Bioinformatics, 31(24), 2015, 3930–3937
doi: 10.1093/bioinformatics/btv486
Advance Access Publication Date: 26 August 2015
Original Paper

Structural bioinformatics

Integrated structure- and ligand-based *in silico* approach to predict inhibition of cytochrome P450 2D6

Virginie Y. Martiny[1,2,†,‡], Pablo Carbonell[3,§], Florent Chevillard[1,¶], Gautier Moroy[1,2], Arnaud B. Nicot[4], Philippe Vayer[5], Bruno O. Villoutreix[1,2] and Maria A. Miteva[1,2,*]

[1]Université Paris Diderot, Sorbonne Paris Cité, UMR-S 973 Inserm, Paris 75013, France, [2]Inserm UMR-S 973, Molécules Thérapeutiques In Silico, Université Paris Diderot, Sorbonne Paris Cité, Paris 75013, France, [3]Research Programme on Biomedical Informatics (GRIB), Department of Experimental and Health Sciences, Universitat Pompeu Fabra, IMIM (Hospital del Mar Medical Research Institute), 08003 Barcelona, Spain, [4]Inserm U1064/ITUN, CHU, 44093 Nantes Cedex, France and [5]BioInformatic Modelling Department, Technologie Servier, 45007 Orléans Cedex1, France

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontana

[†]Present address: Institut de Chimie des Substances Naturelles, CNRS UPR 2301, LabEx LERMIT, Avenue de la Terrasse, 91198 Gif-sur-Yvette, France.

[‡]Present address: Laboratoire «Des maladies rénales rares aux maladies fréquentes, remodelage et réparation», INSERM UMR_S 1155, Hôpital Tenon, 75970 Paris Cedex 20, France.

[§]Present address: SYNBIOCHEM, Manchester Institute of Biotechnology, University of Manchester, Manchester, M1 7DN, UK.

[¶]Present address: Institute of Pharmaceutical Chemistry, Phillips University Marburg, 35037 Marburg, Germany.

Received on May 5, 2015; revised on July 16, 2015; accepted on August 14, 2015

Abstract

Motivation: Cytochrome P450 (CYP) is a superfamily of enzymes responsible for the metabolism of drugs, xenobiotics and endogenous compounds. CYP2D6 metabolizes about 30% of drugs and predicting potential CYP2D6 inhibition is important in early-stage drug discovery.

Results: We developed an original *in silico* approach for the prediction of CYP2D6 inhibition combining the knowledge of the protein structure and its dynamic behavior in response to the binding of various ligands and machine learning modeling. This approach includes structural information for CYP2D6 based on the available crystal structures and molecular dynamic simulations (MD) that we performed to take into account conformational changes of the binding site. We performed modeling using three learning algorithms –support vector machine, RandomForest and NaiveBayesian –and we constructed combined models based on topological information of known CYP2D6 inhibitors and predicted binding energies computed by docking on both X-ray and MD protein conformations. In addition, we identified three MD-derived structures that are capable all together to better discriminate inhibitors and non-inhibitors compared with individual CYP2D6 conformations, thus ensuring complementary ligand profiles. Inhibition models based on classical molecular descriptors and predicted binding energies were able to predict CYP2D6 inhibition with an accuracy of 78% on the training set and 75% on the external validation set.

Contact: maria.miteva@univ-paris-diderot.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cytochrome P450 (CYP) is a superfamily of enzymes responsible for the metabolism of drugs, xenobiotic substances and endogenous compounds (Johansson and Ingelman-Sundberg, 2011; Shimada, 2006; Singh *et al.*, 2011). It has been estimated that about 75% of the marketed drugs are metabolized by CYPs, the major CYP isoforms being 1A2, 2C8, 2C9, 2C19, 2D6 and 3A4 (Singh *et al.*, 2011). These enzymes detoxify the organism from various xenobiotics and activate some prodrugs by oxidation. Yet, oxidation sometimes leads to metabolites that are more active than the administrated drugs or activates pro-carcinogens by creating highly reactive metabolite species. Inhibition of CYP is a complex process because it can correspond to a competitive inhibition in the active site, a modification of the substrate or metabolite flux between the active site and outside of the enzyme or an inhibition by a drug itself or by its metabolites (time-dependent inhibition) leading then to adverse drug–drug interactions (Bode, 2010; Johansson and Ingelman-Sundberg, 2011; Rodriguez-Antona and Ingelman-Sundberg, 2006; Singh *et al.*, 2011). Thus, predicting potential CYP inhibition is important in early-stage drug discovery. CYP2D6 metabolizes about 30% of drugs although it represents only 2% of hepatic CYPs (Singh *et al.*, 2011). Besides liver, intestine and kidney, CYP2D6 is also expressed in brain cells where it may be involved in metabolism of neurotransmitters such as serotonin (Ferguson and Tyndale, 2011). CYP2D6 interacts with many drugs used for regulation of the central nervous system (psychotropics) or the cardiovascular system (anti-rhythmic drugs) (Marechal *et al.*, 2008; Rowland *et al.*, 2006). This isoform is of particular interest as a large amount of drugs can be metabolized only by CYP2D6 (Singh *et al.*, 2011). Moreover, CYP2D6 is highly polymorphic (Ingelman-Sundberg *et al.*, 2007; Martiny and Miteva, 2013) with more than a hundred of allelic variants (Sim and Ingelman-Sundberg, 2010) leading to poor, intermediate, extensive or ultra-rapid metabolism that can render the administrated drug toxic or inefficient (Johansson and Ingelman-Sundberg, 2011; Pinto and Dolan, 2011; Porcelli *et al.*, 2011; Rodriguez-Antona and Ingelman-Sundberg, 2006).

