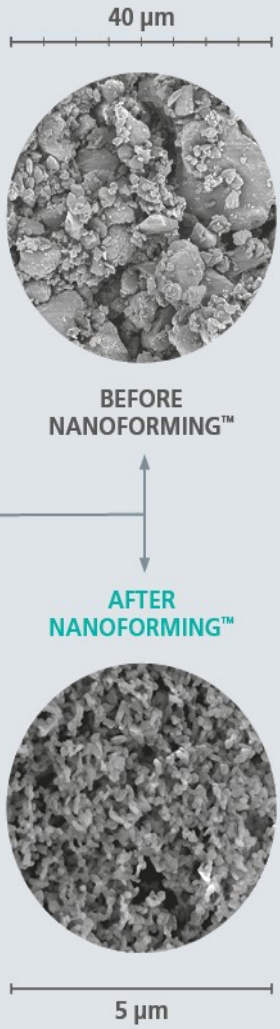
DRUG PARTICLES SMALLER THAN EVER BEFORE



Working with you to enhance drug effectiveness and targeting

Our global experts in nanotechnology and drug particle engineering can help your drugs reach their full therapeutic potential. With the unique ability to produce nanoparticles as small as 10 nm, our award-winning Controlled Expansion of Supercritical Solutions (CESS®) technology can increase dissolution rates and improve bioavailability. Together, we can initiate a new era of novel drug development and enable more patients around the world to benefit from next-generation drug therapies.

nanoform
small is powerful

Contact Nanoform to unlock the potential of your molecules +358 29 370 0150

nanoform.com | info@nanoform.com

@NanoformF Nanoform

Klingberg Joshua (Orcid ID: 0000-0003-4310-4777)

Cawley Adam (Orcid ID: 0000-0002-3442-8617)

Fu Shanlin (Orcid ID: 0000-0002-6238-3612)

Towards Compound Identification of Synthetic Opioids in Non-targeted Screening Using Machine Learning Techniques

Joshua Klingberg[a], Adam Cawley[b], Ronald Shimmon[a] and Shanlin Fu[a]*

[a]*Centre for Forensic Science, University of Technology Sydney, Broadway, NSW 2007, Australia*

[b]*Australian Racing Forensic Laboratory, Racing NSW, Sydney, NSW 2000, Australia*

*Corresponding author. Tel: +61 2 9514 8207. Email: shanlin.fu@uts.edu.au

The constant evolution of the illicit drug market makes the identification of unknown compounds problematic. Obtaining certified reference materials for a broad array of new analogues can be difficult and cost prohibitive. Machine learning provides a promising avenue to putatively identify a compound before confirmation against a standard. In this study, machine learning approaches were used to develop class prediction and retention time prediction models. The developed class prediction model used a Naïve Bayes architecture to classify opioids as belonging to either the fentanyl analogues, AH series or U series, with an accuracy of 89.5%. The model was most accurate for the fentanyl analogues, most likely due to their greater number in the training data. This classification model can provide guidance to an analyst when determining a suspected structure. A retention time prediction model was also trained for a wide array of synthetic opioids. This model utilised Gaussian Process Regression to predict the retention time of analytes based on multiple generated molecular features with 79.7% of the samples predicted within ± 0.1 min of their experimental retention time. Once the suspected structure of an unknown compound is determined, molecular features can be generated and input for the prediction model to compare with experimental retention time. The incorporation of machine learning prediction models into a compound identification workflow can assist putative identifications with greater confidence and ultimately save time and money in the purchase and/or production of superfluous certified reference materials.

Keywords: synthetic opioids; illicit drugs; machine learning; HRMS

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/dta.2976

1. Introduction

Synthetic opioids are of significant concern to society due to their large public health threat. The potential of opioids to be used as performance-enhancing drugs in horses also raises concerns in the equine anti-doping community [1,2].

Recently, there has been an increasing focus on non-targeted analysis methods to detect illicit drugs in high-throughput screening applications, without needing to rely on certified reference materials (CRMs) or library databases [3]. Previous studies have investigated the use of product ion searching to detect novel analogues of known drug classes, exploiting class-specific cleavages giving rise to common product ions [1,2,4,5]. Other studies have looked at more advanced data analysis methods, incorporating mass filtering techniques such as Kendrick Mass Defect analysis to detect structurally related compounds within a complex matrix [1,6]. While these data analysis techniques are useful for detecting the presence of an unknown compound within a sample, further investigation is required to determine the identity of the analyte. Furthermore, while many vendor software packages, such as *Molecular Structure Correlator* (Agilent Technologies) and *Compound Discoverer* (Thermo Fischer Scientific), can help with this process, having a basic understanding of the general classification of an unknown compound can assist with their timely identification.

With the increasing volume of data generated by high-resolution mass spectrometry (HRMS) and the availability of cheaper and more powerful computational processing [7], the use of artificial intelligence approaches, such as machine learning, for toxicological applications has become viable. In general, machine learning algorithms enable a computer to 'learn' information directly from a data set without needing a predetermined equation to use as a model [8]. In this way, the algorithm can adaptively improve its performance over time with experience [9]. Machine learning can be further divided into both supervised and unsupervised learning. Supervised learning involves using a known set of input data (called the training set) with known responses to that data (the output) to train a model that has the ability to predict the response for new input data [7,10]. Supervised learning models are categorised as either 'classification' or 'regression' models. Classification models aim to predict a discrete output, whereas regression models attempt to predict an output within a continuous space, such as a temperature or time range [7]. Unsupervised learning, on the other hand, is useful when attempting to explore data without specific goals or previous knowledge of the information contained within the data [7,11].

Machine learning approaches have been used in the field of toxicology previously, notably within the context of determining structure activity relationships in drug discovery [12,13]. Retention time prediction under a specific chromatographic system has also been explored using regression models from both an environmental [14] and drug analysis perspective [15,16].

This study aimed to develop a machine learning approach to assist with the identification of unknown compounds indicated from non-targeted screening workflows such as those presented by our group [1]. The development of a classification model to predict the subclass of synthetic opioids based on MS^2 dissociation data is presented, alongside the development of a regression model to predict the retention time of suspected compounds from calculated

molecular features. CRMs of novel drug analogues can be difficult and costly to procure. If an analyst has evidence supporting the putative identification of an unknown compound prior to procuring a CRM, it may save time and money for subsequent confirmation of the drug.

2. Experimental

2.1. Solvents and Reagents

All solvents used were liquid chromatography-mass spectrometry (LC-MS) grade. Acetonitrile, ethyl acetate and methanol were obtained from Merck (Darmstadt, Germany). Ammonium acetate and trichloroacetic acid were obtained from Sigma-Aldrich (Castle Hill, NSW, Australia). Acetic acid and hydrochloric acid were obtained from Ajax Chemicals (Sydney, NSW, Australia). Ultrapure-grade water was obtained from a Smart2Pure ultra-pure water system (Thermo Scientific, Langenselbold, Hungary).

