DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

DMD-AR-2020-000320R2

Development of Robust QSAR Models for CYP2C9, CYP2D6, and CYP3A4 Catalysis and Inhibition

Eric Gonzalez*, Sankalp Jain*, Pranav Shah*, Nao Torimoto-Katori, Alexey Zakharov, Đắc-Trung Nguyễn, Srilatha Sakamuru, Ruili Huang, Menghang Xia, R. Scott Obach, Cornelis E.C.A. Hop, Anton Simeonov & Xin Xu.

E.G, S.J, P.S, A.Z, D-T.N, N.T-K, S.S, R.H, M.X, A.S & X.X: : Division of Preclinical Innovation, National Center for Advancing Translational Sciences (NCATS). 9800 Medical Center Dr, Rockville MD, 20850

N.T-K: Discovery Technology Laboratories, Sohyaku. Innovative Research Division, Mitsubishi Tanabe Pharma Corporation. 1000 Kamoshida-cho, Aoba-ku, Yokohama-shi, 227-0033, Japan

R.S.O: Pfizer Inc. Department of Pharmacokinetics, Dynamics and Metabolism, Pfizer. Groton, CT, 06340

C.E.C.A.H: Genentech Inc. Department of Drug Metabolism and Pharmacokinetics, Genentech Inc., San Francisco, CA 94080

*These authors contributed equally.

Running title: Substrate and Inhibitor QSAR models for CYP2C9, 2D6 and 3A4

Corresponding author: Xin Xu, Ph.D.

National Center for Advancing Translational Sciences

Division of Preclinical Innovation

9800 Medical Center Dr

Rockville, MD 20850

Tel: 301-480-9844

E-mail: xin.xu3@nih.gov

Number of text pages: 43

Number of tables: 8

Number of figures: 8

Number of references: 63

Number of words in the Abstract: 250

Number of words in the Introduction: 750

Number of words in the Discussion: 1317

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Abstract

Cytochrome P450 (CYP) enzymes are responsible for the metabolism of >75% of marketed drugs, making it essential to identify the contributions of individual CYPs to the total clearance of a new candidate drug. Overreliance on one CYP for clearance levies a high risk of drug-drug interactions; and considering that several human CYP enzymes are polymorphic, it can also lead to highly variable pharmacokinetics in the clinic. Thus, it would be advantageous to understand the likelihood of new chemical entities to interact with the major CYP enzymes at an early stage in the drug discovery process. Typical screening assays using human liver microsomes do not provide sufficient information to distinguish the specific CYPs responsible for clearance. In this regard, we experimentally assessed the metabolic stability of ~5,000 compounds for the three most prominent xenobiotic metabolizing human CYPs, i.e., CYP2C9, CYP2D6 and CYP3A4, and used the datasets to develop quantitative structure-activity relationship models for the prediction of high clearance substrates for these enzymes. Screening library included the NCATS Pharmaceutical Collection, comprised of clinically approved low molecular weight compounds, and an annotated library consisting of drug-like compounds. To identify inhibitors, the library was screened against a luminescence-based CYP inhibition assay; and through cross-referencing hits from the two assays, we were able to distinguish substrates and inhibitors of these enzymes. The best substrate and inhibitor models (Balanced accuracies ~0.7), as well as the data used to develop these models, have been made publicly available (https://opendata.ncats.nih.gov/adme) to advance drug discovery across all research groups.

Significance Statement:

In drug discovery and development, a drug candidate with an indiscriminate CYP metabolic profile is considered advantageous, since they provide less risk of potential issues with CYP polymorphisms and drug-drug interactions. In this study, we have developed robust substrate and inhibitor QSAR models for the three major xenobiotic metabolizing CYPs, i.e. CYP2C9, CYP2D6, and CYP3A4. The use of these models early in drug discovery will enable project teams to strategize or pivot when necessary, thereby accelerating drug discovery research.

Introduction

Hepatic biotransformation of small molecule therapeutics by cytochrome P450 (CYP) enzymes continues to be the predominant route of metabolic clearance, highly impacting their bioavailability and systemic exposure. An in-depth analysis of clearance mechanisms is important because it will help predict human pharmacokinetics, indicate the probability of drug-drug interactions (DDI), and identify the potential for pharmacokinetic variability due to race, sex, age, and genetic polymorphisms(Roden and George (2002); (Sansone-Parsons et al., 2007). In this regard, it is necessary to determine the contributions of individual CYPs to the total clearance of a compound. When compounds have a high fraction metabolized (F_m) by one enzyme, e.g. S-warfarin and CYP2C9(Kaminsky & Zhang, 1997), variability in enzyme activity or expression can result in unanticipated low clearance, or conversely undergo ultrarapid metabolism, a known issue for CYP2D6 gene variants(Ingelman-Sundberg, 2005). Inhibitors of an enzyme can lead to elevated circulatory concentrations, and toxicity depending on the therapeutic index of the compound, with ensuing black box warnings on the drug label or withdrawal from the market(Di, 2017; Layton et al., 2003). Conversely, enzyme induction can diminish the efficacy of a compound when the increased expression and activity leads to rapid clearance. To address these concerns, compounds are sought that possess a well-distributed metabolism profile across multiple enzymes and clearance mechanisms(Zientek & Youdim, 2015).

The current standard for preliminary estimation of the human metabolic stability of a compound at the discovery stage is the *in vitro* clearance assay using human liver microsomes (HLM), which are enriched with various xenobiotic metabolizing CYPs, including 1A2, 2C9, 2C19, 2D6, and 3A4, among others. However, a simple HLM clearance assay, which monitors

the depletion of a compound over time, does not identify the CYP(s) responsible for the metabolism. Alternatively, assessing the clearance with individual enzymes for each new chemical entity (NCE) would be an inefficient and costly approach, as medicinal chemists generate a multitude of compounds in their exploration of chemical space to develop novel therapies.

One focus at the National Center for Advancing Translational Sciences (NCATS) is to create and disseminate *in silico* tools that facilitate and accelerate translational research. To this end, NCATS, with support from the International Consortium for Innovation and Quality in Pharmaceutical Development (IQC), has endeavored to develop quantitative structure-activity relationship (QSAR) models capable of predicting the specific CYP enzyme(s) responsible for clearance of new, unexplored compounds. Although predictive QSAR models for individual enzymes are commercially available, those highly regarded can be costly, thereby limiting this resource to the broader scientific community that includes small companies, academic research institutes, and non-profit patient focused organizations. Furthermore, commercial models are often developed using small training datasets, and data is typically sourced from literature, which can introduce error through variability in methods from different laboratories with inconsistent expertise and foci. Alternatively, the robust quantitative high-throughput screening technologies at NCATS have enabled production of sizeable databases from standardized protocols(Shah et al., 2016; Veith et al., 2009), which is the foundation for developing predictive QSAR models with improved accuracy.

Here we report the *in vitro* activities of ~5000 low molecular weight compounds with three major CYP enzymes, i.e. CYP2C9, CYP2D6, and CYP3A4, which can be attributed with ~75% of total CYP-mediated metabolism of clinical drugs(F. Peter Guengerich, 2015). We

focused on two primary CYP endpoints used in the discovery stage, i.e. clearance and inhibition. The clearance assay typically used to assess the metabolic stability of a compound is dependent on complete enzymatic turnover, a process consisting of nine steps in the canonical CYP oxidation reaction(F. P. Guengerich, 2018). While informative, the knowledge garnered from this assay is limited to substrates, therefore obliging further studies to identify inhibitors. Notably, competitive inhibition assays that rely on probe conversion, such as P450-Glo™, are alone unable to distinguish between substrates and inhibitors, given that both types of ligands can generate similar readouts through various CYP enzyme binding mechanisms (Figure 1B). By cross-referencing the compounds that exhibit probe inhibition with those that were metabolized, we identified the most probable inhibitors from our datasets.

The successful application of machine learning approaches to develop predictive QSAR models for absorption, distribution, metabolism, elimination and toxicity (ADMET) properties is well recognized(Kearnes et al., 2016; Wenzel et al., 2019), and is the impetus for the work reported herein. Using in-house generated datasets, we developed conventional QSAR models, as well as multi-task models, to predict CYP-substrates and inhibitors. Most importantly, the training datasets, and models with the greatest balanced accuracy, have been published (https://opendata.ncats.nih.gov/adme) to benefit and accelerate drug discovery across all research groups.

Material and Methods

P450-Glo™ assay kits were purchased from Promega Corporation (Madison, WI) for CYP3A4 (V9910), CYP2C9 (V9790) and CYP2D6 (V9890). NADPH Regenerating Solution A (Catalog #: 451220) and B (Catalog #: 451200), Human CYP3A4 (456202), CYP2C9 (456258) and CYP2D6 (456217) Supersomes™ were purchased from Corning Life Sciences (Corning, NY). Ketoconazole, furafylline, sulfaphenazole, quinidine, and albendazole were purchased from Sigma-Aldrich (St. Louis, MO).

Compound Library:

The ~5000 compound library used for this publication encompasses the NCATS Pharmaceutical Collection (NPC)(R. Huang et al., 2011) and an annotated NCATS library. The NPC library contains ~2800 compounds that have been approved for clinical use by US, Canadian, Japanese, and European health regulatory authorities. The NCATS annotated library is comprised of ~2200 diverse drug-like molecules. This annotated library consists of mostly investigational compounds that represent diverse target classes and disease areas. This combined library (~5000 compounds) will henceforth be referred to as the NCATS-ADME library.

High-throughput metabolic stability (clearance) assays:

The substrate depletion assay was employed to determine metabolic stability, using an established mid-density (384-well format) protocol(Shah et al., 2016). The workflow included a robotic system for incubation and sample clean-up, coupled with an automated ultra-high performance liquid chromatography-high resolution mass spectrometry (UHPLC-HRMS) method for sample analysis. Briefly, each 110 µL reaction mixture consisted of 1 µM test article, supersomes, and an NADPH regenerating system in 100 mM phosphate buffer at pH 7.4. The

specific protein and enzyme concentrations, as well as the control compounds utilized are listed in Table 1. Incubations were conducted at 37°C, with mixing, and reaction aliquots were quenched at 0, 5, 10, 15, 30, and 60 min by addition of cold acetonitrile (ACN) with internal standard (IS), i.e. albendazole. Centrifugation at 3000 g, 4 °C for 20 min, was used to clear samples of precipitated protein and debris. Sample analysis in an UHPLC-HRMS instrument, data extraction, and half-life ($t_{1/2}$) determinations were performed as previously described(Shah et al., 2016).

The compounds were binned into clearance categories based on the observed $t_{1/2}$ criteria outlined in Table 2. Data with below limit of quantitation (BLQ), inconclusive (INC), and not found (N/F) designations were excluded from further analysis. The complete dataset, annotated with substrate class, is provided in Supplemental File 2.

