**Supporting Information**

for

**Predicting Bioactivity when there is No Target: Performance of Various Methods in an Open Source Malaria Competition**

Edwin G. Tse, Kiaran Kirk, Davy Guan, Ho Leung Ng, Alice Motion, Murray N. Robertson, Jonathan Cardoso-Silva, Benedict W. J. Irwin, Mario Öeren, Thomas M. Whitehead, Gareth J. Conduit, Alexander D. Wade, Laksh Aithani, Willem P. van Hoorn, Raymond Lui, Slade Matthews, Giovanni Cincilla, Mykola Galushkaand Matthew H. Todd

**Experimental procedures, characterisation data and copies of 1H and 13C NMR spectra for all compounds.**

**Table of Contents**

[General 3](#_Toc38283882)

[Jonathan Cardoso-Silva’s Model 5](#_Toc38283883)

[Molomics’ Model 7](#_Toc38283884)

[Auromind’s Model 10](#_Toc38283885)

[Davy Guan’s Model 12](#_Toc38283886)

[Optibrium’s Model 15](#_Toc38283887)

[Raymond Lui’s Model 18](#_Toc38283888)

[Slade Matthews’s Model 21](#_Toc38283889)

[Ho-Leung Ng’s Model 23](#_Toc38283890)

[Vito Spadavecchio’s Model 25](#_Toc38283891)

[Exscientia’s Model 26](#_Toc38283892)

[Fig. S1 Overall Synthetic Route for the Preparation of Target Compounds X, X, X and X 28](#_Toc38283893)

[Chemistry 28](#_Toc38283894)

[1H and 13C NMR Spectra 41](#_Toc38283895)

[LCMS Traces for the Intermediates of Compound 450 64](#_Toc38283896)

[References 65](#_Toc38283897)

# General

Reagents were purchased from either Sigma–Aldrich, Alfa Aesar, Acros, Merck, Fischer Scientific, Matrix Scientific, Ajax or Fluorochem. Unless otherwise specified, the reagents were used without further purification. Anhydrous conditions: glassware was dried at >130 oC for >12 h, assembled hot and allowed to cool under a high vacuum where appropriate or purged with inert gas. Anhydrous solvents were obtained from the PureSolv system or by drying over activated 3 Å molecular sieves. Nitrogen gas was dried over silica and calcium chloride. Argon gas was used as acquired. The phrase *in vacuo* corresponds to ∼1 mbar on a Schlenk line. Reduced pressure means under rotary evaporation at 40 oC from 900–50 mbar. Flash chromatography was performed on Davisil Grace Davison 40–63 µm (230–400 mesh) silica gel or on a Biotage Isolera One. Analytical thin layer chromatography was performed on Merck Silica Gel 60 F254 precoated aluminium plates (0.2 mm) and visualised with UV irradiation (254 nm) and potassium permanganate, anisaldehyde or ninhydrin staining. High temperature reactions were carried out in silicone oil baths, controlled by temperature probe in the oil bath.

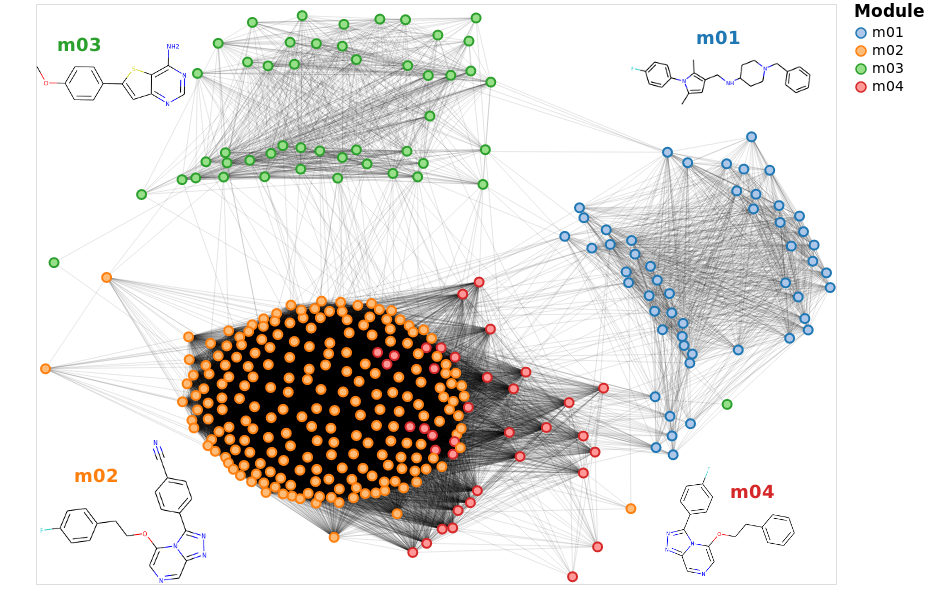
Melting points (m.p.) were recorded on a Stuart SMP10 at 2 oC min-1 (capillaries ø = 1.8–1.9 mm, 100 mm). Infrared spectroscopy was carried out on a Bruker Alpha-E (attenuated total reflectance) without atmospheric compensation and processed using OPUS 7.0 software. Samples were analysed neat. Nuclear magnetic resonance spectroscopy was carried out at 300 K on Bruker spectrometers: either AVANCE III 400 (1H at 400 MHz, 13C at 101 MHz) or AVANCE III 500 (1H at 500 MHz, 13C at 126 MHz). Spectra were processed using Mestrelab Research Mnova. Deuterated solvents (CDCl3, DMSO-d*6*, CD3OD) obtained from the Cambridge Isotope Laboratories. 1H and 13C chemical shifts are reported in parts per million (ppm) with respect to TMS at 0.00 ppm. The chemical shifts of the spectra were calibrated to residual solvent peaks (1H: CHCl3 7.26 ppm, DMSO 2.50 ppm, MeOH 3.31 ppm, TMS 0.00 ppm; 13C: CHCl3 77.16 ppm, DMSO 39.52 ppm, MeOH 49.00 ppm, TMS 0.00 ppm). 1H signal multiplicity is reported as: singlet (s), doublet (d), triplet (t), quartet (q), pentet (p) and combinations thereof, or multiplet (m). Broad signals are designated broad (br). Coupling constants (*J*) are reported in Hertz (Hz). Integrals are relative. app = apparent when the multiplicity was unexpected, e.g. coincidental or unresolved. Low resolution mass spectrometry (*m/z*) was carried out on a Finnigan quadrupole ion trap mass spectrometer using electrospray ionisation (ESI). High resolution mass spectrometry (HRMS) was performed on a Bruker 7T FT-ICR using ESI or APCI. Positive and negative detection is indicated by the charge of the ion, e.g. [M+H]+ indicates positive ion detection. Analytical liquid chromatography-mass spectrometry (LCMS) was performed on an Agilent Inﬁnity 1290 II system consisting of a quaternary pump (G7111A) and a diode array detector WR (G7115A) coupled to a InﬁnityLab LC/MSD (G6125B) using ESI. An Agilent Poroshell 120 EC-C18 column (2.7 µm, 3.0 × 50 mm) was eluted at a ﬂow rate of 1.5 mL/min with a mobile phase of 0.05% formic acid in H2O and 0.05% formic acid in MeCN.

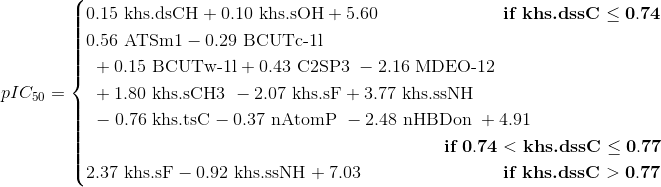
# Jonathan Cardoso-Silva’s Model

A recent two-stage algorithm that combines network analysis and piecewise linear regression modelling, called modSAR[Cardoso-Silva2019], was used to generate modular QSAR models for OSM Series4 data. Fingerprint features are used to create a network representation of the compounds and to split compounds into disjoint modules. Each module is then modelled by a piecewise linear regression model using molecular descriptors.

The original dataset provided containing 440 compounds was deduplicated by their OSM-ID, the median potency value was taken for each duplicated compound. A threshold of 2.5 µM was applied to potency values. Circular topological fingerprints (ECPF4 1024 bits) were generated in RDKit [Landrum] and used to create a network representation of the data while approximately 200 descriptors from Chemistry Development Kit (CDK) version 2.0 [Willighagen2017] were used in the regression models.  Highly correlated CDK descriptors as well as those with near zero variance were removed, resulting in 95 molecular descriptors which were then centered and put on a scale from 0 to 1.

The network of OSM Series4 compounds can be seen in Figure # below. Similarity between compounds was calculated using Tanimoto coefficient (Tc) on their ECPF4 fingerprints, a link is drawn between a pair of molecules if  their similarity is above the threshold Tc >= 0.20. Four modules were identified and are pictured in the figure along with their most representative compound (the compound with the largest number of neighbours in their community).

 A piecewise linear regression model was trained for each module, further splitting the molecules into subgroups which are modelled independently by linear equations [Cardoso-Silva2019-b]. The models were trained in 10-fold Cross validation schema as an example, these were the breakpoints and equations found for compounds in module m01:

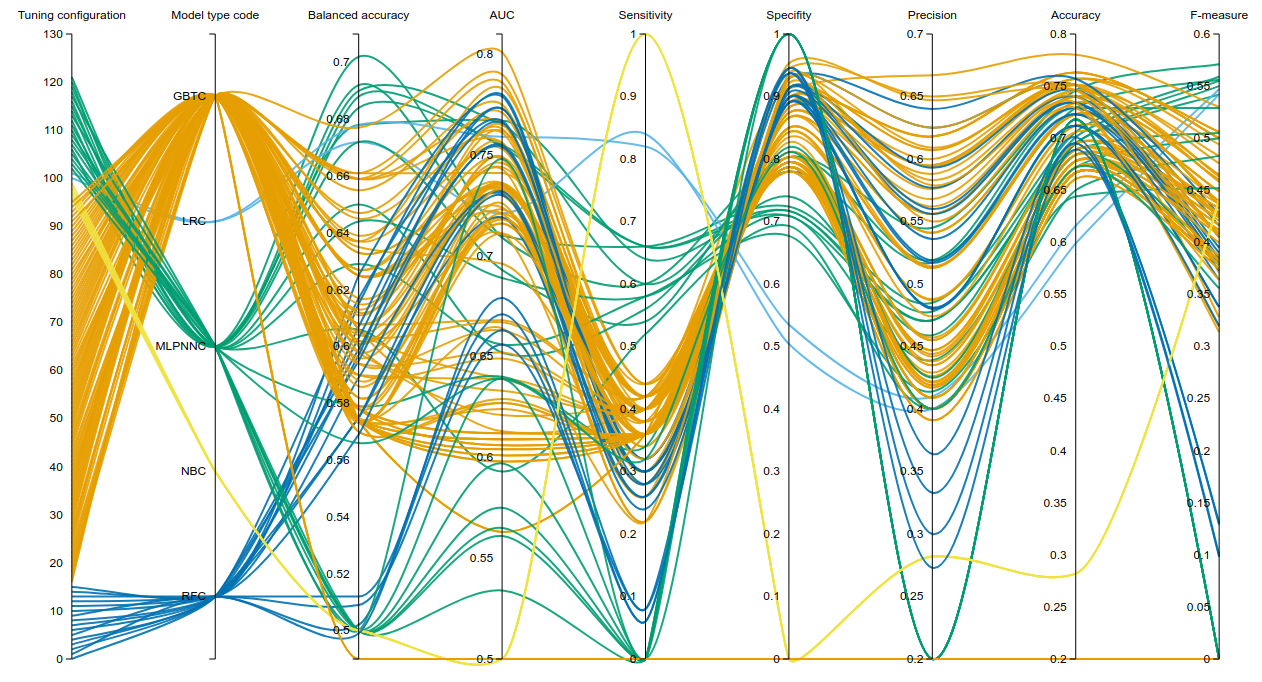


Despite the descriptive capabilities of the method, this model didn’t achieve a high precision on the hidden test set. A possible cause is the choice of descriptors, a different set of molecular descriptors, fingerprints or maybe pseudo-features generated by techniques such as autoencoder could probably yield better results. We could also try to use a different regression technique on the modules such as Random Forest.

