wernstedt A, Traunfellner P, Amberger A, Zschocke J, et al. AG-exclusion zone revisited: Lessons to learn from 91 intronic NF1 3′ splice site mutations outside the canonical AG-dinucleotides. Hum Mutat. 2020;41:1145–56.

56. Brinckmann A, Mischung C, Bässmann I, Kühnisch J, Schuelke M, Tinschert S, et al. Detection of novel NF1 mutations and rapid mutation prescreening with Pyrosequencing. ELECTROPHORESIS. 2007;28:4295–301.

57. Messiaen LM, Callens T, Mortier G, Beysen D, Vandenbroucke I, Roy NV, et al. Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. Hum Mutat. 2000;15:541–55.

58. Evans DG, Bowers N, Burkitt-Wright E, Miles E, Garg S, Scott-Kitching V, et al. Comprehensive RNA Analysis of the NF1 Gene in Classically Affected NF1 Affected Individuals Meeting NIH Criteria has High Sensitivity and Mutation Negative Testing is Reassuring in Isolated Cases With Pigmentary Features Only. EBioMedicine. 2016;7:212–20.

59. Rossum GV, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace;

60. Chollet F, others. Keras. GitHub;

61. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv. 2019;

62. Bergstra J, Yamins D, Cox D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Proc 12th Python Sci Conf. 2013;13–9.

63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011;12:2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html

64. Hunter, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007;9:90–5.

65. Waskom M. seaborn: statistical data visualization. J Open Source Softw. 2021;6:3021.

66. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med [Internet]. 2015;17:405–24. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25741868

67. Pagel KA, Kim R, Moad K, Busby B, Zheng L, Tokheim C, et al. Integrated Informatics Analysis of Cancer-Related Variants. JCO Clin Cancer Inform. 2020;4:CCI.19.00132.

68. Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, et al. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nat Commun. 2017;8:382.

69. Kim S, Jhong J-H, Lee J, Koo J-Y. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10:2.

70. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet. 2013;76:7.20.1-7.20.41.

71. Wu Y, Liu H, Li R, Sun S, Weile J, Roth FP. Improved pathogenicity prediction for rare human missense variants. Am J Hum Genet. 2021;108:1891–906.

72. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.

73. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016;37:235–41.

74. Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G, et al. Optuna. Proc 25th ACM SIGKDD Int Conf Knowl Discov Data Min. 2019;2623–31.

75. Hanley JA. Wiley StatsRef: Statistics Reference Online. 2017;

76. Keilwagen J, Grosse I, Grau J. Area under Precision-Recall Curves for Weighted and Unweighted Data. PLoS ONE. 2014;9:e92209.

77. Hand DJ, Christen P, Kirielle N. F*: an interpretable transformation of the F-measure. Mach Learn. 2021;110:451–6.

78. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol. 2010;6:e1001025.

79. Eilbeck K, Lewis SE. Sequence Ontology annotation guide. Comp Funct Genom. 2004;5:642–7.

80. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv. 2017;

81. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. SN Appl Sci. 2021;3:272.

82. streamlit/streamlit: Streamlit — A faster way to build and share data apps. [Internet]. [cited 2024 Feb 14]. Available from: https://github.com/streamlit/streamlit

83. Trovó-Marqui A, Tajara E. Neurofibromin: a general outlook. Clin Genet. 2006;70:1–13.

84. Scheffzek K, Welti S. Neurofibromatosis Type 1, Molecular and Cellular Biology. 2012;305–26.

85. Rosenbaum T, Wimmer K. Neurofibromatosis type 1 (NF1) and associated tumors. Klinische P Diatrie [Internet]. 2014;226:309–15. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25062113

86. Peltonen S, Kallionpää RA, Peltonen J. Neurofibromatosis type 1 (NF1) gene: Beyond café au lait spots and dermal neurofibromas. Exp Dermatol. 2017;26:645–8.

87. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, Dunnen JT den. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32:557–63.

88. Fokkema IFAC, Kroon M, Hernández JAL, Asscheman D, Lugtenburg I, Hoogenboom J, et al. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet. 2021;29:1796–803.

89. Minkelen R van, Bever Y van, Kromosoeto JNR, Withagen-Hermans CJ, Nieuwlaat A, Halley DJJ, et al. A clinical and genetic overview of 18 years neurofibromatosis type 1 molecular diagnostics in the Netherlands. Clin Genet. 2014;85:318–27.

8 Figures, tables and additional files

Table 2: Overview of hyperparameter space for tuning of our DITTO model. Parameters for model tuning spanned number of layers, activation functions, neurons per layer, kernel initializers, dropout rates, optimizers, and batch size.