A number of modeling studies have been undertaken to understand the molecular basis of CYP2D6–drug interactions and CYP2D6-related metabolism or inhibition (Ai *et al.*, 2015; Cruciani *et al.*, 2005; de Graaf *et al.*, 2006; Kemp *et al.*, 2004; Kirchmair *et al.*, 2012; Kjellander *et al.*, 2007; Livezey *et al.*, 2012; Marechal *et al.*, 2008; Martinez-Sanz *et al.*, 2013; Moroy *et al.*, 2012; Stoll *et al.*, 2011; Tyzack *et al.*, 2013). Most of these analyses have been executed on homology models as the first X-ray structure of human CYP2D6 was solved in 2006 (Rowland *et al.*, 2006). CYP inhibition is considered to be more difficult to predict than sites of metabolism (SOM) due to the lack of reactive sites (Stoll *et al.*, 2011). Recently, a new approach for prediction of CYP inhibition, CypRules, using decision tree algorithm based on compound structural rules has been reported (Shao *et al.*, 2015). However, the extensive flexibility of the CYP2D6 structure, which represents its natural mechanism to accommodate diverse ligands into the active site (Moroy *et al.*, 2012; Wang *et al.*, 2015), can strongly affect the prediction of CYP inhibition when one uses traditional ligand-based Quantitative Structure-Property Relationships (QSAR) modeling (Brändén *et al.*, 2014; Stoll *et al.*, 2011). Indeed, in 2010, Vedani and co-authors (Rossato *et al.*, 2010) adopted a mixed-model protocol in which ligand-based pre-alignment was followed by ligand docking to the CYP2D6 protein structure with flexible side chains which allowed a quantification of ligand binding through multi-dimensional QSAR modeling validated on 56 CYP2D6 binders.

The recently reported CYP2D6 structures in complex with different inhibitors (Wang *et al.*, 2012a, b, 2015) as well as the large number of experimentally validated CYP2D6 binders available in the PubChem BioAssay database (Wang *et al.*, 2012a, b) permit to carry out a comprehensive analysis of CYP2D6 inhibition. In this study, we developed a new integrated *in silico* approach to predict inhibition of CYP2D6. The performed molecular dynamic simulations (MD) provided new insights about the conformational changes in the active site architecture of CYP due to the accommodation of different ligands. We found a set of modeled CYP2D6 conformations, which all together are able to correctly retrieve, after docking, 70% of the known CYP binders in 35% of the screened compound library. The computational protocol integrating docking into this pool and machine learning-based modeling can be successfully applied to predict CYP2D6 inhibitors with 75% of success.

2 Materials and methods

2.1 Molecular dynamics simulations of CYP2D6

Among the X-ray structures of human CYP2D6 available at the Protein Data Bank (PDB) (Bernstein *et al.*, 1977), we retained the only apo X-ray structure PDB ID 2F9Q (Rowland *et al.*, 2006) and one holo structure co-crystallized with prinomastat (a metalloprotease inhibitor and a potent inhibitor of CYP2D6 with an observed $Ki = 0.049\ \mu M$; Wang *et al.*, 2012a, b), PDB ID 3QM4 (Wang *et al.*, 2012a, b), as the other available holo X-ray structures shared very similar conformations with no striking differences in the binding sites [all atom root mean square deviation (RMSD) <1 Å for the PDB IDs: 3QM4, 3TDA, 4WNT, 4WNU, 4WNV, 4WNW, 3TBG]. The apo structure of CYP2D6 PDB ID 2F9Q misses a loop of 10 residues from position 42 to 51. This loop has been built using MODELLER 9.7 (Sali *et al.*, 1995) by generating 100 models and selecting those of lower discrete optimized protein energy score. Residues D230, R231 and M374 are mutated in the X-ray structure and have been replaced by the wild type (WT) residues L230, L231 and V374, respectively, using MODELLER 9.7. These rebuilt apo X-ray structure and the holo X-ray structure were then used for establishing our protocol.

Molecular dynamic simulations using CHARMM c35b1 version (Brooks *et al.*, 1983) have been performed in order to explore the binding site flexibility of the two structures. We used the all-atom PARAM27 force field (Mackerell *et al.*, 2004). The pKa values of the titratable groups were calculated with the Finite Difference Poisson Boltzmann approach using the web server tool Protein Continuum Electrostatics (Miteva *et al.*, 2005) (dielectric values of 11 and 80 for solute and solvent, respectively). We ran 3 MD simulations in complex with the substrates of CYP2D6 propafenone, mexiletine and codeine (Fig. 1) because these drugs have very different chemical structures as compared with the already co-crystalized CYP2D6 inhibitors, and as such those simulations should allow to explore more thoroughly the conformational space of the CYP2D6

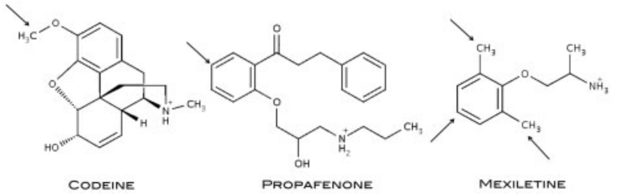

Fig. 1. Structures of substrates of CYP2D6 used for MD simulations. The arrows show the main sites of metabolism

active site. To generate the starting ligand conformations, we carried out preliminary docking experiments with AutoDock 4.2 (Morris *et al.*, 2009) and Vina 1.1.1 (Trott and Olson, 2010) using the holo X-ray structure of the enzyme and after investigating the protonation state of the ligands with MarvinSketch version 5.3 (2010, ChemAxon). The best substrate conformations showing the lowest binding energies with known SOM (Fig. 1) close to the heme Fe (<6 Å) have been chosen as starting conformations for MD simulations. Based on the apo X-ray structure of CYP2D6, we performed four MD simulations: one apo, one in complex with propafenone, one in complex with mexiletine and one in complex with codeine. We also performed two MD simulations based on the holo X-ray structure: one with its co-crystallized ligand and one without bound ligand. Topology and parameters of all the ligands were assigned by using the CgenFF program (Vanommeslaeghe *et al.*, 2012). The solvation was taken into account by the Generalized Born implicit solvent function FACTS (Haberthür and Caflisch, 2008). Non-bonded interactions were truncated with a cut-off distance of 12 Å with a shift function for electrostatics and switch function for the van der Waals interactions. The protein structures were initially minimized using 500 steps of steepest descent algorithm followed by 500 steps of conjugate gradient algorithm. Distances between heavy atoms and hydrogen atoms were constrained using the SHAKE algorithm (Ryckaert *et al.*, 1977) allowing a time step of 2 fs. Each system was heated during 200 ps to reach 300 K and then equilibrated during 400 ps with a temperature window of 300 ± 10 K. The production time was 4 ns for each MD simulation.

