PREDICTION OF CHROMATOGRAPHIC RETENTION TIME IN HIGH RESOLUTION ANTI-DOPING SCREENING DATA USING ARTIFICIAL NEURAL NETWORKS

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Analytical Chemistry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Miller, Thomas; King's College London, Analytical &amp; Environmental Sciences Musenga, Alessandro; King's College London, Analytical &amp; Environmental Sciences Cowan, David; King's College, Forensic Science and Drug Monitoring Barron, Leon; Kings colledge London, Forensic &amp; Analytical Sciences</td>
</tr>
</tbody>
</table>
PREDICTION OF CHROMATOGRAPHIC RETENTION TIME IN HIGH RESOLUTION ANTI-DOPING SCREENING DATA USING ARTIFICIAL NEURAL NETWORKS

Thomas H. Miller, Alessandro Musenga, David A. Cowan, Leon P. Barron*

Analytical & Environmental Sciences Division, School of Biomedical Sciences, King’s College London, 150 Stamford Street, London, SE1 9NH, United Kingdom.

*Corresponding author

Email: leon.barron@kcl.ac.uk

Tel: +44 20 7848 3842

Fax: +44 20 7848 4980

Keywords: artificial neural networks; anti-doping; high resolution mass spectrometry; ultra-high pressure liquid chromatography; retention time
Abstract

The computational generation of gradient retention time data for retrospective detection of suspected sports doping species in post-analysis human urine sample data is presented herein. Retention data for a selection of 86 compounds included in the London 2012 Olympic and Paralympic Games drug testing schedule were used to train, verify and test a range of computational models for this purpose. Spiked urine samples were analysed using solid phase extraction followed by ultra-high pressure gradient liquid chromatography coupled to electrospray ionisation high-resolution mass spectrometry. Most analyte retention times varied \( \leq 0.2 \) min over the relatively short runtime of 10 min. Predicted retention times were within 0.5 min of experimental values for 12 out of 15 blind test compounds (largest error: 0.97 min). Minimising the variance in predictive ability across replicate networks of identical architecture is presented for the first time along with a quantitative discussion of the contribution of each selected molecular descriptor towards the overall predicted value. The performance of neural computing predictions for isobaric compound retention time is also discussed. To the authors’ knowledge, this work is the first to present the development and application of neural networks to the prediction of gradient retention time in archived urine analysis sample data in the field of anti-doping.
Introduction

Drug testing in sport is becoming more and more challenging, since cheats rapidly switch to new or designer drugs to avoid being caught by anti-doping tests. Because of this, in 2011 the World Anti-Doping Agency introduced a new section in the Prohibited List to specify that any substance not approved for human therapeutic use is prohibited. Unfortunately however, there is normally a delay between new substances being used by cheats and the screening methods being implemented to detect them.

New technologies in analytical chemistry now exist which offer a much more flexible and universal approach compared to targeted-type screening analyses. In particular, liquid chromatography with full data capture high resolution mass spectrometry (LC-HRMS) is becoming very popular for drug screening, since it generally permits the addition of new analytes to a screening method with little/no modification to the method. One of the most promising features is the possibility for retrospective analysis of post-acquisition datasets. The flexibility of the HRMS approach was exploited in our laboratory during the London 2012 Olympic and Paralympic Games, where an LC-HRMS screen was employed for the detection of nearly 200 analytes within a single assay. When permitted, in accordance with the International Standard for Laboratories, all sample data acquired by this technique could be re-processed and re-evaluated (for example, where intelligence becomes available about the potential use of new substances). However, several challenges still exist for retrospective detection of drug species in both new and archived samples. Analytical reference standards are not always available, especially for new or analogous compounds which may be synthesised relatively easily and with sufficient efficacy for
practical use by athletes. Moreover, the exact operating conditions of analytical
equipment may not be reproducible at a later date when a repeat analysis is required.
For example, chromatograms generated from separate analyses even using the same
columns may differ due to phase ageing and loss of performance. Therefore, more
effort is required to interpret the original data.

