Chemoinformatics-Based Classification of Prohibited Substances Employed for Doping in Sport

Edward O. Cannon, Andreas Bender,† David S. Palmer, and John B. O. Mitchell*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Received March 30, 2006

Representative molecules from 10 classes of prohibited substances were taken from the World Anti-Doping Agency (WADA) list, augmented by molecules from corresponding activity classes found in the MDDR database. Together with some explicitly allowed compounds, these formed a set of 5245 molecules. Five types of fingerprints were calculated for these substances. The random forest classification method was used to predict membership of each prohibited class on the basis of each type of fingerprint, using 5-fold cross-validation. We also used a k-nearest neighbors (kNN) approach, which worked well for the smallest values of *k*. The most successful classifiers are based on Unity 2D fingerprints and give very similar Matthews correlation coefficients of 0.836 (kNN) and 0.829 (random forest). The kNN classifiers tend to give a higher recall of positives at the expense of lower precision. A naïve Bayesian classifier, however, lies much further toward the extreme of high recall and low precision. Our results suggest that it will be possible to produce a reliable and quantitative assignment of membership or otherwise of each class of prohibited substances. This should aid the fight against the use of bioactive novel compounds as doping agents, while also protecting athletes against unjust disqualification.

INTRODUCTION

The use of performance enhancing substances, more commonly termed doping, threatens both athletes' health and the integrity of their sports. In 1928 the use of stimulants, which are substances that increase alertness and aggression while reducing an individual's tiredness, was outlawed by the International Amateur Athletics Federation (IAAF).[1] By the 1930s, artificial hormones had been produced, but it was not until the 1950s that they were actively being used for doping to stimulate growth, control pain, and increase red blood cell production. A decade later, official doping tests were introduced by several international sports federations such as FIFA, the world governing body of soccer, and the IOC (International Olympic Committee), which established a list of prohibited substances.

Despite this prohibited list, the early 1970s gave rise to a wave of widespread anabolic agent abuse,[2] particularly in sports involving a large strength component. Athletes used these substances to increase muscle size and strength, allowing them to train harder. It did not, however, take long for a test to be constructed that could discriminate between anabolic agents and other substances, resulting in the IOC banning these substances in 1976.[3] The previous lack of standardization between different international sports bodies and governments led in 1999 to the establishment of WADA,[1] an agency set up on the initiative of the IOC to support and monitor the fight against all forms of doping in all sports. WADA provides a World Anti-Doping Code and administers the list of substances prohibited and allowed in sport.

The prohibited list[1] is composed of 11 different classes of banned substance: anabolic agents (S1), hormones and related substances (S2), beta-2 agonists (S3), agents with anti-estrogenic activity (S4), diuretics and masking agents (S5), stimulants (S6), narcotics (S7), cannabinoids (S8), glucocorticosteroids (S9), alcohol (ethanol) (P1), and beta-blockers (P2).

The S1, S2, and S6 banned substances and their performance enhancing effects have been discussed above. Beta-2 agonists (S3) have a powerful anabolic effect when injected into the bloodstream, causing muscle mass to increase and body fat to drop. They are often used to treat asthma patients; these drugs relax and open up airways in the lungs that become restricted during an asthma attack. If an athlete suffers from asthma, the athlete's physician may request a so-called therapeutic use exemption, permitting bronchodilator drugs of this type to be taken by inhalation. Anti-estrogenic agents (S4) such as tamoxifen and clomiphene are used to reduce sodium and water retention allowing bigger muscles to be built, particularly useful in strength-based events. Diuretics (S5), more traditionally used to treat conditions of heart failure or high blood pressure, have come under scrutiny for use in sport for weight loss and elimination of other prohibited drugs from the system. These might be abused in sports where the competitors compete in classes based on body weight, for example in boxing and lightweight rowing, or by jockeys whose horses have been assigned uncomfortably light weights. The use of furosemide for female gymnasts has been reported. Diuretics may also function as masking agents, which are capable of concealing the presence of other prohibited substances in urine.

* Corresponding author phone: +44-1223 762983; fax: +44-1223 763076; e-mail: jbom1@cam.ac.uk.

† Current address: Novartis Institutes for BioMedical Research, Discovery Technologies, Lead Discovery Informatics, 250 Mass Ave., 3C-473, Cambridge, MA 02139.

10.1021/ci0601160 CCC: $33.50
Published on Web 09/15/2006

Narcotics (S7) are taken as performance enhancers since they act as painkillers and may be used to allow an athlete to train for longer or more intense periods. Cannabinoids (S8) such as marijuana produce a sedating and euphoric feeling of wellbeing but may affect perception and judgment. Glucocorticosteroids (S9) have strong anti-inflammatory effects, while beta blockers (P2) have been used to reduce competitors' heart rates, particularly important in sports such as archery, where a slower heartbeat allows higher accuracy.

One strategy used to circumvent antidoping procedures is the design and synthesis of novel drugs, such as 'designer steroids'.[4] This is based on the ideas that, first, the testers are not looking for these substances and, second, that they do not appear explicitly on the banned lists. A spectacular example of this was the production and distribution of the new steroid tetrahydrogestrinone (THG),[4] which was apparently administered to a number of leading sprinters over a period of several months and went undetected until an anonymous informant supplied a sample to the antidoping authorities.

Currently, the WADA[1] list of prohibited substances uses the phrase "and other substances with a similar chemical structure or similar biological effect(s)" as a means of outlawing the use of analogues of known prohibited substances. We believe that chemoinformatics provides the best means of defining quantitatively what "similar chemical structure" means as well as predicting "similar biological effect(s)", and this paper provides an illustration of how that might be achieved.

Thus, there is a clear need to distinguish, in a quantitative manner, banned from allowed substances. One possible way to achieve this is to classify substances using machine learning algorithms in conjunction with molecular descriptors. It is assumed based on the similar property principle[5,6] that molecules with similar features, typically but not exclusively structural, will have similar biological activity. Based on this principle, one expects a drug to be most similar to others in its particular banned class, especially those sharing its receptor, but different from compounds outside that class. Thus it will be possible to classify substances using chemical and possibly biological descriptors, the banned classes ideally occupying distinct regions of chemical space. The selection and description of suitable molecular features and measures of similarity is nontrivial, and we will show that the degree of success in the classification of prohibited substances is dependent on such choices.

We now turn to the question of how a chemoinformatics-based classifier of putative doping agents might be used in practice. The first use of the classifier would be to define the prohibited status of new molecules such as THG, which have been in the past designed in an attempt to get around the rules. As it stands, the "similar chemical structure or similar biological effect(s)" criterion is far from satisfactory, since there is no quantitative definition of "similar" given. The current rules risk wrongly disqualifying athletes on the sole basis of subjectively similar chemical structure. Second, the classifier could be run against large databases of available druglike molecules, such as the MDDR,[7] to define the status of compounds before they come into use in sport.

Third, and more ambitiously, we would hope that the classifier could be combined with as a complete as possible an enumeration of the different receptors responsible for the relevant bioactivities and assays to detect them. Molecules identified by the classifier as putative positives from in silico database screens could then be assayed in vitro to confirm their bioactivity. This combination of computational and experimental methods would be very powerful. Fourth, we suggest that a protocol such as that suggested above would provide a more appropriate legally watertight definition of similarity.

