The Pharmacogenomics Journal (2019) 19:115–126
https://doi.org/10.1038/s41397-018-0044-2

ARTICLE

An optimized prediction framework to assess the functional impact of pharmacogenetic variants

Yitian Zhou[1] · Souren Mkrtchian[1] · Masaki Kumondai[2] · Masahiro Hiratsuka[2] · Volker M. Lauschke[1]

Received: 30 October 2017 / Revised: 27 June 2018 / Accepted: 10 August 2018 / Published online: 12 September 2018

Abstract

Prediction of phenotypic consequences of mutations constitutes an important aspect of precision medicine. Current computational tools mostly rely on evolutionary conservation and have been calibrated on variants associated with disease, which poses conceptual problems for assessment of variants in poorly conserved pharmacogenes. Here, we evaluated the performance of 18 current functionality prediction methods leveraging experimental high-quality activity data from 337 variants in genes involved in drug metabolism and transport and found that these models only achieved probabilities of 0.1–50.6% to make informed conclusions. We therefore developed a functionality prediction framework optimized for pharmacogenetic assessments that significantly outperformed current algorithms. Our model achieved 93% for both sensitivity and specificity for both loss-of-function and functionally neutral variants, and we confirmed its superior performance using cross validation analyses. This novel model holds promise to improve the translation of personal genetic information into biological conclusions and pharmacogenetic recommendations, thereby facilitating the implementation of Next-Generation Sequencing data into clinical diagnostics.

Introduction

In the last decades, rapid progress in sequencing technologies has allowed the deciphering of genomic information on an unprecedented scale. While the initial sequencing of the human genome in the frame of the Human Genome Project cost 2.7 billion USD and took 14 years to complete, costs and times declined to around 1200 USD and 1.5 days for a whole-genome sequence with 30× coverage in 2015 [1] and technology to enable the 100 USD genome has already been announced [2]. As outcomes of these technological advancements, the vast extent of genomic information has propelled medicine by providing information about disease susceptibility, e.g. in cancer [3, 4], type 2 diabetes mellitus [5] or schizophrenia [6], by identifying genes that underlie monogenic disorders [7, 8] and by facilitating the discovery of novel therapeutic targets, particularly in oncology [9].

However, despite these successes of human genomics on a population scale, the translation of personal genomic data into clinically actionable information remains difficult. Each individual harbors on average 23,000–25,000 genetic variants in exons, including 10,000–12,000 variants resulting in amino acid exchanges and around 100 variants resulting in stop-gain mutations, frameshifts or differential splice sites, the vast majority of which are rare with minor allele frequencies (MAF) < 1% [10]. Genes with importance for drug absorption, distribution, metabolism and excretion (ADME) are highly variable [11–13] and such genetic variability has been estimated to account for around 20–30% of the inter-individual differences in drug response [14]. However, while on average around 100 genetic variants are detected across ADME genes in each individual, the overwhelming majority has not been experimentally characterized, which poses a significant challenge for the clinical interpretation of genetic variability and impairs the translation of genomic data into actionable advice [15, 16].

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41397-018-0044-2) contains supplementary material, which is available to authorized users.

✉ Volker M. Lauschke
volker.lauschke@ki.se

[1] Department of Physiology and Pharmacology, Section of Pharmacogenetics, Karolinska Institutet, SE-171 77 Stockholm, Sweden

[2] Laboratory of Pharmacotherapy of Life-Style Related Diseases, Graduate School of Pharmaceutical Sciences, Tohoku University, Sendai, Japan

As systematic experimental analyses in relevant expression systems are hitherto not feasible for these vast numbers of variants, computational methods have been proposed for predicting the functional relevance of identified genetic mutations. In recent years, dozens of algorithms have been presented that aim to distinguish deleterious from neutral variants. These algorithms use a variety of features, such as secondary structure, functional sites, protein stability or sequence conservation, and are mostly based on machine learning techniques, such as support vector machines, artificial neural networks or naïve Bayes classifiers [17–19]. Importantly, computational methods are generally trained on sets of variants with high evolutionary constraints implicated in disease. However, as many ADME genes are generally only poorly conserved, we hypothesize that specialized pharmacogenetic prediction models are needed that have been calibrated on appropriate ADME data sets.

In this study, we used experimental activity data from 337 variants distributed across 43 ADME genes to evaluate current functionality prediction methods and found that standard algorithms are only relatively poor predictors of the functional impact of ADME gene mutations. We thus developed a novel computational functionality prediction model optimized for pharmacogenetic assessments, which substantially outperformed standard algorithms, correctly flagging 93% of experimental loss-of-function (LOF) variants as deleterious and 93% of variants without functional impact as neutral. Thus, the ADME-optimized prediction framework significantly improves in silico functionality assessment of pharmacogenetic variants, thereby facilitating the translation of uncharacterized variants into pharmacogenetic recommendations and providing a further step towards the leveraging of Next-Generation Sequencing data for the personalization of pharmacological treatment.

Methods

In vitro functionality data

We obtained experimental functionality data for 337 single variant alleles from the 43 ADME gene (see Supplementary Table 1 for references). The common variants rs3758581 (CYP2C19 I331V), rs16947 (CYP2D6 R296C) and rs1135840 (CYP2D6 S486T) were considered as neutral. An overview of all analyzed variants, the substrates and expression systems used for characterization and the in silico predictions by all tested algorithms is provided in Supplementary Table 2. Wherever necessary, variant coordinates were translated to a uniform reference genome version. Mutations for which no score could be retrieved by any prediction method were excluded. Variants were considered to have a deleterious impact if they reduced their intrinsic clearance more than 2-fold compared to the wild-type allele (for most genes the **1*, in the case of *NAT1* the **4* allele).

Statistical definitions

True positives (TP) and false negatives (FN) are variants that have a functional impact in vitro and are predicted in silico to be deleterious or neutral, respectively. Conversely, true negatives (TN) and false positives (FP) are defined as mutations that do not affect the functionality of the gene in vitro and are predicted in silico to be neutral or deleterious, respectively. The true positive rate or sensitivity is defined as $\frac{\sum TP}{\sum TP+\sum FN}$, specificity is $\frac{\sum TN}{\sum TN+\sum FP}$ and the false positive rate is defined as $\frac{\sum FP}{\sum TN+\sum FP}$. Furthermore, the positive and negative predictive values are calculated as $\frac{\sum TP}{\sum TP+\sum FP}$ and $\frac{\sum TN}{\sum TN+\sum FN}$, respectively and the total predictive accuracy is $\frac{\sum TP+\sum TN}{\sum TP+\sum TN\sum FP+\sum FN}$.

