Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-018-0107-0

# WhichP450: a multi-class categorical model to predict the major metabolising CYP450 isoform for a compound

**Peter A. Hunt**[1] · **Matthew D. Segall**[1] · **Jonathan D. Tyzack**[2]

Received: 2 October 2017 / Accepted: 15 February 2018


**Abstract**
In the development of novel pharmaceuticals, the knowledge of how many, and which, Cytochrome P450 isoforms are involved in the phase I metabolism of a compound is important. Potential problems can arise if a compound is metabolised predominantly by a single isoform in terms of drug–drug interactions or genetic polymorphisms that would lead to variations in exposure in the general population. Combined with models of regioselectivities of metabolism by each isoform, such a model would also aid in the prediction of the metabolites likely to be formed by P450-mediated metabolism. We describe the generation of a multi-class random forest model to predict which, out of a list of the seven leading Cytochrome P450 isoforms, would be the major metabolising isoforms for a novel compound. The model has a 76% success rate with a top-1 criterion and an 88% success rate for a top-2 criterion and shows significant enrichment over randomised models.

**Keywords** Multi-class classification · Random forests · Cytochrome P450 · Drug–drug interactions · Metabolism

## Introduction

In the development of new chemical entities (NCEs) within the pharmaceutical industry, the knowledge of the metabolic fate of a molecule in vivo is paramount to the successful progression of that compound through the pre-clinical and clinical phases. Phase I metabolism by Cytochrome P450s (CYP450s) modifies a compound, usually by the addition of oxygen, to make it more polar and hence easier to excrete. This modification can lead to reactive or unstable metabolites that can react further with neighbouring proteins or glutathione or rearrange to form further products such as de-alkylated derivatives. The ability of CYP450s to modify an NCE depends on, not only the nature of the NCE, but also on the induction or inhibition of their actions by other xenobiotics. Furthermore, the expression levels and amino acid sequence for certain CYP450s are not homogeneous across the population and so the systemic concentrations of an NCE could vary greatly between individuals in the clinic [1]. Generally, it is desirable for an NCE to be metabolised by more than one isoform in order to minimise the effects of the above dependencies, however activity across too many isoforms would make metabolism and excretion too facile a process and reduce the desired efficacy of the compound. This balance needs to be struck during the research and development of an NCE.

We have previously developed models to predict the atomic sites of metabolism that a CYP450 isoform is likely to attack *if* the NCE is a substrate for that isoform [2]. The regioselectivity of metabolism, and hence the resulting metabolites, can vary significantly between isoforms. Therefore, the prediction of which isoform-specific regioselectivity model(s) are most relevant to the metabolism of an NCE would aid in the application of these models.

In this paper, we will describe the data and methods used to build a model that ranks seven drug-metabolising CYP450 isoforms (CYP3A4, CYP2D6, CYP1A2, CYP2E1, CYP2C8, CYP2C9 and CYP2C19) in order of likelihood that they are a major metabolising isoform for an input compound. These isoforms cover > 85% of those clinically used drugs that are metabolised by CYP450s [3]. The model is

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10822-018-0107-0) contains supplementary material, which is available to authorized users.

✉ Peter A. Hunt
peter@optibrium.com

[1] Optibrium Ltd., F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge CB25 9PB, UK

[2] The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge CB10 1SD, UK

Published online: 20 February 2018

based on carefully curated experimental data that clearly identify the isoform(s) responsible for metabolism, confirmed with in vivo clinical observations. We will test the performance of the model predictions using a fivefold cross-validation procedure and compare against randomised predictions to demonstrate significant enrichment.

Other models to predict interactions with CYP isoforms have previously been described and are summarised in a review by Kirchmair et al. [4]. The modelling techniques previously applied include random forests, support vector machine, or recursive partitioning to classify molecules and partial least squares or multiple linear regression for prediction of $K_i$ or $K_m$. Many of these, such as WhichCYP [5] are based on binding or inhibition data rather than specifically substrate data. While the corresponding data sets are larger, compounds that inhibit or bind to a CYP450 isoform are not necessarily substrates thereof. Other models have utilised higher throughput, in vitro, liver microsomal data generated against the individual isoforms, without reference to the human in vivo clinical data; we have found this to be misleading in many cases due to artefacts generated by assay conditions that are not physiologically relevant. Furthermore, most other models are only able to predict for a smaller number of isoforms than those studied herein (e.g. five isoforms for Percepta [6], WhichCYP [5] or MetaPred [7, 8]) or for a limited set of metabolism transformations (e.g. Percepta [6]).

## Materials and methods

### Data set

The data set used herein was generated during the data gathering for our recently published models of P450 regioselectivity [2], for which we annotated the molecules with which isoforms have a major or minor influence on the metabolites formed. Obviously, an NCE can be metabolised at more than one site and each CYP450 isoform can contribute to the metabolism, to a greater or lesser extent, at each of those sites. In the models described herein and their analyses we have considered only isoforms considered to have a major contribution to the metabolism of a compound. The assigning of any isoform as a major metaboliser for a compound is something of a subjective demarcation based upon the overall extent of metabolism and the views expressed by the original authors in their publications.

The resulting data set contains 465 unique compounds. Each compound may be associated with significant metabolism by multiple isoforms, therefore the data set contains 633 compound/major isoform pairs. The distribution of the number of isoforms associated with the metabolism of each individual compound is shown in Fig. 1a.

The distribution of the numbers of observations for each isoform in the data set is shown below in Fig. 2 and compared with the proportions of drugs metabolised by each isoform, as described by Zanger and Schwab [3]. From this, we can see that the relative proportions in the data set largely agree with those reported, with the only notable exceptions being that CYP2C19, CYP2E1, and CYP1A2 may be slightly overrepresented compared with CYP3A4. This could be because there has been more interest in the CYP2C19, CYP1A2 and CYP2E1 isoforms in the literature due to the association of CYP2C19 with genetic polymorphism [9] and the association of CYP1A2 and CYP2E1 with bioactivation of carcinogens and hepatotoxins [10, 11].