Descriptors are normally classified according to their dimensionality. One-dimensional descriptors represent bulk properties such as molecular weight, volume, and molar refractivity.[8] Two-dimensional descriptors describe properties that can be calculated from the connection table,[9] and three-dimensional descriptors depend on the conformation of the molecule.[10] One-, two-, and optionally three-dimensional descriptors can be combined into an array of feature vectors termed a fingerprint. Fingerprints can then be used to describe the molecule in chemical space. Fingerprints may be binary, as in the cases of Unity 2D[11] fingerprints and also of MACCS keys, designed by MDL for systematic substructure searching.[7,12] A value of 1 represents the presence of a feature in the molecule, while a value of 0 means that the feature is absent. Fingerprints may also contain a list of integer values representing the features present in that molecule. Examples of such fingerprints are typed atom distance (TAD) and typed graph distance (TGD) fingerprints. The TAD fingerprint in MOE[13] uses the Euclidean distance in real space between atoms of two specific types, which requires MOE to generate a putative three-dimensional structure. The TGD fingerprint uses a distance measure between atoms in a molecule defined as the number of bonds separating them in the two-dimensional connection table. We convert these two integer-list fingerprints to binary form by a one-to-one mapping of integers to bits. If the correct distance between two atom types is present, the given integer appears in the list, and hence the corresponding bit in the fingerprint is set to 1. The recent work of Bender et al.[14,15] used the MOLPRINT 2D fingerprint, which works by generating a circular substructure fingerprint for each heavy atom in a molecule using distances from 0 to 2 bonds and keeping a count of the frequency of the atom types. MOLPRINT 2D resembles a number of other circular fingerprints, such as that of Faulon,[16] though it differs in its atom typing.

These descriptor-based methods define the location of molecules in chemical space. Classification tools based on machine learning algorithms are then applied to partition that space. Here, we have used the random forest[17] classifier in conjunction with five fingerprint definitions: Unity 2D, MACCS, all two-dimensional descriptors calculated from MOE (Version 2004.03), TAD, and TGD. We also compare the performance of a naïve Bayesian classifier based on the MOLPRINT 2D atom environment descriptors, as previously described by Bender et al.[14,15]

We are aware of only one previous chemoinformatics study related to drugs in sport; Kontaxakis and Christodoulou trained an artificial neural network to predict the chromatographic retention times of prohibited substances using molecular descriptors.[18]

The rest of this paper is composed of three sections. The Methods section gives details of the data set, the descriptors, and the classifiers. In the Results and Discussion section, we assess the quality of the classifiers using measures which

include the Matthews correlation coefficient (MCC), and the average percentage class correct classification (APCC). The Conclusions section summarizes our findings and their significance.

2. METHODS

(a) Data Set. All methods were applied to a final data set of 5245 molecules derived either explicitly from the prohibited list[1] or taken from corresponding activity classes found in the MDDR[7] database (Version 2003.1). We justify the addition of MDDR molecules on the basis of our understanding of what a WADA prohibited class really is. Such a class is not simply a small prescribed set of proscribed molecules but rather a category of bioactivities. While the numbers of molecules explicitly listed in the classes are small (sometimes very small), these are designed to incorporate analogues of the same bioactivity—hence the use of phrases such as "and other substances with a similar chemical structure or similar biological effect(s)" in the list. Thus we regard it as entirely appropriate to harvest known molecules with the given bioactivities from the MDDR; indeed for many classes it is the only way in which a usable data set could possibly have been obtained.

During the process of reviewing this paper, an anonymous reviewer considered our harvesting of molecules from the MDDR to be unsatisfactory. This reviewer also felt there was an inconsistency in our treatment of the biological and chemical parts of the definition of similarity. "The authors cannot have it both ways: if they use the extended definition of WADA-banned substances to encompass entire MDDR classes then they must also by definition include anything [chemically] similar to a banned substance. This may appear to be a *reductio ad absurdum*, but that is the logical consequence of the authors' approach." Our method of defining the data set effectively regards "similar biological activity" as the key determinant of whether a molecule is a member of a prohibited class. Each of our classifiers is a particular definition of chemical similarity, chosen from a myriad of different possibilities and designed to reproduce as closely as possible the classifications arising from biological similarity. Thus we regard the definition of biological similarity as fixed and mould for ourselves a definition of chemical similarity to match. This reflects an implicit assumption that the doping rules are designed with the intention that a molecule should be prohibited if and only if it has a similar biological effect to other prohibited compounds.

The data set contained the following: 47 anabolic agents (from here on referred to as S1), 272 hormones and related substances (S2), 367 beta-2 agonists (S3), 928 agents with anti-estrogenic activity (S4), 1000 diuretic and masking agents (S5), 804 stimulants (S6), 195 narcotics (S7), 1000 cannabinoids (S8), 26 glucocorticosteroids (S9), and 239 beta-blockers (P2), in addition to 367 explicitly allowed[19] substances. Ethanol was excluded from this study on the basis that one substance could not make up a banned class for the purpose of classification.

More details of the data set are provided in the Supporting Information. The MDDR classes from which it was drawn are given in Table SI1, along with an outline of how it was obtained by filtering an initial data set of 11 565 molecules. This filtering involved removing salts and counterions and deleting the remaining duplicates. The S5 and S8 classes initially incorporated large numbers of molecules found in the MDDR database. As a result, a diversity-based selection was carried out in MOE to choose the 1000 most diverse molecules for the S5 and S8 prohibited classes, based on two-dimensional MOE descriptors. The 1000 most diverse compounds for the S5 and S8 classes were taken forward for classification. Table SI2 (Supporting Information) gives both interclass and intraclass similarity data for our data set, based on Unity 2D fingerprints and the Tanimoto coefficient.[20,21]

Table 1 gives a pictorial representation of the range of molecules in each prohibited class by identifying the molecule in each class that was considered the most similar and the molecule which was the most dissimilar to the rest, using the MOE 2D descriptors. Although many different similarity coefficients exist,[20,21] this analysis is based on the Euclidean distance. The most representative molecule of a class is taken to be that with the smallest mean Euclidean distance to the others in that class; the least representative molecule is that with the largest mean intraclass distance.

(b) Descriptor Generation. Five different molecular fingerprints were calculated. The molecules were first imported into MOE[13] as SMILES strings.[22] The first fingerprint was Unity 2D, calculated in Sybyl 7.1.[11] This fingerprint describes substructural features in the molecules through a Boolean array of length 992 bits. If the substructural feature is present the bit is assigned the value of 1, otherwise its value is 0. The second fingerprint was the MACCS keys,[12] calculated in MOE.

All 146 two-dimensional MOE descriptors were calculated using MOE and comprised our third fingerprint. Our fourth and fifth fingerprints, TAD and TGD, were also generated in MOE. Although the MACCS, TAD, and TGD descriptors are produced by MOE as sets of integers, we used a Perl script to convert them to a binary format; if an integer N appeared in the list generated by MOE, our script set the Nth bit of the binary string to 1. Our fingerprint lengths in numbers of descriptors (bits for binary fingerprints) are Unity 992, MACCS 166, MOE 146, TGD 734, and TAD 1024.

(c) Random Forest Models for Classification. Machine learning is defined as "*The study of computer algorithms that improve automatically through experience*."[23] One of the main areas where machine learning algorithms are used is in data mining,[24] where the objective is to search a large hypothesis space such as chemical feature space and determine the hypothesis that best fits the data. There are two different types of machine learning. The first is unsupervised learning, where the aim is to identify previously unknown similarities in the data. The second method, and the one which we have used here, is supervised learning. With supervised learning the outcome of a particular data entry is already known. This means the class category of the substance is known prior to running the machine learning algorithm in the classification study. The simplest form of supervised learning, which was applied in this work, is binary classification. In a binary classification the data labels can take one of two values, "true" or "false". In this work, substances labeled "true" belong to the specific class of prohibited substance that the classifier is being tested on; otherwise the label is "false".