2.2. Drug Standards

Fentanyl citrate was purchased from Sigma-Aldrich (Castle Hill, NSW, Australia). Hydrochloride salts of 3,4-ethylenedioxy U-47700, 3,4-ethylenedioxy U-51754, 3,4-methylenedioxy U-47700,
4-fluoroisobutyrl fentanyl, 4-chloroisobutyryl fentanyl, acetyl fentanyl, AH-7921, AH-8529, AH-8533, meta-fluoro fentanyl, para-fluoro methoxyacetyl fentanyl, phenyl fentanyl, tetrahydrofuran fentanyl, U-48800, U-51754, UF-17 and β-hydroxythio fentanyl, manufactured by Cayman Chemical (Ann Arbor, MI, USA), were purchased from Sapphire Bioscience (Redfern, NSW, Australia). U-62066 mesylate, along with free base standards of AH-7563, AH-7959,
AH-8507, AH-8532, crotonyl fentanyl, isopropyl U-47700, meta-fluoro methoxyacetyl fentanyl, methacryl fentanyl, N-methyl U-47931E, propyl U-47700, senecioyl fentanyl, thiofentanyl,
U-47109, U-47931E, U-48520, U-49900 and U-50488, manufactured by Cayman Chemical, were also purchased from Sapphire Bioscience. Hydrochloride salts of 4-methoxybutyryl fentanyl, acryl fentanyl, butyryl fentanyl, furanyl fentanyl, U-47700, valeryl fentanyl, 4-fluorobutyryl fentanyl, benzyl fentanyl, N-methyl carfentanil, ocfentanil and ortho-fluoro fentanyl, along with MT-45 dihydrochloride hydrate and neat standards of despropionyl para-fluoro fentanyl, benzodioxole fentanyl, cyclopropyl fentanyl, cyclopentyl fentanyl, phenylpropionyl fentanyl and norfentanyl, manufactured by Chiron Chemicals (Hawthorn, VIC, Australia), were purchased from PM Separations (Capalaba, QLD, Australia). Remifentanil hydrochloride was purchased from GlaxoSmithKline (Boronia, VIC, Australia). Citrate salts of carfentanil, sufentanil and α-methyl fentanyl, along with alfentanil hydrochloride, were purchased from Janssen Pharmaceuticals (North Ryde, NSW, Australia). Desipramine-d_3 was purchased from Grace (Columbia, MD, USA). The chemical information for all opioid standards used in this study can be found in the Supporting Information (Table S1).

2.3. Sample Preparation

2.3.1. Class Prediction Samples

Drug standards were obtained as methanolic standards of varying concentrations. Neat standards of all opioids were diluted in methanol to a concentration of 10 µg/mL. Ten µL of each solution was evaporated to dryness under nitrogen, before being reconstituted in one drop of methanol using a Pasteur pipette (approximately 20 µL) and 100 µL of 10 mM ammonium acetate (pH 4) buffer to give a final concentration of approximately 0.8 µg/mL for analysis. All samples were stored at 4 °C until analysis.

2.3.2. Retention Time Repeatability Studies

Neat mixed standards containing 4-chloroisobutyryl fentanyl, 4-fluoroisobutyryl fentanyl, 4-methoxybutyryl fentanyl, acetyl fentanyl, acryl fentanyl, AH-7563, AH-7921, U-47700, AH-7959, AH-8507, AH-8529, AH-8533, α-methyl fentanyl, butyryl fentanyl, carfentanil, cyclopentyl fentanyl, cyclopropyl fentanyl, fentanyl, furanyl fentanyl, meta-fluoro fentanyl, MT-45, remifentanil, sufentanil, U-50488, U-51754 and valeryl fentanyl were prepared in methanol to give a concentration of 1 µg/mL. Sets of 7 samples were prepared, and the study was repeated 7 times
(n = 49) to evaluate the repeatability of absolute retention times without the possibility of matrix effects.

Blank plasma was obtained from blood samples collected in Lithium Heparin Vacutainers purchased from BD (Mississauga, ON, Canada) from four thoroughbred research horses following approval of the Racing NSW Animal Care and Ethics Committee (ARA 71).

Spiked equine plasma samples (2 mL) were prepared to determine the precision of retention time measurements in a relevant biological matrix. Mixed standards containing the same panel of opioids used in the neat standards were spiked into the plasma samples at a concentration of 10 ng/mL. Spiked samples were prepared in sets of 7, and the extraction was repeated 7 times (n = 49) to determine the repeatability of measured retention times. Desipramine-d_3 (10 ng/mL) was added as an internal standard to compare the repeatability of the absolute retention times (RT) to relative retention times (RRT) in comparison to the internal standard. The RRT was calculated by dividing the RT of the opioid standard by the RT of the internal standard.

2.3.3. Retention Time Prediction Samples

Separate plasma samples were spiked with each of 59 different opioid standards at a concentration of 10 ng/mL. Desipramine-d_3 was again added as an IS at a concentration of 10 ng/mL. Triplicate injections were completed from each sample and the extraction was repeated 3 times for a total of 9 measurements for each standard.

2.3.4. Plasma Extraction Method

Protein precipitation was completed through the addition of 200 µL of trichloroacetic acid (10% in H_2O) to 2 mL samples. The pH of the samples was then adjusted to 3 – 3.5 using hydrochloric acid after which they were centrifuged at 1500 g for 10 minutes. Solid phase

extraction (SPE) was completed using XtrackT® Gravity Flow DAU Extraction Columns (UCT Inc., Bristol, PA, USA). The cartridges were first conditioned with 3 mL of methanol, followed by 3 mL of purified water, after which the samples were loaded. The samples were washed with 3 mL of 0.1 M acetic acid and dried under positive pressure. The cartridges were again conditioned with 3 mL of methanol and dried under positive pressure. The analytes were eluted from the cartridges using 3 mL of solvent containing 3% ammonia and 0.5% methanol in ethyl acetate.

Following SPE, one drop of 0.1 M methanolic hydrochloric acid was added to each of the samples using a Pasteur pipette before the solvent was evaporated under a gentle stream of N_2 at 60 °C. The samples were then reconstituted in one drop of methanol using a Pasteur pipette and 100 µL of an ammonium acetate buffer (pH 3.9), before being transferred to vials for analysis. All samples were stored at 4 °C until analysis.

2.4. Instrumental Analysis

Chromatographic separation was achieved on an Agilent Technologies (Santa Clara, CA, USA) 1290 Infinity II UHPLC, consisting of a high-speed pump (G7120A), multisampler (G7167B) and thermostat and column compartment (G1316A, 35 °C) coupled to an Agilent Technologies 6545 quadrupole time-of-flight (QTOF) mass spectrometer. All data acquisition was conducted using Agilent Technologies MassHunter Workstation (Version B.06.01). A sample volume of 1 µL was injected onto an Agilent Technologies Poroshell 120 EC-C18 LC column (2.1 x 75 mm, 2.7 µm particle size) using a gradient elution with a flow rate of 0.4 mL/min and a total analysis time of 12 min. Mobile phase A consisted of 10 mM ammonium acetate buffer (pH 9) and mobile phase B consisted of 0.1% acetic acid in acetonitrile. Initial mobile phase composition was 75% A, which was held for 0.5 min before being decreased linearly to 67% A over 2 min, before being held for another 1 min. The mobile phase was then further decreased to 55% A over 6 min, before being held for 1 min and returned to 75% A over 0.5 min, before a final 1 min hold. The column was allowed to equilibrate for 2 min before the next sample was injected.

The QTOF was operated in positive electrospray ionisation mode (ESI+) with capillary and fragmentor voltages of 3500 V and 100 V, respectively. The QTOF was calibrated each day before use with an average resolution of 16,679 (FWHM) over the period of analysis at a reference mass of *m/z* 322.048121. The calibration solution was made up using 10 mL of the Agilent Technologies ESI-L tuning mix, diluted with 85.5 mL of acetonitrile and 4.5 mL of ultrapure water. Three microliter of 0.1 mM reference mass solution containing HP-0321 was added to the calibration solution. An AutoMS-MS data acquisition mode was used with mass ranges of 100 – 1000 *m/z* for MS and 50 – 850 *m/z* for MS/MS. Spectra were obtained with an acquisition speed of 10 spectra/s for both MS and MS/MS and collision energies (CE) of 10, 20 and 40 eV were used for CID. A maximum of 5 precursors from the MS scan were selected for CID per cycle, with an abundance threshold of 5000 counts. Active exclusion was used, with precursor being excluded after 2 spectra and released after 0.1 min.

2.5. Data Analysis

All data files were analysed using Agilent Technologies MassHunter Qualitative Analysis Software (Version B.10.0, Build 10.0.10305.0) to generate extracted ion chromatograms (EICs) and mass spectra for use in the statistical analysis software.

2.5.1. Class Prediction Samples

MT-45 and UF-17 were excluded from the classification model as they fall within their own classes and the classification model could not reliably account for these compounds when there is only 1 sample in the training set.