Computational functionality predictions

We compared the functionality assessments of 18 current in silico functionality prediction algorithms, conservation scores and ensemble scores computed using ANNOVAR: [20] SIFT [21], PolyPhen-2 [22], Likelihood ratio tests [23], MutationAssessor [24], FATHMM [25], FATHMM-MKL [26], PROVEAN [27], VEST3 [28], CADD [29], DANN [30], MetaSVM [31], MetaLR [31], GERP++ [32], SiPhy [33], PhyloP [34] (using both vertebrate and mammalian alignments) and PhastCons [35] (using both vertebrate and mammalian alignments).

Development of ADME optimized algorithm

The 337 alleles were randomly partitioned into five subsets for 5-fold cross validations while assuring equal proportions of deleterious and neutral variants (Fig. 1). Thresholds for the individual algorithms were optimized on the basis of the Youden index or informedness function, which can be interpreted as the probability of an informed classification. The Youden index, defined as $I=\text{sensitivity}+\text{specificity}-1$, was calculated for each potential threshold (increments 0.01–0.05) between the highest and lowest possible scores for each respective method. All variants i were classified as deleterious or neutral by each of the k threshold-optimized algorithms. If the computational prediction for var_i aligns with the corresponding experimental result, then score $\text{s}_{\text{k,i}}$

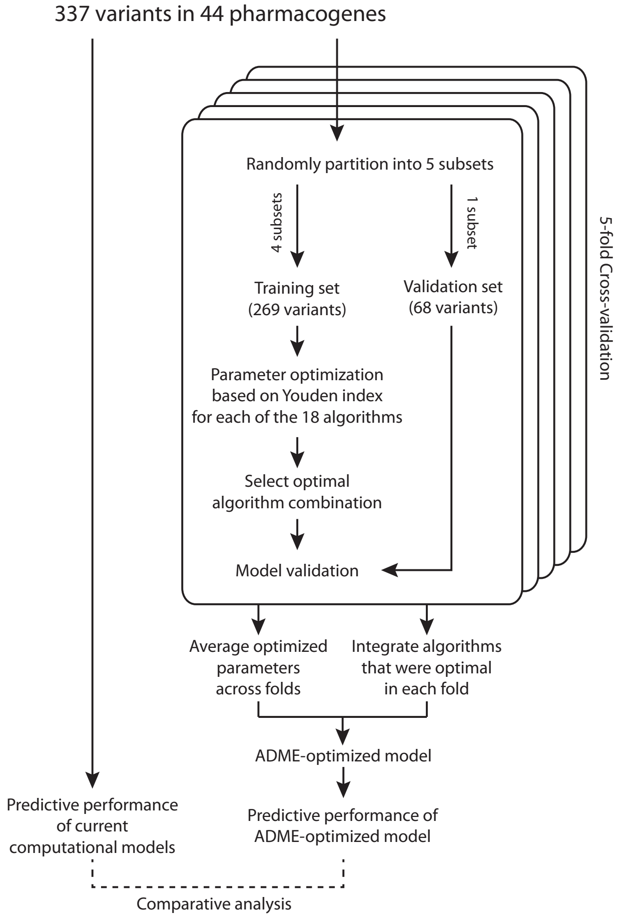


Fig. 1 Schematic depiction of the workflow for the development of the ADME optimized prediction model

$= 1$ otherwise $s_{k,i} = 0$. Subsequently, out of all possible constellations the algorithm combination was selected for the ADME-optimized model for which $\sum_i \sum_l s_{l,i} = max$ with $1 \leq k$. Importantly, the result with this model was validated for each fold using the independent validation set. Overall, optimal results for the pharmacogenetic prediction model were derived by integrating assessments of LRT, MutationAssessor, PROVEAN, VEST3 and CADD. The overall prediction score of the ADME -optimized model is defined as follows: each of the algorithms predicts whether a variant is deleterious or neutral based on its ADME-optimized threshold value (1 = deleterious and 0 = functionally neutral). The final score is derived by averaging the assessments of the individual algorithms (1 or 0). Thus, a score of 1 indicates that all algorithms predicted the variant to be deleterious, a score of 0 that all algorithms predicted the variant to be neutral and a score of e.g. 0.5 that half of the algorithms predicted the variant to be deleterious and half to be neutral. Receiver operating characteristics (ROC) analyses were performed using Prism 6 (GraphPad Software Inc.).

Results

Conventional computational algorithms have a low predictive accuracy when applied to pharmacogenetic variants

We first evaluated the performance of current computational functionality assessment algorithms on pharmacogenetic variants across 43 ADME genes with low evolutionary constraints (Supplementary Table 3). To this end, we derived predictions for 337 pharmacogenetic single nucleotide variants (SNVs) with available high-quality experimental data. These variants cause alterations in the amino acid sequence of their corresponding gene product, which can either cause direct modulation of protein activity, result in changes in protein levels, for instance due to misfolding followed by degradation or entail dysregulation of protein transport. We evaluated eight commonly used functionality prediction algorithms, SIFT, PolyPhen-2, LRT, MutationAssessor, FATHMM, FATHMM-MKL, PROVEAN and VEST3 (Fig. 2a). When using the area under the ROC curve (AUC_{ROC}) as measure for model quality, VEST3, MutationAssessor and PolyPhen-2 exhibited the best performance with AUC_{ROC} values of 0.8, 0.78 and 0.77, respectively, whereas FATHMM performed worst ($AUC_{ROC} = 0.51$; Table 1).

Next, we tested the performance of four models, GERP++, SiPhy, PhyloP and PhastCons using different phylogenetic models (using 7 vertebrates or 20 mammals), resulting in a total of six sores that use evolutionary conservation based on sequence alignments as a measure for functional importance (Fig. 2b). Overall, the predictive power of evolutionary conservation scores ($AUC_{ROC} = 0.58$–0.67) was substantially lower than that of functionality prediction algorithms which base their assessment also on additional features, such as homology alignments or structure-based features ($AUC_{ROC} = 0.51$–0.8; Table 1). These findings suggest that evolutionary conservation alone seems to be a poor indicator of functional impact in poorly conserved loci, such as ADME genes.

We furthermore analyzed the ensemble scores CADD, DANN, MetaSVM and MetaLR that integrate assessments from multiple orthologous methods (Fig. 2c). CADD and DANN performed substantially better than MetaSVM and MetaLR on our data set with the former showing the best predictive performance of all models analyzed ($AUC_{ROC} = 0.81$; Table 1). Importantly, the predictive power of most algorithms on our ADME variant cohort was substantially lower compared to data sets based on pathogenicity-associated variants (Table 1), emphasizing the shortcomings of model parameterization based on genome-wide analyses for pharmacogenetic functionality predictions.

