Structure-based Predictions for Molecular Initiating Events

Andrew John Wedlake

Department of Chemistry

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare College April 2019

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit of 60 000 words.

Andrew John Wedlake

April 2019

Acknowledgements

I would like to thank my supervisor, Professor Jonathan Goodman, for the opportunity to work on this project in his group. His help, advice, optimism, and guidance throughout this PhD has been invaluable. I am grateful to Dr Paul Russell and Dr Steve Gutsell for their assistance, many useful discussions and insightful feedback regarding all my work.

Many thanks to Dr Tim Allen for his day-to-day help, input and discussion of all things MIE. I'd like to thank all other members of the Goodman and Paterson groups, past and present, for creating a great work atmosphere and for patiently listening to my toxicology talks.

I am grateful to Unilever for funding this project, and I would like to thank all the members of SEAC. Particularly, Maria Folia for constructing the random forest models for the data sets, and Elin and Predrag for helping with the many issues of certain software.

I am especially grateful to Jonathan, Paul and Steve for their support and understanding during my concussion-induced medical intermission and subsequent prolonged recovery. I'd also like to thank the University Counselling service for their help during that tough time.

I would like to thank Clare College for their support for the last seven and a half years, in particular the porters, librarians and catering staff. Special mentions must go to the senior tutor, Dr Jacqueline Tasioulas, for her support, and to Bonna Sloan for her cooking, baking and chats throughout these many years at Clare.

Finally, I would like to thank my friends and family for their enthusiasm and support throughout.

Abstract

Toxicity testing of chemicals is currently undergoing its largest ever paradigm shift, moving towards faster, cheaper and more human-relevant methods which focus on mechanistic understanding. An AOP provides a framework for organising biological knowledge and data. The gateway to an AOP is the MIE, and chemistry is key to predicting which chemicals can undergo a MIE. *In silico* predictions of MIEs are a vital tool in a modern, mechanism-focused approach to risk assessment of chemicals.

In this project, new structural alert-based models for receptor binding MIEs have been constructed that create accurate, transparent and interpretable predictions. The alerts have been constructed with an automated workflow that uses Bayesian statistics to iteratively select substructures associated with activity. The models were constructed from balanced data sets taken from human *in vitro* assays in the ChEMBL and ToxCast databases. The new models significantly improve on previous models, with performance metrics comparable to random forest models. Methods for further improving structural alert models are presented, including a method for generalising aromatic atoms in structural alerts to reduce the number of alerts in a model, and construction of a consensus model combining structural alerts with a random forest model. Structural alert models have been constructed for a wide range of biological targets of toxicological interest and the variation in performance across all targets has been explained by considering the proportion of activity cliffs in data sets.

Having significantly improved structural alert models in terms of performance, new methods for assessing confidence in both active and inactive predictions have been developed. These involve considering similarity to relevant chemicals in the training set. The measure of confidence in active predictions allows for applicability of predictions to be evaluated, whilst the measure of confidence in inactive predictions is vital in risk assessment of chemicals.

Moving beyond structural alerts, attempts to describe chemicals in terms of the key interactions made with the biological target have been made. This a step towards describing how the receptor binding MIEs work and then using this knowledge to make better activity predictions. The generalised aromatic structural alerts have been used to predict key receptor binding interactions, which are consistent with interactions derived from crystal structures. Using structural alerts to group chemicals, pharmacophore models have been developed, allowing for activity predictions in terms of general features in three-dimensional space instead of the specific combination of atoms and bonds described by structural alerts.

Contents

Nomenclature

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
5HTR2A	Serotonin 2a receptor
5HTR3A	Serotonin 3a receptor
5-HTT	Serotonin transporter
ACC	Accuracy
AChE	Acetylcholinesterase
ADME	Absorption, distribution, metabolism and excretion
ADRA2A	Alpha-2a adrenergic receptor
ADRB1	Beta-1 adrenergic receptor
AE	Adverse effect
AOP	Adverse outcome pathway
ATP	Adenosine triphosphate
CASP1	Caspase 1
DC	Dendritic cell
DNA	Deoxyribonucleic acid
DRD2	Dopamine D2 receptor
EC$_{50}$	Half maximal effective concentration
ECFP	Extended-connectivity fingerprints
EPA	Environmental Protection Agency
FN	False negative
FP	False positive
GPCR	G protein-coupled receptors
HBA	Hydrogen bond acceptor
HBD	Hydrogen bond donor
hERG	Human ether-a-go-go-related gene
IC$_{50}$	Half maximal inhibitory concentration
ICH	International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use
IUPAC	International Union of Pure and Applied Chemistry
K$_{bapp}$	Apparent binding constant
K$_d$	Receptor/ligand dissociation constant

KE	Key event
KER	Key event relationship
K_i	Receptor/ligand binding affinity
K_{inact}	Enzyme inactivation constant
KNIME	Konstanz Information Miner
LCK	Tyrosine-protein kinase LCK
MACCS	Molecular access system
MAPK1	Mitogen-Activated Protein Kinase 1
MCC	Matthews Correlation Coefficient
MIE	Molecular initiating event
MODI	Modelability index
MoFa	Molecular Fragment miner
MoSS	Molecular Substructure miner
MTP	mitochondrial permeability transition
NAPQI	N-acetyl-p-benzoquinone imine
NECA	5'-N-Ethylcarboxamidoadenosine
NET	Norepinephrine transporter
NPV	Negative predictive value
OECD	Organisation for Economic Co-operation and Development
PBPK	Physiologically Based Pharmacokinetic
PPARγ	Peroxisome proliferator-activated receptor gamma
PPV	Positive predictive value
qAOP	Quantitative adverse outcome pathway
QSAR	Quantitative structure-activity relationship
REACH	Registration, Evaluation, Authorization and Restriction of Chemical
ROS	Reactive oxygen species
SAR	Structure-activity relationship
SE	Sensitivity
SMARTS	Simplified molecular-input line-entry system arbitrary target specification
SMILES	Simplified molecular-input line-entry system
SP	Specificity
SR	Stepwise regression
TN	True negative
ToxCast	The Toxicity Forecaster
TP	True positive
V1AR	Vasopressin V1a receptor

1. Introduction

1.1. Toxicity Testing

Risk assessment of chemicals is essential in ensuring the safety of every person in daily life. Whether it be medicines, food and drink additives, shampoos, skin creams or any other consumable, consumers expect there to be no toxic effects, or at most limited toxic effects. Toxicity testing is currently undergoing its largest ever paradigm shift and, at its core, chemistry has a vital role.[1]

Historically, toxicity testing has been centred on animal testing. These methods had limited mechanistic understanding, focusing instead on apical endpoints, often elicited by exposing organisms to concentrations of chemicals much larger than commercial concentrations for humans. In terms of processes such as metabolic pathways and rates, protein binding, body temperature, amongst many other things, the biology of animals does not match the biology of humans. Hence, results of toxicity tests in animals often are not matched by results in humans.[2] For example, penicillin is toxic to guinea pigs, and coffee, chocolate and avocados are toxic to dogs. Previous studies have found differing but low values for the overall concordance of human and animal hepatotoxicity (40%[3], 55%[4] and 77%[5]). In addition to the unreliability of animal testing, the methods are time consuming and expensive. Consequently, there has been a strong impetus to move towards faster, cheaper and more human-relevant methods of risk assessment which focus on mechanistic understanding.

Problems with historic risk assessment methods have been compounded by recent regulation changes. The European Union's Registration, Evaluation, Authorization and Restriction of Chemical substances (REACH) program[6] came into full force in 2018, requiring companies to register all chemical substances produced or imported in the EU in quantities of greater than one tonne per annum, and identify any toxicological concerns of the chemicals. Combined with the lack of characterisation of the toxicity profile of many chemicals already in the environment, these changes have led to a large number of chemicals needing to be tested. Whilst the primary drive for change in toxicity testing has been due to a desire for better scientific methods, regulation changes have increased pressure on risk assessors.

1.2. Adverse Outcome Pathways and Molecular Initiating Events

In 2007, the American National Research Council published a vision of a new toxicity testing strategy that relies on understanding toxicity pathways – sequences of biological responses that can lead to adverse health effects.[7] Advances in systems biology[8] and in -omics technologies[9,10,11] have allowed for greater understanding of biological processes, paving the way for a more mechanistic approach to toxicity testing.

In 2010, Ankley *et al* proposed the concept of the adverse outcome pathway (AOP).[12] An AOP is a sequence of events from the exposure of an individual to a chemical through to an understanding of the adverse effect (AE) at the individual or population level.[13] It is a flexible framework made of two types of components:

- Key events (KEs) – measurable changes in biological systems.
- Key event relationships (KERs) – the links between key events. There is much flexibility for what can be considered a KER. They can be causal, mechanistic, inferential or correlation based, and can be based on *in vitro*, *in vivo*, or *in silico* data.[12]

AOPs provide a logical framework for organising data in a mechanistic way, with data ranging in biological scale from molecular interactions up to population responses. It can be used to make links between key biological events, and to identify gaps in data.

The AOP framework is outlined in Figure 1.1. Despite often being presented as a linear, unidirectional sequence of steps, AOPs may be more complex, non-linear and branched. For example, there may be positive or negative feedback loops, counter-regulations or modulatory events.[14]

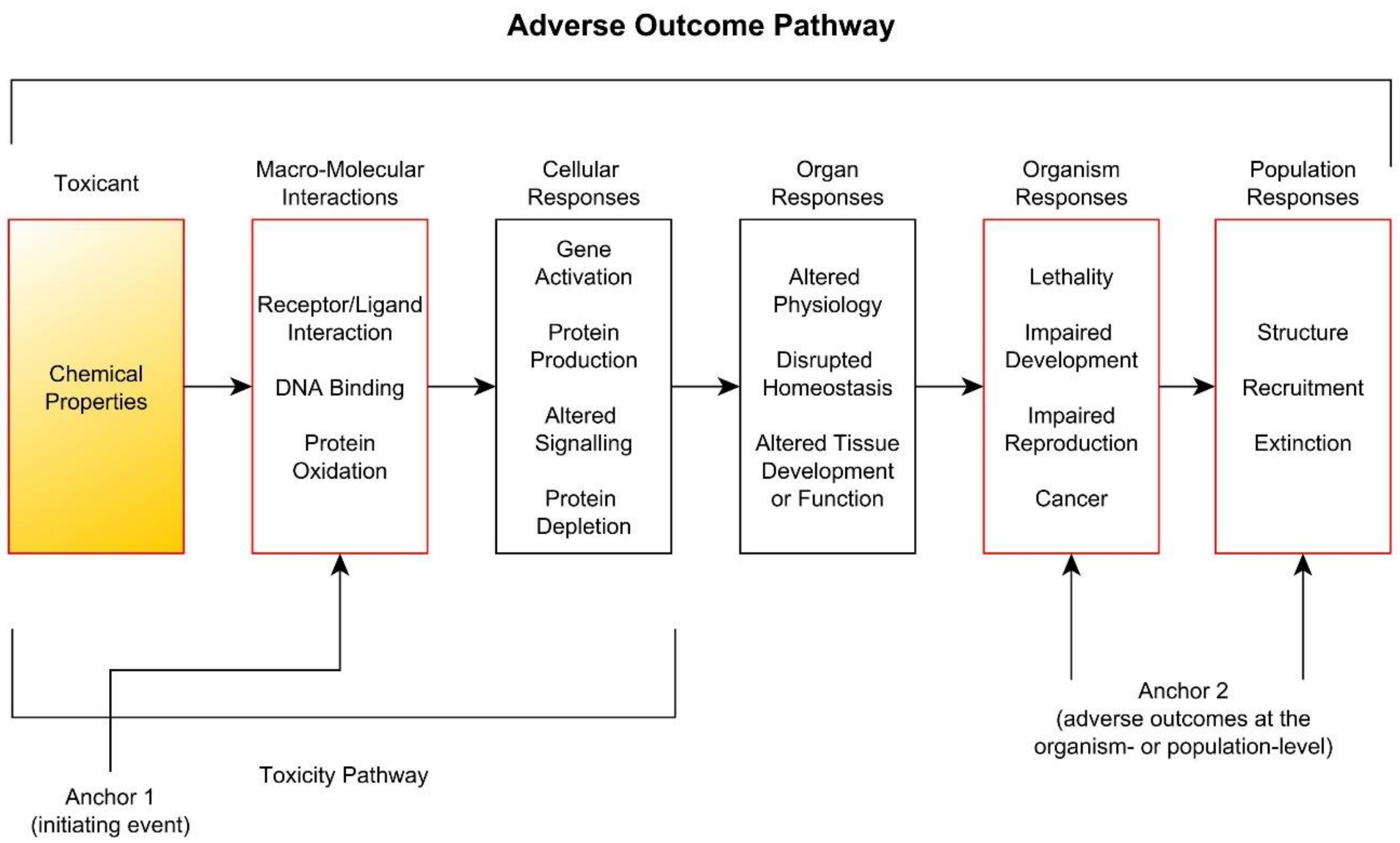


*Figure 1.1***:** *A graphical overview of the Adverse Outcome Pathway framework. A toxicant with certain chemical properties undergoes a molecular initiating event (MIE). The results in a series of key events escalating from cellular responses up to organ-level and organism-level responses, leading to an adverse effect. Understanding which chemical properties are required to elicit a particular MIE requires an understanding of the chemistry of the toxicants. Figure from Allen (2016),*[16] *adapted from Ankley (2010).*[12]

The first key event in an AOP is the molecular initiating event (MIE), defined as the initial interaction between a molecule and a biomolecule or biosystem that can be causally linked to an outcome via a pathway.[15] In theory, knowing the MIEs of a single chemical tells us all AEs of the chemical (if all AOPs for the MIEs are known). In this way, understanding MIEs could be key to understanding the potential toxic effects of a chemical.

The MIE can be seen as the gateway to the AOP. In a large enough dose, chemicals that undergo the MIE cause an AE through an AOP. The AOP is chemically agnostic. That is, the steps from MIE to AE should not be specific to any one chemical. Hence, predicting which chemicals undergo a MIE is key to predicting which chemicals can cause toxicity, and chemistry is key to predicting which chemicals will undergo a MIE.[16] Understanding what structures and chemical features are required to undergo a MIE allows for informed and accurate predictions to be made regarding the MIE.

There are other considerations in addition to the AOP that are required to fill in the full picture starting at chemical exposure and ending in toxicity. The series of steps can be broken down into the steps:

1. Chemical exposure – How much of the chemical is a person exposed to and what part of the body is exposed?
2. Absorption, distribution, metabolism and excretion (ADME) – After exposure to a chemical, what free concentration of chemical reaches the target site?
3. MIE – which chemicals undergo the MIE?
4. AOP(s)
5. Adverse effect

Chemistry has an important role in developing understanding at every step[1], as do *in silico* tools.[17]

1.2.1. Absorption, distribution, metabolism and excretion

Prior to the MIE, an understanding is required of how a chemical can reach the target site, and in what concentration. Physiologically Based Pharmacokinetic (PBPK) models are *in silico* tools which attempt to simplify the many complex processes in the ADME of a chemical.[18] They consist of different compartments, akin to body tissues, connected by a circulating blood flow. By solving differential equations between compartments, a sophisticated estimate of free concentration of a chemical at a target can be calculated.

For some toxic drugs, toxicity is not caused by the chemical itself but by a reactive metabolite. This is particularly common in drugs causing hepatotoxicity as the liver is the main site of metabolism in the body.[19] Paracetamol is the most well-known example of such a drug. A small proportion of paracetamol is metabolised in the liver to N-acetyl-p-benzoquinone imine (NAPQI), which can cause liver injury via a number of MIEs, as shown in Figure 1.2.

Reliably predicting the toxicity of reactive metabolites of a chemical is difficult. First, one must predict the likely sites of metabolism in the initial chemical and which metabolites are likely to form, for which numerous *in silico* tools are available, such as Metaprint2D[20] and Lhasa's Meteor [21]. Kirchmair *et al* have extensively reviewed *in silico* tools for identifying sites of metabolism and likely metabolite products.[22,] Having predicted which metabolites may form, one must then predict if any are likely to undergo any MIEs.

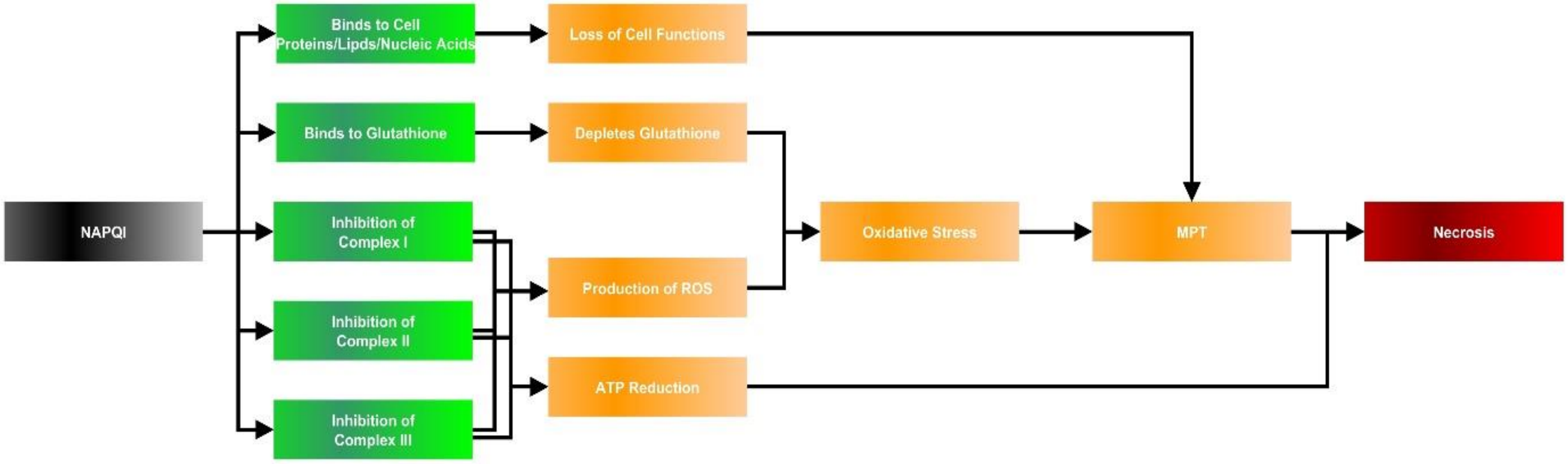


Figure 1.2: A network of adverse outcome pathways for N-acetyl-p-benzoquinone imine (NAPQI) toxicity. NAPQI acts through numerous molecular initiating events (green boxes), causing subsequent key events (orange boxes) along different adverse outcome pathways which all lead to the adverse effect of liver cell necrosis (red box). ROS = reactive oxygen species; ATP = adenosine triphosphate; MTP = mitochondrial permeability transition.

1.2.2. AOP networks and quantitative AOPs

An AOP was originally defined by the Organisation for Economic Co-operation and Development (OECD) to link a single MIE to a single AE. In a reality, biology is more complex. A single MIE may lead to more than one AE through multiple AOPs. A single chemical may undergo more than one MIE which, via different AOPs, lead to the same AE (as seen for NAPQI in Figure 1.2). To convey the true complexity of living systems and *in vivo* models, a network of AOPs is required.[23,24]

Having identified a chemical which undergoes a MIE, an AOP provides a qualitative, chemically-agnostic link to an AE. However, activity at the MIE may not necessarily lead to the AE. For example, there may be compensatory or adaptive mechanisms designed to stop KEs in the AOP from happening. This results in thresholds that need to be overcome for the AE to be observed. The dose of the chemical must be large enough to result in the thresholds being overcome. Accounting for these phenomena requires a quantitative understanding of KERs, although the quantitative understanding does not necessarily have to be between adjacent KEs.

Early attempts to build quantitative AOPs (qAOPs) have been made for skin sensitisation[25], the AOP for which is shown in Figure 1.3. A qAOP has been built linking the MIE of aromatase inhibition to declining population trajectory in fathead minnows.[26] Even for relatively simple AOPs which, in this context, are not complicated by AOP networks, development of qAOPs currently requires significant investment of resources. In the future, as the AOP concept becomes more established and more mechanistic data and models at KE levels become available, qAOPs will become increasingly feasible and important.

Even without quantitative relationships, the AOP framework provides a robust way to organise mechanistic data, which can be used to guide risk assessment in a scientifically-sound, pragmatic way. *In silico* models can cheaply and quickly predict the activity of chemicals at a MIE linked to AEs through AOPs. These predictions can guide risk assessors in picking a select group of chemicals for further testing, reducing the time and resources required.

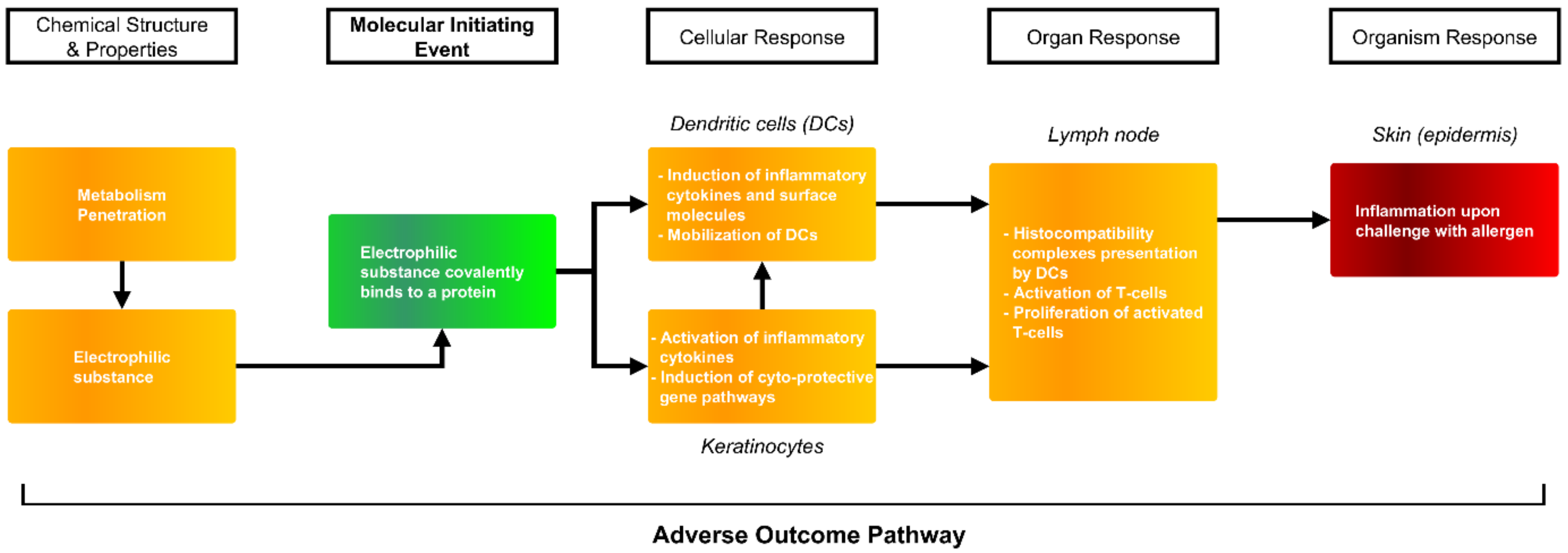

Chemical Structure & Properties
Molecular Initiating Event
Cellular Response
Organ Response
Organism Response
Metabolism Penetration
Electrophilic substance
Electrophilic substance covalently binds to a protein
Dendritic cells (DCs)
- Induction of inflammatory cytokines and surface molecules
- Mobilization of DCs
- Activation of inflammatory cytokines
- Induction of cyto-protective gene pathways
Keratinocytes
Lymph node
- Histocompatibility complexes presentation by DCs
- Activation of T-cells
- Proliferation of activated T-cells
Skin (epidermis)
Inflammation upon challenge with allergen
Adverse Outcome Pathway

*Figure 1.3***:** *The AOP for skin sensitization. Green boxes represent molecular initiating events, orange boxes represent key events, and red boxes represent adverse effects. Figure adapted from the OECD's Adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins (2012).*[107]

1.3. (Quantitative) Structural Activity Relationships

Structure-activity relationship (SAR) models attempt to create an association between some form of biological activity of chemicals and some chemical descriptors. The purpose of SARs is to qualitatively predict chemicals which will be biological active, whilst quantitative structure-activity relationship (QSAR) models make quantitative predictions of biological activity.

The chemical descriptors in the (Q)SAR could be physicochemical properties (such as molecular weight or lipophilicity), common two-dimensional chemical structures (structural alerts) or three-dimensional structures. Generally, SARs try to identify structures which cause biological activity. QSARs try to identify a relationship between biological activity and a quantity associated with the molecule – this could be physicochemical properties or a numerical representation of the structure of the molecule.

OECD's guidelines for (Q)SAR construction[27] suggest five key principles which should be followed to help (Q)SARs gain acceptance in regulatory use:

1. A defined endpoint
2. An unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of-fit, robustness and predictivity
5. A mechanistic interpretation, if possible

Historic applications of (Q)SAR models have focused on predicting apical organism-level endpoints. Biological activity at such endpoints is often due to multiple mechanisms and MIEs, complicating model building. An exception to this is skin sensitisation[28] – an apical endpoint but with a well-defined AOP. Outliers in QSAR models are often chemicals which act through a different mechanism to other chemicals.[29] By constructing (Q)SARs for MIEs, the models are being built for a single mechanism or chemical interaction. Such (Q)SARs are founded in mechanistic understanding without the need to extrapolating over many biological steps to predict an apical endpoint. Thus, MIE-based (Q)SARs are easier to interpret mechanistically than (Q)SARs built for apical end points.

1.3.1. Constructing (Q)SARs

To construct (Q)SARs, a training set of compounds and their biological activity data are collected. Available MIE data includes the half maximal effective concentration (EC_{50}), the half maximal inhibitory concentration (IC_{50}), the receptor/ligand dissociation constant (K_d), and the receptor/ligand binding affinity (K_i). A training set is used to create rules for predicting toxicity. A test set of compounds with known biological activity is then used to assess and justify the rules. Results from the test sets could be used to design updates and changes to the (Q)SARs, leading to a cycle of model improvement, shown in Figure 1.4. The (Q)SARs can be used to make predictions for compounds outside of the training sets for which there is no biological activity data.

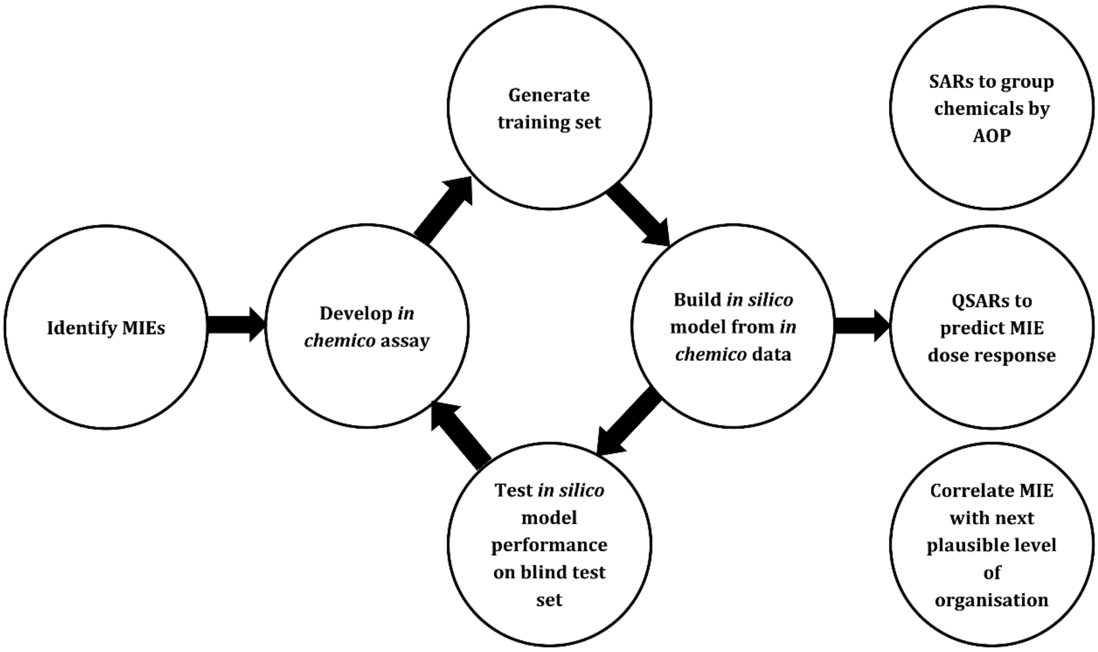

Generate training set
SARs to group chemicals by AOP
Identify MIEs
Develop *in chemico* assay
Build *in silico* model from *in chemico* data
QSARs to predict MIE dose response
Test *in silico* model performance on blind test set
Correlate MIE with next plausible level of organisation

Figure 1.4: A wet/dry cycle for the generation and development of in silico *models. Figure adapted from Gutsell and Russell (2013).*[1]

1.3.2. Machine learning in SARs

Machine learning algorithms have been applied to create SAR models, which are generally high-performing. A wealth of physiochemical properties (such as molecular weight, logP, number of rings, number of hydrogen bond donors, etc.) can be generated *in silico* and complex statistical algorithms can identify the importance of the descriptors or combinations of descriptors. From this, predictions of biological activity are made.

Algorithms that have been applied to (Q)SAR modelling include Random Forest,[30,31] Support Vector Machines,[32,33] K Nearest Neighbors,[34,35] and Neural Networks.[36,37] Neural networks and, more specifically, deep learning have had success in SAR modelling. In the recent Tox21 Data Challenge, the top performing and prizewinning model used a deep learning algorithm with physicochemical properties and fingerprint bit strings as chemicals descriptors.[37]

While the models often perform very well, they lack mechanistic transparency. They are often viewed as "black box" processes, producing accurate predictions, but in a way that is not easy to interpret. Being able to interpret to predictions is particularly important in toxicity testing. As well as knowing that a chemical has been predicted to be biologically active, risk assessors will want to know why it has been predicted to be active.

Random Forest models are often viewed as being the most interpretable of these machine learning methods. They allow identification of descriptors that were most important in predicting classifications can be identified,[38] giving some indication of how predictions have been made.

Random Forest models

Random Forest models consist of many individual decision trees. Each individual tree predicts the activity of a chemical and the most popular prediction becomes the model's overall prediction. The premise is that prediction by committee should be more accurate than the prediction by any individual tree.

A key requirement of this approach is that the individual decision trees are uncorrelated. This should ensure that the overall committee decision is protected from the error of an individual tree. In a Random Forest model, each decision tree is constructed from a random subset of features and by taking a random sample of chemicals from the training set (with replacement). At each branch in the decision tree, a rule regarding one feature is chosen that splits the most active and inactive chemicals within the sample into two different branches, as determined by information gain defined by Shannon entropy.[39] The branches in the decision tree grow until no further information can be gained or until a maximum tree depth is reached. This process is then repeated many times, with a new random sample of chemicals and random subset of features each time.

The interpretability of Random Forest models is a key advantage compared to other classical machine learning approaches. The relative importance of features can be identified by computing how much each feature contributes to information gain at each branch in a decision tree and averaging across all trees in the Random Forest.

Random Forest models have previously been applied to (Q)SAR modelling and biological activity predictions. Of particular relevance to this work, in 2018 Mervin *et al.* used Random Forest algorithms to construct SAR models for 332 protein targets including G protein-coupled receptors (GPCRs), ion channels, enzymes, transporters, and nuclear receptors, taking data from the Chemistry Connect repository.[40,41] Morgan fingerprint were used as inputs for the Random Forest models.

1.3.3. Structural alerts in SARs

Structural alerts are 2D chemical structures used as chemical descriptors in SAR models. Presence of a structural alert within a chemical generally leads to an active prediction.

Structural alerts have been widely used in SARs predicting reactivity-driven MIEs such as skin sensitisation and DNA mutagenicity. In these MIEs, electrophilic compounds covalently bond to skin cell proteins (skin sensitisation) or to DNA (DNA mutagenicity). Structural alerts have been constructed to identify chemical structures which are electrophilic, or which are commonly metabolised to form electrophilic groups. SARs exist for both these MIEs within Lhasa's Derek software, a knowledge-based expert system.[42] Structural alerts within such systems are written by humans using knowledge of chemical mechanisms. Alerts have clear explanations, mechanistic reasoning, and literature references where possible.

Structural alerts have also been used in SARs for receptor binding MIEs.[43,44] In such MIEs, chemicals bind to specific enzyme active sites or receptor binding sites. Interactions are through hydrogen bond formation, hydrophobic interactions, and electrostatic interactions in specific 3D geometries. There are specific steric requirements to fit in the active sites or binding sites. The structural alerts typically define the chemical cores of the binding chemicals, defining a specific scaffold which meets the specific steric requirements. However, there are some limitations to using structural alerts for receptor binding MIEs. Structural alerts do not define requirements outside of the chemical core, meaning they may miss some demands of the chemical required for activity. They are 2D requirements of 3D chemicals, so are simplifications that may miss details that only become apparent when considering different 3D conformations. The applicability of structural alerts for receptor binding is often limited, only applying to chemicals with a specific arrangement of atoms in the chemical core.

The main advantage of structural alerts, both for reactivity-driven and receptor binding MIEs, is that they are mechanistically transparent. Predictions made by structural alerts are easy to interpret. They are computationally simple, particularly as computationally expensive 3D conformation generation is not required for chemicals being tested, allowing for quick processing of many chemicals.

Structural alerts have been combined with MIEs and AOPs to make mechanistically sound predictions for other key toxicological end points. Nelms *et al* have constructed structural alerts for mitochondrial toxicity, finding substructures common to toxicants and assigning MIEs to the alert.[45] Mellor *et al* have created structural alerts for nuclear receptors,[46] and have developed AOPs with binding at the receptors being MIEs leading to the AE of liver steatosis.[47]

The OECD QSAR toolbox[48] provides numerous *in-silico* predictive tools which link chemical characteristics to a potential toxic mechanism or MIE, such as mutagenicity. The toolbox includes structural alerts, for some of which an “alert performance” functionality has recently been added.[49] This provides information from other chemicals which act through the same MIE by the same chemical mechanism to allow for evaluation of the alert-based predictions.

1.3.4. Pharmacophore models

Pharmacophores are a form of (Q)SAR modelling for predicting receptor binding MIEs. They are defined by IUPAC[50] as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response". Rather than identification of specific substructures like in structural alerts, pharmacophores identify combinations of general features in 3D space. These features include hydrogen bond donors or acceptors, ionisable or charged groups, hydrophobic or hydrophilic groups, and aromatic rings. There are two types of pharmacophore model generation: structure-based and ligand-based.

Structure-based pharmacophore models are generated by looking at the receptor binding site and identifying key interactions. This is usually done by observing a crystal structure with a ligand bound to the receptor. The model generated may be specific to the ligand used to generate the crystal structure and the X-ray structure may represent a "tensed" conformation of the ligand due to the crystal packing forces.[51] These factors may lead to structure-based pharmacophores not accurately predict activity of other ligands.

Tskovska *et al* have using structure-based methods to construct a pharmacophore model of PPARγ,[51] shown in Figure 1.5. This is one of the receptors for which Mellor *et al* have developed AOPs leading to liver steatosis.[47]

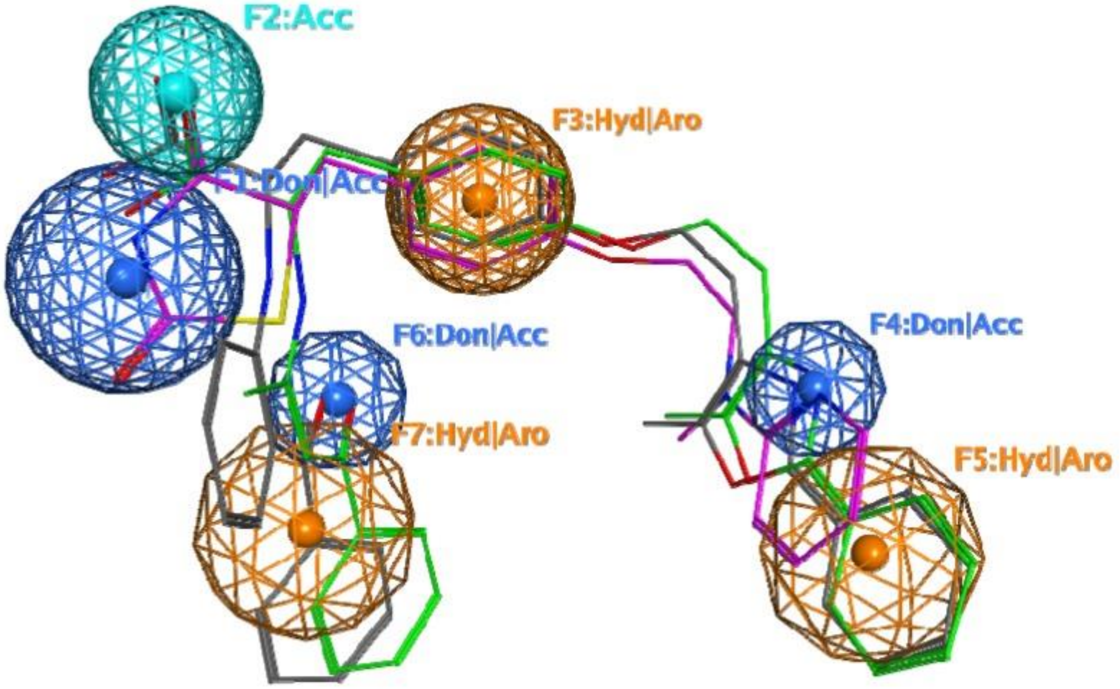


Figure 1.5: Tskovska's pharmacophore model of PPARγ. The pharmacophore model identifies the 3D conformation of required features for binding, in terms of hydrogen bond acceptor (light blue), hydrogen donor and acceptor (dark blue), and hydrophobic and aromatic features (orange). Three full agonists are shown: rosiglitazone with carbon atoms in magenta, another compound with carbon atoms in green, and a third compound with carbon atoms in grey. Image from Tsakovska (2014).[51]

Ligand-based pharmacophore models are generated by superimposing 3D conformations of active chemicals and identifying regions where common features overlap. This approach is often computationally expensive and complicated, requiring extensive conformation generation, overlaying of the possible conformations, and determination of which overlapping features are best for the model. Many *in silico* tools for automating the ligand-based pharmacophore generation process have been designed and are commercially available.[52] Even with the availability of automated tools, the user must first carefully select ligands which elicit activity through the same binding mode. There is however no guarantee that the overlapping conformations of ligands will match the required conformation of the ligands when binding to the receptor in biological systems.

1.3.5. Complementary Models in SARs

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) is a leading international group in establishing guidelines for toxicity testing. In 2014, the ICH published the M7 guideline for assessment of potentially mutagenetic impurities.[53] It is significant as it recognises the potential to accept *in silico* predictions in a field of toxicity instead of *in vitro* studies. The guidelines states that two different (Q)SARs methods that complement each other should be used. One method should be expert rule-based, such as knowledge-derived structural alerts, and the other should be statistical-based, such as machine learning approaches. Inactive predictions from both methods is sufficient evidence to accept an impurity as being of no mutagenic concern without further testing. Barber *et al* have further discussed how complimentary *in silico* predictions can be used together, beyond agreeing on negative predictions.[54]

This landmark guideline establishes the potential of *in silico* predictions to replace *in vitro* studies in mutagenicity. As (Q)SAR models for other MIEs are developed and as confidence increases in their predictions, other fields of toxicity will accept similar use of *in silico* predictions. *In silico* predictions also have applications in toxicity screening, chemical prioritisation, and early stage drug studies.

The ICH M7 guideline also highlights the potential of, and need for, complementary models. For example, structural alert models could be used in combination with machine learning models. Development and improvement of high performing structural alert models for MIEs (beyond mutagenicity) is therefore vital. The impact of these models can be greatly increased by applying them with complementary approaches. A combination approach will increase confidence of *in silico* predictions, which at the very least can be used to guide further *in vitro* tests, and at most can replace *in vitro* and *in vivo* studies.

1.4. Biological Data

1.4.1. Bowes Targets

Bowes *et al* identified 44 biological targets in 2012.[55] Screening a chemical for activity at these targets provides a minimum standard for a broad early assessment of potential toxicity. They were identified by four major pharmaceutical companies (AstraZeneca, GlaxoSmithKline, Novartis and Pfizer) as targets that are tested in at least three of the four companies. The targets cover a wide range of biological functionality, comprising of G protein-coupled receptors (GPCRs), ion channels, enzymes, transporters, and nuclear receptors. The targets are shown in Table 1.1.

The Bowes targets have been identified as being pharmacologically important, often leading to the failure of new drug candidates in clinical trials. They are also important in assessing the safety of consumer goods as activity may lead to systemic toxicity.

The Bowes targets are key MIEs in risk assessment. Constructing SARs for these MIEs is vital in assessing the potential toxicity of chemicals. As important targets, a large amount of data is available for most of the Bowes targets, from which good SARs can be constructed.

However, the Bowes targets represent only a minimum for assessing toxicity of chemicals. Whilst focusing on the Bowes targets is a pragmatic approach for early risk assessment, other biological targets are important. Constructing SARs for other biological targets will provide a more complete assessment of potential toxicity.

Bowes Target	Gene Symbol	Protein type
Acetylcholine receptor subunit α1 or α4	CHRNA4	Ion channel
Acetylcholinesterase	ACHE	Enzyme
Adenosine A2a receptor	ADORA2A	GPCR
Alpha-1a adrenergic receptor	ADRA1A	GPCR
Alpha-2a adrenergic receptor	ADRA2A	GPCR
Androgen receptor	AR	Nuclear receptor
Beta-1 adrenergic receptor	ADRB1	GPCR
Beta-2 adrenergic receptor	ADRB2	GPCR
Cannabinoid CB1 receptor	CNR1	GPCR
Cannabinoid CB2 receptor	CNR2	GPCR
Cholecystokinin A receptor	CCKAR	GPCR
Cyclooxygenase-1	PTGS1	Enzyme
Cyclooxygenase-2	PTGS2	Enzyme
Delta opioid receptor	OPRD1	GPCR
Dopamine D1 receptor	DRD1	GPCR
Dopamine D2 receptor	DRD2	GPCR
Dopamine transporter	SLC6A3	Transporter
Endothelin receptor ET-A	EDNRA	GPCR
GABAA receptor α1 (rat cortex) BZD site	GABRA1	Ion channel
Glucocorticoid receptor	NR3C1	Nuclear receptor
Glutamate (NMDA) receptor subunit zeta 1	GRIN1	Ion channel
Histamine H1 receptor	HRH1	GPCR
Histamine H2 receptor	HRH2	GPCR
Kappa opioid receptor	OPRK1	GPCR
Monoamine oxidase A	MAOA	Enzyme
Mu opioid receptor	OPRM1	GPCR
Muscarinic acetylcholine receptor M1	CHRM1	GPCR
Muscarinic acetylcholine receptor M2	CHRM2	GPCR
Muscarinic acetylcholine receptor M3	CHRM3	GPCR
Norepinephrine transporter	SLC6A2	Transporter
Phosphodiesterase 3A	PDE3A	Ion channel
Phosphodiesterase 4D	PDE4D	Ion channel
Potassium voltage-gated channel subfamily H member 2 (HERG)	KCNH2	Ion channel
Serotonin 1a (5-HT1a) receptor	HTR1A	GPCR
Serotonin 1b (5-HT1b) receptor	HTR1B	GPCR
Serotonin 2a (5-HT2a) receptor	HTR2A	GPCR
Serotonin 2b (5-HT2b) receptor	HTR2B	GPCR
Serotonin 3a (5-HT3a) receptor	HTR3A	Ion channel
Serotonin transporter	SLC6A4	Transporter
Sodium channel protein type V alpha subunit	SCN5A	Ion Channel
Tyrosine-protein kinase LCK	LCK	Ion Channel
Vasopressin V1a receptor	AVPR1A	GPCR
Voltage-gated calcium channel subunit α Cav1.2	CACNA1C	Ion channel
Voltage-gated potassium channel subunit Kv7.1	KCNQ1	Ion channel

Table 1.1: A list of the biological targets identified by Bowes et al *(2012) as recommended targets to be screened for to provide an early assessment of the potential hazard of a chemical. (GPCR = G protein-coupled receptor).*

1.4.2. Biological Relevance

Predictions of the MIEs can be directly linked to AEs via AOPs. The AOP wiki,[56] part of the OECD's AOP Knowledge Bases, is the largest online library of AOPs, collected by crowd-sourcing knowledge. Users can submit AOPs at various levels of development. There are 39 AOPs either completed or under construction on AOP wiki with Bowes targets as MIEs. These are for:

- Serotonin transporter (2)
- Acetylcholinesterase (2)
- Androgen receptor (4)
- Beta-2 adrenergic receptor (1)
- Gamma-aminobutyric acid receptor (2)
- Glucocorticoid receptor (2)
- Glutamate receptor (1)
- Histamine H2 receptor (1)
- Mu-type opioid receptor (2)
- Ether-a-go-go-related gene potassium channel (1)
- Cyclooxygenase-1 (7)
- Sodium channel (5)
- Serotonin transporter (6)
- Voltage-dependent L-type calcium channel (3)

These AOPs provide direct biological relevance for SAR predictions made of the Bowes MIEs, helping to interpret why activity at these targets can lead to toxicity. The AOP for the glutamate receptor taken from AOPwiki is shown in Figure 1.6.

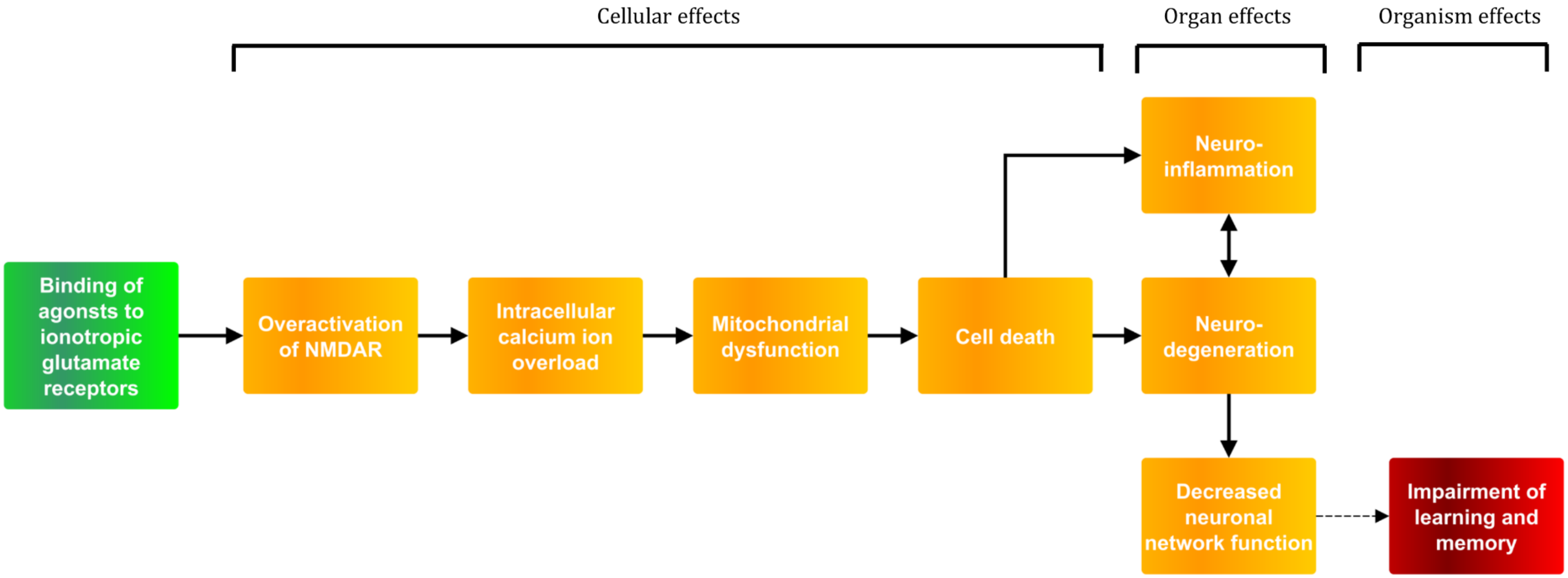

Cellular effects
Organ effects
Organism effects
Binding of agonsts to ionotropic glutamate receptors
Overactivation of NMDAR
Intracellular calcium ion overload
Mitochondrial dysfunction
Cell death
Neuro-inflammation
Neuro-degeneration
Decreased neuronal network function
Impairment of learning and memory

Figure 1.6: An AOP for the glutamate receptor. The green box is the MIE, yellow boxes are key events and the red box is an adverse effect. Full arrows represent key event relationships with good evidence, whilst the dashed arrow represents a key event relationship for which scientific understanding is not completely established. Image is adapted from AOPwiki.[56]

1.5. Computational Methods

1.5.1. Fingerprints and Similarity

Molecular similarity is a vital concept in chemical informatics. Johnson and Maggiora's similarity property principle states that similar compounds will have similar properties,[57] with biological activity being the most frequently studied property. What exactly is meant by similarity between compounds is inherently subjective and therefore difficult to define. Despite this, there have been many different attempts and methods to quantify the similarity between chemicals.[58]

Molecular fingerprints are representations of chemical structures which are often used in similarity searching. Morgan fingerprints were first developed in 1965 as a method for identifying cases of a isomorphisms between chemical structures – the same structure drawn or numbered in a different way.[59] Morgan fingerprints and the closely related extended-connectivity fingerprints (ECFPs)[60] have since been widely accepted and used in chemical informatics in substructure searching, clustering and similarity searching. The ECFP algorithm makes subtle changes to the Morgan algorithm that make it more computationally efficient, but otherwise the methods use the same ideas and steps. Both create bit strings where each bit represents presence of circular substructures. The length of the bit string and the maximum size of the circular substructures are defined by the user.

The steps involved in generating Morgan fingerprints or ECFPs are:

1. *Assignment of initial atom identifiers*. Various properties of the atom, including atom number and connectivity count, are hashed into a single integer atom identifier. This is added to an initial fingerprint set.

2. *Iterative creation of circular identifiers*. Along with the central atom's identifier, the neighbouring atoms' identifiers are hashed into a new identifier representing a circular substructure, which is added to the fingerprint set. This process is iteratively repeated, using the circular substructure's identifier and the neighbouring atoms' identifiers to create a new identifier for a larger circular substructure, which is added to the fingerprint set. The iterative process repeats until the circular substructures reach a maximum size as defined by the user. In Morgan fingerprints, the size of the circular substructure is defined by the radius of the circle, while in ECFP it is defined by the diameter.

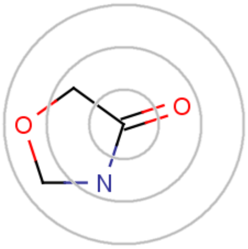


Figure 1.7: Circular substructures of increasing radius, all centred on the same atom. Image taken from ChemAxon's ECFP Documentation[61]*.*

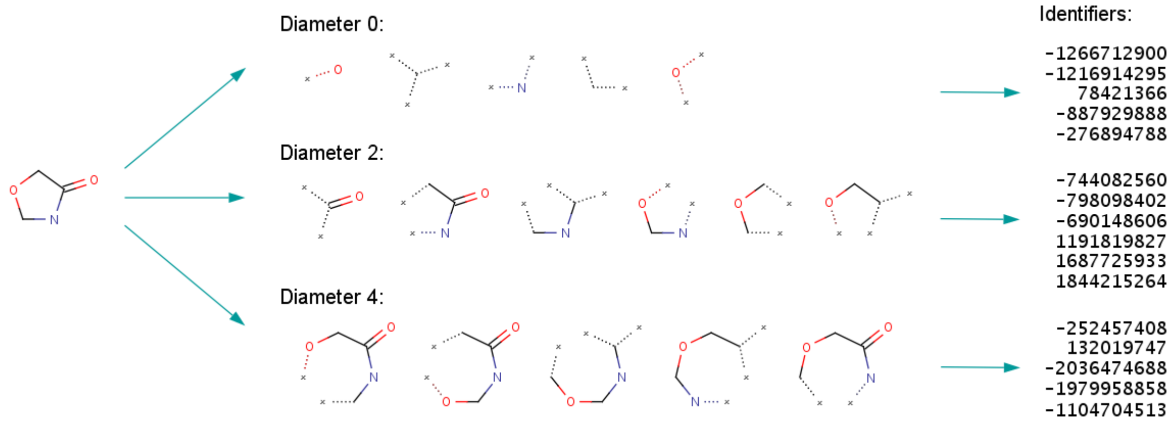


Figure 1.8: A diagrammatic overview of the first two steps of the Morgan and ECFP processes. On the left is the chemical structure. In the middle are the circular substructures of different sizes present in the chemical. Each circular substructure is hashed into an identifier. Note that negative identifiers are possible in the ECFP algorithm but not the Morgan algorithm. Image taken from ChemAxon's ECFP Documentation[61]*.*

3. *Removal of duplicate identifiers.* Cases where multiple identifiers in the fingerprint set represent the same features are found and all but one identifier are removed.

4. *Creation of fingerprint bit string.* The identifiers in the fingerprint set are hashed onto a bit string of length specified by the user. Any hash function that maps arrays of integers randomly and uniformly across the bit string may be used for this step.[60] Each bit therefore represents presence (or absence) of a circular feature. Bit collisions occur where more than one feature is hashed onto the same bit. To create fingerprints that can be compared to each other, the same bit string length and hash function must be used.

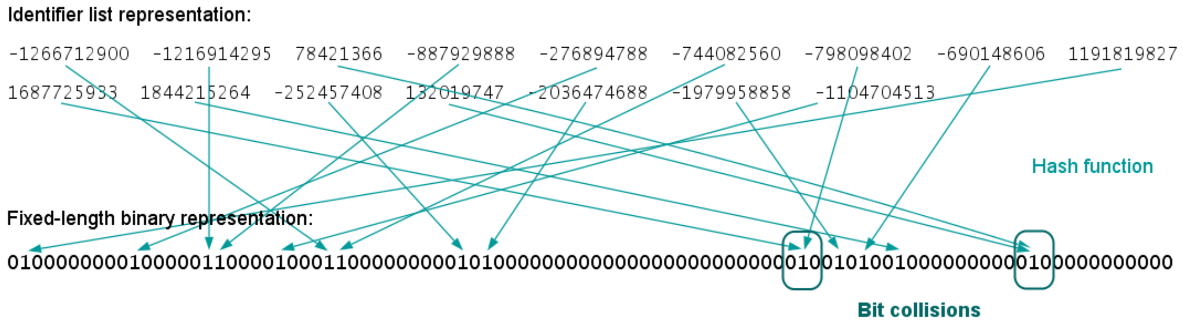


Figure 1.9: Having removed duplicated identifiers, each identifier in the fingerprint set is hashed onto a bit string of length specified by the user. A bit collision occurs where two different identifiers are hashed onto the same bit. Image taken from ChemAxon's ECFP Documentation[61].

The molecular access system (MACCS) fingerprint[62], also known as MACCS keys, is another popular fingerprint used in similarity searching. The MACCS fingerprint is a 166 bit structural key, with each bit representing presence of a pre-specified substructure. Whereas Morgan fingerprints take all circular substructures present in a molecule and hash them onto a fixed-length bit string, MACCS fingerprints only represent presence of pre-defined substructures.

Having created bit string representations of chemicals using fingerprints, numerous methods for calculating similarity between the bit strings are available. The Tanimoto similarity coefficient[63] has been shown to be one of the best metrics for fingerprint-based similarity calculations.[64]

For two fingerprints, fingerprint A and fingerprint B, the Tanimoto similarity coefficient is calculated as:

$$Tanimoto\ similarity = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Where $|A|$ is the number of bits present in fingerprint A, $|B|$ is the number of bits present in fingerprint A, $|A \cap B|$ is the number of bits present in both fingerprint A and fingerprint B, and $|A \cup B|$ is the number of bits present in fingerprint A or fingerprint B.

The Tanimoto similarity coefficients ranges in values from zero for no bits in common to one for all bits present in one string being present in the other string.

Despite being one of the best and most popular metrics, using the Tanimoto coefficient between binary fingerprints to assess similarity is not without imperfections or bias.[65] This includes a tendency to give low similarity values to small compounds resulting in a bias towards small molecules when using Tanimoto similarity for diversity selection.[66]

1.5.2 KNIME

Konstanz Information Miner (KNIME) is a free and open source software for data analytics.[67] Integrated into the KNIME library are a wide range of nodes for chemical informatics purposes.

Molecular Substructure miner (MoSS) is a node in KNIME used for finding maximal common substructures between chemicals.[68] MoSS uses an algorithm called Molecular Fragment miner (MoFa).[69]

MoSS includes an option, known as "ring mining" to treat rings as fixed, indivisible units in the algorithm – when one bond in the ring is added to a fragment in the algorithm, the whole ring is added. The user can specify the size of the rings to apply this to. Ring mining greatly reduces the computational times required to run the program, particularly in structures with many rings such as steroids. It also results in aromatic rings being treated as indivisible units, so broken parts of aromatic rings will not be included in an outputted substructure. Including only parts of aromatic rings would become increasingly problematic when aromatic rings are written as Kekulé structures (alternating single and double bonds).

However, ring mining can lead to inaccurate counts of occurrence of some substructures in the training chemicals. Within the MoSS node when ring mining is used, the occurrence counts of substructures containing part, but not the whole, of a ring will not include chemicals which contain the whole ring. If the substructure were to be used outside of the MoSS node for substructure searching, the chemical containing the substructure as part of a ring would be included as containing the substructure in the search. An example is shown in Figure 1.10.

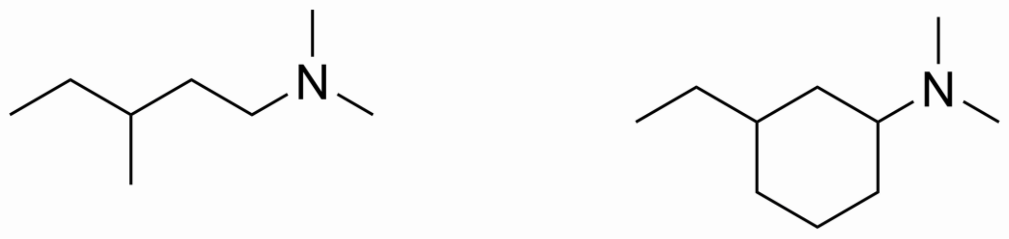


Figure 1.10: Left: an example substructure that could be found by the MoSS program. Right: a chemical which contains the substructure, but this would not be counted within the MoSS program because of ring mining. The substructure contains part of the ring, not the whole ring. Ring mining requires rings of a specified size to be treated as a single, indivisible unit.

Overall, the advantages of using ring mining make it a necessity when finding maximum common substructures with MoSS. The confusion of including only fragments of aromatic rings is avoided, and the increased speed of the algorithm is particularly important in large data sets. However, the user must be aware that the counts of occurrence for a substructure within MoSS may not be accurate if the substructure contains fragments of a ring. Accurate counts can be taken outside of the MoSS program, or the substructures should be used with the requirement that it must contain the entirety of ring.

1.6. Model Validation

To assess and compare performance of different models, a consistent selection of performance statistics has been chosen and used throughout this project. These are outlined here.

Model predictions can be split into four categories, as shown in the confusion matrix:

	Predicted Positive	**Predicted Negative**
Experimental Positive	"True Positive"	"False Negative"
Experimental Negative	"False Positive"	"True Negative"

Sensitivity (SE) is the proportion of experimentally active chemicals correctly predicted as active.

$$SE = \frac{TP}{TP + FN}$$

Specificity (SP) is the proportion of experimentally inactive chemicals correctly predicted as inactive,

$$SP = \frac{TN}{TN + FP}$$

Accuracy (ACC), is the proportion of all chemicals with predicted activity matching experimental activity.

$$ACC = \frac{TP + FP}{TP + TN + FP + FN}$$

Matthews Correlation Coefficient (MCC) is a measure of quality of binary classifications. It can take values between -1 and +1. A value of +1 would be obtained for a model predicting every value correctly, a value of 0 represents a model performing as well as random predictions, and a value of -1 would be obtained by predicting every value incorrectly.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is a particularly useful measure when considering data sets with classes (e.g. actives and inactives) of different sizes. In these cases, other performance statistics alone can be misleading.

Positive predictive value (PPV) is the proportion of positive predictions that are experimentally positive.

$$PPV = \frac{TP}{TP + FP}$$

Negative predictive value (NPV) is the proportion of negative predictions that are experimentally negative.

$$NPV = \frac{TN}{TN + FN}$$

1.7. Previous work and automated algorithms for identification of structural alerts

Knowledge-based structural alert methods, such as the commercially expert systems Lhasa's Derek Nexus[42] and Genetox Expert Alerts from Leadscope,[70] have been used for toxicity predictions. However, constructing these types systems is very time consuming, requires input from experts, and may suffer from human bias. Automated approaches using statistics to identify structural alerts do not suffer from these drawbacks. There have been numerous different approaches to automated generation of structural alerts, each with their own strengths and weaknesses.

Computer Automated Structure Evaluation (CASE) is a fragment-based approach.[71] Chemicals are broken down into linear subunits containing between three and twelve interconnected non-hydrogen atoms. All possible linear fragments between these sizes are derived from each chemical and the occurrence of each fragment in active and inactive chemicals is calculated. These numbers are analysed statistically. If the distribution of active and inactive chemicals is significantly skewed towards active chemicals, the fragment is identified as an activating "biophore". If the distribution is significantly skewed towards inactive chemicals, the fragment is identified as non-activating. Significant skew was initially defined as a distribution that would have had at most a 5% chance of being observed if the occurrence was random, assuming a binomial distribution. Multiple Computer Automated Structure Evaluation (MultiCASE) is similar to CASE, using hierarchical statistical analysis.[72] Where CASE uses all statistically significant fragments, MultiCASE uses the most statistically significant fragment at each iteration, any chemicals containing that fragment are removed from the training set, and the process repeated. For each fragment identified, other correlated fragments and physicochemical properties are used to create a QSAR specific to that biophore. The CASE and MultiCASE approaches are limited by the use of only linear subunits, meaning branching substructures are not accounted for.

Bioalerts is an open source Python library for automatically constructing structural alerts.[73] Substructures are defined by Morgan fingerprints of increasing size. As with CASE, the occurrence of each substructure in the active and inactive chemicals is counted and the probability of that distribution occurring randomly, assuming a binomial distribution, is calculated. A substructure is identified as a structural alert if this probability is below a threshold (for example 5%). The use of Morgan fingerprints means substructures are limited to circular environments, which may not give optimal results.

SARpy uses string mining to automatically construct structural alerts.[74] Molecules are input as SMILES strings and these are fragmented to describe substructures. Substructures are evaluated by likelihood ratio (as used in diagnostic testing), defined as the proportion of active chemicals containing the substructure divided by the proportion of inactive chemicals containing the substructure. A potential limit of likelihood ratio is that it returns a value of infinity for any fragment contained by no inactive chemicals and at least one active chemical. As a result, specific substructures, contained by no inactive chemicals and few active chemicals, will have large likelihood ratios. These overly specific alerts may not generalise well, giving poor predictions outside of the training set. The use of string mining limits substructures to atoms which occur next to each in the SMILES string, making branching difficult to account for.

Each of these algorithms for automatic generation of structural alerts differ in two key ways:

1. How substructures are derived.
2. The statistical approach used to accept or reject a substructure as a structural alert.

The approaches to derivation of substructures are not capable of dealing with branching substructures (fragment-based or string mining approaches) or non-circular environments (fingerprint-based approach). Whilst these approaches may be effective at identifying small substructures, such as those that are electrophilic and capable of causing DNA mutagenicity or skin sensitisation, they would struggle to deal with larger substructures, such as rings with branching features. Hence, these algorithms may not find the optimal substructures and may not be suitable when the optimal substructure is large or branching.

SAR models have different requirements when used for different purposes. For example, in risk assessment, a false negative prediction is the most dangerous type of error and as such, a SAR model for risk assessment should have as high sensitivity as possible, often at the expense of specificity. However, in drug discovery, confidence in active predictions is most important, so a model should have as high specificity as possible, often at the expense of sensitivity. The statistical approaches used in the previously discussed methods for automatically constructing structural alerts do not allow for this type of flexibility.

In this work, maximal common substructure searcher has been used to find the largest substructures common to chemicals in the training set. This does not limit the size or shape of the substructures. A statistical approach to accepting substructures as structural alerts has been used that is flexible, allowing the user to adjust a parameter to change the relative importance of number of actives and inactives in selection of substructures.

Prior work from within the Goodman group

Prior to this work, Allen *et al* have constructed structural alerts for the Bowes targets, published in 2016.[43] For each biological target, data was extracted from ChEMBL[75]. Substructures common to active chemicals were found using a maximal common substructure algorithm. Human analysis, aided by literature searches, was used to select which would be used as structural alerts. A small number of structural alerts were developed for each target (an average of 2.93 per target), but each covered many active chemicals. The ChEMBL database generally contains relatively few inactives for each target. For validation of the structural alerts, an assumption is made to provide additional inactive chemicals: for each target, chemicals present in data sets of other targets in the study are assumed to be inactive at the target of interest if they are not already present in that target's data set.

Concurrent to this work, Allen *et al* published updated structural alert-based models for the same targets in 2018.[44] The same database and methods for extracting data were used as the previous study, including the same method for collecting assumed-negatives for validation. The largest substructure common to 2% of the training set active chemicals was found, coded as a structural alert and chemicals containing the chemical removed from the training set. This is iteratively repeated until only one chemical is contained by the largest common substructures. Different filters are applied to the list of generated structural alerts to create two models for different purposes. A model designed with the highest possible sensitivity (at the cost of specificity) for use in screening chemicals is created by using all alerts that are contained by at least two chemicals in the training set. A second model, designed to have a higher specificity and overall performance, for use in risk assessment is created by using alerts that are contained by at least five chemicals in the training set and which are contained by more active chemicals than (assumed) inactive chemicals in the test set. The overall process for creating these models is summarised in Figure 1.11.

Both the screening and risk assessment models have significantly better performance metrics than the previous work in terms of sensitivity (proportion of experimentally active chemicals correctly predicted), specificity (proportion of experimentally inactive chemicals correctly predicted), accuracy (proportion of all chemicals correctly predicted) and Matthews Correlation Coefficient (MCC). Compared to the previous work, the individual structural alerts used in the new approach are generally larger in size and cover far fewer active chemicals. However, a greater number of structural alerts are used in each target, and the combination of these alerts leads to a model with better overall performance. In this thesis, comparisons will be made to Allen's updated models only, as these are the latest and better performing of the published models. New models will be compared to the "Screening" and "Risk Assessment" models.

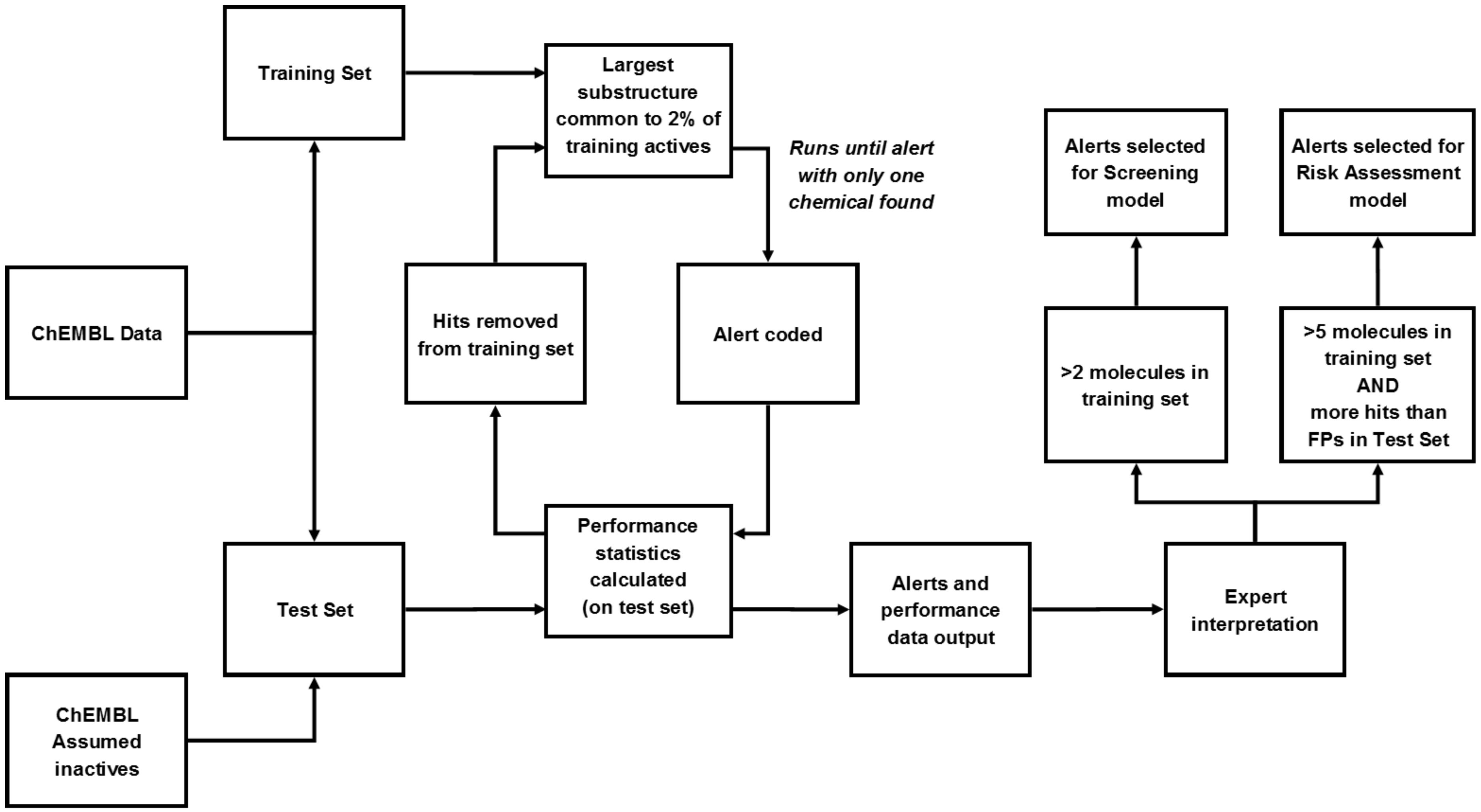

Training Set
ChEMBL Data
ChEMBL Assumed inactives
Test Set
Largest substructure common to 2% of training actives
Hits removed from training set
Runs until alert with only one chemical found
Alert coded
Performance statistics calculated (on test set)
Alerts and performance data output
Expert interpretation
Alerts selected for Screening model
>2 molecules in training set
Alerts selected for Risk Assessment model
>5 molecules in training set AND more hits than FPs in Test Set

Figure 1.11: Overview of the procedure used by Allen et al *to generate structural alert-based models.*[44] *Image adapted from Allen* et al *(2018).*

1.8. Aims of the project

The main aim of this project was to make interpretable predictions for MIEs based on chemical structures. In this thesis, numerous methods and ideas have been investigated to achieve this aim.

Firstly, I aim to construct improved structural alert-based SAR models for MIEs. I hypothesise that combining maximum common substructure searches with a statistical approach to structural alert selection, considering both active and inactive chemical, will improve model performance statistics. These statistics will be compared to performance statistics of previous models. This can be found in Chapter Two.

I aim to investigate the combination of structural alert models with a complementary model in a consensus approach. Where the model predictions are in agreement, there should be an improvement in performance statistics and an increase in confidence in predictions. This aim is explored in Chapter Two.

I aim to use the transparency of structural alert models to evaluate applicability of active predictions. Where a new chemical contains a structural alert and is more similar to the training set active chemicals containing the same structural alert, the active prediction should be considered more applicable. Hence, a greater proportion of active predictions being true positive should be seen as this similarity increases. This work is also found in Chapter Two.