The Find by AutoMS-MS function in the qualitative analysis software was used to extract MS spectra from the neat opioid standards at all 3 CEs (10, 20 and 40 eV). The spectra were exported, along with list of the top 10 most abundant masses, for use in building the class prediction model. The most abundant ions for each synthetic opioid standard were then populated into an Excel spreadsheet from highest to lowest abundance.

2.5.2. Retention Time Repeatability and Prediction Samples

Compound and spectral information for all opioid standards analysed was curated into a database using Agilent Technologies MassHunter Personal Compound Database and Library (PCDL) Manager (Version B.08.00, Build 8209.7 SP1). This PCDL was used in conjunction with the Find by Formula (FbF) function in the qualitative analysis software to generate EICs for all the spiked plasma samples analysed.

For the repeatability studies (2.3.2), absolute RT and RRT were collected and populated within a spreadsheet in Microsoft Excel. The average, standard deviation (SD) and relative standard deviation (%RSD) was calculated within each set of 7 samples and between the 7 sets of samples analysed (n = 49).

The individual spiked samples were analysed (2.3.3) and the average RRT across all 9 measurements was determined for use in training the retention time prediction model.

2.6. Molecular Features

Thirteen molecular features were used in the retention time prediction model. Features were calculated using online software, including the Chemicalize tool (ChemAxon, San Diego, CA, USA) [17] and Pharmaceutical Data Exploration Laboratory [18], or predefined equations. A list of all the molecular features used can be found in Table 1. A pH of 9 was chosen for the determination of logD and logS values as this reflected the pH of the mobile phase used for analysis.

The double bond equivalent (DBE) was calculated using Equation 1 [19], where *a* is number of carbons, *b* is the number of hydrogens, *c* is the number of nitrogens and *f* is the number of halogens in the molecule.

$$DBE = (a + 1) - \frac{b - c + f}{2} \qquad \text{(Equation 1)}$$

The hydrophilic factor (*Hy*) was calculated using the Equation 2 [20]. N_{Hy} is the number of hydrophilic groups in the molecule (or the total number of hydrogen atoms attached to oxygen, nitrogen or sulfur atoms), N_c is the number of carbon atoms in the molecule and *A* is the number of non-hydrogen atoms in the molecule [20].

$$H_y = \frac{(1+N_{Hy})\log_2(1+N_{Hy})+N_C\left(\frac{1}{A}\log_2\frac{1}{A}\right)+\sqrt{\frac{N_{Hy}}{A^2}}}{\log_2(1+A)} \quad \text{(Equation 2)}$$

2.7. Statistical Analysis

All statistical analysis was conducted using Microsoft Excel (Version 16.0.12624.20382) and MathWorks MATLAB® (Version R2019b, 9.7.0.1319299) equipped with the Statistical and Machine Learning Toolbox (Version 11.6). The MATLAB® code for the developed models can be found in the Supporting Information.

2.7.1. Class Prediction Modelling

Product ion data from all 3 CEs (10, 20 and 40 eV) were compared. The compiled data can be found in the Supporting Information (Table S2).

The most abundant ion data was imported into the Classification Learner app within MATLAB. The model response variable was set to be the compound class and *k*-fold cross-validation was used with 50 folds. This method of validation splits the data into a finite number of groups, or folds. When training the model, each iteration uses *k*-1 folds as the training data and remaining group as the test data. This process is then repeated for the remaining groups and an average of the accuracies of each iteration is given at the end [21]. This method of validation provides a suitable estimate of the model accuracy [21].

Several different model types, namely decision trees, discriminant analysis, naïve Bayes classifiers, support vector machines, nearest neighbour classifiers and ensemble classifiers, were investigated in order to determine the most accurate model. Hyperparameter tuning was used to optimise the models and provide the highest accuracy possible. The hyperparameters of a model are the variables that control the training process itself. These hyperparameters are set before the model is trained and can be considered the model settings that need to be optimised [22]. The aim of hyperparameter tuning, therefore, is to find the settings that return the best model performance.

Bayesian optimisation was used to train the model with an acquisition function of 'expected improvement per second plus'. The objective of Bayesian optimisation is to build a probability model of the objective function, in this case the classification error of the model, and use it to select the most promising hyperparameters to train the classification model [22]. This optimisation methods learns with each iteration and uses the results from previous trials to determine the best set of hyperparameters to use for the next trial.

The model was trained and optimised over 30 iterations and the Classification Learner app returned the most accurate model from those iterations. In addition to the overall model accuracy, the F1 score and Matthew's correlation coefficient (MCC) were calculated for the

optimised model. The code for the trained model could then be exported for use in classifying new samples.

2.7.2. Retention Time Prediction

The calculated molecular feature data for each synthetic opioid standard was populated into an Excel spreadsheet, along with the average RRT calculated for each compound. The compiled data can be found in the Supporting Information (Table S3). All the compound data was then imported into the Regression Learner app within MATLAB. The model response variable was set to be the RRT and *k*-fold cross validation was again used with 50 folds.

Several regression models were also investigated to determine the most accurate model, namely regression trees, support vector machines, ensemble of trees and gaussian process regression models. Hyperparameter tuning was also used for the regression models with the same parameters as outlined in 2.7.1. To assess the influence of individual predictors, each molecular feature was sequentially removed, and the model retrained without that predictor. This process was completed 3 times for each feature and an average RMSE value was determined for each feature. The code was again exported so that it could be used to classify new samples.

3. Results and Discussion

3.1. Class Prediction Modelling

Previous work has shown that different subclasses of synthetic opioids display class specific diagnostic ions that can be used for non-targeted screening methods [1,2,4]. By exploiting this same phenomenon, class prediction modelling can be attempted using MS^2 data and the presence of abundant product ions within the resultant spectra. The generic structures of the 3 opioid subclasses included in the class prediction models, namely fentanyl analogues, AH series and U series opioids, can be found in Figure 1 below.

For this study, product ion data from all 3 CEs used for analysis was compared to determine the best input data for training the models. It was found that the MS^2 spectra using a CE of 40 eV gave the highest accuracy when developing the class prediction models. This is likely due to the fact that the higher collision energy caused greater dissociation of the compound structure, breaking it down into smaller units which are more closely related to the core structures of the compounds. The common fragmentation patterns of the opioid subclasses have been presented previously [2].

The use of *k*-fold cross validation is beneficial where smaller data sets are being examined. Using the more simplistic hold-out method requires separating the data set into a training and validation sets, which results in a large proportion of the data not being able to be used for the training of the model [23]. The *k*-fold method, on the other hand, allows the use of all data in the training of the model. By randomly separating the training data into different groups, or folds, training and validation can both be performed on all samples within the data set. The choice for how many folds to use in cross-validation is often determined by the computational power available and the variance allowed in the test error calculated [23]. When

the number of folds is close to, or equal to, the number of samples in the data set, there is less bias in the test error that is calculated [23]. This occurs because the model is essentially being trained and tested against each individual sample, which will result in the best estimation of the overall model accuracy. On the other hand, if the number of folds is much smaller than the number of samples, there will tend to be some variance in reported accuracy of the model each time training is completed as the algorithm will more randomly assign samples to each group. If the composition of each validation group changes, the training process and reported accuracy of the model can change as well.

The downside of using a larger number of folds, however, is that it can slow down the training of the model itself. When using *k*-fold cross-validation, the model must be trained *k*-1 times. The more folds that are used, the more training iterations the model must go through, which can significantly increase the time taken to train the model, especially when large data sets are being used. Therefore, a compromise is made between the variance allowed in the reported error and the computational power available for training the model. For the model applied in this study, a data set containing 57 samples was used. In comparison to many machine learning problems looking at 'big data' this is a rather small sample size. Therefore, the use of 50-folds for cross validation (the most allowed by the classification learner app) provided a reasonably unbiased measure of the overall model error, as *k* was close to the total number of sample present, while still being able to train the model in a reasonable timeframe (within 5-10 minutes).

