The Pharmacogenomics Journal www.nature.com/tpj

ARTICLE OPEN

Check for updates

Computational variant predictors for pharmacogenomics: from evaluation of single alleles to assessment of adverse drug reactions to antidepressants

Jacek Hajto[1], Marcin Piechota[1], Ilse Krätschmer[2], Paula Konowalska[1], Gabriel E. Boyle[3], Douglas M. Fowler[3,4,5], Malgorzata Borczyk[1,6]✉ and Michal Korostynski[1,6]✉

Currently, pharmacogenetics relies on partially annotated star alleles, leaving novel variants and complex haplotypes uninterpretable. Computational scoring frameworks could overcome these limitations. Here, we comprehensively evaluated the ability of existing (CADD, FATHMM-XF, PROVEAN, MutationAssessor, SIFT, PhyloP100, APF, APF2) and novel (PharmGScore and PharmMLScore) variant effect predictors to assess pharmacogenetic alleles in multiple scenarios. Altogether we analyzed 541 PharmVar alleles, high-throughput *CYP2C9* and *CYP2C19* mutational maps, and 200 642 UK Biobank exomes linked with health records containing antidepressant treatment outcomes. Many evaluated tools, especially ensemble frameworks, matched or exceeded star allele classifications (ROC-AUC up to 0.85 for allele definitions, 0.95 in vitro; TPR up to 0.99 for exomes) and accurately predicted severe antidepressant adverse events for carriers of deleterious variants in *CYP2C19* (OR 1.20–1.35). Our findings show that computational predictors deliver star allele accuracy while overcoming their limitations. With additional validation, computational tools could enhance clinical decision frameworks by enabling continuous scoring, incorporating previously unknown variants, and providing genome-wide applicability.

The Pharmacogenomics Journal (2026)26:8 ; https://doi.org/10.1038/s41397-026-00399-0

INTRODUCTION

Linking genomic variation to drug response remains a challenge in precision medicine [1], mostly evident in psychiatry, in which 20–60% of patients do not respond to first-line therapy and drug development is limited [2]. Hundreds of pharmacogenetic variants have already been translated into clinical guidelines [3] and consequently associated with better outcomes and fewer adverse drug reactions (ADRs), which account for five percent of all hospital admissions [4–8].

In particular, star alleles, which are curated lists of specific variants on a haplotype, classified in a substrate-agnostic manner into only four functional levels (increased, normal, decreased, or no function) have facilitated clinical implementations. Unfortunately, this framework cannot accommodate novel and rare variants outside these lists, nor does it capture the full spectrum of the protein's functional outcomes. Importantly, over 90% of annotated pharmacogenetic alleles are rare across ancestries, and 17% of CYP functional diversity remains unexplained by known variants [9–12]. Moreover, the classical star allele approach cannot score most compound haplotypes or efficiently scale to additional pharmacogenes.

Computational predictors help classify variants in diagnostic settings [13]. However, generic tools, particularly conservation-based ones, often underperform on pharmacogenes [14, 15], driving efforts to build dedicated models [16–18], such as APF and its successor APF2 [19, 20]. Built on 337 known pharmacogenetic variants, APF has 93% sensitivity and specificity for deleterious variants. APF2 offers further improvement by adding AlphaMissense deep neural network scores. Remaining limitations of tools dedicated to pharmacogenes include difficulty of use, and lack of support for noncoding variants, especially splice-site or regulatory variants. Current scores cannot combine variant effects across a gene or haplotype or are dedicated to one gene only. To the best of our knowledge, none have yet been compared directly against star allele calls in large-scale human data with linked drug-response phenotypes.

There is a clear need in the field for scalable, automated computational approaches to evaluate pharmacogenetic variants [17, 21, 22], and it remains unknown whether such predictors can match or surpass the performance of star alleles. Here, we address this issue through a multilevel evaluation strategy that provides the most comprehensive evaluation of computational tools for pharmacogenomics to date. We first tested if predictors recover PharmVar star allele functions when multiple SNPs are aggregated per haplotype and showed that PharmGScore exhibited the best performance. Second, we investigated whether the predictors

[1]Laboratory of Pharmacogenomics, Department of Molecular Neuropharmacology, Maj Institute of Pharmacology, Polish Academy of Sciences, Krakow, Poland. [2]Institute of Science and Technology Austria, Klosterneuburg, Austria. [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [4]Department of Bioengineering, University of Washington, Seattle, WA, USA. [5]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [6]These authors contributed equally: Malgorzata Borczyk, Michal Korostynski. ✉email: gosborcz@if-pan.krakow.pl; malgorzata.m.borczyk@gmail.com

Received: 26 July 2025 Revised: 4 December 2025 Accepted: 4 February 2026
Published online: 09 March 2026

generalize well to novel variants independent of traditional training sources, demonstrating high sensitivity of pharmacogene-dedicated scores on deep mutational scanning data for *CYP2C9* and *CYP2C19*. Third, we benchmarked computational tools against a classical pharmacogenetic framework in 200 642 UK Biobank (UKB) exomes, confirming that gene-level burden tests recapitulate star allele phenotypes in population exomes despite incomplete coverage. Finally, to assess applicability to real-world clinical data, we analyzed antidepressant treatment outcomes in the UKB and showed that computational tools predicted severe antidepressant adverse drug reactions with odds ratios of 1.20–1.35, closely matching star allele associations in a clinically meaningful way. Our results provide a robust empirical framework for further method development and show that additive ensemble scoring offers a comprehensive, scalable, and flexible alternative for deriving pharmacogenomic profiles from genome sequencing data.

MATERIALS AND METHODS

Data extraction and preprocessing

Eight pharmacogenes: *CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A5, DPYD, NUDT15* and *SLCO1B1* were examined. For all analyses, single-nucleotide variants located within or up to 60 base pairs from any exon according to Ensembl v106 [23] were retained. All datasets were processed under this definition.

Scores for coding variants were downloaded from dbNSFP v4.3a [24]. For locations with multiple transcripts, the value associated with the Ensembl canonical transcript [23] was chosen, and when absent, available scores were averaged. Whole-genome predictors CADD v1.6 [25] and FATHMM-XF [26] were added, using raw CADD values to match the scale of other tools. APF and APF2 were reconstructed as described [19, 20], substituting VEST4 for the unavailable VEST3. A complete list of predictors is provided in Supplementary Table 1.

Star allele haplotype definitions and assigned clinical functions were downloaded from PharmVar v6.0.10 [27] (Supplementary Table 2). Each star-allele single-nucleotide variant (SNV) was annotated with every predictor and distance to the exon. Missing values and variants lying outside the predefined annotation window were set to not available (NA). For each star allele, a burden score was calculated as the sum of available scores, and minimum values were allocated if all component variants were NA.

Amino acid substitutions from deep mutational scanning data for *CYP2C9* and *CYP2C19* [28, 29] were converted from amino acid to nucleotide changes. Only amino acid changes achievable through a single nucleotide substitution were kept. When a substitution could arise from multiple single nucleotide paths, individual scores were averaged to have one prediction value per mutation.

OQFE exomes in the UKB (data field 23156) were processed by splitting multiallelic sites and retaining variants within the study genes and annotation window. Each variant was annotated with all predictors and NA was assigned if no score existed. Gene-specific burden scores per participant were obtained by summing variant values, with the predictor minimum (neutral score) used when no evaluated variants were observed.

Due to known discrepancies among available tools [30], star alleles were inferred with PGxPOP v1.0 [31] and the Polygenic tool (Intelliseq SA, Krakow, Poland) with updated allele definitions from PharmVar (v. 6.0.10). Concordance was evaluated at haplotype and phenotype levels. For phenotypes, PGxPOP outputs were accepted directly, whereas Polygenic diplotypes were converted with PharmCAT v2.13. European allele frequencies were obtained from PharmGKB [32]. Polygenic-based calls were employed to stratify participants by functional groups.