The full data set was split into five separate training and test set combinations, such that each compound was found in only one of the test sets. A structurally diverse subset of approximately 20% of the data was chosen as the first test set, by maximising the minimum Tanimoto distance between the selected compounds (calculated using 2D fingerprints). The remaining compounds were used as the first training set, as illustrated in Fig. 3. The first test set was marked and the selection process was repeated to select the next 20% of the data set, such that it did not overlap with the first test set and yet was also diverse across the whole data set. The remaining training/test set splits were created in a similar manner.

The distributions of occurrence for the seven isoforms are reasonably consistent across the five training/test set splits, as shown in Fig. 4. Similarly, the distribution of the numbers of major isoforms that contribute significantly to the metabolism of each compound in the test sets are shown in Fig. 1b. Preserving these distributions was not an active feature of the selection process, but it confirms that the selection of test sets has not greatly biased one isoform over another or the distribution of numbers of isoforms associated with each compound.

### Model training

Training sets, each comprising 80% of the whole data set, as detailed above, were used to build five models. In each case the training sets contained one row for each compound/isoform pair, i.e. the same compound could occur multiple times in the training set if it was associated with multiple isoforms. A seven-class random forest classification model [13], where each isoform was a separate class, was produced for each training set. This was performed using the AutoModeller™ within the StarDrop software [14] and 100 trees were generated for each model. The resulting models output probabilities of a novel compound being in each of the isoform classes, where the probabilities were determined by the proportion of trees classifying the compound into that class.

The descriptors used to train the models included whole molecule descriptors, such as the logarithm of the

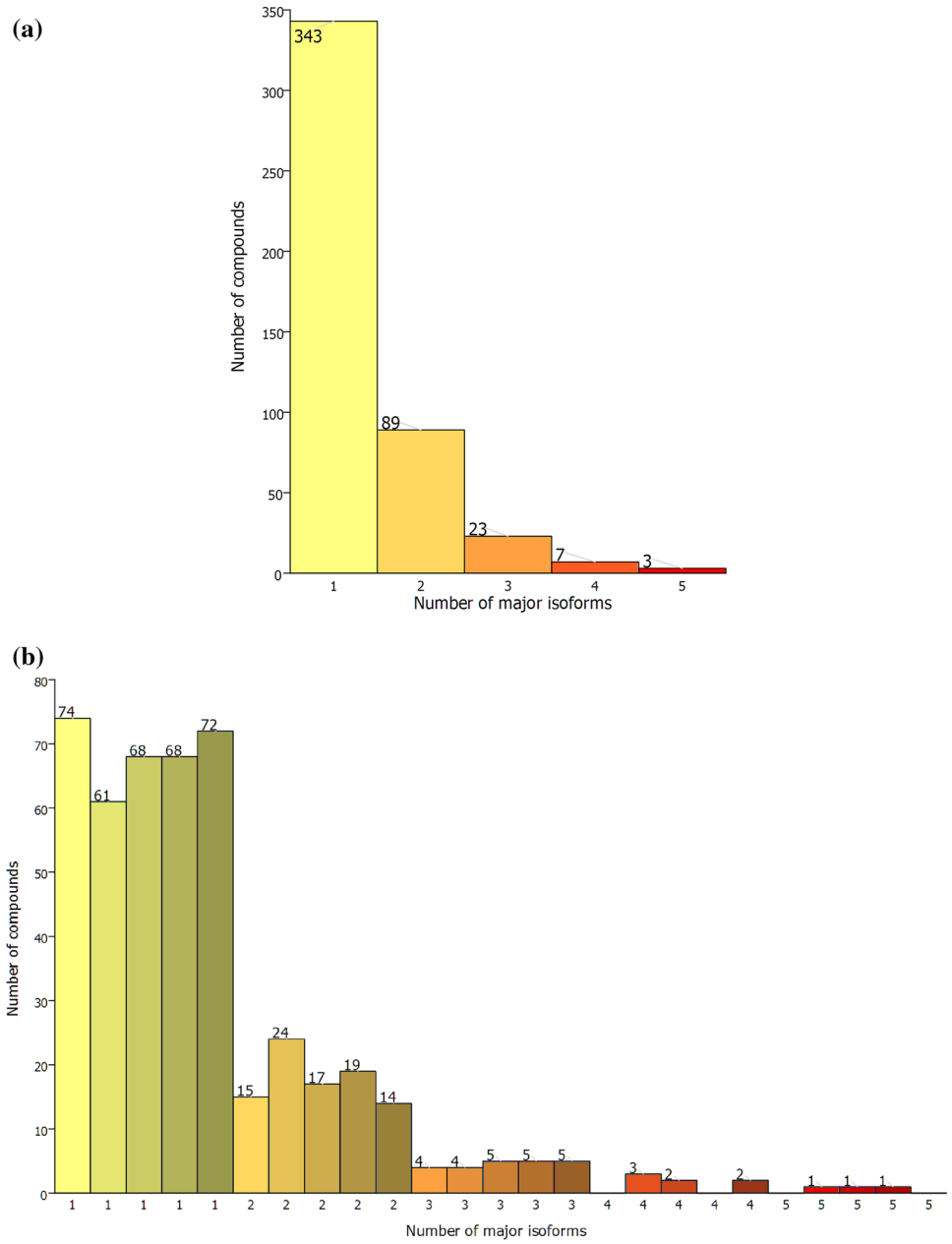


**Fig. 1** Distributions of the numbers of isoforms identified as major metabolising enzymes for each compound: (**a**) shows the distribution across the whole data set; (**b**) shows the distributions for each test set split. The histogram bars in (**b**) are coloured consistently with those in (**a**) for reference and the shading indicates the different test set splits

octanol:water coefficient (logP), molecular weight (MW), the McGowan volume [15] ($V_x$), proportion of rotatable bonds, and counts of 290 structural descriptors encoded as SMARTS.