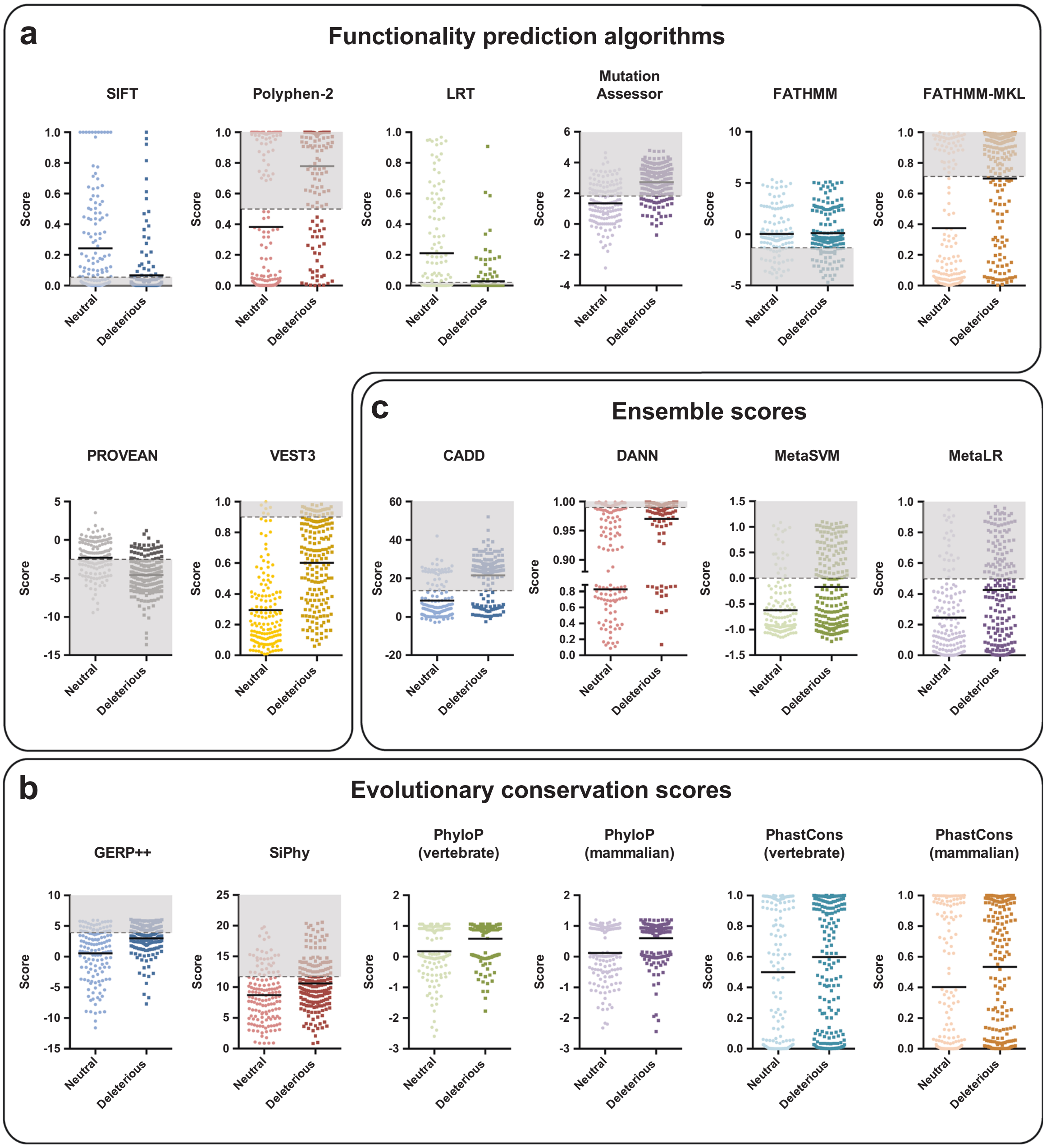


Fig. 2 Overview of the performance of different functionality prediction methods. Variants ($n = 337$) were separated into phenotypically neutral variants (lighter shaded circles) and those that have a relevant impact on substrate metabolism (intrinsic clearance reduced >2-fold; darker shaded squares). **a**–**c** Functionality was predicted using eight common prediction algorithms (**a**), 6 evolutionary conservation scores (**b**) and 4 ensemble scores (**c**). Conventional thresholds of the respective algorithms are depicted as dashed lines and intervals of functionality scores deemed functional are shaded in light gray. The average scores of variants in the neutral and deleterious groups are indicated

Optimization of pharmacogenetic functionality predictions

To improve the predictive power of pharmacogenetic functionality predictions, we structured the problem into two tasks: first, we optimized the classification thresholds of the individual algorithms and, in a second step, we selected the optimal combination of model components.

We decided to optimize parameterization of the algorithms based on the concept of overall informedness, defined as the probability that a prediction is informed (i.e. not by chance) using the Youden index as statistical target metric (see Supplementary Figure 1 for graphical depiction and further explanation). The Youden index J developed as a measure to rate diagnostic tests [36], is defined on the basis of a ROC curve as $J = max_x\{sens(x) + spec(x) - 1\}$

Table 1 Comparison of the predictive performance of functionality prediction tools on pathogenic and pharmacogenetic data sets

Algorithm	Category	Performance on disease-associated data set (AUC_{ROC})	Performance on pharmacogenetic data set (AUC_{ROC})
SIFT	Functionality prediction algorithms	0.76–0.88	0.74
PolyPhen-2		0.79–0.88	0.77
LRT		0.67–0.72	0.75
MutationAssessor		0.8–0.83	0.78
FATHMM		0.87–0.91	0.51
FATHMM-MKL		0.91	0.73
PROVEAN		0.85	0.76
VEST3		0.91	0.8
GERP++	Evolutionary conservation scores	0.67–0.78	0.67
SiPhy		0.69–0.81	0.63
PhyloP (vertebrate)		0.67–0.83	0.64
PhyloP (mammalian)			0.64
PhastCons (vertebrate)		0.67–0.83	0.58
PhastCons (mammalian)			0.61
CADD	Ensemble scores	0.93	0.81
DANN		0.95	0.75
MetaSVM		0.88–0.89	0.68
MetaLR		0.92–0.94	0.68

Performance measures on disease-associated data sets were obtained from refs. [26–31]

across all potential threshold scores *x*. The point x for which the sum of sensitivity and specificity is maximal indicates the optimal threshold value that maximizes the capacity of the test to differentiate between deleterious and neutral variants when sensitivity and specificity are weighted equally, thus avoiding impacts of the unequal distribution of neutral and functionally deleterious variants in our data set [37]. We defined the optimal threshold value for each algorithm or score based on the global maximum of the informedness graph (Fig. 3a). Interestingly, shapes of the informedness functions differed substantially between algorithms. While some algorithms, such as PolyPhen-2 and FATHMM-MKL showed largely stable informedness values across a wide range of threshold scores, others, such as SIFT or PROVEAN, exhibited sharp peaks, indicating drastic differences in the robustness of the method to variation in threshold scores.