2.2 CYP2D6 conformational ensemble generation

For each of the six MD simulations, we extracted structures every 1 ps starting at 1 ns, representing 3000 structures per MD, and a total 18 000 structures (Fig. 2). For each MD simulation, the RMSD between the 3000 extracted structures were calculated for all atoms of the binding site and of the heme moiety. We clustered different conformations of the binding sites by applying hierarchical ascending classification (HAC) on the obtained RMSD matrix using the agglomerative Ward's method as implemented in the R software (RDevelopmentCoreTeam, 2009), and a final RMSD distance of at least 1.5 Å. The protein centroid structures of each cluster were then used for subsequent virtual screening. The binding site volume and druggability scores of CYP2D6 structures were calculated using the DoGSiteScorer (Volkamer *et al.*, 2012) webserver.

2.3 Virtual screening experiments

First, we performed preliminary docking experiments with AutoDock 4.2 (Morris *et al.*, 2009) and Vina 1.1.1 (Trott and Olson, 2010) to probe the docking positions of propafenone, mexiletine and codeine in the two CYP2D6 X-ray structures. We removed the water molecules from the active site for docking and virtual screening because the analysis of the seven structures of CYP2D6 co-crystallized with different inhibitors available in PDB shows that water molecules change significantly their positions depending on the bound ligand, and a key water molecule mediating the heme–inhibitor interactions was not observed. We obtained the best docking results using Vina regarding the SOM positions. Then, VS experiments of 343 inhibitors and 3002 non-inhibitors of CYP2D6 were carried out using Vina to select the binding site conformations best predicting the binding energies among all representative structures obtained by HAC (see Supplementary Materials for inhibitors preparation, Lagorce *et al.*, 2011; Miteva *et al.*, 2010). A grid resolution of 1 Å was used with a number of binding modes of 10 and exhaustiveness of 8 (Trott and Olson, 2010). Thirty-six virtual screening runs were performed on X-ray structures and on protein structures extracted from MD.

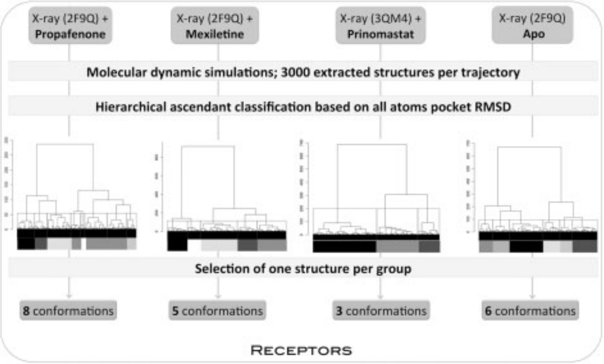
X-ray (2F9Q) + Propafenone
X-ray (2F9Q) + Mexiletine
X-ray (3QM4) + Prinomastat
X-ray (2F9Q) Apo
Molecular dynamic simulations; 3000 extracted structures per trajectory
Hierarchical ascendant classification based on all atoms pocket RMSD
Selection of one structure per group
8 conformations
5 conformations
3 conformations
6 conformations
Receptors

Fig. 2. MD-derived structure classification. This protocol has been applied in the same way for the X-ray structure 3QM4 apo and the X-ray structure 3QM4 in complex with codeine

2.4 Machine learning classification modeling

The 343 inhibitor structures were selected as positive set in order to train machine learning classification models for compounds binding to CYP2D6. In order to build and validate the models, a balanced dataset was built by adding to the positive set an equally sized negative set randomly sampled from 3002 non-inhibitors. An external validation set was built by randomly taking 20% of both positives and negatives in the entire dataset. The remaining 80% was used as training set for the model. To describe topological features of the structures, we used extended connectivity fingerprints (ECFPs) (Carbonell *et al.*, 2013) up to an atom vicinity of 2. To reduce the dimensionality of the resulting input feature matrix, we applied principal component analysis using the statistical package R.

Three machine learning methods were used: a support vector machine (SVM) by using the kernlab R package (ksvm function) (Karatzoglou *et al.*, 2004), a random forest-based predictor by using the RandomForest R package (Liaw and Wiener, 2002) and a NaiveBayesian predictor by using the caret R package (Kuhn, 2008). For each of them we built models using as descriptors the topological features via ECFP and the protein–ligand binding energies calculated on the best performing MD receptor conformations. Performance of each classification model was assessed by the percentage of correctly classified compounds in comparison with the total number of compounds in the set through leave-one-out cross-validation.

3 Results and Discussion

We developed an integrated structure- and ligand-based *in silico* approach able to predict inhibition of CYP2D6. This approach includes structural information for CYP2D6 based on the available crystal structures and molecular dynamic simulations that we performed to take into account conformational changes of the binding site occurring due to the presence of diverse ligands. The approach also includes information from experimental inhibition studies and chemical information from ligands.

3.1 Analysis of CYP2D6 X-ray structures

We selected two X-ray structures of CYP2D6: one apo (PDB ID 2F9Q) and one holo co-crystallized with prinomastat (PDB ID 3QM4). The comparison of the available holo X-ray structures of CYP2D6 (PDB IDs: 3QM4, 3TDA, 4WNT, 4WNU, 4WNV, 4WNW, 3TBG) did not show important differences in the binding site. However, in the absence or in the presence of bound ligands, important conformational changes are observed. The apo structure misses 10 residues and has 3 amino acid substitutions compared with the WT protein, 2 of which were localized in the α helices F and G' (Supplementary Fig. S1A). This structure was repaired using an *in silico* protocol (Supplementary Fig. S1B) and the only difference observed was a larger volume for the fixed protein, increasing from 657 to 676 Å^3, due to loop optimization between the helices F and G'. In our study, the structure called apo is the rebuilt one.

