Mechanistically-based QSARs to Describe Metabolic Constants in Mammals

Alessandra Pirovano,1 Mark A.J. Huijbregts,1 Ad M.J. Ragas,1,2 Karin Veltman1 and A. Jan Hendriks1

1Radboud University Nijmegen, Institute for Wetland and Water Research, Department of Environmental Science, Nijmegen, The Netherlands; 2School of Science, Open University, Heerlen, The Netherlands

Summary — Biotransformation is one of the processes which influence the bioaccumulation of chemicals. The enzymatic action of metabolism involves two processes, i.e. the binding of the substrate to the enzyme followed by a catalytic reaction, which are described by the Michaelis–Menten constant ($K_m$) and the maximum rate ($V_{max}$). Here, we developed Quantitative Structure–Activity Relationships (QSARs) for Log($1/K_m$) and Log($V_{max}$) for substrates of four enzyme classes. We focused on oxidations catalysed by alcohol dehydrogenase (ADH), aldehyde dehydrogenase (ALDH), flavin-containing monooxygenase (FMO) and cytochrome P450 (CYP) in mammals. The chemicals investigated were xenobiotics, including alcohols, aldehydes, pesticides and drugs. We applied general linear models for this purpose, employing descriptors related to partitioning, geometric characteristics, and electronic properties of the substrates, which can be interpreted mechanistically. The explained variance of the QSARs varied between 20% and 70%, and it was larger for Log($1/K_m$) than for Log($V_{max}$). The increase of $1/K_m$ with compound logP and size suggests that weak interactions are important, e.g. by substrate binding via desolvation processes. The importance of electronic factors for $1/K_m$ was described in relation to the catalytic mechanism of the enzymes. $V_{max}$ was particularly influenced by electronic properties, such as dipole moment and energy of the lowest unoccupied molecular orbital. This can be explained by the nature of the catalysis, characterised by the cleavage and formation of covalent or ionic bonds (strong interactions). The present study may be helpful to understand the underlying principles of the chemical specific activity of four important oxidising enzymes.

Key words: biotransformation, $K_m$, QSAR, $V_{max}$

Address for correspondence: Alessandra Pirovano, Radboud University Nijmegen, Institute for Wetland and Water Research, Department of Environmental Science, Heyendaalweg 135, 6525 AJ Nijmegen, The Netherlands.
E-mail: A.Pirovano@science.ru.nl

Introduction

The bioaccumulation potential of chemicals in organisms is a vital element in environmental risk assessment (1). The accumulation of a chemical is the result of a series of physiological and physical processes: absorption, distribution, metabolism and excretion (ADME). Metabolism, also referred to as biotransformation in the case of xenobiotics (1), occurs via enzymatic reactions involving two processes. Firstly, the chemical needs to reach the enzyme and bind to it; secondly, a catalytic reaction has to take place. The latter process is described by the maximum rate of reaction ($V_{max}$) at saturating substrate concentration (2). Alternatively, $V_{max}$ can be expressed as turnover number ($k_{cat}$, with units of time$^{-1}$), which represents the number of substrate molecules converted into product per enzyme molecule per unit time, when the enzyme is saturated with substrate (3). The other parameter used to characterise an enzymatic reaction is the Michaelis–Menten constant ($K_m$), which is the substrate concentration at half $V_{max}$. $K_m$ is equal to the ratio ($k_{cat}$ + $k_{-1}$)/$k_{1}$, where $k_{-1}$ and $k_{1}$ are constants, respectively, for breakdown and formation of the complex enzyme–substrate (ES) (4). If $k_{cat}$ is smaller than $k_{-1}$, $K_m$ is assumed to be equal to the dissociation constant $K_d$ for the ES complex. In this case, $1/K_m$ reflects the affinity of the enzyme for its substrate: a low $K_m$ (or high $1/K_m$) corresponds to high binding affinity. For reactions that exhibit Michaelis–Menten kinetics and at non-saturating substrate concentrations, the ratio $V_{max}/K_m$ provides an estimation of the intrinsic clearance (CL$_{int}$) (5–6). CL$_{int}$, which is a measure of enzyme activity toward a compound, can be extrapolated to an equivalent whole-body metabolic rate, required for risk assessment (7). Several studies (8, 9) have shown the importance of Quantitative Structure–Activity Relationships (QSARs) for the investigation of $K_m$ and $V_{max}$, most of which focused on drugs oxidised by cytochrome P450 (CYP). The binding to the enzyme, represented by $1/K_m$, was shown to be mainly related to compound hydrophobicity (2, 10), probably due to desolvation effects, although electronic and geometric
factors, such as polarity and size, can also be important (9). The rate appears to be influenced by electronic properties, such as frontier orbital energies or hydrogen bonding properties (11–13). In fact, catalytic processes are characterised by the cleavage and formation of covalent bonds (2). However, the above-mentioned studies focused on a particular series of P450 substrates, implying applicability only for specific combinations of chemicals and P450 enzymes. Recently, Pirovano et al. (14) studied the relationships between $1/K_m$ and hydrophobicity, i.e. the octanol–water partitioning coefficient ($K_{ow}$), for a broader set of chemicals and oxidising enzymes in mammals. The chemicals investigated were xenobiotics such as alcohols, aldehydes, drugs and pesticides. The enzymes examined, in addition to CYP, were alcohol dehydrogenase (ADH), aldehyde dehydrogenase (ALDH) and flavin-containing monoxygenase (FMO).

