

# In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted $k$ -Nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragment Analysis of Inhibitors versus Noninhibitors

Berith F. Jensen,<sup>†,‡</sup> Christian Vind,<sup>§</sup> Søren B. Padkjær,<sup>||</sup> Per B. Brockhoff,<sup>‡</sup> and Hanne H. F. Refsgaard<sup>\*,†</sup>

Exploratory ADME, Diabetes Research Unit, Novo Nordisk A/S, 2760 Måløv, Denmark, Scientific Computing, Novo Nordisk A/S, 2760 Måløv, Denmark, Protein Structure and Biophysics, Novo Nordisk A/S, 2760 Måløv, Denmark, and Informatics and Mathematical Modelling, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Received March 22, 2006

Inhibition of cytochrome P450 (CYP) enzymes is unwanted because of the risk of severe side effects due to drug–drug interactions. We present two in silico Gaussian kernel weighted  $k$ -nearest neighbor models based on extended connectivity fingerprints that classify CYP2D6 and CYP3A4 inhibition. Data used for modeling consisted of diverse sets of 1153 and 1382 drug candidates tested for CYP2D6 and CYP3A4 inhibition in human liver microsomes. For CYP2D6, 82% of the classified test set compounds were predicted to the correct class. For CYP3A4, 88% of the classified compounds were correctly classified. CYP2D6 and CYP3A4 inhibition were additionally classified for an external test set on 14 drugs, and multidimensional scaling plots showed that the drugs in the external test set were in the periphery of the training sets. Furthermore, fragment analyses were performed and structural fragments frequent in CYP2D6 and CYP3A4 inhibitors and noninhibitors are presented.

## Introduction

Metabolism determines the fate of a compound entering the body, ultimately controlling whether or not the compound exerts a toxic effect. Ideally, drugs and other xenobiotics are broken down to harmless soluble metabolites that are easily excreted through the urine or bile.<sup>1</sup> Cytochromes P450 (CYPs<sup>a</sup>) are the primary enzymes responsible for human drug metabolism. The isoenzyme CYP3A4 is the most abundant hepatic cytochrome P450 and is estimated to be involved in the metabolism of more than 50% of all marketed drugs.<sup>2</sup> CYP3A4 can also be inhibited by other xenobiotics, such as flavonoids, which are consumed in large quantities in the human diet.<sup>3</sup> CYP2D6 is a polymorphic member of the P450 super-family, which has been studied extensively, partially because 7–10% of the Caucasians and 1% of the Asians lack CYP2D6 activity and partially because it catalyzes the oxidation of many broadly prescribed pharmaceuticals, including antiarrhythmics, antidepressants, antipsychotics, beta-blockers, and analgesics.<sup>1,4</sup>

Because a huge variety of drugs are metabolized by CYP3A4 and CYP2D6, inhibition of these enzymes by one drug might lead to a decreased clearance of another drug when two or more drugs are administered simultaneously. Such unexpected drug–drug interactions can potentially have fatal consequences for the patient, and new chemical entities (NCEs) should be investigated for CYP inhibition as early as possible in drug research.<sup>5</sup>

In silico approaches for predicting the CYP inhibition potential of drugs are attractive as they may be applied to entire chemical libraries at the outset of the drug discovery process, usually at very small cost. In that way, the in silico models offer considerable potential for reducing the number of experimental studies required for compound selection and for improving the success rate. Furthermore, predictions can be made on virtual compounds.

Different in silico models for the classification of CYP2D6<sup>1,6–8</sup> and CYP3A4<sup>5–7,9–13</sup> inhibition have been published. The CYP2D6 models are based on 100 to 1810 diverse compounds, with the most predictive model being the one published by O'Brien and de Groot.<sup>8</sup> This model, built on 1810 compounds and validated with a test set consisting of 600 molecules, has an accuracy as high as 99% when 23% of the compounds are not classified. The published CYP3A4 models are based on 218 to 4000 diverse compounds, and the accuracies cover a field from 66–94%. In both CYP3A4 and CYP2D6 models there is generally a marked difference in the descriptors chosen and the applied statistical methods.

In this work, we report in silico models that classify CYP2D6 and CYP3A4 inhibition based on two diverse data sets of NCEs tested in 20  $\mu$ M concentrations for inhibition in human liver microsomes with dextromethorphan and erythromycin as substrate probes. The models are built using a novel Gaussian kernel weighted  $k$ -nearest neighbor ( $k$ -NN) algorithm based on Tanimoto similarity<sup>14</sup> searches on extended connectivity fingerprints (ECFP) and functional class fingerprints (FCFP)<sup>15</sup> from SciTegic. The classification method provides a probability of class memberships for each molecule when predicting new drug candidates. It was decided that the probability should be above 60% for classification because a classification result based on a probability of class membership below 60% would be unreliable and could not be used to decide whether or not a potential drug candidate should be eliminated.

\* To whom correspondence should be addressed. Tel.: + 45 44 43 03 67. Fax: + 45 44 66 39 39. E-mail: hare@novonordisk.com.

<sup>†</sup> Diabetes Research Unit, Novo Nordisk A/S.

<sup>‡</sup> Technical University of Denmark.

<sup>§</sup> Scientific Computing, Novo Nordisk A/S.

<sup>||</sup> Protein Structure and Biophysics, Novo Nordisk A/S.

<sup>a</sup> Abbreviations: CYP, cytochrome P450; ADME, absorption distribution metabolism and elimination; CYP2D6, cytochrome P450 2D6; CYP3A4, cytochrome P450 3A4; ECFP, extended connectivity fingerprints; FCFP, functional class fingerprints;  $k$ -NN,  $k$ -nearest neighbor; LOOCV, leave-one-out cross validation; MDS, multidimensional scaling; NCE, new chemical entities; CDF, cumulative distribution function.

**Table 1.** Training and Test Set Divided into Two Classes with Respect to CYP2D6 Inhibition in Human Liver Microsomes

	training set	test set
inhibition < 50%	552	183
inhibition ≥ 50%	313	105
sum	865	288

**Table 2.** Training and Test Set Divided into Two Classes with Respect to CYP3A4 Inhibition in Human Liver Microsomes

	training set	test set
inhibition < 50%	766	255
inhibition ≥ 50%	271	90
sum	1037	345

## Results

**Training Sets.** The training sets consisted of a total of 865 compounds, which were tested for inhibition of CYP2D6 in human liver microsomes with dextromethorphan as substrate, and a total of 1037 compounds tested for inhibition of CYP3A4 in human liver microsomes with erythromycin as substrate; 571 of the compounds were tested for both CYP2D6 and CYP3A4 inhibition. The distribution of compounds in the two classes is shown in Tables 1 and 2.

The compounds were from 20 Novo Nordisk discovery projects and had been tested over a 4–5 year period. The complete structures of the compounds cannot be disclosed at this time because they are in the discovery stage at Novo Nordisk. To address the structural diversity of the compounds in training sets, a cluster analysis using 166 MACCS structural keys and a threshold defined by a Tanimoto coefficient<sup>14</sup> of 0.85 was performed. The 865 compounds tested for CYP2D6 inhibition were split into 589 clusters, the five highest populated clusters contained 31, 15, 14, 11, and 11 members, respectively, the next 117 clusters contained 2–10 members, and 467 clusters were singletons. The 1037 compounds tested for CYP3A4 inhibition were split into 741 clusters, and the five highest populated clusters contained 26, 23, 12, 11, and 10 members, respectively. The next 141 clusters contained 2–7 members, and 595 clusters were singletons. The results suggest that the two training sets were diverse within the MACCS chemical space they represent.












**Structural Fragments.** All compounds in the two training sets were decomposed into ring fragments and functional groups as described in the method section. Structural groups for which we found statistical difference in frequency between the two classes with statistical difference beyond a 0.005 test level, are shown in Tables 3 and 4. The frequency of a fragment in a CYP inhibition class was calculated as

$$\text{frequency of a fragment} = \frac{(N_{\text{fragment\_class}} \times N_{\text{total}})}{(N_{\text{fragment\_total}} \times N_{\text{class}})} \quad (1)$$

where  $N_{\text{fragment\_class}}$  is the number of compounds containing the fragment in a CYP inhibition class,  $N_{\text{total}}$  is the total number of compounds,  $N_{\text{fragment\_total}}$  is the total number of compounds containing the fragment, and  $N_{\text{class}}$  is the number of compounds in the CYP inhibition class. The data sets were unbalanced between inhibitors and noninhibitors, but as  $N_{\text{class}}$  was included in eq 1, the effect of the biased data set was eliminated. The frequencies of fragments in a CYP inhibition class were statistically compared by standard binomial techniques.














