RANKING STRATEGIES TO SUPPORT TOXICITY PREDICTION: A CASE STUDY ON POTENTIAL LXR BINDERS

Anna Palczewska[ac]*, Simona Kovarich[b], Andrea Ciacci[b], Elena Fioravanzo[b], Arianna Bassan[b],Daniel Neagu[a]

[a] Department of Computer Science, University of Bradford, UK

[b] S-IN Soluzioni Informatiche S.r.l., via G. Ferrari 14, 36100 Vicenza, Italy

[c] School of Computing, Creative Technologies & Engineering, Leeds Beckett University, UK

Abstract

The current paradigm of toxicity testing is set within a framework of Mode-of-Action (MoA)/Adverse Outcome Pathway (AOP) investigations, where novel methodologies alternative to animal testing play a crucial role, and allow to consider causal links between molecular initiating events (MIEs), further key events and an adverse outcome. *In silico* (computational) models are developed to support toxicity assessment within the MoA/AOP framework. This paper focuses on the evaluation of potential binding to the Liver X Receptor (LXR), as this has been identified among the MIEs leading to liver steatosis within an AOP framework addressing repeated dose and target-organ toxicity.

The objective of this study was the development of a priority setting strategy, by means of *in silico* approaches and chemometric tools, to allow for the screening and ranking of chemicals according to their toxicity potential. As a case study, the present paper outlines the methodologies and procedures that have been developed in the context of the COSMOS/ety assessment project [4], which developed computational methods in view of supporting cosmetics safety assessment, to rank chemicals based on their potential binding to LXR. Chemicals are ranked based on molecular and QSAR modelling outcomes. The contribution in this paper is threefold: the QSAR model for LXR dataset, an application of molecular modeling approaches, which have been developed and optimized for drug discovery, in the context of toxicology, and finally ranking chemicals based on diverse modelling outcomes. The novelty in this paper consists of the employment of linear (logistic

*Corresponding author.
Email address:a.palczewska@leedsbeckett.ac.uk

regression) and non-linear (Random Forest) models in the context of ranking chemicals. The results show that these methods can be successfully applied for prioritization of compounds of major concern for potential liver toxicity, and that they perform better than the ranking methods reported in the literature to date (such as total ordering or data fusion).

Keywords

ranking; QSAR; in silico; LXR; toxicity prediction;

1. Introduction

Toxicity testing and safety assessment strategies in the 21st century moved towards a novel paradigm in toxicology: no longer based on a complex array of *in vivo* studies evaluating apical adverse outcomes, it is focused on the development of Mode-of-Action (MoA)/Adverse Outcome Pathway (AOP) investigation, which provides information on the causal links between a molecular initiating event (MIE), intermediate key events (KEs) and an adverse outcome (AO) of regulatory concern. In parallel, recent scientific advances in biology and biotechnology (e.g., omics technologies, bioinformatics and computational toxicology) set the basis for a new toxicity-testing system, based on the use of new approach methodologies (NAM), such as High-throughput screening (HTS) () assays, toxicokinetic and toxicodinamic (TK-TD) studies, (Q)SAR and read-across [1].

In the MoA/AOP framework addressing repeated dose and target-organ toxicity, liver steatosis has been recognized as one of the first manifestations of liver toxicity. Liver steatosis is characterized by excessive accumulation of triglycerides in lipid droplets in the hepatocytes and results from the disturbance in the homeostasis of hepatic lipids. The development of steatosis can be attributed to many different causes. Among others, the interaction of exogenous chemicals with nuclear receptors (NRs) involved in the homeostasis of fatty acids metabolism is one of the molecular initiating events of increasing concern [2]. A variety of nuclear receptors could play a role in liver

steatosis, including LXR (liver X receptor), PXR (Pregnane X Receptor), AhR (Aryl hydrocarbon receptor), ER (estrogen receptor) and PPARα and PPARγ (peroxisome proliferator-activated receptor isoforms α and γ respectively). The role of their activation in the MoA/AOP leading to liver steatosis has been described in detail elsewhere [2, 3].

Within the COSMOS project [4], one of seven projects forming the SEURAT-1 European research initiative [5] with emphasis on cosmetics safety assessment, which ran from January 2011 to December 2015, alternative *in silico* models have been developed to support toxicity prediction in the MoA/AOP framework. More specifically, different *in silico* methodologies, including (Q)SAR and molecular modelling, have been employed for the evaluation of potential binding to NRs involved in the development of liver steatosis. One of the objectives of the COSMOS project was the integration of different developed models, based on multiple approaches, to define a priority setting procedure for the screening and ranking of chemicals according to their toxicity potential. Screening chemicals sharing a common molecular initiating event could also support read-across and grouping strategies.

The present paper outlines the methodologies and procedures that have been developed in the context of the COSMOS project to rank chemicals based on their potential binding to Liver X Receptor (LXR). In more detail, different methodologies, including ranking methods, consensus modeling and data fusion techniques, have been employed and compared to combine *in silico* responses generated by different models predicting LXR binding potential.

Firstly, different molecular modelling (MM) approaches, which are usually applied in drug discovery, were explored and combined in order to characterize the ligand binding domain of the receptor and to define the essential features leading to LXR binding: ensemble docking, e-Pharmacophore and Fingerprint-based similarity. These selected models were applied to our LXR dataset and their outcomes were collected for further ranking procedures. Secondly, a new QSAR model for LXR binding prediction was developed. This Partial Least Square – Discriminant

Analysis classification model was developed and validated using a commercial software based on different combinations of MOSES [6] molecular descriptors. Finally, we employed a screening workflow for liver steatosis alerts to predict potential binders to nuclear receptors. The outcomes from all these procedures were used to rank chemicals on their LXR binding potential.

In this paper we consider two different ranking scenarios. In the first use case we rank chemicals based on their outputs from molecular modelling approaches (docking, e-pharmacophore and two similarity measure) that are continuous values representing the power of the binding affinity. In the second use case we rank chemicals based on outputs obtained from three *in silico* methods including: ensemble docking, QSAR model and liver steatosis alerts which are categorical values.

We compare three ranking methods: ordering, consensus modelling and data fusion. As ordering and data fusion are known methods for screening chemicals, combinatorial modelling was used for the first time in the context of ranking. The novelty here is the use of statistical and machine learning models to rank chemicals based on their predicted outcomes by other modelling methods. Two models logistic regression and random forest were proposed and compared to rank chemicals for potential toxicity. Both models return the probability of binding to the LXR receptor and chemicals are ranked based on these probabilities.

To evaluate and compare ranking methods the AUC and Enchrichment Factor (EF) measures are used. The comparison of ranking performance by EF using in the top 1 – 20% range measures the number of correctly identifying positives. This is driven by the use of molecular modelling which are optimized to identify the most active compounds. The results show that consensus methods can be successfully applied for prioritization of compounds of major concern for liver toxicity and they perform better than other ranking methods (such as total ordering or data fusion) reported in the literature.

Additionally, a workflow for ranking was implemented in the KNIME Analytic Platform [7], and shared publicly through the COSMOS KNIME Webportal [8], to increase model transparency and applicability.

This paper is organized as follows: Section 2 presents the dataset and all modeling methods we used in our study to rank binders to LXR. These involve description of the dataset, modeling approaches and validation metrics. Section 3 outlines the results from the comparison of the ranking methods. We consider two use cases for ranking methods combining different models based on their predictive outcome type. Section 4 concludes our study and points to some further work.

2. Materials and Methods

This section describes the dataset used in our study and all methodologies employed to estimate binding potential and ranking chemicals.

2.1. The Dataset

LXR agonists are widely studied as potential therapy agents for a variety of diseases, such as atherosclerosis, diabetes and pulmonary inflammation. Experimental data on IC_{50} LXRβ binding affinity were collected from the literature (the dataset, including references, is reported in Appendix A), leading to the creation of a dataset of 356 compounds, mainly drugs or drug candidates, which consisted of groups of congeneric series sharing a common scaffold. Examples of structures and scaffolds included in the dataset are provided in Figure 1.

The collected “LXR binders” cover a wide range of binding affinity, with IC_{50} values spanning from 1 nM to greater than 10000 nM. Based on the analysis of the distribution of binding affinity data and in line with the classification criteria applied by Zhao and co-workers in [9], arbitrary IC_{50} thresholds for LXR binding potential were defined, and LXR binders were assigned to four classes: "low active" (IC_{50} >1000 nM), "moderate active " ($100<IC_{50}\leq1000$ nM), "active" ($20<IC_{50}\leq100$ nM), and "very active" ($IC_{50}\leq20$ nM). There were 136 chemicals in the very active class, 83

chemicals in “active” class, 80 in the moderate class and 57 in the low active class. Figure 2 presents the distribution of LXR binding affinity data. No particular trend or specific features were observed among the congeneric series and activity classification.

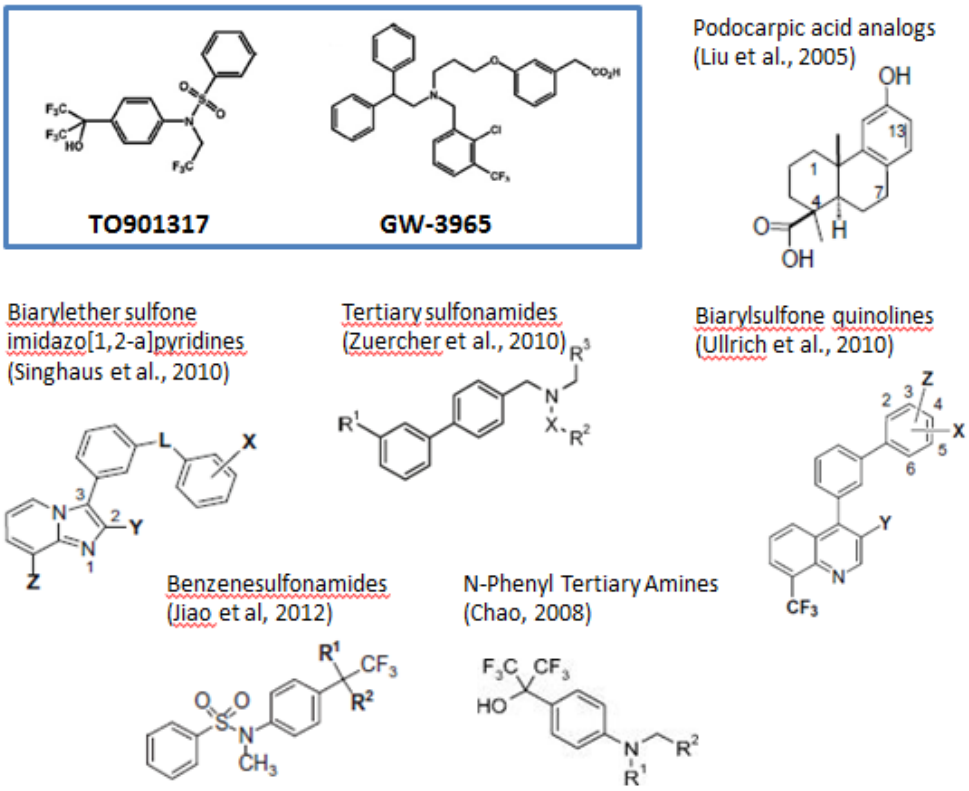


Figure 1 Examples of structures and scaffolds included in the LXR binders dataset (Full references reported in Appendix A). Compounds in the blue box (i.e., TO901317 and GW-3965) are synthetic LXR agonists commonly used as reference chemicals and positive controls in the biological assays.