In Chapter Three, I aim to expand the scope of existing models by making predictions for a wider range of MIEs of biological relevance to chemical risk assessment. Having constructed these models, I aim to explain any variation in model performance. I hypothesise biological targets with more data points should generally give better performing models.

Structural alert models return inactive predictions for chemicals containing no structural alerts. I aim to increase confidence in these predictions of inactivity derived from structural alert-based SAR models. Methods for increasing confidence in inactive predictions have been applied to structural alerts for reactivity-driven MIEs.[76] I discuss these further in Chapter Four and predict that applying similar methods to structural alerts for receptor binding MIEs will increase confidence in inactive predictions, although reactivity-driven MIEs and receptor binding MIEs are mechanistically different. The negative predictions identified as confident negative predictions should have a higher proportion of true negative chemicals than those which are not considered confident.

In Chapter Five, I aim to create a method for constructing generalised aromatic structural alerts. I predict that these generalised alerts will reduce the number of structural alerts in each model. Furthermore, the generalised alerts will provide insight into the features required for receptor binding and, hence, provide information regarding the mechanism of binding for each biological target.

Finally, I aim to expand upon the structural alerts by using them to construct 3D pharmacophore models. Pharmacophore models will be constructed from chemicals grouped by structural alerts. I predict that this will give models that are high performing and more general, allowing for better predictions in new areas of chemical space.

2. Automated workflow for construction of structural alert-based structure-activity relationships

2.1. Creation of New Balanced Data Sets

ChEMBL is a publicly available, manually curated database of bioactive chemicals.[75] It contains 14 million activity values for more than 1.6 million different compounds from 1.2 million assays.[77] The data is extracted from scientific literature, deposited data sets and from other databases. The chemicals are mostly pharmaceuticals or of interest to drug design. For each report of activity, ChEMBL includes a wealth of data, including quantitative measurements of activity, what assay was used for the measurement, and a "confidence score". This confidence score relates to how certain the assay is measuring the assigned biological target and also reflects the type of target assigned. For example, a confidence score of nine relates to a direct single protein target being assigned, a confidence score of eight relates to a homologous single protein target being assigned, whilst a confidence score of seven relates to assignment of a direct protein complex subunit.

ChEMBL includes different measurements of biological activity through receptor binding, including:

- The half maximal inhibitory concentration (IC50) – the concentration of inhibitor required to half the reaction rate of an enzyme-catalysed reaction.
- The half maximal effective concentration (EC50) – the concentration of a compound that gives half-maximal response. For antagonists, this is the same as the IC50.
- Inhibition constant (K_i) – the equilibrium constant of the dissociation of inhibitor-enzyme complex.
- Dissociation constant (K_d) – the equilibrium constant of the dissociation of a ligand-enzyme complex (not limited to inhibitors).

These measurements therefore relate to different aspects of receptor binding. IC50 and K_i measure antagonism, whilst EC50 and K_d measure both agonism and antagonism. Agonists and antagonists cause opposite changes in biological activity but often bind at the same receptor binding site with binding modes that differ only slightly.[78] K_i and K_d are measures of affinity – the capability of a ligand to bind to a receptor – whilst IC50 and EC50 are measures of efficacy – the change in activity of the receptor. Affinity and efficacy are not equivalent, but both measurements provide data regarding receptor binding. In order to create the best possible models, as much data as possible was required. Following precedent,[44,79] all data for these different aspects of receptor binding was combined in this work. A qualitative activity cut-off of 10 µM was used, as used in preceding work.[44,79,80] This cut-off represents both highly and marginally active compounds. For most biological targets, ChEMBL has much more data on active chemicals than inactive chemicals with this cut-off.

Previously, Allen *et al*[43,44] have constructed models for Bowes targets[55] using ChEMBL data. For validation of the models, additional inactive chemicals are required. An assumption is made to provide additional inactive chemicals: for each target, chemicals present in data sets of other targets in the study are assumed to be inactive at the target of interest if they are not already present in that target's data set. Whilst this is a reasonable assumption, it would be preferable to use only data points that have been directly tested at the biological target so that no assumptions need to be made.

The Toxicity Forecaster[81] (ToxCast) is a program run by the Environmental Protection Agency (EPA). It uses high throughput screening methods to rapidly generate data for large numbers of chemicals, providing publicly available data for over 9 000 chemicals across 1 000 assays. Many of the chemicals run in the early stages of ToxCast were pesticides, insecticides or other known toxicants of interest. They were often run across many assays, eliciting activity at a small proportion of these. As a result, ToxCast has more data on inactive chemicals than active chemicals for most targets.

Recent work has highlighted a phenomenon affecting ToxCast known as the cytotoxic burst phenomenon.[82] In high throughput platforms, large numbers of assays for different targets give active responses when chemicals approach cytotoxic concentrations, suggesting a cytotoxic mechanism affects the entire cell, as opposed to the chemical activating each target individually. This leads to false reports of experimental activity at the biological targets, and one should be aware of this when analysing ToxCast data. However, the cytotoxic burst has not been directly accounted for in this work and instead has been treated as experimental error associated with ToxCast data.

In this section, new data sets have been made by combining data from ChEMBL with data from ToxCast. The new data sets have balanced numbers of active and inactive datapoints.

2.1.1. Method

For each target, bioactivity data for Homo sapiens was downloaded from ChEMBL (data extracted April 2018). Activity reports were filtered to remove any with a confidence score of less than eight, meaning that it comes from assays which assign the single protein target directly or through a homologous single protein target. Hence, all data comes from human *in vitro* assays.

Only activities reported with "Standard Units" of nM were kept, leaving reports of EC50, IC50, K_i and K_d. RDKit[83] Salt Stripper was used to remove common salts and counter ions from chemicals (e.g. chloride, bromide, sodium, magnesium, and nitrate ions, and many more). It was assumed these common salts are not involved in the MIE, and variation in activity when the counter ions are changed was due to experimental errors. All chemicals with more than 100 atoms were removed. The common substructure algorithm can get stuck when there are several large similar molecules, such as large proteins with repeating structures. The 100 atom threshold was found to remove the troublesome large chemicals whilst keeping the majority of chemicals.

The SMILES strings were re-written to be canonical using RDKit, such that the chemical strings are written in a consistent way across all entries from both ChEMBL and ToxCast.

For each chemical, mean activity was taken – values of activity that were reported as "greater than" a certain value were removed for these calculations. The mean of the activities was taken for each chemical, as to use all the data available. However, if this work is to be used in risk assessment in the future, the most active report for each chemical could be used to give a "worst case scenario" prediction. Chemicals with a mean activity of 10 000 nM or lower were assigned as active; those with over 10 μM were assigned as inactive. This activity cut-off is consistent with Allen's 2016 and 2018 models.[43,44] However, other studies have used different activity limits for different biological targets.[84] In future, applying different activity limits in this work could be investigated.

For each target, human data was downloaded from the ToxCast Dashboard (data extracted November 2016). Toxcast's in-built assignment of binary activity[85] was used for data from the ToxCast database. This involves using an algorithm to assign the best model to an activity-concentration series, and from this, a binary activity was assigned. In the future, AC50 values could be extracted from ToxCast and a more transparent activity cut-off could be used.

As with ChEMBL data, common salts and counter ions were stripped using RDKit Salt Stripper, chemicals with greater than 100 atoms were removed, and SMILES strings were re-written using RDKit. If a chemical has contrasting reports of being both active and inactive in different assays it was considered active.

Data from ChEMBL and ToxCast were combined into one data set. Where chemicals have contrasting activity reports between ChEMBL and ToxCast, the activity from ChEMBL was used, as to prioritise ChEMBL's human run assays over ToxCast's machine run high throughput assays. However, in future risk assessment work, the most active report could be used to once again give a "worst case scenario".

Initially, the holdout method was used for constructing data sets, with chemicals split randomly with roughly 75% forming the training set and 25% forming the test set. Later, four-fold cross-validation was also used and the results compared to the results obtained by holdout.

2.1.2. New data sets for the Bowes targets

The Bowes targets represent key MIE for toxicology predictions. Of the 44 targets, 24 have data from human *in-vitro* assays in ToxCast. These also have data in the ChEMBL database. New data sets have been constructed for the selection of 24 Bowes targets, and are summarised in Table 2.1.

Combining the data from both databases gives balanced data sets in terms of number of active and inactive chemicals, with ChEMBL providing most of the actives and ToxCast providing most of the inactives. However, ChEMBL and ToxCast generally cover different areas of chemical space, with ChEMBL mostly containing data for pharmaceutical chemicals and ToxCast containing mostly containing data for pesticides and other reactive chemicals which generally are not particularly structurally similar to the ChEMBL chemicals. Hence, the combined data sets could be viewed as having a different kind of imbalance, with the inactive chemicals largely being chemically dissimilar from the active chemicals. Nevertheless, combining the databases provides much more useful data sets for model construction than either imbalanced database individually.

All data points in the data sets come directly from assays testing activity at the biological target. Unlike previous methods, no assumptions are needed to find further inactive data points, and so no additional uncertainty is introduced into the models constructed from the data sets.

The new data sets contain only data from Homo sapiens *in vitro* assays, so models and predictions built from the data will be relevant to humans without the need of cross-species extrapolation, although *in vitro* to *in vivo* extrapolation will be required and has its own challenges.[86]

	Training Sets						Test Sets					
	ChEMBL		ToxCast		Both		ChEMBL		ToxCast		Both	
Target	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives
Acetylcholinesterase	1945	746	45	645	14	76	595	250	13	222	2	24
Adenosine A2a receptor	2903	118	46	1366	8	98	974	42	10	430	2	29
Alpha-2a adrenergic receptor	598	32	42	672	9	65	182	6	9	215	5	24
Androgen receptor	1340	171	114	3668	535	1630	450	65	34	1219	164	531
Beta-1 adrenergic receptor	923	72	33	659	4	75	281	26	15	230	5	20
Beta-2 adrenergic receptor	1425	765	28	664	9	72	476	275	5	219	2	19
Delta opioid receptor	2194	180	38	651	8	73	749	61	15	231	2	23
Dopamine D1 receptor	948	104	65	1287	16	93	296	24	19	458	8	25
Dopamine D2 receptor	4206	128	47	656	9	71	1418	40	12	220	4	22
Dopamine transporter	1725	111	122	1250	23	84	586	35	49	410	6	27
Endothelin receptor ET-A	950	62	10	724	2	73	320	19	3	242	0	32
Glucocorticoid receptor	1510	56	523	4625	241	562	508	22	153	1530	83	178
hERG	3581	1669	21	647	15	105	1263	587	10	202	5	36
Histamine H1 receptor	923	76	34	676	6	70	298	44	13	213	2	27
Mu opioid receptor	2582	316	68	1308	11	99	915	95	30	462	5	28
Muscarinic acetylcholine receptor M1	1440	170	35	694	9	73	516	45	13	238	2	22
Muscarinic acetylcholine receptor M2	1163	108	45	1307	17	91	395	32	16	463	0	32
Muscarinic acetylcholine receptor M3	1140	108	46	659	4	72	326	36	16	215	5	24
Norepinephrine transporter	2097	89	109	1265	17	95	646	32	37	439	6	21
Serotonin 2a (5-HT2a) receptor	2752	35	34	662	11	74	951	12	6	232	3	19
Serotonin 3a (5-HT3a) receptor	332	26	15	748	1	4	97	11	7	266	0	0
Serotonin transporter	3015	113	26	675	16	64	976	33	7	225	3	25
Tyrosine-protein kinase LCK	1281	175	2	225	0	5	446	51	3	66	0	2
Vasopressin V1a receptor	451	21	4	689	1	77	160	5	2	241	1	26

Table 2.1: Summary of the new data sets created using data from both ChEMBL and ToxCast databases. These data sets are for twenty-four of the Bowes targets which had human in vitro data in both databases. Actives in ChEMBL are chemicals with a mean activity of less than 10 µM and inactives as chemicals with a mean activity of greater than 10 µM. ToxCast's inbuilt definitions of activity are used for chemicals in ToxCast. As shown in the table, for all targets, ChEMBL provides most of the active data and ToxCast provides most of the inactive data. Whilst the differing sources of data may introduce a different kind of imbalance, the new data sets are balanced in terms of number of active and inactive chemicals.

2.2. An Automated workflow for construction of structural alert-based structure-activity relationships

2.2.1. Bayesian Statistics

In Bayesian statistics, probability is a description of how certain you are that something is true.[87,88] If you are very sure of something, new data is unlikely to change your mind. Bayesian statistics is a broad subject which has been applied in many different ways for many purposes. Madigan *et al.* have previously reviewed some applications of Bayesian statistics in pharmacology.[89]

In this chapter, SAR models were constructed by iteratively selecting common substructures occurring in training chemicals to be coded as structural alerts. At each iteration, one could observe how many active and inactive chemicals contain different substructures. How could this be used to systematically pick the best performing structural alert?

Simply using the ratio of occurrence in actives to occurrence in inactives, as used in SARpy,[74] would result in near-exclusive selection of substructures which occur in no inactives. For example, a substructure which occurred in two actives and zero inactives would be picked as a structural alert ahead of a substructure which occurred in 200 actives and one inactive.

As discussed in Section 1.7, previous statistical approaches have used a significance level test with a single binomial distribution to identify activating or non-activating substructures. If the probability of randomly producing the observed distribution of active and inactive chemicals is sufficiently low, the substructure is identified as either activating or non-activating. This approach does not allow for any adjustment in relative weighting of active and inactive chemicals.

An alternative approach is, for a given distribution of active and inactive chemicals contained by a substructure, to compare the probabilities of the distribution being given by two models - one model being biased towards active chemicals and the other model being random. The greater the probability given by the biased model compared to the random model, the more activating a substructure is. Changing how biased the biased model is changes the relative weight of active and inactive chemicals - the greater the bias, the more active chemicals required per inactive chemical. This comparison in probability from competing models can be done by calculating Bayes Factor in Bayesian statistics. Using Bayesian statistics for this model comparison allows for more flexibility in how the problem is approached. Bayesian statistics do not require the bias of the biased model to be fixed but allow it to be treated as unknown parameter to be fit to a distribution. However, if the bias model is fixed, the equation derived from Bayes statistics is equivalent to the Neyman-Pearson lemma in classical statistics.[90] The key difference from the

Neyman-Pearson lemma is that the fixed model and the prior likelihood of each model occurring have been explicitly stated, and these can be changed later. In this work, only the effect of using different fixed biases will be explored. The model comparison will be set up as a Bayesian equation so that the bias can easily be fitted to a distribution in future.

2.2.1.1. Bayes Theorem

Bayes Theorem, when considering the appropriateness of a model, *M_i*, for given data, *D*, is written as

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{p(D)}$$

Where: $p(M_i|D)$ is probability of model *i* occurring given data

$p(D|M_i)$ is probability of data occurring for the model *i*

$p(M_i)$ is probability of the model *i* occurring

$p(D)$ is probability of the data occurring

When a second possible model, *M_j*, is considered, this becomes:

$$\frac{p(M_i|D)}{p\left(M_j|D\right)} = \frac{p(D|M_i)}{p(D|M_j)}\frac{p(M_i)}{p(M_j)}$$

Where Bayes Factor is defined as:

$$Bayes\ Factor = \frac{p(D|M_i)}{p(D|M_j)}$$

$p(D|M_i)$ and $p(D|M_j)$ can be calculated, and so Bayes Factor is a value that can be calculated.

Assuming that $p(M_i) = p(M_j)$, i.e. that before any data has been considered, both models are equally likely to occur:

$$\frac{p(M_i|D)}{p\left(M_j|D\right)} = Bayes\ Factor$$

Hence, the likelihood of two models occurring for given data can be compared by calculating Bayes Factor. Bayes Factor indicates how many times more likely one model, *M_i* is compared to another model, *M_j*.

2.2.1.2. Using Bayes Factor to pick structural alerts

For each substructure, the given data (D) is the number of actives containing the substructure and number of inactives containing the substructure.

Bayes factor is calculated, comparing between two Binomial models defined as:

- M_{bias} – a model which is bias towards active predictions. M_{bias} predicts active with a probability of θ and inactive with a probability of (1 – θ), where 0.5 < θ < 1.
- M_{random} – a model which predicts activity randomly. It predicts active with a probability of 0.5 and inactive with a probability of 0.5

$$Bayes\ Factor = \frac{p(D|M_{bias})}{p(D|M_{random})}$$

$$Bayes\ Factor = \frac{\theta^{actives}(1-\theta)^{inactives}}{0.5^{actives}0.5^{inactives}}$$

Where "$actives$" is the number of actives containing the substructure, and "$inactives$" is the number of inactives containing the substructure.

The value for theta (θ) must be selected by the user and will result in different priorities when selecting from a list of substructures. This is explored further in later sections.

Taking the logarithm of the previous equation made it easier to handle large values of $actives$ and $inactives$:

$$\log(Bayes\ Factor) = actives \times \log\theta + inactives \times \log(1-\theta) - (actives + inactives)\log 0.5$$

Simply put, Bayes Factor can be viewed as a scoring system for each substructure. The more actives containing a substructure, the higher the value of Bayes Factor. The more inactives containing a substructure, the lower the value of Bayes Factor. Adjusting the value of θ changes the relative scoring of active and inactive – a greater value of θ will result in greater increases in Bayes Factor from active chemicals and greater decreases in Bayes Factor from inactive chemicals.

2.2.2. Methods

A training set of chemicals (in SMILES format) and binary activities was inputted into the workflow. The maximum common substructures occurring in at least two of the active chemicals were found using the MoSS node[68] in KNIME.[67] MoSS will only output substructures which occur in less than a specified percentage of the inactive chemicals. This value was a parameter which can be selected by the user.

MoSS outputs the common substructures and how many times each occurs in the active and inactive chemicals, according to the MoSS algorithm. However, these values are slightly inaccurate due to ring mining used in the algorithm. Re-calculating accurate counts for all substructures output by MoSS would be too time consuming as often many thousands of substructures were output. Instead, Bayes Factor was calculated for each substructure using the occurrence in actives and inactives calculated by MoSS, and only the substructures with the 65 largest values are kept. It was assumed here that the inaccuracies in the counts given by the MoSS algorithm were not so large that the actual best performing substructure was not in the top 65 substructures. Accurate values for occurrence of active and inactive chemicals were calculated for the 65 substructures, and Bayes Factor recalculated. Only the substructure with the highest value of Bayes Factor was kept. When two substructures had the same Bayes Factor, the substructure which occurs in more active chemicals was chosen.

The user decided the lower bounds for a structural alert in terms of number of actives and inactive chemicals, and the lower bounds Bayes Factor was calculated using these values. If the remaining substructure had a Bayes Factor larger than the lower bounds and was contained by more actives than the minimum required number, it was added to the list of structural alerts. Any active chemicals containing the substructure were removed from the training set and the whole process was repeated iteratively until no substructures satisfied the lower bounds for an alert.

This iterative process produced a list of independent structural alerts. Chemicals containing a structural alert were predicted to be active, and those containing no alerts were predicted to be inactive.

The resulting model was applied to both the training set and test set, and performance statistics were calculated for both.

The process is summarised in Figure 2.1.

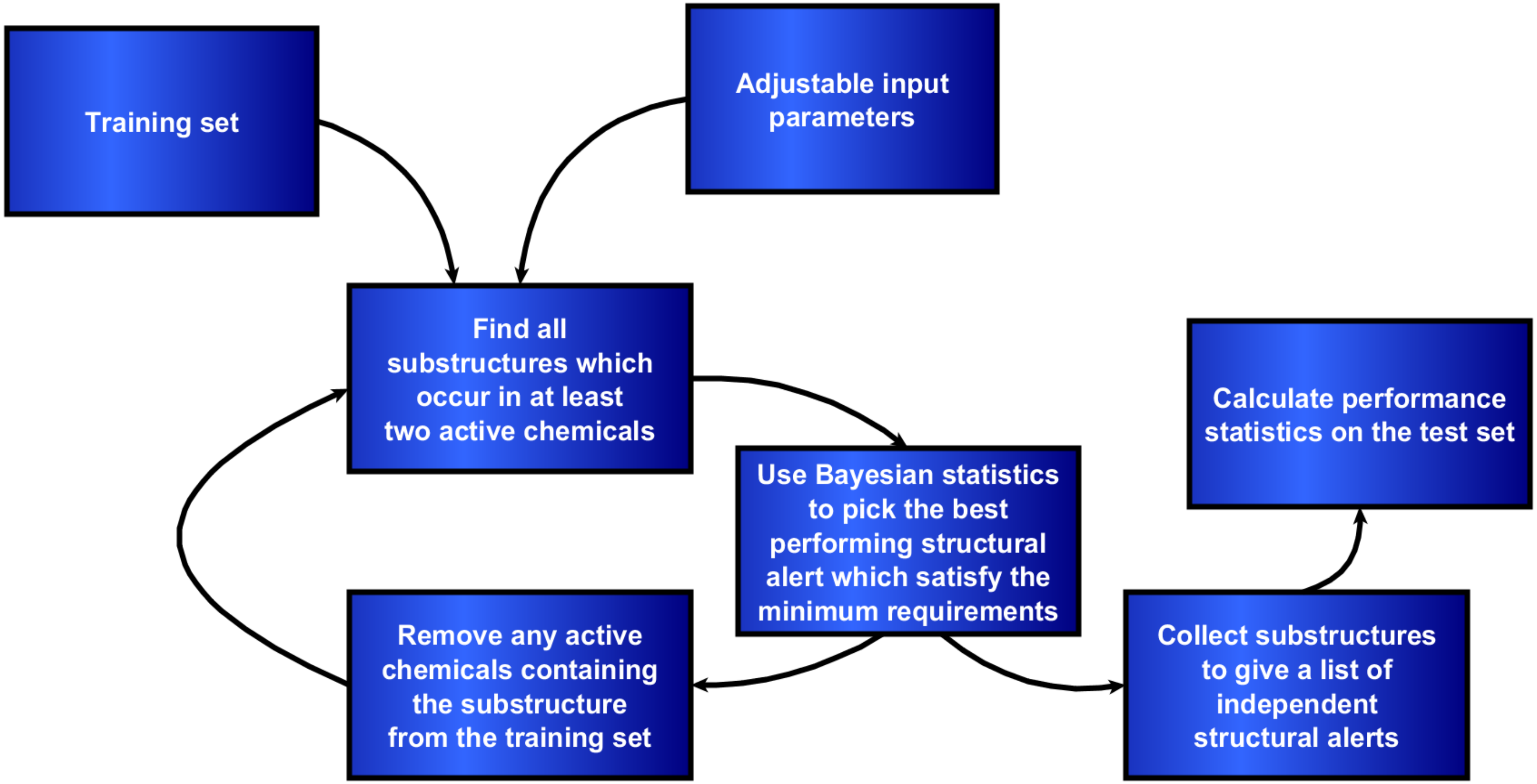

Training set
Adjustable input parameters
Find all substructures which occur in at least two active chemicals
Use Bayesian statistics to pick the best performing structural alert which satisfy the minimum requirements
Calculate performance statistics on the test set
Remove any active chemicals containing the substructure from the training set
Collect substructures to give a list of independent structural alerts

Figure 2.1: An overview of the automated workflow for creating structural alert-based SAR models. The adjustable performance metrics are theta value in Bayesian statistics, lower bounds for a structural alert, and maximum occurrence of a substructure in the inactive chemicals.

2.2.2.1. Adjustable parameters

There are three adjustable parameters in the workflow:

1. **Theta value in Bayesian statistics.** The “model bias towards active predictions” being compared to the random model in the Bayes Factor calculation. The bias model predicts active with a probability of “theta”, where $0.5 < \text{theta} < 1$. The user can choose the value of theta.
2. **Lower bounds for a structural alert**. The user can select minimum requirements, in terms of the occurrence in actives and the occurrence in inactives. Bayes Factor is calculated using these values. A substructure must have a Bayes Factor greater than or equal to this value and must be contained by the minimum number of actives to be considered as a structural alert.
3. **Maximum percentage occurrence of a substructure in the inactive chemicals**. This is a parameter used by the MoSS node. Choosing a larger value results in many more common substructures being identified by MoSS, but these additional substructures (contained by more inactive chemicals) may not be the statistically best performing substructure, so may not which structural alerts are selected. This results in significantly longer computational time required to run the workflow, particularly for large data sets, but may not significantly change the resultant structural alert models. Therefore, tuning of the parameter is required to find a suitable balance between computational efficiency and evaluating as many substructures as possible.

Different models have been created to show the effects of varying each of these parameters individually, whilst maintaining the other parameters.

2.2.2.2. Example Models

Two different sets of parameters have been used to showcase how the automated workflow can be used to create different models for different purposes.

- **“Screening” model.** Parameters: theta = 0.51; 15% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert of two actives and one inactive.
- **“Risk assessment” model**. Parameters: theta = 0.95; 1% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert of two actives and one inactive.

2.2.3. Results and discussion

2.2.3.1. Effect of varying parameters

Theta in Bayes Factor calculation

The model bias towards active predictions has a probability of theta of predicting active, where 0.5 < theta < 1. Changing the value of theta changes what the automated workflow will prioritise when choosing which of the substructures will be a structural alert in each iteration.

As theta is decreased, Bayes Factor becomes largest for the substructure with the largest occurrence in the actives and is less affected by the occurrence in the inactives. As the value of theta approaches 0.5, the equation for Bayes Factor approaches a value of 1 regardless of number of actives or inactives containing the substructure, so substructures are ordered by the secondary filter of number of actives.

As the value of theta is increased, Bayes Factor becomes largest for the substructure with the fewest occurrences in the inactives and is less affected by the occurrence in actives. As the value of theta approaches 1, the equation for Bayes Factor contains a $0^{inactives}$ term, so will only return non-zero values for substructures contained by no inactive chemicals.

An example of the effect of changing theta is shown in Table 2.2.

			Bayes Factor	
Substructure	**Actives**	**Inactives**	**With theta of 0.51**	**With theta of 0.95**
"A"	6	1	1.1	4.7
"B"	60	10	2.7	**5.3E+06**
"C"	120	40	**4.8**	2.8E-07

Table 2.2: An example of how different theta values lead to different substructures being picked in the automated workflow for construction of structural alert-based models. With a low value of theta, substructure "C" would have the highest Bayes Factor, but with a high value of theta, substructure "B" – occurring in fewer actives but also fewer inactives – would have the highest Bayes factor.

A value of theta can be chosen by the user to match their purpose. For example, for the purposes of screening large numbers of chemicals and not missing any active chemicals, a low value of theta would be used. This would result in more true positives (higher sensitivity) at the expense of more false positives (lower specificity).

For the purpose of identifying potential pharmaceutical lead compounds, confidence in actives is most important and as such the user would want to ensure there are as few false positives as possible. In this case, a high value of theta should be used, leading to a model which identifies fewer false positives (higher specificity) at the expense of fewer true positives (lower sensitivity).

The effect of changing theta can be seen in Figure 2.2: as theta increases, sensitivity decreases and specificity increases. At a theta value of 0.75, the otherwise smooth trends in performance metrics appear to hit a bump. This is because choosing the top performing substructure is a discrete process. Whilst changing theta leads to continuous changes in the values of Bayes Factor for each substructure, there will not necessarily be a change in which substructure is statistically considered best in each iteration. Hence, there will not always be a continuous change in performance metrics.

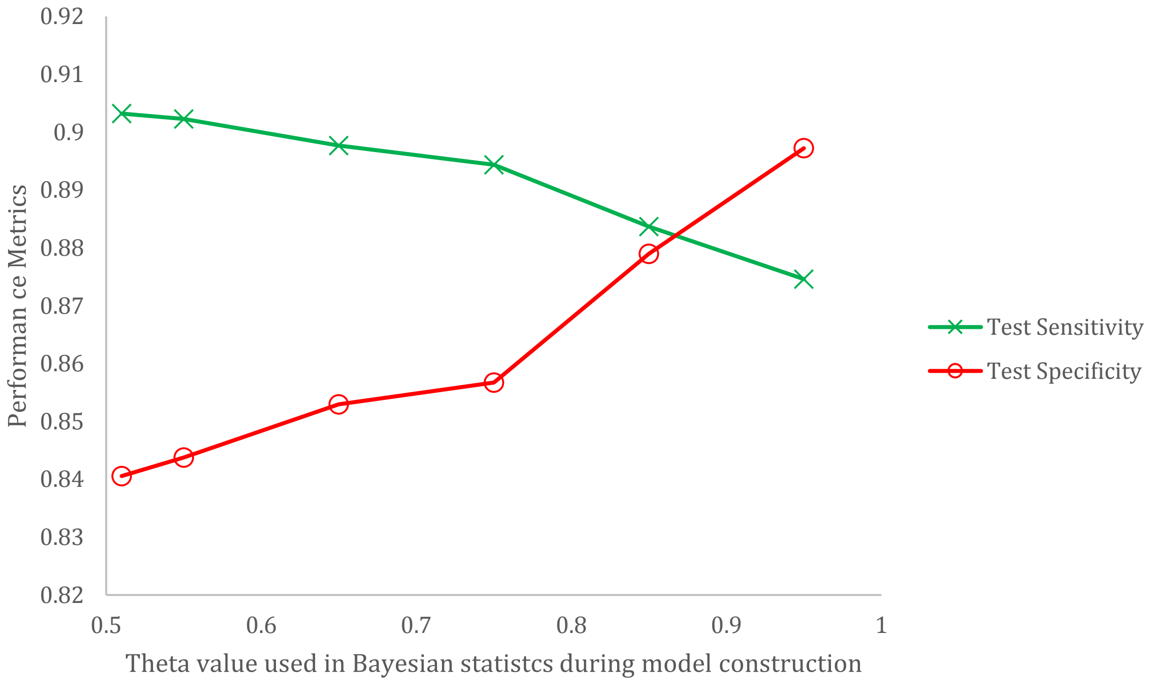


Figure 2.2: The variation of performance metrics (sensitivity and specificity) as the theta value used in the Bayes statistics in model creation is changed. Models are built for the selection of 24 Bowes targets and the means of the performance metrics are calculated across all targets. The other parameters in model construction are kept constant at 5% maximum occurrence of an alert in the inactive chemicals and lower bounds for an alert of two actives and one inactive.

Increasing theta results in selection of structural alerts that contain fewer training set false positives, but also fewer true positives. The structural alerts will be more specific and tend to be larger in size. A larger number of these structural alerts are required to cover the training set true positives, so models built with larger theta values will contain more structural alerts, as shown in Figure 2.3. As with Figure 2.2, a slight deviation from the smooth trend is seen at a theta value of 0.75 for the same reasons.

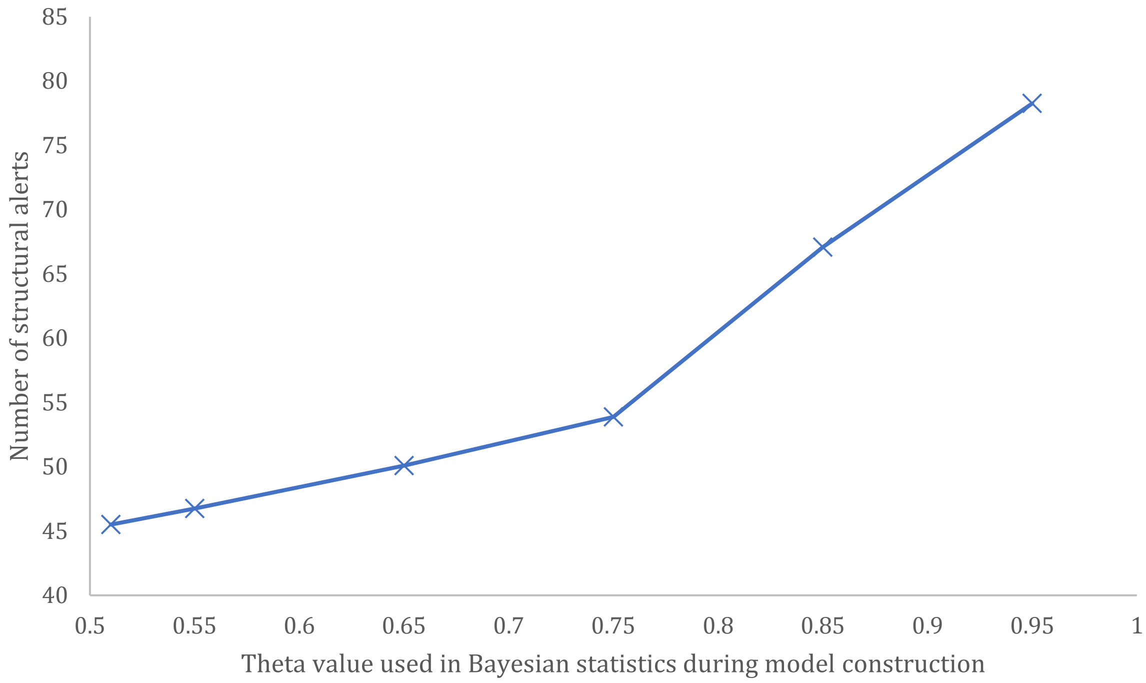


Figure 2.3: The number of structural alerts in a model metrics as the theta value used in the Bayes statistics in model creation is changed. Models are built for the selection of 24 Bowes targets and the means of the performance metrics are calculated across all targets. The other parameters in model construction are kept constant at 5% maximum occurrence of an alert in the inactive chemicals and lower bounds for an alert of two actives and one inactive.

Lower bounds for a structural alert

The user can choose the lower bounds for a structural alert, in terms of the minimum required number of actives and the maximum number of inactives for that minimum number of actives. A minimum required Bayes Factor is calculated for these numbers of actives and inactives. A structural alert must have a Bayes Factor greater than or equal to the minimum required Bayes Factor, and it must be contained by at least the minimum required actives. Less stringent lower bounds can be implemented by increasing the minimum required actives or by decreasing the maximum number of inactives for the minimum required actives.

Increasing the maximum number of inactives in the lower bounds means more structural alerts will be allowed, increasing the number of true positives covered by the overall model and hence increasing sensitivity. The additional alerts will also cover more false positives, so specificity will decrease. These trends are shown in Figure 2.4.

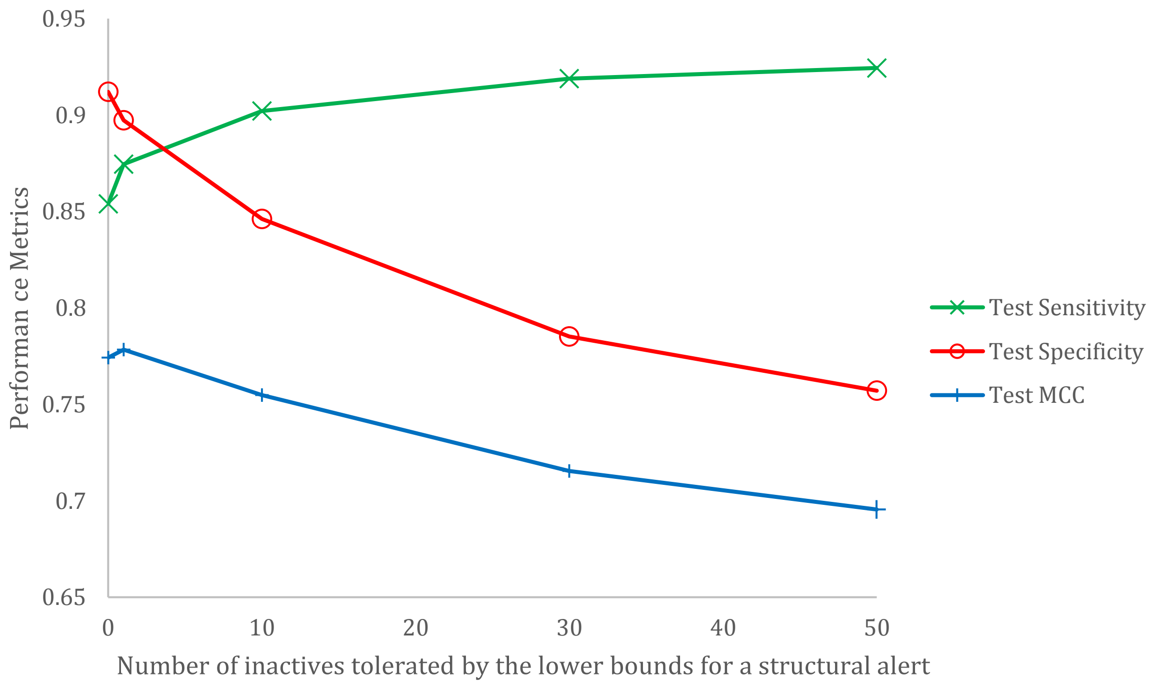


Figure 2.4: Variation of performance metrics with the number of inactives in the lower bounds for a structural alert. The lower bounds require two actives and a varying maximum of inactive chemicals. Increasing the maximum number of inactives increases sensitivity but decreases specificity Models are built for the selection of 24 Bowes Targets and the means of the performance metrics are calculated across all targets. The other parameters in model construction are kept constant with a theta value of 0.95 and 5% maximum occurrence of an alert in the inactive chemicals.

Increasing the minimum required actives in the lower will have the opposite effect to increasing the tolerated inactives in the lower bounds. Fewer structural alerts will be allowed, decreasing sensitivity, but fewer false positives will be covered so specificity will increase. These trends are shown in Figure 2.5.

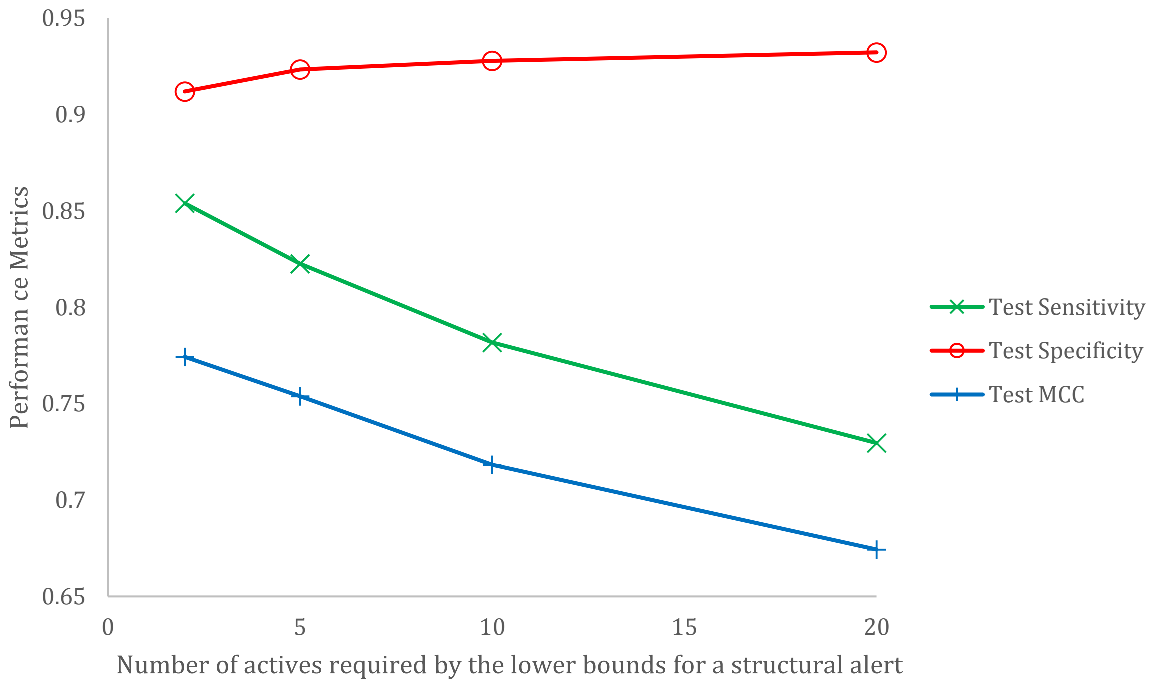


Figure 2.5: Variation of performance metrics (Matthews correlation coefficient, sensitivity and specificity) as the lower bounds for a structural alert are changed. Increasing the minimum number of actives contained by a structural alert increases specificity but decreases sensitivity. In each case, for the lower bounds for an alert are a varying number of active chemicals and no inactive chemicals. Models are built for the selection of 24 Bowes targets and the means of the performance metrics are calculated. The other parameters in model construction are kept constant with a theta value of 0.95 and 5% maximum occurrence of an alert in the inactive chemicals.

In the results shown in Figure 2.5, as the minimum required actives is increased from two, the specificity increases only slightly but the sensitivity decreases greatly. Hence, in this work a minimum requirement of two actives is used throughout.

Maximum percentage occurrence of a structural alert in the inactive chemicals

The primary purpose of varying the maximum percentage occurrence of a structural alert in the inactive chemicals is to change the computational time of the workflow – using a smaller value will lead to shorter computational times.

Varying this parameter also has some effect on the models built. Increasing the maximum percentage occurrence of a structural alert in the inactive chemicals means a substructure which occurs in a greater number of training inactive chemicals can be selected as a structural alert, providing it has a higher Bayes Factor than all other substructures.

The parameter "maximum percentage occurrence of a structural alert in the inactive chemicals" has a larger effect on models built with lower theta values.

For models built with low theta values, Bayes Factor tends to be greatest for substructures with the most occurrence in the active chemicals and is less affected by number of false positives. Substructures with higher maximum percentage occurrence in the inactive chemicals can be selected as structural alerts if the increase in occurrence in inactive chemicals is matched with an increase in occurrence in the active chemicals. These alerts will lead to models which will hit more true positives, but also more false positives. This can be seen in Figure 2.6, with sensitivity increasing but specificity decreasing as the maximum percentage occurrence in the inactive chemicals increases. There is also a slight decrease in MCC.

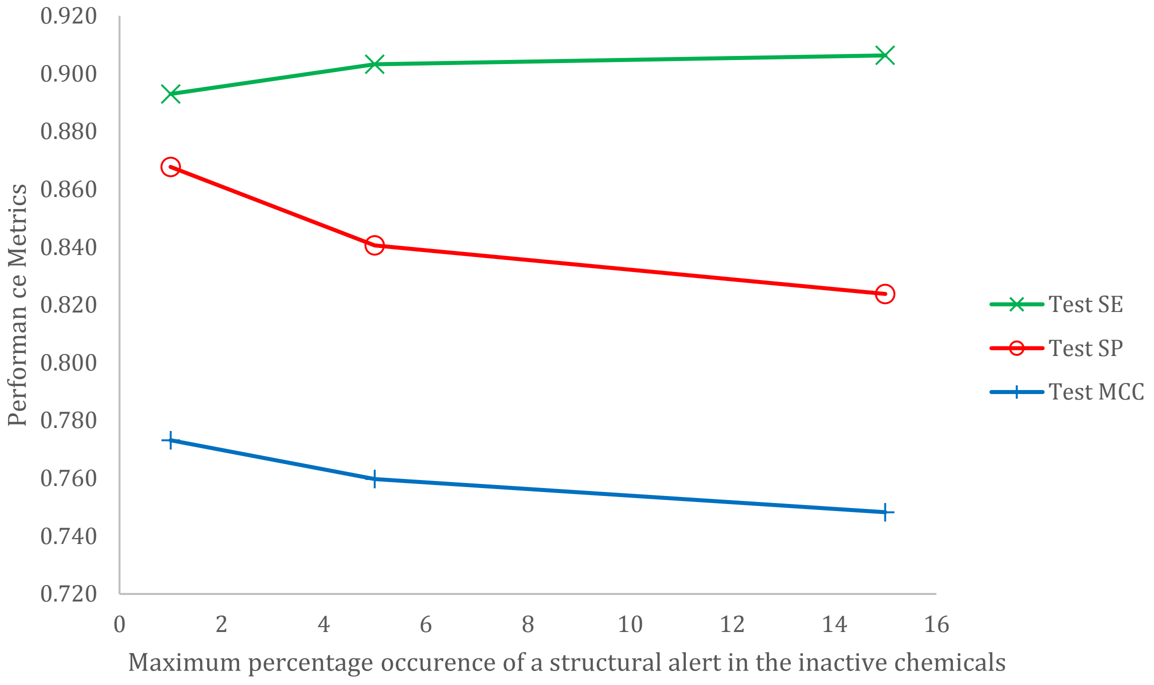


Figure 2.6: The variation of performance metrics (Matthews correlation coefficient, sensitivity and specificity) with maximum percentage occurrence in active chemicals for a model with theta of 0.51. All models have the same lower bounds for a structural alert (two actives and one inactive). Models are built for the selection of 24 Bowes Targets and the means of the performance metrics are calculated across all targets.

When models are built with high theta values, Bayes Factor tends to be greatest for substructures which are contained by few false positives. As such, increasing the maximum percentage occurrence of a structural alert in the inactive chemicals has only a small effect on performance metrics. Increasing this parameter still leads to an increase in sensitivity and a decrease in specificity in most targets. However, these changes are smaller in magnitude than for models with low values of theta, as can be seen in Figure 2.7.

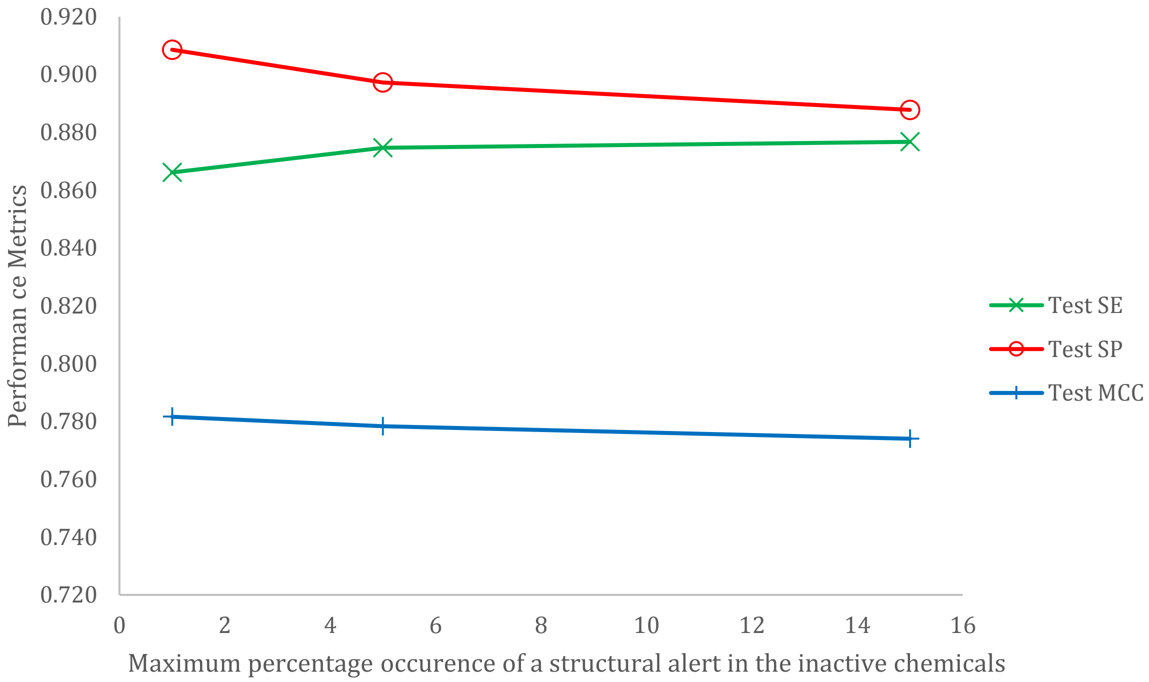


Figure 2.7: The variation of performance metrics (Matthews correlation coefficient, sensitivity and specificity) with maximum percentage occurrence in active chemicals for a model with theta of 0.95. All models have the same lower bounds for a structural alert (two actives, one inactive). Models are built for the selection of 24 Bowes Targets and the means of the performance metrics are calculated.

2.2.3.2. Example Models

Two different sets of parameters have been used to create two models which give examples of how the automated workflow can be used for different purposes.

Parameters were chosen to create a model for the purposes of screening large numbers of chemicals and not missing any active chemicals. A set of parameters was chosen to create a model which maximises sensitivity at the expensive of specificity. This is known as the "screening" model. The parameters are: theta of 0.51; 15% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert: two actives, one inactive.

A different set of parameters was chosen to create a model which prioritises minimising false positives and maximising specificity whilst creating structural alerts to cover as many of the active chemicals as possible. This is known as the "risk assessment" model. The parameters are: theta of 0.95; 1% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert: two actives, one inactive.

As well as showing how the workflow can be used for different purposes, setting up the models in this way facilitates comparisons between Allen's "Screening" and "Risk assessment" models.[44]

The performance of the two models on the Bowes targets is shown in Tables 2.3 and 2.4.

		Training set								Test set							
Target	**Alerts**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
Acetylcholinesterase	166	1792	83	212	1384	89.4%	94.3%	91.5%	0.830	499	50	111	446	81.8%	89.9%	85.4%	0.713
Adenosine A2a receptor	72	2855	94	102	1488	96.6%	94.1%	95.7%	0.905	935	44	51	457	94.8%	91.2%	93.6%	0.858
Alpha-2a adrenergic receptor	68	596	25	53	744	91.8%	96.7%	94.5%	0.890	157	12	39	233	80.1%	95.1%	88.4%	0.769
Androgen receptor	115	1499	85	490	5384	75.4%	98.4%	92.3%	0.798	436	42	212	1773	67.3%	97.7%	89.7%	0.723
Beta-1 adrenergic receptor	49	913	36	47	770	95.1%	95.5%	95.3%	0.905	258	22	43	254	85.7%	92.0%	88.7%	0.777
Beta-2 adrenergic receptor	135	1252	86	210	1415	85.6%	94.3%	90.0%	0.803	363	63	120	450	75.2%	87.7%	81.6%	0.635
Delta opioid receptor	41	2189	166	51	738	97.7%	81.6%	93.1%	0.828	735	59	31	256	96.0%	81.3%	91.7%	0.795
Dopamine D1 receptor	72	881	56	148	1428	85.6%	96.2%	91.9%	0.832	247	25	76	482	76.5%	95.1%	87.8%	0.743
Dopamine D2 receptor	68	4195	136	67	719	98.4%	84.1%	96.0%	0.854	1383	59	51	223	96.4%	79.1%	93.6%	0.764
Dopamine transporter	70	1745	91	125	1354	93.3%	93.7%	93.5%	0.868	568	35	73	437	88.6%	92.6%	90.3%	0.805
Endothelin receptor ET-A	24	937	50	25	809	97.4%	94.2%	95.9%	0.918	305	18	18	275	94.4%	93.9%	94.2%	0.883
Glucocorticoid receptor	123	1814	93	460	5150	79.8%	98.2%	92.6%	0.823	537	55	207	1675	72.2%	96.8%	89.4%	0.742
hERG	456	3099	229	518	2192	85.7%	90.5%	87.6%	0.751	878	145	400	680	68.7%	82.4%	74.1%	0.499
Histamine H1 receptor	66	908	29	55	793	94.3%	96.5%	95.3%	0.906	273	18	40	266	87.2%	93.7%	90.3%	0.808
Mu opioid receptor	56	2576	117	85	1606	96.8%	93.2%	95.4%	0.903	889	42	61	543	93.6%	92.8%	93.3%	0.859
Muscarinic acetylcholine receptor M1	77	1397	72	87	865	94.1%	92.3%	93.4%	0.862	479	42	52	263	90.2%	86.2%	88.8%	0.759
Muscarinic acetylcholine receptor M2	54	1155	79	70	1427	94.3%	94.8%	94.5%	0.890	374	35	37	492	91.0%	93.4%	92.3%	0.844
Muscarinic acetylcholine receptor M3	69	1136	65	54	774	95.5%	92.3%	94.1%	0.879	307	22	40	253	88.5%	92.0%	90.0%	0.801
Norepinephrine transporter	69	2101	94	122	1355	94.5%	93.5%	94.1%	0.877	625	32	64	460	90.7%	93.5%	91.9%	0.836
Serotonin 2a (5-HT2a) receptor	81	2737	59	60	712	97.9%	92.3%	96.7%	0.902	928	30	32	233	96.7%	88.6%	94.9%	0.850
Serotonin 3a (5-HT3a) receptor	28	316	11	32	767	90.8%	98.6%	96.2%	0.910	87	4	17	273	83.7%	98.6%	94.5%	0.859
Serotonin transporter	48	3004	105	53	747	98.3%	87.7%	96.0%	0.879	944	37	42	246	95.7%	86.9%	93.8%	0.822
Tyrosine-protein kinase LCK	55	1239	35	44	370	96.6%	91.4%	95.3%	0.873	412	21	37	98	91.8%	82.4%	89.8%	0.709
Vasopressin V1a receptor	15	446	21	10	766	97.8%	97.3%	97.5%	0.947	150	6	13	266	92.0%	97.8%	95.6%	0.907
Average	**87**	**1699**	**80**	**133**	**1407**	**92.6%**	**93.4%**	**94.1%**	**0.868**	**532**	**38**	**78**	**460**	**86.6%**	**90.9%**	**90.2%**	**0.782**

Table 2.3: The results of the "Risk assessment" model created by the automated workflow for construction of structural alert-based models. Parameters for the automated workflow are selected to give a model with higher specificity at the expense of sensitivity. The parameters used are: theta 0.95; 1% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert: two actives, one inactive

		Training set								Test set							
Target	**Alerts**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
Acetylcholinesterase	67	1895	433	109	1034	94.6%	70.5%	84.4%	0.684	555	157	55	339	91.0%	68.3%	80.8%	0.616
Adenosine A2a receptor	45	2889	159	68	1423	97.7%	89.9%	95.0%	0.889	950	58	36	443	96.3%	88.4%	93.7%	0.857
Alpha-2a adrenergic receptor	30	610	152	39	617	94.0%	80.2%	86.5%	0.742	172	50	24	195	87.8%	79.6%	83.2%	0.669
Androgen receptor	102	1506	119	483	5350	75.7%	97.8%	91.9%	0.788	442	46	206	1769	68.2%	97.5%	89.8%	0.725
Beta-1 adrenergic receptor	14	934	84	26	722	97.3%	89.6%	93.8%	0.876	277	33	24	243	92.0%	88.0%	90.1%	0.802
Beta-2 adrenergic receptor	98	1278	146	184	1355	87.4%	90.3%	88.9%	0.777	376	86	107	427	77.8%	83.2%	80.6%	0.612
Delta opioid receptor	24	2192	176	48	728	97.9%	80.5%	92.9%	0.823	745	69	21	246	97.3%	78.1%	91.7%	0.794
Dopamine D1 receptor	47	910	110	119	1374	88.4%	92.6%	90.9%	0.811	263	41	60	466	81.4%	91.9%	87.8%	0.742
Dopamine D2 receptor	30	4224	194	38	661	99.1%	77.3%	95.5%	0.830	1408	71	26	211	98.2%	74.8%	94.3%	0.784
Dopamine transporter	50	1775	248	95	1197	94.9%	82.8%	89.7%	0.791	581	81	60	391	90.6%	82.8%	87.3%	0.740
Endothelin receptor ET-A	10	948	76	14	783	98.5%	91.2%	95.1%	0.903	311	27	12	266	96.3%	90.8%	93.7%	0.874
Glucocorticoid receptor	92	1843	250	431	4993	81.0%	95.2%	90.9%	0.782	556	85	188	1645	74.7%	95.1%	89.0%	0.731
hERG	99	3469	1392	148	1029	95.9%	42.5%	74.5%	0.475	1170	511	108	314	91.5%	38.1%	70.6%	0.361
Histamine H1 receptor	35	926	69	37	753	96.2%	91.6%	94.1%	0.881	286	27	27	257	91.4%	90.5%	91.0%	0.819
Mu opioid receptor	38	2597	220	64	1503	97.6%	87.2%	93.5%	0.865	905	70	45	515	95.3%	88.0%	92.5%	0.840
Muscarinic acetylcholine receptor M1	27	1433	192	51	745	96.6%	79.5%	90.0%	0.789	505	69	26	236	95.1%	77.4%	88.6%	0.752
Muscarinic acetylcholine receptor M2	30	1165	137	60	1369	95.1%	90.9%	92.8%	0.856	385	51	26	476	93.7%	90.3%	91.8%	0.836
Muscarinic acetylcholine receptor M3	30	1149	131	41	708	96.6%	84.4%	91.5%	0.826	322	51	25	224	92.8%	81.5%	87.8%	0.753
Norepinephrine transporter	40	2125	159	98	1290	95.6%	89.0%	93.0%	0.853	643	47	46	445	93.3%	90.4%	92.1%	0.838
Serotonin 2a (5-HT2a) receptor	33	2761	165	36	606	98.7%	78.6%	94.4%	0.828	941	64	19	199	98.0%	75.7%	93.2%	0.791
Serotonin 3a (5-HT3a) receptor	11	326	58	22	720	93.7%	92.5%	92.9%	0.840	92	21	12	256	88.5%	92.4%	91.3%	0.789
Serotonin transporter	33	3022	226	35	626	98.9%	73.5%	93.3%	0.797	952	75	34	208	96.6%	73.5%	91.4%	0.742
Tyrosine-protein kinase LCK	25	1254	115	29	290	97.7%	71.6%	91.5%	0.756	425	35	24	84	94.7%	70.6%	89.6%	0.677
Vasopressin V1a receptor	9	447	77	9	710	98.0%	90.2%	93.1%	0.861	151	27	12	245	92.6%	90.1%	91.0%	0.814
Average	**42**	**1737**	**212**	**95**	**1274**	**94.5%**	**83.7%**	**91.2%**	**0.805**	**559**	**77**	**51**	**421**	**90.6%**	**82.4%**	**88.9%**	**0.748**

Table 2.4: The results of the "Screening" model created by the automated workflow for construction of structural alert-based models. Parameters for the automated workflow are selected to give a model with higher sensitivity at the expense of specificity. The parameters used are: theta 0.51; 15% maximum occurrence of a substructure in the inactive chemicals; lower bounds for a structural alert: two actives, one inactive.

Both models give very impressive performance metrics in the test sets. The risk assessment and screening models have a mean MCC in the test sets of 0.782 and 0.748 respectively, indicating an excellent match between model predictions and experimental activities. The screening model has a higher sensitivity than the risk assessment model, but lower specificity. This highlights how different sets of parameters for the same automated workflow can create models suited for different purposes. In this case, the parameters that are changed are the theta value in the Bayes Factor calculation and the maximum percentage of a substructure in the inactive chemicals.

Both models used lower bounds for a structural alert of two actives and one inactive. This was found to be a “sweet spot” in terms of a balance between sensitivity and specificity. A more stringent lower bound requirement (i.e. more actives or fewer inactives) would increase specificity but has a larger reduction in sensitivity. A less stringent lower bound requirement (i.e. more inactives tolerated) would increase sensitivity but has a larger reduction in specificity. Hence, an even larger sensitivity can be obtained than that presented in the “screening” parameters by using less a less stringent lower bounds, but that the larger decreases in specificity lead to decreases in MCC. Less stringent lower bounds could also be considered as over-training the model.

Compared to the screening model, the risk assessment model creates structural alerts which are larger in size and more specific, covering fewer chemicals per alert. A greater number of these structural alerts are required to cover the active chemicals, but fewer false positives are hit, resulting in higher specificity and similar sensitivity.

The interpretability and transparency of the active predictions is a key advantage to the structural alert models. If a new chemical is predicted to be active by one of the models, the user can see which structural alert(s) it contains, which training set chemicals contain that structural alert, and hence understand why the model has made an active prediction. Furthermore, the structures of the chemicals in the training set containing the alert can be seen and can be compared to the structure of the new chemical.

2.3. Four-fold cross validation

In all previous results hold-out validation has been used, with data being split randomly between a training set (roughly 75% of chemicals) and a test set (roughly 25% of chemicals). Models are trained on the training set and then applied to the test set to calculate performance statistics. The same training and test set has been used throughout this work. One might worry that the performance metrics only occur when using this split of data only. A different validation method has been trialled to see if similar performance metrics are obtained.

Four-fold cross validation involves splitting the data into four groups. Each group in turn acts as the test set, with the other three groups being combined to make a training set. This approach takes four times as long to train and test models, as four different models must be created, but it allows inconsistencies in the data set or in model construction to be spotted. Model performance varying significantly across the four groups in four-fold cross validation would suggest an inconsistent method and would decrease confidence in the constructed models.

2.3.1. Data and methods

The same data that was extracted from ChEMBL and ToxCast previously has been used here. Active and inactive chemicals were separated and each partitioned randomly into four groups, ensuring a similar balance of active and inactive chemicals in each group. Each group in turn was assigned as the test set, with the other three groups combined to make a training set.

For each training set combination, the automated workflow for structural alert-based model construction was applied with the "Risk Assessment" parameters used previously: 0.95 theta, 1% maximum occurrence of a structural alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. The models were applied to the relevant test set and performance statistics calculated.

2.3.2. Results and discussion

The workflow for construction of structural alert-based models has been applied to the data sets using four-fold cross validation. The mean and standard deviation of each performance metric in the training and test sets have been calculated and are shown in Table 2.5.

As with hold-out validation, the performance metrics are very high for almost all biological targets. A mean MCC of 0.775 across all targets with four-fold cross validation indicates very high performing models and agrees, within one standard deviation, with the value of 0.782 obtained when using hold-out validation.

The magnitude of the standard deviation is very small for all metrics for almost all targets. The exception is the human ether-a-go-go-related gene (hERG) for which larger standard deviations are seen, particularly in test and training specificity. All other metrics for other targets have very low standard deviations.

It can be concluded that the method for constructing models is consistent across targets. The hERG is the only exception, likely because its data is particularly difficult to model, as reflected in the low mean performance metrics. This is consistent with literature reports of the tendency of hERG to bind a molecules with a wide range of chemical structures.[91,92]

The very low standard deviations in performance metrics in four-fold cross validation suggest that there is little inconsistency in data across the four groups. For these targets and these models, hold-out validation does equally well at calculating performance statistics, but requires only a quarter of the time. Thus, only hold-out validation will be used for the rest of the work.

		Training Set								Test Set							
Target	Alerts	SE	σ_{SE}	SP	σ_{SP}	ACC	σ_{ACC}	MCC	σ_{MCC}	SE	σ_{SE}	SP	σ_{SP}	ACC	σ_{ACC}	MCC	σ_{MCC}
Acetylcholinesterase	151.8	0.887	0.005	0.943	0.004	0.911	0.004	0.824	0.008	0.802	0.016	0.905	0.010	0.846	0.013	0.700	0.026
Adenosine A2a receptor	64.5	0.968	0.003	0.935	0.002	0.957	0.001	0.905	0.002	0.946	0.007	0.916	0.016	0.936	0.005	0.859	0.012
Alpha-2a adrenergic receptor	63.0	0.914	0.008	0.971	0.003	0.945	0.003	0.891	0.005	0.769	0.018	0.943	0.016	0.864	0.008	0.730	0.017
Androgen receptor	114.5	0.755	0.007	0.985	0.001	0.923	0.002	0.799	0.007	0.669	0.004	0.976	0.004	0.894	0.004	0.718	0.011
Beta-1 adrenergic receptor	49.3	0.943	0.003	0.950	0.007	0.946	0.003	0.892	0.006	0.880	0.026	0.927	0.008	0.902	0.016	0.805	0.031
Beta-2 adrenergic receptor	141.8	0.858	0.007	0.946	0.002	0.903	0.003	0.808	0.005	0.711	0.027	0.886	0.008	0.800	0.018	0.607	0.034
Delta opioid receptor	57.5	0.973	0.002	0.867	0.023	0.943	0.005	0.859	0.014	0.943	0.016	0.839	0.063	0.913	0.008	0.789	0.025
Dopamine D1 receptor	73.8	0.857	0.004	0.969	0.005	0.923	0.004	0.841	0.008	0.743	0.029	0.938	0.005	0.859	0.011	0.707	0.022
Dopamine D2 receptor	92.0	0.982	0.001	0.858	0.006	0.961	0.001	0.858	0.005	0.962	0.007	0.829	0.028	0.940	0.008	0.785	0.028
Dopamine transporter	74.0	0.936	0.002	0.934	0.005	0.935	0.001	0.868	0.002	0.892	0.008	0.906	0.011	0.898	0.001	0.795	0.003
Endothelin receptor ET-A	23.0	0.967	0.006	0.944	0.004	0.956	0.004	0.911	0.008	0.942	0.014	0.936	0.018	0.939	0.002	0.878	0.003
Glucocorticoid receptor	126.5	0.801	0.008	0.984	0.001	0.928	0.002	0.828	0.005	0.726	0.014	0.969	0.003	0.896	0.005	0.747	0.014
hERG	338.0	0.884	0.024	0.733	0.119	0.824	0.033	0.631	0.078	0.772	0.054	0.672	0.103	0.732	0.011	0.447	0.042
Histamine H1 receptor	59.5	0.944	0.007	0.957	0.009	0.950	0.005	0.899	0.011	0.875	0.034	0.923	0.014	0.898	0.017	0.797	0.032
Mu opioid receptor	65.0	0.966	0.002	0.937	0.006	0.955	0.001	0.905	0.002	0.940	0.002	0.924	0.014	0.934	0.004	0.862	0.009
Muscarinic acetylcholine receptor M1	72.5	0.944	0.008	0.911	0.007	0.931	0.006	0.855	0.012	0.894	0.014	0.882	0.017	0.889	0.010	0.769	0.020
Muscarinic acetylcholine receptor M2	50.3	0.942	0.001	0.948	0.003	0.945	0.001	0.889	0.002	0.899	0.024	0.941	0.008	0.922	0.014	0.842	0.028
Muscarinic acetylcholine receptor M3	60.8	0.948	0.005	0.935	0.019	0.943	0.006	0.882	0.013	0.889	0.026	0.908	0.022	0.897	0.011	0.793	0.020
Norepinephrine transporter	70.5	0.943	0.003	0.941	0.005	0.943	0.002	0.881	0.004	0.909	0.013	0.919	0.014	0.913	0.005	0.822	0.008
Serotonin 2a (5-HT2a) receptor	75.8	0.983	0.001	0.930	0.009	0.972	0.003	0.916	0.008	0.955	0.009	0.896	0.024	0.942	0.005	0.833	0.014
Serotonin 3a (5-HT3a) receptor	27.0	0.900	0.011	0.990	0.005	0.963	0.002	0.911	0.004	0.799	0.032	0.974	0.010	0.922	0.014	0.810	0.036
Serotonin transporter	58.8	0.985	0.003	0.900	0.013	0.967	0.002	0.901	0.005	0.969	0.008	0.872	0.025	0.948	0.004	0.847	0.010
Tyrosine-protein kinase LCK	65.0	0.969	0.002	0.917	0.018	0.957	0.004	0.881	0.013	0.923	0.004	0.857	0.024	0.907	0.007	0.752	0.020
Vasopressin V1a receptor	16.3	0.970	0.008	0.975	0.003	0.973	0.001	0.943	0.002	0.942	0.030	0.965	0.010	0.957	0.011	0.907	0.024
Average	**82.9**	**0.926**	**0.006**	**0.932**	**0.012**	**0.940**	**0.004**	**0.866**	**0.010**	**0.865**	**0.018**	**0.904**	**0.020**	**0.898**	**0.009**	**0.775**	**0.020**

Table 2.5: The results of using the workflow for construction of structural alert-based models (parameters: 0.95 theta, 1% maximum occurrence of a structural alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive) with four-fold cross validation. Mean values and standard deviation (σ) for each performance metric are calculated in the training and test sets. Performance metrics are sensitivity (SE), specificity (SP), accuracy (ACC) and Matthews Correlation Coefficient (MCC).

2.4. Comparison to other models

2.4.1. Allen's models

Allen *et al* first constructed structural alert-based models for Bowes' targets in 2016,[43] and updated the models in 2018.[44] The updated models are similar to models developed in this work as both use an iterative cycle of creating structural alerts and removing chemicals from the training set. A major difference between the method presented here and Allen's approach is a consideration of substructures' occurrence in the inactive chemicals, not just in the active chemicals. The use of Bayesian statistics allows both variables to be considered, so false positives in the training set can be minimised whilst trying to build a model which correctly predicts true positives.

Allen has constructed models for the twenty-four biological targets for which models have been built in this work. However, the models have been constructed using different data sets so direct comparison is difficult.

Allen's model is built on ChEMBL data only. This is unbalanced data, with most biological targets having significantly more actives than inactives. In construction of Allen's models, the production of structural alerts from the training set does not consider occurrence of substructures in the inactive chemicals, so the imbalance in data is unimportant. However, an assumption is made to add inactive chemicals to the test set for the purpose of validation. For each target, chemicals which have been tested at other Bowes targets, but not the target of interest, are assumed to be inactive, giving approximately 11 000 additional negatives for each receptor. Such an assumption is likely to be valid for pharmaceutical chemicals because they are likely to have been tested at all targets during trials, but only active bioactivity data tends to be report in publications (from which ChEMBL collects data). However, the negative assumption will not be valid in all cases and there will be uncertainty about all assumed negatives. This uncertainty is not a major problem for Allen's methods, where assumed negatives are used only for validation. The new approach, however, uses both training set positives and negatives in selecting the best substructures to be structural alerts. Uncertainty in negatives would therefore have a direct effect on model construction and could be problematic. Hence, importance of constructing new data sets with no assumptions regarding inactive data.

As Allen's models and the new models are built and tested on different data sets, direct comparison is difficult. Both data sets take the majority of their active chemicals from ChEMBL, and the methods are both structural alert-based, so some comparisons should still be made. The average performances of the models in their respective test sets are shown in Table 2.6.

		Mean Test Set Performance Metrics							
Model	**Alerts**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
Allen Risk Assessment	53.0	250	153	109	15465	64.4%	98.7%	97.9%	0.646
Allen Screening	86.7	308	572	51	10698	80.9%	94.8%	94.6%	0.585
Wedlake Risk Assessment	86.5	532	38	78	460	86.6%	90.9%	90.2%	0.782
Wedlake Screening	42.5	559	77	51	421	90.6%	82.4%	88.9%	0.748

Table 2.6: Comparison of the performance metrics of Allen's updated models from 2018[44] and the models developed in this work, labelled "Wedlake" models. "Screening" models are designed to have as large a sensitivity (SE) as possible, whilst the "Risk Assessment" models are designed to have the best overall performance. Performance metrics are calculated across the twenty-four Bowes targets for which both methods have created models. Different data sets are used for the Wedlake and Allen models and so direct comparisons of performance are difficult. However, Matthews Correlation Coefficient (MCC) is commonly considered the best single measure of overall model performance and a clear increase can be seen in both Wedlake models compared to Allen models. Both Wedlake models have higher SE than both Allen models. Allen models have higher specificity (SP) and accuracy (ACC), but these values are inflated by the use of assumed negative data. Overall, the performance statistics suggest the Wedlake models are significant improvements on the Allen model.

Allen's models are tested using the assumed negatives, hence the large excess of negative data. For these unbalanced data sets, MCC provides the fairest indication of overall performance. The MCCs of both new models are significantly higher than the Allen model, indicating better overall performance of models.

The specificity of both Allen models is higher, but there is uncertainty with the negative predictions due to the inclusion of assumed negatives. The vast number of the assumed negatives dominates the accuracy, resulting in higher, potentially misleading accuracy values for the Allen models. Sensitivity of the new models is greater than that of the Allen models.

In the Allen models the Screening model differs from the Risk Assessment model by using less stringent minimum requirements for structural alerts, resulting in a greater number of alerts being used to achieve a larger sensitivity. In contrast, in the new models the Screening model uses fewer structural alerts than the Risk Assessment model to achieve a greater sensitivity. The same minimum requirement for structural alerts is used in both models. Instead, other parameters are changed, resulting in different substructures being chosen as structural alerts. The Screening model contains larger, less specific structural alerts which are contained by a greater number of actives.

Overall, the new models represent a significant improvement on the Allen models in terms of performance. The use of Bayesian statistics to pick structural alerts allows the consideration of the occurrence of a substructure in both the active chemicals and in the inactive chemicals. The new, balanced data sets constructed in this work mean that no assumptions about data need to be made.