Table 1: Most and Least Representative Molecules in Each Prohibited Class

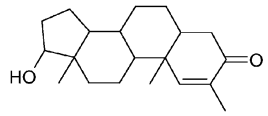
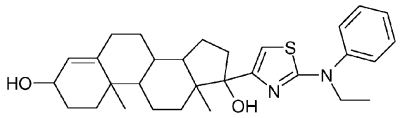
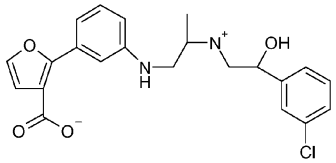
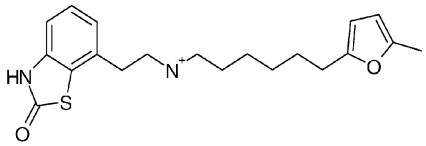
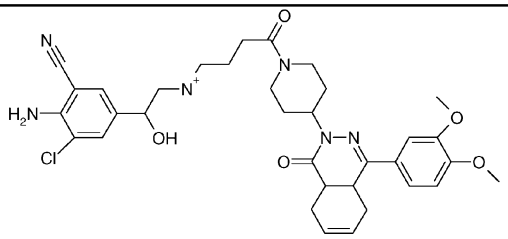
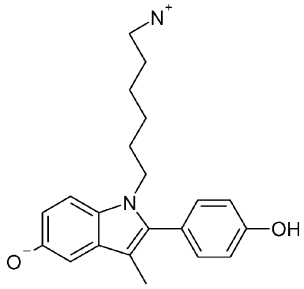
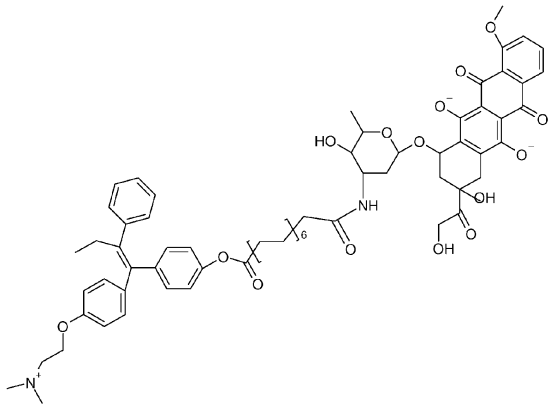
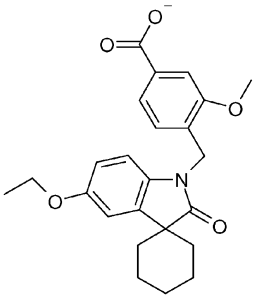
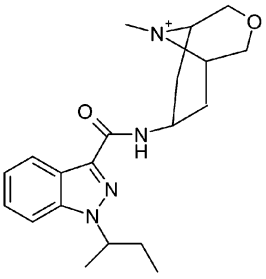
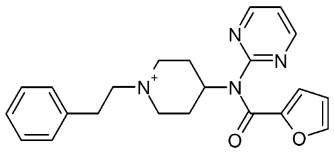
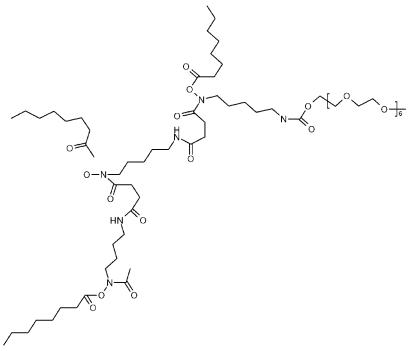
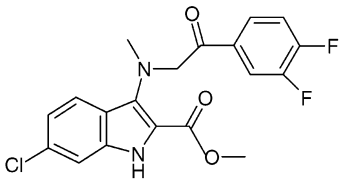
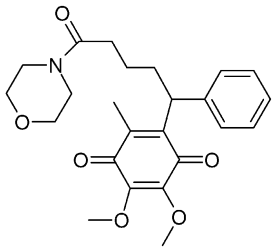
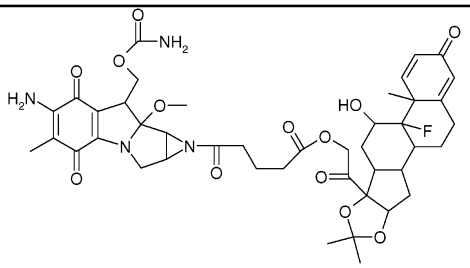
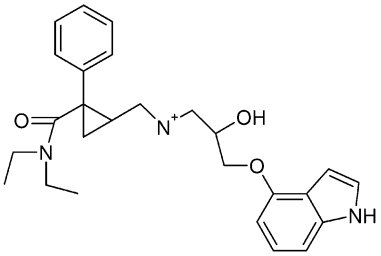
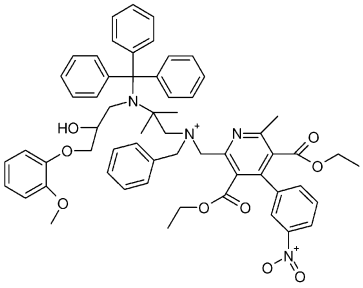
Class	Most Representative	Least Representative
S1		
S2		$C_{81}H_{122}N_{23}O_{21}$
S3		
S4		
S5		Cyclic polypeptide with 92 residues
S6		Linear polypeptide with 51 residues

Table 1 (Continued)

Class	Most Representative	Least Representative
S7		
S8		Dendritic polypeptide with 26 residues
S9		
P2		

We used random forest as a machine learning algorithm to classify the prohibited substances based on whether they were considered part of a banned class, using the fingerprints outlined above as input to the classifiers.

The random forest algorithm has been shown to be successful as a classification tool in work such as that by Svetnik et al.[25] Random forest is an ensemble of decision trees which can handle high dimensional data well and has the ability to ignore irrelevant descriptors. Random forest operates by generating a user-defined number of decision trees. At each node of each tree, a split is made based on the best partitioning available from any of a random selection of *mtry* descriptors from the set of molecular descriptors previously generated for the data set of molecules. Random forest produces one output per molecule per tree, classifying the molecule into either the category of members or that of nonmembers for a particular banned class. The outputs are then aggregated together, and the category of each molecule is predicted on the basis of that for which the majority vote is obtained.

In this work we used random forest to classify the molecules as members or nonmembers of each banned class. These random forest classifications were undertaken in the computing and statistical package R[26] with 5-fold cross-validation. We used a random forest classifier based on the default parameter values for each fingerprint definition. We also investigated optimizing the parameters of the random forest classifier by varying *mtry* between 10 and 100 (in steps of 10) and *ntree* between 100 and 1000 (in steps of 100). The parameter *mtry* is the number of descriptors randomly selected as candidates at each node in the tree, and *ntree* is the number of trees in the ensemble. In the classifiers where we varied *mtry*, we used the default *ntree* value of 500. Conversely, where we varied *ntree* we used the default *mtry*,

which is the square root of the length of the fingerprint (rounded down to an integer). Finally, we combined the 'best' *ntree* and *mtry* values for each given fingerprint definition in 'pseudo-optimized' random forest classifiers. This gave 22 random forest classifiers for each of the five fingerprint definitions; each of these 110 classifiers was applied to each of the 10 classes of prohibited substances.

(d) k-Nearest Neighbor Classification. We also generated simple models based on k-nearest neighbor (kNN) classification, implemented in R. There is evidence in the literature that kNN is a useful nonlinear method for assigning molecular similarities.[27] The class membership of a given molecule in the test set was assigned on the basis of the majority vote of its k nearest neighbors in the training set. Proximity was measured by the Euclidean distance in the appropriate descriptor space. In the event of a tied vote, the casting vote was assigned randomly; any ties for the position of *k*th nearest neighbor led to the inclusion of all tied candidates in the 'electoral college'. We used an analogous 5-fold cross-validation approach to that of the random forest models, with 80% of the data used as the training set and 20% as the test set in each run, and the performance being evaluated over sets of five runs in which each molecule in the data set is predicted once.