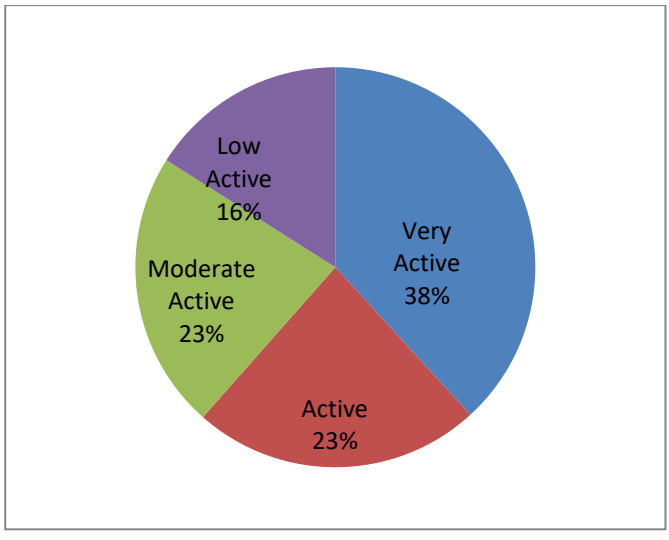


Figure 2. Distribution of chemicals within each class.

The dataset of LXR binders was enriched with decoy molecules, i.e. molecules that are presumed to be inactive against a target (they will not likely bind to the target). Decoys are commonly used to validate the performance of molecular modelling studies, for example molecular docking, which was used in this study. One-thousand decoy molecules were selected from Schrödinger 1K Drug-Like Ligand Decoys Set [10]. This collection of decoys was created by selecting 1000 ligands from a one million compound library that were chosen to exhibit "drug-like" properties [10, 11]. Within this case study, the resulting dataset of 1000 decoys and 356 binders (1356 molecules in total) was used for the following purposes:

i) to assess the ability (evaluated in terms of EF – Enrichment Factor) of the developed *in silico* models (i.e., molecular modelling approaches, QSAR models and structural alerts) to identify LXR binders.

ii) to assess the performance of the different ranking approaches (i.e., total order ranking methods, consensus models and data fusion).

The full dataset is provided as Supplementary Research Data.

2.2. In silico modelling approaches for LXR binding

2.2.1. Molecular modelling studies for LXR binding

Molecular modelling studies were performed to analyse and predict the LXR binding potential. Available 3D crystal structures of LXRβ complexed with structurally distinct ligands were collected from the PDB database [11, 12] and analysed in order to characterize the ligand binding domain (LBD) of the receptors. Different molecular modelling (MM) approaches were explored and combined in order to characterize the ligand binding domain of the receptor and to define the essential features leading to LXR binding:

a) Ensemble docking

Protein–ligand docking is a powerful tool to study and provide a proper understanding of protein–ligand interactions. Docking is typically used in different stages of drug design, to facilitate for example the design of potentially active leads. Molecular docking, in practice, has two essential requirements: molecular structures of ligands of interest and the X-ray structure of the protein target under consideration. To account for protein structural variations we used an algorithm referred to as ensemble docking. The algorithm can simultaneously dock a ligand into an ensemble of protein structures and automatically select a ligand-protein pair that returned the best score. The PDB code of LXR proteins used in the ensemble docking are: 1P8D, 1PQ6, 1PQ9, 3L0E, 4DK7. Default settings for flexible docking procedure with the Glide Standard Precision [10,13-14] protocol were selected**.**

b) e-Pharmacophore

This is a hybrid method combining ligand- and structure-based methodologies with the aim of locating key pharmacophoric features from a docked ligand [15, 16]. The best docking pose of the T0901317 obtained with Glide Extra Precision (XP) [17] protocol was chosen as reference ligand. The e-Pharmacophore script, a module of Phase software [18-19], was used on the XP docking results of reference compound (see Figure 1) for the pharmacophore generation. Default settings and the standard set of six pharmacophore features were used: hydrogen bond acceptor, hydrogen bond donor, hydrophobic region, positive ionisable and negative ionisable region, aromatic ring. The resulting e-pharmacophore, composed of five sites: three hydrophobic regions and two aromatic rings, was used as query for virtual screening of the LXR binders dataset. Screened molecules should match at least three sites. Dataset hits were ranked based on the default Fitness Score [19]. Fitness score is a linear combination of three terms: the alignment score (RMS deviation between the site point positions in the matching conformation and the site point positions in the pharmacophore with a penalty for a partial matching), the vector score (average cosine between vector features in the matching conformation and the vector features in the reference conformation)

and the volume score (ratio of the common volume occupied by the matching conformer and the reference conformer, to the total volume, the volume occupied by both).

c) Fingerprints-based similarity

Fingerprints (FPs) are high dimensional vectors of bits that encode the presence or absence of a set of chemical features in a molecule. FP values depend on the atom typing scheme used, which can range from the graph representation (atoms and bonds are equivalent) to E-state atom types (each atom is influenced by its neighbouring atoms). FP variation is influenced also by the method used to map the graphical substructures (i.e., linear paths, circular growth of the molecular fragments, atom-pairs, triplets). Similarity or distance matrices computed from fingerprints can be used to screen structures by similarity to one or more reference structures. The results are returned as properties that contain the similarity to the reference structure ranging from 0 to 1. In the present work Molprint2d fingerprints [20] were used. Similarity was measured by a Tanimoto metric. Two different reference molecules were selected as templates and the following names are used: FP30 is relative to T0901317 and FP145 is relative to the most active molecule (ID 145 of LXR binders dataset).

2.2.2. QSAR model for LXR binding prediction

QSAR classification models were developed for the prediction of LXR binding potential. The dataset used for model development consisted of 97 chemicals selected from the LXR binders dataset (n=356 compounds). The dataset included 50 “very active” LXR binders (IC_{50}<20 nM), which were assigned to the ACTIVE class, and 47 “low active” LXR binders (IC_{50}>1000 nM), which were assigned to the INACTIVE class. This dataset was obtained after an under-sampling, based on structural diversity, of the compounds from the two classes. Such a dataset was constructed to be balanced between ACTIVE and INACTIVE records, and to identify the structural features (encoded by the modelling molecular descriptors) discriminating the very active from the

low active compounds. The under-sampling was performed in KNIME using an RDKit node for diversity picking (the picking is done using the MaxMin algorithm), based on 1-2D MOSES descriptors, and lead to the selection of 50 ACTIVE compounds out of 136.

To develop a QSAR model we decided to use Partial Least Square – Discriminant Analysis (PLS-DA). It is one of the most used classification technique in QSAR [59]. Its main advantage in use of this model is the possibility to control overfitting by choosing the right number of latent variables and, at the same time, to easily interpret the obtained model in terms of outliers and role played by the molecular descriptors in classification.

PLS-DA) classification models were developed and validated using SIMCA [21] based on combinations of different MOSES molecular descriptors (i.e., physico-chemical and 1D-2D-3D descriptors). Models were internally validated by means of 7-fold full cross-validation, and externally validated by means of an *a priori* 30% splitting of the dataset into training (78 compounds) and test set (19 compounds). One PLS-DA classification model was finally selected based on classification accuracy (optimizing sensitivity), external predictivity, model interpretability and reproducibility (Table 1). The model was based on three latent variables derived from seven MOSES 2D descriptors (selected based on the parameter Variable Influence on Projection, that maximizes Q^2_{CV}), which encode basic electronic properties, hydrophobicity, molecular shape and complexity: HDon_O (Number of oxygen atom-based hydrogen bonding donors), Polariz (polarizability), NRotBond (Number of rotatable bonds), NAtoms (Number of atoms), NStereo (Number of tetrahedral stereo centers), Complexity (Molecular complexity), Rgyr (radius of gyration). The molecular descriptors were centered and scalded by the software SIMCA before modelling.

Two approaches were used to define the applicability domain of the model: i) Leverage approach, ii) Similarity approach (based on Euclidean distances). Applicability domain thresholds (ADT) were calculated for each approach:

- ADT for Leverage = 0.231
- ADT for Similarity = 16.55

Compounds with leverage/similarity values greater than the specified thresholds are considered outside the applicability domain of the model and their predictions could be not reliable since model extrapolations.

The PLS-DA was also implemented as KNIME workflow (published on KNIME Webportal [8] and in SEURAT-1 Tools & Methods Catalogue [61]) to generate binary predictions for LXR binding potential, namely "1= ACTIVE" (i.e., LXR binders) and "0= INACTIVE" (i.e., not or weak LXR binder) (see Figure 3).

	No.	Overall Accuracy	Sensitivity	Specificity	False Negatives
Training set	78	86% [a] (87% [b])	93% [a] (90% [b])	79% [a] (84% [b])	3 [a] (4 [b])
Test set	19	79% [a,b]	80% [a] (70% [b])	78% [a] (89% [b])	2 [a] (3 [b])

Table 1. Summary of the selected PLS-DA classification model for LXR binding. [a] Statistics of the original QSAR model developed with SIMCA; [b] Statistics of the QSAR model implemented in KNIME.