For CYP2D6 structural groups, including benzene, acetamides, carboxylic acids, phenol fragments, sulfone, and carbonic

**Table 3.** Occurrence and Frequency of Ring Fragments and Functional Groups in the CYP2D6 Training Set; 571 of the 865 Compounds Were Also Represented in the CYP3A4 Training Set

Structure group	AutoNom Name	Non-inhibitors	Inhibitors	F(1)	F(2)
	Benzene	505	215	1.10	0.83 <sup>a</sup>
	Acetamide	260	117	1.08	0.86 <sup>b</sup>
	Propan-2-one	102	36	1.16	0.72 <sup>b</sup>
	Pyridine	64	20	1.19	0.66 <sup>b</sup>
	Acetic acid	55	10	1.33	0.43 <sup>a</sup>
	Propen-2-ol	43	5	1.40	0.29 <sup>a</sup>
	Methanesulfonylmethane	37	2	1.49	0.14 <sup>a</sup>
	Carbamic acidisopropyl ester	42	11	1.24	0.57 <sup>b</sup>
	N-Ethyl-acetamide	25	3	1.40	0.30 <sup>a</sup>
	Diethyl-isopropyl-amine	54	75	0.66	1.61 <sup>a</sup>
	3,5-Dimethyl-[1,2,4]oxadiazole	6	21	0.35	2.15 <sup>a</sup>

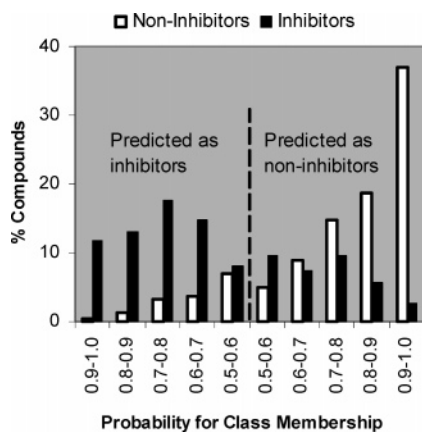
<sup>a</sup> Significant difference on a 0.001 test level from the frequency of the noninhibitors. <sup>b</sup> Significant difference on a 0.005 test level from the frequency of the noninhibitors.

**Table 4.** Occurrence and Frequency of Ring Fragments and Functional Groups in the CYP3A4 Training Set; 571 of the 1037 Compounds Were Also Represented in the CYP2D6 Training Set

Structure group	AutoNom Name	Non-inhibitors	Inhibitors	F(1)	F(2)
	Diethyl-isopropyl-amine	105	10	1.24	0.33 <sup>a</sup>
	Triethyl-amine	96	15	1.17	0.52 <sup>a</sup>
	Pyrrolidine	87	5	1.28	0.21 <sup>a</sup>
	Acetic acid	79	2	1.32	0.09 <sup>a</sup>
	Acetyl-urea	34	2	1.28	0.21 <sup>a</sup>
	Benzene	423	125	1.04	0.87 <sup>b</sup>
	Benzene	28	2	1.26	0.26 <sup>a</sup>
	1-Cyclopentyl-piperazine	22	0	1.35	0.00 <sup>a</sup>
	Acetamide	345	150	0.94	1.16 <sup>b</sup>
	Propen-2-ol	22	30	0.57	2.21 <sup>a</sup>
	Methanesulfonylmethane	21	28	0.58	2.19 <sup>a</sup>
	Eth-(E)-ylidene-vinyl-amine	0	17	0.00	3.83 <sup>a</sup>
	Benzene	67	71	0.66	1.97 <sup>a</sup>

<sup>a</sup> Significant difference on a 0.001 test level from the frequency of the noninhibitors. <sup>b</sup> Significant difference on a 0.005 test level from the frequency of the noninhibitors.

acid esters, were more frequent in noninhibitors than in inhibitors, whereas tertiary amines and oxadiazole fragments were more frequent in inhibitors than in noninhibitors (Table 3). For CYP3A4 structural groups, such as tertiary amines, carboxylic acids, pyridine fragments, pyrrolidine, cyclopentyl-



**Figure 1.** Distribution of the predictions with respect to probability of class membership when cross validating the CYP2D6 inhibition model based on FCFP and 865 drug candidates.

**Table 5.** Result of Cross Validation of the CYP2D6 Gaussian Kernel Weighted  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	180 (+24)	79 (+30)	54	313
measured CYP2D6 noninhibitor	48 (+38)	438 (+28)	66	552
sum	228 (+62)	517 (+58)	120	865

<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

piperazine, and one- or three-benzene groups, were frequent in noninhibitors, whereas the chemical structures, acetamides, phenol fragments, and sulfone and pyridine fragments, were found to be frequent in CYP3A4 inhibitors (Table 4).

**Construction of CYP2D6 Model.** The optimal parameters of the CYP2D6 kernel weighted  $k$ -NN model were chosen using leave-one-out cross validation (LOOCV) on the 865 compounds in the training set, as described in Materials and Methods. It was decided to use FCFP\_6 as descriptors, 20 neighbors, a dynamic smoothing factor set to 0.5, and an uncertainty term of 0.2. The distribution of predictions with respect to probability of class membership is depicted in Figure 1, while the classification result is shown in Table 5. Figure 1 shows that the noninhibitors in general were predicted with higher certainty than the inhibitors. A higher percentage of the noninhibitors was predicted with a probability of class membership of 0.70 or more, and fewer compounds were falsely predicted as inhibitors. Table 5 shows that 120 of the 865 compounds in the training set, corresponding to 14%, were not classified. As mentioned above, the prediction results were unbalanced between inhibitors and noninhibitors; 69% of the classified inhibitors and 90% of the classified noninhibitors were classified correctly. Overall, 83% of the classified compounds were predicted to the right class.

A permutation test<sup>16</sup> was made to secure that the predictions were based on the variance described by the fingerprints. The  $y$ -values were randomized four times, the four new models gave a mean prediction accuracy on 55% (standard deviation = 0.02), and 23% (standard deviation = 0.02) of the compounds were not classified due to the 60% threshold. The result shows that the predictions made by the original model were not based on an accidental correlation.

To investigate the power of the new Gaussian weighted  $k$ -NN method, traditionally binary  $k$ -NN was applied to the same data set. Table 6 shows that, when using traditionally binary  $k$ -NN,

**Table 6.** Result of Cross Validation of the Traditional CYP2D6 Binary  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	137 (+35)	89 (+52)	87	313
measured CYP2D6 noninhibitor	45 (+32)	430 (+45)	77	552
sum	182 (+67)	519 (+97)	164	865

<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

**Table 7.** Result of Test Set Validation of the CYP2D6 Gaussian Kernel Weighted  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	53 (+10)	37 (+5)	15	105
measured CYP2D6 noninhibitor	11 (+9)	159 (+4)	13	183
sum	64 (+19)	196 (+9)	28	288

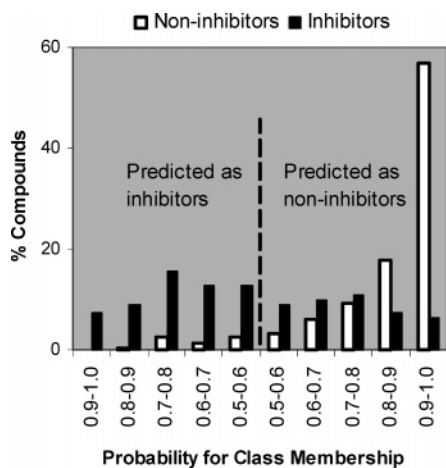
<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

164 of the 865 compounds in the training set corresponding to 19% were not classified and 81% of the classified compounds were predicted to the right class. The results indicated that the new Gaussian kernel weighted  $k$ -NN method applied to the data reported here was more predictive than the traditional binary  $k$ -NN method.