Predictive computing techniques, such as artificial neural networks (ANNs), can
interpret underlying trends in complex datasets by learning from case examples. There
are several types of ANN, but the feed-forward multi-layer type is commonly the
most used across a range of scientific applications such as analytical chemistry,
environmental science and molecular biology.6–10 By interconnecting a set of input
data with a series of hidden layer neurons, statistical weights and biases between them
are systematically optimised towards producing a minimised error overall output. In
the training phase, the ANN requires a known true value as a comparator and once an
acceptable number of training cycles (or epochs) is determined, the optimised ANN
can be used to predict the same output where experimentally-derived data is
unavailable (i.e. a blind test). In this case, information about a set of analytes or
chromatographic system (as inputs) and the experimentally-derived retention data for
standards within spiked, matrix-matched samples (as an output) could provide a
useful body of information for an ANN to predict retention time retrospectively under
the conditions observed at the time.

Research to characterise retention behaviour has used linear solvation energy
relationships (LSERs). These use solute descriptors and linear regression modelling to
predict a retention parameter for a particular set of compounds. Both algorithm-based
and predictive ANN approaches have incorporated the use of LSERs as inputs to
estimate retention (usually as retention factor, $k$) across a number of chromatographic modes,\textsuperscript{11-14} column formats,\textsuperscript{15-17} for method optimisation purposes,\textsuperscript{18-22} to predict retention behaviour across column formats, or to estimate the retention of unknowns.\textsuperscript{23-25} Each LSER solute descriptor must be calculated and each coefficient is usually obtained via measured retention data for a set of representative compounds. Thus, the time to produce an accurate LSER model can be significant (as well as especially challenging for ionisable compounds in particular\textsuperscript{15,18}). As LSER coefficients for retention modelling also depend on observed retention on a stationary phase, factors such as the age of the phase may lead to observable differences and lower accuracy over time. Furthermore, the exact composition of a changing mobile phase under gradient conditions is very difficult to characterise and requires the inclusion of extra variables such as purity and individual system dwell time for example. Whilst the investigation of LSERs in retention modelling advances, ANN-based predictive approaches using alternative descriptors may therefore hold promise for more immediate application.

The aim of this work was to investigate the use of ANNs for the prediction of retention time in archived urine analysis data from the London 2012 Olympic and Paralympic Games using a method employing ultra-high pressure liquid chromatography coupled to HRMS. In particular, the selection of alternative molecular descriptor data to those used in LSERs is presented, along with a study of the ANN dependency on these descriptors and the variance across replicated ANNs. Lastly, this work aimed to assess whether ANNs could be used to discriminate structurally similar and/or isobaric species.
Experimental

Reagents

For a detailed account of all reagents used in the experimental please see the accompanying supplementary information (SI), Table S2. For those compounds where no certificate of analysis was available, a solution of each compound was infused separately into the HRMS and full scan data acquired without any applied collisionally induced dissociation. The internal calibrator was caffeine. Identity was considered confirmed where m/z values within 2 ppm of the theoretical value were achieved along with a consistent fragmentation pattern.

Urine sample preparation

Drug-free urine samples were collected from 15 healthy volunteers, anonymised, divided into 1 mL aliquots and stored at -10 °C until analysis. Each aliquot was then spiked with all reference compounds. Following addition of internal standards and glucuronide hydrolysis reagents (see SI, 2.0), formic acid was added to the samples. This solution was then extracted on Bond-Elut Plexa PCX (60 mg, 3 mL barrel) solid phase extraction (SPE) cartridges (Agilent Technologies, Lake Forest, CA). Cartridges were pre-conditioned with 0.5 mL of methanol and 0.5 mL of 2 % v/v formic acid. After loading, cartridges were washed with 1 mL of each of 2 % formic acid, water and 20:80 v/v methanol:water. Elution was performed in 3 mL 3 % (v/v) ammonium hydroxide in methanol:acetonitrile (50:50, v/v). Extracts were dried under nitrogen at 60 °C and then reconstituted to 100 µL of 0.3 % v/v formic acid in 95:5 v/v water:acetonitrile.
Instrumental conditions