# Molomics’ Model

As the objective of the requested model was to predict activity of series-4 compounds, only molecules of this series were used for model development. The data was curated according to Molomics standard protocol. A modelling set and an external set were generated for both regression and classification modelling.

General 2D descriptors and ECFC4 structural fingerprints were calculated. Constant or correlated descriptors were removed. Z-score normalization was applied to all the descriptors.

Many regression and classification models were developed using different Machine Learning methods (i.e. Random Forest, Gradient Boosted Tree, Logistic Regression, Multi-layer Perceptron Neural Network, Naïve Bayesian). The most important parameters of each of these models were tuned using response-stratified 10-fold Cross Validation. According to QSAR Best Practices [Tropsha2010], only models fulfilling strict acceptability criteria were considered to be eligible. No regression model resulted eligible. 3 of the most promising classification models were selected and submitted to a statistical significance analysis using bootstrapping (*n*=100) and Y-randomization (*n*=100) sampling. The final model was selected on the basis of the statistical significance analysis together with internal & external validation performance, resulting in a logistic regression model using a stochastic average gradient as solver [Schmidt2017], a uniform regularisation and a learning step size = 0.01. 

***Figure #****: Performance metrics of Molomics’ classification models. GBTC: Gradient Boosted Trees; LRC: Logistic Regression; MLPNNC: Multi-Layer Perceptron Neural Network; NBC: Naïve Bayesian; RFC: Random Forest.*

Nine possible Applicability Domain metrics were evaluated for the selected model according to a published unified approach [Dragos2009].

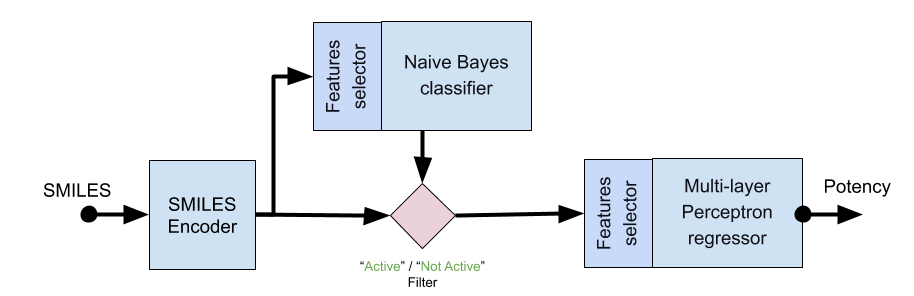
The best of them resulted to be a structure-based minimum distance towards the modelling set. Using this Applicability Domain allows to provide a confidence for each prediction.

Compound suggestions were generated using a Collective Intelligence approach through Molomics Technology (not published). This allowed to design molecules with optimal *in silico* Pfal activity, solubility and Caco-2 cell permeability. The 2 most desirable compounds (M295, M4250), selected through a consensus decision-making process, were discarded by OSM due to their difficulty in synthesis. Alternatively, compounds M2726 and M4625 were proposed for synthesis and experimental testing. 

***Figure #****: Molecules generated by Molomics technology through Collective Intelligence.*

According to OSM rules of transparency and collaboration, the model was finally deployed through Molomics Technology and made publicly available (molomics.com/osm) and exploitable for the design of additional Series-4 candidates through Collective Intelligence.

# Auromind’s Model

The model developed by Auromind for predicting "Potency vs Parasite (uMol)" for the Series 4 compounds consists of three modules integrated into a single pipeline.  ***Figure #****: The architecture of Auromind model for predicting* *"Potency vs Parasite (uMol)" for the Series 4 compounds.*

The first module is represented by an encoder isolated out of the pre-trained variational auto-encoder [Garciarena2018] (VAE). It transforms an input SMILES into a chemical compound fingerprint consisting of a 1024 real values vector. Often this vector is called “latent”-vector and a space formed by these vectors a "latent"-space accordingly. VAE positions similar compounds nearby to each other in the “latent” space. It helps to design machine learning algorithms, which utilise similarity in the “latent” space to predict various chemical properties. The VAE [armchem2019] used in this study has been trained on the 1.7m compounds selected from the ChEMBLv24 database [Davies2015]. The accuracy (of SMILES reconstruction) obtained as a result of 10 cross-validations was equal to 0.837±0.006. In addition to accuracy, the VAE performance was measured by Hamming [Norouzi2012] (0.682±0.051) and Levenshtein [Miller2019] (0.677±0.114) distances. Both belong to a family of editing distances and gives a different perspective on the assessment of the VAE performance.

The second module in the developed pipeline is a Gaussian Naive Bayes [Rish2001] (GNB) classifier. It performs a selection of active compounds (which potency value less or equals 2.5 thresholds). To achieve better accuracy the initial features space was reduced to 100 using the Chi-squared test. GNB has been selected as the best performing model (with ROC-AUC equals to 0.741±0.024) using a grid-search among other classification models.

The third module in the developed pipeline is a Multi-layer Perceptron [Ramchoun2001] (MLP) regressor, which receives chemical compounds already classified as "active" and predicts an actual potency value in a range between 0 and 2.5. The approach of filtering compounds helped to deal with experimental samples where IC50 values were greater than the max concentration that was tested in the assay. MPL was selected as the best using a grid-search performed among other regression models. To achieve better performance were selected 40 features using a univariate linear regression test. It was observed that all of the tested models were suffering from overfitting and demonstrated relatively poor performance. MPL showed the best result R2 (coefficient of determination) equals 0.141±0.017 among others with the regularization parameter equals 1. Such poor performance caused by an extremely limited number of training samples (205), preventing adequately establishing a mechanism of predicting a real value of potency. These results can be potentially improved with an increasing number of training samples.

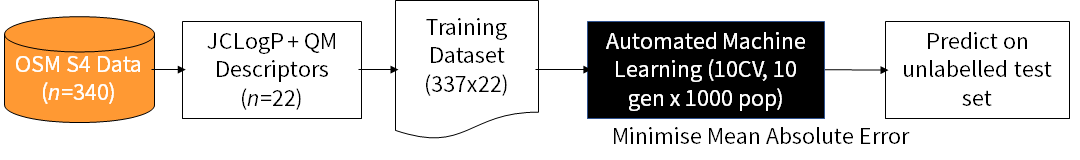
# Davy Guan’s Model

The aim of phenotypic drug discovery in this competition to characterise the effects of chemical agents on *in vivo* assays instead of a specific biological target resonates with the effort in toxicology to improve predictive modelling of animal assays to reduce, refine or replace their use [Wittwehr2017]. This meant we could utilise our experience and modelling methodology developments from that domain to address the same challenges arising from limited data availability and the nature of predicting in vivo outcomes in this project [Guan2018].

The provided 440 chemical OSM S4 dataset was deduplicated by averaging the potency values for each replicated ligand sharing the same OSM code to yield a 340 chemical training set in SMILES string format. This training set was further curated with the removal of salts and solvents, charge neutralisation, and the addition of explicit hydrogens to output in 2D SDF using ChemAxon Standardizer 18.22.0, 2019, ChemAxon (https://www.chemaxon.com). These 2D structures were input into a KNIME workflow [Berthold2008] consisting of the Open Babel [O’Boyle2011] and RDKit [Landrum] nodes to prepare initial 3D structures optimised using the Universal Force Field. The MaPhi descriptor program (manuscript in preparation, DOI: http://doi.org/10.5281/zenodo.1407646) acted as an interface for all subsequent quantum mechanical descriptor calculations. Geometry optimisation was conducted with the semiempirical PM7 method in gas phase implemented in MOPAC2016 [Stewart2016] followed by a final geometry optimisation stage at the Hartree Fock with 3 corrections (Hf-3c) [Sure2013] level of theory with implicit aqueous solvation using the CPCM implementation in Orca 4.2.0 [Neese2012][Neese2018][Barone1998]. 21 quantum mechanical descriptors derived from the electronic properties were calculated at the Hf-3c level of theory in implicit aqueous solvation with 337 chemicals achieving convergence (99.1% yield). Water-octanol partition coefficients were predicted using the JCLogP model in JChem for Excel 19.7.26, 2019, ChemAxon ([https://www.chemaxon.com](https://www.chemaxon.com/)), which was selected as an additional bioavailability descriptor after comparison to other partition coefficient models [Guan2020]. A 2.5 uM threshold was applied to the potency values in this dataset to produce a 22 descriptor training set for regression modelling.

The TPOT Python library [Olson2016] was used to automate model pipeline construction from dataset scaling and transformation to machine learning algorithm selection and hyperparameter optimisation. This library was initially configured to optimise 1000 models for the lowest Mean Absolute Error (MAE) from 10-fold cross validation over 50 generations, however, overfitting was detected for predicting the test dataset. The default setting of ten generations was selected instead with no overfitting detected. Additional dataset variants were also prepared with selected Mordred descriptors [Moriwaki2018] in collaboration with Raymond Lui with no apparent performance improvement. The final OSM Round 2 predictive model found 0.548 uM MAE in 10-fold cross validation following TPOT pipeline optimisation with 0.82 precision for the unlabelled test set. This regression model achieved the best predictive performance in the academic category which led into the second phase of the Open Source Malaria Round 2 Predictive Challenge.

