Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis

Claire L. Mellor, Fabian P. Steinmetz and Mark T.D. Cronin*

*Corresponding author

Mark T.D Cronin

School of Pharmacy and Biomolecular Sciences
Liverpool John Moores University
Byrom Street
Liverpool
L3 3AF
England

Tel. +44 151 231 2402;

e-mail address: M.T.Cronin@ljmu.ac.uk

Keywords

Nuclear receptors, hepatic steatosis, structural alerts, workflow.

TABLE OF CONTENTS GRAPHIC HERE

ABSTRACT

In silico models are essential for the development of integrated alternative methods to identify organ level toxicity and lead towards the replacement of animal testing. These models include (quantitative) structure-activity relationships ((Q)SARs) and, importantly, the identification of structural alerts associated with defined toxicological endpoints. Structural alerts are able both to predict toxicity directly and assist in the formation of categories to facilitate read-across. They are particularly important to decipher the myriad mechanisms of action that result in organ level toxicity. The aim of this study was to develop novel structural alerts for nuclear receptor (NR) ligands that are associated with inducing hepatic steatosis and to show the vast amount of current data that are available. Current knowledge on NR agonists was extended with data from the ChEMBL database (12,713 chemicals in total) of bioactive molecules and from studying NR ligand-binding interactions within the protein data base (PBD, 624 human NR structure files). A computational structural alerts based workflow was developed using KNIME from these data using molecular fragments and other relevant chemical features. In total 214 structural features were recorded computationally as SMARTS strings and, therefore, they can be used for grouping and screening during drug development and hazard assessment and provide knowledge to anchor adverse outcome pathways (AOPs) via there molecular initiating effects (MIE).

INTRODUCTION

Nuclear receptors (NR) belong to a large superfamily of ligand-inducible transcription factors that, upon activation, mediate the expression of their target genes.[1] The ligands associated with NR activation are usually lipophilic, small in size and include the following chemical classes: endogenous steroids, oxysterols, thyroid hormones as well as various lipids and retinoids.[2,3] NRs are essential for the regulation of specific target genes that are involved in development, metabolism, reproduction and other vital physiological processes. Upon ligand-induced activation, NRs elicit a rapid cellular adaptation to environmental changes via the induction of the required genes and pathways.[4]

Due to their involvement in many essential processes within the body, the search for novel ligands for nuclear receptor(s) (NR(s)) has intensified in order to identify possible preventative / therapeutic treatments for a wide range of diseases including diabetes, cancer, cardiovascular diseases, atherosclerosis, neurodegenerative diseases and obesity.[3,4] For example, the oestrogen receptor (ER) antagonist tamoxifen is used for the treatment of ER positive breast cancer.[5-7] As NR ligands are widely used it is imperative their safety is considered, as there are reports of NR ligands leading to drug induced liver injury (DILI), such as liver steatosis, due to the bio-activation of drugs (or metabolites) and / or the induction of hepatotoxic pathways. [8-11]

Traditional approaches to determine safety have required the use of animal tests. However, the promotion of what is termed "21st Century Toxicology" has led to the move from traditional animal testing safety assessment methods to the use of integrated alternative strategies which utilise toxicokinetics, computational models and *in vitro* testing.[10,12-14, 15] The shift in the mind set occurring within toxicology has given rise to the concept of the Adverse Outcome Pathway (AOP) framework.[13-14, 16-18] An AOP describes the causal linkage between a molecular

initiating event (MIE) and an adverse outcome at individual or population level.[12,18] The AOP developed within the current AOP Wiki knowledge base (AOP-KB) for hepatic steatosis defines liver toxicity as the adverse effect and nuclear receptor binding being the MIE, thus this knowledge provides the starting point for computational methods.[14,18,19]

Computational methods include the use of (quantitative) structure-activity relationships ((Q)SARs) as well as other approaches including biokinetic models. QSARs require the use of mathematical models in order to predict biological activity of chemicals from their structure or physico-chemical properties. An SAR is a qualitative link between a certain molecular substructure to a specific biological activity.[20] Structural alert(s) (SA(s)) derived from SARs can aid in the formation of categories with similar chemicals that are associated to share the same SAR. The assessment of these category members can, in turn, allow for the better definition of the domain of the SA.[20,21] SA are common structural fragments that are associated with a specific toxicity which very often have mechanistic rational to support these links – with reference to an AOP this is in terms of the MIE. [20-22] SAs are already used to screen potential lead chemicals for idiosyncratic toxicity within industry settings.[24] Thus, through knowledge of the AOP, they can form the first step in understanding the links from a specific chemical to its possible mechanistic pathways and those organs that may be affected.[20] For the AOP concept to be applied, particularly to support category formation and read-across, an SA associated with a MIE for a particular adverse pathway must be elucidated and described.

The use of new (toxicological and informatics) approaches can help to aide in the formation of SAs; for instance, applying freely available software and utilising the growing number of open access databases of toxicological information.[25] For this study, the new methods approaches used included the ChEMBL database of bioactive molecules (with >1.5 million compounds and 9,000 biological targets), KoNstanz Inforamtion MinEr (KNIME) software (allows the analysis/ mining of data and can be used to build predictive models) and the protein data bank

(PBD), (a database of > 100,000 crystallographic structures of proteins e.g. receptors which can be analysed alongside software such as PyMOL and Marvin Beans, a ChemAxon suite of programmes that allow the visualisation and drawing of chemicals, all of which are freely available).[26-30]