Four of our fingerprints, Unity, MACCS, TAD, and TGD, were stored in binary format, and we did not carry out any scaling of their descriptors. The MOE descriptors were scaled to zero mean and unit variance before carrying out the kNN calculations. Our models used $k = 1$, 2, 3, 4, 5, 10, 15, and 20. This resulted in eight kNN classifiers per fingerprint definition, 40 in total.

In our preliminary work, artificial neural networks and support vector machines performed significantly less well than the random forest and kNN classifiers reported here, and we did not pursue their use.

(e) Naïve Bayesian Classifier. The naïve Bayesian classifier uses the feature vector for a specific molecule to predict the probability that it is part of the particular prohibited class under test. The category (that is, either the members or the nonmembers of the given WADA class) with the higher probability from the product of the individual probabilities of the elements of each feature vector is assumed to be the category that the substance originated from

$$\mathrm{P}(\mathrm{CA}_v|\mathrm{F}) = \frac{\mathrm{P}(\mathrm{CA}_v)\mathrm{P}(\mathrm{F}|\mathrm{CA}_v)}{\mathrm{P}(\mathrm{F})} \tag{1}$$

$\mathrm{P}(\mathrm{CA}_v)$ is the probability of category v (the two "categories" here being members and nonmembers of the WADA prohibited class under test); P(**F**) is the feature vector probability; and $\mathrm{P}(\mathrm{F}|\mathrm{CA}_v)$ is the probability of F given a category CA_v.

Feature selection is employed to select the number of features present in the smaller of the two data sets used for classification (here the members the banned class to be predicted, since it always contained far fewer compounds than the nonmembers). While feature selection has a profound impact on classification performance,[14,15] more recently the impact of different Laplacian corrections in combination with the Bayes classifier was investigated.[28] As with the previous classifiers, our work used 5-fold cross-validation. We generated a single naïve Bayesian classifier and applied it to each of the 10 classes of prohibited substances. This brought the total number of classifiers used to 151, that is (110 + 40 + 1).

(f) Measures of Performance. For each of the classifiers operating on each of the 10 banned classes, a 2 × 2 confusion matrix was generated. This matrix gave the numbers of (1) true positives (t_p), members of the banned class classified correctly; (2) true negatives (t_n), nonmembers classified correctly; (3) false positives (f_p), nonmembers wrongly classified as being in the banned class; and (4) false negatives (f_n), members wrongly classified as nonmembers.

False positives can arise in one of two ways, remembering that molecules are tested against one prohibited class at a time. The first is that an explicitly allowed molecule is predicted as positive. The second is that a molecule from the 'wrong' prohibited class is predicted as positive against the class under test.

The recall of the classification of members of the banned class (positives) is given by

$$R_p = 100\% \times \frac{t_p}{t_p + f_n} \tag{2}$$

and the precision by

$$P_p = 100\% \times \frac{t_p}{t_p + f_p} \tag{3}$$

We analogously define the recall for nonmembers (negatives) by

$$\mathrm{R}_n = 100\% \times \frac{t_n}{t_n + f_p} \tag{4}$$

and the precision for nonmembers by

$$\mathrm{P}_n = 100\% \times \frac{t_n}{t_n + f_n} \tag{5}$$

Using this information, we calculated two measures to express the quality of each classification. The first, and our preferred measure, was the Matthews correlation coefficient (MCC)

$$\mathrm{MCC} = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{6}$$

This coefficient returns a value between −1 and 1. The higher the value of the MCC, the more reliable is the classification result. A few classifiers predicted all molecules to be nonmembers of classes S1, S7, or S9; in these cases we used the value of 0 for the MCC for that classifier within that class, implying that no meaningful classification was taking place. This is formally equivalent to replacing the denominator in eq 6 to give a slightly modified definition of the MCC

$$\mathrm{MCC}^* = \frac{t_p t_n - f_p f_n}{\mathrm{MAX}[1, \sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}]} \tag{7}$$

In fact, Baldi et al.[29] noted that, as $(t_p + f_p)$ tends to zero, the limiting value of the MCC is indeed zero. This observation may be a more mathematically elegant resolution of the

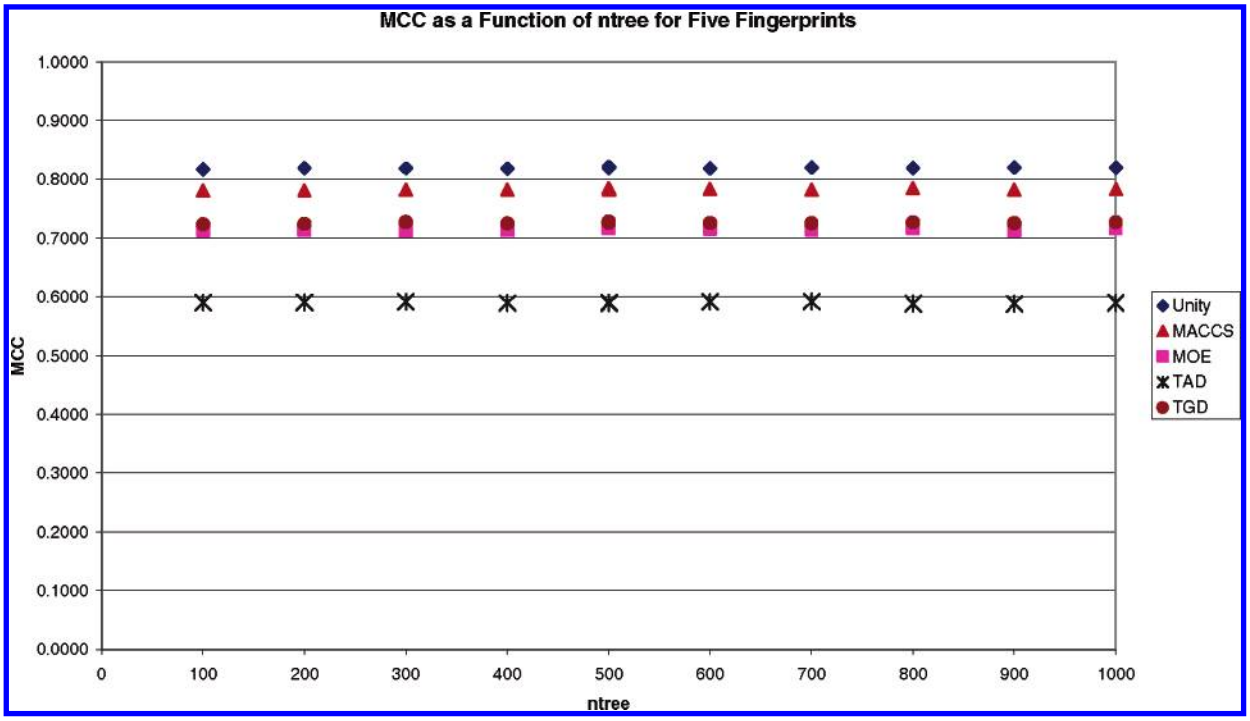


Figure 1. Variation of the MCC with random forest parameter *ntree* for the five fingerprints. In each case, the default *mtry* is used (see Table 2). The five default models from Table 2 appear at *ntree* = 500 on this graph, though each was repeated as part of the scan—giving slightly different values.

apparent difficulty than either our MCC* or Burset and Guigo's "average conditional probability".[30] An overall MCC was calculated for each classifier by summing the t_p, t_n, f_p, and f_n values over the 10 classes and then applying eq 6. The overall MCC was our primary criterion for assessing classification performance, though the MCC (in the above-mentioned few cases, strictly MCC*) values for each banned class were also of interest. The MCC helpfully combines assessment of both recall and precision into a single number.