2.4.2. Random Forest models

The previous methodology presented in this work is a novel way of constructing structural alert-based models. To compare this new approach to existing, accepted methods, classical machine learning approaches were investigated using physicochemical features and structural descriptors as inputs. These descriptors describe many different features of a chemical structure. Understanding how these many features, individually or as combinations, relate to activity at a biological target is very difficult. However, machine learning algorithms are capable of identifying such correlations if and where they exist. Random Forest, Neural Network, Support Vector Machine and k-Nearest Neighbour were all initially explored, but Random Forest models were chosen to be prioritised over these other approaches due to their greater interpretability. Interpretability of predictions is particularly important in the context of toxicity and risk assessment. This work was done by Maria Folia, working at the Safety and Environmental Assurance Centre, Unilever.

To provide a direct comparison to the models constructed by the automated workflow, the Random Forest models were built using the same ChEMBL and ToxCast training set and performance statistics calculated using the same test set.

2.4.2.1. Method

RDKit[83] was used to create 200 physiochemical descriptors for each chemical including molecular, topological, van der Waals surface area and lipophilicity descriptors. The model was built using the RandomForestClassifier from the sklearn[93] package using the default settings apart from two hyperparameters: the number of trees and the maximum depth of the trees, which were tuned using GridSearchCV.

The Random Forest models have been applied to the test set and performance statistics were calculated.

2.4.2.2. Results and Discussion

Models have been constructed for the same set of twenty-four Bowes targets as the structural alert-based models. The same training and test sets have been used to allow for direct comparison between methods. The performance of the Random Forest models in the test sets is shown in Table 2.7.

	Test Set							
Target	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
Acetylcholinesterase	559	110	51	386	91.6%	77.8%	85.4%	0.707
Adenosine A2a receptor	967	55	19	446	98.1%	89.0%	95.0%	0.888
Alpha-2a adrenergic receptor	175	12	21	233	89.3%	95.1%	92.5%	0.849
Androgen receptor	424	39	224	1776	65.4%	97.9%	89.3%	0.713
Beta-1 adrenergic receptor	278	33	23	243	92.4%	88.0%	90.3%	0.806
Beta-2 adrenergic receptor	368	59	115	454	76.2%	88.5%	82.5%	0.653
Delta opioid receptor	753	74	13	241	98.3%	76.5%	92.0%	0.802
Dopamine D1 receptor	262	39	61	468	81.1%	92.3%	88.0%	0.745
Dopamine D2 receptor	1424	65	10	217	99.3%	77.0%	95.6%	0.834
Dopamine transporter	596	49	45	423	93.0%	89.6%	91.6%	0.827
Endothelin receptor ET-A	314	27	9	266	97.2%	90.8%	94.2%	0.884
Glucocorticoid receptor	549	64	195	1666	73.8%	96.3%	89.5%	0.745
hERG	1188	384	90	441	93.0%	53.5%	77.5%	0.522
Histamine H1 receptor	300	35	13	249	95.8%	87.7%	92.0%	0.841
Mu opioid receptor	911	51	39	534	95.9%	91.3%	94.1%	0.875
Muscarinic acetylcholine receptor M1	508	53	23	252	95.7%	82.6%	90.9%	0.802
Muscarinic acetylcholine receptor M2	386	46	25	481	93.9%	91.3%	92.4%	0.848
Muscarinic acetylcholine receptor M3	330	33	17	242	95.1%	88.0%	92.0%	0.837
Norepinephrine transporter	655	42	34	450	95.1%	91.5%	93.6%	0.867
Serotonin 2a (5-HT2a) receptor	955	47	5	216	99.5%	82.1%	95.7%	0.871
Serotonin 3a (5-HT3a) receptor	94	6	10	271	90.4%	97.8%	95.8%	0.893
Serotonin transporter	973	49	13	234	98.7%	82.7%	95.1%	0.855
Tyrosine-protein kinase LCK	428	23	21	96	95.3%	80.7%	92.3%	0.765
Vasopressin V1a receptor	152	15	11	257	93.3%	94.5%	94.0%	0.873
Average	**565**	**59**	**45**	**439**	**91.6%**	**86.8%**	**91.3%**	**0.804**

Table 2.7: Performance of Random Forest models on the test set. The models have been constructed from the same training set and tested on the same test set as the structural alert-based models built with the automated workflow.

Within the test sets, the performance metrics of the models created by the automated workflow are very similar to those of the Random Forest models, with the Random Forest models performing slightly better overall according to mean MCC (0.804 compared to 0.782 for the risk assessment model) and accuracy (91.3% compared to 90.2% for the risk assessment model).

The different methods generally give similar performance for each receptor. Notably, all methods create models which do not perform well on the hERG data, despite the large size of the data set (the training set contains 3 617 actives and 2 421 inactives). The poor performance for the human hERG being seen in both models supports the idea that it is not due to shortcomings in the structural alert method or the Random Forest method, but rather due to some inherent difficulty in modelling the data itself.

Model	Test SE	Test SP	Test ACC	Test MCC
Risk Assessment	86.6%	90.9%	90.2%	0.782
Screening	90.6%	82.4%	88.9%	0.748
Random Forest	91.6%	86.8%	91.3%	0.804

Table 2.8: Direct comparison of the average performance statistics of the models created by the automated workflow and the Random Forest models. Averages are taken over the same selection of 24 Bowes targets.

The similarity in performance between the new structural alert-based method and the Random Forest method provides credibility to the new method. The Random Forest models slightly outperform the structural alert models, but the difference between performance statistics is minimal. With similar performance metrics, the key difference between the approaches is the interpretably of predictions.

Random Forest models can identify which physicochemical features are most important in making activity predictions for each target. However, in terms of understanding why an activity prediction is made for a chemical, Random Forest models are difficult to interpret, even if they are considered more interpretable than other “black boxes” machine learning classifiers.

The transparency in the structural alert models results in predictions that are easy to interpret. The user can see both the activity prediction and why the activity prediction has been made. Understanding why activity predictions are made is particularly important in toxicity testing. While the Random Forest model may perform slightly better in terms of average performance metrics, the major advantage of the structural alert-based methods is the greater interpretability of model predictions.

Whilst it is good to make comparison between models, it is important to remember that the models are not in competition with each other. Rather, they can be used together to make predictions of biological activity with greater confidence.

2.5. Consensus approach

The ICH M7 guideline[53] for predicting potential mutagenicity of impurities acknowledges the potential for *in silico* predictions to replace *in vitro* studies. It requires two complementary (Q)SAR models to be applied together – an expert rule-based method and a statistical-based method. This is a significant landmark for the use of *in silico* (Q)SARs in risk assessment. It also shows the importance of using complementary models together to increase confidence in predictions.

Here, a consensus approach has been developed, using the models created by the automated workflow for construction of structural alert-based models together with the Random Forest models. Where the structure-based structural alert model and the physiochemical feature-based Random Forest model both agree on an activity prediction, one would have more confidence in the prediction.

2.5.1. Method

For the selection of twenty-four Bowes targets, the Random Forest model and a structural alert-based model created by the automated workflow (parameters: theta 0.95, 5% maximum occurrence in inactives, lower bounds of two actives and one inactive) have been combined in a consensus model. Predictions where the models agree are kept but where the models disagree, the chemicals were removed, and the predictions were labelled "inconclusive".

The consensus model has been applied to the test sets of the Bowes targets and performance metrics calculated.

2.5.2. Results and discussion

The results for applying the consensus model to the Bowes targets are shown in Table 2.9. Averaging over the targets, there is a 92.4% agreement in predictions between the Random Forest and structural alert models, showing high concurrence between the models. Averaged across all targets, all performance metrics increase in comparison to both models individually. Average MCC increases by 0.056 compared to the Random Forest model alone and by 0.082 compared to the structural alerts alone.

In particular, hERG – which both methods created relatively poor models for – shows very large changes in MCC, increasing by 0.195 compared to Random Forest and by 0.214 compared to structural alerts. Whilst 30% of chemicals returned inconclusive predictions, the 70% for which there was consensus in predictions showed vast improvements in performance.

In this work, two complementary models have been combined to make predictions for receptor binding MIEs. One model is based on statistically-derived structural alerts and the other is based on statistical links between physicochemical properties. The two models derive predictions from different chemical descriptors using different algorithms, and therefore can be considered orthogonal approaches.

Using the two models together improves overall performance and increases confidence in the predictions. It also shows how the structural alert-based model may be used in approaches like that required by the ICH M7 guideline, increasing the relevance of this work.

Target	Agreement	TP	FP	FN	TN	Inc.	SE	Δ vs SA	Δ vs RF	SP	Δ vs SA	Δ vs RF	ACC	Δ vs SA	Δ vs RF	MCC	Δ vs SA	Δ vs RF
Acetylcholinesterase	84.9%	510	63	38	328	167	93.1%	7.3%	1.4%	83.9%	8.3%	6.1%	89.2%	8.1%	3.8%	0.778	0.159	0.071
Adenosine A2a receptor	95.3%	934	37	12	434	70	98.7%	3.3%	0.7%	92.1%	1.9%	3.1%	96.5%	2.9%	1.5%	0.922	0.064	0.034
Alpha-2a adrenergic receptor	91.6%	154	6	18	226	37	89.5%	9.4%	0.2%	97.4%	2.7%	2.3%	94.1%	5.9%	1.5%	0.879	0.115	0.031
Androgen receptor	95.2%	401	11	189	1745	117	68.0%	0.7%	2.5%	99.4%	1.7%	1.5%	91.5%	1.8%	2.2%	0.768	0.045	0.055
Beta-1 adrenergic receptor	95.8%	270	24	20	239	24	93.1%	2.4%	0.7%	90.9%	1.0%	2.8%	92.0%	1.7%	1.7%	0.840	0.035	0.035
Beta-2 adrenergic receptor	88.5%	334	36	91	420	115	78.6%	4.5%	2.4%	92.1%	5.8%	3.6%	85.6%	5.2%	3.1%	0.716	0.105	0.062
Delta opioid receptor	93.6%	728	44	8	232	69	98.9%	3.2%	0.6%	84.1%	0.9%	7.6%	94.9%	2.8%	2.9%	0.869	0.064	0.067
Dopamine D1 receptor	92.3%	241	14	55	456	64	81.4%	4.9%	0.3%	97.0%	2.1%	4.7%	91.0%	3.3%	3.0%	0.810	0.070	0.066
Dopamine D2 receptor	95.6%	1379	48	7	206	76	99.5%	3.1%	0.2%	81.1%	2.0%	4.2%	96.6%	3.1%	1.0%	0.867	0.105	0.033
Dopamine transporter	90.0%	553	20	33	396	111	94.4%	6.2%	1.4%	95.2%	5.1%	5.6%	94.7%	5.8%	3.2%	0.892	0.115	0.065
Endothelin receptor ET-A	94.3%	303	14	8	256	35	97.4%	3.3%	0.2%	94.8%	3.0%	4.0%	96.2%	3.2%	2.1%	0.924	0.064	0.040
Glucocorticoid receptor	93.5%	507	18	163	1625	161	75.7%	3.2%	1.9%	98.9%	2.3%	2.6%	92.2%	2.8%	2.6%	0.808	0.068	0.063
hERG	70.1%	859	126	67	422	629	92.8%	23.8%	-0.2%	77.0%	-5.4%	23.6%	86.9%	12.6%	9.4%	0.716	0.214	0.195
Histamine H1 receptor	94.0%	283	21	12	245	36	95.9%	5.2%	0.1%	92.1%	0.9%	4.4%	94.1%	3.2%	2.2%	0.882	0.063	0.042
Mu opioid receptor	95.4%	882	29	32	521	71	96.5%	2.9%	0.6%	94.7%	1.9%	3.4%	95.8%	2.5%	1.7%	0.911	0.052	0.036
Muscarinic acetylcholine receptor M1	92.0%	476	32	18	243	67	96.4%	5.8%	0.7%	88.4%	1.8%	5.7%	93.5%	4.4%	2.6%	0.858	0.091	0.055
Muscarinic acetylcholine receptor M2	93.5%	368	21	17	471	61	95.6%	4.1%	1.7%	95.7%	1.6%	4.5%	95.7%	2.7%	3.2%	0.912	0.055	0.064
Muscarinic acetylcholine receptor M3	94.4%	316	18	16	237	35	95.2%	3.8%	0.1%	92.9%	1.3%	4.9%	94.2%	2.7%	2.2%	0.882	0.054	0.045
Norepinephrine transporter	93.1%	619	20	29	432	81	95.5%	5.0%	0.5%	95.6%	3.3%	4.1%	95.5%	4.3%	2.0%	0.908	0.086	0.041
Serotonin 2a (5-HT2a) receptor	95.6%	931	26	4	208	54	99.6%	2.5%	0.1%	88.9%	1.8%	6.8%	97.4%	2.5%	1.7%	0.919	0.070	0.047
Serotonin 3a (5-HT3a) receptor	96.1%	90	3	9	264	15	90.9%	3.4%	0.5%	98.9%	2.5%	1.0%	96.7%	2.8%	0.9%	0.916	0.069	0.023
Serotonin transporter	94.6%	940	25	8	227	69	99.2%	3.3%	0.5%	90.1%	1.4%	7.4%	97.3%	3.0%	2.1%	0.916	0.080	0.060
Tyrosine-protein kinase LCK	94.7%	412	17	17	92	30	96.0%	3.4%	0.7%	84.4%	2.1%	3.7%	93.7%	3.2%	1.4%	0.804	0.080	0.040
Vasopressin V1a receptor	93.8%	147	1	8	252	27	94.8%	2.8%	1.6%	99.6%	1.8%	5.1%	97.8%	2.2%	3.8%	0.953	0.047	0.080
Average	**92.4%**	**527**	**28**	**37**	**424**	**93**	**92.4%**	**4.9%**	**0.8%**	**91.9%**	**2.2%**	**5.1%**	**93.9%**	**3.9%**	**2.6%**	**0.860**	**0.082**	**0.056**

Table 2.9: Performance of the consensus model on the test sets. “Agreement” is the proportion of chemicals for which the structural alert-based model and the Random Forest model give the same prediction. Where the models give different predictions for a chemical, it is labelled “inconclusive” (Inc.). The structural alert-based model created by the automated workflow with parameters: theta 0.95, 5% maximum occurrence in inactives, lower bounds of two actives and one inactive

2.6. Applicability of active predictions

One of the five key principles from OECD's guidelines for (Q)SAR model construction for regulatory purposes is a defined domain of applicability.[27] Abiding by these principles will help the structural alert models gain acceptance for use in risk assessment.

The applicability of an active prediction of a new chemical containing a structural alert can be assessed by looking at the structures of the training chemicals which contain the alert and comparing to the structure of the new chemical. The structural alert was built from these training chemicals and so the active prediction is derived from these chemicals. From these chemicals, an applicability domain can be defined.

An applicability domain could be defined by using the range of values of a selection of common physicochemical features, such as molecular weight and lipophilicity (i.e. log P). But how do these features relate to receptor binding mechanisms? They are not easily interpretable quantities in this context and hence are not good for assessing applicability of structural alert predictions.

Rather, a better way of judging applicability is to directly compare the structure of the new chemical and the structures of the training chemicals containing the alert. If an expert considers the new chemical to be similar to these training chemicals, the prediction is applicable. This is difficult to define because similarity is inherently subjective, but Tanimoto similarity based on Morgan fingerprints can be used to guide the user.

In this section, using similarity between a new chemical and training chemicals containing the same alert to define applicability domains has been investigated. Similarity has been quantified using Tanimoto similarity based on Morgan fingerprints.

2.6.1. Method

The same data sets as outlined previously for the 24 Bowes targets with ChEMBL and ToxCast data have been used here.

Structural alerts have been generated from the training sets using the automated workflow as outlined previously. The parameters used were: theta 0.95, 5% maximum occurrence of an alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive.

The structural alerts have been applied to the test sets, predicting test chemicals to be active if they contain an alert. When a test chemical was found to contain a structural alert, Tanimoto similarity based on Morgan fingerprints (radius two atoms and string length 4 096 bits) to training active chemicals containing the same alert was calculated. How this correlated with accuracy of active predictions has been investigated.

The proposed process for judging applicability of active predictions is shown in Figure 2.8.

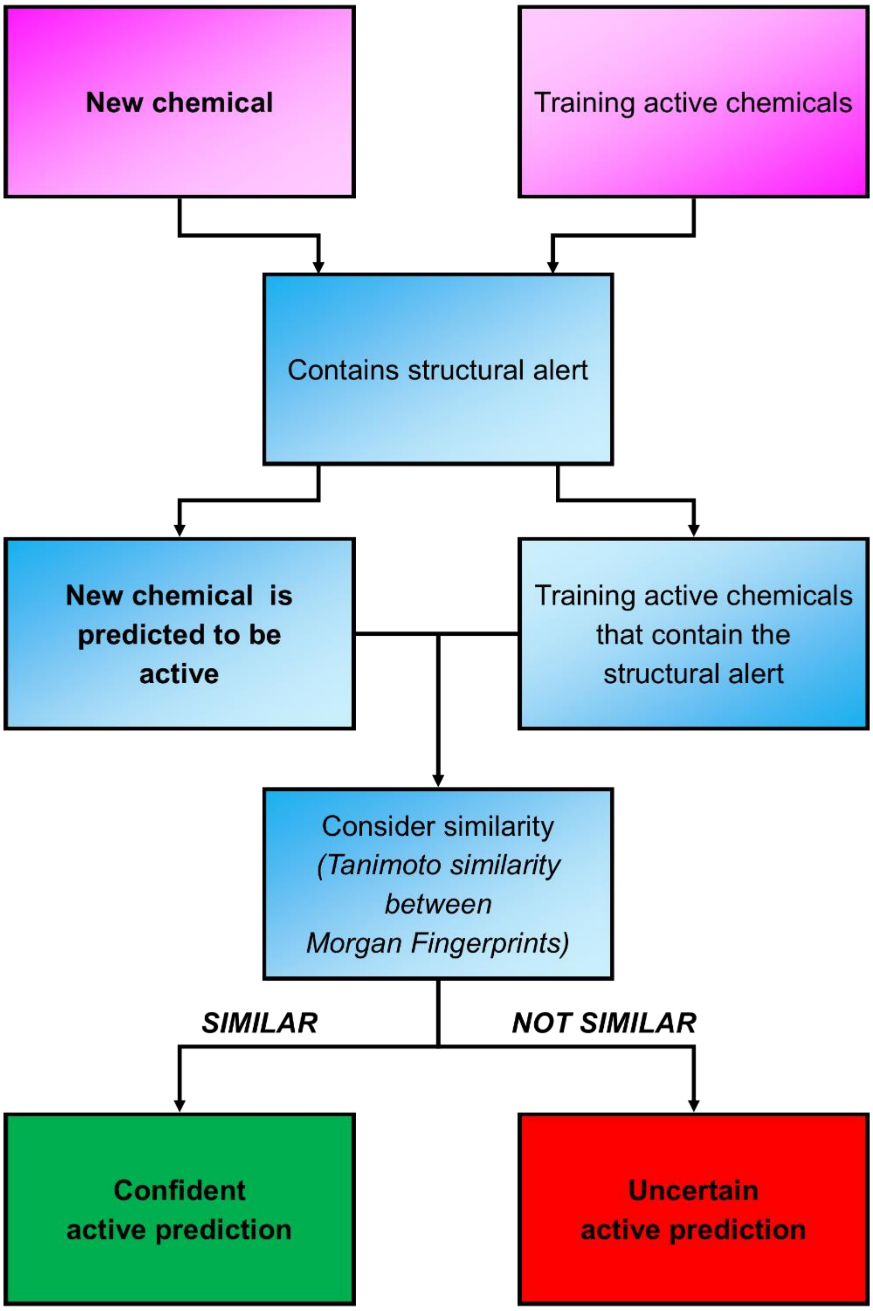

New chemical
Training active chemicals
Contains structural alert
New chemical is predicted to be active
Training active chemicals that contain the structural alert
Consider similarity (Tanimoto similarity between Morgan Fingerprints)
SIMILAR
NOT SIMILAR
Confident active prediction
Uncertain active prediction

Figure 2.8: The process for judging applicability of an active prediction. Where a test chemical contains a structural alert, the training active chemicals which contain the same alert are found. Similarity between the test chemical and these training active chemicals is used to judge confidence in the active prediction. In this figure, pink boxes are input chemicals, blue boxes are key steps in the process, green boxes represent high confidence predictions and red represents low confidence predictions.

2.6.2. Results and discussion

The Tanimoto similarity coefficients (based on Morgan fingerprints) between a test chemical containing an alert and the active chemicals containing the same alert were calculated. The largest values were kept, and chemicals put in groups according to this value. The proportion of false positive predictions in each group was calculated and the mean across all test sets is shown in Figure 2.9.

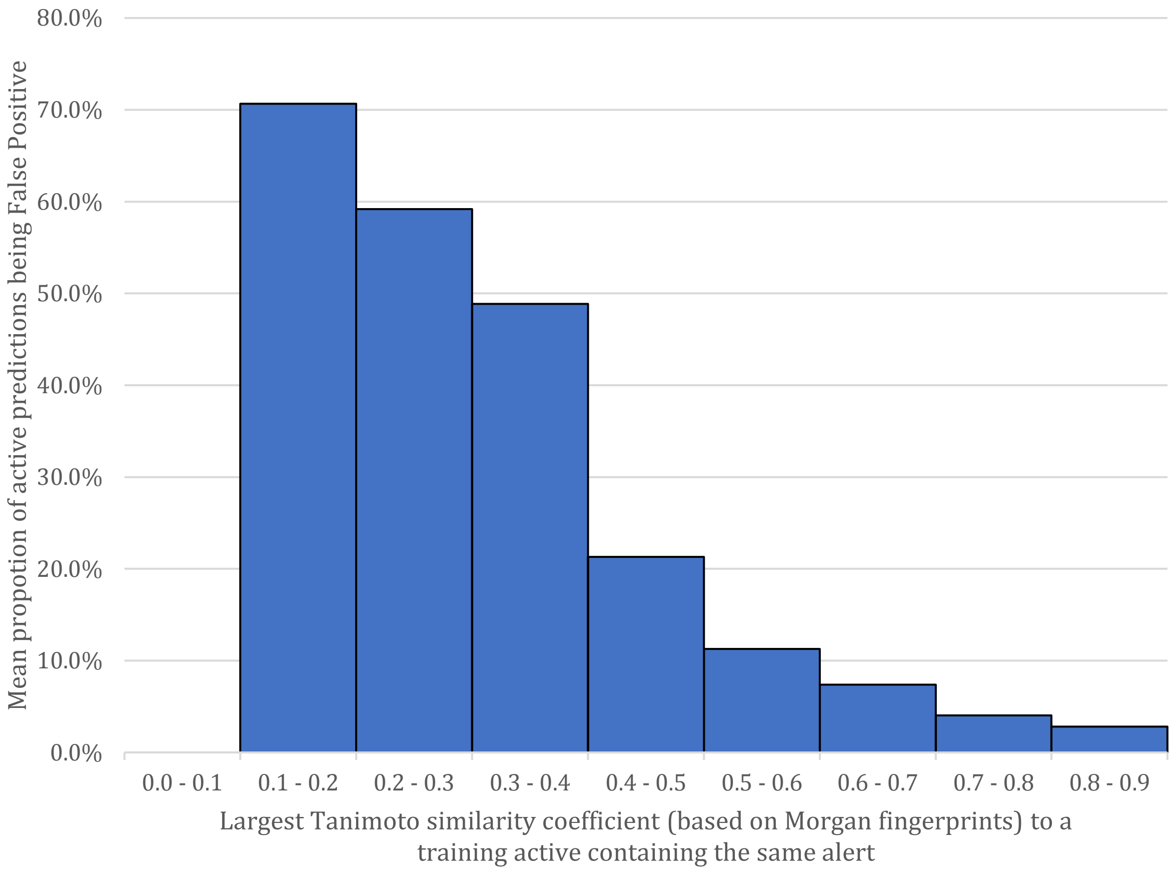


Figure 2.9: The variation of the mean proportion of false positives in active predictions (from structural alerts) of test set chemicals across groups defined by the maximum Tanimoto similarity (based on Morgan fingerprints) to training active chemicals containing the same alert as the test chemical. Structural alerts are created from training sets by the automated workflow with parameters: theta 0.95, 5% maximum occurrence of an alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. The data shown is the mean across the data sets of the twenty-four Bowes targets with human data in ToxCast and ChEMBL. No chemicals had a maximum similarity of less than 0.1 to a training active containing the same alert - having the same structural alert substructure results in a similarity of at least 0.1 in all cases here.

Test chemicals with low similarity values are very dissimilar to the active chemicals from which the structural alert was derived. Thus, the structural alert may not be applicable to these test chemicals. This is reflected in the results, as test chemicals with low similarity values contain a large proportion of false positives. As Tanimoto similarity (based on Morgan fingerprints) to the training chemicals containing the alert increases, the proportion of false positives decreases. This clearly demonstrates that applicability of a structural alert for a test chemical can be well characterised by considering similarity to the training chemicals containing the same structural alert. Generally, the greater the maximum Tanimoto similarity coefficient to the training active chemicals containing the same structural alert, the greater the confidence in the active prediction.

Ideally, applicability would be determined by an expert directly looking at the similarity of the structures of the test chemical and the training chemicals containing the alert. Tanimoto similarity coefficients (based from Morgan fingerprints) will help guide the expert in this decision. Examples of doing this are shown are in Figures 2.10a–e. In these examples, an alert-containing test chemical and the training active alert-containing chemical with the largest Tanimoto similarity (based on Morgan fingerprints) are compared to each other to assess the applicability of the structural alert to the test chemical.

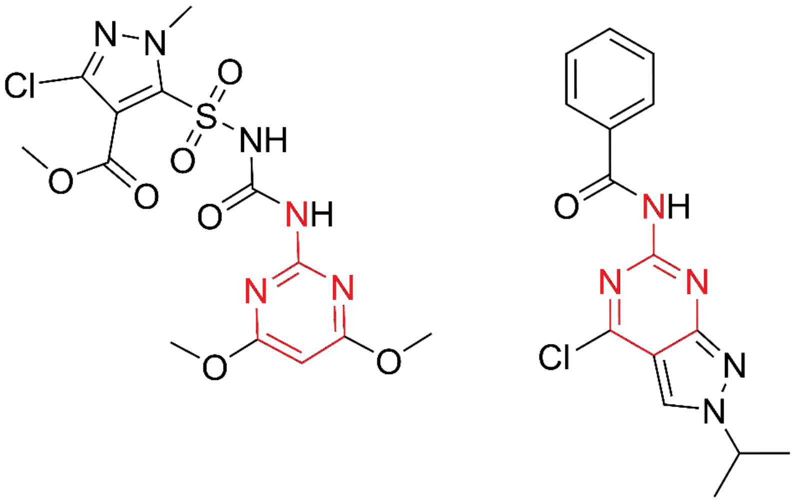


Figure 2.10a: Dissimilar chemicals. Left: test chemical; right: the most similar training active to the test chemical according to Tanimoto similarity (based on Morgan fingerprints) – the coefficient is 0.223. The structural alert, highlighted in red, is contained by 1 103 training actives and 44 inactives. In this case, the two chemicals contain the same structural alert but otherwise differ greatly. The test chemical bears little resemblance to any of the training active chemicals that lead to identification of the structural alert, so the active prediction should not be considered applicable to this test chemical. The test chemical is found to be inactive.

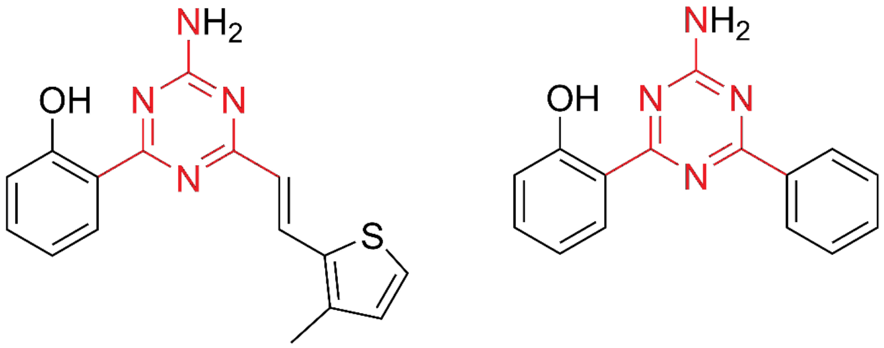


Figure 2.10b: Somewhat similar chemicals. Left: test chemical; right: the most similar training active to the test chemical according to Tanimoto similarity (based on Morgan fingerprints) – the coefficient is 0.468. The structural alert, highlighted in red, is contained by eight training actives and no inactives. The two chemicals contain the same structural alert as a central structure, and the same group on the left side of the structure. On the right side of the structure, the test chemical has a different side-group from the training active. Both side groups are similar in that they end in an aromatic ring, although the ring in the test chemical is held slightly further from the central structure by an inflexible alkene. Overall, the chemicals are similar enough that the prediction should be considered applicable to the test chemical. It is found to be active.

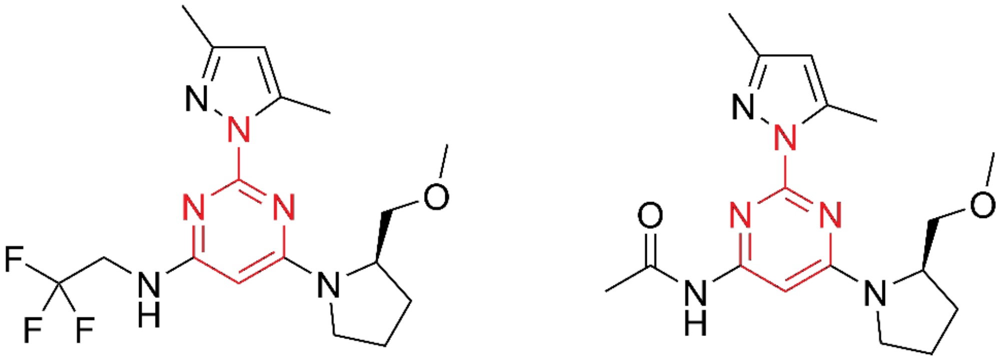


Figure 2.10c: Highly similar chemicals. Left: test chemical; right: the most similar training active to the test chemical according to Tanimoto similarity (based on Morgan fingerprints) – the coefficient is 0.703. The structural alert, highlighted in red, is contained by 1 103 training actives and 44 inactives. The two chemicals differ only by the group in the bottom left of the structures, with the training chemical containing an amide and the test chemical containing an amine with an unreactive trifluoro ethyl- group. The amine in the test chemical and the amide in the training chemical will have different effects on the electronics of the aromatic ring. Otherwise, the two chemicals are highly similar and so the active prediction should be considered applicable to the test chemical. The test chemical is found to be active.

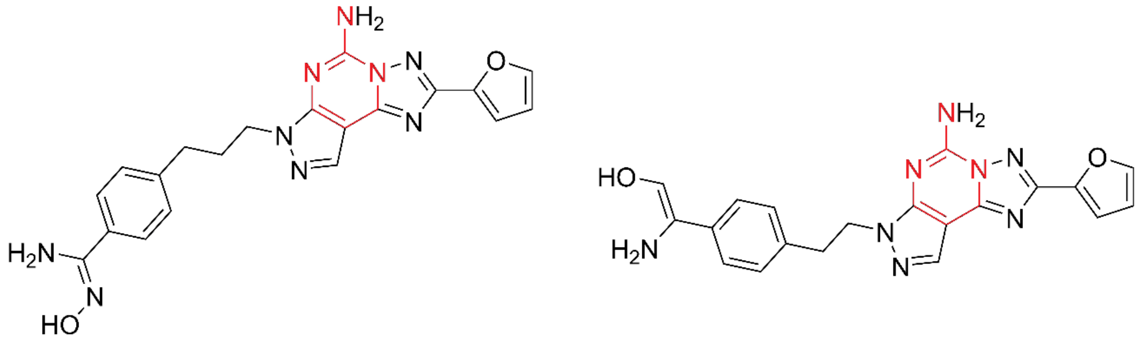


Figure 2.10d: Very highly similar chemicals. Left: test chemical; right: the most similar training active to the test chemical according to Tanimoto similarity (based on Morgan fingerprints) – the coefficient is 0.903. The structural alert, highlighted in red, is contained by 1 103 training actives and 44 inactives. The two chemicals are very highly similar, differing only by the length of an alkane linker between identical groups. The active prediction can be confidently assigned as applicable. The test chemical is found to be active.

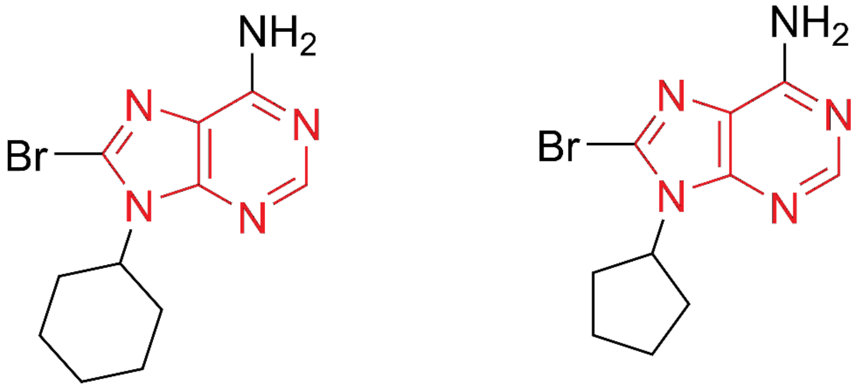


Figure 2.10e: Very highly similar chemicals, but in this example the test chemical is inactive. Left: test chemical; right: the most similar training active to the test chemical according to Tanimoto similarity (based on Morgan fingerprints) – the coefficient is 0.971. The structural alert, highlighted in red, is contained by 617 training actives and 52 inactives. The two chemicals are very highly similar, differing only by the size of a carbon ring. The active prediction can be confidently assigned as applicable, but the test chemical is found to be active. This shows that the applicability process is not perfect, even with very highly similar chemicals. Activity cliffs (defined as very similar chemicals with vastly different activities) such as this are very difficult to predict. The test chemical has a K_I *of 15 000 nM so, whilst following outside of the activity cut-off of 10 000 nM, it is still weakly active.*

To give the structural alert-based models a clearly defined applicability domain, as required by OECD guidelines for (Q)SARs, a cut-off can be included for the minimum Tanimoto similarity coefficient (based on Morgan fingerprints) between a test chemical containing an alert and the active chemicals containing the same alert. Chemicals which contain an alert but fall below the cut-off are considered "out of domain" active predictions.

Different cut-offs have been trialled for the minimum required Tanimoto similarity coefficient (based on Morgan fingerprints) between a test chemical containing an alert and the active chemicals containing the same alert.

The results are shown in Figure 2.11. They show a smooth relationship and so it is difficult to pick a single cut-off. Also considering the results shown in Figure 2.9, a sensible choice for a cut-off seems to be a minimum required Tanimoto similarity coefficient (based on Morgan fingerprints) of 0.4 between the test chemical containing an alert and the active chemicals containing the same alert. At this cut-off, a mean of 6.2% of active predictions in the test sets are considered "out of domain" and 57.8% of them are false positives.

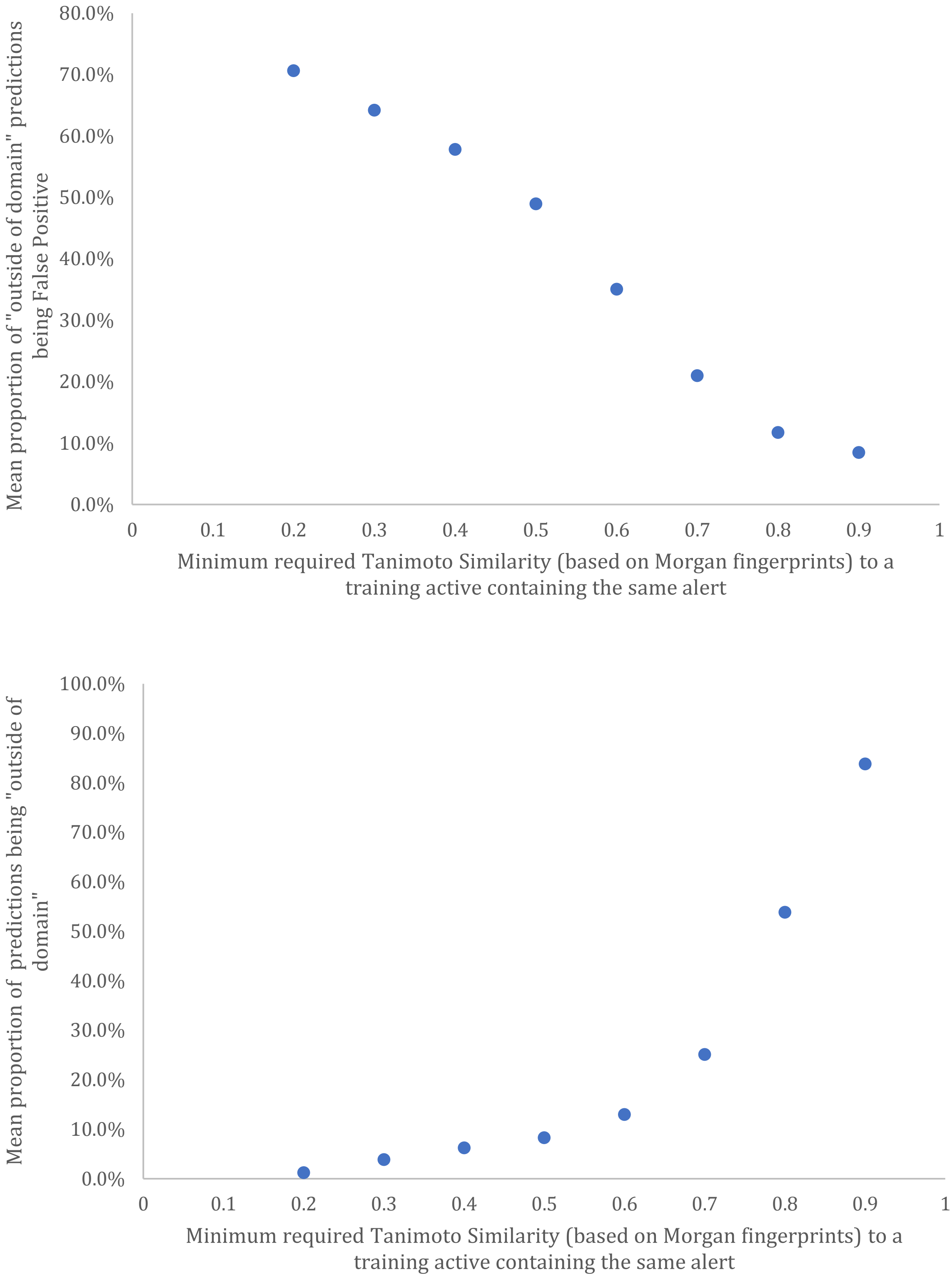


Figure 2.11: The effects of changing the minimum required Tanimoto similarity (based on Morgan fingerprints (radius two atoms, string length 4 096 bits)) between a test chemical containing a structural alert to a training active containing the same alert. The top figure shows proportion of active predictions which do not meet the requirement and the percentage of these predictions which are false positive is shown in the bottom figure.

The effects of applying a minimum required Tanimoto similarity coefficient (based on Morgan fingerprints) of 0.4 between the test chemical containing an alert and the active chemicals containing the same alert is shown in Table 2.10. A small but significant overall increase in PPV is observed for all biological targets when applying this cut-off. The average PPV of the "out of domain" predictions is very low, with almost 60% of the predictions being false positives. The overall increase in PPV is small as only a small proportion of active predictions in the test set are out of domain. Regardless, establishing this measure of confidence is vitally important for assessing the applicability of active predictions of chemicals outside of these data sets. It will help the methods gain acceptance for use in risk assessment.

The structural alerts developed here are substructures which have statistically been found to be associated with activity, but they are only fragments of chemicals. Tanimoto similarity based on Morgan fingerprints provide a measure of how similar the full structure of the test chemical is to the training active chemicals which contain the same structural alert. Where test chemicals have high similarity to training active chemicals containing the same alert, one can have more confidence in the active prediction.

	Out of domain			In domain					
Target	**TP**	**FP**	**PPV**	**Applicable TP**	**Applicable FP**	**PPV**	**No Domains PPV**	**ΔPPV**	**% outside of domain**
Acetylcholinesterase	4	9	0.308	513	52	0.908	0.894	0.014	2.2%
Adenosine A2a receptor	9	12	0.429	932	37	0.962	0.951	0.011	2.1%
Alpha-2a adrenergic receptor	15	7	0.682	142	6	0.959	0.924	0.036	12.9%
Androgen receptor	10	20	0.333	385	20	0.951	0.908	0.043	6.9%
Beta-1 adrenergic receptor	4	3	0.571	269	25	0.915	0.907	0.008	2.3%
Beta-2 adrenergic receptor	21	32	0.396	337	38	0.899	0.836	0.062	12.4%
Delta opioid receptor	9	17	0.346	724	36	0.953	0.933	0.020	3.3%
Dopamine D1 receptor	18	18	0.500	229	8	0.966	0.905	0.061	13.2%
Dopamine D2 receptor	14	29	0.326	1368	30	0.979	0.959	0.019	3.0%
Dopamine transporter	21	25	0.457	544	22	0.961	0.923	0.038	7.5%
Endothelin receptor ET-A	0	9	0.000	304	15	0.953	0.927	0.026	2.7%
Glucocorticoid receptor	8	24	0.250	506	24	0.955	0.915	0.040	5.7%
hERG	51	27	0.654	831	118	0.876	0.859	0.017	7.6%
Histamine H1 receptor	14	10	0.583	270	15	0.947	0.919	0.028	7.8%
Mu opioid receptor	11	21	0.344	878	21	0.977	0.955	0.022	3.4%
Muscarinic acetylcholine receptor M1	20	25	0.444	461	16	0.966	0.921	0.045	8.6%
Muscarinic acetylcholine receptor M2	14	13	0.519	362	18	0.953	0.924	0.029	6.6%
Muscarinic acetylcholine receptor M3	10	10	0.500	307	13	0.959	0.932	0.027	5.9%
Norepinephrine transporter	21	27	0.438	603	11	0.982	0.943	0.039	7.3%
Serotonin 2a (5-HT2a) receptor	17	26	0.395	915	8	0.991	0.965	0.027	4.5%
Serotonin 3a (5-HT3a) receptor	1	5	0.167	90	5	0.947	0.901	0.046	5.9%
Serotonin transporter	14	22	0.389	931	10	0.989	0.967	0.022	3.7%
Tyrosine-protein kinase LCK	34	8	0.810	382	13	0.967	0.952	0.015	9.6%
Vasopressin V1a receptor	2	5	0.286	148	1	0.993	0.962	0.032	4.5%
Average	**14.3**	**16.8**	**0.422**	**518.0**	**23.4**	**0.955**	**0.924**	**0.030**	**6.2%**

Table 2.10: The effect of applying an applicability cut-off to active predictions. Test chemicals which contain a structural alert are required to have a Tanimoto similarity (based on Morgan fingerprints) of at least 0.4 to a training active containing the same alert to be considered "in domain". TP = True Positive; FP = False Positive; PPV = Positive predictive value (proportion of positive predictions which are TP).

2.6.3. Relative similarity in judging applicability

Previously, applicability domains have been constructed by calculating Tanimoto similarity between a new chemical and training active chemicals containing the same structural alert as the new chemical. A single cut-off of 0.4 has been applied for all structural alerts from all biological targets. This value was found by tuning different cut-off values in the test sets. However, one may also want to consider the similarity of the alert-containing training chemicals to one another when deciding if an active prediction for a test chemical should be considered applicable. For example, if for one structural alert, the alert-containing training actives are highly similar to each other, a cut-off of 0.4 may be considered too low.

Here, a selection of structural alert examples were investigated to see if choosing different cut-offs for different alerts was sensible. Each structural alert was applied to the training and test sets to find chemicals containing the alert. These alert-containing chemicals were split into four groups: training actives, training inactives, test actives, and test inactives. For each chemical, the largest Tanimoto similarity to the alert-containing training active chemicals was calculated. If the chemical was itself an alert-containing training active chemical, maximum similarity to all other alert-containing training active chemicals was calculated (i.e. similarity between the chemical and itself was not considered). The distribution of similarity values in the training chemicals was examined to see if a clear cut-off could be made. Ideally, there would be a clear cut-off which separates training active chemicals from training inactive chemicals and this cut-off should be applicable to the test set chemicals.

Androgen Receptor - Alert 1

This structural alert occurred in 461 training active chemicals and 30 training inactive chemicals. As it is contained by many chemicals, it is a good case study. The structure of the alert is shown in Figure 2.12. The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives is shown in Figure 2.13.

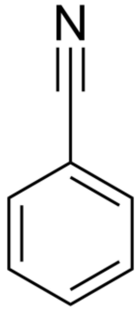


Figure 2.12: The structure of the structural alert "Androgen Receptor - Alert 1".

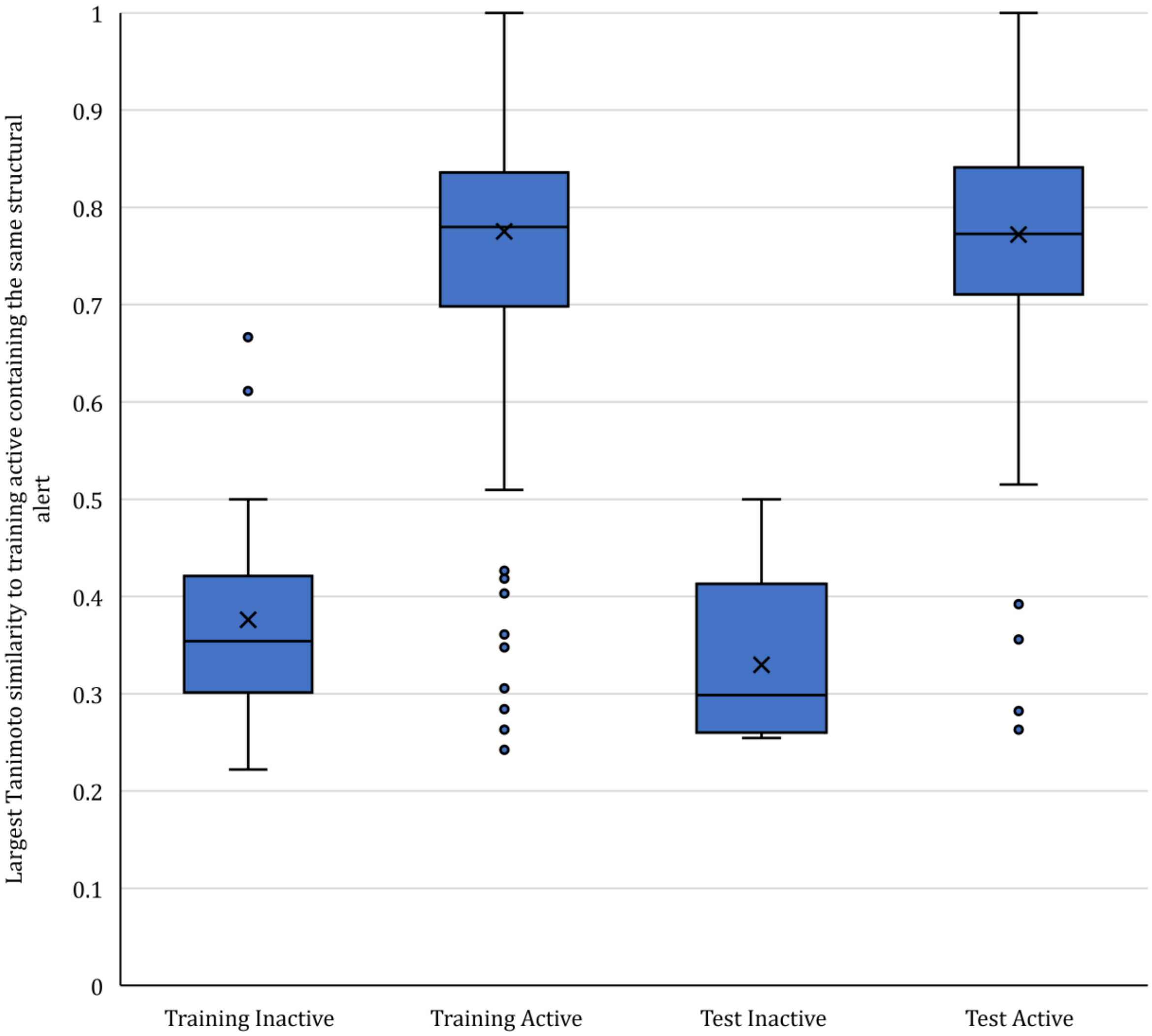


Figure 2.13: The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives for "Alert 1" for the androgen receptor. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a similarity 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

For this alert, a clear difference between the distribution of the training active chemicals and the distribution of the training inactive chemicals can be seen, with the inactive chemicals having a distribution centred around a lower similarity. The whiskers of the box-and-whisker plot, defined as any chemicals with a similarity 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile, suggest a cut-off of at least 0.5 Tanimoto similarity to an alert-containing training active chemical for an active prediction to be considered applicable. Above this cut-off, 99.6% of training chemicals are active (450 active and 2 inactive) and below this cut off, 28.2% of training chemicals are active (11 active and 28 inactive). Similar results are seen by applying the cut off to the test chemicals - 99.3% of test chemicals above the cut-off are active (143 active and 1 inactive) and 46.7% of test chemicals below the cut-off are active (8 active and 7 inactive). With this alert, we see that a clear cut-off can be observed and applied to the predictions from the alert to increase PPV.

Beta-2 adrenergic receptor - Alert 1

This structural alert occurred in 845 training active chemicals and 64 training inactive chemicals. The structure of the alert is shown in Figure 2.14. This alert is also contained by many chemicals, but the substructure is smaller and more flexible than the previous structural alert. The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives is shown in Figure 2.15.

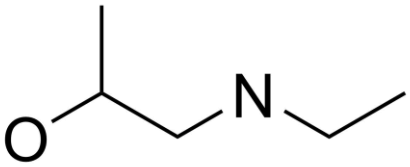


Figure 2.14: The structure of the structural alert "Beta-2 adrenergic receptor - Alert 1".

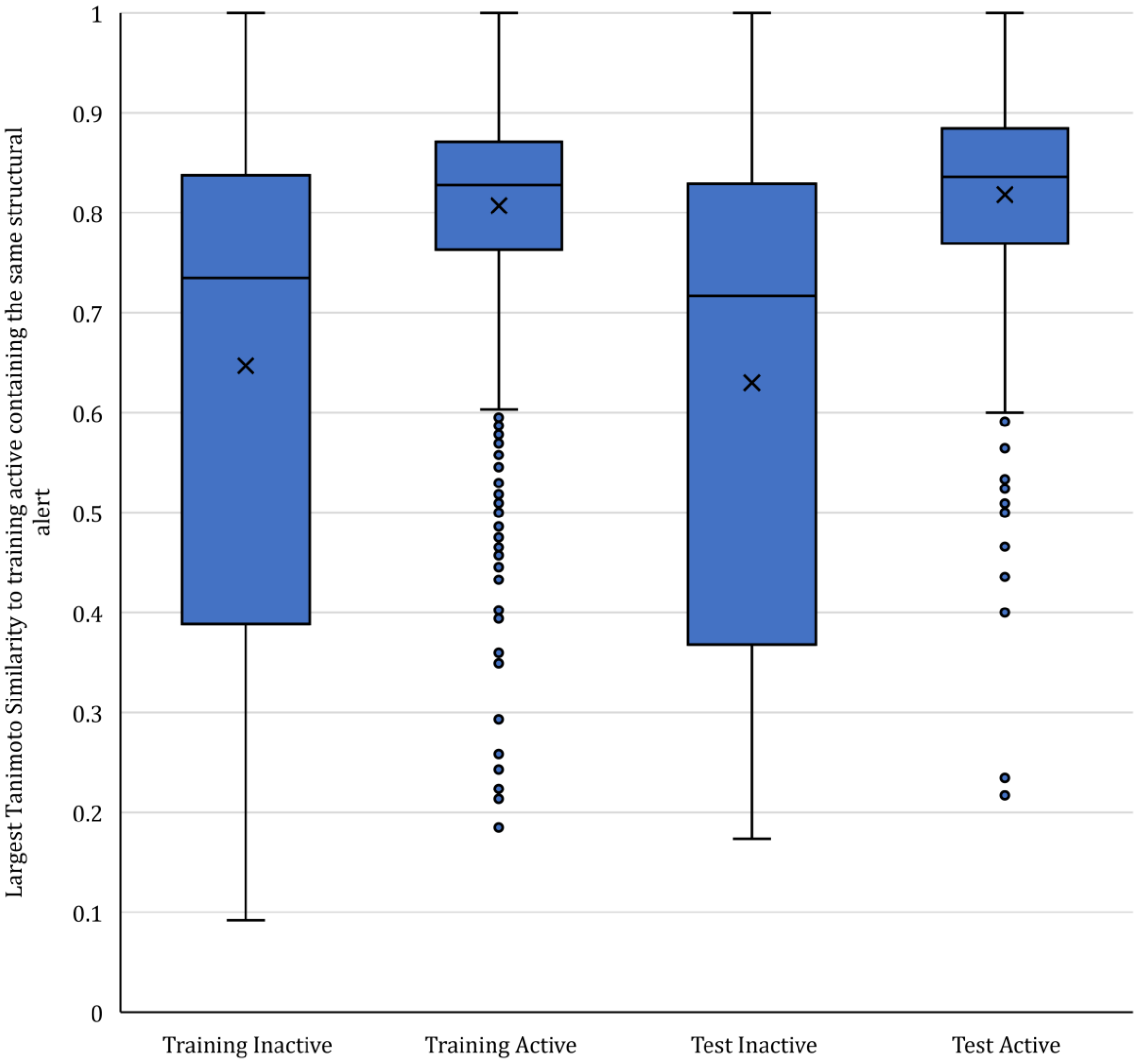


Figure 2.13: The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives for "Alert 1" for the beta-2 adrenergic receptor. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a similarity 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

For this alert, there is not a clear difference between the distribution of the training active chemicals and the distribution of the training inactive chemicals. The distribution of the training active chemicals is skewed towards large similarity values but there are many active chemicals lower than the lower whisker. The distribution of the training inactive chemicals covers a large spread of values from 0.9 to 1.0. For this alert, no clear cut-off can be applied to separate the active chemicals from the inactive. However, using the bottom of the whisker of the training active chemicals would give a cut-off of 0.6. Above this cut-off, 94.8% of training chemicals are active (784 active and 43 inactive) and below this cut off, 74.4% of training chemicals are active (61 active and 21 inactive). Similar results are seen by applying the cut off to the test chemicals - 92.7% of test chemicals above the cut-off are active (267 active and 21 inactive) and 55.2% of test chemicals below the cut-off are active (16 active and 13 inactive). As with the previous alert, applying a cut-off improves PPV. However, this cut-off of 0.6 Tanimoto similarity to an alert-containing training active chemical could be considered a high bar, requiring a large degree of structural similarity beyond the structural alert substructure. Furthermore, PPV is only slightly increased - using the structural alert with no cut-offs gives a PPV of 93.0% in training chemicals and 89.3% in test chemicals. For this alert, it is difficult to assign a single clear cut-off. This may be a result of the flexible nature of the structural alert substructure.

hERG - Alert 4

This structural alert occurred in 93 training active chemicals and 15 training inactive chemicals. This alert is contained by fewer training chemicals than the previous two structural alerts. The structure of the alert is shown in Figure 2.16. The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives is shown in Figure 2.17.

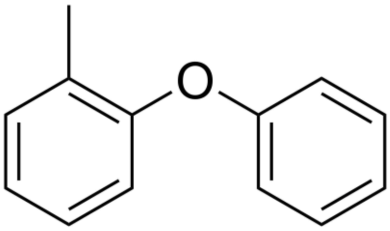


Figure 2.16: The structure of the structural alert "hERG - Alert 4".

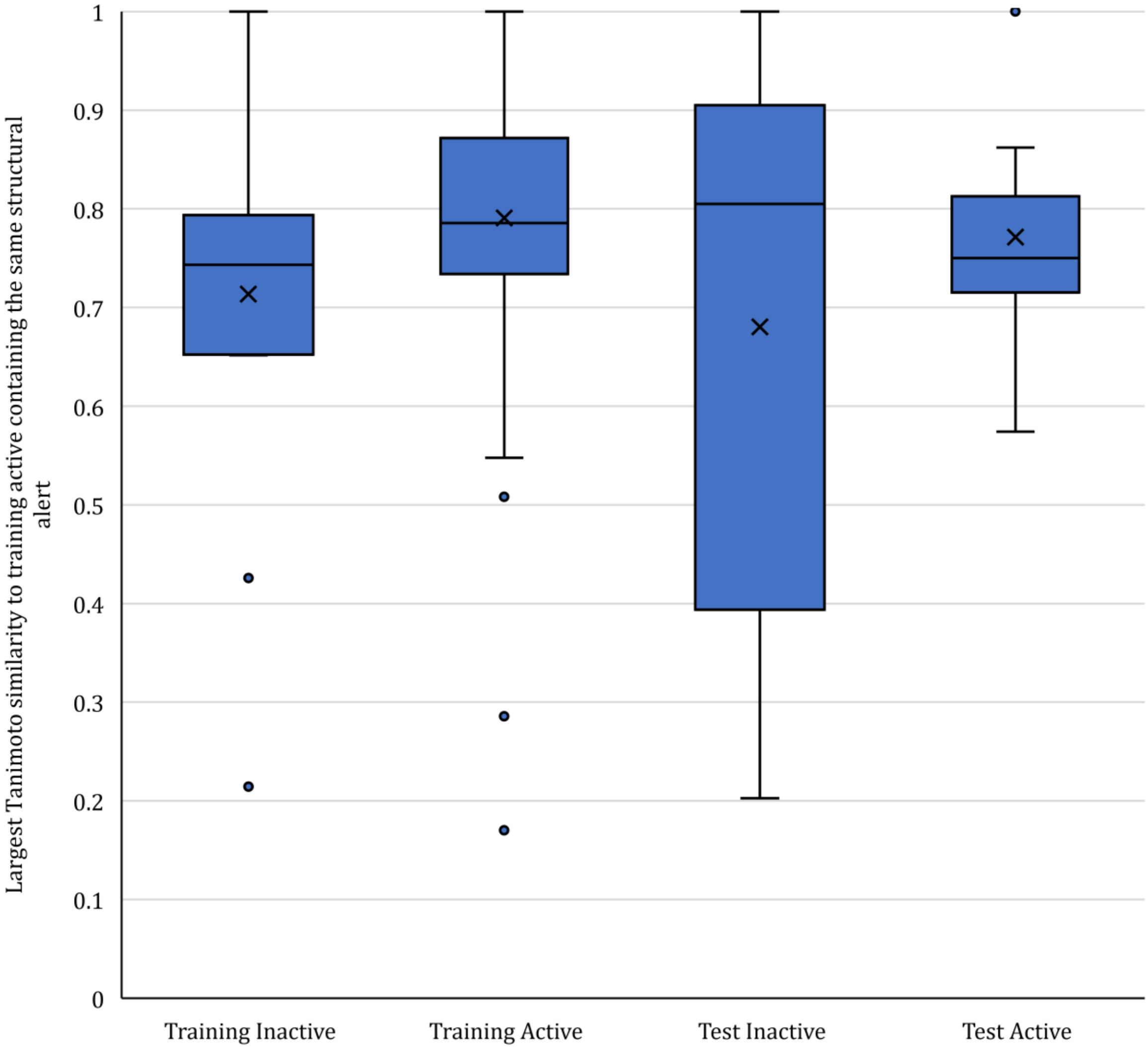


Figure 2.14: The distribution of maximum Tanimoto similarity (between Morgan fingerprints) to alert-containing training actives for "Alert 1" for hERG. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a similarity 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

For this alert, there is no clear difference between the distribution of the training active chemicals and the distribution of the training inactive chemicals. The similarity values for all but two of the inactive chemicals falls between the whiskers of the training active chemicals. Using the bottom whisker of the training active chemicals gives a cut-off of 0.55. Above this cut-off, 87.1% of training chemicals are active (88 active and 13 inactive) and below this cut off, 71.4% of training chemicals are active (5 active and 2 inactive). Here we see the cut-off providing little information in the training chemicals. However, applying the cut-off to the test chemicals gives good results - 88.2% of test chemicals above the cut-off are active (30 active and 4 inactive) and 0% of test chemicals below the cut-off are active (0 active and 1 inactive). Unfortunately, these good results in the test set are not conclusive as there are too few alert-containing test inactive chemicals. This structural alert highlights the difficulty in picking a similarity cut-off for each alert, particularly when there is little data.

In this section, applicability of an active prediction from a structural alert has been judged by considering Tanimoto similarity to training active chemicals containing the same alert. The cut-off of 0.4 found by tuning cut-off values across the test sets of all targets provides a good benchmark value and has been shown to improve PPV. The three examples structural alerts considered show the importance of considering a case-by-case cut-off in similarity, but also the complexities involved. Where the distributions of similarity values of active and inactive chemicals can be clearly distinguished from each other, as with the first alert (androgen receptor - alert 1), a clear cut-off can be identified. This may not give the same cut-off value as found by considering all alerts together. However, the distributions of active and inactive chemicals may not be easily separable, particularly if there is limited data. Whilst a case-by-case consideration of applicability of each alert may be time-demanding, it should be valuable as the global cut-off of 0.4 may not be applicable to all alerts.

2.7. Conclusions

Using human *in vitro* data from both ChEMBL and ToxCast, new data sets have been created for twenty-four Bowes Targets. These targets represent MIEs that are very significant in risk assessment. Unlike previous methods, the data sets are balanced in terms of number of active and inactive chemicals, and no assumptions have been made regarding inactive data points.

Using freely available software, an automated workflow has been designed which builds structural alert-based models for predicting activity. The workflow uses Bayesian statistics to select substructures common to multiple training chemicals to be structural alerts. There are adjustable parameters which give the workflow flexibility, allowing it to be used for different purposes.

The workflow has been applied to the new data sets, creating models with very impressive performance metrics. Two different sets of parameters have been given as examples to show the versatility of the workflow and to compare to previous structural alert-based models for the same targets. On average across the 24 targets, both example models correctly predict over 88% of chemicals in the test sets. Mean values of MCC in the test sets of 0.782 and 0.748 indicate excellent overall performance of the models. The new models are a significant improvement on previous structural alert-based models for the same targets.

Random Forest models have been built for the data sets by Maria Folia (Unilever). These also show excellent predictivity in the test sets, with 91% accuracy and a mean MCC of 0.804. The performance of the new structural alert-based models is similar to the performance of these Random Forest models. The key advantage of the structural alert-based models compared to other models, such as the Random Forest ones, is that the predictions are transparent and easily interpretable.

The structural alert-based models have been combined with the Random Forest models in a consensus model. The two models can be considered as orthogonal methods, making activity predictions using different input data and different algorithms. Where the models agree in predictions, one would have more confidence in the prediction. Compared to both models individually, the consensus approach greatly increases overall performance. The development of the consensus model is potentially very significant. It shows how the models for receptor binding MIEs could be used in risk assessment, comparable to how *in silico* (Q)SARs are already used in predicting mutagenicity according to the ICH M7 guideline.

Generally, when a test chemical contains an alert, we will have more confidence in active prediction if the test chemical is similar to the active training chemicals containing the same alert. In this work, it has been shown that confidence in positive predictions correlates well with the largest Tanimoto similarity (based on Morgan fingerprints) between the test chemical and the training active chemicals containing the same alert. This has been used to define an applicability domain for the structural alerts. For an active prediction to be considered “applicable”, a test chemical containing an alert must have a Tanimoto similarity coefficient (based on Morgan fingerprints) of 0.4 to at least one active training chemical containing the same structural alert. With the addition of applicability domains, the structural alert-based models fulfil the five key priorities set out by OECD for use of (Q)SARs for regulatory purposes.

3. Expanding the scope of the workflow

3.1. Additional biological targets

With the aim of expanding the scope of the work beyond the 24 Bowes Targets previously modelled, the automated workflow was applied to new biological targets. Targets of interest were identified in house at Unilever as providing valuable toxicological information for risk assessment, derived from the biological targets list in Sipes *et al.*'s 2013 paper.[94] From this list of targets, 66 were not Bowes Targets and had data in both ToxCast and ChEMBL databases. In this chapter, models were built for these biological targets.

Whilst testing chemicals for biological activity at all of these additional targets in *in vitro* assays may not be cost effective in early stages of risk assessment, making *in silico* predictions through SAR models would be quick and easy (once models are constructed). These activity predictions for the additional targets will allow for a much broader assessment of potential toxicity.

3.1.1. Data sets

Data sets have been constructed here using the same procedure outlined for the Bowes Targets (Section 2.1.1.).

For each target, bioactivity data for Homo sapiens was downloaded from ChEMBL (data extracted November 2018). Activity reports were filtered to remove any with a confidence score of less than eight. Only activities reported with Standard Units of nM were kept, leaving reports of EC50, IC50, K_i and K_d, and on rare occasions K_{bapp} (apparent binding constant) and K_{inact} (enzyme inactivation constant). RDKit[83] Salt Stripper was used to remove common salts and counter ions from chemicals. All chemicals with more than 100 atoms were removed. The SMILES strings were re-written to be canonical using RDKit, such that the format is consistent across all reports. For each chemical, mean activity was taken – values of activity reported as "greater than" a certain value were removed for these calculations. Chemicals with a mean activity of 10 000 nM or lower were assigned as active; those with over 10 000 nM were assigned as inactive.

For each target, Homo sapiens data was downloaded from the ToxCast Dashboard, using ToxCast's in-built binary activity assignments. As with ChEMBL data, common salts and counter ions were stripped using RDKit Salt Stripper, chemicals with greater than 100 atoms were removed, and SMILES format was re-written using RDKit. If a chemical has contrasting reports of being both active and inactive in different assays it is considered active.

Data from ChEMBL and ToxCast were combined into one data set. Where chemicals have contrasting activity reports between ChEMBL and ToxCast, the activity from ChEMBL is used.

The chemicals were randomly split, with roughly 75% forming the training set, and 25% forming the test set.

A summary of the data sets for each target is shown in Table 3.1.

	Training Sets						Test Sets					
	ChEMBL		ToxCast		Both		ChEMBL		ToxCast		Both	
Target Gene	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives	Actives	Inactives
AGTR1	588	119	13	759	0	1	196	43	8	260	1	0
AKT1	2057	144	14	755	1	1	690	54	4	263	0	4
BACE1	4409	550	64	1425	1	0	1516	175	26	455	0	1
BCHE	976	426	60	1178	2	9	341	150	20	381	1	1
CASP1	1002	1626	7	719	1	83	355	543	3	203	1	25
CASP10	4	0	4	793	0	0	4	0	2	243	0	0
CASP2	26	12	11	781	0	0	6	8	4	246	0	0
CASP3	876	594	5	757	0	3	292	201	4	273	0	0
CASP5	15	11	29	1228	0	0	4	1	10	390	0	0
CASP8	233	48	13	801	0	0	79	15	5	268	0	0
CHRM5	472	77	36	660	4	73	149	24	15	227	3	24
CHUK	224	22	3	782	0	0	88	10	1	256	0	0
CSF1R	987	23	13	766	0	0	332	5	4	259	0	0
CSNK1D	496	18	28	758	0	0	178	3	8	248	0	0
EDNRB	604	151	0	779	0	1	205	60	0	245	0	0
ELANE	1554	289	37	655	3	74	526	98	14	233	0	26
EPHA2	382	18	5	811	0	0	139	4	2	269	0	0
EPHB2	34	13	2	761	0	0	7	5	2	277	0	0
FGFR1	1612	122	24	798	0	0	522	28	5	259	0	0
FKBP1A	253	69	0	671	0	1	101	28	0	239	0	0
FLT1	778	82	10	1413	1	101	296	12	2	434	1	38
FLT4	479	8	9	821	0	0	181	2	5	252	0	0
FYN	305	41	12	698	4	67	95	14	2	226	2	31
GSK3B	1911	170	21	753	1	7	613	71	4	256	0	0
HDAC3	784	91	10	797	1	1	250	25	6	226	0	1
IGF1R	1867	75	9	745	1	1	603	30	4	282	0	0
INSR	653	45	5	784	0	0	230	11	0	253	0	0
KDR	5647	278	181	922	3	1	1930	79	55	300	0	0
LTB4R	239	9	17	776	1	0	89	3	4	244	0	0
LYN	335	16	9	776	0	1	106	4	4	252	0	0
MAPK1	4671	7601	4	661	25	91	1505	2467	2	228	2	29
MAPK3	48	7	10	735	1	64	21	6	4	232	2	39
MAPK9	911	44	5	778	0	1	312	8	0	258	0	0
MAPKAPK2	599	112	16	759	0	2	209	26	5	260	0	0
MET	2170	86	8	749	1	0	689	29	3	281	0	0

MMP13	1791	92	32	717	1	1	550	25	13	277	1	0
MMP2	2159	480	25	774	1	7	741	149	10	268	2	0
MMP3	1284	147	25	622	0	7	440	48	8	214	2	0
MMP9	1643	196	268	1105	29	83	523	69	108	372	11	24
NEK2	215	26	9	758	0	1	71	4	2	271	1	0
NR1I3	39	9	336	2212	2	2	9	0	126	685	1	1
P2RY1	431	10	6	826	0	0	120	2	3	265	0	0
PAK4	286	24	9	807	0	1	83	3	2	267	0	1
PDE4A	439	28	41	755	0	0	155	7	18	227	0	0
PDE5A	1142	148	54	666	2	74	340	28	12	233	1	27
PIK3CA	3524	154	17	1387	1	6	1179	45	3	493	0	1
PPARG	2365	250	874	4949	38	309	757	90	315	1559	13	128
PPP1CA	51	20	16	689	0	0	12	7	3	236	0	0
PPP2CA	3	1	5	761	0	0	3	0	1	275	0	0
PTEN	0	0	30	1479	0	0	0	0	6	485	0	0
PTPN1	1118	864	19	770	0	4	330	299	4	244	0	1
PTPN11	246	110	20	804	0	1	79	44	9	253	0	0
PTPN13	11	10	21	750	0	0	5	1	10	261	0	0
PTPN14	0	0	4	770	0	0	2	0	6	485	0	0
PTPN2	244	132	14	776	0	0	75	55	6	246	0	0
RAF1	1009	36	8	779	0	0	333	15	1	254	0	0
RARA	164	2	87	2463	2	0	69	5	32	779	2	0
RARB	205	8	19	2516	3	0	65	3	6	821	0	0
ROCK1	978	56	2	785	0	2	315	22	0	253	0	0
RPS6KA5	163	4	4	773	0	0	53	2	4	260	0	0
SIRT2	253	206	19	771	0	0	77	71	12	240	0	0
SIRT3	93	47	15	764	0	0	37	7	6	257	0	0
SRC	2071	373	8	763	1	4	622	128	2	263	1	0
TACR2	564	21	49	1273	13	91	228	8	19	492	3	32
TBXA2R	731	57	39	1352	0	1	211	33	21	461	0	0
TEK	576	41	13	819	0	1	196	20	3	251	0	0

Table 3.1: The additional biological targets and the number of active and inactive chemicals in the data sets. Data is taken from human in vitro *data from ChEMBL and ToxCast databases. Actives in ChEMBL are chemicals with a mean activity of less than 10 000 nM and inactives as chemicals with a mean activity of greater than 10 000 nM. ToxCast's inbuilt definitions of activity are used for chemicals extracted from ToxCast.*

As also seen with the Bowes targets previously, for most targets ChEMBL provides far more active chemicals than inactive, with only Caspase 1 (CASP1) and Mitogen-Activated Protein Kinase 1 (MAPK1) having more inactives than actives. In contrast, ToxCast provides far more inactive chemicals than active chemicals for all targets. Alone, each database gives imbalanced data that would be difficult to model.

