Development and Validation of Pharmacoinformatic Similarity-based Tools for Safety Assessment of Chemicals

Kevin Pinto Gil

TESI DOCTORAL UPF / 2019

DIRECTOR DE LA TESI

Dr. Manuel Pastor Maeso,

DEPARTAMENT CEXS

The research leading to these results has received support from the Innovative Medicines Initiative (IMI) resources of which are composed of financial contribution from the European Union’s projects 1) FP7 IMI1 Intelligence Led Assessment of Pharmaceuticals in the Environment (iPiE) under grant agreement 115735 and 2) H2020 IMI2 Enhancing TRANslational SAFEty Assessment (eTRANSAFE) under grant agreement 777365, through Integrative Knowledge Management and EFPIA companies’ in-kind contribution.

Agradecimientos / Acknowledgements

No ha sido un camino lleno de rosas u orquídeas. Muchos puentes disulfuro se han roto, pudiendo llegar a elevarse a categoría de problema. Pipi, muchas cebollas has puesto en mi camino para sacarme todas esas gotas llenas de sentimientos llamadas lágrimas. Sí, a veces me he sentido como un eosinófilo, creyéndome débil, comparándome con mis otros compañeros, el resto de glóbulos blancos, sin recordar que yo también tengo mis funciones imprescindibles. Sí, a veces, he estado dentro de ese bucle sin aparente salida, pero entonces aparecían mis DNA polimerasas para sacarme, la gente que realmente creía en mí.

En primer lugar, gracias a Manuel Pastor por haberme dado la oportunidad de hacer el doctorado en su grupo. Gracias a Ferran Sanz por el soporte económico.

A Carina, Chus, Miguel y Alfons por vuestra ayuda cuando os la he pedido, siempre disponibles y con una sonrisa.

Gracias a mis primeros excompañeros de doctorado, Pau Carrió (fue breve, pero me lo pase genial contigo), Oriol López (mil gracias por todo, por tus consejos, por preocuparte por mí, un ejemplo a seguir sin duda).

Gracias a todas las personas que han pasado por el laboratorio, Nacho, Bet, Biel, Eric, Ismael, César y José Carlos. Todos habéis contribuido en mi aprendizaje muchísimo, aunque no lo creáis. Gracias por escuchar mis confesiones y por esas risas. Especialmente gracias a ti JC, gracias por ayudarme un montón guiándome. Echaré de menos esas comidas en el Shoko y Moncho's.

Gracias y mil gracias a vosotros, Sergi, Montse, Laura, Aida, Ana, Elena, Dani e Yvonne. Con vosotros descubrí que la ciencia puede ser divertida y maravillosa si la compartes con gente tan buena como vosotros, fueron unos tiempos que nunca olvidaré. Yvonne, siempre te he admirado y nunca tendré palabras suficientes de agradecimiento por todo lo que has hecho y sigues haciendo por mí. Personas tan buenas y bonitas como tú son difíciles de encontrar, y yo soy muy afortunado de haberte encontrado.

A veces cuando menos te lo esperas, te reencuentras con personas como Dani que cambian tu rumbo totalmente en la vida. Gracias Dani por reaparecer en ese mismo instante a esa misma hora y por seguir estando ahí. Gracias a ti, he conocido a muchas personas más, entre ellas a mi Pauet y Manu, a Yura. Os quiero.

A mis amigos y amigas de la infancia. Que decir de vosotrxs. Sin duda alguna os llevo y os llevaré siempre en mis

pensamientos y en lo más hondo de mi corazón. Alicia, Zaida, Lourdes, Alexia, Javi. Con vosotrxs, he pasado media vida y lo que nos queda. Zaida, tu fuiste la culpable de que estudiara farmacia y mira dónde he llegado. Alicia, eres única y siempre, siempre, has estado a mi lado. Alexia, que persona tan bonita eres por dentro y por fuera. Esas risas cada vez que nos vemos no nos las quitarán nunca nadie. De ti también he aprendido muchísimo. ¿Marga, quién nos iba a decir que nos reencontraríamos otra vez y de esta forma?, gracias por ser como eres conmigo. Javi, muchísimas gracias por estar a mi lado, por aconsejarme, por escucharme y por tener siempre una sonrisa. Lourdes, te dejo para el final porque sabes lo que significas para mí, hemos pasado y aprendido tantas cosas juntos … desde aquel trabajo sobre Murcia, fiestas de pijamas, Edimburgo, Rubí y Hospitalet. Eres tan especial para mí. Te quiero muchísimo y lo sabes. Os quiero a todxs.

A vosotros, Karoru, Raquel y César. No podría olvidarme de vosotros porque formáis una parte importante de mi vida. Gracias por aportarme tanto. Por esas visitas a Edimburgo Karoru, por la LOCURA.

A todxs vosotrxs, Vane, Sonia, Sara, Marta, Fabiola, Rosa, Jose Luis, Meri, Toni Picornell, Beatriz, Montse, Maria, Alejandra, Mari, Juan. Gracias por darme un pedacito de vuestros corazones y creer en mí.

A toda mi familia, por creer en mí, no sabéis como de agradecido estoy, especialmente a mi tía Lola, gracias a ti empecé este doctorado. Os quiero.

A mis padres, porque este camino sin vosotros no hubiera sido posible, me habéis cuidado y seguís haciéndolo a mis 32 años. Cuanto lloré cuando me separé de vosotros, pero aquí me tenéis otra vez, no os libraréis de mí tan fácilmente. Mama, Papa, os quiero. Mama, Wo Ai Ni.

A mis hermanos, David y Carolina. Que orgulloso estoy de vosotros. Gracias por todo el apoyo incondicional que he recibido y que sigo recibiendo día a día. Que suerte tener esta familia tan bonita. Si soy así hoy en día, sin duda alguna vosotros tenéis gran parte de culpa. Os estimo tantíssim.

A mis sobrinos Dídac y Gael. Sois pequeñitos aún pero no hacen falta las palabras para sentir el amor incondicional que me dais. Os amo.

A mis abuelas y abuelos, especialmente ellas, Teresa y Maruja. Ellas son y han sido siempre mis ángeles de la guarda. Desde pequeñito me han mimado y cuidado como solo las abuelas saben hacerlo. Por creer en mi cuando nadie lo hacía y por seguir haciéndolo. Gran parte de la

persona que soy ahora es mérito vuestro. Sois mi vida entera. Os quiero muchísimo.

Mil gracias a Thomas, contigo empecé el camino de mi doctorado y el camino de ser completamente libre sin tener que vivir ocultando quién realmente soy. Aunque ahora ya no estemos más juntos, eres y serás una de las personas más importantes de mi vida entera. Me has apoyado en todo, sacándome una sonrisa cuando yo realmente estaba ahogado, me has ayudado en lo máximo posible sin pedir nada a cambio. He aprendido tanto de ti, que ni te lo imaginas. Personas tan bonitas como tú, son muy difíciles de olvidar.

A Gaëlle, mi Kitty, por aparecer en mi vida hace 4 años. Otra de las personas que me han ayudado a creer más en mí. Me has apoyado en los momentos difíciles de mi vida. Siempre estaré agradecido por todo lo que has hecho por mí. Brindemos por Jaume Serra y por Barcelona.

A mis compis de gimnasia artística. Que ilusión poder empezar un deporte que siempre has querido hacer desde pequeñito cuando eres mayor, y disfrutarlo conociendo a personas tan bonitas como vosotrxs. Clara, Adri, Eva, Adrià, Nacho, Javi, Diego, Zazil, Laura, Noa, Jordi, Bernat, Andrés, y muchos más que van y vienen. Gracias por todo. Este

deporte y vosotros me habéis dado luz, donde había oscuridad.

Gracias a Remi Cuchillo e Irene Cadavid por aparecer en mi vida, sois tan especiales. Remi, lo que unió JEDI (bueno y más cosas jajaja) que no lo separe nadie. Irene, que risas nos hemos pegado y cuantas historias hemos creado.

Thanks to Harris and Pattama Wapeesittipan. I love you guys for a thousand years more. Pattama, I am truly grateful to have met you. Thanks for sharing your thai food knowledge with me, but also for being so humble and good to me.

Thanks to George Gerogiokas for being so entropic. I learnt a lot from you. It is a gift to have someone in your life who is smiling always. But also, to have someone who is sharing his life with one of the most important persons in my life. Thanks, thanks, thanks.

A mi mejor amiga Maica Llavero, mi ejemplo a seguir. ¿Hay algo que te inquieta, te atormenta o te perturba? No te quedes atada a tu dolor, a tu miedo, a tu incertidumbre. Gracias por todo lo que me has dado y me sigues dando. ¿Quién nos iba a decir que ese viaje a Edimburgo nos iba a marcar tanto? Las experiencias vividas junto a ti son inolvidables, pero si realmente he apostado por esto, sin duda alguna ha sido por ti. De todas las personas que he

conocido en mi vida, sin duda alguna tú eres una de las más especiales. Eres mi Natural Killer, ese linfocito tan particular que puede reconocer lo malo donde los otros no pueden. Una vez me escribiste, serás feliz haciendo lo que hagas, y te digo que eso es lo que intento y seguiré intentando hacer. Como tú un día me prometiste, yo te prometo que nunca seré de esas personas que se olvide de ti, porque olvidarte es imposible. Te quiero y te requiero.

Gracias a la música, tú, fiel compañera que me has acompañado siempre en los momentos tristes y alegres. Me has y sigues inspirando tanto.

A mis gatines, Tiger & Rayban, siempre os llevaré dentro de mi corazón, no os olvido no. Stewie y niña, sois tremendos, cuanto cariño dais. A mi primera gata Nina, hace ya unos cuantos años que ya no estás con nosotros, pero me has dado tantas alegrías. A mi gato Toni Elías, eres el más especial. Un día desapareciste de mi vida sin rastro y sin saber si estás bien. Ni un día he dejado de pensar en ti. Me diste lo mejor y lo peor en este mundo entero. A ti, allá donde estés, en los recovecos de mi corazón siempre estarás. Nunca has dejado de existir para mí.

Gracias a Edimburgo, Singapur, pero sobre todo a Barcelona por todo lo que me habéis dado. Cuanta libertad, seguridad, cultura, experiencias, … Gracias.

Por último, quiero dar las gracias a todas esas personas que he ido conociendo por este camino que se llama vida.

“El futuro tiene muchos nombres. Para los débiles es lo inalcanzable. Para los temerosos, lo desconocido. Para los valientes es la oportunidad.” Victor Hugo

Gracias, Gracias y tres veces gracias.

Abstract

Despite the investment of vast amounts of money over the last decades, drug discovery and development remains an inefficient process, which can be stopped at different steps, leading to the loss of all resources invested. For this reason, there is an urgent need to develop methods for relating chemical structural information and *in vitro* bioactivity to toxicity outcomes in early drug development stages.

This thesis describes novel *in silico* prediction methods, using novel similarity metrics and prediction tools adapted to the Chemical Safety Assessment (CSA) of drugs. Their use is illustrated by their application to liver toxicity endpoints. The proposed approach involves five steps: (1) Data collection, (2) best similarity metrics identification, (3) read across similarity validation, (4) QSAR modelling, and (5) implementation.

Resumen

A pesar de las inmensas cantidades de dinero invertidas en las últimas décadas, el descubrimiento y desarrollo de nuevos fármacos sigue siendo un proceso ineficiente, que puede detenerse en diferentes fases, implicando una pérdida importante de todos los recursos invertidos. Por esta razón, existe una necesidad muy urgente de desarrollar métodos que relacionen la información estructural química y la bioactividad *in vitro* con los resultados de toxicidad en las primeras etapas de desarrollo de fármacos.

Esta tesis describe nuevos métodos de predicción *in silico*, utilizando nuevas métricas de semejanza y herramientas de predicción adaptadas a la Evaluación de Seguridad Química de medicamentos. Su uso se caracteriza por su aplicación en toxicidad hepática. El enfoque propuesto implica cinco pasos: (1) Recopilación de datos, (2) identificación de métricas de mejor semejanza, (3) validación de la semejanza usando Read across, (4) modelado QSAR e (5) implementación.

Keywords

Drug development, safety assessment, hepatotoxicity, chemical similarity, reproducibility, DILI

Diseño de fármacos, evaluación de la seguridad, hepatotoxicidad, semejanza química, reproducibilidad, DILI

Preface

The present work describes the development of new tools for predicting toxicity of drug candidates starting from pre-existing information. Drug Induced Liver Injury (DILI) was used as an example through the manuscript to illustrate their application. These methods can be used to obtain i) similar compounds by similarity searching and ii) building better predictive models for *in silico* toxicology that could help the expert toxicologist identify DILI in early drug development phases. Our methods do not aim to replace but to reduce the use of animal models to prioritize candidates, to remove those which could lead to DILI. This work was developed in close collaboration with the European projects iPiE and eTRANSAFE, in which both academia and pharmaceutical companies are deeply involved. The methods developed could be used and applied in the future with a real impact in the pharmaceutical industry, contributing to the 3R principles (replacement, reduction and refinement), helping to reduce animal testing and costs as well as speeding up the safety assessment procedures.

Table of contents

1. INTRODUCTION

1.1. Drug Development Process

Where do drugs come from? The path leading to the creation of new drugs is long, risky, and complicated. It involves the analysis of available chemical and biological data suggesting new compounds likely to possess the required properties of safety and efficacy [1].

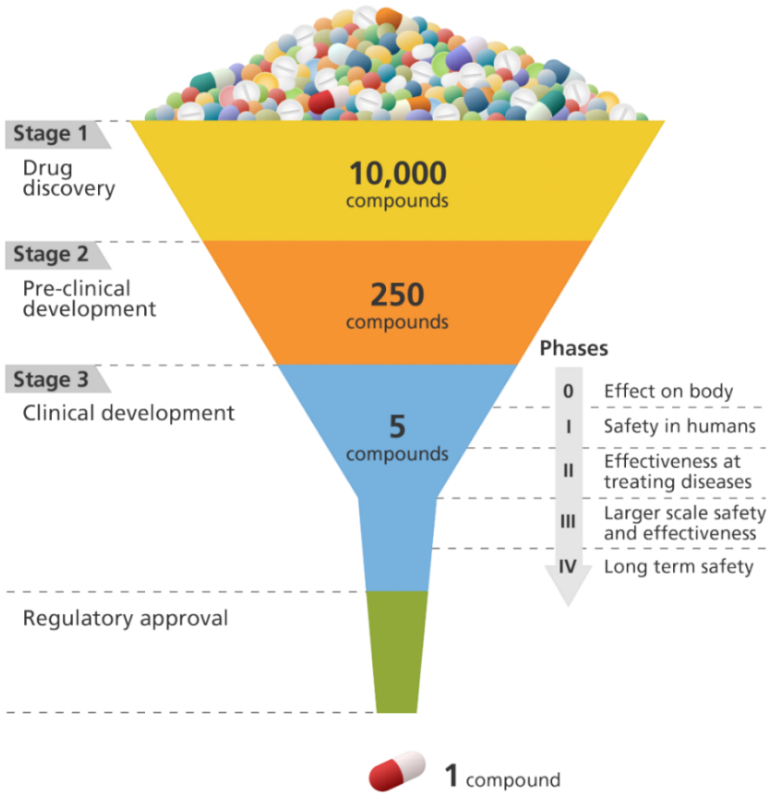


Fig. 1 **| The Drug Development Timeline**, extracted from [2] , indicates how many compounds are tested at different stages of the pipeline to yield on average a single approved drug.

In a biomedical research context, *in silico* computational models are often applied alongside experimental techniques to predict biological properties of chemical compounds [3]. Modern information technologies have made big data available in safety sciences, i.e., extremely large data sets that may be analyzed only computationally to reveal patterns, trends and associations [4]. Models obtained by such means can be used for predicting, rationalizing and estimating physico-chemical properties of molecules and their interactions with macromolecules, thereby allowing a more rational approach to drug development [5] as shown in Figures 1-2.

Given the involvement of a particular macromolecule in a human disease, its action can be modulated with a small organic molecule so as to obtain a therapeutic effect. Once a macromolecular target related to a certain disease is identified, small molecule binders (hits) for that particular target are found and optimized (hit to lead) [6]. Promising molecules are subjected to additional assays *in vitro* and *in vivo* to collect efficacy, toxicity and pharmacokinetic data [3]. The next stage involves clinical trials in humans. Phase I is designed to test the molecule on healthy people to determine whether it is safe [7,8]. Phase II takes into account the efficacy and safety of a molecule [9]. Finally, Phase III is focused in evaluating efficacy, effectiveness and safety to determine if a molecule has a therapeutic effect [6].

Ultimately, a molecule that clears all these hurdles had to be approved by regulatory agencies like the Food and Drug Administration (FDA) or European Medicinal Agency (EMA) before it can be commercialized [9]. Later on, there is a post-marketing surveillance, a pharmacovigilance phase for monitoring overlooked adverse effects or long-term effects associated to the drug usage on the population [10].

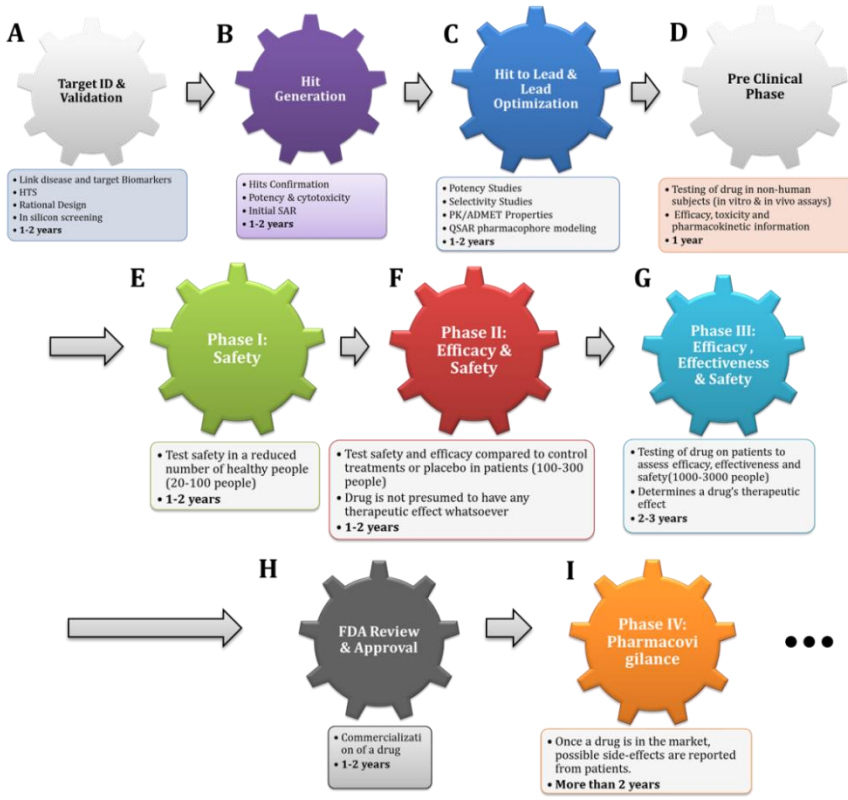


Fig. 2 | **The Drug Discovery pipeline**. The Drug Discovery process involves steps A to C. Clinical trials involves steps E to G. After step H a drug is commercialized but can be removed from the market if unexpected side-effects are reported. Extracted and modified from [11]

1.2. Overview of Toxicology

In the drug development process, safety assessment plays a central role. The therapeutic usefulness of new candidates depends on their safety as much as on efficacy. For this reason, toxicological evaluation is applied at different steps. It is important to define different aspects about toxicology.