## Model validation

The independent test sets comprising 20% of the whole data set were used to test the corresponding models. Each of the test sets had one row per compound and the target column for prediction contained a list of the isoforms considered as major isoforms contributing to the metabolism of that compound. The rank-ordered list of probabilities of an isoform contributing to the metabolism of each test compound, output by a model, was compared with the list of observed isoforms to assess the performance of the model.

The top-*k* accuracy of each model was assessed, whereby a successful prediction was deemed to be one where at least one isoform predicted in the top *k* ranked isoforms matched any of the observed isoforms listed in the target column. The distribution of the numbers of major isoforms associated with the compounds in each of the test sets is shown in Fig. 1b. Obviously, the more major isoforms that are listed for a test compound, the easier it is to obtain a successful prediction; however, the distributions for the number of isoforms per compound are reasonably consistent across the

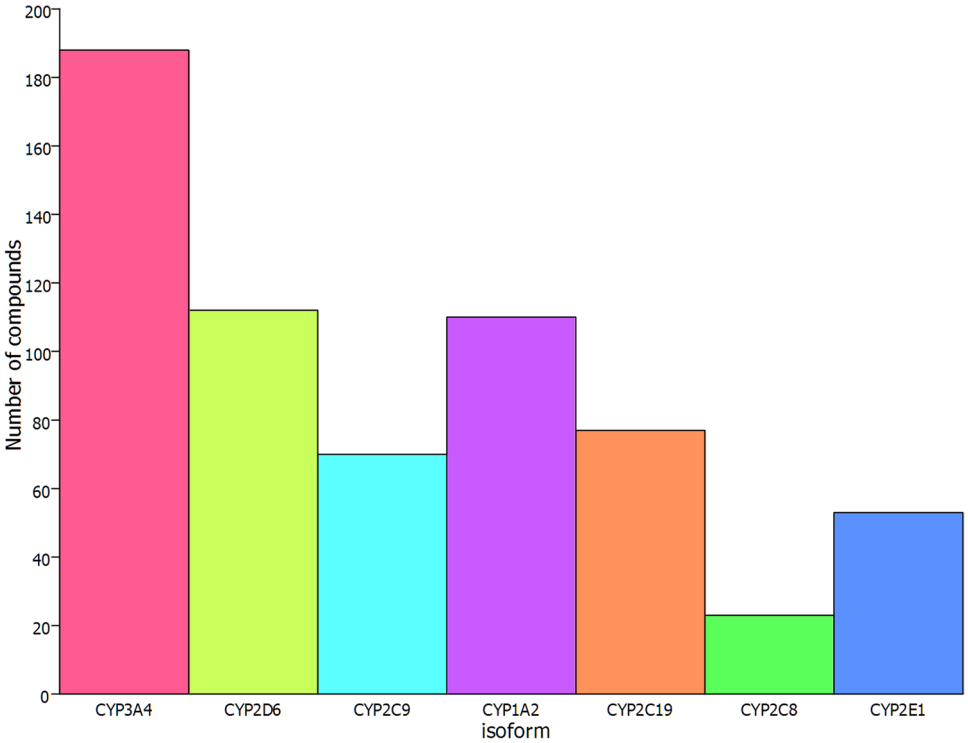


**Fig. 2** The numbers of compounds metabolised by each isoform in the data set. The isoforms are listed in the order of the frequency with which they are observed to metabolise marketed drugs, from the most common (left) to least common (right), as determined by Zanger and Schwab [3]

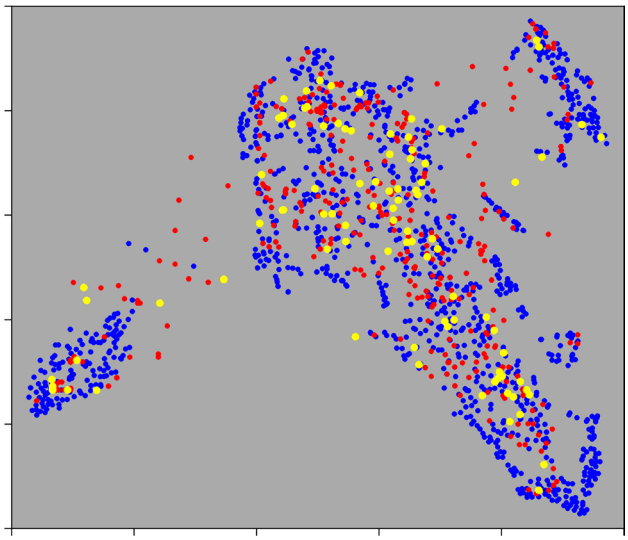

**Fig. 3** A chemical space representation of approximately 1300 launched drugs (blue points) with the compounds in the full CYP450 metabolism data set overlaid (red and yellow points). The compounds from one of the test sets are indicated by the yellow points. This shows that the data set used herein covers a large majority of small molecule drug space. In this chemical space plot, the proximity of two points represents the structural similarity between the corresponding compounds defined using a Tanimoto index based on a 2D path-based fingerprint. The distribution of points is generated using the t-distributed stochastic neighbour embedding algorithm [12]

five sets and hence the performance statistics should also be consistent.

As multiple major isoforms are associated with many compounds, it is also useful to assess how well a model orders the list of predicted isoforms, such that the major isoforms are at the top of the list. A way of assessing this ranking is to compute an area under the curve (AUC) of a receiver operating characteristic (ROC) curve. This can be calculated for each compound in the test set by running through the ordered prediction list and counting a true positive for each correct isoform found and a false positive for every incorrect isoform. An AUC of one for a compound would indicate a perfect prediction, and AUC of 0.5 corresponds to the expected performance of random selection and less than 0.5 would be worse than random. The average AUC for the compounds in the test set gives a measure of the overall performance of the model. This protocol is analogous to the method used to assess our published CYP450 regioselectivity models [2].