By combining the two databases, data sets have been created that are balanced in terms of similar numbers of active and inactive chemicals. All data comes directly from human *in vitro* assays for the biological target and so predictions based on the data will be relevant to humans without the need for cross-species extrapolation.

However, one should be aware that ChEMBL and ToxCast tend to cover different areas of chemical space. ChEMBL largely contains pharmaceuticals whilst much of ToxCast is made up of insecticides, pesticides and other reactive chemicals which have been tested across many assays.

3.1.2. Methods

3.1.2.1. Structural alerts

The automated workflow for construction of structural alert-based models has been applied to the data sets for the new biological targets. The same procedure was used as previously (Section 2.2.2).

A training set of chemicals (in SMILES format) and binary activities was inputted into the workflow. The maximum common substructures occurring in at least two of the active chemicals were found using the MoSS node[68] in KNIME.[67] MoSS will only output substructures which occur in less than a certain percentage of the inactive chemicals. This value was a parameter which can be selected by the user – choosing a larger value will result in the workflow taking longer to run.

MoSS outputs the common substructures and how many times each occurs in the active and inactive chemicals, according to the MoSS algorithm. However, these values are slightly inaccurate due to ring mining used in the algorithm. Re-calculating accurate counts for all substructures output by MoSS would be too time consuming as often many thousands of substructures were output. Instead, Bayes Factor was calculated for each substructure using the occurrence in actives and inactives calculated by MoSS, and only the substructures with the 65 largest values are kept. It was assumed here that the inaccuracies in the counts given by the MoSS algorithm were not so large that the actual best performing substructure was not in the top 65 substructures. Accurate values for occurrence of active and inactive chemicals were calculated for the 65 substructures, and Bayes Factor recalculated. Only the substructure with the highest value of Bayes Factor was kept. When two substructures had the same Bayes Factor, the substructure which occurs in more active chemicals was chosen.

The user decided the lower bounds for a structural alert in terms of number of actives and inactive chemicals, and the lower bounds Bayes Factor was calculated using these values. If the remaining substructure had a Bayes Factor larger than the lower bounds and was contained by more actives than the minimum required number, it was added to the list of structural alerts. Any active chemicals containing the substructure were removed from the training set and the whole process was repeated iteratively until no substructures satisfied the lower bounds for an alert.

This iterative process produced a list of independent structural alerts. Chemicals containing a structural alert were predicted to be active, and those containing no alerts were predicted to be inactive.

The resulting model was applied to both the training set and test set, and performance statistics were calculated for both.

3.1.2.2. Random forest

Maria Folia (Unilever) has constructed random forest models using the same training sets as the structural-alert based models.

The same procedure has been used to create the random forest models as previously for the Bowes targets (section 2.4.2.1).

RDKit[83] was used to create 200 physiochemical descriptors for each chemical including molecular, topological, van der Waals surface area (VSA) and lipophilicity descriptors. The model was built using the RandomForestClassifier from the sklearn package and kept the default settings apart from two hyperparameters, the number of trees (n_estimators) and the maximum depth of the trees (max_depth) which were tuned using GridSearchCV.

The random forest models have been applied to the test set and performance statistics were calculated.

3.1.3. Results and discussion

Structural alert models have been created for the new targets using the same two sets of parameters as for the Bowes Targets: a “risk assessment” model (theta 0.95, 1% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive) and a “screening” model (theta 0.51, 15% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive). The performance of these models for each biological target are shown in Tables 3.2 and 3.3.

Random forest models have been created for the same data sets by Maria Folia. The same methodology was used as for the construction of random forest models for the Bowes targets. The performance of this model for each biological target is shown in Table 3.4.

Target Gene Symbol	Alerts	Training set								Test set							
		TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC
AGTR1	27	559	11	42	868	93.0%	98.7%	96.4%	0.926	179	5	26	298	87.3%	98.3%	93.9%	0.874
AKT1	62	2013	72	59	828	97.2%	92.0%	95.6%	0.895	655	42	39	279	94.4%	86.9%	92.0%	0.815
BACE1	140	4350	196	124	1779	97.2%	90.1%	95.0%	0.882	1436	77	106	554	93.1%	87.8%	91.6%	0.799
BCHE	101	901	54	137	1559	86.8%	96.7%	92.8%	0.848	284	37	78	495	78.5%	93.0%	87.1%	0.732
CASP1	82	717	36	293	2392	71.0%	98.5%	90.4%	0.766	229	40	130	731	63.8%	94.8%	85.0%	0.641
CASP10	1	4	0	4	793	50.0%	100.0%	99.5%	0.705	3	0	3	243	50.0%	100.0%	98.8%	0.703
CASP2	3	24	2	13	791	64.9%	99.7%	98.2%	0.765	4	3	6	251	40.0%	98.8%	96.6%	0.461
CASP3	54	757	40	124	1314	85.9%	97.0%	92.7%	0.846	241	19	55	455	81.4%	96.0%	90.4%	0.796
CASP5	7	24	3	20	1236	54.5%	99.8%	98.2%	0.689	3	2	11	389	21.4%	99.5%	96.8%	0.346
CASP8	5	232	15	14	834	94.3%	98.2%	97.4%	0.924	77	7	7	276	91.7%	97.5%	96.2%	0.892
CHRM5	48	465	23	47	787	90.8%	97.2%	94.7%	0.888	140	17	27	258	83.8%	93.8%	90.0%	0.787
CHUK	21	202	12	25	792	89.0%	98.5%	96.4%	0.894	74	8	15	258	83.1%	97.0%	93.5%	0.824
CSF1R	49	960	22	40	767	96.0%	97.2%	96.5%	0.930	297	10	39	254	88.4%	96.2%	91.8%	0.840
CSNK1D	43	479	10	45	766	91.4%	98.7%	95.8%	0.913	147	9	39	242	79.0%	96.4%	89.0%	0.779
EDNRB	28	579	21	25	910	95.9%	97.7%	97.0%	0.937	193	16	12	289	94.1%	94.8%	94.5%	0.886
ELANE	57	1525	27	69	991	95.7%	97.3%	96.3%	0.924	493	23	47	334	91.3%	93.6%	92.2%	0.840
EPHA2	16	358	9	29	820	92.5%	98.9%	96.9%	0.928	123	4	18	269	87.2%	98.5%	94.7%	0.881
EPHB2	10	27	2	9	772	75.0%	99.7%	98.6%	0.829	3	3	6	279	33.3%	98.9%	96.9%	0.393
FGFR1	69	1549	57	87	863	94.7%	93.8%	94.4%	0.879	479	15	48	273	90.9%	94.8%	92.3%	0.838
FKBP1A	11	237	9	16	732	93.7%	98.8%	97.5%	0.933	92	6	9	261	91.1%	97.8%	95.9%	0.897
FLT1	48	739	30	50	1566	93.7%	98.1%	96.6%	0.924	269	9	30	475	90.0%	98.1%	95.0%	0.895
FLT4	48	452	14	36	815	92.6%	98.3%	96.2%	0.918	151	8	35	246	81.2%	96.9%	90.2%	0.802
FYN	42	262	12	59	794	81.6%	98.5%	93.7%	0.843	72	6	27	265	72.7%	97.8%	91.1%	0.765
GSK3B	125	1816	63	117	867	93.9%	93.2%	93.7%	0.860	522	54	95	273	84.6%	83.5%	84.2%	0.664
HDAC3	39	760	43	35	846	95.6%	95.2%	95.4%	0.907	235	17	21	235	91.8%	93.3%	92.5%	0.851
IGF1R	54	1838	34	39	787	97.9%	95.9%	97.3%	0.936	580	18	27	294	95.6%	94.2%	95.1%	0.892
INSR	34	636	15	22	814	96.7%	98.2%	97.5%	0.950	207	6	23	258	90.0%	97.7%	94.1%	0.884
KDR	151	5628	148	203	1053	96.5%	87.7%	95.0%	0.827	1878	65	107	314	94.6%	82.8%	92.7%	0.743
LTB4R	16	237	4	20	781	92.2%	99.5%	97.7%	0.938	81	3	12	244	87.1%	98.8%	95.6%	0.888
LYN	28	317	16	27	777	92.2%	98.0%	96.2%	0.910	86	12	24	244	78.2%	95.3%	90.2%	0.761
MAPK1	453	3023	103	1677	8250	64.3%	98.8%	86.4%	0.710	723	132	786	2592	47.9%	95.2%	78.3%	0.514
MAPK3	6	39	0	20	806	66.1%	100.0%	97.7%	0.803	14	1	13	276	51.9%	99.6%	95.4%	0.676
MAPK9	40	882	23	34	800	96.3%	97.2%	96.7%	0.934	287	12	25	254	92.0%	95.5%	93.6%	0.873
MAPKAPK2	49	557	22	58	851	90.6%	97.5%	94.6%	0.889	175	21	39	265	81.8%	92.7%	88.0%	0.754
MET	62	2123	52	56	783	97.4%	93.8%	96.4%	0.911	646	35	46	275	93.4%	88.7%	91.9%	0.813
MMP13	52	1770	43	54	767	97.0%	94.7%	96.3%	0.914	527	26	37	276	93.4%	91.4%	92.7%	0.842
MMP2	71	2091	87	94	1174	95.7%	93.1%	94.7%	0.887	706	30	47	387	93.8%	92.8%	93.4%	0.858
MMP3	38	1266	35	43	741	96.7%	95.5%	96.3%	0.920	418	23	32	239	92.9%	91.2%	92.3%	0.835

MMP9	103	1755	76	185	1308	90.5%	94.5%	92.1%	0.842	525	45	117	420	81.8%	90.3%	85.4%	0.712
NEK2	28	181	8	43	777	80.8%	99.0%	94.9%	0.850	50	6	24	269	67.6%	97.8%	91.4%	0.728
NR1I3	58	183	25	194	2198	48.5%	98.9%	91.6%	0.615	22	19	114	667	16.2%	97.2%	83.8%	0.229
P2RY1	7	432	13	5	823	98.9%	98.4%	98.6%	0.969	117	4	6	263	95.1%	98.5%	97.4%	0.940
PAK4	29	264	6	31	826	89.5%	99.3%	96.7%	0.914	70	4	15	267	82.4%	98.5%	94.7%	0.850
PDE4A	23	440	20	40	763	91.7%	97.4%	95.2%	0.899	152	11	21	223	87.9%	95.3%	92.1%	0.839
PDE5A	51	1106	55	92	833	92.3%	93.8%	93.0%	0.857	306	18	47	270	86.7%	93.7%	89.9%	0.800
PIK3CA	69	3486	78	56	1469	98.4%	95.0%	97.4%	0.938	1151	28	31	511	97.4%	94.8%	96.6%	0.920
PPARG	227	2675	186	602	5322	81.6%	96.6%	91.0%	0.808	763	129	322	1648	70.3%	92.7%	84.2%	0.661
PPP1CA	13	51	3	16	706	76.1%	99.6%	97.6%	0.836	6	1	9	242	40.0%	99.6%	96.1%	0.570
PPP2CA	1	3	1	5	761	37.5%	99.9%	99.2%	0.527	2	0	2	275	50.0%	100.0%	99.3%	0.705
PTEN	2	4	1	26	1478	13.3%	99.9%	98.2%	0.322	0	4	6	481	0.0%	99.2%	98.0%	-0.010
PTPN1	148	963	57	174	1581	84.7%	96.5%	91.7%	0.828	224	42	110	502	67.1%	92.3%	82.7%	0.627
PTPN11	46	186	5	80	910	69.9%	99.5%	92.8%	0.787	34	14	54	283	38.6%	95.3%	82.3%	0.431
PTPN13	5	14	2	18	758	43.8%	99.7%	97.5%	0.609	3	2	12	260	20.0%	99.2%	94.9%	0.327
PTPN14	1	2	0	2	770	50.0%	100.0%	99.7%	0.706	0	0	4	266	0.0%	100.0%	98.5%	-
PTPN2	20	207	10	51	898	80.2%	98.9%	94.8%	0.844	53	7	28	294	65.4%	97.7%	90.8%	0.709
RAF1	29	982	21	35	794	96.6%	97.4%	96.9%	0.938	319	10	15	259	95.5%	96.3%	95.9%	0.916
RARA	23	190	17	63	2448	75.1%	99.3%	97.1%	0.815	63	19	40	765	61.2%	97.6%	93.3%	0.650
RARB	10	201	14	26	2510	88.5%	99.4%	98.5%	0.902	53	10	18	814	74.6%	98.8%	96.9%	0.776
ROCK1	53	932	26	48	817	95.1%	96.9%	95.9%	0.919	280	12	35	263	88.9%	95.6%	92.0%	0.843
RPS6KA5	21	142	4	25	773	85.0%	99.5%	96.9%	0.892	35	4	22	259	61.4%	98.5%	91.9%	0.700
SIRT2	35	223	11	49	966	82.0%	98.9%	95.2%	0.855	61	13	28	298	68.5%	95.8%	89.8%	0.689
SIRT3	8	82	7	26	804	75.9%	99.1%	96.4%	0.817	34	6	9	258	79.1%	97.7%	95.1%	0.792
SRC	96	1987	89	93	1051	95.5%	92.2%	94.3%	0.877	563	47	62	344	90.1%	88.0%	89.3%	0.775
TACR2	26	575	17	51	1368	91.9%	98.8%	96.6%	0.921	223	13	27	519	89.2%	97.6%	94.9%	0.881
TBXA2R	41	729	29	41	1381	94.7%	97.9%	96.8%	0.930	207	23	25	471	89.2%	95.3%	93.4%	0.848
TEK	32	553	11	36	850	93.9%	98.7%	96.8%	0.933	174	13	25	258	87.4%	95.2%	91.9%	0.834
Average	**51**	**893**	**33**	**91**	**1209**	**84.1%**	**97.4%**	**95.8%**	**0.852**	**276**	**20**	**51**	**390**	**74.1%**	**95.4%**	**92.2%**	**0.740**

Table 3.2: Performance of "Risk Assessment" models on the data sets of the additional biological targets. Models created using the automated workflow for creation of structural alert-based models, with parameters: theta 0.95, 1% maximum occurrence of an alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. TP = true positives; FP = false positives; FN = false negatives; TN = true negatives; SE = sensitivity, SP = specificity; ACC = accuracy; MCC = Matthews correlation coefficient.

Target Gene Symbol	Alerts	Training set								Test set							
		TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC
AGTR1	26	561	21	40	858	93.3%	97.6%	95.9%	0.914	182	8	23	295	88.8%	97.4%	93.9%	0.873
AKT1	33	2025	200	47	700	97.7%	77.8%	91.7%	0.800	664	76	30	245	95.7%	76.3%	89.6%	0.753
BACE1	45	4405	452	69	1523	98.5%	77.1%	91.9%	0.808	1486	150	56	481	96.4%	76.2%	90.5%	0.764
BCHE	52	954	179	84	1434	91.9%	88.9%	90.1%	0.797	316	81	46	451	87.3%	84.8%	85.8%	0.712
CASP1	64	852	147	158	2281	84.4%	93.9%	91.1%	0.786	283	69	76	702	78.8%	91.1%	87.2%	0.703
CASP10	1	4	0	4	793	50.0%	100.0%	99.5%	0.705	3	0	3	243	50.0%	100.0%	98.8%	0.703
CASP2	3	24	2	13	791	64.9%	99.7%	98.2%	0.765	4	4	6	250	40.0%	98.4%	96.2%	0.428
CASP3	36	788	111	93	1243	89.4%	91.8%	90.9%	0.810	260	45	36	429	87.8%	90.5%	89.5%	0.779
CASP5	7	26	4	18	1235	59.1%	99.7%	98.3%	0.708	3	4	11	387	21.4%	99.0%	96.3%	0.286
CASP8	4	232	15	14	834	94.3%	98.2%	97.4%	0.924	77	7	7	276	91.7%	97.5%	96.2%	0.892
CHRM5	26	476	57	36	753	93.0%	93.0%	93.0%	0.853	151	24	16	251	90.4%	91.3%	91.0%	0.810
CHUK	16	208	33	19	771	91.6%	95.9%	95.0%	0.857	73	11	16	255	82.0%	95.9%	92.4%	0.794
CSF1R	24	976	122	24	667	97.6%	84.5%	91.8%	0.838	321	40	15	224	95.5%	84.8%	90.8%	0.815
CSNK1D	23	478	69	46	707	91.2%	91.1%	91.2%	0.818	158	26	28	225	84.9%	89.6%	87.6%	0.747
EDNRB	23	584	79	20	852	96.7%	91.5%	93.6%	0.870	196	34	9	271	95.6%	88.9%	91.6%	0.832
ELANE	28	1551	195	43	823	97.3%	80.8%	90.9%	0.810	509	68	31	289	94.3%	81.0%	89.0%	0.769
EPHA2	13	359	13	28	816	92.8%	98.4%	96.6%	0.922	123	7	18	266	87.2%	97.4%	94.0%	0.865
EPHB2	6	28	8	8	766	77.8%	99.0%	98.0%	0.767	2	5	7	277	22.2%	98.2%	95.9%	0.231
FGFR1	31	1595	215	41	705	97.5%	76.6%	90.0%	0.783	507	52	20	236	96.2%	81.9%	91.2%	0.805
FKBP1A	4	247	30	6	711	97.6%	96.0%	96.4%	0.909	99	12	2	255	98.0%	95.5%	96.2%	0.909
FLT1	23	751	204	38	1392	95.2%	87.2%	89.9%	0.791	278	43	21	441	93.0%	91.1%	91.8%	0.831
FLT4	20	461	117	27	712	94.5%	85.9%	89.1%	0.782	171	35	15	219	91.9%	86.2%	88.6%	0.774
FYN	20	277	115	44	691	86.3%	85.7%	85.9%	0.683	81	36	18	235	81.8%	86.7%	85.4%	0.652
GSK3B	50	1869	321	64	609	96.7%	65.5%	86.6%	0.687	574	128	43	199	93.0%	60.9%	81.9%	0.587
HDAC3	13	769	130	26	759	96.7%	85.4%	90.7%	0.822	247	41	9	211	96.5%	83.7%	90.2%	0.809
IGF1R	28	1854	135	23	686	98.8%	83.6%	94.1%	0.861	590	62	17	250	97.2%	80.1%	91.4%	0.807
INSR	15	639	105	19	724	97.1%	87.3%	91.7%	0.839	206	38	24	226	89.6%	85.6%	87.4%	0.750
KDR	67	5702	370	129	831	97.8%	69.2%	92.9%	0.734	1926	118	59	261	97.0%	68.9%	92.5%	0.707
LTB4R	12	236	25	21	760	91.8%	96.8%	95.6%	0.882	83	9	10	238	89.2%	96.4%	94.4%	0.859
LYN	19	326	102	18	691	94.8%	87.1%	89.4%	0.777	97	31	13	225	88.2%	87.9%	88.0%	0.731
MAPK1	356	3033	380	1667	7973	64.5%	95.5%	84.3%	0.655	760	222	749	2502	50.4%	91.9%	77.1%	0.479
MAPK3	6	39	0	20	806	66.1%	100.0%	97.7%	0.803	14	1	13	276	51.9%	99.6%	95.4%	0.676
MAPK9	32	888	77	28	746	96.9%	90.6%	94.0%	0.880	293	30	19	236	93.9%	88.7%	91.5%	0.829
MAPKAPK2	29	585	88	30	785	95.1%	89.9%	92.1%	0.841	192	38	22	248	89.7%	86.7%	88.0%	0.759
MET	30	2138	157	41	678	98.1%	81.2%	93.4%	0.833	673	78	19	232	97.3%	74.8%	90.3%	0.769
MMP13	23	1784	133	40	677	97.8%	83.6%	93.4%	0.844	546	57	18	245	96.8%	81.1%	91.3%	0.808
MMP2	44	2111	194	74	1067	96.6%	84.6%	92.2%	0.831	721	71	32	346	95.8%	83.0%	91.2%	0.806
MMP3	21	1276	98	33	678	97.5%	87.4%	93.7%	0.865	431	42	19	220	95.8%	84.0%	91.4%	0.814

MMP9	52	1801	249	139	1135	92.8%	82.0%	88.3%	0.759	547	95	95	370	85.2%	79.6%	82.8%	0.648
NEK2	22	189	18	35	767	84.4%	97.7%	94.7%	0.845	55	10	19	265	74.3%	96.4%	91.7%	0.742
NR1I3	53	196	38	181	2185	52.0%	98.3%	91.6%	0.619	24	22	112	664	17.6%	96.8%	83.7%	0.233
P2RY1	4	432	101	5	735	98.9%	87.9%	91.7%	0.835	119	34	4	233	96.7%	87.3%	90.3%	0.800
PAK4	23	269	26	26	806	91.2%	96.9%	95.4%	0.881	75	13	10	258	88.2%	95.2%	93.5%	0.825
PDE4A	15	447	64	33	719	93.1%	91.8%	92.3%	0.840	161	23	12	211	93.1%	90.2%	91.4%	0.827
PDE5A	31	1138	171	60	717	95.0%	80.7%	88.9%	0.775	333	50	20	238	94.3%	82.6%	89.1%	0.781
PIK3CA	34	3513	195	29	1352	99.2%	87.4%	95.6%	0.896	1159	69	23	470	98.1%	87.2%	94.7%	0.875
PPARG	157	2728	614	549	4894	83.2%	88.9%	86.8%	0.718	816	254	269	1523	75.2%	85.7%	81.7%	0.611
PPP1CA	12	51	4	16	705	76.1%	99.4%	97.4%	0.827	6	1	9	242	40.0%	99.6%	96.1%	0.570
PPP2CA	1	3	1	5	761	37.5%	99.9%	99.2%	0.527	2	0	2	275	50.0%	100.0%	99.3%	0.705
PTEN	2	4	1	26	1478	13.3%	99.9%	98.2%	0.322	0	4	6	481	0.0%	99.2%	98.0%	-0.010
PTPN1	75	1041	280	96	1358	91.6%	82.9%	86.5%	0.733	272	112	62	432	81.4%	79.4%	80.2%	0.596
PTPN11	36	202	32	64	883	75.9%	96.5%	91.9%	0.759	48	23	40	274	54.5%	92.3%	83.6%	0.507
PTPN13	4	13	2	19	758	40.6%	99.7%	97.3%	0.583	3	2	12	260	20.0%	99.2%	94.9%	0.327
PTPN14	1	2	0	2	770	50.0%	100.0%	99.7%	0.706	0	0	4	266	0%	100%	98.5%	-
PTPN2	15	223	50	35	858	86.4%	94.5%	92.7%	0.793	59	23	22	278	72.8%	92.4%	88.2%	0.649
RAF1	13	995	130	22	685	97.8%	84.0%	91.7%	0.836	330	51	4	218	98.8%	81.0%	90.9%	0.823
RARA	19	189	22	64	2443	74.7%	99.1%	96.8%	0.801	63	19	40	765	61.2%	97.6%	93.3%	0.650
RARB	7	205	21	22	2503	90.3%	99.2%	98.4%	0.897	56	11	15	813	78.9%	98.7%	97.1%	0.796
ROCK1	27	958	135	22	708	97.8%	84.0%	91.4%	0.832	298	38	17	237	94.6%	86.2%	90.7%	0.814
RPS6KA5	19	146	11	21	766	87.4%	98.6%	96.6%	0.881	37	6	20	257	64.9%	97.7%	91.9%	0.703
SIRT2	23	230	31	42	946	84.6%	96.8%	94.2%	0.826	65	17	24	294	73.0%	94.5%	89.8%	0.696
SIRT3	4	93	19	15	792	86.1%	97.7%	96.3%	0.825	35	9	8	255	81.4%	96.6%	94.5%	0.772
SRC	46	2030	245	50	895	97.6%	78.5%	90.8%	0.799	596	93	29	298	95.4%	76.2%	88.0%	0.745
TACR2	17	578	62	48	1323	92.3%	95.5%	94.5%	0.873	231	27	19	505	92.4%	94.9%	94.1%	0.866
TBXA2R	20	737	173	33	1237	95.7%	87.7%	90.6%	0.809	211	71	21	423	90.9%	85.6%	87.3%	0.733
TEK	24	560	80	29	781	95.1%	90.7%	92.5%	0.848	179	24	20	247	89.9%	91.1%	90.6%	0.809
Average	**30.7**	**910**	**113**	**73**	**1129**	**86.1%**	**90.5%**	**93.2%**	**0.795**	**289**	**44**	**38**	**366**	**77.8%**	**89.3%**	**90.9%**	**0.708**

Table 3.3: Performance of "Screening" models on the data sets of the additional biological targets. Models created using the automated workflow for creation of structural alert-based models, with parameters: theta 0.51, 15% maximum occurrence of an alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. TP = true positives; FP = false positives; FN = false negatives; TN = true negatives; SE = sensitivity, SP = specificity; ACC = accuracy; MCC = Matthews correlation coefficient.

	Test set							
Target Gene Symbol	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
AGTR1	172	8	33	295	83.9%	97.4%	91.9%	0.834
AKT1	672	49	22	272	96.8%	84.7%	93.0%	0.836
BACE1	1506	130	36	501	97.7%	79.4%	92.4%	0.811
BCHE	313	59	49	473	86.5%	88.9%	87.9%	0.751
CASP1	255	40	104	731	71.0%	94.8%	87.3%	0.698
CASP10	2	0	4	243	33.3%	100.0%	98.4%	0.573
CASP2	3	3	7	251	30.0%	98.8%	96.2%	0.369
CASP3	240	19	56	455	81.1%	96.0%	90.3%	0.794
CASP5	5	1	9	390	35.7%	99.7%	97.5%	0.536
CASP8	75	5	9	278	89.3%	98.2%	96.2%	0.891
CHRM5	153	24	14	251	91.6%	91.3%	91.4%	0.820
CHUK	82	7	7	259	92.1%	97.4%	96.1%	0.895
CSF1R	325	16	11	248	96.7%	93.9%	95.5%	0.909
CSNK1D	163	18	23	233	87.6%	92.8%	90.6%	0.808
EDNRB	198	15	7	290	96.6%	95.1%	95.7%	0.911
ELANE	519	43	21	314	96.1%	88.0%	92.9%	0.851
EPHA2	124	6	17	267	87.9%	97.8%	94.4%	0.876
EPHB2	2	0	7	282	22.2%	100.0%	97.6%	0.466
FGFR1	506	27	21	261	96.0%	90.6%	94.1%	0.871
FKBP1A	96	11	5	256	95.0%	95.9%	95.7%	0.894
FLT1	283	17	16	467	94.6%	96.5%	95.8%	0.911
FLT4	171	12	15	242	91.9%	95.3%	93.9%	0.874
FYN	80	9	19	262	80.8%	96.7%	92.4%	0.803
GSK3B	604	79	13	248	97.9%	75.8%	90.3%	0.784
HDAC3	240	20	16	232	93.8%	92.1%	92.9%	0.858
IGF1R	587	30	20	282	96.7%	90.4%	94.6%	0.878
INSR	218	8	12	256	94.8%	97.0%	96.0%	0.919
KDR	1939	137	46	242	97.7%	63.9%	92.3%	0.690
LTB4R	83	6	10	241	89.2%	97.6%	95.3%	0.880
LYN	100	12	10	244	90.9%	95.3%	94.0%	0.858
MAPK1	622	16	887	2708	41.2%	99.4%	78.7%	0.544
MAPK3	12	0	15	277	44.4%	100.0%	95.1%	0.649
MAPK9	307	27	5	239	98.4%	89.8%	94.5%	0.891
MAPKAPK2	193	24	21	262	90.2%	91.6%	91.0%	0.817
MET	683	51	9	259	98.7%	83.5%	94.0%	0.859
MMP13	547	48	17	254	97.0%	84.1%	92.5%	0.833
MMP2	719	70	34	347	95.5%	83.2%	91.1%	0.804
MMP3	434	43	16	219	96.4%	83.6%	91.7%	0.821

MMP9	556	84	86	381	86.6%	81.9%	84.6%	0.685
NEK2	59	7	15	268	79.7%	97.5%	93.7%	0.806
NR1I3	10	1	126	685	7.4%	99.9%	84.5%	0.233
P2RY1	117	9	6	258	95.1%	96.6%	96.2%	0.912
PAK4	80	5	5	266	94.1%	98.2%	97.2%	0.923
PDE4A	159	13	14	221	91.9%	94.4%	93.4%	0.864
PDE5A	335	31	18	257	94.9%	89.2%	92.4%	0.846
PIK3CA	1174	61	8	478	99.3%	88.7%	96.0%	0.907
PPARG	785	105	300	1672	72.4%	94.1%	85.8%	0.696
PPP1CA	4	2	11	241	26.7%	99.2%	95.0%	0.401
PPP2CA	1	0	3	275	25.0%	100.0%	98.9%	0.497
PTEN	0	4	6	481	0.0%	99.2%	98.0%	-0.010
PTPN1	269	73	65	471	80.5%	86.6%	84.3%	0.668
PTPN11	44	14	44	283	50.0%	95.3%	84.9%	0.532
PTPN13	0	0	15	262	0.0%	100.0%	94.6%	-
PTPN14	0	0	4	266	0.0%	100.0%	98.5%	-
PTPN2	53	19	28	282	65.4%	93.7%	87.7%	0.618
RAF1	333	17	1	252	99.7%	93.7%	97.0%	0.941
RARA	63	5	40	779	61.2%	99.4%	94.9%	0.729
RARB	57	1	14	823	80.3%	99.9%	98.3%	0.880
ROCK1	305	25	10	250	96.8%	90.9%	94.1%	0.882
RPS6KA5	42	7	15	256	73.7%	97.3%	93.1%	0.755
SIRT2	58	10	31	301	65.2%	96.8%	89.8%	0.686
SIRT3	35	6	8	258	81.4%	97.7%	95.4%	0.807
SRC	609	76	16	315	97.4%	80.6%	90.9%	0.810
TACR2	230	13	20	519	92.0%	97.6%	95.8%	0.902
TBXA2R	211	32	21	462	90.9%	93.5%	92.7%	0.835
TEK	187	11	12	260	94.0%	95.9%	95.1%	0.900
Average	**288**	**26**	**39**	**384**	**76.7%**	**93.2%**	**93.1%**	**0.762**

Table 3.4: Performance of the random forest models on the data sets of the additional biological targets. TP = true positives; FP = false positives; FN = false negatives; TN = true negatives; SE = sensitivity, SP = specificity; ACC = accuracy; MCC = Matthews correlation coefficient.

The structural alert models for the additional biological targets perform very well in the test sets, with mean accuracy of 92% and 91%, and mean MCC of 0.740 model and 0.708 for the risk assessment and screening models respectively. Many of the new targets have fewer data points than the Bowes targets, but this is unsurprising as the Bowes targets are much studied and widely accepted as targets of interest. Despite this, the majority of the models for the new targets perform similarly to the models of the Bowes target. The "risk assessment" and "screening" models for the new targets have a mean accuracy of 92% and 91%, and a mean MCC of 0.782 and 0.748 respectively.

Maria Folia (Unilever) has constructed random forest models for the additional targets, again using the same training and test sets as the structural alert models. A mean accuracy of 93% and mean MCC of 0.762 indicate excellent model performance in the test sets of these additional targets. The model performance is similar to the Bowes targets models (average MCC of 0.802).

The performance of the structural alert-based models and the random forest models for the new targets is similar, as it was for the Bowes targets. The main advantages of the structural alert models over random forest models is greater transparency and easily interpretable predictions, as discussed in Section 2.4.2.2.

Construction of structural alert models for the additional 66 biological targets greatly increases the scope of the work, allowing activity predictions to be made for a wider range of targets of different biological interest. The high performance of these models shows that the structural alert approach is not limited to the well-studied Bowes targets.

3.2. Explaining trends in performance

Why do the models for some targets perform poorly compared to others? In terms of test set MCC, targets which have poor performing structural alert-based models tend to also have poor performing random forest models. Is the data for these targets inherently harder to model? Is there a way of measuring how difficult data may be to model? With the Bowes targets and the additional biological targets, there is enough data to spot trends in performance statistics across different data sets and models.

Two methods were used to investigate the trends in performance. The first method quantifies the structural similarity of actives to other actives and of actives to inactives, then calculates the difference. The second method is inspired by work by Tropsha,[95] identifying activity cliffs between structurally similar chemicals. Both methods use Morgan fingerprints and Tanimoto coefficients to quantify similarity between pairs of chemicals.

3.2.1. Variation with data set size

It was initially hypothesised that variation in performance was due to size of the training set, with small data sets being harder to model due to a lack of data making it harder to spot similarities and patterns in active chemicals.

Performance of a model was quantified using the test set MCC. The results of a model with the following parameters was used: theta value 0.95, 5% maximum occurrence in inactives and lower bounds for a structural alert of two actives and one inactive. Figure 3.1 shows the variation of test set MCC with number of actives in the training set of a target.

As hypothesised, the smallest data sets give poorest model performance. The results suggest a cut-off of around 200 actives in the training set should be imposed. Below this cut-off there is simply not enough data for model construction. The cut-off of 200 training set actives has been included in Figure 3.1.

It should be noted that test sets are a third of the size of the training set, and so targets with small training sets will have very small test sets. There will be a large variance when performance is evaluated on such small test sets. However, models were still built for these targets and their performance data included to see how well the algorithms performed when using small data sets and to see if any conclusions about data set size could be made.

Above 200 actives in the training set, there is no correlation between size of data set and model performance. For example, hERG (3 617 actives in training set) and MAPK1 (4 700 actives in training set) have two of the largest data sets but produce two of the poorest performing models (test set MCCs of 0.502 and 0.454 respectively).

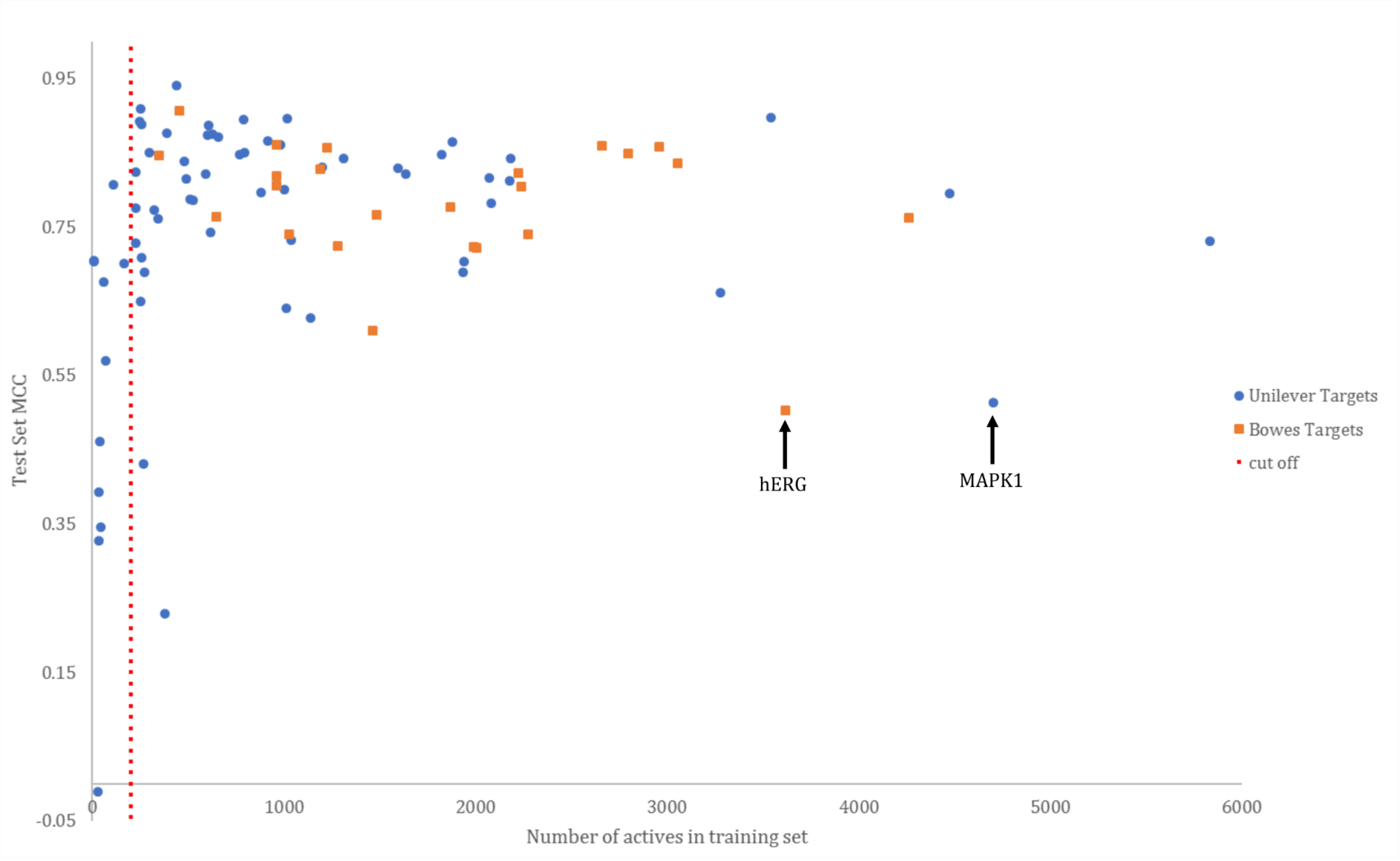


Figure 3.1: Variation of test set MCC with the number of actives in the training set of the target

3.2.2. Distinction metric

In theory, structural alerts create the best performing models when there are groups of structurally similar active chemicals which are distinctly different from the inactive chemicals. It was hypothesised that structural alerts will have difficulty modelling data where groups of structurally similar active chemicals are not structurally dissimilar from inactive chemicals. An attempt to quantify this structural distinction between groups of active chemicals and the inactive chemicals has been made here. For each active chemical, the distance in chemical space to the nearest active chemicals was calculated, as was the distance to the nearest inactive chemicals. The difference between these two values was the distinction between nearest active and inactive chemicals. This idea has been visualised in Figure 3.2.

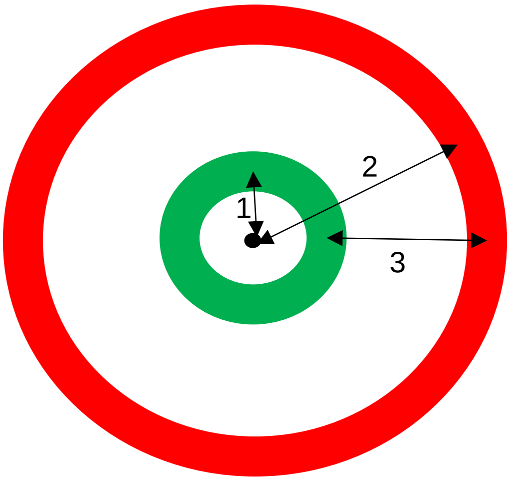


Figure 3.2: This figure is a representation of chemical space. The black dot represents an active chemical, the green band represents the most structurally similar active chemicals and the red band represents the most structurally similar inactive chemicals. The difference between active and inactive space, arrow "3", is calculated by subtracting the distance to the nearest actives, arrow "1", from the distance to the nearest inactives, arrow "2". To create high performing structural alerts, arrow "3" should be large.

3.2.2.1. Method

Morgan fingerprints were created for all chemicals within the training set, with a bit string length of 1 024 and a radius of two atoms. This was done using RDKit nodes within KNIME. For a particular active chemical, the Tanimoto similarity coefficient to every other active was calculated. The mean of the highest ten values – the ten nearest active neighbours – was calculated giving the active:active similarity. For the same active chemical, the Tanimoto coefficient to the inactive chemicals was calculated. The mean of the highest ten values – the ten nearest inactive neighbours – was calculated giving the active:inactive similarity. The calculations were repeated for each active chemical in the training set, and the means taken for both values. The mean active:inactive similarity was subtracted from the mean active:active similarity, giving a single "distinction metric" for a target's training set.

The distinction metric was calculated for the training set of all targets – the selection of 24 Bowes Targets and the additional targets of biological interest.

3.2.2.2. Results and Discussion

For each target, distinction metric was calculated from the training set data. Structural alert models were applied to the test set and MCC value calculated (structural alert models constructed by the automated workflow with parameters: theta value 0.95, 5% maximum occurrence in inactives and lower bounds for a structural alert of two actives and one inactive). These two values were plotted against each other and this graph is shown in Figure 3.3.

The cut-off of 200 training actives suggested previously was then applied to the data sets, removing the data from any targets that fall below the cut-off. The graph of test set MCC against distinction metric was replotted, shown in Figure 3.4.

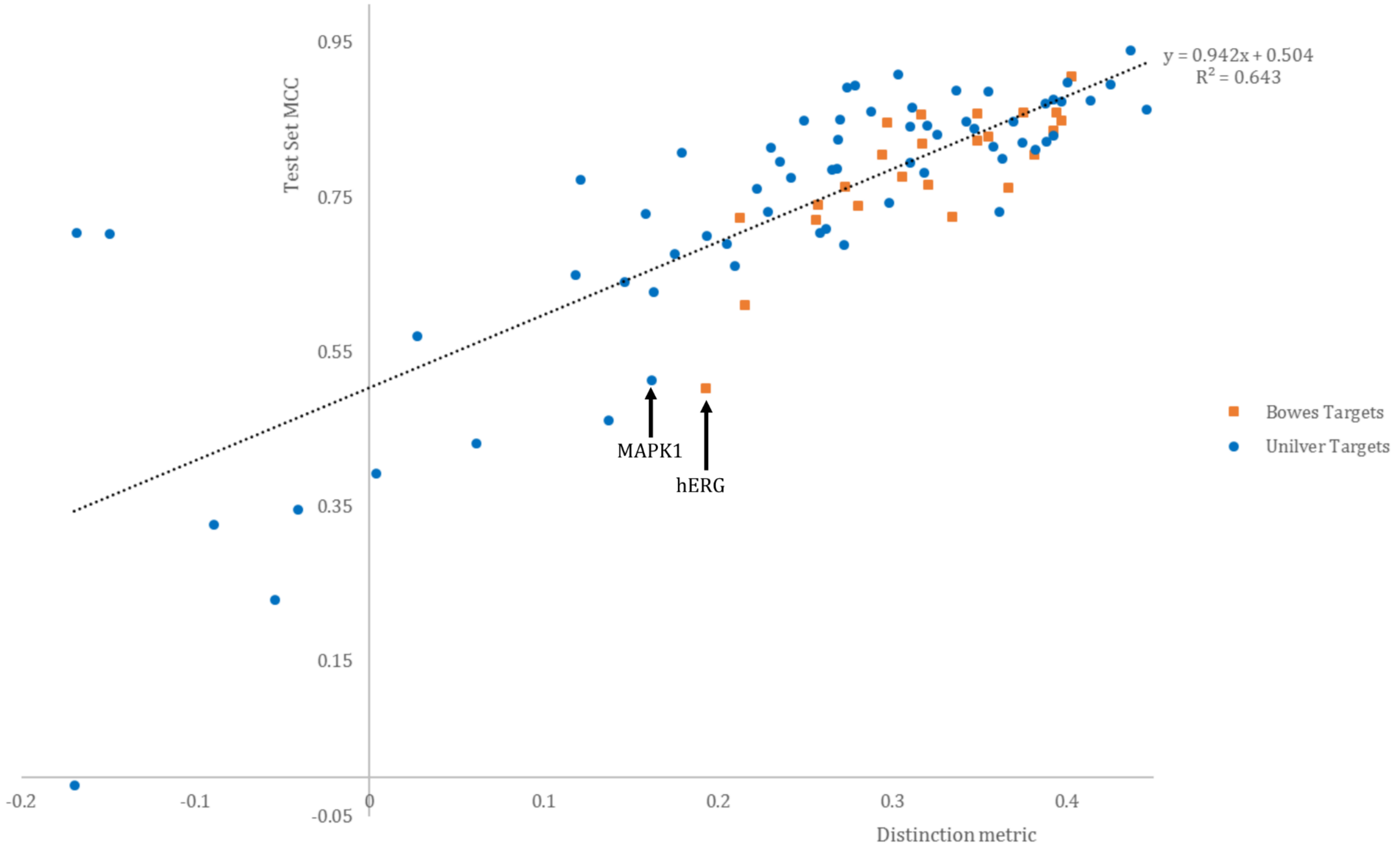


Figure 3.3: Variation of test set MCC with distinction metric.

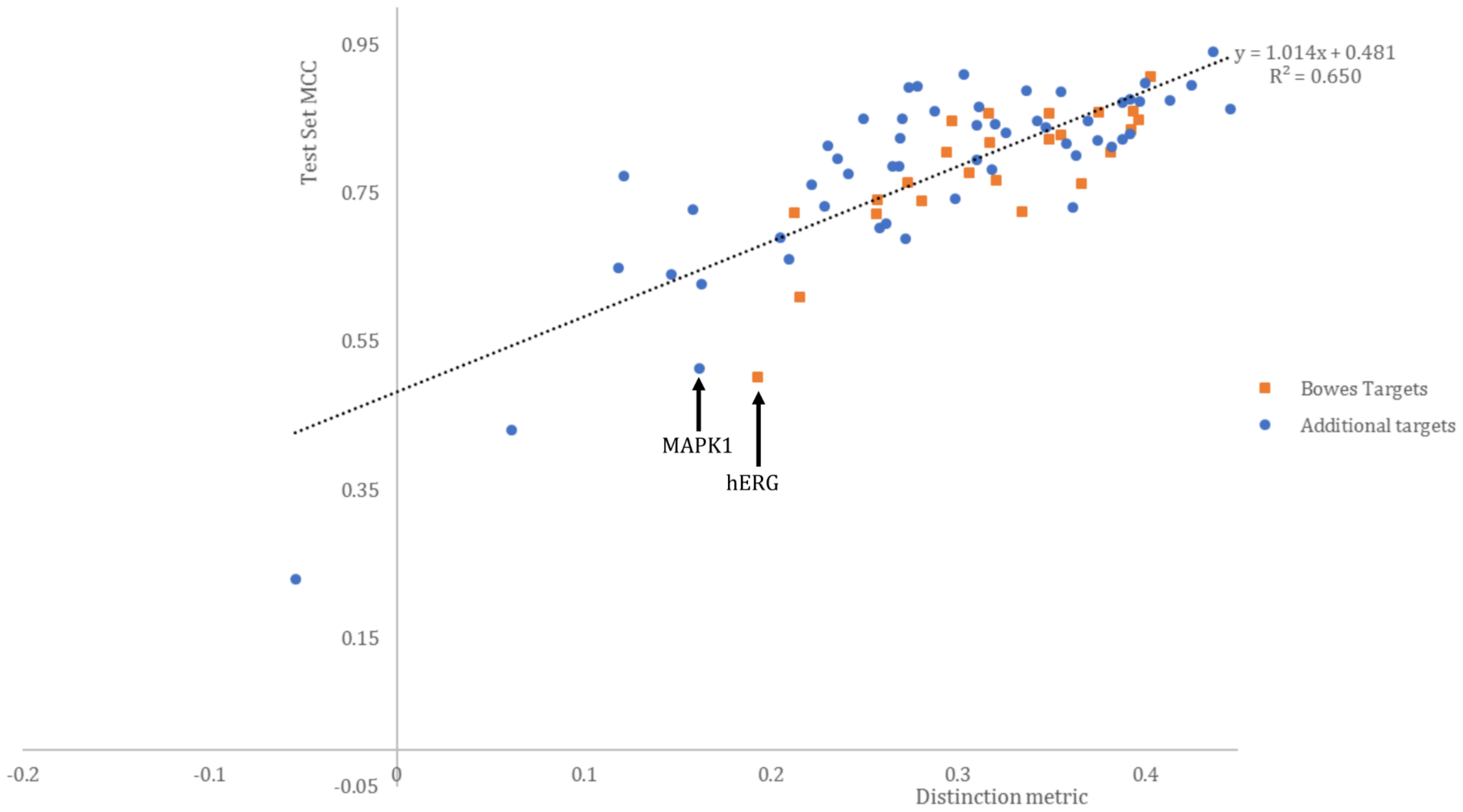


Figure 3.4: Variation of test set MCC with distinction metric, removing any targets with fewer than 200 active chemicals in the training set

There is a good correlation between test set MCC and the distinction metric, with an R^2 value of 0.634 when all targets are included. Some targets have negative values for the distinction metric – the ten nearest inactive chemicals are on average more similar to an individual than the ten nearest active chemicals – and these have test set MCCs which have the largest deviation from the line of best fit. Most of these targets have small data sets, falling below the cut-off of 200 actives in the training set, and have large variation in MCC as a result.

Removing targets which fall below the 200 actives cut-off gives a slightly better correlation, with an R^2 value of 0.636. However, the correlation appears to be weighted on a small number of targets with low distinction metric and low MCC. Despite this, the distinction metric does a fair job of explaining variation in model performance and gives far more insight than considering size of data set alone.

A potential source of error might arise from comparing to the ten nearest neighbours in all data sets. Better correlations may be achieved by comparing to a different number of nearest neighbours. This number could be a different fixed value or could scale with data set size.

Despite this possible source of error, a good correlation is seen between test set MCC and the distinction metric, indicating there is some merit in using fingerprints and structural similarity to predict model performance.

3.2.3. Dataset Modelability

Activity cliffs are very similar compounds with vastly different activities. They are often difficult to predict and cause challenges for SAR modelling.[96] Tropsha has previously proposed a modelability index (MODI) as an attempt to predict how feasible it is to create high performing QSAR models for a data set.[95] It applies a concept similar to activity cliffs, finding pairs of nearest neighbour chemicals where there is a change in activity. These nearest neighbour pairs could indicate activity cliffs or isolated chemicals for which there are no reports of similar within the data set. Both will be challenging for SAR models to predict. MODI is the proportion of nearest neighbour pairs for which there is no change in activity, calculated for the active and inactive chemicals separately and then averaged. Data sets with larger values of MODI should be less problematic to model.

Tropsha defines the nearest neighbour as the compound with "the smallest Euclidean distance from a given compound estimated in the entire descriptor space".[95] This definition may be appropriate when considering models constructed using many different descriptors, such as random forest models, but a structural-based alternative to the use of multi-dimension descriptor space would be more appropriate for the structural alert-based models. In this work, Morgan fingerprints have been generated for all compounds in a data set and the nearest neighbour to a given compound has been identified as the compound with the highest Tanimoto similarity coefficient between fingerprints.

3.2.3.1. Method

Morgan fingerprints were generated for all chemicals in a target's training set with the RDKit[83] Fingerprint node in KNIME[67] using a bit string length of 1 024 and a radius of two atoms. For each training chemical, Tanimoto similarity coefficients between these fingerprints were calculated to all other training chemicals. The chemical with the highest Tanimoto similarity was identified as the nearest neighbour.

For binary data sets like the ones used in this work, MODI is calculated as follows:

$$MODI = \frac{1}{2}(\frac{N_a^{same}}{N_a^{total}} + \frac{N_i^{same}}{N_i^{total}})$$

Where:

- N_a^{same} is the number of active compounds that have their nearest neighbour also being active compound
- N_a^{total} is the total number of actives compounds
- N_i^{same} is the number of inactive compounds that have their nearest neighbour being an inactive compound
- N_i^{total} is the total number of inactive compounds

MODI was calculated for the training sets of the selection of 24 Bowes targets and the 66 additional targets.

3.2.3.2. Results and Discussion

The variation of test set MCC from models built by the automated workflow (parameters: theta value 0.95, 5% maximum occurrence in inactives and lower bounds for a structural alert of two actives and one inactive) with MODI is shown in Figure 3.5.

When considering all targets, a strong correlation can be observed between MODI and test set MCC, with an R^2 value of 0.799. As shown in Figure 3.6, removing targets with less than 200 actives in the training set improves the correlation. The R^2 value increases to 0.884, indicating a very strong correlation between MODI and model performance.

Such a high correlation gives us confidence in explaining why poor model performance is observed for some targets. Despite hERG and MAPK1 having large data sets, they have comparatively low MODI values (0.769 and 0.708 respectively) indicating a higher proportion of pairs of nearest neighbour pair with different activity. These pairs could be activity cliffs or indicative of reports of active chemicals with no similar chemicals in the data set, although this is unlikely with such large data sets. This data is inherently harder to model, and the poor performance of models is not due to problems with the methodology of the automated workflow for constructing structural alert-based models. Better predictivity for these targets might be possible with different, more complex modelling techniques capable of accounting for activity cliffs, such as docking studies or pharmacophore modelling. However, even these may struggle with troublesome targets like hERG, which is inhibited by a wide variety of structurally dissimilar chemicals.[91] This makes constructing SARs for hERG particularly troublesome. Furthermore, the binding modes of hERG are considered "an unsolved mystery",[92] making receptor structure-based approaches to biological activity prediction (e.g. docking) as difficult as ligand-based approaches.

Importantly, the MODI calculation involves finding the proportion of activity cliffs in the active and inactive chemicals separately and then averaging the two. This allows MODI to account for imbalanced data, giving a better prediction of overall model performance.

There is a tighter correlation between test set MCC and MODI than between test set MCC and the distinction metric. As such, MODI will be used for the rest of this study, but distinction metric will not.

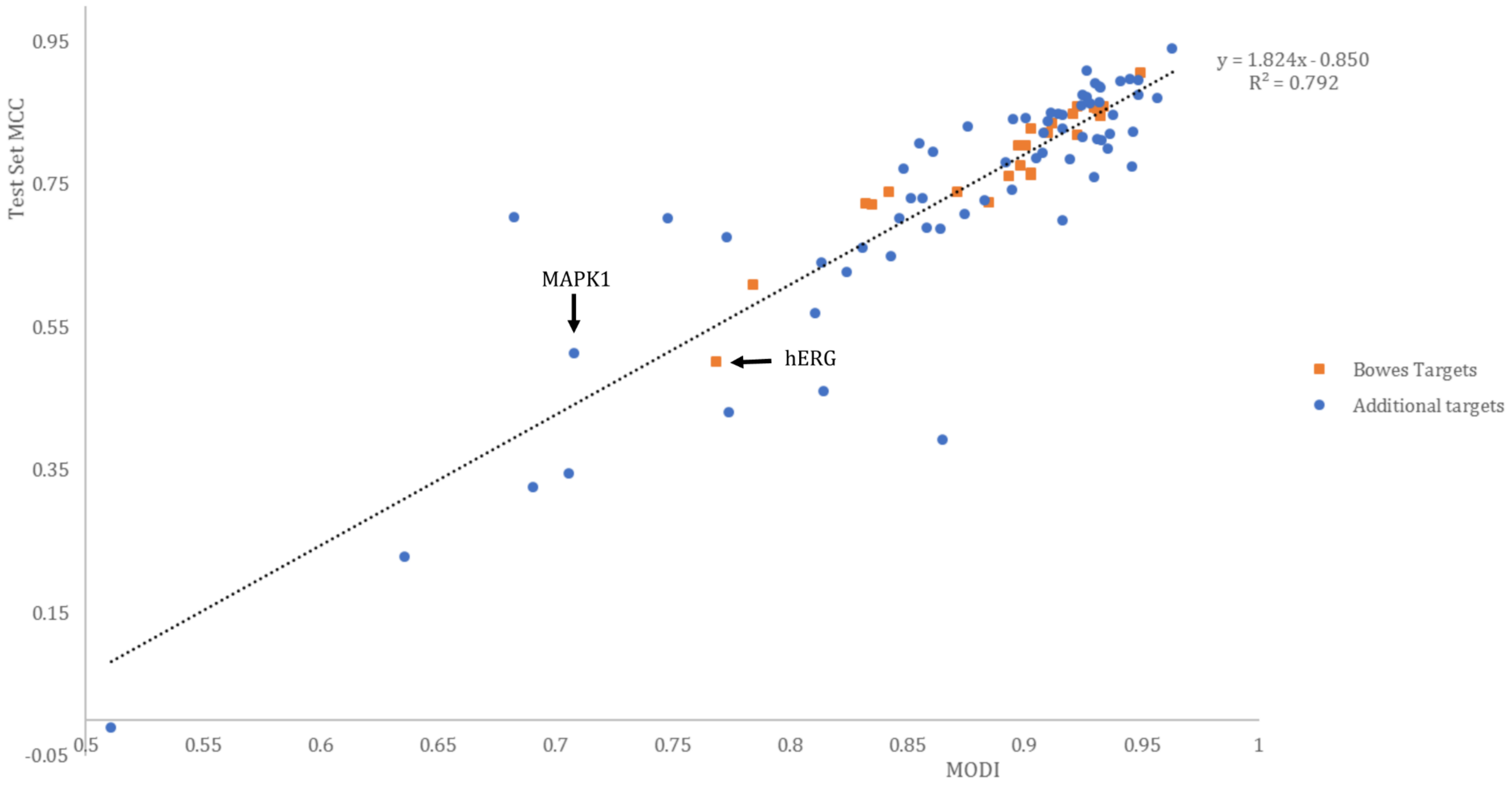


Figure 3.5: Variation of test set MCC from models built by the automated workflow (parameters: theta value 0.95, 5% maximum occurrence in inactives and lower bounds for a structural alert of two actives and one inactive) with MODI

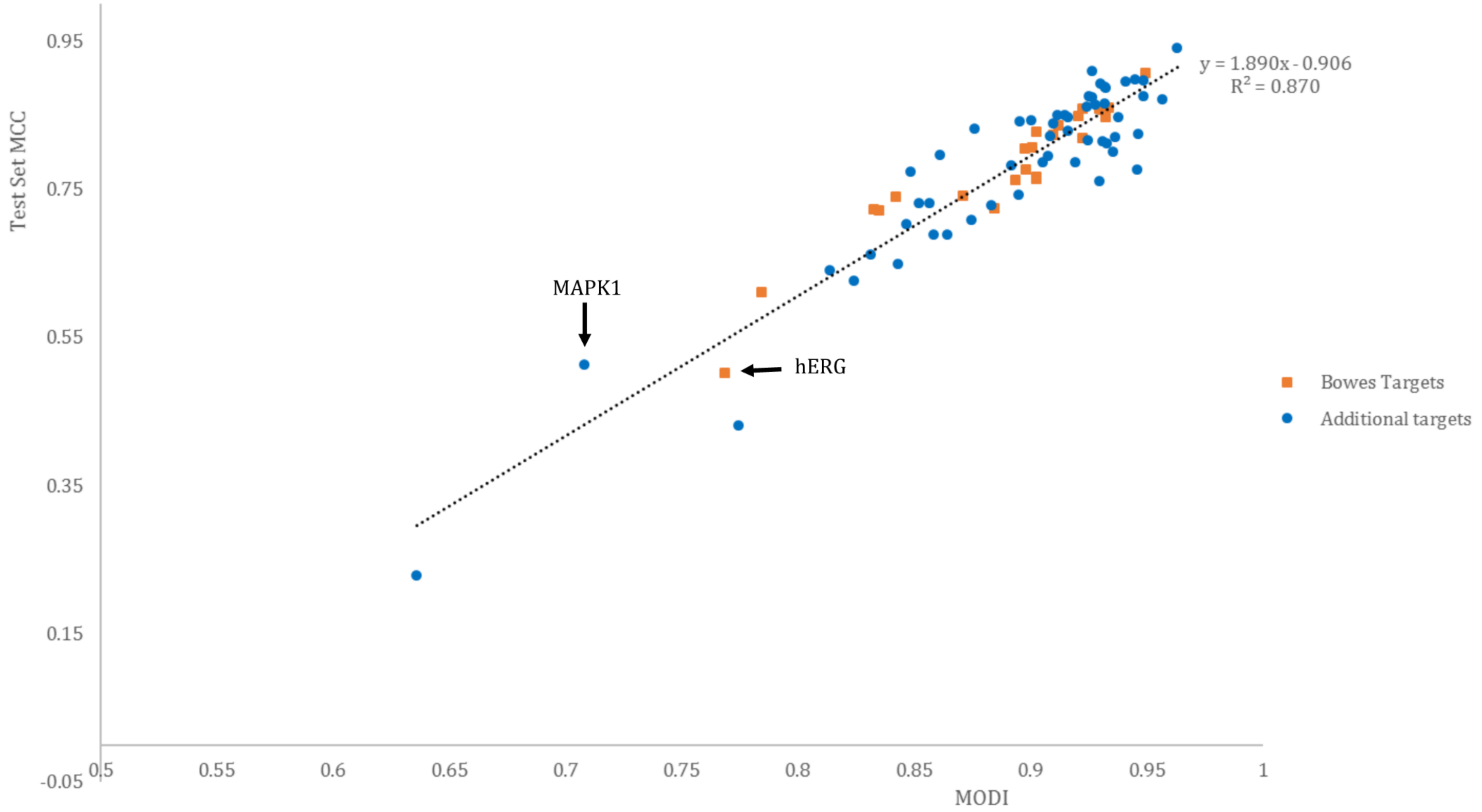


Figure 3.6: Variation of test set MCC from models built by the automated workflow (parameters: theta value 0.95, 5% maximum occurrence in inactives and lower bounds for a structural alert of two actives and one inactive) with MODI, removing targets with less than 200 active chemicals in the training set.

3.2.3.3. Random forest models

A high correlation between MODI and the performance of the random forest models is seen, as shown in Figure 3.7, and with data set with fewer than 200 training set actives removed in Figure 3.8. When only considering data sets with at least 200 training set actives, an extremely high R^2 value of 0.893 is observed for the correlation. This demonstrates the versatility of the fingerprint-based MODI, predicting the model performance for both the random forest models and the structural alert models. It also shows the similarity in performances of the structural alert-based models and the random forest models for individual biological targets.

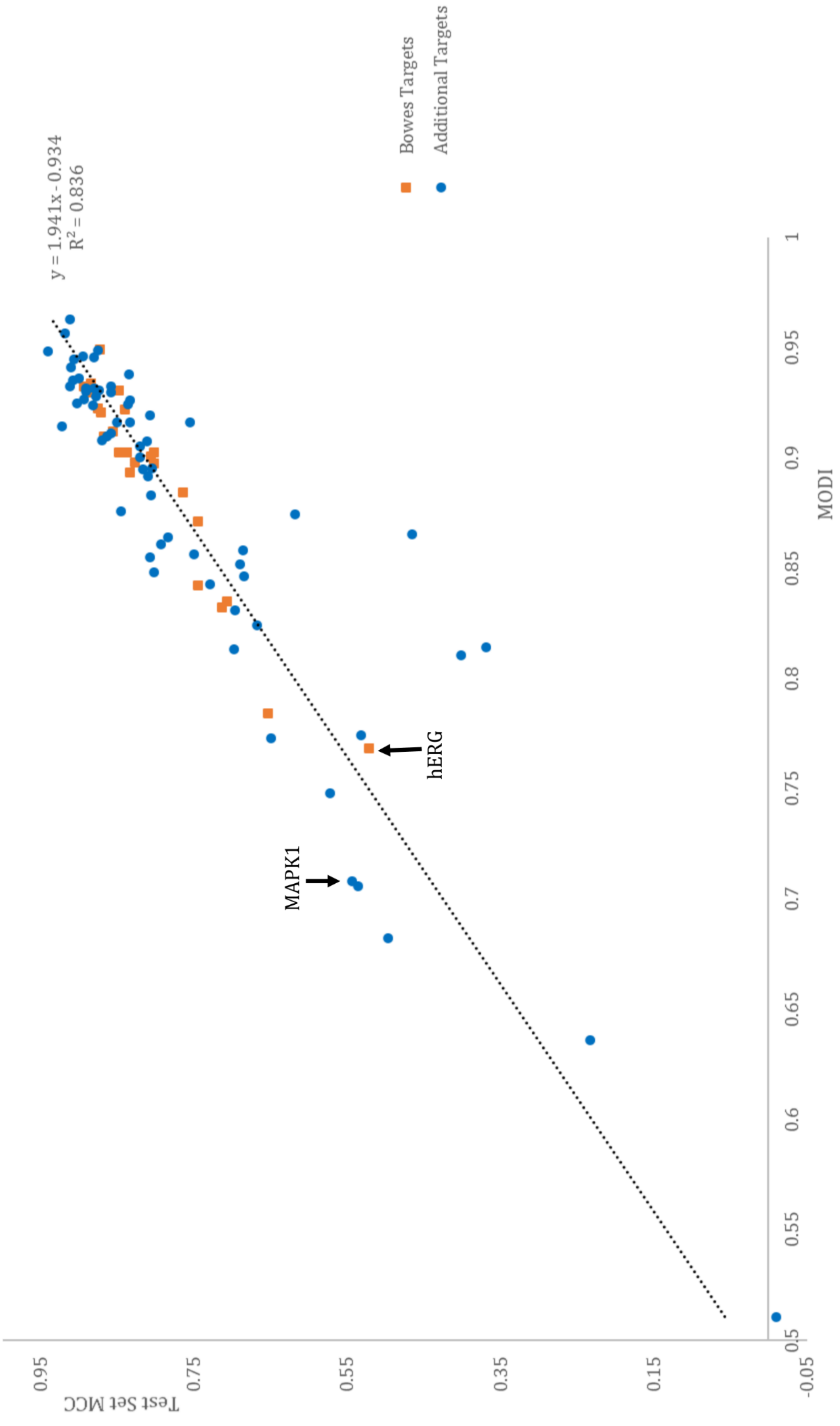


Figure 3.7: Performance of random forest models for all targets against MODI.

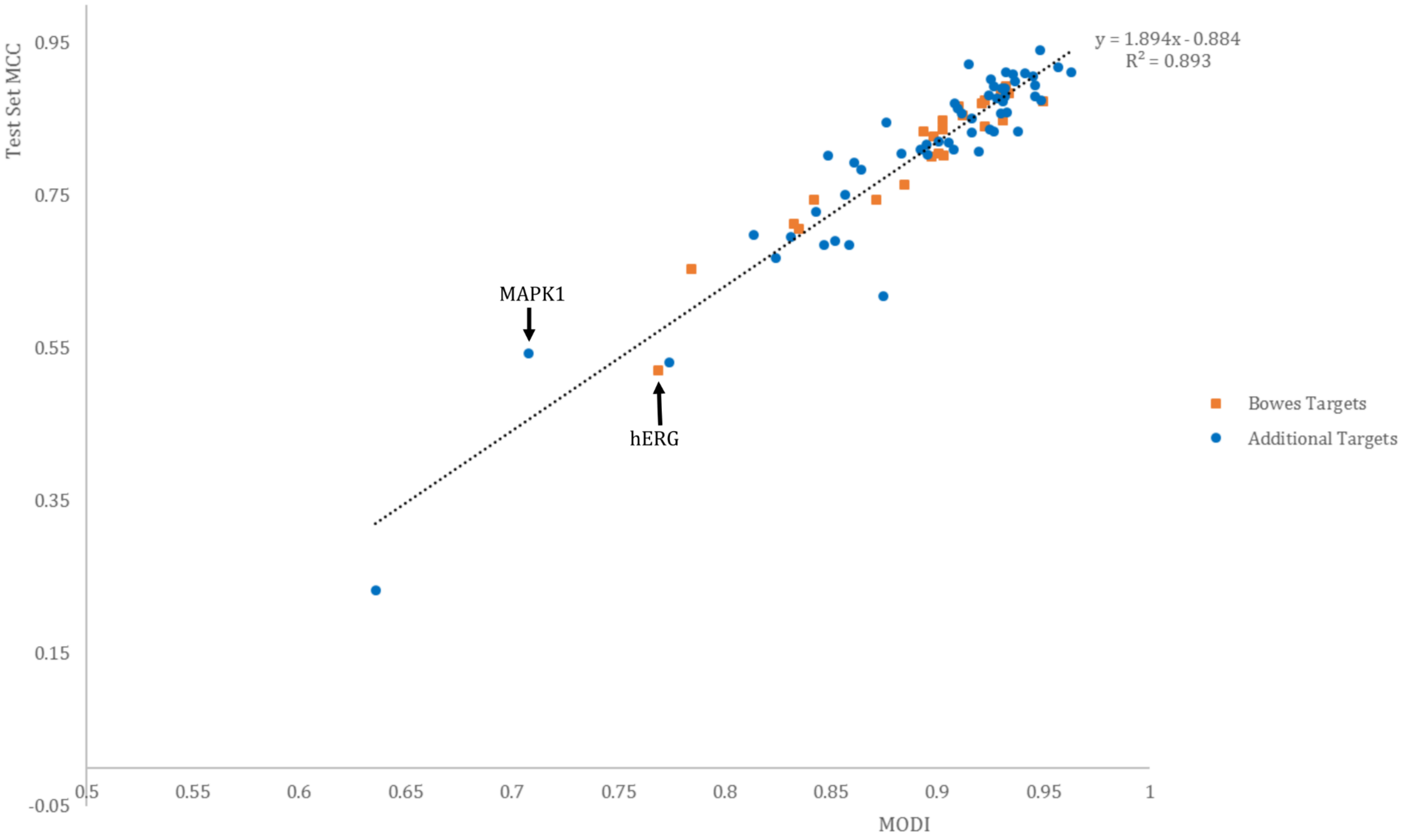


Figure 3.8: Performance of random forest models against MODI, removing targets with less than 200 actives in the training set.

3.2.3.4. Application of MODI

Having explained trends in performance of the existing models, we can now predict performance of future models before constructing them by evaluating the data set from which they will be made. Firstly, the data set should have greater than 200 active chemicals. Fingerprint-based MODI can then be calculated quickly, and model performance can be predicted from this. This is particularly valuable for larger data sets for which model construction is computationally expensive.

3.3. Bowes targets without ToxCast data

Models have been created for 24 Bowes targets using data from ChEMBL and ToxCast. There is no human data in ToxCast for the remaining Bowes targets, but they are still important MIEs. Despite there being fewer data points available for these targets, activity predictions would still be useful.

3.3.1. Data Sets

For each target, Homo sapiens bioactivity data was downloaded from ChEMBL (data extracted November 2018). Activity reports were filtered to remove any with a confidence score of less than eight. Only activities reported with Standard Units of nM were kept, leaving reports of EC50, IC50, K_i and K_d. RDKit[83] Salt Stripper was used to remove common salts and counter ions from chemicals. All chemicals with more than 100 atoms were removed. The SMILES strings were re-written to be canonical using RDKit, such that the format was consistent across all reports.

Where chemicals had exact values of activity, means of the values were taken. Chemicals with a mean activity of 10 000 nM or lower were assigned as active; those with over 10 000 nM were assigned as inactive. Some values of activity are reported as “greater than” a certain value instead of as exact values. In these cases, if the activity for a chemical was reported as greater than 10 000 nM and the chemical had no other reports of activity, the chemical was assigned as inactive.

Data points were split randomly, with 75% forming a training set and 25% forming a test set.

A summary of the data sets is shown in Table 3.5. For three of the targets, no human data with a minimum confidence score of eight was found. Of the seventeen targets with data, three have fewer training actives than the cut-off of 200 suggested previously. Most of the data sets have significant imbalances in data, with many more actives than inactives.