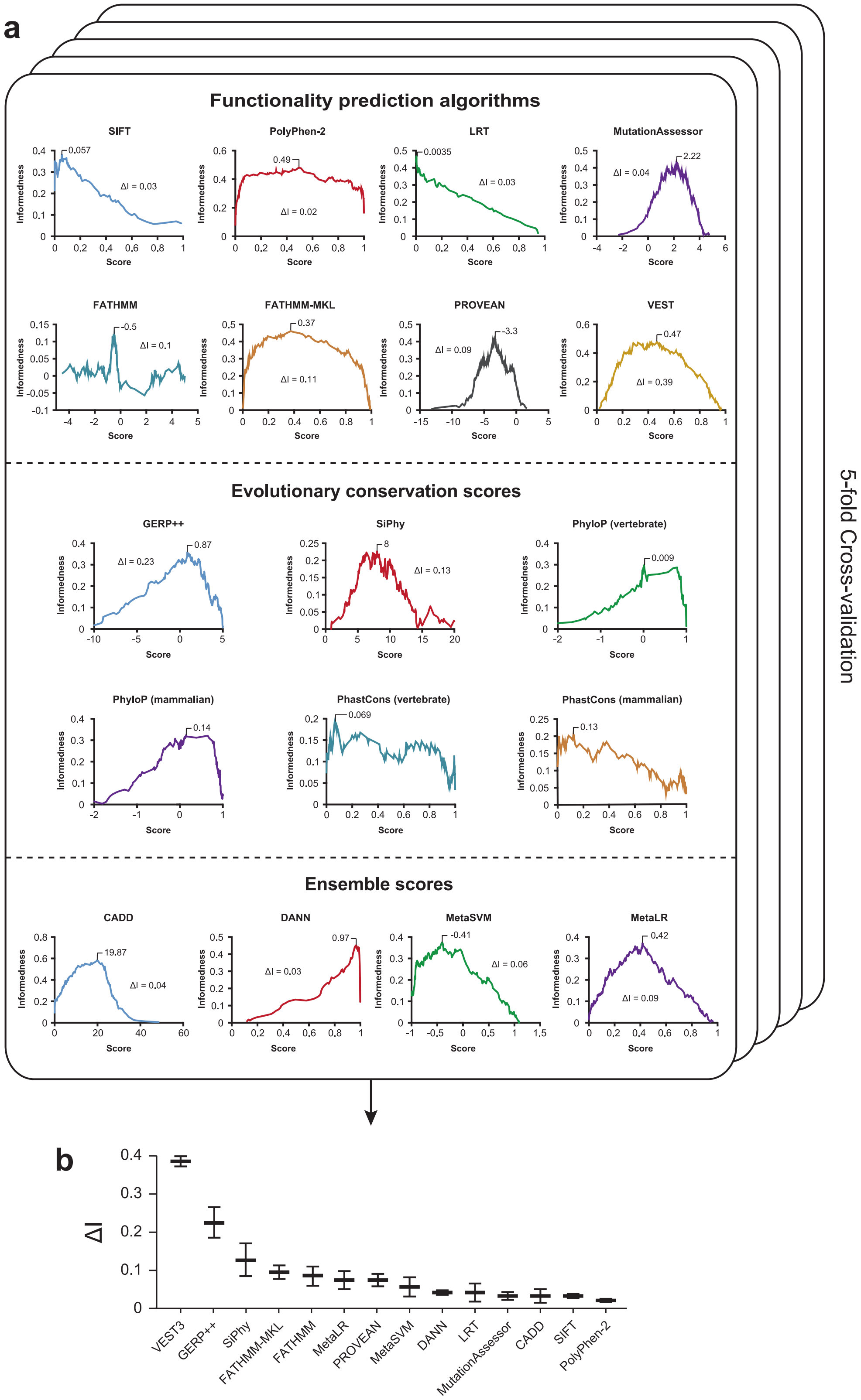
To evaluate the sensitivity of this approach to variability in training set variants we performed 5-fold cross validations in which we partitioned the variants into five equally sized subsets. Of these five subsets, four are used for model training and one is used for independent validation. This process is iterated five times with each of the five subsamples serving once as validation data set. For most algorithms, including PROVEAN (|coefficient of variation| = 0.006), DANN (|CV| = 0.012) and VEST3 (|CV| = 0.054), the optimal threshold differed only marginally between folds, demonstrating the robustness of the threshold optimization (Supplementary Table 4). In contrast, optimal thresholds were substantially different across folds for PhyloP (|CV| = 3.12) and FATHMM (|CV|=2.33). Interestingly, the added value of threshold optimization differed substantially across prediction tools (Fig. 3b and Table 2). While threshold optimization did only marginally improve the informedness of PolyPhen-2, SIFT or CADD ($\Delta I < 0.03$), the performance of other algorithms, such as SiPhy ($\Delta I = 0.13$), GERP++($\Delta I = 0.22$) and VEST3 ($\Delta I = 0.38$), were highly improved.

When integrating the individual predictions for each variant into a consensus decision by averaging the ADME-optimized thresholds across folds, the resulting model achieved 82% sensitivity and 62% specificity. To improve this predictive accuracy, we evaluated the predictive performance for all possible combinations of threshold-optimized algorithms. Importantly, optimal model constituents were highly similar between folds (Supplementary Table 4) and, based on these findings, we integrated the LRT, MutationAssessor, PROVEAN, VEST3 and CADD using ADME-optimized parameters (Table 2) into our pharmacogenetic prediction framework.

Performance of ADME optimized prediction framework

In the training data sets the ADME optimized prediction framework achieved overall sensitivity and specificity of 80 ± 2% S.D. and 80 ± 3% S.D., respectively, thus outperforming all previously reported functionality prediction algorithms, conservation or ensemble scores. This superior performance of the ADME optimized model was validated using the independent variants from each training set, achieving sensitivity and specificity of 79 ± 10% S.D. and 81 ± 11% S.D., respectively (Fig. 4).