**Validation of CYP2D6 Model.** The model was validated on a test set consisting of 288 drug candidates. The average Tanimoto similarity to the training set of all 288 compounds and 20 nearest neighbors was 0.53. The result is depicted in Table 7, which shows that 28 of the 288 compounds, corresponding to 10%, were not classified. Of the classified compounds, 82% were predicted to the right class. In accordance with the LOOCV result, the distribution of correctly predicted inhibitors and noninhibitors was biased; in the validation result, 59% of the classified inhibitors and 94% of the classified noninhibitors were correctly classified.

**Construction of CYP3A4 Model.** The optimal parameters of the CYP3A4 kernel weighted  $k$ -NN model were chosen using LOOCV on the 1037 compounds in the training set. Based on the cross validation, it was decided to use ECFP\_6 as descriptors, 20 neighbors, a dynamic smoothing factor set to 0.4, and an uncertainty term of 0.2. The distribution of predictions with respect to probability of class membership is depicted in Figure 2, while the classification result is shown in Table 8. A high percentage of the noninhibitors was predicted with a probability of class membership on 0.80 or more, and few compounds were falsely predicted as inhibitors. Table 8 shows that 103 of the 1037 compounds in the training set, corresponding to 10%, were not classified. Again, the prediction results were unbalanced between inhibitors and noninhibitors; 75% of the classified inhibitors and 96% of the classified noninhibitors were classified correctly, and overall, 87% of the classified compounds were predicted to the right class.

A permutation test was made to make sure that the predictions were based on the variance described by the fingerprints. The four new models gave a mean prediction accuracy on 68% (standard deviation = 0.01), and 15% (standard deviation = 0.02) of the compounds were not classified due to the 60% threshold. The result shows that the predictions made by the original model were not based on an accidental correlation.



**Figure 2.** Distribution of the predictions with respect to probability of class membership when cross validating the CYP3A4 inhibition model based on ECFP and 1037 drug candidates.

**Table 8.** Result of Cross Validation of the CYP3A4 Gaussian Kernel Weighted  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	120 (+34)	93 (+24)	58	271
measured CYP2D6 noninhibitor	32 (+20)	689 (+25)	45	766
sum	152 (+54)	782 (+49)	103	1037

<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

**Table 9.** Result of Cross Validation of the Traditional CYP3A4 Binary  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	87 (+26)	124 (+34)	60	271
measured CYP2D6 noninhibitor	23 (+16)	681 (+46)	62	766
sum	110 (+42)	805 (+80)	122	1037

<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

Traditionally, binary  $k$ -NN was also applied to the same CYP3A4 data set. The result is shown in Table 9. The table shows that when using traditionally binary  $k$ -NN, 122 of the 1037 compounds in the training set were not classified, corresponding to 12%. Of the classified compounds, 84% were predicted to the right class. The results indicated that the new Gaussian kernel weighted  $k$ -NN method applied to the data reported here was more predictive than the traditional binary  $k$ -NN method.

**Validation of CYP3A4 Model.** The model was validated with a test set consisting of 345 drug candidates. The average Tanimoto similarity to the training set of all 345 compounds and 20 nearest neighbors was 0.48. The results are listed in Table 10, which shows that 47 of the 345 compounds, corresponding to 14%, were not classified. Of the classified compounds, 88% were predicted to the right class. In accordance with the LOOCV result, the distribution of correctly predicted inhibitors and noninhibitors was unbalanced; in the validation result, 65% of the classified inhibitors and 94% of the classified noninhibitors were correctly classified.

**Assessment Using an External Test Set.** For further assessment of the two models, an external test set of 14

**Table 10.** Result of Test Set Validation of the CYP3A4 Gaussian Kernel Weighted  $k$ -NN Model<sup>a</sup>

	pred. inhibitor	pred. noninhibitor	nonclassified	sum
measured CYP2D6 inhibitor	44 (+11)	24 (+11)	22	90
measured CYP2D6 noninhibitor	13 (+13)	217 (+12)	25	255
sum	57 (+24)	241 (+23)	47	345

<sup>a</sup> The numbers in brackets illustrate how the nonclassified compounds would be classified.

**Table 11.** Prediction of Percent Inhibition of Compounds in an External Test Set Consisting of Known Inhibitors of Either CYP2D6 or CYP3A4<sup>a</sup>

reference drug	known inhibitor of	predicted CYP2D6 inhibition <sup>b</sup>	predicted CYP3A4 inhibition <sup>b</sup>
ajmalicine <sup>17</sup>	CYP2D6	52% (inh)	31% (noninh)
azelastine <sup>18</sup>	CYP2D6	45% (noninh)	27% (noninh)
fluoxetine <sup>19</sup>	CYP2D6	27% (noninh)	20% (noninh)
paroxetine <sup>19</sup>	CYP2D6	31% (noninh)	48% (noninh)
perphenazine <sup>19</sup>	CYP2D6	54% (inh)	28% (noninh)
quinidine <sup>19</sup>	CYP2D6	79% (inh)	13% (noninh)
sertraline <sup>19</sup>	CYP2D6	62% (inh)	39% (noninh)
clotrimazole <sup>10,20,21</sup>	CYP3A4	28% (noninh)	56% (inh)
cyclosporin A <sup>21</sup>	CYP3A4	26% (noninh)	37% (noninh)
indinavir <sup>19</sup>	CYP3A4	53% (inh)	55% (inh)
itraconazole <sup>19-20</sup>	CYP3A4	44% (noninh)	35% (noninh)
ketoconazole <sup>10,19-21</sup>	CYP3A4	40% (noninh)	44% (noninh)
miconazole <sup>10,20,21</sup>	CYP3A4	47% (noninh)	39% (noninh)
troleandomycin <sup>21</sup>	CYP3A4	40% (noninh)	38% (noninh)

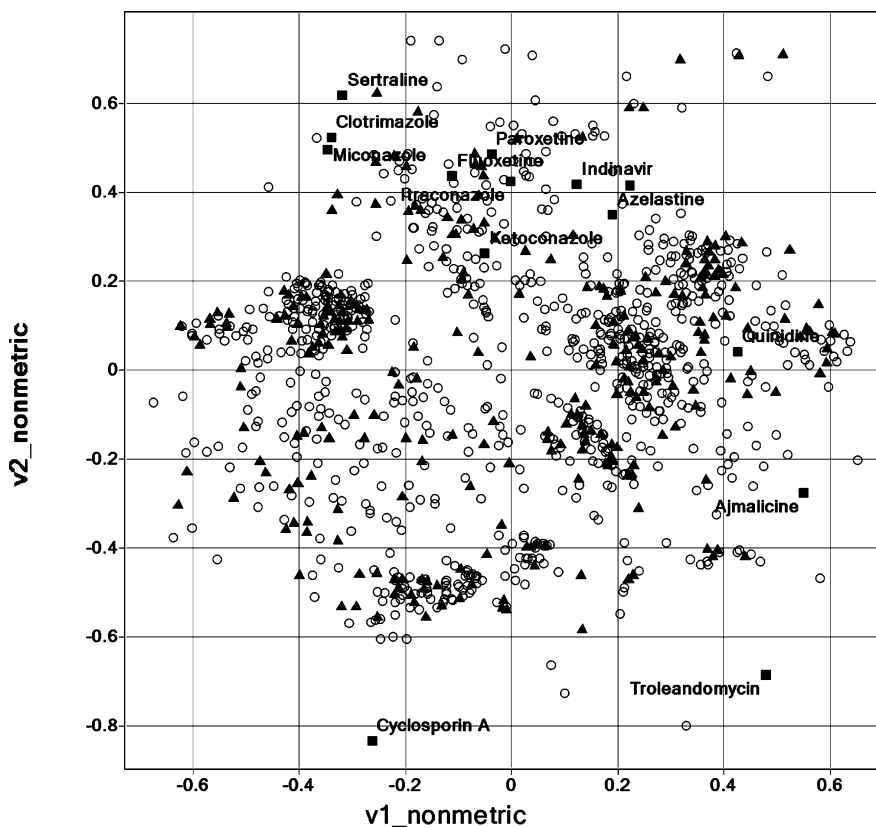
<sup>a</sup> The average Tanimoto coefficients to the 20 nearest neighbors in the training set were 0.20 and 0.24 for the CYP2D6 and CYP3A4 models, respectively. <sup>b</sup> inh = inhibitor; noninh = noninhibitor.

published inhibitors of either CYP2D6 or CYP3A4 were classified by use of both classification models. The prediction results of the external test set are shown in Table 11.