The second measure was the average percentage class correct (APCC)

$$\mathrm{APCC} = (R_p + R_n)/2 \qquad (8)$$

APCC is a measure of recall, being the average of the percentage recalls of positives and negatives. An overall APCC was calculated for each classifier by summing the t_p, t_n, f_p, and f_n values over the 10 classes and then applying eq 8.

Measuring performance using a single number is difficult—APCC may favor classifiers giving more false positives (the negative category is much the larger, so R_p rises by more than R_n falls when positives are overpredicted). It will be shown below that this effect favors the naïve Bayesian classifier here. The MCC uses all of t_p, t_n, f_p, and f_n and is usually regarded as a more balanced measure than the APCC or other similar quantities. We do not claim that the MCC always provides a uniquely fair assessment; indeed there may be considerable debate as to what "fair" means, especially at the interface of science and sport.[31] The relative merits of different measures of classification performance have been discussed at length by Baldi et al.[29]

3. RESULTS AND DISCUSSION

The performance of the random forest classifiers using default parameters was analyzed for each of the five fingerprints (Tables 2 and SI3(a)). Considering the values of t_p, t_n, f_p, and f_n aggregated over the 10 classes, the classifier based on Unity fingerprints gave the best model (MCC=0.821), followed by MACCS (0.782). TGD (0.728) and MOE 2D (0.717) also performed well. The TAD classifier (0.590) was the least successful.

Figure 1 shows the MCC values obtained for different values for *ntree* for the various fingerprint definitions. The results show that varying the *ntree* parameter in random forest made little or no difference to the performance of the classifiers, in agreement with earlier work of Svetnik.[25] Even 100 trees per forest seems to be an adequate number, and gains from further increasing *ntree* are small. On occasions, the stochastic nature of the method led to small deteriorations in performance from using more trees, though random forests do not suffer from overfitting in the way that a neural network would.[17] Variations between fingerprints are larger than those resulting from changing *ntree*, and the following relative order of performance is maintained:

$$\text{Unity} > \text{MACCS} > \text{TGD} \approx \text{MOE} > \text{TAD}$$

The results of varying *mtry* are shown in Figure 2. The results are not very sensitive to *mtry* for moderate to large values for *mtry*. However, for values significantly smaller than the square root of the fingerprint length, which is the default setting, some deterioration is seen for TAD and TGD, two of the longer fingerprints, though the effect is very modest for Unity. For MACCS and MOE, the fingerprints are too short for our scan to cover values significantly less than the default value of 12. The rankings of the fingerprints suggested by this graph are similar to before, leading to the same order as before for fixed *mtry* (rather than the default *mtry* in each case):

$$\text{Unity} > \text{MACCS} > \text{TGD} \approx \text{MOE} > \text{TAD}$$

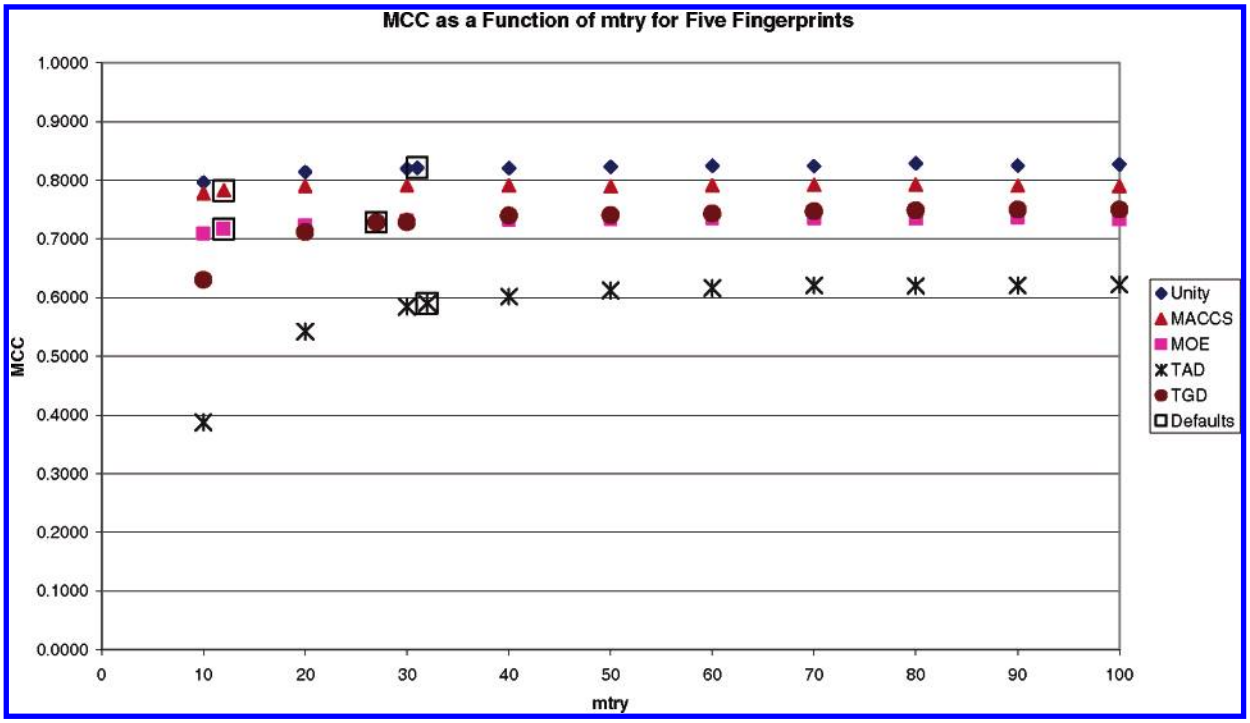


Figure 2. Variation of the MCC with random forest parameter *mtry* for the five fingerprints. In each case, the default value of *ntree* (500) was used. The default values of *mtry* vary as the square root of the fingerprint length and are marked by open boxes on this plot.

Table 2: Classification Performance of the Five Random Forest Default Models, and (in Bold Type) the Best, According to MCC, Random Forest Classifier in This Work[a]

fingerprint	ntree	mtry	t_p	t_n	f_p	f_n	R_p	P_p	R_n	P_n	APCC	MCC
Unity	500	31	3548	47420	152	1330	72.73	95.89	99.68	97.27	86.21	0.8214
MACCS	500	12	3295	47376	196	1583	67.55	94.39	99.59	96.77	83.57	0.7823
TGD	500	27	3013	47259	313	1865	61.77	90.59	99.34	96.20	80.55	0.7283
MOE 2D	500	12	2892	47307	265	1986	59.29	91.61	99.44	95.97	79.36	0.7172
TAD	500	32	2115	47259	313	2763	43.36	87.11	99.34	94.48	71.35	0.5902
Unity	**500**	**80**	**3625**	**47398**	**174**	**1253**	**74.31**	**95.42**	**99.63**	**97.42**	**86.97**	**0.8286**
MOLPRINT	Naïve Bayesian		4400	39963	7609	478	90.20	36.64	84.01	98.82	87.10	0.5129

[a] The table gives the random forest parameters *ntree* and *mtry*, the numbers of true and false positives and negatives (t_p, t_n, f_p, and f_n) aggregated over all 10 classes of prohibited substances, the percentage recall and precision achieved for positives and negatives (R_p, P_p, R_n, and P_n), and the APCC and MCC. The MOLPRINT 2D naïve Bayesian classifier is also included. A complete listing of results for all random forest classifiers is given in the Supporting Information, Table SI3(a).

The best random forest classifier found in this work, by a small margin, was that with *mtry* = 80, giving an MCC of 0.829 (see penultimate row of Table 2).

The five pseudo-optimized random forest classifiers produced results that were, in each case, very close to the MCC of the best classifier for that fingerprint, the largest difference being 0.005. In only one case (TAD) did the pseudo-optimized classifier actually perform slightly better than the previous best for that fingerprint.