The aim of the second phase of the Open Source Malaria Round 2 Predictive Challenge was to generate two molecules, one of which with a triazolopyrazine scaffold and one with a novel, dissimilar scaffold. A 290,000 chemical generative modelling dataset was prepared with 250,000 ZINC molecules, 340 OSM S4 compounds, and 40,000 compounds screened against *P. falciparum* in the ChEMBL database. This dataset only contained SMILES structures. The ChemGE program [Yoshikawa2018] was selected to expediently generate novel compounds based on the representative triazolopyrazine scaffold. This was achieved by incorporating an additional calculated triazolopyrazine similarity index comparing generated molecules to the reference triazolopyrazine scaffold in addition to the default configuration calculating synthetic accessibility, lipophilicity, and an aromatic ring penalty for scoring all generated molecules. A 22,000 chemical dataset with known *P. falciparum* potency values was composed from the ChEMBL database to prepare an additional model based on Mordred descriptors in case the solvated quantum mechanical descriptors could not be calculated in time for the deadline. The dataset curation methodology and KNIME workflow to generate 3D structures from the first phase was applied followed with the calculation of 1825 descriptors in Mordred 1.1.1. This dataset was used for generating a regression model using TPOT to optimise 100 models over 10 generations. Both models were used to generate consensus predictions by averaging predicted potencies for generated molecules with subsequent ranking based on JCLogS solubility values calculated using JChem for Excel 19.7.26, 2019, ChemAxon ([https://www.chemaxon.com](https://www.chemaxon.com/)).



**Figure XXX#.** Abridged modelling methodology for predicting the unlabelled test dataset in the first phase of the OSM Round 2 Predictive Challenge

# Optibrium’s Model

We used the Alchemite method (developed by Intellegens and Optibrium) [Whitehead2019] which performs multiple deep imputation on the entire dataset using an iterated deep neural network. Additional pKa descriptors (most basic site pKa, most acidic site pKa) were calculated using the StarDrop 6.6 pKa model [Hunt2020] ([Predicting pKa Using a Combination of Quantum Mechanical and Machine Learning Methods](https://www.optibrium.com/community/publications/publications-a-presentations/472-predictingpkaquantummachine)), this uses quantum mechanical descriptors. The master chemical dataset was kept partitioned as individual assays from different labs because mixing results from multiple assays is not desirable due to multiple definitions of inactivity cutoff and scales of resolution. A single Alchemite model was trained comprising each assay type over different isoforms from different labs as well as single shot inhibition and ion regulation activity columns. No hyperparameter optimisation was performed as the dataset was quite small, default general parameters were used with 3 iteration layers, iterative mixing fraction of 0.95, 50 input columns (from both descriptors or other endpoints) per output column. 330 StarDrop AutoModeller descriptors were generated for each compound.

Models were built using pIC50 and pEC50 values which were log-transformed from the original data. The coefficient of determination R2 values for the training and test sets are shown in **Table X.** Some of the assays apparently had very consistent data and both training and test models showed excellent performance, e.g. Pfal (K1) IC50 (Avery), Pfal IC50 (Syngene), Pfal IC50 (GSK) with test R2 values as high as 0.95. Unfortunately the model for Pfal IC50 (Dundee) was less good, which was the final assay used to evaluate the test compounds. The data suggest there could be more inconsistency in the Dundee assays that other labs results, or rather that the GSK, Syngene labs were high precision results, or that the activity of compounds tested in are easily modelled.

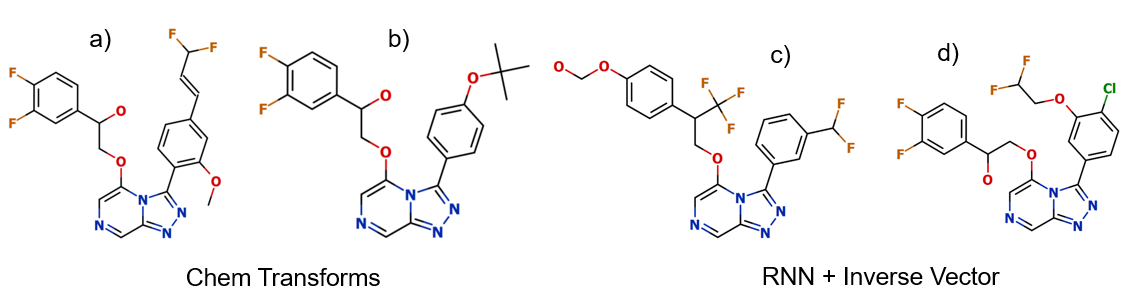
Alchemite produced **accurate error bars** for both training and test results. This meant it was possible to discard predictions which were known to be inaccurate. This was essential at the compound generation stage, where only confident predictions about compounds were trusted. We believe this is the reason why the Alchemite predicted pIC50 value of 6.4 for the *tert*-butyl Optibrium suggested compound matched the experimental value of 6.2 so closely and led to a successfully active compound where other methods which could not consider uncertainty struggled (a small known variance versus a potentially large unknown variance), however further data would be needed to be confident in this hypothesis.

**Table X**: Coefficient of determination score for independent assay columns which had enough test data to create a meaningful value (>4 points). Dundee was the assay conditions used for the final measurements.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Endpoint | Train R2 | Test R2 | Endpoint | Train R2 | Test R2 |
| EC50 Chembl (uM) | 0.72 | 0.16 | **Pfal IC50 (Dundee)** | 0.59 | 0.32 |
| Ion Regulation Activity | 0.76 | 0.25 | Pfal IC50 (GSK) | 0.95 | 0.8 |
| PfaI EC50 (Inh) | 0.73 | 0.21 | Pfal IC50 (Ralph) | 0.79 | 0.17 |
| Pfal (K1) IC50 (Avery) | 0.95 | 0.93 | Pfal IC50 (Syngene) | 0.79 | 0.80 |
| Pfal IC50 (Avery) | 0.86 | 0.70 | Single Shot Inhibition % | 0.79 | 0.70 |

Compound suggestions were generated in two ways:

1. A “bottom up” method using med chem expansions around promising compounds in the original series to generate around 90,000 ideas. This expansion was performed using the Nova module in StarDrop. All solutions were passed through the Alchemite model and the best were assessed using a multi-parameter score function from StarDrop which attempted to maximise activity in all assays and confidence in those assays, as well as increase solubility.
2. A “top-down” approach, in which a recurrent neural network (RNN) was trained on most of ChEMBL to generate SMILES from a vector input of descriptors. This vector comes from inverse solving the Alchemite model for vector which is predicted by the model to be confidently active. This vector was then turned into SMILES by the RNN, and the outputs were screened for synthesis and reactivity problems by hand.

The final compounds were suggested, two from each generation method, with compound b) being selected for synthesis due to a) and c) having potential HF and OCO reactivity problems respectively. 

# 

# Raymond Lui’s Model

A machine learning regression model approach was used to generate quantitative structure-activity relationship (QSAR) correlations between physicochemical/electronic descriptors and PfATP4 inhibitory potency.

*Dataset curation.* A training set of 340 unique molecules with known PfATP4 inhibitory potencies was curated from the provided OSM data; duplicates containing the same OSM ID were consolidated by taking their average potency, and any final values greater than 2.5 µM were set to 2.5 µM to define an active target range for the regression model. Molecules were expressed in SMILES and further curated in ChemAxon Standardizer v18.22.0 (ChemAxon Ltd., 2019; [https://www.chemaxon.com](https://www.chemaxon.com/)) by removing salts and solvents, neutralising any charged fragments, and adding explicit hydrogens.

*Molecular descriptor calculation.* Three-dimensional structural geometries were initially constructed in OpenBabel [O'Boyle2011] and RDKit [Landrum] using the Universal Force Field (UFF), then further optimised in MOPAC2016 [Stewart2016] using the PM7 method in the gas phase. 1,572 two- and three-dimensional physicochemical descriptors were then calculated in Mordred v1.1.1 [Moriwaki2018] for Python v3.6. Another 21 quantum mechanically-derived three-dimensional electronic descriptors were calculated in Orca v4.2.0 [Neese2012,Neese2018] after further structural optimisation using the Hartree Fock with 3 corrections (HF-3c) method [Sure2013] solvated in water with the Conductor-like Polarizable Continuum Model (CPCM) [Barone1998]. Structural curation, optimisation, and descriptor calculation was performed in collaboration with Davy Guan.

*Baseline QSAR model optimisation.* A genetic search algorithm was used to develop two baseline models correlating the provided PfATP4 inhibitory potencies with each of the physicochemical and electronic descriptor sets. A population of modelling pipelines each consisting of components including data preprocessors, machine learning algorithms, and their hyperparameters is firstly initiated. The genetic search algorithm then iteratively evolves these pipelines over multiple generations by randomly mutating the values and combinations of the aforementioned components to minimise the 10-fold cross-validation (10-fCV) mean absolute error (MAE), eventually returning the lowest MAE pipeline model. The Tree-based Pipeline Optimisation Tool (TPOT) v0.10.2 [Olson2016] for Python v3.7 was used to perform searches up to 50 generations with population sizes up to 1,000 pipelines.

*Descriptor ranking analysis.* A permutation feature importance approach was used to quantify the importance of each physicochemical and electronic descriptor to its respective QSAR model when predicting PfATP4 inhibitory potency. The importance weight of each descriptor was calculated as the MAE gained (i.e. performance lost) by the model when it predicts with that descriptor column converted to noise via random permutations. Feature importance analyses were performed using eli5 v0.10.1 (<https://www.github.com/TeamHG-Memex/eli5>; accessed October 2019) for Python v3.7, with importance weights computed as the average ΔMAE over 50 permutations for each descriptor.

*Serial descriptor selection.* The dimensionality of the 1,572 physicochemical descriptor training set was reduced in order to obtain a smaller and optimal descriptor-to-molecule ratio and improve model generalisability. A serial descriptor selection protocol was developed by iteratively performing the aforementioned genetic search algorithm and permutation feature importance sequence, with each iteration taking the top 50% ranked descriptors from the previous iteration to form a new QSAR model - similar to a serial dilution process. Five iterations were performed resulting in five smaller physicochemical models with 786, 393, 197, 99, and 50 descriptors.

*Final QSAR model ensemble.* The 50 physicochemical descriptor subset was chosen as it returned the lowest 10-fCV MAE of 0.364 μM. In order to diversify the molecular representations available for QSAR generation, the top 9 electronic descriptors from the original descriptor ranking analysis were also included. Seven final models, three using the 50 physicochemical descriptors and four using the 59 physicochemical and electronic descriptors, were each optimised using TPOT for up to 50 generations with population sizes up to 500 pipelines. The PfATP4 inhibitory potency predictions of the seven pipelines (10-fCV MAE ranging from 0.334-0.385 μM) were ensembled by computing the average.

All descriptor data, modelling code, pipeline details, and experimental outputs from this submission are available at <https://github.com/luiraym/PfATP4-Potency-Predictor>.

# 

# Slade Matthews’s Model

A series representative model types was selected from the different classes of classifier from the Weka stable to cover a wide variety of model types in the search for the best performer. Simple linear regression from the Functions category, Additive regression from the Meta classifier category, M5Rules from the Rules category, and Random Forest from the Trees category were chosen for performance evaluation using 10-fold cross validation. Target values were thresholded to 2.5 µM. Weka preprocessing consisted of removing useless and string type attributes followed by attribute selection using the Pearson correlation to the class in the CorrelationAttributeEval filter selecting for 200 attributes followed by a random sort. The model with the lowest root mean square error (RMSE) and the highest Person product moment (r) in 10-fold CV on the training data was a Random Forest model (Table XXX). This model was saved for evaluation on the test set. Test set predictions on the Random Forest model were submitted to the OSM competition for assessment against held out potency data. All of the predicted potency values were very high, range (1.696 – 2.016), and subsequently the model was not able to predict potent ligands from within the test set. This may be due to the use of a training set with many values set to a ceiling of 2.5 as the maximum relevant value to the modelling problem accounting for 66.4% of the training data. This skewed the model toward predicting higher values. These results show that more complex modelling strategies are required to build a model capable of predicting the binding potency for the *Pf*ATP4 ion pump target.