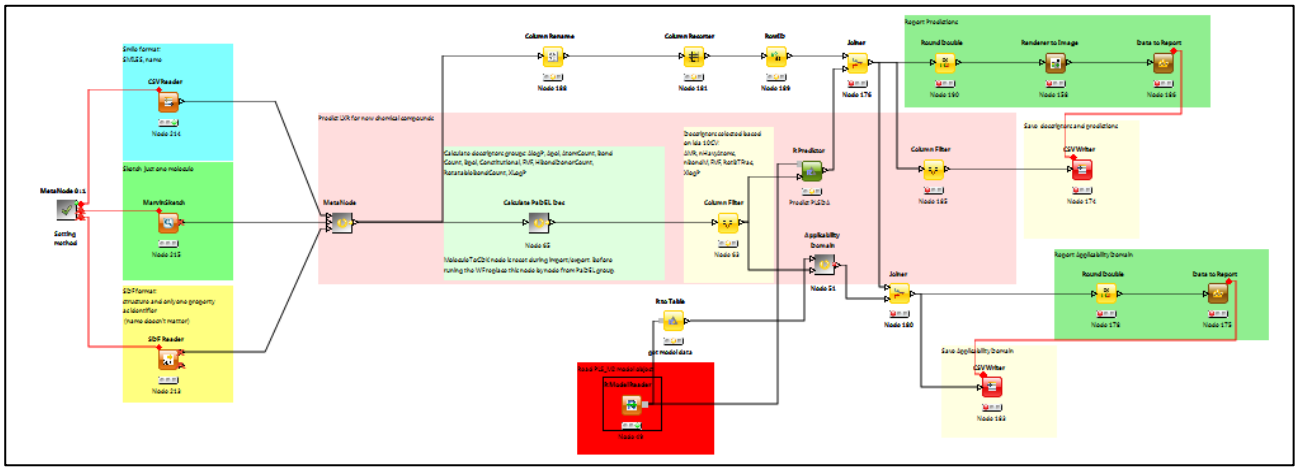


Figure 3. KNIME workflow LXR binding prediction (PLS-DA)

2.2.3. KNIME workflow for nuclear receptor-mediated liver steatosis alerts

The KNIME workflow (WF) [25, 26, 61] identifying ligands for the LXR receptor was used to generate binary predictions for LXR binding potential, namely "1=ACTIVE" (i.e., presence of an alert for LXR binding) and "0=INACTIVE" (i.e., no alerts identified). This workflow was developed in the COSMOS project and includes a set of rules or structural features and physico-chemical ranges to identify potential NR ligands [19, 20]. These rules were built based on nuclear receptors ligands and relative binding data extracted from ChEMBL_19 [22] and the Protein Data Base [11]. The ChEMBL data includes K_i (inhibition constant), K_d (dissociation constant), AC_{50} (50% activity in molar units) and EC_{50} (50% effect concentration in molar units). These compounds, when regarded as active (pChEMBL $\geq$ 5), were described in terms of physico-chemical descriptors by means of CDK within KNIME [7] and similarity towards each other to obtain potentially important substructures (IDEA consult [23]). The Protein Data Base was searched for crystal structures of relevant protein-ligand interactions using PyMOL [24].

2.3. Methods for ranking/screening chemicals

Ranking methods belong to Multi-criteria Decision Making (MCDM), a discipline dealing with decisions that involves the choice of a best alternative from several potential candidates in a decision, subject to several criteria or attributes [27]. Mathematics applied to decision making provides methods to quantify or prioritize different judgements that are typically subjective. Ordering (or ranking) is one of the possible ways to analyse data and to get an overview over the elements of a system, where the elements are commonly described by several variables. The general workflow for ranking is summarized in the scheme illustrated in Figure 4. Initial data are transformed to the same domain often [0,1], then ranking methods are applied. There are numbers of ranking algorithms that order elements within the dataset. For example, utility and desirability are measures from the total ordering. Pareto order is another type of order that introduces a relation

of incomparability between elements. In this paper, we consider three different ranking methods: i) Total Order Ranking, ii) Consensus modelling, iii) Data fusion. Below we provide a detailed description for these approaches and we will test them to rank chemicals based on potential LXR binding. Our input data consist of a combination of predicted values measuring the LXR binding possibility.

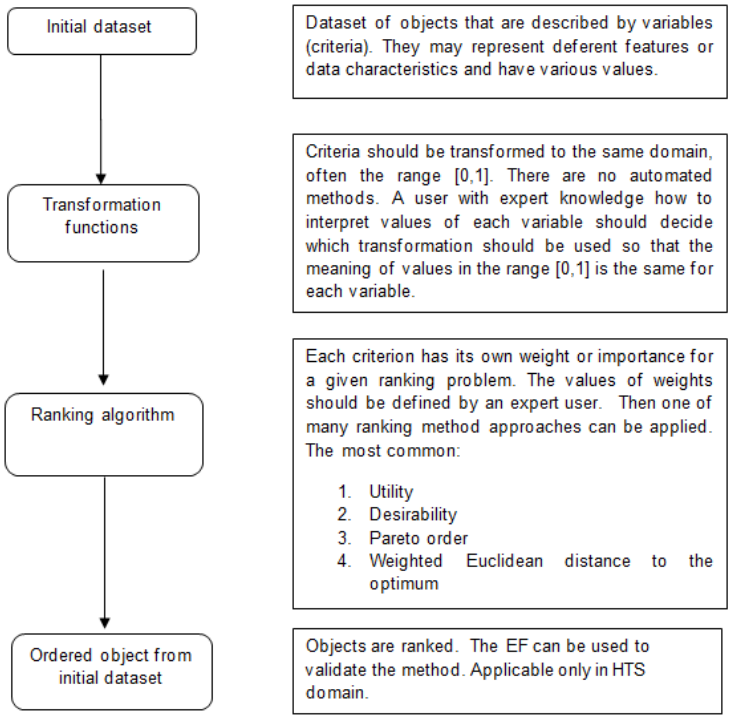


Figure 4. Scheme describing a standard ranking procedure.

2.3.1. Data normalization

Data transformation is required prior applying ranking methods. All input variable (criteria) should be unified to the same domain. In the case of LXR potential binders, the criteria can be scaled using a linear transformation. As the output class is known we can transform data to have an increasing order where min corresponds to inactive class and max to active class. We used these two normalization methods:

- Min-Max normalization [0,1] is a process of taking data and transforming it to a value between 0.0 and 1.0. The lowest (min) value is set to 0.0 and the highest (max) value is set to 1.0

$$\text{Normalized}(x_i) = \frac{x_i - \min}{\max - \min},$$

where min and max are calculated for a given variable/criteria.

- Z-score scaling (N[0,1] distribution) is a process of converting data into a number of standard deviations that the test score being converted is above or below the mean:

$$z = \frac{x - \text{mean}}{\text{st. dev}}$$

All ranking methods have been applied to both normalizations. Min-max and z-score are affine transformations so they do not change results for consensus models and total order methods, the only difference can be noticed for data fusion approaches

2.3.2 Ranking methods

2.3.2.1 Total Order Ranking

Total order ranking (TOR) methods are scalar techniques that can be used to rank objects on the basis of more than one criterion. The different criteria values are combined into a global ranking index, and objects are ordered sequentially according to the numerical value of the ranking index. Since criteria are not always in agreement (i.e., criteria can be conflicting), there is a need to find an overall optimum that can deviate from the optima of one or more of the single criteria [28]. In this paper, the two methods known as utility and desirability (see below) are used for ranking. These methods were successfully implemented in the Decision Analysis by Ranking Techniques (DART) software [29] that can be used for ranking chemicals according to their environmental and toxicological concern.

Utility is defined as an arithmetic mean of a number of criteria (c) taken into account during the ranking/screening process: $U_i = \frac{\sum_{j=1}^{p} c_{ij}}{p}$

W_Utility is defined as a weighted mean of a number of criteria (c) taken into account during the ranking/screening process: $U_i = \sum_{j=1}^{p} w_j c_{ij}$

Desirability is a geometric mean of a number of criteria (c) taken into account during the ranking/screening process: $D_i = \sqrt[p]{\prod_{j=1}^{p} c_{ij}}$,

W_Desirability is a weighted geometric mean of a number of criteria (c) taken into account during the ranking/screening process: $D_i = c_{i1}^{w_1} c_{i2}^{w_2} \dots c_{ip}^{w_p}$,

where $\sum_{i=1}^{p} w_i = 1$ and all criteria (c) must be in the same domain.

The weights can be calculated if the output class is known as follows:

$$w_i = \frac{cor_i}{\sum_{i=1}^{p} cor_i}$$

for i=1,...,p and where cor_i is a Pearson correlation coefficient between each criteria and the output class. Otherwise a user decides what is the importance of each criterion for a given ranking problem and assigns weights manually.

The total order methods can be applied in a general case for the single endpoint or combination of various endpoints to define level of a toxicity concern. In this case each criterion should be transformed to the range [0,1] with the same meaning for each variable/criteria.

We applied the four ranking approaches, namely Utility, Weighted Utility, Desirability and Weighted Desirability, to rank compounds according to their LXR binding potential, and results are discussed in the next section.

2.3.2.2. Consensus modelling

In order to classify a chemical compound to be a potential LXR binder (active/inactive), binary classification models aggregating the results from different modelling approaches were developed. This is a novel approach in the context of ranking, most of classifiers used in the literature are used for predicting a given outcome and their results are averaged to build a consensus model [30]. In this paper we develop yet another classifier that takes the output of other modelling methods as an input variable. Two models are proposed based on the different methodologies reviewed: logistic regression and random forest. The logistic regression is a parametric model: a linear model is fitted to estimate a quantity which, after transformation by logistic function equals the probability that an observation will belong to a particular class. The random forest is an efficient non-linear tool: a combination of tree predictors such that each tree is built independently from the others. Notably, random forest copes well with variables with non-linear scale. The ranking score was then defined by a value that represents the possibility that a given chemical compound is in the "active" class. More detailed description of chosen models is provided below.