Mellor et al. (2015) reviewed the NRs linked to liver injury, identifying ten NRs that can cause the onset of hepatic steatosis, these are summarised in Table 1.[18] Each of these NRs is associated with a definable mechanism of action and / or toxicity pathway that could form the basis of an AOP. A MIE is definable for each NR, therefore with analysis of suitable data for NR binding, this raises the possibility of defining a suite of SA which could form the basis of toxicity prediction or grouping. Thus, the aim of this study was to develop a set of SAs for the NRs associated with hepatic steatosis listed in Table 1. This was achieved with reference to the MIEs for the NR and utilised the ChEMBL database as a source of information. A workflow was created to collate knowledge that can predict binding to the NR listed in Table 1.[26] The workflow was developed in a two-step process. The first step involved identifying the physico-chemical properties that define the chemical space/ domain of agonists of the NR through the calculation of descriptors from studying the chemical structure. The second step involved the identification of structural features associated with NR binding which were then coded into SMARTS strings so they could be implemented in the workflow. The identification of structural features was performed by studying the ligand-binding interactions of the agonists to their respective NR using the PDB files viewed in PyMol and by studying the literature associated with these files (referenced within PDB).[28,29,31] The workflow can be used for hazard assessment, screening of potential ligands for chemical leads or grouping. The workflow is freely available to use and download from COSMOS Space (http://knimewebportal.cosmostox.eu) To view a web tutorial describing the use of the workflow please use the following link (https://www.youtube.com/watch?v=ggkU6lZfDfY).

TABLE 1 HERE

METHODS

Identifying chemical structures with relevant NR binding data from ChEMBL

The 10 NR listed in Table 1 were searched for within the online ChEMBL database (Version19) using their names and standard nomenclature identifiers in order to find agonists.[26] To identify agonists, the names and / or nomenclature of the NRs, as reported in Table 1, were entered into the search bar within the online ChEMBL website with *Homo sapiens* selected as the species of interest. Data retrieved were downloaded to comma-separated values (.csv) files and later saved in an Excel spreadsheet. Those data with pChEMBL values were selected and all other chemicals without these values were removed. The pChEMBL value is an approach to standardise different types of activity values.[32] In addition, the following information was obtained from ChEMBL: the chemical name, molecular formula, SMILES string of each agonist, the assay type (e.g. receptor activation), activity value (reported as Ki, Kd, AC50, IC50, and EC50) and other relevant information regarding the assay. Only agonistic Ki, Kd, AC50 or EC50 values were utilised to remove data relating to inhibition of receptors (e.g. those with IC50 values / Ki values) and to ensure receptor activation data were considered. Chemicals were then ordered by pChEMBL value (highest to lowest) and those with a pChEMBL of <5 removed. The pChEMBL of > 5 was used as the threshold for activity as when studying the ChEMBL data this was the point at which most of chemicals were considered active and has been utilised previously.[33,34] Duplicate chemicals were removed at this point. Agonists with pChEMBL values above the threshold value were studied using the Marvin Bean (version 1.6) chemical visualisation software in order to find common structural features associated with NR activation; these features were recorded as SMARTS strings.[30,31]

Analysis of ligand-binding information from the Protein Data Bank (PDB)

The PDB was investigated in conjunction with the list of agonists and associated data obtained from ChEMBL to further study the ligand-binding of agonists to their respective NR. The NR names /nomenclature (Table 1) were searched for on the PDB website and human NR files containing information about agonistic binding to human NR structures were selected. The selected files were viewed to study their ligand-protein interactions using PyMOL (version 1.3) and the linked publications provided within the PDB were read to find the key amino acid residues on the NR binding site that have been shown to interact with specific functional groups on the ligand. The functional groups on the agonists needed for these essential binding interactions with the NR binding site such as hydroxyl moiety group were then drawn as SMILES strings and added to form the rules of the workflow. [28, 29, 31]

Calculation of Molecular Descriptors

Molecular properties and other descriptors for the agonists were calculated using the CDK node in KNIME (version 3.2).[27] SMILES strings for the ligands were retrieved from ChEMBL and then cleaned (removal of salts, inorganics and mixtures). The SMILES were entered into the CDK node and all available descriptors (33 available within CDK node) were calculated. Descriptors were identified that described features relevant to ligand-protein interactions and gave a specific range of values for each NR. In total eight descriptors were used: molecular weight (MW, describing molecular size), the calculated logarithm of the octanol-water partition coefficient (xlogP, lipophilicity), vertex adjacency matrix (VAIM, molecular size and complexity), number of hydrogen bond acceptors / donors (HBA/ HBD, binding interactions), eccentric connectivity index (ECI, structural complexity), topical polar surface area (TPSA, relative polarity) and the number of rotatable bonds (RB, molecular flexibility and entropy).

Development of Workflows

KNIME (version 3.2) was used to build a structural feature-based workflow to screen for ligands predicted to bind to NR that are associated with the onset of hepatic steatosis (Table 1). The workflows executed rules based on physico-chemical properties and structural features established through studying relevant pChEMBL values and the structural information within PDB. The main KNIME workflow is an amalgamation of eight smaller workflows for each NR making screening chemicals fast and easy for users. Each of these individual workflows is made up of two steps. The first calculates physico-chemical properties of the chemicals being screened to identify if the chemicals are in the chemical space defined previously (descriptor ranges applied). The second step runs the chemicals being screened against the structural features found to be essential for receptor binding that have been developed for each NR. In summary the workflow firstly identifies if a compound is in the chemical space associated with being an agonistic binder, then whether it has the structural features required for binding, which is an informed method of grouping for receptor mediated effects.

The workflow developed allows users to enter one chemical to be screened for NR binding (either via a .csv file containing the SMILES string for the chemical or by drawing the chemical structure using the drawing tool available) or via a batch process (using a .csv file containing the SMILES for all the chemicals being screened). The output of the workflow is a table of all the chemicals that were identified as binders to one or more of the NR listed in Table 1 (note: RAR and RXR are combined and CAR is not present within the workflow, see the CAR section in the Results), the NR that they are predicted to bind to are listed. If a batch is run and no chemicals are identified as possible binders, a message will appear to let the user know that their chemicals are deemed not to be a binder to the NRs listed.

RESULTS

Data and information obtained from ChEMBL and PDB

Data and other information about ligand binding were extracted from ChEMBL and the PDB for the ten NR listed in Table 1. The number of agonists obtained from ChEMBL, those deemed to be active, the range of pChEMBL values found for each NR and the number of human PDB files (which contain crystallographic representations of the NRs binding with agonists) found associated with each NR is summarised in Table 2.