Another method to evaluate the predictive capabilities of the models is to compare their performances against those expected for random selections of isoforms. The success rates for four approaches for 'random' estimation, which include different levels of prior knowledge, were used for comparison with the test set results for the models. The simplest approach ('Uniform Random') assumed no prior information and therefore chooses isoforms in random order with equal probability (i.e. 1/7). In this case, the expected top-*k* performance of such a Uniform Random model can be calculated analytically (as described in the Supplementary Information and labelled as 'Expected Uniform Random'

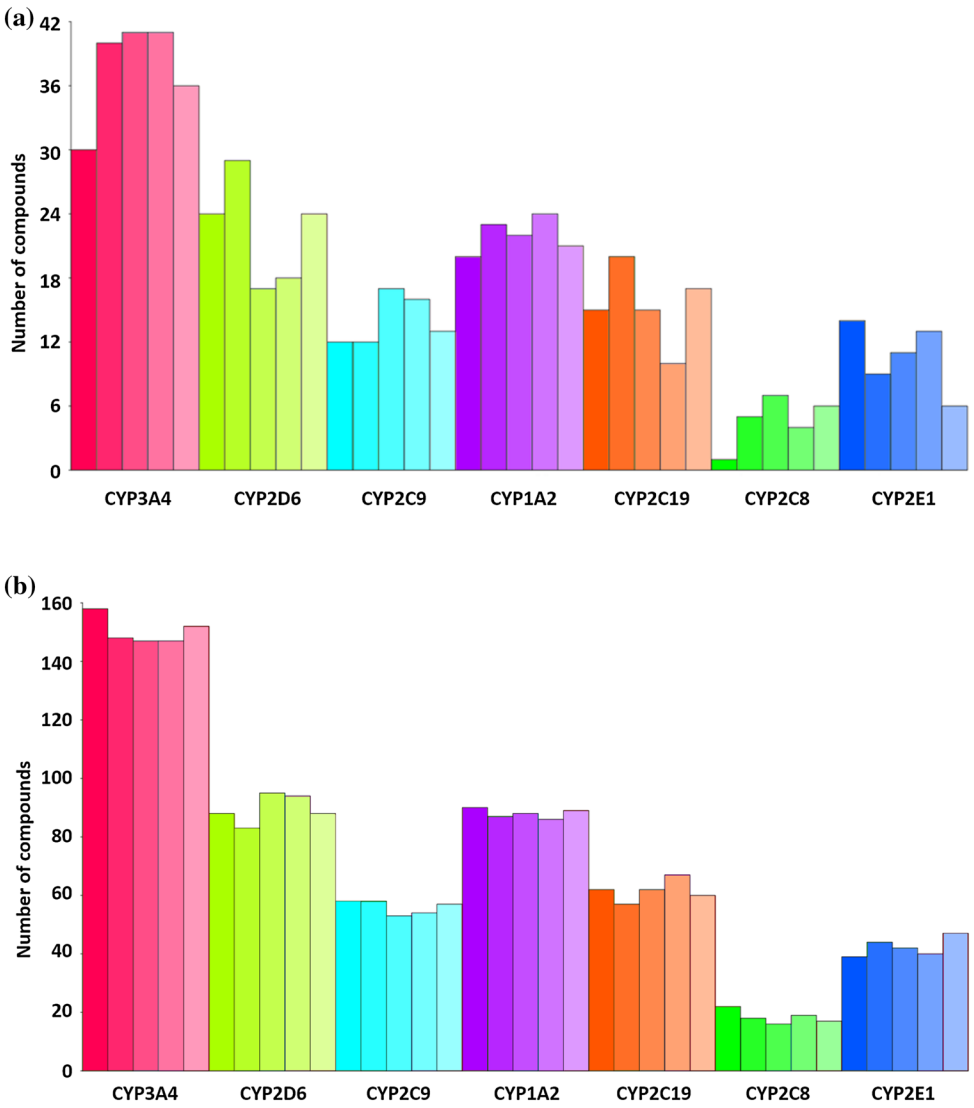


**Fig. 4** Distributions of the numbers of compounds metabolised by each isoform across the different set splits: **a** test sets; **b** training sets; with the graduated colours indicating the five different set splits. The CYP450s are coloured consistently with Fig. 2 and ordered by the frequency with which they are observed to metabolise marketed drugs, from the most common (left) to least common (right), as determined by Zanger et al. [3]

below), based on the distribution of the number of major isoforms observed for each compound in the test sets as shown in Fig. 1b.

Clearly, we know that the probabilities of a compound being metabolised by each isoform are not uniform; so, the second 'Guided Random' method, biased the random order of predicted isoforms by the frequency of occurrence of each isoform as a metabolising enzyme in the training set.

As an alternative to making random 'predictions' of the rank order of isoforms, the predictions from the models can be kept the same, but the connection between a compound structure and its observed isoforms in the test set can be randomised. Therefore, the third random method ('Y-shuffled') took the frequency of occurrence of each isoform within the observed lists of each test set and constructed a new target column which maintained the same frequency of isoform occurrence and also the same distribution of lengths of the observed lists of major isoforms, to ensure a fair comparison.

An additional form of prior knowledge is the correlation between observations of isoforms within the data set. Therefore, the final random method ('Y-scrambled') maintains the observed lists of isoforms but randomised their associations with the compound structures in the test set, again keeping the model predictions the same. Each randomisation was performed 500 times for each data set split and the performances were averaged to avoid the potential for spurious chance results.

Finally, after development of the models described herein, CYP metabolism data on a further 29 compounds were abstracted from the literature. These compounds were not included in any of the model building and constitute an independent test of the models.