Table XXXX: Training performance for Weka models built on 200 Mordred descriptors calculated from 3D structures optimized at the semi-empirical PM7 level of theory with implicit water solvation (bold indicates lowest RMSE in training).

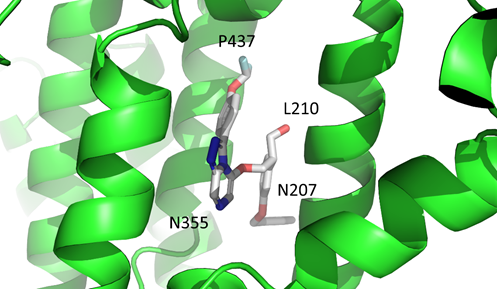
|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Mean Absolute Error | RMSE | Correlation Coefficient |
| Simple Linear Regression | 0.7289 | 0.863 | 0.3576 |
| Additive Regression | 0.6943 | 0.8981 | 0.3934 |
| M5Rules | 0.7059 | 0.9458 | 0.304 |
| Random Forest | 0.6659 | 0.8057 | 0.4958 |

# 

# Ho-Leung Ng’s Model

We sought to build on our relatively successful predictions made in the first round of the competition which used a structure-based approach of modeling inhibitors binding to the presumptive target, *Plasmodium falciparum* ATP4 [Jiménez-Díaz2014], an ATP-driven sodium transporter. Our primary goal was to create a structural model that would be easily interpretable and actionable by chemists. Towards that goal, we refined our ATP4 homology and predicted a consensus ligand binding mode. With these models, we calculated 1D, 2D, and 3D features and predicted the activity of the untested molecules using gradient boosted ensembles [Natekin2013].

To improve the prediction of the drug binding site and poses, we docked the ten most potent Series 4 compounds into our 2016 homology model of ATP4 using smina [Koes2013] with the Vinardo scoring function [Quiroga2016]. Docking was performed with residues within 4 Å of the ligand allowed to sample different rotamers and the smina completeness parameter set to 16. The top three poses for each docked molecule were kept for further analysis. When the top poses were overlaid, we observed one common pose occurred far more than any other. This pose was identified for eight of the ten ligands. It corresponds approximately a 180° rotation from the pose we predicted in 2016. The ligand pose we identified in 2016 was only observed for 2-4 ligands and was the second most frequently occurring pose. The pose for OSM-S-532 (Figure #) was chosen as the template for redocking and scoring the Series 4 molecules.



**Figure #**. Docked pose of OSM-S-532 into homology model of ATP4.

We improved the model of ATP4 bound to OSM-S-532 with multiple rounds of simulated annealing energy minimization in Yasara [Krieger2015].  Models were energy minimized first with the Nova2 force field, with the docked ligand fixed and without solvent to relax side chains at the binding site. This was followed by two rounds of energy minimization using the Yamber3 force field [Krieger2004] with explicit solvation and the ligand free to move.

The energy minimized model of OSM-S-532 bound to ATP4 was used as a reference for redocking the Series 4 molecules using POSIT (OpenEye Scientific Software). POSIT docks molecules favoring poses similar to a reference pose. SAR models were built using pIC50 and pEC50 values. A serious challenge was handling activity data for weakly active molecules, usually notated as IC50 >10 mM.  The POSIT docking classification score, “Great”, “Good”, “Mediocre”, and “None” was used to assign IC50 values to weakly active molecules of 30, 90, 270, and 1000 mM respectively. This is a rather arbitrary assignment but was considered to be better than the initial IC50 assignment of >10 mM. Quantitative 3D docking scores from idock [Li2012] (based on Autodock Vina [Trott2010]) and RF-Score v3 [Li2015] were added as features for SAR scoring. Over 1800 additional conventional 1D, 2D, and 3D SAR features were calculated with Mordred [Moriwaki2018].

To perform regression, we chose to use gradient boosted tree ensembles rather than neural networks. In our experience, neural networks have a higher tendency to overfit when training data is limited and can require extensive trial and error optimization for best performance. Gradient boosted ensemble regression between molecular features and measured activities was performed with XGBoost [Chen2016]. The 10-fold cross validation R2 was 0.33 averaged over ten independent runs.

# Vito Spadavecchio’s Model

# An ensemble machine learning model was created to predict the classification (active or inactive) of compounds in an *in vitro* assay for the presumptive target *Plasmodium falciparum* ATP4. Compounds were represented as SMILES patterns and transformed to 2048-bit fingerprints (ECPF4) using RDKit [Landrum]. The ensemble model was constructed from a classification and regression model. The classification model utilised logistic regression (l2 penalty, *lbfgs* solver) and used the scikit-learn library [Landrum]. The feed-forward neural network regression model (ReLU activation, 4 layers) written with the Keras API (<https://keras.io/>)  and was trained using the Adam optimiser and mean squared error loss function. For both models, compounds in the training set had input IC50 values transformed to Log(IC50). All compounds with *in vitro* activity <1 μM were labelled as ‘active’ and all other compounds labelled as ‘inactive’. A total of 440 compounds were used in the course of  training of both models. The classification model was tuned to minimise the Matthews correlation coefficient (MCC), while the regression model minimised the mean unsigned error (MUE) between prediction and test set. Both models were refined using a 85/15 train/test split across a 10-fold cross validation. The mean MCC for the classification model 0.45 +/- 0.07, and the mean MUE for the regression model was 0.463 +/- 0.043 (95% CI for both errors). For the unlabelled test set, all compounds which were both predicted as ‘active’ by the classifier and had a linear regression prediction less than 0.35 μM were labelled as being ‘active’ in the test set. Both models were written using Python 3.6.

# Exscientia’s Model

Molecular features were generated with ECFP4 fingerprints [Rogers2010]. Due to the small sample size of the training dataset and the curse of dimensionality [Bellman1961] of chemical space (we might be working with 2000+ features), some featurisation preprocessing and model hyperparameter tuning was required to obtain a generalisable model.

Firstly, common bit fragments (fragments that appeared in more than 15 of the 34 test set molecules) were removed from the featurization. Since the test set lies in highly clustered chemical space defined by these common features, removing them prevents the model from overfitting to noise in these features.

Secondly, models were found to only be generalisable when subjected to heavy regularisation. In practice, this meant large regularisation constants in regressions, small tree depth and tree pruning in random forests, and early stopping in boosting models. These results are presented in the table below.

**Table X:** Coefficient of determination (R2) scores for various models

|  |  |  |  |
| --- | --- | --- | --- |
| **Cross-validation split** | **Random Forest** | **AdaBoost** | **Ridge Regression** |
| 0 | 0.381 | 0.276 | 0.332 |
| 1 | 0.368 | 0.246 | 0.404 |
| 2 | 0.303 | 0.273 | 0.190 |
| 3 | 0.368 | 0.368 | 0.353 |
| 4 | 0.167 | 0.189 | 0.292 |
| mean | 0.317 | 0.270 | 0.314 |

aAll models were implemented in scikit-learn.

The final model chosen was a ridge regression with the regularisation parameter set to unity.

# Fig. S1 Overall Synthetic Route for the Preparation of Target Compounds X, X, X and X

# Chemistry

**General Procedure 1: Condensation of hydrazinylpyrazine with aldehyde.** Compound 23 (1 equiv.) was dissolved in EtOH (112 mM). Aldehyde (1 equiv.) was added and the reaction stirred at rt overnight. The suspension was filtered and washed with cold EtOH to give the corresponding hydrazone that was used without further purification.

**General Procedure 2: Cyclisation of hydrazone to triazolopyrazine core.** The product from General Procedure 1 (1 equiv.) was dissolved in CH2Cl2 (112 mM). PhI(OAc)2 (1 equiv.) was added and the reaction stirred at rt overnight. The reaction was quenched with sat. NaHCO3 solution, diluted with CH2Cl2 and the organic layer was separated. The aqueous layer was extracted with CH2Cl2 (2 x) and the combined organic layers were washed with sat. NaHCO3 solution, brine, dried (MgSO4), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding triazolopyrazine core.

**General Procedure 3: Reduction of esters to alcohols.** Ester (1 equiv.) was dissolved in anhydrous THF (566 mM) and cooled to 0 oC. LiAlH4 (1 M in THF, 2 equiv.) was added dropwise and the reaction mixture stirred for 10 min at 0 oC, then at rt. Upon completion, the reaction was diluted with THF and cooled to 0 oC. H2O (1 mL/1 g of LiAlH4) was added followed by 15% aq. NaOH (1 mL/1 g of LiAlH4) and H2O (3 mL/1 g of LiAlH4). The mixture was allowed to warm to rt and stirred for 15 min. MgSO4 was added and the reaction mixture was filtered through a pad of celite and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding alcohol.

**General Procedure 4: Nucleophilic displacement of triazolopyrazine core chlorine with alcohol.** Alcohol (1.0 equiv.) was added to PhMe (168 mM) along with triazolopyrazine core (1.0 equiv.), KOH (3.0 equiv.) and 18-crown-6 (0.1 equiv.). The reaction was stirred at rt until completion as indicated by TLC (100% EtOAc). The reaction was diluted with H2O, then extracted with EtOAc (3 x). The combined organic layers were washed with H2O until the aqueous layer became neutral, followed by brine, dried (MgSO4), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica to give the corresponding ether-linked product.

**2-Chloro-6-hydrazinylpyrazine (23).** 2,6-Dichloropyrazine (14.3 g, 96.1 mmol, 1 equiv.) was dissolved in EtOH (140 mL, 0.69 M). Hydrazine monohydrate (9.43 mL, 192 mmol, 2 equiv.) was added and the reaction heated at reflux overnight. The solvent was removed under reduced pressure and the residue partitioned between H2O and EtOAc. The organic layer was separated and the aqueous layer extracted with EtOAc (3 x). The combined organic layers were washed with brine and concentrated under reduced pressure to give **23** as a yellow powder (12.1 g, 87%). Used without further purification. m.p. 133–135 oC (lit.[1] 136–139 oC); 1H NMR (300 MHz; CDCl3) δ: 8.13 (s, 1H), 7.89 (s, 1H), 6.38 (s, 1H), 3.86 (s, 2H); 13C NMR (75 MHz; CDCl3) δ: 156.4, 146.7, 132.4, 129.0. Spectroscopic data matched those in the literature.[1]

**(*E*)-2-(2-(4-(*tert*-Butoxy)benzylidene)hydrazinyl)-6-chloropyrazine (435).** Prepared according to General Procedure 1 from: compound **23** (750 mg, 5.19 mmol, 1 equiv.) and 4-(*tert*-butoxy)benzaldehyde (904 µL, 5.19 mmol, 1 equiv.) to give **435** as a pearlescent pale yellow powder (1.29 g, 82%). m.p. 207–211 oC; 1H NMR (500 MHz, DMSO-*d6*) δ: 11.47 (s, 1H), 8.53 (s, 1H), 8.03 (s, 1H), 8.02 (s, 1H), 7.65 (d, *J* = 8.6 Hz, 2H), 7.02 (d, *J* = 8.6 Hz, 2H), 1.33 (s, 9H); 13C NMR (126 MHz, DMSO-*d6*) δ: 156.5, 152.4, 145.6, 142.5, 132.1, 129.1, 128.7, 127.6, 123.5, 78.6, 28.6; *m/z* (ESI+) 305 ([M+H]+, 100%).