As a result of the decision that compounds with predicted percent inhibition between 40 and 60% could not be classified, the CYP2D6 model was only able to classify 8 of the 14 compounds in the external test set, while 10 compounds were classified by the CYP3A4 model. The relatively low number of predicted compounds can be explained by the low average Tanimoto similarity between the 14 compounds and the 20 nearest neighbors in the CYP2D6 and CYP3A4 training set, which were 0.20 and 0.14, respectively. The low Tanimoto coefficients for the external test set compared to the two internal test sets indicated that the external test set differs more from the two training sets than the internal test sets.

The CYP2D6 model predicted two of the remaining four CYP2D6 classified inhibitors and predicted all noninhibitors to the right class. The CYP3A4 model was not able to predict any of the CYP3A4 inhibitors with a probability above 60% but predicted all noninhibitors to the right class as well.

**Illustration of Data by MDS Plots.** To illustrate how data were distributed, multidimensional scaling<sup>22</sup> (MDS) plots, based on a matrix containing the Tanimoto distances between all compounds, were made. The MDS plots show where in the chemical space the internal and external test sets were compared to the CYP2D6 and CYP3A4 training sets (Figures 3 and 4). The two plots stress that the external data set was in the periphery of the two training sets, while the internal test sets were more or less covered by the chemical space of the training sets. In Figures 5 and 6, the compounds that were falsely predicted by the two models are plotted.



**Figure 3.** Multidimensional scaling (MDS) plot of CYP2D6 data: circles = training set (865 objects), triangles = internal test set (288 objects), and squares = external test set (14 objects).

Figure 5 shows that the compounds from the external test set that were falsely predicted by the CYP2D6 model were grouped in a sparse area of the training set in the top of the plot, while the falsely predicted internal test drugs were distributed evenly throughout the plot. In Figure 6, the compounds from the external test set that were falsely predicted by the CYP3A4 were grouped in two clusters.

Figures 7 and 8 show how the nonclassified test compounds were distributed in the chemical space that the training set represented. Most of the compounds from the external test set that could not be predicted by the CYP2D6 model were found in sparse areas of the training set throughout the plot (Figure 7), while the nonclassified internal test drugs were distributed evenly throughout the plot. However, the compounds from the external test set that could not be predicted by the CYP3A4 model were more widely distributed.

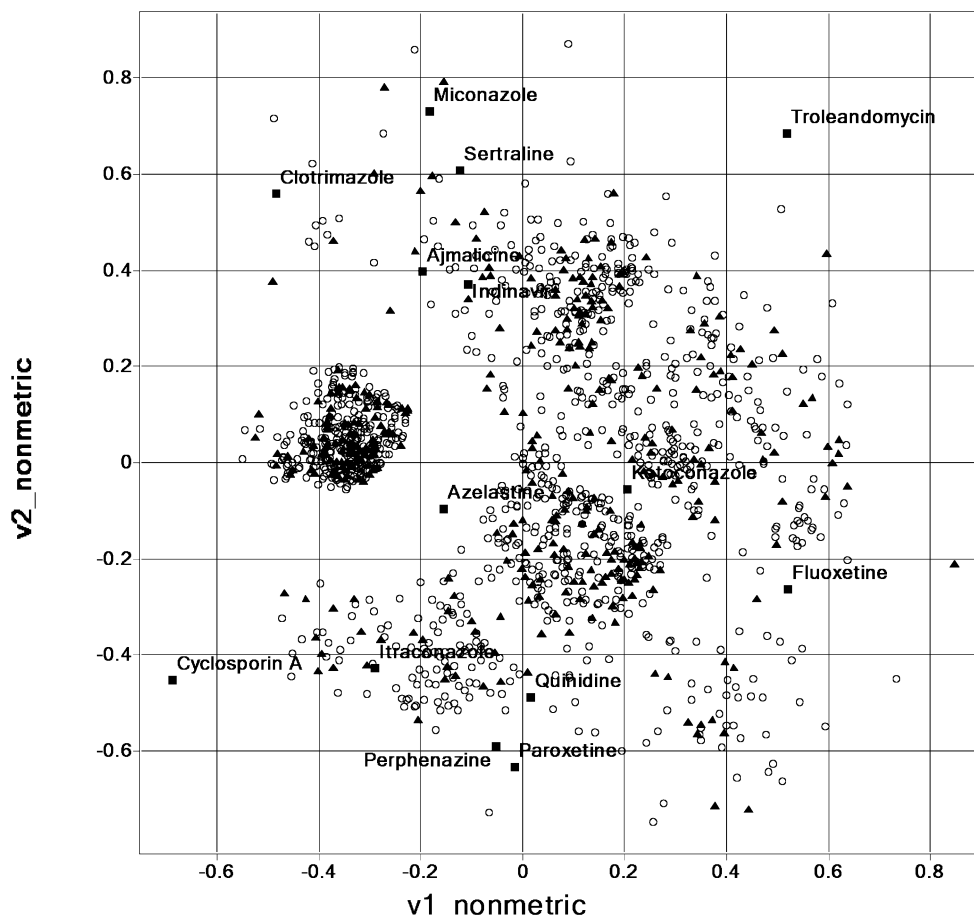
## Discussion

In this work, *in silico* models that classify CYP2D6 and CYP3A4 inhibition based on two diverse data sets are presented. The compounds in both data sets were tested by a clustering analysis and found to be diverse in the chemical space they represent. It is unlikely that a model can cover all of the chemistry space if it is built using only compounds from Novo Nordisk projects. Thus, when applying the models on new drug candidates one has to make sure that the compounds in the training set are representative of these compounds. This can be done by evaluating the average similarity between the test compound and the *k*-NNs in the training set. If the similarity is low the Gaussian kernel weighted *k*-NN model used in this study will most probably give in-doubt classifications. This is expected to be the case, as test compounds that end up in sparsely populated parts of the chemical space, with high probability,

will have *k* equally weighted nearest neighbors with arbitrary distributed CYP inhibition rates. The problem occurred in this study when the two *in silico* models were assessed by 14 published inhibitors of either CYP2D6 or CYP3A4. The average Tanimoto similarity between the 14 compounds and the 20 nearest neighbors in the CYP2D6 and 3A4 training set were 0.20 and 0.14, respectively. The 14 compounds were poorly predicted by the models, and based on the low similarity between test compounds and training set, one would not expect reliable predictions.

All compounds in the two training sets were decomposed into ring fragments and functional groups. To our knowledge, it is the first time frequent structural fragments of noninhibitors and inhibitors of CYP2D6 and CYP3A4 are presented. Interestingly, carboxyl acid fragments were more frequent in noninhibitors than inhibitors of both CYP2D6 and CYP3A4. As one would expect, it was not the same structural fragments that were the most frequent in the inhibitors of the two isoenzymes. A tertiary amine fragment was found frequently in CYP2D6 inhibitors and in CYP3A4 noninhibitors, whereas a phenol fragment and sulfone were found frequently in CYP2D6 noninhibitors and CYP3A4 inhibitors (Tables 3 and 4). It is generally accepted that substrates for CYP2D6 are basic and substrates for CYP3A4 are neutral or basic.<sup>23,24</sup> This corresponds well together with the finding that the carboxylic acid fragment was frequent in noninhibitors of the two isoenzymes. It is described that CYP2D6 substrates contain at least one basic nitrogen,<sup>23</sup> which could be part of the explanation for why a tertiary amine fragment was found frequently in CYP2D6 inhibitors (Table 3).

Physical-chemical descriptors have successfully been included in *in silico* models for CYP2D6<sup>6,7,25</sup> and CYP3A4<sup>6-7,9,25</sup> inhibition, but only few studies with applications of fingerprints



**Figure 4.** MDS plot of CYP3A4 data: circles = training set (1037 objects), triangles = internal test set (345 objects), and squares = external test set (14 objects).