Antidepressant-related phenotype extraction is provided in the Supplementary Methods. Briefly, data were parsed for drugs used for depression, including treatment-resistant depression (e.g., amitriptyline, citalopram, fluoxetine, sertraline, paroxetine, venlafaxine, mirtazapine, trazodone, quetiapine, clomipramine, imipramine, duloxetine, aripiprazole, bupropion, escitalopram), using both prescription codes and free-text entries. Prescriptions per participant were aggregated into continuous therapy courses, defined by the absence of dose or drug changes. Severe adverse drug reactions were identified through hospital and death registry ICD-10 codes T43.0-T43.2 and Y49.0-Y49.2. Participants with such codes were classified as intentional poisoning ($n = 481$) or accidental poisoning ($n = 121$), the latter requiring no record of drug-related self harm.

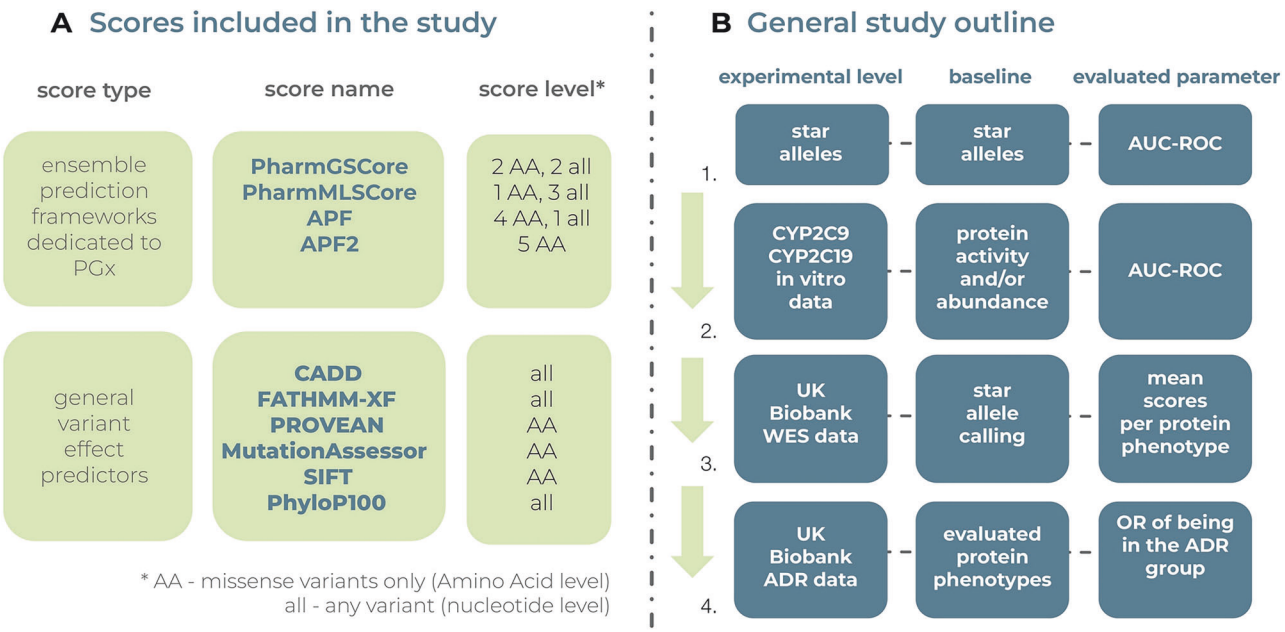

Fig. 1 High-level overview of the study. A Scores included in the analyses. PGx-specific prediction frameworks (PharmGScore, PharmMLScore, APF, and APF2) are ensemble scores that integrate multiple general variant predictors. Six general variant predictors were selected for comparison: CADD, FATHMM-XF, PROVEAN, MutationAssessor, SIFT, and PhyloP100. Some single and ensemble component scores evaluate only missense variants (marked as 'AA' in the 'score level' column), while others can score any variant in the whole genome (marked as 'all' in the 'score level' column). **B** General Study Outline. The study evaluates computational predictors across four levels of complexity: 1. Comparison with curated star allele definitions from the PharmVar database. 2. Validation against high-throughput in vitro measurements of *CYP2C9* and *CYP2C19* protein abundance and enzymatic activity. 3. Assessment of concordance between computational predictions and classical star-allele calls derived from UKB whole-exome sequencing for key pharmacogenes. 4. Application of both star allele-based and computational approaches to UKB clinical records to predict antidepressant treatment switches and severe adverse drug reactions (ADRs).

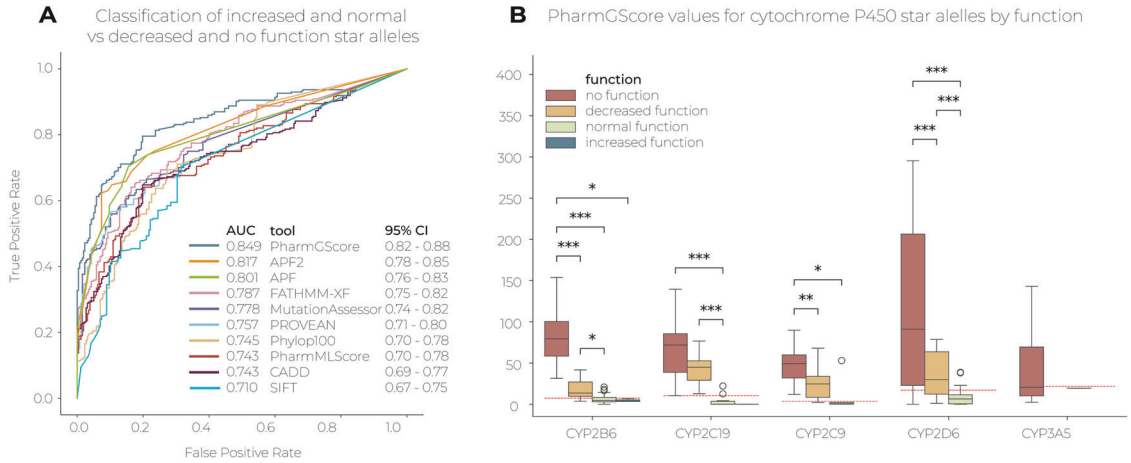

Fig. 2 Effective prediction of star allele phenotypes for genes in PharmVar database using computational tools. A ROC curves for the ten evaluated algorithms are displayed, with corresponding AUC values and 95% CIs provided on the graph for each score. For ROC analysis, star alleles with normal or increased function were assigned a value of 0, while those with no or decreased function were assigned a value of 1. ROC curves for individual phenotype classifications are presented in Supplementary Figure 4. **B** PharmGScore distribution for cytochrome P450 (CYP) star alleles grouped by functional phenotype (for other tools, see Supplementary Figure 5). Differences between phenotypes for each gene were determined with one-way ANOVA and pairwise t-tests with Bonferroni correction (full statistical results are available in Supplementary Table 4). Bonferroni corrected *p*-values: ns: not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Red lines indicate algorithmically determined cutoffs for damaging variants for each gene.

Methods

We set up two novel variant-evaluating frameworks, PharmGScore and PharmMLScore, with full descriptions provided in the Supplementary Methods. Briefly, PharmGScore was created from four complementary predictors (CADD, FATHMM-XF, Mutation Assessor, PROVEAN). We annotated every SNV located within 60 bp of an exon (Supplementary Figure 2) with those scores, gene-wise quantile-normalised their values to a 0–150 scale and averaged to yield one score per variant. Pre-computed PharmGScore values are available (see Data and code availability). PharmMLScore was developed using 91 four-score ensembles, each including CADD and FATHMM-XF plus two additional high-coverage predictors selected from dbNSFP. Variant scores were phred-scaled, capped at 40, rescaled to 0–1, summed per star allele, and combined with variant counts to form five input features. A feed-forward neural network was trained with gene-wise leave-one-out cross-validation to distinguish normal/increased from decreased/no function alleles. The best model included CADD, FATHMM-XF, SIFT and PhyloP100 and was selected as the final PharmMLScore.