By comparing the two X-ray structures, apo and holo (Supplementary Fig. S2), a striking difference appears: the apo structure (Supplementary Fig. S2A) displays a long and regular F α-helix, while the holo structure (Supplementary Fig. S2B) displays an α-helix broken into two pieces, named the F and F' segments (Supplementary Fig. S2C). By comparing these two X-ray structures with other CYP2 members, CYP2A6 PDB ID 1Z10, CYP2C8 PDB ID 2NNJ and CYP2C9 PDB ID 1OG2 (Supplementary Fig. S2D), it is seen that proteins share a similar fold, but considerable differences appear around the helix F. Indeed, the helix F' present in the holo X-ray structure of CYP2D6 is a common feature found in the CYP2 family (Wang *et al.*, 2012a,b, 2015), while the helix F' is not found in the apo CYP2D6 structure. The helix G of both X-ray structures of CYP2D6 (Supplementary Fig. S2C) is shortened regarding the other members of CYP2 family (Supplementary Fig. S2D). The formation of the helix F' segment in the holo CYP2D6 structures as well as the displacement of the surrounding secondary structures demonstrate the adaptation of the protein to the upcoming ligands.

Regarding the binding site, there are 4 key residues F120, E216, D301 and F483 involved in the enzyme–substrate interaction (Marechal *et al.*, 2008) (Supplementary Fig. S3). Residues E216 and D301 (Supplementary Fig. S3A) are negatively charged and attract the basic nitrogen of CYP2D6 substrates (Kirton *et al.*, 2002; Paine *et al.*, 2003; Wang *et al.*, 2015). The residue E216 participates mainly in the interaction with the ligand (Paine *et al.*, 2003), while D301 also plays a structural role (Hanna *et al.*, 2001) by interacting with the backbone of F120 (Supplementary Fig. S3A and B). The latter controls the orientation of substrates toward the heme moiety (Flanagan *et al.*, 2004) via π-stacking interaction (Supplementary Fig. S3B) with an aromatic ring of the CYP2D6 substrates (Supplementary Fig. S3C).

By comparing the orientation of key residues between the apo and holo X-ray structures, we observe that D301 orientation (Supplementary Fig. S3A) is conserved maintaining its interaction with F120. The orientation of F120 is also conserved, thus interacting by π–π stacking with prinomastat (Supplementary Fig. S3B). However, F483 and E216 have a completely different orientation, resulting in a larger volume of the binding pocket of 676 Å^3 in the apo structure to 712 Å^3 in the holo one. Thus, the CYP2D6 apo X-ray structure (Rowland *et al.*, 2006) displays a closed active site that cannot accommodate its known ligands. CYP2D6 binds compounds with a basic nitrogen positively charged and oxidizes atoms at a distance of 5–8 Å from the heme Fe (Hritz *et al.*, 2008; Ito *et al.*, 2008; Marechal *et al.*, 2008; Rowland *et al.*, 2006; Wang *et al.*, 2015). In order to explore further the conformational space of the CYP2D6 active site, we performed MD simulations with three representative and diverse CYP2D6 substrates (propafenone, mexiletine and codeine; Fig. 1) different from the already co-crystallized CYP2D6 ligands (Wang *et al.*, 2015).

3.2 Docking

We performed docking experiments with propafenone, mexiletine and codeine. The positions of the SOMs in the docking poses guided the selection of the best poses. Docking of propafenone (Fig. 3A) shows a possible H-bond between the positively charged nitrogen and E216. One of the two aromatic rings interacts with F483 by π-stacking, and the second aromatic ring is involved in interaction with F120, the SOM being close to the heme Fe. The predicted binding energy is $-8.2\,\mathrm{kcal\,mol^{-1}}$, which shows the relevance of this ligand conformation as a starting pose for the MD simulations. Docking of mexiletine (Fig. 3B) suggested interactions between the positively charged nitrogen atom and E216 and a T-stacking with F120. The distance between the SOM and the heme Fe is within 6 Å and the predicted energy is $-7.0\,\mathrm{kcal\,mol^{-1}}$. The codeine (Fig. 3C) best docking pose proposed that the SOM is close to the heme Fe and the aromatic rings system interacts with F120. Although the positively charged nitrogen atom of codeine is not oriented toward E216, the predicted binding energy is still favorable with a value of $-8.9\,\mathrm{kcal\,mol^{-1}}$. Being a large and rigid compound, docking of codeine in the rigid binding site was not completely successful, highlighting the importance of considering protein flexibility for such cases. Previously, the key role of F120 was also suggested by a docking study of four inhibitors of CYP2D6 showing face-to-face and edge-to-face π interactions (Livezey *et al.*, 2012).

The poses with the lowest predicted binding energies for each substrate showing the SOM within 6 Å around the heme Fe have been selected as starting conformations for the MD simulations.

3.3 MD simulations to identify diverse binding site conformations

As previously demonstrated, MD simulations are a pertinent approach to study the dynamic behavior of CYP family (Nair and Miners, 2014). After docking, we ran MD simulations. Four MD simulations have been performed on the apo structure, three in complex with the three ligands (Fig. 1) and one without bound ligand. Starting from the holo X-ray structure, we ran two MD simulations,

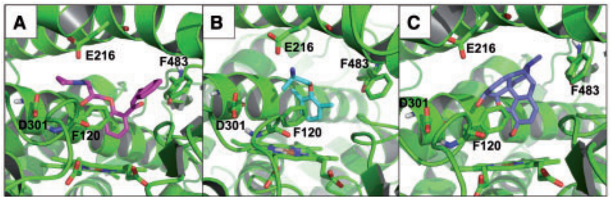

Fig. 3. Docking of ligands into the apo X-ray CYP2D6 (shown in green) (PDB ID 2F9Q) prior to MD. (**A**) Propafenone shown in sticks in magenta atom type. (**B**) Mexiletine shown in sticks in cyan atom type. (**C**) Codeine shown in sticks in mauve atom type

one with the co-crystallized ligand pinomastat and another without a bound ligand, and thus, in total, six MD simulations were carried out. During the entire MD simulations the bound propafenone, mexiletine or codeine remained stable in the active site showing small fluctuations of the distances to the Fe heme.

Different strategies can be employed in order to select MD-derived representative structures for further analysis. For example, Hritz *et al.* (2008) extracted 2500 conformations of CYP2D6 from 10 MD simulation runs of 1 ns with 5 different ligands bound in the active site and docked a limited number of ligands into all 2500 conformations. Considering so many structures for docking without any classification is not efficient because many redundant structures of the binding site can be present. In this study, we successfully reduced the number of structures by applying HAC based on RMSD of all atoms of the binding sites to the extracted structures. Initially, 3000 structures were extracted for each MD simulation, and thus 18000 structures in total were generated. After our procedure, we obtained 6 structures from the trajectory based on the apo 2F9Q, 8 structures from the trajectory based on the 2F9Q complexed with propafenone, 5 structures from the trajectory based on the 2F9Q complexed with mexiletine, 5 structures from the trajectory based on the 2F9Q complexed with codeine, 7 structures from the trajectory based on the apo 3QM4 and 3 structures from the trajectory based on the 3QM4 in complex with its co-crystallized ligand prinomastat (in total 34 MD-derived structures).