**(*E*)-2-Chloro-6-(2-(4-isopropylbenzylidene)hydrazinyl)pyrazine (429).** Prepared according to General Procedure 1 from: compound **23** (1.00 g, 6.92 mmol) and 4-isopropylbenzaldehyde (1.05 mL, 6.92 mmol) to give **429** as a yellow powder (1.49 g, 78%). m.p. 231–233 oC; 1H NMR (500 MHz, DMSO-*d6*) δ: 11.51 (s, 1H), 8.54 (s, 1H), 8.04 (s, 2H), 7.65 (d, *J* = 8.2 Hz, 2H), 7.30 (d, *J* = 8.2 Hz, 2H), 2.92 (p, *J* = 6.9 Hz, 1H), 1.21 (d, *J* = 6.9 Hz, 6H); 13C NMR (126 MHz, DMSO-*d6*) δ: 152.4, 150.1, 145.6, 142.8, 132.2, 132.1, 128.7, 126.8, 126.7, 33.4, 23.7; *m/z* (ESI+) 275 ([M+H]+, 100%).

**(*E*)-2-Chloro-6-(2-(4-chlorobenzylidene)hydrazinyl)pyrazine (433).** Prepared according to General Procedure 1 from: compound **23** (1.00 g, 6.92 mmol) and 4-chlorobenzaldehyde (972 mg, 6.92 mmol) to give **433** as a yellow powder (1.53 g, 82 %). m.p. 230–233 oC;1H NMR (500 MHz, DMSO-*d6*) δ: 11.64 (s, 1H), 8.58 (s, 1H), 8.07 (s, 1H), 8.05 (s, 1H), 7.77 (d, *J* = 8.5 Hz, 2H), 7.48 (d, *J* = 8.5 Hz, 2H); 13C NMR (126 MHz, DMSO-*d6*) δ: 152.2, 145.5, 141.3, 133.8, 133.4, 132.6, 128.9, 128.3 (one obscured signal); *m/z* (ESI+) 267 ([M+H]+, 100%).

**(*E*)-2-Chloro-6-(2-(4-(difluoromethoxy)benzylidene)hydrazinyl)pyrazine (24).** Prepared according to General Procedure 1 from: compound **23** (2.00 g, 13.8 mmol, 1 equiv.) and 4-(difluoromethoxy)benzaldehyde (1.83 mL, 13.8 mmol, 1 equiv.) to give **24** as a pale yellow powder (2.56 g, 62%). m.p. 197–200 oC; ﻿1H NMR (300 MHz, DMSO-*d6*) δ: 11.55 (s, 1H), 8.56 (s, 1H), 8.05 (d, *J =* 5.2 Hz, 2H), 7.79 (d, *J =* 8.5 Hz, 2H), 7.29 (t, *J =* 73.9 Hz, 1H), 7.22 (d, *J =* 8.4 Hz, 2H); 13C NMR (75 MHz, DMSO-*d6*) δ: 152.3, 151.7, 145.6, 141.6, 132.4, 131.5, 128.8, 128.3, 118.8, 116.2 (t, *J =* 257.9 Hz); *m/z* (ESI+) 299 ([M+H]+, 100%).

**3-(4-(*tert*-Butoxy)phenyl)-5-chloro-[1,2,4]triazolo[4,3-*a*]pyrazine (437).** Prepared according to General Procedure 2 from: compound **435** (750 mg, 2.46 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **437** as a pale yellow powder (595 mg, 80%). m.p. 143–147 oC;1H NMR (500 MHz, CDCl3) δ: 9.31 (s, 1H), 7.85 (s, 1H), 7.52 (d, *J* = 8.7 Hz, 2H), 7.12 (d, *J* = 8.7 Hz, 2H), 1.43 (s, 9H); 13C NMR (126 MHz, CDCl3) δ: 158.2, 148.4, 147.3, 143.1, 132.4, 129.8, 123.0, 122.2, 121.0, 79.7, 29.0; *m/z* (ESI+) 303 ([M+H]+, 100%).

**5-Chloro-3-(4-isopropylphenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (432).** Prepared according to General Procedure 2 from: compound **429** (750 mg, 2.73 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **432** as a pale orange powder (662 mg, 89 %). m.p. 133–137 oC;1H NMR (500 MHz, CDCl3) δ: 9.32 (s, 1H), 7.85 (s, 1H), 7.54 (d, *J* = 8.3 Hz, 2H), 7.37 (d, *J* = 8.1 Hz, 2H), 3.02 (p, *J* = 6.9 Hz, 1H), 1.32 (d, *J* = 6.9 Hz, 6H); 13C NMR (126 MHz, CDCl3) δ: 152.0, 148.6, 147.3, 143.1, 131.5, 129.7, 126.2, 124.0, 122.2, 34.3, 24.0; *m/z* (ESI+) 273 ([M+H]+, 100%).

**5-Chloro-3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (89).** Prepared according to General Procedure 2 from: compound **433** (750 mg, 2.81 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **89** as an orange powder (634 mg, 85%). m.p. 180–183 oC;1H NMR (300 MHz, CDCl3) δ: 9.29 (s, 1H), 7.86 (s, 1H), 7.55 (d, *J* = 8.6 Hz, 2H), 7.48 (d, *J* = 8.5 Hz, 2H); 13C NMR (75 MHz, CDCl3) δ: 147.2, 142.9, 137.2, 132.6, 129.8, 128.3, 125.1, 121.8 (one obscured signal); *m/z* (ESI+) 265 ([M+H]+, 100%).

**5-Chloro-3-(4-(difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (25).** Prepared according to General Procedure 2 from: compound **24** (2.00 g, 6.70 mmol). The crude product was dissolved in EtOH (~300 mL) and H2O was added to precipitate the product. The solid was filtered, washed with H2O, then Et2O and dried *in vacuo* to give **25** as pale yellow needles (1.08 g, 54%). m.p. 130–133 oC; ﻿1H NMR (300 MHz, CDCl3) δ: 9.34 (s, 1H), 7.88 (s, 1H), 7.65 (d, *J =* 7.6 Hz, 2H), 7.27 (d, *J =* 7.7 Hz, 2H), 6.64 (t, *J =* 73.1 Hz, 1H); 13C NMR (75 MHz, CDCl3) δ: 153.2, 147.4, 147.3, 143.2, 133.2, 130.0, 123.8, 122.0, 118.8, 115.6 (t, *J =* 261.4 Hz); *m/z* (ESI+) 297 ([M+H]+, 100%).

**Methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (434).** 2-(3,4-Difluorophenyl)-2-hydroxyacetic acid (850 mg, 4.52 mmol, 1 equiv.) and *p*-TsOH monohydrate (17.2 mg, 0.09 mmol, 0.02 equiv.) were dissolved in MeOH (3.12 mL, 1.45 M) and the reaction heated to reflux (80 oC). The reaction was cooled to rt and the solvent removed. EtOAc was added to the residue and the organic layer washed with H2O, sat. NaHCO3 solution, brine, dried (MgSO4), filtered and concentrated under reduced pressure to give **434** as clear colourless oil that solidified on standing (738 mg, 81%). No further purification required. m.p. 44–49 oC; 1H NMR (500 MHz, CDCl3) δ: 7.30 – 7.24 (m, 1H), 7.23 – 7.02 (m, 2H), 5.14 (d, *J* = 5.0 Hz, 1H), 3.78 (s, 3H), 3.52 (d, *J* = 5.1 Hz, 1H); 13C NMR (126 MHz, CDCl3) δ: 173.6, 151.5 (d, *J* = 12.6 Hz), 149.5 (d, *J* = 12.6 Hz), 135.1, 122.8, 117.5 (d, *J* = 17.5 Hz), 115.8 (d, *J* = 18.4 Hz), 71.8, 53.5; *m/z* (ESI+) 225 ([M+Na]+, 100%).

**Methyl 2-(3,4-difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)acetate (438).** Compound **434** (600 mg, 2.97 mmol, 1.0 equiv.) was dissolved in CH2Cl2 (10.7 mL, 277 mM), *p*-TsOH (102 mg, 0.59 mmol, 0.2 equiv.) and 3,4-dihydro-2*H*-pyran (0.3 mL, 3.26 mmol, 1.1 equiv.) were added and the reaction stirred at rt. The reaction was quenched with ice cold H2O and the organic layer separated. The aqueous layer was extracted with CH2Cl2 (3 x) and the combined organic layers washed with H2O, brine, dried (MgSO4), filtered and concentrated under reduced pressure to give the crude product, which purified by automated flash chromatography on silica (6–50% ethyl acetate in hexanes) to give **438** as a viscous dark orange oil (395 mg, 46%). 1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers) δ: 7.38 – 7.28 (m, 2H), 7.24 – 7.10 (m, 4H), 5.27 (s, 1H), 5.18 (s, 1H), 4.86 (t, *J* = 3.0 Hz, 1H), 4.57 (t, *J* = 3.4 Hz, 1H), 3.72 (s, 6H), 3.55 – 3.44 (m, 4H), 1.93 – 1.36 (m, 12H); 13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers) δ: 171.3, 170.7, 150.7 (dd, *J* = 249.4, 12.0 Hz), 150.6 (dd, *J* = 249.3, 13.1 Hz), 150.5 (dd, *J* = 249.3, 13.4 Hz), 150.4 (dd, *J* = 248.5, 12.6 Hz), 133.8 (dd, *J* = 5.4, 4.2 Hz), 133.7 – 132.9 (m), 123.8 (dd, *J* = 6.6, 3.6 Hz), 123.4 (dd, *J* = 6.4, 3.7 Hz), 117.5 (t, *J* = 18.1 Hz, 2C), 116.6 (d, *J* = 18.0 Hz), 116.4 (d, *J* = 18.4 Hz), 97.4, 97.0, 75.8, 74.5, 62.5, 62.3, 52.62, 52.58, 30.3, 30.2, 25.3 (2C), 19.1, 18.8; *m/z* (ESI+) 309 ([M+Na]+, 100%).