Importantly, when analyzing all 337 pharmacogenetic variants using the developed ADME-optimized prediction framework, we found that the score of the ADME-optimized prediction model correlates well with the extent of functional impact of the variant in question ($R^2 = 0.9$, $p = 2.9 \times 10^{-5}$; Fig. 5a). For LOF variants (<10% activity of WT) the model yielded scores of 0.84 ± 0.02 s.e.m., which continuously decreased with increased functionality in vitro up to 0.19 ± 0.02 s.e.m. for functionally neutral variants (>90% activity of WT). When translating these scores into dichotomous functionality predictions, the model achieved



93% sensitivity (101/109 variants) for LOF variants that decreased activity >10-fold whereas variants with only mild functional effects were recognized with 55–70% sensitivity (Fig. 5b). Conversely, prediction specificity for variants that exhibited >90% of the functional activity of the WT allele, was 93% (66/71 variants), whereas the specificity for

Fig. 3 Pharmacogenetic threshold optimization results in substantially higher probabilities to make informed decisions. **a** The degree of informedness is plotted as a function of threshold score for eight functionality prediction algorithms, six evolutionalry conservation scores and four ensemble scores. The threshold score corresponding to the global maximum of informedness is indicated. ΔI denotes the gain in informedness between using the pharmacogenetically optimized threshold and the conventional threshold provided in the literature. Results are depicted for one of the five folds in our cross-validation analysis. **b** Averaging the ΔI values of the five folds demonstrates that the increases in informedness due to ADME-specific parameterization differ substantially between algorithms and are stable across folds. As no standard thresholds for PhyloP and PhastCons are provided in the literature, no ΔI values for these conservation scores are shown. Error bars indicate S.D.

Table 2 Overview of computational method parameters to assess the functionality of pharmacogenetic variants

		Conventional			ADME optimized		
Algorithm	Category	Threshold	Sensitivity (%)	Specificity (%)	Threshold	Sensitivity (%)	Specificity (%)
SIFT	Functionality prediction algorithms	<0.05	80.7	54.2	<0.0376	75.6	57.6
PolyPhen-2		>0.447	80.8	63	>0.3841	83	61.6
LRT		<0.001	66.3	72.3	<0.0025	77.3	65.2
MutationAssessor		>1.9	79	63.7	>2.0566	74	67.8
FATHMM		<−1.5	18.2	81.9	<0.486	69.9	27.1
FATHMM-MKL		>0.73	64.2	68	>0.3982	77.4	63.3
PROVEAN		<−2.5	80.7	56.9	<−3.286	72.2	72.2
VEST3		>0.9	14.3	95.9	>0.4534	67.6	78.8
GERP++	Evolutionary conservation scores	>4.4	28.4	84.4	>1.2482	84.2	47.6
SiPhy		>12.17	32.1	78.2	>7.2442	51.9	72.7
PhyloP (vertebrate)		NA	NA	NA	>0.5216	70.5	53.7
PhyloP (mammalian)		NA	NA	NA	>0.0461	77.4	49
PhastCons (vertebrate)		NA	NA	NA	>0.07	81.1	34.7
PhastCons (mammalian)		NA	NA	NA	>0.1872	67.4	49.7
CADD	Ensemble scores	>15	75.8	74.8	>19.19	74.2	78.9
DANN		>0.99	68.9	70.1	>0.9688	85.8	54.4
MetaSVM		>0	43.4	86.3	>−0.3371	51.6	78.1
MetaLR		>0.5	41.2	84.2	>0.4039	52.2	76.7

Sensitivity and specificity of each prediction method is shown for conventional disease dataset-based parameterization and ADME optimized parameters. Threshold values are in arbitrary units, values for sensitivity and specificity are provided in percentage (%)

variants with 50–100% activity was only 56–82% . Overall, these performance metrics resulted in a predictive accuracy of 93% for LOF and functionally neutral variants, compared to 84% for CADD, the score with the next highest accuracy.

Overall, the ADME optimized model achieved the highest extent of informedness for LOF and neutral variants ($I_{\mathrm{ADME}} = 0.86$), followed by CADD ($I_{\mathrm{CADD}} = 0.65$) and LRT ($I_{\mathrm{LRT}} = 0.63$; Fig. 5c). Similarly, when all variants are considered and classified dichotomously, the ADME model substantially outperformed current models ($I_{\mathrm{ADME}} = 0.6$ followed by $I_{\mathrm{CADD}} = 0.51$). In contrast, VEST3 and FATHMM only yielded overall values of $I_{\mathrm{VEST}} = 0.11$ and $I_{\mathrm{FATHMM}} = 0.01$, respectively. Besides the increased predictive power, the integrated ADME model successfully derived assessments for all variants, while some individual algorithms were unable to predict the functional impact of up to 5% of all variants analyzed (Fig. 5d).

Lastly, we analyzed whether the predictive performance of the ADME optimized prediction model depended on the frequency of the respective variant. The majority of the 337 variants analyzed in this study were rare ($n = 285$) or very rare ($n = 232$) with MAF < 1% or MAF < 0.1%, respectively. Notably, the predictive power of the model for LOF and functionally neutral variants was better for very rare ($I_{\mathrm{MAF}<1\%} = 0.87$) and rare mutations ($I_{0.1\% \leq \mathrm{MAF} < 1\%} = 1$) compared to common variants ($I_{\mathrm{MAF} \geq 1\%} = 0.45$; Fig. 5e). Similar trends were observed when all variants were considered either in our model (Fig. 5f) or in individually tested algorithms (Supplementary Figure 2). While our results correlated significantly with data from REVEL ($R^2 = 0.5$; Supplementary Figure 3), a prediction method to analyze the pathogenicity of rare missense variants [38], the ADME optimized prediction framework performed substantially better for the prediction of pharmacogenetic variants: When using the threshold score that resulted in the best Youden