TABLE 2 HERE

Descriptor ranges applied

Descriptors were calculated within KNIME using the CDK node. Descriptors were calculated for all agonists collected from ChEMBL that were identified as active (pChEMBL $\geq$ 5). Eight descriptors were chosen in total and these were selected as they gave information relevant for ligand binding/ ligand shape and so define the chemical space for the properties needed to bind to the NR of interest. A summary of the ranges used for the molecular descriptors and applied for each NR within the workflow is presented Table 3 below.

TABLE 3 HERE

Ligand-protein binding information and building of SA

From studying the ligand binding interactions found within the crystallographic PDB files of human NR to known agonists, key structural features that were shown to be essential were identified. These structural features were classed as essential as they occurred in many of the

PDB files showing agonistic binding for the NR, also the papers associated with each PDB file made reference to these important structural features, therefore this knowledge was built upon and added to using knowledge obtained from ChEMBL (chemical structure obtained and activity values for known agonists of each NR). The structural features that were developed for the workflow are summarised below for each NR studied. N.B the tables showing structural features found for each NR are only shown for the AHR receptor – all others are found within the supporting information.

AHR

The PDB files associated with agonistic binding to the human AHR receptor were studied along with the shape of the agonists obtained from ChEMBL. From these the ligand-binding patterns and those chemical features present in all known AHR agonists were identified. It was found that ligands must form interactions (usually via hydrogen bonds) with the key residues Met328, Tyr353 and Phe367 found within the AHR binding pocket in order to activate the AHR NR.[35] These structural features were then coded into SMARTS strings.[31] The AHR workflow was split into two parts, firstly the chemical must contain at least one of the backbone ring structures as reported in Table 4 (showing the SMARTS strings and visual representations) as these were observed as being essential to fitting into the binding pocket. Secondly the chemical must contain either one of the oxygen functional groups seen in Table 5 or substitutes for oxygen (nitrogen/chlorine groups) reported in Table 6. The oxygen/nitrogen functional groups were observed to be essential to form hydrogen bonds between the ligand and the ligand binding pocket of the AHR.

TABLE 4 HERE

TABLE 5 HERE

TABLE 6 HERE

CAR

When searching for data associated with the CAR NR within the ChEMBL database, only 40 chemical structures could be found. Furthermore, no pChEBML values were assigned to these chemicals. As the quantity and quality of the data available for CAR were limited, this NR was excluded from model development. This should be noted as a subject for further investigation in future to develop structural alerts for this NR and also for the development of AOPs.

ER

The binding of ER agonists was observed (within the PDB files containing crystallographic representation of agonists binding to the human ER). It was found that a ligand must interact with the key residues Arg346, Glu 305 and H13475 within the ER binding domain in order for ER activation.[36] The bonds formed in order for this interaction to occur involved hydrophobic van der Waals interactions within the lipophilic pocket. The structural features of ligands that occurred in the PDB files were coded into SMARTS patterns. The binding of ER agonists was shown to be similar to other steroid hormone NRs with the exception that binding was found to be different for ER agonists with a higher molecular weight. Therefore the ER workflow first splits the chemicals being screened based on MW within the range ($700 \leq MW \leq 2250$), the MW range was selected based on the MW of the known binders within ChEMBL. Those chemicals with a MW within this range were checked against the steroid structure check (Supplementary data - Table S1). Those chemicals with a MW less than 700 pass through the usual descriptor checks and then proceed to the structural feature screening. Similar to the AHR rules, the chemical must contain at least one of the essential scaffold ring SA (Supplementary data - Table S2) and one of the oxygen functional groups (Supplementary data - Table S3) or nitrogen functional group (Supplementary data - Table S4). The oxygen and nitrogen functional

groups were found to form essential hydrogen bonds between the ligands and the ligand binding pocket.

FXR

Structural features implemented for FXR screening are expressed in Supplementary data Table S5 and S6. The residuals of arginine and histidine, sometimes incorporating water molecules, form hydrogen bonds with carboxylic groups (3BEJ). The threonine, asparagine and glutamic acid residues may form further hydrogen bonds, in particular to oxygens (4II6, 3BEJ). Sub-structural patterns in FXR ligands are mainly defined by oxygens, and to a lesser extent, nitrogens, sulphurs and halogens, and the manner in which they are attached to aromatic and aliphatic ring structures. Many ligands do not have significant structural resemblance to the endogenous ligands, such as chenodeoxycholic acid.[37]

GR

The conclusions from the PDB files and literature searches revealed that ligands that bind with high affinity to GR contain a ketone group (or other similar substitute group) which forms hydrogen bonds between the ARG-611 and Gln-570 amino acid residues on the ligand binding pocket of the GR.[38] The hydrogen atom from the 17b-hydroxyl group has a partial positive charge which allows it to interact and form bonds with highly electronegative atoms that are bound to an amino acid residue. These essential features were coded into SMARTS strings. The first step within the GR workflow splits the chemicals depending on MW. Chemicals that had MW within this range ($610 \leq MW \leq 1200$) went through one check to look for the ring structure as described in supplementary data, Table S7. Those chemicals with a MW less than 610 are screened against the descriptor ranges (Table 3) and to identify essential structural features. The chemical must contain a backbone ring structure (Supplementary data - Table S8) and must also contain either an oxygen group (Supplementary data - Table S9) or a nitrogen group (Supplementary data – Table S10). The binding observed for GR actives was similar to

that observed for other steroid based NR. They have specific ring structures with many oxygen / nitrogen functional groups that help to form strong hydrogen bonds between the ligand and ligand binding pocket.