The classification accuracies for the different models trained can be found in Table 2. The overall accuracy of the model can be defined as the total number of correct prediction (i.e. total true positives and true negatives) in relation to the total number of samples in the training set [24]. After investigating several different types of classification models available in the MATLAB classification learner app, it was found that the naïve Bayes model provided the best option for class prediction. While the ensemble model provided a slightly higher overall accuracy, it resulted in a loss of accuracy for the fentanyl analogues and no change in the prediction accuracy of the AH series, with the increase in overall accuracy coming from the U series compounds. In addition, an ensemble model is a more complicated model and took significantly longer to train than the naïve Bayes. The compromise between a slightly lower overall accuracy and simpler/faster model to use and re-train as new data is obtained means that a naïve Bayes model may be more suitable for routine use.

As the name suggests, a naïve Bayes classifier uses Bayes theorem to determine the appropriate output based on the given data [25]. A naïve Bayes classifier also involves several assumptions. The first of these assumptions is that the input variables (predictors) are independent of each other, i.e. the presence of one feature does not affect the others [26]. Secondly, this model type assumes that all the predictors have an equal effect on the outcome [26]. While these assumptions are not always true for every data set, naïve Bayes classifiers still provide high classification accuracy [27].

The confusion matrix produced from the trained model is displayed in Figure 2. Using the hyperparameter tuning incorporated in the classification learner app, it was found that a kernel naïve Bayes structure performed the best with a gaussian kernel type. The minimum

classification error plot showing the change in error with each training iteration can be found in the Supporting Information (Figure S1). The model was retrained using fewer training iterations, however there was no difference in the overall accuracy of the model (89.5%) or the confusion matrix produced (Figures S2 and S3). If the model was retrained in the future to incorporate additional compounds, the use of fewer training iterations may save time in the construction of the updated model, however care should be taken to ensure that the optimal parameters are still reached. The overall accuracy of the trained model was found to be 89.5% and it can be seen in Figure 2 that the model was most accurate for the fentanyl analogues. The confusion matrix shows that there is also some bias towards the fentanyl analogues, with the model preferentially classifying this group over the others. This is likely due to the fact that there were significantly more fentanyl analogue standards used in the training of the model than the other two subclasses. This is somewhat unavoidable given that this group tends to be much more prevalent and diverse than the AH and U series opioids. In the context of routine drug screening, having a model that is more accurate for the detection of fentanyl analogues may, in fact, be beneficial. These compounds are the most prevalent of the synthetic opioids with the United Nations Office on Drugs and Crime reporting that the majority of the 22 new opioids reported to their early warning advisory were fentanyl analogues [28]. Additionally, in Europe, until the end of 2018 they accounted for the majority of the novel opioid analogues being reported to the European Monitoring Centre for Drugs and Drug Addiction [29]. This suggests that the inherent biases present in the model may reflect the current landscape of the illicit opioid market.

It is possible, however, to artificially introduce biases into the training of the model using a cost matrix. While this may reduce the overall accuracy of the model, it may help to balance the false classifications from the AH and U series compounds. This compromise should therefore be evaluated on a case-by-case basis, depending on the intended application of the model. As the drug market continues to develop with the production of new compounds, it may be possible to further refine the model if more AH and U series opioids are developed which can be incorporated into the training of the model.

In addition to the overall accuracy of the model, the F1 score and MCC were calculated based on the confusion matrix presented in Figure 2. It has been suggested that, when a dataset is unbalanced (i.e. there are more samples belonging to one class than the others), accuracy alone may not be a reliable measure of model performance [30,31]. This is because the accuracy measurement may provide an overoptimistic estimation based on the classification ability of the majority class [30,31]. The F1 score, however, is a better metric when the classes are unbalanced and takes into account both the precision (also known as the positive predictive value) and recall (also known as sensitivity) of the model [24]. The MCC can also be a reliable metric as it will only produce a high score if the model prediction obtained good results for all four of the confusion matrix categories, namely true positive, true negative, false positive and false negative [31]. While both the F1 score and MCC are designed for use on binary datasets (i.e. where there are only two classes present), they can be applied to multiclass models through the use of 'micro averaging' and weighted averages. Micro averaging treats the entire data set as an aggregate result and does not consider the individual classes. In this case, all of the confusion matrix categories mentioned above would be added together for

each class and used to calculate the F1 score and MCC. On the other hand, the F1 score and MCC can be calculated for each class individually and a weighted average taken, taking into account any class imbalance. Table 3 shows the F1 score and MCC values calculated for the optimised model using both calculation methods. A score of 1 for both of these metrics would indicate a perfect correlation. It can be seen from these scores that they all return high values, with the micro averaged F1 score equalling the overall accuracy of the model, which supports the high accuracy and suitability of the trained model.

As new opioids belonging to any of the 3 classes are identified, their MS^2 data can be collected and incorporated into the model, ideally leading to an increase in prediction accuracy over time as new data is generated. Additionally, if new synthetic opioid subclasses are identified, such as the benzimidazole class recently reported by Blanckaert *et al.* [29], product ion information can be collated and the model retrained to expand the scope of compounds covered. In this way, the classification model can continue to be developed and adapted over time to stay up to date with developments in the illicit drug market.

3.2. Retention Time Repeatability Studies

Before retention time prediction can be attempted, it is important to establish the precision of RT values under the applied chromatographic conditions. If significant variation in the RT of a given compound is observed, the efficacy of a retention time prediction algorithm would be severely limited.

In order to first determine the repeatability of RT values without the possibility of matrix effects, mixed neat standards were analysed using the developed chromatographic method. A representative panel of 26 different synthetic opioids, including compounds from each of the different subclasses, were chosen for the repeatability study. Sets of samples were analysed across multiple days to account for both intra and inter-day variability and the average RT, SD, and %RSD were calculated. It was found that the %RSD for all the representative compounds was within 4%, suggesting a high degree of precision. Importantly, the absolute SD of all the compounds was ≤ 0.120 min. This falls within $\pm$ 0.2 min required by the Association of Official Racing Chemists (AORC) for the identification of compounds [32].

While the precision of the RT values in neat standards shows promise, in order to determine the applicability to realistic samples, spiked plasma samples were analysed to determine if matrix effects would affect this. These samples were also prepared at a lower concentration (10 ng/mL) to further simulate a more realistic case scenario. In addition, an internal standard of desipramine-d_3 was included to evaluate if the use of RRT values could further improve the precision of measurements. It was found that there is no significant difference in the absolute RT measurements between the neat and spiked samples, with most compounds still displaying %RSD values within 4%. As might be expected, the absolute SD of the RRT measurements were much smaller than for the absolute RT. Generally, the RRT measurements showed improved precision between 0.1% and 0.5%. Based on these results, it is suggested that an internal standard is included when developing retention time prediction models, so that RRT values can be used. The results of the repeatability study in both neat standards and matrix spikes can be found in the supporting information (Tables S4 and S5).

While the internal standard used in this study was desipramine-d_3, in practice, any compound with a suitable retention time could be used to determine RRT values for training and implementing the prediction model.

3.3. Retention Time Prediction

When training a regression model, the accuracy of the model can be evaluated by considering the root mean square error (RMSE) of the output variables. The mean square error is average square of the difference between the predicted and actual target variables [33]. This metric, however, is measured in units that are the square of the target output, therefore the RMSE is often preferred as it is easier to interpret in the context of the developed model [21,33]. Put simply, the RMSE measures the standard deviation present in the residuals of the model. The residuals display the difference (error) between the experimental value and the predicted value given by the model. The less variation there is between the experimental and predicted values, the smaller the residuals will be and, therefore, a more accurate model will give a lower RMSE.

In this study, four different regression models were evaluated to determine the best model architecture for this application. The observed RMSE values for each model type are presented in Table 4. Since regression models are predicting values in a continuous space, variation can be seen between the RMSE values for the same model type over different training periods. Therefore, the RMSE values presented in Table 4 show an average over 3 models trained on the same data.