For all separations, a Waters Acquity UPLC® ultra-high pressure liquid chromatographic system was used (Waters, Milford, MA, USA). Separations were performed on a Waters Acquity BEH-C\textsubscript{18} column (2.1 × 50 mm, 1.7 µm). Gradient elution was performed as follows: 95:5 water:acetonitrile (both in 0.3 % formic acid) for 0.5 min; then to 80:20 for 3.0 min; a linear ramp to 75:25 for 2.0 min; to 43:67 for 1.5 min; and finally to 10:90 for 1 min. Total time for the gradient elution was 8 min, followed by 2 min for re-equilibration. Separations were performed at 0.3 mL/min and at 30 °C throughout. The injection volume was 10 µL. The syringe and injector were washed sequentially with 0.2 % formic acid in acetonitrile/water (30:70, v/v) and methanol/acetonitrile (90:10, v/v) up to 5 times each before every run to avoid carryover.

Detection was performed using fast polarity switching high resolution mass spectrometry on an Exactive instrument (Thermo Fisher Scientific, San Jose, USA) equipped with a heated electrospray ionisation (HESI-II) source. Enhanced resolution mode was employed at 25,000 FWHM resolution. Three events occurred during each acquisition cycle by performing a full scan both in positive and negative ionisation mode (both with disabled CID) followed by a full scan in positive ionisation mode only with CID (HCD collision energy 30 eV). For a full description of other HRMS conditions, see the SI.

Molecular descriptors, neural network prediction models and architectures

Eighteen molecular descriptors were generated for each compound in the optimised network. Of these, 15 descriptors were computed using Parameter Client freeware
(Virtual Computational Chemistry Laboratory, Munich, Germany) using canonical simplified molecular-input line entry system strings (SMILES). Descriptors including $pK_a$, Ghose-Crippen and Moriguchi log $P$ (AlogP or MlogP), number of double bonds (nDB), number of four to nine-membered rings (nR04-nR09) and number of carbon or oxygen atoms (nC or nO) were used for network optimisation. A full list of other descriptors is detailed in Table S1 and Figure 3. To investigate the use of predicted $pK_a$ values, Percepta PhysChem Profiler (ACD Labs, ON, Canada) software was used. All predicted retention times ($t_r^p$) were performed using licensed ANN software (Trajan Software Ltd., Lincolnshire, UK) and compared with experimentally determined retention time ($t_r^E$). A selection of ANN types and architectures were investigated including linear, radial basis function (RBF), probabilistic neural networks (PNN) and multi-layer perceptrons (MLPs).

**Results and Discussion**

**Experimentally determined chromatographic retention time and reproducibility**

This screening method was used during the 2012 Olympic Games for the detection of nearly 200 compounds in our ISO 17025 accredited laboratory. All of these species were detectable at concentrations corresponding to 50 % of the WADA minimum required performance level$^{26}$ in force at that time in spiked, drug-free urine. On average, peak widths at baseline were of the order of ~0.15 min with some notable exceptions. For example, peak widths for morphine and etilefrine were 0.3 and 0.4 min respectively, with some variability between urine samples. In order to examine the potential usefulness of computational tools for the retrospective generation of retention times in archived data, the reliability of generated
chromatographic retention data was first evaluated. An in-depth evaluation of other method validation experiments including selectivity, sensitivity and limits of detection is given elsewhere. The evaluation of retention time reproducibility was performed using 15 different urine samples analysed in groups of 5 with groups being analysed on different days. Rather than expressing the reproducibility as a relative standard deviation, the maximum variation in retention time was preferred, since the aim was to establish a suitable detection window. The maximum within-day $t_r^E$ variability was 0.22, 0.23 and 0.32 min for the 3 groups of 5 urine samples respectively. Considering all 15 urine samples together (3 groups, 3 different days), the maximum retention time variability was 0.35 min (i.e. ±0.175 min). Data analyses during screening were normally performed by reviewing extracted ion chromatograms with an m/z window of ±5 ppm and a retention time window of ±0.5 min. Athlete urine samples were always run using a “bracketed” approach to take into account phase ageing and to ensure system performance throughout the entire batch acquisition.