We evaluated the performance of ten selected scores (CADD, FATHMM-XF, PROVEAN, MutationAssessor, PhyloP100, SIFT, APF, APF2, PharmMLScore, PharmGScore) on star alleles with known functional consequences using receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) metrics with 95% confidence interval (CI) computed with the bootstrap method. Alleles of reduced activity (decreased and no function) were contrasted with functional ones (normal and increased), and the optimal decision thresholds were defined by maximizing Youden's index. Pairwise analyses for all functional groups were performed analogously. We compared score values between all functional groups with one-way ANOVA and pairwise t-tests excluding variants with missing values.

For deep mutational scanning data of *CYP2C9* and *CYP2C19*, we categorized variants into three groups: damaging (<0.2), medium-impact (0.2–0.8), and benign or increased function (>0.8) and generated ROC curves for each score (classification of <0.2 vs >0.8).

In WES analysis, UKB participants were stratified by star allele-based enzyme activity. For each of the ten predictors, within these strata we calculated True and False Positive Rates (TPR/FPR) defined as the proportion of correctly and incorrectly classified individuals, respectively.

For the clinical outcome analysis, UKB participants were classified based on *CYP2C19* score values for each tool into two groups: without damaging variants and with damaging variants. When considering star alleles, carriers of normal or increased function alleles were assigned to the first group, and carriers of decreased or no function alleles to the second. Associations between *CYP2C19* activity, the incidence of ADRs, and the number of pharmacological therapies, were assessed using χ^2 tests.

RESULTS

In this study, we evaluated computational tools for pharmacogenomic variant scoring across varying data complexity levels, including UKB clinical data (Fig. 1).

Prediction of the functional impact of star alleles with known consequences

We first examined how computational tools evaluate whole star allele definitions when assuming an additive framework. We compared ten variant scoring tools using pharmacogenetic haplotypes from PharmVar, focusing on star allele definitions. The dataset included 541 star alleles (544 unique SNVs, 4.26 SNVs per allele) across 8 genes (*CYP2B6, CYP2C19, CYP2C9, CYP2D6, CYP3A5, DPYD, NUDT15, SLCO1B1*), with 9 variants linked to increased protein function, 248 to normal, 92 to decreased, and 192 to no function (Supplementary Figure 3). We assessed each tool's ability to distinguish normal or increased function from decreased or no function star alleles. PharmGScore performed best (ROC-AUC = 0.849), followed by APF2 (0.817) and APF (0.801). (Fig. 2A and Supplementary Table 3).

For pairwise comparisons of functional consequences, classification was most effective for normal vs. no function alleles (mean AUC = 0.782, Supplementary Figure 4 and Supplementary Table 3). Distinguishing increased function vs. normal function alleles was least effective (mean AUC = 0.413). Overall, highly damaging variants were classified correctly in most cases, while increased function variants were commonly misclassified as damaging. PharmGScore performed best in four of six pairwise comparisons, CADD in decreased function vs. no function (AUC = 0.71), and PharmMLScore in increased function vs. normal function (AUC = 0.58). For genes with few star alleles (*NUDT15*, *CYP3A5*, *SLCO1B1*), the tools' performance and interpretability were limited.

Predictions vary by gene (Fig. 2B, Supplementary Figure 5), so universal deleteriousness thresholds are impractical. We propose gene- and tool-specific thresholds using Youden's index (Supplementary Table 3), applied, for consistency, even to tools with predefined score thresholds (e.g., CADD, APF, APF2).

We analyzed score distributions for haplotypes across five Cytochrome P450 genes (PharmGScore in Fig. 2B and others in Supplementary Figure 5). PharmGScore showed significantly higher values for no function vs. normal function star alleles in four genes, with

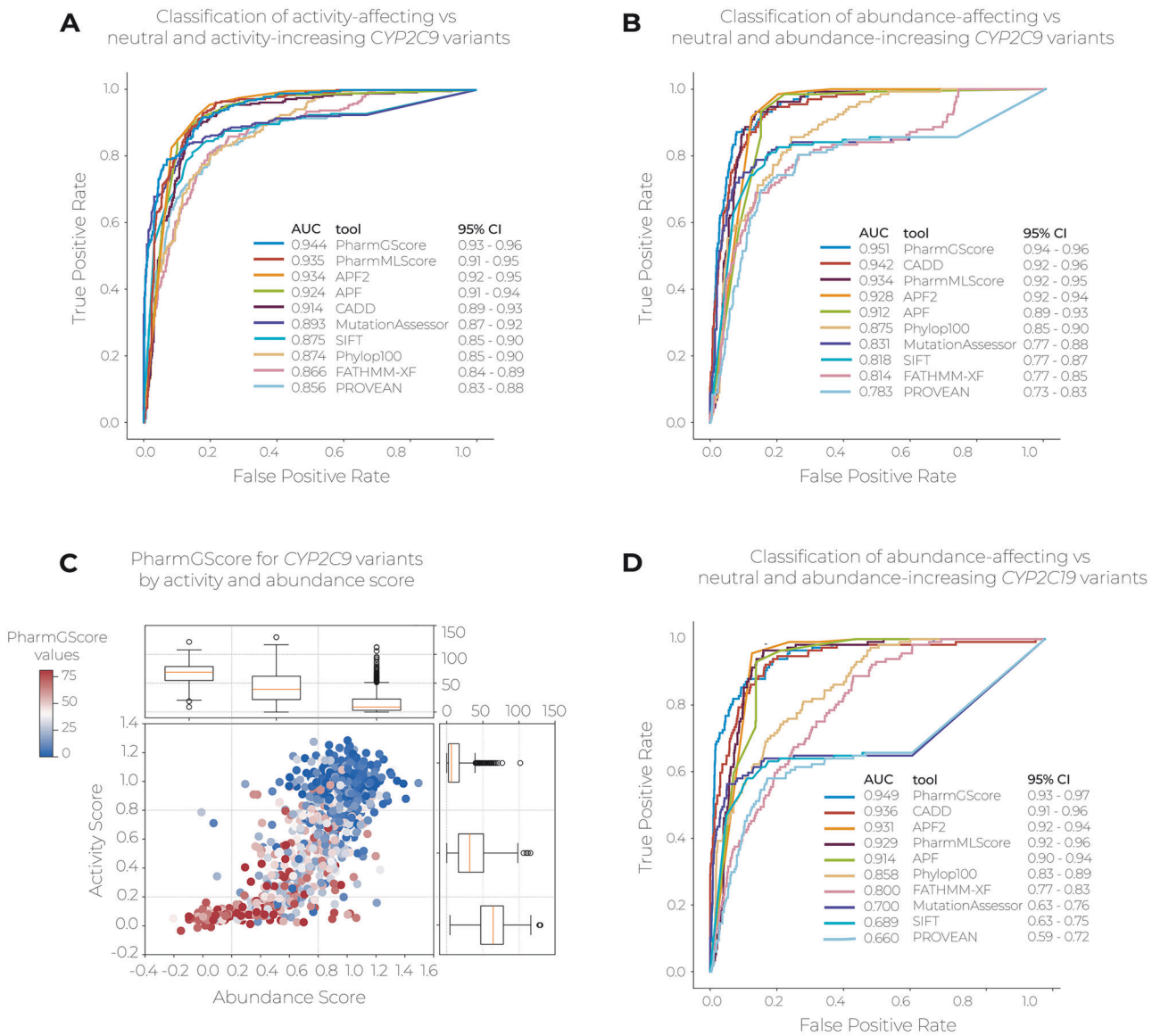

Fig. 3 Effective prediction of known and novel deleterious SNVs using computational tools. A ROCs for the classification of *CYP2C9* activity-affecting variants by the scores evaluated in this study. Variants with activity scores greater than 0.8 were assigned a value of 0 (normal or increased function), while those with scores less than 0.2 were assigned a value of 1 (decreased or no function). For variants not evaluated by a score, a neutral consequence (0) was assumed (PROVEAN, SIFT, and MutationAssessor do not evaluate nonsense variants; APF2, PhyloP100, MutationAssessor, SIFT, and PROVEAN do not evaluate synonymous variants). **B** Same as A, but for the classification of *CYP2C9* abundance-affecting variants. **C** Scatterplot depicting *CYP2C9* activity and abundance scores, colored by PharmGScore, for 1 010 SNVs with both measurements available. **D** Same as A, but for the classification of *CYP2C19* abundance-affecting variants.