P450-Glo qHTS:

The P450-Glo inhibition assay is a luminescent technique used to detect CYP activity, through the liberation of luciferin from CYP probe substrates. P450-Glo assays were performed using a previously described method with minor modifications(Veith et al., 2009). All assays were optimized by incubating positive control compounds at both RT and 37°C conditions. Since no difference in compound activities were found at RT and 37°C for CYP2D6 and CYP3A4, assays for these two enzymes were run at RT. Briefly, 2 µL of CYP-substrate mix was dispensed into medium-binding white/solid 1,536-well plates using a Flying Reagent Dispenser (FRD, Aurora Discovery, Carlsbad, CA) with the exception of adding Bovine Serum Albumin (BSA) to the mixture for CYP2C9. The initial optimization assays for CYP2C9 yielded lower signal to

background ratios and higher well to well variation. To improve signal and prevent adhesion of protein to tubes of the plate dispenser, 0.4% BSA was added to the CYP2C9 enzyme assays. 23 nL of each positive control (columns 1-4) and test compound (columns 5-48) dissolved in DMSO was transferred to the assay plates using a Wako Pintool station (Wako Automation, San Diego, CA). The positive controls used in these experiments are listed in Table 3. After the control/test compounds were transferred, the assay plates were incubated at room temperature (RT) for 10 min before the addition of 2 μL NADPH regeneration solution using an FRD. The reaction incubation continued at either RT or 37 °C for 60 min, and then quenched by FRD addition of 4 μL of the detection reagent. After a 20 min incubation at room temperature, the luminescence intensity was measured and quantified using a ViewLux plate reader (PerkinElmer, Shelton, CT). Data was expressed as relative luminescence units.

The concentration-response activity data for each compound, relative to control, was fit with the four parameter Hill equation to obtain percent activity and potency values. The complete P450-Glo qHTS datasets have been deposited to PubChem, with the following assay IDs (AIDs): 1645841(CYP3A4), 1645840 (CYP2D6), and 1645842 (CYP2C9). Compounds were classified as a "hit" in the P450-Glo screens when the inhibition efficacy was > 65% and the potency <10 μM.

Training set (TR) and test set (TS) preparation for QSAR:

The NCATS-ADME library was pre-processed to eliminate entries containing duplicates, inorganic compounds, non-covalent complexes and mixtures. Furthermore, salts and compounds containing organometals were removed. The chemical structures were then standardized using Francis Atkinson Standardizer tool. To estimate the statistical performance in a robust way, we used a 5-fold cross-validation routine. The final dataset (Table 4) was then split 5-folds while

retaining the initial ratio of active/inactive (stratified sampling). For each fold, $4/5^{th}$ of the dataset was used as the training set and remaining $1/5^{th}$ as the test set, sliding over folds.

Parsing of substrates and inhibitors by process of elimination:

The "hits" in the P450-Glo datasets were cross-referenced with the substrate classifications from the clearance assay. The compounds were binned into 4 different categories using the classification criteria outlined in Table 5.

Molecular descriptor calculation:

The following sets of descriptors were calculated for each of the datasets:

1. The combination of fingerprints with five physicochemical properties, i.e. molecular weight (MW), atom-based calculated partition coefficient (SlogP)(Wildman & Crippen, 1999), topological polar surface area (TPSA), number of H-bond donor and number of H-bond acceptor, was reported to provide superior performance for prediction of P450-mediated properties(A. Zakharov et al., 2019). Hence, we used Avalon fingerprints (1024 bits) and Morgan fingerprints (calculated using RDKit(Landrum)) in combination with the abovementioned five physicochemical properties.

2. Dragon Descriptors- The Dragon package provided us with 3840 descriptors(Inc; Solutions). Constant value descriptors (0 throughout) and descriptors with low variance (<0.4) were removed. For the final modeling exercise, we used 1164 descriptors.

Machine learning- Stratified bagging Random Forest and multi-task deep neural networks:

Random Forest (with default parameters) was used as a base-classifier(Breiman, 2001). The number of trees was arbitrarily set to 100 (default), since it has been shown that the optimal

number of trees usually falls between 64-128, and increasing the number of trees does not necessarily improve model performance(Oshiro et al., 2012). The problem of data imbalance was overcome using under-sampling stratified bagging (SB)(He & Garcia, 2009; Kotsiantis, 2008; Tetko et al., 2013), which has been proven to be one of the best performing methods for dealing with imbalanced datasets(Jain et al., 2018; Tetko et al., 2013). SB is a machine learning technique that is based on an ensemble of models developed using multiple training datasets sampled from the original training set. This technique uses minority class samples to create the training set of positive samples using the traditional bagging approach (resampling with replacement), and then randomly selects the same number of samples from majority class. Thus, the total bagging training set size was double the minority class. Several models are then calculated and averaged to produce a final ensemble model(Tetko et al., 2013). Because of random sampling, about 37% of the compounds are left out in each run, creating "out-of-the-bag" sets which are used for testing the performance of the final model(Tetko et al., 2013). Although a small set of samples are selected each time, majority of compounds contribute to the overall bagging procedure, given that datasets were generated randomly. Further, an earlier study by Tetko et al.(Tetko et al., 2013) showed that larger numbers of models per ensemble (e.g. 128, 256, 512 and 1024) did not significantly increase the balanced accuracy of models. Thus, in this study we built a total of 64 models per ensemble. All models using RF in combination with stratified bagging were developed and deployed by using the data analytics platform KNIME(Berthold et al., 2008).

The performance of multi-task deep neural network (DNN) method on our datasets was also evaluated. DNN has gained prestige and has been widely applied across different domains of science and technology(Korotcov et al., 2017; A. V. Zakharov et al., 2019). DNN is a

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

variation of an artificial neural network (ANN) that consists of several sequential hidden layers. Each layer is represented by a linear vector transformation, Wx+b (where W is a matrix of tunable weights and b is a bias vector), followed by a nonlinear transformation function, i.e., sigmoid. In this study, multi-task DNN models (MT-DNN v1) were developed using the multi-layer feedforward neural networks implemented in Keras using the Tensorflow backend. The loss function was minimized using the Adam algorithm. All models developed in this study were evaluated by 5-fold cross-validation(Tropsha, 2010).

Model performance assessment:

The performance of each classification model was assessed based on sensitivity (Eq. 1), specificity (Eq. 2), accuracy (Eq. 3), balanced accuracy (BACC, Eq. 4), and the Matthews correlation coefficient (MCC, Eq. 5). Accuracy may be misleading for a highly imbalanced dataset, which makes BACC and MCC more appropriate performance measures to compare different classifiers, given their ability to handle skewed datasets.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (1)$$

$$Specificity = \frac{TN}{(TN+FP)} \quad (2)$$

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (3)$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{(TP)}{(TP+FN)} + \frac{(TN)}{(TN+FP)}\right) \quad (4)$$

$$MCC = \frac{\{(TP*TN)-(FP*FN)\}}{\{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)\}^{1/2}} \quad (5)$$

TP: true positives; TN: true negatives; FP: false positives; FN: false negatives.

Results

Data curation for the clearance assays:

The compound library was annotated with structural information using the simplified molecular-input line-entry system (SMILES) and LyChI (NCATS) notation formats. During compound batch review, a small percentage (~3 %) were found to have inconsistent structural annotations, which created missing data amongst the three CYP enzyme datasets when pivoted by either SMILES or LyChI. The majority of mis-annotations were introduced through the vendor supplied information, which is difficult to detect at purchase, or receipt, considering that the bulk of compounds included in the study were procured within large commercial compound libraries. Structural information was verified with the vendor and the library annotations were updated accordingly.

Of the ~5000 compounds screened in the clearance assay, 80% were assigned a usable $t_{1/2}$ within each CYP dataset, following exclusion of the compounds with BLQ, INC, and N/F designations. The unreliable data generated in this study is typical of high-throughput MS-based assays, which are afflicted by erratic pipetting errors common to liquid handlers performing incubations 384-well format, and weak mass spec signal, due to inefficient ionization or unmonitored adduct formation.

CYP substrates identified in high throughput clearance assays:

As expected, the highest number of substrates ($t_{1/2}$ <30 mins) were identified in the CYP3A4 screen (45%), followed by CYP2D6 (33%) and CYP2C9 (27%) (Table 6). Overall, only an 11% substrate overlap was observed between all three enzymes.

Physicochemical distribution of high and low-clearance compounds:

Molecular properties of compounds, such as SlogP, TPSA, and MW, were calculated using an in-house compound dataset annotation tool, known as NCATS Find(NCATS). For all three enzymes, a large proportion of substrates ($t_{1/2}$ <30 mins) fell within the 250-550 MW range, had SlogP values between 2-6, TPSA values less than 100, 0-2 hydrogen bond donors (data not shown) and 1-8 hydrogen bond acceptors (data not shown). No major difference was observed in the physicochemical property distributions between substrates of the enzymes except for CYP2D6 substrates which displayed lower molecular weight and lower TPSA compared to CYP2C9 and CYP3A4 (Figure 2). Furthermore, a direct correlation between the calculated $t_{1/2}$ values and the above-mentioned molecular descriptors was not apparent. Additionally, we did not find the established charged preferences of CYP2C9 (Acids) and CYP2D6 (basic amines)(Kerns & Di, 2008) to be distinguishing physicochemical features in our dataset (Data not shown).

CYPs inhibitors and activators identified in qHTS assays:

The highest percentage of P450-Glo “hits” were obtained in the CYP3A4 screen (29%), followed by CYP2C9 (23%) and CYP2D6 (19%) (Table 7). In contrast to the clearance assay, only 5% overlap was found amongst the three enzymes. It should be noted that both inhibitors and substrates decrease the luminescent signal in this assay by occluding the probe from the substrate pocket, and therefore a P450-Glo “hit” may not be a CYP inhibitor in the true sense.

The P450-Glo "hits" used in this study for the development of predictive QSAR models exclude compounds that increase CYP metabolism of the probe substrate, observed through an elevated luminescence readout. As current knowledge would lead us to expect, CYP3A4 exhibited the greatest number of molecules (118) that increased luciferin production. Notably, from the 37 compounds that stimulated CYP2C9 metabolism, two molecules were indiscriminate against CYP3A4, i.e. Proscillaridin and Hematopoietic prostaglandin D synthase (HPDGS)-inhibitor-1 (although at varying AC_{50} values). While the 50% active concentration (AC_{50}) range of CYP2C9, between 0.025 and ~45 μM, was comparable to that of CYP3A4, spanning from 0.001 to ~39 μM, the sole two molecules that increased CYP2D6 activity both had a potency of ~30 μM. Furthermore, only CYP2D6 and CYP3A4 had a single discrete compound with increased probe metabolism that also exhibited high clearance, Tenatoprazole (a putative proton pump inhibitor) and SCHEMBL17791590 (an aldehyde dehydrogenase inhibitor), respectively. Although, predictive QSAR models for activators were not feasible, given that a larger dataset of activator compounds would be needed to power the model, the complete list of activating compounds is provided has been provided in Supplemental File 2.

Parsing of substrates and inhibitors by process of elimination:

Although the P450-Glo assay alone cannot distinguish inhibitors and substrates, when stratified with the clearance assay, the parsing of probable inhibitors and substrates is feasible. The number of putative inhibitors identified in the P450-Glo assay decreased significantly (77%, 32% and 66% for CYP3A4, 2C9 and 2D6 respectively). The Venn diagrams in Figure 3A & B show the overlap of substrates and inhibitors, highlighting the broad substrate/inhibitor recognition capabilities of the three enzymes. The predisposition of CYP3A4 to metabolize xenobiotics is apparent, with the number of substrates being three times greater than that of

inhibitors. While CYP2D6 also exhibited a higher tendency to metabolize compounds, at substrate-inhibitor ratio of ~2, CYP2C9 is clearly more susceptible to inhibition, with a corresponding ratio of ~0.5. Nonetheless, the annotation of these enzymes as xenobiotic metabolizers is validated through the observation of notably higher counts and overlap amongst substrates, as compared to the parsed inhibitors.