In the present study, we extended our analysis to other descriptors, which were chosen on the basis of mechanistic considerations. Furthermore, not only did we investigate descriptors for $1/K_m$ but also for $V_{max}$. The aim of the current study was to develop QSARs with $\log(1/K_m)$ and $\log V_{max}$ as endpoints for ADH, ALDH, FMO and CYP enzymes in mammals. General linear models were built with descriptors related to partitioning, as well as geometric and electronic properties of the substrates.

Materials and Methods

An extensive version of the Materials and Methods section is reported in the Supplementary Information, which is available at: http://www.eco-itn.eu.

Experimental data set

Data collection

We considered the following enzymes: alcohol dehydrogenase (ADH), aldehyde dehydrogenase (ALDH), flavin-containing monoxygenase (FMO) and cytochrome P450 (CYP). $K_m$ and catalytic rates (expressed either as $V_{max}$ or $k_{cat}$) were taken from BRENDA enzyme database (15) and published reviews (8, 16–18). Constants, measured for mammals in in vitro assays of purified, non-recombinant, hepatic enzymes, were selected. All the data were checked in the original papers and are reported in Table S1 in the Supplementary Information. The substrates collated were mainly drugs and compounds found in the environment. Each compound was assigned to a relevant chemical class by using ECOSAR v 1.0, a program present in EPI Suite (19).

Data treatment

Michaelis–Menten constants ($K_m$) were expressed in $\mu$M. Since catalytic rates were reported in heterogeneous units and with different constants (as $V_{max}$ or as $k_{cat}$), it was necessary to standardise the data. We expressed all rates as $V_{max}$ with $\mu$mol/min/$mg_{PROTEIN}$ as units. For CYP enzymes, $V_{max}$ was stated with respect to microsomal protein weight, i.e. $mg_{PROTEIN} = mg_{MICR PROT}$. We transformed $k_{cat}$ (min$^{-1}$) into $V_{max}$ values ($\mu$mol/min/$mg_{PROTEIN}$) by multiplying $k_{cat}$ by the specific content of the enzyme ($E$; $\mu$mol/$mg_{MICR PROT}$) for CYP (11) and dividing $k_{cat}$ by the molecular weight of the enzyme ($M_r$; $mg_{ENZ}/\mu$mol) for the other enzymes. If $M_r$ or $E$ values were not reported in the paper from which we collected $k_{cat}$, we used average values from other studies (Table S2 in the Supplementary Information).

$K_m$ and $V_{max}$ data for different substrates were combined into four data sets, one for each enzyme family. Each substrate was characterised by a single value of $1/K_m$ or $V_{max}$; if multiple values were available for one substrate, we calculated the geometric mean of the experimental $1/K_m$ or $V_{max}$ values, as well as the geometric standard deviation.

Descriptors and QSAR models

Descriptor calculation and selection

We compiled a list of physicochemical descriptors based on mechanistic considerations. We anticipated $1/K_m$ and $V_{max}$ to be related to the partitioning, geometric and electronic properties of the substrates of P450 (11, 16, 20, 21). Therefore, we collected the descriptors (18 in total) used in the QSARs for $\log(1/K_m)$ or $\log V_{max}$ in the above-mentioned studies. We hypothesised that they could be applied to all four enzyme classes, as they were among the descriptors commonly used to describe biological responses to xenobiotics (22). The descriptors were computed with Chemaxon (http://www.chemaxon.com) through the OCHEM platform (23), and with MOPAC2009 (Hamiltonian AM1) (24) by using Vega ZZ v2.4.0 (25). For the calculation of all descriptors, the molecular conformations were optimised with MOPAC. A correlation matrix was calculated to detect collinear descriptors, i.e. with correlation coefficients ($R$) > 0.8 or < -0.8. Among the collinear descriptors, we retained the one that we considered easiest to interpret mechanistically.

The final set of descriptors is reported in Table 1, together with a brief explanation and the software used to compute them. The descriptors were autoscaled to zero-mean and unit-variance to ensure equal contribution of all variables in the models.
**Model development**

General linear models (GLM) were developed for Log(1/K_m) and LogV_max with the software R v.2.15.1 (26). We used the R package 'bestglm' (27) to select the best subset of variables for the linear regression after an exhaustive search. In order to avoid over-fitting, we set the maximum number of variables to be included in the subsets at six (28). We selected the best model as the one with the lowest Akaike's Information Criterion (AIC) value. We performed a final check for collinearity of the descriptors in the individual QSARs by using variance inflation factors (VIFs), calculated with the R package ‘car’ (29). The threshold for collinearity was VIF > 3 (30). If all the variables had VIFs < 3, the QSAR was accepted; otherwise, the variable with the highest VIF was removed from the data set and the ‘bestglm’ method was performed again, until all the VIF values were smaller than the threshold (31). The models were internally validated with the leave-one-out (LOO) procedure with WEKA v.3.6.7 (32).