Figure 3 and Tables 3 and SI3(b) show the results achieved by the kNN classifiers. Two trends are apparent. One is that Unity tends to perform best for a given *k*, followed closely by MACCS, though the detailed rank order varies with *k*. The other is that performance falls off with *k*, the best results for each fingerprint coming from the $k = 1$ classifier. Interestingly, the loss in recall of positives upon increasing *k* does not bring with it any significant improvement in precision. The Unity $k = 1$ classifier is the most successful one seen in this work. Thus the best (according to the MCC) predictor that we have found of the WADA-prohibited (or allowed) status of a compound is simply to assign it the same status as the nearest neighbor in the Unity descriptor space of the training set. In general, the better kNN classifiers tend to predict more positives, obtaining a higher recall, but lower precision, than the better random forest ones. We note that, in our test calculations, these kNN classifiers based on Euclidean distance narrowly outperformed alternatives which used the Tanimoto coefficient. The latter were, therefore, not pursued further.

The APCC measure, which is more sensitive to the recall of positives than the MCC, gave a somewhat different ranking of classifiers. Although both measures ranked the Unity $k = 1$ kNN classifier as the best, the next 21 places in the MCC rankings were filled by Unity random forest classifiers. In contrast, the top seven classifiers in the APCC list were kNN ones, and the naïve Bayesian classifier had a higher APCC than any random forest one. We see a tendency for the choice of fingerprint to dominate the top end of the MCC-based classifier rankings, while the choice of classification algorithm plays a bigger role at the top end of the APCC-based rankings.

The naïve Bayesian classifier (see the last row of Table 2) was unusual in that it predicted many more positives than the other classifiers (having the highest numbers of both true

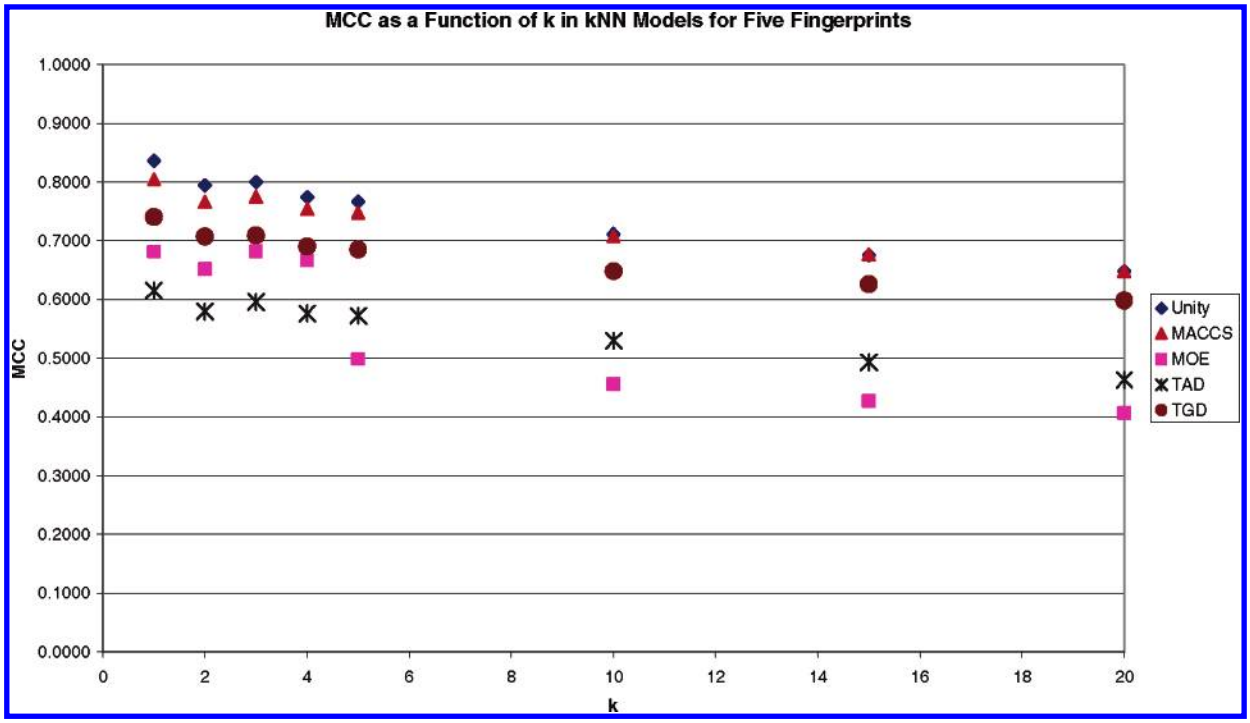


Figure 3. Performance (MCC) of the k nearest neighbor classifiers (k = 1, 2, 3, 4, 5, 10, 15, 20) for the five fingerprints.

Table 3: Classification Performance of the Five k-Nearest Neighbor Classifiers for k = 1 (Best Classifier in Bold), and for Unity for Other Values of k[a]

fingerprint	k	t_p	t_n	f_p	f_n	R_p	P_p	R_n	P_n	APCC	MCC
Unity	1	4148	46855	717	730	85.03	85.26	98.49	98.47	91.76	0.8363
MACCS	1	4027	46688	884	851	82.55	82.00	98.14	98.21	90.35	0.8045
TGD	1	3741	46406	1166	1137	76.69	76.24	97.55	97.61	87.12	0.7404
MOE 2D	1	3482	46139	1433	1396	71.38	70.84	96.99	97.06	84.18	0.6814
TAD	1	3176	45868	1704	1702	65.11	65.08	96.42	96.42	80.76	0.6152
Unity	2	3929	46719	853	949	80.55	82.16	98.21	98.01	89.38	0.7946
Unity	3	3848	46878	694	1030	78.88	84.72	98.54	97.85	88.71	0.7996
Unity	4	3682	46839	733	1196	75.48	83.40	98.46	97.51	86.97	0.7735
Unity	5	3599	46884	688	1279	73.78	83.95	98.55	97.34	86.17	0.7668
Unity	10	3190	46899	673	1688	65.40	82.58	98.59	96.53	81.99	0.7114
Unity	15	2892	46970	602	1986	59.29	82.77	98.73	95.94	79.01	0.6758
Unity	20	2668	47025	547	2210	54.69	82.99	98.85	95.51	76.77	0.6483

[a] The table gives the numbers of true and false positives and negatives (t_p, t_n, f_p, and f_n) aggregated over all 10 classes of prohibited substances, the percentage recall and precision achieved for positives and negatives (R_p, P_p, R_n, and P_n), and the APCC and MCC. A complete listing of results for all kNN classifiers is given in the Supporting Information, Table SI3(b).

and false positives). It predicted 2.5 times more positives than were actually present; thus it got almost all the real positives right but at the considerable cost of mispredicting very many real negatives as predicted positives. So, although it produced an excellent recall of 90.2% of positives (the best of any classifier), its precision was relatively poor (36.6%, the worst of any classifier). Overall, its MCC ranked it 144th out of 151 classifiers, but its APCC ranked it as high as 8th.

The value of such a classifier is very dependent on the relative costs of false positives and false negatives. If one used a classifier to make real decisions to punish or exclude athletes, the cost of a false positive (leading to unjust disqualification) is high, both in terms of public relations and possible legal action, while the cost of a false negative would be lower. In this scenario, our random forest classifiers would be better than kNN, which would in turn be a much better choice than the naïve Bayesian classifier. The best random forest classifier would be a good conservative choice, with a very small probability of a false positive—it gives 174 in total. Out of 52 450 assignments (5245 molecules × 10 classes), 47 572 ought correctly to be classified as negatives; the aforementioned best random forest classifier misclassifies only 0.37% of these (i.e., 174) as false positives. Given that it is rare for a molecule to be a false positive in more than one class, this implies an average probability (considering all banned classes) of about 3% of a molecule being wrongly assigned as prohibited (174/5245 is an upper bound).