LXR

LXR actives were studied and the sub-structural features were coded into SMARTS strings (Supplementary data - Tables S11 and S12). A potential ligand contains a ring backbone, which may have interactions with phenylalanine, tryptophan and histidine residues, in particular π-π stacking. Furthermore, the compound must also contain functional groups, in particular terminal oxygens, interacting with arginine or threonine residues and the secondary amine of a leucine (PDB: 3LOE, 4NQA, 4DK7), as can be seen in Figure 1 showing hydrogen bonding between the ARG319 and LEU330 residues of the LXR binding pocket to the oxygen groups of the ligand.

FIGURE 1 HERE

PPAR

PPAR actives were studied and the sub-structural features were coded into SMARTS strings. The chemical must not contain a steroid backbone (Supplementary data - Table S13) but must contain one of the specific “diaromatic” scaffold and one of the specific functional groups in order to be an active. Additional alerts describe fatty acid- and retinoid-like compounds, which may have moderate PPAR affinity (Supplementary data - Tables S14, S15 and S16).

PXR

It was found that ligands of the PXR must form interactions (usually hydrogen bonds) with the key residues Ser 208, Ser247, GLn285, His407 and Arg410 within the PXR binding pocket in order for PXR activation to occur.[39] The sub-structural features of PXR actives studied were coded into SMARTS strings. Similar to the other steroidal NR, the chemical must contain at least one of the essential scaffold ring SA (Supplementary data - Table S17) and one of the oxygen functional groups (Supplementary data - Table S18) or nitrogen functional group (Supplementary data - Table S19). The oxygen and nitrogen functional groups were found to form hydrogen bonds between the ligands and the ligand binding pocket.

RAR/RXR

After observing the RAR and RXR receptors separately it was noted that their actives had very similar binding patterns and it was decided to combine them into one workflow. Generally RAR/RXR ligands are lipophilic, but there are a few compounds which are active without being lipophilic (XLogP < 2.2), e.g. n-phosphono-L-phenylalanyl-L alanylglycinamide with an XLogP of -2.4. As these compounds have peptide-like bonds, XLogP exception rules were created (Supplementary data - Table S21). To narrow down the compounds passing through this alert, such as inactive amino sugars, a further filter (Supplementary data - Table S22) was used. As shown in Figure 2, there are certain groups (in particular double bond oxygens), binding to arginine and serine residues, e.g. the hydrogen bond between ARG278 or SER289 and an oxygen of a ligand's carboxylic group within the RAR domain. The responsible structural features are described in the alerts (Supplementary data - Table S23). Furthermore RAR/RXR ligands contain at least one ring structure, which could be aromatic or aliphatic, e.g. cyclohexene of retinoic acid, as expressed in the SA (Supplementary data - Table S24). [25, 40]

FIGURE 2 HERE

Testing the screening workflow

The ChEMBL chemicals deemed to be active via their pChEMBL value were used to test if all of the chemicals that are known agonists for the NR of interest (Table 1) are identified by the screening workflow. The results demonstrated that all of the chemicals that have been identified as binders within ChEMBL were successfully predicted as binders to their associated NR showing that the workflow was accurate at identifying chemical's as being binders .

DISCUSSION

21st Century Toxicology relies heavily on the development of alternative testing methods (computational, biokinetics, *in vitro*) as opposed to the traditional extensive animal methods used previously. Alternative approaches now favour the inclusion of computational models, however, traditional *in silico* models (QSARS/SARs) have struggled in the past to deal with organ level toxicity. Despite this, recently there has been some improvement through the use of SA, especially focussed on MIEs.[21-23] In general, SA are well developed for MIEs depending on the reactivity of a xenobiotic with a biological macromolecules, for instance the formation of covalent bond as demonstrated by the many profilers (e.g. for protein or DNA binding) implemented in the OECD QSAR Toolbox. It remains much more difficult to develop profilers for receptor mediated toxicity, with the current state of the art being MIE-derived 2D descriptors.[21-23,41,42]Whilst these issues are recognised, encouraging recent studies have shown that it is possible that useful information and models, including profilers, can be developed for receptor mediated toxicity.[43,44]

Whilst it is becoming common place to code 2D interactions e.g. protein / DNA binding as molecular fragments, the next challenge lies with the grouping of receptor mediated effects. Ultimately the modelling of most receptor-ligand interactions must address the use of molecular modelling and other types of molecular design software, and a framework for undertaking this task has been presented recently.[43] Despite the simplicity of a 2D approach,

progress can be made rapidly,[23] and such profilers are amenable to use in e.g. the OECD QSAR Toolbox.[45] In this study the issue of the capture of information relating to MIE has been addressed, in part, by the use of structural alerts based workflows. SA can be used both as a direct predictor of toxicity and also for grouping chemicals for read across. Through the development of AOPs, SA can be used collectively if they have the same MIE, our understanding of this MIE can then provide a linkage to mechanistic pathways and the adverse effects induced via these pathways. AOPs are now integral to risk assessment, therefore, AOP development is important and the role SA play in their implementation is essential.

There are many (Q)SAR models available for the prediction of NR mediated effects.[46] Therefore, the purpose of this study was not to repeat previously undertaken work but rather to build on existing knowledge to create a new set of SA/ structural features that can be implemented in an *in silico* workflow. This investigation has focused on NR previously linked to the onset of hepatic steatosis.[18] Within this study the use of new generation resources (PBD, pChEMBL, KNIME, PyMOL, Marvin Beans) has been a key element. This demonstrated how existing data may be used in future studies to create knowledge regarding toxicological interactions. A total of 12,713 chemicals were identified in ChEMBL that were linked to NR and could be used in this study (with a pChEMBL $\geq$5). In addition 624 human PDB files showed binding information of ligands to the NR of interest for this study. These figures show the vast amount of current data that are available and, when linked to AOPs, have the potential to provide a goldmine of information.

Generally all workflows in this study can be divided into two essential steps. The first step involves the screening for ligand-specific physico-chemical descriptors and the second step involves the use of sub-structural features. The sub-structural features produced for most of the NR workflows follow a similar pattern of scrutinising for key scaffolding structures (e.g. ring structures) and then further screening for essential functional groups. However, there are a few

exceptions such as for RXR and PPAR (which have some exclusion rules) and for GR, AR, ER and PXR (which have high MW filter to account for those ligands that were larger and had different receptor binding patterns compared the low MW ligands).