**2-(3,4-Difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)ethan-1-ol (439).** Prepared according to General Procedure 3 from: compound **438** (300 mg, 1.05 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **439** as a viscous pale yellow oil (150 mg, 55%). 1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers) δ: 7.28 – 6.92 (m, 6H), 4.92 – 4.81 (m, 1H), 4.81 – 4.72 (m, 1H), 4.68 (dd, *J* = 6.8, 4.6 Hz, 1H), 4.50 (dd, *J* = 5.6, 2.8 Hz, 1H), 4.00 (dt, *J* = 11.0, 5.2 Hz, 1H), 3.73 – 3.62 (m, 4H), 3.56 (tt, *J* = 10.2, 4.6 Hz, 2H), 3.32 (dt, *J* = 10.9, 4.7 Hz, 1H), 3.05 – 2.96 (m, 1H), 2.18 – 2.07 (m, 1H), 1.91 – 1.36 (m, 12H); 13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers) δ: 150.5 (dd, *J* = 248.8, 12.8 Hz), 150.4 (dd, *J* = 248.2, 12.7 Hz), 150.1 (dd, *J* = 248.3, 12.7 Hz), 149.9 (dd, *J* = 247.6, 12.6 Hz), 137.7 – 136.7 (m), 136.3 – 135.5 (m), 122.9 (dd, *J* = 6.3, 3.6 Hz), 122.7 (dd, *J* = 6.3, 3.6 Hz), 117.4 (d, *J* = 17.3 Hz), 117.2 (d, *J* = 17.2 Hz), 115.9 (d, *J* = 17.7 Hz), 115.8 (d, *J* = 17.8 Hz), 99.6, 98.1, 79.6, 78.8, 67.5, 66.5, 63.9, 62.9, 31.1, 30.7, 25.30, 25.26, 20.3, 19.6; *m/z* (ESI+) 281 ([M+Na]+, 100%).

**(6-Methylpyridin-3-yl)methanol (430).** Prepared according to General Procedure 3 from: 6-methylnicotinic acid (750 mg, 5.47 mmol); purified by automated flash chromatography on silica (1–10% MeOH in CH2Cl2) to give **430** as a yellow oil (103 mg, 15%). 1H NMR (500 MHz, CDCl3) δ: 8.42 (d, *J* = 2.0 Hz, 1H), 7.61 (dd, *J* = 7.9, 2.2 Hz, 1H), 7.14 (d, *J* = 7.9 Hz, 1H), 4.66 (s, 2H), 2.53 (s, 3H) (alcohol OH signal not seen); 13C NMR (126 MHz, CDCl3) δ: 157.6, 147.8, 135.8, 133.6, 123.3, 62.8, 30.1; *m/z* (ESI+) 146 ([M+Na]+, 100%). Spectroscopic data matched those in the literature.[2]

***tert*-Butyl (3-(2-hydroxyethyl)benzyl)carbamate (452).** Prepared according to General Procedure 3 from: 2-(3-(((*tert*-butoxycarbonyl)amino)methyl)phenyl)acetic acid (500 mg, 1.88 mmol); purified by automated flash chromatography on silica (1–10% MeOH in CH2Cl2) to give **452** as a viscous clear colourless oil (332 mg, 70%). 1H NMR (500 MHz, CDCl3) δ: 7.27 (t, *J* = 7.8 Hz, 1H), 7.18 – 7.05 (m, 3H), 4.90 (br s, 1H), 4.29 (br d, *J* = 5.5 Hz, 2H), 3.84 (t, *J* = 6.6 Hz, 2H), 2.84 (t, *J* = 6.6 Hz, 2H), 1.46 (s, 9H) (alcohol OH signal not seen); 13C NMR (126 MHz, CDCl3) δ: 156.1, 139.3, 139.1, 129.0, 128.2, 128.1, 125.7, 79.7, 63.7, 44.7, 39.2, 28.5; *m/z* (ESI+) 274 ([M+Na]+, 100%).

**3-(4-(*tert*-Butoxy)phenyl)-5-(2-(3,4-difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)ethoxy)-[1,2,4]triazolo[4,3-*a*]pyrazine (444).** Prepared according to General Procedure 4 from: compound **439** (100 mg, 0.39 mmol) and compound **437** (117 mg, 0.39 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **444** as an orange powder (95.9 mg, 47%). m.p. 72–75 oC; 1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers) δ: 9.02 (s, 1H), 9.01 (s, 1H), 7.62 (dd, *J* = 8.6, 1.5 Hz, 4H), 7.38 (s, 1H), 7.29 (s, 1H), 7.14 (dd, *J* = 8.6, 1.5 Hz, 4H), 7.09 – 7.01 (m, 1H), 7.00 (dt, *J* = 9.9, 8.2 Hz, 1H), 6.95 (ddd, *J* = 10.6, 7.6, 2.1 Hz, 1H), 6.80 (dq, *J* = 6.4, 2.0 Hz, 1H), 6.73 (ddd, *J* = 10.2, 7.6, 2.1 Hz, 1H), 6.60 (dq, *J* = 6.1, 1.9 Hz, 1H), 4.89 (t, *J* = 5.5 Hz, 1H), 4.70 (dd, *J* = 6.7, 4.8 Hz, 1H), 4.44 (dd, *J* = 9.9, 5.3 Hz, 1H), 4.37 (q, *J* = 5.1, 4.3 Hz, 2H), 4.29 (dd, *J* = 9.8, 6.9 Hz, 1H), 4.22 (ddd, *J* = 19.1, 9.9, 5.3 Hz, 2H), 3.77 (ddd, *J* = 10.7, 7.1, 3.3 Hz, 1H), 3.47 – 3.40 (m, 1H), 3.37 (ddd, *J* = 11.7, 9.1, 3.0 Hz, 1H), 3.25 (dt, *J* = 11.0, 4.6 Hz, 1H), 1.87 – 1.45 (m, 12H), 1.44 (s, 9H), 1.43 (s, 9H); 13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers) δ: 157.8, 157.7, 150.5 (dd, *J* = 249.7, 12.6 Hz), 150.42 (dd, *J* = 250.0, 12.5 Hz), 150.36 (dd, *J* = 249.0, 13.0 Hz), 150.2 (dd, *J* = 248.7, 12.5 Hz), 147.90, 147.88, 147.3, 147.2, 144.04, 143.99, 137.1, 137.0, 136.2 – 135.5 (m), 134.6 (t, *J* = 4.2 Hz), 131.7, 131.6, 123.3 (dd, *J* = 6.3, 3.5 Hz), 123.1, 123.0, 122.6 (dd, *J* = 6.2, 3.6 Hz), 122.3, 122.1, 117.6 (d, *J* = 17.4 Hz), 117.4 (d, *J* = 17.4 Hz), 116.1 (d, *J* = 17.7 Hz), 115.8 (d, *J* = 17.9 Hz), 109.1, 108.7, 99.3, 96.6, 79.6, 74.0, 73.7, 73.4, 63.2, 62.4, 30.6, 30.5, 29.04, 29.03, 25.3, 25.2, 19.7, 19.1 (two obscured signals); *m/z* (ESI+) 525 ([M+H]+, 100%).

**2-((3-(4-(*tert*-Butoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)-1-(3,4-difluorophenyl)ethan-1-ol (445).** Compound **444** (70.0 mg, 0.13 mmol, 1 equiv.) was dissolved in EtOH (1.63 mL, 82 mM). CuCl2.2H2O (1.14 mg, 6.67 µmol, 5 mol%) was added and the reaction heated at reflux. The solvent was removed and EtOAc was added. The mixture was washed with H2O (3 x), then brine, and the organic layer was dried (MgSO4), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **445** as a pale orange powder (49.2 mg, 84%). m.p. 75–79 oC; 1H NMR (500 MHz, CDCl3) δ: 9.05 (s, 1H), 7.66 (d, *J* = 8.6 Hz, 2H), 7.17 (d, *J* = 8.6 Hz, 2H), 7.17 – 7.06 (m, 1H), 7.03 (ddd, *J* = 10.2, 7.4, 1.6 Hz, 1H), 6.93 – 6.87 (m, 1H), 4.78 (dt, *J* = 7.4, 3.3 Hz, 1H), 4.26 (dd, *J* = 9.3, 3.3 Hz, 1H), 4.19 – 4.08 (m, 1H), 2.00 (d, *J* = 3.7 Hz, 1H), 1.42 (s, 9H) (alcohol OH signal not seen); 13C NMR (126 MHz, CDCl3) δ: 158.0, 150.6 (dd, *J* = 249.5, 12.7 Hz), 150.4 (dd, *J* = 249.7, 12.6 Hz), 147.8, 146.9, 143.8, 137.3, 135.2 – 135.0 (m), 131.7, 123.2, 122.6, 122.2 (dd, *J* = 6.4, 3.6 Hz), 117.8 (d, *J* = 17.4 Hz), 115.4 (d, *J* = 18.2 Hz), 108.5, 79.9, 75.0, 70.6, 29.0; *m/z* (ESI+) 441 ([M+H]+, 100%); HRMS (ESI+) found 441.1739 [M+H]+, C23H22F2N4O3H+ requires 441.1738.

**3-(4-Isopropylphenyl)-5-((6-methylpyridin-3-yl)methoxy)-[1,2,4]triazolo[4,3-*a*]pyrazine (440).** Prepared according to General Procedure 4 from: compound **430** (100 mg, 0.81 mmol) and compound **432** (221 mg, 0.81 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give a white powder (181 mg, 62%). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H2O) to give a white powder (40.6 mg, 14%). Repurified by automated reversed-phase flash chromatography on silica (5–75% MeOH in H2O) to give **440** as a white powder (16.3 mg, 6%). m.p. decomposed >150 oC; 1H NMR (400 MHz, CD3OD) δ: 9.00 (s, 1H), 8.33 (d, *J* = 2.0 Hz, 1H), 7.65 (s, 1H), 7.54 – 7.48 (m, 3H), 7.23 (d, *J* = 8.0 Hz, 1H), 7.12 (d, *J* = 8.2 Hz, 2H), 5.32 (s, 2H), 2.87 (p, *J* = 6.9 Hz, 1H), 2.53 (s, 3H), 1.20 (d, *J* = 6.9 Hz, 6H); 13C NMR (101 MHz, CD3OD) δ: 160.0, 153.0, 152.3, 149.9, 148.9, 146.0, 139.2, 136.6, 131.9, 128.7, 126.6, 125.8, 124.8, 110.2, 71.3, 35.2, 24.2, 23.7; *m/z* (ESI+) 360 ([M+H]+, 100%); HRMS (ESI+) found 360.1832 [M+H]+, C21H21N5OH+ requires 360.1824.