1) Logistic regression is a probabilistic statistical classification model [31] that can be used to predict a binary response. The probabilities describing the possible outcomes of a single trial are modelled as a linear function of the explanatory (predictor) variables using a logistic function. While linear regression uses ordinary least squares to find a best fitting line, and comes up with coefficients that predict the change in the dependent variable for one unit change in the independent variable, logistic regression estimates the probability of an event occurring. The regression coefficients are estimated using maximum likelihood estimation and they represent the change in the logit for each unit change in the predictor. The probability of the event occurring (i.e., probability for a compound to be in the active class) was taken as a ranking score.

2) Random Forest (RF) model introduced by Breiman [32] is a collection of tree predictors. Each tree is grown according to the following procedure [33]:

- the bootstrap phase selects randomly a subset of the training dataset – a local training set for growing the tree; the remaining samples in the training dataset form a so-called out-of-bag (OOB) set and are used to estimate the RF's goodness-of-fit.
- the growing phase grows the tree by splitting the local training set at each node according to the value of one variable from a randomly selected subset of variables (called the best split method) using the classification and regression tree (CART) method [34].
- each tree is grown to the largest extent possible; there is no pruning.

The bootstrap and growing phases require an input of random quantities. It is assumed that these quantities are independent between trees and identically distributed. Consequently, each tree can be viewed as sampled independently from the ensemble of all tree predictors for a given training dataset. For prediction, an instance is run through each tree in a forest down to a terminal node which assigns it a class. Predictions supplied by the trees undergo a voting process: the forest returns a class with the maximum number of votes. Draws are resolved through a random selection. The percentage of trees voting towards the active class was taken as a ranking score.

The consensus modelling approaches that are proposed here can be applied in a general case to aggregate various models for a single endpoint when the binary output class is given. In this paper, we applied these consensus models for ranking chemicals based on their LXR binding potential.

2.3.2.3. Data fusion

Data fusion methods [35] were employed as an additional approach for ranking chemicals. The scores obtained from the three independent MM methods (docking, e-pharmacophore and fingerprint-based similarity as described in section 2.2.1) cannot be directly added or averaged to obtain a single score. In fact, the docking score is provided in units of kilocalories per mole (with more negative values representing higher LXR binding affinity), the fingerprint scores lie on the range [0, 1] and e-pharmacophore fitness is always > 0 with the greatest values representing higher

LXR binding affinity. Furthermore, simply scaling the scores from the methods does not properly account for the variability and dynamic range of the different methods. Here we converted the scores for each ligand to a standard score (i.e., Z-score) as defined in Eq 1. Z-scores are particularly useful because they indicate by how many standard deviations a value is above or below the mean of a distribution. The Z-score is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

$$Z_{mi} = \frac{S_{mi} - \mu_m}{\sigma_m} \quad (1)$$

In Eq 1, Z_{mi} is the Z-score obtained for ligand i from method m. S_{mi} is the score of the i_{th} ligand in the database for method m, μ_m is the mean score of all the compounds (actives plus decoys) in the database, and σ is the standard deviation of the distribution of the scores obtained from the method m. The sign on the docking scores was inverted to maintain consistency with fingerprints and fitness, where a more positive score is better.

Importantly, the input data should be normalized before the data fusion methods could be applied. The normalized variables represent the ranking scores rs_i that can be further combined to give a final score. In chemistry data fusion is often used to combine various database searches according to chemical compound similarities. In this paper we consider the following methods:

- Z2: average Z-scores for two (best) methods
- Z3: average Z-scores for three (best) methods
- SUM RANK = $\sum_{i=1}^{p} rs_i$ where p is a number of criteria.
- MAX RANK $\max(rs_1, rs_2, \ldots, rs_p)$ where p is a number of criteria.

The scores obtained from different data fusion methods were sorted to rank the compounds and to calculate the enrichment factors (EF).

2.3.Validation methods

To validate and compare different *in silico* modelling approaches (i.e., molecular modelling approaches, QSAR model and alerts), the enrichment factor (EF) was used as reference parameter.

The enrichment factor is a key parameter used in molecular modelling studies to evaluate the quality of the docking and scoring compared to a random selection [35]. The enrichment factor (EF) is defined as:

$$EF = \frac{A_S|T|}{|S|A_T}, \quad (2)$$

where A_s is the number of active compounds in a sampled set, T is the total number of all chemicals, S is the number of chemicals in sampled set and A_T is the total number of active compounds.

In general, the enrichment factor is compared to random screening; the maximum enrichment is determined by the total number of active compounds and the total number of molecules in the database. In this study, there are 299 “ACTIVE” compounds (including LXR binders classified as “moderate active”, “active” and “very active”) among a total of 1356 molecules (including 1000 decoys, 57 “low active” LXR binders and 299 actives). Thus, the achievable maximum EF is 1356/299 = 4.5. If only 5% of active compounds were found among the top ranked 2% of the whole dataset, then the enrichment factor would be equal to 2.5, which corresponds to a random selection at the 2% of the dataset. The enrichment factor is calculated considering the top 20% chemicals. In this paper, we reported EF at 1, 2, 5, 10 and 20%.

To compare different modelling methods, another metric was employed: the area under Accumulation Curve (AC) analysis for EF curves [36]. The AC curve represents a function that plots the true positive rate as a function of the fraction of the data classified as positive at a given threshold of screened compounds. The area under the curve (AUC) represents a quantification of

the curve and facilitates an easier comparison of results. The received value is in the range [0.0, 1.0], where 0.5 indicates a random performance.

3. Results

In the present paper, a case-study exemplifying the use of different ranking approaches (i.e. total order ranking, consensus modelling and data fusion) that combines results from different sources is presented to rank chemicals based on their potential binding to Liver X Receptor (LXR). The application of these methodologies took into account the following issues: i) different type of input data, which depended on the output of the different *in silico* models used to predict LXR binding (i.e., continuous values from MM approaches, and categorical values from the QSAR classification model as well as the KNIME WF for NR-mediated liver steatosis alerts); ii) different normalization techniques, namely Min-max and z-score normalization.

Different *in silico* modelling approaches developed for LXR binding have been employed, compared and integrated to identify the best performing procedure for prioritizing chemicals according to their LXR binding potential. Combinations of models for LXR binding were considered in two use cases for their integration by ranking approaches. A summary is provided in Table 2.

3.1. Use case 1 - ranking based on Molecular Modelling methods

The original LXR binders dataset (n=1356) was cleaned due to the presence of chemicals having missing values for some of the employed MM approaches leading to a final dataset of 1104 chemicals, with 286 labelled “ACTIVES” and 818 “INACTIVES”.

Use cases	***In silico* models**	**Input data type**	**Ranking methods**
1**) Ranking based only on MM methods**, using the enrichment factor (EF) as reference parameter to assess their ability to identify LXR	- Ensemble Docking (ED) - e-Pharmacophore (eP) - Fingerprint similarity-30 (FP-30) - Fingerprint similarity-145 (FP-145)	Continuous values	Consensus modeling - RF - LogReg Total Order Ranking - Utility - Desirability

binders.			- W_Utility - W_Desirability Data fusion - Z2 - Z3 - MAXRANK - SUMRANK
2) **Ranking based on different *in silico* approaches**, using the enrichment factor (EF) as reference parameter to assess their ability to identify LXR binders.	- Ensemble Docking (ED) - PLS-DA model - WF for NR-mediated liver steatosis alerts	Categorical values	Consensus modeling - RF - LogReg Total Order Ranking - Utility - Desirability - W_Utility - W_Desirability Data fusion - Z2 - Z3 - MAXRANK - SUMRANK

Table 2 **Use cases for applying ranking approaches based on LXR binding potential.**

As a preliminary step, a comparison among the four MM models was performed based on i) the enrichment factor (Figure 5) and ii) the correlation (by means of Pearson correlation coefficient) between the values obtained from different MM approaches and the ACTIVE/INACTIVE class. For the calculation of the correlation, the Ensemble Docking scores were multiplied by minus one (-1) to have the same increasing order of scores as the other approaches. The following Pearson correlation coefficients were obtained: Ensemble Docking ("docking_score"), r = 0.71; e-Pharmacophore ("Fitness"), r = 0.18; Fingerprint similarity-30 ("FPsimilarity30"), r = 0.27; Fingerprint similarity-145 ("FPsimilarity145"), r = 0.55. The Ensemble Docking, followed by FPsimilarity145, resulted to be the best performing approach based on both EF and correlation analysis.

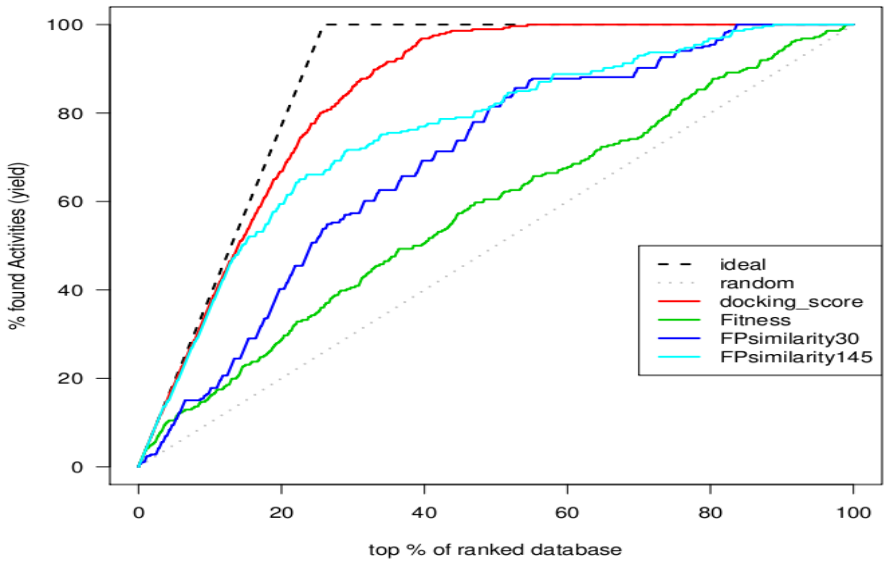


Figure 5. EF curves for different MM models.