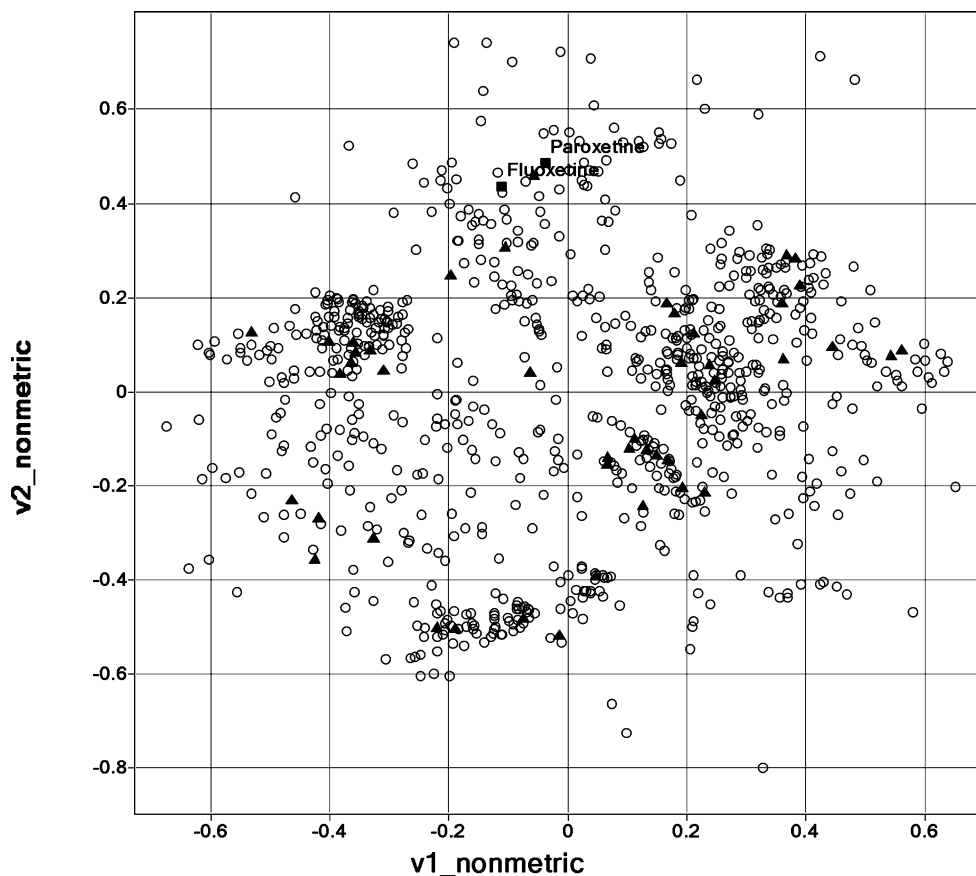
are published.<sup>8,11,12</sup> In the present study, we used ECFP from SciTegic for modeling and found that FCFP\_6 gave the most predictive CYP2D6 model, while ECPF\_6 were more suitable for classification of CYP3A4 inhibition. In FCFP, atoms are characterized by functional types. For instance, in FCFP, all halogens give the same atom bit code, whereas in ECFP, the atoms are characterized by chemical elements and different halogens have different atom bit codes.<sup>15</sup> FCFP\_6 contain fewer features than ECFP\_6, and there is no structural explanation to why FCFP\_6 gave the most predictive CYP2D6 model, while ECPF\_6 was found more suitable for classification of CYP3A4 inhibition. It should be mentioned that there was not much difference in the model performance between the two kinds of fingerprints. In a study of the inhibition of *Escherichia coli* dihydrofolate reductase, Rogers et al.<sup>15</sup> investigated differences between FCFP\_6 and ECFP\_6. They find better results when the model building is based on FCFP\_6 than ECFP\_6. They suggest that it is likely that the extra abstraction obtained by FCFP\_6 is not advantageous for predicting samples similar to the training data but become valuable when extrapolating to molecules that are quite different from the training data. The cluster analysis of the two training sets used in the present study showed that there were more clusters compared to compounds in the CYP2D6 training set, indicating that this data set was more diverse than the CYP3A4 training set. This could be a part of the explanation to why different kinds of fingerprints were found to be optimal in the two *in silico* models.

The models presented in this work were built using a Gaussian kernel weighted *k*-NN algorithm based on a Tanimoto similarity search on ECFP. *k*-NN statistics and other nonlinear methods have previously been found very useful in QSAR modeling.<sup>15,26–31</sup>

The similar property principle was first presented explicitly by Johnson and Maggiora<sup>32</sup> and states that molecules that are structurally similar are likely to have similar properties. However, there are many exceptions to the principle,<sup>33</sup> as even minor structural variations can have a drastic effect on the levels of activity in a set of analogues. Nevertheless, the principle is in general applicable and there is substantial evidence to support its use in lead-discovery programs.<sup>34–38</sup>

The classification method provides a probability of class memberships for each molecule when predicting new drug candidates. In the present study, it was decided that the probability should be above 60% for a drug to be classified. The consequence of this was that 10–14% of the compounds in the internal test set were to be not classified. It is not clear why these compounds could not be classified; it was compounds from different projects, some found in the periphery of the training set and some in the center (see Figures 7 and 8). The average similarities between the nonclassified compounds and the 20 nearest neighbors in the training set were 0.50 and 0.48 for CYP2D6 and CYP3A4, respectively. This was almost the same as the average similarities between all test compounds and the 20 nearest neighbors in the training set, which were 0.53 and 0.48 for CYP2D6 and CYP3A4, respectively.

Overall, the models were predictive when validated with internal test sets; here 82% for CYP2D6 and 88% for CYP3A4 of the predicted compounds in the internal test set were classified correctly. However, the results clearly showed that in both models noninhibitors were classified with higher certainty than inhibitors. Assessed by internal test sets, only 59% and 65% of the classified inhibitors were predicted correctly by the CYP2D6 and CYP3A4 models, respectively, while the same number for



**Figure 5.** MDS plot of falsely predicted compounds in the CYP2D6 model: circles = training set (865 objects), triangles = false predicted internal test compounds, and squares = false predicted external test compounds.

noninhibitors was 94% in both models. One could think that the different classification accuracies were due to the fact that there were 2–3 times as many noninhibitors compared to inhibitors in both training sets. To test this hypothesis, a training set of 500 compounds with equal numbers of inhibitors and noninhibitors was randomly selected from the original CYP2D6 training set of 865 compounds. Subsequently, the original internal CYP2D6 test set was classified by the new model. The distribution of correct classified inhibitors and noninhibitors did not change, showing that the biased training sets used in this study were not the cause of the unbalanced classification results between inhibitors and noninhibitors.

One can discuss the value of a model with very biased output. However, despite the unbalanced classification results obtained in the present study, it is reasonable to use the developed models when predicting CYP inhibition potential of new drug candidates, as the models only have few false positives. In both models, only 4% of the classified compounds were false positives, meaning that few potential drugs will be eliminated before synthesis.

## Conclusion

In this study, Gaussian kernel weighted  $k$ -NN models to predict CYP2D6 and CYP3A4 inhibition were constructed to predict CYP2D6 and CYP3A4 inhibition. Assessment using both cross validation and an internal test set resulted in the CYP2D6 classification model predicting 82–83% of the classified compounds correctly, while 10–14% of the compounds were not classified. Similarly, the CYP3A4 classification model was assessed with cross validation and an internal test set and predicted 87–88% of the classified compounds correctly, while 10–14% of the compounds were not classified.

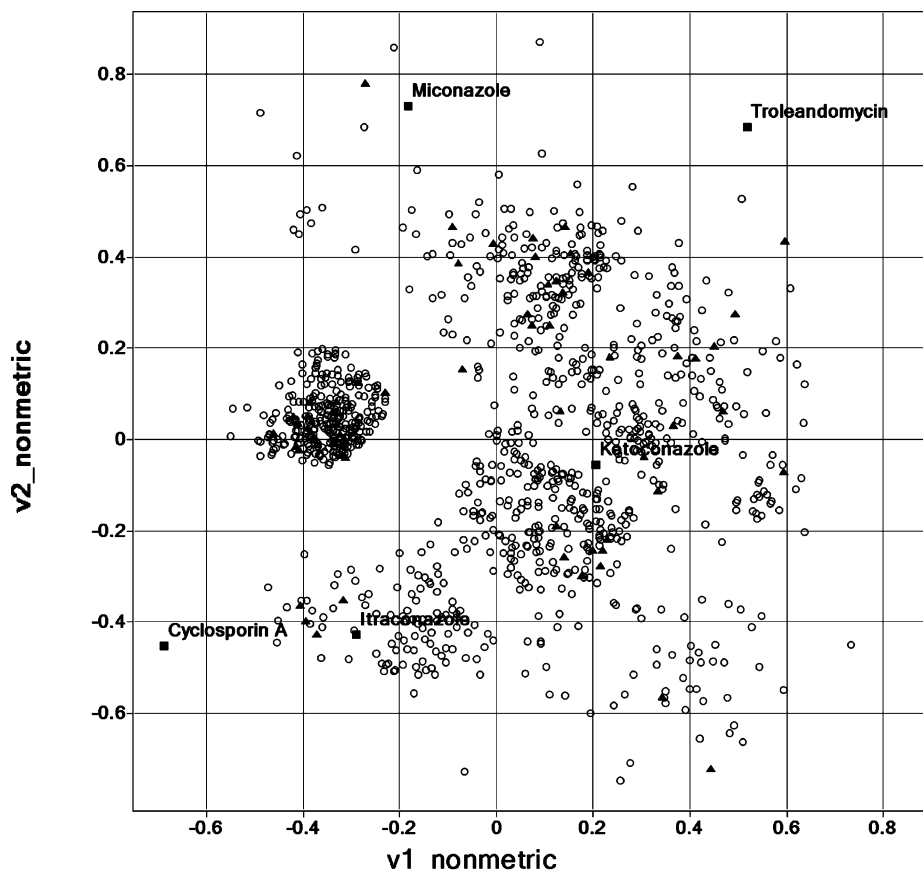
It was found that the in silico models based on ECFP could be used to select CYP2D6 and CYP3A4 noninhibitors in an early stage of discovery projects.