Figure 3C displays an example of a chemical space plot for all CYP2C9 data based on visual clustering(Optibrium). We find that the compounds are widely scattered pointing to the diversity of our dataset.

It is important to note this approach overlooks compounds which have divergent enzymatic mechanisms based on the presence of additional ligands, by which binding alone leads to the oxidation of the molecule but can lead to a purely inhibitory enzyme-ligand complex in the presence of the probe (Figure 1 schemes Bi, Bii middle, and Bii right, can be true for the same compound). The compounds in reference cannot be distinguished from those in Category 1 per the rationale used (Table 5), and for the purposes of this study will remain in that category given that they fall within the high clearance threshold. Ketoconazole is a prime example of this case, as it is a well-established substrate while also considered a potent inhibitor of CYP catalytic activity (Boulenc et al., 2016; https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/018533s041lbl.pdf, 2020). We reviewed the literature for several of the category 1 compounds and found that some historical compounds can be further annotated with this dataset, such as the assignment of tripelenamine as a substrate for CYP2D6. While it is not surprising that this first generation antihistamine is cleared by the enzyme, the reported metabolism of this archaic compound is limited(Chaudhuri et al., 1976; Yeh, 1991), and is solely categorized as an inhibitor of CYP2D6 in the Drug

Interactions Flockhart Table ™ (Flockhart), illustrating the presence of missing substrate/inhibitor annotations for familiar older compounds.

Predictive models - SB and MT-DNN:

Once data analysis and curation were complete, we focused our attention on building classification models, which can effectively distinguish actives from inactives using a machine-learning approach. For this, a panel of classifiers was trained on all datasets using different combinations of descriptors. To avoid bias that might occur due to the splitting schemes employed, all models were evaluated in a five-fold external cross-validation scheme. Considering the average prediction performance across 5-folds, models for all six datasets (3 substrate + 3 inhibitor) showed BACC values close to or above 70%. For the CYP3A4 substrate dataset, a DNN-based model with dragon descriptors performed the best (BACC = 76%; MCC= 0.51), closely followed by SB with a combination of Morgan fingerprints and five physicochemical properties (BACC = 75% MCC= 0.49), which was found to be the best performing method for the five remaining datasets. Taking a consensus between different descriptor combinations and/or machine-learning approaches did not improve the model performance (data not shown). Considering that the two approaches did not yield significantly different BACC and MCC values (Figure 4; Table 8), SB with Morgan fingerprints and five physicochemical properties was chosen as the default model for all datasets due to its accessibility (i.e. open-source). Table S1 in Supplemental File 1 reports prediction performance measures for all datasets used in this study.

Applicability Domain assessment:

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

The applicability domain (AD) of a QSAR model defines the limitation in its structural domain and response space. In other words, this principle for model validation restricts the applicability of a model to reliably predict test compounds that are structurally similar to training compounds used while building the model. Historically, several approaches have been proposed to calculate the applicability of a QSAR model(Patel et al., 2018; Sahigara et al., 2013; Sushko et al., 2010; Yun et al., 2017). In this study, for estimation of the model's AD, the Tanimoto similarity was assessed(Bajusz et al., 2015; TT, 1958) between test set compounds and its nearest neighbor in the training set, using Morgan fingerprints. The calculations were performed separately for all six datasets. For each fold within each dataset, we filtered out compounds that were below a certain similarity threshold and further calculated the BACC and the coverage of predictions as the percentage of compounds that fall within the model's AD. The distribution of BACC, and corresponding coverage values, for test sets (5-fold average) versus AD cut-offs are presented in Figure 5, using the CYP2C9 substrate dataset as an example. The data reveals a positive trend between AD and prediction accuracy, wherein the AD threshold value increases with the prediction accuracy of the model. The coverage of prediction correlates inversely with AD, as shown by the dramatic decrease in coverage as AD values rise.

The best prediction results for CYP2C9 substrates were achieved with an AD equal to 0.8, resulting in a BACC of 0.79, although with a very low coverage value of ~1%. The coverage achieved with an AD cut-off of 0.7 was not significantly better. The optimal ratio of both the accuracy of prediction and coverage was achieved with an AD cut-off value of 0.6. Similar results were obtained for all other datasets [Supplemental File 1, Table S2]. Given the clear trend between the accuracy of prediction and AD values, this approach can be used to establish the confidence level of predictions.

Analysis of uncertainty of prediction/class probability:

In addition to the category, the classification approach provides an output for class probability, a numerical value between 0 and 1, which corresponds to the probability of a compound being active. Class probability is an estimation of the reliability of predictions and is referred to as 'uncertainty of prediction'. Values close to 1 indicate active compounds whereas values close to 0 indicate inactive compounds. Analysis of class probability showed most of the misclassification was in the class probability range of 0.5 - 0.6. In the case of the CYP2C9 substrate dataset, the models predicted more than 80% of the compounds correctly for the class probability range between 0 - 0.4 and 0.7 - 1 (Figure 6). The same trend was observed for all six prediction models (Supplemental File 1, Table S3), reinforcing increased confidence in model predictions when the 0.5-0.6 class probability range is excluded.

Comparison with the reference tools/models:

Following completion of the models, an external validation test set was sought to ascertain their utility; a challenging endeavor considering that clearance and inhibition data for individual enzymes is limited and scattered in literature. In addition, we pursued to compare our model performance with other open-source models that exist in literature. While a few open-source websites offer CYP-specific substrate and inhibitor models, most, were developed using compounds from literature, i.e. essentially a subset of the NPC. Given that our models were developed on the entire NPC dataset, the effort to compare model performance metrics was abandoned.

The focus was then shifted to comparing against commercial models. ADMET Predictor from Simulations Plus is one of the leading software packages for ADMET predictions and is

regularly used at NCATS. The software includes substrate and inhibitor classification models for 9 CYP enzymes using data obtained from Biovia metabolite database, Drugbank, and other literature sources. The total number of compounds used to build the CYP2C9, CYP2D6 and CYP3A4 substrate models ranged between 1400 to 1600, whereas the inhibition models were developed using ~700 compounds. To compare against ADMET Predictor, our models developed herein were re-trained using only the NPC library. Since SB with Morgan fingerprints and five physicochemical properties showed superior performance in comparison to other techniques, we used this combination to develop prediction models on the NPC library. These models were then used to predict the NCATS annotated library and model performances were compared against predictions from ADMET Predictor. Model statistics on the training set (NPC) can be found in the Supplemental File 1, Table S4. As shown in Figure 7, the models developed from this work outperformed those from ADMET Predictor (AP) both in terms of BACC and MCC.

To ascertain the robustness of our model, we identified singletons in the NCATS annotated library and compared prediction results/model statistics for those compounds. We found 615 singletons in the NCATS annotated library and once again, our models outperformed ADMET Predictor (Supplemental File 1, Table S4). Although our models exhibited superior performance on this test set as compared to ADMET Predictor, it must be noted that the data used by the models in ADMET predictor may not have been generated in assays that are same or similar to those employed in this study. Therefore, these results are provided only for a comparative assessment and must be cautiously inferred.

Discussion

Human liver microsomes (HLM) are the gold standard for studying phase I/CYP-mediated metabolism. An abundance of HLM data exists in literature, and several groups have used this data to publish QSAR models(Hu et al., 2010; Lee et al., 2007; Liu et al., 2015; Sakiyama et al., 2008; Zakharov et al., 2012). The majority of established, respected models, and (importantly) source datasets, are proprietary, which limits their public accessibility. Although several HLM clearance models exist, the QSAR knowledge for individual CYPs remains limited. A joint effort by NCATS and members from the IQ Consortium commenced the mission to publish a database of clearance values, which will not only help advance drug design efforts, but also provide a better understanding of the structure-activity relationship for major CYP enzymes. The benefits gained by the scientific community from this effort include: 1) Enhancing lead optimization by guiding structure modification; 2) Improving hit selection by high-throughput and computational screening; and 3) Enabling advanced computational human metabolic models for individual metabolic enzymes.

At ~4000 molecules, it is the largest library of compounds screened for individual CYP enzymes. Notably, the database includes the majority of investigational and health authority approved drugs, making it the most publicly available, comprehensive list of CYP2C9, 2D6, and 3A4 substrates and inhibitors for clinically used small molecules, which has been founded on single-source empirical data. The complex kinetics of cytochrome P450 enzymes creates a multitude of enzyme-ligand scenarios, which could make designations of substrate or inhibitor ambiguous. Stemming from observations of non-linear kinetics with CYP-mediated reactions, Korzekwa et al. provided some of the first evidence that enzymes bind multiple ligands simultaneously(Korzekwa et al., 1998). The issue of CYP cooperativity and allosteric

interactions have since been reviewed extensively(Davydov & Halpert, 2008; Denisov et al.). However, it is still necessary to evaluate the CYP-ligand binding empirically, as predictive models do not exist for all the possibilities. Importantly, the datasets and models generated from this effort cannot be extended beyond the simple categorical assignment of substrates and inhibitors, considering that an extensive amount of additional investigation is necessary to fully characterize binding modes(F. P. Guengerich et al.), which is more appropriately evaluated spectroscopically using the shift in the quintessential P450 absorbance band. Further, CYP inhibition models are complexed by multi-ligand interactions, where different probe substrates may lead to alternate structure-activity relationships. Nonetheless, we consider the datasets and models reported herein are widely applicable, as they were developed using P450-Glo system, which is a commonly used assay in the drug discovery screening paradigm.

The predictive machine learning models developed from these studies are fortified by the use of reliable data, which is unhindered by assay and lab-to-lab variability, an inherent affliction of most commercial and open-source models that is introduced by sourcing data from compiled literature. We employed SB and multi-task deep learning models to classify compounds as substrates or inhibitors for three predominant xenobiotic metabolizing enzymes (CYP3A4, CYP2C9 and CYP2D6). Despite the imbalance of the datasets in our study, especially the inhibitor datasets, we were able to achieve classification accuracies (BACC) around 70% (Figure 4; Supplemental Table S1). Comparison with the widely used commercial software, i.e. ADMET Predictor, demonstrates the value of our model and the quality of our data. Since 2012, chemists at NCATS have synthesized >20,000 compounds for more than 250 drug discovery projects that cover a wide range of disease areas, pharmacological targets and cellular pathways. A high degree of similarity in physicochemical properties (molecular weight, Slog P, TPSA, HBA and

HBD) was observed between this dataset (Siramshetty et al., 2020) and our NCATS ADME 5K library. Thus, our models can be used during the compound design phase as well after synthesis as a filtering mechanism to rank order compounds for phenotyping and CYP-inhibition assays in drug discovery.