For each model, the fitting is evaluated based on the coefficient of determination (R^2), the adjusted R^2 (R^2_adj), the Root Mean Squared Error (RMSE), and the p-value from the F-test (p). We report the LOO cross-validated R^2 (Q^2_LOO) and RMSE (RMSE_LOO) to assess the internal predictivity of the models.

**Additional regressions**

In our previous work on the relationship between 1/K_m and hydrophobicity (14), we observed two groups of compounds that were outliers: 22 substituted benzaldehydes for ALDH (Table S3 in the Supplementary Information) and 52 ‘non-specific’ chemicals for CYP (mainly Neutral Organics, according to the ECOSAR classification). Therefore, in this work, we have also investigated the possible influence of these classes of compound in the QSARs. We developed two additional sets of QSARs for both ALDH and CYP: one with all the compounds except the group of outliers (ALDH_1 and CYP_1), and another with only the group of outliers (ALDH_2 and CYP_2). For the QSARs with the 22 substituted benzaldehydes, the maximum number of variables to be selected by the algorithm was set to four, due to the relatively low number of compounds. We have also developed an overall regression for Log(1/K_m), by merging all data from the four data sets and adding a qualitative variable called ‘Enzyme’ with four categories (ADH, ALDH, FMO, CYP) representing the enzyme group of the data point.

**Results**

The QSARs developed for Log(1/K_m) and LogV_max are presented in Tables 2 and 3, respectively, with the standardised regression coefficients (i.e. the regression coefficients that do not depend on the units and were obtained by using the auto-scaled descriptors). The non-standardised regression coefficients can be found in Tables S4 and S5 in the Supplementary Information. As an example, Figure 1 represents the measured versus the predicted values for Log(1/K_m) and LogV_max for ADH.

**Table 1: Descriptors used to develop the QSARs**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Units</th>
<th>Description</th>
<th>Type</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>logP</td>
<td>[—]</td>
<td>Calculated octanol water partitioning coefficient</td>
<td>Partitioning</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>A</td>
<td>[Å^2]</td>
<td>Van der Waals’ surface area, calculated at pH 7.4</td>
<td>Geometric</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>a/d^2</td>
<td>[—]</td>
<td>area/depth^2</td>
<td>Geometric</td>
<td>MOPAC</td>
</tr>
<tr>
<td>l/w</td>
<td>[—]</td>
<td>length/width</td>
<td>Geometric</td>
<td>MOPAC</td>
</tr>
<tr>
<td>apK_a1</td>
<td>[—]</td>
<td>Strongest acidic pK_a</td>
<td>Electronic</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>bpK_a1</td>
<td>[—]</td>
<td>Strongest basic pK_a</td>
<td>Electronic</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>HBD</td>
<td>[—]</td>
<td>Hydrogen bond donor, calculated at pH 7.4</td>
<td>Electronic</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>HBA</td>
<td>[—]</td>
<td>Hydrogen bond acceptor, calculated at pH 7.4</td>
<td>Electronic</td>
<td>Chemaxon</td>
</tr>
<tr>
<td>ν</td>
<td>[Debye]</td>
<td>Dipole moment</td>
<td>Electronic</td>
<td>MOPAC</td>
</tr>
<tr>
<td>E_HOMO</td>
<td>[eV]</td>
<td>Energy of the highest occupied molecular orbital (HOMO)</td>
<td>Electronic</td>
<td>MOPAC</td>
</tr>
<tr>
<td>E_LUMO</td>
<td>[eV]</td>
<td>Energy of the lowest unoccupied molecular orbital (LUMO)</td>
<td>Electronic</td>
<td>MOPAC</td>
</tr>
<tr>
<td>ΔE_L-H</td>
<td>[eV]</td>
<td>ΔE_L-H = E_LUMO - E_HOMO</td>
<td>Electronic</td>
<td>MOPAC</td>
</tr>
<tr>
<td>H_f</td>
<td>[kcal/mol]</td>
<td>Final heat of formation</td>
<td>Electronic</td>
<td>MOPAC</td>
</tr>
</tbody>
</table>

\(^{a}\)The length, width and depth of a molecule represent molecular dimensions measured orthogonally relative to the main molecular plane (35).
Table 2: The variables selected and their standardised regression coefficients, together with the regression statistics for Log(1/Km)