**Table 1** Top-$k$ ($k = 1$–3) and AUC results on the independent test sets for the models built and tested with each of the training/test set splits

| Set split | Top-1% performance | Top-2% performance | Top-3% performance | Average AUC |
|---|---|---|---|---|
| 1 | 71 | 82.8 | 90.3 | 0.86 |
| 2 | 81.7 | 87.1 | 89.2 | 0.87 |
| 3 | 68.8 | 83.9 | 93.5 | 0.86 |
| 4 | 80.6 | 95.7 | 98.9 | 0.93 |
| 5 | 79.6 | 92.5 | 94.6 | 0.91 |
| Average | 76.3 | 88.4 | 93.3 | 0.89 |

## Results and discussion

The top-$k$ and AUC results for the five models built and tested using each of the training/validation set splits, are detailed in Table 1. From this, we can see that the performance of the different models is stable with respect to the data set split. The top-1 prediction success is 76%, on average, and the average top-2 prediction success rate is 88%.

Comparisons of the results for the five models with the different random methods are shown in Table 2. The top-$k$ results are illustrated in Fig. 5, from which we can see that the model performances are much higher than the values obtained by any of the random methods.

Similarly, the AUC results for the models and random methods are shown in Fig. 6, where the clear difference in performance suggests that the classification models are highly predictive compared to any of the random methodologies. Indeed, a one-tailed t-test indicates that the models are significantly more predictive than any of the random methods ($p < 0.0005$).

It is also interesting to confirm that prior knowledge of either the biological system or of the distributions within the specific set split improves the ability to predict over simple uniform random selection. As expected, the unbiased Uniform Random methodology performs the worst,

**Table 2** Summary of the statistics for the top-$k$ and AUC performances of the models and the four random methods. For each, the average and standard deviation over the five data set splits is shown

| Method | Models | Expected uniform random | Uniform random | Guided random | Y-shuffled | Y-scrambled |
|---|---|---|---|---|---|---|
| Top-1 (%) | 76.3 ± 5.4 | 19.4 ± 1.1 | 19.5 ± 4.0 | 25.1 ± 4.3 | 28.2 ± 4.0 | 29.0 ± 4.1 |
| Top-2 (%) | 88.4 ± 4.9 | 36.4 ± 1.5 | 36.5 ± 4.9 | 45.4 ± 4.8 | 49.5 ± 4.3 | 50.0 ± 4.3 |
| Top-3 (%) | 93.3 ± 3.4 | 51.6 ± 1.7 | 51.7 ± 5.1 | 62.1 ± 4.7 | 64.6 ± 4.2 | 64.5 ± 4.1 |
| AUC | 0.89 ± 0.03 | 0.5 ± 0.0 | 0.5 ± 0.0 | 0.59 ± 0.01 | 0.61 ± 0.02 | 0.61 ± 0.02 |

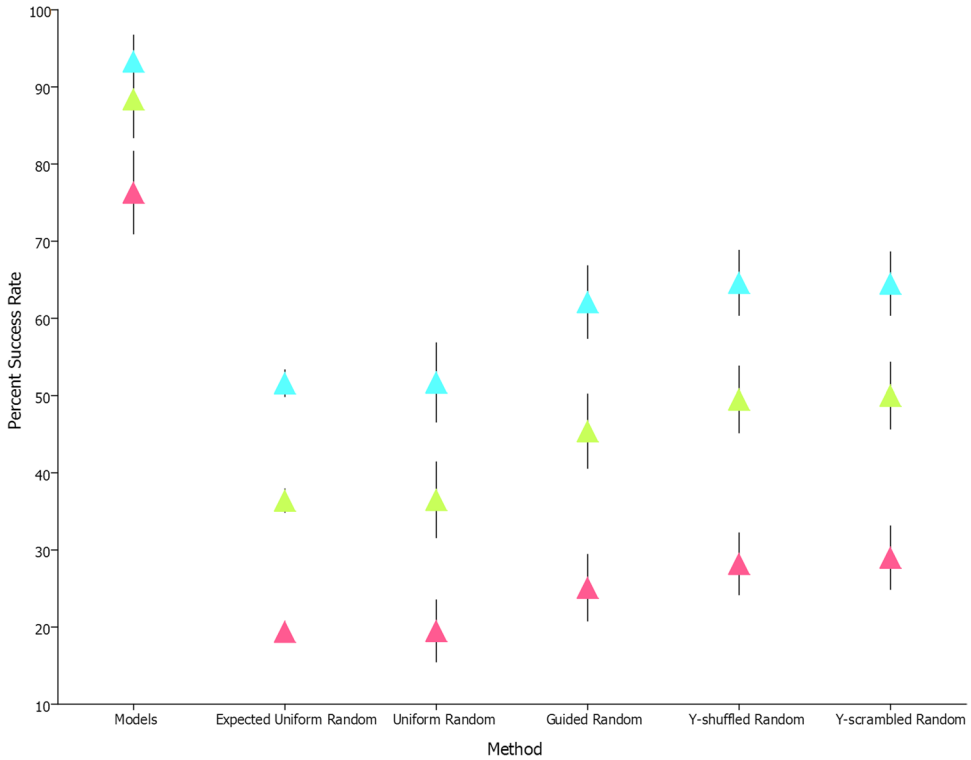


**Fig. 5** Percentage success rates for the models versus the random estimates for top-1 (pink points), top-2 (green points), and top-3 (cyan points) criteria. For each method, the average result is shown with whiskers illustrating the standard deviation over the five data set splits