The Gaussian Process Regression (GPR) model provided the best accuracy for the RT prediction. GPR models are non-parametric models, which use a Bayesian approach to regression problems [34]. One advantage of GPR approaches is that the prediction is probabilistic, meaning that the estimate for a given point contains uncertainty information as well [35]. It has been suggested that this type of model is suitable for complex regression problems, with high-dimensional data (i.e. large number of predictor variables) and small samples sizes [36].

Once evaluation of the different model types had been completed, the effect of individual predictors on the overall accuracy of the model was explored. These accuracies were then plotted against the overall accuracy of a model trained using all predictors (Figure 3). It can be seen from Figure 3 that 5 of the predictors resulted in a large increase in RMSE when removed, indicating that they are important predictors for determining retention time. A further 3 indicators showed a smaller increase in RMSE, or no change, indicating that they do not have as large of an effect, but can still influence the accuracy of the model. For 5 of the predictors, however, namely nO, nDB, nR06, logS and Hy, it was found that the accuracy of the model improves when those features are excluded. This indicates that these features are not good predictors of retention time and may harm the overall accuracy of the model. Following this determination, a new GPR model was trained, which omitted these 5 predictors, and was determined to have an average RMSE of 0.084348. Interestingly, when training the adjusted model without these predictors, no variation in the RMSE was observed between training attempts, unlike the other models. The difference between using RRT and

absolute RT values for the model output was also evaluated for the optimised model. The model returned an RMSE value of 0.4432 using the absolute RT values. The difference in the error can be explained by the absolute RT values being larger than the RRT, resulting in a larger standard deviation being measured for the residuals. When the trained model was evaluated, however, a greater variance was seen between the predicted and experimental RT values. The outputs from the model trained using absolute RT values can be found in the supporting information (Figures S4 and S5).

Figure 4 presents the predicted response (in this case RRT) compared to the experimentally determined 'true' RRT to provide a visual representation of the model accuracy. The majority of the data points are in good agreement with the regression line. The hyperparameter optimisation completed on the GPR model found that a basis function of zero and a nonisotropic squared exponential kernel function provided the best model accuracy. The sigma value used in the Gaussian processes varied slightly between the different training attempts, however the overall RMSE produced was the same. The specific hyperparameters used in the optimised model were determined automatically by the training algorithm. While these hyperparameters can be useful for comparison between similar models, they do not impact the routine implementation of the model once trained. The minimum mean square error (MSE) plot showing the change in the model error over the sequential training iterations can be found in the Supporting Information (Figure S6).

Another visual method that can be used to examine the errors present within the module is by examining the residuals produced from the predictions (Figure 5). When the residuals are clustered around the lower end of the y-axis, it indicates a higher model accuracy, as a residual of 0 means that the model has correctly predicted the RRT of the given sample. In the case of this model, 79.7% of the samples produced residuals within ± 0.1 min. The RMSE of 0. 084348 calculated for this model refers to the standard deviation of the residuals shown in Figure 5. This means that samples predicted within approximately ± 0.1 min fall within the expected variation of the model. This further reinforces the suitability of the trained model to the application of retention time prediction. Figure 5 also shows that the residuals are relatively symmetrical and there are no clear patterns present. This indicates that there is no significant bias in the model. If an observable trend were displayed in the residual plot, such as an increase in error as the experimental RRT increased, it would indicate that there was a problem with the algorithm the model was using for prediction.

In the same way as the classification model, the GPR model can be re-trained with new data to expand it to include more compounds. It is important, however, that any new samples which are to be subjected to the prediction model, or used to re-train the model, are analysed using the same chromatographic conditions as the original training data. This means that if a laboratory incorporates retention time prediction into their data analysis workflow, the model used needs to be trained and optimised to suit the analytical methods being used.

A limitation of retention time prediction modelling using molecular features is that a suspected compound identity must be known before the model can be applied. In order to generate the predictors that are used by the model, a suspected structure is needed. This does not diminish the usefulness of the model when incorporated into a rigorous compound

identification workflow, however. A classification model, such as the one presented in section 3.1, can be used to give a general indication of the type of compound present in the suspected sample. This can help inform further identification processes, such as the use of vendor software like *Molecular Structure Correlator* or *Compound Discoverer*, in order to putatively identify an unknown compound. Once this putative identification has been achieved, the required molecular features for the compound can be generated and input to the retention time prediction model. The predicted RRT can then be compared to the experimental RRT to provide further evidence to support the identity of the unknown. In cases where multiple possible identities are determined for a specific unknown, the molecular features of all possible identities can be generated, and the predicted RRT values compared to determine the most likely identity of the unknown compound. By using models such as these, laboratories can perform a putative identification of an unknown compound with a higher degree of confidence, which can save time and money by limiting the purchase and/or production of erroneous CRMs.

A limitation of the models presented in this study is the lack of authentic samples with which to validate the applicability of the models. Future studies should strive to demonstrate the accuracy of models such as these on authentic administration samples of synthetic opioids. This limitation does not preclude the inclusion of these models in a complementary targeted/non-targeted screening workflow, as the intention of models such as these is to provide preliminary intelligence to assist an analyst in the identification of an unknown component within a sample.

4. Conclusion

The use of machine learning to assist with the identification of unknown compounds has shown significant potential. The classification model developed in this study showed a high degree of accuracy for the prediction of opioid subclasses. This model can be further developed and refined as new compounds are produced to encompass a broad spectrum of compounds. While the developed model showed some bias towards the classification of fentanyl analogues, cost matrices can be introduced to counteract this bias. These can be applied on a case-by-case basis depending on the priorities of the laboratory. Additionally, the retention time prediction model showed good correlation between predicted and experimental RRT values. The use of an internal standard to correct for any intra- and inter-day variations resulted in improved precision compared to absolute RT values alone. The developed models can be incorporated into a compound identification workflow and expanded and optimised based on the requirements of an individual laboratory.

Acknowledgements. This research is supported by an Australian Government Research Training Program Scholarship awarded to J Klingberg. The authors thank Ms Lauren McClure at ARFL for assistance in the operation and maintenance of the LC-QTOF-MS instrument. The Poroshell LC column used in this study was a gift from Agilent Technologies. The authors would also like to thank Emmanuel Blanchard and MathWorks for their help developing the MATLAB code.