CYP2D6 and *CYP2B6* best resolved. FATHMM-XF, APF, CADD, and PharmMLScore also effectively distinguished functional groups (Supplementary Figure 5 and Supplementary Table 4). *CYP3A5* was the hardest to score (not significant for all tools). Its non-functional allele *CYP3A5*3* (92% frequency in Europeans) has non-coding splice site variants, misclassified as benign by CADD and FATHMM-XF, indicating a need for specialized scoring for high-frequency impaired-function alleles. *CYP2C9* required lower deleteriousness thresholds than *CYP2B6* and *CYP2D6*. Overall, ensemble scores outperformed individual scores (70% of gene-level separations correct for ensemble tools and 43% for single predictors, Supplementary Table 4).

Variant predictor performance for enzymatic activity and protein abundance measurement

Datasets beyond observational studies allow for the evaluation of computational tools on novel variants. This previously unexplored area is important due to the limited number and biased frequency (towards common variants) of known alleles. We tested ten tools using high-throughput in vitro functional data for *CYP2C9* and *CYP2C19* missense, synonymous, and nonsense variants [28, 29]. Less than 10% of these variants appear in gnomAD, providing an independent genotype-phenotype map for mostly novel variants.

We analyzed 1 329 *CYP2C9* variants with activity scores and 1 459 with abundance scores (1 096 and 1 195 missense, 212 and 244 synonymous, 21 and 20 stop-gain) building ROC curves for classification of "near-normal" (score > 0.8) versus "low" (score < 0.2) phenotypes. All predictors performed well, with AUC ≥ 0.86 for activity (Fig. 3A) and AUC ≥ 0.78 for abundance (Fig. 3B). Pharmacogene-specific ensemble scores (PharmGScore, PharmMLScore, APF, APF2) and CADD achieved AUC > 0.90 in both assays, performing best at detecting damaging variants. PharmGScore's median values decreased stepwise with activity

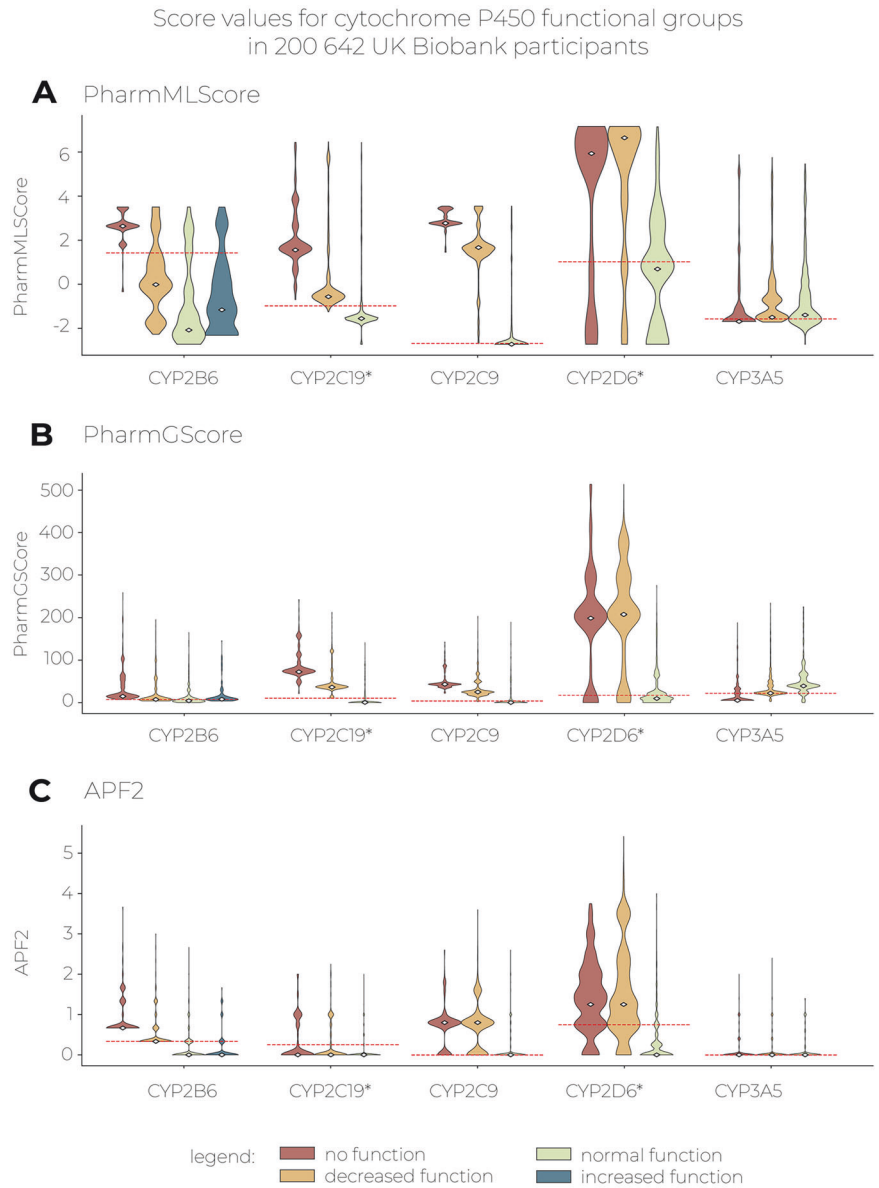

Fig. 4 Violin plots of computational tool score distribution cytochrome P450 genes for UKB participants, stratified by metabolizer status. The top three scores based on True Positive Rate (TPR) are shown: **A** PharmMLScore, **B** PharmGScore, **C** APF2. Protein functional groups were derived from star alleles identified using the Polygenic tool. The dotted line represents the cut-off threshold determined by Youden's index, based on PharmVar database data. White dots with black outlines denote median values. The width of each violin is proportional to the number of participants with a given score value.

(<0.2 = 64.33; 0.2–0.8 = 31.66; >0.8 = 6.41) and abundance (<0.2 = 68.91; 0.2–0.8 = 21.60; >0.8 = 3.09) (Fig. 3C and Supplementary Figure 6). For 1 958 *CYP2C19* variants with abundance scores, AUCs ranged from 0.66 (PROVEAN) to 0.95 (PharmGScore) (Fig. 3D). Missense-only methods (PROVEAN, SIFT, MutationAssessor) failed to classify 39 damaging nonsense variants. Because our ensembles weight missense and nucleotide-level features equally, we tested missense-only performance in *CYP2C9* and *CYP2C19* data (Supplementary Figure 7) and observed only modest differences (AUCs: MutationAssessor 0.927, SIFT 0.900, CADD 0.879, FATHMM-XF 0.838; PROVEAN 0.847), indicating that nucleotide-level predictors capture most of the nonsynonymous signal. In summary, CADD and pharmacogene-specific ensemble methods consistently achieved AUC > 0.9, demonstrating robustness for novel damaging variants.