***tert*-Butyl (3-(2-((3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)ethyl)benzyl)carbamate (453).** Prepared according to General Procedure 4 from: compound **452** (150 mg, 0.60 mmol) and compound **89** (158 mg, 0.60 mmol); purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give **453** as a light brown powder (209 mg, 73%). 1H NMR (500 MHz, CDCl3) δ: 8.98 (s, 1H), 7.59 (d, *J* = 8.5 Hz, 2H), 7.39 (d, *J* = 8.5 Hz, 2H), 7.30 (s, 1H), 7.19 (t, *J* = 7.5 Hz, 1H), 7.14 (d, *J* = 7.6 Hz, 1H), 6.86 (br s, 1H), 6.75 (d, *J* = 7.4 Hz, 1H), 4.86 (br s, 1H), 4.43 (t, *J* = 6.6 Hz, 2H), 4.22 (d, *J* = 5.6 Hz, 2H), 2.94 (t, *J* = 6.5 Hz, 2H), 1.43 (s, 9H); 13C NMR (126 MHz, CDCl3) δ: 156.0, 147.9, 146.3, 143.9, 139.7, 136.6, 136.5, 136.4, 132.1, 129.1, 128.1, 127.7, 127.6, 126.33, 126.26, 108.5, 79.7, 71.2, 44.6, 34.5, 28.5; *m/z* (ESI+) 480 ([M+H]+, 100%), 502 ([M+Na]+, 52%).

**(3-(2-((3-(4-Chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)ethyl)phenyl)methanamine (454).** Compound **453** (150 mg, 0.31 mmol, 1.00 equiv.) was dissolved in CH2Cl2 (0.91 mL, 345 mM). TFA (0.27 mL, 3.50 mmol, 11.2 equiv.) was added and the reaction stirred at rt overnight. The solvent was removed and the residue directly purified by automated flash chromatography on silica (1–10% MeOH in CH2Cl2, then 100% MeOH) to give a sticky brown solid (135 mg, >100%). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H2O) to give **454** as a white powder (77.0 mg, 65%). m.p. 74–77 oC; 1H NMR (400 MHz, CD3OD) δ: 8.97 (s, 1H), 7.68 (d, *J* = 8.5 Hz, 2H), 7.54 (s, 1H), 7.50 (d, *J* = 8.5 Hz, 2H), 7.32 – 7.25 (m, 2H), 7.07 (s, 1H), 6.98 – 6.89 (m, 1H), 4.60 (t, *J* = 6.2 Hz, 2H), 4.03 (s, 2H), 3.02 (t, *J* = 6.2 Hz, 2H) (amine NH2 signal not seen); 13C NMR (126 MHz, CDCl3) δ: 149.0, 147.6, 145.9, 139.7, 137.5, 136.2, 134.7, 133.5, 130.6, 130.4, 130.1, 129.2, 128.1, 127.6, 110.0, 72.4, 44.2, 35.2; *mj/z* (ESI+) 380 ([M+H]+, 100%); HRMS (ESI+) found 380.1282 [M+H]+, C20H18ClN5OH+ requires 380.1278.

**1-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)propan-2-ol (450).** To a mixture of lactic acid (0.83 mL, 11.1 mmol, 1.0 equiv.), 3,4-dihydro-2*H*-pyran (3.04 mL, 33.3 mmol, 3.0 equiv.) and CH2Cl2 (2.98 mL, 3.73 M) in an ice bath was added pyridinium *p*-toluenesulfonate (279 mg, 1.11 mmol, 0.1 equiv.) and pyridine (1 drop). The reaction was stirred at rt overnight. CH2Cl2 was added and the solution washed with 5% NaHCO3, H2O (2 x), dried (MgSO4), filtered and concentrated under reduced pressure to give a mixture of **X** and **X** as a clear colourless oil (1.09 g, 38%). Used without further purification. *m/z* (ESI+) 209 (**X**, [M+Na]+, 100%); *m/z* (ESI+) 281 (**X**, [M+Na]+, 100%). The mixture of **X** and **X** (700 mg, 2.71 mmol) was subjected to General Procedure 3 and purified by automated flash chromatography on silica (12–100% EtOAc to hexanes) to give a ~1:2 mixture of **X** and **X** as a clear colourless oil (96.3 mg, 17%). 1H NMR (500 MHz, CDCl3) δ: 4.72 (dd, *J* = 5.1, 2.8 Hz, 1H), 4.56 (dd, *J* = 4.3, 2.8 Hz, 2H), 3.96 – 3.90 (m, 1H), 3.86 (dtd, *J* = 11.2, 7.6, 7.1, 3.5 Hz, 3H), 3.74 (dt, *J* = 9.6, 6.7 Hz, 2H), 3.64 (t, *J* = 6.5 Hz, 4H), 3.58 (dd, *J* = 11.6, 3.5 Hz, 1H), 3.54 – 3.43 (m, 4H), 3.39 (dt, *J* = 9.6, 6.5 Hz, 2H), 1.92 – 1.33 (m, 33H), 1.21 (d, *J* = 6.4 Hz, 3H); 13C NMR (126 MHz, CDCl3) δ: 99.2, 99.1, 75.1, 67.6, 66.3, 63.4, 63.0, 62.6, 32.7, 31.2, 30.9, 29.6, 25.6, 25.5, 22.6, 20.2, 19.8, 17.8; *m/z* (ESI+) 183 (**X**, [M+Na]+, 100%); *m/z* (ESI+) 211 (**X**, [M+Na]+, 100%). The mixture of **X** and **X** (70.0 mg, 0.44 mmol) was subjected to General Procedure 4 with compound **25** (130 mg, 0.44 mmol) and purified by automated flash chromatography on silica (25–100% EtOAc in hexanes) to give a ~1:1 mixture of **X** and **X** as a brown powder (87.0 mg, 47%). 1H NMR (500 MHz, CDCl3) δ: 9.05 (s, 1H), 9.03 (s, 1H), 7.73 (t, *J* = 9.0 Hz, 4H), 7.33 (s, 1H), 7.29 (s, 1H), 7.29 – 7.20 (m, 4H), 6.64 (t, *J* = 73.3 Hz, 1H), 6.63 (t, *J* = 73.3 Hz, 1H), 4.52 (t, *J* = 3.7 Hz, 1H), 4.24 – 4.14 (m, 3H), 4.16 – 4.09 (m, 2H), 3.97 – 3.88 (m, 1H), 3.82 (dtd, *J* = 11.7, 8.8, 8.2, 3.8 Hz, 2H), 3.66 (dt, *J* = 9.7, 6.6 Hz, 1H), 3.54 – 3.40 (m, 2H), 3.29 (dt, *J* = 9.5, 6.2 Hz, 1H), 1.85 – 1.30 (m, 18H), 1.07 (d, *J* = 6.5 Hz, 3H); 13C NMR (126 MHz, CDCl3) δ: 152.7, 152.5, 147.91, 147.88, 146.3, 144.3, 144.2, 136.6, 136.3, 132.6, 132.5, 125.0, 124.9, 118.9, 118.5, 115.9 (t, *J* = 260.7 Hz), 115.7 (t, *J* = 261.1 Hz), 108.5, 108.2, 99.2, 98.7, 74.0, 71.0, 70.8, 67.2, 62.7, 62.4, 53.6, 30.85, 30.82, 29.2, 28.4, 25.5, 25.4, 22.8, 19.9, 19.3, 18.2; *m/z* (ESI+) 421 (**X**, [M+H]+, 100%); *m/z* (ESI+) 449 (**X**, [M+H]+, 100%). The mixture of **X** and **X** (56.0 mg, 0.13 mmol, 1 equiv.) was dissolved in EtOH (1.62 mL, 82 mM). CuCl2.2H2O (1.14 mg, 6.66 µmol, 5 mol%) was added and the reaction heated at reflux. The solvent was removed and EtOAc was added. The mixture was washed with H2O (3 x), then brine, and the organic layer was dried (MgSO4), filtered and concentrated under reduced pressure to give the crude product, which was purified by automated flash chromatography on silica (1–15% MeOH in CH2Cl2) to give a brown oil (39.8 mg, 89%). Repurified by automated reversed-phase flash chromatography on silica (5–100% MeOH in H2O) to give **450** as a white powder (10.3 mg, 23%). m.p. 137–140 oC; 1H NMR (400 MHz, CDCl3) δ: 9.07 (s, 1H), 7.75 (d, *J* = 8.7 Hz, 2H), 7.32 (s, 1H), 7.29 (d, *J* = 8.6 Hz, 2H), 6.62 (t, *J* = 73.0 Hz, 1H), 4.18 – 4.12 (m, 1H), 4.04 – 3.90 (m, 2H), 1.06 (d, *J* = 6.3 Hz, 3H); 13C NMR (126 MHz, CDCl3) δ: 152.5, 147.9, 146.1, 144.0, 136.9, 132.5, 125.3, 119.1, 115.6 (t, *J* = 262.1 Hz), 108.5, 75.6, 65.5, 18.7; *m/z* (ESI+) 337 ([M+H]+, 100%); HRMS (ESI+) found 337.1107 [M+H]+, C15H14F2N4O3H+ requires 337.1112.

**5-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)pentan-1-ol (450rpf3).** Isolated from the same reaction as for 450 to give 450rpf3 as a white powder (16.1 mg, 36%). m.p. 104–108 oC; 1H NMR (500 MHz, CDCl3) δ: 9.01 (s, 1H), 7.70 (d, *J* = 8.7 Hz, 2H), 7.28 (s, 1H), 7.24 (d, *J* = 8.6 Hz, 2H), 6.65 (t, *J* = 73.2 Hz, 1H), 4.20 (t, *J* = 6.0 Hz, 2H), 3.54 (t, *J* = 6.3 Hz, 2H), 1.79 – 1.54 (m, 2H), 1.53 – 1.30 (m, 2H), 1.14 (ddd, *J* = 11.7, 4.6, 2.5 Hz, 2H); 13C NMR (126 MHz, CDCl3) δ: 152.5, 147.9, 146.3, 144.3, 136.3, 132.6, 125.1, 118.6, 115.7 (t, *J* = 261.3 Hz), 108.2, 71.0, 62.5, 32.0, 28.5, 22.3; *m/z* (ESI+) 365 ([M+H]+, 100%); HRMS (ESI+) found 365.1424 [M+H]+, C17H18F2N4O3H+ requires 365.1425.