5. References

1. Klingberg J, Cawley A, Shimmon R, Fouracre C, Pasin D, Fu S. Finding the Proverbial Needle: Non-targeted Screening of Synthetic Opioids in Equine Plasma. *Drug Test Anal.* Accepted 29th June 2020.
2. Klingberg J, Cawley A, Shimmon R, Fu S. Collision-Induced Dissociation Studies of Synthetic Opioids for Non-targeted Analysis. *Front Chem.* 2019;7(331).
3. Pasin D, Cawley A, Bidny S, Fu SL. Current applications of high-resolution mass spectrometry for the analysis of new psychoactive substances: a critical review. *Anal Bioanal Chem.* 2017;409(25):5821-5836.
4. Noble C, Dalsgaard PW, Johansen SS, Linnet K. Application of a screening method for fentanyl and its analogues using UHPLC-QTOF-MS with data-independent acquisition (DIA) in MSE mode and retrospective analysis of authentic forensic blood samples. *Drug Test Anal.* 2017;10(4):651-662.
5. Pasin D, Cawley A, Bidny S, Fu S. Characterization of hallucinogenic phenethylamines using high-resolution mass spectrometry for non-targeted screening purposes. *Drug Test Anal.* 2017;9(10):1620-1629.
6. Anstett A, Chu F, Alonso DE, Smith RW. Characterization of 2C-phenethylamines using high-resolution mass spectrometry and Kendrick mass defect filters. *Forensic Chem.* 2018;7:47-55.
7. Margagliotti G, Bollé T. Machine learning & forensic science. *Forensic Sci Int.* 2019;298:138-139.
8. MATLAB. Introducing Machine Learning. In: *Machine Learning with MATLAB.* Online: MathWorks; 2016. 92991v00
9. Mitchell TM. *Machine Learning.* New York: McGraw-Hill; 1997.
10. MATLAB. Applying Supervised Learning. In: *Machine Learning with MATLAB.* Online: MathWorks; 2016. 80827v00
11. MATLAB. Applying Unsupervised Learning. In: *Machine Learning with MATLAB.* Online: MathWorks; 2016. 80823v00
12. Ekins S. *Computational Toxicology : Risk Assessment for Chemicals.* Newark, United States: John Wiley & Sons, Incorporated; 2018.
13. Luechtefeld T, Rowlands C, Hartung T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicol Res-UK.* 2018;7(5):732-744.
14. Pyke JS, Black G, Chen K, Anumol T, Young TM. *Simultaneous Targeted Quantitation and Suspect Screening of Environmental Contaminants in Sewage Sludge.* Online: Agilent Technologies;2019. 5994-0750EN.
15. Miller TH, Musenga A, Cowan DA, Barron LP. Prediction of Chromatographic Retention Time in High-Resolution Anti-Doping Screening Data Using Artificial Neural Networks. *Anal Chem.* 2013;85(21):10330-10337.
16. Mollerup CB, Mardal M, Dalsgaard PW, Linnet K, Barron LP. Prediction of collision cross section and retention time for broad scope screening in gradient reversed-phase liquid chromatography-ion mobility-high resolution accurate mass spectrometry. *J Chromatogr A.* 2018;1542:82-88.
17. ChemAxon. Chemicalize. 2020; https://chemicalize.com/app.
18. Wei YC. PaDEL-Descriptor. 2014; http://www.yapcwsoft.com/dd/padeldescriptor/.
19. Lambert JB. *Organic structural spectroscopy.* 2nd ed. ed. Upper Saddle River, N.J: Pearson Prentice Hall; 2011.
20. Molecular Descriptors Guide. In: U.S. Environmental Protection Agency, ed. 1.0.2 ed. Online2008.
21. Ozdemir S. *Principles of Data Science.* Packt Publishing; 2016.
22. Koehrsen W. A Conceptual Explanation of Bayesian Hyperparameter Optimization for Machine Learning. *Towards Data Science* 2018; https://towardsdatascience.com/a-

conceptual-explanation-of-bayesian-model-based-hyperparameter-optimization-for-machine-learning-b8172278050f. Accessed 24th April, 2020.
23. James G, Witten D, Hastie T, Tibshirani R. Resampling Methods. In: *An Introduction to Statistical Learning: with Applications in R.* New York, NY: Springer New York; 2013:175-201. 978-1-4614-7138-7
24. Huilgol P. Accuracy vs. F1 Score. *Analytics Vidhya* 2019; https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2. Accessed 9th July, 2020.
25. Cichosz P. Naive Bayes classifier. In: *Data Mining Algorithms : Explained Using R.* Somerset, United Kingdom: John Wiley & Sons, Incorporated; 2015. 9781118950807
26. Gandhi R. Naive Bayes Classifier. *Towards Data Science* 2018; https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c. Accessed 9th April, 2020.
27. Webb GI. Naïve Bayes. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning and Data Mining.* Boston, MA: Springer US; 2017:895-896. 978-1-4899-7687-1
28. United Nations Office on Drugs and Crime. *World Drug Report 2019.* Vienna: United Nations Office on Drugs and Crime; June 2019.
29. Blanckaert P, Cannaert A, Van Uytfanghe K, et al. Report on a novel emerging class of highly potent benzimidazole NPS opioids: Chemical and in vitro functional characterization of isotonitazene. *Drug Test Anal.* 2020;12:422-430.
30. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21(1):6.
31. Bekkar M, Djemaa HK, Alitouche TA. Evaluation measures for models assessment over imbalanced data sets. *J Informa Eng Appl.* 2013;3(10).
32. Association of Official Racing Chemists. *AORC Guidelines for the Minimum Criteria for Identification by Chromatography and Mass Spectrometry.* Online2016. http://www.aorc-online.org/documents/aorc-ms-criteria-modified-23-aug-16/.
33. Rajdeep D, Manpreet Singh G, Nick P. Mean Squared Error and Root Mean Squared Error. In: *Machine Learning with Spark - Second Edition.* Packt Publishing; 2017. 1785889931
34. Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J Math Psychol.* 2018;85:1-16.
35. Rasmussen CE. *Gaussian processes for machine learning.* Cambridge, Mass: MIT Press; 2006.
36. Yang D, Zhang X, Pan R, Wang Y, Chen Z. A novel Gaussian process regression model for state-of-health estimation of lithium-ion battery using charging curve. *J Power Sources.* 2018;384:387-395.

Table 1. Molecular features used for retention time prediction

Feature	Abbreviation	Source
Stongest basic pKa	pKa	Chemicalize [17]
Number of carbon atoms	nC	Chemicalize [17]
Number of oxygen atoms	nO	Chemicalize [17]
Number of double bonds	nDB	Chemicalize [17]
Number of 6-membered rings	nR06	Chemicalize [17]
Number of benzene rings	nBNZ	Chemicalize [17]
Partition coefficient	logP	Chemicalize [17]
Distribution coefficient (pH 9)	logD (pH 9)	Chemicalize [17]
Aqueous solubility (pH 9)	logS (pH 9)	Chemicalize [17]
Ghose-Crippen logP	AlogP	Pharmaceutical Data Exploration Laboratory [18]
Moriguchi logP	MlogP	Pharmaceutical Data Exploration Laboratory [18]
Double bond equivalents	DBE	Equation 1 [19]
Hydrophilic factor	Hy	Equation 2 [20]

Table 2. Accuracy of each class prediction model trained.

Model	Overall Accuracy (%)	Class Accuracy (%)		
		Fentanyls	AH Series	U Series
Decision Tree	87.7	94.3	57.1	86.7
Discriminant Analysis	86.0	85.7	71.4	93.3
Naïve Bayes	89.5	100.0	57.1	80.0
Support Vector Machine	84.2	97.1	71.4	60.0
Nearest Neighbour	87.7	100.0	71.4	66.7
Ensemble	91.2	97.1	57.1	93.3

Table 3. Accuracy of each class prediction model trained.

Metric	Micro Average	Weighted Average
F1 Score	0.895	0.889
Matthew's Correlation Coefficient	0.842	0.801

Table 4. Accuracy of each retention time prediction model trained, measured by the root mean square error (RMSE).

Model	RMSE (mean ± range)
Regression Tree	0.177630 ± 0.0000
Support Vector Machine	0.123920 ± 0.01810
Ensemble of Trees	0.158627 ± 0.01540
Gaussian Process Regression	0.096608 ± 0.018173
Gaussian Process Regression (omitted features)	0.084348 ± 0.00000