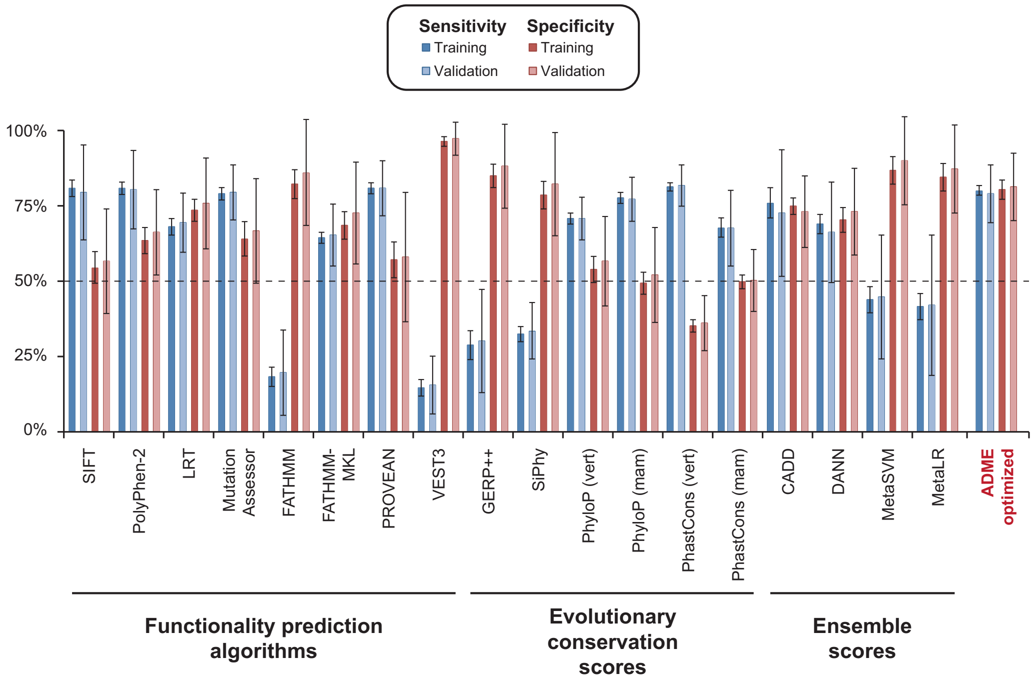


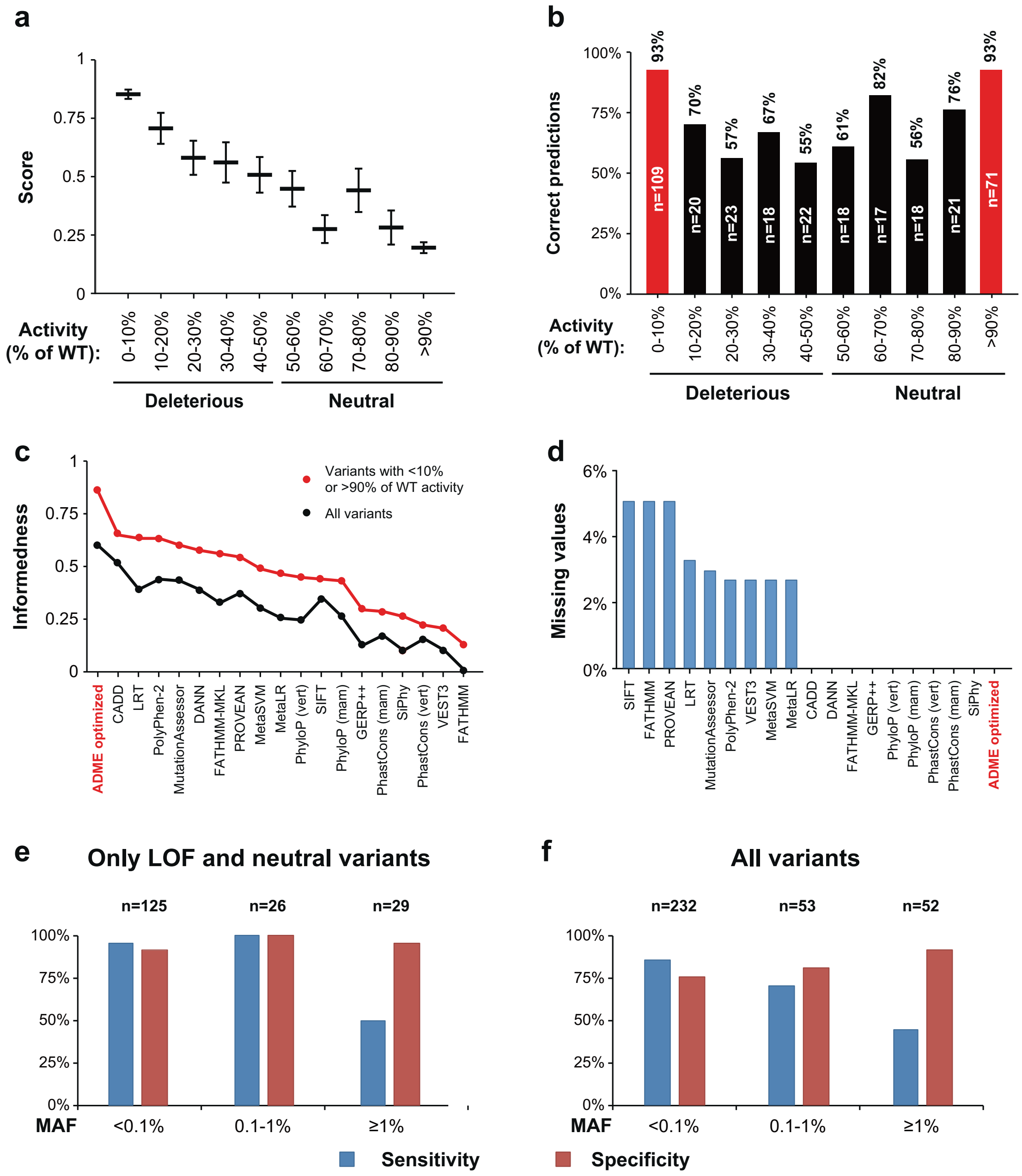
Fig. 4 The ADME optimized model outperforms conventional methods for the functionality prediction of pharmacogenetic variants. Column plot showing the sensitivity (shades of blue) and specificity (shades of red) of commonly used functionality prediction algorithms, ensemble scores and evolutionary conversation scores as well as of the ADME optimized prediction model presented here. Notably, the ADME optimized model was the only method achieving both sensitivity and specificity of >80% in both training and validation data set. Error bars indicate S.D. across folds

index for disease associated variants (0.5), REVEL achieved informedness values of 0.36 and 0.61 on our pharmacogenetic data set when considering all or only LOF and functionally neutral variants, respectively. In contrast, on the same variants the ADME-optimized model achieved informedness levels of 0.6 and 0.86, respectively (Supplementary Table 5). These findings emphasize the usefulness of the ADME optimized prediction model for the functional interpretation of pharmacogenetic variants with low frequencies, which, due to their large numbers, are difficult to systematically characterize in vitro.

Discussion

Despite the abundance of genomic data generated in the frame of multiple completed and ongoing population-scale sequencing projects, the understanding of personal genomic data and translation into clinically actionable information is still very limited. Functional interpretation of identified mutations relies either on clinical or experimental data, which is only available for a small subset of well-characterized genetic variants, or on computational prediction tools. A vast number of algorithms and scores have been presented that predict the likelihood of whether a genetic variant has a functional impact based on sequence homology, structural features, preexisting annotations or, most importantly, evolutionary constraints [39] and these tools have been reasonably successful in predicting mutations associated with disease [40, 41]. However, the predictive quality of these algorithms on specific classes of genes with lower evolutionary constraints that are often not directly disease-associated, has not been evaluated.