Clustering is a powerful approach that allows the grouping of 'similar' compounds, to distinguish chemical series within a dataset of diverse compounds, analyze the SAR, and identify 'regions' of chemistry that may yield good properties. Manual inspection of our datasets revealed great structural diversity. With the aim to quantify as well as qualitatively describe this structural diversity, we performed (i) clustering analysis based on Morgan fingerprints (KNIME), as well as (ii) clustering based on maximum common substructure (StarDrop). From 3584 compounds that encompassed the substrate data across the three enzymes, 1829 different clusters were identified using Morgan fingerprints. From those, 1067 were singletons, and only 24 clusters contained ≥ 10 compounds. The most populated cluster contained 30 compounds. Clustering based on maximum common substructure algorithm, as implemented in StarDrop (similarity threshold = 0.70), also revealed high structural diversity in the dataset, yielding 2059 singletons and a max cluster size of 33 compounds. Additionally, the Murcko(Bemis & Murcko, 1996) scaffold algorithm used, identified over 2617 different Murcko scaffolds, within the aforementioned 3584 compound dataset. The Murcko analysis produced an average scaffold-to-compound ratio of 0.73, once again signifying the large structural diversity of our dataset. Analysis showing the most frequent scaffolds is presented in Figure 8. Benzene scaffolds were found with very low frequency (~6% of the dataset), and no other scaffolds reached prevalence values above 0.5%.

Considering the central role CYP450 enzymes play in the clearance of small molecule therapeutics, evaluation of the specific CYP enzymes that catalyze the metabolism of new chemical entities is essential at the preclinical phase of the drug discovery and development process. For more than 20 years, health authorities have provided guidance towards the characterization of *in vitro* drug interactions involving CYPs, as well as other enzymes involved in the human disposition of xenobiotics(S.-M. Huang et al., 2008; Prueksaritanont et al., 2013). The focus is on reducing the risk that a novel molecule will act as a DDI "perpetrator" in the clinic, through the interactions, i.e. substrate and inhibitor, with CYP enzymes. However, DDI assessments are typically conducted long after the concept has been synthesized and now poised as a developmental molecule. Although the models provided in this report are not adequate to replace the studies necessary to predict clinical DDI, the prediction of CYP substrate and inhibitors can be fully exploited early in the discovery phase for ranking or selecting compounds for experimental validation.

One of the first examples of CYP polymorphism was reported by Eichelbaum et al. (Eichelbaum et al., 1975), who showed that N-oxidation of sparteine was subject to a high degree of inter individual variability. Since then, it has been well established that almost all drug metabolizing CYPs enzymes are polymorphic. Johansson and colleagues(Johansson & Ingelman-Sundberg, 2011), have summarized the most important CYP alleles related to drug toxicity, and the classes of drugs most commonly affected by these polymorphisms. The polymorphic variability in pharmacokinetics, which drives pharmacodynamics, can lead to toxicity or inefficacy; both results being detrimental to patients. The use of our models in early drug discovery could enable the flagging of compounds/series which rely heavily on one of these three enzymes for clearance, as well as those with inhibition potential against these enzymes,

prompting medicinal chemists to produce compounds that avoid potential future development problems. The application of this level of detailed information early in the drug discovery process will be invaluable.

In summary, we report the first systemic attempt to profile and generate substrate and inhibitor database of this scope and size, for major CYP450 enzymes. This collaborative effort between NCATS and IQ Consortium yielded several useful tools including: 1) a high-throughput automated incubation method for metabolic stability screening, 2) two different automated data acquisition methods with two different mass spectrometry systems, 3) automated method of assigning $t_{1/2}$ via the 'Validator' software (code publicly available(Shah et al., 2016)), 4) large publicly available datasets (>4000 compounds) for three major CYP enzymes, and 5) robust predictive models for CYP substrates and inhibitors (https://opendata.ncats.nih.gov/adme). We look forward to the prospect that the knowledge gained, and the tools developed, from this venture will accelerate drug translational research in academia, small biotech, and pharmaceutical companies.

Acknowledgements

The authors would like to acknowledge compound management especially Paul Shinn and Misha Itkin for their support. The authors would also like to thank Jorge Neyra for his help with implementing the QSAR models. The authors would also like to acknowledge all working group members from the IQ Consortium especially Dr. Fabio Broccatelli, Dr. Susanne Winiwarter, Dr. Prashant Desai and Dr. Matthew Cerny for their valuable insights.

Authorship Contributions:

Participated in research design: Gonzalez, Shah, Torimoto-Katori, Nguyễn, Zakharov, Obach, Hop & Xu

Conducted experiments: Shah, Torimoto-Katori, Gonzalez & Sakamuru.

Contributed new reagents and analytic tools: Xia and Xu.

Performed data analysis: Shah, Gonzalez, Jain, Zakharov & Huang.

Wrote or contributed to the writing of the manuscript: Gonzalez, Jain, Shah, Zakharov, Nguyễn, Sakamuru, Torimoto-Katori, Huang, Xia, Obach, Hop, Simeonov & Xu.

All authors reviewed the manuscript.

References

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform, 7*, 20. doi:10.1186/s13321-015-0069-3

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry, 39*(15), 2887-2893. Retrieved from <Go to ISI>://MEDLINE:8709122

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications*, 319-326. doi:Doi 10.1145/1656274.1656280

Boulenc, X., Nicolas, O., Hermabessiere, S., Zobouyan, I., Martin, V., Donazzolo, Y., & Ollier, C. (2016). CYP3A4-based drug-drug interaction: CYP3A4 substrates' pharmacokinetic properties and ketoconazole dose regimen effect. *European journal of drug metabolism and pharmacokinetics, 41*(1), 45-54. Retrieved from <Go to ISI>://MEDLINE:25374256

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Chaudhuri, N. K., Servando, O. A., Manniello, M. J., Luders, R. C., Chao, D. K., & Bartlett, M. F. (1976). Metabolism of tripelennamine in man. *Drug metabolism and disposition: the biological fate of chemicals, 4*(4), 372-378. Retrieved from <Go to ISI>://MEDLINE:8293

Davydov, D. R., & Halpert, J. R. (2008). Allosteric P450 mechanisms: multiple binding sites, multiple conformers or both? *Expert opinion on drug metabolism & toxicology, 4*(12), 1523-1535. doi:10.1517/17425250802500028

Denisov, I. G., Frank Dj Fau - Sligar, S. G., & Sligar, S. G. Cooperative properties of cytochromes P450. (1879-016X (Electronic)).

Di, L. (2017). Reaction phenotyping to assess victim drug-drug interaction risks. *Expert opinion on drug discovery, 12*(11), 1105-1115. Retrieved from <Go to ISI>://MEDLINE:28820269

Eichelbaum, M., Spannbrucker, N., & Dengler, H. J. (1975). Proceedings: N-oxidation of sparteine in man and its interindividual differences. *Naunyn-Schmiedeberg's archives of pharmacology, 287 Suppl*, R94. Retrieved from <Go to ISI>://MEDLINE:1143521

Flockhart, D. A. Drug Interactions Flockhart Table. https://drug-interactions.medicine.iu.edu/MainTable.aspx.

Guengerich, F. P. (2015). Human Cytochrome P450 Enzymes. In P. R. Ortiz de Montellano (Ed.), *Cytochrome P450: Structure, Mechanism, and Biochemistry* (pp. 523-785). Cham: Springer International Publishing.

Guengerich, F. P. (2018). Mechanisms of Cytochrome P450-Catalyzed Oxidations. *ACS Catal, 8*(12), 10964-10976. doi:10.1021/acscatal.8b03401

Guengerich, F. P., Wilkey, C. J., & Phan, T. T. N. Human cytochrome P450 enzymes bind drugs and other substrates mainly through conformational-selection modes. (1083-351X (Electronic)).

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi:10.1109/TKDE.2008.239

https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/018533s041lbl.pdf. (2020). Retrieved from https://www.accessdata.fda.gov/drugsatfda_docs/label/2014/018533s041lbl.pdf

Hu, Y., Unwalla, R., Denny, R. A., Bikker, J., Di, L., & Humblet, C. (2010). Development of QSAR models for microsomal stability: identification of good and bad structural features for rat, human and mouse microsomal stability. *J Comput Aided Mol Des, 24*(1), 23-35. doi:10.1007/s10822-009-9309-9

Huang, R., Southall, N., Wang, Y., Yasgar, A., Shinn, P., Jadhav, A., Nguyen, D. T., & Austin, C. P. (2011). The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs

enabling repurposing and chemical genomics. *Sci Transl Med, 3*(80), 80ps16. doi:3/80/80ps16 [pii]

10.1126/scitranslmed.3001862

Huang, S.-M., Strong, J. M., Zhang, L., Reynolds, K. S., Nallani, S., Temple, R., Abraham, S., Habet, S. A., Baweja, R. K., Burckart, G. J., Chung, S., Colangelo, P., Frucht, D., Green, M. D., Hepp, P., Karnaukhova, E., Ko, H.-S., Lee, J.-I., Marroum, P. J., Norden, J. M., Qiu, W., Rahman, A., Sobel, S., Stifano, T., Thummel, K., Wei, X.-X., Yasuda, S., Zheng, J. H., Zhao, H., & Lesko, L. J. (2008). New Era in Drug Interaction Evaluation: US Food and Drug Administration Update on CYP Enzymes, Transporters, and the Guidance Process. *The Journal of Clinical Pharmacology, 48*(6), 662-670. doi:https://doi.org/10.1177/0091270007312153

Inc, T. Dragon. http://www.talete.mi.it/products/dragon_description.htm.

Ingelman-Sundberg, M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *The pharmacogenomics journal, 5*(1), 6-13. Retrieved from <Go to ISI>://MEDLINE:15492763

Jain, S., Kotsampasakou, E., & Ecker, G. F. (2018). Comparing the performance of meta-classifiers-a case study on selected imbalanced data sets relevant for prediction of liver toxicity. *Journal of computer-aided molecular design, 32*(5), 583-590. Retrieved from <Go to ISI>://MEDLINE:29626291

Johansson, I., & Ingelman-Sundberg, M. (2011). Genetic polymorphism and toxicology--with emphasis on cytochrome p450. *Toxicological sciences : an official journal of the Society of Toxicology, 120*(1), 1-13. Retrieved from <Go to ISI>://MEDLINE:21149643

Kaminsky, L. S., & Zhang, Z. Y. (1997). Human P450 metabolism of warfarin. *Pharmacol Ther, 73*(1), 67-74. doi:10.1016/s0163-7258(96)00140-4

Kearnes, S., Goldman, B., & Pande, V. (2016). Modeling Industrial ADMET Data with Multitask Networks. *Stat*, arXiv:1606.08793.

Kerns, E. H., & Di, L. (2008). Drug-Like Properties: Concepts, Structure Design and Methods. *Drug-Like Properties: Concepts, Structure Design and Methods*, 1-528. Retrieved from <Go to ISI>://WOS:000311101200044

Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular pharmaceutics, 14*(12), 4462-4475. Retrieved from <Go to ISI>://MEDLINE:29096442

Korzekwa, K. R., Krishnamachary, N., Shou, M., Ogai, A., Parise, R. A., Rettie, A. E., Gonzalez, F. J., & Tracy, T. S. (1998). Evaluation of atypical cytochrome P450 kinetics with two-substrate models: evidence that multiple substrates can simultaneously bind to cytochrome P450 active sites. *Biochemistry, 37*(12), 4137-4147. Retrieved from <Go to ISI>://MEDLINE:9521735

Kotsiantis, S. B. (2008). Handling imbalanced data sets with a modification of Decorate algorithm. *International Journal of Computer Applications in Technology, 33*(2-3), 91-98. doi:Doi 10.1504/Ijcat.2008.021931

Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org. Retrieved from http://www.rdkit.org

Layton, D., Key, C., & Shakir, S. A. W. (2003). Prolongation of the QT interval and cardiac arrhythmias associated with cisapride: limitations of the pharmacoepidemiological studies conducted and proposals for the future. *Pharmacoepidemiology and drug safety, 12*(1), 31-40. Retrieved from <Go to ISI>://MEDLINE:12616845

Lee, P. H., Cucurull-Sanchez, L., Lu, J., & Du, Y. J. (2007). Development of in silico models for human liver microsomal stability. *J Comput Aided Mol Des, 21*(12), 665-673. doi:10.1007/s10822-007-9124-0

Liu, R., Schyman, P., & Wallqvist, A. (2015). Critically Assessing the Predictive Power of QSAR Models for Human Liver Microsomal Stability. *J Chem Inf Model, 55*(8), 1566-1575. doi:10.1021/acs.jcim.5b00255

NCATS. LyChI. https://github.com/ncats/lychi/. doi:https://github.com/ncats/lychi

NCATS. Resolver. https://tripod.nih.gov/servlet/resolver/.