Toxicology is the study of the adverse effects of chemicals or physical agents on living organisms, in our case, on humans. A toxicologist is trained to examine and judge the nature of those effects on human health [12]. Then, a toxicological research plays an important role examining the behaviour of chemicals inside the human/animal body. For example, to look at cellular, biochemical and molecular mechanisms of action assess the probability of their occurrence [12]. More and more, our society is increasingly dependent on chemicals. Toxicologists are an important part of decision-making processes for evaluating potential hazards. The main toxicologist's duties are descriptive, mechanistic and regulatory, all vitally important for chemical risk assessment [12].

1.3. Safety Assessment

Safety plays a critical role in the above-mentioned drug development process since the therapeutic usefulness of new

candidates depends on their safety as much as on efficacy. Safety testing is one of the main concerns in the drug development process, since safety liabilities can produce that a candidate fails at any stage, thus producing an enormous waste of time and money.

For instance, the selectivity of a ligand for a target is an important early consideration. Promiscuous ligands are more frequently associated with adverse drug reactions and clinical trials failures and also the well-known “anti-targets” (e.g. hERG), where compounds that binds to a different target to the specific one. Some selectivity within a family of related proteins is often desirable although exquisite selectivity is rarely attained. Hence, safety is continuously evaluated at all stages of drug development. But a more dramatic scenario would be discovering a toxic effect during clinical and pharmacovigilance phases having a huge impact with severe adverse effects on patients, and huge amount of money losses. This was the case of Thalidomide, withdrawn for causing severe birth defects. This must be avoided as much as possible. Therefore, a huge effort of safety assessments closer to the initial drug development stages is taken seriously with the aim to fail early and cheap.

Recently, there has been a change of paradigm in the field of toxicity from a focus on observations to the mechanistic understanding of the toxicity. One of the consequences of this

improved understanding is the possibility of predicting the toxicity of new compounds, from different chemical or biological properties of the compounds instead of merely observing them. These predictive models require a better understanding of the toxicity mechanism. Predicting mechanisms plus predicting potential targets can suggest ways to test *in vitro*, being a good way to guide the safety assessment testing strategy. This does not aim to replace completely the animal model but used to prioritize candidates, to remove those which could cause toxicity at early stages of development. In this work we provide new approaches to tackling predictive toxicology that could help toxicologist to answer questions to be able to prioritize candidates but also to capture better the mechanisms of toxicity.

Animal testing has been used to evaluate the toxicity of the individual chemical ingredients such as acute toxicity, skin/eye irritancy, potential for skin sensitization, and so on [4,13–18].

Safety assessment is important for consumer chemicals as well as substances used in cosmetics and food industry. The Registration, Evaluation, Authorization and Restriction of Chemicals Regulation (REACH) [4,13–18] regulation initiative, since March 2013 [19] does not allow commercialization of EU cosmetics using ingredients tested on animals. REACH has been dealing with finding alternative

methods to safety studies in experimental animals [20]. In this scenario, *in silico* methods must be considered a valuable tool for the application of the well-known 3R principles (replacement, reduction and refinement), helping reducing animal testing and costs as well as to speed up the safety assessment procedures. These *in silico* methods basically depends on the availability and quality of the data provided [20] and there is a need for improving and optimising *in silico* methods.

1.4. Drug-Induced Liver Injury (DILI) in Safety Assessment

Drug-induced Liver Injury (DILI) is the main cause of liver disfunction which may lead from mild non-specific symptoms to more severe signs like hepatitis, cholestasis, cirrhosis and jaundice [21]. Among the adverse drug reactions (ADRs) causing drug attrition in clinicals trials or drug withdrawal from the market, DILI plays a major role [22–24]. Early identification of DILI is beneficial for both public institutions and the pharma industry. However, the lack of translatability from *in vitro* and animal models on one side, and the poor performance of *in silico* models when applied in the real world on the other side, makes early DILI identification difficult to achieve [25,26]. Animal models cannot capture all aspects of human physiology represented by the different mechanisms of action causing DILI [27–29]. On the other hand, despite *in*

vitro models on human cells or tissues, pharmacodynamic and pharmacokinetic (PD/PK) aspects are not considered, relying on solutions such as 3D organs and PBPK (Physiologically based pharmacokinetic) modelling to approximate and predict their effects in drug metabolism [30,31].

1.5. *In Silico* Modelling in DILI Safety Assessment

In silico models are a cheap and fast tool to assess DILI in early stages of drug discovery, and can be applied to any compound, including virtual ones [32,33]. In general, *in silico* methods can be classified in statistical-based, knowledge-based and structure-based approaches [32,34,35].

Knowledge-based models look for structural alerts in molecules that were previously identified to produce an adverse outcome. Structure-based (often called ligand-based) models use only the structure of small organic compounds or the 3D structure of macromolecules to predict for example binding affinities [36]. In contrast, statistical-based models can predict the DILI outcome of an unknown compound by fitting a series of molecules with known activity to a molecular representation of the molecules, typically chemical descriptors or fingerprints. Reported models generally are built using QSAR. This method starts by training

a machine-learning (ML) algorithm (see section 1.7.3.) with positive and negative samples represented structurally by a set of descriptors or fingerprints. A few of these models extend their molecular representations by adding target activity information, which resulted in a slight improvement [37,38]. While knowledge-based models offer a clearer insight into the mechanism of action, they suffer from lower performance due to the lack of sensitivity (see section 3.6.3.) inherent to the fact that DILI assessment is limited to alerts defined by the toxicologist [39].

In general, the performance of statistical models is higher, but depends considerably on both the quantity and quality of the data [40].

1.6. Chemical and Biological Similarity

Several safety assessment methods are applied in the drug development process playing an important role to label candidates as toxic/not toxic, and one of them is by chemical similarity.

Similarity, or the state of being similar, is a concept that has been pursued in many fields, such as mathematics, computing, linguistics, music, psychology, chemistry, biology, etc. In our case, we are interested in chemical and biological similarity. Many sources suggested several types of structural

representations to measure the similarity between two molecules [41]. Similar compounds tend to have similar properties [42], this is the rational basis of the bioisosterism concept, which justifies which type of similarity is relevant in a biomedical context. Bioisosteric compounds have similar biological properties and therefore, the similarity metrics which are relevant represent this bioisosterism. Bioisosters could be chemical substituents or groups with similar physical or chemical properties which produce similar biological properties to another chemical compound. There are classical and non-classical bioisosters (see Figure 3) [43]. In drug design, they are used to enhance the desired biological or physical properties of a compound.

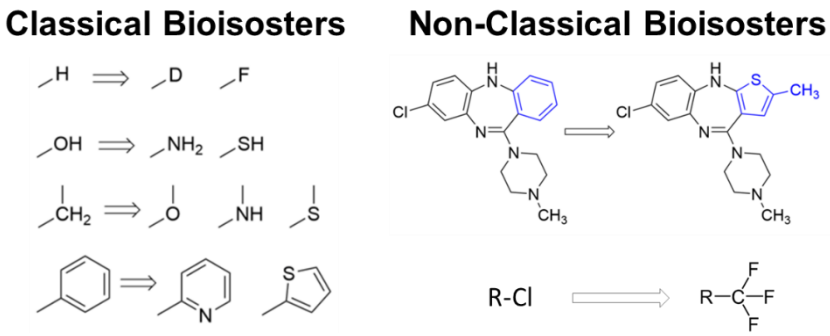


Fig. 3 | **Classical versus non-classical bioisosters** [44,45].

The evaluation of similarity requires identifying source chemicals for which data is available and that are similar to a target chemical for which no data are available. These predictions can be made to fill data gaps. As the Nobel Prize-winning pharmacologist James Black advocated in 1988, *"the*

most fruitful basis for the discovery of a new drug is to start with an old drug" [46]. For instance, sildenafil (Viagra) and vardenafil (Levitra) both are PDE5 inhibitors with similar medical uses, i.e. drugs that are indicated against erectile disfunction [47]. However, if we perform a similarity substructure search by SMILES arbitrary target specification (SMARTS) [48], a language that allows you specifying substructures with rules that are straightforward extensions of SMILES [49], it would miss vardenafil on the pyrazolopyrimidine ring of sildenafil [47].

But how can a computer state if two molecules are similar? There are several methods for computing similarity that are relatively trivial and can be coded computationally nowadays [52]. Those are implemented in most of the existing toxicological databases, such as AIM, AMBIT, CBRA, CIIP, QSAR Toolbox, Toxmatch, Toxread, which are described in detail in [50].

The most widely used similarity methods use molecular fingerprints [52]. They map a substructure onto a binary string [51]. This can be performed using a key-based approach. They can ask binary questions with yes/no answers about the contents of a molecule such as: Does it have polar or non-polar groups? Is there an aromatic ring? Is there any hydrophobic/hydrophilic interaction? Depending on the answer a 1 (for yes) or a 0 (for no) at a specific position in the

fingerprint will be assigned. Useful key-based fingerprints are often several hundred bits long [52]. One of the drawbacks is that most molecules contain few of the substructures defined by the key, so most bits are set to 0 meaning most fingerprints will be inefficient sparse vectors containing many zeros [52].

Alternatively, people use hash-based fingerprinting methods [52]. Briefly, they represent molecules with SMILES [49] of unique patterns (bonds, atoms, etc). Then a hash function is fed with every pattern returning integers between 1 to N. The hash value of a pattern is then used to set a bit in the fingerprint [52]. One of the benefits of hash-based fingerprints is that they contain a lot of structural information in a small key size, hence they are faster at comparing fingerprints leading to quicker computations in comparison with key-based fingerprints [52].

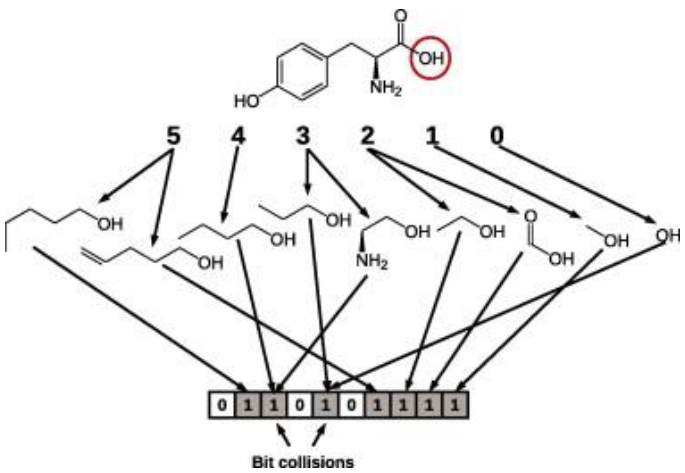


Fig. 4 | **Binary fingerprints representation of a molecule**. Extracted from [52].

These fingerprints are said to be too ‘black’ because bit collisions are permitted, and they occur when two different patterns have the same hash value (see Figure 4). Excessive bit collisions decrease the information content [52].

1.7. Statistical Models in Computational Toxicology

Because of the advances in both data quality and machine-learning techniques, statistical-based models have become of great interest recently. Many QSAR models for DILI prediction have been reported using different datasets and techniques [29,53–61]. In general, balanced models based on datasets with hundreds of compounds reach accuracies around 65%. Other models with better accuracies are either unbalanced in their sensitivity-specificity (see section 3.6.3.) or built with small datasets. This lack of predictability can be attributed to two principal factors; (1) the various mechanisms causing DILI [62], and (2) the fact that the classification of compounds as DILI positives and negatives fulfils different criteria depending on the dataset/study, and these criteria sometimes rely on statistical analysis of reports obtained from literature mining. In this regard, efforts to create high quality DILI datasets have been performed during the past years.

Remarkably, DILIrank [63] compiles 1036 compounds ranked by DILI risk in humans with confirmed causal evidence. Mulliner et al. [59] created a database compiling compounds from different sources including clinical, postmarked, and preclinical data. While DILI assignment is not as rigorous as in DILIrank, it comprises 3712 compounds that cover a huge chemical space. Also, as different species are considered, this dataset is appropriate to study/model translational aspects of DILI.

As mentioned previously, one of the challenges in DILI modelling is the diversity of mechanisms leading to hepatotoxicity. To overcome this problem, Liu et al. [55] built hepatic adverse effects QSAR models using the SIDER database [64], which contains a compilation of adverse effects caused by commercial drugs. These models were validated against databases tagged with DILI activity, reaching ~65% of accuracy in external validation. Mulliner et al. classified hepatotoxic compounds into three hierarchical levels from general hepatotoxicity at the first level, to more detailed endpoints discriminating clinical chemistry and morphological findings at the second level, and hepatocellular and hepatobiliary findings at the third level. Subsequently, QSAR models were built for each endpoint obtaining reasonable performance for some of them (~65% of accuracy).

Another key element when applying QSAR models in the real world is the applicability domain (AD) [65]. In simple words, AD techniques call for the capability of a QSAR model to return a reliable prediction for a given instance. AD determination is not used in the majority of the models reported, and when used, it is based on the descriptor distribution not only limiting the chemical space, but also subject to ambiguous interpretation [65,66]. AD is important for assessing where models can be applied with confidence.

1.7.1. Read Across

In the read-across approach [50,67–74], endpoint information for one chemical (the source chemical) is used to infer the same endpoint for another chemical (the target chemical), which is considered to be "similar" in some way (see *section 1.5.*). In principle, read across (RA) can be used to assess physicochemical properties, toxicity, environmental fate and ecotoxicity. For any of these endpoints, it may be performed in a qualitative or quantitative manner [43] as described in Figure 5. There are many RA databases [75–89].

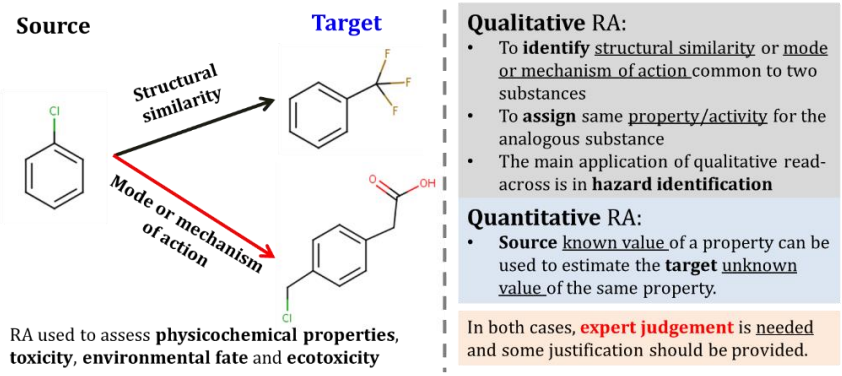


Fig. 5 **| The Read-Across approach** [43].

To compute chemical similarity computationally, there are various tools available. In Python, there is the well-known open source RDKit tool [90–92]. This tool can represent structures and measure the similarity between molecules using features such as Morgan or pharmacophore fingerprints, but also physiochemical properties among others [52]. This toolkit utilizes different metrics for quantifying similarity between two structural representations, e.g., Tanimoto coefficient, cosine coefficient, Tversky index, etc. [93] which can be useful for different types of models.

The most widely used metric to evaluate similarity of two fingerprints, *a* and *b* is the Tanimoto coefficient called Jaccard elsewhere (see equation 1, and for more information see Table 3 in Methods).

$$J_{A,B} = \frac{c}{(a + b - c)} \qquad \text{Eq. 1}$$

The maximum value is 1 when all the bits set in *a* are set in *b*. When no bits are shared between *a* and b, 0 value is assigned. J values of 0.1-0.4 means there is no relationship between them. In general, it is considered that J values over 0.7 for validated keys or hash functions are significantly similar [52].

1.7.2. QSAR methodology

Quantitative Structure-Activity Relationship (QSAR) models aim to infer knowledge analyzing the relationship between the structure of compounds and their biological properties. Descriptors are normally computed from the chemical structure (1D, 2D) or a conformation (3D). QSAR models [54,85] are an alternative to using chemical/biological similarity searches to find compounds having similar properties/descriptors to a real molecule such as a drug which is already in the market. QSAR models use machine learning methods as the ones described above [54,85]. As can be seen from Figure 6, correlation does not imply causation. Thus, correlated properties correlating well with a property of interest, fail to be linked. Besides, we need expert judgement to build reliable models.

Fig. 6 | **The QSAR Fallacy.** Figure extracted from [94].

When building QSAR models, the original dataset is often randomly split into a training set and a validation set. We use cross validation methods, typically Leave-One-Out, to make internal validation of the model and build our applicability domain. Finally, an external validation is performed with the validation set to test if the model performs well.

1.7.3. Machine Learning

In Machine Learning (ML), computer programs have the ability of automatically learning and improving from experience without the necessity of being explicitly programmed [95]. ML is present in artificial intelligence as well as data mining and is getting more and more extensively used. ML methods are categorized as supervised or unsupervised. On the one hand, supervised methods use

labelled information to make predictions. Usually, they work for both qualitative and quantitative using classifier and regressor estimators, respectively. On the other hand, the unsupervised ones use unclassified or unlabelled training data. The most widely used techniques in computational toxicology are:

- Decision Trees: Unsupervised ML method using a tree approach where each branch represents the results of the test, and each leaf the class label, after computing all attributes [96].