3.4 Selection of the best MD-derived structures via virtual screening experiments

Although a number of studies have focused on SOM or inhibition prediction for CYP using data mining ligand-based approaches (Kirchmair *et al.*, 2012), only a few structure-based models have demonstrated their ability to distinguish active from inactive compounds for CYP2D6. Many of those structure-based analyses were based on a very limited number of compounds for 2D6 (de Graaf *et al.*, 2006; Rossato *et al.*, 2010). Our curated dataset contained 343 active and 3002 inactive compounds (Supplementary Fig. S4). We performed virtual screening experiments of this dataset on 34 MD-derived structures and we compared the enrichment results with those of the two experimental structures (Supplementary Fig. S5). The best results were obtained on 6 MD-derived structures: one MD structure from the apo 2F9Q trajectory named MD1 and one from the 3QM4 trajectory without bound ligand named MD2 (Supplementary Fig. S5A), two MD structures from the trajectory of 2F9Q in complex with propafenone named MD3 and MD4 (Supplementary Fig. S5B) and two MD structures from the trajectory of 2F9Q in complex with mexiletine named MD5 and MD6 (Supplementary Fig. S5C). MD1 and MD2 (Supplementary Fig. S5A) retrieved active compounds similarly to the experimental structures. MD3, MD4, MD5 and MD6 (Supplementary Fig. S5C) discriminate better the active compounds than the two experimental structures. From 1–5% of the ranked compounds library (Supplementary Fig. S5B), MD3 performed twice better than the X-ray structures. Similarly, MD5 and MD6 retrieved more active compounds than the X-ray structures in the early stage of enrichment. From a structural point of view, MD1 (Fig. 4A) shows the same features as 2F9Q, meaning that the F' helix is absent. However, the G' helix is shorter and displaced into the inner side of the protein, resulting in a lower active site volume than in the X-ray structure. MD2 (Fig. 4B) shows the same features as the holo X-ray structures 3QM4, with α F' helix present. The α G' helix is shorter and the α helices F, G and A are displaced into the outer side of the protein, resulting in larger binding site volumes (Supplementary Table S1). Results are more surprising for MD3 and MD4 (Fig. 4C). Indeed, none of them contains the F' helix segment and the α G' helix has disappeared in both cases while the volumes of the cavities are larger. Regarding MD5 (Fig. 4D), the α G' helix is shorter and the α F' helix seems to be present. MD6 (Fig. 4E) shows a shorter α G' helix and a broken α F helix into two α F and F' helices, as observed in the 3QM4 structure (Fig. 4F). Interestingly, we were able to find structures derived from a trajectory initially started from the apo 2FQ9 structure with docked mexiletine that did not contain an F' helix (such as MD6) which reproduce structural features of the X-ray holo structure displaying the broken α F helix into α F and F' segments. Among the MD-derived structures showing a good discrimination between active and inactive compounds, we also found MD3 and MD4 with no G' helix present. We analyzed the secondary structures of the MD-derived structures and the five residues that form the α G' helix using the program STRIDE (Heinig and Frishman, 2004). STRIDE was able to identify the α G' helix for the structures MD1 (Fig. 4A), MD5 (Fig. 4D) and MD6 (Fig. 4E). Observing the structures MD3 and MD4, the α G' helix seems to disappear. This is supported by STRIDE, which assigned a 3:10 helix for MD3 structure (Fig. 4C) and a random coil for MD4 structure (Fig. 4C) without helix. It may be possible that the α G' helix goes easily through transitory unstructured states along the MD trajectory taking into account its small number of residues. We also investigated the appearance of the α F' helix in MD6 structure using STRIDE. The α F' helix is composed of five residues as the α G' helix. STRIDE assigns an α helix for MD6 structure, which is consistent with our observation. For the other structures, no α F' helix was found. These results demonstrate that our MD simulations correctly suggested that the helix F can be easily broken due to ligand binding, well supported by the available X-ray structures of CYP2D6 (Wang *et al.*, 2012a, b, 2015). Our protocol allowed to

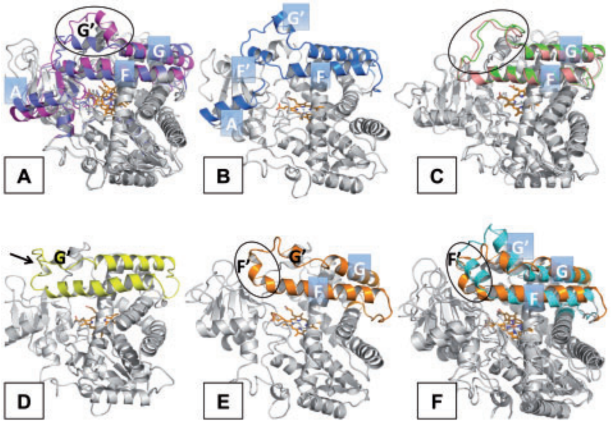

Fig. 4. MD-derived structures with the heme moiety shown in orange sticks. (**A**) Superimposition of MD1 colored in mauve and the X-ray apo structure 2F9Q colored in magenta. The circle shows the displacement of the α G′ helix. (**B**) MD2 structure colored in blue (extracted from the apo 3QM4 trajectory). (**C**) Superimposition of MD3 and MD4 structures colored in green and pink, respectively (extracted from the 2F9Q complexed with propafenone trajectory). The circle shows the absence of α G′ helix. (**D**) MD5 structure colored in yellow (extracted from the 2F9Q complexed with mexiletine trajectory). The black arrow shows the secondary structure, which is found to be the beginning of the α F′ helix. (**E**) MD6 structure colored in orange (extracted from the 2F9Q complexed with mexiletine trajectory). The circle shows the appearance of the F′ helix. (**F**) Superimposition of the MD6 structure colored in orange and the holo X-ray structure 3QM4 colored in cyan. The black circle shows α F′ helices