To further explore possible factors influencing predictor performance, we evaluated tool performance on *CYP2C9* missense variants, stratified by evolutionary conservation, amino acid substitution biochemical impact, and allele frequency. Pharmacogene-specific tools and CADD achieved higher AUC values compared to other predictors for variants in regions of low evolutionary conservation (particularly at the vertebrate scale). All tools performed better on substitutions with high severity, whereas differences related to allele frequency were not significant (Supplementary Figure 8, Supplementary Table 5).

Computational prediction of pharmacogenetic variant consequences in next-generation sequencing data

We hypothesized that in population-level exome data, computational predictors recapitulate star allele phenotypes. To test this, we assessed ten predictors using star allele calls from 200 642 UKB WES sequences, called with Polygenic and PGxPOP tools. After confirming concordance and alignment with European star allele frequencies [32] (Supplementary Figures 9 and 10), Polygenic

Table 1. Association between predicted deleterious variants in *CYP2C19* and depression treatment outcomes in the UKB WES data.

	>2 treatment changes			Severe ADRs		
Score	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**
Star alleles	1.02	0.97–1.08	0.39	**1.20**	**1.01–1.42**	**0.03**
PharmGScore	1.03	0.98–1.09	0.20	**1.24**	**1.05–1.47**	**0.01**
PharmMLScore	1.03	0.98–1.08	0.25	**1.23**	**1.05–1.46**	**0.01**
APF	0.98	0.90–1.06	0.55	1.22	0.94–1.59	0.13
APF2	1.06	0.91–1.22	0.46	1.33	0.86–2.06	0.20
CADD	**1.11**	**1.01–1.22**	**0.02**	**1.34**	**1.00–1.78**	**0.05**
FATHMM-XF	0.99	0.92–1.07	0.84	1.12	0.87–1.45	0.38
MutationAssessor	1.00	0.93–1.09	0.90	1.17	0.90–1.51	0.24
PhyloP100	1.14	0.95–1.37	0.14	1.00	0.53–1.86	0.99
PROVEAN	0.97	0.90–1.06	0.53	1.25	0.97–1.62	0.09
SIFT	1.09	0.95–1.26	0.21	1.41	0.92–2.16	0.11

Variants in *CYP2C19* were evaluated using computational scores, with chi-squared tests assessing associations with two or more depression treatment changes or ADRs to antidepressants.
OR, Odds Ratio.

results were used. Score values for five CYP genes were plotted by predicted protein phenotypes (Fig. 4 and Supplementary Figure 11). For *CYP2C19* and *CYP2D6*, increased function variants could not be classified due to WES data limitations (see Discussion). We calculated true positive (TPR) and false positive (FPR) rates per gene and functional category (Supplementary Table 6). No function star alleles were well predicted (average TPR = 0.61, FPR = 0.27), while increased metabolizer alleles had higher false calls (TPR = 0.64, FPR = 0.50). For *CYP2C19*, PharmMLScore and PharmGScore outperformed the other scores, while APF and APF2 struggled with splice site variants in the no function *2 allele. PharmMLScore showed the highest sensitivity (0.81), followed by PharmGScore (0.78) and APF2 (0.72). PROVEAN and MutationAssessor, focusing on amino acid substitutions, had the highest specificities (0.16 and 0.20, respectively).

Star allele classification may misidentify individuals by overlooking rare or novel variants. For example, in the UKB cohort, we identified a *CYP2C19* missense variant (p.Glu354Lys) in 15 participants, with a high CADD score (24.8) and low protein abundance (0.19), suggesting a damaging effect.

Prediction of severe adverse events to antidepressants using star alleles and computational predictors

Building on the performance of ensemble tools established above, we analyzed UKB health records for antidepressant outcomes. Finally, we assessed whether predicted damage stratifies antidepressant treatment switching and severe ADR codes comparably to star allele calls. Prior work linked *CYP2C19* poor metabolism to switching, side-effects, and non-response to escitalopram [33, 34]. Accordingly, we examined *CYP2C19*, assessing prescription switches and severe ADRs. We labelled individuals as carriers or non-carriers of damaging *CYP2C19* variants for each score (see Methods).

Therapy change was any medication or dose switch, and ≥2 switches served as our non-response proxy. Among 75 256 antidepressant users, 29 343 met this criterion. Neither star alleles nor most predictors were significantly associated with damaging variants and therapy switches (Table 1, Fig. 5A). CADD alone reached nominal significance (p = 0.023). Notably, seven of ten tools showed OR > 1 but with wide intervals. PharmGScore and PharmMLScore both resembled the star allele estimate.

We next analyzed 602 antidepressant poisoning diagnoses, including 89 participants with Y49.0-Y49.2 and 517 with T43.0-T43.2 ICD-10 codes. Again, odds ratios were computed for carriers of damaging *CYP2C19* variants versus non-carriers. Damaging *CYP2C19* status increased poisoning risk according to star allele-based prediction (OR = 1.20 Table 1, Fig. 5B) as well as PharmGScore (1.24) and PharmMLScore (1.23). CADD gave a higher but less precise result (1.34). Overall, nine of ten predictors indicated OR > 1 for carriers of damaging *CYP2C19* variants to be in the ADR group. As hospitalizations due to antidepressant poisonings include intentional overdoses, we restricted the analysis to participants without any self-harm diagnostic codes ($n = 121$, Supplementary Figure 12). This resulted in similar but not significant results for star alleles (OR = 1.26, CI 0.76 - 1.83, p-value = 0.22), PharmGScore (OR = 1.19, CI 0.81 - 1.72, p-value = 0.37) and PharmMLScore (OR = 1.22, CI 0.84 - 1.77, p-value = 0.29) but not for CADD (OR = 0.88, CI 0.41 - 1.90, p-value = 0.75).

DISCUSSION

Pharmacogenetic testing for common genotypes has proven effective but remains insufficient for complete interpretation of next-generation sequencing data. Notably, even a single unannotated variant can significantly alter an individual's drug metabolism profile [9, 10, 12]. Population-based biobanks, whole-haplotype and deep mutational scanning datasets combined with machine learning offer promising advancements [3, 35]. Here, we integrated the above approaches with star allele-based protein phenotyping and benchmarked ten variant scoring tools across eight pharmacogenes, with a particular focus on *CYP2C19* due to its important role in psychiatric pharmacotherapy [36]. To the best of our knowledge, this study represents the first comprehensive effort to evaluate entire pharmacogenetic sequences across multiple genes and four increasingly complex data levels.