Optibrium. StarDrop. www.optibrium.com/stardrop. Retrieved from www.optibrium.com/stardrop

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). *How Many Trees in a Random Forest?*, Berlin, Heidelberg.

Patel, M., Chilton, M. L., Sartini, A., Gibson, L., Barber, C., Covey-Crump, L., Przybylak, K. R., Cronin, M. T. D., & Madden, J. C. (2018). Assessment and Reproducibility of Quantitative Structure–Activity Relationship Models by the Nonexpert. *Journal of Chemical Information and Modeling, 58*(3), 673-682. doi:10.1021/acs.jcim.7b00523

Prueksaritanont, T., Chu, X., Gibson, C., Cui, D., Yee, K. L., Ballard, J., Cabalu, T., & Hochman, J. (2013). Drug–Drug Interaction Studies: Regulatory Guidance and An Industry Perspective. *The AAPS Journal, 15*(3), 629-645. doi:10.1208/s12248-013-9470-x

Roden, D. M., & George, A. L., Jr. (2002). The genetic basis of variability in drug responses. *Nature reviews Drug discovery, 1*(1), 37-44. Retrieved from <Go to ISI>://MEDLINE:12119608

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics, 5*(1), 27. doi:10.1186/1758-2946-5-27

Sakiyama, Y., Yuki, H., Moriya, T., Hattori, K., Suzuki, M., Shimada, K., & Honma, T. (2008). Predicting human liver microsomal stability with machine learning techniques. *Journal of molecular graphics & modelling, 26*(6), 907-915. Retrieved from <Go to ISI>://MEDLINE:17683964

Sansone-Parsons, A., Krishna, G., Simon, J., Soni, P., Kantesaria, B., Herron, J., & Stoltz, R. (2007). Effects of age, gender, and race/ethnicity on the pharmacokinetics of posaconazole in healthy volunteers. *Antimicrobial agents and chemotherapy, 51*(2), 495-502. Retrieved from <Go to ISI>://MEDLINE:17101682

Shah, P., Kerns, E., Nguyen, D.-T., Obach, R. S., Wang, A. Q., Zakharov, A., McKew, J., Simeonov, A., Hop, C. E. C. A., & Xu, X. (2016). An Automated High-Throughput Metabolic Stability Assay Using an Integrated High-Resolution Accurate Mass Method and Automated Data Analysis Software. *Drug metabolism and disposition: the biological fate of chemicals, 44*(10), 1653-1661. doi:10.1124/dmd.116.072017

Siramshetty, V. B., Shah, P., Kerns, E., Nguyen, K., Yu, K. R., Kabir, M., Williams, J., Neyra, J., Southall, N., Nguyễn, Đ.-T., & Xu, X. (2020). Retrospective assessment of rat liver microsomal stability at NCATS: data and QSAR models. *Scientific Reports, 10*(1), 20713. doi:10.1038/s41598-020-77327-0

Solutions, K. Dragon. https://chm.kode-solutions.net/products_dragon.php.

Sushko, I., Novotarskyi, S., Korner, R., Pandey, A. K., Cherkasov, A., Lo, J. Z., Gramatica, P., Hansen, K., Schroeter, T., Muller, K. R., Xi, L. L., Liu, H. X., Yao, X. J., Oberg, T., Hormozdiari, F., Dao, P. H., Sahinalp, C., Todeschini, R., Polishchuk, P., Artemenko, A., Kuz'min, V., Martin, T. M., Young, D. M., Fourches, D., Muratov, E., Tropsha, A., Baskin, I., Horvath, D., Marcou, G., Muller, C., Varnek, A., Prokopenko, V. V., & Tetko, I. V. (2010). Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *Journal of Chemical Information and Modeling, 50*(12), 2094-2111. doi:10.1021/ci100253r

Tetko, I. V., Novotarskyi, S., Sushko, I., Ivanov, V., Petrenko, A. E., Dieden, R., Lebon, F., & Mathieu, B. (2013). Development of dimethyl sulfoxide solubility models using 163,000 molecules: using a domain applicability metric to select more reliable predictions. *Journal of Chemical Information and Modeling, 53*(8), 1990-2000. Retrieved from <Go to ISI>://MEDLINE:23855787

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular informatics, 29*(6-7), 476-488. Retrieved from <Go to ISI>://MEDLINE:27463326

TT, T. (Ed.) (1958). *An elementary mathematical theory of classification and prediction*.

Veith, H., Southall, N., Huang, R., James, T., Fayne, D., Artemenko, N., Shen, M., Inglese, J., Austin, C. P., Lloyd, D. G., & Auld, D. S. (2009). Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat Biotechnol, 27*(11), 1050-1055. doi:10.1038/nbt.1581

Wenzel, J., Matter, H., & Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling, 59*(3), 1253-1268. doi:10.1021/acs.jcim.8b00785

Wildman, S. A., & Crippen, G. (1999). Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci., 39*, 868-873.

Yeh, S. Y. (1991). Metabolic profile of tripelennamine in humans. *Journal of pharmaceutical sciences, 80*(9), 815-819. Retrieved from <Go to ISI>://MEDLINE:1800700

Yun, Y. H., Wu, D. M., Li, G. Y., Zhang, Q. Y., Yang, X., Li, Q. F., Cao, D. S., & Xu, Q. S. (2017). A strategy on the definition of applicability domain of model based on population analysis. *Chemometrics and Intelligent Laboratory Systems, 170*, 77-83. doi:10.1016/j.chemolab.2017.09.007

Zakharov, A., Gonzalez, E., Shah, P., Nguyen, D. T., Southall, N., Torimoto-Katori, N., Sakamuru, S., Xia, M. H., Zhao, T. G., Obach, R. S., Hop, C., Simeonov, A., & Xu, X. (2019). AI-driven QSAR modeling of P450-mediated drug metabolism. *Abstracts of Papers of the American Chemical Society, 257*. Retrieved from <Go to ISI>://WOS:000478860504830

Zakharov, A. V., Peach, M. L., Sitzmann, M., Filippov, I. V., McCartney, H. J., Smith, L. H., Pugliese, A., & Nicklaus, M. C. (2012). Computational tools and resources for metabolism-related property predictions. 2. Application to prediction of half-life time in human liver microsomes. *Future Med Chem, 4*(15), 1933-1944. doi:10.4155/fmc.12.152

Zakharov, A. V., Zhao, T., Nguyen, D.-T., Peryea, T., Sheils, T., Yasgar, A., Huang, R., Southall, N., & Simeonov, A. (2019). Novel Consensus Architecture To Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *Journal of Chemical Information and Modeling, 59*(11), 4613-4624. Retrieved from <Go to ISI>://MEDLINE:31584270

Zientek, M. A., & Youdim, K. (2015). Reaction phenotyping: advances in the experimental strategies used to characterize the contribution of drug-metabolizing enzymes. *Drug metabolism and disposition: the biological fate of chemicals, 43*(1), 163-181. Retrieved from <Go to ISI>://MEDLINE:25297949

Footnotes:

This research was supported by the Intramural Research Program of the National Institutes of Health, National Center for Advancing Translational Sciences.

The authors declare no conflict of interest.

Figure Legends

Figure 1: Reaction schemes for most common scenarios in metabolic clearance and P450-Glo assays. A. The clearance assay will identify the substrates that proceed through typical Michaelis-Menten kinetics (1), or through multi-ligand processes (2), but is unable to identify either competitive (3) or mechanism-based (4) inhibitors. B. Rationale for categorizing P450-Glo assay hits as either substrate or inhibitor based on cross-referencing the observations from the two assays. A test article is to be able to occlude the pro-luciferin probe by competing for the substrate pocket (i and ii left) but may also generate an inhibitory multi-ligand complex (ii middle and right). Alternatively, the test article may simply be a poor ligand (iii left), form a non-inhibitory multi-ligand complex (iii right), or exhibit efficient clearance that does not impede probe metabolism (iv).

Figure 2: Property distributions including (A) molecular weight, (B) TPSA and (C) SlogP for the entire dataset compared to substrates ($t_{1/2}$ <30 mins) of CYP3A4, CYP2C9 and CYP2D6 in the supersome clearance assay.

Figure 3: Substrate (A) and inhibitor (B) overlap amongst three enzymes (C) Chemical space plot example for CYP2C9.

Figure 4: 5-Fold cross-validation results from SB with Morgan fingerprints- (A) AUC and (B) Balanced accuracy (BACC).

Figure 5: Distribution of prediction results for test set over AD cut-offs and coverage values using CYP2C9 substrate dataset as an example.

Figure 6: Distribution of substrates and inhibitors over class-probability prediction values.

Figure 7: Comparison of model performance on NCATS annotated library.

Figure 8: Plot presenting the most frequent Bemis-Murcko scaffolds in our dataset and relative percentages of frequencies within.