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>logP</th>
<th>A</th>
<th>a/d²</th>
<th>l/w</th>
<th>apK_a1</th>
<th>bpK_a1</th>
<th>HBD</th>
<th>HBA</th>
<th>ν</th>
<th>E_HOMO</th>
<th>E_LUMO</th>
<th>ΔE_L–H</th>
<th>Hf</th>
<th>Intercc.</th>
<th>ALDH</th>
<th>FMO</th>
<th>CYP</th>
<th>n</th>
<th>R²</th>
<th>R²(adj)</th>
<th>RMSE</th>
<th>p</th>
<th>Q²_LOO</th>
<th>RMSE_LOO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH</td>
<td>0.98 (± 0.16)</td>
<td>-0.24 (± 0.13)*</td>
<td>-0.34 (± 0.15)</td>
<td>-0.25 (± 0.15)*</td>
<td>-0.36 (± 0.13)</td>
<td>-2.66 (± 0.12)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>34</td>
<td>0.73</td>
<td>0.68</td>
<td>0.62</td>
<td>&lt; 0.01</td>
<td>0.57</td>
<td>0.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDH</td>
<td>0.64 (± 0.18)</td>
<td>0.82 (± 0.19)</td>
<td>-0.30 (± 0.14)</td>
<td>0.54 (± 0.20)</td>
<td>0.51 (± 0.21)</td>
<td>0.39 (± 0.15)</td>
<td>-0.18 (± 0.13)*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>77</td>
<td>0.56</td>
<td>0.52</td>
<td>1.06</td>
<td>&lt; 0.01</td>
<td>0.47</td>
<td>1.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMO</td>
<td>0.25 (± 0.08)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-0.42 (± 0.07)</td>
<td>-0.30 (± 0.07)</td>
<td>-1.99 (± 0.06)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>149</td>
<td>0.39</td>
<td>0.37</td>
<td>0.63</td>
<td>&lt; 0.01</td>
<td>0.35</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP</td>
<td>0.30 (± 0.07)</td>
<td>0.20 (± 0.07)</td>
<td>-0.18 (± 0.07)</td>
<td>-0.10 (± 0.07)*</td>
<td>-0.36 (± 0.07)</td>
<td>-2.73 (± 0.06)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>121</td>
<td>0.40</td>
<td>0.37</td>
<td>0.68</td>
<td>&lt; 0.01</td>
<td>0.30</td>
<td>0.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>0.41 (± 0.06)</td>
<td>-0.09 (± 0.05)*</td>
<td>0.19 (± 0.05)</td>
<td>—</td>
<td>-0.14 (± 0.07)*</td>
<td>-0.21 (± 0.08)</td>
<td>-1.91 (± 0.19)</td>
<td>1.83 (± 0.22)</td>
<td>-0.28 (± 0.21)*</td>
<td>-0.84 (± 0.20)</td>
<td>381</td>
<td>0.59</td>
<td>0.58</td>
<td>0.94</td>
<td>&lt; 0.01</td>
<td>0.57</td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additional regressions

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>logP</th>
<th>A</th>
<th>a/d²</th>
<th>l/w</th>
<th>apK_a1</th>
<th>bpK_a1</th>
<th>HBD</th>
<th>HBA</th>
<th>ν</th>
<th>E_HOMO</th>
<th>E_LUMO</th>
<th>ΔE_L–H</th>
<th>Hf</th>
<th>Intercc.</th>
<th>ALDH</th>
<th>FMO</th>
<th>CYP</th>
<th>n</th>
<th>R²</th>
<th>R²(adj)</th>
<th>RMSE</th>
<th>p</th>
<th>Q²_LOO</th>
<th>RMSE_LOO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALDH₁</td>
<td>0.65 (± 0.15)</td>
<td>0.63 (± 0.14)</td>
<td>0.24 (± 0.12)*</td>
<td>-0.36 (± 0.11)</td>
<td>0.27 (± 0.12)</td>
<td>-0.28 (± 0.12)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-0.12 (± 0.11)*</td>
<td>—</td>
<td>—</td>
<td>55</td>
<td>0.77</td>
<td>0.74</td>
<td>0.74</td>
<td>&lt; 0.01</td>
<td>0.72</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDH₂</td>
<td>-0.57 (± 0.33)*</td>
<td>1.02 (± 0.31)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.52 (± 0.31)*</td>
<td>-0.34 (± 0.27)*</td>
<td>—</td>
<td>—</td>
<td>22</td>
<td>0.53</td>
<td>0.46</td>
<td>1.16</td>
<td>&lt; 0.01</td>
<td>0.36</td>
<td>1.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP₁</td>
<td>0.32 (± 0.07)</td>
<td>0.36 (± 0.07)</td>
<td>-0.09 (± 0.06)*</td>
<td>-0.09 (± 0.06)*</td>
<td>-0.25 (± 0.07)</td>
<td>0.12 (± 0.06)</td>
<td>-2.90 (± 0.05)</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>65</td>
<td>0.73</td>
<td>0.71</td>
<td>0.39</td>
<td>&lt; 0.01</td>
<td>0.60</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP₂</td>
<td>0.53 (± 0.10)</td>
<td>-0.13 (± 0.10)*</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-0.32 (± 0.10)</td>
<td>-2.65 (± 0.09)</td>
<td>—</td>
<td>—</td>
<td>56</td>
<td>0.52</td>
<td>0.50</td>
<td>0.68</td>
<td>&lt; 0.01</td>
<td>0.45</td>
<td>0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The probability (p) value of the coefficient is greater than 0.05.
For symbols see Table 1. The K_m values were expressed as μM. The most important descriptor of each regression is shown in bold.
Table 3: The variables selected and their standardised regression coefficients, together with the regression statistics for Log $V_{\text{max}}$