Within the literature it remains unclear what role the different nuclear receptor subtypes play in terms of activation of the pathways associated with each NR.[47] As the binding of ligands to the different receptor subtypes was observed to overlap, it was decided to combine the subtypes into one screening workflow. It would have been challenging to develop structural features for one specific receptor subtype as many ligands are able to bind to different subtypes albeit sometimes with different affinities although this would remain a long term goal. For example, the ER agonist raloxifene has PChEMBL value of 10.52 for the ERα and a PChEMBL value of 8.8 for the ERβ therefore, it can be seen that these molecules share similar MIE's across two different NR. Also it cannot be determined if one ligand only binds to one receptor subtype due to the constraints of the data available in the ChEMBL database. Therefore a NR workflow, such as for ER, is a combined workflow incorporating all receptor subtypes, such as ERα and ERβ. It was noted that the ligands of some of the NR were similar, particularly those that are specific for retinoids (e.g. RAR and RXR ligands) and steroids (e.g. AR, ER and GR ligands). This means that ligands may have the ability to activate many NRs (to a certain extent). As predictions for promiscuous receptors can be difficult, the full set of predictions is given within the output file of the screening workflow.

This generalistic approach has the advantage of rapidly identifying chemicals that have a similar structure to previously known NR binders and has been tailored specifically towards human NR binding. However, because this method focuses only on qualitative identification of alerts associated with the initial event in an AOP, it cannot be used as a standalone for hazard identification (identifying the potential for harm) or risk assessment (the likelihood of harm associated with specific patterns and levels of exposure). The ability to predict an AO will

depend upon the scientific confidence in the predictive models that link quantitation of upstream key events (e.g., the MIE) to downstream key events in the AOP and the ability to predict risk involves integrating knowledge of toxicokinetics, biological activity, dose response with predicted or modelled exposures.[12, 48] Nevertheless, the workflow can be very useful in reducing, refining or replacing traditional animal toxicity testing. It can be used to develop categories of substances for use in read across to enable inference from measured human health and/or environmental properties/endpoints from reference substance(s) within the group to fill data gaps for substances that lack data for such properties / endpoints. In addition, the workflow can also be used for prioritisation to differentiate chemicals that may require further testing as part of an integrated testing strategy from those that do not show structural alerts for specific NR pathways.[49] For example, the workflow could potentially be applied to initiatives such as the USEPA's Endocrine Disruptor Screening Program as the initial step in setting priorities for further in vitro or in vivo screening for oestrogen, androgen or thyroid activities; substances that do not trigger SAs would be deprioritised.[50]

CONCLUSIONS

214 structural features were developed from MIEs associated with AOPs and combined with eight different descriptors to create a decision based workflow for each NR. The individual NR workflows have been amalgamated into one large screening workflow for all NRs investigated and with the focus being the NRs associated with the onset of hepatic steatosis. This study highlights that modern technologies (PDB, CheMBL, KNIME) provide new opportunities, due to their extensive data, to build alerts and use the information potentially contained with AOPs. This study is the first to produce a SA based workflow of this size for a receptor mediated toxicity, in this case linked to hepatic steatosis as the target organ adverse effect through the AOP. The workflow produced has addressed the problem of grouping chemicals that have hepatic steatosis as their endpoint, a previously difficult task.

FUNDING INFORMATION

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n° 266835 (COSMOS Project) and from Cosmetics Europe. More information is available at www.cosmostox.eu.

ASSOCIATED CONTENT

Supporting Information contains the structural features and alerts developed for each NR. The Supporting Information is available free of charge on the ACS Publications website at….

ABBREVIATIONS

Adverse Outcome Pathway, AOP; Molecular Initiating Event, MIE; Aryl Hydrocarbon Receptor, AHR; Constitutive Androstane Receptor, CAR; oEstrogen Receptor, ER; Farnesoid X Receptor , FXR; Glucocorticoid Receptor , GR; Liver X Receptor , LXR; Peroxisome Proliferator-Activated Receptor, PPAR; Pregnane X Receptor, PXR; Retinoic Acid Receptor , RAR; Retinoid X receptor, RXR; Nuclear Receptor, NR; Structural Alerts, SA; KoNstanz Inforamtion MinEr, KNIME; Protein Data Bank, PDB; octanol-water partition coefficient, xlogP; Vertex AdIacency Matrix, VAIM; Hydrogen Bond Acceptors / Donors, HBA/ HBD; Eccentric Connectivity Index, ECI; Topical polar surface area, TPSA; Rotatable Bonds; RB; Comma-Separated Values, CSV.

ACKNOWLEDGEMENTS

The authors of this paper would like to thank Iva Lukac for her help and contribution to the production of NR images.

REFERENCES

1. Wang, O., and LeCluyse, E. L. (2003) Role of orphan nuclear receptors in the regulation of drug metabolising enzymes. *Clin. Pharmacokinet. 42*, 1331-1357.
2. Francis, G. A., Fayard, E., Picard, F., and Auwerx, J. (2003) Nuclear receptors and the control of metabolism. *Annu. Rev. Physiol. 65*, 261–311.
3. Maglich, J. M., Sluder, A., Guan, X., Shi, Y., McKee, D. D., Carrick, K., Kamder, K., Willson, T. M., and Moore, J. T. (2001) Comparison of complete nuclear receptor sets from the human, *Caenorhabditis elegans* and *Drosophila genomes*. *Genome Biol. 2*, research0029.1- research0029.7.
4. Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schutz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P., and Evans, R. M. (1995) The nuclear receptor superfamily: the second decade. *Cell 83*, 835–39.
5. Barkhem T. L., Carlsson, B., Nilsson Y., Enmark, E., Gustafsson, J., and Nilsson, S. (1998) Differential response of estrogen receptor alpha and estrogen receptor beta to partial estrogen agonists/antagonists. *Mol. Pharmacol. 54*, 105-112.
6. Brown, K. (2002) Breast cancer chemoprevention: risk-benefit effects of the antioestrogen tamoxifen. *Expert Opin. Drug Saf. 1*, 253–267.
7. Jordan, V.C. (2003) Tamoxifen: a most unlikely pioneering medicine. *Nat. Rev. Drug. Discov. 2*, 205–213.
8. Damstra, T., Barlow, S., Bergman, A., Kavlock, R., and van Der, K. G. (2002) *International Programme on Chemical Safety Global Assessment: The State-of-the-Science of Endocrine Disruptors*. Geneva:World Health Organization.