As discussed in the Introduction, it would also be possible to use such a classifier to carry out virtual screening on possible drugs and metabolites, with hits being flagged up for further investigation by assays. In this case, no immediate punitive action would be taken against an athlete unless bioactivity was confirmed by an experimental result. Here, a high confidence of not missing compounds would be desirable, kNN would be better than random forest, and the naïve Bayesian classifier would be somewhat more competitive with other classifiers.

Typically, there is more than one receptor that may give rise to the bioactivity associated with a particular WADA

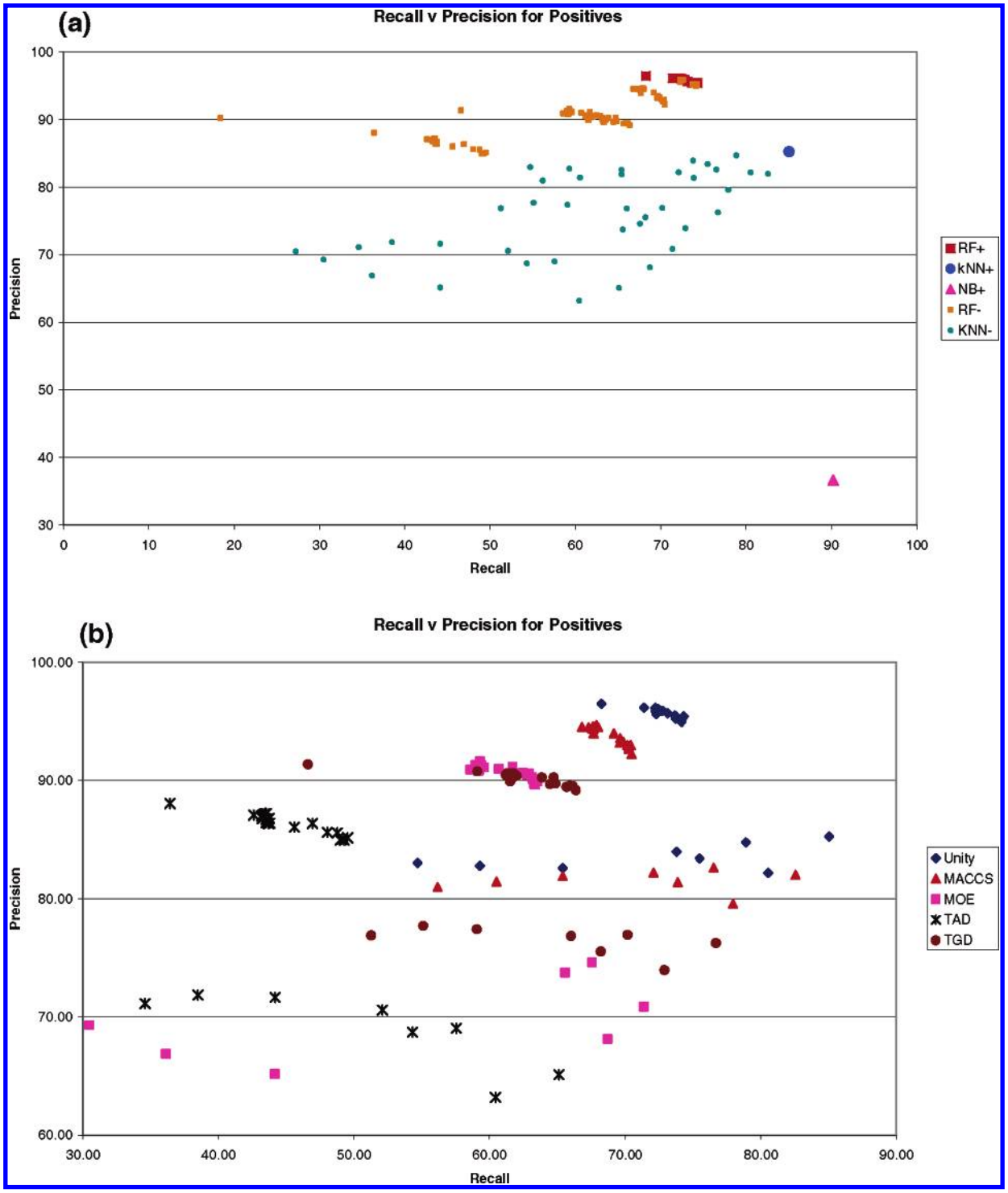


Figure 4. (a) The tradeoff of recall versus precision for positives. Each point represents a single classifier. Ten random forest classifiers (red squares), one kNN classifier (blue circle), and the naïve Bayesian classifier (pink triangle) are Pareto optimal. Classifiers which are not Pareto optimal are shown as orange squares for random forest and aquamarine circles for kNN. (b) An enlargement of the top right with points colored by fingerprint. All Pareto optimal random forest and kNN classifiers are based on Unity fingerprints.

prohibited class. Thus, the data set for each class is likely to consist of more than a single cluster in chemical space, and our classifiers are being asked to recognize more than a single pharmacophore. In this context, it is not surprising that the kNN methods with small values of k do well. These are very local models, with the classification assignment being based on the properties of a small number of neighboring molecules, rather than on the overall properties of the potentially diverse set sharing the same WADA class label. Thus, kNN is well suited to classes comprising a plurality of clusters. We believe that the use of the classification algorithm to "scaffold hop" in this way is important since fingerprints like Unity, which are a conglomerate of atom and feature counts and path-based descriptors, are thought to be relatively poor at finding different structural scaffolds with the same biological effect. This was demonstrated quite impressively for the related Daylight fingerprints.[32]

Figure 4(a) illustrates the tradeoff of recall versus precision for positives for all 151 classifiers. Ten random forest classifiers based on Unity fingerprints, the $k = 1$ Unity kNN

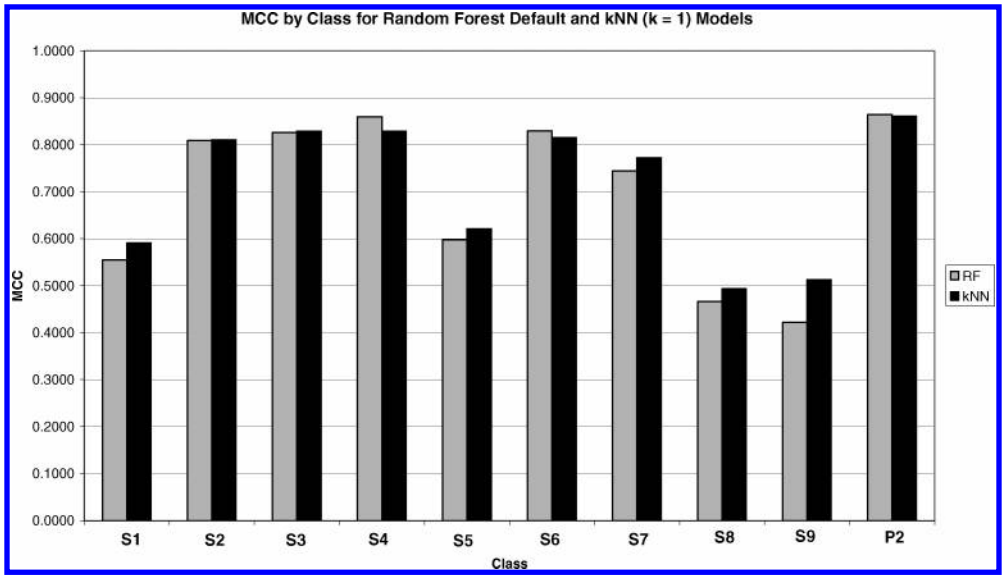


Figure 5. The difficulty of classifying molecules from each prohibited class is illustrated by the class-specific MCC values based on the aggregated numbers of true and false positives and negatives over the five random forest default classifiers (grey) and over the five kNN ($k = 1$) classifiers (black). The underlying numerical data are given in the Supporting Information, Table SI4.

classifier, and the naïve Bayesian classifier are Pareto optimal (a Pareto optimal classifier is one which is not outperformed on both recall and precision by the same alternative classifier). The diagram shows that the improvement of about 5% in recall on going from the $k = 1$ Unity kNN classifier to the naïve Bayesian classifier comes at the cost of a loss of in precision of nearly 49%. Thus, the kNN classifier is preferred. Figure 4(b) confirms that Unity, and, to a lesser extent, MACCS fingerprints produce the most effective random forest and kNN classifiers, which appear toward the top right of the recall versus precision plot.