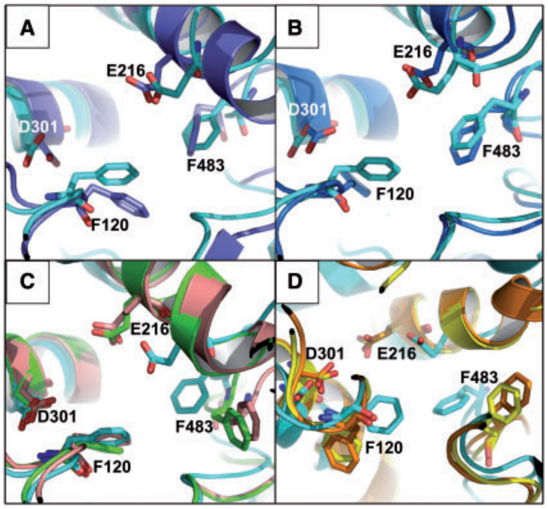

Fig. 5. Key residues of binding sites of the MD-derived structures in comparison with the holo X-ray structure colored in cyan. (**A**) MD1 is colored in mauve; (**B**) MD2 is colored in blue; (**C**) MD3 is colored in green and MD4 in salmon; (**D**) MD5 is colored in yellow and MD6 in orange

Table 1. RMSD of MD-derived structures compared with the holo X-ray structure-binding site

Structure	MD1	MD2	MD3	MD4	MD5	MD6
RMSD (Å)	1.53	1.49	1.81	1.95	1.69	1.90

extract diverse conformations of the binding sites, four of them showing a better or similar discrimination of the active compounds than the X-ray structures.

3.5 Analysis of the best MD-derived binding site structures

We analyzed the differences in the binding site that could be involved in discriminating inhibitors and non-inhibitors. MD1 structure shows enrichment close to the X-ray holo structure (Supplementary Fig. S5A) and displays active site key residues slightly displaced (Fig. 5A), in particular the two Phe residues. MD2 structure (Fig. 5B) as MD1 structure has slightly displaced key residues and changed the orientation of F120. MD3 and MD4 structures (Fig. 5C) show conserved D301 and F120 positions, but a displacement of E216. The most striking difference is due to F483, which exhibits a completely different orientation. The MD5 and MD6 structures (Fig. 5D) show significant structural variations. Indeed, although D301 has a well-conserved orientation, E216 is displaced and both Phe residues have completely different orientations compared with the X-ray structure. RMSD calculations between the MD and X-ray structures confirm these observations (Table 1). Highest RMSD values are obtained for MD3–MD6 structures. MD3, MD4, MD5 and MD6 structures (Fig. 5C and D) show better enrichments of active compounds than the X-ray structure (Supplementary Fig. S5B and C) also having a different orientation of F483. It seems that although both F120 and F483 are responsible for the orientation of the ligands in the cavity, a displacement of F483 allows a better interaction with various ligands. It is also possible to see that a different orientation of F120 in addition to the F483 displacement (MD5 and MD6 structures) (Fig. 5D) facilitates the interactions with bulky ligands. Thus, we have identified MD structures capable of binding small ligands into MD3 and MD4 structures (Fig. 5C), or bulky ligands into MD5 and MD6 structures (Fig. 5D).

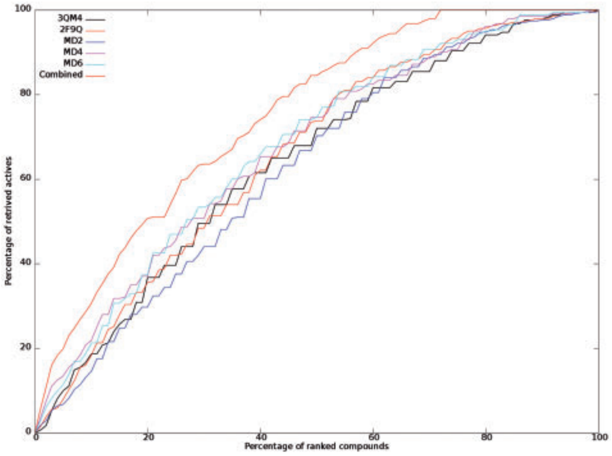

Fig. 6. Enrichment curves obtained using docking into the structures: the X-ray ones: 2FQ9, 3QM4; the MD structures: MD2, MD4 and MD6. The combined enrichment is obtained by taking the best binding energies predicted by docking into MD2, MD4 and MD6

3.6 Small set of MD-derived structures best retrieving the CYP2D6 binders

We attempted to find a small pool of different structures able to bind different active compounds of CYP2D6. Keeping in mind the observed phenomenon of ‘ligand memory’ of the corresponding binding site (Rueda *et al.*, 2012), we considered three MD structures, MD2, MD4 and MD6, extracted from MD trajectories with the three different substrates, showing interesting differences in the binding site. These MD structures exhibit differences, particularly nearer to helices F’ and G’ (Fig. 4), and also different conformations of key binding site residues (Fig. 5). They all have a good druggability score (Supplementary Table S1), as well as similar or better enrichment of CYP2D6 binders than the X-ray structures (Supplementary Fig. S5). Thus, our MD-derived structures seem to present complementary binding pocket profiles permitting to accommodate different ligands. We calculated the combined enrichment of the three structures, MD2, MD4 and MD6, meaning that for each percent of the chemical library, we counted all different active molecules that are retrieved by these three structures. The combined enrichment of the three MD structures indeed shows better results. In fact, 70% of CYP2D6 binders were found in 35% of the ranked dataset when combining docking into the three MD structures (Fig. 6). These results demonstrated that we identified three complementary conformations of the active site that can accommodate the binding of diverse ligands.