Tables

Table 1: Summary of enzyme concentrations, cofactor activities and controls used in the metabolic stability assays

Matrix	Final Protein Concentration [mg/mL]	Total CYP content [nM]	Cytochrome c Reductase activity [nmol/(min x mg protein)]	Cytochrome b_5 content [pmol/mg protein]	High clearance controls	Moderate clearance controls	Low clearance controls
CYP3A4	~0.2	30	2900	1090	Buspirone, Loperamide	Ketoconazole	Antipyrine, Carbamazepine
CYP2C9	~0.12	45	985	710	Glyburide, Glimepiride	Tamoxifen	Antipyrine, Meloxicam
CYP2D6	~0.38	60	3000	-	Bufuralol, Desipramine, Amitriptyline	Mexiletine,	Codeine

Table 2: Categorization of clearance data

Observed $t_{1/2}$	**Category**	**Substrate Class**
$t_{1/2} \leq 30$ min	Unstable	1
$t_{1/2} > 30$ min	Stable	0
low signal [a]	Below Limit of Quantification (BLQ)	Blank
unable to reasonably fit line to data [b]	Inconclusive (INC)	
analyte not detected in mass spectrometer	Not Found (N/F)	

[a] *y-intercept of line fitting the Ln(analyte/IS) vs time plot is ≤ -9.0*

[b] *Albendazole also assigned as INC due to its use as IS*

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Table 3: Summary of incubation conditions and positive controls used in the P450-Glo assays

Enzyme	**Inhibitor**	**Dilution Format**	**Inhibitor Concentration**	**Incubation conditions**
CYP3A4	Ketoconazole	16 concentrations/ 2-fold dilution in duplicates	57 µM-1.8 nM	1 h/RT
CYP2C9	Sulfaphenazole		57 µM-1.8 nM	1 h/ 37˚C/ 0.4% BSA
CYP2D6	Quinidine		1.4 µM-0.04 nM	1 h/RT

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Table 4: Summary of substrate and inhibitor datasets used in this study

Type	Dataset name	Total number of compounds	Number of actives	Number of inactives	Imbalance ratio (inactives/actives)
Substrate Data	CYP2C9	3966	1126	2840	3:1
	CYP2D6	3946	1318	2628	2:1
	CYP3A4	3974	1883	2091	1:1
Inhibitor Data	CYP2C9	3288	570	2718	5:1
	CYP2D6	3187	367	2820	8:1
	CYP3A4	2794	340	2454	7:1

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Table 5: Parsing rationale for substrate and inhibitors

Category	Clearance / P450-Glo	Classification	Parsing Rationale
1	+ / +	Substrate	Exhibiting activity in both assays, the compound is a clear ligand for the enzyme(s). It is unclear whether the parent, product, or both are responsible for the inhibition.
2	- / +	Inhibitor	The compound is able to inhibit the enzyme metabolism of a probe substrate, but is not itself cleared, indicating that the parent molecule is an effective inhibitor.
3	- / -	Non-competitor	The lack of activity in both assays signifies that either binding does not occur, or that the interaction does not generate a catalytically competent or inhibitory complex.
4	+ / -	Substrate	Although a clear substrate, the binding kinetics of the parent compound and its metabolites do not preclude the concomitant metabolism of the P450-Glo probe.

Table 6: Percentage of high clearance compounds across three CYP enzymes.

	CYP3A4	CYP2C9	CYP2D6
CYP3A4	47%	21%	19%
CYP2C9		28%	13%
CYP2D6			33%

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Table 7: Percentage of P450-Glo hits across three enzymes

	CYP3A4	CYP2C9	CYP2D6
CYP3A4	35%	19%	13%
CYP2C9		29%	10%
CYP2D6			25%

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

Table 8: Summary of Cross Validation Results

		BACC					
		Substrates			Inhibitors		
Classifiers	**Descriptors**	**CYP2C9**	**CYP2D6**	**CYP3A4**	**CYP2C9**	**CYP2D6**	**CYP3A4**
SB	Morgan + PhysChem	0.69	0.71	0.75	0.74	0.64	0.68
SB	Avalon + PhysChem	0.67	0.69	0.72	0.73	0.63	0.64
SB	Dragon_Normalized	0.68	0.69	0.76	0.72	0.62	0.67
DNN	Morgan + PhysChem	0.62	0.67	0.72	0.63	0.58	0.58
DNN	Avalon + PhysChem	0.64	0.66	0.68	0.63	0.59	0.58
DNN	Dragon_Normalized	0.66	0.68	0.72	0.63	0.57	0.59

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

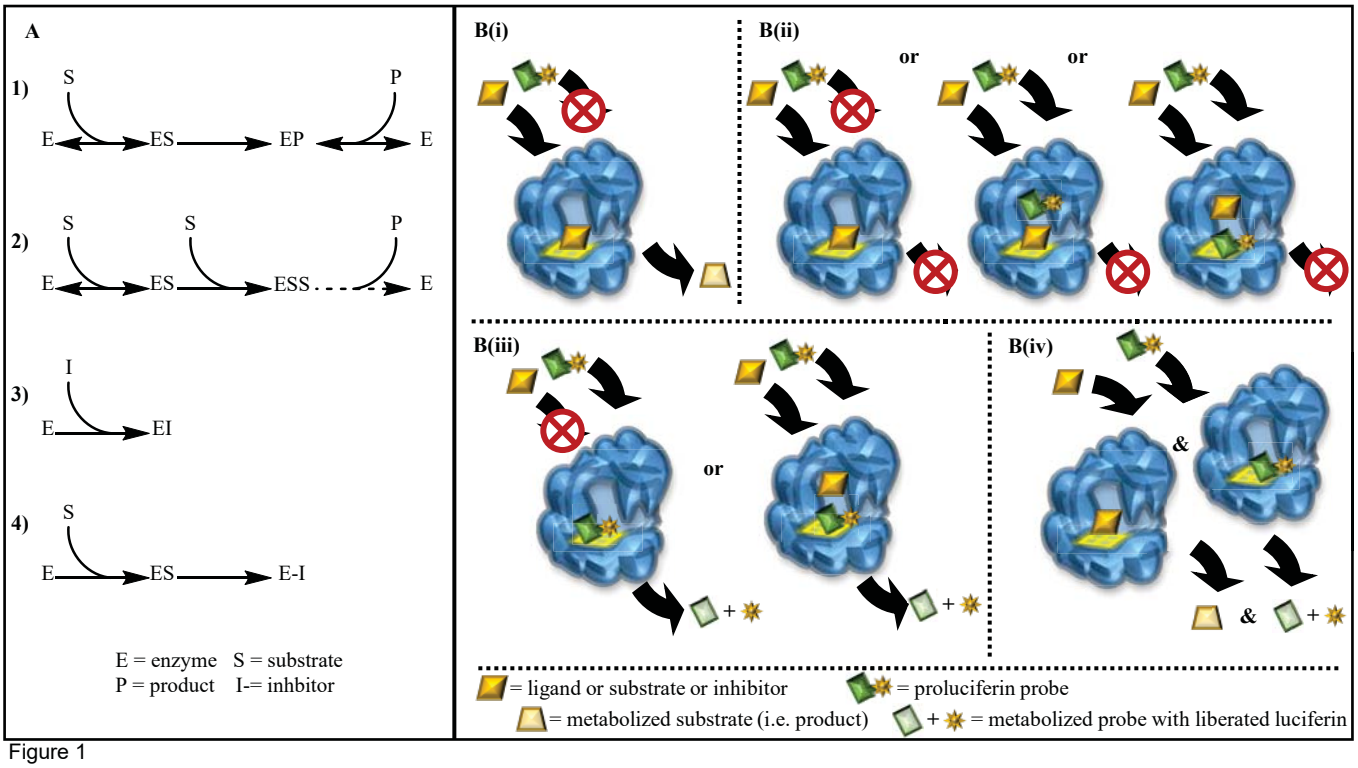


Figure 1

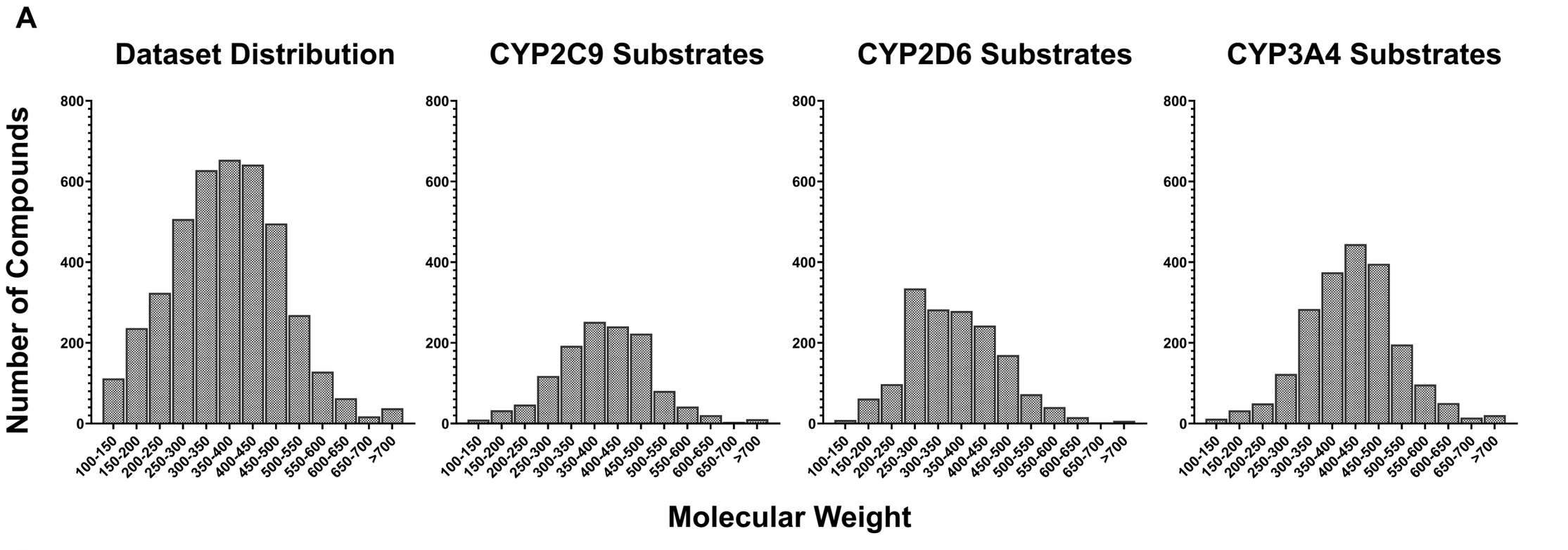
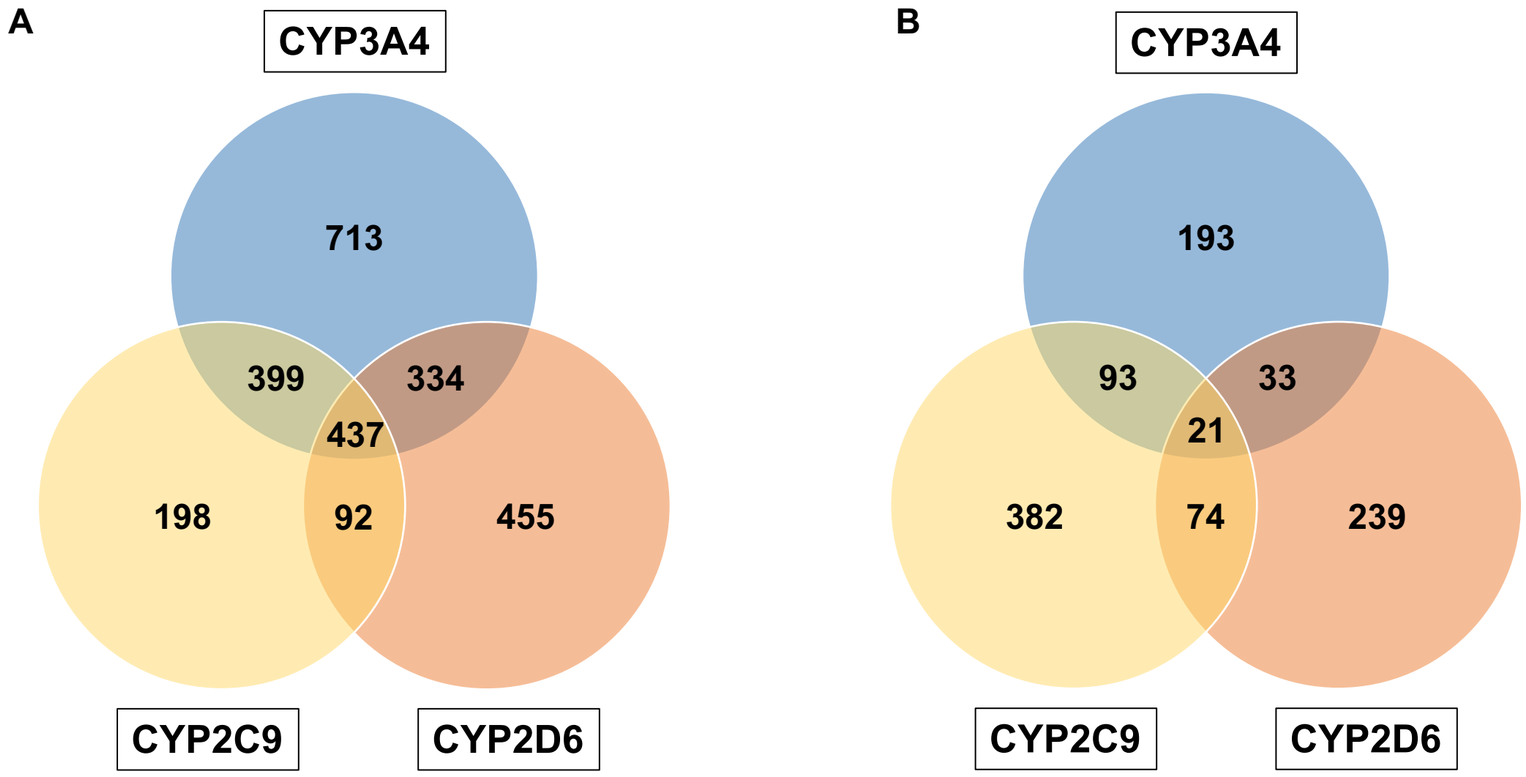