## Materials and Methods

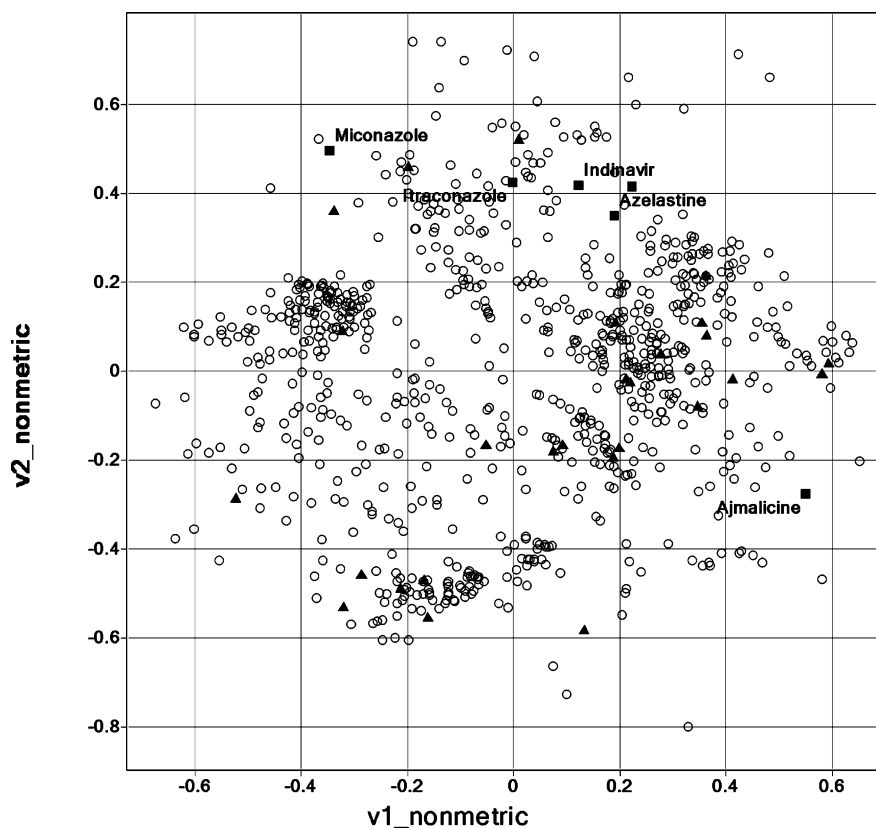
**Materials.** For both CYP inhibition screening assays, mixed pools of human liver microsomes were obtained from Gentest BD Biosciences (Woburn, MA). Dextromethorphan[ $O$ -methyl- $^{14}C$ ] was purchased from ARC (St. Louis, MO). Quinidine sulfate dehydrate, ketoconazole, erythromycin, troleandomycin, nicotinamide adenine dinucleotide phosphate (NADPH), charcoal, and trichloroacetic acid solution were purchased from Sigma-Aldrich (St. Louis, MO). Erythromycin[ $N$ -methyl- $^{14}C$ ] and Ultima Gold were from Perkin-Elmer, (Boston, MA).

**Inhibition Assay.** The in vitro CYP-subtype activity in the absence and presence of a test compound (the potential CYP-subtype inhibitor) was determined using a CYP-subtype selective substrate, dextromethorphan for CYP2D6 and erythromycin for CYP3A4. Incubations were utilized (200  $\mu$ L, 37  $^{\circ}$ C, 10 min) containing human liver microsomes (HLM, 0.1 mg), [ $O$ -methyl- $^{14}C$ ]-dextromethorphan in the CYP2D6 assay (total concn: 3  $\mu$ M = Km) or [ $N$ -methyl- $^{14}C$ ] erythromycin and erythromycin in the CYP3A4 assay (total concn: 20  $\mu$ M = Km), the cofactor NADPH (1 mM), and the test compound (20  $\mu$ M). All incubations were performed in triplicate. The HLM preparations used consisted of a pool of  $\geq 15$  donors to obtain an average concentration of CYP subtypes, which are known to differ markedly between individuals.

The metabolic conversion of [ $O$ -methyl- $^{14}C$ ]-dextromethorphan or [ $N$ -methyl- $^{14}C$ ] erythromycin was assessed by activated charcoal extraction, followed by liquid scintillation counting of the super-



**Figure 6.** MDS plot of falsely predicted compounds in the CYP3A4 model: circles = training set (1037 objects), triangles = false predicted internal test compounds, and squares = false predicted external test compounds.

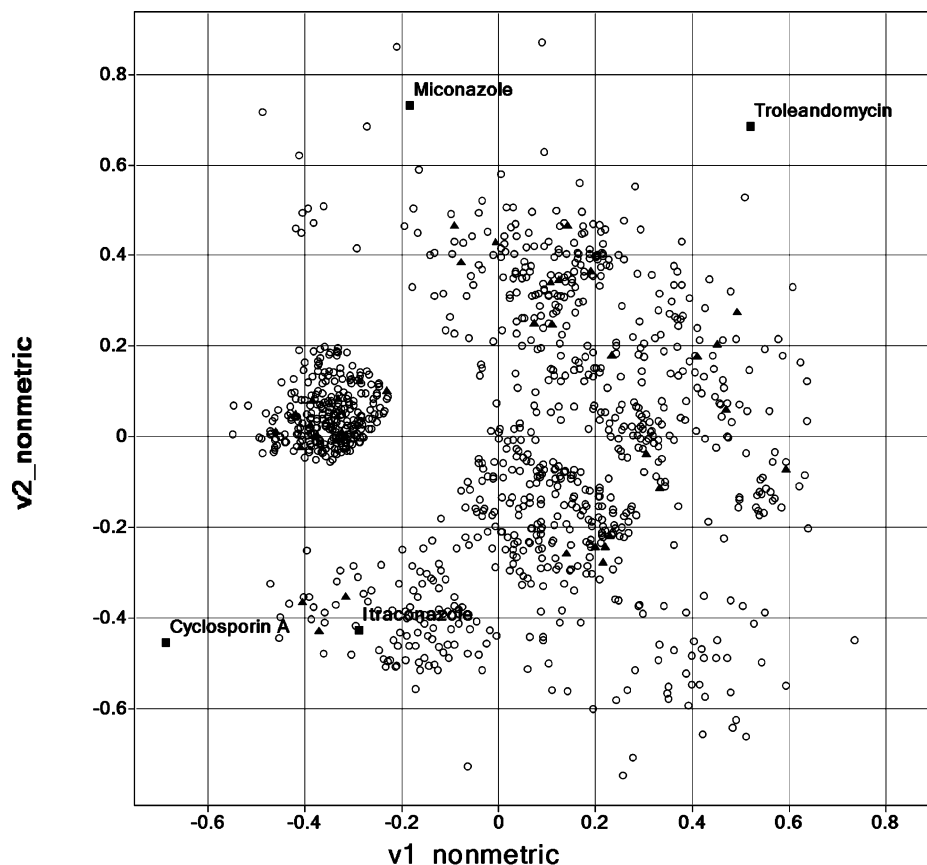


**Figure 7.** MDS plot of nonclassified compounds in the CYP2D6 model: circles = training set (865 objects), triangles = non-classified internal test compounds, and squares = nonclassified external test compounds.

nant. The thereby measured  $^{14}\text{C}$ -formaldehyde reflected the metabolism of the selective substrate by CYP subtype.