Selection of molecular descriptors, network type and architecture

By generating SMILES strings for each compound, an initial set of >200 diverse molecular descriptors were generated. Whilst the instrumental screening method incorporated a larger number of compounds in practice, the number studied here was limited to those compounds with literature reported, experimentally-derived $pK_a$ values (as this descriptor was not initially available within the Parameter Client). Given that the majority of these compounds were fully/partly ionised in the mobile phase, this descriptor was considered important for the potential prediction of retention times. Multiple network types and architectures were examined. On the whole, it was observed that use of a larger numbers of inputs did not offer advantages.
in terms of prediction accuracy and a much smaller set was ultimately more practical and yielded better correlations. As a result, some descriptors were not included in the model which might have otherwise been expected under reversed-phase chromatography conditions such as analyte topological surface area, number of atoms containing lone pairs of electrons (e.g. oxygen or nitrogen atoms), or molecular weight.

Whether using a large or small input descriptor set, a linear-type network offered no detectable correlation ($R^2$ and slope both <0.1 in all cases) between $t_r^E$ and $t_r^P$ and irrespective of whether inputs were used for training, verification or blind testing. As such, this network type was removed from consideration. Alternative network types on the other hand offered better performance and the best correlation coefficients achieved for each type ranged from $R^2=0.86$ (RBF) to 0.98 (4-layer MLP) with the slope ($m$) ranging from 0.97 (3-layer MLP) to 1.03 (PNN), again for all 86 compounds. The best network shown in Figure 1(a) was a feed-forward, back propagation-type MLP with 2 hidden layers of 5 and 4 nodes respectively. This 18:5:4:1 MLP architecture was trained using 18 molecular descriptors using 61 compounds. Verification was performed with 10 compounds and training of the network was ceased when a minimum for residual error was observed (2,000 epochs here). The network was “blind” tested using 15 additional compounds displaying structural variance and a spread of input data and $t_r^E$ data across the 8 min runtime. Data for $t_r^E$ was not used by the ANN for test compounds (see Table S2) and performance was observed after training and verification. When the dataset was divided into training, verification and blind testing data, their respective $m$ and $R^2$ values remained satisfactory and were also better than the corresponding plots for
PNN, RBF and 3-layer MLP networks. By examining the residual errors in the entire dataset (Figure 1(b)), it was observed that retention time could be predicted within 0.5 min of $t_r^E$ for 80 out of 86 compounds (~93 %). This window was initially selected to cover the window employed during experimental screening outlined above. In particular, $t_r^P$ for 97 %, 90 % and 80 % of the training, verification and test compounds respectively lay within this window. In fact for the test set alone, this level of accuracy was maintained even to ±0.3 min of $t_r^E$. Therefore, these results show that this type of network offered promising predictive ability for most compounds in a relatively rapid gradient separation. The largest residual errors were observed for MDA, clenbuterol and $p$-methylamphetamine at -0.947 min (used within the test set), -0.721 min (in verification set) and -0.704 min (in training set) respectively. Therefore, all species could be predicted within 1 min of $t_r^E$. However, it should be noted that this ANN was specifically tailored for this chromatographic system and method. If applied to other separation systems and modes, it is likely that similar descriptor and ANN optimisation experiments would be necessary.

Consistency in $t_r^P$ using replicate networks of identical type and architecture