Apart from evaluating existing scores and building on previous work [16, 17, 19, 20], we developed two novel ensemble approaches which can evaluate both coding and non-coding variants and are designed to analyze whole haplotype sequences. PharmGScore was designed to overcome some known limitations of current tools [15]. It integrates a diverse set of existing scores and ensures a broad scoring range to distinguish moderately to highly damaging variants. Across various tasks, PharmGScore consistently ranked among the top performers. To facilitate its adoption by the pharmacogenomic community, we provide precomputed PharmGScore values for all SNVs across human exons (±60 bp). PharmMLScore was developed to model the

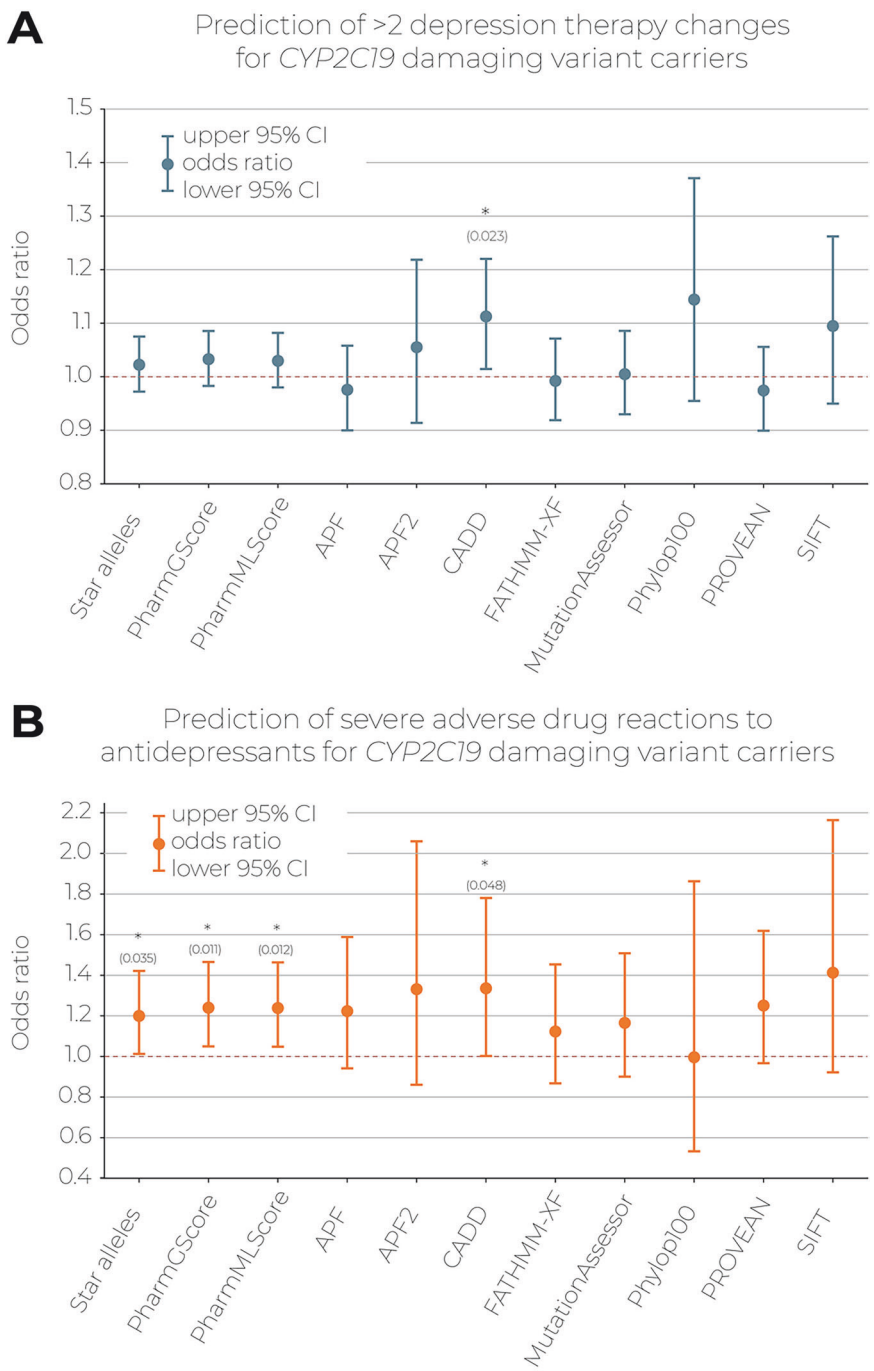

Fig. 5 Impact of predicted damaging *CYP2C19* variants on depression treatment outcomes in 200 642 UKB exome-sequenced participants. Participants were stratified by each computational tool into those with no damaging variants vs. those with damaging variants in *CYP2C19*. **A** Treatment changes. Blue dots represent the odds ratio (OR) of experiencing >2 antidepressant switches in carriers vs. non-carriers, with whiskers indicating 95% CI. **B** Severe ADR. Orange dots depict the OR for at least one severe ADR; whiskers show 95% CI. The dashed line indicates no association (OR = 1) between damaged *CYPC19* and treatment outcome. $P < 0.05$ values are marked with "*"; exact p-values are shown in brackets.

relationship between input scores and star allele function while mitigating overfitting on limited datasets, a known limitation for variant effect prediction development [37]. This neural network-based model evaluates entire haplotypes within each gene. Although it did not outperform other tools on the PharmVar star allele dataset, it demonstrated promising performance in real-world data analyses. We anticipate that its advantages will become more evident as larger and more diverse training datasets emerge [35]. In particular, neural network approaches are well suited for modeling substrate-specific outcomes, which are increasingly recognized as major challenges in pharmacogenomics [38]. Both frameworks are designed to support new genes, though annotated variants are required to establish gene-specific thresholds.

Earlier work applied a transfer learning approach to *CYP2D6* haplotypes, achieving notable success in predicting uncurated star alleles [16]. In contrast, other tools, both general (e.g., CADD) [25] and pharmacogene-specific (e.g., APF, APF2) [19, 20], are not

designed to evaluate multiple variants. An additive burden model simplifies assessing compound heterozygous effects, however benign variants should be assigned very low values to minimize their impact on the total burden. In this study, we demonstrate that, for many scores, this additive burden assumption largely recapitulates star allele-based assignments.

Most previous studies have evaluated the classification of benign versus damaging variants [16, 19, 20]. Here, we also assessed the ability of each tool to score variants with intermediate and increased function, revealing reduced performance of in silico predictors for these categories. To address this limitation, we propose moving beyond the traditional discretized classification into metabolizer classes. For *CYP2C9* and *CYP2C19*, we demonstrated that PharmGScore values effectively correlate with abundance and activity scores at the intermediate levels. However, translating these continuous scores into actionable clinical guidelines remains an ongoing challenge. Another approach may be through improved machine learning models: in a study of 190 pharmacogenetic variants, a random forest classifier distinguished efficiently between no, normal, decreased and increased function variants [17].

In general, all four pharmacogene-dedicated predictors (PharmGScore, PharmMLScore, APF, APF2) and CADD demonstrated strong performance in evaluating damaging variants from deep mutational scanning datasets. Our finding holds significant promise for psychiatric pharmacogenomics, given the critical roles of CYP2C9 and CYP2C19 in antidepressant metabolism [39]. Moreover, this analysis circumvented the issue of circularity between training and evaluation processes [37] by testing the tools on previously unseen variants.

In WES data, PharmMLScore, PharmGScore and APF2 demonstrated the highest TPR for recovering star allele-predicted phenotypes. We acknowledge that WES misses certain pharmacogenetic variants, such as the common *CYP2C19**17 increased function polymorphism. However, WES remains the most prevalent form of clinical sequencing data and among the primary star allele definitions in PharmVar only three variants fall outside a 60 bp window from exons. Furthermore, previous studies [40] have shown that WES captures the majority of pharmacogenetic variation. Moreover, limiting the analysis to exon-adjacent noncoding variants reduces noise from intronic variants [41]. Therefore, we argue that benchmarking computational tools using WES data is currently the most practical approach.

Depression, the most prevalent mental disorder, is a key focus for pharmacogenomic research as individual variability in response to antidepressants has a significant genetic basis [42, 43]. *CYP2C19* poor metabolizers exhibit altered antidepressant response rates and an increased risk of side effects [33, 44]. In a study of 2 087 patients, significant variability in serum escitalopram concentrations persisted within the *1/*1 group [45]. We also identified a potentially damaging *CYP2C19* missense variant (minor allele frequency ~0.000075), not reported in curated databases, highlighting the need for computational tools in variant analysis. In our antidepressant switching analysis, star allele-based calls failed to yield positive predictions, likely due to phenotypic noise. In contrast, the CADD-based approach achieved nominal significance, possibly due to the detection of damaging missense variants. Four approaches, star alleles, PharmGScore, PharmMLScore and CADD, reached significance in predicting diagnostic codes associated with antidepressant toxicity. Importantly, the estimated ORs for star alleles, PharmGScore and PharmMLScore remained nearly identical where ADR codes were restricted only to accidental poisonings, confirming the potential role of computational-based predictors in assessing increased risk of severe ADRs.