3.7 Structure-based QSAR classification models for CYP2D6

The CYP2D6 X-ray structures and the best performing MD structures in terms of distinguishing active compounds were considered in order to train classification models that take into account protein structure information. We performed modeling using three learning algorithms: SVM, RandomForest and NaiveBayesian. We built combined models based on topological information of compound structures using ECFP descriptors and the predicted binding energies computed on both X-ray and MD protein conformations that best retrieved the active compounds (Table 2). Accuracies are based on a

Table 2. Performance of the combined QSAR models for different CYP2D6 conformations

Structure	Method	Accuracy of QSAR model without binding energy information (LOO cross validation	Accuracy of QSAR model including binding energy information (LOO cross validation)	Accuracy of QSAR models including binding energy information or CypRules models (external sets)
X-Ray apo (2F9Q)	SVM	76.3%	77.85%	72.4%
	RandomForest	75.2%	75.00%	75.8%
	NaiveBayesian	72.1%	73.80%	74.1%
	CypRules			70.9%
X-Ray holo (3QM4)	SVM	77.8%	78.4%	74.6%
	RandomForest	75.3%	74.6%	75.2%
	NaiveBayesian	72.7%	72.8%	72.3%
	CypRules			70.7%
MD2	SVM	77.6%	78.0%	74.6%
	RandomForest	74.2%	74.7%	75.1%
	NaiveBayesian	70.2%	73.1%	70.2%
	CypRules			68.26%
MD4	SVM	78.3%	78.6%	74.7%
	RandomForest	74.8%	76.0%	75.6%
	NaiveBayesian	71.2%	72.2%	72.4%
	CypRules			70.5%
MD6	SVM	77.3%	78.9%	74.4%
	RandomForest	74.5%	74.5%	74.0%
	NaiveBayesian	73.1%	73.3%	73.9%
	CypRules			70.4%

Note: The performance of CypRules, a method that uses compound structural information (Shao *et al.*, 2015), is shown for the same external sets.

leave-one-out (LOO) cross-validation. Each test was repeated randomly 10 times with average accuracies reported in Table 2. Interestingly, all SVM models that included binding energies slightly improved the prediction with an increased accuracy (the percentage of correctly predicted inhibitors and non-inhibitors), for example from 77.3% to 78.9% for MD6 structure. Regarding the external dataset, the accuracy of our models is competitive with the recently reported rule-based tool CypRules predicting CYP2D6 inhibition (Shao *et al.*, 2015) (Table 2). Our inhibitor training set involves 343 compounds, which cover most of the chemical diversity of conventional drugs metabolized by CYP2D6 (molecular weight 170 to 512, XlogP 0.4 to 6). The inhibition of CYP2D6 is mainly competitive and our training set does not contain irreversible inhibitors of CYP2D6. Thus, we believe that our approach covers CYP2D6 competitive inhibitors. For compounds that do not follow competitive inhibition, the mechanisms of interactions with the active site could be quite different. Irreversible inhibitors could have a different MD behavior from those characterized in this study and may require additional binding site conformations. In future, new models may be created based on different CYP2D6 inhibitory mechanisms.

Our model shows that taking into account structural information and conformational changes of CYP2D6 binding site allows to improve traditional QSAR models and to better understand the mechanism of inhibition. Furthermore, CYP2D6 is the most polymorphic isoform of CYP with more than 105 allelic variants identified to date (http://www.cypalleles.ki.se/) and structure-based approaches will be essential to account for patient mutations in order to better predict CYP2D6 drug metabolism and inhibition for personalized medications.

4 Conclusion

We report an original *in silico* approach for CYP2D6 inhibition based on the knowledge of the protein structure and dynamic behavior due to various ligand binding combined with machine learning modeling validated on a large number of active and inactive compounds. We explored a large portion of the conformational space of CYP2D6 using MD simulations with three different substrates that do not have an experimentally known bioactive conformation and identified MD structures displaying better performance than the X-ray apo and holo structures with regard to distinguishing between active and inactive compounds. In addition, we identified three MD-derived structures that are capable all together to better retrieve the active compounds compared with individual CYP2D6 conformations, confirming that these three binding site conformations have different and complementary substrate profiles. Our models predicting CYP2D6 inhibition showed an accuracy of 75% on the external validation set, results competitive with other recently reported prediction models.

Funding

Support from Servier, (INSERM) the French National Research Institute and the University Paris Diderot is greatly appreciated. V.M. is supported by the ED 563 MTCI and the ERC advanced grant of Prof. P. Ronco, Inserm 1155, Tenon Hospital. P.C. is supported by UPFellows program with the support of the Marie Curie COFUND program and by BBSRC/Engineering and Physical Sciences Research Council (EPSRC) grant BB/M017702/1.

Conflict of interest: none declared.

References

Ai,N. *et al.* (2015) In silico methods for predicting drug-drug interactions with cytochrome P-450s, transporters and beyond. *Adv. Drug Deliv. Rev.*, **86**, 46–60.

Bernstein,F.C. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Bode,C. (2010) The nasty surprise of a complex drug-drug interaction. *Drug Discov. Today*, **15**, 391–395.

Brändén,G. *et al.* (2014) Structure-based ligand design to overcome CYP inhibition in drug discovery projects. *Drug Discov. Today*, **19**, 905–911.

Brooks,B.R. *et al.* (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.*, **4**, 187–217.

Carbonell,P. *et al.* (2013) Stereo signature molecular descriptor. *J. Chem. Inf. Model*, **53**, 887–897.

Cruciani,G. *et al.* (2005) MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.*, **48**, 6970–6979.

de Graaf,C. *et al.* (2006) Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.*, **49**, 2417–2430.

Ferguson,C.S. and Tyndale,R.F. (2011) Cytochrome P450 enzymes in the brain: emerging evidence of biological significance. *Trends Pharmacol. Sci.*, **32**, 708–714.

Flanagan,J.U. *et al.* (2004) Phe120 contributes to the regiospecificity of cytochrome P450 2D6: mutation leads to the formation of a novel dextromethorphan metabolite. *Biochem J.*, **380**, 353–360.

Haberthür,U. and Caflisch,A. (2008) FACTS: fast analytical continuum treatment of solvation. *J. Comput. Chem.*, **29**, 701–715.

Hanna,I.H. *et al.* (2001) Diversity in mechanisms of substrate oxidation by cytochrome P450 2D6. Lack of an allosteric role of NADPH-cytochrome P450 reductase in catalytic regioselectivity. *J. Biol. Chem.*, **276**, 39553–39561.

Heinig,M. and Frishman,D. (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.*, **32**, W500–W502.

Hritz,J. *et al.* (2008) Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: a combined approach of molecular dynamics and ligand docking. *J. Med. Chem.*, **51**, 7469–7477.