Here, we benchmarked 18 commonly used prediction methods on a pharmacogenetic data set encompassing 337 variants with available high-quality experimental characterization data using functional assays, which have been suggested as gold standard sets for the benchmarking of computational tools [42]. We focused on pharmacokinetic genes involved in drug metabolism and transport as these can be genetically highly polymorphic and are subject to low evolutionary constraints. In contrast, drug targets are highly heterogeneous regarding their evolutionary conservation and are commonly associated with congenital diseases [43]. Importantly, we found that performance of tested algorithms on this ADME data set was substantially lower than on data sets comprising of pathogenic variants (Table 1). Of the different methods tested, evolutionary conservation scores exhibited overall the worst performance, supporting our hypothesis that selective constraints are unreliable measures for assessing the functionality of variants in genes with low evolutionary pressure, such as



ADME genes [44]. Given that most algorithms rely on evolutionary conservation as a core feature, these findings suggest that ADME gene-specific parameter optimization and integration of orthogonal approaches represent an appealing rationale to improve the pharmacogenetic predictions.

After optimization our model significantly outperformed all individual functionality prediction methods achieving a predictive accuracy of 93% for LOF and functionally neutral variants, compared to 84% for CADD, the second best algorithm. Interestingly, we achieved the best overall performance not by integrating the individually best performing algorithms. For instance, LRT ranked only as 5 with an accuracy of 81.8% but the LRT score was integrated into the most predictive ADME model. This finding is in agreement with the performance of the model on human disease alleles for which the overlap between LRT and other methods has been shown to be low [23].

Deviations between in vitro data and in silico predictions can be allotted to both computational and experimental

Fig. 5 The ADME optimized model provides quantitative estimates of functional variant effects. **a** The score provided by the ADME optimized prediction model correlates quantitatively with the level of gene product functionality determined experimentally in vitro ($R^2 = 0.9$, $p = 2.9 \times 10^{-5}$). Highest scores are provided for LOF variants with <10% of WT functionality (0.84 ± 0.02 s.e.m.), while variants that do not affect gene product functionality receive lowest scores (0.19 ± 0.02 s.e.m.). Data is plotted as mean ± s.e.m. **b** 93% of variants that resulted in severely decreased functionality in vitro (<10% activity of WT) were correctly classified as deleterious, whereas variants whose effect on functionality was only moderate (decreased functionality variants; 10–50% activity of WT), were flagged with lower probabilities. Similarly, variants that showed equivalent activity than WT (>90%) were more likely to be flagged as functionally neutral (93% specificity) than variants with 50–90% of activity. **c** Levels of informedness are shown for all variants (black) and variants with <10% and >90% of WT activity (red curves corresponding to red columns in **b**). Note that the ADME optimized prediction framework achieved the highest values of informedness, irrespective of which variants were considered. **d** Overview of the fraction of variants for which no prediction could be obtained by the individual algorithms. While SIFT, FATHMM and PROVEAN did not return predictions for 5% of variants, CADD, DANN, SiPhy, PhastCons, PhyloP, GERP and the ADME optimized model provided assessments for all non-synonymous variants analyzed here. **e**, **f** Column plot depicting sensitivity and specificity of the ADME optimized prediction model for LOF and functionally neutral variants (**e**) or all variants (**f**) depending on their minor allele frequencies (MAF). Note that predictive measures are higher for very rare (MAF < 0.1%) and rare variants (0.1% ≤ MAF < 1%) compared to common variants (MAF ≥ 1%). vert vertebrate, mam mammalian

factors [45]. Firstly, the use of sensitivity and specificity as statistical summary metrics requires dichotomous variant classification, which relies on the definition of an activity threshold below which variants are considered as deleterious (here 50% of WT) and modulation of this cutoff will influence the number of discrepancies. On our pharmacogenetic data set the sensitivity and specificity of predictions was substantially higher for variants that caused >10-fold reduction or no reduction (activity ≥ 90%) in protein functionality, respectively, compared to variants that only had moderate effects (Fig. 5b), indicating that the choice of a more stringent threshold would further improve predictive performance.

Secondly, inter-experimental variability can change the classification of a variant, particularly for variants that result only in moderate decreases of protein activity; a problem which can only be overcome by stringent experimental replications. Furthermore, variants that result in substrate-specific functionality changes can be missed when probing functionality using a limited number of assays (Supplementary Table 2). We observed substrate-dependent differences for *CYP2D6*49*, which significantly reduces enzyme activity towards the CYP2D6 substrates dextromethorphan and bufuralol but does not affect the clearance of tamoxifen [46, 47]. Similarly, *CYP2C8*10* and *CYP2C8*13* exhibited reduced amodiaquine *N*-deethylation activity while their paclitaxel hydroxylation kinetics remained unaffected [48].

Lastly, discrepancies can occur between the functional impact of a variant in vitro and in vivo. One such example is *CYP2D6*35*, which shows reduced tamoxifen hydroxylation in vitro [47] but has not been associated with reduced activity in vivo [49]. Similarly, *CYP2A6*8* is unlikely to affect catalytic activity in vivo [50] but strongly impairs nicotine and coumarin metabolism in vitro [51]. Our ADME optimized prediction model clearly flagged both alleles as functionally neutral (Fig. 2a), thus correctly predicting the functional consequence in vivo. However, for the sake of consistency and clarity we trained our model exclusively with quantitative and homogeneous experimental in vitro data and did not introduce more heterogeneous and variable data from patient phenotyping.

The presented prediction framework improved both sensitivity and specificity of functionality predictions for variants in poorly conserved genes compared to preexisting assays. However, while the model is capable of predicting the functionality of genetic variations beyond missense mutations, such as indels, frameshifts and synonymous variants, comprehensive investigations into the performance regarding these variant classes are currently not feasible due to the small number of such pharmacogenetic variants with available experimental characterization data.

In summary, we have developed and validated a functionality prediction framework for genetic variants in ADME genes that significantly outperforms current methods using multiple quality metrics, is not limited to previously encountered mutations and can be easily applied to novel variants through the use of the established ANNOVAR platform. Importantly, the model not only informs about the likelihood that the variant in question has deleterious effects on the functionality of the gene product but also provides quantitative estimates of its effect on gene function. Thus, it presents a versatile tool that aspires to improve the prediction of phenotypic consequences of variants discovered in genomic sequencing projects, thereby facilitating the translation of the entire spectrum of patient's genetic variability into pharmacogenetic recommendations.