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>logP</th>
<th>A</th>
<th>a/d</th>
<th>$\mu$</th>
<th>apK$_{a1}$</th>
<th>bpK$_{a1}$</th>
<th>HBD</th>
<th>HBA</th>
<th>$\nu$</th>
<th>$E_{\text{HOMO}}$</th>
<th>$E_{\text{LUMO}}$</th>
<th>$\Delta E_{L-H}$</th>
<th>Hf</th>
<th>Interc.</th>
<th>n</th>
<th>$R^2$</th>
<th>$R^2_{\text{adj}}$</th>
<th>RMSE</th>
<th>p</th>
<th>$Q^2_{\text{LOO}}$</th>
<th>RMSE$_{\text{LOO}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.44 (±0.08)</td>
<td>-0.44 (±0.08)</td>
<td>-0.25 (±0.09)</td>
<td>0.36 (±0.09)</td>
<td>-0.38 (±0.07)</td>
<td></td>
<td></td>
<td>33</td>
<td>0.53</td>
<td>0.48</td>
<td>0.40</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>ALDH</td>
<td>-0.35 (±0.09)</td>
<td>0.37 (±0.11)</td>
<td></td>
<td>0.10 (±0.07)</td>
<td></td>
<td></td>
<td></td>
<td>-0.17 (±0.09)</td>
<td>-0.17 (±0.09)</td>
<td>0.23 (±0.09)</td>
<td>0.16 (±0.09)</td>
<td>-0.44 (±0.07)</td>
<td></td>
<td></td>
<td>74</td>
<td>0.25</td>
<td>0.19</td>
<td>0.55</td>
<td>&lt;0.01</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td>FMO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
<td></td>
<td></td>
<td>-0.07 (±0.03)</td>
<td>-0.07 (±0.03)</td>
<td>-0.15 (±0.03)</td>
<td>0.04 (±0.03)</td>
<td>-0.19 (±0.03)</td>
<td></td>
<td></td>
<td>94</td>
<td>0.30</td>
<td>0.25</td>
<td>0.26</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>CYP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.14 (±0.06)</td>
<td>-0.14 (±0.06)</td>
<td>-0.09 (±0.06)</td>
<td>-0.09 (±0.06)</td>
<td>-0.16 (±0.05)</td>
<td>0.19 (±0.06)</td>
<td>0.21 (±0.06)</td>
<td>-1.45 (±0.05)</td>
<td>121</td>
<td>0.26</td>
<td>0.22</td>
<td>0.55</td>
</tr>
</tbody>
</table>

**Additional regressions**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>logP</th>
<th>A</th>
<th>a/d</th>
<th>$\mu$</th>
<th>apK$_{a1}$</th>
<th>bpK$_{a1}$</th>
<th>HBD</th>
<th>HBA</th>
<th>$\nu$</th>
<th>$E_{\text{HOMO}}$</th>
<th>$E_{\text{LUMO}}$</th>
<th>$\Delta E_{L-H}$</th>
<th>Hf</th>
<th>Interc.</th>
<th>n</th>
<th>$R^2$</th>
<th>$R^2_{\text{adj}}$</th>
<th>RMSE</th>
<th>p</th>
<th>$Q^2_{\text{LOO}}$</th>
<th>RMSE$_{\text{LOO}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALDH$_1$</td>
<td>-0.37 (±0.09)</td>
<td>0.25 (±0.10)</td>
<td></td>
<td>0.15 (±0.09)</td>
<td></td>
<td></td>
<td></td>
<td>-0.17 (±0.07)</td>
<td>-0.27 (±0.09)</td>
<td>-0.29 (±0.06)</td>
<td></td>
<td></td>
<td>52</td>
<td>0.25</td>
<td>0.17</td>
<td>0.44</td>
<td>0.02</td>
<td>0.12</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALDH$_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.19 (±0.07)</td>
<td>0.34 (±0.16)</td>
<td>0.34 (±0.16)</td>
<td></td>
<td></td>
<td>22</td>
<td>0.29</td>
<td>0.22</td>
<td>0.64</td>
<td>0.04</td>
<td>0.12</td>
<td>0.74</td>
</tr>
<tr>
<td>CYP$_1$</td>
<td>-0.19 (±0.07)</td>
<td>0.34 (±0.09)</td>
<td></td>
<td>0.19 (±0.09)</td>
<td></td>
<td></td>
<td></td>
<td>0.23 (±0.10)</td>
<td>-0.42 (±0.11)</td>
<td>0.24 (±0.10)</td>
<td></td>
<td></td>
<td>65</td>
<td>0.37</td>
<td>0.30</td>
<td>0.54</td>
<td>&lt;0.01</td>
<td>0.24</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP$_2$</td>
<td>0.26 (±0.10)</td>
<td>0.18 (±0.06)</td>
<td></td>
<td>-0.11 (±0.06)</td>
<td>0.14 (±0.07)</td>
<td>-0.29 (±0.08)</td>
<td></td>
<td>0.19 (±0.09)</td>
<td></td>
<td>0.21 (±0.06)</td>
<td>-1.45 (±0.05)</td>
<td>-1.23 (±0.09)</td>
<td></td>
<td></td>
<td>56</td>
<td>0.43</td>
<td>0.36</td>
<td>0.40</td>
<td>&lt;0.01</td>
<td>0.24</td>
<td>0.47</td>
</tr>
</tbody>
</table>