Ingelman-Sundberg,M. *et al.* (2007) Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacol. Ther.*, **116**, 496–526.

Ito,Y. *et al.* (2008) Analysis of CYP2D6 substrate interactions by computational methods. *J. Mol. Graph. Model.*, **26**, 947–956.

Johansson,I. and Ingelman-Sundberg,M. (2011) Genetic polymorphism and toxicology—with emphasis on cytochrome p450. *Toxicol. Sci.*, **120**, 1–13.

Karatzoglou,A. *et al.* (2004) kernlab—an S4 package for kernel methods in R. *J. Stat. Softw.*, **11**, 1–20.

Kemp,C.A. *et al.* (2004) Validation of model of cytochrome P450 2D6: an in silico tool for predicting metabolism and inhibition. *J. Med. Chem.*, **47**, 5340–5346.

Kirchmair,J. *et al.* (2012) Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.*, **52**, 617–648.

Kirton,S.B. *et al.* (2002) Impact of incorporating the 2C5 crystal structure into comparative models of cytochrome P450 2D6. *Proteins*, **49**, 216–231.

Kjellander,B. *et al.* (2007) Exploration of enzyme-ligand interactions in CYP2D6 & 3A4 homology models and crystal structures using a novel computational approach. *J. Chem. Inf. Model.*, **47**, 1234–1247.

Kuhn,M. (2008) Building predictive models in R using the caret package. *J. Stat. Softw.*, **28**, 1–26.

Lagorce,D. *et al.* (2011) The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics*, **27**, 2018–2020.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News* 2, **3**, 18–22.

Livezey,M. *et al.* (2012) Molecular analysis and modeling of inactivation of human CYP2D6 by four mechanism based inactivators. *Drug Metab. Lett.*, **6**, 7–14.

Mackerell,A.D. *et al.* (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, **25**, 1400–1415.

Marechal,J.D. *et al.* (2008) Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. *Br. J. Pharmacol.*, **153** (Suppl. 1), S82–S89.

Martinez-Sanz,J. *et al.* (2013) New QSAR models for human cytochromes P450, 1A2, 2D6 and 3A4 implicated in the metabolism of drugs. Relevance of dataset on model development. *Mol. Informatics*, **32**, 573–577.

Martiny,V.Y. and Miteva,M.A. (2013) Advances in molecular modeling of human cytochrome P450 polymorphism. *J Mol Biol.*, **425**, 3978–3992.

Miteva,M.A. *et al.* (2010) Frog2: efficient 3D conformation ensemble generator for small compounds. *Nucleic Acids Res.*, **38**, W622–W627.

Miteva,M.A. *et al.* (2005) PCE: web tools to compute protein continuum electrostatics. *Nucleic Acids Res.*, **33**, W372–W375.

Moroy,G. *et al.* (2012) Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discov. Today*, **17**, 44–55.

Morris,G.M. *et al.* (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.*, **30**, 2785–2791.

Nair,P.C. and Miners,J.O. (2014) Molecular dynamics simulations: from structure function relationships to drug discovery. *In Silico Pharmacol.*, **2**, 4.

Paine,M.J. *et al.* (2003) Residues glutamate 216 and aspartate 301 are key determinants of substrate specificity and product regioselectivity in cytochrome P450 2D6. *J. Biol. Chem.*, **278**, 4021–4027.

Pinto,N. and Dolan,M.E. (2011) Clinically relevant genetic variations in drug metabolizing enzymes. *Curr. Drug Metab.*, **12**, 487–497.

Porcelli,S. *et al.* (2011) Genetic polymorphisms of cytochrome P450 enzymes and antidepressant metabolism. *Expert Opin. Drug Metab. Toxicol.*, **7**, 1101–1115.

RDevelopmentCoreTeam. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodriguez-Antona,C. and Ingelman-Sundberg,M. (2006) Cytochrome P450 pharmacogenetics and cancer. *Oncogene*, **25**, 1679–1691.

Rossato,G. *et al.* (2010) Probing small-molecule binding to cytochrome P450 2D6 and 2C9: an in silico protocol for generating toxicity alerts. *ChemMedChem*, **5**, 2088–2101.

Rowland,P. *et al.* (2006) Crystal structure of human cytochrome P450 2D6. *J. Biol. Chem.*, **281**, 7614–7622.

Rueda,M. *et al.* (2012) ALiBERO: evolving a team of complementary pocket conformations rather than a single leader. *J. Chem. Inf. Model.*, **52**, 2705–2714.

Ryckaert,J.P. *et al.* (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, **23**, 327–341.

Sali,A. *et al.* (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318–326.

Shao,C.Y. *et al.* (2015) CypRules: a rule-based P450 inhibition prediction server. *Bioinformatics*, **31**, 1869–1871.

Shimada,T. (2006) Xenobiotic-metabolizing enzymes involved in activation and detoxification of carcinogenic polycyclic aromatic hydrocarbons. *Drug Metab. Pharmacokinet.*, **21**, 257–276.

Sim,S.C. and Ingelman-Sundberg,M. (2010) The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum. Genomics*, **4**, 278–281.

Singh,D. *et al.* (2011) Novel advances in cytochrome P450 research. *Drug Discov. Today*, **16**, 793–799.

Stoll,F. *et al.* (2011) Utility of protein structures in overcoming ADMET-related issues of drug-like compounds. *Drug Discov. Today*, **16**, 530–538.

Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.

Tyzack,J.D. *et al.* (2013) Prediction of cytochrome P450 xenobiotic metabolism: tethered docking and reactivity derived from ligand molecular orbital analysis. *J. Chem. Inf. Model.*, **53**, 1294–1305.

Vanommeslaeghe,K. *et al.* (2012) Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.*, **52**, 3155–3168.

Volkamer,A. *et al.* (2012) Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.*, **52**, 360–372.

Wang,A. *et al.* (2012a) Crystal structure of human cytochrome P450 2D6 with prinomastat bound. *J. Biol. Chem.*, **287**, 10834–10843.

Wang,A. *et al.* (2015) Contributions of ionic interactions and protein dynamics to cytochrome P450 2D6 (CYP2D6) substrate and inhibitor binding. *J. Biol. Chem.*, **290**, 5092–5104.

Wang,Y. *et al.* (2012b) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.