One of the main issues with ANNs is that training follows a random process towards minimising the overall error. However at larger numbers of training epochs, networks may become over-trained and whilst training set errors may be minimised, the verification set error can increase (as it is not used by the software itself during the training process). The user often manually defines where training ceases by observing where the combined verification set residual error is at its lowest. In some cases this verification error is lowest at lower numbers of training epochs, but where training errors may remain high. Therefore, the variability in network connection weightings
may be very different from network to network (even with identical architectures) and a balance is required potentially towards achieving some consistency. In this case, replicate networks (n=10) using the optimised architecture were built to examine consistency and the range of $t_r^P$ values for the verification and test sets were plotted as Figure 2. Also shown alongside each $t_r^P$ entry in Figure 2 is the $t_r^E$ variance of the analytical method for comparison purposes. By stopping training between 2,000-3,900 epochs in each case (where verification and training error combined were lowest), it was found that relatively consistent input weightings were achieved and $t_r^P$ performance could be maintained. To our knowledge, this has not been shown previously in predictive chromatographic retention studies. However, and despite their accuracies and consistencies, $t_r^E$ data for 6/15 blind test set compounds still lay outside the range of $t_r^P$ values generated by all 10 networks. Of these, and of particular interest, were three structurally-related species: dimethylamphetamine, MDA and phentermine. As synthetic compounds of this general type often only have slight variations in their molecular structure, this approach could be very useful to predict retention time for an array of related species where a reference standard is unavailable. The average $t_r^P$ errors across all 10 networks for these three compounds were +0.97, -0.62 and -0.24 min ($t_r^E$=2.76, 2.50 and 2.91 min) respectively. The lowest respective errors in $t_r^P$ achieved by any one network were +0.59, -0.57 and -0.07 min. By comparison, and upon inspection of the average $t_r^P$ error for other related species within the training set, compounds such as MDMA, benzphetamine, fenfluramine and methamphetamine all performed reasonably well (average absolute error of 0.08-0.18 min from $t_r^E$ for all 10 networks). However, and as noted earlier, average $t_r^P$ error for $p$-methylamphetamine was the worst case in the training set at an average of -0.44 min across all networks. In the verification set, only one related
compound, amphetamine, yielded 0.13 min error in $t_r^P$ ($t_r^E=2.37$ min). Therefore, these figures suggest that dimethylamphetamine and MDA may be unusual cases and that ANN predictive ability for unknowns may not, on average, be poorer for these types of compound. With regard to the general test set, all verification and test compound retention times could be predicted within 1 min of $t_r^E$ (with average±standard deviation of $t_r^P$ error for all 25 compounds = 0.3±0.2 min for all 10 networks). In comparison to the maximum measured experimental variance above (0.35 min), the potential for using multiple networks to predict retention time shows promise. Therefore, the window for ANN accuracy could be set at ±0.5 min in total (again, also comparable to the experimental retention time window used in practice during the Olympic & Paralympic Games). Each network required a few minutes to build, train, verify and test. Therefore, this *in silico* approach could represent a significant saving in terms of time, effort and cost when used in combination with a semi-targeted post data acquisition analysis of an athlete urine sample. Samples from the Olympic Games are stored for a total of 8 years in case re-analysis is required at a later date. However, this is not the case for most other sporting events where negative urine samples are discarded after 3 months unless probable cause for prolonged storage is justified. As several isobaric compounds can exist, HRMS data alone may be considered insufficient for this purpose. Therefore, the benefits of ANN predictions could aid in the investigation of doping before analysis, before the reporting stage and after the reporting stage. When used in combination with HRMS, this gives added value especially in directing which analytical reference materials are to be synthesised.

*Molecular descriptor contribution towards $t_r^P$ and variance across replicate networks*
Using the optimised MLP network, the contribution of each descriptor towards the generation of $t_r^P$ was calculated for a single compound, $A$, by determining the absolute deviation, $D_A$, arising from sequential removal of one descriptor, $i$, at a time.$^{27}$ Data were inputted into Equation 2 below to calculate the overall variance, $V_A$, arising for that compound. Values for $V_A$ for each compound (denoted $A=1-61$) were then used to determine the overall percentage change in accuracy of predicted retention time for all compounds, $\Delta t_r^P(\%)$, for the removal of that descriptor according to Equation 3.

\[
V_A = \frac{(D_A)_i}{\sum_{i=1-18} (D_A)_i} \quad (2)
\]

\[
\Delta t_r^P(\%) = \frac{100V_A}{\sum_{A=1-61} V_A} \quad (3)
\]