Target	Training Actives	Training Inactives	Test Actives	Test Inactives
Acetylcholine receptor subunit α1 or α4	0	0	0	0
Alpha-1a adrenergic receptor	1217	101	418	27
Calcium Voltage-Gated Channel Subunit Alpha1 C	81	106	46	44
Cannabinoid CB1 receptor	3157	1738	1024	570
Cannabinoid CB2 receptor	4204	846	1473	302
Cholecystokinin A receptor	308	10	101	4
Cyclooxygenase-1	553	1262	207	411
Cyclooxygenase-2	1751	1004	597	364
GABAA receptor α1 (rat cortex) BZD site	0	0	0	0
Glutamate (NMDA) Receptor	0	0	0	0
Histamine H2 receptor	228	184	69	65
Kappa opioid receptor	3068	521	1040	178
Monoamine oxidase A	771	997	270	351
Phosphodiesterase 3A	155	143	62	36
Phosphodiesterase 4D	449	126	155	48
Serotonin 1a (5-HT1a) receptor	2686	217	978	59
Serotonin 1b (5-HT1b) receptor	799	82	253	34
Serotonin 2b (5-HT2b) receptor	1128	70	422	18
Sodium channel protein type V alpha subunit	355	816	124	243
Voltage-gated potassium channel subunit Kv7.1	25	32	5	7

Table 3.5: The Bowes targets which have no human data available in ToxCast. Human in vitro *data is extracted from ChEMBL and randomly split, with roughly 75% of chemicals forming the training set and the remaining 25% forming the test set.*

3.3.2. Results and Discussion

Using the data sets extracted from ChEMBL, structural alert-based models have been constructed using the automated workflow with “Risk Assessment” parameters (theta 0.95, 1% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive) and with “Screening” parameters (theta 0.51, 15% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive).

The results of the Risk Assessment model are shown in Table 3.6. MODI values have been calculated for each training set and, using the best fit line between test set MCC and MODI in the other targets, test set MCCs predicted.

Three of the targets have fewer training actives than the cut-off of 200 so would not be expected to form good models. The other data sets are mostly imbalanced, with many more active chemicals than inactive chemicals. In data sets with few inactives, falsely predicting the activity of the inactive chemicals will lead to very low specificity. As MCC is designed to account for imbalances in data, false positive predictions in data sets with few inactive chemicals results in large decreases in MCC. Cholecystokinin A receptor is an example of a such a target. Despite correctly predicting 95% of the active chemicals in the test set, three of the four inactive chemicals are falsely predicted to be active, leading to a very low MCC value of 0.165.

The results of the Screening model are shown in Table 3.7. This model is designed to increase sensitivity at the expense of specificity. In data sets with few inactive chemicals, the decreased specificity has a larger effect on MCC values than the increase in sensitivity. Hence, lower MCC values are seen in the test set of each of the biological targets and the mean MCC decreases compared to the Risk Assessment models.

	Training Set					Test Set			
Target	**Actives**	**Inactives**	**MODI**	**MODI predicted MCC**	**Alerts**	**SE**	**SP**	**ACC**	**MCC**
Alpha-1a adrenergic receptor	1217	101	0.78	0.557	32	94.0%	44.4%	91.0%	0.333
Calcium Voltage-Gated Channel Subunit Alpha1 C	81	106	0.793	0.583	11	52.2%	93.2%	72.2%	0.495
Cannabinoid CB1 receptor	3157	1738	0.858	0.711	172	84.4%	79.1%	82.5%	0.626
Cannabinoid CB2 receptor	4204	846	0.857	0.708	140	92.9%	70.9%	89.1%	0.624
Cholecystokinin A receptor	308	10	0.518	0.043	7	95.0%	25.0%	92.4%	0.165
Cyclooxygenase-1	553	1262	0.709	0.417	94	41.1%	94.6%	76.7%	0.445
Cyclooxygenase-2	1751	1004	0.786	0.570	161	73.9%	73.9%	73.9%	0.467
Histamine H2 receptor	228	184	0.796	0.589	32	65.2%	81.5%	73.1%	0.473
Kappa opioid receptor	3068	521	0.806	0.608	103	92.3%	68.0%	88.8%	0.574
Monoamine oxidase A	771	997	0.798	0.592	97	57.8%	92.6%	77.5%	0.549
Phosphodiesterase 3A	155	143	0.802	0.600	16	67.7%	94.4%	77.6%	0.603
Phosphodiesterase 4D	449	126	0.826	0.647	28	81.3%	91.7%	83.7%	0.646
Serotonin 1a (5-HT1a) receptor	2686	217	0.837	0.670	50	96.6%	35.6%	93.2%	0.336
Serotonin 1b (5-HT1b) receptor	799	82	0.832	0.660	26	95.3%	67.6%	92.0%	0.621
Serotonin 2b (5-HT2b) receptor	1128	70	0.723	0.445	42	93.1%	16.7%	90.0%	0.075
Sodium channel protein type V alpha subunit	355	816	0.797	0.590	51	54.8%	94.7%	81.2%	0.564
Voltage-gated potassium channel subunit Kv7.1	25	32	0.762	0.522	5	20.0%	71.4%	50.0%	-0.098
Average	**1232**	**486**	**0.78**	**0.560**	**62.8**	**74.0%**	**70.3%**	**81.5%**	**0.441**

Table 3.6: Performance of structural alert-based models for Bowes targets using ChEMBL data only. Models were built by the automated workflow with the "Risk Assessment" parameters: theta 0.95, 1% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive. "MODI predicted MCC" has been calculated by fitting MODI values to the best-fit line of Figure 3.6.

	Training Set					Test Set			
Target	**Actives**	**Inactives**	**MODI**	**MODI predicted MCC**	**Alerts**	**SE**	**SP**	**ACC**	**MCC**
Alpha-1a adrenergic receptor	1217	101	0.78	0.557	26	98.1%	7.4%	92.6%	0.088
Calcium Voltage-Gated Channel Subunit Alpha1 C	81	106	0.793	0.583	15	69.6%	58.1%	64.0%	0.279
Cannabinoid CB1 receptor	3157	1738	0.858	0.711	125	95.4%	31.3%	73.0%	0.368
Cannabinoid CB2 receptor	4204	846	0.857	0.708	78	98.2%	20.9%	85.4%	0.325
Cholecystokinin A receptor	308	10	0.518	0.043	12	94.1%	25.0%	91.4%	0.146
Cyclooxygenase-1	553	1262	0.709	0.417	138	86.0%	36.8%	53.3%	0.237
Cyclooxygenase-2	1751	1004	0.786	0.570	123	95.3%	30.6%	70.8%	0.357
Histamine H2 receptor	228	184	0.796	0.589	36	79.7%	38.1%	59.8%	0.196
Kappa opioid receptor	3068	521	0.806	0.608	62	97.9%	16.7%	86.6%	0.253
Monoamine oxidase A	771	997	0.798	0.592	109	91.5%	26.7%	55.2%	0.232
Phosphodiesterase 3A	155	143	0.802	0.600	19	90.3%	38.9%	71.4%	0.349
Phosphodiesterase 4D	449	126	0.826	0.647	21	91.6%	50.0%	81.8%	0.458
Serotonin 1a (5-HT1a) receptor	2686	217	0.837	0.670	44	99.1%	8.6%	94.0%	0.153
Serotonin 1b (5-HT1b) receptor	799	82	0.832	0.660	17	98.4%	51.5%	93.0%	0.612
Serotonin 2b (5-HT2b) receptor	1128	70	0.723	0.445	29	96.9%	5.6%	93.2%	0.028
Sodium channel protein type V alpha subunit	355	816	0.797	0.590	56	86.3%	61.5%	70.1%	0.458
Voltage-gated potassium channel subunit Kv7.1	25	32	0.762	0.522	5	20.0%	71.4%	50.0%	-0.098
Average	**1232**	**486**	**0.781**	**0.560**	**53.8**	**87.5%**	**34.1%**	**75.6%**	**0.261**

Table 3.7: Performance of structural alert-based models for Bowes targets using ChEMBL data only. Models were built by the automated workflow with the "Screening" parameters: theta 0.51, 15% maximum occurrence of an alert in the inactive chemicals, lower bounds for an alert of two actives and one inactive. "MODI predicted MCC" has been calculated by fitting MODI values to the best-fit line of Figure 3.6.

In selecting which substructure should be a structural alert, the calculation of Bayes Factor uses absolute values of number of inactive (and active) chemicals containing the substructures, not the proportion of total inactives. Hence, a theta value which produces models with high MCC values in balanced data sets is unlikely to produce models with high MCC values in data sets with few inactive chemicals. A larger theta value could be used to create a model which hits fewer false positives in training and would likely have a higher test set MCC, but this would be overfitting the model to limited data. The best way to improve the performance of the models in these data sets is not to tune the parameters to avoid the small number of inactives and give the highest possible MCC, but to increase the number of inactives in the training and test sets.

Even when ignoring targets with fewer than 200 actives or with particularly large excesses of active chemicals compared to inactive chemicals, poor test set MCC values are seen. These values are, however, similar to the values predicted by calculating MODI, suggesting that even with imbalanced data, MODI does a good job of predicting test set MCC. The relatively low MODI values indicate a higher proportion of nearest neighbour pairs with differing activity in these data sets compared to the data sets for targets with data from both ChEMBL and ToxCast. It should be noted that the MODI calculation involves calculating proportion of nearest neighbour pairs with differing activity in both the actives and the inactive separately, then averaging the two values. Thus, in imbalanced data sets with large excesses of active chemicals, the small number of inactive chemicals can have large contributions to MODI values, as well as to MCC values. With few inactive chemicals and many active chemicals, it is more likely that an inactive chemical's nearest neighbour is an active chemical, resulting in low MODI.

Hence, the best way to improve MCC and MODI values, and to judge the performance of these models more fairly, is to use data sets with a good balance of active and inactive data.

Based on their models, Rosenkranz and Cunningham concluded that SAR models can tolerate imbalance in data between 33% actives and 75% actives.[97] These cut-offs were derived using a fragment-based machine learning method so may be relevant to the models created here, although Rosenkranz and Cunningham's models were applied to mutagenicity predictions. When these cut-offs are applied to the models generated here, along with the minimum 200 training actives cut-off, only five targets remain: cannabinoid CB1 receptor, cyclooxygenase-2, histamine H2 receptor, monoamine oxidase A and phosphodiesterase 3A. The average performance across the Risk Assessment models of these five targets is better than average across all targets, with a mean accuracy of 77.0% and a mean MCC of 0.544.

The results for these data sets demonstrate the importance of using balanced data sets (in terms of number of active and inactive chemicals) in model construction.

3.4. Conclusions

The automated workflow has been used to build new structural alert-based models for 66 biological targets which are not Bowes targets. The models generally perform extremely well. In the "risk assessment" models, on average 92.2% of predictions are correct and a mean MCC of 0.740 is seen across the targets. These results are similar to the performance of random forest models, which, averaging across all targets, correctly predict 93.1% of chemicals and have a mean MCC of 0.762. The main advantage of the structural alert-based models over random forest models is that the predictions are considerably easier to interpret, allowing the user to understand why the prediction has been made. By constructing models for a larger number of biological targets, all identified by Unilever to be important MIEs in risk assessment, the scope of the project has been greatly expanded beyond the Bowes Targets.

The variation in performance of the models for different targets cannot be explained by only considering the size of the data sets. Performance has been explained in terms of two metrics: a distinction metric which quantifies similarity of active chemicals to other active chemicals and to inactive chemicals, and a MODI based on proportion of activity cliffs in the data, inspired by work by Tropsha.[95] MODI correlates extremely well with test set MCC, giving a linear relationship with a R^2 value of 0.88. Being able to explain variation in performance in terms of a property of the data sets provides additional confidence in the model construction methods. The poor performance of both structural alert-based models and random forest models for some targets (e.g. hERG and MAPK1) is due to a high proportion of activity cliffs in the data, not due to some random error in the models.

The correlation between MODI and model performance also allows the user to quickly predict how good a model will be before beginning the computational expensive process of model construction.

Models have been created for seventeen of the Bowes targets which lack human data in ToxCast, but the remaining three Bowes targets contain no data suitable for this study in either ChEMBL or ToxCast. Data sets have been curated from ChEMBL data only, but many of these are imbalanced, with many more actives than inactives. The automated workflow for construction of structural alert-based models has been applied to these data sets, but the resultant models perform relatively poorly on the test sets. The poor performance statistics are often a result of a lack of inactive data points in the test set. In these cases, falsely predicting the activity of a small number of inactives results in large drops in specificity and MCC. This is particularly problematic in the "screening" models, which are designed to have large sensitivities at the cost of lower specificity, resulting in a very low average MCC of 0.261. The "Risk Assessment" models, designed to have a higher specificity, have a better overall performance with an average MCC of 0.441. An even greater MCC could be obtained by changing the parameters of the workflow, but this would be overfitting the models. The results obtained for these models highlight the importance of using data balanced in terms of numbers of actives and inactives to construct models.

4. Confidence in negative predictions

4.1. Introduction

In risk assessment, a false negative prediction can be the most dangerous type of erroneous prediction as it could potentially lead to hazardous chemicals being exposed to consumers. Hence, confidence in the negative predictions of *in silico* models is of vital importance.

Structural alert models predict activity by identifying substructures that are common to active chemicals. Chemicals that contain a structural alert are predicted to be active and chemicals containing no structural alerts are predicted to be inactive. However, is absence of a structural alert enough to predict inactivity with confidence?

If a new chemical is predicted active by the structural alert-based model, the confidence in the active prediction can be assessed by looking at the structural alert(s) contained by the chemical. As shown previously in section 2.6, one can look at the structure of the chemicals in the training set that contain the structural alert and consider how similar the new chemical is to the relevant training chemicals. One would have more confidence if the new chemical is similar to the active training chemicals containing the alert and dissimilar to the inactives.

If a new chemical contains no structural alerts, it is predicted inactive but no further information is given with the current models. This new chemical could be an inactive chemical with features that the model has seen in the training sets when building alerts, but it could also be an active chemical with new features that structural alerts would not recognise. Simply reporting absence of structural alert gives us no confidence in distinguishing between these two scenarios. Can more information be provided about negative predictions so that the user can have more trust in the prediction?

4.1.1. Lhasa methods

Derek Nexus[42] is an expert knowledge-based software designed by Lhasa.[98] It gives predictions for toxicity endpoints including bacteria mutagenicity and skin sensitisation using structural alerts created by experts from literature knowledge, public data and proprietary data. Bacteria mutagenicity and skin sensitisation have reactivity-driven MIEs for which presence of electrophilic groups leads to activity. This is the fundamental difference to the structural alert-based models constructed in this work, which predict receptor binding MIEs. To be active for a receptor binding MIE, a particular combination of features, such as hydrogen bond donors, hydrogen bond acceptors, ionisable or charged groups, aromatic rings, hydrophilic and hydrophobic groups, are required in a particular three-dimensional geometry. Despite being built for different types of MIEs, both models use structural alerts, and so ideas used to improve the Lhasa models could be applied to the models in this work.

Williams *et al*[76] have applied two methods to the Derek Nexus bacteria mutagenicity models to improve confidence in negative predictions: exclusion rules and classification of features.

Each structural alert has its own exclusion rules. These are features which negate the activity prediction of the structural alert – if a chemical contains the structural alert but also contains an exclusion rule, it is predicted to be inactive. This differs from a standard negative prediction which is due to absence of any structural alerts. The negative predictions due to exclusion rules have been made by expert knowledge, rather than absence of an alert. Williams *et al* investigated the difference between confidence in the two negative predictions and concluded that the exclusion rules should not be used to improved confidence in negative predictions for their mutagenicity alerts. They found no improvement in negative predictivity in inactive predictions due to exclusion rules than in inactive predictions due to lack of alert alone.

The features (structural fragments) of a new chemical containing no structural alerts are evaluated and compared to features seen in the training chemicals. If the new chemical contains features that are often present in known false negative compounds, it is considered to have "misclassified" features. If the new chemical contains features that are not present in the library of structural features, it is considered to have "unclassified" features. Chemicals with no misclassified or unclassified features were found to have the largest negative predictivity (NPV) – the proportion of predicted negatives which are true negatives – and hence increased confidence in negative predictions. Lhasa has included these categories of negative predictions into Derek Nexus. Building on the work by Williams *et al,* Chilton *et al* have also applied these categories of negative predictions to skin sensitisation,[99] which, like mutagenicity, is caused by reactivity driven MIEs.

In this chapter, methods inspired by those used by Lhasa for reactivity driven MIEs have been applied to the structural alerts for the receptor binding MIEs developed in this project. The aim of this work is to provide more confidence in negative predictions.

4.2. Structural alert exclusion rules

The structural alerts for mutagenicity in Derek Nexus each have a set of exclusion rules which overrule active predictions in certain circumstances. For example, chemicals containing epoxides are generally mutagenetic, and so a structural alert exists for them. However, tri- or tetra-substituted epoxides do not lead to mutagenicity because there is too much steric bulk around the epoxide, blocking the approach of any nucleophile. As a result, an exclusion rule for the epoxide structural alert is written for tri- and tetra-substituted epoxides. This example is shown in Figure 4.1.

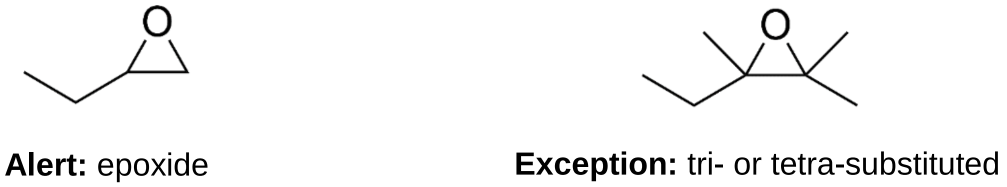


Figure 4.1: Left: a structural alert created by Lhasa for mutagenicity. Right: an "exception" to the alert. Tri- or tetra-substituted epoxides are too sterically hindered to be reactive and so are generally not active in the Ames test.

The idea of exclusion rules has been applied to the structural alerts created by the automated workflow for the Bowes targets. Features which appear to negate the activity prediction of a structural alert were identified by finding additions to the structural alert that occur in multiple inactive chemicals and no active chemicals. These are coded as exclusion rules. An example is shown in Figure 4.2.

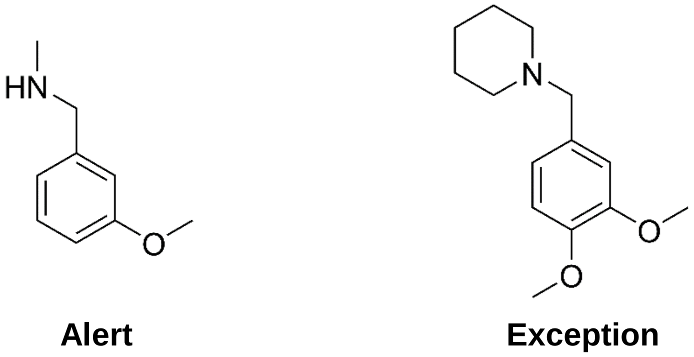


Figure 4.2. Left: a structural alert for acetylcholinesterase, contained by 151 true positives and 74 false positives. Right: a proposed "exception" to the alert, contained by six false positives and no true positives in the training set.

In Derek Nexus, exclusion rules are created based on expert knowledge and with clear explanations. The exclusion rules are derived from the explanations. In contrast, the exclusion rules for the structural alerts for receptor binding MIEs are derived directly from observations of the data. Clear explanations are not provided with these exclusion rules. To provide a clear explanation of the exclusion rules, knowledge of how the chemicals containing an alert bind to the receptor is required.

Even without a clear explanation of mechanism, the exclusion rules provide a rationale behind the negative predictions that is not present when making predictions from lack of structural alerts alone.

In this work, exclusion rules have been constructed for the structural alerts for receptor binding. The performance of negative predictions due to the exclusion rules was then compared to the performance of negative predictions due to absence of any alerts.

4.2.1. Method

All chemicals in the training set of a biological target containing a particular structural alert were found. The MoSS[100] node in KNIME[67] was applied to these chemicals to find substructures which occur in at least two of the alert-containing inactive chemicals and no active chemicals. Only substructures directly containing the positive structural alert were kept and were coded as "exclusion rules" to the alert. This requirement meant exception rules will be directly related to the structural alert itself. This process of building exclusion rules was repeated for all structural alerts of the target.

If a chemical contained the structural alert but also contained an exclusion rule to that alert, it was not predicted to be active.

Seven targets were chosen to be used in this study. These targets were acetylcholinesterase, androgen receptor, beta-2 adrenergic receptor, dopamine D2 receptor, glucocorticoid receptor, hERG and mu opioid receptor. These targets contained the greatest number of false positives, meaning there was more data from which exclusion rules can be constructed.

Two sets of structural alerts created by the automated workflow with two different sets of parameters have been included for each target:

1. Theta value of 0.95, 5% maximum occurrence of a structural alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. These parameters result in a model with high specificity and structural alerts that are generally more specific.
2. Theta value of 0.51, 5% maximum occurrence of a structural alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive. These parameters result in a model with higher sensitivity but lower specificity.

For each target and for each set of structural alerts, exclusion rules have been constructed.

4.2.2. Results and discussion

The structural alert models with exclusion rules have been applied to the test sets of the relevant targets and the performance of the models is shown in Table 4.1. In all models, the number of false positives in the test set decreases, and the specificity increases. In most models, the number of true positives also decreases, leading to a decrease in sensitivity. Despite being trained from only false negatives in the training set, the exclusion rules affect both true negatives and false negatives in the test sets of some models. This shows that the exclusion rules are not completely accurate. Despite increasing specificity in all models, the concurrent decrease in sensitivity results in only small increases to test set MCC, and in some cases, decreases.

A direct comparison of the effect on only negative predictions is shown in Table 4.2. Only a small proportion of negative predictions come from the exclusion rules. Structural alerts have been built using Bayesian statistics to cover active chemicals whilst avoiding inactive chemicals. As a result, most alerts have few false positives in the training set from which to build exclusion rules. Models built using a higher value for the theta parameter will contain structural alerts that are generally larger in size, more specific and are contained by fewer false positives. These will be harder to build exclusion rules for. Hence, there are generally fewer predictions due to exclusion rules for models built for the same target when using a theta value of 0.95 than when using a theta value of 0.51.

For most models, the negative predictive value (NPV) – proportion of negative predictions which are true negatives – for negative predictions from exclusion rules is lower than the NPV for negative predictions due to lack of alerts alone. It can be concluded that there is no additional confidence in the negative predictions due to exclusion rules than in the negative predictions due to lack of structural alerts alone. The predictions by exclusions rules are for chemicals which are very similar to active chemicals – all of these chemicals contain a substructure which has been identified as statistically likely to be contained by active chemicals. Predicting the activity of chemicals which lie on the border between activity and inactivity is inherently difficult, and so there should not be much confidence in the negative predictions from exclusion rules. The same conclusion was reached by Williams *et al* in their work on exclusion rules for structural alerts for mutagenicity.[76]

		Alerts Only								Alerts with exclusion rules								Change			
Target	Theta	TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC	ΔSE	ΔSP	ACC	ΔMCC
Acetylcholinesterase	0.51	555	157	55	339	0.910	0.683	0.808	0.616	545	121	65	375	0.893	0.756	0.832	0.660	-0.016	0.073	0.024	0.044
Acetylcholinesterase	0.95	517	61	93	435	0.848	0.877	0.861	0.721	514	60	96	436	0.843	0.879	0.859	0.718	-0.005	0.002	-0.002	-0.003
Androgen receptor	0.51	442	46	206	1769	0.682	0.975	0.898	0.725	442	41	206	1774	0.682	0.977	0.900	0.731	0.000	0.003	0.002	0.006
Androgen receptor	0.95	436	42	212	1773	0.673	0.977	0.897	0.723	436	39	212	1776	0.673	0.979	0.898	0.727	0.000	0.002	0.001	0.004
Beta-2 adrenergic receptor	0.51	376	86	107	427	0.778	0.832	0.806	0.612	374	83	109	430	0.774	0.838	0.807	0.614	-0.004	0.006	0.001	0.002
Beta-2 adrenergic receptor	0.95	358	70	125	443	0.741	0.864	0.804	0.611	356	66	127	447	0.737	0.871	0.806	0.615	-0.004	0.008	0.002	0.005
Dopamine D2 receptor	0.51	1391	75	43	207	0.970	0.734	0.931	0.740	1364	63	70	219	0.951	0.777	0.922	0.721	-0.019	0.043	-0.009	-0.019
Dopamine D2 receptor	0.95	1382	59	52	223	0.964	0.791	0.935	0.762	1378	53	56	229	0.961	0.812	0.936	0.770	-0.003	0.021	0.001	0.008
Glucocorticoid receptor	0.51	556	85	188	1645	0.747	0.951	0.890	0.731	548	66	196	1664	0.737	0.962	0.894	0.741	-0.011	0.011	0.004	0.011
Glucocorticoid receptor	0.95	539	59	205	1671	0.724	0.966	0.893	0.739	538	57	206	1673	0.723	0.967	0.894	0.741	-0.001	0.001	0.000	0.001
hERG	0.51	1125	441	153	384	0.880	0.465	0.718	0.387	1108	418	170	407	0.867	0.493	0.720	0.394	-0.013	0.028	0.003	0.007
hERG	0.95	882	145	396	680	0.690	0.824	0.743	0.502	882	142	396	683	0.690	0.828	0.744	0.506	0.000	0.004	0.001	0.004
Mu opioid receptor	0.51	907	93	43	492	0.955	0.841	0.911	0.811	905	61	45	524	0.953	0.896	0.931	0.853	-0.002	0.055	0.020	0.042
Mu opioid receptor	0.95	889	42	61	543	0.936	0.928	0.933	0.859	886	41	64	544	0.933	0.930	0.932	0.857	-0.003	0.002	-0.001	-0.002

Table 4.1: Performance of the structural alert models with and without exclusion rules. For each target, the results for structural alerts generated by the automated workflow using two different sets of parameters are presented. One set of parameters uses a theta value of 0.51, meaning substructures which are contained by more true positives, but also more false positives, are chosen as structural alerts. The other set of parameters used a theta value of 0.95, meaning more specific substructures which are contained by fewer false positives, but also fewer true positives, are chosen as structural alerts. Both sets of parameters use maximum occurrence in the inactive chemicals of 5% and lower bounds for a structural alert of two actives and one inactive.

		Lack of structural alert only			Exclusion rules			
Target	**Theta**	**FN**	**TN**	**NPV**	**FN**	**TN**	**NPV**	**ΔNPV**
Acetylcholinesterase	0.51	55	339	0.860	10	36	0.783	-0.078
Acetylcholinesterase	0.95	93	435	0.824	3	1	0.250	-0.574
Androgen receptor	0.51	206	1769	0.896	0	5	1.000	0.104
Androgen receptor	0.95	212	1773	0.893	0	3	1.000	0.107
Beta-2 adrenergic receptor	0.51	107	427	0.800	2	3	0.600	-0.200
Beta-2 adrenergic receptor	0.95	125	443	0.780	2	4	0.667	-0.113
Dopamine D2 receptor	0.51	43	207	0.828	27	12	0.308	-0.520
Dopamine D2 receptor	0.95	52	223	0.811	4	6	0.600	-0.211
Glucocorticoid receptor	0.51	188	1645	0.897	8	19	0.704	-0.194
Glucocorticoid receptor	0.95	205	1671	0.891	1	2	0.667	-0.224
hERG	0.51	153	384	0.715	17	23	0.575	-0.140
hERG	0.95	396	680	0.632	0	3	1.000	0.368
Mu opioid receptor	0.51	43	492	0.920	2	32	0.941	0.022
Mu opioid receptor	0.95	61	543	0.899	3	1	0.250	-0.649

Table 4.2: Comparison of the negative predictivity (NPV) – the proportion of negative predictions which are true negatives – for predictions made by exclusion rules and for predictions made by lack of structural alert only. For each target, the results for structural alerts generated by the automated workflow using two different sets of parameters are presented. One set of parameters uses a theta value of 0.51, meaning substructures which are contained by more true positives, but also more false positives are chosen as structural alerts. The other set of parameters used a theta value of 0.95, meaning more specific substructures which are contained by fewer false positives but also fewer true positives are chosen as structural alerts. Both sets of parameters use maximum occurrence in the inactive chemicals of 5% and lower bounds for a structural alert of two actives and one inactive.

4.3. Classification of negative predictions by features

Chemicals that are predicted to be inactive by the structural alert models could be split into two groups: chemicals which have features that have been seen in the inactives used in model construction, and chemicals with unknown features which are unlike the inactives used in model construction. One may have more confidence in the inactivity prediction of chemicals in the former group as the model has seen the features and built structural alerts that avoid containing them. Hence, we need some idea of when a new chemical is like or unlike training data.

In this work, "chemical features" have been defined as small groups of connected atoms, i.e. small substructures. In Morgan fingerprints, circular atom neighbourhoods are hashed into a bit string. The user specifies radius of the circular atom neighbourhood and the length of the bit string. Circular atom neighbourhoods will be generated at the radius specified by the user, and at every atom radius lower than this down to one atom. Bit collisions occur when multiple circular atom neighbourhoods within the same chemical are hashed onto the same bit. Fingerprints with a longer bit string will have a lower probability of bit collisions. In absence of bit collisions, each bit in a Morgan fingerprint represents the presence of a small substructure. Thus, each bit in a Morgan fingerprint represents a chemical feature.

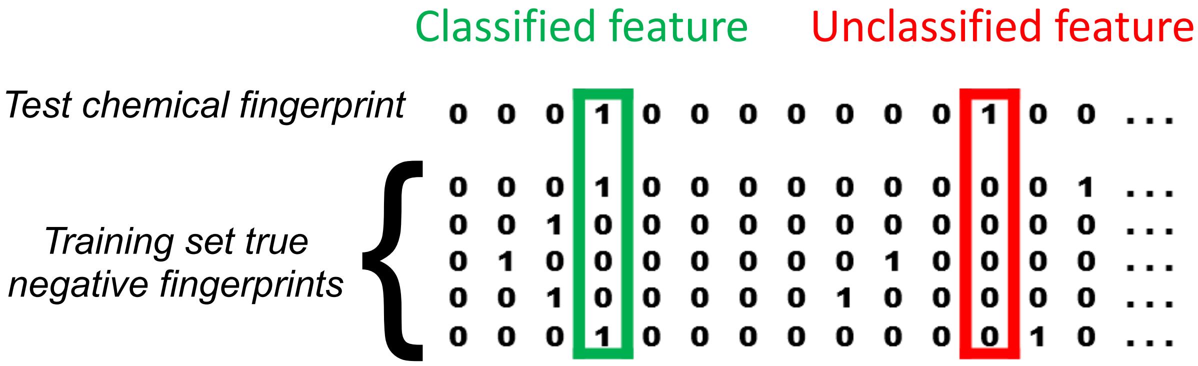
A Morgan fingerprint can be generated for a chemical which has been predicted to be inactive, giving a bit string that represents the features present. This can be compared to the Morgan fingerprints of the training chemicals to determine if all features of the new chemical have been seen in the inactives involved in model construction. Features which have been seen in inactive chemicals in the training set are considered "classified" features, and those which have not been seen in the training inactive chemicals are considered "unclassified".

Other fingerprints, such as the MACCS fingerprints could be used instead of Morgan fingerprints.

In their work in assessing confidence in negative predictions from mutagenicity structural alerts, Williams *et al* found a higher NPV for chemicals containing only classified features.[76] Similar results were found for structural alerts for skin sensitisation by Chilton *et al.*[99] In this section, it is investigated whether an increase in NPV is also observed for the structural alerts for receptor binding MIEs. The choice of bit string length and radius of circular atom neighbourhoods in the Morgan fingerprint were also investigated.

4.3.1. Method

For a particular target, structural alerts (created using the training set) are applied to the test set. Only chemicals which are predicted to be inactive are kept and Morgan fingerprints are generated for these. Morgan fingerprints are also generated for all inactive training set true negatives (TN) chemicals using the same bit string length and the same radius of fingerprint. For each test chemical predicted to be inactive, the positions of present bits ("1") within the fingerprint are found. If the bit is present in the any training TN, it is considered a "classified feature". If the bit is absent in all training set chemicals, it is considered an "unclassified feature". If a test chemical contains no structural alerts and has one or more unclassified features, it is categorised as "inactive with unclassified features". Test chemicals containing no structural alerts and no unclassified features are categorised as "inactive with classified features". The classification-by-features process is represented graphically in Figure 4.3.


Classified feature
Unclassified feature
Test chemical fingerprint
Training set true negative fingerprints

```
0 0 0 1 0 0 0 0 0 0 0 1 0 0 ...

0 0 0 1 0 0 0 0 0 0 0 0 0 1 ...
0 0 1 0 0 0 0 0 0 0 0 0 0 0 ...
0 1 0 0 0 0 0 0 0 1 0 0 0 0 ...
0 0 1 0 0 0 0 0 1 0 0 0 0 0 ...
0 0 0 1 0 0 0 0 0 0 0 0 1 0 ...
```

Figure 4.3: a graphical representation of the classification-by-features process. Each row of numbers is a Morgan fingerprint. The top row represents the test set chemical and the other rows represent the training set true negative chemicals. Where a test set feature is present in any of the training set true negatives, it is considered a "classified feature". If the test set feature is not present in any of the training set true negatives, it is considered an "unclassified feature".

For each target that the classification-by-features process was applied to, structural alerts built by the automated workflow with the "Risk Assessment" parameters were used in each case (theta 0.95, 1% maximum occurrence in inactives, structural alert lower bounds of two actives and one inactive). These parameters were chosen as they give the greatest number of false negatives for each target.

Initially, the MACCS 166 fingerprints have also been used instead of Morgan fingerprints in the classification-by-features process. The RDKit[83] implementation of the MACCS 166 Fingerprints was used, generated using the Fingerprint node in KNIME.[67]

The effect of using different bit string lengths in generating the Morgan fingerprints was investigated.

4.3.2. Results and discussion

4.3.2.1. MACCS Fingerprints

The MACCS 166 fingerprints were initially used for the classification-by-features process for three targets: acetylcholinesterase, alpha-2a adrenergic receptor and dopamine D2 receptor. The results shown in Table 4.3. The results show that very few chemicals have unclassified features when using this fingerprint.

The MACCS 166 fingerprint indicates the presence or absence of 166 features. Whilst it may be suitable for other purposes such as substructure searching, the MACCS 166 fingerprints are not suitable for this purpose because they do not allow enough distinction between chemicals. Here, each bit present in a test chemical is compared to the same bit position in all true negatives in the training set. With at least 700 true negatives in the training sets of each target, it is likely that most of the 166 features are present at least once. Hence, most features will be considered classified features. The MACCS fingerprint might be suitable in data sets with fewer chemicals, but they are not suitable in these data sets.

	Classified Chemicals			Unclassified Chemicals					
Target	TN	FN	NPV	TN	FN	NPV	NPV without classification	ΔNPV with classification	% unclassified chemicals
AChE	445	111	80.0%	1	0	100.0%	80.1%	0.0%	0.2%
ADRA2A	230	38	85.8%	3	1	75.0%	85.7%	0.2%	1.5%
DRD2	223	51	81.4%	0	0	-	81.4%	0.0%	0.0%

Table 4.3: Effect of classification algorithm when using MACCS 166 fingerprints. Data is shown for three targets: acetylcholinesterase (AChE), alpha-2a adrenergic receptor (ADRA2A), and dopamine D2 receptor (DRD2). TN = True negatives; FN = False negatives; NPV = negative predictive value.

4.3.2.2. Morgan Fingerprints

Effect of fingerprint string length

The results of the classification-by-features for four different targets using Morgan fingerprints with radius two atoms and varying bit string length are shown in Table 4.4. These targets are acetylcholinesterase, the alpha-2a adrenergic receptor, the dopamine D2 receptor, and the vasopressin V1a receptor.

From here on, different features in different chemicals being hashed to the same bit position in different fingerprints will be referred to as a "bit clash". This differs from a "bit collision", which is different features in the same chemical being mapped to the same bit position in the same fingerprint. Increasing the length of the Morgan fingerprint strings reduces the probability of both bit clashes and bit collisions occurring. The average number of "on" bits of the test chemicals of each target has been included in Table 4.4 to give an indication of the sparsity of the fingerprints.

A bit collision in a test chemical would result in two features sharing the same bit in the test fingerprint, which could lead to one of the features wrongly being considered classified if one feature is present in the training chemicals but the other is not. A bit clash between a test chemical's feature and a different feature in a training true negative would also result in the test feature wrongly being considered classified. In the classification-by-features process, each bit in the test chemical is compared to the same bit in all training true negatives, of which there are at least 700 in the four targets here. Hence, a bit clash is approximately 700 times more likely to lead to erroneous classification of a test set feature than a bit collision, and bit collisions, by comparison, are not a significant source of error.

Thus, there are two main reasons why a feature present in a test set chemical would be labelled as classified:

1. The feature is present in at least one of the training set chemicals under consideration
2. A different feature is present in one of the training set chemicals and is hashed onto the same bit position in a Morgan fingerprint string as the feature of the test chemical (bit clash).

To interpret the results and theory of the classification-by-features process, we must first investigate the number of bit clashes at different fingerprint string lengths, and the effects bit clashes have on performance.

The probability of a test bit not having any bit clashes with the training chemicals can be calculated if we assume a random distribution of bits in a bit string, and we assume that each training chemical's bit string is independent of the other training chemicals' bit strings. Whilst the former is a valid assumption, the latter is not because many training chemicals will contain the same features. Making these assumptions would lead to an underestimate of the probability of no bit clashes between a test bit and the training chemicals.

A more pragmatic approach to assessing the likelihood and effects of bit clashes at different fingerprint string lengths is shown below.

		Morgan fingerprint			Unclassified Chemicals			Classified Chemicals				
Target	Training TN	Radius	Bits	Mean "on" bits	TN	FN	NPV	TN	FN	NPV	NPV pre-classification	ΔNPV with classification
AChE	1384	2	1024	36.0	3	0	100.0%	443	111	80.0%	80.1%	-0.1%
AChE	1384	2	2048	36.3	31	11	73.8%	415	100	80.6%	80.1%	0.5%
AChE	1384	2	4096	36.6	155	37	80.7%	291	74	79.7%	80.1%	-0.3%
AChE	1384	2	8000	36.7	233	58	80.1%	213	53	80.1%	80.1%	0.0%
ADRA2A	744	2	1024	29.2	6	2	75.0%	227	37	86.0%	85.7%	0.3%
ADRA2A	744	2	2048	29.4	73	19	79.3%	160	20	88.9%	85.7%	3.2%
ADRA2A	744	2	4096	29.7	139	30	82.2%	94	9	91.3%	85.7%	5.6%
ADRA2A	744	2	8000	29.7	165	35	82.5%	68	4	94.4%	85.7%	8.8%
DRD2	719	2	1024	29.8	7	1	87.5%	216	50	81.2%	81.4%	-0.2%
DRD2	719	2	2048	30.0	56	16	77.8%	167	35	82.7%	81.4%	1.3%
DRD2	719	2	4096	30.2	124	35	78.0%	99	16	86.1%	81.4%	4.7%
DRD2	719	2	8000	30.3	148	41	78.3%	75	10	88.2%	81.4%	6.8%
V1AR	766	2	1024	29.6	4	1	80.0%	262	12	95.6%	95.3%	0.3%
V1AR	766	2	2048	29.8	59	2	96.7%	207	11	95.0%	95.3%	-0.4%
V1AR	766	2	4096	30.1	141	10	93.4%	125	3	97.7%	95.3%	2.3%
V1AR	766	2	8000	30.1	185	11	94.4%	81	2	97.6%	95.3%	2.2%

Table 4.4: Effect of classify-by-features algorithm for four different targets when using Morgan fingerprints of varying bit string length. The mean number of "on" bits in the Morgan fingerprints of the test chemicals is included in the results. The targets are acetylcholinesterase (AChE), alpha-2a adrenergic receptor (ADRA2A), dopamine D2 receptor (DRD2), and vasopressin V1a receptor (V1AR). TN = True negatives; FN = False negatives; NPV = negative predictive value (proportion of negative predictions which are TNs).

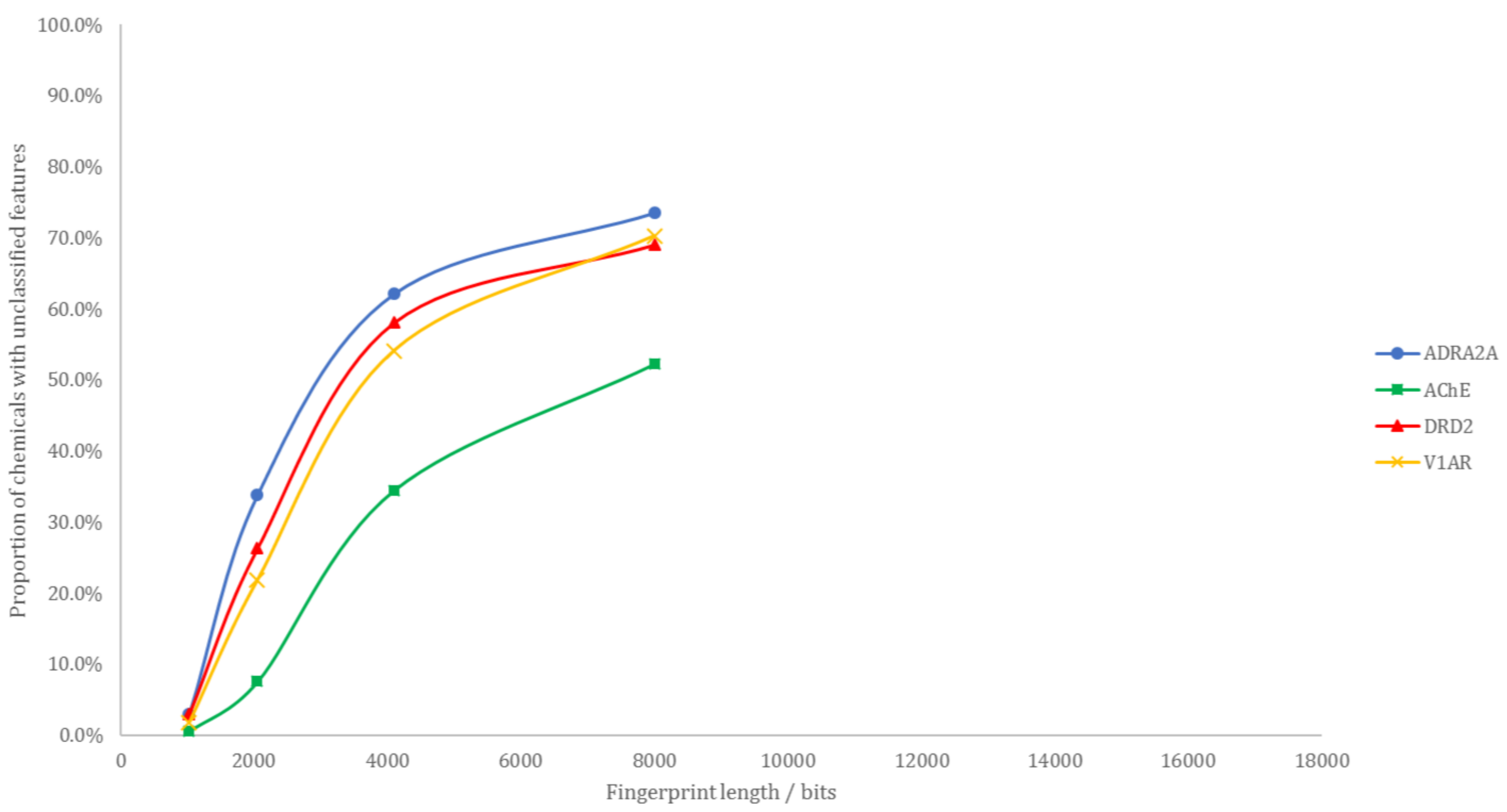


Figure 4.4: The proportion of chemicals with unclassified features as fingerprint length is increased. When the proportion of unclassified features no longer increases with increases in fingerprint length, there will be no bit clashes. Extrapolation of these graphs gives an estimate of a fingerprint length of roughly 12 000 - 18 000 bits required for there to be no bit clashes in these data sets. The targets shown are acetylcholinesterase (AChE), alpha-2a adrenergic receptor (ADRA2A), dopamine D2 receptor (DRD2), and vasopressin V1a receptor (V1AR).

The effect of bit string length on the proportion of chemicals with unclassified features is shown in Figure 4.4. At low bit string lengths, most chemicals are considered classified – with fingerprints of 1 024 bits, less than 3% of chemicals are considered unclassified in all four targets. When bit string length is increased to 8 000 bits, many more chemicals contain unclassified features – between 52% and 74% for the four targets. This increase in unclassified chemicals is due to bit clashes. At small bit string lengths, there is a higher probability of bit clashes, so a higher probability of features being considered classified and therefore more classified chemicals. Increasing the length of the bit string reduces the likelihood of bit clashes, resulting in more features being considered unclassified and therefore more unclassified chemicals.

Increasing the bit string length beyond 8 000 bits should eventually lead to a fingerprint long enough that there are no bit clashes. With this fingerprint, any classified features will be considered classified because that feature is present in at least one training set chemical rather than due to a bit clash. Further increasing the bit string length would have no change on the number of classified features. By extrapolating to the point where the curves in Figure 4.4 would be flat, the bit string length at which there is an insignificant number of bit clashes could be predicted as between 12 000 and 18 000 bits. This is a very rough prediction from only four points in each curve, but it gives an idea of the length of fingerprint that would be required to evaluate the effect of classifying features without errors due to bit clashes. However, the workflow for running the classify-by-features process in the targets' data sets is very memory intensive and time consuming. Using fingerprint strings lengths of greater than 8 000 bits would require greater computational power than a desktop computer, or construction of a more efficient program. For now, conclusions will be made from fingerprints up to 8 000 bits in length, and if it is deemed to be pragmatic to run calculations with larger bit strings, this will be done later.

Extrapolating the graphs to fingerprints string lengths of between 12 000 and 18 000 bits would result in approximately 80% of chemicals being considered unclassified. With only 20% of chemicals in these test sets being considered classified, one might question how useful this classification is. It could indicate that the Morgan fingerprints are too detailed for this purpose.

The number of true negatives in the training set must also be considered when discussing bit clashes. Data sets with a larger number of chemicals will generally have a larger number of total features present in the data set. This will lead to a greater number of bit clashes when the features are hashed to a bit string of fixed length. Of the four targets shown in Figure 4.4, acetylcholinesterase has the largest number of training true negatives and so has the greatest number of bit clashes at all fingerprint lengths.

In absence of results from impractically long fingerprints, the efficacy of the classification-by-features process must be evaluated whilst being aware of errors caused by bit clashes. Bit clashes will lead to random error, effecting both true negatives and false negatives with equal probability. Hence, a greater number of bit clashes will result in the change in NPV in the classified chemicals approaching zero. For each target, as fingerprint length increases, random error from bit clashes will decrease and the change in NPV in the classified chemicals will approach the value it would take without bit clashes. This trend can be seen most clearly in the results for the alpha-2a adrenergic receptor and dopamine D2 receptors, with ΔNPV of 8.8% and 6.8% respectively with fingerprints of 8 000 bit length. A general increase in NPV, allowing for error, can also be seen for the vasopressin V1a receptor, but a lower ΔNPV 2.2% is seen at 8 000 bit string length. As fingerprint string length increases for the acetylcholinesterase data, no clear increase in NPV is seen, with ΔNPV staying at around 0%. Acetylcholinesterase has a larger number of training true negatives than the other targets, leading to more random error due to bit clashes. However, at a bit string length of 8 000, more than half of the test chemicals are unclassified chemicals and so are unaffected by bit clashes. It can be concluded that the classification process is not helping to identify false negatives in the acetylcholinesterase data, but in the other three targets, unclassified chemicals are more likely to be false negatives than chemicals with only classified features.

Effect of fingerprint radius

Increasing the radius of the Morgan fingerprints would result in larger circular environments being considered features, as well as those already considered at a radius of two atoms. Hence, more features will be hashed onto the fingerprint bit string and, keeping bit string length constant, there will be more bit clashes. To use a Morgan fingerprint of radius three atoms with no bit clashes in the data sets used here, very large string lengths will be required, and these will be too computationally expensive. As such, only fingerprints with radius two have been used here.

Performance at fixed string lengths

To explore why the change in NPV varies between different targets, the classification-by-features process has been applied to further targets beyond the preliminary study. The targets investigated here are:

- acetylcholinesterase
- alpha-2a adrenergic receptor
- beta-1 adrenergic receptor
- dopamine D2 receptor
- norepinephrine transporter
- serotonin 2a receptor
- serotonin 3a receptor
- serotonin transporter
- tyrosine-protein kinase LCK
- vasopressin V1a receptor

Morgan fingerprints of radius two atoms and 4 096 bit string length were used for all of these targets. A longer string length would have led to fewer bit clashes and less random error but would have been computationally more expensive (calculations with fingerprint lengths of 4 096 bits took between a few hours and a day to run, whilst lengths of 8 000 could take several days).

The performance of the structural alert-based models used before the classification process are shown in Table 4.5, and the results of the classification process are shown in Table 4.6.

With the exception of acetylcholinesterase, the NPV for chemicals with only classified features is greater than the overall NPV in all targets, suggesting that the classification-by-features process is helping to distinguish true negatives from false negatives. Particularly large increases in NPV are seen in tyrosine-protein kinase LCK (8.5%), the serotonin transporter (8.0%), and the serotonin 2a receptor (7.8%). The targets with the most training true negatives tend to have more features in these chemicals, and so there are more bit clashes when using the same length bit string. As a result, these targets have changes in NPV closer to zero. Conversely, the target with the least training true negative has the largest increase in NPV.

		Training set								Test set							
Target	**Alerts**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**	**TP**	**FP**	**FN**	**TN**	**SE**	**SP**	**ACC**	**MCC**
LCK	55	1239	35	44	370	96.6%	91.4%	95.3%	0.873	412	21	37	98	91.8%	82.4%	89.8%	0.709
5HTR2A	81	2737	59	60	712	97.9%	92.3%	96.7%	0.902	928	30	32	233	96.7%	88.6%	94.9%	0.850
DRD2	68	4195	136	67	719	98.4%	84.1%	96.0%	0.854	1383	59	51	223	96.4%	79.1%	93.6%	0.764
ADRA2A	68	596	25	53	744	91.8%	96.7%	94.5%	0.890	157	12	39	233	80.1%	95.1%	88.4%	0.769
5-HTT	48	3004	105	53	747	98.3%	87.7%	96.0%	0.879	944	37	42	246	95.7%	86.9%	93.8%	0.822
V1AR	15	446	21	10	766	97.8%	97.3%	97.5%	0.947	150	6	13	266	92.0%	97.8%	95.6%	0.907
5HTR3A	28	316	11	32	767	90.8%	98.6%	96.2%	0.910	87	4	17	273	83.7%	98.6%	94.5%	0.859
ADRB1	49	913	36	47	770	95.1%	95.5%	95.3%	0.905	258	22	43	254	85.7%	92.0%	88.7%	0.777
NET	69	2101	94	122	1355	94.5%	93.5%	94.1%	0.877	625	32	64	460	90.7%	93.5%	91.9%	0.836
AChE	166	1792	83	212	1384	89.4%	94.3%	91.5%	0.830	499	50	111	446	81.8%	89.9%	85.4%	0.713

Table 4.5: The performance of the structural alert-based models used in the classification-by-features process. All were generated with "Risk Assessment" parameters: theta 0.95, 1% maximum occurrence in the inactives, and lower bounds for an alert of two actives and one inactive. The targets are: tyrosine-protein kinase LCK (LCK), serotonin 2a receptor (5HTR2A), dopamine D2 receptor (DRD2), alpha-2a adrenergic receptor (ADRA2A), serotonin transporter (5-HTT), vasopressin V1a receptor (V1AR), serotonin 3a receptor (5HTR3A), beta-1 adrenergic receptor (ADRB1), norepinephrine transporter (NET), and acetylcholinesterase (AChE).

		Classified Chemicals			Unclassified Chemicals					
Target	Training TN	TN	FN	NPV	TN	FN	NPV	NPV without classification	ΔNPV with classification	% unclassified chemicals
LCK	370	30	7	81.1%	68	30	69.4%	72.6%	8.5%	72.6%
5HTR2A	712	112	5	95.7%	121	27	81.8%	87.9%	7.8%	55.8%
DRD2	719	99	16	86.1%	124	35	78.0%	81.4%	4.7%	58.0%
ADRA2A	744	94	9	91.3%	139	30	82.2%	85.7%	5.6%	62.1%
5-HTT	747	123	8	93.9%	132	34	79.5%	85.9%	8.0%	55.9%
V1AR	766	125	3	97.7%	141	10	93.4%	95.3%	2.3%	54.1%
5HTR3A	767	129	4	97.0%	144	13	91.7%	94.1%	2.9%	54.1%
ADRB1	770	120	12	90.9%	134	31	81.2%	85.5%	5.4%	55.6%
NET	1355	319	37	89.6%	141	27	83.9%	87.8%	1.8%	32.1%
AChE	1384	291	74	79.7%	155	37	80.7%	80.1%	-0.3%	34.5%

Table 4.6: The results of the classification process on a wide selection of targets. Morgan fingerprints of radius two atoms and 4 096 bit string length were used for all results shown here. The targets are: tyrosine-protein kinase LCK (LCK), serotonin 2a receptor (5HTR2A), dopamine D2 receptor (DRD2), alpha-2a adrenergic receptor (ADRA2A), serotonin transporter (5-HTT), vasopressin V1a receptor (V1AR), serotonin 3a receptor (5HTR3A), beta-1 adrenergic receptor (ADRB1), norepinephrine transporter (NET), and acetylcholinesterase (AChE).

Explaining the results of the classification-by-features process

To use the classification-by-features process to increase confidence in negative predictions, it is necessary to explain why some targets give better results than others, and why acetylcholinesterase does not give an increase in NPV when all other targets do.

Consider simply taking the Tanimoto similarity coefficient between each test chemical's fingerprint and the training true negative chemicals' fingerprints. If a test chemical has a high similarity coefficient relative to any single training true negative, most of its features will be present in that one training chemical. The test chemical will therefore be less likely to contain any unclassified features. Conversely, a test chemical with a low similarity score relative to all training true negatives is more likely to contain unclassified features. Therefore, we expect to see a correlation between a test chemical's maximum Tanimoto similarity coefficient to training true negatives and the probability of the chemical being considered classified. This trend is not expected to be exact because:

- The feature comparison across all training inactives is more complex than a similarity coefficient to just one chemical.
- Presence of bit clashes leads to some features being considered classified even though the feature is not present in any of the training inactives (as discussed previously).

Morgan fingerprints of radius two atoms and string length of 4 096 bits were generated for all chemicals. The Tanimoto similarity coefficients between each test set chemical predicted to be inactive and the training set true negatives were calculated and the maximum value was found. the distribution of the maximum Tanimoto similarity coefficients in both the test set true negatives and the test set false negatives was plotted graphically. This value was also plotted against the number of unclassified features for each test chemical.

These graphs are shown for tyrosine-protein kinase LCK (the target with largest increase in NPV) and for acetylcholinesterase (the target with no change in NPV) in Figures 4.5a and 4.5b respectively.

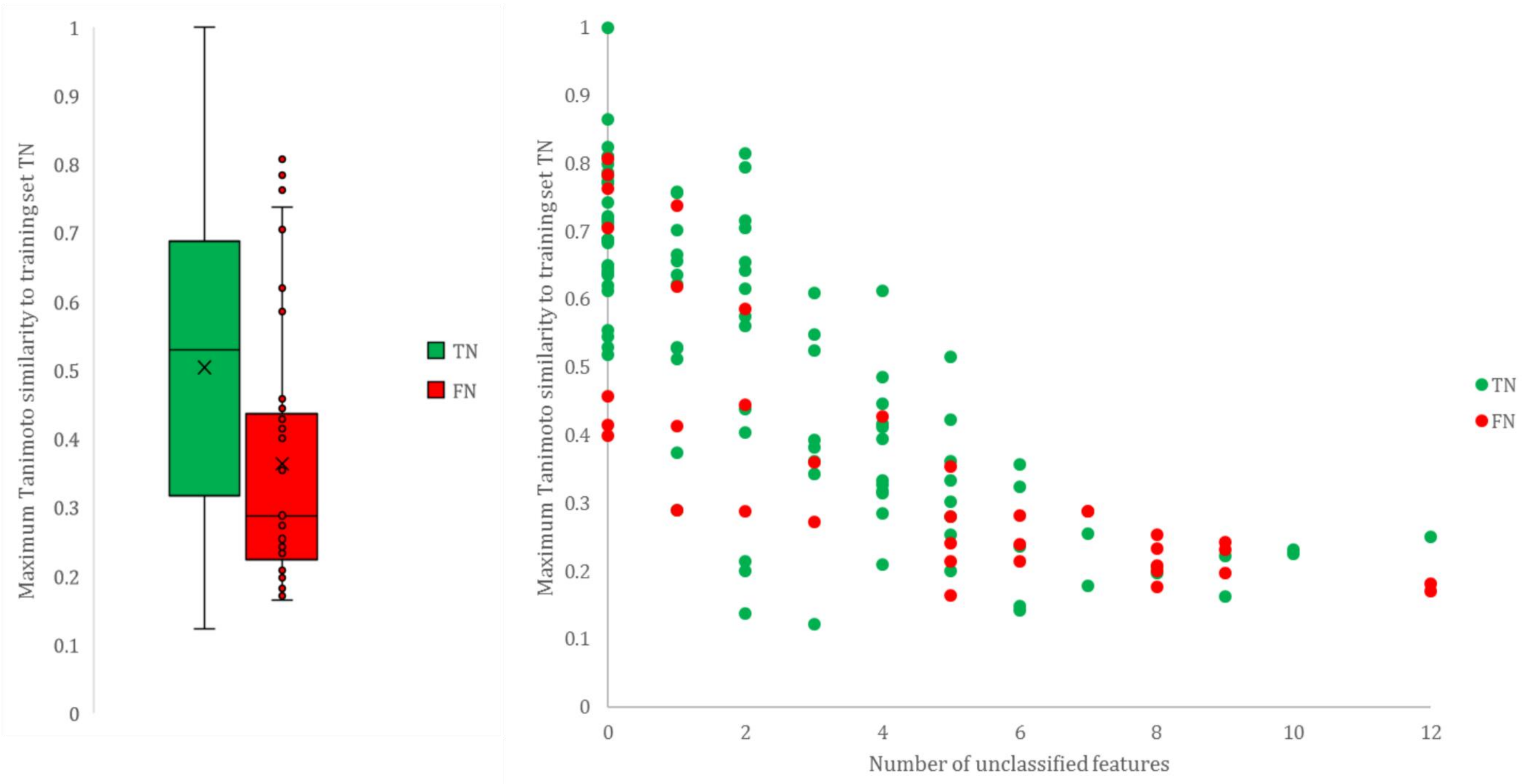


Figure 4.5a: Data for tyrosine-protien kinase LCK. Left: variation of maximum Tanimoto similarity (based on Morgan fingerprints) to training set true negatives from each chemical in the test set predicted to be inactive, showing the different skew of data in the true negatives (TN) from the false negatives (FN). Right: how this similarity value varies with number of unclassified features identified in classification process. Morgan fingerprints of radius two atoms and 4 096 bit string length were used.

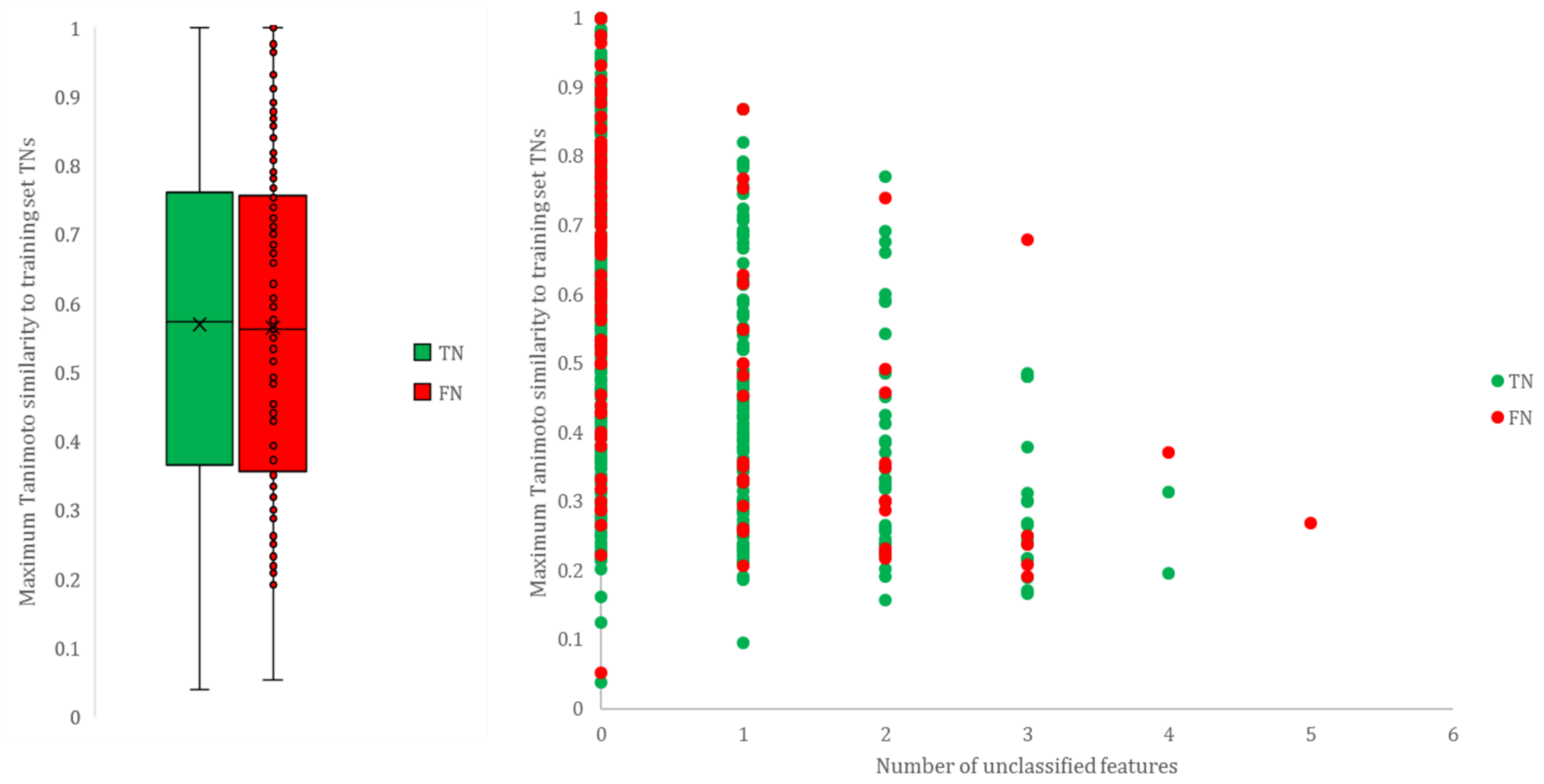


Figure 4.5b: Data for acetylcholinesterase. Left: variation of maximum Tanimoto similarity (based on Morgan fingerprints) to training set true negatives from each chemical in the test set predicted to be inactive, showing a similar skew of data in the true negatives (TN) from the false negatives (FN). Right: how this similarity value varies with number of unclassified features identified in classification process. Morgan fingerprints of radius two atoms and 4 096 bit string length were used.

In both targets, chemicals with lower maximum Tanimoto similarity coefficient (based on Morgan fingerprints) to training set true negatives generally contain more unclassified features, and, hence, are more likely to be unclassified chemicals. This trend is complicated by bit clashes leading to random errors, and so is clearer for tyrosine-protein kinase LCK, which has fewer training true negatives than acetylcholinesterase and therefore fewer bit clashes.

In both targets, there is no significant skew in the distribution of Tanimoto coefficients (based on Morgan fingerprints) of the test set true negatives. The targets differ in the distribution of the false negatives. The false negatives in tyrosine-protein kinase LCK are strongly skewed towards low Tanimoto coefficients. As a result, the false negatives are more likely to be identified as having unclassified features. The false negatives in acetylcholinesterase are not skewed and have a near-identical distribution of Tanimoto coefficients as the true negatives. It is not possible to distinguish the true negatives from false negatives according to these distributions, and this is reflected in the results of the classification-by-features process.

In acetylcholinesterase there are some training inactive chemicals which are very similar to test active chemicals. These are activity cliffs. These similar active and inactive chemicals might come from the same chemical series – chemicals with the same structural backbones but with differing side groups leading to significant changes in activity.

The classification-by-features process appears to be identifying chemicals which are most similar to training true negatives as "classified chemicals", regardless of whether they are active or inactive. Whilst, the classification-by-features process has resulted in improvements in NPV for most targets, similar improvements in NPV may be possible by a different method that simply considers Tanimoto similarity coefficients (based on Morgan fingerprints) to training set true negatives. Such a method would be less computational expensive and less affected by bit clashes.

The classification-by-features method presented here has been inspired by a method developed by Lhasa for reactivity-driven MIEs. In these MIEs, presence of a single electrophilic feature can lead to activity. Hence, identifying an unclassified feature through the classification-by-feature process allows identification of features that could be electrophilic, leading to unexpected activity in a chemical predicted to be inactive. In reactivity-drive MIEs, "local" changes in chemical structure can lead to changes of activity.

The structural alerts developed in this work have been created for biological targets which are activated through receptor binding-based mechanisms. For a chemical to undergo a receptor binding-based MIE, it must have the right combination of features (hydrogen bond donors, hydrogen bond acceptors, ionisable or charged features, aromatic rings, etc) in a specific configuration in three-dimensional space. Presence of an unknown feature alone may not be enough to lead to unpredicted activity. The classification-by-features process allows the identification of unknown features, but it does not indicate if the unknown feature is in a position in the chemical to affect receptor binding. Receptor binding MIEs are less sensitive to “local” changes in chemical structure.

4.4. Classification of negative predictions by similarity

In the classification-by-features section, it was concluded that equally good results could be achieved by simply considering Tanimoto similarity coefficients (based on Morgan fingerprints) between a predicted inactive and the training set true negatives. The key hypothesis of this method is that inactive chemicals that are more similar to training set true negatives are more likely to be true negatives.

Acetylcholinesterase was a problematic target for the classification-by-features method. The presence of activity cliffs meant some training true negatives were very similar to active chemicals, resulting in false negative chemicals being consider "classified". To identify similar cases, similarity between predicted inactive chemicals and training set actives will also be considered. Test chemicals with a high Tanimoto similarity (based on Morgan fingerprints) to a training active chemical are more likely to be false negative.

4.4.1. Method

Morgan fingerprints with a radius of two atoms and a string length of 4 096 bits were generated for all chemicals. Two rules were applied to test chemicals which were predicted to be negative by the structural alert-based models:

1. The test chemical must not exceed a maximum Tanimoto similarity coefficient (between fingerprints) to any training active chemical. Chemicals which do not meet this requirement are considered too similar to known active chemicals and therefore more likely to be a false negative prediction. These chemicals were labelled “no alert but like actives”.
2. The test chemical must have a Tanimoto similarity coefficient (between fingerprints) to at least one training true negative that exceeds a minimum value. Chemicals which do not meet this requirement are considered too dissimilar to the training inactives and were labelled “out of domain”.

Chemicals which met these two requirements were considered “classified” chemicals.

Different values for both requirements were trialled to find the effect of varying each, and to find the best performing combination.

The maximum Tanimoto similarity coefficient to any training active was varied without applying the second requirement.

The minimum Tanimoto similarity coefficient to at least one training true negative was varied after applying the first requirement with a maximum Tanimoto similarity coefficient of 0.7 to any training active chemical.

Structural alert-based models created by the automated workflow with “Risk Assessment” parameters (0.95 theta, 1% maximum occurrence in the inactive chemicals, and lower bounds of two active and one inactive) were applied to the test sets of the 24 Bowes targets with ToxCast data. The classification-by-similarity process was applied to the chemicals predicted to be inactive by the models.

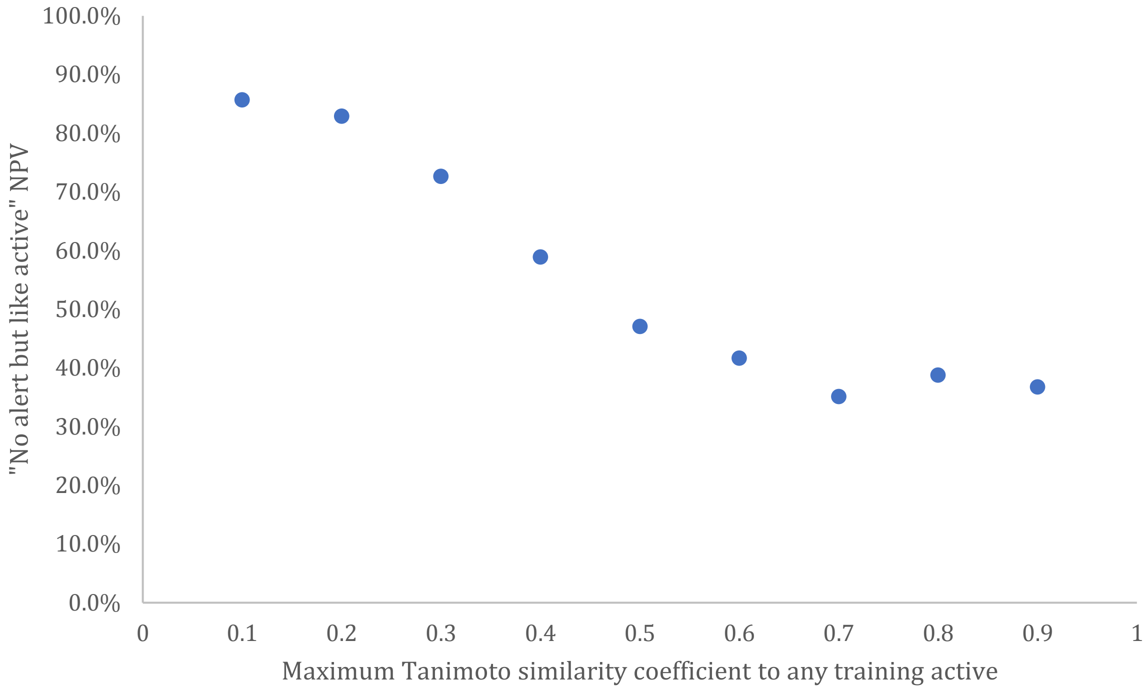
4.4.2. Results and discussion

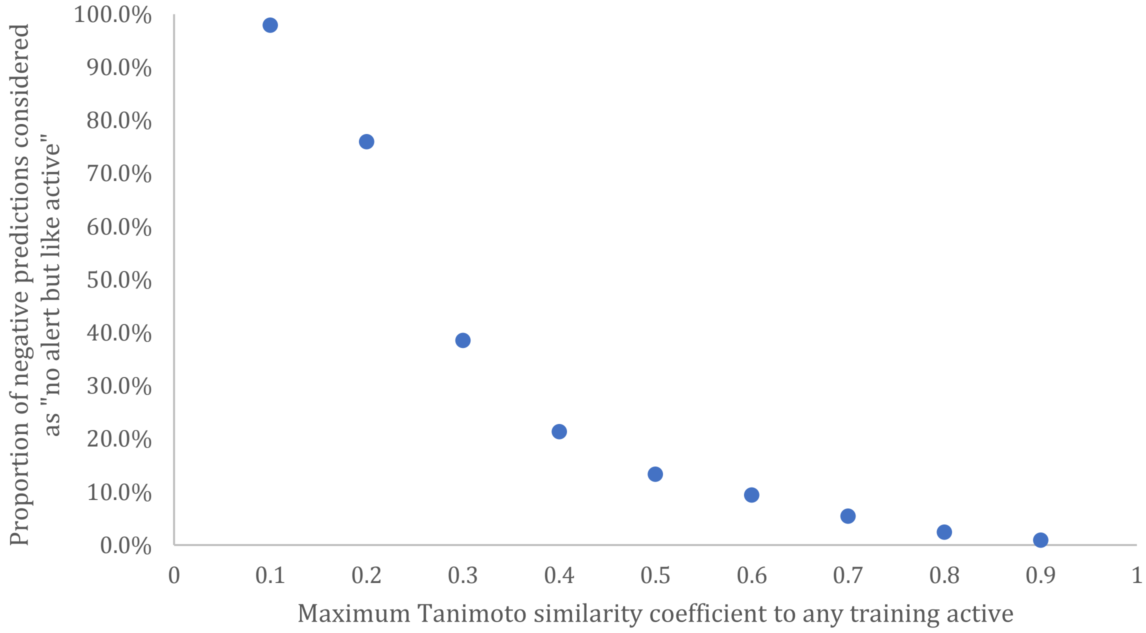
4.4.2.1. Variation of maximum similarity to a training active

The classification-by-similarity process was applied with varying maximum Tanimoto similarity coefficient based on Morgan fingerprints to any training active. There was no requirement for similarity to a known inactive. Hence, negative predictions were put in two groups: “No alert but like active” if the chemical exceeded a maximum similarity to any training active, or “classified” if it did not. The purpose of this was to find a sensible limit for similarity to training actives.