Figure 2



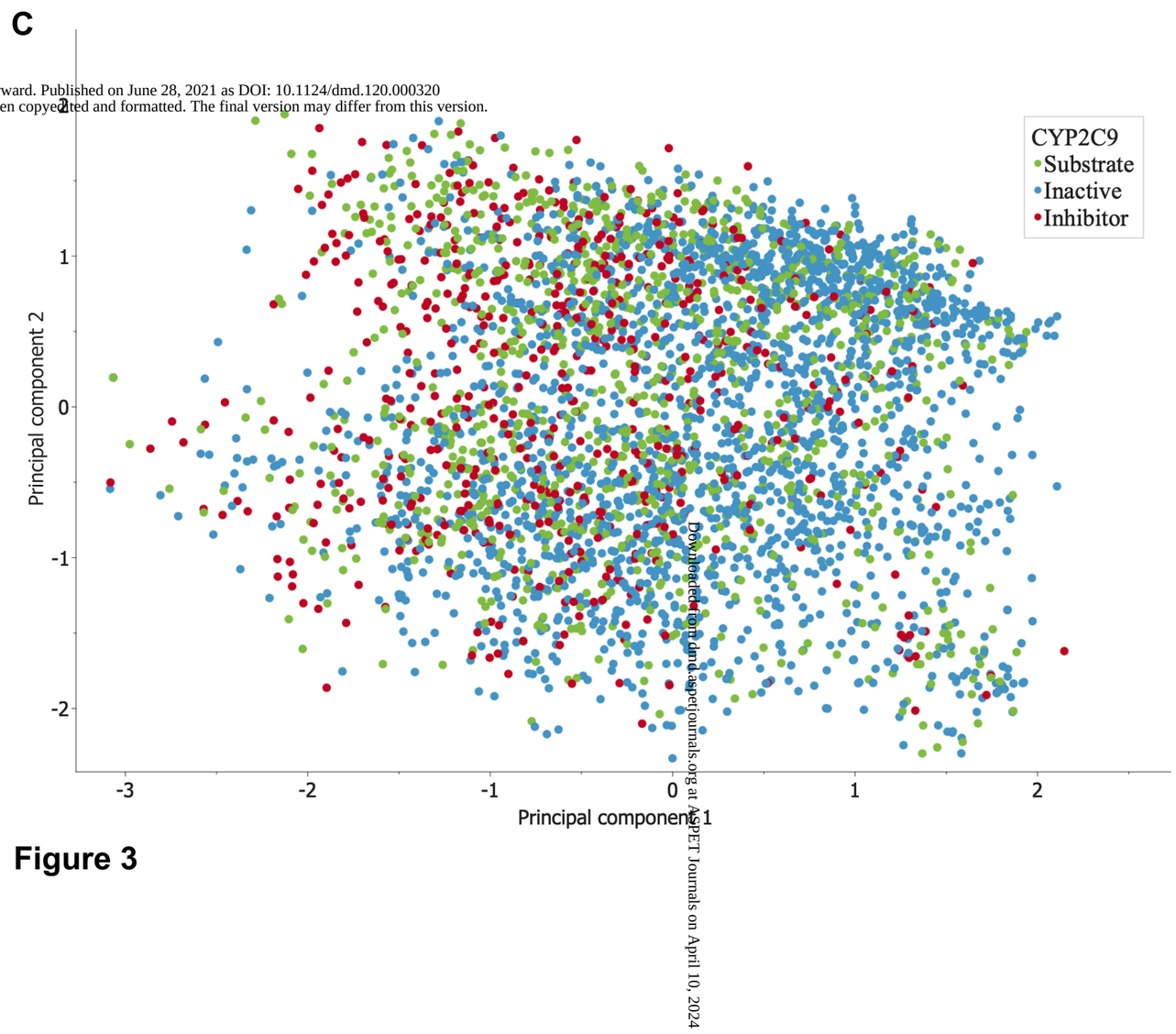


Figure 3

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

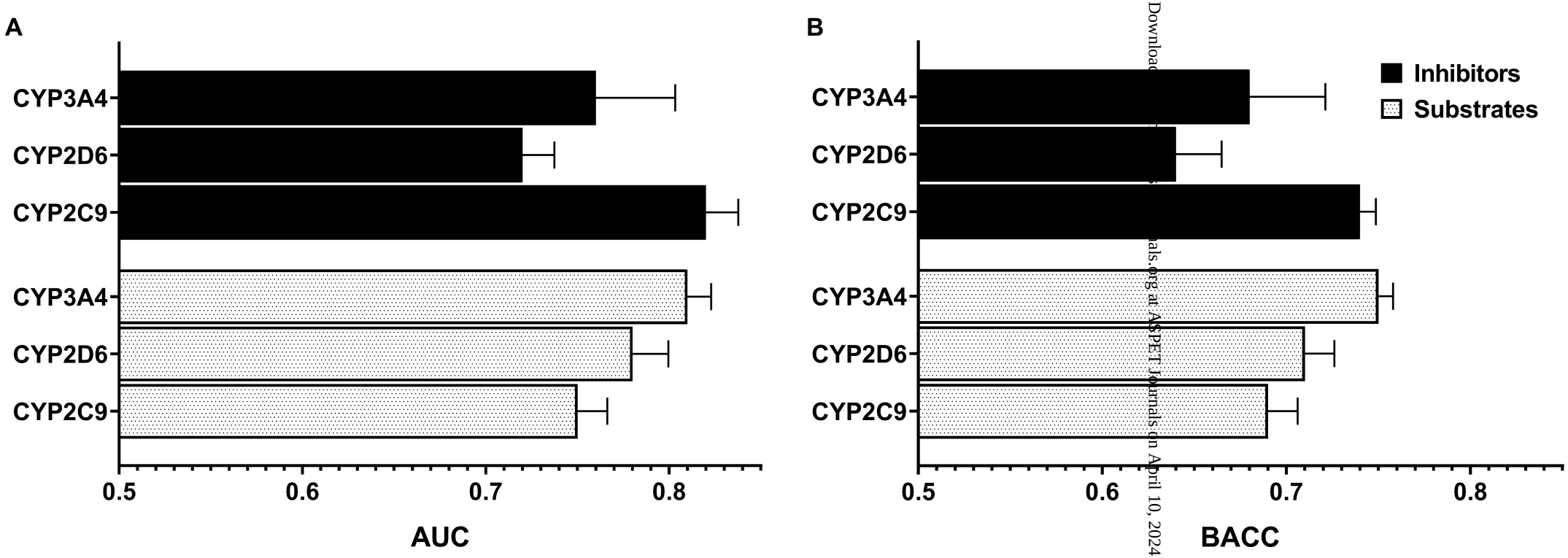


Figure 4

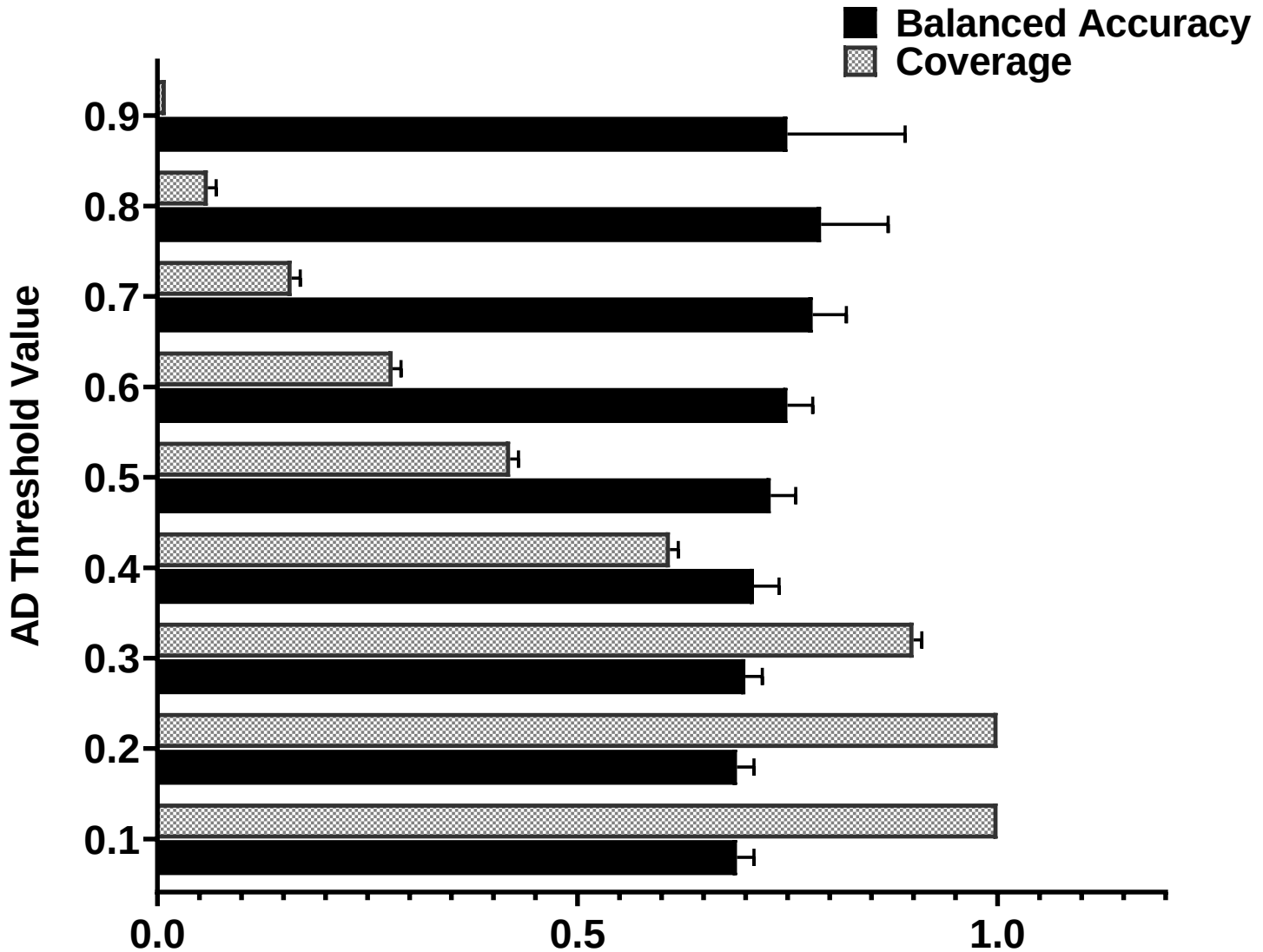

Balanced Accuracy
Coverage
AD Threshold Value
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.5
1.0