3.1.1. Development of the Consensus models: LogReg and RF

As a first step, the two consensus models based on different approaches, i.e., Random Forest (RF) and Logistic Regression (LogReg), were developed using as input variables the results generated from the four MM models and the dataset of 1104 chemicals to train the two models. The 10-fold cross-validation was then used to test the robustness of the consensus models and the obtained statistics are reported in Table 4.

Model	TP	TN	FP	FN	ACC	ERR
LogReg	249	750	68	**37**	0.90	0.10
RF	243	774	44	**43**	0.92	0.08

Table 4 Statistics of logistic regression (LogReg) and Random Forest (RF).

Although the Random Forest shows a higher accuracy than the LogReg model, it is characterized by a higher number of false negatives (i.e. active compounds classified as inactive). This is related to the fact that the RF model is affected more by the unbalanced dataset used to train the models, where the major class is the inactive one.

The RF model was developed in R using random forest model from the caret package [60]. To build a model we used 10 trees as the input data has the small number of variables. Similarly the LogReg model was build using generalized linear model logit from the caret package. The outcomes of these models were used to rank chemicals for the LXR binding potential.

3.1.2. Comparing ranking approaches based on MM methods

As illustrated in Table 2, ten ranking methods based on three different ranking approaches (consensus modeling, total ordering and data fusion) were applied. The weight for W_Utility and W_Desirability were calculated from the Pearson correlation values reported before for the four MM methods and according to the equations reported in Section 2.3.1. The scores for LogReg represent the probability of being in the active class; the scores for random forest represent the percentage of trees voting towards the active class.

The enrichment factor (EF) was used as reference parameter to assess the ability of the models to identify LXR binders. Ensemble docking, which produced the best score for ranking among all MM approaches, was used as reference for the assessment and comparison of different ranking methods.

Because the normalizations used in our case are affine transformations of the input variables we developed separate models for each normalization type using the entire dataset for model generation as described below.

3.1.2.1. Min-max normalization

All criteria were normalized by min-max, and then the ten ranking methods were applied. The corresponding enrichment factors curves are presented in Figure 6a, where docking, random guess and the ideal model are also shown. The EF for random guess is equal to 1 so any method that gives EF > 1 is considered to be better than random. Table 5 presents the EFs calculated for a sample of max 20% top elements in the ranking.

The AUC analysis was used to compare various ranking methods (Table 6). Three approaches are characterized by AUC values (based on EF curves) greater than AUC for Docking, i.e. the two Consensus models (i.e., Random Forest and Logistic Regression) and the Weighted Utility. Among these three methods, in this case the Random Forest resulted as the best performing approach.

3.1.1.1. Z-score normalization

Z-score normalization was applied to the ten ranking methods. Due to negative values of z-scores, W_Utility and W_Desirability were not calculated. The enrichment factors were calculated for the remaining eight methods and respective curves are presented on Figure 6b. Table 5 presents the EFs calculated for a sample of max 20% top elements in the ranking.

As performed for the min-max normalization, the AUC analysis was used to compare various ranking methods (Table 6). In this case, two approaches are characterized by AUC values (based on EF curves) greater than the corresponding AUC for Docking, namely Random Forest and Logistic Regression; the Random Forest proved to be the best performing ranking approach.

Ranking	Min-Max					Z-Score				
	1%	2%	5%	10%	20%	1%	2%	5%	10%	20%
Docking	3.86	3.86	3.72	3.69	3.34	3.86	3.86	3.72	3.69	3.34
RF_scores	3.66	3.69	3.65	3.76	3.49	3.25	3.34	3.65	3.72	3.51
LogReg_scores	3.86	3.86	3.86	3.69	3.44	3.86	3.86	3.86	3.69	3.44
Utility	3.16	3.15	2.94	2.94	3.13	2.46	2.99	2.6	2.81	3.15
Desirability	3.16	3.34	2.53	2.74	3.04	3.86	3.86	3.58	3.43	3.23
W_Utility	3.86	3.86	3.86	3.58	3.30	-	-	-	-	-
W_Desirability	3.86	3.86	3.44	3.51	3.13	-	-	-	-	-
z2	3.86	3.67	3.30	3.06	3.07	2.46	2.99	2.73	2.81	3.2
z3	3.15	3.16	2.95	2.95	3.16	2.46	3.16	2.66	2.87	3.18
maxRank	3.86	3.86	3.43	3.36	3.18	3.15	3.16	2.62	2.92	3.18
sumRank	3.16	3.15	2.94	2.94	3.13	2.46	2.99	2.6	2.81	3.15

Table 5. EFs calculated for a sample of 1%, 2%, 5%, 10% and 20% top elements in the ranking for min-max and z-score normalizations.

Normalization	Docking	Consensus modeling		Total Order Ranking				Data Fusion			
		RF	LogReg	Utility	Des	W_Ut	W_Des	z2	z3	maxRank	sumRank
min-max	0.80	**0.87**	**0.84**	0.70	0.62	0.81	0.68	0.63	0.70	0.73	0.70
z-score	0.80	**0.87**	**0.84**	0.69	--	0.79	--	0.71	0.71	0.68	0.69

Table 6. AUC values for Docking and different ranking methods based on min-max and z-score normalizations.

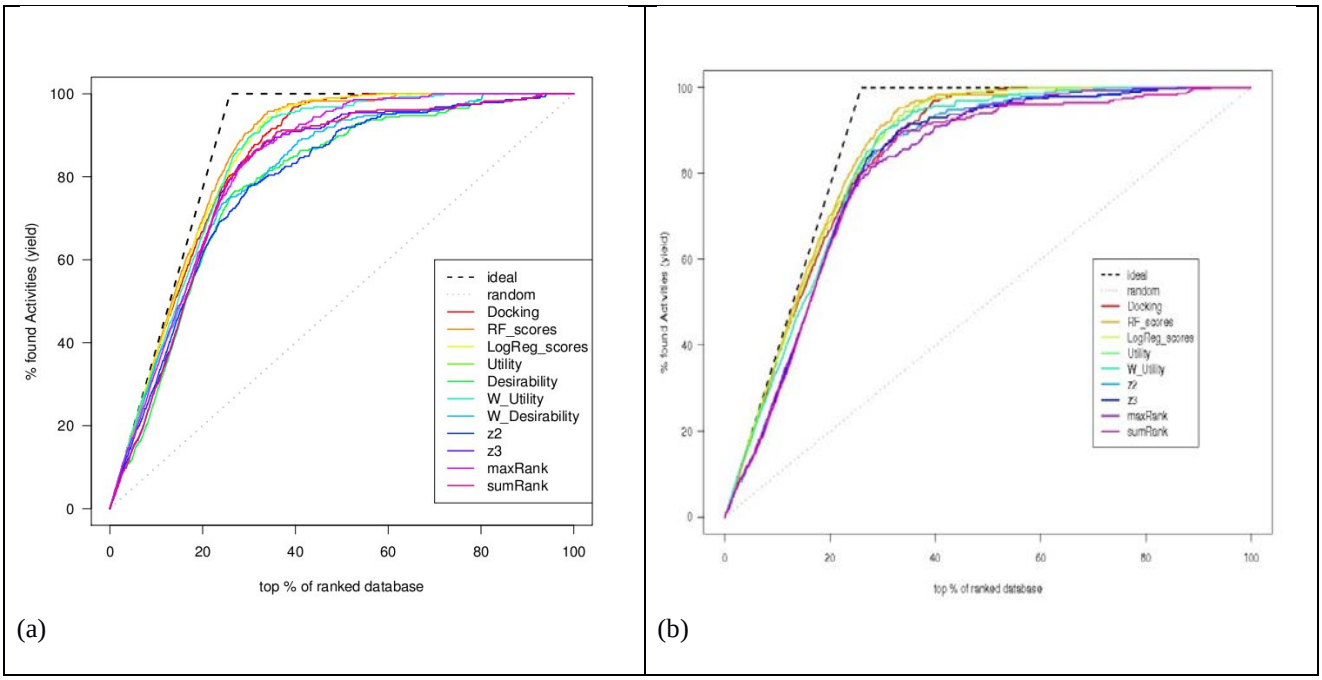


Figure 6. EF curves for different ranking methods and Docking (as reference). a) min-max normalization; b) z-score normalization

The results show that consensus models perform better than the other ranking methods, especially data fusion.

3.2. Use case 2 - ranking based on different *in silico* approaches

The original LXR binders dataset (n=1356) was cleaned due to the presence of chemicals having missing values in the results from the considered *in silico* approaches, i.e. Ensemble Docking (best performing approach among MM methods), the PLS-DA classification QSAR and the NR-mediated liver steatosis alerts (only the LXR pathway was considered). Compounds for which the PLS-DA classification QSAR gave unreliable prediction were also removed. This was obtained using the

applicability domain thresholds as described in Section 2.2.2. The final dataset consists of 870 chemicals, including 286 "ACTIVES" and 584 "INACTIVES".

Similarly to the Use case 1, a comparison among the different *in silico* approaches for LXR binding was performed based on i) the enrichment factor (Figure 7) and ii) the correlation between the values obtained from different approaches and the ACTIVE/INACTIVE class. For the calculation of the correlation, the Ensemble Docking scores were multiplied by (-1) to have the same increasing order of scores as the other approaches. The following Pearson correlation coefficients were obtained: Ensemble Docking ("docking_score"), r = 0.73; NR-mediated liver steatosis alerts ("NRass_WF_LXR_alerts"), r = 0.40; PLS-DA classification QSAR ("PLS_class"), r = 0.23.

Both EF and Pearson correlation showed that Ensemble Docking (followed by LXR alerts) is the best approach for ranking chemicals based on their potential of being active or inactive for LXR binding. Considering the endpoint type, i.e. binding affinity, it was expected that molecular modelling could provide better results than the other two modelling approaches, which are based on 2D information.