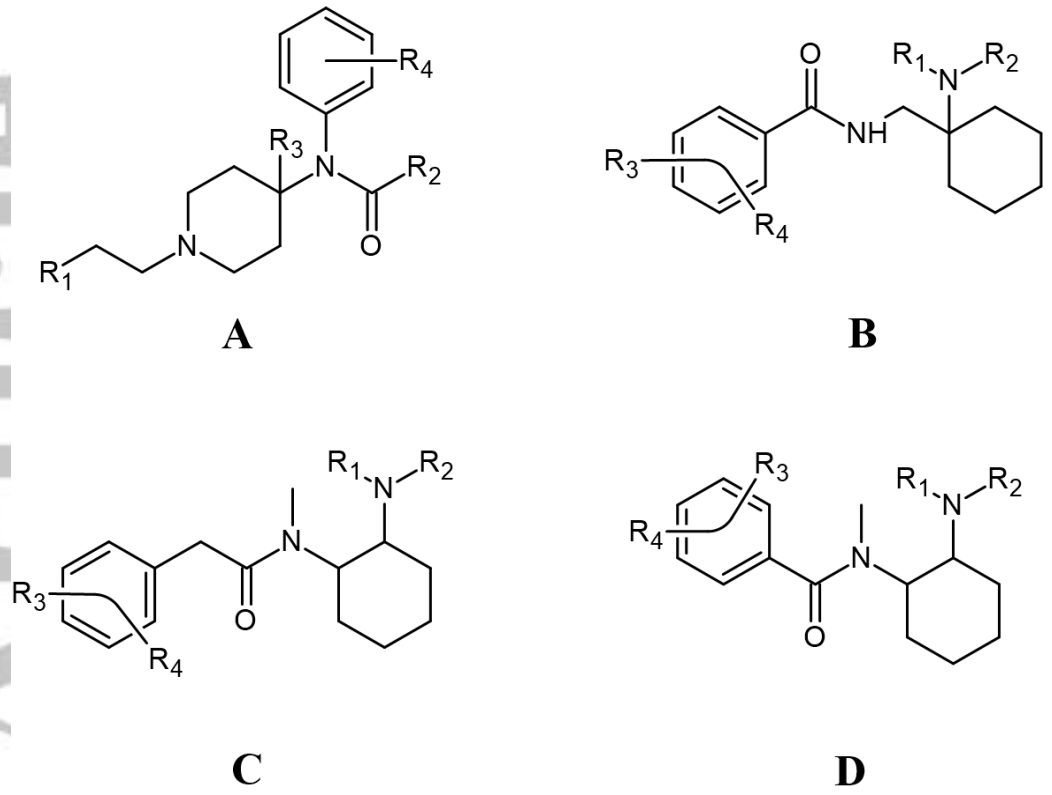


Figure 1. Generic structures of the opioid subclasses used in the class prediction model, including fentanyl analogues (A), AH series (B) and U series with (C) and without (D) a methylene spacer.

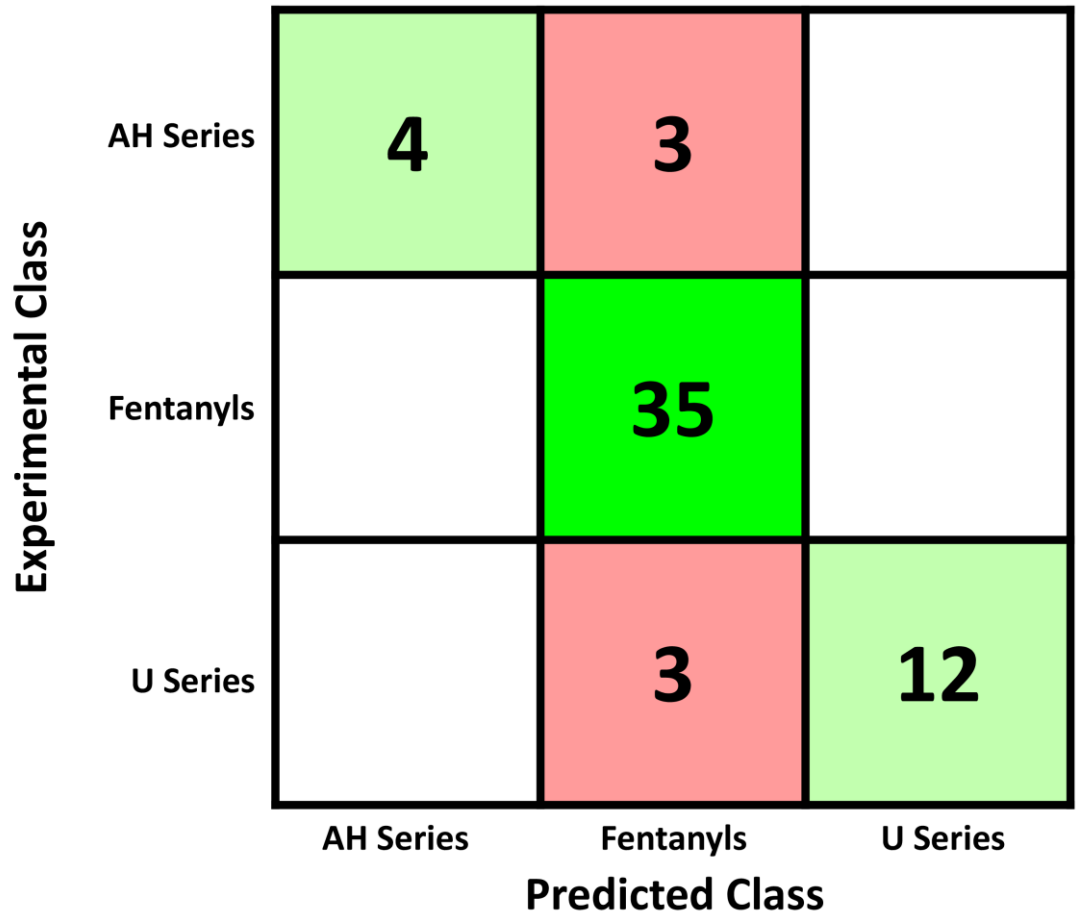


Figure 2. Confusion matrix showing the prediction accuracy of the developed Naive Bayes model for each opioid subclass.

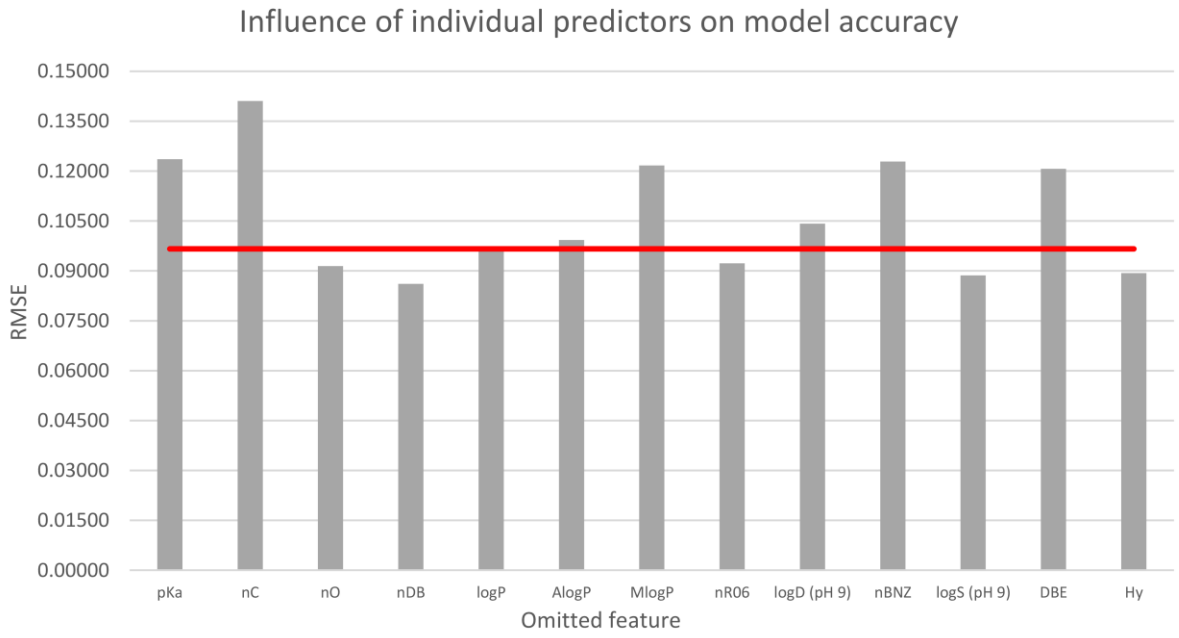


Figure3. Model accuracy following removal of individual predictors in comparison to the accuracy of a model trained with all features (red). A lower RMSE indicates a higher model accuracy.