The contribution of each descriptor towards $t_r^P$ was different. Using the optimised network, the top 5 contributing molecular descriptors (in decreasing order) to $t_r^P$ for all compounds using the best network were nC (17.5 % change in $t_r^P$ when removed), nR09 (10.0 % change), nO (9.9 % change), nDB (8.8 % change) and ALogP (8.2 % change). Overall, some of these $\Delta t_r^P$ values were unsurprising given that silica-C$_{18}$ was used as a stationary phase. For experimentally-derived descriptors however, removal of pK$_a$ only resulted in a change of 3.8 %, but subsequent removal of AlogD and MlogD resulted in $\Delta t_r^P$ of $\leq$3 % each. As expected, descriptors such as the number nR04 having little or no incidence resulted in small $\Delta t_r^P$ when removed. Examination of the data relative to $t_r^E$ revealed that predictions were on the whole worse with the removal of selected descriptors.

To further understand the variability across replicate networks, this experiment was repeated for 9 additional replicate networks. The contribution and variance of each descriptor is shown as Figure 3. Taking the mean values, the top 5 contributing
molecular descriptors were the same as reported above, albeit in a slightly different order (from highest to lowest: nC, nDB, nO, AlogP and nR09). This indicated that ANNs could be replicated with acceptable consistency in their operation. Interestingly, the largest measured variance in $t_r^p$ across replicate ANNs for removal of a single descriptor was also observed as the most highly contributing descriptor, nC, at ~11%. Removal of descriptors which resulted in the lowest % $t_r^p$ variance were also those which were used least by ANNs during prediction. However for the rest of the descriptor set, no trend was apparent to conclude that a more heavily used input would result in a higher variance when omitted from replicate networks. Furthermore, it was observed again that AlogP was repeatedly used in preference to MlogP in $t_r$ predictions. MlogP has been shown previously to be a less accurate descriptor in comparison to other available computational models and in respect to experimentally-derived logP. Here, it was observed that MlogP data offered poorer discrimination between some species and especially those of isobaric mass. It was also apparent that this had a consequential effect on the use of AlogD over MlogD as a preferred input descriptor of the two. However, as a null % change was not observed by removing MlogP from the dataset in any network, it was still used to some degree by ANNs.

Substitution of experimentally-derived $pK_a$ with predicted $pK_a$

In practice, retention on reversed-phase media is likely to be heavily dependent on analyte logP and $pK_a$ (and subsequently logD). The molecular description software used here could generate predicted analyte logP, but not $pK_a$. The initial use of experimentally-derived $pK_a$ data from the literature overcame this problem and resulted in satisfactory $t_r^p$ accuracy as shown earlier. However, the availability of (or indeed the ability to generate) experimental $pK_a$ significantly limits the usefulness of
this ANN approach if it is to be used in post-acquisition data mining. The use of a separate pKₐ prediction software package was therefore investigated as a possible alternative. In an initial experiment, experimentally-derived pKₐ data were removed from the input dataset and replaced directly with predicted values into the pre-trained ANN. With this substitution, a very slight reduction in tᵣᵣ accuracy was observed, but it still remained acceptable (not shown). The network in Figure 1(a) generated an overall tᵣᵣ / tᵣₑ correlation with a slope of 1.00 (R² = 0.98) compared with 0.95 (R²=0.98) in this case across all 86 analytes. This obviously indicated some minor discrepancies between numerical values in these datasets. To further understand the importance of predicted versus experimentally-derived pKₐ as ANN inputs, a new network of identical architecture was retrained using computationally-derived pKₐ values to observe whether prediction accuracy could be restored. From this, it was seen that tᵣᵣ accuracy was marginally worse using computed pKₐ (Figure S3), but still remained acceptable overall (overall slope=0.94; R²=0.95). In previous work, Livingstone also showed that predicted pKₐ data can display inaccuracies up to 1.5 log units (using a larger compound set than here). Thus, in ANN predictions of test compound retention time, experimentally-derived data should be used where possible, but predicted values could potentially be substituted with awareness that accuracy may be slightly affected.