- K-nearest neighbours (KNN): unsupervised ML method which uses as an input the k closest source values near the labelled point of interest, and for classification the most frequent label is used among the k points [95].

- Principal Component Analysis (PCA): Unsupervised ML method that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables, the so-called principal components [97].

- Partial Least Squares (PLS): supervised ML method, PLS regression and PLS discriminant

analysis (PLS-DA) where the Y is quantitative or categorical, respectively. PLS is a statistical method that relates to principal components regression, and projects the predicted and the observable variables into a new space [98].

- Random Forest (RF): supervised ML method, an ensemble learning method for classification, regression that builds a multitude of decision trees at training time. The output is for a classification or a mean prediction of the individual trees for qualitative and quantitative method respectively [99].

In this work we are going to use some of the statistical methods to predict DILI. PLS for supervised RA in a similarity context (see results sections *4.2.* and *4.3*), and RF and PLS in our AE QSAR models (see results section *4.4.*)

1.8. Reproducibility

In computational toxicology, the concept of reproducibility is increasingly used. Because of we are dealing with virtual data, is “easier” to obtain reproducible results, e.g. running a model which gives the same result every time perform the analysis, in other words, it can be easier controlled than when you are

doing an experiment with animals which can vary the result for many external and internal factors.

Reproducibility is the agreement among the replicate results carried out with the same methodology, for example in different locations by different people [100].

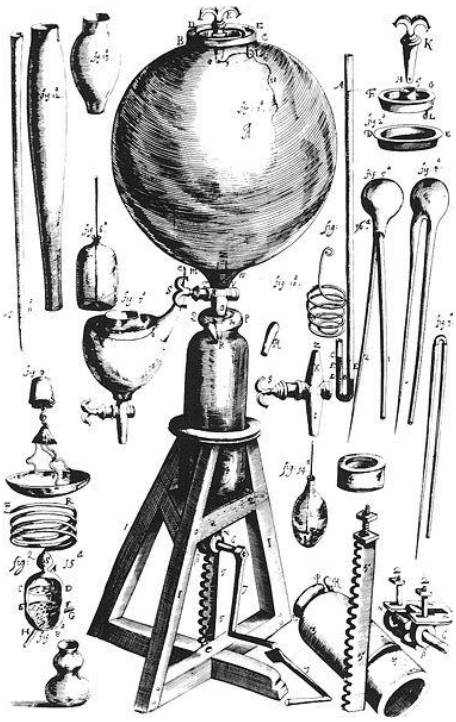


Fig. 7 | **Reproducibility**. Boyle's air pump was, in terms of the 17th century, a complicated and expensive scientific apparatus, making reproducibility of results difficult. Extracted from [101].

If we go back to the 17th century in England, the reputed Irish scientific Robert Boyle, published the air pump design to generate and study the vacuum (see Figure 7). Boyle was the

first one to stress the importance of reproducibility by repeating the same experiment repeatedly so that reproducible data can be credible to the scientific community.

Nowadays, irreproducible research is a challenge. In October 2018, Nature's International Journal of Science published [102]:

"*Science moves forward by corroboration when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study.*

There is growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all have a part in tackling reproducibility. Nature has taken substantive steps to improve the transparency and robustness in what we publish, and to promote awareness within the scientific community."

Indeed, reproducibility is important to be taken into account and as the Nature article above mentions, we need to move forward by corroboration, verifying other's results. To do that, we need to publish successes and failures. To make

reproducible research example would be the laboratory notebooks, either paper or electronic ones, codes, data, and well-organized text files need to exist.

This thesis tends to be as much reproducible as possible, and for this reason, we developed everything in Jupyter Notebooks with a specific Conda environment to be installed and run wherever, allowing the end-user obtaining reproducible results.

2. OBJECTIVES

The general objective of the thesis is:

- Develop novel similarity-based tools (RA and QSAR) adapted to the Chemical Safety Assessment (CSA) of drugs.

The specific objectives are:

- Collect datasets representing the chemical space typically covered in drug development annotated with a representative toxicity endpoint (DILI)
- Identify the most suitable descriptors and metrics, benchmarking their performance in CSA related applications
- Validate the best similarity metrics obtained, performing a systematic comparison with a collection of different similarity descriptors and using RA for predicting DILI
- Develop new QSAR modelling strategies, combining chemical and existing knowledge to predict DILI.
- Implement the RA and QSAR modelling in data-processing software tools (Jupyter notebooks) suitable for applications in both academic and industrial environments for reproducibility purposes

3. METHODS

3.1. Software

To facilitate the work and data processing done in this thesis, a **Conda** [103] **environment** with all dependencies and extra packages (see a summary on Table 1) was created. Conda environment is a directory containing a collection of Conda packages installed. This environment can be downloaded and installed following the instructions written in [104]. We created this Conda environment because we want to work in a place where we can get reproducible results controlling the versions of the programs used. The operating system used was the Scientific Linux 64-bit (CentOS 7). A collection of jupyter notebooks for similarity searching and QSAR modelling were also created [104].

Table 1 | Python and external packages loaded within the Conda environment.

Software	Version	Reference
Python	3.6.7	[105]
Pandas	0.24.2	[106]
Scikit-learn	0.20.3	[107]
RDKit	2018.09.2.0	[90]
nonconformist	1.2.5	[108]
scipy	1.0.0	[109]
numpy	1.16.2	[110]

Software	Version	Reference
Standardiser	2014	[111]
molVS	0.1.1	[112]
moka	3.0	[113,114]
Flame	0.2	[115]

3.2. Structure Retrieval and Curation Procedure

As can be seen from Figure 8, compound structures are retrieved using an in-house workflow involving the following steps:

A. Structure resolver [116]: from chemical identifiers (CAS RN, chemical name, DrugBank ID, etc) a structure is retrieved as InChI.

B. Two-dimensional structures are computed using RDKit [90].

C. Structures are canonicalized using molVS [112] and then, neutralized and standardized using standardizer [111].

D. Reasonable three-dimensional structures are obtained using RDKit [90] ETKDG. The ionization state of every compound is adjusted to pH 7.4 using MoKa [113,114].

E. For the resulting structures parent InChIs and InChI keys (standard and non-standard) plus parent SMILES are computed. Duplicates are checked and dropped by using the non-standard InChI key as identifier keeping

the information of dropped molecules where differences are found.

F. Finally, molecules are saved in machine-readable (Python pickle of RDKit mols) [117], SDFile, and human readable SMILES Tables (tabular, TSV).

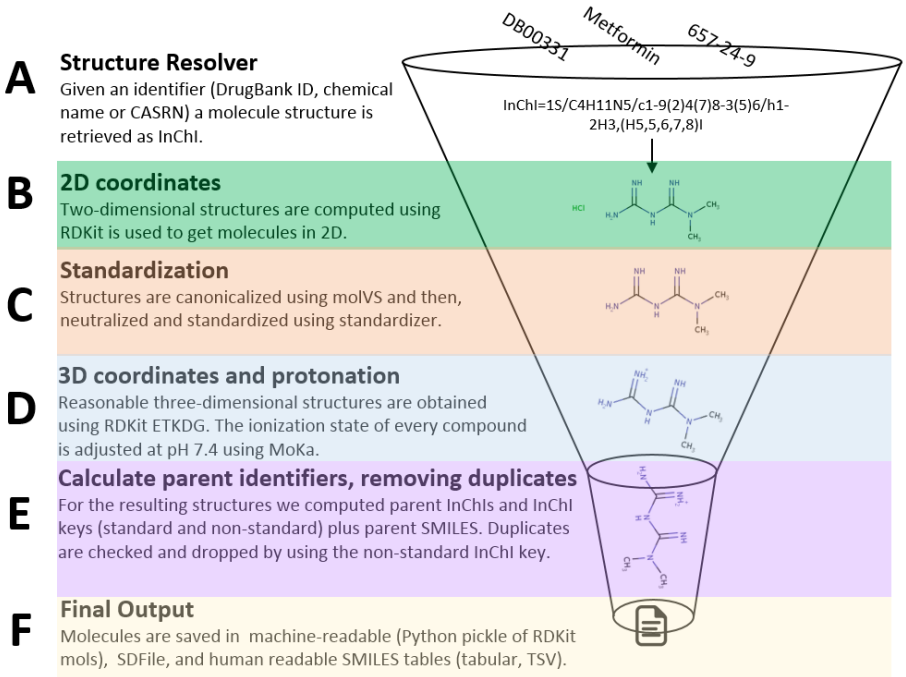


Fig. 8 | **Normalization protocol.**

3.3. Data Extraction and Normalization

ATC dataset. The ATC dataset [118] contains marketed drugs (active substances) classified in a hierarchy with five different levels [119]: 1st) fourteen anatomical/pharmacological, 2nd) pharmacological or

therapeutic groups, 3rd and 4th) chemical, pharmacological or therapeutic subgroups and 5th) the chemical substance. Then, the 5th level was used to obtain 3D protonated structures.

To do so, the general normalization protocol described in methods *section 3.2.* was applied to the ATC dataset. This dataset was used for different purposes: i) Merging structures from external datasets by ATC code or by other similarity method such as Jaccard or by another identifier (parent non-standard InChI Key), and ii) clustering compounds by ATC Ontology.

Table 2 | Effect of the normalization workflow on the size of the ATC dataset.

Normalization Step	**Molecules**
Original	4580
2D	2151
standardised	2128
3D	2126
Protonated 7.4	2126
Final	2126

2126 out of 4580 structures with ATC code were obtained. In the whole process, mixtures we removed, for instance, N02AJ06 corresponding to paracetamol plus codeine. In

addition, macromolecules such as peptides, proteins (e.g. L03AC01 corresponding to the interleukin Aldesleukin) were also eliminated. We finally saved in machine-readable formats (Python pickle of RDKit mols), SDFile, and human readable SMILES tables (tabular, TSV), the ATC dataset (4580 molecules) containing 2126 molecules with structures and 2454 molecules without structures.

SIDER dataset. SIDER [120–122] is a database which contains information about adverse drug reactions of marketed drugs. SIDER connects 1430 drugs with 5868 different reported adverse effect (AE) terms. Last version was downloaded (4.1, October 21, 2015). It was used to extract compounds with clinical annotations on hepatic adverse effects. MedDRA ontology [123,124] was applied to SIDER compounds with associated Preferred Terms (PT) and filtered by the primary Organ Class (SOC) level of Hepatobiliary disease. 568 compounds in total show at least one hepatotoxic related AE, while 369 of the remaining compounds were free of any hepatotoxic adverse effect. Finally, only AEs present in at least 40 compounds were considered to build QSAR models (see Figure 9).

This threshold on the number of compounds was defined arbitrarily as the minimum number of compounds needed to obtain reliable QSAR models, resulting in a total of 19 AE QSAR models. Building QSAR models requires datasets

containing both active and inactive compounds, therefore, for a given AE, negatives were picked from other compounds showing any other hepatobiliary AE (to strength selectivity) and from other SIDER compounds not showing any hepatobiliary AE (to increase chemical diversity).

These structures of SIDER compounds were generated from those on the ATC dataset, using the ATC code as the index. This dataset is accessible in machine-readable formats (Python pickle of RDKit mols), SDFile, and human readable SMILES Tables (tabular, TSV).

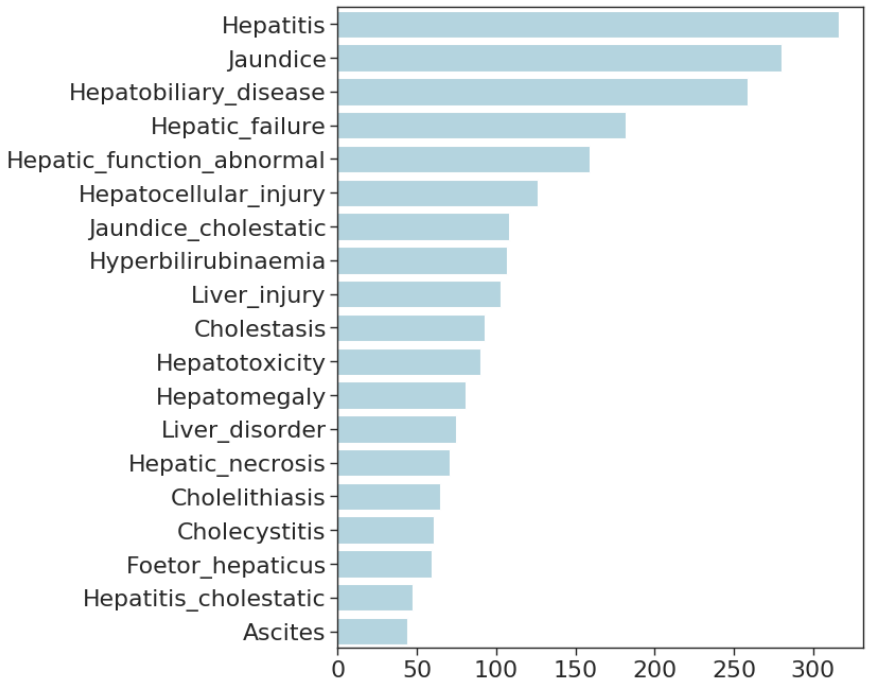


Fig. 9 | **Adverse Effects Distribution.** AEs distribution in SIDER obtained by selecting those with a Hepatobiliary Disorders, as defined by the MedDRA ontology. A threshold of 40 compounds is arbitrary selected as

the minimum number of compounds to obtain reliable models, resulting in a total of 19 adverse effects QSAR models. y axis are the AEs. x axis is the total number of compounds where the AE appears.

Mulliner dataset. Mulliner dataset was extracted from [59] containing 3712 compounds with preclinical and clinical DILI annotations. We filtered this dataset removing compounds without human clinical annotations. 2172 out of 3712 compounds available in the dataset were labelled as human (H_HT endpoint) DILI positive.

DILIrank dataset [63]. DILIrank is an open access project at the FDA's National Centre for Toxicological Research to study DILI. It contains 1036 compounds labelled with different DILI concern levels: i) 192 vMost-DILI-concern, ii) 278 vLess-DILI-concern, iii) 312 v-No-DILI-concern, and iv) 254 Ambiguous-DILI-concern. Drugs with an ambiguous call label were removed.

Pfizer dataset. The Pfizer dataset contains 626 compounds with different DILI annotations extracted from [125]. From the original Pfizer dataset, only those with evidence of human hepatotoxicity (HH label) (273 compounds), and those with no evidence of human hepatotoxicity (NE label) (152 compounds) were considered in this work.

Among these compounds, 16 compounds had same CASRN with a different name and 1 compound with both same

CASRN and name. Those molecules were manually checked. We fixed incorrect CASRN such as Naltrexone and Methotrexate which were clearly different molecules. We removed molecules with different hepatotoxicity label and duplicates. Finally, we obtained 407 out of 626 molecules from the original dataset.

We used our ad-hoc tool made within the group to resolve the structure by CAS or chemical name. 406 structures were retrieved successfully. We eliminated Auranofin with CASRN 34031-32-8 due to the fact that our tool could not retrieve the structure and it contained gold metal which is not treatable with our methods.

Structures were canonicalized using molVS and then neutralized and standardized using standardizer. 35 entries did not pass the protocol because they contained more than one molecule. After a manual check, 6 molecules were eliminated, cyanocobalamine, ammonium chloride, sodium chloride, FeSO4, arsenic trioxide and cisplatin. Then, two-dimensional structures (395) were computed using RDKit. Next, reasonable three-dimensional structures (395) were obtained using RDKit. The ionization state of every compound was adjusted to pH 7.4 using MoKa. Moreover, parent InChIs and InChI keys (standard and non-standard) plus parent smiles were computed.

Duplicates were checked and dropped in every step by using the non-standard InChI key as identifier keeping the information of dropped molecules where differences were found. 4 duplicates were identified in 2D computation step, such as streptomycin sulphate and streptomycin. Finally, 395 molecules were saved in machine-readable (Python pickle of RDKit mols), SDFile, and human readable SMILES Tables (tabular, TSV).

O'Brien dataset [126]**.** The 40-compound dataset was extracted from the tables provided by [55] in supporting information section.

DrugBank database [45]**.** The latest release of DrugBank (version 5.1.3, released 2019-04-02) was downloaded. This database includes 2587 approved small molecule drugs, 1287 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6305 experimental drugs among others. In this work, only the approved small molecule drugs set was considered.

3.4. DILI Sets Toxicity Criteria

Labelling a compound as DILI positive is not an easy task. In the datasets described above, expert toxicologists use different criteria to assign DILI as positive.

DILIrank. Includes verification of DILI causality provided by experts.

Pfizer. DILI positive drugs include those withdrawn from the market mainly due to hepatotoxicity [125], those not marketed in the United States due to hepatotoxicity, those that received black box warnings from the Food and Drug Administration (FDA) due to hepatotoxicity, those that are marketed with hepatotoxicity warnings in their labels and others that had more than 10 well-known associations to liver injury.

Obrien. The criteria taken by O´Brien et al. are the following: severe human hepatotoxicity was attributed to drugs producing more than 1% frequency of increased serum Alanine Aminotransferase (ALT) plus two of either (1) jaundice, (2) more than three reports of liver failure, or (3) a black box warning. Moderate human hepatotoxicity was ascribed to drugs producing 0.1–1% frequency of increased serum ALT plus either jaundice or a label of occurrence of adverse effect. Non-toxic or minimally toxic drugs were defined as those with less than 0.1% frequency of increased ALT and associated with no clinical symptoms. The fourth category consists of drugs that were not known to be hepatotoxic but were known to have other organ toxicities [126]**.**

Mulliner. A compound was classified as positive when a finding associated with a particular endpoint and dose was reported or at least one associated endpoint at a lower level was positive. Accordingly, a compound was classified as negative if no associated findings or all lower level endpoints were negative [59].

3.5. Dataset Structure Generation

The structure retrieval and normalization procedure (see Figure 8 in section *3.2.*) was applied to the ATC dataset to obtain the final structures. Then, ATC codes were used as a key identifier for merging with the other datasets described in section *3.3.* that did not contain structures. If the ATC code was not available, the structure retrieval and normalization protocol was applied to obtain the associated structures.