Acknowledgements This study was supported by the European Union's Horizon 2020 research and innovation program U-PGx under grant agreement no. 668353 and by the Swedish Research Council [grant agreement numbers 2016-01153 and 2016-01154].

Author contributions Y.Z. collected the variants and analyzed the data. Y.Z. and S.M. performed the computational functionality analyses. M.K. and M.H. compiled in vitro functionality data. V.M. L. designed the study, analyzed the data and wrote the manuscript. All authors discussed and agreed on the final version of the manuscript.

Compliance with ethical standards

Conflict of interest V.M.L. is a co-founder and owner of HepaPredict AB. The remaining authors declare that they have no conflict of interest.

References

1. Wetterstrand KA DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) http://www.genome.gov/sequencingcostsdata. Accessed 14 Aug 2017.
2. Illumina Press Release. https://www.illumina.com/company/news-center/press-releases/press-release-details.html?newsid=2236383.
3. Stadler ZK, Thom P, Robson ME, Weitzel JN, Kauff ND, Hurley KE, et al. Genome-wide association studies of cancer. J Clin Oncol. 2010;28:4255–67.
4. Foulkes WD, Knoppers BM, Turnbull C. Population genetic testing for cancer susceptibility: founder mutations to genomes. Nat Rev Clin Oncol. 2015;13:41–54.
5. McCarthy MI. Genomics, type 2 diabetes, and obesity. New Engl J Med. 2010;363:2339–50.
6. Giusti-Rodríguez P, Sullivan PF. The genomics of schizophrenia: update and implications. J Clin Investig. 2013;123:4557–63.
7. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Clin Oncol. 2013;14:681–91.
8. Sawyer SL, Hartley T, Dyment DA, Beaulieu CL, Schwartzentruber J, Smith A, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. Clin Genet. 2015;89:275–84.
9. Hyman DM, Taylor BS, Baselga J. Implementing genome-driven oncology. Cell. 2017;168:584–99.
10. Consortium GP, Auton A, Brooks LD, Kang HM, College B, Harvard BIoMa, et al. An integrated map of genetic variation from 1092 human genomes. Nature. 2012;491:56–65.
11. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012;337:100–4.
12. Fujikura K, Ingelman-Sundberg M, Lauschke VM. Genetic variation in the human cytochrome P450 supergene family. Pharm Genom. 2015;25:584–94.
13. Bush WS, Crosslin DR, Owusu-Obeng A, Wallace J, Almoguera B, Basford MA, et al. Genetic variation among 82 pharmacogenes: the PGRNseq data from the eMERGE network. Clin Pharmacol Ther. 2016;100:160–9.
14. Sim SC, Kacevska M, Ingelman-Sundberg M. Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. Pharm J. 2013;13:1–11.
15. Lauschke VM, Ingelman-Sundberg M. Precision medicine and rare genetic variants. Trends Pharmacol Sci. 2016;37:85–86.
16. 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;19:20–29.
17. Peterson TA, Doughty E, Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. J Mol Biol. 2013;425:4047–63.
18. Trost B, Kusalik A. Computational prediction of eukaryotic phosphorylation sites. Bioinformatics. 2011;27:2927–35.
19. Kulshreshtha S, Chaudhary V, Goswami GK, Mathur N. Computational approaches for predicting mutant protein stability. J Comput Aided Mol Des. 2016;30:401–12.
20. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.
21. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11:863–74.
22. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
23. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.
24. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118–e118.
25. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using Hidden Markov models. Hum Mutat. 2012;34:57–65.
26. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.
27. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE. 2012;7:e46688–46613.
28. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genom. 2013;14(Suppl 3):S3.
29. Kircher M, Witten DM, Jain P, O'Roak B J, Cooper G M, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
30. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31:761–3.
31. 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. 2015;24:2125–37.
32. 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:e1001025–1001013.
33. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25:i54–i62.
34. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.
36. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3:32–35.

37. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2:37–63.
38. 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 Human Genet. 2016;99:877–85.
39. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genom Hum Genet. 2006;7:61–80.
40. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011;32:358–68.
41. Martelotto LG, Ng CK, De Filippo MR, Zhang Y, Piscuoglio S, Lim RS, et al. Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome Biol. 2014;15:453–420.
42. Mahmood K, Jung C-h, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genom. 2017;11:10.
43. Sun J, Zhu K, Zheng W, Xu H. A comparative study of disease genes and drug targets in the human protein interactome. BMC Bioinforma. 2015;16Suppl 5(Suppl 5):S1.
44. Lauschke VM, Milani L, Ingelman-Sundberg M. Pharmacogenomic biomarkers for improved drug therapy—recent progress and future developments. AAPS J. 2017;20:4.
45. Gallion J, Koire A, Katsonis P, Schoenegge A-M, Bouvier M, Lichtarge O. Predicting phenotype from genotype: improving accuracy through more robust experimental and computational modeling. Hum Mutat. 2017;38:569–80.
46. Sakuyama K, Sasaki T, Ujiie S, Obata K, Mizugaki M, Ishikawa M, et al. Functional characterization of 17 CYP2D6 allelic variants (CYP2D6.2, 10, 14A-B, 18, 27, 36, 39, 47-51, 53-55, and 57). Drug Metab Dispos. 2008;36:2460–7.
47. Muroi Y, Saito T, Takahashi M, Sakuyama K, Niinuma Y, Ito M, et al. Functional characterization of wild-type and 49 CYP2D6 allelic variants for N-desmethyltamoxifen 4-hydroxylation activity. Drug Metab Pharmacokinet. 2014;29:360–6.
48. Tsukada C, Saito T, Maekawa M, Mano N, Oda A, Hirasawa N, et al. Functional characterization of 12 allelic variants of CYP2C8 by assessment of paclitaxel 6 alpha-hydroxylation and amodiaquine N-deethylation. Drug Metab Pharmacokinet. 2015;30: 366–73.
49. Gaedigk A, Ryder DL, Bradford LD, Lceder JS. CYP2D6 poor metabolizer status can be ruled out by a single genotyping assay for the-1584G promoter polymorphism. Clin Chem. 2003;49:1008–11.
50. Xu C, Rao YS, Xu B, Hoffmann E, Jones J, Sellers EM, et al. An in vivo pilot study characterizing the new CYP2A6*7, *8, and *10 alleles. Biochem Biophys Res Commun. 2002;290: 318–24.
51. Hosono H, Kumondai M, Maekawa M, Yamaguchi H, Mano N, Oda A, et al. Functional Characterization of 34 CYP2A6 Allelic Variants by Assessment of Nicotine C-Oxidation and Coumarin 7-Hydroxylation Activities. Drug Metab Dispos. 2017;45:279–85.