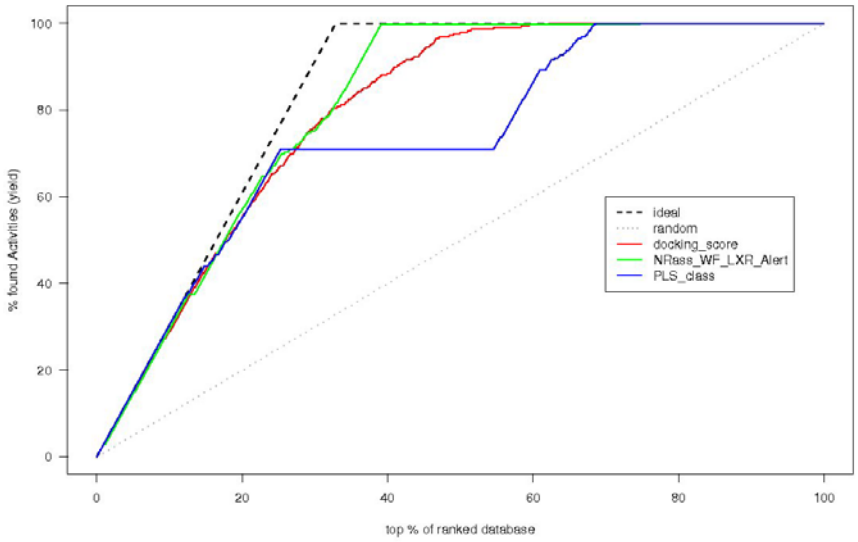


Figure 7. EF curves for different *in silico* models: 1) Ensemble Docking ("docking_score"), 2) NR-mediated liver steatosis alerts ("NRass_WF_LXR_Alert"), 3) PLS-DA classification QSAR ("PLS_class").

3.2.1. Development of the consensus models: LogReg and RF

Two consensus models based on different approaches, i.e., Random Forest (RF) and Logistic Regression (LogReg), were developed using as input variables the results generated from the three different *in silico* models and the dataset of 870 chemicals was used to train the two models. The 10-fold cross-validation was then used to test the robustness of the consensus models and the obtained statistics are reported in Table 7. In this case the random forest model deals better with false negatives than the logistic regression and is characterized by comparable overall accuracy.

Model	TP	TN	FP	FN	ACC	ERR
LogReg	229	537	47	**57**	**0.88**	0.12
RF	237	516	68	**49**	**0.87**	0.13

Table 7. Statistics of logistic regression (LogReg) and Random Forest (RF).

Similarly to the scenario 1, the RF and LogReg models was developed in R using the caret package. Ten trees were used for the RF model development. The outcomes of these models were used to rank chemicals for the LXR binding potential.

3.2.2. Comparing ranking approaches based on different in silico methods

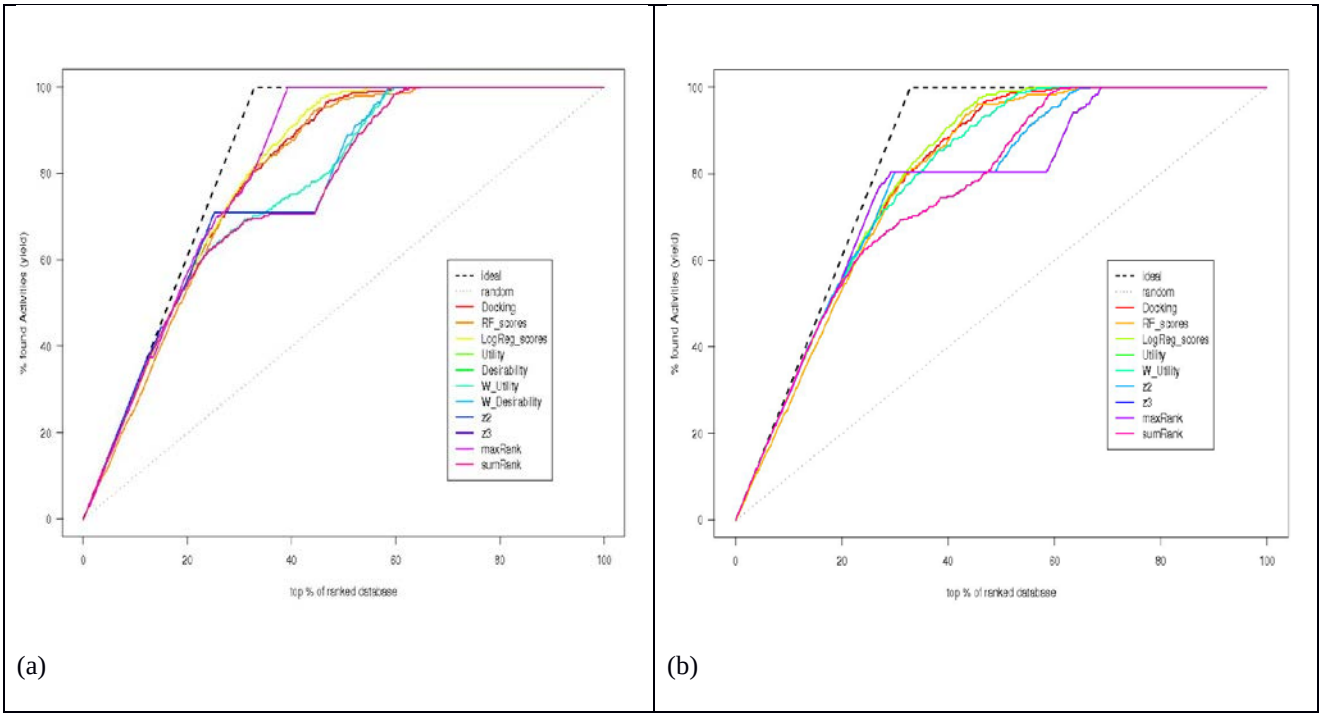
As described above, ten ranking methods, based on three different ranking approaches (consensus, total ordering and data fusion), were applied (see Table 2).

The enrichment factor (EF) was used as reference parameter to assess their ability to identify LXR binders and Docking was used as reference for the assessment and comparison of different ranking methods. Separate models for each normalization type, namely min-max and z-score normalization, were developed using the entire dataset for model generation. Based on min-max and z-score normalizations, only Logistic Regression model performed better than docking. Both consensus

models perform better than the other ranking approaches (Figure 8 and Table 8). Table 9 presents the EFs calculated for a sample of max 20% top elements in the ranking.

Normalization	Docking	Consensus modeling		Total Order Ranking				Data Fusion			
		RF	LogReg	Utility	Des	W_Ut	W_Des	z2	z3	maxRank	sumRank
min-max	0.75	0.71	**0.77**	0.63	0.63	0.63	0.63	0.36	0.63	0.50	0.63
z-score	0.75	0.71	**0.77**	0.63	--	0.73	--	0.68	0.63	0.63	0.63

Table 8. AUC values for Docking and different ranking methods based on min-max and z-score normalizations.



(a) (b)

Figure 8. EF curves for different ranking methods and Docking (as reference). a) min-max normalization; b) z-score normalization

Ranking	Min-Max					Z-Score				
	1%	2%	5%	10%	20%	1%	2%	5%	10%	20%
Docking	3.04	3.04	2.97	2.9	2.76	3.04	3.04	2.97	2.9	2.76
RF_scores	2.69	2.52	2.55	2.59	2.67	2.69	2.69	2.69	2.62	2.66
LogReg_scores	3.04	3.04	2.97	2.87	2.78	3.04	3.04	2.97	2.87	2.78
Utility	3.04	3.04	2.97	2.9	2.71	3.04	3.04	2.97	2.9	2.71
Desirability	3.04	3.04	2.97	2.9	2.71	3.04	3.04	2.97	2.9	2.8
W_Utility	3.04	3.04	2.97	2.9	2.71	-	-	-	-	-
W_Desirability	3.04	3.04	2.97	2.9	2.71	-	-	-	-	-
z2	1.59	1.59	1.59	1.59	1.59	3.04	3.04	2.97	2.9	2.8
z3	3.04	3.04	2.97	2.9	2.71	3.04	3.04	2.97	2.9	2.71
maxRank	1.18	1.18	1.18	1.18	1.18	3.04	3.04	2.97	2.9	2.66
sumRank	3.04	3.04	2.97	2.9	2.71	3.04	3.04	2.97	2.9	2.71

Table 9. EFs calculated for a sample of 1%, 2%, 5%, 10% and 20% top elements in the ranking for min-max and z-score normalizations.

3.3. KNIME Workflow for ranking

In order to allow for a wider applicability, reproducibility and transparency of the presented ranking methodology, a KNIME workflow was developed within the COSMOS project implementing some of the methods for ranking chemicals based on multiple *in silico* predictions of LXR binding potency. The following four methods were implemented in the workflow as R snippets: utility and desirability, belonging to the Total Order Ranking methods, and sumRank and maxRank, belonging to the data fusion methods. Based on the input data type, two workflows were developed: i) one workflow to be applied with already normalized input data (numerical values, i.e. predictions, in the range <0,1>, representing the probability of chemicals to be potential LXR binders); ii) the second workflow including a preliminary normalization process, to be applied with input data/predictions converted in the range <min, max>, where “min” corresponds to the lowest LXR binding probability and “max” to the highest LXR binding probability; these values are further normalized using the min-max normalization method to values in the range <0,1>. In both workflows, the ranks are collected in one table and the Sorter node is used to sort chemicals based on the selected

ranking method. Top-ranked chemicals are those of major concern, i.e. higher LXR binding potential based on different predictions and estimation methodologies.

The workflow is freely available through the COSMOS KNIME WebPortal [8] while related documentation can be found within the COSMOS Space [37].

4. Conclusions

The present paper illustrates one of the outcomes of the COSMOS project related to the development of *in silico* tools to support long-term and target organ toxicity prediction. The main aim was to demonstrate the power of integrated modelling framework for chemical prioritization. Specifically, some of the case studies reported in the paper exemplify the use of ranking approaches for screening and ranking (i.e., sorting) of chemicals based on their potential to bind to Liver X Receptor (LXR), as this has been identified among the molecular initiating events leading to liver steatosis. The novelty of this work is twofold: the integration of molecular modelling methods with (Q)SAR approaches for predicting binding potential, and the application of consensus model for ranking chemicals. The case studies were based on a dataset of 1356 molecules, including a set of 356 LXR binders extracted from the literature (mainly drugs or drug candidates with measured IC_{50} LXRβ binding affinity data) and 1000 Decoy molecules selected from Schrödinger 1K Drug-Like Ligand Decoys Set (molecules presumed to be inactive against LXR). We studied different combinations of *in silico* models for the prediction of LXR binding potential, which leads to the following conclusions:

- Among different *in silico* approaches we tested (e.g., docking, fingerprints, pharmacophore, QSAR and structural alerts), Ensemble Docking was the best-performing approach for the identification of potential LXR binders. This result can be explained by the structure of the LXR receptor, for which the docking performs exceptionally well, and by the endpoint type, i.e. binding affinity, where 3D information encoded by docking allows for a better

prediction performance than the other employed modelling approaches (QSAR and structural alerts), which are based only on 2D information.