3.6. Similarity Analysis

3.6.1. Similarity Metrics

Table 3 lists the metrics used in results *section 4.2.* They required bit-vector biological or chemical fingerprints (e.g. Morgan, AEs, …) or continuous chemical or biological properties (e.g. lipophilicity, solubility, molecular weight, protein binding, …) to compare chemicals. This was used for

comparing DILI toxic / DILI no toxic properties of chemicals (e.g. biological assays with a positive or negative outcomes).

Table 3 | Metrics to assess similarity between two molecules. Extracted from [93].

Metrics	**Equation for continues variables**	**Equation for dichotomous variables**
Tanimoto or Jaccard	$J_{A,B} = \frac{\left[\sum_{j=1}^{n} X_{jA} X_{jB}\right]}{\left[\sum_{j=1}^{n}(X_{jA})^2 + \sum_{j=1}^{n}(X_{jB})^2 - \sum_{j=1}^{n} X_{jA} X_{jB}\right]}$	$J_{A,B} = \frac{c}{(a+b-c)}$
Euclidean	$E_{A,B} = \left[\sum_{j=1}^{n}(X_{jA} - X_{jB})^2\right]^{1/2}$	$E_{A,B} = \sqrt[2]{a+b-2c}$
Manhattan	$M_{A,B} = \sum_{j=1}^{n}\lvert X_{jA} - X_{jB}\rvert$	$M_{A,B} = a+b-2c$
Cosine	$C_{A,B} = \left[\sum_{j=1}^{n} X_{jA} X_{jB}\right] \Big/ \left[\sum_{j=1}^{n}(X_{jA})^2 + \sum_{j=1}^{n}(X_{jB})^2\right]^{1/2}$	$C_{A,B} = \frac{c}{\sqrt[2]{a+b}}$

A and **B** for molecule **A** and **B** respectively. Xj means the j-th feature of a molecule A or B. **a** and **b** are the number of bits in molecule A and B respectively, while **c** is the number of bits that are in both molecules.

Similarity is calculated using the metrics shown in Table 3 using the following equation:

$$Sim = 1 - \frac{distance}{dmax(percentile95)} \qquad \text{Eq. 2}$$

where distance is $J_{A,B}$, $E_{A,B}$, $M_{A,B}$ and $C_{A,B}$ (see Table 3), *dmax* is the maximum distance value calculated as the percentile 95 using the reference dataset distances, in other words, the maximum distance indicating the value under which 95% of distances are observed in the reference dataset. *Sim* value will range between 0 and 1, where 1 corresponds to identical molecules whose distance is 0.

3.6.2. Read Across Similarity Metrics Assessment

The correlation matrix pandas function [106] was used to calculate the similarity scores using four different metrics (Jaccard, Cosine, Euclidean and City-Block) (available via scipy [127]) applying equation 2 described above for reading across. All these metrics can handle dichotomous and non-dichotomous variables. Many times, we need to handle data plenty of noise hampering the toxicity evaluation [128]. For this reason, to check whether biased metrics generated by supervised classifiers describe the bioisosterism better than raw (non-supervised) similarity metrics, we followed two strategies, i) unbiased RA and ii) biased RA, which are pictured at the top and at the bottom Figure 10, respectively. As illustrated in Figure 10, the first part (A and B) is shared in both strategies: A dataset containing experimental toxicity information (DILI) was used. Morgan fingerprints (2048) were calculated for each molecule. In order to see how the

performance decreases, 5000 random binary numbers were added as noise, using numpy and a random seed of 1987 to be reproducible afterwards. The matrix was transposed to obtain compounds as columns instead of rows. Then, the corresponding correlation matrix for the similarity metrics mentioned above was calculated, with/out adding random noise, over [0,5000] in increments of 50 until 300, then adding 400, 500, 1000 and 5000 obtaining 11 matrices of descriptors.

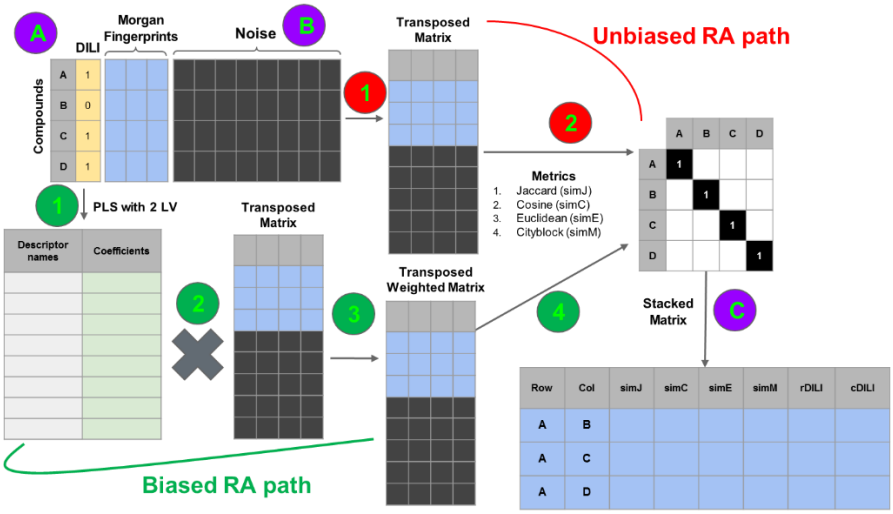


Fig. 10 | **Read Across similarity metrics assessment workflow.** Both biased and unbiased RA share A, B and C steps. A) Morgan Fingerprints Calculation, B) adding random noise and C) calculating the stacked matrix allowing the identification of bioisosters, activity cliffs and grouping. Unbiased RA: Red path with 2 steps: 1) Transpose Matrix, 2) Correlation matrix calculation using four different distance metrics (see Table 4). Biased RA: Green path with four steps: 1) PLS coefficients calculation with two Latent Variables. 2) Multiply the transposed descriptors matrix by the PLS coefficients to obtain 3) the transposed weighted matrix incorporating DILI

biological information. 4) Correlation matrix calculation using 4 different distance metrics (see Table 4).

3.6.2.1. Unbiased RA

As can be seen from the top right (purple-red path) in Figure 10, the previous correlation matrix was stacked, compounds with the same row and column by identifier were dropped (here: non-Standard InChI Key), to finally be merged with the original data set to obtain information on DILI toxicity and ATC category.

Then, a grid search was performed by allowing the similarity cut-off to vary in the interval [0,1] by increments of 0.05 (e.g. 0-1, 0.05-1, ...,0.9-1) to calculate the quality parameters described below.

3.6.2.2. Biased RA

As illustrated at the bottom left (purple-red path) of Figure 10, to avoid the random noise penalization, and to add biological information to the chemical descriptors, a PLS (explained below) with two Latent Variables was run with the 11 matrices of descriptors as X and DILI experimental value as Y. Then, PLS coefficients were retrieved to see the feature importance (Figure 10). Note that, the PLS coefficients were multiplied

by the corresponding transposed matrix, obtaining a transposed weighted matrix, giving higher importance to the variables that better explain DILI. Then, the correlation matrix was stacked, and compounds with the same row and column by chemical identifier were dropped (here: non-Standard InChI Key) to finally be merged with the original data set to obtain information on DILI toxicity and ATC category.

3.6.3. Quality Similarity Assessment

For the quality assessment, a grid searching was performed by allowing the similarity cut-off to vary in [0,1] with increments of 0.05 (e.g. 0-1, 0.05-1, ...,0.9-1) to calculate the quality parameters summarized in Table 4.

Table 4 | Qualitative assessment parameters equations.

Quality Parameter	***Keyword***	***Formula***
Sensitivity	Sens	$\frac{TP}{TP + FN}$
Specificity	Spec	$\frac{TN}{FP + TN}$
Matthews Correlation Coefficient	MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Activity Cliffs	ACff	$\frac{FP + FN}{TP + TN + FP + FN}$

Quality Parameter	*Keyword*	*Formula*
Bioisosters	Biois	$\frac{TP + TN}{TP + TN + FP + FN}$

3.6.4. Similarity Majority Voting Approach

For assessing the similarity between a query compound (so called “target”) and compounds in a dataset (“so called “source”) DILI consensus values were computed by assigning the value of (1) to positive and (0) to negative compounds calculating the mean of the top 10 similar compounds. If this mean value > 0.55 we assign positive (1), if mean value < 0.45 we assign negative (0), else: out of domain (2).

3.7. SIDER Overlapping for Validation Sets

All DILI sets into 2 datasets were divided by checking the overlapping with SIDER dataset. Overlapping molecules were kept, with similarity distance greater or equal to 0.9 to any SIDER compound, calculating distances using Morgan fingerprints with radius 3 and Jaccard similarity.

Table 5 lists the number of overlapping and non-overlapping SIDER molecules in all DILIsets. Non-overlapping datasets were used as external validation sets for QSAR modelling and Most DILIrank dataset as benchmark dataset was used for similarity metrics analysis.

Table 5 | Overlapping and non-overlapping SIDER datasets.

Clinical datasets with experimental DILI value			**Biological datasets**		**SIDER Overlap**			**SIDER Non-Overlap**		
dataset	**mols**	**AEs**	**datasets**	**mols**	**pos**	**neg**	**Total**	**pos**	**neg**	**Total**
SIDER	937	19	Mulliner	2172	505	107	612	933	627	1560
			lessDILIrank	663	278	223	501	63	99	162
			mostDILIrank	528	128	223	351	78	99	177
			Pfizer	378	150	81	231	92	55	147
			Obrien	40	0	0	0	25	15	40

Figure 11 display two PCA scores plots, showing a non-SIDER and SIDER overlapping compounds between DILIsets. In Figure 11A, is clearly represented the chemical space covered by all DILI sets, where Mulliner, mostDILIrank and lessDILIrank are the ones covering a large part of the SIDER chemical space. Conversely, Pfizer and O'Brien datasets covered on a small part of the SIDER chemical space.

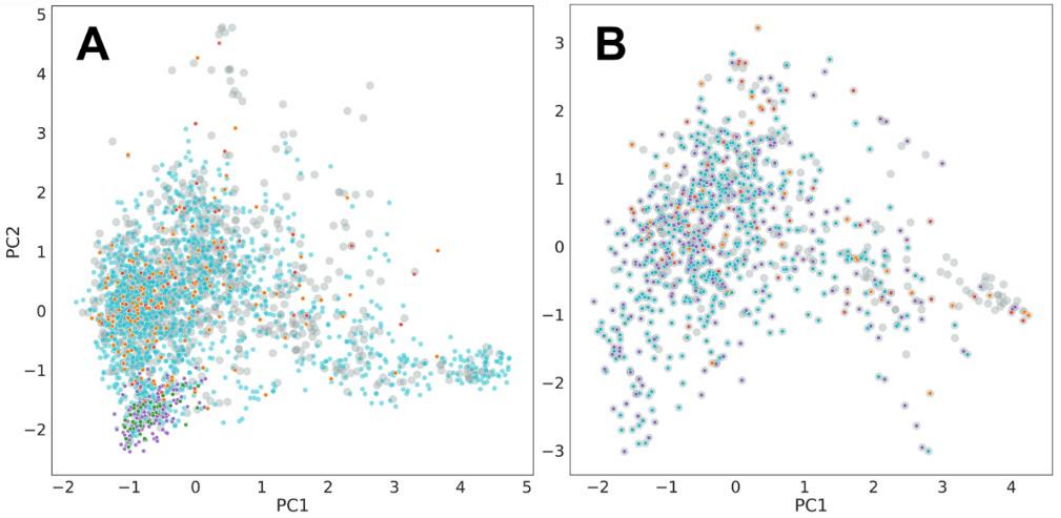


Fig. 11 | **SIDER versus DILIsets overlapping PCA.** A) Non-SIDER overlapping representation. B) SIDER overlapping analysis. Principal component analysis using Morgan Fingerprints (2048). Big grey dots representing SIDER dataset. Red and Orange dots for Less and Most DILIrank sets respectively. Purple dots for Pfizer dataset. Finally, O'Brien dataset as green dots.

Combination Dataset [45]**.** We built a combination dataset for validation with the aim of testing how predictions would work in the real world which contains either DILIrank, Mulliner, O'brien and Pfizer molecules without duplicates and non-overlapping DrugBank small molecule drugs (1405). First, we removed some molecules from DILIrank general dataset containing 20 'nan' and 218 Ambiguous DILI-concern values. Then, for the other labels, vLess-DILI-Concern and vMost-DILI-Concern, and vNo-DILI-concern, 1 and 0 values were assigned respectively. Second, we combined DILIrank, Mulliner, O'brien and Pfizer datasets. We found 609

duplicates out of 3281 molecules. We removed duplicates and we curated DILI activity value applying majority voting. We obtained 2672 molecules. Finally, we combined the previous dataset with non-overlapping DrugBank molecules (1405), which contains 4077 molecules.

3.8. Modelling

3.8.1. Descriptors

For building models, we need to calculate properties of the molecules. We used the RDKit tool to generate the 200 molecular descriptors used when building QSAR models.

3.8.2. Partial Least Squares (PLS)

The scikit-learn Partial Least Squares Regression (PLS-R) [129] using Nonlinear Iterative Partial Least Squares (NIPALS) algorithm was used. Coefficients were used to weight our descriptors in RA analysis and to rank AEs in the QSAR analysis, building a PLS model using the predictions from every model as descriptors (X) and the experimental DILI toxicity value (Y).

3.8.3. Conformal Random Forest (C-RF)

A python customized version of the scikit-learn Random Forest Classifier (RF-C) using conformal predictions was used to build the models using Flame [115]. The number of trees (n_estimators), and maximum features (features) were optimised using a grid search algorithm to obtain optimum values according to Out Of Bag (OOB) criterion. Then, a comparable evaluation of the model predictive quality was performed by applying 5 Kfold cross-validation. The applicability domain of the models was assessed by using the conformal approach [130]. The performance of the qualitative models has been assessed computing the sensitivity, specificity and Matthews Correlation Coefficient, as described in Table 4, but also the conformal coverage which is the sample percentage within the applicability domain of the conformal estimator at a significance given. In other words, lets assume that a data set contains 1000 compounds, if a conformal estimator is valid at 0.9 confidence level, then a maximum of 100 compounds are allowed to be miss-predicted [130]. The conformal coverage would be the compounds predicted percentage within this applicability domain. Finally, all models were validated by predicting the test series and computing the external sensitivity, specificity, MCC and Applicability domain parameters.

3.8.4. AE QSAR Models

AE QSAR models were built using FLAME modelling framework [115]. Random forest under conformal prediction framework was selected as machine learning algorithm. Detailed information on the models is available in annex (Table S7). Finally, RDkit descriptors were used as molecule representation. For a given AE, negatives were taken from other SIDER compounds with different hepatobiliary AE and also from compounds not showing any hepatic AE from SIDER. Later, at model building, negative instances are subsampled to even positive instances.

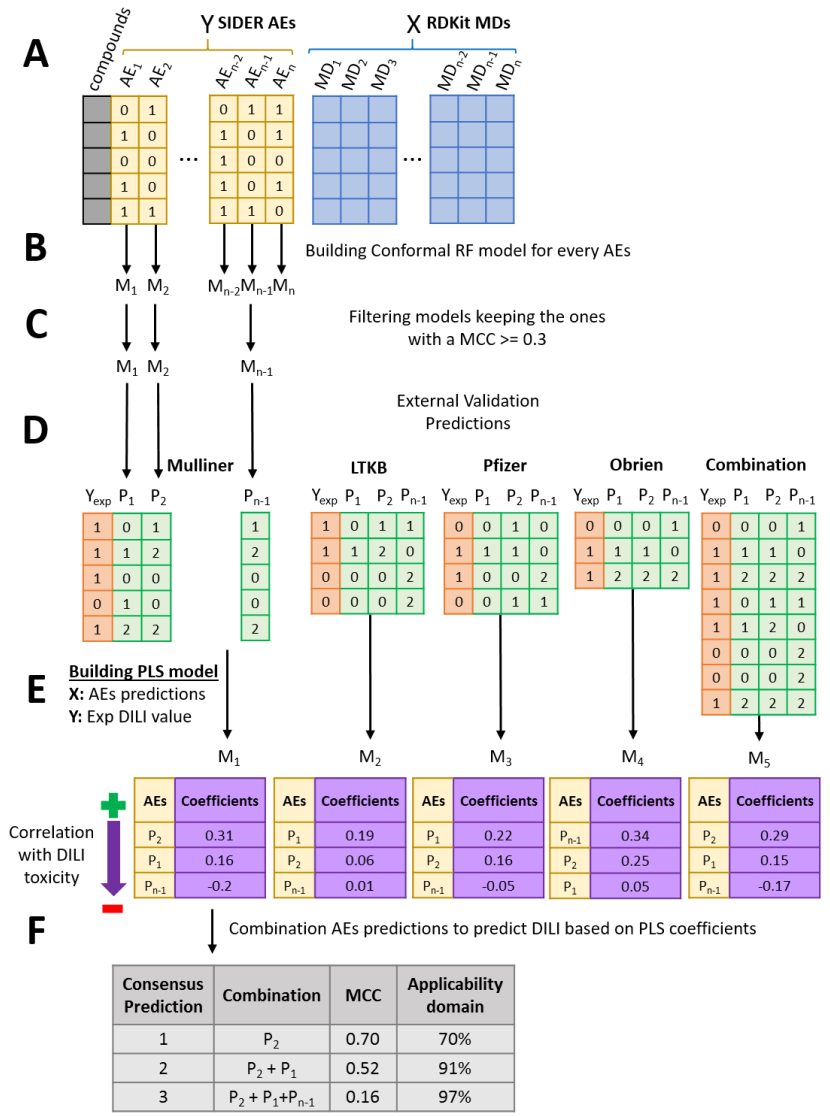
3.8.5. Expert Models

The “expert models” are machine-learning models fitted using AE predictions as explanatory variables and DILI outcome as response variables. Expert models were built using the same modelling settings used in AE QSAR models. In this case we used predictions from AEs QSAR models as explanatory variables and DILI label as response variable. Compounds from DILI databases not being present in SIDER were considered to train the model.