The specificity of the assays was validated by incubation performed in the absence and presence of known inhibitors:





**Figure 8.** MDS plot of nonclassified compounds in the CYP3A4 model: circles = training set (1037 objects), triangles = nonclassified internal test compounds, and squares = nonclassified external test compounds.

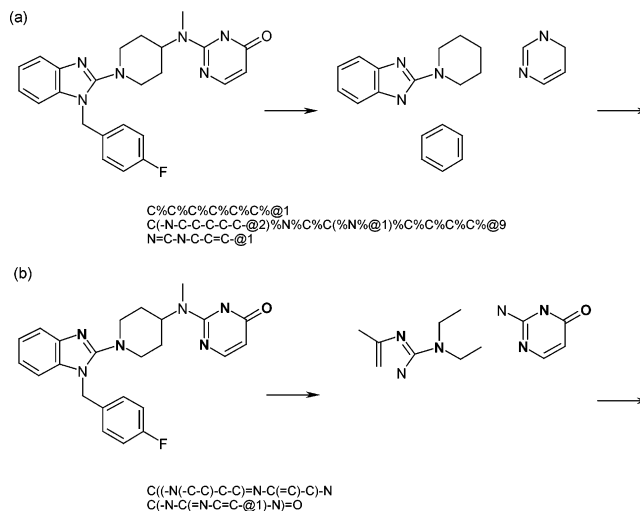
quinidine for CYP2D6 and keconazole and troleandomycin for CYP3A4. For test compounds where >100% inhibition was determined, % inhibition was set to 100%.

**Training Set and Test Set.** The CYP2D6 data set consisted of 1153 drug candidates and the CYP3A4 data set of 1382 compounds, both from 20 Novo Nordisk discovery projects. Before the model constructions were performed, the data sets were sorted by project and then sorted with respect to percent CYP inhibition. Subsequently, data were divided into a training set and a test set. Each data set was divided so that every fifth compound was selected for the test set and the rest of the compounds for the training set. The selection method was chosen to obtain a training set where all Novo Nordisk projects and levels of inhibition were represented.

**Cluster Analysis.** To address the structural diversity of the compounds in the training set, a cluster analysis using the Jarvis–Patrick clustering method,<sup>39</sup> 166 MACCS structural keys,<sup>40</sup> and a threshold defined by a Tanimoto coefficient<sup>14</sup> of 0.85 was performed using the fingerprint database clustering procedure in MOE.<sup>41</sup>

**Structural Fragments.** Two decomposition schemes were developed to define molecular equivalence classes.<sup>42</sup> The first class converted a H-depleted molecular graph into a list of ring fragments by removing all atoms and associated bonds that were not in a ring. The second class used standard definitions for donors and acceptors<sup>43</sup> and included their beta neighborhood into the associated fragment. Overlapping fragments were considered as one union fragment. For both decomposition schemes, each H-depleted molecular graph was transformed into fragments and represented in a canonical line notation similar to SLN,<sup>44</sup> as illustrated in Figure 9 for an antihistamine, mizolastine. All algorithms were implemented as Cheshire scripts.<sup>45</sup> It was subsequently feasible to apply standard pivoting techniques in a spreadsheet on records associating compound id, fragment string, and CYP inhibition class.

**Descriptors.** The in silico Gaussian kernel weighted *k*-NN algorithm is based on Tanimoto similarity searches using either SciTegic's ECFP or FCFP.<sup>15</sup> In ECFP and FCFP, the initial code



**Figure 9.** Decomposition of Mizolastine with respect to the ring scheme (a) and the donor/acceptor scheme (b).

assigned to an atom is based on the number of connections, the element type, the charge, and the mass for ECFPs and on six generalized atom types (hydrogen-bond donor, hydrogen-bond acceptor, positively ionizable, negatively ionizable, aromatic, and halogen) for FCFPs. This code, in combination with the bond information and with the codes of its immediate neighbor atoms, was hashed to produce the next order code, which was mapped into an address space of size 232, and the process was iterated until the required level of description had been achieved.<sup>46</sup> SciTegic's ECFP is a fairly novel type of fingerprint, which uses an adapted version of the Morgan algorithm to generate a feature vector that indicates the presence of structural features among a vast set of potential features. Please consult Klon et al.<sup>47</sup> or Rogers et al.<sup>15</sup> for further details. In this study, ECFP\_2, ECFP\_4, ECFP\_6, FCFP\_2,

FCFP\_4, and FCFP\_6 fingerprints were tested. The numeric code denotes the diameter in bonds up to which features were generated.

**Statistical Method.** A novel Gaussian kernel weighted  $k$ -NN regression and classification algorithm has been used in the present study. The algorithm was implemented using the software package Pipeline Pilot from SciTegic<sup>48</sup> and the open-source statistics package R<sup>49</sup>. The algorithm augments the standard  $k$ -NN algorithm by taking the Tanimoto similarity to the nearest neighbors in ECFP space into account and was inspired by work by Shepard,<sup>50</sup> Lowe,<sup>51</sup> Atkeson et al.,<sup>52</sup> and Harper et al.<sup>27</sup> The algorithm, when used for classification, was split into the following two steps: (1) Use Gaussian kernel weighted  $k$ -NN regression to estimate the regression value and to estimate the regression estimate's associated uncertainty for the test molecule. (2) Calculate the Gaussian cumulative distribution function (CDF) at 50% inhibition, with mean and standard deviation given by the regression value and uncertainty, respectively, estimated in step 1.

The regression step, step 1, was split into the following steps: (1) For the test molecule, find the  $k$  nearest neighbors, their measured inhibition  $y_i$ , and their corresponding similarities  $s_i$  measured as Tanimoto similarity in the ECFP space spanned by the training set. (2) Assign weights to each of the  $k$  nearest neighbors based on their similarities, using a Gaussian kernel to map the similarity to a weight

$$w_i = K(s_i) = [(2\Pi)^{-1/2}/\sigma] \times \exp[-(1 - s_i)^2/(2 \times \sigma^2)] \quad (2)$$

The kernel bandwidth  $\sigma$  is chosen dynamically for the test molecule based on the nearest neighbor only

$$\sigma = (1 - s_j) \times [\text{dynamic smoothing factor}] + 0.000001 \quad (3)$$

The [dynamic smoothing factor] was chosen from the training set using LOOCV. The extra term, 0.000 001, had been added to avoid singularities. (3) Estimate the Gaussian kernel weighted  $k$ -NN regression as a weighted average, using  $w_i$ , of the measured inhibition  $y_i$

$$y_{\text{estimated}} = \sum_{i=1}^k w_i \times y_i / \sum_{i=1}^k w_i \quad (4)$$

(4) Having estimated the regression value, it remains to estimate the regression estimate's associated uncertainty. Therefore, assignment of a new set of weights to each of the  $k$  nearest neighbors based on their similarities occurs using a Gaussian kernel to map the similarity to a weight

$$w_{i,\text{uncertainty}} = K(s_i) = [(2\Pi)^{-1/2}/\sigma] \times \exp[-(1 - s_i)^2/(2 \times \sigma^2)] \quad (5)$$

The kernel bandwidth  $\sigma$  was chosen dynamically for the test molecule based on the nearest neighbor *plus an uncertainty term*

$$\sigma = (1 - s_j) \times [\text{dynamic smoothing factor}] + [\text{uncertainty term}] \quad (6)$$

The [uncertainty term] was chosen from the training set using LOOCV. (5) Estimate the regression estimate's associated uncertainty as the square root of the weighted average of the squared deviations

$$\text{error\_bar}_{\text{estimated}} = \sqrt{\sum_{i=1}^k w_{i,\text{uncertainty}} \times (y_i - y_{\text{estimated}})^2 / \sum_{i=1}^k w_{i,\text{uncertainty}}} \quad (7)$$

Given the predicted regression value  $y_{\text{estimated}}$  and the estimated uncertainty  $\text{error\_bar}_{\text{estimated}}$  and assuming that the uncertainty of the prediction follows a normal distribution, the estimates can easily

be converted to probabilities like  $P(\text{class} = \text{inhibitor} | x) \sim P(y \geq 50\% | x)$  using the Gaussian CDF as described in step 2 above.