# 1H and 13C NMR Spectra

**2-Chloro-6-hydrazinylpyrazine (23)**

1H NMR (300 MHz, CDCl3)



13C NMR (75 MHz, CDCl3)



**(*E*)-2-(2-(4-(*tert*-Butoxy)benzylidene)hydrazinyl)-6-chloropyrazine (435)**

1H NMR (500 MHz, DMSO-*d6*)



****

13C NMR (126 MHz, DMSO-*d6*)

****

****

**(*E*)-2-Chloro-6-(2-(4-isopropylbenzylidene)hydrazinyl)pyrazine (429)**

1H NMR (500 MHz, DMSO-*d6*)



****

13C NMR (126 MHz, DMSO-*d6*)

****

****

**(*E*)-2-Chloro-6-(2-(4-chlorobenzylidene)hydrazinyl)pyrazine (433)**

1H NMR (500 MHz, DMSO-*d6*)





13C NMR (126 MHz, DMSO-*d6*)

****



**(*E*)-2-Chloro-6-(2-(4-(difluoromethoxy)benzylidene)hydrazinyl)pyrazine (24)**

1H NMR (300 MHz, DMSO-*d6*)





13C NMR (75 MHz, DMSO-*d6*)

****



**3-(4-(*tert*-Butoxy)phenyl)-5-chloro-[1,2,4]triazolo[4,3-*a*]pyrazine (437)**

1H NMR (500 MHz, CDCl3)



13C NMR (126 MHz, CDCl3)

****



**5-Chloro-3-(4-isopropylphenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (432)**

1H NMR (500 MHz, CDCl3)





13C NMR (126 MHz, CDCl3)

****



**5-Chloro-3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (89)**

1H NMR (300 MHz, CDCl3)





13C NMR (75 MHz, CDCl3)

****



**5-Chloro-3-(4-(difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazine (25)**

1H NMR (300 MHz, CDCl3)





13C NMR (75 MHz, CDCl3)





**Methyl 2-(3,4-difluorophenyl)-2-hydroxyacetate (434)**

1H NMR (500 MHz, CDCl3)



****

13C NMR (126 MHz, CDCl3)

****

****

**Methyl 2-(3,4-difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)acetate (438)**

1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers)



****

13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers)

****

****

**2-(3,4-Difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)ethan-1-ol (439)**

1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers)



****

13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers)

****

****

**(6-Methylpyridin-3-yl)methanol (430)**

1H NMR (500 MHz, CDCl3)





13C NMR (126 MHz, CDCl3)

******



***tert*-Butyl (3-(2-hydroxyethyl)benzyl)carbamate (452)**

1H NMR (500 MHz, CDCl3)

****

13C NMR (126 MHz, CDCl3)

****

**2-((Tetrahydro-2*H*-pyran-2-yl)oxy)propan-1-ol (456)**

1H NMR (500 MHz, CDCl3)





+

13C NMR (126 MHz, CDCl3)

****



+

**3-(4-(*tert*-Butoxy)phenyl)-5-(2-(3,4-difluorophenyl)-2-((tetrahydro-2*H*-pyran-2-yl)oxy)ethoxy)-[1,2,4]triazolo[4,3-*a*]pyrazine (444)**

1H NMR (500 MHz, CDCl3, present as a mixture of diastereomers)



****13C NMR (126 MHz, CDCl3, present as a mixture of diastereomers)



**2-((3-(4-(*tert*-Butoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)-1-(3,4-difluorophenyl)ethan-1-ol (445)**

1H NMR (500 MHz, CDCl3)



13C NMR (126 MHz, CDCl3)

****



**3-(4-Isopropylphenyl)-5-((6-methylpyridin-3-yl)methoxy)-[1,2,4]triazolo[4,3-*a*]pyrazine (440)**

1H NMR (400 MHz, CD3OD)



******13C NMR (101 MHz, CD3OD)



***tert*-Butyl (3-(2-((3-(4-chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)ethyl)benzyl)carbamate (453)**

1H NMR (500 MHz, CDCl3)



13C NMR (126 MHz, CDCl3)



**(3-(2-((3-(4-Chlorophenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)ethyl)phenyl)methanamine (454)**

1H NMR (400 MHz, CD3OD)



****13C NMR (126 MHz, CD3OD)



**3-(4-(Difluoromethoxy)phenyl)-5-(2-((tetrahydro-2*H*-pyran-2-yl)oxy)propoxy)-[1,2,4]triazolo[4,3-*a*]pyrazine (457)**

1H NMR (500 MHz, CDCl3)



+

****13C NMR (126 MHz, CDCl3)



+

**1-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)propan-2-ol (450)**

1H NMR (400 MHz, CDCl3)



13C NMR (126 MHz, CDCl3)

****



**5-((3-(4-(Difluoromethoxy)phenyl)-[1,2,4]triazolo[4,3-*a*]pyrazin-5-yl)oxy)pentan-1-ol (450rpf3)**

1H NMR (500 MHz, CDCl3)

****

******13C NMR (126 MHz, CDCl3)

****

# LCMS Traces for the Intermediates of Compound 450



[M+Na]+

[M+Na]+



[M+Na]+

[M+Na]+



[M+H]+

[M+H]+

# References

[1] Bradac, J. *et al.* Heterocycles. 167. Telesubstitution and Other Transformations of Imidazo[1,2-a]- and s-Triazolo[4,3-a]pyrazines. *J. Org. Chem.* **42**, 4197–4201 (1977).

[2] Kyba, E. P., Liu, S. T., Chockalingam, K. & Reddy, B. R. A General Synthesis of Substituted Fluorenones and Zazfluorenones. *J. Org. Chem.* **53**, 3513–3521 (1988).

[Cardoso-Silva2019] Cardoso-Silva, J., Papageorgiou, L. G. & Tsoka, S. Network-Based Piecewise Linear Regression for QSAR Modelling. *J. Comput. Aided Mol. Des.* **33**, 831–844 (2019).

[Landrum] Landrum, G. RDKit: Open-source cheminformatics. Available at: [https://www.rdkit.org](https://www.rdkit.org/). (Accessed: October 2019).

[Willighagen2017] Willighagen E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminform.* **9**, No. 33 (2017).

[Cardoso-Silva2019-b] Cardoso-Silva J., Papadatos G., Papageorgiou L. G. & Tsoka S. Optimal Piecewise Linear Regression Algorithm for QSAR Modelling. *Mol. Inf.* **38**, 1800028 (2019).

[Tropsha2010] Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inform.* **29**, 476–488 (2010).

[Schmidt2017] Schmidt, M., Le Roux, N. & Bach, F. Minimizing Finite Sums with the Stochastic Average Gradient. *Math. Program.* **162**, 83–112 (2017).

[Dragos2009] Dragos, H., Gilles, M. & Alexandre, V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J. Chem. Inf. Model.* **49**, 1762–1776 (2009).

[Garciarena2018] Garciarena, U., Santana, R. & Mendiburu, A. Expanding Variational Autoencoders for Learning and Exploiting Latent Representations in Search Distributions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, Kyoto, July 15-19, 2018; 849–856.

[armchem2019] Tool for creating a deep learning model for predicting properties of chemical compounds based on SMILES. Available at: <https://github.com/mmgalushka/armchem>. (Accessed: August 2019).

[Davies2015] Davies, M. *et al.* ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **43**, W612–W620 (2015).

[Norouzi2012] Norouzi, M., Fleet, D. J. & Salakhutdinov, R. R. Hamming distance metric learning. In *Advances in Neural Information Processing Systems 25*; Pereira, F.; Burges, C. J. C.; Bottou, L.; Weinberger, K. Q., Eds.; Curran Associates, Inc., 1061–1069 (2012).

[Miller2019] Miller, F. P., Vandome, A. F. & McBrewster, J. Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau-Levenshtein Distance, Spell Checker, Hamming Distance. Orlando: Alpha Press; 2009.

[Rish2001] Rish, I. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Seattle, WA, August 4, 2001; 41–46.

[Ramchoun2001] Ramchoun, H., Idrissi, M. A. J., Ghanou, Y. & Ettaouil, M. Multilayer Perceptron: Architecture Optimization and Training. *IJIMAI*, **4**, 26–30 (2016).

[Wittwehr2017] Wittwehr, C. *et al.* How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology. *Toxicol. Sci.* **155**, 326–336 (2017).

[Guan2018] Guan, D., Fan, K., Spence, I. & Matthews, S. Combining Machine Learning Models of *In Vitro* and *In Vivo* Bioassays Improves Rat Carcinogenicity Prediction. *Regul. Toxicol. Pharmacol.* **94**, 8–15 (2018).

[Berthold2008] Berthold, M. R. *et al.* In *KNIME: The Konstanz Information Miner*, Berlin, Heidelberg, Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; 319–326.

[O’Boyle2011] O'Boyle, N. M. *et al.* Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **3**, No. 33 (2011)

[Stewart2016] Stewart, J. J. P. *MOPAC2016*, Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016.

[Sure2013] Sure, R. & Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **34**, 1672–1685 (2013).

[Neese2012] Neese, F. The ORCA Program System. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 73–78 (2012).

[Neese2018] Neese, F. Software Update: The ORCA Program System, Version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1327 (2018).

[Barone1998] Barone, V. & Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **102**, 1995–2001 (1998).

[Guan2020] Guan, D., Lui, R. & Matthews, S. LogP Prediction Performance with the SMD Solvation Model and the M06 Density Functional Family for SAMPL6 Blind Prediction Challenge Molecules. *J. Comput. Aided Mol. Des.* **34**, 511–522(2020).

[Olson2016] Olson, R. S. *et al.* In *Automating Biomedical Data Science Through Tree-Based Pipeline Optimization*, Applications of Evolutionary Computation, Cham, 2016; Squillero, G.; Burelli, P., Eds. Springer International Publishing: Cham, 2016; 123–137.

[Moriwaki2018] Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **10**, No. 4 (2018).

[Yoshikawa2018] Yoshikawa, N. *et al.* Population-Based De Novo Molecule Generation, Using Grammatical Evolution. *Chem. Lett.* **47**, 1431–1434 (2018).

[Whitehead2019] Whitehead, T. M., Irwin, B. W. J., Hunt, P., Segall, M. D. & Conduit, G. J. Imputation of Assay Bioactivity Data Using Deep Learning. *J. Chem. Inf. Model*. **59**, 1197–1204 (2019).

[Hunt2020] ­­Hunt, P. A. *et al.* Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. *J. Chem. Inf. Model.* Just Accepted (2020).

[Jiménez-Díaz2014] Jiménez-Díaz, M. B. *et al.* (+)-SJ733, a Clinical Candidate for Malaria that acts through ATP4 to Induce Rapid Host-Mediated Clearance of *Plasmodium*. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E5455–E5462 (2014).

[Natekin2013] Natekin, A. & Knoll, A. Gradient Boosting Machines, A Tutorial. *Front. Neurorobot.* **7**, No. 21 (2013).

[Koes2013] Koes, D. R., Baumgartner M. P. & Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).

[Quiroga2016] Quiroga, R. & Villarreal, M. A. Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS ONE* **11**, e0155183 (2016).

[Krieger2015] Krieger, E. & Vriend, G. New Ways to Boost Molecular Dynamics Simulations. *J. Comput. Chem.* **36**, 996-1007 (2015).

[Krieger2004] Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A. & Vriend, G. Making Optimal use of Empirical Energy Functions: Force-Field Parameterization in Crystal Space. *Proteins* **57**, 678–683 (2004).

[Li2012] Li, H., Leung, K.-S. & Wong, M.-H. idock: A multithreaded virtual screening tool for flexible ligand docking. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, CA, May 9-12, 2012; 77–84.

[Trott2010] Trott, O. & Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **31**, 455-461 (2010).

[Li2015] Li, H., Leung, K.-S., Wong, M.-H. & Ballester, P. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *J. Mol. Inf.* **34**, 115–126 (2015).

[Chen2016] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, August 13-17, 2016; 785–794.

[Rogers2010] Rogers, D & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **50**, 742–754 (2010).

[Bellman1961] Bellman, R. E. *Adaptive Control Processes: A Guided Tour* (Princeton Legacy Library, 1961).