*a The probability (p) value of the coefficient is greater than 0.05.
For symbols see Table 1. The $V_{\text{max}}$ values were expressed as $\mu$mol/min/mgPROT. The most important descriptor of each regression is shown in bold.*
Figure 1: Measured versus predicted values for Log(1/Km) and LogVmax, for compounds metabolised by ADH in mammals.

The solid lines indicate the 1:1 bisector, and the dashed lines indicate ± 2 log units error. Laboratory measurements (dots) for each compound: Log transformed geometrical mean of a) 1/Km [μM\(^{-1}\)] and b) Vmax [μmol/min/mg\text{PROT}], with the geometric standard deviation (horizontal bar).
Log(1/K_m)

Significant correlations (p < 0.05) were obtained for all QSARs for Log(1/K_m) (Table 2), whose R^2_adj and Q^2_LOO varied from 0.37 to 0.74 and from 0.30 to 0.72, respectively. The most common descriptors were area (A), octanol–water partitioning coefficient (logP), and difference between frontier orbital energies (ΔE_{L-H}), all three appearing in six out of nine QSARs. The area had positive regression coefficients, ranging from 0.25 to 1.02. The coefficients of logP and ΔE_{L-H} had positive and negative signs, respectively; in all cases, except for ALDH_3 (QSAR with only the 22 substituted benzaldehydes) and ALDH (only for ΔE_{L-H}). These three descriptors were the most important ones, i.e. with the highest standardised coefficients, in most of the QSARs: the area for ALDH, ALDH_2, and CYP_1 (QSAR without the ’remaining chemicals’); logP for ADH, ALDH_1 (QSAR without the 22 substituted benzaldehydes), and the overall regression; ΔE_{L-H} for CYP. The hydrogen bond acceptor (HBA) had the highest standardised regression coefficient (~0.42) in the QSAR for FMO.

LogV_max

Correlations significant at the 0.05 level were obtained for all QSARs for LogV_max (Table 3). The goodness of fit and the internal predictivity were lower for LogV_max, if compared to Log(1/K_m), with R^2_adj and Q^2_LOO varying from 0.17 to 0.48 and from 0.12 to 0.41, respectively. The most common descriptor, appearing in six out of nine QSARs, was the dipole moment (υ), with coefficients ranging from –0.42 to 0.36. It was also the most important descriptor in the QSARs for ALDH_1 and CYP_1. The area (A) featured in four QSARs with a positive regression coefficient; it had the highest standardised coefficient (0.37) in the QSARs for ALDH and ALDH_2. LogP occurred in three QSARs with a negative regression coefficient, and, with a standardised coefficient of –0.27, it was the most important descriptor for ALDH_1, together with the dipole moment. Among the other descriptors, apK_c1, HBA, and E_{LUMO} had the highest correlation coefficients for FMO (~0.15), CYP_2 (~0.29), and ADH (~0.44), respectively. H_f was the most important descriptor for CYP, with a standardised coefficient of 0.21.

Discussion

Regressions

In this study, QSAR models were developed for Log(1/K_m) and LogV_max of four groups of mammalian enzymes. We used relevant physicochemical descriptors reflecting the hydrophobic, geometric, and electronic properties of the chemicals. Common features were found within the QSARs for Log(1/K_m) and LogV_max, despite the different reaction types of the four enzymes considered. Log(1/K_m) was largely controlled by hydrophobicity (logP), as well as area (A) and frontier orbital energy (ΔE_{L-H}), while the rate (V_max) was mainly influenced by electronic parameters, such as dipole moment (υ), hydrogen bonding properties (HBD and HBA), and energy of the lowest unoccupied molecular orbital (E_{LUMO}). The difference in the molecular properties controlling Log(1/K_m) and LogV_max was expected from the nature of the processes underlying these two constants. The inverse of K_m is usually assumed to be equal to the affinity constant for enzyme binding, which is generally a desolvation process; thus, it is controlled mainly by hydrophobicity. Yet, this equivalence is valid only if the enzymatic process is composed of two steps — formation of the ES complex and successive catalysis — and if the latter is lower than the dissociation of the substrate from the enzyme. The V_max represents the catalytic process, which is characterised by the cleavage and formation of covalent bonds; thus, it is more influenced by the electronic properties of the substrates (2).