Advancing functional prediction methods requires a deeper understanding of the biological reasons why pharmacogenes remain challenging to interpret. Drug-response variants differ evolutionarily and functionally from disease-causing variants, as they are less constrained by purifying selection [46, 47]. Our detailed analysis suggests that predictors capable of accurately identifying functionally relevant variants in less conserved regions tend to achieve the best overall results in pharmacogenomics.

Several limitations of this work must be acknowledged. First, our analysis excluded structural variants, making increased-function variants of *CYP2D6* undetectable. Although known no-function *CYP2D6* copy number variants were evaluated, as they are associated with damaging SNVs [16], a systematic analysis of structural variation is still needed, particularly for *CYP2D6*. Tools such as PyPGx [46], Stargazer [47], and StellarPGx [48] can efficiently detect structural variants and are a promising direction in further development of computational tools for even more comprehensive variant effect prediction. Second, our evaluation did not account for substrate-specific differences when determining phenotype scores, which is a limitation for the current star allele system as well. Third, the small number of known star alleles for certain genes, such as *NUDT15, CYP3A5*, and *SLCO1B1* limits the interpretation of the results. Finally, beyond variant scoring, the integration of non-genetic factors, such as patient lifestyle and comorbidities, drug properties and interactions, and environmental and demographic factors, will be needed to improve clinical relevance.

Others have proposed an iterative pipeline integrating classical pharmacogenetic variant calling with computational tools [21]. Here, we provide the first evidence that, with further development, computational predictors can outperform star allele-based approaches by matching their quality and being superior in flexibility and inclusion of rare damaging variants.

DATA AVAILABILITY

Pre-computed PharmGScore tables for all possible SNVs in human exons ±60 bp are available at https://doi.org/10.5281/zenodo.15926746. Star-allele definitions were downloaded from PharmVar v6.0.10 [27]. Other scores were downloaded from: dbNSFP (https://sites.google.com/site/jpopgen/dbNSFP), CADD (https://cadd.gs.washington.edu), and FATHMM-XF (https://fathmm.biocompute.org.uk/fathmm-xf/). CYP2C9/CYP2C19 activity and abundance scan data are archived on MaveDB (urn:mavedb:00000062-a, urn:mavedb:00000062-b, urn:mavedb:00000095-a). The UKB data used in this research is available to approved researchers and has been accessed through the application 62979. All analyses were performed using Hail (v. 0.2.128), a Python-based (v. 3.10) library for working with genomic data (https://hail.is/), and the PyTorch (v. 2.2.1) library for machine learning methods.

CODE AVAILABILITY

All the projects' code is available on GitHub (https://github.com/ippas/ifpan-pharmgscore-manuscript).

REFERENCES

1. Lauschke VM, Ingelman-Sundberg M. Emerging strategies to bridge the gap between pharmacogenomic research and its clinical implementation. Npj Genomic Med [Internet]. 2020 Mar 5 [cited 2025 July 8];5(1). Available from: https://www.nature.com/articles/s41525-020-0119-2
2. Howes OD, Thase ME, Pillinger T. Treatment resistance in psychiatry: state of the art and new directions. Mol Psychiatry. 2022;27(1):58–72.
3. Pirmohamed M. Pharmacogenomics: current status and future perspectives. Nat Rev Genet. 2023;24(6):350–62.
4. Casajús A, Zubiaur P, Méndez M, Campodónico D, Gómez A, Navares-Gómez M, et al. Genotype-Guided prescription of azathioprine reduces the incidence of adverse drug reactions in TPMT intermediate metabolizers to a similar incidence as normal metabolizers. Adv Ther. 2022;39(4):1743–53.
5. Laatikainen O, Sneck S, Bloigu R, Lahtinen M, Lauri T, Turpeinen M. Hospitalizations due to adverse drug events in the elderly—a retrospective register study. Front Pharmacol [Internet]. 2016 Oct 5 [cited 2025 July 9];7. Available from: http://journal.frontiersin.org/article/10.3389/fphar.2016.00358
6. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA. 1998;279(15):1200.