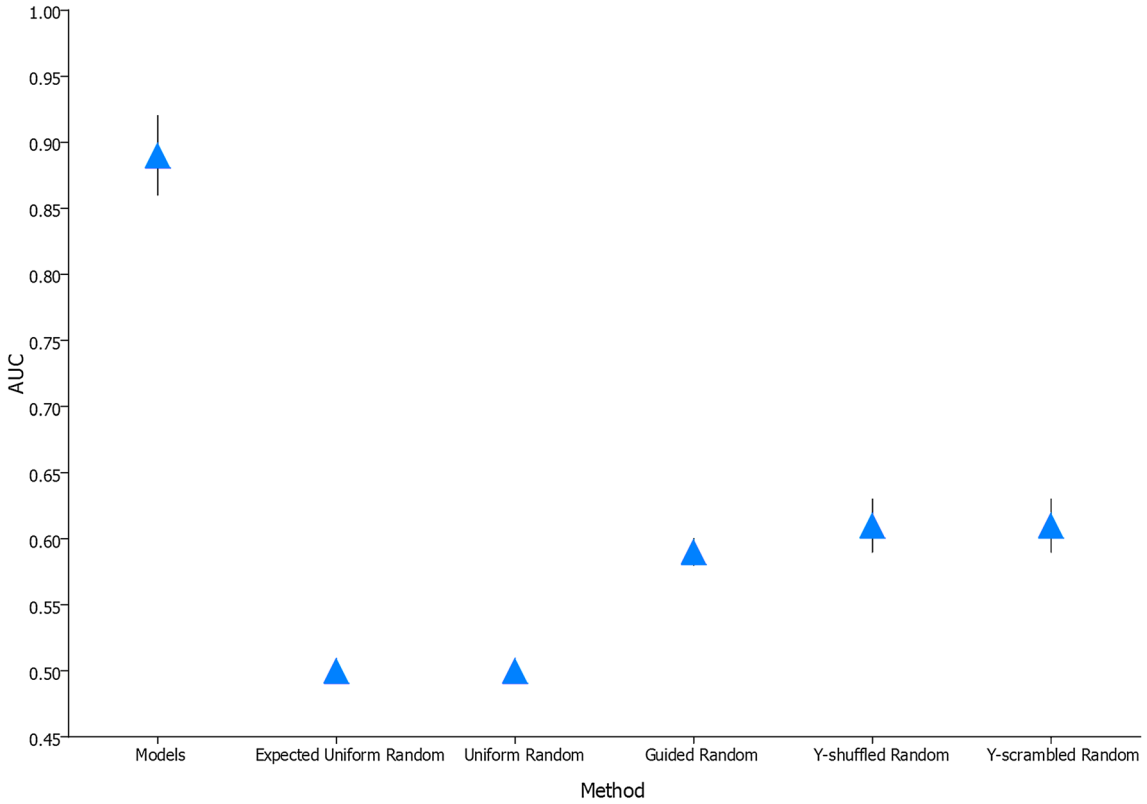


**Fig. 6** A comparison of the average AUCs for the five models compared to the random methods. For each method, the average result is shown with whiskers illustrating the standard deviation over the five data set splits

with top-1 and top-2 success rates of around 20 and 37% respectively, whereas the Guided Random method fared better with around 25 and 45% respectively. The Y-scrambled and Y-shuffled performances were indistinguishable from each other and only slightly better in their average values than Guided Random, with top-1 and top-2 performances of 29 and 50% respectively, although the ranges of performances on the different set splits overlapped for these three measures. It is also good to see that the performance of the Uniform Random method is in good agreement with the analytically-derived expected value.

In this analysis, we have considered only isoforms identified as having a major contribution to the metabolism of a compound. However, one could also envisage that a predicted isoform may not be a major isoform but does metabolise that compound to a lesser extent. In these cases, a false positive for prediction of a major isoform may be correct, if we are willing to accept predictions of these 'minor' isoforms. The minor isoforms were also annotated while compiling the data set (see the Supplementary Information) and therefore this experiment was also undertaken. As noted earlier, the more isoforms that are considered as major metabolising enzymes for a compound, the easier it is to predict one of those isoforms in the top-*k* criteria and, as expected, the success rates for the model predictions increased to above 88% for the top-1 criterion, whilst the Guided Random top-1 successes only rose to 46%. Conversely, the average AUC values dropped slightly for all the models; however, the AUC performance of the random models was still markedly different to those of the real models. A more detailed analysis of these results is provided in the Supplementary Information.

The performance of the models on the additional, 29-compound independent test set is summarised in Table 3. By chance, this set contains a large proportion of compounds metabolised by CYP3A4, hence simply choosing this as the major isoform one would have a top-1 success of 72.4%. However, our models predict on average 80.0% correct for top-1 and are above 91% accurate for the top-2 metric, again noticeably better than either the uniform ($16.2 \pm 6.8$% top-1 and $31.6 \pm 8.7$% top-2 performance), or biased ($25.9 \pm 7.8$%

**Table 3** Summary of the statistics for the top-*k* and AUC performances of the models for the 29-compound test set. The average and standard deviation over the five data set splits is shown

| | Top-1 (%) | Top-2 (%) | Top-3 (%) | AUC |
|---|---|---|---|---|
| 29 Compound set | $80.0 \pm 3.4$ | $91.8 \pm 2.8$ | $95.9 \pm 3.4$ | $0.89 \pm 0.19$ |

top-1 and 47.4 ± 8.6% top-2) random selection, would be expected to achieve respectively.

The five random forest models created with the different training/test set splits use slightly different descriptor sets and hence find different descriptors to be important in determining category membership and discrimination in each model. However, one can examine the five models to see which descriptors are consistently found to be important in all models, and these are shown in Table 4. Each Random Forest model comprises 100 trees and the importance of each descriptor is calculated by first calculating the accuracy of each tree in the ensemble over all data points not in the bootstrapped sample used to construct the individual tree (so-called 'out-of-bag' samples). The values of a single descriptor are then randomly permuted and the accuracy of the single tree is calculated again over the out-of-bag samples. Because this randomisation effectively voids the effect of the descriptor, the descriptor's importance is determined by averaging the percentage decrease in prediction accuracy over all trees after performing the randomisation.