The rationale for using different weighing schemes for the estimation of the regression value and for the estimation of the uncertainty, respectively, is that the actual measurements inherently contain some noise. The [uncertainty term] makes it possible to represent this noise in the model, by allowing more distant neighbors more influence on the uncertainty estimation than in the regression value estimation. In this way, a more reliable uncertainty estimation is obtained.

**Validation.** The parameters  $k$ , [dynamic smoothing factor], and [uncertainty term] and the type of ECFP of the models were chosen using LOOCV.<sup>53</sup> Permutation tests were made to make sure that the predictions made by the original models were based on the variance described by the fingerprints. The  $y$ -values from each training set were randomized four times, and four new models were generated. The models based on randomized data were then validated using LOOCV. The models were subsequently assessed using an internal CYP2D6 and CYP3A4 test set of 288 and 345 compounds, respectively. Furthermore, an external test set of 14 published inhibitors of either CYP2D6 or CYP3A4 were classified by use of the two classification models.

**Acknowledgment.** We thank Anni Schmidt and Karen Stentoft for excellent experimental work.

## References

- (1) Susnow, R. G.; Dixon, S. L. Use of Robust Classification Techniques for the Prediction of Human Cytochrome P450 2D6 Inhibition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1308–1315.
- (2) Wrighton, S. A.; Schuetz, E. G.; Thummel, K. E.; Shen, D. D.; Korzekwa, K. R.; Watkins, P. B. The human CYP3A subfamily: Practical considerations. *Drug Metab. Rev.* **2000**, *32*, 339–361.
- (3) Fukuda, K.; Otha, T.; Oshima, Y.; Ohashi, N.; Yoshikawa, M.; Yamazoe, Y. Specific CYP3A4 inhibitors in grapefruit juice: Furocoumarin dimers as components of drug interactions. *Pharmacogenetics* **1997**, *7*, 391–396.
- (4) Jones, B. C.; Tyman, C. A.; Smith, D. A. Identification of the cytochrome P450 isoforms involved in the *O*-demethylation of 4-nitroanisole in human liver microsomes. *Xenobiotica* **1997**, *27*, 1025–1037.
- (5) Kriegl, J. M.; Eriksson, L.; Arnhold, T.; Beck, B.; Johansson, E.; Fox, T. Multivariate modelling of cytochrome P450 3A4 inhibition. *Eur. J. Pharm. Sci.* **2005**, *24*, 451–463.
- (6) Ekins, S.; Berbaum, J.; Harrison, R. K. Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab. Dispos.* **2003**, *31*, 1077–1080.
- (7) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb. Sci.* **2005**, *24*, 491–502.
- (8) O'Brien, S. E.; de Groot, M. J. Greater than the sum of its parts: Combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48*, 1287–1291.
- (9) Zuegge, J.; Fechner, U.; Roche, O.; Parrott, N. J.; Engkvist, O.; Schneider, G. A fast virtual screening filter for cytochrome P450 inhibition liability of compound libraries. *Quant. Struct.-Act. Relat.* **2002**, *21*, 249–256.
- (10) Balakin, K.; Ekins, S.; Bugrim, A.; Ivanenkov, Y. A.; Korolev, D.; Nikolsky, Y. V.; Skorenko, A. V.; Ivashchenko, A. A.; Savchuk, N. P.; Nikolskaya, T. Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metab. Dispos.* **2004**, *32*, 1183–1189.
- (11) Molnár, L.; Keserü, G. M. A neural network based virtual screening of cytochrome P450 3A4 inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 419–421.
- (12) Arimoto, R.; Prasad, M. A.; Gifford, E. M. Development of CYP3A4 inhibition models: Comparisons of machine-learning techniques and molecular descriptors. *J. Biomol. Screening* **2005**, *10*, 197–205.
- (13) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Sci.* **2004**, *44*, 1971–1978.
- (14) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (15) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.
- (16) Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson, L. Model validation by permutation tests: Applications to variable selection. *J. Chemom.* **1996**, *10*, 521–532.

- (17) Usia, T.; Watabe, T.; Kadota, S.; Tezuka, Y. Cytochrome P450 2D6 (CYP2D6) inhibitory constituents of *Catharanthus roseus*. *Biol. Pharm. Bull.* **2005**, *28* (6), 1021–1024.
- (18) Nakajimi, M.; Ohyama, K.; Nakamura, S.; Shimada, N.; Yamazaki, H.; Yokoi, T. Oxidation catalyzed by human P-450 enzymes. *Drug Metab. Dispos.* **1999**, *27*, 792–797.
- (19) <http://medicine.iupui.edu/flockhart/table.htm>.
- (20) Pelkonen, O.; Mäenpää, J.; Taavitsainen, P.; Rautio, A.; Raunio, H. Inhibition of human P450 (CYP) enzymes. *Xenobiotica* **1998**, *28*, 1203–1253.
- (21) Wanchana, S.; Yamashita, F.; Hashida, M. QSAR analysis of the inhibition of recombinant CYP 3A4 activity by structurally diverse compounds using a genetic algorithm-combined partial least squares method. *Pharm. Res.* **2003**, *20*, 1401–1408.
- (22) Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **1964**, *29*, 115–129.
- (23) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates. Part 2: Properties of cytochrome P450 substrates. *Drug Discovery Today* **1997**, *2*, 479–486.
- (24) Lewis, D. F. V. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics. *Biochem. Pharmacol.* **2000**, *60*, 293–306.
- (25) Balakin, K.; Ekins, S.; Bugrim, A.; Ivanenkov, Y. A.; Korolev, D.; Nikolsky, Y. V.; Ivashchenko, A. A.; Savchuk, N. P.; Nikolskaya, T. Quantitative structure–metabolism relationship modeling of metabolic *N*-nealkylation reaction rates. *Drug Metab. Dispos.* **2004**, *32*, 1111–1120.
- (26) Labute, P. *Binary QSAR: A new method for quantitative structure activity relationships*, Proceedings of the 1999 Pacific Symposium; World Scientific Publishing: Singapore, 1999.
- (27) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, R. A. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295–1300.
- (28) Labute, P.; Nilar, S.; Williams, C. A probabilistic approach to high throughput drug discovery. *Comb. Chem. High Throughput Screening* **2002**, *5*, 135–145.
- (29) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Model.* **2004**, *44*, 1177–1185.
- (30) Jensen, B. F.; Refsgaard, H. H. F.; Bro, R.; Brockhoff, P. B.; Classification of membrane permeability: A methodological investigation. *QSAR Comb. Sci.* **2005**, *24*, 449–457.
- (31) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46*, 193–200.
- (32) Johnson, M. A., Maggiora, G. M. Eds. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
- (33) Kubinyi, H. Similarity and dissimilarity: A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 225–252.
- (34) Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (35) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (36) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activities? *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (37) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.
- (38) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, *48*, 240–248.
- (39) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comp.* **1973**, *C22*, 1025–1034.
- (40) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
- (41) MOE (Molecular Operating Environment) software available from Chemical Computing Group ([www.chemcomp.com](http://www.chemcomp.com)).
- (42) Xu, Y.; Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.
- (43) Definitions taken from UNITY4.2 available from Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144–2319.
- (44) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71–79.
- (45) Cheshire is a part of MDL Chemistry Rules Interface available from MDL Information Systems, Inc., San Leandro, CA, 94577.
- (46) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- (47) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 11, 2743–2749.
- (48) Pipeline Pilot by SciTegic, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123–1365, at [www.scitegic.com](http://www.scitegic.com).
- (49) R by “The R Project for Statistical Computing” at <http://www.r-project.org/>.
- (50) Shepard, D. A two-dimensional interpolation function for irregularly-spaced data, Proceedings of the 1968 23rd ACM National Conference, August 27–29, 1968; ACM Press: New York, 1968; pp 517–524.
- (51) Lowe, D. G. Similarity metric learning for a variable-kernel classifier. *Neural Comput.* **1995**, *7*, 72–85.
- (52) Atkeson, C. G.; Moore, A. W.; Schaal, S. Locally weighted learning. *Artif. Intell. Rev.* **1997**, *11*, 11–73.
- (53) Wold, S. Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics* **1978**, *20*, 397–405.

JM060333S