The variability explained by the QSARs ranged from 20% to 70% (R^2_adj in Tables 2 and 3). The correlations improved substantially for Log(1/K_m) by leaving out distinct substance groups such as substituted benzaldehydes. Weak correlations may indicate that the underlying catalytic reactions are complex and only partly related to the physicochemical descriptors chosen (33). The fit of the QSARs could be improved by using theoretical molecular descriptors, i.e. calculated by mathematical formulae or computational algorithms (34), which are able to represent other aspects of molecular structures, such as topological indices and functional group counts. Yet, we did not include these descriptors in the present paper, because the objective was to allow for the mechanistic interpretation of the QSARs.

The QSARs in the present work had lower R^2 values in comparison to the QSARs for CYP developed in other studies, whose R^2 values were around 0.8–0.9 (8, 9, 11, 16). Yet, the latter data sets typically included homologous series of about ten structurally-related compounds, metabolised by a given isoenzyme in one mammalian species. Thus, those models are applicable only to very specific combinations of compounds, isoenzymes and species, for which a similar behaviour can be expected.

The data sets consisted of compounds assigned to ECOSAR classes and known to be substrates of the enzymes considered in this study. The applicability domains of the QSARs are defined by the range (min. and max.) of the values of the descriptors used.
to build the model (35), which are reported in Tables S6 and S7 in the Supplementary Information for Log(1/Km) and LogV_{max} respectively.

The experimental data come from different laboratories, often employing different protocols (36), e.g. for pH and temperature conditions, which can affect enzyme activity (3). In addition, the rates were reported in the papers either as V_{max} or k_{cat} values. The latter were transformed into V_{max} (Supplementary Information, Table S2) by using the conversion factors reported in the papers from which we collected k_{cat}, when available; otherwise, we used average values obtained in other studies. Consequently, the input data are subject to variation, implying uncertainty in the QSARs. Furthermore, we merged data measured for different mammalian species (human, horse, rat, mouse, pig and rabbit) and isoenzymes (any of the several forms of an enzyme, all of which catalyse the same reaction but are characterised by different properties). This can be another source of unexplained variation; however, the focus of the present work was on general features of the metabolic process.

We built four general QSARs each for Log(1/Km) and LogV_{max}, one for every enzyme. In our previous study (14) on the relationships between hydrophobicity and Log(1/Km), we found an improvement in the regressions after the removal of two groups of influential chemicals: 22 substituted benzaldehydes for ADH and 56 ‘remaining chemicals’ (chemicals belonging to non-specific ECOSAR classes, mainly Neutral Organics) for CYP. Hence, in the present study, we developed four additional QSAR sub-models each for Log(1/Km) and LogV_{max}, one without and one with the only groups of influential chemicals. For ALDH, the fitting increased with respect to the general QSAR only for the sub-model built for Log(1/Km), excluding the substituted benzaldehydes (ALDH1). For both endpoints, the most important descriptor was the area for the substituted benzaldehydes and logP for the other aldehydes. For CYP, this subdivision led to QSAR sub-models with improved fitting for both Log(1/Km) and LogV_{max}, although for the latter the Q^2_{LOO} values were low (around 0.2). It appears that the enzymatic constants can be dependent on chemical class. The ‘remaining chemicals’ for CYP may have different abilities to fit into and interact with the enzyme active site.

Mechanistic explanation

The QSARs developed in this work were generally in line with previous studies on enzyme metabolism, mainly concerning CYP enzymes (9). In the following paragraphs, the influencing descriptors in the QSARs are explained in relation to the catalytic cycles of the enzymes. Liver ADH catalyses the reversible transformation of alcohols to the corresponding aldehydes or ketones. ALDH enzymes oxidise a wide range of aldehydes to their corresponding carboxylic acids (37). FMO oxygenates various xenobiotics, such as pesticides and drugs, containing a nucleophilic heteroatom (usually sulphur and nitrogen) (18). The oxygen abstraction takes place before binding via a nucleophilic attack by the substrate. The CYP enzymes usually catalyse monooxygenase reactions involving the insertion of an oxygen atom into a substrate (11).

The hydrophobicity (logP) featured in many QSARs for Log(1/Km), for which it had a positive correlation coefficient, with the exception of the QSAR for ‘substituted benzaldehydes’. The increase of 1/Km with compound hydrophobicity is likely to indicate the importance of weak interactions such as substrate binding via desolvation processes, i.e. displacement of water molecules due to the binding of the substrate in the active site (38). The different behaviour of substituted benzaldehydes was observed in our previous work relating Log(1/Km) to compound hydrophobicity (equations shown in Supplementary Information Tables S8 and S9). In the work of Klyosov (39), correlations between the Km of aldehydes and their hydrophobicity (expressed in terms of Hansch constant, π) were found for all aldehydes tested, except substituted benzaldehydes. In our QSARs for LogV_{max}, logP featured only in three QSARs, which is in accordance with the common understanding that rates are not likely to be influenced by partitioning properties. In addition, logP had a negative coefficient for LogV_{max}, indicating that hydrophobicity disfavours the catalysis of the substrates.