- Normalization approaches, either based on min-max or z-score methods, did not significantly affect the performance of the ranking approaches in the 1–20% range.
- Among the ranking methods tested in the use case 1 (MM approaches), the two consensus modelling approaches, namely Random Forest and Logistic Regression showed better results than the use of docking alone to rank chemicals based on their LXR binding potential.
- Among the ranking methods tested in the use case 2 (in silico approaches), only Logistic Regression showed better results than the use of docking as a single measure to rank chemicals based on their LXR binding potential.
- Both use cases proved that combined methods for screening chemicals give better evidence than using a single approach.
- Both consensus models give better results than the other ranking approaches.
- The implementation of ranking models as KNIME workflows increased the transparency and allowed for further applicability of the developed prioritization strategies.

The application of molecular modeling approaches, which have been developed and optimized for drug discovery, in the context of toxicology, e.g. for ranking purposes, is promising, especially when dealing with endpoints involving 3D interactions of ligands (e.g., exogenous chemicals) toward specific targets (e.g., receptors), as in the presented case study of LXR binding. However, their practical implementation presents several issues, especially concerning the assessment of model predictivity and applicability domain. Common metrics usually applied for (Q)SARs to assess prediction reliability and applicability domain are not directly applicable to molecular modelling approaches. In the presented case studies the approaches were validated by EF using the top 1–20% range, thus focusing only on the potentially most active compounds. Calculation of the

applicability domain was handled using a set of decoy molecules with structural and physico-chemical properties similar to LXR binders. Finally, the superior performance of the MM approaches was found to be at the expense of a higher complexity than traditional (Q)SAR approaches.

Using ranking in chemical screening allows for identification of chemicals that can be potentially dangerous or having given properties. If the true outcome is unknown for these selected chemicals, further experimental testing should be conducted. In the present study, chemicals are ranked based on their potency to bind to the LXR receptor. It was not possible to identify a direct relationship between LXR binding (in terms of IC_{50} data) and liver toxicity due to the absence of liver toxicity data for the studied compounds. Thus, it is not possible to conclude that top-ranked compounds are the most toxic. However, since LXR binding has been recognized as a molecular initiating event (MIE) leading to liver steatosis, we could conclude that the top-ranked compounds, which exhibit higher LXR binding potency (lowest IC_{50} values), could be those of higher concern for liver toxicity, thus focusing further investigation on these prioritized compounds (e.g., experimental testing).

In this paper we demonstrated the application and power of statistical and non-statistical modeling techniques in the context of chemical screening. We also proposed a framework for integration of various models for the same endpoint. The results show that consensus models outperform other standard ranking approaches and can be successfully used for chemical prioritization.

The use of ranking methodologies, which in the presented case studies was limited to the integration of different models predicting LXR binding potential, can be further extended to combine data (predicted as well as experimental) across multiple endpoints. Other case studies developed within the COSMOS project (not reported in the present paper) were based on the integration of results on LXR binding potential and liver toxicity potential (based on *in silico* predictions generated by different models). In these additional case studies, ranking methods allowed for the prioritization of

compounds of major concern for liver toxicity possibly acting through a specific MoA (e.g., from LXR binding to liver steatosis).

Overall, the procedure presented here, which makes use of *in silico* modelling approaches and chemometric tools, is part of an integrated strategy which combines multiple methodologies alternative to *in vivo* animal testing (e.g., *in silico*, *in vitro*, mechanistic information) to support repeated dose and target organ toxicity prediction in the MoA/AOP framework.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) COSMOS Project under grant agreement n° 266835 and from Cosmetics Europe. Acknowledgements go also to COSMOS partners involved in the modelling studies: Fabian P. Steinmetz and Claire L. Mellor (from Liverpool John Moores University, England), Merilin Al Sharif, Petko Alov, Ilza Pejeva, Ivanka Tsakovska, and Vessela Vitcheva (from Institute of Biophysics and Biomedical Engineering, Bulgaria).

References

[1] National Research Council (NRC), Toxicity Testing in the 21st Century: A Vision and a Strategy. The National Academies Press Washington, DC, 2007.

[2] B. Landesmann, M. Goumeou, S. Munn & M. Whelan, Description of Prototype Modes-of-Action Related to Repeated Dose Toxicity. (Publications Office of the European Union). http://publications.jrc.ec.europa.eu/repository/handle/111111111/27015, 2012

[3] M. Al Sharif , P. Alov, V. Vitcheva, I. Pajeva , I. Tsakovska, Modes-of-Action Related to Repeated Dose Toxicity: Tissue-Specific Biological Roles of PPARγ Ligand-Dependent Dysregulation in Nonalcoholic Fatty Liver Disease. Hindawi Publishing Corporation - PPAR Research. Volume 2014, Article ID 432647, 2014

[4] COSMOS (Integrated In Silico Models for the Prediction of Human Repeated Dose Toxicity of Cosmetics to Optimise Safety) https://www.cosmostox.eu, (accessed 23.07.2018)

[5] SEURAT-1 (Safety Evaluation Ultimately Replacing Animal Testing) https://www.seurat-1.eu, (accessed 23.08.2018)

[6] MOSES.Descriptors (2011) MOSES.Descriptors 1.0. Molecular Networks GmbH, Erlangen, Germany www.molecular-networks.com

[7] KNIME, https://www.knime.org, (accessed 07.07.2018).

[8] COSMOS KNIME Webportal. https://knimewebportal.cosmostox.eu (accessed 23.08.2018).

[9] W. Zhao, et al. Three-Dimensional Pharmacophore Modeling of Liver-X Receptor Agonists. J. Chem. Inf. Model. 51,pp. 2147–2155, 2011.

[10] Glide, version 6.9, Schrödinger, LLC, New York, NY, 2015.

[11] PDB database. Protein Data Bank http://www.rcsb.org (accessed 06.07.2018).

[12] H.M. Berman et al. The Protein Data Bank. Nucleic Acids Res. 28, 2000, pp. 235–242.

[13] Friesner, R. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, J. Med. Chem. 2004, 47, 1739-1749.

[14] Halgren, T. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. J. Med. Chem. 2004, 47, 1750–1759.

[15] N.K Salam, R. Nuti, W. Sherman, Novel Method for Generating Structure-Based Pharmacophores Using Energetic Analysis. J. Chem. Inf. Model. 49, 2009, pp. 2356–2368.

[16] K. Loving, N.K. Salam, W. Sherman, Energetic analysis of fragment docking and application to structure-based pharmacophore hypothesis generation. J. Comput. Aided Mol. Des. 23, 2009, pp. 541–554.

[17] Friesner, R. A. et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. J. Med. Chem. 2006, 49, 6177–6196.

[18] Phase, version 4.5, Schrödinger, LLC, New York, NY, 2015.

[19] Dixon, S. L. et al. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening. 1. Methodology and Preliminary Results. J. Comput. Aided Mol. Des., 2006, 20, 647-671. www.schrodinger.com.

[20] J. Duan, S.L. Dixon, J.F. Lowrie and W. Sherman, Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. J. Mol. Graph. Model. 29(2), 2010, pp. 157–170.

[21] SIMCA by UMETRICS, ver. 13.0.3.0. http://www.umetrics.com/products/simca

[22] ChEMBL www.ebi.ac.uk/chembl, (accessed 07.07.2014)

[23] IDEAconsult Ltd. www.ideaconsult.net/products, (accessed 13.05.2014)

[24] PyMOL, http://www.pymol.org, accessed (01.07.2014)

[25] C.L. Mellor, F.P. Steinmetz, M.T. Cronin, Using Molecular Initiating Events to Develop a Structural Alert Based Screening Workflow for Nuclear Receptor Ligands Associated with Hepatic Steatosis. Chem Res Toxicol., 29(2), 2016, pp. 203-212. doi: 10.1021/acs.chemrestox.5b00480.

[26] F.P. Steinmetz, C.L. Mellor, T. Meinl, M.T. Cronin, Screening Chemicals for Receptor-Mediated Toxicological and Pharmacological Endpoints: Using Public Data to Build Screening Tools within a KNIME Workflow. Mol Inform. 34(2-3), 2015, pp. 171-8. doi: 10.1002/minf.201400188. b 2015 Feb 20.

[27] E. Triantaphyllou, Multi-Criteria Decision Making: A Comparative Study. Dordrecht, The Netherlands: Kluwer Academic Publishers (now Springer). 2000, pp. 320.

[28] M. Pavan and R. Todeschini, Optimization: Multi-criteria Decision Making methods, in Comprehensive Chemometrics, B. Walczak, R.T. Ferré, and S. Brown, eds., 2008

[29] DART software. Available at: https://eurl-ecvam.jrc.ec.europa.eu/laboratoriesresearch /predictive_toxicology/qsar_tools/DART

[30] H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Öberg, P. Dao, A. Cherkasov and I.V. Tetko, Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. Journal of chemical information and modeling, 48 (4), 2008, pp. 766-84.

[31] S.H. Walker, D.B. Duncan. Estimation of the probability of an event as a function of several independent variables. Biometrika. 54, 1967, pp. 167–178.

[32] L. Breiman, Random forests, Machine Learning, 45 (1), 2001, pp. 5–32.

[33] L. Breiman L and A. Cutler Random forests. http://www.stat.berkeley.edu/~breiman /RandomForests/, 2008, (accessed 23.08.2018)

[34] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees, ser. Statistics/Probability Series. Belmont, California, U.S.A.: Wadsworth Publishing Company, 1984

[35] G.M. Sastry, V.S. Inakollu,W. Sherman, Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. J. Chem. Inf. Model. 53, 2013, pp. 1531−1542.