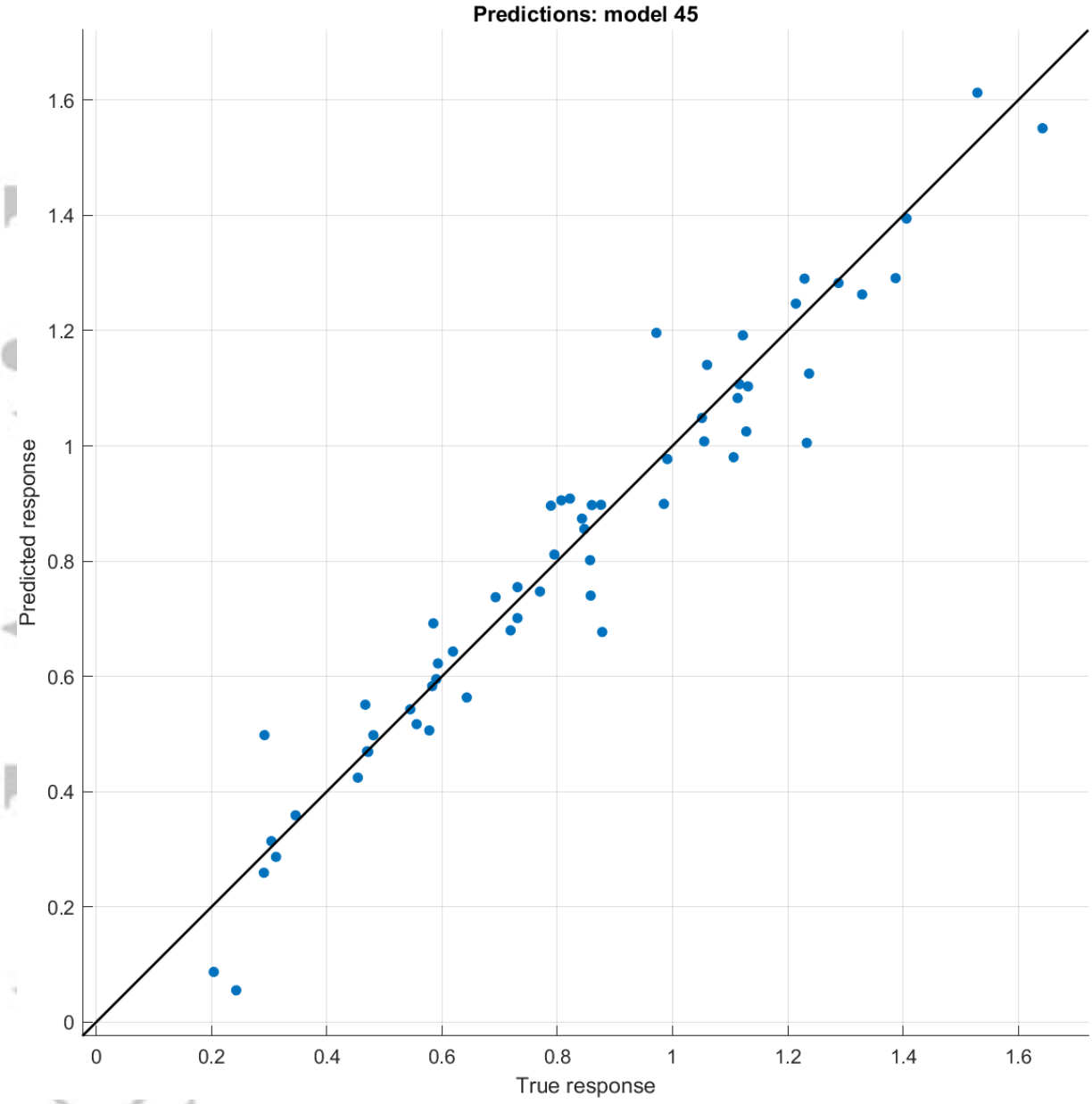


Figure 4. Predicted (RRTp) vs. experimental (RRTe) for the developed Gaussian Process Regression model

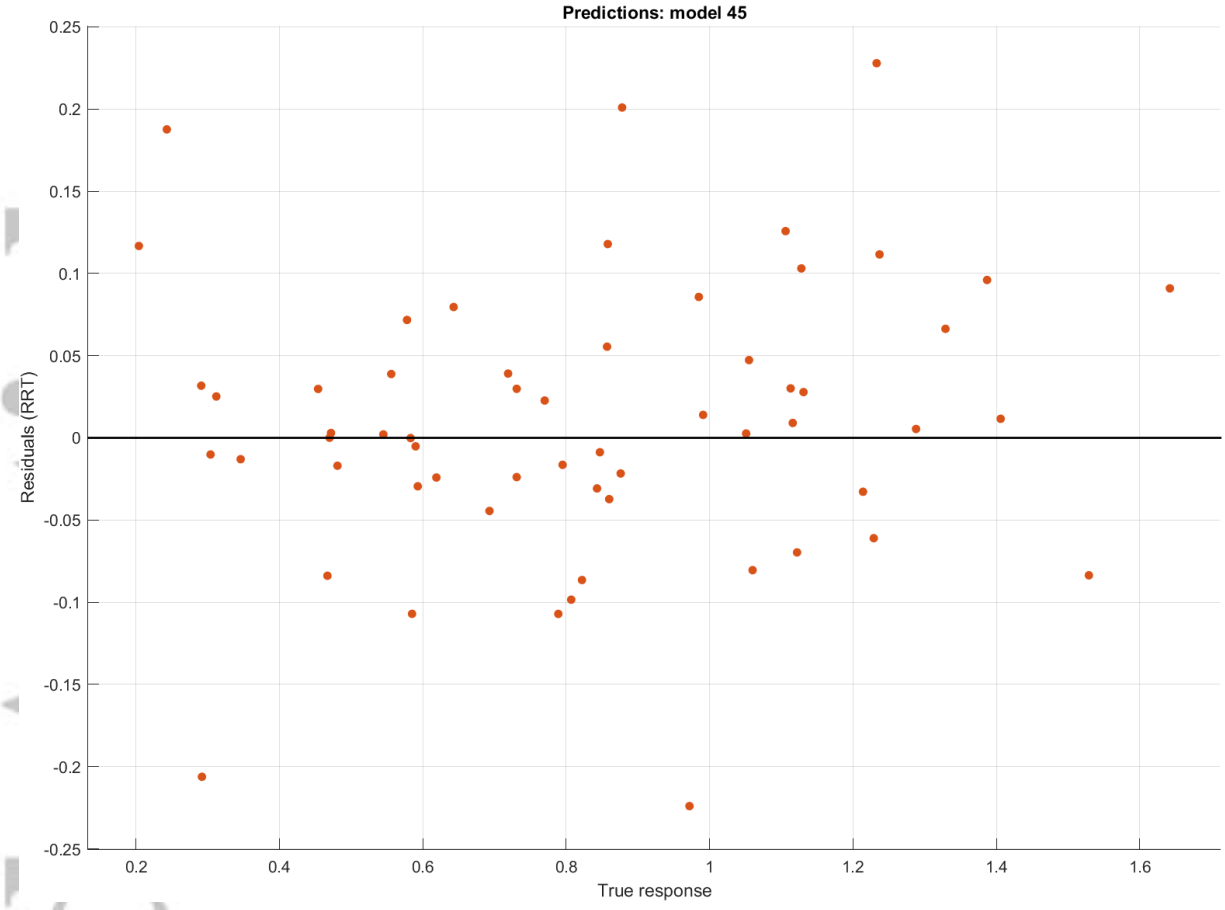
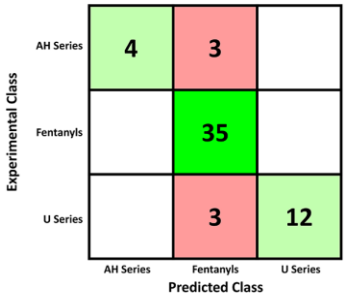


Figure 5. Residuals produced from the retention time prediction model

Machine learning approaches for the identification of unknown compounds in equine plasma is presented. The optimisation of a classification model to predict opioid subclasses and a regression model to predict experimental retention times is presented.



Towards Compound Identification in Non-targeted Screening Using Machine Learning Techniques

Joshua Klingberg, Adam Cawley, Ronald Shimmon, Shanlin Fu*