3.8.6. Optimizing AEs Models by Progressively Combining Predictions

Relevant AEs predictions were used for predicting clinical DILI toxicity. That was done to increase the quality of diagnosing hepatotoxicity predictions. Thus, SIDER dataset was used to build conformal Random Forest AEs models (Figure 12B) using RDKit chemical descriptors. Those models were filtered by keeping the ones with MCC greater or equal to 0.3 (Figure 12C). Dataset showing no SIDER overlapping was used as validation sets (Figure 12D).

Then, we built a PLS model (Fig. 12E) using predictions from every AE model as descriptors (X) and the experimental clinical DILI toxicity value (Y) in order to get a ranking of biological descriptors using the PLS coefficients output. A comparison among all sources is performed as could be seen from Figure 18 in results section. Finally, the biological relevant AEs predictions were used to increase the quality of hepatotoxicity assessment using a consensus approach described above (Figure 12F).



Consensus Prediction	Combination	MCC	Applicability domain
1	P_2	0.70	70%
2	$P_2 + P_1$	0.52	91%
3	$P_2 + P_1 + P_{n-1}$	0.16	97%

Fig. 12 | **Detailed modelling workflow used to develop AEs QSARs models.** A) retrieval of training dataset per AE from SIDER. B) Model building for every AE using RDKit chemical descriptors. C) Selection of high-quality models (MCC >= 0.3). D) AE predictions on validation sets. E) PLS analysis of predictions. We obtained PLS scores account for AE importance on DILI assessment. F) Combination of relevant AE predictions to increase the accuracy of hepatotoxicity assignment using a consensus approach (majority voting).

4. RESULTS

Overview of Results

The main aim of this thesis is to develop novel similarity-based tools adapted to the Chemical Safety Assessment (CSA) of drugs. Liver toxicity will be used as an example of toxicity endpoint due to its particular relevance in the drug development pipeline.

The proposed approach involves five steps: (1) Data collection, (2) identification of the best similarity metrics, (3) validation of the similarity metrics, applying them in RA (4) development of QSAR modelling strategies (5) Implement the RA and QSAR models in suitable software tools

(1) To do so, we started collecting datasets (see Table 1) annotated with their DILI properties from diverse open access data sources. These datasets were curated and processed to obtain robust and reliable models.

(2) We benchmarked the similarity metrics performance in RA to compare prediction based on similarity with experimental annotations. This made it possible for us to identify bioisosters molecules and activity cliffs. The strategy here extends the classic concept of structural similarity incorporating additional

properties, like, for example, experimental data describing biological properties (related to hepatotoxicity) of the compounds.

(3) We validated these best similarity metrics obtained, performing a systematic comparison with a collection of similarity descriptors (morgan fingerpints, pharmacophoric fingerprints, biofingerprints and RDKit molecular descriptors) and use RA to predict DILI.

(4) We developed QSAR modelling strategies. We proposed two approaches to evaluate DILI: i) Optimizing biological models (AEs) by progressively combining predictions, finding the optimal combination using a consensus approach (majority voting) and ii) expert models using biological predictions as descriptors.

(5) In order to guarantee reproducible results, we developed all this work in Jupyter Notebooks which works with the software packages described in Table 1 from methods section. These notebooks facilitate the selection of the best similarity metrics and descriptors to obtain similar compounds from a reference dataset and can be used for balancing datasets or to create homologous categorized

datasets among many other things. But also, to build QSAR models to predict DILI.

4.1. Data Collection

We collected different datasets for supporting the selection of the best similarity metrics and for building the QSAR models. Our aim was to obtain data representative of the situations faced in clinical research, more specifically, clinical data containing information about liver toxicity.

Table 6 lists benchmarking datasets collected. Information about the origin and normalization protocol can be found in methods section 2.2. All these datasets were extracted from open access sources (see Table 1). We included these datasets to represent part of the chemical universe, because they contain highly valuable information to be used for modelling Hepatotoxicity.

Table 6 | Datasets collection covering part of the drug development universe.

Drug development level	Dataset	Reference
Chemical	ATC	[119]
Clinical	Mulliner	[131]
Clinical	SIDER	[120–122]

Drug development level	Dataset	Reference
Clinical	DILIrank	[132,133]
Clinical	Pfizer	[133]
Clinical	Obrien	[133]

4.2. Best Similarity Metrics Identification

The metrics used in this work were listed on Table 3 (see methods *section 3.6.1.*). These metrics were selected because they are the most well-known used for similarity assessment and they can deal with binary and non-binary variables [93]. Frequently, we need to handle data with plenty of noise which difficult the toxicity evaluation.

For this reason, to check whether biased metrics generated by supervised classifiers better describe the bioisosterism than raw (non-biased) similarity metrics, we test the potential advantages of biased metrics and the resilience of the metrics to the presence in the descriptors set of non-relevant information. We planned to choose the best similarity metrics by benchmarking the similarity metrics performance in RA to compare predicted with experimental annotations in order to identify the ability of the methods for identifying bioisosters molecules and activity cliffs. In the quality assessment (see section 3.6.2.) we performed a grid searching by allowing

similarity cut-off to vary over [0,1] in increments of 0.05 (e.g. 0-1, 0.05-1, ...,0.9-1) to calculate the quality parameters summarized on Table 4 (see methods *section 3.6.3.*). The Matthews Correlation Coefficient (MCC) was the main quality parameter used to select the best similarity metric. The strategy here extends the classic concept of structural similarity incorporating additional properties, like, e.g., experimental data describing biological properties (related to hepatotoxicity) of the compounds.

For this analysis, we used the overlapping molecules (for more information about overlapping analysis, see methods section *3.7.*) between SIDER and DILIrank dataset containing 100 vMost DILI-concern (as experimental DILI positives) and 152 vNon DILI-concern (as experimental DILI negatives). From this dataset we obtained SIDER information (AEs, ATC category, …) plus DILI experimental toxicity.

The simplistic concept of reading across to predict hepatotoxicity of compounds using inferences based on their biological and/or chemical similarity to other compounds was performed. In order to see how the performance decreases, we added 5000 random binary numbers as noise. Following the unbiased and biased protocol described in methods section 3.6.2 (see Figure 10), we followed both the unbiased and biased RA protocols described in above performing i) unbiased RA (see methods section 3.6.2.1.), ii) biased RA

(see methods section 3.6.2.2.) using the most widely used descriptor (2048 Morgan Fingerprints with radius 3) as molecule representation and iii) the similarity metrics comparison between both unbiased and biased RA, identifying the best RA method to be applied in the Validation section.

4.2.1. Unbiased RA

Unbiased methods are a valuable tool when one does not have any toxicity experimental information. In this section we want to show the performance obtained by using a non-biased RA method, to be compared afterwards with the biased one. We calculated the quality parameters as described in methods section 3.6.2.1. We consider MCC equal or greater than 0.3 to have acceptable quality. As can be seen in Figures 13A and 13B, as we add random noise the performance decreases in both MCC and the number of molecules present in the similarity range chosen for Euclidean and City-Block but Jaccard and Cosine are not so sensitive to random noise. After adding 50 random numbers, the Matthews correlation coefficient (MCC) is clearly penalized by the noise.

For example, in the 0.5-1 similarity cut-off range, where one can find quite similar molecules (20 to 200) (see Figure 13B), MCC ranges from 0.1 to 0.9 without adding random noise. On

the other hand, when adding 100 random numbers, only cosine metric still gets a good MCC, whilst the remaining obtaining 0 as MCC.

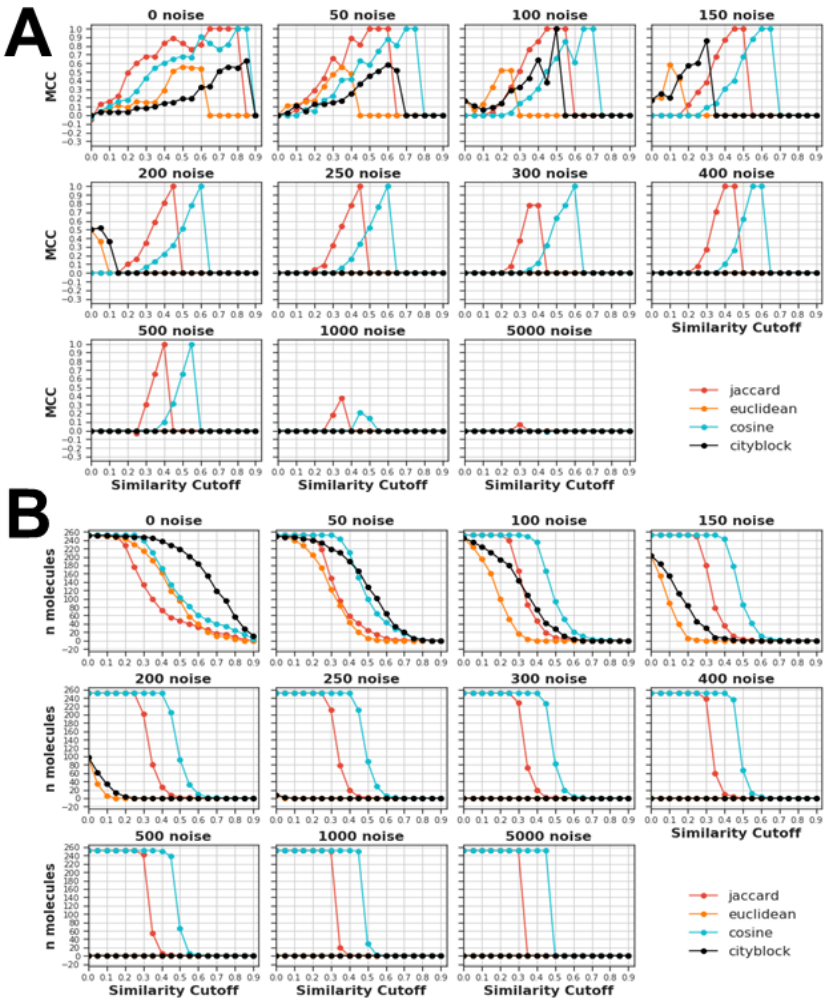


Fig.13 **| Performance of Unbiased RA**. A) MCC in y axis and similarity cut-off from 0 to 1 in x axis. B) n molecules in y axis and similarity cut-off from 0 to 1 x axis. A and B) Metrics comparison using different descriptors. Jaccard in red, Euclidean in orange, Cosine in blue and city-block in black.

4.2.2. Biased RA

Our hypothesis is that biased methods can mitigate the negative effect of non-relevant variables and contribute to a better performance by giving more importance to the variables which correlate with the biological outcome. In this exercise we simulated this effect with the addition of random variables. To test this hypothesis, we ran a PLS with 2 Latent Variables to the 11 matrixes of descriptors as X and DILI experimental value as Y.

Next, we retrieved the PLS coefficients to see the feature importance as can be seen in Figure 14. Note that, we multiplied the PLS coefficients by corresponding transposed matrix, obtaining a transposed weighted matrix, giving higher importance to the variables that explain better DILI. Then, we stacked the correlation matrix, dropped compounds with same source and target InChI Keys, to finally merge with the original dataset to obtain information about DILI toxicity plus ATC category information.

We calculated the quality parameters as described in methods section 3.6.2.1. We consider MCC equal or greater than 0.3 to have acceptable quality. As can be seen in Figures 15A and 15B, as we add random noise the performance decreases in both MCC and the number of molecules present in the similarity range chosen for

Euclidean and City-Block but Jaccard and Cosine are not as sensitive to random noise.

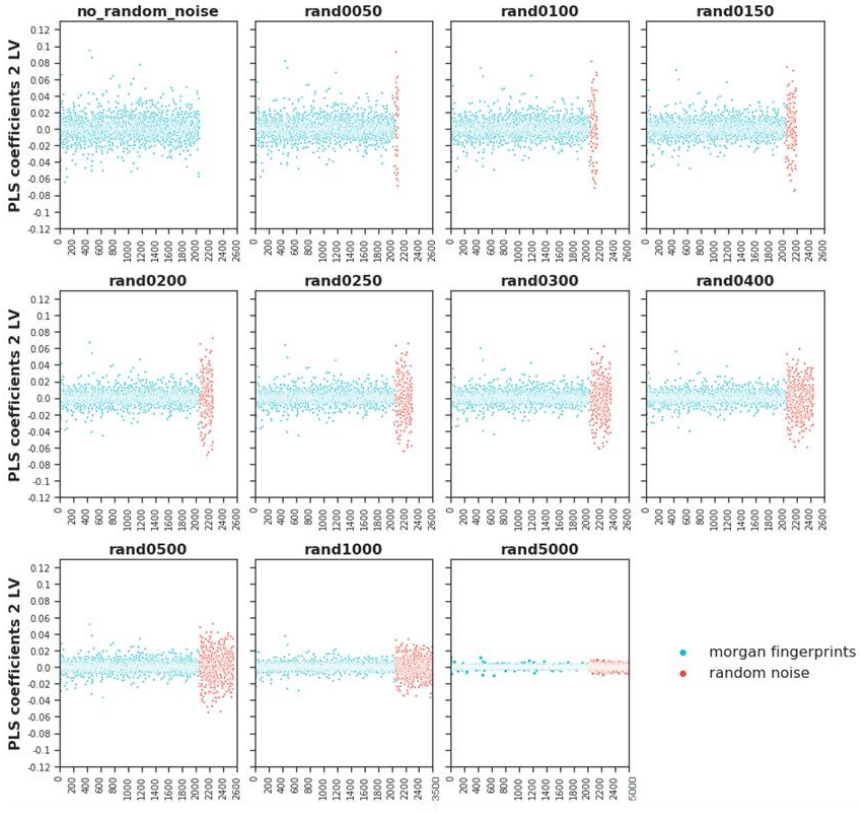


Fig. 14 **| PLS coefficients performance in Biased RA.** Random noise (in red) was added to the 2048 bit Morgan fingerprints (in blue) vector from 0 to 5000.

Analyzing the effect of a biased metric by comparing the results shown in section 4.2.1. obtained in the unbiased methods we could observe that only after adding 200 random numbers is the Matthews correlation coefficient (MCC) penalized by the noise, showing that the negative effect of

non-relevant variables was mitigated, contributing to a better performance.

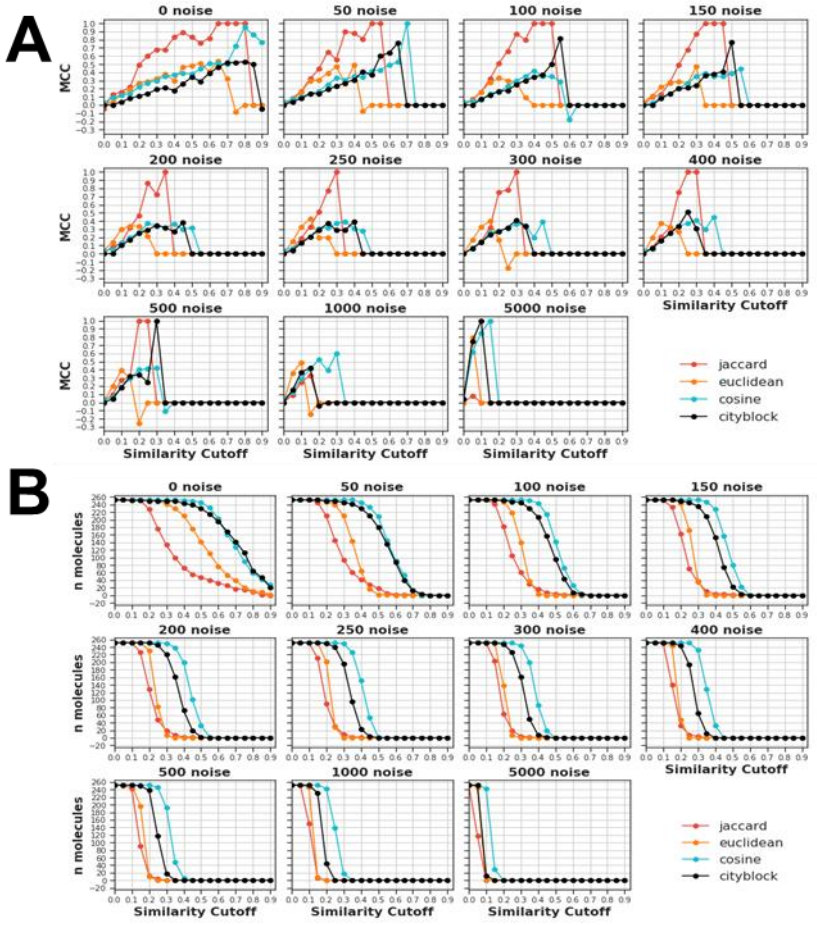


Figure 15 **| Performance of Biased RA similarity.** A) MCC in y axis and similarity cut-off from 0 to 1 in x axis. B) n molecules in y axis and similarity cut-off from 0 to 1 x axis. A and B) Metrics comparison using different descriptors. Jaccard in red, Euclidean in orange, Cosine in blue and city-block in black.

For example, in the 0.5-1 similarity cut-off range, where one can find quite similar molecules (40 to 250 molecules), MCC ranges from 0.3 to 0.9 without adding random noise. On the

other hand, when adding more than 200 random numbers, cosine and city-block metrics still gets an acceptable MCC, whilst the remaining ones obtains 0 as MCC.

4.3. Similarity Validation

4.3.1. Similarity Metrics Validation

As we saw in the previous section, biased RA was performing better than the unbiased one. We used the biased RA method not only because it performed better but also because we wanted to reduce the negative effect of non-relevant variables giving more importance to the variables which correlate with DILI endpoint using the PLS coefficients. We dealt with different binary and non-binary descriptors. Therefore, using the whole information learnt, we validated these similarity metrics, performing a systematic comparison with the most widely used descriptors (see descriptors Table 7) alone but also combined (e.g. Morgan fingerprints plus RDkit descriptors plus bio fingerprints). This was done to see if we will get better information about similarity, for example: are Morgan fingerprints (2048 variables) giving you better similarity explanation than RDkit molecular descriptors (200 variables)? Will a combination give us better results as we will expect? Will biological information add something better to distinguish activity cliffs?

Table 7 | Chemical and Biological molecular descriptors.

Descriptors	type	Description
Morgan fingerprints	chemical	fingerprints
Pharmacophores	chemical	fingerprints
RDKit MD	Physico-chemical	Molecular descriptors
Adverse effects (AEs)	In vivo Clinical	Biofingerprints

To do so, we calculated 4 different descriptors (see Table 7). We used the standard scaler to scale the matrix producing variables with average 0 and variance 1. Then, we ran PLS using different descriptors combination to see the feature importance as we performed above in biased RA.