Figure 5

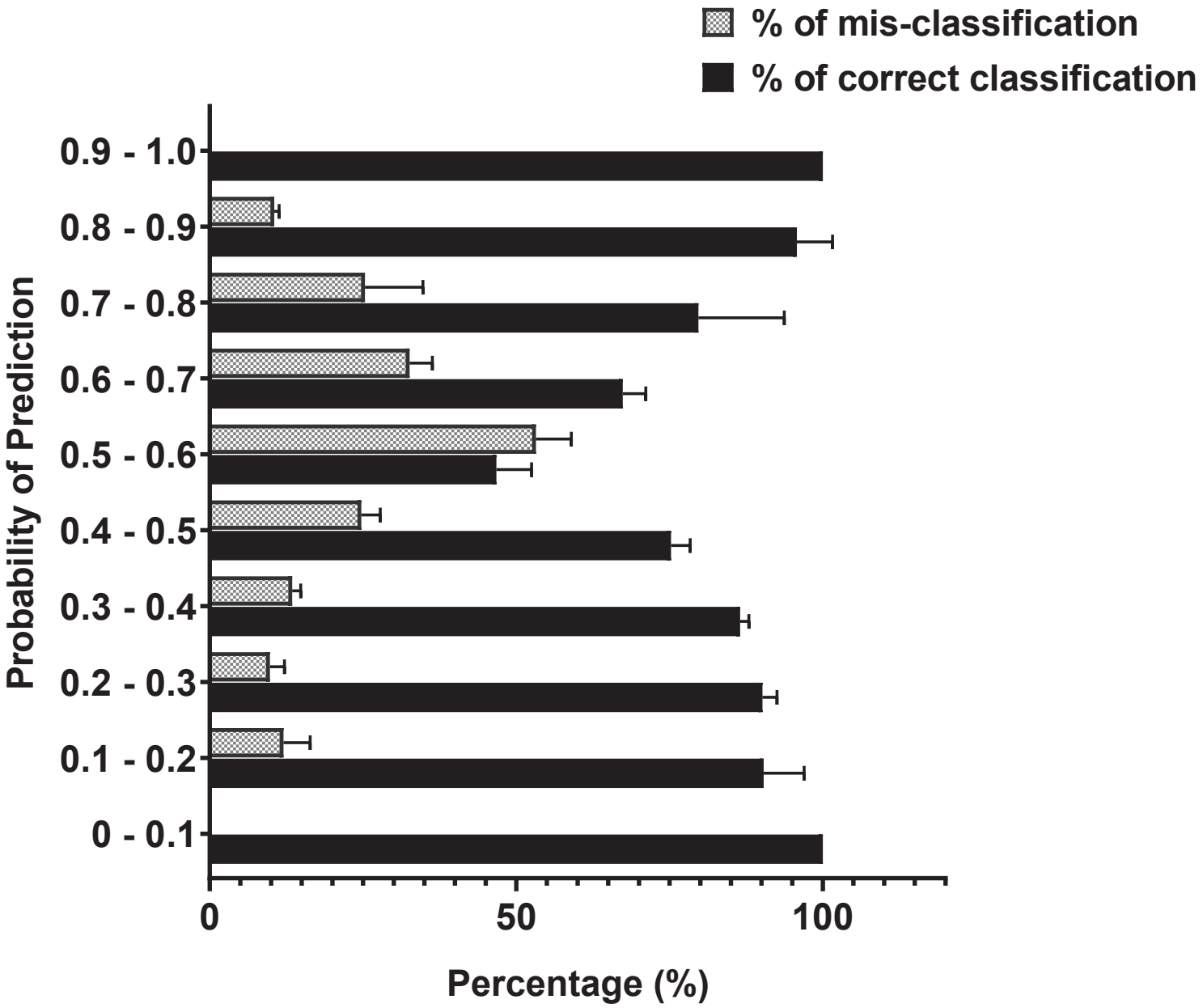

% of mis-classification
% of correct classification
Probability of Prediction
0.9 - 1.0
0.8 - 0.9
0.7 - 0.8
0.6 - 0.7
0.5 - 0.6
0.4 - 0.5
0.3 - 0.4
0.2 - 0.3
0.1 - 0.2
0 - 0.1
0
50
100
Percentage (%)

Figure 6

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

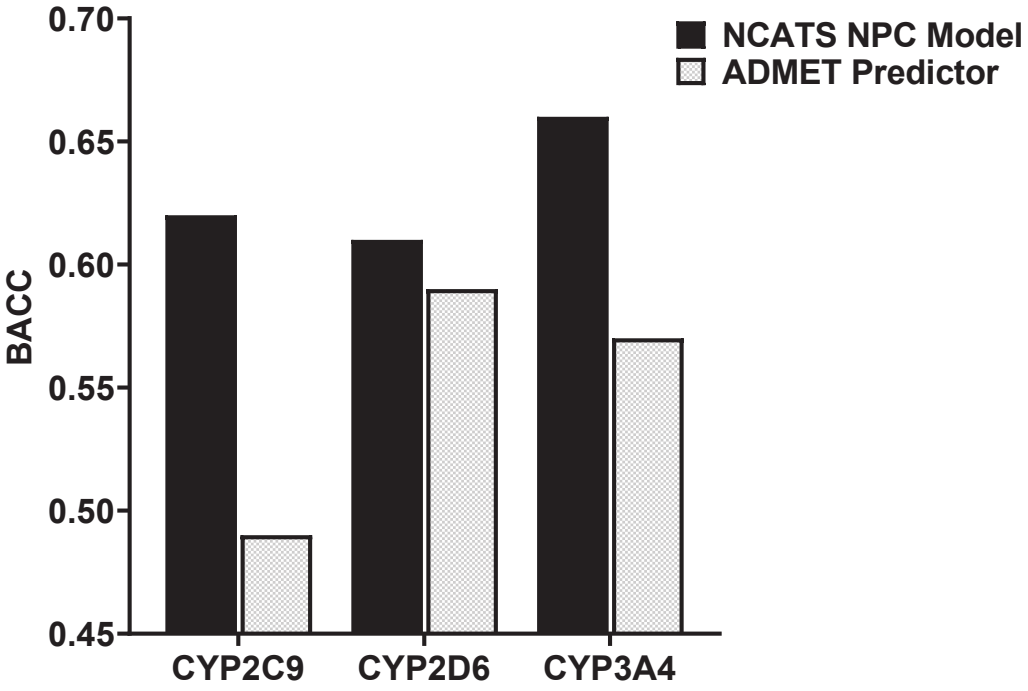


Figure 7

DMD Fast Forward. Published on June 28, 2021 as DOI: 10.1124/dmd.120.000320
This article has not been copyedited and formatted. The final version may differ from this version.

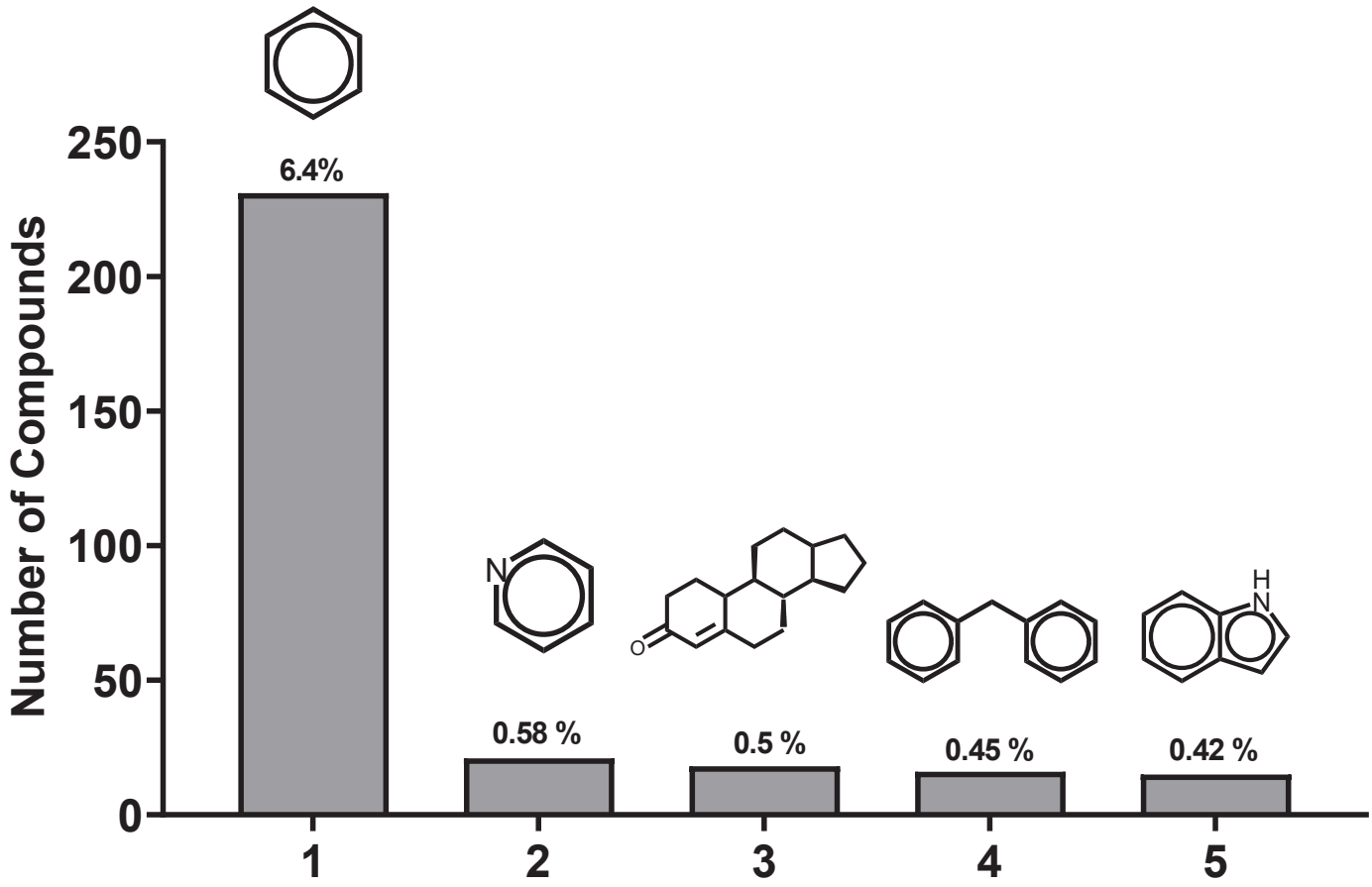


Figure 8