Figure 5 shows the degree of difficulty of classifying molecules from each prohibited class, illustrated by the class-specific MCC values averaged over the five random forest default classifiers and also over the five kNN ($k = 1$) classifiers. The corresponding numerical data are given in Table SI4 (Supporting Information). There is some considerable variation in difficulty between the classes; six (P2, S4, S3, S6, S2, and S7) appear relatively 'easy' to predict and four (S9, S8, S1, and S5) relatively 'difficult'. However, this is the only place in our paper where we report classification performance across data sets of different sizes. This variation introduces a significant caveat to the discussion of classification performance on the individual classes.

The two smallest classes (S9, $N = 26$; S1, $N = 47$) are among the hardest to predict, but there is no significant overall correlation between class size and prediction difficulty as judged by the class-specific MCC. Our S5 and S8 data sets were largely derived using diversity selection from the MDDR, so the classifiers' relatively poor performance in predicting them may be related to their greater diversity. The higher kNN classifiers (above $k = 5$) are able to predict virtually no positives for S9, which has typically only 21 members in the training set for each cross-validation; in this case the criterion that a majority of the k neighbors be members is extremely exacting. However, S1—although larger by less than a factor of 2—does have positives predicted even up to $k = 20$. We expect prediction difficulty to depend both on the diversity within a class and on its degree of separation from other classes in chemical space.

4. CONCLUSIONS

We have shown that it is possible to identify banned substances taken from the 2005 prohibited list of the World Anti-Doping Agency. The most successful classifiers are based on Unity fingerprints and give very similar Matthews correlation coefficients of 0.836 (kNN) and 0.829 (random forest). The kNN classifiers tend to give a higher recall of positives at the expense of lower precision. The naïve Bayesian classifier of Bender et al.[14,15] gives a slightly higher recall than any of our other classifiers but at the cost of much lower precision.

Which method is preferred then depends on the relative cost of false positives versus false negatives. The best random forest classifier would be a good choice if a chemoinformatics method were to be used to guide decisions on the disqualification and suspension of athletes, since it has a low false positive rate. The best kNN method gives a higher recall of positives at the cost of lower precision and would be preferred if a classifier were to be used to flag up compounds for further investigation.

Our results suggest that it will be possible to produce a reliable and quantitative assignment of membership or otherwise of each class of prohibited substances. This should aid the fight against the use of bioactive novel compounds as doping agents, while also protecting athletes against unjust disqualification.

ACKNOWLEDGMENT

We thank the EPSRC, Unilever plc, the Gates Cambridge Trust, and the Pfizer Institute for Pharmaceutical Materials Science for funding and Tripos, Inc. for the provision of Sybyl 7.1.

Supporting Information Available: Outline of how the final data set of 5245 molecules was obtained and listing of the WADA and MDDR classes from which it was drawn (Table SI1), interclass and intraclass similarity data for our data set, based on Unity 2D fingerprints and the Tanimoto coefficient (Table SI2), complete listing of results for all 151 classifiers, aggregated over the 10 WADA classes (Table SI3), and class-

specific results, aggregated over classifiers, showing the degree of difficulty of classifying molecules from each prohibited class (Table SI4). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) World Anti-Doping Agency (WADA), Stock Exchange Tower, 800 Place Victoria, (Suite 1700), P.O. Box 120, Montreal, Quebec H4Z 1B7, Canada. http://www.wada-ama.org/ (accessed Oct 7, 2005).

(2) Todd, J.; Todd, T. In *Doping in Elite Sport: The Politics of Drugs in the Olympic Movement*; Wilson, W., Derse, E., Eds.; Human Kinetics, Champaign, IL, U.S.A., 2001; Chapter 1, pp 65−128.

(3) Brooks, R. V.; Firth, R. G.; Sumner, N. A. Detection of anabolic steroids by radioimmunoassay. *Br. J. Sports Med.* **1975**, *9*, 89−92.

(4) Death, A. K.; McGrath, K. C. Y.; Kazlauskas, R.; Handelsman, D. J. Tetrahydrogestrinone is a potent androgen and progestin. *J. Clin. Endocrin. Metab*. **2004**, *89*, 2498−2500.

(5) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(6) Bender, A.; Jenkins, J. L.; Li, Q.; Adams, S. E.; Cannon, E. O.; Glen, R. C. Molecular Similarity: Advances in Methods, Applications and Validations in Virtual Screening and QSAR. *Annu. Rep. Comput. Chem*. **2006**, *2*, 145−173.

(7) Elsevier MDL, 2440 Camino Ramon, San Ramon, CA 94583, U.S.A. http://www.mdli.com (accessed June 14, 2006).

(8) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(9) Estrada, E.; Uriarte, E. Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.

(10) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.

(11) Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144-2319, U.S.A. http://www.tripos.com (accessed June 14, 2006).

(12) McGregor, M. J.; Pallai, P. V. Clustering large databases of compounds using the MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(13) Chemical Computing Group, Suite 910, 1010 Sherbrooke St. W, Montreal, Quebec H3A 2R7, Canada. http://www.chemComp.com (accessed June 14, 2006).

(14) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching using Atom Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

(15) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(16) Faulon, J.-L. Stochastic Generator of Chemical Structure. 1. Application to the Structure Elucidation of Large Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204−1218.

(17) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(18) Kontaxakis, S. G.; Christodoulou, M. A. A Neural Network System for Doping Detection in Athletes, *Proceedings of the 4th International Conference on Technology and Automation*, Thessaloniki, Greece, October 2002.

(19) The list of 367 explicitly allowed substances was downloaded from the Web site of the Canadian Centre for Ethics in Sport. http://www.cces.ca/pdfs/CCES-PUB-SubstanceClassification-E.pdf. (accessed Mar 27, 2006).

(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(21) Leach, A. R.; Gillett, V. J. *An Introduction to Chemoinformatics*; Kluwer Academic Publishers: Dordrecht, Netherlands, 2003.

(22) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(23) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.

(24) Burbridge, R.; Buxton, B. An Introduction to Support Vector Machines for Data Mining, YOR12 Conference, Nottingham, U.K., 2001. http://www.orsoc.org.uk/conf/previous/yor12/streams/yor12_data_mining.htm. (accessed Jan 9, 2006).

(25) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1947−1958.

(26) R Development Core Team R. A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria, 2005; ISBN 3-900051-07-0. http://www.R-project.org (accessed June 14, 2006).

(27) Itskowitz, P.; Tropsha, A. k Nearest Neighbors QSAR modeling as a variational problem: Theory and applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.

(28) Bender, A. Ph.D. thesis, University of Cambridge, U.K., 2005.

(29) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−424.

(30) Burset, M.; Guigo, R. Evaluation of gene structure prediction programs. *Genomics* **1996**, *34*, 353−367.

(31) Duckworth, F. C.; Lewis, A. J. A fair method for resetting the target in interrupted one-day cricket matches. *J. Oper. Res. Soc.* **1998**, *49*, 220−227.

(32) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536−1548.

CI0601160