We finally discarded pharmacophoric descriptors and the combinations containing them because they are too sparse (nearly 30000 fingerprints) and they did not show an acceptable predictive ability of the model nor the quality of the biased metric, therefore they cannot explain better DILI endpoint than the other descriptors. Following the biased RA protocol, we calculated the quality parameters as described in methods section 3.6.2.2. We consider MCC equal or greater than 0.3 to have an acceptable quality.

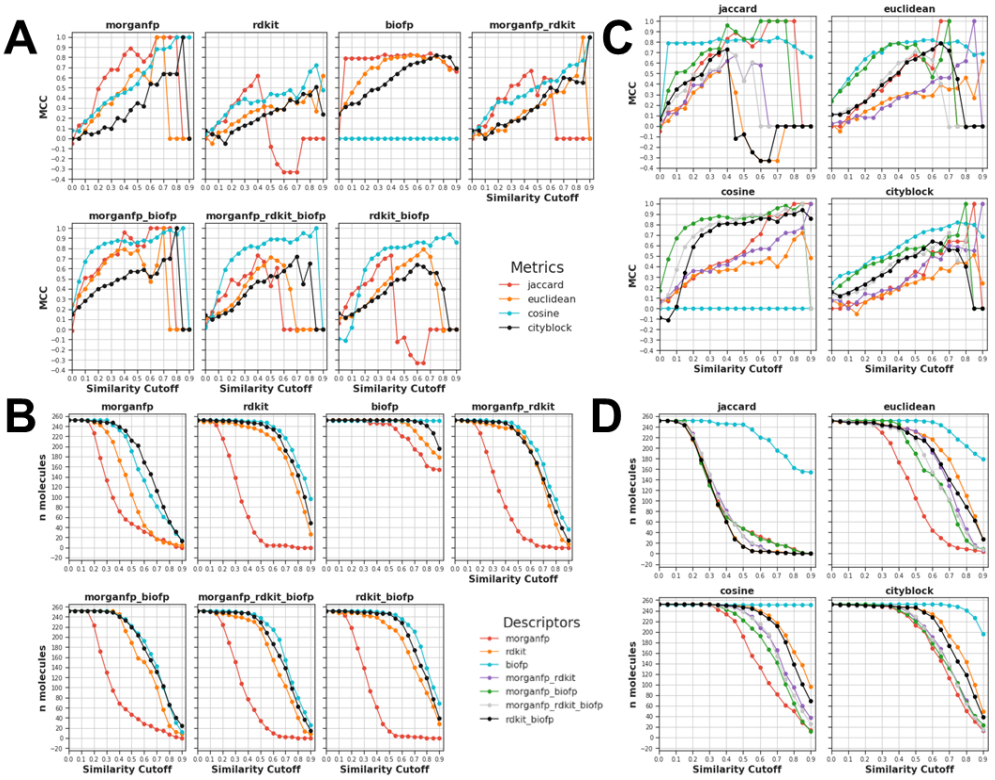


Fig. 16 **| Descriptors combination performance using Biased RA similarity.** A and C) MCC in y axis and similarity cut-off from 0 to 1 in x axis. B and D) n molecules in y axis and similarity cut-off from 0 to 1 x axis. A and B) Metrics comparison using different descriptors. Jaccard in red, Euclidean in orange, Cosine in blue and City-block in black. C and D) Descriptors comparison using different metrics. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

Figures 16A and 16B show the 7 descriptors combination graphs using the 4 similarity metrics described in methods *section 3.6.1,* showing the MCC and number of molecules in a similarity range cut-off respectively. In general, we can see

a similar trend in the 7 descriptors combinations for both MCC and number of molecules. Surprisingly, the Jaccard similarity is the only one which gets slightly lower performance, showing that for non-binary descriptors does not perform as well as the other ones. Cosine similarity fluctuates in some cases such as when using only biofingerprints have MCC values around 0 in all similarity ranges. Additionally, we represented the same results in an alternative manner showing a descriptors comparison using the 4 similarity metrics mentioned before as illustrated in Figures 16C and 16D, representing the MCC and the number of molecules in the similarity range given respectively. Here, it is clearly depicted that biofingerprints on their own shows different trend for all metrics used. Comparing both representations, we can conclude that Euclidean and City-block are the best similarity metrics because the obtain a MCC higher than 0.3 maintaining the number of molecules tested in ranges from 100 to 260 in similarity cut-off ranges from [0.4-1] to [0.7-1].

4.3.2. Similarity RA Consensus Examples

We ran all these analysis with the aim of picking the best similarity metric which allows us to get a good similarity between a query compound (so called “target”) and compounds in a dataset (“so called “source”). To check the performance of the different metrics we run the similarity test using diverse example targets, and descriptors combinations.

As described in methods section 3.6.2, we filtered the similarity stacked matrix, sorting values by Jaccard similarity over the 7 descriptor combinations. Next, we obtained the 10 first source most similar compounds to the target compound as illustrated through Figures 17 and 18. Figure 17 shows one example sorted by Jaccard similarity calculated with Morgan fingerprints and Figure 18 shows a chemical structure representation. To see more examples sorting by Jaccard using the rest of combinations see Figures S1-S6 in annex.

We can also observe the behaviour of the remaining metrics for this particular target compound, in this case clarithromycin which is DILI positive. Applying the consensus rules described in methods section 3.6.4, we obtain a DILI positive consensus value with a similarity cut-off higher than 0.6 and applicability domain 70% ±10%. Here, Nystatin which is DILI negative obtained for some descriptors combination (such as RDKit) higher similarity values than 0.6 which can be considered as possible activity cliff but if we have a look at the biofingerprints similarity and it clearly obtains 0 similarity score in all of metrics used, then we can discard this compound. It seems that here the best combinations are i) morgan fingerprints plus biofingerprints and the one adding ii) morgan fingerprints, biofingerprints and RDKit molecular descriptors.

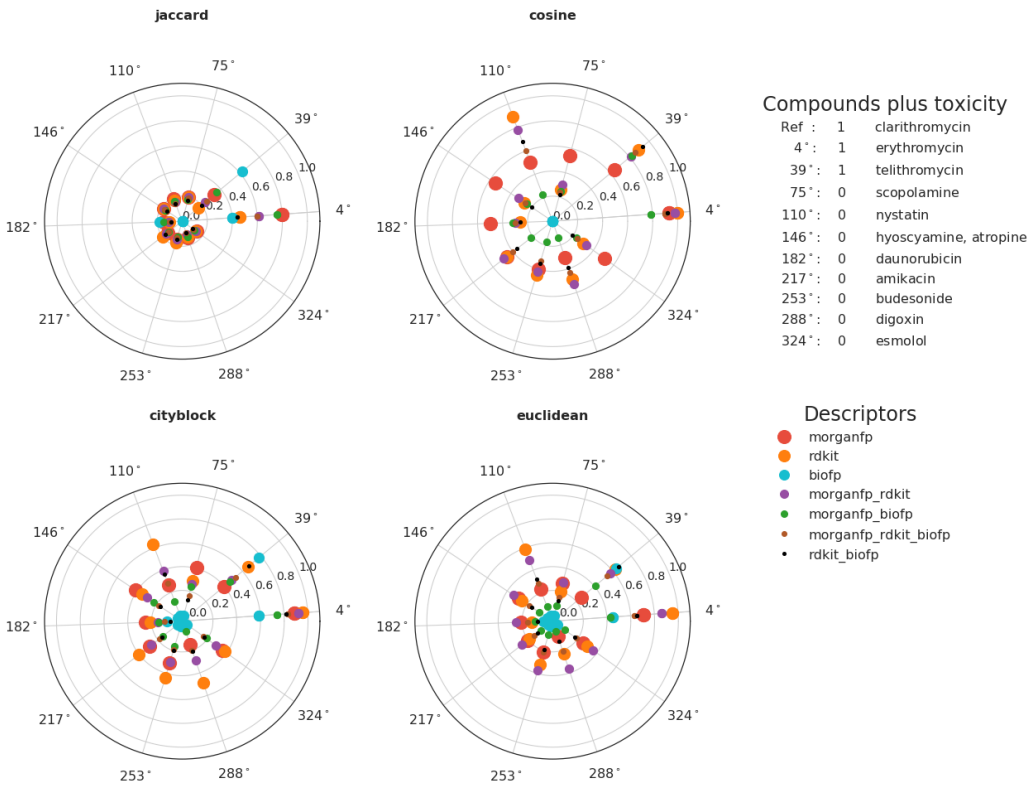


Fig. 17 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using morgan fingerprints. Legend) 0 and 1 mean DILI negative and positive, respectively. Clarithromycin is the reference compound and is DILI positive. 10 best clarithromycin similar compounds represented in radial plot format. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

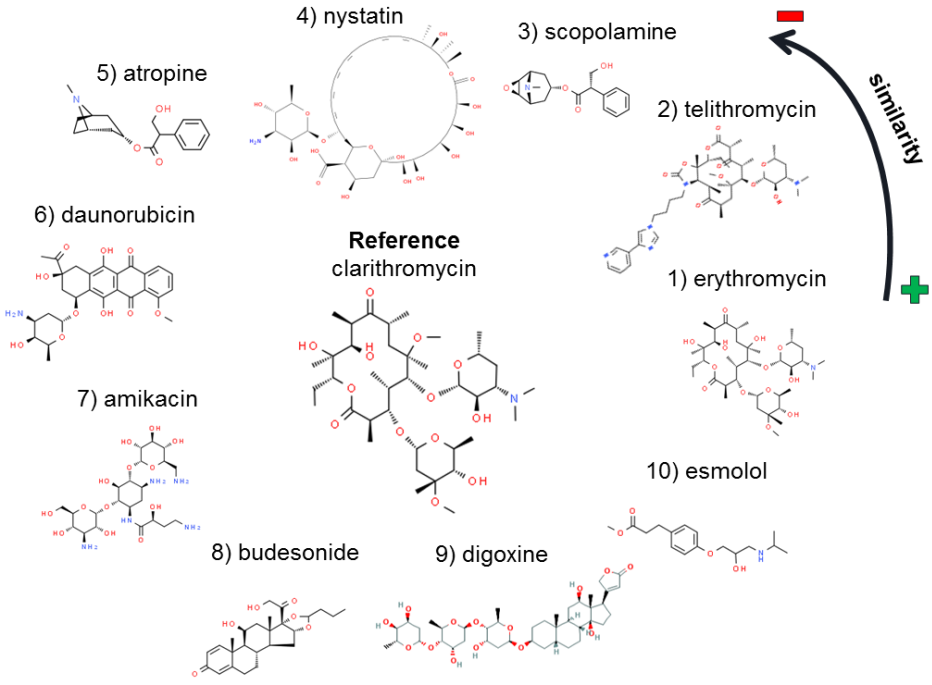


Fig. 18 **| Similarity chemical structures representation.** Ten best similar structures to clarithromycin (reference one) representation being erythromycin the most similar and esmolol the less one.

4.4. Adverse Effect QSAR Models

Another way to obtain DILI predictions is taking advantage of QSAR models. So far, we generated many similarity biased read across models to obtain bioisosters of a target molecule, making the assumption that similar compounds have similar biological properties. In this work we used a similar approach employed by Liu. et al. [23] by building QSAR models to predict hepatic adverse effects with the final aim of using their predictions to build expert models to incriminate DILI.

We used the SIDER dataset containing AEs (see section 3.3 methods) to build 19 Random Forest AE QSAR models under the conformal prediction framework (see section 3.8.3.) following the protocol described in Figure 12 from methods using RDkit descriptors (200 descriptors) as a molecular representation. Conformal significance of the model was set to 0.20 (80% of confidence). Quality statistics for internal 5-fold cross-validation for each model are summarized in Table 8.

Table 8 | Performance of AEs models in internal validation.

Endpoint	Sensitivity	Specificity	MCC[a]	Coverage	Accuracy
Jaundice cholestatic	0.81	0.74	0.55	0.54	0.78
Hyperbilirubinaemia	0.71	0.83	0.54	0.6	0.77
Hepatic failure	0.78	0.75	0.53	0.63	0.77
Cholelithiasis	0.74	0.77	0.51	0.41	0.75
Cholestasis	0.74	0.71	0.45	0.55	0.73
Hepatitis	0.74	0.7	0.44	0.55	0.72
Jaundice	0.76	0.66	0.43	0.55	0.71
Hepatobiliary disease	0.75	0.65	0.41	0.48	0.71
Hepatomegaly	0.69	0.67	0.36	0.44	0.68
Hepatotoxicity	0.66	0.68	0.34	0.61	0.67

Endpoint	Sensitivity	Specificity	MCC[a]	Coverage	Accuracy
Hepatic function abnormal	0.68	0.63	0.31	0.46	0.66

[a]Matthews correlation coefficient.

Only models with a Matthews correlation coefficient (MCC) [134] equal or greater than 0.3 were considered. Model accuracies range between 66-78% while MCCs vary from 0.32 to 0.58 supporting balanced sensitivities and specificities. Model coverage (samples within the AD, see methods section *3.8.3.*) is between 0.41 and 0.57, being penalized by the small dataset, the diverse chemical space covered, and the strict 5-fold internal cross-validation method. The use of conformal prediction [108,130] was critical as model performance decreases critically when not used (annex S1 Table).

4.4.1. QSAR AEs Model Validation

Once the AEs models were built as shown in the previous section, we need to validate them so as they can be used for predicting DILI. The AEs models were externally validated using databases where compounds are labelled by their DILI activity. DILIrank [63], Mulliner [59], Pfizer [125] and O'Brien [126] databases were considered in this study. We took advantage of the criteria used by Liu et al. [55] to normalize DILI positives and negatives across the different databases

(see methods section 3.3). Only compounds not being present in SIDER were considered for the external validation. Two different methods for combining the results to generate an aggregated prediction were used; (1) by considering DILI positive if any adverse effect is predicted, and (2) by using a consensus approach defined as follows: if the number of positive AE assignments is equal or greater than the number of negative ones, the compound is classified as DILI positive, otherwise it was considered DILI negative. In both validation rules, if all AE predictions fall out of the AD, then the compound is classified as out of the AD. In order to maximize the predictive power of both considered rules, AE models were selected by iteratively adding AE predictions from most to less predictive AE (see next section) and computing the consensus performance for each dataset. AE selection leads to a considerable improvement of performance power for both rules (see Figures S7-S11) but, differences in the effect of AE addition among datasets are noticeable: DILIrank and Mulliner datasets show a slight improvement in prediction performance obtaining a reasonable compromise between predictive ability and coverage with 4 and 7 AEs respectively. In the case of Pfizer and O'Brien datasets, the predictive power is considerably higher when optimization is applied. Table 9 shows the comparison between optimized and non-optimized rules for the different external sets considered in this work.

Consensus rules classify a compound as DILI positive if number of positive AEs assignments is equal or greater than the number of negative ones, otherwise the compound is considered DILI negative. On the other hand, “any positive” rule labels a compound as positive if any of the predicted AEs is predicted.

Table 9 | Performance of consensus and “any positive” rules in DILI prediction

		Non optimized				Optimized			
		MCC[a]	**Sens**	**Spec**	**nAE**	**MCC[a]**	**Sens**	**Spec**	**nAE**
Consensus Rule	**Most DILIrank**	0.26	0.61	0.66	11	0.49	0.73	0.78	4
	Pfizer	0.02	0.65	0.37	11	0.36	0.81	0.55	2
	O'brien	0.17	0.75	0.43	11	0.37	0.82	0.53	4
	Mulliner	0.24	0.75	0.50	11	0.27	0.74	0.52	4
	DrugBank + others	0.17	0.53	0.63	11	0.23	0.56	0.67	3
Any positive rule	**Most DILIrank**	0.26	0.57	0.70	11	0.49	0.7	0.79	4
	Pfizer	0.09	0.67	0.43	11	0.36	0.81	0.55	2
	O'brien	0.47	0.80	0.67	11	0.37	0.82	0.53	4
	Mulliner	0.18	0.67	0.52	11	0.23	0.7	0.53	4
	DrugBank + others	0.14	0.48	0.67	11	0.23	0.55	0.68	3

[a]Matthews correlation coefficient

4.4.2. Adverse Effect Analysis

Both DILI labelling rules give the same importance to all AEs. Not all models have the same predictive power and not all AEs are equally important to label a compound as DILI positive or negative. We performed partial least squares (PLS) as described in methods section 3.8.2 and correlation analyses on compounds shared by SIDER (real AEs) and other databases in order to gain insight into the AE importance to incriminate DILI for each dataset, as reflected by the PLS coefficient values. Also, a similar analysis was performed using AE predictions on compounds do not present in SIDER in order to evaluate the predictive power of AE models for each dataset. Figure 19 shows the relative importance of AEs for the different datasets (PLS coefficients) in the DILI classification.

In general, there is an agreement of most important AEs between datasets indicating that our models are able to capture their relative importance when evaluating a compound on its DILI potential. Hepatitis and jaundice are on average the most relevant predicted AE when classifying a compound as DILI positive. This can be attributed to the frequency and importance of these AEs in study reports. Interestingly, Jaundice cholestatic, the best AE model in internal validation, appears overrepresented in both the PLS and correlation analysis pointing out the importance of model

quality besides the AE importance. Detailed information on PLS scores and correlation analysis can be found in annex S2, S3, and S4 Tables.