Performance of ANNs for isobaric and isomeric compounds

If the ANN approach is to work in any realistic scenario with HRMS detection, then tᵣᵣ accuracy should be high for compounds which are both isobaric and which display high degrees of structural similarity. From the list of compounds used in this study, several such cases existed (Table 1). Nadolol and metipranolol, both non-selective
beta-blockers, display identical mass and molecular formula, but differ slightly in their molecular structure and logP. Experimentally, this translated to a two-fold difference in their retention times on this C_{18} phase. In this case, the optimised ANN performed well by predicting retention within 0.4 min for both compounds despite their high structural and pK_{a} similarity. Differences in logP could have accounted for this performance, as it was used here in the calculation of logD. Hydrophilic factor was calculated based on the number of atoms in a molecule, the number of carbon atoms and the number of hydrophilic groups (such as –OH or –NH for example). Unsaturation indices are calculated based on the number of atoms containing proton deficiencies (e.g. aromatic rings or multiple bonds). Therefore, these latter two descriptors could be considered loosely correlated with logP on a fundamental level. Therefore, it might be somewhat unsurprising that this resulted in higher accuracy in this case given the nature of the separation mode.

In a second similar case involving the isobaric compounds, phentermine and methamphetamine t^p_{r} accuracy remained high (±0.22 min from t^E_{r}), but the order of elution was incorrectly predicted. Furthermore, t^p_{r} for a third isobaric compound, p-methylamphetamine was predicted as similar to that of phentermine and methamphetamine, but was less accurate to its t^E_{r} at -0.7 min. All three of these compounds had identical MlogP data, but only slightly differing AlogP and pK_{a} values. For the majority of the other 13 compounds eluting between methamphetamine and p-methylamphetamine and where descriptor diversity was more pronounced, this was less of a problem and t^p_{r} inaccuracy remained <0.5 min.
As a third case example, a set of diastereoisomers were chosen: ephedrine and pseudoephedrine. With identical mass, molecular formula, atomic connections, AlogP and MlogP, it was somewhat unsurprising that they remained experimentally unresolved on this stationary phase. Whilst their pKₐ values were 0.3 units apart, and therefore a little further apart than of those species in the previous example, this had little effect on $t_r^{P}$ (within 0.15 min of $t_r^{E}$ for both species). Whilst further work on the use of ANN using resolved diastereoisomers is required, it should be noted that the SMILES strings used here employs a string of characters that specified the structure of a selected molecule including its spatial arrangements. Optical isomerism in SMILES strings is designated by the ‘@’ symbol. Where a single ‘@’is used, this represents the substituent groups giving rise to the chirality being positioned anti-clockwise around the chiral atom whereas two symbols (‘@@’) indicates clockwise orientation of the groups around the chiral atom. Therefore, as this may lead to a difference in molecular description, then this could be used by the ANN to differentiate such species. Naturally, this would also depend on the importance of the molecular descriptor itself to the generation of $t_r^{P}$ (similar to Figure 3).

**Conclusion**

This work showed that ANNs could be used to predict chromatographic retention times for 93 % of all selected doping-related compounds to within 0.5 min of their true value and to within 1 min for all other compounds. All compounds were detected in a urine extract matrix and were separated under gradient elution conditions. When applied to the prediction of unknowns alone, the same level of accuracy was maintained. Variance across replicate networks of identical architecture revealed that descriptors were used to similar degrees towards prediction of retention time. This
approach could be used with quantitative structure-activity relationships alone, but it is recommended that analyte $pK_a$ should be experimentally derived for ANN training. Ultimately, prediction of retention times in archived screening data would simplify and aid data reprocessing as a complementary tool to retrospective analysis both in the identification of unknowns and where reference materials were not originally included in the analytical screen. Therefore, the combination of full data capture HRMS and \textit{in silico} predictive approaches could improve the capability for semi-targeted urine sample screening before, during and after major sporting events.

**Acknowledgements**

The authors would like to thank Dr David Barlow for providing constructive comments on earlier drafts of this manuscript.

**Supplementary Information**

Tables containing the raw data relating to experimental and predicted retention times for all compounds and experiments herein; molecular descriptors and associated ANN input data; and all ANN models and architectures. This material is available free of charge via the Internet at [http://pubs.acs.org](http://pubs.acs.org).