7. Swen JJ, Van Der Wouden CH, Manson LE, Abdullah-Koolmees H, Blagec K, Blagus T, et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. The Lancet. 2023;401(10374):347–56.
8. Wang X, Tang B, Zhou M, Liu L, Feng X, Wang X, et al. Efficacy and safety of genotype-guided warfarin dosing versus non-genotype-guided warfarin dosing strategies: A systematic review and meta-analysis of 27 randomized controlled trials. Thromb Res. 2022;210:42–52.
9. Ingelman-Sundberg M, Mkrtchian S, Zhou Y, Lauschke VM. Integrating rare genetic variants into pharmacogenetic drug response predictions. Hum Genomics [Internet]. 2018 Dec [cited 2025 July 9];12(1). Available from: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-018-0157-3
10. Kozyra M, Ingelman-Sundberg M, Lauschke VM. Rare genetic variants in cellular transporters, metabolic enzymes, and nuclear receptors can be important determinants of interindividual differences in drug response. Genet Med. 2017;19(1):20–9.
11. Wright GEB, Carleton B, Hayden MR, Ross CJD. The global spectrum of protein-coding pharmacogenomic diversity. Pharmacogenomics J. 2018;18(1):187–95.
12. Zhou Y, Lauschke VM. The genetic landscape of major drug metabolizing cytochrome P450 genes—an updated analysis of population-scale sequencing data. Pharmacogenomics J. 2022;22(5–6):284–93.
13. 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. 2015;17(5):405–24.
14. Li B, Seligman C, Thusberg J, Miller JL, Auer J, Whirl-Carrillo M, et al. In silico comparative characterization of pharmacogenomic missense variants. BMC Genomics [Internet]. 2014 May [cited 2025 July 9];15(S4). Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-S4-S4
15. Zhou Y, Lauschke VM. Computational tools to assess the functional consequences of rare and noncoding pharmacogenetic variability. Clin Pharmacol Ther. 2021;110(3):626–36.
16. McInnes G, Dalton R, Sangkuhl K, Whirl-Carrillo M, Lee SB, Tsao PS, et al. Transfer learning enables prediction of CYP2D6 haplotype function. Schlessinger A, editor. PLOS Comput Biol. 2020;16(11):e1008399.
17. Pandi MT, Koromina M, Tsafaridis I, Patsilinakos S, Christoforou E, Van Der Spek PJ, et al. A novel machine learning-based approach for the computational functional assessment of pharmacogenomic variants. Hum Genomics [Internet]. 2021 Dec [cited 2025 July 9];15(1). Available from: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-021-00352-1
18. Scheinfeldt LB, Brangan A, Kusic DM, Kumar S, Gharani N. Common treatment, common variant: Evolutionary prediction of functional pharmacogenomic variants. J Pers Med. 2021;11(2):131.
19. Zhou Y, Mkrtchian S, Kumondai M, Hiratsuka M, Lauschke VM. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. Pharmacogenomics J. 2019;19(2):115–26.
20. Zhou Y, Pirmann S, Lauschke VM. APF2: an improved ensemble method for pharmacogenomic variant effect prediction. Pharmacogenomics J [Internet]. 2024 June [cited 2025 July 9];24(3). Available from: https://www.nature.com/articles/s41397-024-00338-x
21. Tremmel R, Pirmann S, Zhou Y, Lauschke VM. Translating pharmacogenomic sequencing data into drug response predictions—How to interpret variants of unknown significance. Br J Clin Pharmacol. 2025;91(2):252–63.
22. Ji Y, Shaaban S. Interrogating pharmacogenetics using next-generation sequencing. J Appl Lab Med. 2024;9(1):50–60.
23. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Res. 2025;53(D1):D948–57.
24. 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 [Internet]. 2020. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00803-9 Dec [cited 2025 July 9];12(1).
25. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886–94.
26. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Hancock J, editor. Bioinformatics. 2018;34(3):511–3.
27. Gaedigk A, Casey ST, Whirl-Carrillo M, Miller NA, Klein TE. Pharmacogene variation consortium: A global resource and repository for pharmacogene variation. Clin Pharmacol Ther. 2021;110(3):542–5.
28. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. Am J Hum Genet. 2021;108:1735–51.
29. Boyle GE, Sitko KA, Galloway JG, Haddox HK, Bianchi AH, Dixon A, et al. Deep mutational scanning of CYP2C19 in human cells reveals a substrate specificity-abundance tradeoff. Landry C, editor. GENETICS [Internet]. 2024 Nov 6 [cited 2025 July 9];228(3). Available from: https://academic.oup.com/genetics/article/doi/10.1093/genetics/iyae156/7774808
30. Tafazoli A, Van Der Lee M, Swen JJ, Zeller A, Wawrusiewicz-Kurylonek N, Mei H, et al. Development of an extensive workflow for comprehensive clinical pharmacogenomic profiling: lessons from a pilot study on 100 whole exome sequencing data. Pharmacogenomics J. 2022;22(5–6):276–83.
31. McInnes G, Lavertu A, Sangkuhl K, Klein TE, Whirl-Carrillo M, Altman RB. Pharmacogenetics at Scale: An Analysis of the UK Biobank. Clin Pharmacol Ther. 2021;109(6):1528–37.
32. Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, et al. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2021;110(3):563–72.
33. Kamp M, Lo CWH, Kokkinidis G, Chauhan M, Gillett AC, AMBER Research Team, et al. Sociodemographic, clinical, and genetic factors associated with self-reported antidepressant response outcomes in the UK Biobank. Psychol Med [Internet]. 2025 [cited 2025 July 9];55. Available from: https://www.cambridge.org/core/product/identifier/S0033291725000388/type/journal_article
34. Wong WLE, Fabbri C, Laplace B, Li D, Van Westrhenen R, Lewis CM, et al. The effects of CYP2C19 genotype on proxies of SSRI antidepressant response in the UK biobank. Pharmaceuticals. 2023;16(9):1277.
35. Zhou Y, Tremmel R, Schaeffeler E, Schwab M, Lauschke VM. Challenges and opportunities associated with rare-variant pharmacogenomics. Trends Pharmacol Sci. 2022;43(10):852–65.
36. Shubbar Q, Alchakee A, Issa KW, Adi AJ, Shorbagi AI, Saber-Ayad M From genes to drugs: CYP2C19 and pharmacogenetics in clinical practice. Front Pharmacol [Internet]. 2024 Feb 14 [cited 2025 July 15];15. Available from: https://www.frontiersin.org/articles/10.3389/fphar.2024.1326776/full
37. Livesey BJ, Marsh JA. Interpreting protein variant effects with computational predictors and deep mutational scanning. Dis Model Mech [Internet]. 2022 June 1 [cited 2025 July 13];15(6). Available from: https://journals.biologists.com/dmm/article/15/6/dmm049510/275742/Interpreting-protein-variant-effects-with
38. Tremmel R, Honore A, Park Y, Zhou Y, Xiao M, Lauschke VM. Machine learning models for pharmacogenomic variant effect predictions – recent developments and future frontiers. Pharmacogenomics. 2025;26(5–6):171–82.
39. Porcelli S, Fabbri C, Spina E, Serretti A, De Ronchi D. Genetic polymorphisms of cytochrome P450 enzymes and antidepressant metabolism. Expert Opin Drug Metab Toxicol. 2011;7(9):1101–15.
40. Van Der Lee M, Allard WG, Bollen S, Santen GWE, Ruivenkamp CAL, Hoffer MJV, et al. Repurposing of diagnostic whole exome sequencing data of 1,583 individuals for clinical pharmacogenetics. Clin Pharmacol Ther. 2020;107(3):617–27.
41. Barbosa P, Savisaar R, Carmo-Fonseca M, Fonseca A. Computational prediction of human deep intronic variation. GigaScience [Internet]. 2022 Dec 28 [cited 2025 July 19];12. Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giad085/7329463
42. Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep [Internet]. 2018. Available from: https://www.nature.com/articles/s41598-018-23584-z Apr 3 [cited 2025 July 16];8(1).
43. Tansey KE, Guipponi M, Hu X, Domenici E, Lewis G, Malafosse A, et al. Contribution of common genetic variants to antidepressant response. Biol Psychiatry. 2013;73(7):679–82.
44. Calabrò M, Fabbri C, Kasper S, Zohar J, Souery D, Montgomery S, et al. Metabolizing status of CYP2C19 in response and side effects to medications for depression: Results from a naturalistic study. Eur Neuropsychopharmacol. 2022;56:100–11.
45. Jukić MM, Haslemo T, Molden E, Ingelman-Sundberg M. Impact of *CYP2C19* genotype on escitalopram exposure and therapeutic failure: A retrospective study based on 2087 patients. Am J Psychiatry. 2018;175(5):463–70.
46. Lee SB, Shin JY, Kwon NJ, Kim C, Seo JS. ClinPharmSeq: A targeted sequencing panel for clinical pharmacogenetics implementation. Boon-Peng H, editor PLOS ONE. 2022;17(7):e0272129.
47. Lee S, Wheeler MM, Thummel KE, Nickerson DA. Calling star alleles with stargazer in 28 pharmacogenes with whole genome sequences. Clin Pharmacol Ther. 2019 Dec;106(6):1328–37.
48. Twesigomwe D, Drögemöller BI, Wright GEB, Siddiqui A, Da Rocha J, Lombard Z, et al. StellarPGx: A nextflow pipeline for calling star alleles in cytochrome P450 genes. Clin Pharmacol Ther. 2021;110(3):741–9.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under Application Number 62979. We are grateful to the UK Biobank and all its voluntary participants. This work used data provided by patients and collected by the NHS as part of their care and support.

This study was funded by the National Science Center, Poland: PRELUDIUM BIS-3 grant no. 2021/43/O/NZ7/01187 (development and benchmarking of variant scores) and SONATINA 5 grant 2021/40/C/NZ2/00218 (UKB analyses). Additional support came from the statutory funds of the Maj Institute of Pharmacology PAS. We gratefully acknowledge Poland's high-performance Infrastructure PLGrid ACK Cyfronet AGH, for providing computer facilities and support within computational grant no PLG/2022/015861. DMF and GEB were funded by NIH grants NIH R35GM152106 and UM1HG011969.

AUTHOR CONTRIBUTIONS

MK, MP, MB, IK and JH designed the research. JH, and PK conducted the analyses. GEB and DMF analyzed data from deep mutational scanning. MB and JH prepared the figures and tables, MB drafted the manuscript, JH wrote parts of the manuscript, MK, IK, DMF provided critical revisions.

COMPETING INTERESTS

MP and MK are founders of the bioinformatics company Intelliseq SA. The remaining authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The UK Biobank has obtained approval from the North West Multi-center Research Ethics Committee as a Research Tissue Bank approval (references: 11/NW/0382, 16/NW/0274, and 21/NW/0157). This approval covers all researchers using the UK Biobank resource under an approved application; therefore, separate ethical approval was not required for this study. All participants provided written informed consent to participate in the study.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41397-026-00399-0.

Correspondence and requests for materials should be addressed to Malgorzata Borczyk or Michal Korostynski.