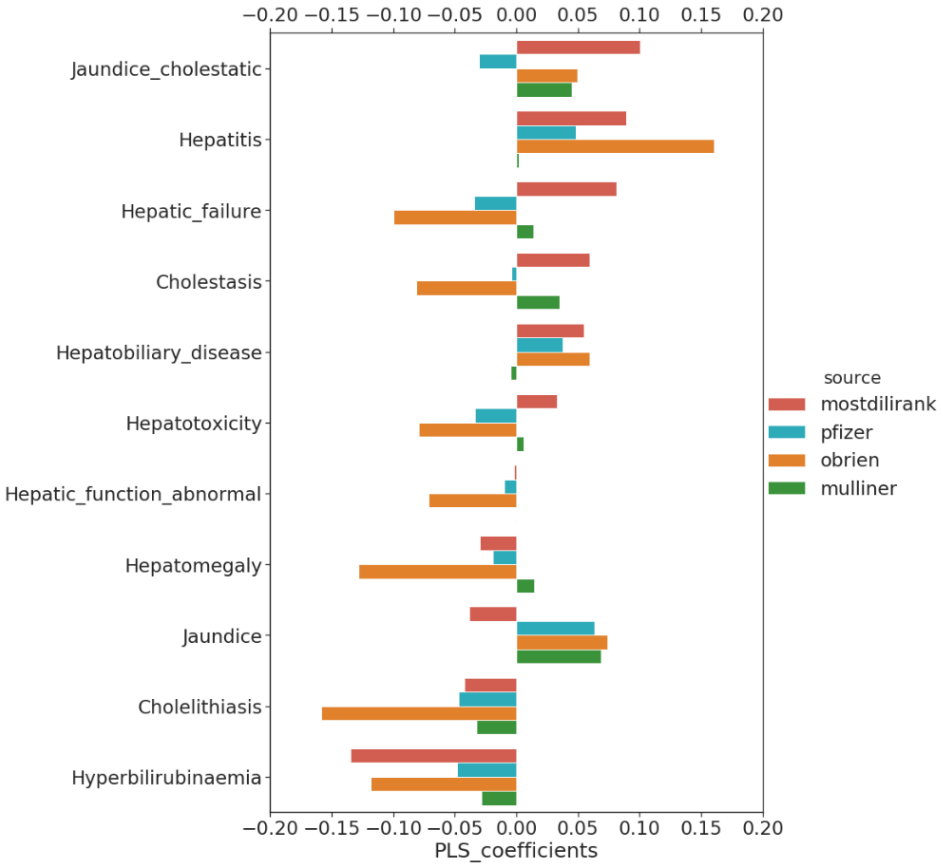


Fig. 19 | **AEs PLS coefficients importance.** PLS score plot for each AE in the different datasets considered in this work. Higher scores imply more correlation between AEs and DILI labelling.

4.4.3. Expert Models

Even if the model aggregation rules described in the previous sections work relatively well for the DILI assignment, it is

burdened by the arbitrariness in the number of AE used and the simplicity of the rule approaches. To overcome this problem, we introduce the concept of “expert models”. These “expert models” are machine-learning models using AE predictions as explanatory variables and DILI outcome as response variables. Unlike the consensus rule approach, “expert models” can learn relationships between AEs related to DILI classification, and provide confidence in DILI classification thanks to the use of conformal prediction. “Expert models” were created for each dataset using AE predictions of compounds not being present in SIDER and externally validated using remaining datasets (compounds not being present in SIDER nor training set). Table 10 shows the performance of “expert models” for internal validation (5-fold) and external-validation. Mulliner expert model shows the higher absolute values in performance external validation (MCC of 0.61 on average, not weighted by dataset size) with the highest coverage (0.42). Pfizer expert model shows a similar performance to Mulliner expert model, but the coverage decreases to 0.15. Finally, DILIrank expert model reaches a MCC of 0.49 with 0.30 of coverage.

Table 10 | Performance of expert models in internal and external validation**.**

		Quality parameters	DILIrank	Pfizer	Mulliner
Internal validation		**MCC**	0.67	0	0.37
		Sensitivity	0.82	0	0.59
		Specificity	0.85	1	0.78
		Coverage	0.35	0.07	0.31
		Accuracy	0.84	0.37	0.69
		Conformal	yes	yes	yes
		Significance	0.1	0.1	0.1
External sets	**DILIrank**	**MCC**	-	0.48	0.63
		Sensitivity	-	0.6	0.9
		Specificity	-	0.86	0.77
		Coverage	-	0.16	0.34
		Accuracy	-	0.75	0.81
		Conformal	-	yes	yes
		Significance	-	0.1	0.1
	Pfizer	**MCC**	0.32	-	0.38
		Sensitivity	0.73	-	0.65
		Specificity	0.59	-	0.73
		Coverage	0.25	-	0.4
		Accuracy	0.66	-	0.69
		Conformal	yes	-	yes

		Quality parameters	DILIrank	Pfizer	Mulliner
		Significance	0.1	-	0.1
	O'brien	**MCC**	0.78	1	0.83
		Sensitivity	0.9	1	0.91
		Specificity	1	1	1
		Coverage	0.35	0.13	0.52
		Accuracy	0.91	1	0.93
		Conformal	yes	yes	yes
		Significance	0.1	0.1	0.1
	Mulliner	**MCC**	0.36	0.25	-
		Sensitivity	0.67	0.72	-
		Specificity	0.69	0.53	-
		Coverage	0.29	0.16	
		Accuracy	0.68	0.65	-
		Conformal	yes	yes	-
		Significance	0.1	0.1	-
Average		**MCC**	0.49	0.58	0.61
		Sensitivity	0.77	0.77	0.82
		Specificity	0.76	0.80	0.83
		Coverage	0.30	0.15	0.42
		Accuracy	0.75	0.80	0.81

[a]Matthews correlation coefficient.

[b]Per one of samples within the applicability domain.

Although expert models show high performance at DILI labelling, whether or not they offer advantages to regular QSAR models has to be assessed. QSAR models were built using the same datasets used for expert models but taking molecule descriptors instead of the predicted AEs. Results evince that despite a slightly lower performance in internal-validation, statistics in external-validation of expert models are significantly better. Coverage for QSAR models is higher, probably due to the higher confidence given by the number of variables used to train the model (200 in QSAR versus 11 in expert models). However, increasing the confidence in the predictions (from 85% to 90%) did not result in a better performance although the lower coverage. Detailed information on QSAR models' statistics is available in Annex (S5 Table).

Despite expert models showing great performance and robustness, machine-learning models often suffer of lack of performance when applied out of the safe space of building environment. For this reason, we built a dataset containing DILIrank, Mulliner, O'brien and Pfizer molecules without duplicates and no overlapping DrugBank small molecule drugs (1405 extra negatives) and used expert models to predict DILI outcome (only compounds not used in training for each model considered). This test is intended to assess the performance in a not biased dataset (DrugBank) where DILI outcome is not balanced. Table 11 shows an equilibrated

sensitivity/specificity for all three expert models, demonstrating that our models are not over-predicting DILI.

Table 11 | Performance of expert models real world validation set.

		Most DILIrank	Pfizer	Mulliner
Drugbank set	**MCC** [a]	0.31	0.23	0.33
	Sensitivity	0.55	0.66	0.69
	Specificity	0.75	0.58	0.66
	Coverage [b]	0.29	0.11	0.47
	Accuracy	0.68	0.61	0.67

[a]Matthews correlation coefficient.

[b]Per one of samples within the applicability domain.

4.5. Implementation

In order to guarantee reproducible results, we developed everything using Jupyter notebooks in a Conda environment which can be downloaded and installed with all with all the dependencies and extra packages used. Two Jupyter notebooks, one for unbiased RA and another for biased RA for the similarity assessment, facilitated the selection of the best similarity metrics and descriptors to obtain similar compounds from a reference dataset that can be used for balancing datasets or to create homologous categorized datasets among many other things. On the other hand, for

building QSAR models to predict DILI, we take advantage of the whole tools inside the Flame framework [115] and created two Jupyter notebooks, one using the optimizing AEs rule and another for building the Expert models.

All datasets and AEs models are available on request. These are provided as tar files and can easily be imported in Flame ready to run predictions.

4.5.1. Installation of the Jupyter Notebooks

To install the Jupyter notebooks first install the Conda environment found in my github [104] and follow the instructions written in Flame github [115]**.** In the step of creating the conda environment with all the dependencies and extra packages (numpy, RDKit...), instead of using the flame environment.yml, replace it with the one we provided (thesis_environment.yml), which assures us to use all the versions of the programs used here. Nevertheless, if you want to use the last versions released of whole programs, just install flame as it is and then install the other ones.

5. DISCUSSION

This thesis is focused on the prediction of the biological properties of compounds by applying the concept of the bioisosterism. Most computational methods used for the prediction of toxicity endpoints assume that similar compounds have similar biological properties. This principle is at the core of computational methods like read-across (RA) or quantitative structure–activity relationships (QSARs), there are many caveats like “activity cliffs” which have been pointed out.

Early identification of DILI remains a major concern for the regulatory agencies and the pharma industry. The standard animal model is moving towards a more ethical and mechanistic based toxicological assessment [135]. Within non-animal testing methods, *in silico* method play a key role in the transition to a complete non-animal testing scenario [136], not only by providing assessment methods but also for processing, analyzing and filtering the increasing quantity of biological data generated by high-throughput screening assays [137] and OMICS techniques [138].

5.1. Similarity RA

One of the central hypothesis of this thesis is that the source of the problems is the definition of what we mean by “similar”

and how we compute similarity. Our group has published diverse similarity metrics, relevant in the context of safety assessment [139]. Also, there is much interest in similarity metrics due to the large implication for developing structural alerts and read-across [13]. Although acceptance of read-across by regulators remains difficult [13], read-across plays a pivotal role in hazard assessment of chemicals. Indeed, it is the only currently available non-animal alternative method for many regulatory environments and for many toxicological endpoints [4]. In other contexts similarity metrics are also of critical importance for the definition of the applicability domain of any *in silico* method and the quantification of the uncertainties associated to any prediction [17].

For benchmarking the similarity metrics performance, we used RA as a validation tool, in order to compare predicted with experimental annotations. This made possible to identify bioisosteric molecules and discard activity cliffs. The strategy here extends the classic concept of structural similarity incorporating additional properties, like, for example, experimental data describing biological properties (related to hepatotoxicity) of the compounds. We have performed two RA methods, one unbiased and another biased using the very well-known Morgan fingerprints with radius 3. In order to test the effect of using non-biologically relevant descriptors, we tested the effect of adding 5000 binary random numbers to see how sensible were both methods. The results shown

that biased RA methods perform better, producing acceptable results even after adding 200 noise variables in comparison with the unbiased method which were highly penalized since the beginning.

After that we used PLS which was a supervised machine learning technique that allow us to add information about DILI experimental value, in other words, we are giving more importance to the variables that explain better DILI. Our analysis showed that Jaccard and Cosine produces almost always the same results and City-block and Euclidean as well. But we wanted to go further, and we validated our metrics with different highly used descriptors, i) Morgan fingerprints (2048), ii) pharmacophoric fingerprints (30000), iii) hepatotoxic AEs (11) from sider as biofingerprints, and iv) molecular descriptors from RDKit.(200). All of them contain dichotomous variables, except RDKit which contains non-dichotomous variables. We applied the biased RA protocol and we concluded that the descriptors themselves give different information about structural similarity. For example, Morgan fingerprints give us better information about structure than RDkit which give us better information about ADME properties, or biofingerprints on hepatotoxicity.

Strikingly, if we combined them all we can distinguish between bioisosters better and discard activity cliffs. Hence, it is possible to assign with more confidence a DILI toxicity

value. This was illustrated by the example shown in Figure 17 in radial plot format. City-Block and Euclidean were the best similarity metrics for all descriptors combination using biased RA obtaining a MCC higher than 0.3. They maintained the number of molecules tested in a coverage space of 100 to 260 molecules within the similarity cut-off ranges from [0.4-1] to [0.7-1]. It is worth combining descriptors and fingerprints giving extra information such as the combination of morgan fingerprint plus RDKit plus biofingerprints, so that the interpretability of the results can be increased.

The polar plot was a nice method for representing similar source compounds to the target one as can be seen from Figure 17. If we obtain only source compounds toxicity, we compute the target compound toxicity by applying the majority voting algorithm, taking into account compounds with similarity higher than 0.7. The same method can be used when one compound contains target toxicity and has similar source compounds. One can assign the toxicity of the target compound to most similar source compounds using a threshold of similarity greater than 0.7.

5.2. QSAR Models

Another way to obtain DILI predictions is taking advantage of QSAR models as we did with similarity biased read across method to get bioisosters of a target molecule, making the

assumption that similar compounds have similar biological properties.

We have built QSAR models for reported hepatic AEs using SIDER database [120–122]. Only models with a MCC equal or greater than 0.3 were considered for their application. This selection resulted in 11 high performance models (Table 9). Due to the similar procedure, some of the endpoints are the same as in with Liu et al. DILIps [133]. Despite the lower performance of our models in internal validation, likely to be produced by the stricter 5-fold internal validation, there is better performance in external validation we obtained when applying the consensus rule (Table 8).

Labelling a compound as DILI positive is not an easy task [140,141]. DILI can be caused by different mechanisms and these mechanisms are not always captured by in-vitro and in-vivo models [62]. Additionally, the lack of clinical biomarkers allowing to causally connect DILI with molecular mechanisms, makes harder the classification of DILI drugs according to their mode of action. In this work, we rely on reported clinical AEs of drugs. These reports are not free of biasing and for example, the overrepresentation of hepatitis and jaundice might be related to the more frequent diagnostic they comprise in comparison to other AEs. We analysed the AE importance (both real and predicted AEs) in DILI assignment and found it to be correlated with the frequency of AEs.

Effectively, DILI classification in many studies depends on the number of AE annotations, no matter which AE is reported [59,125,126](see methods). It may be thought that the higher importance of hepatitis and jaundice in the AE models predictions might be related to a higher number of predictions inside of the AD. However, the distribution of predictions within the AD, are not significantly higher for hepatitis or jaundice in comparison to other AEs (see annex S12-S14 Figures), suggesting that AE models presented herein successfully capture the AE importance. It is important to recall that AE prediction importance is also related to the model performance, as shown by the higher relative importance of predicted ‘Jaundice cholestatic’ in comparison to real assignments. Correlation between real AEs and DILI assignment vary among datasets. For example, the correlation between Hepatitis and DILI for DILIrank dataset (verified most DILI concern label) is 0.77 while for Mulliner dataset is 0.31 (see annex S3 Table). This reflects the differences in the assessment of DILI potential among datasets/studies. For example, Mulliner consider a compound to be DILI positive if any hepatic-related finding is found [59], while in DILIrank dataset, a more strict criteria is taken. This is reflected in the fact that all compounds present in both DILIrank and Mulliner datasets with inconsistent DILI label are always DILI positive in Mulliner dataset (see annex Table S6).

Bearing in mind that DILI labelling was assigned using different criteria in diverse datasets, we created the expert models to not only be able to predict DILI for a given compound, but also to do it in the way it would be done by the related expert (study, database). These experts models, able to assess DILI using predictions from QSAR AE predictions, show better performance than any reported DILI model. Kotsampasakou et al. [142] summarized the performance of previously reported models in literature, even if it is difficult to compare among different models because of the different ways to assess model performance (internal or external validation), the size of validation sets, and statistics considered. Expert models reported in this work perform as well as models considering only internal validation statistics or used small external validation data sets, and clearly outperforms models validated under similar conditions. It has to be emphasized that expert models were fitted using AE predictions of compounds not being present in SIDER, and were externally validated using compounds of other datasets, which can be seen as a two-step external validation.

Moreover, we built QSAR models for each dataset and validated them by predicting DILI on the others. The lower performance of these models based on RDkit descriptors (200 variables) puts in value the predictive power of expert models which are trained on predictions from 11 AE models (11 variables). Additionally, we created an imbalanced

dataset containing all DILI datasets used in this work, enriched with DILI negative compounds taken from DrugBank database to assess the performance of expert models when query compounds belong to a wide chemical space and the number of positive and negative instances are not balanced. Expert models showed balanced sensitivities-specificities (Table 10), not falling into DILI over-prediction.

The quality of these models is strongly linked to the use of conformal prediction as applicability domain technique. Conformal predictors not only assess for the reliability of a prediction but creates a framework to build models with clearly defined uncertainties that ultimately, are required in a risk assessment context. These uncertainties can be combined with other methodologies in a weight-of-evidence approach to provide a toxicologist with a balanced view on a chemical of interest.

6. CONCLUSION

- Benchmarking the similarity metrics performance in RA allowed us to compare predicted with experimental annotations. This allowed to identify bioisosters molecules and discard activity cliffs.

- Biased RA methods performed better than unbiased RA, obtaining good results even after adding 200 random variables in comparison with the unbiased method whose performance suffered since the beginning (50 random noise).

- Combining descriptors allowed us to better distinguish between bioisosters and discard activity cliffs. Hence, the increased confidence in the assignment of a DILI toxicity value.

- City-Block and Euclidean were good similarity metrics for all combination descriptors in a biased RA.

- We built QSAR models to predict hepatic adverse effects using an approach similar to the one employed by Liu. et al. [23]. These models were used to predict adverse effects of compounds from known DILI databases. The importance of hepatic adverse effects was analyzed in both real and predicted AEs in order

to optimize a consensus rule to assess DILI from adverse effect predictions.

- We used these predictions to build “expert models” able to assess DILI by finding associations among predicted adverse effects. These “expert models” can capture the subjective criteria taken in each study to discriminate between hepatotoxic and non-hepatotoxic compounds and were externally validated by predicting DILI on the other datasets.

7. BIBLIOGRAPHY

[1] Merz KM, Ringe D, Reynolds CH. Drug design : structure- and ligand-based approaches. 1st ed. Cambridge, UK: Cambridge University Press; 2010.

[2] How are drugs designed and developed? | Facts | yourgenome.org. https://www.yourgenome.org/facts/how-are-drugs-designed-and-developed (accessed September 9, 2019).

[3] Kapetanovic IM. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact 2008;171:165–76. doi:10.1016/j.cbi.2006.12.006.

[4] Hartung T. Food for Thought Making Big Sense from Big Data in Toxicology by Read-Across. ALTEX 2016;33:83–93.

[5] Jorgensen WL. The many roles of computation in drug discovery. Science 2004;303:1813–8. doi:10.1126/science.1096361.

[6] Rishton GM. Nonleadlikeness and leadlikeness in biochemical screening. Drug Discov Today 2003;8:86–96. doi:10.1016/S1359644602025722.

[7] Lemmens T, Elliott C. Justice for the Professional Guinea Pig. Am J Bioeth 2001;1:51–3. doi:10.1162/152651601300169095.

[8] Shamoo AE, Resnik DB. Strategies to Minimize Risks

and Exploitation in Phase One Trials on Healthy Subjects*. Am J Bioeth 2006;6:W1–13. doi:10.1080/15265160600686281.

[9] Brown SR, Gregory WM, Twelves CJ, Buyse M, Collinson F, Parmar M, et al. Designing phase II trials in cancer: a systematic review and guidance. Br J Cancer 2011;105:194–9. doi:10.1038/bjc.2011.235.