[36] J.F. Truchon and C.I. Bayly, Evaluating virtual screening methods: good and bad metrics for the ”early recognition” problem. J. Chem. Inf. Model. 47 (2), 2007, pp. 488–508.

[37] COSMOS Space. http://cosmosspace.cosmostox.eu, 2015, (accessed on 23.08.2018)

[38] S. Svensson, et al. Crystal structure of the heterodimeric complex of LXRα and RXRβ ligand-binding domains in a fully agonistic conformation. EMBO J. 22, 2003, pp. 4625–4633.

[39] E.Y. Chao et al. Structure-Guided Design of N-Phenyl Tertiary Amines as Transrepression-Selective Liver X Receptor Modulators with Anti-Inflammatory Activity. J. Med. Chem. 51, 2008, pp. 5758–5765.

[40] X. Fradera et al. X-Ray Structures of the LXRα LBD in Its Homodimeric Form and Implications for Heterodimer Signaling. J. Mol. Biol. 399, 2010, pp. 120–132.

[41] R.C. Bernotas et al. 4-(3-Aryloxyaryl)quinoline sulfones are potent liver X receptor agonists. Bioorg. Med. Chem. Lett. 20, 2010, pp. 209–212.

[42] S. Hoerer, A. Schmid, A. Heckel, A., R, M. Budzinski, H. Nar, Crystal Structure of the Human Liver X Receptor β Ligand-binding Domain in Complex with a Synthetic Agonist. J. Mol. Biol. 334, 2003, pp. 853–861.

[43] W.J. Zuercher et al. Discovery of Tertiary Sulfonamides as Potent Liver X Receptor Antagonists. J. Med. Chem. 53, 2010, pp. 3412–3416.

[44] M. Färnegårdh et al. The Three-dimensional Structure of the Liver X Receptor β Reveals a Flexible Ligand-binding Pocket That Can Accommodate Fundamentally Different Ligands. J. Biol. Chem. 278, 2003, pp. 38821–38828.

[45] W. Liu et al. Design, synthesis, and structure-activity relationship of podocarpic acid amides as liver X receptor agonists for potential treatment of atherosclerosis. Bioorg. Med. Chem. Lett. 15, 2005, pp. 4574–4578.

[46] J.W. Szewczyk et al. SAR studies: designing potent and selective LXR agonists. Bioorg. Med. Chem. Lett. 16, 2006, pp. 3055–3060.

[47] B. Hu et al. Discovery of phenyl acetic acid substituted quinolines as novel liver X receptor agonists for the treatment of atherosclerosis. J. Med. Chem. 49, 2006, pp. 6151–6154.

[48] B. Hu et al. Further modification on phenyl acetic acid based quinolines as liver X receptor modulators. Bioorg. Med. Chem. 15, 2007, pp. 3321–3333.

[49] J. Wrobel et al. Indazole-based liver X receptor (LXR) modulators with maintained atherosclerotic lesion reduction activity but diminished stimulation of hepatic triglyceride synthesis. J. Med. Chem. 51, 2008, pp. 7161–7168.

[50] B. Hu et al. Carboxylic acid based quinolines as liver X receptor modulators that have LXRbeta receptor binding selectivity. Bioorg. Med. Chem. Lett. 18, 2008, pp. 54–59.

[51] R.C Bernotas et al. Biarylether amide quinolines as liver X receptor agonists. Bioorg. Med. Chem. 17, 2009, pp. 1663–1670.

[52] B. Hu et al. Discovery and SAR of cinnolines/quinolines as liver X receptor (LXR) agonists with binding selectivity for LXRbeta. Bioorg. Med. Chem. 17, 2009, pp. 3519–3527.

[53] J.M. Travins et al. 1-(3-Aryloxyaryl)benzimidazole sulfones are liver X receptor agonists. Bioorg. Med. Chem. Lett. 20, 2010, pp. 526–530.

[54] J.W. Ullrich et al. Synthesis of 4-(3-biaryl)quinoline sulfones as potent liver X receptor agonists. Bioorg. Med. Chem. Lett. 20, 2010, pp. 2903–2907.

[55] R.R. Singhaus et al. 3-(3-Aryloxyaryl)imidazo[1,2-a]pyridine sulfones as liver X receptor agonists. Bioorg. Med. Chem. Lett. 20, 2010, pp. 521–525.

[56] B. Hu et al. Identification of phenylsulfone-substituted quinoxaline (WYE-672) as a tissue selective liver X-receptor (LXR) agonist. J. Med. Chem. 53, 2010, 3296–3304.

[57] X. Jiao et al. Discovery and optimization of a series of liver X receptor antagonists. Bioorg. Med. Chem. Lett. 22, 2012, pp. 5966–5970.

[58] D.J. Kopecky et al. Discovery of a new binding mode for a series of liver X receptor agonists. Bioorg. Med. Chem. Lett. 22, 2012, pp. 2407–2410.

[59] R.G. Brereton and G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away. J. Chemometrics, 28:, 2014 pp. 213-225

[60] Caret: Classification and Regression Training
https://cran.rproject.org/web/packages/caret/index.html (accessed 21.12.2018)

[61] SEURAT-1 Tools & Methods Catalogue,
http://publications.jrc.ec.europa.eu/repository/bitstream/JRC102532/method%20catalogue%20pubsy-final.pdf (accessed 21.12.2018)

APPENDIX A - THE LXR DATASET

Among the reviewed literature, we identified 21 key papers reporting experimental data for LXR binding affinity (affinity toward both α and β isoforms of the receptor). These papers often included additional data obtained from a variety of functional biological assays which measured LXR agonism and/or antagonism, as well as molecular responses induced upon the activation of LXR, such as the induction of LXR target genes (e.g., SREBP1c) and cellular triglyceride (TG) accumulation. A list of the selected papers is provided in Table A.1.

Source dataset	Dataset	Endpoint *
Farnegardh et al., 2003[44]	TO-901317 and GW 3965	LXRα/β binding affinity
Hoerer et al., 2003[42]	TO-901317	LXRβ binding affinity
Svensson et al., 2003[38]	TO-901317	LXRα binding affinity
Liu et al., 2005[45]	podocarpid acid analogs (n=25)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA)
Szewczyk et al., 2006[46]	heterocyclic LXR agonists (n=29)	LXRα/β binding affinity LXRα/β agonist activity (βLac TA) PPARα/β/γ binding affinity
Hu et al., 2006[47]	phenyl acetic acid substituted quinolones (n=12)	LXRα/β binding affinity
Hu et al., 2007[48]	phenyl acetic acid substituted quinolones (n=27)	LXRα/β binding affinity
Wrobel et al., 2008[49]	substituted 2-benzyl-3-aryl-7-trifluoromethylindazoles (n=17)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA, ABCA1 and SREBP1c expression)

Source dataset	Dataset	Endpoint *
Chao et al., 2008[39]	N-Phenyl Tertiary Amines (n=21)	LXRα/β binding affinity TG accumulation
Hu et al., 2008[50]	Carboxylic acid based quinolines (n=22)	LXRα/β binding affinity
Bernotas et al., 2009[51]	Biarylether amide quinolines (n=17)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA, ABCA1 expression)
Hu et al., 2009[52]	cinnolines/quinolines (n=27)	LXRα/β binding affinity
Fradera et al. 2010[40]	GW 3965 and SureCN2898933	LXRα/β binding affinity
Bernotas et al., 2010[41]	4-(3-Aryloxyaryl)quinoline sulfones (n=23)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA, ABCA1 expression)
Travins et al 2010[53]	1-(3-Aryloxyaryl) benzimidazole sulfones (n=31)	LXRα/β binding affinity LXRα/β agonist activity (ABCA1 expression) TG accumulation
Ullrich et al 2010[54]	series of 4-(3-biaryl)quinolines with sulfone substituents on the terminal aryl ring (n=18)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA, ABCA1 and SREBP1c expression) TG accumulation
Singhaus et al. 2010[55]	3-(3-Aryloxyaryl)imidazo[1,2-a]pyridine sulfones (n=32)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA)
Hu et al., 2010[56]	phenyl sulfone substituted quinoxaline (n=15)	LXRα/β binding affinity LXRα/β agonist activity (Gal4 TA, ABCA1 expression) TG accumulation

Source dataset	Dataset	Endpoint *
Zuercher et al., 2010[43]	tertiary sulfonamides (n=14)	LXRβ binding affinity LXRα/β agonist activity (Gal4 TA)
Jiao et al., 2012[57]	benzenesulfonamides (n=52)	LXRβ binding affinity LXRα/β agonist activity (Gal4 TA)
Kopecky et al., 2012[58]	pyrrole derivatives (n=9)	LXRα/β binding affinity

Table A.1. List of selected literature sources reporting experimental data for LXR binding affinity and activation.

The dataset provided as a supplementary research data consists of 13 columns containing the following information: ID (internal), SMILES, pIC50 LXRbeta BA (collected from literature), Activity_classification (based on IC50 LXRbeta), status (active/decoy), 7 columns containing results from molecular modelling described in Table A.2 and a source.

Scoring function	Approach	Software	Outcome
docking_score	Molecular Docking	Glide Schrödinger	-
Fitness	Pharmacophore Matching	Phase Schrödinger	-
FPsimilarity30	Tanimoto similarity using binary fingerprints (Molprint2d) using as reference cmpd 30 of LXR dataset	Canvas Schrödinger	-
FPsimilarity145	Tanimoto similarity using binary fingerprints (Molprint2d) using as reference cmpd 145 of LXR dataset	Canvas Schrödinger	-
NRass_WF_LXR_Alert	Structural Alerts for LXR binding by applying the KNIME WF: NRassHepatSteat_03092014_FS	KNIME	1= presence LXR alert. 0= no LXR alerts

PLS_class	PLS-DA QSAR based on MOSES descriptors (HDon_O, Polariz,NRotBond, NAtoms, NStereo, Complexity, Rgyr)	KNIME	1= active. 0=inactive
PLS_AD	Applicability Domain of PLS QSAR model based on Similarity	KNIME	reliable (into AD); unreliable (out AD)

Table A.1. List of columns in LXR dataset.

.