The results are shown in Figure 4.6.a and Figure 4.6.b.

The lowest NPV in the “no alert but like active group”, and, hence, largest proportion of false positive predictions, was seen when a value of 0.7 was used as the maximum Tanimoto similarity (based from Morgan fingerprints) to any training active chemical. At this limit, an average across data sets of 5.5% of negative predictions were considered “no alert but like active”.

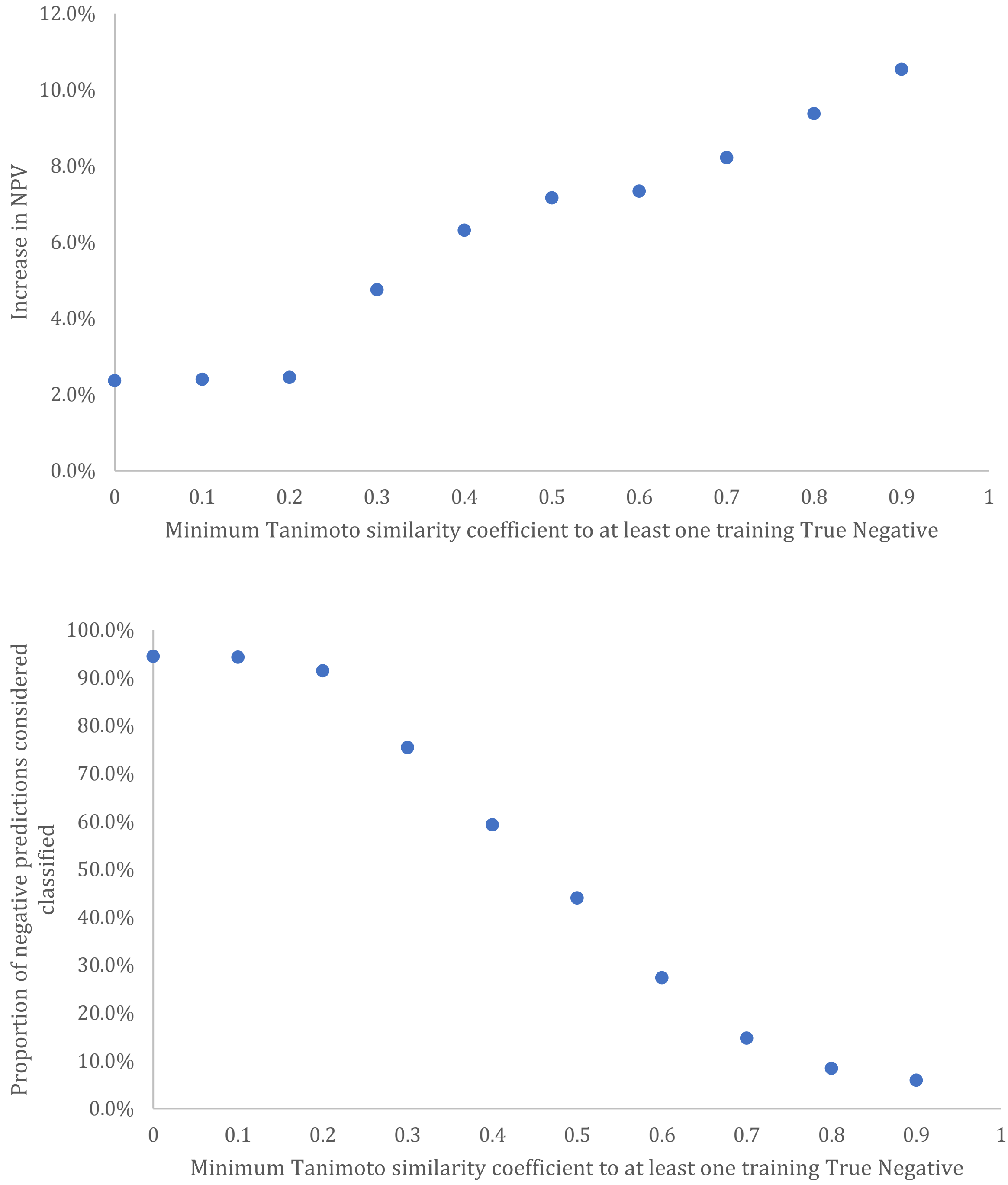




Figures 4.6.a (top) and 4.6.b (bottom): Structural alerts constructed with the automated workflow (“Risk Assessment” parameters) were applied to the test sets of the 24 Bowes targets with ToxCast data. The classification-by-similarity process was applied to the test chemicals predicted to be inactive, varying the maximum Tanimoto similarity coefficient based from Morgan fingerprints to any training active chemical and no minimum similarity to training true negatives. The variation of NPV in the “no alert but like active” group is shown in the top figure, and the variation of the proportion of negative predictions in this category shown in the lower figure, taking mean values across all test sets.

4.4.2.2. Variation of minimum similarity to a training true negative

After applying the first requirement with a maximum Tanimoto similarity coefficient (based on Morgan fingerprints) of 0.7 to any training active chemical, the minimum Tanimoto similarity coefficient to at least one training true negative was varied. At each different, the NPV for the chemicals considered “classified” in each data set is calculated and compared to the NPV when using structural alerts only. The mean change in NPV is taken across all data sets, and the variation is shown in Figure 4.7a. The variation of mean proportion of chemicals considered “classified” is shown in Figure 4.7.b.



Figures 4.7.a (top) and 4.7.b (bottom): Structural alerts constructed with the automated workflow ("Risk Assessment" parameters) were applied to the test sets of the 24 Bowes targets with ToxCast data. The classification-by-similarity process was applied to the test chemicals predicted to be inactive, using a maximum Tanimoto similarity coefficient of 0.7 to any training active chemical and a varying minimum Tanimoto similarity coefficient to training true negatives. The variation of change in NPV is shown in the top figure, and the variation of proportion of chemicals considered classified is shown in the lower figure, taking mean values across all test sets.

The mean NPV when using only structural alerts was 85.9%. When a maximum Tanimoto similarity coefficient (between Morgan fingerprints) of 0.7 to any training active chemical is required, the mean NPV increases to 88.2%. At a minimum required Tanimoto similarity coefficient (between Morgan fingerprints) of 0.1 to at least one training true negative, there is no change in NPV as all chemicals meet this requirement. As the minimum required values increases, the mean NPV increases at every step, up to 96.4% at 0.9 (larger minimum required values were not used). This suggests that test chemicals that are more similar to training true negative are more likely to be true negatives, agreeing with Johnson and Maggiora's similarity property principle.[57]

As the minimum required Tanimoto similarity coefficient to at least one training true negative increases, the mean proportion of negative-predicted test chemicals considered to be classified decreases with an S-shaped curve.

To evaluate confidence in negative predictions, the user could look at the largest Tanimoto similarity coefficient to training true negatives – the larger the value, the more likely the test chemical is to be a true negative. However, less than 15% of test chemicals have a value of 0.7 or greater. A cut-off of a minimum Tanimoto similarity coefficient to at least one training true negative of 0.4 provides a good balance of high NPV and high proportion of chemicals considered classified with the test sets.

4.4.2.3. Overall process

The classification-by-similarity process when using this cut-off is graphically shown in Figure 4.8.

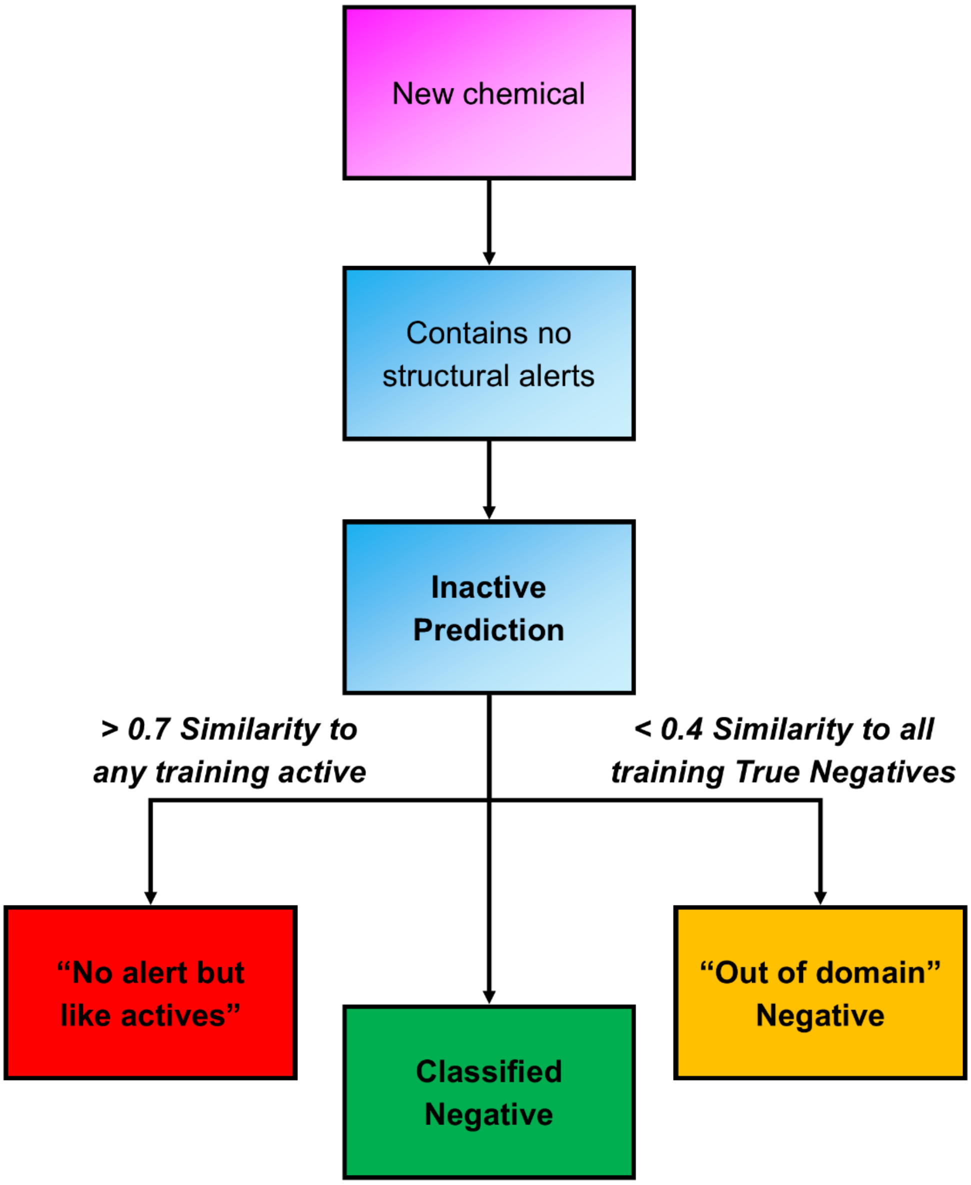

New chemical
Contains no structural alerts
Inactive Prediction
> 0.7 Similarity to any training active
< 0.4 Similarity to all training True Negatives
"No alert but like actives"
Classified Negative
"Out of domain" Negative

Figure 4.8: The three classifications of negative predictions, as defined by Tanimoto similarity coefficients between the Morgan fingerprints of a test chemical and the training chemicals. In this image, pink boxes are input chemicals, blue boxes are key steps in the process, green boxes represent high confidence predictions, yellow represent medium confidence predictions, and red represents low confidence predictions.

The performance of the structural alert-based models used before the classification-by-similarity process is shown in Table 4.7, and the results of the classification-by-similarity process with the suggested cut-offs for each target are shown in Table 4.8.

Target	Alerts	Training set								Test set							
		TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC
Acetylcholinesterase	166	1792	83	212	1384	89.4%	94.3%	91.5%	0.830	499	50	111	446	81.8%	89.9%	85.4%	0.713
Adenosine A2a receptor	72	2855	94	102	1488	96.6%	94.1%	95.7%	0.905	935	44	51	457	94.8%	91.2%	93.6%	0.858
Alpha-2a adrenergic receptor	68	596	25	53	744	91.8%	96.7%	94.5%	0.890	157	12	39	233	80.1%	95.1%	88.4%	0.769
Androgen receptor	115	1499	85	490	5384	75.4%	98.4%	92.3%	0.798	436	42	212	1773	67.3%	97.7%	89.7%	0.723
Beta-1 adrenergic receptor	49	913	36	47	770	95.1%	95.5%	95.3%	0.905	258	22	43	254	85.7%	92.0%	88.7%	0.777
Beta-2 adrenergic receptor	135	1252	86	210	1415	85.6%	94.3%	90.0%	0.803	363	63	120	450	75.2%	87.7%	81.6%	0.635
Delta opioid receptor	41	2189	166	51	738	97.7%	81.6%	93.1%	0.828	735	59	31	256	96.0%	81.3%	91.7%	0.795
Dopamine D1 receptor	72	881	56	148	1428	85.6%	96.2%	91.9%	0.832	247	25	76	482	76.5%	95.1%	87.8%	0.743
Dopamine D2 receptor	68	4195	136	67	719	98.4%	84.1%	96.0%	0.854	1383	59	51	223	96.4%	79.1%	93.6%	0.764
Dopamine transporter	70	1745	91	125	1354	93.3%	93.7%	93.5%	0.868	568	35	73	437	88.6%	92.6%	90.3%	0.805
Endothelin receptor ET-A	24	937	50	25	809	97.4%	94.2%	95.9%	0.918	305	18	18	275	94.4%	93.9%	94.2%	0.883
Glucocorticoid receptor	123	1814	93	460	5150	79.8%	98.2%	92.6%	0.823	537	55	207	1675	72.2%	96.8%	89.4%	0.742
hERG	456	3099	229	518	2192	85.7%	90.5%	87.6%	0.751	878	145	400	680	68.7%	82.4%	74.1%	0.499
Histamine H1 receptor	66	908	29	55	793	94.3%	96.5%	95.3%	0.906	273	18	40	266	87.2%	93.7%	90.3%	0.808
Mu opioid receptor	56	2576	117	85	1606	96.8%	93.2%	95.4%	0.903	889	42	61	543	93.6%	92.8%	93.3%	0.859
Muscarinic acetylcholine receptor M1	77	1397	72	87	865	94.1%	92.3%	93.4%	0.862	479	42	52	263	90.2%	86.2%	88.8%	0.759
Muscarinic acetylcholine receptor M2	54	1155	79	70	1427	94.3%	94.8%	94.5%	0.890	374	35	37	492	91.0%	93.4%	92.3%	0.844
Muscarinic acetylcholine receptor M3	69	1136	65	54	774	95.5%	92.3%	94.1%	0.879	307	22	40	253	88.5%	92.0%	90.0%	0.801
Norepinephrine transporter	69	2101	94	122	1355	94.5%	93.5%	94.1%	0.877	625	32	64	460	90.7%	93.5%	91.9%	0.836
Serotonin 2a (5-HT2a) receptor	81	2737	59	60	712	97.9%	92.3%	96.7%	0.902	928	30	32	233	96.7%	88.6%	94.9%	0.850
Serotonin 3a (5-HT3a) receptor	28	316	11	32	767	90.8%	98.6%	96.2%	0.910	87	4	17	273	83.7%	98.6%	94.5%	0.859
Serotonin transporter	48	3004	105	53	747	98.3%	87.7%	96.0%	0.879	944	37	42	246	95.7%	86.9%	93.8%	0.822
Tyrosine-protein kinase LCK	55	1239	35	44	370	96.6%	91.4%	95.3%	0.873	412	21	37	98	91.8%	82.4%	89.8%	0.709
Vasopressin V1a receptor	15	446	21	10	766	97.8%	97.3%	97.5%	0.947	150	6	13	266	92.0%	97.8%	95.6%	0.907
Average	**87**	**1699**	**80**	**133**	**1407**	**92.6%**	**93.4%**	**94.1%**	**0.868**	**532**	**38**	**78**	**460**	**86.6%**	**90.9%**	**90.2%**	**0.782**

Table 4.7: The performance of the structural alert-based models used in the classification-by-similarity process. All were generated with "Risk Assessment".

	Out of domain			No alert but like actives			Classified negatives							
Target	TN	FN	NPV	TN	FN	NPV	TN	FN	NPV	Pre-classified NPV	ΔNPV	Proportion Classified	MCC	ΔMCC
Acetylcholinesterase	130	28	82.3%	49	38	56.3%	267	45	85.6%	80.1%	5.5%	56.0%	0.762	0.049
Adenosine A2a receptor	127	19	87.0%	5	18	21.7%	325	14	95.9%	90.0%	5.9%	66.7%	0.890	0.032
Alpha-2a adrenergic receptor	99	24	80.5%	0	2	0.0%	134	13	91.2%	85.7%	5.5%	54.0%	0.841	0.072
Androgen receptor	304	59	83.7%	84	31	73.0%	1385	122	91.9%	89.3%	2.6%	75.9%	0.791	0.067
Beta-1 adrenergic receptor	92	14	86.8%	6	15	28.6%	156	14	91.8%	85.5%	6.2%	57.2%	0.832	0.055
Beta-2 adrenergic receptor	197	66	74.9%	4	9	30.8%	249	45	84.7%	78.9%	5.7%	51.6%	0.693	0.058
Delta opioid receptor	105	16	86.8%	2	9	18.2%	149	6	96.1%	89.2%	6.9%	54.0%	0.793	-0.002
Dopamine D1 receptor	139	46	75.1%	8	7	53.3%	335	23	93.6%	86.4%	7.2%	64.2%	0.845	0.102
Dopamine D2 receptor	88	28	75.9%	2	15	11.8%	133	8	94.3%	81.4%	12.9%	51.5%	0.787	0.023
Dopamine transporter	114	30	79.2%	3	8	27.3%	320	35	90.1%	85.7%	4.5%	69.6%	0.843	0.038
Endothelin receptor ET-A	111	3	97.4%	3	10	23.1%	161	5	97.0%	93.9%	3.1%	56.7%	0.899	0.016
Glucocorticoid receptor	302	50	85.8%	67	21	76.1%	1306	136	90.6%	89.0%	1.6%	76.6%	0.785	0.043
hERG	157	95	62.3%	151	177	46.0%	372	128	74.4%	63.0%	11.4%	46.3%	0.597	0.098
Histamine H1 receptor	98	28	77.8%	7	4	63.6%	161	8	95.3%	86.9%	8.3%	55.2%	0.881	0.073
Mu opioid receptor	167	22	88.4%	4	14	22.2%	372	25	93.7%	89.9%	3.8%	65.7%	0.881	0.023
Muscarinic acetylcholine receptor M1	88	21	80.7%	9	12	42.9%	166	19	89.7%	83.5%	6.2%	58.7%	0.788	0.028
Muscarinic acetylcholine receptor M2	136	16	89.5%	6	8	42.9%	350	13	96.4%	93.0%	3.4%	68.6%	0.877	0.033
Muscarinic acetylcholine receptor M3	104	22	82.5%	3	7	30.0%	146	11	93.0%	86.3%	6.6%	53.6%	0.849	0.048
Norepinephrine transporter	114	25	82.0%	3	11	21.4%	343	28	92.5%	87.8%	4.7%	70.8%	0.874	0.038
Serotonin 2a (5-HT2a) receptor	90	21	81.1%	1	10	9.1%	142	1	99.3%	87.9%	11.4%	54.0%	0.890	0.040
Serotonin 3a (5-HT3a) receptor	105	14	88.2%	4	1	80.0%	164	2	98.8%	94.1%	4.7%	57.2%	0.949	0.090
Serotonin transporter	89	24	78.8%	2	8	20.0%	155	10	93.9%	85.4%	8.5%	57.3%	0.848	0.026
Tyrosine-protein kinase LCK	37	21	63.8%	5	6	45.5%	56	10	84.8%	72.6%	12.3%	48.9%	0.750	0.042
Vasopressin V1a receptor	122	5	96.1%	0	5	0.0%	144	3	98.0%	95.3%	2.6%	52.7%	0.941	0.034
Average			**81.9%**			**35.2%**			**92.2%**	**85.9%**	**6.3%**	**59.3%**	**0.829**	**0.047**

Table 4.8: Results of the classification-by-similarity method on the test set chemicals predicted to be negative. “Out of domain” chemicals are those which do not have a Tanimoto similarity (between Morgan fingerprints) of at least 0.4 to any training true negative. “No alert but similar to active” chemicals are those which have a Tanimoto similarity coefficient (between Morgan fingerprints) of 0.7 or greater to any training active chemical. “Classified negatives” are chemicals that contain no alerts and which remain after removing the other two categories. Test set MCC has been calculated using classified negatives and all positive predictions, and compared to MCC when using all negative predictions and positive predictions.

The classification-by-similarity process leads to increases in NPV for all targets. An average increase in NPV of 6.3% is seen across all targets, which is a significant increase in the confidence of negative predictions. The increase in NPV observed here is greater than that observed for the classification-by-features method for nine of the ten targets which have been used in both methods, with alpha-2a adrenergic receptor being the one exception – an increase of 5.5% for classification-by-similarity compared to an increase of 5.6% for classification-by-features.

In the classification-by-similarity method, a larger proportion of chemicals were considered classified than in the classification-by-features method with a fingerprint string length of 4 096 bits. Acetylcholinesterase is an exception, with a smaller proportion of classified chemicals in the classification-by-features method. This is due to the presence of many bit clashes in acetylcholinesterase causing most chemicals (65%) to be classified.

Compared to classification-by-features, the classification-by-similarity method results in a larger proportion of chemical being considered classified negatives, and there is a larger NPV in the classified chemicals. Therefore, classification-by-similarity is the superior method.

The similarity to training active chemicals allows identification of chemicals that are similar to known active chemicals despite not containing any structural alerts. These chemicals are more likely to be active and so should be in their own category of negative prediction. A risk assessor should have low confidence in negative predictions which are in the “no alerts but like actives” category.

The requirement of minimum similarity to true negative chemicals in the training set means only chemicals which bare some resemblance to inactive chemicals seen in training of the model are considered classified. Chemicals which do not meet this requirement could be considered as too different to the training chemicals and therefore “out of domain” compared to the inactives used in training the model. Whilst these chemicals contain no structural alerts and are not similar to training active chemicals (according to Tanimoto similarity between fingerprints), they might be from a different region of chemical space to the training chemicals.

Chemicals considered to be “Classified Negatives” contain no structural alerts, are not similar to training active chemicals and are somewhat similar to at least one training inactive chemical. A risk assessor should have high confidence in this category of negative prediction.

In future work, rather than splitting negative predictions into three discrete categories, one could investigate creating a quantitative assessment of confidence in negative predictions by considering the largest Tanimoto similarity between fingerprints of the test chemical and the training actives, and between the test chemical and the training true negatives.

4.5. Conclusions

Increasing confidence in negative predictions is of vital importance, particularly in risk assessment. Williams *et al* have investigated two methods for increasing confidence in negative predictions from structural alert models for mutagenicity,[76] related to reactivity-driven MIEs. Methods inspired by that work have been applied to the structural alerts for Bowes targets, which are receptor binding MIEs.

Exclusion rules have been created for the structural alerts for Bowes Targets. These are features added to a structural alert to give a substructure which occurs only in inactive chemicals in the training set. There was found to be no increase in confidence in the negative predictions due to exclusion rules than in the negative predictions due to lack of structural alerts alone. Predicting the activity of chemicals which lie on the border between activity and inactivity is inherently difficult, so there is limited confidence in the negative predictions from exclusion rules.

Whether features, as defined by fingerprints, in test chemicals were present in the training set inactive chemicals was used to assess confidence in negative predictions. Test chemicals containing only features that are present in the training true negative chemicals were considered "classified" chemicals.

When MACCS fingerprints were used to define features, too few chemicals were considered unclassified for this approach to be useful. MACCS fingerprints code only 166 features and this was not enough to distinguish between chemicals.

Features were coded using bits in a Morgan fingerprint, but bit clashes were found to be a major problem when using short fingerprint strings. A greater confidence in negative predictions was seen in the classified chemicals in nine out of ten biological targets, but not in acetylcholinesterase. However, a large proportion of chemicals were considered unclassified. The performance of this method has been explained using the maximum Tanimoto similarity coefficient between the fingerprints of each test chemical and the training true negatives. Chemicals with a larger maximum Tanimoto similarity between fingerprints are more likely to be classified chemicals. The distribution of these values in the false negatives compared to the distribution in the true negatives has been used to explain the variation of NPV in the biological targets.

A new method for classifying negative predictions has been designed using Tanimoto similarity coefficients between Morgan fingerprints of a test chemical and the training set chemicals. Bit clashes are insignificant with this method as fingerprints are compared in a one-to-one way instead of the one-to-many way used in classification-by-features. Chemicals which are predicted to be negative by the structural alert models are split into three categories:

- “Similar to active” – chemicals which have a Tanimoto similarity coefficient of greater than 0.7 to any training active chemical. Whilst these chemicals do not contain any structural alerts, they are highly similar to known active chemicals, so are more likely to be false negatives. A risk assessor should have low confidence in this category of negative prediction. Averaging across the Bowes targets, only 35% of negative predictions in this category were true negatives.
- “Out of domain” – chemicals which do not have a Tanimoto similarity between fingerprints of at least 0.4 to any true negative in the training set. These chemicals do not contain any structural alerts and are not similar to training active chemicals, but they are not similar to any of the inactive chemicals seen in training the model. A risk assessor should have a fair amount of confidence in this category of negative prediction but be aware that chemicals in this category may be from a different region of chemical space to the training chemicals. Averaging across the Bowes targets, 82% of negative predictions in this category were true negatives.
- “Classified negative” – chemicals which have a Tanimoto similarity coefficient of at least 0.4 to any true negative in the training set and do not have a Tanimoto similarity of coefficient of greater than 0.7 to any training active. These chemicals do not contain any structural alerts, are not similar to any training active chemicals, and have a good similarity to a known inactive chemical. One should have high confidence in this category of negative prediction. Averaging across the Bowes targets, 92% of negative predictions in this category were true negatives.

In the “classified negatives” category of negative prediction, NPV was increased in the test sets of all targets. The increase was greater than that seen in the classification-by-features method, and more chemicals were considered classified. Hence, the classification-by-similarity method is the superior method for receptor binding MIEs, and it greatly increases the confidence in negative predictions.

Previous studies have used a method similar to the classify-by-features method to increase NPV in structural alert based-models for the reactivity driven MIEs related to mutagenicity[76] and skin sensitisation.[99] In these MIEs, presence of a single electrophilic feature can lead to activity. “Local” changes in molecular structure can lead to changes in activity. The classify-by-features approach identifies presence of unknown features, which represent unknown local changes, to increase confidence in negative predictions for reactivity drive MIEs.

Here, it has been shown that the classify-by-similarity method is more effective than the classify-by-features method for increasing NPV in structural alert-based models for receptor binding

MIEs. In these MIEs, chemicals need a specific combination of features in three-dimensional space. Activity is the result of several interactions across the molecule as a whole. Hence, the classify-by-similarity metric which considers similarities between molecules as a whole (a "global" metric) has been more effective in assessing confidence in negative predictions.

5. Generalisation of aromatic structural alerts

5.1. Reformatting structural alerts

The structural alerts generated by the automatic workflow are written in a format created by the MoSS node.[68] The substructures are written in a SMILES format, which can be read and used by all the relevant nodes within KNIME. However, this format could be problematic when used in other programs and for other purposes. It would be better to have the alerts written in a SMARTS format – a format widely accepted for writing substructures.

In theory, SMILES strings should be valid SMARTS strings, so no changes should be needed when treating a SMILES string as a SMARTS string. However, this is not the case for the structural alerts as outputted by the MoSS node, evidenced by a drop in the number of molecules contained by some alerts when the alerts are treated as SMARTS strings instead of SMILES strings. Whilst there is not a problem when using the structural alerts SMILES strings internally within KNIME (where the relevant nodes are capable of treating substructures as SMILES strings), there may be a problem if they were to be used externally as SMARTS strings (where programs may only be capable of treating substructures as SMARTS strings).

The issue is due to treatment of atoms at the end of straight chains within the substructure. Take for example carbon: MoSS will write "C" at the end of a straight chain to mean any carbon – aliphatic or aromatic. In SMARTS format, "C" means only an aliphatic carbon, not an aromatic carbon. Instead, "[C,c]" should be written, meaning aromatic or aliphatic carbon. The same issue arises for nitrogen atoms. This change needs to be made to the structural alerts as outputted by MoSS and the structural alerts should be treated as SMARTS so that they can be universally used by all programs. An example is shown in Figure 5.1.

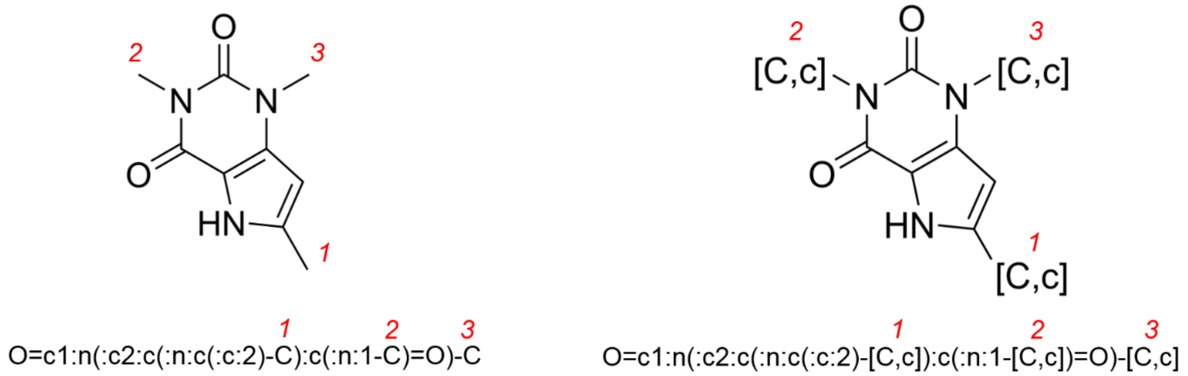

2
3
1
1 2 3
O=c1:n(:c2:c(:n:c(:c:2)-C):c(:n:1-C)=O)-C
2
3
[C,c]
[C,c]
1
[C,c]
1 2 3
O=c1:n(:c2:c(:n:c(:c:2)-[C,c]):c(:n:1-[C,c])=O)-[C,c]

Figure 5.1. An example of a structural alert with three terminal carbons, labelled "1", "2", and "3". On the left is the alert when written in SMILES format. On the right is the alert when written in SMARTS format.

Treating the MoSS formatted alerts as SMARTS strings without making this change will result in the alerts missing molecules that should be predicted positive by these alerts – molecules with aromatic structures at the end of the chains of the structural alert.

At the time, no existing programs directly dealing with this issue could be found. Using KNIME and java scripts, a new workflow has been created to make the outlined changes to the structural alerts.

Terminal carbons were identified within the SMILES string as being written as:

1. "-C)" anywhere in the string
2. "C-" at the start of the string
3. "-C" at the end of the string

The same rules apply for identifying terminal nitrogen atoms.

The MoSS formatted structural alert strings were inputted into the workflow. Each character in the string was split into an individual cell in an array, apart from "Cl" (chlorine atom) or "Br" (bromine atom) for which both characters occupy one cell. The sequence of cells was checked for cases where consecutive cells meet any of the three rules for terminal carbons. Where terminal carbons are identified, the cell containing "C" was replaced by "[C,c]". Similarly, terminal nitrogen atoms were identified and replaced with "[N,n]". The entire series of cells was then concatenated sequentially to give the new SMARTS string.

Structural alerts created by the automated workflow for the Bowes Targets were used to check this process. The structural alert SMILES strings were changed to corrected SMARTS strings using the process outlined above. Unlike the previous strings, the new SMARTS strings no longer miss molecules when treated as SMARTS strings. The predictions from structural alerts are consistent between the original SMILES strings used as SMILES (within KNIME nodes) and the SMARTS strings used as SMARTS strings (within any program).

Structural alert-based models can be generated using the automated workflow in SMILES format. The structural alerts can then be converted to SMARTS format using the workflow outlined here to give alerts in a universally accepted format that can be used in for all purposes.

5.2. Generalising aromatic structural alerts

For some targets, active chemicals with closely related aromatic substructures are found in the same data set. Researchers may have tried adding or removing heteroatoms in different positions within an aromatic system with the goal of investigating how each position affects activity (whilst maintaining aromaticity). Changing heteroatoms within an aromatic system has little effect on the three-dimensional shape of the system but can have an effect on the electronics of the ring and on the ability to form hydrogen bonds. Heteroatoms with lone pairs perpendicular to the aromatic ring, such as oxygen and certain nitrogen atoms (where the lone pair is not involved in aromatic bonding), can act as hydrogen bond acceptors. By investigating which positions within the aromatic system require such heteroatoms, we can attempt to identify where hydrogen bonds may be formed with the target binding site. This gives us understanding of how the MIE of receptor binding works. This ligand-based understanding can be compared with target-based understanding, such as X-ray crystal structures.

It would seem sensible to have the closely related aromatic substructures, such as those shown in Figure 5.2, contained within the same alert.

In SMARTS strings, the character “a” means any aromatic atom can be found at a particular position in the substructure. This feature allows one SMARTS string to contain several closely related aromatic substructures.

The algorithm of the MoSS node uses SMILES strings to build maximum common substructures. Hence, the SMARTS features cannot be used in the maximum common substructure search directly. We can however take the structural alerts written as SMILES strings (as generated by the automated workflow), convert them into SMARTS strings, and then use features of SMARTS strings to combine closely related aromatic substructures into one alert.

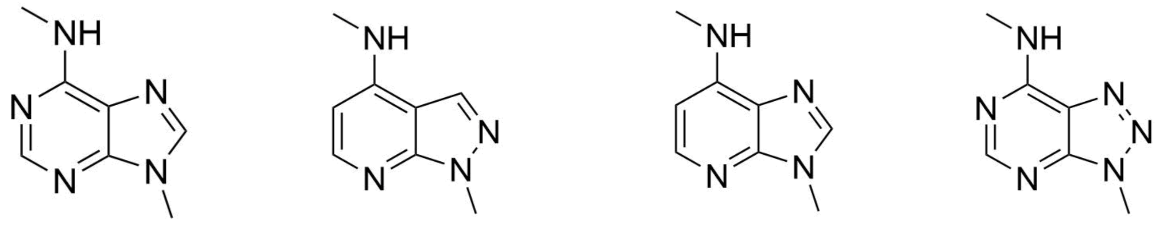


Figure 5.2. A series of closely related aromatic substructures found in the adenosine A2a receptor training set. The far-left substructure is identified as a structural alert, being contained by 375 active chemicals and 25 inactive chemicals. The other three substructures are not directly identified as structural alerts but are collectively contained by fifteen active chemicals.

A workflow has been designed for the aromatic generalisation process, starting with an existing structural alert and generalising aromatic atoms only if there is sufficient evidence within the data set to justify doing so. “Sufficient evidence” was defined by using Bayesian statistics, as done previously in the automated workflow for construction of structural alert-based models (see Section 2.2.1).

5.2.1. Method

The structural alerts in SMILES formats, as outputted by the automated workflow, were inputted into a new workflow. Alerts containing inorganic elements were removed – inorganic elements were identified as any that require square brackets to refer to the atom in SMILES notation. Structural alerts containing no aromatic features were identified as ones with no lower-case letters and were removed (note that non-aromatic substructures including Cl and Br would be exceptions to this rule, but this is unimportant as they will be unaffected by the rest of the workflow). Each character in the string was split into an individual cell in an array, apart from "Cl" (chlorine atom) or "Br" (bromine atom) for which both characters occupy one cell. As outlined in the previous section, terminal carbon and nitrogen atoms within the substructure were identified and converted to a SMARTS consistent representation ([C,c] and [N,n] respectively).

All aromatic atoms – identified as cells which were entirely lower case (so Cl and Br will not be included) – were replaced by the character "a" (meaning a general aromatic atom of any element) giving entirely generalised aromatic systems. This substructure was labelled as the "parent alert".

At this point, a loop began. Each "a" in the parent alert was individually returned to the original atom in that position whilst leaving the other "a"s, giving multiple substructures known as the "children alerts". The number of active and inactive chemicals containing the parent alert and each of the children alerts was found, from which Bayes Factor was calculated. If the parent alert has a Bayes Factor greater than all of the children alerts, it was output, and the loop was ended. If any of the children alerts have a Bayes Factor greater than or equal to the parent alert's Bayes Factor, the child alert with the greatest Bayes Factor becomes the new parent alert and was returned to the beginning of the loop. The loop was repeated until a parent alert was output or if the best performing child alert contained no "a"s, in which the case the original alert with no generalised aromatic atoms was output.

This process is shown graphically in Figure 5.3.

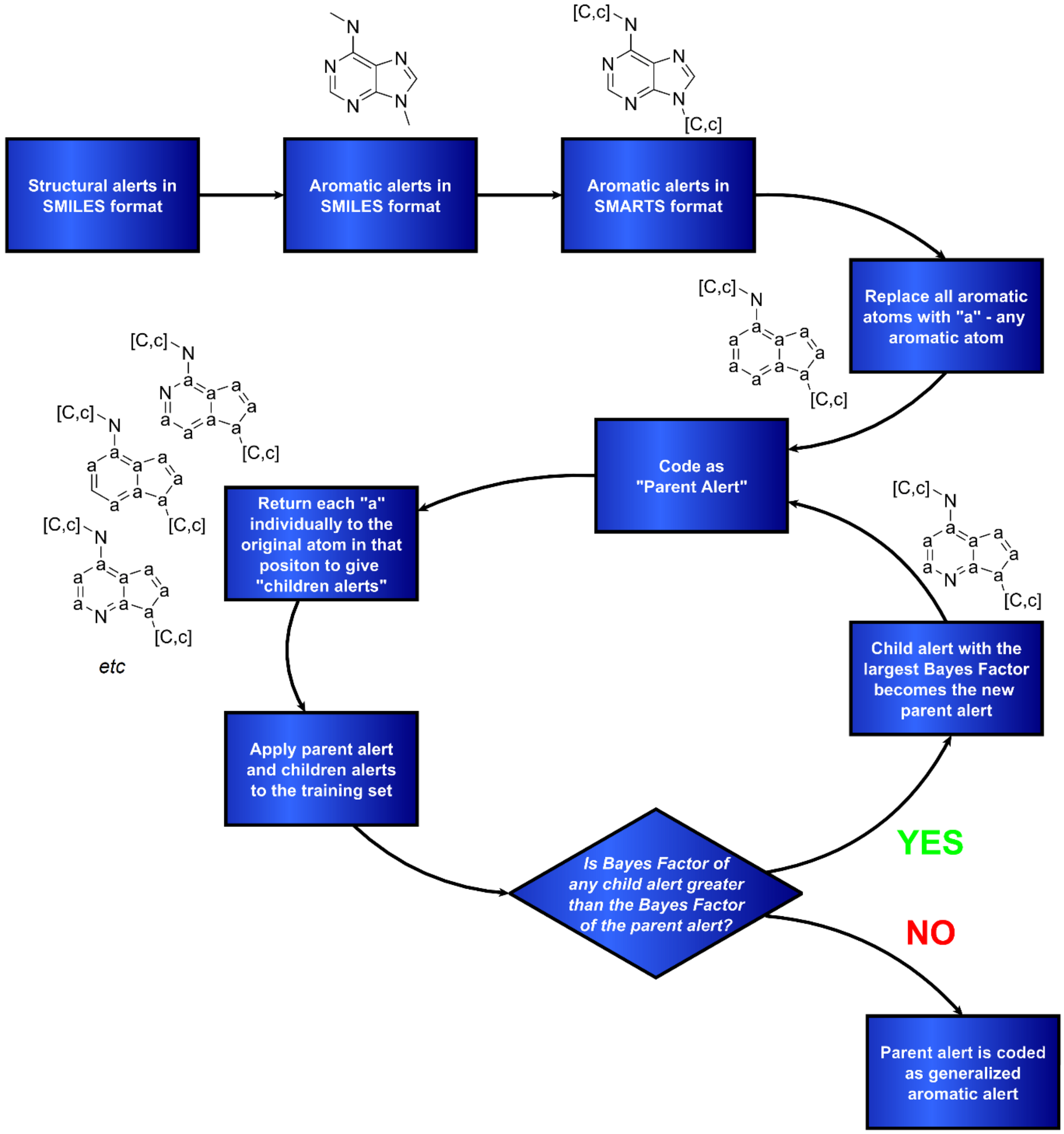

Structural alerts in SMILES format
Aromatic alerts in SMILES format
Aromatic alerts in SMARTS format
Replace all aromatic atoms with "a" - any aromatic atom
Code as "Parent Alert"
Return each "a" individually to the original atom in that positon to give "children alerts"
etc
Apply parent alert and children alerts to the training set
Is Bayes Factor of any child alert greater than the Bayes Factor of the parent alert?
YES
Child alert with the largest Bayes Factor becomes the new parent alert
NO
Parent alert is coded as generalized aromatic alert

Figure 5.3. An overview of the generalising aromatic alerts workflow. A structural alert is shown to give an example of how the workflow works.

5.2.1. Results and Discussion

The algorithm takes a structural alert and, if there is sufficient evidence, creates a SMARTS alert that will also be contained by additional closely related aromatic structures.

The outlined algorithm could be viewed as a method of steepest ascent, giving an estimation of the best performing (in terms of Bayes Factor) possible placement of generalised aromatic atoms within the substructure. To be certain that the output generalised structural alert is the best possible performing substructure, every possible combination of generalised aromatic atoms would have to be checked, which would be much more computationally expensive.

The user must select a value of theta in the Bayes Factor calculations for this algorithm. The considerations discussed previously (Section 2.2.3.1) regarding the choice of theta used in the automated workflow for creating structural alert-based also apply here – lower theta values will maximise true positives but also give more false positives than higher theta values, which will minimise false positives but give fewer true positives.

An example illustrating the generalisation of a structural alert is show in Figure 5.4. The structural alert is the same one as shown in Figure 5.2. All aromatic atoms in a structural alert are replaced with generalised aromatic atoms, potentially increasing the number of active and inactive chemicals containing the alert. The method of steepest ascent is followed to find which atoms should be kept as general aromatic atoms. In the example shown in Figure 5.4, only three positions have sufficient evidence to be kept as general aromatic atoms. All three of the closely related aromatic substructures shown in Figure 5.2 are contained by the final generalised aromatic alert.

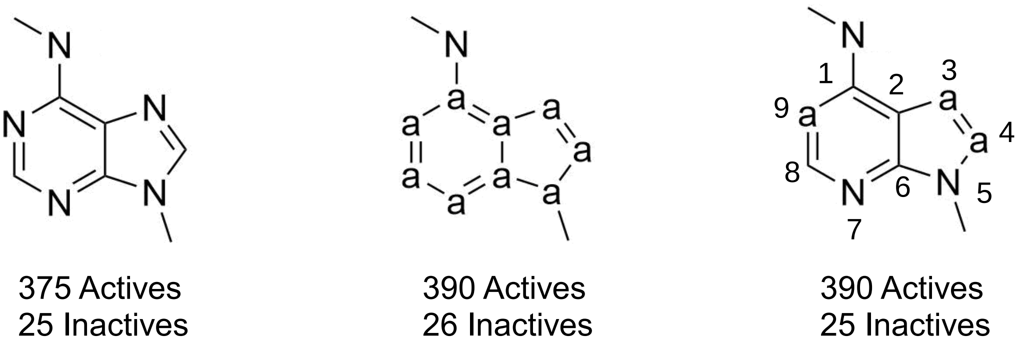


Figure 5.4. The left-hand substructure is a structural alert for the adenosine A2a receptor. The middle substructure is the fully generalised "parent alert" created from the structural alert. The right-hand substructure is the final result of the generalised aromatic structural alert algorithm. In this example, the final generalised structural alert is contained by the same number of active chemicals as the fully generalised "parent alert" but is contained by fewer inactive chemicals.

The process of generalising the alerts provides us with additional information as to how the chemicals may interact with the target's binding site. For example, take the substructures shown in Figure 5.4. In the original structural alert, the aromatic nitrogen atoms at positions labelled 3, 7 and 9 have lone pairs capable of acting as hydrogen bond acceptors (the lone pair of the nitrogen atom at position 5 is involved in the aromatic system and unavailable for forming hydrogen bonds). Looking at the non-generalised alert alone, we cannot confidently discern which aromatic atoms, if any, are involved in the binding mode. Looking at the generalised aromatic alert and the additional chemicals containing the alert, we see active chemicals in which position 9 is occupied by carbon, suggesting a hydrogen bond may not be formed from this position. Positions 1, 2, 6 and 8 are fixed as carbons so are not involved in hydrogen bonding. Position 7 is fixed as a nitrogen atom, suggesting it might be acting as a hydrogen bond acceptor. In fact, the one extra inactive chemical contained by the fully aromatised "parent alert" is a chemical where position 7 is a carbon, supporting the idea that a nitrogen is required in this position to act as a hydrogen bond acceptor. Position 5 is required to be a nitrogen to keep the structure aromatic. At least one of position 3 and 4 must be a nitrogen – there are cases where both positions are nitrogen atoms, and cases were one position is a nitrogen atom and the other a carbon atom, but no cases where both positions are carbon atoms. This suggests that if a hydrogen bond is formed at this side of the substructure, the hydrogen bond donor of the biological target is flexible enough to reach either position.

The nitrogen-based side group off position 1 is not in the aromatic ring and so is not affected by the generalisation process. The generalisation process therefore provides us no additional data on whether it will be involved in interactions or not – we have no data from active or inactive chemicals containing an atom other than nitrogen. The nitrogen lone pair is present in all chemicals and could be a hydrogen bond acceptor.

The predicted interactions are shown graphically in Figure 5.5.

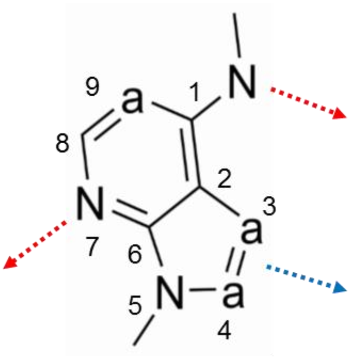


Figure 5.5: The predicted interactions of a generalised structural alert for the adenosine A2a receptor. The predicted interactions are derived by looking at the position of hydrogen bond acceptors in chemicals containing the generalised substructure. Red arrows indicate a hydrogen bond formed with good geometry and blue arrows indicate either a flexible or a non-essential hydrogen bond.

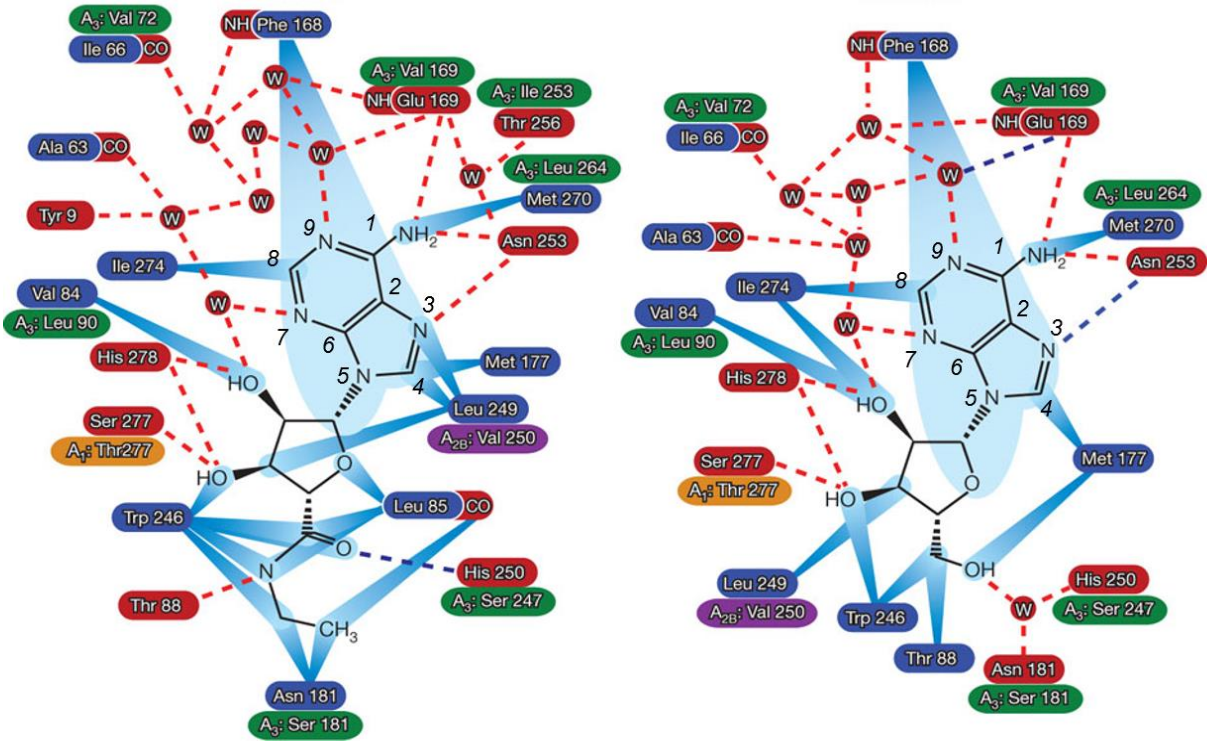


Figure 5.6: Key binding interactions derived from crystal structures for the adenosine A2a receptor bound to 5'-N-ethylcarboxamidoadenosine (NECA) (left) and adenosine (right). Red dashed lines represent hydrogen bonds with favourable geometry and blue dashed lines represent hydrogen bonds with unfavourable geometry (donor-acceptor distance of greater than 3.6 Å). Blue rays represent van der Waals contacts. Red "W" circles are water molecules. Amino acid residues within 3.9 Å of the chemical are highlighted surrounding it: residues highlighted in red form hydrogen bonds, and residues highlighted in blue form van der Waals contacts. Figure is taken from Lebon (2011).[101]

The information regarding the binding mode inferred from the generalised structural alert can be compared to crystal structures. Key binding interactions derived from crystal structures for the adenosine A2a receptor when bound to 5'-N-Ethylcarboxamidoadenosine (NECA) and adenosine were derived by Lebon *et al.*[101] These are shown in Figure 5.6. These chemicals are similar to the generalised structural alert previously discussed and shown in Figure 5.5. They key difference is the nitrogen group on position 1 in the structural alert includes a carbon which is not present in the chemicals. In chemicals containing the alert, that carbon can lead to larger groups, creating extra steric demands that are not present in NECA or adenosine, although it may be possible that the additional carbon side-group can rotate away from steric clashes. This caveat should be remembered when making comparisons between the crystal structures and the proposed interactions of the generalised substructure. Despite this, the hydrogen bond predicted to be formed from the nitrogen group on position 1 is seen in NECA and adenosine.

The nitrogen in position 3 is shown to form a hydrogen bond with good geometry (donor-acceptor distance of less than 3.6 Å) in NECA but with poor geometry (donor-acceptor distance of greater than 3.6 Å) in adenosine. The observation in adenosine seems to agree with the prediction that the hydrogen bond is either flexible or non-essential.

The nitrogen in position 7, predicted to form an essential hydrogen bond, is seen to form a hydrogen bond with water in NECA and adenosine. This is an isolated water molecule, so may be an important structural water.

Nitrogen in position 9, identified as a non-essential nitrogen in the generalised substructure, is seen in NECA and adenosine to form hydrogen bonds to a water molecule in a group of water molecules. This water fills an empty space in the receptor pocket, so the interaction may not be essential. The carbon side group in the generalised substructures – which is not observed in NECA or adenosine – might sit in this pocket, displacing the water to which a hydrogen bond is formed, supporting the idea that this is a non-essential interaction.

Overall, the crystal structures for NECA and adenosine appear to be consistent with the interactions predicted by looking at the generalised structural alert for the adenosine A2a receptor. This example shows how generalised structural alerts can be used to elucidate key interactions for receptor binding.

5.3. Effect of generalising aromatic substructures on model performance

As a generalised alert is contained by more active chemicals and has a higher Bayes Factor than the structural alert used for its derivation, one might expect generalising an entire set of structural alerts for a target would improve overall model performance.

Method

The aromatic generalisation process was applied to structural alerts for three targets: adenosine A2a receptor, acetylcholinesterase, and dopamine D2 receptor. The new generalised alerts replace the original structural alerts. For each target, two sets of structural alerts were used with different theta values:

- theta value 0.95, 5% maximum occurrence in inactive chemicals, lower bounds for an alert of two actives and one inactive
- theta value 0.51, 5% maximum occurrence in inactive chemicals, lower bounds for an alert of two actives and one inactive

The generalisation process was applied to the structural alert models, using the same theta as the workflow used to construct the structural alert models.

Results and Discussion

The performance metrics of the generalised and non-generalised models are shown in Table 5.1. In each case, generalising the structural alerts has resulted in many more false positives than true positives in both training and test sets. Consequently, accuracy and MCC decrease. Acetylcholinesterase with theta of 0.95 is an exception, for which the changes are very small.

				Training set								Test set							
Target	Theta	Model	Alerts	TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC
ADRA2A	0.95	Not Generalised	57	2879	121	78	1461	97.4%	92.4%	95.6%	0.903	941	49	45	452	95.4%	90.2%	93.7%	0.858
ADRA2A	0.95	Generalised	57	2912	300	45	1282	98.5%	81.0%	92.4%	0.833	967	100	19	401	98.1%	80.0%	92.0%	0.82
ADRA2A	0.51	Not Generalised	45	2889	159	68	1423	97.7%	89.9%	95.0%	0.889	950	58	36	443	96.3%	88.4%	93.7%	0.857
ADRA2A	0.51	Generalised	45	2951	889	6	693	99.8%	43.8%	80.3%	0.576	984	289	2	212	99.8%	42.3%	80.4%	0.567
AChE	0.95	Not Generalised	109	1846	314	158	1153	92.1%	78.6%	86.4%	0.72	523	121	87	375	85.7%	75.6%	81.2%	0.619
AChE	0.95	Generalised	109	1846	321	158	1146	92.1%	78.1%	86.2%	0.716	525	121	85	375	86.1%	75.6%	81.4%	0.622
AChE	0.51	Not Generalised	67	1895	433	109	1034	94.6%	70.5%	84.4%	0.684	555	157	55	339	91.0%	68.3%	80.8%	0.616
AChE	0.51	Generalised	67	1972	996	32	471	98.4%	32.1%	70.4%	0.428	592	343	18	153	97.0%	30.8%	67.4%	0.384
DRD2	0.95	Not Generalised	69	4197	136	65	719	98.5%	84.1%	96.1%	0.855	1382	59	52	223	96.4%	79.1%	93.5%	0.762
DRD2	0.95	Generalised	69	4228	248	34	607	99.2%	71.0%	94.5%	0.791	1414	98	20	184	98.6%	65.2%	93.1%	0.731
DRD2	0.51	Not Generalised	40	4209	184	53	671	98.8%	78.5%	95.4%	0.827	1391	75	43	207	97.0%	73.4%	93.1%	0.74
DRD2	0.51	Generalised	40	4231	263	31	592	99.3%	69.2%	94.3%	0.782	1418	101	16	181	98.9%	64.2%	93.2%	0.733

Table 5.1. The effect of aromatic generalisation of structural alerts on the structural alerts created using the automated workflow for construction of structural alert-based models. Two different theta value have been used in generating the structural alerts and the same theta values were used in the aromatic generalisation process. The process has been applied to three targets: adenosine A2a receptor (ADRA2A), acetylcholinesterase (AChE) and dopamine D2 receptor (D2R).

The structural alerts were originally created in a way that they were independent of each other, but this is not case after they have been generalised. The additional actives covered by the generalised alerts may already be covered by different alerts, leading to no increase in true positives when all alerts are used together. The generalised substructure may also contain more inactives compared to the original alert if there is a concurrent increase in the number of actives containing the generalised substructure that results in overall increase in Bayes factor. However, if all additional actives contain other structural alerts but the additional inactives do not contain other alerts, there will be an overall increase in false positives but no, or little, overall increase in true positives.

Using a lower theta value in generalising alerts should lead to more true positives but also more false positives for each alert. This can be seen in the results when comparing models constructed with low theta to models constructed with high theta. However, many of the extra true positives are not independent from the other alerts, leading to a larger increase in false positives than true positives for the overall model. Larger increases in false positives are seen in models with low theta compared to those with high theta.

5.3.1. Stepwise Regression

There is significant overlap in terms of the actives covered by different generalised alerts. The generalised alerts are not independent from each other, and some may be completely redundant, covering no unique actives when applied in combination with the other alerts. We could attempt to identify overlapping structural alerts by looking directly at the substructures. However, statistics provides a technique for identifying and removing redundant variables without having to examine structures – stepwise regression.

Method

A table was generated of each chemical in the training set, the binary activity of the chemical, and the structural alert each chemical contains. This data was fitted to a binomial generalised linear model where the alerts were the independent variables and the activity was the dependent variable (assumed to be a function of the alerts). Stepwise regression identified and removed structural alerts that were considered to have no significant unique contributions to predicting activity. Fitting of the data to a binomial generalised linear model and subsequent stepwise regression was done in RStudio.[102]

Results and Discussion

The stepwise regression process has been applied to both non-generalised and generalised alerts shown in Table 5.1.

The performance of the original alerts and the generalised alerts, before and after stepwise regression, was calculated and is shown in Table 5.2.

				Training set								Test set							
Target	Theta	Model	Alerts	TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC
ADRA2A	0.95	Not Generalised	57	2879	121	78	1461	97.4%	92.4%	95.6%	0.903	941	49	45	452	95.4%	90.2%	93.7%	0.858
ADRA2A	0.95	Not Generalised (SR)	55	2879	121	78	1461	97.4%	92.4%	95.6%	0.903	941	49	45	452	95.4%	90.2%	93.7%	0.858
ADRA2A	0.95	Generalised	57	2912	300	45	1282	98.5%	81.0%	92.4%	0.833	967	100	19	401	98.1%	80.0%	92.0%	0.82
ADRA2A	0.95	Generalised (SR)	44	2912	300	45	1282	98.5%	81.0%	92.4%	0.833	967	100	19	401	98.1%	80.0%	92.0%	0.82
ADRA2A	0.51	Not Generalised	45	2889	159	68	1423	97.7%	89.9%	95.0%	0.889	950	58	36	443	96.3%	88.4%	93.7%	0.857
ADRA2A	0.51	Not Generalised (SR)	45	2889	159	68	1423	97.7%	89.9%	95.0%	0.889	950	58	36	443	96.3%	88.4%	93.7%	0.857
ADRA2A	0.51	Generalised	45	2951	889	6	693	99.8%	43.8%	80.3%	0.576	984	289	2	212	99.8%	42.3%	80.4%	0.567
ADRA2A	0.51	Generalised (SR)	32	2951	889	6	693	99.8%	43.8%	80.3%	0.576	984	289	2	212	99.8%	42.3%	80.4%	0.567
AChE	0.95	Not Generalised	109	1846	314	158	1153	92.1%	78.6%	86.4%	0.72	523	121	87	375	85.7%	75.6%	81.2%	0.619
AChE	0.95	Not Generalised (SR)	108	1846	314	158	1153	92.1%	78.6%	86.4%	0.72	523	121	87	375	85.7%	75.6%	81.2%	0.619
AChE	0.95	Generalised	109	1846	321	158	1146	92.1%	78.1%	86.2%	0.716	525	121	85	375	86.1%	75.6%	81.4%	0.622
AChE	0.95	Generalised (SR)	105	1846	321	158	1146	92.1%	78.1%	86.2%	0.716	525	121	85	375	86.1%	75.6%	81.4%	0.622
AChE	0.51	Not Generalised	67	1895	433	109	1034	94.6%	70.5%	84.4%	0.684	555	157	55	339	91.0%	68.3%	80.8%	0.616
AChE	0.51	Not Generalised (SR)	67	1895	433	109	1034	94.6%	70.5%	84.4%	0.684	555	157	55	339	91.0%	68.3%	80.8%	0.616
AChE	0.51	Generalised	67	1972	996	32	471	98.4%	32.1%	70.4%	0.428	592	343	18	153	97.0%	30.8%	67.4%	0.384
AChE	0.51	Generalised (SR)	64	1972	996	32	471	98.4%	32.1%	70.4%	0.428	592	343	18	153	97.0%	30.8%	67.4%	0.384
DRD2	0.95	Not Generalised	69	4197	136	65	719	98.5%	84.1%	96.1%	0.855	1382	59	52	223	96.4%	79.1%	93.5%	0.762
DRD2	0.95	Not Generalised (SR)	68	4197	136	65	719	98.5%	84.1%	96.1%	0.855	1382	59	52	223	96.4%	79.1%	93.5%	0.762
DRD2	0.95	Generalised	69	4228	248	34	607	99.2%	71.0%	94.5%	0.791	1414	98	20	184	98.6%	65.2%	93.1%	0.731
DRD2	0.95	Generalised (SR)	55	4226	236	36	619	99.2%	72.4%	94.7%	0.799	1414	94	20	188	98.6%	66.7%	93.4%	0.741
DRD2	0.51	Not Generalised	40	4209	184	53	671	98.8%	78.5%	95.4%	0.827	1391	75	43	207	97.0%	73.4%	93.1%	0.74
DRD2	0.51	Not Generalised (SR)	40	4209	184	55	671	98.8%	78.5%	95.4%	0.827	1391	75	43	207	97.0%	73.4%	93.1%	0.74
DRD2	0.51	Generalised	40	4231	263	31	592	99.3%	69.2%	94.3%	0.782	1418	101	16	181	98.9%	64.2%	93.2%	0.733
DRD2	0.51	Generalised (SR)	30	4226	248	36	607	99.2%	71.0%	94.4%	0.79	1414	95	20	187	98.6%	66.3%	93.3%	0.739

Table 5.2: The effect of aromatic generalisation of structural alerts created using the automated workflow for construction of structural alert-based models. Stepwise regression (SR) is applied to both the original alerts and the aromatic generalised alerts, resulting in removal of redundant alerts. Two different theta value have been used in generating the structural alerts and the same theta values were used in the aromatic generalisation process. The process has been applied to three targets: adenosine A2a receptor (ADRA2A), acetylcholinesterase (AChE) and dopamine D2 receptor (D2R).

When stepwise regression was applied to the non-generalised alerts, little change to the models is seen. No more than two alerts are removed and in all cases the performance is unchanged. This shows that the iterative structural alert selection method in the automated workflow does an excellent job of choosing independent substructures.

There were larger decreases in the number of alerts when regression was applied to the generalised alerts, due to the generalisation process creating alerts which were not independent of each other. Large decreases in number of alerts were seen for the adenosine A2a and dopamine D2 receptors. This suggests the presence of closely related aromatic structures which can be represented by fewer generalised alerts than non-generalised alerts. However, there were only small decreases in number of alerts for the acetylcholinesterase models, suggesting there were few aromatic structures which could be generalised.

Despite decreases in the number of structural alerts, stepwise regression only increased model performance in the generalised alerts of the dopamine D2 receptor. In the models for the other two targets, no change in any performance statistic was seen. The removed alerts were redundant, being contained by no unique actives or inactives.

There are clear benefits to using generalised aromatic alerts, in terms of interpreting how the substructure might be involved in the binding mode, and in representing similar structures in fewer alerts. However directly applying the generalisation process to the original structural alerts did not improve the performance of the overall model as the resulting alerts were no longer independent from each other. Stepwise regression allows identification of redundant structural alerts, but it did not result in significant increases in performance.

5.4. Integrating generalised aromatic structures into the automated workflow for structural alert-based models

In the automated workflow for structural alert generation, after each structural alert is chosen, any active chemicals containing the alert are removed from the training set. This iterative method leads to alerts which are independent from each other. To create independent generalised alerts, the generalisation process can be included within the iterative steps of the automated workflow.

5.4.1. Method

The data sets used here are those for the Bowes targets with data from ToxCast and ChEMBL, created previously (Section 2.1).

A training set was input into the workflow and the MoSS node[68] was used to generate maximum common substructures. For each substructure, Bayes factor was calculated with a user-chosen theta value. Only the 65 substructures with the highest Bayes factors were kept as it was too computationally expensive to use more (aromatic generalisation can take up to five minutes per aromatic substructure and doing this more than 65 times for up to 135 cycles of picking structural alerts becomes time consuming). They were converted to SMARTS formats and each substructure was run through the aromatic generalisation algorithm, as outlined previously. The result of this was the replacement of atoms in aromatic substructures with generalised aromatic atoms where an increase in Bayes Factor was observed. For the 65 substructures, some of which may have generalised aromatic structures, accurate occurrences in the active and inactive chemicals were found, and Bayes Factor was calculated. The substructure with the largest Bayes Factor was coded as a structural alert if it meets the user's lower bounds for an alert (minimum number of active chemicals contained by the alert and minimum Bayes Factor). All active chemicals containing the structural alert were removed from the training set and the process was repeated until no substructures met the lower bounds for an alert. The list of structural alerts was output and applied to the test set to give performance statistics.

A graphical overview of the method is shown in Figure 5.7.

This workflow has been applied to the data sets for the Bowes Targets using parameters of: theta value 0.95, 5% maximum occurrence of a substructure in the inactives, and lower bounds for an alert of two actives and one inactive.

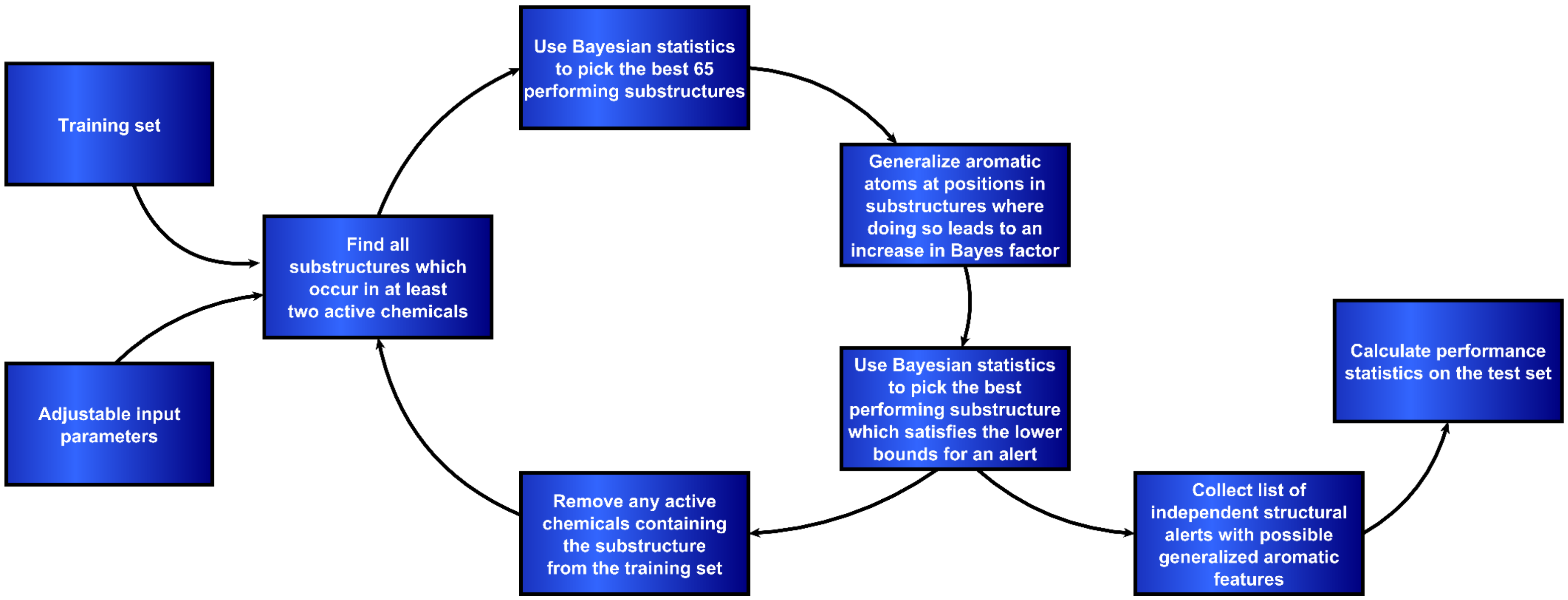

Training set
Adjustable input parameters
Find all substructures which occur in at least two active chemicals
Use Bayesian statistics to pick the best 65 performing substructures
Generalize aromatic atoms at positions in substructures where doing so leads to an increase in Bayes factor
Use Bayesian statistics to pick the best performing substructure which satisfies the lower bounds for an alert
Remove any active chemicals containing the substructure from the training set
Collect list of independent structural alerts with possible generalized aromatic features
Calculate performance statistics on the test set

Figure 5.7. Graphical overview of the automated workflow that integrates generalisation of aromatic substructures into construction of independent structural alerts.

5.4.2. Results and discussion

The automated workflow with integrated generalisation of aromatic substructures was applied to Bowes Targets. However, this workflow is computationally expensive in terms of memory, and so it was not possible to generate data for the largest data set – hERG. The results for the other targets are shown in Table 5.3.

The overall performances of the models were not significantly changed, as shown by the small changes in test set MCC. Importantly, the number of alerts for each target's model decreases. On average, the new alerts were contained by more active chemicals, giving the user more confidence in active predictions derived from these alerts.

Particularly large decreases in number of alerts were seen in three targets: adenosine A2a receptor, dopamine D2 receptor, and tyrosine-protein kinase LCK. The largest increases in test set MCC were also seen for these targets. The active chemicals for these targets likely contain closely related aromatic structures which were better modelled by a single generalised structural alert than multiple non-generalised alerts.

The mu opioid receptor is the only target for which there was an increase in the number of alerts. In this target, the alert chosen in the first iteration was the same alert as the one chosen in the original automated workflow, but part of the substructure was aromatically generalised. The partially generalised alert was contained by more active chemicals than the original. These additional active chemicals were removed from the training set and as a result, a different series of subsequent structural alerts was chosen in the next iterations. At each iteration, the substructure with the largest Bayes factor in the generalised alert workflow was chosen, but unusually, this resulted in more alerts to cover a smaller number of remaining active chemicals (although it hits less false positive chemicals in doing so). Whilst this is unlikely, it is not impossible. The increase in structural alerts in the mu opioid receptor therefore represents an unlikely exception to the general decrease in structural alerts seen in all other receptors.

Overfitting is the use of more terms in a model than is necessary.[103] Hence, a model using fewer structural alerts to cover the same number of active chemicals is less overfitted. By reducing the number of alerts in the structural alert models without changing performance, the models become less overfitted. In this way, the generalised structural alert-based models are improvements on the non-generalised structural alert models.