It is interesting to see descriptors such as the McGowan volume [15] and the count of basic nitrogens in the most important list, as one can rationalise the effects that these would have for CYP2E1 and CYP3A4, which accommodate small and large molecules respectively, and for CYP2D6, which has a preference for basic molecules through an ion-pair interaction with residue ASP301. The other hydrophobic descriptors may indicate steric requirements (such as the propensity of CYP2C9 or CYP2C19 for aromatic acids) or simply be an indication of the various levels of flexibility that the different CYP active sites may require.

## Conclusions

We have demonstrated the generation of a multi-class model that is able to predict which of seven CYP450 isoforms is most likely to be a major metabolising isoform of a compound, should it be a substrate for CYP450-mediated metabolism. By rank ordering the probabilities of the seven isoforms, the model can also suggest which other isoforms are the next most likely, should a compound be metabolised by more than one isoform. Where different Phase I regioselectivity predictions are made for a compound, based on different CYP450 isoforms, these models will enable a more accurate Phase I metabolite profile prediction.

Another important aspect of CYP450-related metabolism comes from the avoidance of drug–drug interactions (DDIs). DDIs can occur where co-administered drugs are predominately cleared by, or bind to, the same CYP450 isoform and, as such, can interfere with the clearance of one another from the systemic circulation, creating drug concentrations in excess of their normal levels and potentially causing toxicity. This situation can also occur naturally in a patient due to either genetic polymorphisms in, or different levels of expression of, particular CYP450s. Examples of relatively common polymorphisms include those leading to poor metabolism phenotypes for CYP2D6 in Caucasian and

**Table 4** 18 Important descriptors found to be consistently important descriptors across the five different set splits

| Descriptor | Explanation |
|---|---|
| BasicGroup | Number of basic nitrogen |
| q192 | Number of sp2 non-aromatic atoms separated from another ring atom by four bonds |
| nC(sp3) | Number of sp3 carbons |
| NRB | Number of single bonds to heavy atoms |
| sssCH | Number of sp3 carbons with exactly one hydrogen |
| ringat | Number of cyclic atoms |
| q481 | Number of aliphatic atoms in a ring with two singly bonded atoms connected and separated from an aromatic atom by four bonds |
| C2 | Number of sp3 carbons with no or exactly one hydrogen and connected to aliphatic carbons only |
| CH1Aa | Number of aliphatic sp3 carbons with exactly two hydrogens |
| nNprot | Number of protonated nitrogens at pH 7.4 |
| dssC | Number of sp2 carbons with no hydrogens |
| Vx | The McGowan volume [15] |
| q453 | Number of atoms with three explicit connections separated by four bonds from an atom other than carbon |
| sOH | Number of aliphatic sp3 oxygens connected to one hydrogen |
| AroRingAttachment | Number of aromatic atoms with three explicit bonds |
| HAO | Number of oxygens not in nitro groups |
| H1a | Number of aliphatic carbons with one hydrogen |
| nC(sp2) | Number of sp2 carbons not connected to an sp2 oxygen |

CYP2C19 in Asian populations [16]. Again, the knowledge that a particular compound may be affected by these variations in the clinic would help prioritise other compounds without such potential liabilities.

We have recently published an accurate quantum mechanical method to predict the regioselectivities of the seven CYP450 isoforms considered in this paper [2]. In the future, the models described herein could be combined with those predicting regioselectivity of metabolism by the different isoforms to predict, at least qualitatively, metabolite profiles of novel compounds metabolised by CYP450 enzymes, as illustrated for Venlafaxine [17] in Fig. 7. In order to achieve accurate results, it may also be necessary to consider the relative abundances of the different isoforms in the liver and possibly in other tissues.

Another future improvement may be to retrain the models including a class of compounds which are not significantly metabolised by a CYP450. Such a 'no P450' class would be useful for determining when predictions of metabolism by CYP450 isoforms will not be relevant for determining the metabolic fate or elimination of a potential drug.

## Experimental

All results were generated using StarDrop v6.3 and the analysis used bespoke python scripts.

## Associated content

### Supporting information available

The data sets described herein and the calculations and methods used to derive the expected random probabilities for compounds with more than one major isoform; the