9. Glass, C. K., and Rosenfeld, M. G. (2000) Coregulator exchange in transcriptional functions of nuclear receptors. *Genes Dev. 14*, 121–141.
10. National Research Council (NRC) (2007) *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: National Academies Press.
11. Sonoda, J., Pei, L., and Evans, R. M. (2008) Nuclear receptors: decoding metabolic disease. *FEBS Lett. 582*, 2–9.
12. Patlewicz, G., Simon, T., Rowlands, J. C., Budinsky, R. A., and Becker R. A. (2015) Proposing a scientific confidence framework to help support the application of adverse outcome pathways for regulatory purposes. *Regul. Toxicol. Pharmacol. 71*, 463-477.
13. Vinken, M. (2013) The adverse outcome pathway concept: A pragmatic tool in toxicology. *Toxicology 312*, 158–165.
14. Vinken, M. (2015) Adverse outcome pathways and drug-induced liver injury. *Chem. Res. Toxicol. 28*, 1391−1397.
15. Hartung, T. (2009) Toxicology for the twenty-first century. Nature, 460, 208-201.
16. Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L., Schmieder, P. K., Serrano, P. K., Tietge, J. E., and Villeneuve, D. L. (2010) Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem. 29*, 730–741.
17. Groh, K. J., Carvalho, R. N., Chipman, J. K., Denslow, N. D., Halder, M., Murphy, C. A., Roelofs, D., Rolaki, A., Schirmer, K., and Watanabe K.H. (2015) Development and application of the adverse outcome pathway framework for understanding and predicting chronic toxicity: I. Challenges and research needs in ecotoxicology. *Chemosphere 120*, 764–777.

18. Mellor, C. L., Steinmetz, F. P., and Cronin M. T. D. (2015) The identification of nuclear receptors associated with hepatic steatosis to develop and extend adverse outcome pathways. *Crit. Rev. Toxicol.* 1-15 , http://dx.doi.org/10.3109/10408444.2015.1089471.
19. AOP WIKI (2015) AOP wiki NR linked to hepatic steatosis, https://aopkb.org/aopwiki/index.php/Aop:34. Accessed on 5th May 2015.
20. Worth, A. P. (2004) The tiered approach to toxicity assessment based on the integrated use of alternative (non-animal) tests. In *Predicting Chemical Toxicity and Fate* (Cronin, M. T. D., and Livingstone, D. J.) pp391-412, CRC Press, Boca Raton FL.
21. Nelms, M. D., Mellor, C. L., Cronin, M. T. D., Madden, J. C., and Enoch S. J. (2015) The development of an in silico profiler for mitochondrial toxicity. *Chem. Res. Toxicol. 28*, 1891−1902.
22. Gutsell, S., and Russell, P. (2013) The role of chemistry in developing understanding of adverse outcome pathways and their application in risk assessment. *Toxicol. Res. 2*, 299−307.
23. Allen, T. E. H., Goodman, J. M., Gutsell, S., and Russell, P. (2014) Defining molecular initiating events in the Adverse Outcome Pathway framework for risk assessment. *Chem. Res. Toxicol. 27*, 2100-2112.
24. Stepan, A. F., Walker, D. P., Bauman, J., Price, D. A., Baillie, T. A., Kalgutkar, A. S., and Aleo, M. D. (2011) Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol. 24*, 1345–1410.
25. Steinmetz, F. P., Mellor, C. L., Meinl, T., Cronin, M. T. D. (2015) Screening chemicals for receptor-mediated toxicological and pharmacological endpoints: Using public data to build screening tools within a KNIME workflow. *Mol. Inform. 34*,171-178.

26. ChEMBL (2015) https://www.ebi.ac.uk/chembl/. Accessed 5th May 2015.
27. KNIME (2015) www.knime.org, accessed 5th May 2015.
28. PDB (2015) www.rcsb.org, accessed 5th May 2015.
29. PyMOL (2015) http://www.pymol.org, accessed 5th May 2015.
30. ChemAxon (2015) https://www.chemaxon.com/download/marvin-suite/#mbeans. Accessed 5th May 2015.
31. Daylight (2015) www.daylight.com, accessed 5th May.
32. Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J. P. (2013) The ChEMBL bioactivity database: an update. *Nucleic Acids Res. 42*, 1-8.
33. Chichester, C., Digles, D., Siebes, R., Loizou, A., Groth, P., and Harland, L. (2015) Drug discovery FAQs: workflows for answering multidomain drug discover questions. Drug Discovery Today, 20, 399-405.
34. Kruger, F.A., Gaulton, A., Nowotka, M., and Overington, J.P. (2014) PPDMs-a resource for mapping small molecule bioactivities from ChEMBL to Pfam-A protein domains. Bioinformatics, 31(5), 776-778.
35. Shiizaki, K., Ohsako, S., Kawanishi., M., and Yagi, T. (2014) Identification of amino acid residues in the ligand-binding domain of the aryl hydrocarbon receptor causing the species-specific response to omeprazole: possible determinants for binding putative endogenous ligands. *Mol. Pharmacol. 85*, 279-89.
36. Roberts, L. R., Armor, D., Barker, C., Bent, A., Bess, K., Brown, A., Favor, D. A., Ellis, D., Irving, S. L., MacKenny, M., Phillips, C., Pullen, N., Stennett, A., Strand, L., and Styles, M. (2011) Sulfonamides as selective oestrogen receptor β agonists. *Bioorg. Med. Chem. Lett. 21*, 5680–5683.