Target	Alerts	Training								Test								ΔAlerts	ΔTest MCC
		TP	FP	FN	TN	SE	SP	ACC	MCC	TP	FP	FN	TN	SE	SP	ACC	MCC		
Acetylcholinesterase	135	1839	150	165	1317	91.8%	89.8%	90.9%	0.814	518	62	92	434	84.9%	87.5%	86.1%	0.721	-3	0.000
Adenosine A2a receptor	42	2890	126	67	1456	97.7%	92.0%	95.7%	0.906	954	51	32	450	96.8%	89.8%	94.4%	0.874	-15	0.016
Alpha-2a adrenergic receptor	63	595	26	54	743	91.7%	96.6%	94.4%	0.887	158	11	38	234	80.6%	95.5%	88.9%	0.778	-4	0.014
Androgen receptor	107	1499	89	490	5380	75.4%	98.4%	92.2%	0.797	435	42	213	1773	67.1%	97.7%	89.6%	0.722	-8	-0.001
Beta-1 adrenergic receptor	20	934	85	26	721	97.3%	89.5%	93.7%	0.875	277	38	24	238	92.0%	86.2%	89.3%	0.785	-6	-0.020
Beta-2 adrenergic receptor	131	1255	89	207	1412	85.8%	94.1%	90.0%	0.803	358	72	125	441	74.1%	86.0%	80.2%	0.606	-4	-0.004
Delta opioid receptor	48	2186	118	54	786	97.6%	86.9%	94.5%	0.865	732	53	34	262	95.6%	83.2%	92.0%	0.802	-3	-0.002
Dopamine D1 receptor	68	881	55	148	1429	85.6%	96.3%	91.9%	0.833	248	26	75	481	76.8%	94.9%	87.8%	0.743	-4	0.003
Dopamine D2 receptor	44	4208	158	54	697	98.7%	81.5%	95.9%	0.846	1402	61	32	221	97.8%	78.4%	94.6%	0.796	-25	0.034
Dopamine transporter	64	1749	119	121	1326	93.5%	91.8%	92.8%	0.853	565	46	76	426	88.1%	90.3%	89.0%	0.779	-1	0.002
Endothelin receptor ET-A	14	941	60	21	799	97.8%	93.0%	95.6%	0.911	304	24	19	269	94.1%	91.8%	93.0%	0.860	-1	0.000
Glucocorticoid receptor	112	1828	104	446	5139	80.4%	98.0%	92.7%	0.824	555	60	189	1670	74.6%	96.5%	89.9%	0.755	-12	0.016
hERG																			
Histamine H1 receptor	36	915	67	48	755	95.0%	91.8%	93.6%	0.870	283	29	30	255	90.4%	89.8%	90.1%	0.802	-2	-0.017
Mu opioid receptor	60	2576	102	85	1621	96.8%	94.1%	95.7%	0.910	883	36	67	549	92.9%	93.8%	93.3%	0.860	4	0.001
Muscarinic acetylcholine receptor M1	53	1418	99	66	838	95.6%	89.4%	93.2%	0.856	490	44	41	261	92.3%	85.6%	89.8%	0.780	-5	0.013
Muscarinic acetylcholine receptor M2	49	1156	82	69	1424	94.4%	94.6%	94.5%	0.888	376	31	35	496	91.5%	94.1%	93.0%	0.857	-2	0.000
Muscarinic acetylcholine receptor M3	47	1132	72	58	767	95.1%	91.4%	93.6%	0.868	318	25	29	250	91.6%	90.9%	91.3%	0.824	-2	-0.004
Norepinephrine transporter	57	2106	121	117	1328	94.7%	91.6%	93.5%	0.864	630	45	59	447	91.4%	90.9%	91.2%	0.820	-7	-0.003
Serotonin 2a (5-HT2a) receptor	49	2756	88	41	683	98.5%	88.6%	96.4%	0.892	932	32	28	231	97.1%	87.8%	95.1%	0.854	-3	0.005
Serotonin 3a (5-HT3a) receptor	27	317	10	31	768	91.1%	98.7%	96.4%	0.914	88	5	16	272	84.6%	98.2%	94.5%	0.859	0	0.012
Serotonin transporter	53	3003	95	54	757	98.2%	88.8%	96.2%	0.887	945	33	41	250	95.8%	88.3%	94.2%	0.834	-9	-0.002
Tyrosine-protein kinase LCK	27	1247	82	36	323	97.2%	79.8%	93.0%	0.803	427	25	22	94	95.1%	79.0%	91.7%	0.748	-23	0.023
Vasopressin V1a receptor	15	446	21	10	766	97.8%	97.3%	97.5%	0.947	150	6	13	266	92.0%	97.8%	95.6%	0.907	0	0.000
Average	**57**	**1647**	**88**	**107**	**1358**	**93.4%**	**91.9%**	**94.1%**	**0.866**	**523**	**37**	**58**	**447**	**88.6%**	**90.2%**	**91.1%**	**0.799**	**-6**	**0.004**

Table 5.3: Performance metrics for the models built by integrating generalisation of aromatic substructures into the automated workflow for construction of independent structural alerts. The model has been applied to 23 Bowes targets with ToxCast and ChEMBL data available. A model was not constructed for hERG as applying the workflow to this particularly large data set is too memory exhaustive. The number of structural alerts and test set MCC values have been compared to the same values for models constructed without generalisation of aromatic substructures, using the same parameters (theta 0.95, 5% maximum occurrence in the inactive chemicals, and lower bounds for an alert of two actives and one inactive).

5.5. Conclusions

A workflow has been designed to convert the format of the structural alerts, as outputted by the MoSS node in KNIME, to a widely accepted SMARTS. This allows the structural alerts to be used for external purposes, increasing the impact of the models.

A process for generalising aromatic structural alerts has been outlined. Specific aromatic atoms within a structural alert are replaced with generalised aromatic atoms where there is sufficient data to support doing so, guided by use of Bayesian statistics.

Examining the generalised structural alerts gives more information on the mechanisms involved in the chemicals binding to the target. Predictions of key receptor binding interactions have been made by examining a generalised structural alert for the adenosine A2a receptor. These predictions were found to be consistent with binding interactions derived from crystal structures for similar chemicals. This process helps to understand how a receptor-binding MIE may occur, and this understanding helps to make better models and predictions for the MIE.

Generalising aromatic structural alerts allows more chemicals with similar structures to be represented by the same alert, in theory reducing the number of alerts required to model the data and reducing overfitting. However, directly applying the aromatic generalisation process to the original structural alert models did not improve performance as the resulting alerts were no longer independent of each other. Stepwise regression was used to remove alerts which became redundant after the generalising process, but this was still not an optimal way to produce high performing, independent, generalised alerts.

The aromatic generalisation process has been incorporated into the automated workflow for generating structural alerts, generalising aromatic substructures before selecting the best substructure to become a structural alert in each iteration. This results in construction of models containing independent, aromatically generalised structural alerts.

Applying this new workflow to the Bowes targets data sets created models with performance statistics comparable to the non-generalised structural alert models, but with significantly fewer structural alerts for each model. The use of fewer structural alerts is indicative of a less overfitted model. In this way, the generalised structural alert-based models represent important improvements over the non-generalised structural alert-based models.

6. Pharmacophore models from structural alerts

6.1. Introduction

Receptor binding MIEs can be caused by non-covalent interactions between chemicals and biological targets, such as binding sites on proteins or active sites on enzymes. These interactions include ionic attractions, hydrogen bonds, van der Waals forces, pi stacking of aromatic rings, and hydrophilic and hydrophobic interactions. Within the receptor binding site or active site, there are often specific steric requirements, and so a specific three-dimensional conformation of interaction features is required to undergo the receptor binding MIE. Pharmacophore models attempt to describe the required features and their required position in three-dimensional space. Predictions of biological activity of chemicals are made by seeing how well the chemical can fit to the conformation of features described by the pharmacophore.

It is often possible to bind at the same receptor through different sets of interactions, known as binding modes. When creating a pharmacophore model from ligands, chemicals should be chosen that act through the same binding mode, and these should be as diverse as possible so that the most important common features can be discerned from other features.

The structural alerts for receptor binding MIEs created in this project are two-dimensional substructures which have statistically been found to be associated with binding activity. The structural alerts often describe a central structural scaffold of a series of chemicals. Whilst a structural alert does not define any features outside of the substructure, it often defines the specific conformation of these features. The chemical groups which make up the features outside of the alert, and the groups that link the features to the alert, vary between the active chemicals. However, the features and their relative three-dimensional positions should not change significantly if they are involved in important interactions in the same binding mode. In the Chapter “generalising aromatic substructures” (Chapter 5), a protocol is described which helps to identify the features within the structural alert that are required for activity, where data is available.

There is clear synergy between the concepts in structural alerts and pharmacophore modelling. Pharmacophores expand upon the structural alerts for receptor binding MIEs, moving from statistically identified, two-dimensional substructures to three-dimensional arrangements of features involved in receptor binding interactions. Pharmacophores attempt to directly describe the mechanism of a receptor binding mode. Pharmacophore modelling is computationally expensive as three-dimensional conformations need to be generated and overlaid in a way that leads to overlap of features. It also requires careful selection of chemicals which act through the same binding mode. Structural alerts often define a scaffold common to many chemicals, holding other features in a specific conformation. Overlaying active chemicals by a common structural alert provides a good starting point for identification of common features. Furthermore, chemicals containing the same structural alert are likely to act through the same binding mode. Therefore, chemicals containing the same structural alert provide an excellent starting point for pharmacophore generation.

The disadvantage of using chemicals that share a structural alert is a lack of diversity in the chemicals. If the structural alert contains many features within the substructure itself, it will not be possible to identify which of the features are involved in receptor binding interactions and all are equally likely to be included in the pharmacophore model. The generalising aromatic structural alert process (Chapter 5) helps to remove this issue for aromatic atoms where possible, identifying which atoms are present in all chemicals and hence likely to be required for activity. Using generalised structural alerts is therefore important in pharmacophore generation.

Even with the generalised structural alerts, one might have concerns about a pharmacophore built from chemicals containing one structural alert being too specific to that alert. Having built the pharmacophore on a first structural alert, the model can be applied to chemicals containing other alerts. If multiple chemicals containing a different alert fit the pharmacophore model well, it is likely this new alert and the first alert describe the same binding mode. Chemicals containing these alerts can be combined into a more diverse set of chemicals and used to build an updated pharmacophore model.

In this chapter, pharmacophore modelling of chemicals with the same structural alerts has been explored.

6.2. Method

An overview of the methodology involved in pharmacophore construction is shown in Figure 6.1. Many automated algorithms for ligand-based pharmacophore model construction are commercially available,[52] and it would be pragmatic to use one here. In this work, the HipHop[104] algorithm for pharmacophore generation has been used via the "Common Feature Pharmacophore Generation From a Set of Ligands" feature in Biovia's Discovery Studio (19.1 Client).[105]

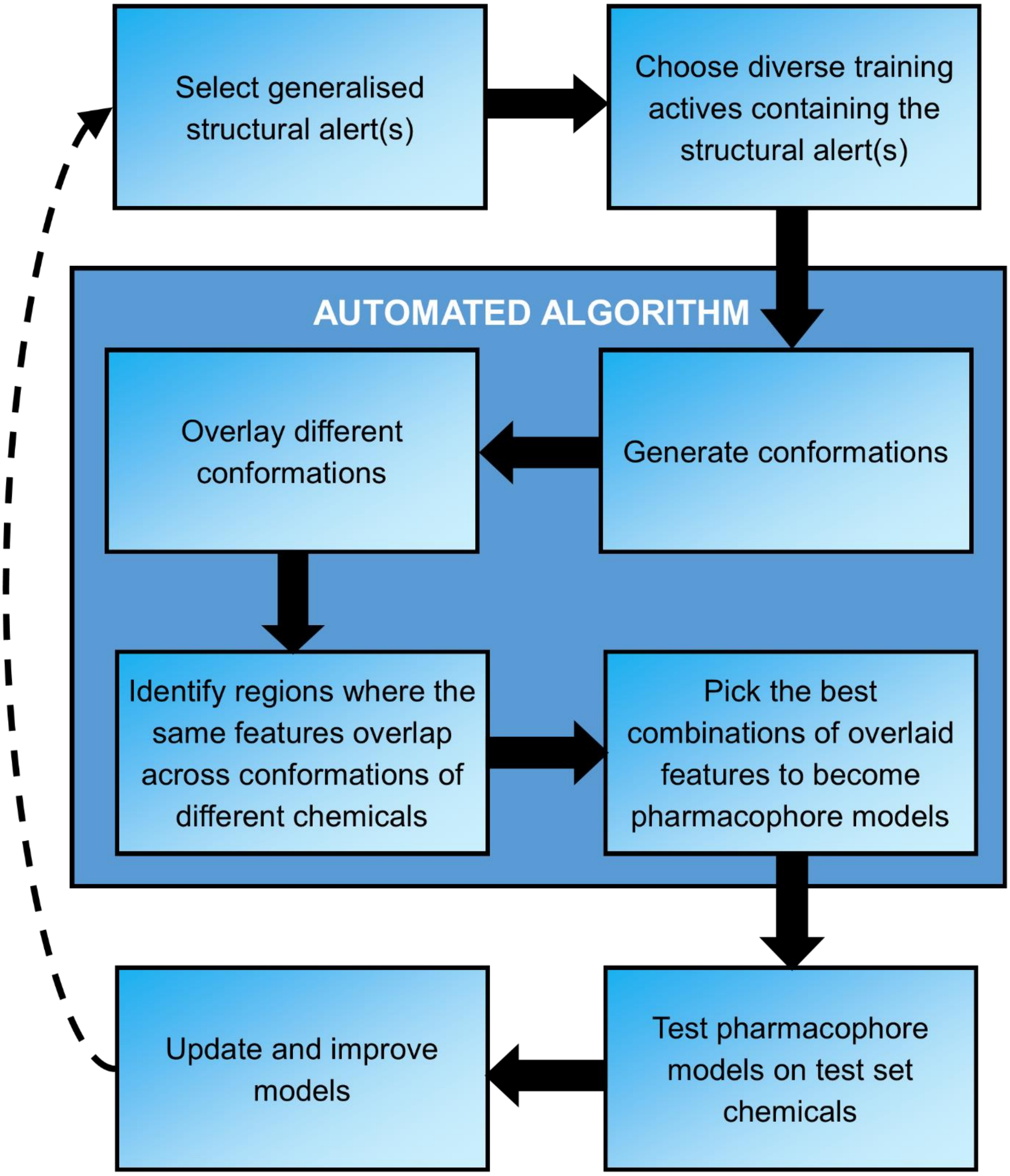

Select generalised structural alert(s)
Choose diverse training actives containing the structural alert(s)
AUTOMATED ALGORITHM
Overlay different conformations
Generate conformations
Identify regions where the same features overlap across conformations of different chemicals
Pick the best combinations of overlaid features to become pharmacophore models
Update and improve models
Test pharmacophore models on test set chemicals

Figure 6.1: A graphical overview of the proposed pharmacophore construction process.

In this work, the adenosine A2a receptor was chosen as the biological target to be focused on because many of the structural alerts could be generalised by the aromatic generalisation process, and it is a target with available crystal structures to which pharmacophore models can be compared. The same ChEMBL/ToxCast training set and test set created previously (Section 2.1) were used here.

The structural alerts created by the automated workflow (Section 2.2) with the "Risk Assessment" parameters (theta 0.95, 1% maximum occurrence of an alert in the inactive chemicals, and lower bounds for an alert of two actives and one inactive) were used. Compared to alerts created by the workflow with different parameters, these were the most specific structural alerts, containing the fewest false positives. The aromatic generalisation process (chapter 5), with a theta value of 0.95, was applied to these alerts to create generalised structural alerts.

A structural alert (generalised or not) was chosen to be the basis of pharmacophore construction. The structural alert (or alerts) was applied to the training set to find all active chemicals containing the alert. A diverse subset of ten chemicals was chosen from these chemicals with the RDKit Fingerprint Diversity node within KNIME,[83] which picked diverse chemicals using Tanimoto distances between fingerprints. The picking was done using the MaxMin algorithm.[104] Ten chemicals were chosen, which gives a sample size large enough to provide diversity between alert-containing chemicals but not so large that the pharmacophore construction algorithm was too computationally expensive.

The following steps were completed within the "Common Feature Pharmacophore Generation from a Set of Ligands" feature in Biovia's Discovery Studio (19.1 client). A maximum of 500 conformations were generated for each chemical within an energy threshold of 30 kcal/mol using the "BEST" algorithm. In each chemical, atoms which can act as pharmacophore features were identified. The features used here were hydrogen bond donors (HBDs), hydrogen bond acceptors (HBAs), aromatic rings, positive and negative ionisable groups, and hydrophobic groups. Using the HipHop algorithm,[104] the generated conformations were overlaid to find regions where features overlap across different conformations, from which pharmacophore models were built. Features must be at least 0.2 Å apart. Pharmacophore models were generated for all possible combinations of overlapping features from different conformations. Each model was ranked based on how well the molecules map onto the proposed pharmacophore, and the highest-ranking models were output.

There are additional options to define which features are required in any output pharmacophores and how many molecules must completely or partially map to the pharmacophore, giving the user more control over which pharmacophores are outputted. In this work the requirements of the pharmacophore model were:

- pharmacophore models must have at least four features
- each feature in the pharmacophore can miss a maximum of one training chemical
- at least nine of the ten active molecules must map to all features
- no chemicals can miss all features entirely.

The pharmacophore model is tested by seeing how well chemicals fit to it. For each pharmacophore, four groups of chemicals are used for testing:

1) The ten training active chemicals from which the model was built
2) 50 test active chemicals containing the alert(s) from which the model was built. If less than 50 are in this category, all chemicals are used.
3) 50 test active chemicals which do not contain the alert(s) from which the model was built
4) 50 test inactive chemicals, with the additional requirement that each must contain at least one HBA, HBD, and aromatic ring – inactive chemicals containing none of these will not be capable of fitting to the model and so would not be a fair comparison.

The 50 chemicals selected in from each category are picked with the RDKit Fingerprint Diversity node.

Chemicals are fitted to the pharmacophore model using the "Ligand Pharmacophore Mapping" function in Biovia's Discovery Studio (19.1 client). Within this function, conformations are generated for each chemical using the same setting as used previously in pharmacophore generation and the conformations are fitted to the pharmacophore model using Kabsch's fitting algorithm.[106] "Flexible fitting" is used, allowing conformations to be slightly manipulated (within an energy limit) to better fit the pharmacophore's features. Each conformation is given a score for how well it fits the model, and the best scoring conformation for each chemical is outputted.

After constructing each pharmacophore model, a cut-off was applied to the fit values. Chemicals that have a fit value higher than the cut-off are predicted to be active and chemicals with a lower fit value are predicted to be inactive. The cut-off was the lowest fit value that was not an outlier in the training active chemicals from which the pharmacophore was generated. Outliers were defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile.

6.3. Results and Discussion

A selection of pharmacophore models which demonstrate the success and limitations of this approach are shown in this section.

Each model is built from ten training active chemicals containing a structural alert. The models are tested on the four groups outlined above. A good model should result in high fit values for the training active chemicals with the alerts and the test active chemicals with the alert, whilst fit values in the inactive chemicals should be low. If the pharmacophore model describes the key features required for a particular binding mode and it is not too specific to the structural alert from which it is built, some active chemicals which do not contain the alert but act through the same binding mode should also have high fit scores.

6.3.1. Pharmacophores from non-generalised structural alerts

6.3.1.1. Alert054

Alert054 is shown in Figure 6.2. The structural alert has a large central core and specifies the position of three side groups, creating a defined scaffold for other features outside of the alert. Hence it should form a good template for a pharmacophore model. The two nitrogen atoms with available lone pairs (the two top right nitrogen atoms in Figure 6.2) which can act as HBDs are in all training chemicals and therefore are likely to be in the pharmacophore model.

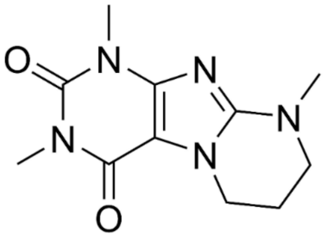


Figure 6.2: "Alert054" – A structural alert for the adenosine A2a receptor, created by the automated workflow for construction of structural alert-based models ("Risk Assessment" parameters). In the training set, this alert is contained by 33 active chemicals and two inactive chemicals.

The ten training active chemicals containing Alert054 from which the pharmacophore model was built are shown in Figure 6.7.

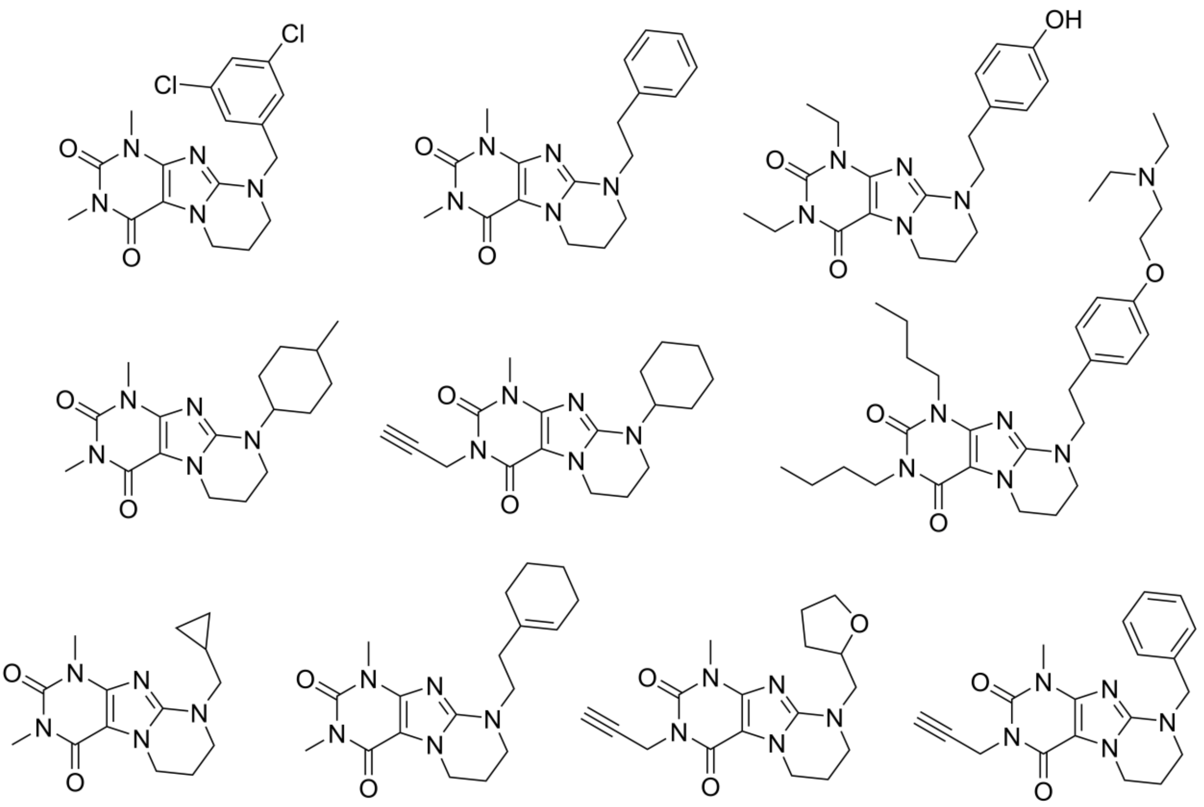


Figure 6.3: The ten training active chemicals from which the pharmacophore model was built. These chemicals are chosen from all training active chemicals containing Alert054 by RDKit's Fingerprint Diversity node in KNIME.

A pharmacophore model, shown in Figure 6.4, has been built from ten diverse active chemicals containing Alert054. The two oxygen atoms in the alert form two HBAs in the pharmacophore model. These provide two fixed points common to all chemicals containing Alert054, ensuring that all training chemicals are in the same alignment when fitted to the model. Looking at the best fitting conformations of the training chemicals to the pharmacophore model, we can see that all chemicals are aligned with each other. With this, we have more confidence that the pharmacophore algorithm has accurately overlaid all training chemicals and built a model which accurately represents the data. However, neither of the two nitrogen atoms present in the structural alert which can act as HBAs are identified as HBAs in the pharmacophore model. This is likely due to the maximum limit of ten features in a single model, but it shows that the HipHop algorithm may miss some common features. Looking at the chemicals in Figure 6.4, it seems that the pharmacophore model would have a better fit to most chemicals if the top-left hydrophobic group was lower in position, although the best-fitting chemical may not fit that new model as well. It seems like the pharmacophore model outputted is too strongly based on to this best-fitting chemical.

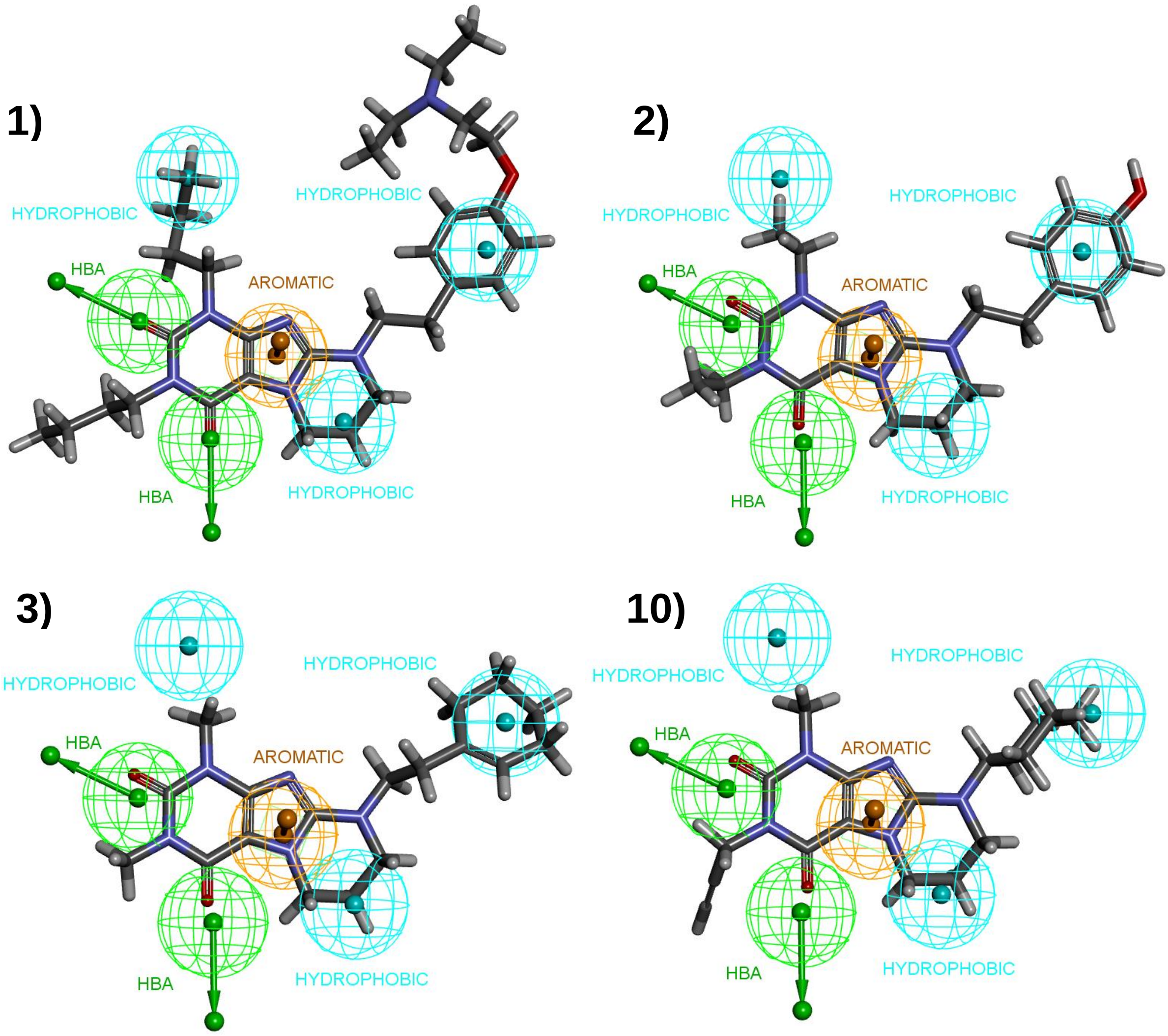


Figure 6.4: Pharmacophore model for chemicals containing Alert054 with four training set chemicals (carbon atoms in grey, nitrogen in blue, oxygen in red, and hydrogen in white). The number of the chemical indicates its rank in terms of best fit to model of the ten training chemicals – 1) is the best-fitting chemical and 10) is the worst-fitting. In the pharmacophore model, green zones represent position of hydrogen bond acceptors, orange zones represent aromatic ring, and light blue represents hydrophobic regions.

Figure 6.5 shows the fit values of chemicals to the pharmacophore model. The training active chemicals (all containing Alert054) from which the model is built fit the model well, with high fit values. A similar distribution of fit values is seen in the thirty test chemicals containing the alert, suggesting the pharmacophore model is not only applicable to the chemicals on which it was trained. However, high fit values for these chemicals are expected as most of the pharmacophore features are defined within the structural alert.

The distribution of fit values is significantly lower in the test active chemicals which do not contain Alert054. Many of these chemicals might act through a different binding mode and so would not fit well to this pharmacophore. However, some of these chemicals have higher fit values and these may act through the same binding mode as the chemicals containing Alert054. The higher fit values of such chemicals provide support to the pharmacophore model, suggesting that it is not too overfitted to only chemicals containing the structural alert.

The inactive test chemicals have a similar distribution to the active chemicals that do not contain Alert054, but with a slightly lower mean and median. The inactive chemical with the largest fit value has an experimental K_i of 13 500 nM and so, whilst inactive according to the activity cut-off of 10 000 nM used in defining the data sets, it is only just below the cut-off and shows weak activity. This result does not necessarily indicate a problem with the model but shows the difficulty of applying a binary cut-off to a quantitative measure like activity.

The high fit values for chemicals containing Alert054 and the clear distinction from the distribution of the inactive chemicals show that this pharmacophore model describes chemicals containing the structural alert well. The high fit values of some active chemicals without the alert point towards the model not being too overfitted to the structural alert.

In order to make activity predictions from the pharmacophore model, a cut-off in fit scores has been applied to the data shown in Figure 6.5. The cut-off is the lowest non-outlier fit value in the training active chemicals containing the alert. Chemicals with a fit score of 0.666 or greater are predicted to be active, chemicals with a fit score lower than 0.666 are predicted to be inactive. With this cut-off, 100% of active chemicals with the alert are correctly predicted active. 18% of active chemicals without the alert are predicted active. Most of the active chemicals without the alert are hypothesised to act through different binding modes and so are not expected to have high fit scores. 16% of inactive chemicals are incorrectly predicted to be active, which is a fairly high proportion of the inactive sample. Whilst this pharmacophore appears to be performing well, a more selective pharmacophore which incorrectly predicts a lower proportion of inactive chemicals would be obtained if the model were to include the two missing HBA features identified in the structural alert.

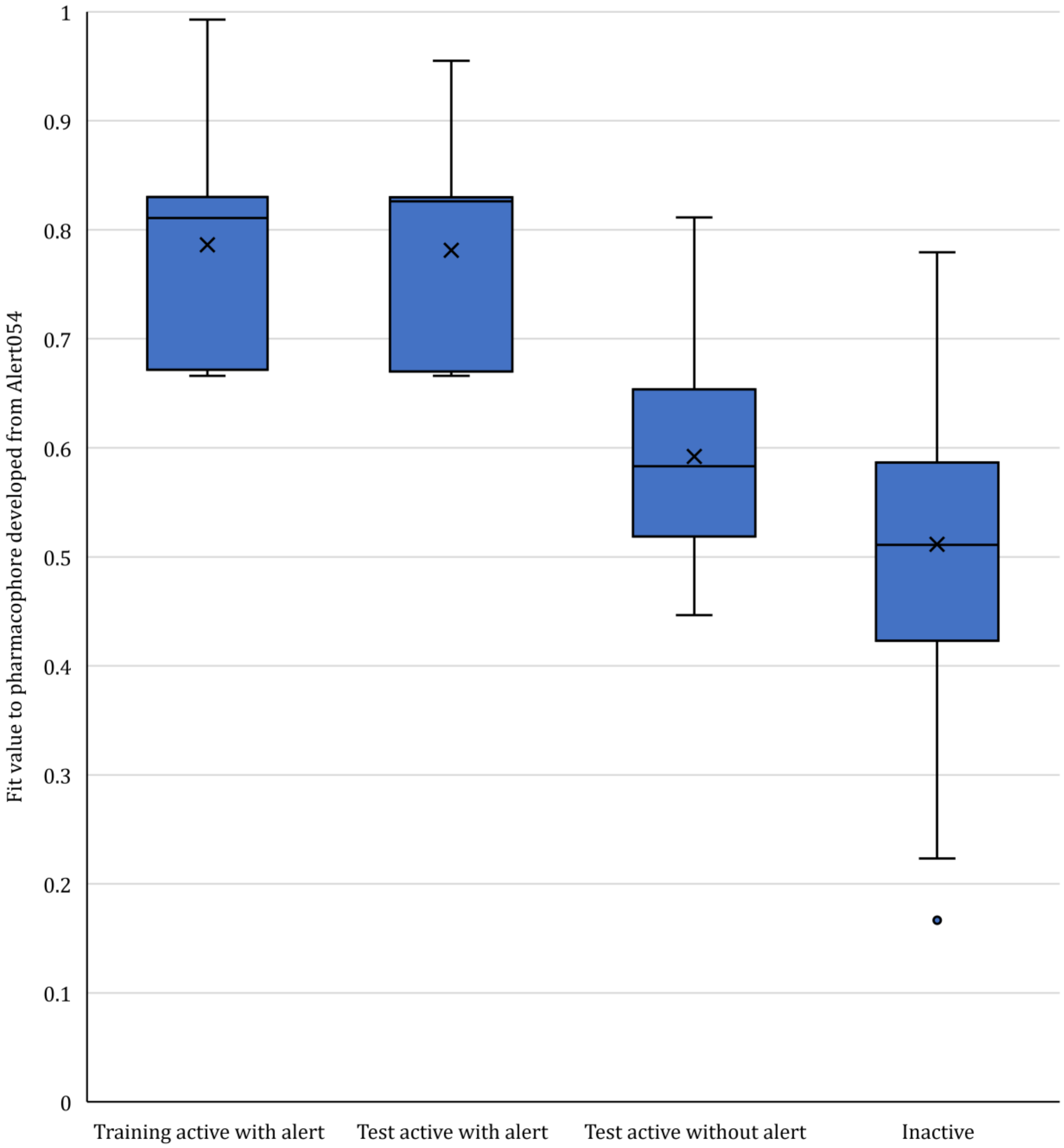


Figure 6.5: The range of "fit values" for chemicals in different groups to the pharmacophore developed from Alert054. Fit values range from 1.0 for a perfect fit of a chemical to a pharmacophore, to 0.0 for no features fit. The groups of chemicals, from left to right in the figure, are: the 10 training active chemicals from which the pharmacophore model was built, 25 test chemicals which contain Alert054, 50 test active chemicals which do not contain the alert, and 50 inactive test chemicals which contain at least one HBA, HBD, and aromatic ring. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

6.3.1.2. Alert021

Alert021 is shown in Figure 6.6. The alert has not been generalised. It defines two aromatic rings (a furan-based structure and a pyridine-based structure) but does not define where side groups branch from the structure. Hence, the chemicals containing Alert021 are less defined by the structural alert than Alert054.

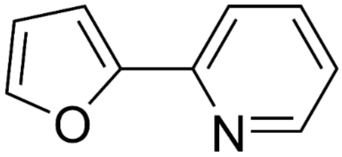


Figure 6.6: "Alert021" - a structural alert for the adenosine A2a receptor, created by the automated workflow for construction of structural alert-based models ("Risk Assessment" parameters). In the training set, this alert is contained by 22 active chemicals and no inactive chemicals.

The ten training active chemicals containing Alert021 from which the pharmacophore model was built are shown in Figure 6.7.

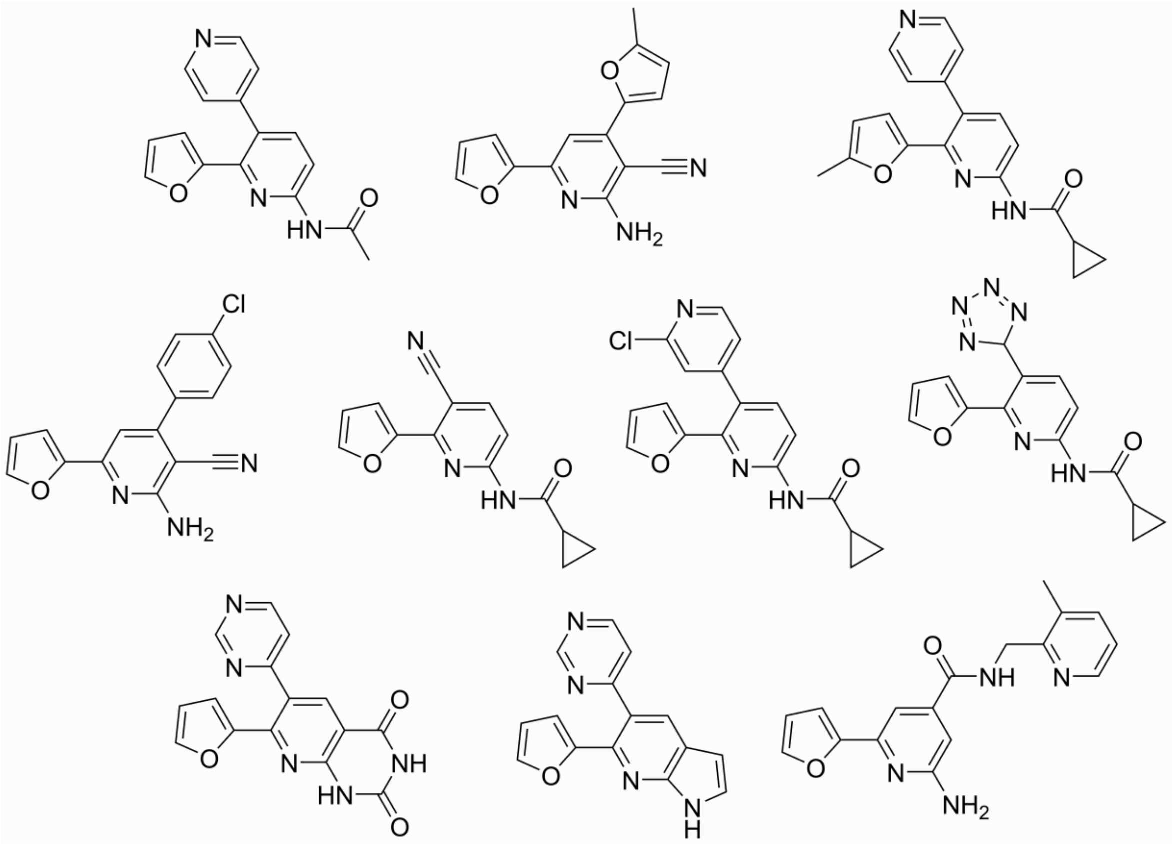


Figure 6.7: The ten training active chemicals from which the pharmacophore model was built. These chemicals are chosen from all training active chemicals containing Alert021 by RDKit's Fingerprint Diversity node in KNIME.

The pharmacophore model is shown with four chemicals in their best-fitting conformations in Figure 6.8. When chemicals are fitted to the model, nine of the ten training chemicals align in the pharmacophore in the same way – with the pyridine as central aromatic ring and furan as an aromatic ring on the left. The one molecule that does not align with the others, shown in Figure 6.8(10), has a different orientation with a different aromatic ring as the left-hand aromatic feature. From this mismatching orientation, the chemical could be rotated 180° (around a North-Western axis as drawn in the figure) to give an orientation that aligns with the other chemicals and has the same fit value to the pharmacophore model. The two orientations have the same fit values and the computer algorithm, when picking between the orientations, does not consider the orientations of the other chemicals and hence has no reason to prefer the orientation which aligns with the other chemicals.

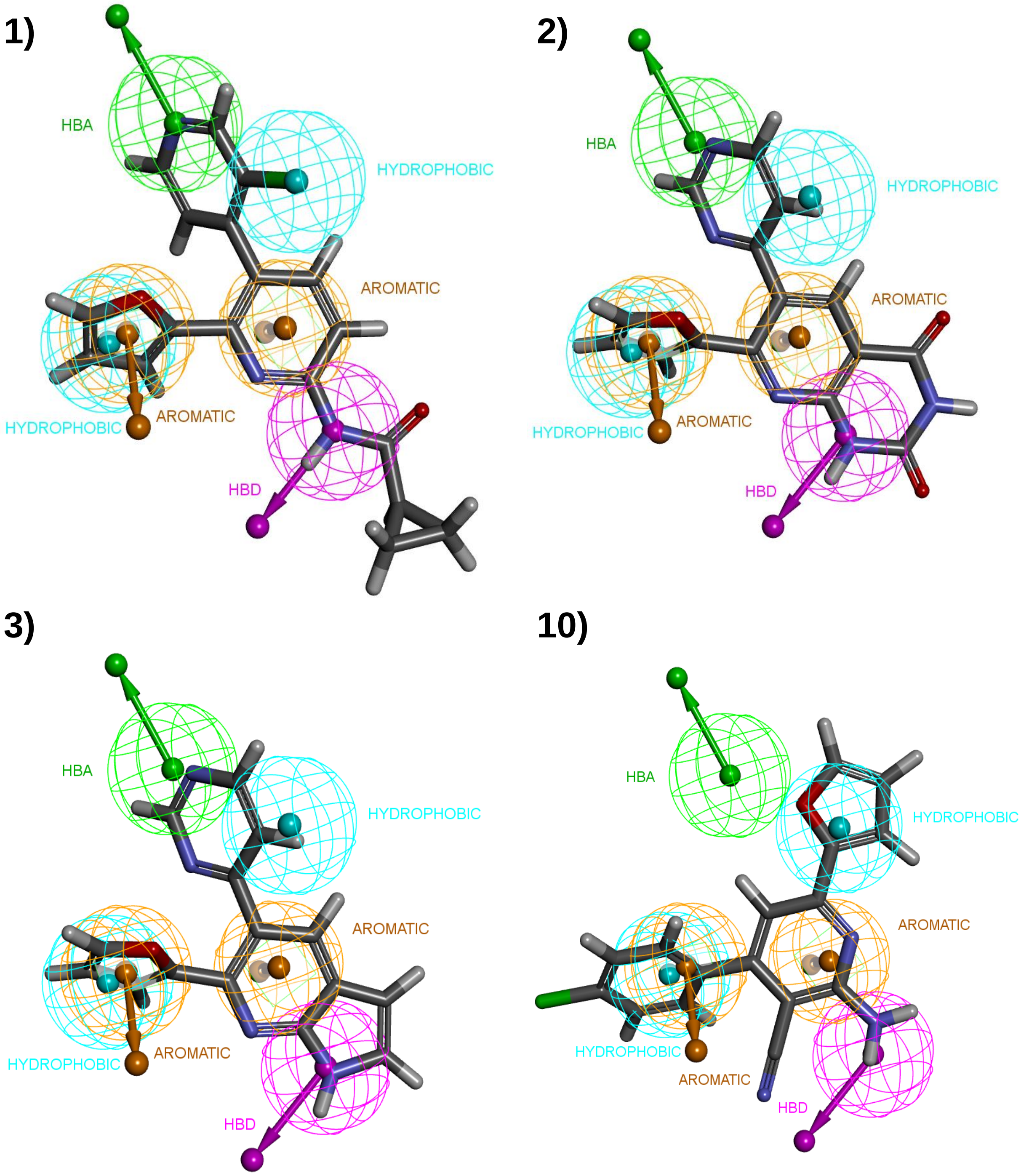


Figure 6.8: Pharmacophore model for chemicals containing Alert021 with four training set chemicals (carbon atoms in grey, nitrogen in blue, oxygen in red, hydrogen in white, and chlorine in green). The number of the chemical indicates its rank in terms of best fit to model of the ten training chemicals; 1) is the best-fitting chemical and 10) is the worst-fitting. In the pharmacophore model, green zones represent position of hydrogen bond acceptors, purple zones represent hydrogen bond donors, orange zones represent aromatic ring, and light blue represents hydrophobic regions.

Figure 6.9 shows the fit values of chemicals to the pharmacophore model. The training active chemicals (all containing Alert054) from which the model is built fit the model well, with high fit values in a narrow range. One chemical, shown in Figure 6.8(1), has a fit value of essentially 1, while all other training actives have fit values between 0.66 and 0.82, suggesting that the pharmacophore algorithm is built from one chemical and is too strongly based on that one chemical. The twenty test chemicals containing Alert021 have a narrow distribution of fit values, all falling between 0.66 and 0.82. No chemicals in the sample of inactive chemicals have fit values that lie within this range, suggesting that the pharmacophore is selective for active chemicals. Disregarding the single chemical with the near perfect fit, the pharmacophore model appears to be capable of identifying active chemicals containing Alert021 in a narrow range of fit values (0.66 to 0.82) and distinguishing them from the inactive chemicals. It is therefore a good pharmacophore model. The pharmacophore model (two aromatic rings, two hydrophobic regions, a HBA and a HBD) provides more information regarding shared features of the chemicals containing Alert021 than the alert alone (a pyridine and furan ring). This demonstrates how the pharmacophore model can be a significant improvement on the structural alerts.

In order to make activity predictions from the pharmacophore model, a cut-off in fit scores has been applied to the data shown in Figure 6.9. Chemicals with a fit score of 0.65 or greater are predicted to be active, chemicals with a fit score lower than 0.65 are predicted to be inactive. With this cut-off, 100% of active chemicals with the alert are correctly predicted active. 28% of active chemicals without the alert are predicted active and only 8% of inactive chemicals are incorrectly predicted to be active. This pharmacophore is making active predictions for all active chemicals with the alert, a large proportion of active chemicals which do not contain the alert, and only a small proportion of the inactive sample. From this, we can conclude that this pharmacophore model is performing well.

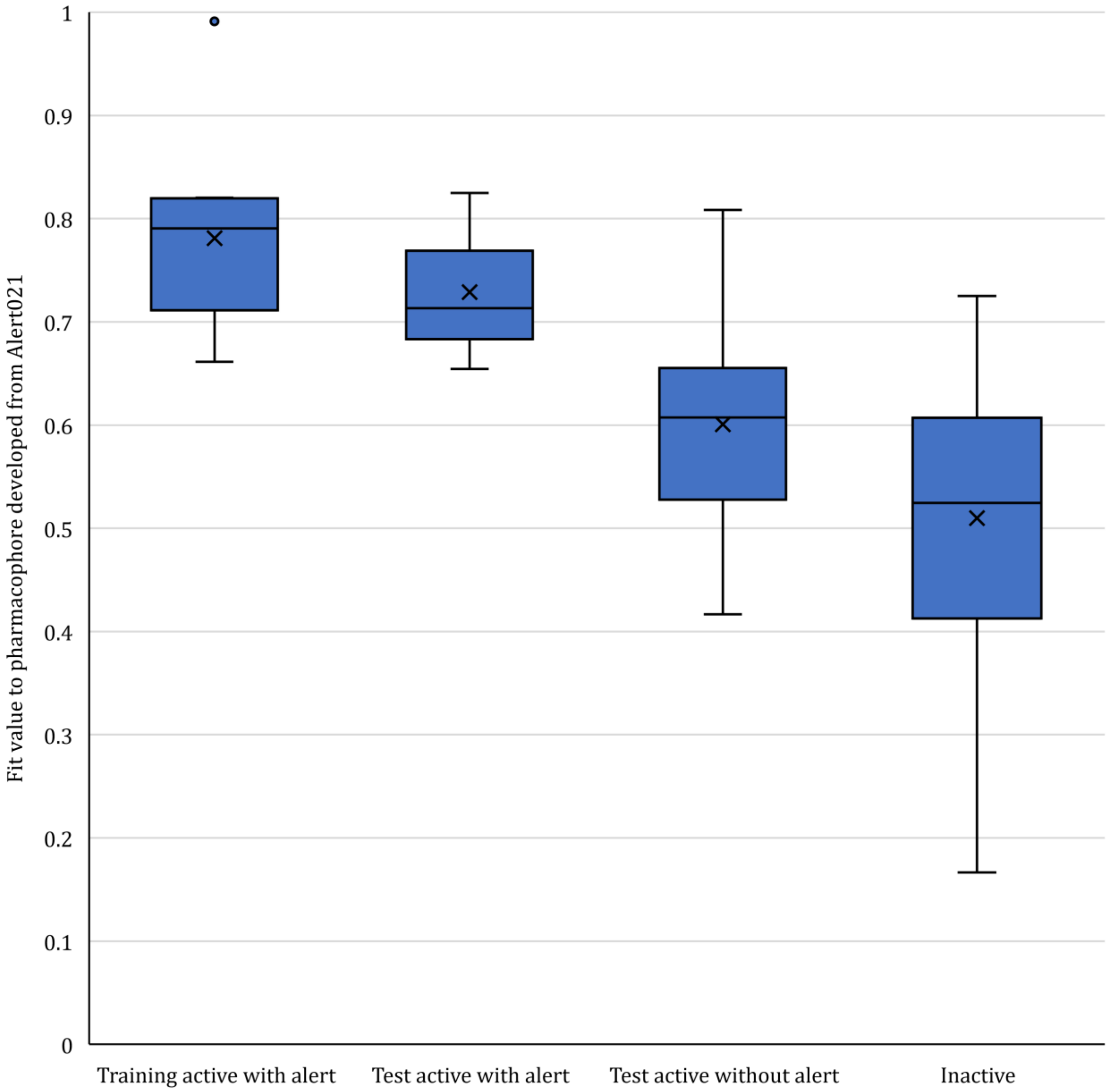


Figure 6.9: The range of "fit values" for chemicals in different groups to the pharmacophore developed from Alert021. Fit values range from 1 for a perfect fit of a chemical to a pharmacophore, to 0 for no features fit. The groups of chemicals, from left to right in the figure, are: the 10 training active chemicals from which the pharmacophore model was built, 20 test chemicals which contain Alert021, 50 test active chemicals which do not contain the alert, and 50 inactive test chemicals which contain at least one HBA, HBD, and aromatic ring. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

Applying the Alert021 pharmacophore to other structural alerts

More than a quarter of the sample of active chemicals which do not contain Alert021 have fit values greater than the cut-off suggested previously. These chemicals may cause activity through the same binding mode. If many chemicals containing the same structural alert have high fit values to the pharmacophore built from Alert021, the structural alert and Alert021 will describe chemicals acting through the same binding mode.

To investigate such chemicals further, the pharmacophore built from Alert021 has been applied to ten chemicals contained by each structural alert. Only structural alerts containing at least ten active chemicals in the training set are used here so that there is a significant sample size for the alerts used. Each set of ten chemicals is selected by the RDKit Fingerprint Diversity node from the training active chemicals containing each alert. The distribution of fit values is shown in Figure 6.10.

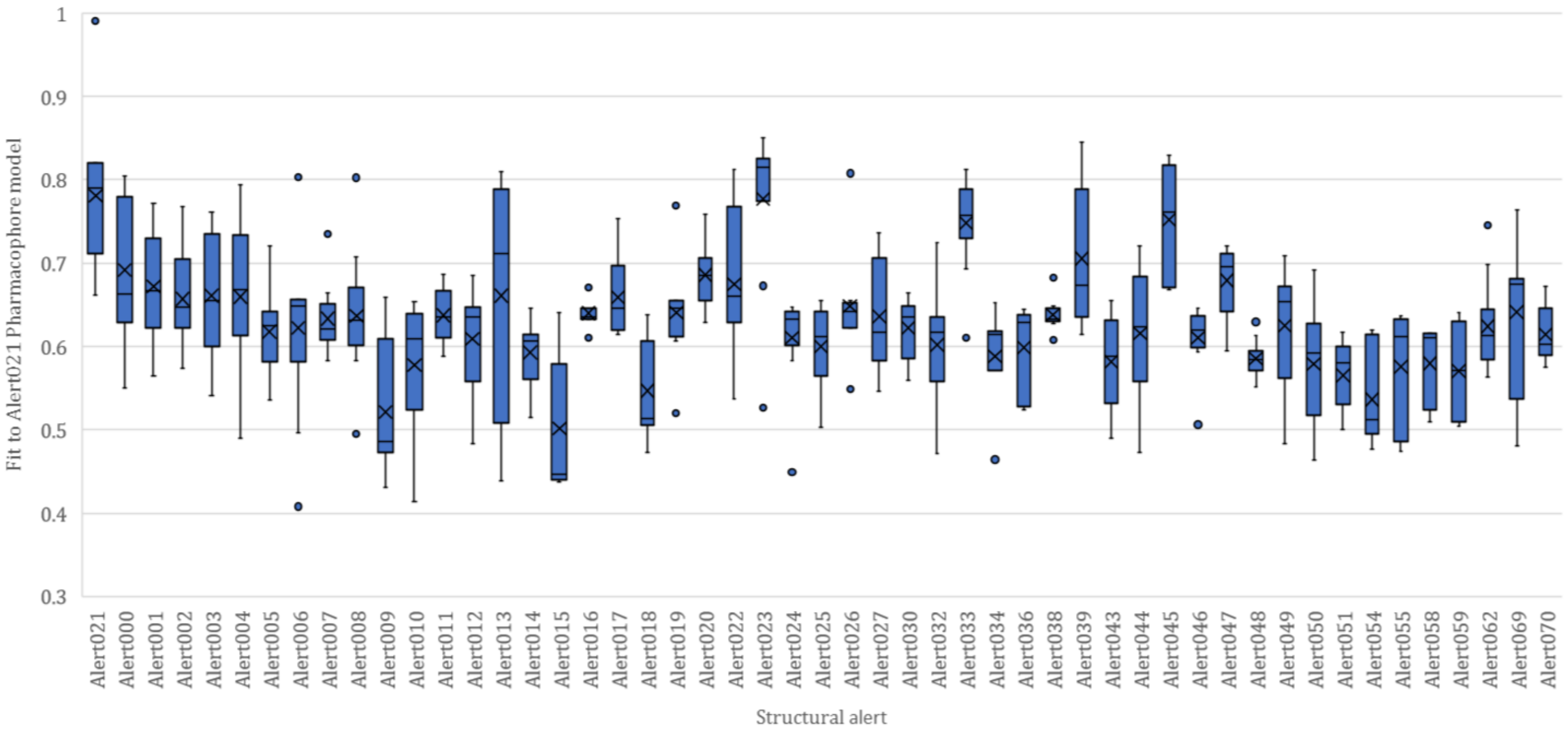


Figure 6.10: The distribution of fit values to the pharmacophore built from Alert021 for chemicals containing each structural alert for the adenosine A2a receptor. The distribution in Alert021 is shown on the far left. For each alert, ten chemicals are chosen using RDKit's Fingerprint Diversity node from the training active chemicals containing the structural alert. The structural alerts are generated using the automated workflow for construction of structural alert-based models with "Risk Assessment" parameters. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

For most structural alerts, the distribution of fit values is lower than that of Alert021. The chemicals containing these alerts likely do not act through the same binding mode.

However, many chemicals have fit values over the cut-off of 0.65 suggested previously. Numerous structural alerts have most chemicals fitting the model well, with the distributions of fit values having high median and means. Alert023, Alert033, Alert039, and Alert045 have the highest mean and median fit values after Alert021. The structures of each of these alerts and the best-fitting chemical containing each alert are shown in Figure 6.11.

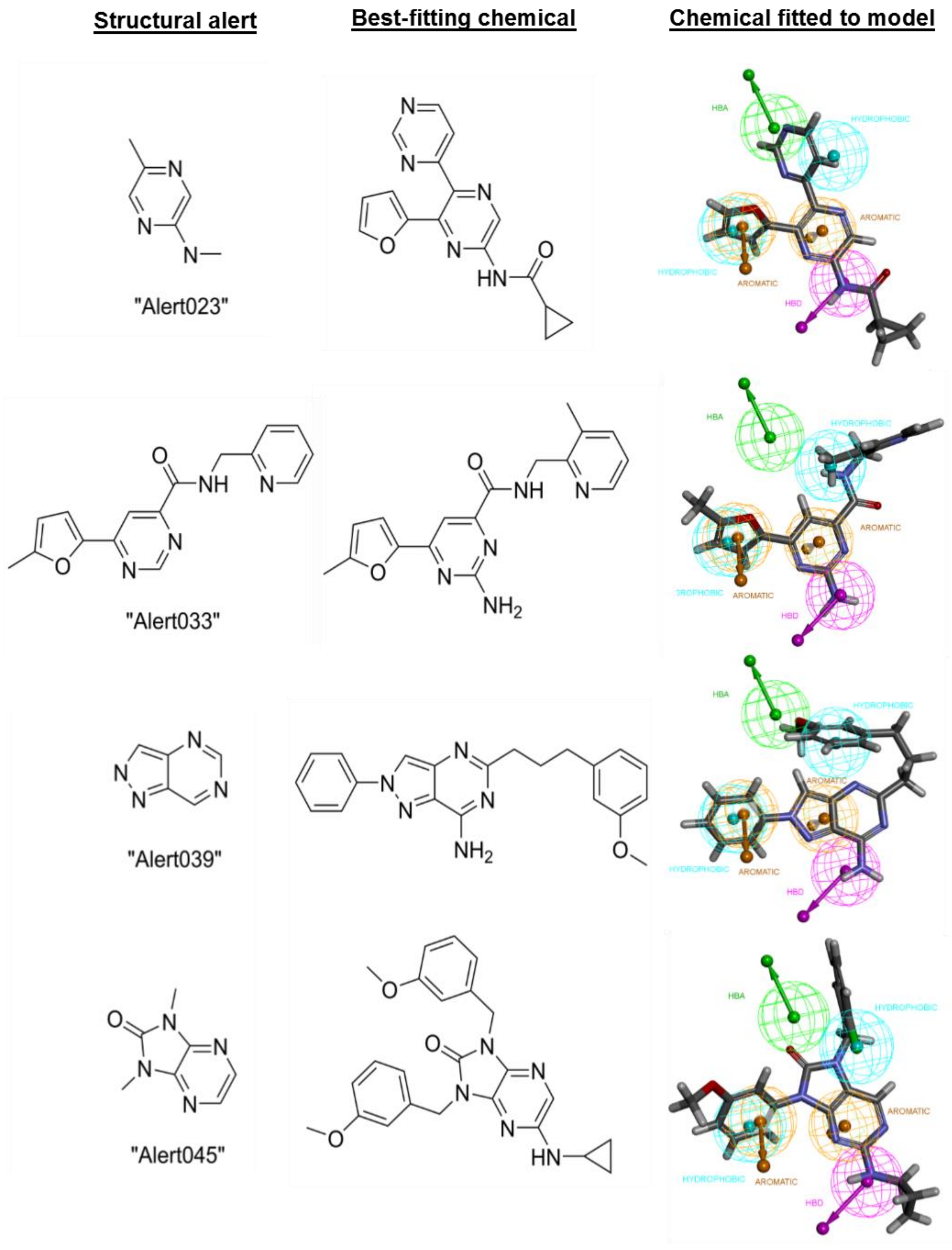


Figure 6.11: Four structural alerts for the adenosine A2a receptor, the chemical with the highest fit value to the pharmacophore model built from Alert021, and the best-fitting conformation of that chemical to the pharmacophore model.

Looking at the structural alert alone, the structural alerts do not appear to show much similarity to Alert021 or each other. However, looking at the best-fitting chemicals and comparing to the Alert021 training chemicals in Figure 6.7, some similarities can be seen. Compared to chemicals containing Alert021, the chemicals containing Alert023 and Alert033 have an additional nitrogen atom in the central aromatic ring. These chemicals would all be contained by Generalised Alert021. The similarity of the chemicals containing Alert039 and Alert045 to the Alet021 chemicals are especially clear when three dimensional conformations are generated and aligned to the pharmacophore model.

These alerts likely define chemicals which elicit activity through the same binding mode. The chemicals contained by these alerts, combined with the chemicals containing Alert021, provide a larger, more diverse pool of chemicals acting through the same binding mode. They can be used to update and refine the pharmacophore for the binding mode. For example, the best-fitting chemicals from Alert033, Alert039 or Alert045 do not hit the HBA feature in the pharmacophore model suggesting that the feature might not be necessary for receptor binding. This feature could be adjusted or removed in updated pharmacophore models.

The results for this pharmacophore model demonstrate how pharmacophore models expand upon structural alerts. The structural alerts represent two-dimensional fragments which are statistically associated with activity, whereas pharmacophores consider the whole chemical. In the example shown here, considering the whole, three-dimensional molecule has allowed similarities to be identified between chemicals that were not identified by the structural alerts alone. Consideration of the whole, three-dimensional molecule is required to understand the mechanism of the receptor binding modes. Pharmacophore models are step towards this.

6.3.2. Effect of using generalised aromatic alerts

6.3.2.1. Non-Generalised Alert016

Alert016 is shown in Figure 6.12. The structural alert has a large, flat aromatic core and specifies the position of two side groups, creating a defined, rigid scaffold for other features outside of the alert.

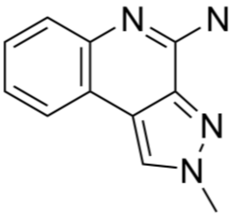


Figure 6.12: "Alert016" – A structural alert for the adenosine A2a receptor, created by the automated workflow for construction of structural alert-based models ("Risk Assessment" parameters). In the training set, this alert is contained by 18 active chemicals and two inactive chemicals.

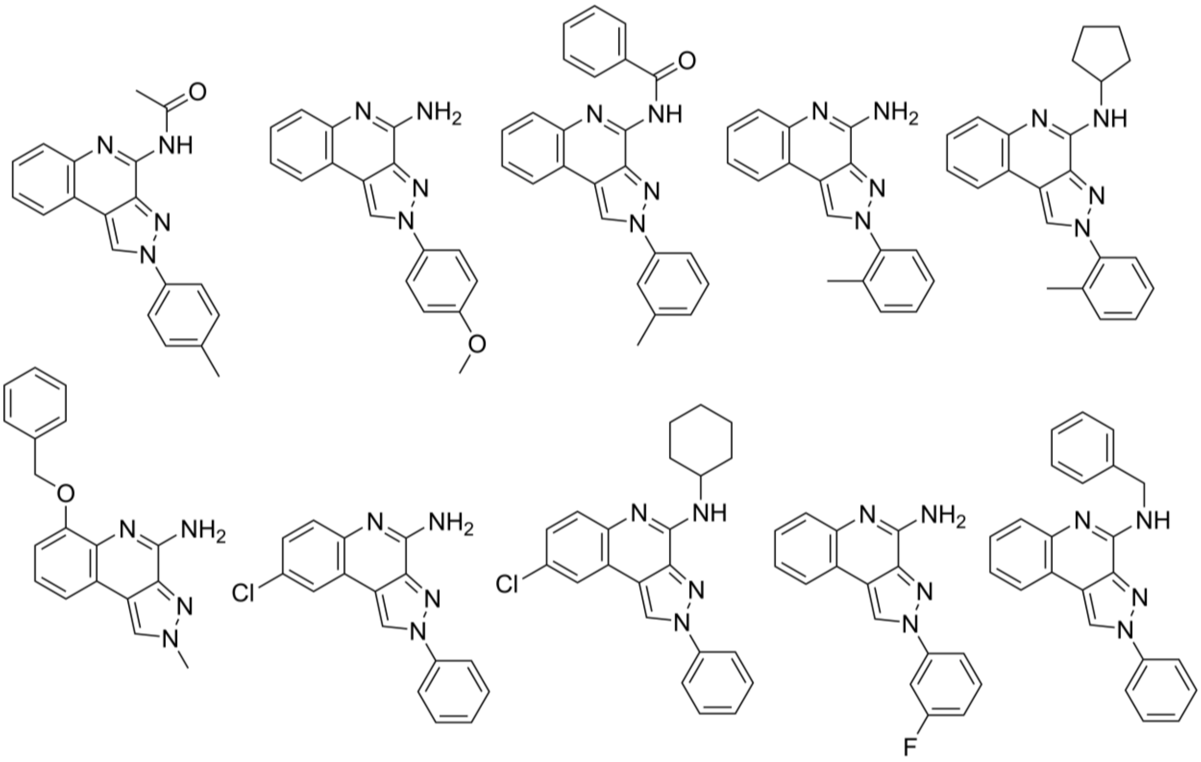


Figure 6.13: The ten active training chemicals containing Alert016 from which a pharmacophore was built. These chemicals are chosen from all training active chemicals containing Alert016 by RDKit's Fingerprint Diversity node in KNIME.

Of the active chemicals containing Alert016 in the training set, ten are chosen using RDKit's Fingerprint Diversity node and these are shown in Figure 6.13. A pharmacophore model was built from these chemicals. Figure 6.14 shows the pharmacophore model with four of the training chemicals fitted – these are the three best fitting chemicals and the single worst fitting chemical. Nine of the ten chemicals are aligned in the same orientation when fitted to pharmacophore model. The model has identified three features within the structural alert (aromatic ring, hydrophobic region, and HBA). In the nine aligning chemicals, the three features within the structural alert fit to these same features in the pharmacophore model, ensuring the chemicals take the same orientation. The one chemical that has a different orientation is shown in Figure 6.14(10) and has a lower fit value than the other training chemicals. This chemical is the only training chemical to lack an aromatic ring branching from the bottom right of Alert016 (see bottom left chemical in Figure 6.13) and takes a different orientation to fit both aromatic features in the pharmacophore model. It might be the case that this chemical has a different orientation in the receptor binding site, but it is more likely that the chemical has the same orientation as others containing Alert016 and the aromatic ring is not an important feature.

Despite occurring in all training chemicals and having a lone pair, the nitrogen within the six-membered ring of Alert016 is not identified as a HBA by the pharmacophore algorithm. This suggests that pharmacophore generation algorithm may not be picking out all important features. The pharmacophore model shown is the highest-ranking model according to the algorithm, but other models are generated and the expected HBA feature may be present in these.

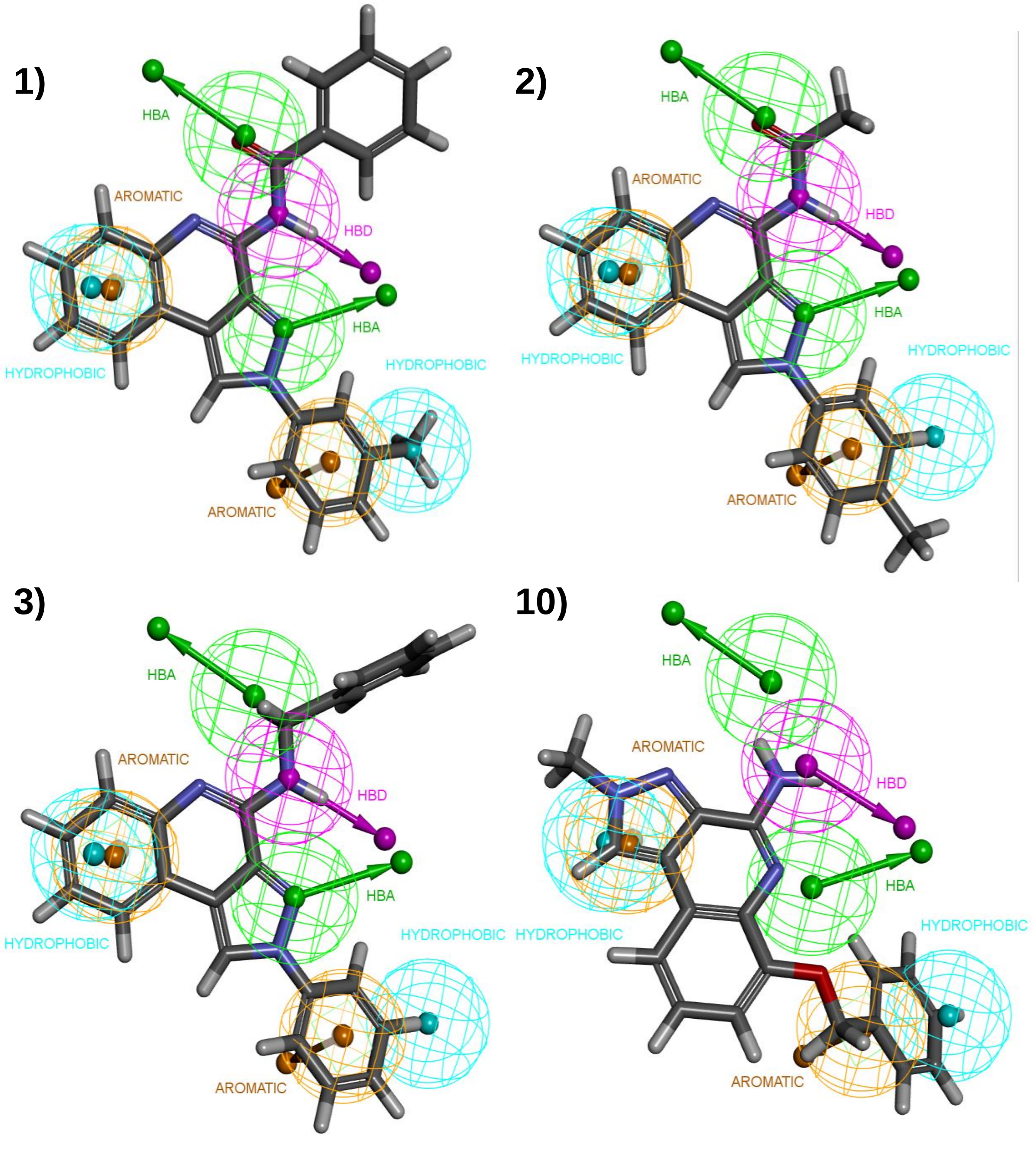


Figure 6.14: Pharmacophore model for chemicals containing Alert016 with four training set chemicals (carbon atoms in grey, nitrogen in blue, oxygen in red, and hydrogen in white). The number of the chemical indicates its rank in terms of best fit to model of the ten training chemicals – 1) is the best-fitting chemical and 10) is the worst-fitting. In the pharmacophore model, green zones represent position of hydrogen bond acceptors, purple zones represent hydrogen bond donors, orange zones represent aromatic ring, and light blue represents hydrophobic regions.

Figure 6.15 shows the fit values of chemicals to the pharmacophore model. The training active chemicals (all containing Alert016), from which the model is built, fit the model well, with high fit values in a narrow range. A similar distribution of fit values is seen in the twenty test chemicals containing the alert, suggesting the pharmacophore model is not only applicable to the chemicals on which it was trained. As with the previous models, one training active has a fit value of essentially 1.0, which is greater than all other active chemicals (training and test), suggesting the model is too strongly based on this one chemical.

The distribution of fit values is significantly lower in the test active chemicals which do not contain Alert016. Many of these chemicals likely act through a different binding mode and so would not fit well to this pharmacophore. However, some of these chemicals have higher fit values and these may act through the same binding mode as the chemicals containing Alert016. The higher fit values of such chemicals provide support to the pharmacophore model, suggesting that it is not significantly overfitted to only chemicals containing the structural alert. The inactive test chemicals have a similar distribution to the active chemicals that do not contain Alert016, but with a slightly lower mean and median.

The high fit values for chemicals containing Alert016 and the clear distinction from the distribution of the inactive chemicals show that this pharmacophore model describes chemicals containing Alert016 well. The high fit values of some active chemicals which do not contain the alert suggest that the model is not too overfitted to the structural alert, but there are few of these, so the model is likely to be fairly specific to Alert016. Overall, this appears to be a good model.

A cut-off for activity of a fit value of 0.71 was applied to the data. This results in 90% of training active chemicals with the structural alert being correctly predicted as active – the one chemical incorrectly predicted is the one that takes a different alignment to the other nine chemicals when fitted to the model. This could either be an outlier, or indicative of the model identifying some incorrect features. All test active chemicals with the alert are correctly predicted as actives. 12% of test active chemicals without the alert are predicted to be active. Only 2% of inactive chemicals are predicted to be active, indicating that the model is very selective. With such a selective model, we can be confident that the 12% of test active chemicals with the alert predicted to be active are likely to be acting through a similar binding mode to chemicals containing Alert021. Whilst the one outlying training active chemical with the alert could indicate flaws, the model is still selective for active chemicals and performing well.