**References**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
Table 1. Performance of ANNs for predictions of $t_r$ for isobaric and isomeric compounds.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Monoisotopic Mass (Da) (Measured $m/z$ for [M+H]$^+$ ion)</th>
<th>Molecular Formula</th>
<th>pK_a $^{ref}$</th>
<th>$t_r^i$ (min)</th>
<th>$t_r^p$ (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metipranolol</td>
<td>309.1940 (310.2013)</td>
<td>C$<em>{17}$H$</em>{27}$NO$_4$</td>
<td>2.86 (2.35)</td>
<td>9.18$^{xy}$</td>
<td>5.62 6.02</td>
</tr>
<tr>
<td>Nadolol</td>
<td>309.1940 (310.2013)</td>
<td>C$<em>{17}$H$</em>{27}$NO$_4$</td>
<td>1.15 (1.36)</td>
<td>9.67$^{xy}$</td>
<td>2.81 2.52</td>
</tr>
<tr>
<td>Phentermine</td>
<td>149.1205 (150.1277)</td>
<td>C$<em>{10}$H$</em>{15}$N</td>
<td>1.84 (2.55)</td>
<td>10.1$^{xy}$</td>
<td>2.91 2.69</td>
</tr>
<tr>
<td>Methamphetamine</td>
<td>149.1205 (150.1277)</td>
<td>C$<em>{10}$H$</em>{15}$N</td>
<td>2.07 (2.55)</td>
<td>9.87$^{xy}$</td>
<td>2.59 2.80</td>
</tr>
<tr>
<td>$p$-methylamphetamine</td>
<td>149.1205 (150.1277)</td>
<td>C$<em>{10}$H$</em>{15}$N</td>
<td>2.12 (2.55)</td>
<td>10.0$^{xy}$</td>
<td>3.45 2.75</td>
</tr>
<tr>
<td>Ephedrine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMILE: O[C@H][c1]ccc1)<a href="NC">C@@H</a> C</td>
<td>165.1154 (166.1226)</td>
<td>C$<em>{10}$H$</em>{15}$NO</td>
<td>1.24 (1.66)</td>
<td>9.60$^{xy}$</td>
<td>2.00 1.84</td>
</tr>
<tr>
<td>Pseudoephedrine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMILE: O[C@@H][c1]ccc1)<a href="NC">C@@H</a> C</td>
<td>165.1154 (166.1226)</td>
<td>C$<em>{10}$H$</em>{15}$NO</td>
<td>1.24 (1.66)</td>
<td>9.80$^{xy}$</td>
<td>2.00 1.85</td>
</tr>
</tbody>
</table>
Figure 1. (a) Correlation between experimentally-derived ($t_{R}^{E}$) and predicted ($t_{R}^{P}$) retention times using the optimised 18:5:4:1 multilayer perceptron (inset) trained for 2000 epochs; and (b) residual errors in $t_{R}^{P}$ using the optimised network for all analytes ($n = 86$). All raw data represented in Table S2 in the SI.
Figure 2. Paired comparisons of measured $t_r^E$ variability (left-hand box) from $n=15$ volunteer urine data versus $t_r^P$ variability for $n=10$ replicate 18:5:4:1 MLP networks (right-hand box). Data is presented for the verification (marked V) and blind-test compounds (marked T) only. Boxes represent 25-75th percentile, whiskers represent 10-90th percentile and dots are outliers. Thin lines represent the median and thick lines represent the mean. *Experimental variance not determined for benzthiazide and polythiazide and reported values represent a single measurement. For all raw data (including for the training set) please refer to Tables S3 & S4 in the SI.
Figure 3. Percentage change in $t_r^P$ relative to $t_r^E$ upon systematic removal of each molecular descriptor from 10 replicate 15:5:4:1 MLP networks. Boxes include data from the 25th-75th percentile, the median (thin line), and the mean (thick line). Error bars include the 5th and 95th percentile and dots represent outliers. For all molecular descriptor definitions, please see experimental section or Table S1 in the SI.