Geometric properties of the substrates were included in several QSARs, the most frequent being the molecular area (A), always with a positive regression coefficient. The area was often the most important descriptor for Log(1/Km), and its contribution might be explained in two possible ways. First, larger dimensions increase the possibility of interactions with the binding site, which is an effect purely related to size. In addition, the area can be an indicator of compound hydrophobicity, as large molecules are often more hydrophobic. Thus, in the QSARs for Log(1/Km), the presence of the area reconfirmed the hydrophobic nature of the binding sites of the enzymes. For FMO and CYP, the area featured in the QSARs for 1/Km, but the most important descriptors were related to electronic properties. In these cases, 1/Km may not be an indicator of binding, as it describes stronger interactions. The catalytic mechanism of FMO involves a nucleophilic attack, which takes place before binding (18). CYP enzymes have a catalytic mechanism with many steps occurring between substrate binding and oxygenation (40). It was shown that Km values may be sensitive to kinetic
perturbations at catalytic steps taking place after substrate binding; thus, 1/Km values may not be good approximations of affinity constants (41). The electronic descriptors related to protonation (apK_a and bpK_a) featured in many QSARs, especially the acidic dissociation constant, which had negative regression coefficients for Log(1/Km). This means that 1/Km is higher for more acidic compounds (i.e. with lower pK_a). The ionisation constant was a relevant factor also in QSARs for microbial biodegradation (42), due to the importance of protonation for enzyme–substrate interactions, as well as for penetration of the compound through the lipid bilayer. Electronic descriptors, such as HBD, HBA and dipole moment (ν), featured quite often, especially in the QSARs for LogVmax. This indicates that hydrogen bonding and polarity may play a significant role in the substrate–enzyme interactions.

In our study, we included frontier orbital parameters associated with metabolic properties: the energy of the lowest unoccupied and of the highest occupied molecular orbital, i.e. E_LUMO and E_HOMO, respectively, together with their difference (ΔE_L-H). E_LUMO and E_HOMO measure the ability of a molecule to accept and to donate an electron pair, respectively; thus, they describe the electrophilicity and the nucleophilicity of the substrate (43). The difference ΔEL–H is a stability index: the higher ΔEL–H, the higher the compound reactivity in chemical reactions. In fact, it is the relative difference between the nucleophile and electrophile orbitals that governs the reactivity of a given nucleophile–electrophile interaction (44). ELUMO appeared only in the QSAR of LogVmax for FMO, with a positive coefficient, as expected from its catalytic cycle. The substrates of FMO are nucleophiles, i.e. electron donors (18), and the higher the HOMO energy, the greater is the ability of the chemical to act as an electron donor. ELUMO and ΔEL–H featured in QSARs both for Log(1/Km), generally with a negative correlation coefficient, and for LogVmax with a positive correlation. This could be explained with the kinetics of the Michaelis–Menten reactions. Both Km and Vmax can be expressed in terms of kcat: Vmax is the product of kcat and total enzyme concentration, and Km is equal to the ratio (kcat + k-1)/kcat, where k-1 and k1 are constants, respectively, for breakdown and formation of the complex enzyme–substrate (4). The more reactive the molecule (i.e. the higher ΔEL–H), the higher is the catalytic rate (kcat), therefore the lower 1/Km (negative coefficient) and the higher Vmax (positive coefficient). The presence of ELUMO in the QSARs for LogVmax for ADH, ALDH, and CYP indicates that their substrates are likely electrophilic in nature, as it can be expected from their metabolic reactions. For ADH, a network of hydrogen bonding facilitates the deprotonation of the alcohol substrate bound to the active site of the enzyme (45). The ALDH catalytic mechanism involves a nucleophilic attack on the carbonyl group (C=O) of the aldehydes (46), which are reactive electrophilic compounds. At the CYP active site, the oxidation of chemicals is carried out by an electron-deficient complex (FeO3⁺), which abstracts either a hydrogen atom or an electron from the substrate (40). CYP enzymes would then behave as Lewis bases (nucleophiles) or Brønsted bases (H-acceptors). In fact, together with ELUMO, also pK_a and hydrogen bonding properties were important in the QSARs for LogVmax in CYP.

Conclusions

The QSARs developed in this study for Log(1/Km) and LogVmax of four important oxidising enzymes included physicochemical descriptors, which can be calculated and interpreted in a straightforward way. The processes underlying biotransformation were discussed from a mechanistic point of view, which may be useful in future research aimed at the prediction of the clearance of chemicals.

Supplementary Information

Supplementary Information Tables S1–S9 are available at http://www.eco-itn.eu. In addition, the data sets collected for this study can be accessed at http://ochem.eu.

Acknowledgements

Alessandra Pirovano was supported by the EU through the Environmental Cheminformatics (ECO) project (FP7-PEOPLE-ITN-2008, No. 238701). Karin Veltman was supported by the EU through a Marie-Curie Intra-European Fellowship (FP7-PEOPLE-2010-IEF, No. 273104). This research was partly funded by the EU through the TOXTRAIN project (implementation of a TOXicity assessment Tool for pRActical evaluation of life-cycle Impacts of techNologies; FP7-PEOPLE-2011-IAPP, No. 285286).

References