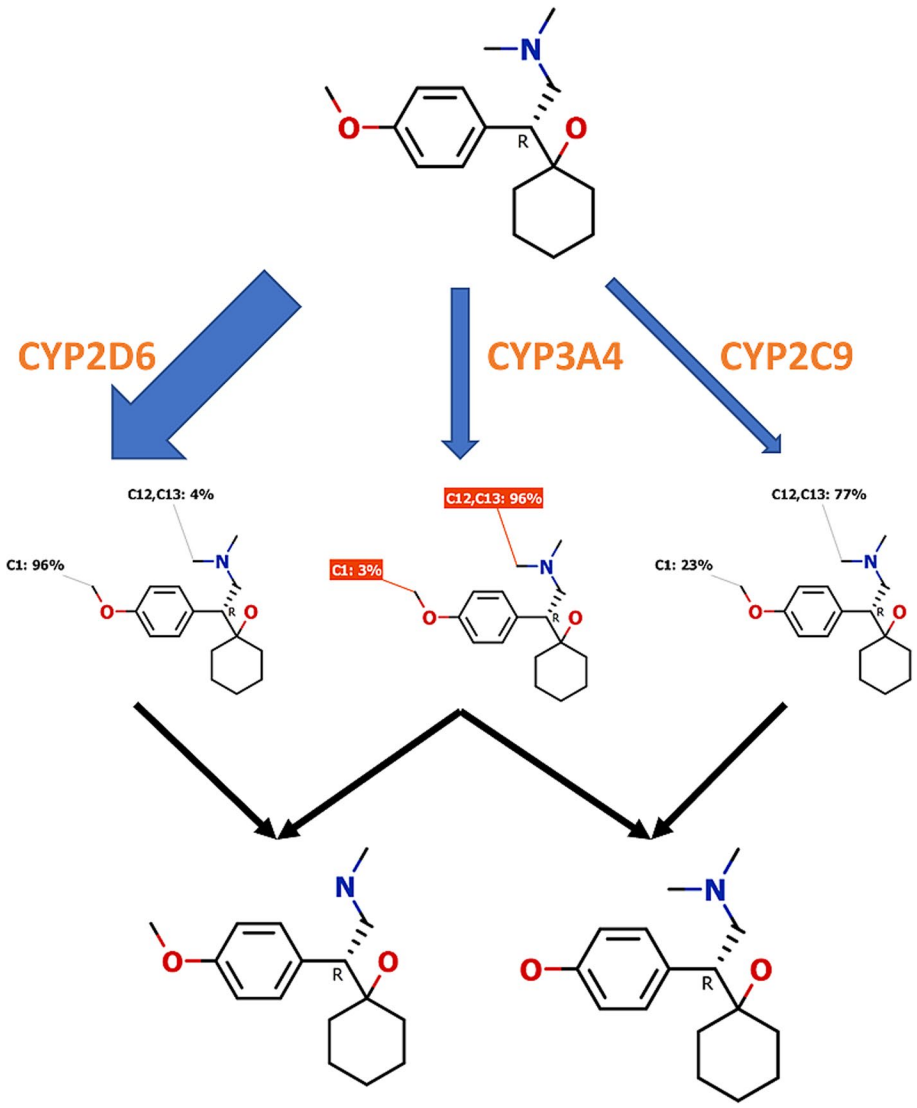


**Fig. 7** Prediction of significant P450 metabolites for Venlafaxine. The blue arrows indicate the predicted major isoforms responsible for metabolism of Venlafaxine; the width of the arrows is proportional to the probabilities output by the random forest models. Below these are predictions of the regioselectivity of metabolism by each of the major isoforms, using the models described in Tyzack et al. [2]. At the bottom, the corresponding predicted metabolites are shown. The predicted isoform probabilities for Venlafaxine are [CYP2D6 = 0.55; CYP3A4 = 0.18; CYP2C9 = 0.12; CYP2C19 = 0.07; CYP1A2 = 0.06; CYP2C8 = 0.02; CYP2E1 = 0]. Experimentally, the major observed isoform is CYP2D6 whilst CYP3A4, CYP2C9 and CYP2C19 are found to have minor contributions to metabolism of Venlafaxine [17]. The predicted metabolites correspond to those observed clinically

equivalent model performance when minor isoform information is included; and the t-test calculations for significance. This material is available free of charge via the internet.

**Acknowledgements** This research has received funding from the Union Seventh Framework Programme 2013 under the Grant agreement no. 602156.

## Compliance with ethical standards

**Conflict of interest** Matthew Segall and Peter Hunt are current employees of Optibrium Ltd., which develops the StarDrop software in which the methods described herein are implemented. Jonathan Tyzack is a former employee of Optibrium Ltd.

## References

1. Kantae V, Krekels EHJ, Esdonk MJV et al (2017) Integration of pharmacometabolomics with pharmacokinetics and pharmacodynamics: towards personalized drug therapy. Metabolomics 13:9. https://doi.org/10.1007/s11306-016-1143-1
2. Tyzack JD, Hunt PA, Segall MD (2016) Predicting regioselectivity and lability of cytochrome P450 metabolism using quantum mechanical simulations. J Chem Inf Model 56:2180–2193
3. Zanger UM, Schwab M (2013) Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol Ther 138:103–141
4. Kirchmair J, Göller AH, Lang D, Kunze J, Testa B, Wilson ID, Glen RC, Schneider G (2015) Predicting drug metabolism: experiment and/or computation? Nat Rev Drug Discov 14:387–404
5. Rostkowski M, Spjuth O, Rydberg P (2013) WhichCyp: prediction of cytochromes P450. Inhib Bioinform 29:2051–2052
6. ACD/Labs Percepta platform (2017) http://www.acdlabs.com/products/percepta/ Accessed 6 Sept 2017
7. MetaPred website (2017) http://crdd.osdd.net/raghava/metapred/ Accessed 6 Sept 2017
8. Mishra NM, Agarwal S, Raghava GPS (2010) Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. BMC Pharmacol 10:8
9. Hagymási K, Müllner K, Herszényi L, Tulassay Z (2011) Update on the Pharmacogenomics of proton pump inhibitors. Pharmacogenomics 12:873–888
10. García-Suástegui WA, Ramos-Chávez LA, Rubio-Osornio M, Calvillo-Velasco M, Atzin-Méndez JA, Guevara J, Silva-Adaya D (2017) The role of CYP2E1 in the drug metabolism or bioactivation in the brain. Oxid Med Cell Longev 2017(4680732):14
11. Wang B, Zhou SF (2009) Synthetic and natural compounds that interact with human cytochrome P450 1A2 and implications in drug development. Curr Med Chem 16:4066–4218
12. van der Maaten L, Hinton G (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 9:2579–2605
13. Breiman L (2011) Random forests. Mach Learn 45:5–32
14. StarDrop landing page on the Optibrium website (2017) http://optibrium.com/stardrop/. Accessed 6 Sept 2017
15. Abraham MH, McGowan JC (1987) The use of characteristic volumes to measure cavity terms in reversed-phase liquid-chromatography. Chromatographia 23:243–246
16. Preissner SC, Hoffmann MF, Preissner R, Dunkel M, Gewiess A, Preissner S (2013) Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. PLoS ONE 8:e82562
17. Fogelman SM, Schmider J, Venkatakrishnan K, von Moltke LL, Harmatz JS, Shader RI, Greenblatt DJ (1999) O- and N-demethylation of Venlafaxine in vitro by human liver microsomes and by microsomes from cDNA-Transfected cells: effect of metabolic inhibitors and. SSRI Antidepressants Neuropsychopharmacol 20:480–490