37. Akwabi-Ameyaw, A., Caravella, J. A., Chen, L., Creech, K. L., Deaton, D. N., Madauss, K. P., Marr, H. B., Miller, A. B., Navas, F., Parks, D. J., Spearing, P. K., Todd, D., Williams, S. P., and Wisely, G. B. (2011) Conformationally constrained farnesoid X receptor (FXR) agonists: alternative replacements of the stilbene. *Bioorg. Med. Chem. Lett. 21*, 6154-6160.
38. Biggadike, K., Bledsoe, R. K., Coe, D. M., Cooper, T. W., House, D., Iannone, M. A., Macdonald, S. J., Madauss, K. P., McLay, I. M., Shipley, T. J., Taylor, S. J., Tran, T. B., Uings, I. J., Weller, V., Williams, S. P. (2009). Design and x-ray crystal structures of high-potency nonsteroidal glucocorticoid agonists exploiting a novel binding site on the receptor. *Proc. Natl. Acad. Sci. USA 106*, 18114-18119.
39. Watkins, R. E., Wisely, G. B., Moore, L. B., Collins, J. L., Lambert, M. H., Williams, S. P., Willson, T. M., Kliewer, S. A., and Redinbo, M. R. (2001) The human nuclear xenobiotic receptor PXR: Structural determinants of directed promiscuity. *Science 292*, 2329-2333.
40. Klaholz, B. P., Mitschler, A., and Moras, D (2000) Structural basis for isotype selecitivity of the human retinoic acid nuclear receptor. J. Mol. Biol. 302, 155−170.
41. Enoch, S. J., and Cronin, M. T. D. (2010) A review of the electrophilic reaction chemistry involved in covalent DNA binding. *Crit. Rev. Toxicol. 40*, 728-748.
42. Enoch, S. J., Ellison, C. M., Schultz, T. W., and Cronin, M. T. D. (2011) A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. *Crit. Rev. Toxicol. 41*, 783-802.
43. Tsakovska, I., Al Sharif, M., Alov, P., Diukendjieva, A., Fioravanzo, E., Cronin, M. T. D., and Pajeva, I. (2014) 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. 15*, 7651-7666.

44. Fratev, F., Tsakovska, I., Al Sharif, M., Mihaylova, E., and Pajeva, I. (2015) Structural and dynamical insight into PPARγ antagonism: in silico study of the ligand-receptor interactions of non-covalent antagonists. I*nt. J. Mol. Sci. 16*, 15405-15424.

45. Diderich, R., (2010) Tools for category formation and read-across: Overview of the OECD (Q)SAR Application Toolbox, in Cronin, M. T. D., and Madden, J. C. (eds) *In Silico Toxicology. Principles and Applications.* RSC Publishing, Cambridge, pp. 385-407.

46. A PubMed (2015) search for QSAR models nuclear receptor binding gave 227 results. http://www.ncbi.nlm.nih.gov/pubmed/?term=QSAR+models+nuclear+receptor+binding. Accessed September 9th 2015.

47. DeLisle, R. K., Yu, S., Nair, A. C., and Welsh, W. J. (2001) Homology modeling of the estrogen receptor subtype β (ER-β) and calculation of ligand binding affinities. *J. Mol. Graph. Mod. 20*, 155–167.

48. Villeneuve, D.L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T.H., LaLone, C.A., Landesmann, B., Lettieri, T., Munn, S., Nepelska, M., Ottinger, M.A., Vergauwen, L., and Maurice Whelan, M. (2014) Adverse Outcome Pathway Development II: Best Practices. Toxicol. Sci. (2014) 142 (2): 321-330.

49. Hartung, T., Luechtefeld, T., Maertens, A., Kleensang, A. (2013) Integrated testing strategies for safety assessments. ALTEX. 2013;30(1):3-18.

50. EPA, Endocrine Disruption Screening Program. http://www.epa.gov/endocrine-disruption/use-high-throughput-assays-and-computational-tools-endocrine-disruptor. Accessed December 21st 2015.

Table 1: Nuclear receptors associated with hepatic steatosis and abbreviations as defined by Mellor et al 2015.[17]

Nuclear receptor name	Abbreviation	Nomenclature Identification
Aryl Hydrocarbon Receptor	AHR	bHLHe76
Constitutive Androstane Receptor	CAR	NR1I3
oEstrogen Receptor	ER	NR3A1/2
Farnesoid X Receptor	FXR	NR1H4/5
Glucocorticoid Receptor	GR	NR3C1
Liver X Receptor	LXR	NR1H2/3
Peroxisome Proliferator-Activated Receptor	PPAR	NR1C1-3
Pregnane X Receptor	PXR	NR1I2
Retinoic Acid Receptor	RAR	NR1B1-3
Retinoid X receptor	RXR	-

Table 2: Summary of the data and information obtained from ChEMBL and the PDB for the different NR agonists

Nuclear receptor	Number of agonists obtained from ChEMBL		Range of pChembl values for chemicals	Number of PDB files found that contain human NR structures
	Total	with pChEMBL ≥ 5		
AHR	219	170	4.0 - 9.35	20
CAR	-	-	-	-
ER	7528 (4586α) (2942β)	1489 (791α) (698β)	4.14 - 11.00	249
FXR	799	602	4.21 - 8.7	23
GR	2029	2021	4 - 10	62
LXR	1536 (749α) (787β)	812 (368α) (444β)	4.09 – 9.00	16
PPAR	13358 (4034α) (3040β)(6284γ)	5700 (1999α) (1196β)(2505γ)	4.00 – 10.74	166
PXR	463	135	4.00 – 9.15	68
RAR	2511 (848α) (878β)(785γ)	855 (258α) (325β)(272γ)	4.55 – 10.4	20
RXR	2380 (1845α) (263β)(272γ)	950 (563α) (189β)(198γ)	4.68 – 10.1	109