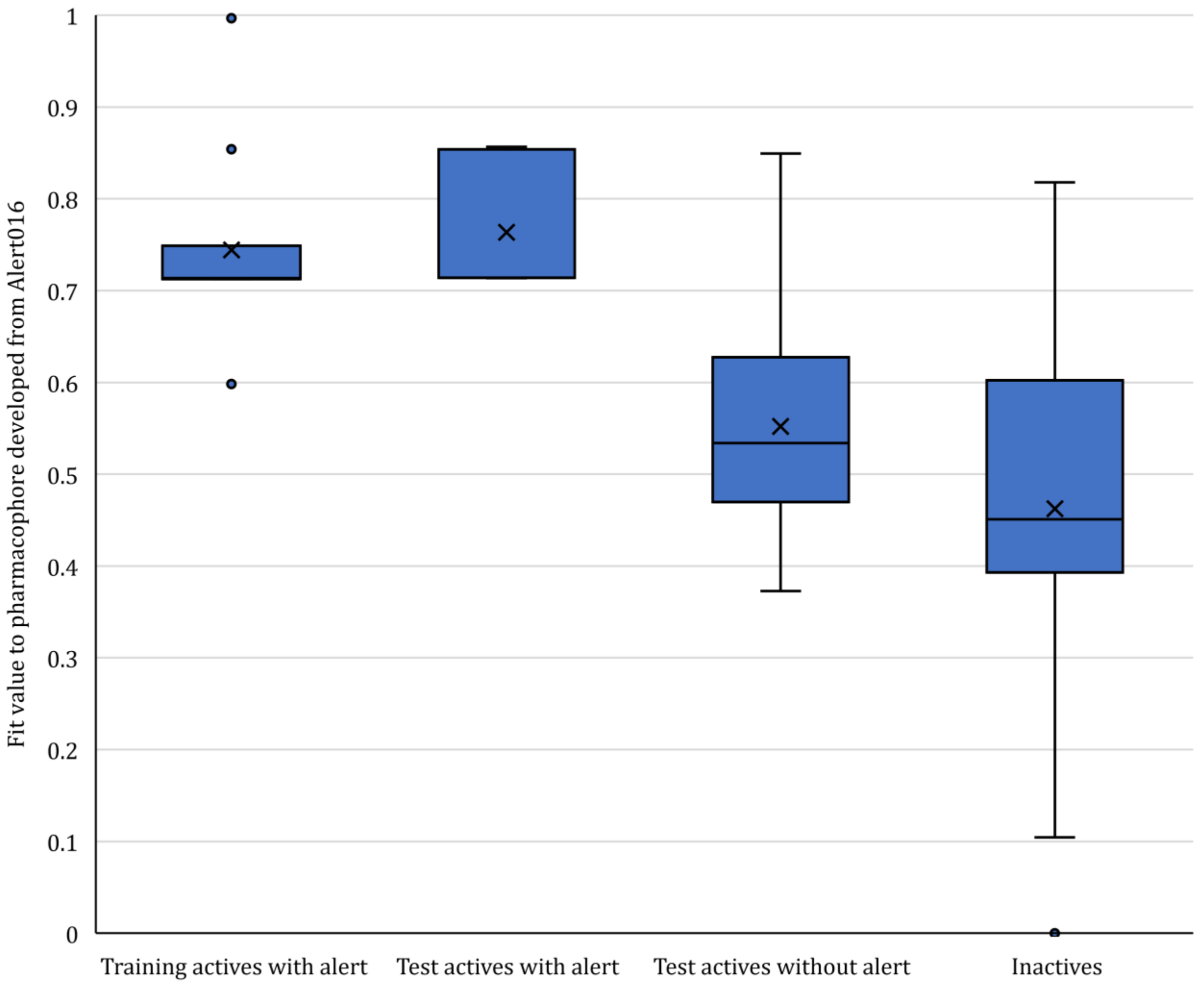


Figure 6.15: The range of "fit values" for chemicals in different groups to the pharmacophore developed from Alert016. Fit values range from 1.0 for a perfect fit of a chemical to a pharmacophore, to 0.0 for no features fit. The groups of chemicals, from left to right in the figure, are: the 10 training active chemicals from which the pharmacophore model was built, 20 test chemicals which contain Alert016, 50 test active chemicals which do not contain the alert, and 50 inactive test chemicals which contain at least one HBA, HBD, and aromatic ring. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

6.3.2.2. Generalised Alert016

The generalisation of aromatic substructures process has been applied to Alert016, giving the generalised alert shown in Figure 6.16. Generalisation of the alert significantly increases the number of training set chemicals containing the alert, up to 86 actives and three inactives from 18 actives and two inactives. The two aromatic nitrogen atoms with available lone pairs in Alert016 are present in the generalised alert, suggesting they may be involved as HBAs in important interactions in the binding mode. Five positions, formerly carbon atoms, have been replaced with general aromatic atoms. Active training chemicals are observed where these positions are nitrogen atoms. These nitrogen atoms, although some are able to act as HBAs, are not observed in all chemicals containing the structure and are therefore unlikely to be involved in important interactions. These are additional features for the pharmacophore generation algorithm to sort through, but they should ultimately be ignored in the best pharmacophore model.

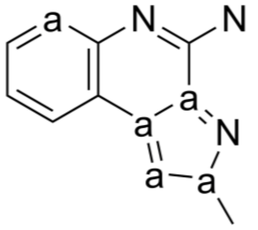


Figure 6.16: "Generalised Alert016" – A structural alert for the adenosine A2a receptor, created by the automated workflow for construction of structural alert-based models ("Risk Assessment" parameters) and then generalised with the "generalisation of aromatic substructures" method (theta 0.95). In the substructure, "a" represents any aromatic atom. In the training set, this alert is contained by 86 active chemicals and three inactive chemicals.

As previously, ten training active chemicals containing the structural alert are chosen with RDKit's Fingerprint Diversity node. These are shown in Figure 6.17. The generalised alert is contained by more active training chemicals than the non-generalised alert, giving a larger pool of chemicals from which to pick a more diverse subset. The generalisation of the structural alert also allows for more diverse substructures within the alert itself. In the ten chemicals, nitrogen atoms are observed in each of the generalised atom positions at least once.

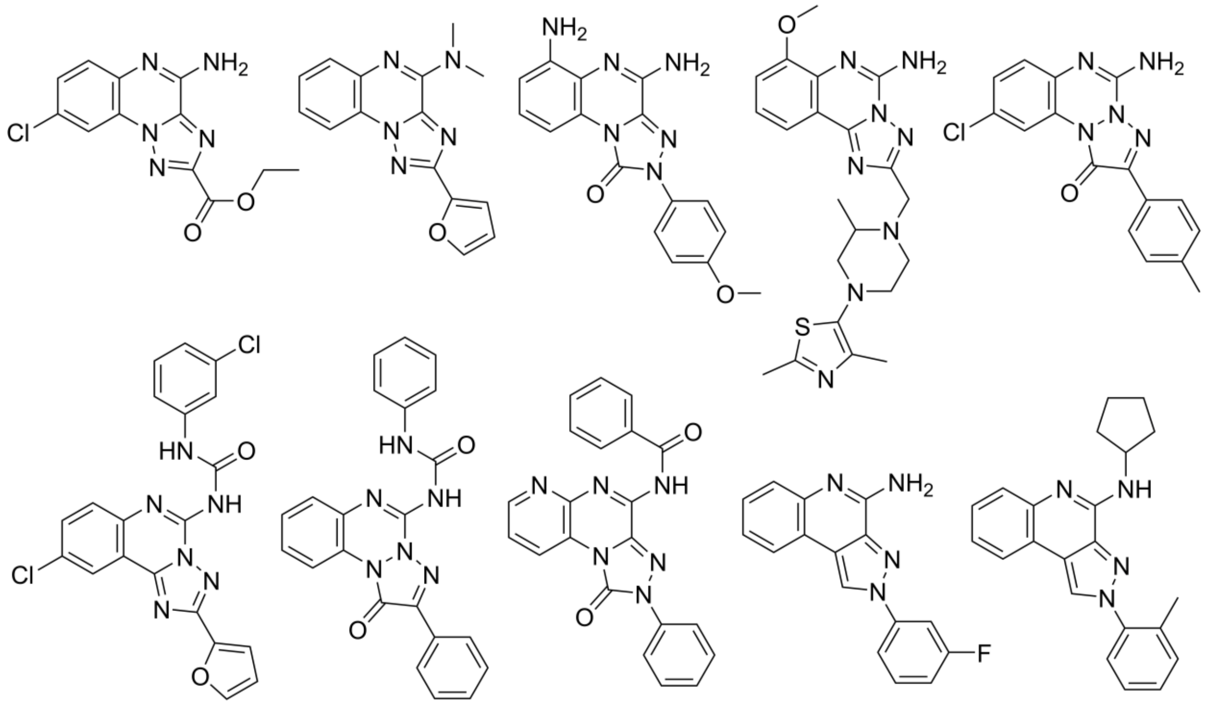


Figure 6.17: The ten active training chemicals containing Generalised Alert016 from which a pharmacophore is built. These chemicals are chosen from all training active chemicals containing Alert016 by RDKit's Fingerprint Diversity node in KNIME.

The pharmacophore model built from the subset of diverse chemicals containing Generalised Alert016 is shown in Figure 6.18 with the six best-fitting chemicals from that subset. The chemicals are shown in the conformation output by the pharmacophore generation algorithm, which is the conformation that best fits the model features.

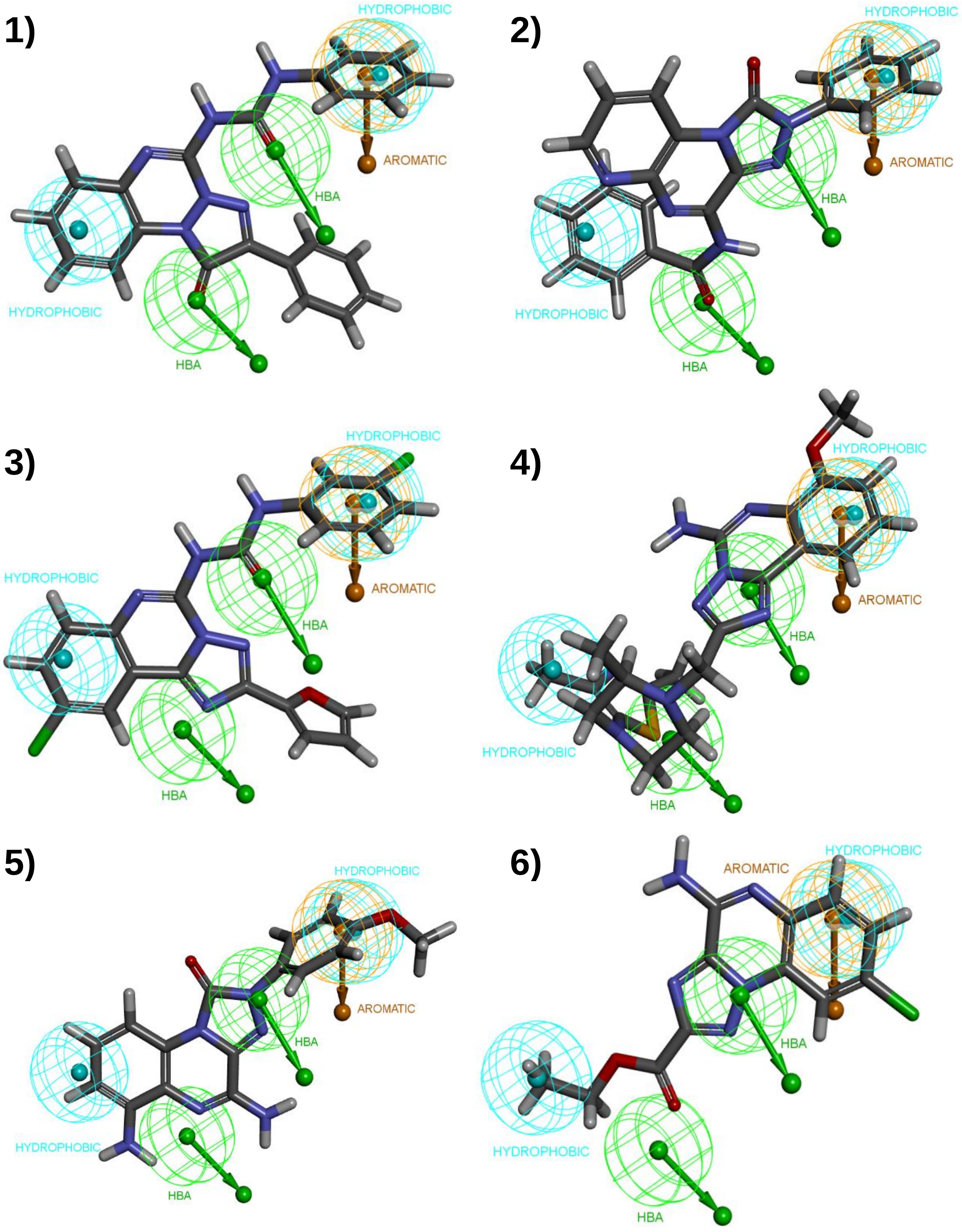


Figure 6.18: Pharmacophore model for chemicals containing generalised Alert016 with six training set chemicals (carbon atoms in grey, nitrogen in blue, oxygen in red, hydrogen in white, and chlorine in green). The number of the chemical indicates its rank in terms of best fit to model of the ten training chemicals – 1) is the best-fitting chemical, 2) is the second-best fitting chemical, etc. In the pharmacophore model, green zones represent position of hydrogen bond acceptors, orange zones represent aromatic ring, and light blue represents hydrophobic regions.

In the best-fitting chemical, the pharmacophore model identifies only one feature within the structural alert: a hydrophobic region. This is not a feature that is selective to the structural alert; many other groups can act as a hydrophobic region. There is no combination of features in the new pharmacophore model that are filled by the structural alert substructure in all chemicals, and hence there is no reason for the chemicals to adopt aligning orientations when fitted to the model. Only one training chemical aligns with the best fitting chemical, with remaining chemicals all adopting different orientations, some of which are shown in Figure 6.18. With so many different orientations, the pharmacophore model is not picking up the important features which are shared across the training chemicals. The generalised structural alert is a central scaffold shared across all chemicals, but it is not identified by the pharmacophore model, which instead fits the chemicals to an outer hydrophobic region on one side, and a hydrophobic and aromatic feature on the other outer side. These outer, non-specific features are contained by all training chemicals, giving them all high fit values to the model even if they do not contain the two HBAs which make up the rest of the pharmacophore model. As a result, the pharmacophore generation algorithm considers this model to be a good representation of the training chemicals even though it has not truly described the features in the central structural alert substructure. This pharmacophore model, whilst fitting the training molecules (when in differing orientations) is not defining the specific common binding mode.

As with previous pharmacophore models, the model based on Generalised Alert016 has been tested on the diverse subset of training active chemicals with the alert, test active chemicals with the alert, test actives without the alert, and test inactives. The distribution of fit values is shown in Figure 6.19.

Unlike the previous pharmacophore models, there is no clear distinction between the distribution of fit values in chemicals containing the structural alert and the active chemicals without the alert or the inactive chemicals. The active chemicals containing Generalised Alert016 have high fit values to the model, but so do most of the active chemicals which do not contain the alert, and a large proportion of the inactive chemicals. The pharmacophore model is not specific to the binding mode of the chemicals containing the structural alert and does not distinguish active molecules from inactive.

The minimum fit value of the training active chemicals with the alert is 0.578. Using this as an activity cut-off, active predictions are made for: 100% of training active chemicals with the alert, 94% of test active chemicals with the alert, 80% of active chemicals without the alert, and 64% of inactive chemicals. This is a poor pharmacophore model which is not selective for active chemicals.

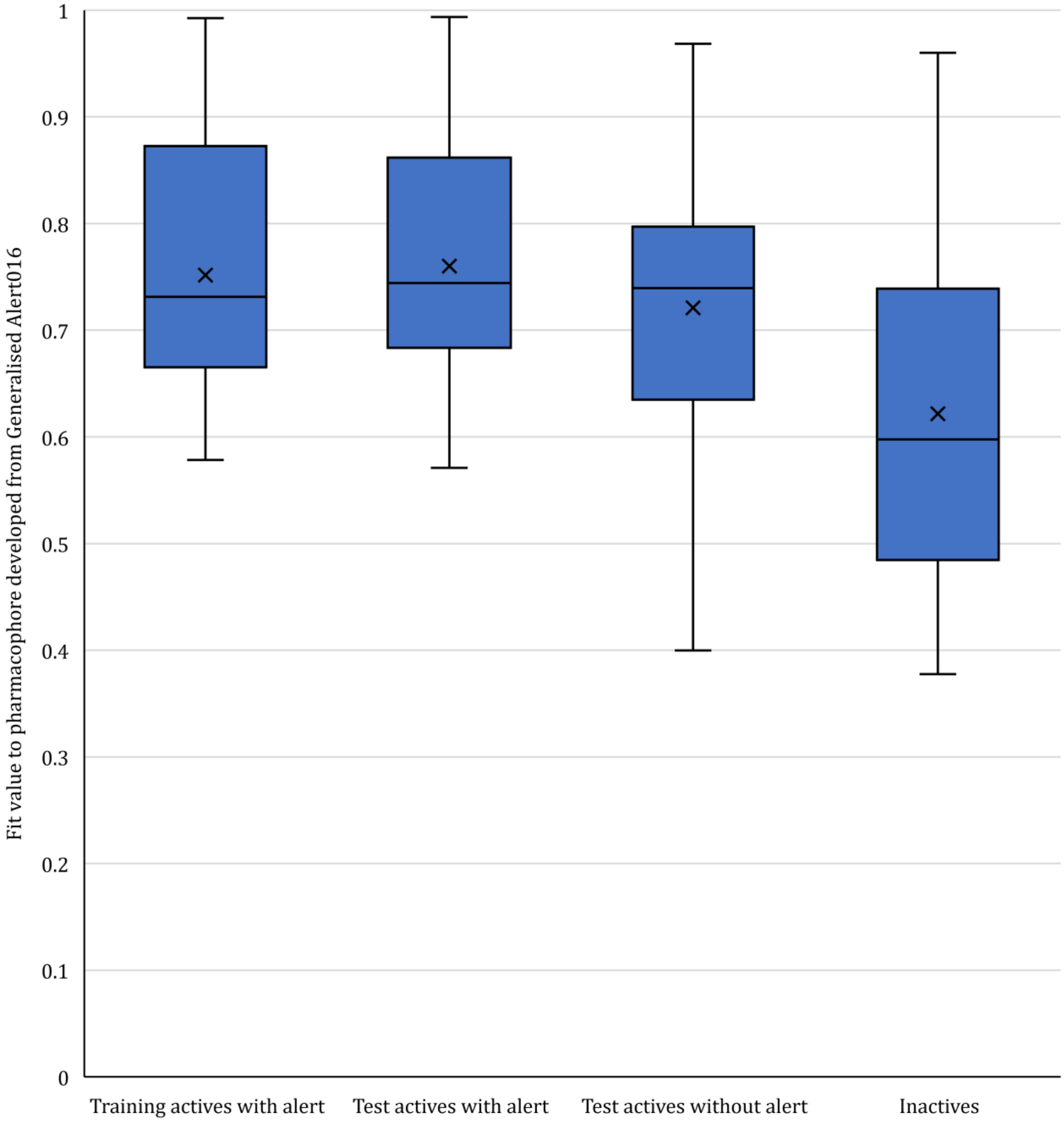


Figure 6.19: The range of "fit values" for chemicals in different groups to the pharmacophore developed from Generalised Alert016. Fit values range from 1.0 for a perfect fit of a chemical to a pharmacophore, to 0.0 for no features fit. The groups of chemicals, from left to right in the figure, are: the 10 training active chemicals from which the pharmacophore model was built, 50 test chemicals which contain Generalised Alert016, 50 test active chemicals which do not contain the alert, and 50 inactive test chemicals which contain at least one HBA, HBD, and aromatic ring. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

It was hoped that using generalised alerts would provide additional information to help the pharmacophore generation algorithm identify which of the features within the structural alert are required in the model, but instead it has resulted in construction of poor pharmacophore models. Instead of aiding model construction, the additional information provided by generalised alerts may have been problematic for the pharmacophore generation algorithm.

In the case of Alert016, additional groups capable of acting as HBAs are introduced in different positions within the structural alert which should have been ignored in the final model, giving a model similar to the one developed from non-generalised Alert016. The pharmacophore for non-generalised Alert016 has been applied to the same groups of chemicals used in testing Generalised Alert016, and the distribution of fit values is shown in Figure 6.20. The distribution of fit values in the active chemicals containing Generalised Alert016 to the pharmacophore model is clearly distinct from the distribution of fit values in the inactive chemicals and in the active chemicals without the generalised alert.

There is one clear outlier in the training active chemicals containing Generalised Alert016, with a low fit value of 0.42. This is the one chemical with no hydrogen on the amine outside of the aromatic system. In the pharmacophore model, this amine has been identified as a HBD but it could be identified as a HBA to give a model which gives a better fit of this outlying chemical without affecting the fit values of the other chemicals (where the nitrogen is in an amide, the carbonyl could act as a HBA). This example demonstrates how additional information gained from the larger pool of chemicals containing a generalised structural alert could have been used to develop a better pharmacophore model.

Applying the cut-off of 0.71 previously derived from non-generalised Alert016 to this data results in active predictions for: 60% of training active chemicals containing generalised Alert016, 65% of test active chemicals containing generalised Alert016, 2% of test active which do not contain generalised Alert016, and 6% of inactive chemicals.

However, using some pragmatic flexibility with the cut-off gives improved results. A slightly lower of cut-off of 0.70 results in active predictions for 90% of training active chemicals containing generalised Alert016, active predictions for 90% of test active chemicals containing generalised Alert016, active predictions for 4% of test active which do not contain generalised Alert016, and active predictions for 8% of inactive chemicals.

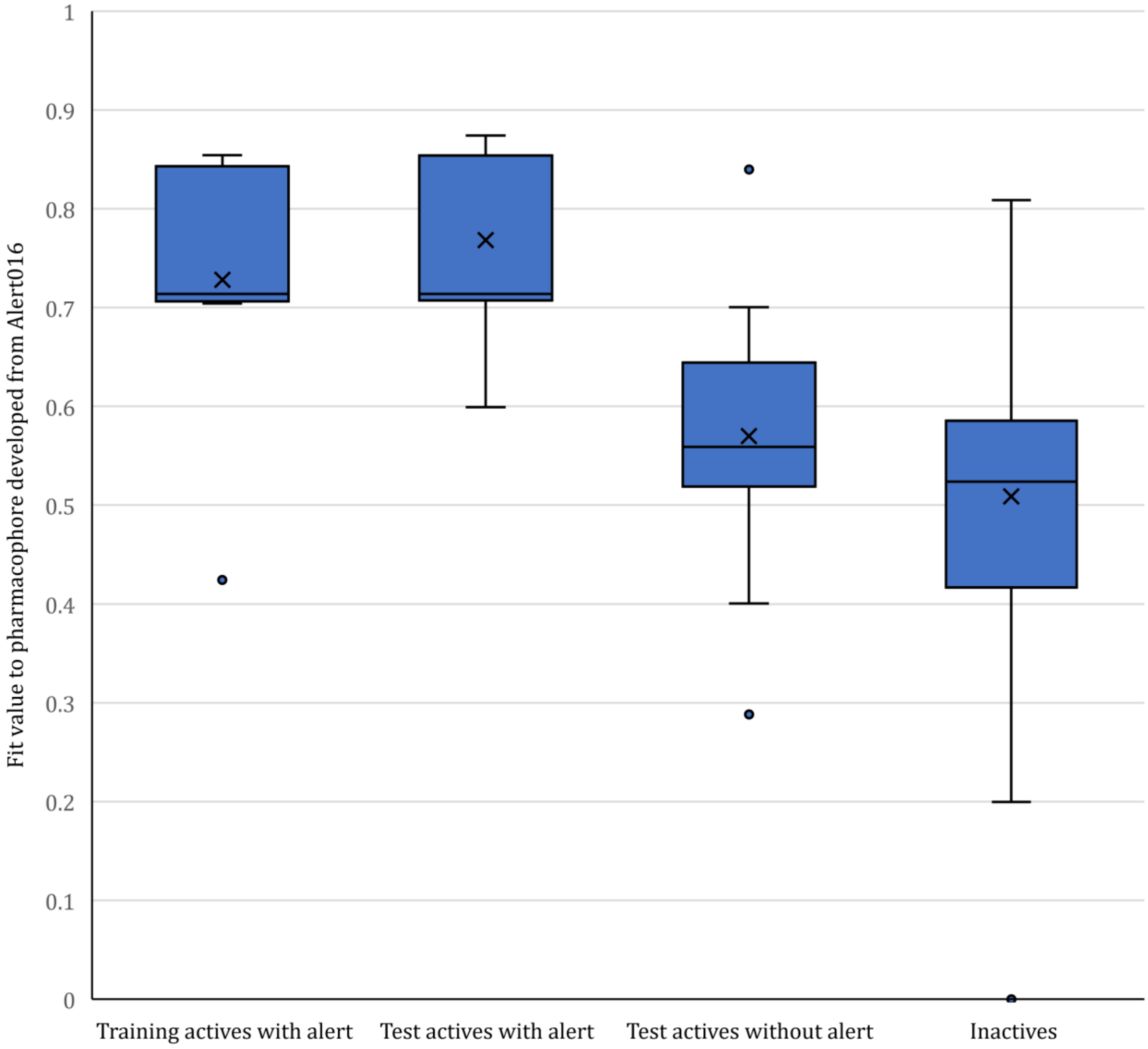


Figure 6.20: The range of "fit values" for chemicals to the pharmacophore developed from Alert016 before aromatic generalisation. The groups of chemicals are the same used in testing the pharmacophore constructed from Generalised Alert016. Fit values range from 1.0 for a perfect fit of a chemical to a pharmacophore, to 0.0 for no features fit. The groups of chemicals, from left to right in the figure, are: the ten training active chemicals from which the pharmacophore model for Generalised Alert016 was built, 50 test chemicals which contain Generalised Alert016, 50 test active chemicals which do not contain the alert, and 50 inactive test chemicals which contain at least one HBA, HBD, and aromatic ring. In the plot, boxes represent chemicals within the lower and upper quartiles, and whiskers represent the lowest and highest values that are not outliers. Outliers, shown as dots, are defined as any chemicals with a fit value 1.5 times the interquartile range lower than the lower quartile or greater than the upper quartile. The line within the box is the median value and the cross is the mean value.

These results show a much better performance of the pharmacophore derived from non-generalised Alert016 than the pharmacophore derived from the generalised alert. Compared to the results obtained for the data sets used in testing non-generalised Alert016, a much lower proportion of test actives not containing the alert are seen here – 4% compared to 14%. This indicates that a large proportion of the chemicals classed as "test actives not containing Alert016" contained generalised Alert016, supporting the hypothesis that non-alert-containing actives with high fit values likely act through the same binding mode as the alert-containing actives from which the pharmacophores were built.

The high fit values of chemicals containing Generalised Alert016 to the pharmacophore model constructed from non-generalised Alert016, and the clear distinction between fit values of the chemicals with the generalised alert and the inactive chemicals, suggest that this pharmacophore model is a good model for all the chemicals containing the generalised alert. This is a stark contrast to the poor results of the pharmacophore model developed from chemicals containing Generalised Alert016.

Chemicals containing the non-generalised alerts have the exact same distribution of features within the substructure defined by the structural alert, and there is a high probability that at least some of these features are picked out by the pharmacophore generation algorithm. This would result in the substructure aligning to these features in the models in the same way across all chemicals, leading to the chemicals adopting a similar orientation to the model. When generalised alerts are used, some features are defined within the alert and will be consistent across all chemicals, but where generalised aromatic atoms are present, some chemicals will have additional features which should not be included in the final pharmacophore model. However, the pharmacophore generation algorithm is not able to work this out. Instead, it creates a model where general, outer features are defined but specific, central features are not. Hence, the substructure common to all chemicals not being aligned across chemicals fit to the model. These models are too general and do not distinguish chemicals containing the generalised alert from other chemicals. The models do not define the specific combination and conformation of features that is required for eliciting activity through the specific binding mode. Therefore, using generalised structural alerts to identify chemicals to be used in the pharmacophore generation algorithm results in poorer models than using non-generalised structural alerts.

6.3.3. Improving pharmacophore generation

It appears that the best pharmacophore models are created when the structural alignment of the pharmacophore generation includes or focuses on the atoms covered by the structural alert. As a result, all chemicals align to the pharmacophore model in the same way. Poor pharmacophore models have been generated from structural alerts when the structural alert substructure is not aligned across chemicals fitted to the model. Whilst the HipHop algorithm cannot be easily changed, other changes in the pharmacophore generation process can be made to ensure the structural alert is aligned across all chemicals. Some changes are suggested here.

- The HipHop algorithm creates many different pharmacophore models from different combinations and conformations of features, ranking each internally. With default settings in Discovery Studio, only the top ten ranking models are output, and these are often the same combination of features in slightly different conformations. With different settings, more models could be output. The user could then search through these different models to find pharmacophores where the features within the structural alert are defined, resulting in the substructure being aligned in the training chemicals. These pharmacophores may not be the best-fitting according to the internal ranking of HipHop within the training chemicals, but the model will be better at differentiating chemicals with the structural alert from inactive chemicals or active chemicals which act through different binding modes.
- The user can define a minimum number of each type of feature required to be present in outputted pharmacophore models, although no minimum was required in this work. The number of each type of feature present within the structural alert can be counted and set as the minimum required in the pharmacophore model. This helps to filter the pharmacophore models constructed by the HipHop algorithm to find ones where the features within the alert are defined.
- The structural alert can be added as a “custom feature” in Discovery Studio and then set as a required feature in the HipHop algorithm. In theory, all chemicals would be required to have the structural alert in the same orientation and the HipHop algorithm would identify features outside of the alert and their positions relative to the alert. The custom feature could then be removed and features that are present within the structural alert could be manually added to the pharmacophore model. This has been attempted for a number of structural alerts but in each case no pharmacophore models were generated by the HipHop algorithm.
- In creating pharmacophore models here, only active chemicals have been input into the HipHop algorithm. Inactive chemicals could also be included, and the algorithm will try to

avoid models to which the inactive chemicals have high fit values. This should help the algorithm identify the features that are unique to the active chemicals compared to the inactive chemicals, reducing the likelihood of the non-specific pharmacophore models being highly ranked within the HipHop algorithm.

Even in cases where most training chemicals are aligned by non-generalised structural alerts in the pharmacophore model, the pharmacophores were not perfect. The three models based on non-generalised alerts presented here were strongly influenced by one training chemical in each case, and some features present and overlaid in all training chemicals were not identified as features within the model. These problems might be solved by using some of the above suggested changes. However, these problems, combined with the problems with generalised structural alerts, might be indicative of short comings with the HipHop algorithm. In future work, other available automatic pharmacophore algorithms could be used, or a new process designed which involves overlaying or tethering common substructures between chemicals.

In this work, binary activity data has been used, so all chemicals are viewed as equally active and are equally weighted in the pharmacophore generation algorithm. In the work presented here, each model was strongly based on a single chemical, but in each case, that chemical is not the most potent. This information was not used in pharmacophore generation and the inclusion of this data might result in improved models. In future work, the experimental potency values could be used instead of binary activity, providing information as to which chemicals and features should be weighted highest.

6.3.4. Use of Pharmacophores in Risk Assessment

In this chapter, examples have been given to show how pharmacophore models can be constructed from structural alerts. These show the potential for pharmacophores to expand upon structural alerts: structural alerts can identify similarities in fragments of chemicals, but these fragments describe only specific combinations of atoms and bonds, whilst pharmacophores are less specific and can identify similarities in features across an entire chemical structure.

Despite issues with the automated algorithm for generating the pharmacophores, each model showed good selectivity, almost always making active predictions for all active chemicals containing the alert the model was built from, a small proportion of active chemicals which did not contain that alert, and few inactive chemicals.

In the future, a combination of many pharmacophore models being used to make activity predictions for a particular biological target is envisioned, much like a combination of structural alerts has been successful applied in this work previously. Cut-offs in fit values would be defined for each pharmacophore to predict active chemicals.

As a starting point, a pharmacophore model could be constructed from each structural alert. As seen with Alert021, different structural alerts may contain chemicals which can be modelled by the same pharmacophore. Identifying these cases and combining these chemicals will greatly reduce the number of pharmacophores needed to model the entirety of the data.

Being able to construct pharmacophores from generalised structural alerts is key to this vision. Generalised alerts contain more diverse structures which act through the same binding mode. The independent, generalised structural alert models constructed in section 5.4 would provide a good basis for pharmacophore construction. However, it was found to not be possible to construct good pharmacophore models from generalised alerts with the HipHop algorithm. Before further progress can be made in building a combination of pharmacophores for use in risk assessment, this problem must be overcome.

6.3.5. Conclusions

For a chemical to be active at a receptor binding MIE, it needs to have a specific combination of features in a specific conformation in three-dimensional space. The structural alerts for receptor binding MIEs often describe a specific scaffold holding features within and outside of the alert in a specific conformation. These structural alerts have a synergy with pharmacophore models, which attempt to predict activity by modelling the required specific combination and conformation of features.

Pharmacophore models have been constructed for the adenosine A2a receptor from chemicals containing the same structural alert. Three good pharmacophore models have been constructed from three different structural alerts which have not been aromatically generalised. For each of these models, active chemicals containing the relevant alert had a distribution of high fit values which was clearly distinct from the distribution of lower fit values in the inactive chemicals. The distribution of fit values in the active chemicals that did not contain the relevant alert was also generally low, although a small proportion had fit values similar to the active chemicals containing the alert. Such chemicals may act through the same binding mode, suggesting that the pharmacophore models are not too overfitted to the structural alerts from which they are based. In particular, the pharmacophore for "Alert021" shows how pharmacophores can expand upon the knowledge derived by structural alerts. Structural alerts define a common substructure within a larger chemical, but pharmacophore models can identify the common features shared across the whole chemical.

Despite seeming to perform well, there were still some issues with these models. All three models had one training chemical with a near perfect fit to the model, significantly larger than the other training chemicals' fit values, suggesting the model is strongly based on a single chemical in each case. Looking at experimental results, this chemical was not the most potent binder. Even though the structural alert substructures aligned in the training chemicals when fitted to the pharmacophores, the models did not pick out all features present in the structural alert. This could be due to the upper limit on features present in a pharmacophore model, in which case the algorithm must arbitrarily pick a selection of features present in the structural alert. This selection of features may not represent the true nature of the receptor binding mode.

The generalisation of aromatic substructure process (Chapter 5) provides more information regarding which aromatic heteroatoms in a structural alert are required for activity. This helps identify which features within the alert are involved in the binding mode. Aromatic generalisation also increases the number of active chemicals containing the alert. A more diverse subset can be selected from the larger pool of active chemicals containing the alert, which should lead to generation of a better pharmacophore. However, when a pharmacophore model was built from the generalised alert, it was a poor model. There was no longer a clear distinction between the distribution of fit values in active chemicals with the generalised alerts and the other chemicals (active or inactive). Rather than improving the pharmacophore models, the additional information provided by using generalised alerts has made the models worse. Instead of only picking the features within the generalised structural alert that are shared across all chemicals, the pharmacophore model picks no features from the structural alert at all. There is no alignment in orientations of training chemicals to the resultant pharmacophore and the model does not define the features specific to the binding mode. Consequently, the pharmacophore model cannot clearly distinguish active chemicals containing the alert from other active chemicals or inactive chemicals and is therefore a poor model.

The excellent pharmacophore models developed from structural alerts without generalisation shows the potential of combining structural alerts and pharmacophores. Generalisation of aromatic structural alerts should further expand this potential. However, using generalised alerts has led to problems with the HipHop algorithm. Some ideas for producing better pharmacophore models with the HipHop algorithm from the generalised alerts have been suggested for future exploration, although the best solution may be to use a different algorithm entirely.

7. Conclusions

Toxicity testing of chemicals is currently undergoing its largest ever paradigm shift, moving towards faster, cheaper and more human-relevant methods which focus on mechanistic understanding. An AOP provides a framework for organising biological knowledge and data. The gateway to an AOP is the MIE, and chemistry is key to predicting which chemicals can undergo a MIE. *In silico* predictions of MIEs are a vital tool in a modern, mechanism-focused approach to risk assessment of chemicals.

Improvements to Structural Alert Models

Structural alert-based SAR models have been constructed for Bowes targets[55] that significantly improve upon previous models.[44] An automated workflow has been designed for constructing these models, using Bayesian statistics to select substructures common to active chemicals but not inactive chemicals. Models have been constructed from data sets with balanced numbers of active and inactive chemicals, with all data coming from human *in vitro* assays. The new structural alert models have very impressive performance metrics, similar to random forest models applied to the same data sets. The key advantage of the structural alert-based models is that the predictions are transparent and easily interpretable. This is particularly important in toxicity testing, where risk assessors want to know not just whether a chemical is predicted to be active, but also why the prediction has been made.

Importantly, the new structural alert models and the random forest models should not be viewed as "in competition", but as two independent, complementary SAR models to be used together to aid risk assessors in assessing potential toxicity. The two models have been combined in a consensus approach, which increases confidence in predictions and overall performance. The development of the consensus model is significant as it shows how the models for receptor binding MIEs could be used in risk assessment, comparable to how *in silico* (Q)SARs are already used in predicting mutagenicity according to the ICH M7 guideline.

The automated workflow has been used to build new structural alert-based models for 66 additional biological targets that are not Bowes targets, allowing for a broader assessment of a chemical's potential toxicity and expanding the scope of this project.

The variation in performance of the models on different target's data sets has been explained by using an approach inspired by Tropsha's "modelability index".[95] Being able to explain variation in performance in terms of a property of the data sets provides additional confidence in the model construction methods.

A process for generalising aromatic substructures has been outlined. Specific aromatic atoms within a structural alert are replaced with generalised aromatic atoms where there is sufficient data to support doing so, guided by use of Bayesian statistics.

Generalising aromatic structural alerts allows more chemicals with similar structures to be represented by the same alert. The process for generalising aromatic substructure has been incorporated into the automated workflow for construction of structural alert-based models. This constructs models which have a slight increase in average MCC in the test sets of the Bowes targets and a significant decrease in the number of structural alerts in the models compared to the non-generalised structural alert models. The use of fewer structural alerts is indicative of a less overfitted model. In this way, the generalised structural alert-based models represent important improvements over the non-generalised structural alert-based models.

These methods result in structural alert models that are a significant improvement on previous structure-based predictive tools for receptor binding MIE.

Confidence in Predictions

Having constructed improved structural alert models, new methods have been introduced that effectively measure confidence in all predictions from the models.

Confidence in an active prediction has been shown to correlate with the largest Tanimoto similarity (based on Morgan fingerprints) between the test chemical and the training active chemicals containing the same alert. The more similar an alert-containing chemical is to the training active chemicals containing the same alert, the more confidence in the active prediction. This provides a continuous measure for evaluating confidence in active predictions, allowing applicability to be assessed. A cut-off has been applied to the continuous measure of confidence to define an applicability domain for the structural alerts. Applying these applicability domains to the test sets of the Bowes targets significantly increased PPV. With the addition of applicability domains, the structural alert-based models satisfy the five key priorities set out by OECD for use of (Q)SARs for regulatory purposes.

Confidence in negative predictions is particularly important in risk assessment, where an erroneous negative prediction can potentially lead to consumers being exposed to hazardous chemicals. A new method for classifying negative predictions has been designed using Tanimoto similarity coefficients between Morgan fingerprints of a test chemical and the training set chemicals. Negative predictions have been split into three categories based on these similarities: “Similar to active” inactives, “out of domain” inactives, and “classified” inactives. Within the classified inactives category, NPV was increased in the test sets of all targets.

These new methods greatly increase the relevance and applicability of the structural alert models, particularly when they are applied to new data sets. With these methods, it is now possible to identify when the structural alert models are extrapolating from the training chemicals to different areas of chemical space, and an appropriate measure of confidence is returned with the resulting predictions. This is particularly important when applying the models to new chemicals in risk assessment.

Together, the methods in this project give a new, high-performing structure-based predictive model with improved applicability to risk assessment. A new scheme for using structural alert models to make activity predictions for receptor binding MIEs is shown in Figure 7.1.

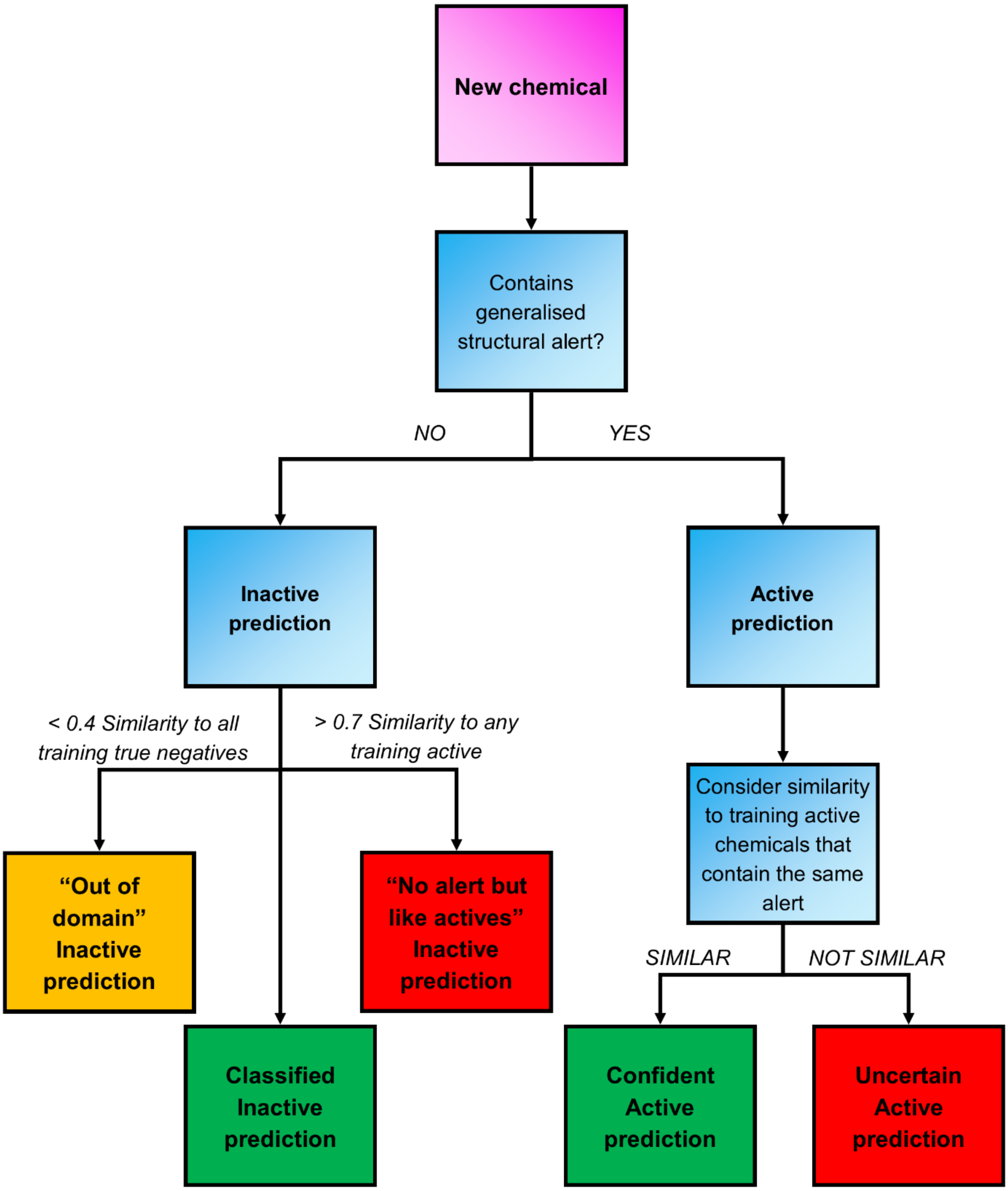

New chemical
Contains generalised structural alert?
NO
YES
Inactive prediction
Active prediction
< 0.4 Similarity to all training true negatives
> 0.7 Similarity to any training active
Consider similarity to training active chemicals that contain the same alert
"Out of domain" Inactive prediction
"No alert but like actives" Inactive prediction
SIMILAR
NOT SIMILAR
Classified Inactive prediction
Confident Active prediction
Uncertain Active prediction

Figure 7.1: The new scheme for making activity predictions of chemicals with structural alert-based models for receptor binding MIEs. The generalisation of aromatic substructure process has been integrated into the automated workflow for construction of structural alert-based models. Confidence in active and inactive predictions are then assessed using different methods. In this image, pink boxes are input chemicals, blue boxes are key steps in the process, green boxes represent high confidence predictions, yellow represent medium confidence predictions, and red represents low confidence predictions.

Receptor Binding MIEs and Reactivity Driven MIEs

In this project, structural alerts have been used to create high-performing SAR models for receptor binding MIEs. However, structural alerts only identify fragments of chemicals. The methods used for assessing confidence in predictions highlight the importance of combining structural alerts with considerations of “global” similarity – considerations of similarity between entire structures of chemicals.

In assessing confidence in active predictions, considering global similarity indicates whether the rest of the chemical beyond the structural alert substructure shares features with the active chemicals that gave rise to the structural alert.

In assessing confidence in inactive predictions for receptor binding MIEs, consideration of global similarity to training chemicals was found to be more effective than methods for identifying local changes in chemicals, like those used by Williams *et al* in reactivity driven MIEs related to mutagenicity.[76]

Despite structural alerts being used to make predictions for both receptor binding MIEs and reactivity driven MIEs, this project highlights the key differences between the mechanisms of the MIEs. To be active at a receptor binding MIE, a chemical needs a specific combination of features in a specific arrangement in three-dimensional space, whereas to be active at a reactivity driven MIE a chemical requires a single electrophilic feature. Thus, methods that apply to one type of MIE may not be applicable to the other MIE. Structural alerts for reactivity driven MIEs identify electrophilic groups and require less context in terms of how the alert relates to the rest of a chemical’s structure. Structural alerts for receptor binding MIEs identify common scaffolds or fragments associated with activity and judging the applicability of these alert requires a consideration of the rest of a chemical’s structure.

Beyond Structural Alerts

Whilst correlating well with biological activity, structural alerts define only specific combinations of atoms and bonds, which can be limiting when making activity predictions for chemicals from different regions of chemical space. The generalisation of aromatic structural alerts process is an important way of reducing this limitation. Further methods have been explored to identify the key features shared by chemicals containing the same structural alert, and the key interactions these features make with the biological targets. This would allow an understanding of how a receptor binding MIE may occur, and this understanding helps to make better models for the MIE, from which better, more general predictions can be made.

Examining the generalised aromatic structural alerts gives more information on the mechanisms involved in the chemicals binding to the target. By looking at all chemicals covered by the same generalised alerts, one can identify which common features within the alert are necessary for activity. An example of this was shown for a generalised structural alert for the adenosine A2a receptor. Predictions of key receptor binding interactions where made by examining the generalised alert and were found to be consistent with binding interactions derived from crystal structures for similar chemicals.

Ligand-based pharmacophore models identify a common three-dimensional arrangement of shared features of biologically active chemicals in an attempt to identify the specific arrangement of features required to activate a receptor binding MIE. Structural alerts have been used as a basis for grouping chemicals from which pharmacophores were constructed using the HipHop algorithm.[104] Three good pharmacophore models have been constructed from three different non-generalised structural alerts.

Despite these models seeming to perform well, there were still some issues with construction of pharmacophores. The models overfit to a single chemical in each case and they did not pick out all overlapping features. Contrary to expectations, poor models were constructed when using generalised structural alerts to group chemicals. Instead of producing better, more general models, the increase in diversity of chemicals became problematic for the HipHop algorithm. Ideas for correcting these issues have been suggested and will be explored in future work.

Even with some issues, the success of the models from non-generalised structural alerts shows the potential of building pharmacophores from structural alerts. Pharmacophores expand upon the two-dimensional substructures described by structural alerts, instead describing three-dimensional similarities across the whole chemical structure – a global consideration of similarity, albeit a more complex one than Tanimoto similarity between Morgan fingerprints (as used previously in this project). They are a step towards describing chemicals in terms of the key interactions made with the biological target. Thus, further development of this pharmacophore approach could result in better models founded in the mechanism of the MIE.

Final Remarks

Overall, the work presented within this project fulfils the main aim of making interpretable predictions for MIEs based on chemical structures. High performing structural alert-based SAR models have been constructed that make accurate, transparent, and easily interpretable predictions for receptor binding MIEs. The addition of new methods for assessing confidence and applicability of both active and inactive predictions are vital in applying the models to new chemicals. This project makes significant contributions and advancements to the topic of structure-based predictions for MIEs, which will be particularly important for use in assessing toxicity of chemicals.

References

1 S. Gutsell and P. J. Russell, The role of chemistry in developing understanding of adverse outcome pathways and their application in risk assessment, *Toxicol. Res. (Camb).*, 2013, **2**, 299–307.

2 E. Gottmann, S. Kramer, B. Pfahringer and C. Helma, Data quality in predictive toxicology: Reproducibility of rodent carcinogenicity experiments, *Environ. Health Perspect.*, 2001, **109**, 509–514.

3 H. Olson, G. Betton, D. Robinson, K. Thomas, A. Monro, G. Kolaja, P. Lilly, J. Sanders, G. Sipes, W. Bracken, M. Dorato, K. Van Deun, P. Smith, B. Berger and A. Heller, Concordance of the toxicity of pharmaceuticals in humans and in animals., *Regul. Toxicol. Pharmacol.*, 2000, **32**, 56–67.

4 D. Fourches, J. C. Barnes, N. C. Day, P. Bradley, J. Z. Reed and A. Tropsha, Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species, *Chem. Res. Toxicol.*, 2010, **23**, 171–183.

5 D. Mulliner, F. Schmidt, M. Stolte, H.-P. Spirkl, A. Czich and A. Amberg, Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope, *Chem. Res. Toxicol.*, 2016, **29**, 757–767.

6 Health and Safety Executive REACH Homepage, http://www.hse.gov.uk/reach/index.htm, (accessed 12 April 2019).

7 National Research Council, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, The National Academies Press, Washington, DC, 2007.

8 H. Kitano, Systems Biology: A Brief Overview, *Science (80-.).*, 2002, **295**, 1662 LP – 1664.

9 M. J. Aardema and J. T. MacGregor, Toxicology and genetic toxicology in the new era of 'toxicogenomics': Impact of '-omics' technologies, *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, 2002, **499**, 13–25.

10 P. R. Bushel, M. Deveau, R. S. Thomas, M. Husain, D. Krewski, S. Auerbach, I. D. Moffat, C. L. Yauk, A. Williams and J. A. Bourdon-Lacombe, Technical guide for applications of gene expression profiling in human health risk assessment of environmental chemicals, *Regul. Toxicol. Pharmacol.*, 2015, **72**, 292–309.

11 E. J. Perkins, T. W. Collette, G. Hodges, T. H. Hutchinson, A. Cossins, F. Falciani, N. Garcia-Reyero, S. Plaistow, P. Graystock, D. Taylor, M. Hecker, D. Becker, P. Kille, M. Viant, A. Lange,

D. Knapen, A. Schroeder, J. Colbourne, E. Butler, G. Ankley, S. Gutsell, K. Chipman, S. F. Owen, M. Cronin, I. Katsiadaki, S. Marshall, E. K. Brockmeier and K. E. Tollefsen, The Role of Omics in the Application of Adverse Outcome Pathways for Chemical Risk Assessment, *Toxicol. Sci.*, 2017, **158**, 252–262.

12 G. T. Ankley, R. S. Bennett, R. J. Erickson, D. J. Hoff, M. W. Hornung, R. D. Johnson, D. R. Mount, J. W. Nichols, C. L. Russom, P. K. Schmieder, J. A. Serrrano, J. E. Tietge and D. L. Villeneuve, Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment., *Environ. Toxicol. Chem.*, 2010, **29**, 730–41.

13 OECD, Proposal for a Template, and Guidance on Developing and Assessing the Completeness of Adverse Outcome Pathways, Appendix I, Collection of Working Definitions, 2012, http://www.oecd.org/chemicalsafety/testing/49963554.pdf

14 B. van Ravenzwaay, D. Drasdo, A. Ghallab, A. Terron, A. Limonciel, R. Reif, T. Hartung, H. Kamp, B. Hardy, C. Cadenas, B. van der Burg, A. Mantovani, H. Vrieling, E. Willighagen, T. Braunbeck, C. Evelo, S. Höhme, S. Escher, S. Dooley, J. Beltman, R. Graepel, F. Y. Bois, B. Zdrazil, R. Hassan, P. Jennings, C. Fisher, A. Forsby, D. Fluri, E. Danen, P. Godoy, R. Marchan, E. Mombelli, G. Ecker, J. G. Hengstler, D. Gadaleta, M. Leist, D. Kroese, S. H. Bennekou, A. Braeuning, M. Martens, F. Oesch, S. Schildknecht, I. Gardner, J. Kelm, C. van Thriel, A. H. Meijer, O. Taboureau, B. van de Water, M. Schwarz, F. Sanz, M. Vinken, T. Waldmann and B. Koch, Adverse outcome pathways: opportunities, limitations and open questions, *Arch. Toxicol.*, 2017, **91**, 3477–3505.

15 T. E. H. Allen, J. M. Goodman, S. Gutsell and P. J. Russell, Defining molecular initiating events in the adverse outcome pathway framework for risk assessment, *Chem. Res. Toxicol.*, 2014, **27**, 2100–2112.

16 T. E. H. Allen, J. M. Goodman, S. Gutsell and P. J. Russell, A History of the Molecular Initiating Event, *Chem. Res. Toxicol.*, 2016, **29**, 2060–2070.

17 M. T. D. Cronin and A.-N. Richarz, Relationship Between Adverse Outcome Pathways and Chemistry-Based In Silico Models to Predict Toxicology, *Appl. Vitr. Toxicol.*, 2017, **3**, 286–297.

18 H. M. Jones and K. Rowland-Yeo, Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development, *CPT Pharmacometrics Syst. Pharmacol.*, 2013, **2**, 1–12.

19 S. Russmann, G. A. Kullak-Ublick and I. Grattagliano, Current concepts of mechanisms in drug-induced hepatotoxicity., *Curr. Med. Chem.*, 2009, **16**, 3041–53.

20 L. Carlsson, O. Spjuth, S. Adams, R. C. Glen and S. Boyer, Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse, *BMC Bioinformatics*, , DOI:10.1186/1471-2105-11-362.

21 Lhasa's Meteor Nexus, http://www.lhasalimited.org/products/meteor-nexus.htm, (accessed 12 April 2019).

22 J. Kirchmair, M. J. Williamson, J. D. Tyzack, L. Tan, P. J. Bond, A. Bender and R. C. Glen, Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms, *J. Chem. Inf. Model.*, 2012, **52**, 617–648.

23 M. M. Angrish, J. M. O'Brien, X. Zhang, N. Pollesch, D. L. Villeneuve, L. Margiotta-Casaluci, L. C. Smith, D. Knapen, S. Munn, M. C. Fortin, I. Katsiadaki and M. Leonard, Adverse outcome pathway networks I: Development and applications, *Environ. Toxicol. Chem.*, 2018, **37**, 1723–1733.

24 D. L. Villeneuve, M. M. Angrish, M. C. Fortin, I. Katsiadaki, M. Leonard, L. Margiotta-Casaluci, S. Munn, J. M. O'Brien, N. L. Pollesch, L. C. Smith, X. Zhang and D. Knapen, Adverse outcome pathway networks II: Network analytics, *Environ. Toxicol. Chem.*, 2018, **37**, 1734–1748.

25 G. Maxwell, C. MacKay, R. Cubberley, M. Davies, N. Gellatly, S. Glavin, T. Gouin, S. Jacquoilleot, C. Moore, R. Pendlington, O. Saib, D. Sheffield, R. Stark and V. Summerfield, Applying the skin sensitisation adverse outcome pathway (AOP) to quantitative risk assessment, *Toxicol. Vitr.*, 2014, **28**, 8–12.

26 R. B. Conolly, G. T. Ankley, W. Cheng, M. L. Mayo, D. H. Miller, E. J. Perkins, D. L. Villeneuve and K. H. Watanabe, Quantitative Adverse Outcome Pathways and Their Application to Predictive Toxicology, *Environ. Sci. Technol.*, 2017, **51**, 4661–4672.

27 Organization for Economic Co-operation and Development, *Principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models*, 2007.

28 G. Patlewicz, A. O. Aptula, D. W. Roberts and E. Uriarte, A minireview of available skin sensitization (Q)SARs/expert systems, *QSAR Comb. Sci.*, 2008, **27**, 60–76.

29 M. T. D. Cronin and T. W. Schultz, Pitfalls in QSAR, *J. Mol. Struct. THEOCHEM*, 2003, **622**, 39–51.

30 L. Breiman, Random Forests, *Mach. Learn.*, 2001, **45**, 5–32.

31 V. Svetnik, A. Liaw, C. Tong and T. Wang, in *Svetnik V., Liaw A., Tong C., Wang T. (2004) Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of*

Pharmaceutical Molecules. In: Roli F., Kittler J., Windeatt T. (eds) Multiple Classifier Systems. MCS 2004. Lecture Notes i.

32 C. Cortes and V. Vapnik, Support-vector networks, *Mach. Learn.*, 1995, **20**, 273–297.

33 R. Burbidge, M. Trotter, B. Buxton and S. Holden, Drug design by machine learning: Support vector machines for pharmaceutical data analysis, *Comput. Chem.*, 2001, **26**, 5–14.

34 N. S. Altman, An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, *Am. Stat.*, 1992, **46**, 175–185.

35 W. Zheng and A. Tropsha, Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 185–194.

36 J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci.*, 1982, **79**, 2554–2558.

37 A. Mayr, G. Klambauer, T. Unterthiner and S. Hochreiter, DeepTox: Toxicity Prediction using Deep Learning, *Front. Environ. Sci.*, 2016, **3**, 80.

38 C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, Conditional variable importance for random forests., *BMC Bioinformatics*, 2008, **9**, 307.

39 C.E. Shannon , vol. 27, pp. , July, October, A Mathematical Theory of Communication, *Bell Syst. Tech. J.*, 1948, **27**, 379–423,623–656.

40 L. H. Mervin, A. M. Afzal, L. Brive, O. Engkvist and A. Bender, Extending in silico protein target prediction models to include functional effects, *Front. Pharmacol.*, 2018, **9**, 1–13.

41 S. Muresan, P. Petrov, C. Southan, M. J. Kjellberg, T. Kogej, C. Tyrchan, P. Varkonyi and P. H. Xie, Making every SAR point count: The development of Chemistry Connect for the large-scale integration of structure and bioactivity data, *Drug Discov. Today*, 2011, **16**, 1019–1030.

42 Lhasa's Derek Nexus, http://www.lhasalimited.org/products/derek-nexus.htm, (accessed 12 April 2019).

43 T. E. H. Allen, S. Liggi, J. M. Goodman, S. Gutsell and P. J. Russell, Using Molecular Initiating Events to Generate 2D Structure-Activity Relationships for Toxicity Screening, *Chem. Res. Toxicol.*, 2016, **29**, 1611–1627.

44 T. E. H. Allen, J. M. Goodman, S. Gutsell and P. J. Russell, Using 2D Structural Alerts to Define Chemical Categories for Molecular Initiating Events, *Toxicol. Sci.*, 2018, **165**, 213–223.

45 M. D. Nelms, C. L. Mellor, M. T. D. Cronin, J. C. Madden and S. J. Enoch, Development of an in Silico Profiler for Mitochondrial Toxicity, *Chem. Res. Toxicol.*, 2015, **28**, 1891–1902.

46 C. L. Mellor, F. P. Steinmetz and M. T. D. Cronin, Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis, *Chem. Res. Toxicol.*, 2016, **29**, 203–212.

47 C. L. Mellor, F. P. Steinmetz and M. T. D. Cronin, The identification of nuclear receptors associated with hepatic steatosis to develop and extend adverse outcome pathways, *Crit. Rev. Toxicol.*, 2016, **46**, 138–152.

48 OECD QSAR Toolbox, http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm, (accessed 12 April 2019).

49 D. Yordanova, T. W. Schultz, C. Kuseva, H. Ivanova, T. Pavlov, G. Chankov, Y. Karakolev, A. Gissi, T. Sobanski and O. G. Mekenyan, Alert performance: A new functionality in the OECD QSAR Toolbox, *Comput. Toxicol.*, 2019, **10**, 26–37.

50 C. G. Wermuth, C. R. Ganellin, P. Lindberg and L. A. Mitscher, Chapter 36. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997, *Annu. Rep. Med. Chem.*, 1998, **33**, 385–395.

51 I. Tsakovska, M. Al Sharif, P. Alov, A. Diukendjieva, E. Fioravanzo, M. T. D. Cronin and I. Pajeva, Molecular modelling study of the PPARγ receptor in relation to the mode of action/adverse outcome pathway framework for liver steatosis, *Int. J. Mol. Sci.*, 2014, **15**, 7651–7666.

52 S. Y. Yang, Pharmacophore modeling and applications in drug discovery: Challenges and recent advances, *Drug Discov. Today*, 2010, **15**, 444–450.

53 International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use, *ICH M7 Assessment and Control Of DNA Reactive (Mutagenic) Impurities In Pharmaceuticals To Limit Potential Carcinogenic Risk*, 2014.

54 C. Barber, A. Amberg, L. Custer, K. Dobo, S. Glowienke, J. Van Gompel, S. Gutsell, J. Harvey, M. Honma, M. Kenyon, N. Kruhlak, W. Muster, L. Stavitskaya, A. Teasdale, J. Vessey and J. Wichard, Establishing best practise in the application of expert review of mutagenicity under ICH M7, *Regul. Toxicol. Pharmacol.*, 2015, **73**, 367–377.

55 J. Bowes, A. J. Brown, J. Hamon, W. Jarolimek, A. Sridhar, G. Waldron and S. Whitebread, Reducing safety-related drug attrition: the use of in vitro pharmacological profiling, *Nat. Rev. Drug Discov.*, 2012, **11**, 909–922.

56 AOP Wiki, https://aopwiki.org/, (accessed 12 April 2019).

57 A. M. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, John Willey & Sons, 1990.

58 G. Maggiora, M. Vogt, D. Stumpfe and J. Bajorath, Molecular similarity in medicinal chemistry, *J. Med. Chem.*, 2014, **57**, 3186–3204.

59 H. L. Morgan, The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service, *J. Chem. Doc.*, 1965, **5**, 107–113.

60 D. Rogers and M. Hahn, Extended-Connectivity Fingerprints, *Acs.org*, 2010, **50**, 742–754.

61 ChemAxon Extended Connectivity Fingerprint Documentation, https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP, (accessed 5 March 2019).

62 MACCS keys. MDL Information Systems. Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse, Reoptimization of MDL Keys for Use in Drug Discovery, J. Chem. Inf. Comput. Sci., 2002, 42, 1273–1280

63 Rogers DJ and Tanimoto TT, A Computer Program for Classifying Plants, *Science*, 1960, **132**, 1115–1118.

64 D. Bajusz, A. Rácz and K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?, *J. Cheminform.*, 2015, **7**, 1–13.

65 A. Bender and R. C. Glen, Molecular similarity: a key technique in molecular informatics, *Org. Biomol. Chem.*, 2004, **2**, 3204–3218.

66 J. D. Holliday, N. Salim, M. Whittle and P. Willett, Analysis and display of the size dependence of chemical similarity coefficients, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 819–828.

67 KNIME, https://www.knime.org/, (accessed 4 April 2018).

68 Node description for MoSS, https://www.knime.org/files/nodedetails/_chemistry_mining_MoSS.html, (accessed 12 April 2019).

69 C. Borgelt and M. R. Berthold, Mining molecular fragments: finding relevant substructures of molecules, 2003, 51–58.

70 G. Roberts, G. J. Myatt, W. P. Johnson, K. P. Cross and P. E. Blower, LeadScope : Software for

Exploring Large Sets of Screening Data, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 1302–1314.

71 G. Klopman, Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules, *J. Am. Chem. Soc.*, 1984, **106**, 7315–7321.

72 G. Klopman, MULTICASE 1. A Hierarchical Computer Automated Structure Evaluation Program, *Quant. Struct. Relationships*, 1992, **11**, 176–184.

73 I. Cortes-Ciriano, Bioalerts: A python library for the derivation of structural alerts from bioactivity and toxicity data sets, *J. Cheminform.*, 2016, **8**, 4–9.

74 T. Ferrari, D. Cattaneo, G. Gini, N. Golbamaki Bakhtyari, A. Manganaro and E. Benfenati, Automatic knowledge extraction from chemical structures: The case of mutagenicity prediction, *SAR QSAR Environ. Res.*, 2013, **24**, 631–649.

75 EBI ChEMBL, https://www.ebi.ac.uk/chembl/, (accessed 27 April 2018).

76 R. V Williams, A. Amberg, A. Brigo, L. Coquin, A. Giddings, S. Glowienke, N. Greene, R. Jolly, R. Kemper, C. O'Leary-Steele, A. Parenty, H. P. Spirkl, S. A. Stalford, S. K. Weiner and J. Wichard, It's difficult, but important, to make negative predictions, *Regul. Toxicol. Pharmacol.*, 2016, **76**, 79–86.

77 A. Gaulton, A. Hersey, M. L. Nowotka, A. Patricia Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magarinos, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, The ChEMBL database in 2017, *Nucleic Acids Res.*, 2017, **45**, D945–D954.

78 W. L. Duax and J. F. Griffin, Structural features which distinguish estrogen agonists and antagonists, *J. Steroid Biochem.*, , DOI:10.1016/0022-4731(87)90318-9.

79 A. Koutsoukas, R. Lowe, Y. Kalantarmotamedi, H. Y. Mussa, W. Klaffke, J. B. O. Mitchell, R. C. Glen and A. Bender, In silico target predictions: Defining a benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window, *J. Chem. Inf. Model.*, 2013, **53**, 1957–1966.

80 L. H. Mervin, A. M. Afzal, G. Drakakis, R. Lewis, O. Engkvist and A. Bender, Target prediction utilising negative bioactivity data covering large chemical space, *J. Cheminform.*, 2015, **7**, 1–16.

81 R. Kavlock, The future of toxicity testing—The NRC vision and the EPA's ToxCast program national center for computational toxicology, *Neurotoxicol. Teratol.*, 2009, **31**, 237.

82 T. B. Knudsen, N. Sipes, K. Houck, R. Judson, R. Huang, D. Reif, K. Crofton, M. Martin, D. Rotroff, D. Filer, M. Xia, J. Wambaugh, J. Phuong, I. Shah, S. Little, N. Kleinstreuer, R. Woodrow Setzer, P. Kothya, A. M. Richard, D. Smith and R. S. Thomas, Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space, *Toxicol. Sci.*, 2016, **152**, 323–339.

83 Rational Discovery Kit, www.rdkit.org, (accessed 12 April 2019).

84 T. I. Oprea, C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, A. Jadhav, L. J. Jensen, G. L. Johnson, A. Karlson, A. R. Leach, A. Ma'ayan, A. Malovannaya, S. Mani, S. L. Mathias, M. T. McManus, T. F. Meehan, C. Von Mering, D. Muthas, D. T. Nguyen, J. P. Overington, G. Papadatos, J. Qin, C. Reich, B. L. Roth, S. C. Schürer, A. Simeonov, L. A. Sklar, N. Southall, S. Tomita, I. Tudose, O. Ursu, D. Vidović, A. Waller, D. Westergaard, J. J. Yang and G. Zahoránszky-Köhalmi, Unexplored therapeutic opportunities in the human genome, *Nat. Rev. Drug Discov.*, 2018, **17**, 317–332.

85 D. L. Filer, P. Kothiya, W. R. Setzer, R. S. Judson and M. T. Martin, *The ToxCast Analysis Pipeline : An R Package for Processing and Modeling Chemical Screening Data*, 2014.

86 M. Yoon, J. L. Campbell, M. E. Andersen and H. J. Clewell, Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results, *Crit. Rev. Toxicol.*, 2012, **42**, 633–652.

87 D. Barber, *Bayesian Reasoning and Machine Learning*, Cambridge University Press, 2011.

88 R. Van de Schoot, D. Kaplan, J. Denissen, J. B. Asendorpf, F. J. Neyer and M. A. G. van Aken, A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research, *Child Dev.*, 2014, **85**, 842–860.

89 D. Madigan, P. Ryan, S. Simpson and I. Zorych, in *Bayesian Statistics 9*, Oxford University Press, 2011, pp. 421 – 438.

90 W. G. Phillips and J. M. Rejda-Heath, Thiazole carboxanilide fungicides: A new structure-activity relationship for succinate dehydrogenase inhibitors, *Pestic. Sci.*, 1993, **38**, 1–7.

91 B. T. Priest, I. M. Bell and M. L. Garcia, Role of hERG potassium channel assays in drug development, *Channels*, 2008, **2**, 87–93.

92 S. Kalyaanamoorthy and K. H. Barakat, Binding modes of hERG blockers: An unsolved mystery in the drug design arena, *Expert Opin. Drug Discov.*, 2018, **13**, 207–210.

93 scikit-learn, https://scikit-learn.org/stable/, (accessed 12 April 2019).

94 N. S. Sipes, M. T. Martin, P. Kothiya, D. M. Reif, R. S. Judson, A. M. Richard, K. A. Houck, D. J.

Dix, R. J. Kavlock and T. B. Knudsen, Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays, *Chem. Res. Toxicol.*, 2013, **26**, 878–895.

95 A. Golbraikh, E. Muratov, D. Fourches and A. Tropsha, Data set modelability by QSAR, *J. Chem. Inf. Model.*, 2014, **54**, 1–4.

96 G. M. Maggiora, On Outliers and Activity CliffsWhy QSAR Often Disappoints, *J. Chem. Inf. Model.*, 2006, **46**, 1535–1535.

97 H. S. Rosenkranz and A. R. Cunningham, SAR modeling of unbalanced data sets., *SAR QSAR Environ. Res.*, 2001, **12**, 267–274.

98 C. A. Marchant, K. A. Briggs and A. Long, In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic, *Toxicol. Mech. Methods*, 2008, **18**, 177–187.

99 M. L. Chilton, D. S. Macmillan, T. Steger-Hartmann, J. Hillegass, P. Bellion, A. Vuorinen, S. Etter, B. P. C. Smith, A. White, P. Sterchele, A. De Smedt, M. Glogovac, S. Glowienke, D. O'Brien and R. Parakhia, Making reliable negative predictions of human skin sensitisation using an in silico fragmentation approach, *Regul. Toxicol. Pharmacol.*, 2018, **95**, 227–235.

100 C. Borgelt, T. Meinl and M. Berthold, MoSS: a program for molecular substructure mining, *Proc. 1st Int. Work. open source data Min.*, 2005, 6–15.

101 G. Lebon, T. Warne, P. C. Edwards, K. Bennett, C. J. Langmead, A. G. W. Leslie and C. G. Tate, Agonist-bound adenosine A2Areceptor structures reveal common features of GPCR activation, *Nature*, 2011, **474**, 521–526.

102 RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/ (accessed 12th April 2019)

103 D. M. Hawkins, The Problem of Overfitting, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1–12.

104 M. Ashton, J. Barnard, F. Casset, M. Charlton, G. Downs, D. Gorse, J. Holliday, R. Lahana and P. Willett, Identification of diverse database subsets using property-based and fragment-based molecular descriptions, *Quant. Struct. Relationships*, 2002, **21**, 598–604.

105 Ref. Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017, San Diego: Dassault Systèmes, 2016.

106 W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Crystallogr. Sect. A*, 1978, **34**, 827–828.

107 OECD, The Adverse Outcome Pathway for Skin Sensitization Initiated by Covalent Binding

to Proteins – Part 1 Series on Testing and Assessment No.168, http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2012)10/part1&doclanguage=en, (accessed 21 June 2016).