Table 3. DITTO performance metrics.

Table 4. Findings for variants that had discordant classifications. Annotations are shown for each of the five *NF1* variants with discordant classification between ClinVar and DITTO. Annotations are all based on the ENST00000356175 *NF1* transcript.

Table 1: Breakdown of high and low impact categorized variants used for training and testing.

Figure 1. An overview of the training and testing methodology, and the architecture of DITTO model. a) 2,177,684 variants were extracted from ClinVar, filtered, and annotated for 117 distinct features using OpenCRAVAT. Following preprocessing and featuring engineering, we performed a

train:validation:test variant split. Training and validation data was used to tune and train a neural network classifier that generates a deleteriousness score (0,1) for variant damage prediction. **b)** We developed a Nextflow pipeline to read VCF files, annotate variants using OpenCRAVAT, and run DITTO. **c)** The DITTO architecture is a single hidden layer neural network with 239 neurons in the input layer, 161 neurons in the hidden layer, and 1 output layer with 96,162 parameters.

Figure 2: Explainability in DITTO predictions: a) Application of SHAP plots aids in explainability of DITTO predictions. **b)** The cumulative SHAP derived feature importance plot for the overall model derived from 10,000 random variants from training. The top features are GnomAD (gnomad and gnomad3 allele frequencies), cadd.phred (CADD score) [43], funseq2.score (FunSeq ranking of non-coding regulatory cancer variants) [44], clinpred.score (ClinPred missense variant predictions) [45], synonymous_variant (variant consequence), cosmic_gene.occurences (count of number of times the gene appears in Cosmic database) [46,47], coding_Yes (likelihood of variant altering the coding sequence), Lrt.omega (evolutionary conservation of the amino acid residue) [48] and spliceai.ds_dl (SpliceAI score) [31]. **c)** The SHAP plot provides an indication of specific feature contributions to the model's prediction for an example variant with a DITTO score of 1.

Figure 3: Comparison of the performance of DITTO, CADD, ClinPred, MetaSVM, GERP, Revel, SpliceAI, and Alphamissense. DITTO performed better overall than the other methods tested. **a)** Receiver Operating Characteristic (ROC) and **b)** Precision Recall (PR) curves for each classifier are shown. **c)** As measured by F1 score DITTO outperformed other methods across all chromosomes. Underlying data for the plot is provided in Table S5.

Figure 4: F1 scores shown across variant consequences and per chromosome. The x-axis represents the chromosome, and the y-axis represents the F1 score. **a)** Performance across different methods tested for all variants. **b-e)** Scores for the fraction of variants identified as **b)** missense, **c)** intergenic, **d)** splice site, and **e)** intronic. F1 scores used to generate these plots, and for other variant consequence types not shown in this Figure due to space constraints, are provided in Table S5.

Figure 5. SHAP plots showing top contributing features for DITTO predictions across each variant consequence. An overall and variant consequence specific plot is shown for: **a)** missense **b)** intergenic **c)** splice site, and **d)** intronic variants. Plots are limited to display only the top 15 features. The importance of features can be seen to differ amongst variant consequence although some trends, such as importance of allele frequencies, are shared.

Figure 6: DITTO webapp. In the illustrated example, a variant on Chromosome 5 is seen to be associated with eight transcripts overlapping 3 genes (*EIF4E1B*, *MIR481* and *SNCB*). DITTO assigns a score of 1 to transcript "ENST00000318682.11 (*EIF4E1B*)", indicating a high probability of pathogenicity. For transcript "ENST00000580852.1 (*MIR481*)", DITTO assigns a score of 0.12, which leans towards a benign interpretation. Transcript-aware analysis offers valuable insights into the variable effects of a variant, highlighting potential impact across different biological contexts.

Figure 7: Results from testing of DITTO on the NF1 LOVD-derived dataset. a) A scatterplot showing the distribution of variants across the NF1 gene (x-axis) by DITTO score (y-axis), colored by ClinVar classification (blue for benign/likely benign, green for uncertain significance, red for pathogenic/likely pathogenic, and conflicting interpretations in gold). Intron-exon structure is shown by blue bar on top. The red dotted line represents a 0.91 cutoff for classification of a variant as deleterious with high confidence. **b)** Violin plots comparing DITTO scores against ClinVar class. **c)** A Sankey diagram showing the number of variants mapping from ClinVar classification to DITTO classification (see Table S6). Colors represent the same class across all plots.

Supplementary material is attached to the manuscript.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supptables.xlsx