Note: α, β and γ values given in parentheses are the number of chemicals found that are associated with binding to either the α, β or γ subunit of the NR

Table 3: The descriptor ranges used for all NR and implemented within the workflow

Physico-chemical property	Value								
	AHR	**CAR**	**ER**	**FXR**	**GR**	**LXR**	**PPAR**	**PXR**	**RAR/RXR**
VAIM	4.5-6.5	-	4 - 7.5	-	4 - 8.5	4.7 - 7	5 - 7	5 - 7	5 - 7
HBD	$\leq$ 6	-	$\leq$ 10	-	$\leq$ 15	-	-	$\leq$ 5	-
MW	180 - 900	-	140 - 700	$\geq$ 900	180 - 610	$\leq$ 750	< 800	300 - 610	$\leq$ 550
HBA	$\leq$ 10	-	$\leq$ 15	-	$\leq$ 15	-	-	$\leq$ 10	-
XLogP	$\leq$ 8	-	$\leq$ -2	-	$\leq$ -1	$\leq$ 2	-	$\leq$ 0	-
ECI	-	-	-	150 - 2400	-	-	-	-	-
RB	-	-	-	3- 11	-	-	-	-	3 - 30
TPSA	-	-	-	15 - 200	-	5 - 150	1.2 - 20	-	-

Table 4: SMARTS strings and chemical structure of backbone ring for AHR actives.

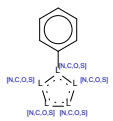
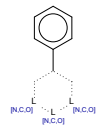
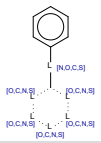
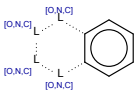
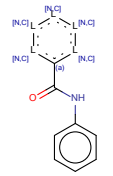
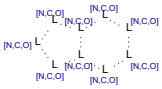
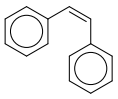
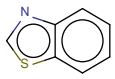
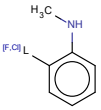
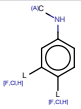
SMARTS string	Structual Feature
[#7,#6,#8,#16]1[#7,#6,#8,#16][#7,#6,#8,#16][#7,#6,#8,#16]([#7,#6,#8,#16]1)-c1ccccc1	
[#6]~1~[#7,#6,#8]~[#7,#6,#8]~[#7,#6,#8]~[#6]~[#6]~1-c1ccccc1	
[#8,#6,#7,#16]~1~[#8,#6,#7,#16]~[#8,#6,#7,#16]~[#6](~[#8,#6,#7,#16]~[#8,#6,#7,#16]~1)-[#7,#8,#6,#16]-c1ccccc1	
[#8,#7,#6]~1~[#8,#7,#6]~[#8,#7,#6]~c2ccccc2~[#8,#7,#6]~1	
O=[#6](-[#7]-c1ccccc1)-c1[#7,#6][#7,#6][#7,#6][#7,#6][#7,#6]1	
[#7,#6,#8]~1~[#7,#6,#8]~[#7,#6,#8]~2~[#7,#6,#8]~[#7,#6,#8]~[#7,#6,#8]~[#7,#6,#8]~[#7,#6,#8]~2~[#7,#6,#8]~1	
C(=C\c1ccccc1)\c1ccccc1	
c1nc2ccccc2s1	
[#6]-[#7]-c1ccccc1-[#9,#17]	
[#6;A][#7]-c1ccc(-[#9,#17,#1])c(-[#9,#17,#1])c1	

Table 5: SMARTS strings and chemical structure of oxygen group for AHR actives.

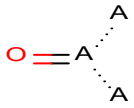
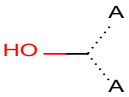
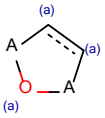
SMARTS string	Structual Feature
~(~*)=O	
~[#6](~)-[#8]	
c1c*o*1	

Table 6: SMARTS strings and chemical structure of oxygen substitute (nitrogen/ chlorine) for AHR actives.

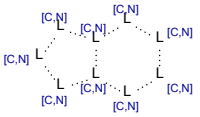
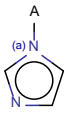
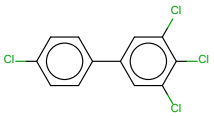
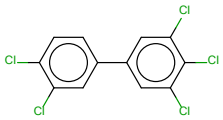
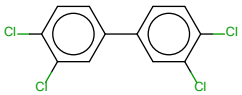
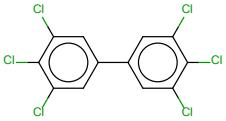
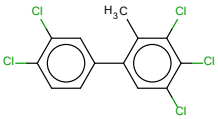
SMARTS string	Structual Feature
[#6,#7]~1~[#6,#7]~[#6,#7]~2~[#6,#7]~[#6,#7]~[#6,#7]~[#6,#7]~[#6,#7]~[#6,#7]~2~[#6,#7]~1	
[#7;a]~1~*~*~*~*~*~1	
*n1ccnc1	
Clc1ccc(cc1)-c1cc(Cl)c(Cl)c(Cl)c1	
Clc1ccc(cc1Cl)-c1cc(Cl)c(Cl)c(Cl)c1	
Clc1ccc(cc1Cl)-c1ccc(Cl)c(Cl)c1	
Clc1cc(cc(Cl)c1Cl)-c1cc(Cl)c(Cl)c(Cl)c1	
Cc1c(Cl)c(Cl)c(Cl)cc1-c1ccc(Cl)c(Cl)c1	

FIGURE TITLES

Figure 1: Ligand-protein interaction of 4NQA (PDB, 2014), showing potential hydrogen bond formation of oxygen groups on the ligand to key residues ARG278 and SER289 within the LXR binding domain

Figure 2: Ligand-protein interaction of 2LBD (PDB, 2014), showing potential hydrogen bond formation between the ligand and the key residues LEU330 and ARG319 within the RAR binding pocket.