[10] Vlahović-Palčevski V, Mentzer D. Postmarketing Surveillance. Handb. Exp. Pharmacol., vol. 205, 2011, p. 339–51. doi:10.1007/978-3-642-20195-0_17.

[11] Drug Discovery Pipeline. https://steveblank.files.wordpress.com/2013/08/drug-discovery-pipeline.jpg (accessed October 30, 2019).

[12] Klaassen CD. The Basic Science of Poisons. 2001.

[13] Ball N, Cronin MTD, Shen J, Blackburn K, Booth ED, Bouhifd M, et al. Toward Good Read-Across Practice (GRAP) guidance. ALTEX 2016;33:1–18. doi:10.14573/altex.1601251.

[14] Luechtefeld T. Analysis of public oral toxicity data from REACH registrations 2008-2014. ALTEX 2016;33:111–22. doi:10.14573/altex.1510054.

[15] Luechtefeld T, Maertens A, P. Russo D, Rovida C, Zhu H, Hartung T. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008-2014. ALTEX 2016;33:95–109. doi:10.14573/altex.1510052.

[16] Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu

H, Hartung T. Analysis of Draize Eye Irritation Testing and its Prediction by Mining Publicly Available 2008-2014 REACH Data 1. ALTEX 2016;33:1–18.

[17] Patlewicz G, Ball N, Boogaard PJ, Becker R a., Hubesch B. Building scientific confidence in the development and evaluation of read-across. Regul Toxicol Pharmacol 2015;72:117–33. doi:10.1016/j.yrtph.2015.03.015.

[18] Zhu H, Bouhifd M, Kleinstreuer N, Kroese ED, Liu Z, Luechtefeld T, et al. Supporting Read-Across Using Biological Data. ALTEX 2016;33:1–18. doi:10.14573/altex.1510052.

[19] Regulation TC. Interface between REACH and Cosmetics regulations. 2013.

[20] Lilienblum W, Dekant W, Foth H, Gebel T, Hengstler JG, Kahl R, et al. Alternative methods to safety studies in experimental animals: Role in the risk assessment of chemicals under the new European Chemicals Legislation (REACH). Arch Toxicol 2008;82:211–36. doi:10.1007/s00204-008-0279-9.

[21] Saini N, Bakshi S, Sharma S. In-silico approach for drug induced liver injury prediction: Recent advances. Toxicol Lett 2018;295:288–95. doi:10.1016/j.toxlet.2018.06.1216.

[22] Cheng A, Dixon SL. In silico models for the prediction of dose-dependent human hepatotoxicity. J Comput Aided Mol Des 2003;17:811–23.

doi:10.1023/B:JCAM.0000021834.50768.c6.

[23] Watkins PB. Drug Safety Sciences and the Bottleneck in Drug Development. Clin Pharmacol Ther 2011;89:788–90. doi:10.1038/clpt.2011.63.

[24] Fung M, Thornton A, Mybeck K, Wu JH, Hornbuckle K, Muniz E. Evaluation of the Characteristics of Safety Withdrawal of Prescription Drugs from Worldwide Pharmaceutical Markets-1960 to 1999. Drug Inf J 2001;35:293–317. doi:10.1177/009286150103500134.

[25] Kaplowitz N. Idiosyncratic drug hepatotoxicity. Nat Rev Drug Discov 2005;4:489–99. doi:10.1038/nrd1750.

[26] Olson H, Betton G, Robinson D, Thomas K, Monro A, Kolaja G, et al. Concordance of the Toxicity of Pharmaceuticals in Humans and in Animals. Regul Toxicol Pharmacol 2000;32:56–67. doi:10.1006/rtph.2000.1399.

[27] Sullins J. Transcending the meat: immersive technologies and computer mediated bodies. J Exp Theor Artif Intell 2000;12:13–22. doi:10.1080/095281300146281.

[28] Driessen M, Vitins AP, Pennings JLA, Kienhuis AS, Water B van de, van der Ven LTM. A transcriptomics-based hepatotoxicity comparison between the zebrafish embryo and established human and rodent in vitro and in vivo models using cyclosporine A, amiodarone and acetaminophen. Toxicol Lett 2015;232:403–12. doi:10.1016/j.toxlet.2014.11.020.

[29] Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A. Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species. Chem Res Toxicol 2010;23:171–83. doi:10.1021/tx900326k.

[30] Ashammakhi N, Elkhammas E, Hasan A. Translating advances in organ-on-a-chip technology for supporting organs. J Biomed Mater Res Part B Appl Biomater 2019;107:2006–18. doi:10.1002/jbm.b.34292.

[31] Zhao P, Zhang L, Grillo JA, Liu Q, Bullock JM, Moon YJ, et al. Applications of Physiologically Based Pharmacokinetic (PBPK) Modeling and Simulation During Regulatory Review. Clin Pharmacol Ther 2011;89:259–67. doi:10.1038/clpt.2010.298.

[32] Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H. Chemical predictive modelling to improve compound quality. Nat Rev Drug Discov 2013;12:948–62. doi:10.1038/nrd4128.

[33] Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, et al. Predicting Hepatotoxicity Using ToxCast *in Vitro* Bioactivity and Chemical Structure. Chem Res Toxicol 2015;28:738–51. doi:10.1021/tx500501h.

[34] Marchant CA, Fisk L, Note RR, Patel ML, Suárez D. An expert system approach to the assessment of hepatotoxic potential. Chem Biodivers 2009;6:2107–14. doi:10.1002/cbdv.200900133.

[35] Pizzo F, Lombardo A, Manganaro A, Benfenati E. A

New Structure-Activity Relationship (SAR) Model for Predicting Drug-Induced Liver Injury, Based on Statistical and Expert-Based Structural Alerts. Front Pharmacol 2016;7:442. doi:10.3389/FPHAR.2016.00442.

[36] Anderson AC. The Process of Structure-Based Drug Design. Chem Biol 2003;10:787–97. doi:10.1016/j.chembiol.2003.09.002.

[37] Wu L, Liu Z, Auerbach S, Huang R, Chen M, McEuen K, et al. Integrating Drug's Mode of Action into Quantitative Structure–Activity Relationships for Improved Prediction of Drug-Induced Liver Injury. J Chem Inf Model 2017;57:1000–6. doi:10.1021/acs.jcim.6b00719.

[38] Alves VM, Golbraikh A, Capuzzi SJ, Liu K, Lam WI, Korn DR, et al. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. J Chem Inf Model 2018;58:1214–23. doi:10.1021/acs.jcim.8b00124.

[39] Przybylak KR, Cronin MT. *In silico* models for drug-induced liver injury – current status. Expert Opin Drug Metab Toxicol 2012;8:201–17. doi:10.1517/17425255.2012.648613.

[40] Sun H, Xia M, Austin CP, Huang R. Paradigm Shift in Toxicity Testing and Modeling. AAPS J 2012;14:473–80. doi:10.1208/s12248-012-9358-1.

[41] Kumar A, Zhang KYJ. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. Front Chem 2018;6:315. doi:10.3389/fchem.2018.00315.

[42] Nikolova N, Jaworska J. Approaches to Measure Chemical Similarity– a Review. QSAR Comb Sci 2003;22:1006–26. doi:10.1002/qsar.200330831.

[43] Grouping of Chemicals: Chemical Categories and Read-Across - OECD. http://www.oecd.org/chemicalsafety/risk-assessment/groupingofchemicalschemicalcategoriesandread-across.htm (accessed November 2, 2019).

[44] File:Phenyl methylthiophene replacement.png - Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Phenyl_methylthiophene_replacement.png (accessed November 6, 2019).

[45] File:Classical bioisosteres 2.png - Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Classical_bioisosteres_2.png (accessed November 6, 2019).

[46] How can computational chemistry help find new drugs from old?. https://www.cresset-group.com/about/news/new-from-old/ (accessed September 9, 2019).

[47] Toque HA, Priviero FB, Zemse SM, Antunes E, Teixeira CE, Webb RC. Effect of the phosphodiesterase 5

inhibitors sildenafil, tadalafil and vardenafil on rat anococcygeus muscle: Functional and biochemical aspects. Clin Exp Pharmacol Physiol 2009;36:358–66. doi:10.1111/j.1440-1681.2008.05071.x.

[48] Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed November 3, 2019).

[49] Daylight Theory: SMILES. https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed November 3, 2019).

[50] Grace P, George H, Prachi P, Imran S. Navigating through the minefield of read-across tools: A review of in silico tools for grouping. Comput Toxicol (Amsterdam, Netherlands) 2017;3:1–18. doi:10.1016/j.comtox.2017.05.003.

[51] Broder AZ, Broder AZ. Some applications of Rabin's fingerprinting method. Seq II METHODS Commun Secur Comput Sci 1993:143--152.

[52] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63. doi:10.1016/J.YMETH.2014.08.005.

[53] Zhang C, Cheng F, Li W, Liu G, Lee PW, Tang Y. *In silico* Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. Mol Inform 2016;35:136–44. doi:10.1002/minf.201500055.

[54] Chen M, Hong H, Fang H, Kelly R, Zhou G, Borlak J, et al. Quantitative Structure-Activity Relationship Models for Predicting Drug-Induced Liver Injury Based on FDA-Approved Drug Labeling Annotation and Using a Large Collection of Drugs. Toxicol Sci 2013;136:242–9. doi:10.1093/toxsci/kft189.

[55] Liu Z, Shi Q, Ding D, Kelly R, Fang H, Tong W. Translating Clinical Findings into Knowledge in Drug Safety Evaluation - Drug Induced Liver Injury Prediction System (DILIps). PLoS Comput Biol 2011;7:e1002310. doi:10.1371/journal.pcbi.1002310.

[56] Muller C, Pekthong D, Alexandre E, Marcou G, Horvath D, Richert L, et al. Prediction of drug induced liver injury using molecular and biological descriptors. Comb Chem High Throughput Screen 2015;18:315–22.

[57] Liew CY, Lim YC, Yap CW. Mixed learning algorithms and features ensemble in hepatotoxicity prediction. J Comput Aided Mol Des 2011;25:855–71. doi:10.1007/s10822-011-9468-3.

[58] Rodgers AD, Zhu H, Fourches D, Rusyn I, Tropsha A. Modeling Liver-Related Adverse Effects of Drugs Using *k* Nearest Neighbor Quantitative Structure−Activity Relationship Method. Chem Res Toxicol 2010;23:724–32. doi:10.1021/tx900451r.

[59] Mulliner D, Schmidt F, Stolte M, Spirkl H-P, Czich A, Amberg A. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope.

Chem Res Toxicol 2016;29:757–67. doi:10.1021/acs.chemrestox.5b00465.

[60] Chen M, Bisgin H, Tong L, Hong H, Fang H, Borlak J, et al. Toward predictive models for drug-induced liver injury in humans: are we there yet? Biomark Med 2014;8:201–13. doi:10.2217/bmm.13.146.

[61] Ekins S, Williams AJ, Xu JJ. A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury. Drug Metab Dispos 2010;38:2302–8. doi:10.1124/dmd.110.035113.

[62] Fraser K, Bruckner DM, Dordick JS. Advancing Predictive Hepatotoxicity at the Intersection of Experimental, *in Silico* , and Artificial Intelligence Technologies. Chem Res Toxicol 2018;31:412–30. doi:10.1021/acs.chemrestox.8b00054.

[63] Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. Drug Discov Today 2016;21:648–53. doi:10.1016/j.drudis.2016.02.015.

[64] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44:D1075–9. doi:10.1093/nar/gkv1075.

[65] Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 2012;17:4791–810.

doi:10.3390/molecules17054791.

[66] Norinder U, Carlsson L, Boyer S, Eklund M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. J Chem Inf Model 2014;54:1596–603. doi:10.1021/ci5001168.

[67] Varsou D-D, Tsiliki G, Nymark P, Kohonen P, Grafström R, Sarimveis H. toxFlow: A Web-Based Application for Read-Across Toxicity Prediction Using Omics and Physicochemical Data. J Chem Inf Model 2018;58:543–9. doi:10.1021/acs.jcim.7b00160.

[68] Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, et al. Integrative chemical-biological read-across approach for chemical hazard classification. Chem Res Toxicol 2013;26:1199–208. doi:10.1021/tx400110f.

[69] Helman G, Shah I, Patlewicz G. Extending the Generalised Read-Across approach (GenRA): A systematic analysis of the impact of physicochemical property information on read-across performance. Comput Toxicol 2018;8:34–50. doi:10.1016/J.COMTOX.2018.07.001.

[70] Shah I, Liu J, Judson RS, Thomas RS, Patlewicz G. Systematically evaluating read-across prediction and performance using a local validity approach characterized by chemical structure and bioactivity information. Regul Toxicol Pharmacol 2016;79:12–24.

doi:10.1016/J.YRTPH.2016.05.008.

[71] Schultz TW, Richarz A-N, Cronin MTD. Assessing uncertainty in read-across: Questions to evaluate toxicity predictions based on knowledge gained from case studies. Comput Toxicol 2019;9:1–11. doi:10.1016/J.COMTOX.2018.10.003.

[72] Zhu H, Bouhifd M, Donley E, Egnash L, Kleinstreuer N, Kroese ED, et al. Supporting read-across using biological data. ALTEX 2016;33:167–82. doi:10.14573/altex.1601252.

[73] Ball N, Cronin MTD, Shen J, Blackburn K, Booth ED, Bouhifd M, et al. Toward Good Read-Across Practice (GRAP) guidance. ALTEX 2016;33:149–66. doi:10.14573/altex.1601251.

[74] Stuard SB, Heinonen T. Relevance and Application of Read-Across - Mini Review of European Consensus Platform for Alternatives and Scandinavian Society for Cell Toxicology 2017 Workshop Session. Basic Clin Pharmacol Toxicol 2018;123:37–41. doi:10.1111/bcpt.13006.

[75] RepDose Database Fraunhofer ITEM QSAR. https://repdose.item.fraunhofer.de/ (accessed September 9, 2019).

[76] AMBIT – Cheminformatics data management system – Cefic-Lri. http://cefic-lri.org/toolbox/ambit/ (accessed September 9, 2019).

[77] US EPA O. Distributed Structure-Searchable Toxicity

(DSSTox) Database.

[78] SciFinder^{n} | CAS. https://www.cas.org/products/scifinder-n (accessed September 9, 2019).

[79] Leadscope, Inc. : Leadscope - Chemoinformatics Platform for Drug Discovery. http://www.leadscope.com/ (accessed September 9, 2019).

[80] ChemIDplus Advanced - Chemical information with searchable synonyms, structures, and formulas. https://chem.nlm.nih.gov/chemidplus/chemidheavy.jsp (accessed September 9, 2019).

[81] US EPA O. Analog Identification Methodology (AIM) Tool.

[82] Yordanova D, Kuseva C, Tankova K, Pavlov T, Chankov G, Chapkanov A, et al. Using metabolic information for categorization and read-across in the OECD QSAR Toolbox. Comput Toxicol 2019;12:100102. doi:10.1016/j.comtox.2019.100102.

[83] Yordanova D, Schultz TW, Kuseva C, Tankova K, Ivanova H, Dermen I, et al. Automated and standardized workflows in the OECD QSAR Toolbox. Comput Toxicol 2019;10:89–104. doi:10.1016/j.comtox.2019.01.006.

[84] Yordanova D, Schultz TW, Kuseva C, Ivanova H, Pavlov T, Chankov G, et al. Alert performance: A new functionality in the OECD QSAR Toolbox. Comput

Toxicol 2019;10:26–37. doi:10.1016/j.comtox.2018.12.003.

[85] Kuseva C, Schultz TW, Yordanova D, Ivanova H, Tankova K, Pavlov T, et al. Category consistency in the OECD QSAR Toolbox: Assessment and reporting tool to justify read-across. Comput Toxicol 2019;11:65–71. doi:10.1016/j.comtox.2019.03.002.

[86] Kuseva C, Schultz TW, Yordanova D, Tankova K, Kutsarova S, Pavlov T, et al. The implementation of RAAF in the OECD QSAR Toolbox. Regul Toxicol Pharmacol 2019;105:51–61. doi:10.1016/j.yrtph.2019.03.018.

[87] US EPA O. OncoLogic™ - A Computer System to Evaluate the Carcinogenic Potential of Chemicals.

[88] Toxtree tool | EU Science Hub. https://ec.europa.eu/jrc/en/scientific-tool/toxtree-tool (accessed September 9, 2019).

[89] Aggregated Computational Toxicology Resource (ACTOR). https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7B2CD5A640-2523-4628-868B-DC8E6460125F%7D (accessed September 9, 2019).

[90] G L. RDKit, version x.x: open-source cheminformatics. http://www.rdkit.org (accessed April 20, 2019).

[91] RDKit fingerprints. https://www.rdkit.org/docs/GettingStartedInPython.html#fingerprinting-and-molecular-similarity (accessed April

20, 2019).

[92] RDKit molecular descriptors. https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors.html#module-rdkit.Chem.Descriptors (accessed April 20, 2019).

[93] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 2015;7:20. doi:10.1186/s13321-015-0069-3.

[94] Freakonomics: Correlation ≠ Causation (Money can't buy elections) | The Graphic Recorder. http://www.thegraphicrecorder.com/2012/01/18/freakonomics-correlation-≠-causation-money-cant-buy-elections/ (accessed September 9, 2019).

[95] Bishop CM. Pattern Recognition and Machine Learning. Cambridge: Springer; 2006.

[96] Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Cent Eur J Oper Res 2018;26:135–59. doi:10.1007/s10100-017-0479-6.

[97] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst 1987;2:37–52. doi:10.1016/0169-7439(87)80084-9.

[98] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 2001;58:109–30. doi:10.1016/S0169-7439(01)00155-1.

[99] Breiman L. Random Forests. vol. 45. 2001.

[100] Leek JT, Peng RD. Opinion: Reproducible research can

still be wrong: adopting a prevention approach. Proc Natl Acad Sci U S A 2015;112:1645–6. doi:10.1073/pnas.1421412111.

[101] File:Boyle air pump.jpg - Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Boyle_air_pump.jpg (accessed September 9, 2019).

[102] Lockyer N. Nature. [Macmillan Journals Ltd., etc.];

[103] Conda 4.7.12. https://docs.conda.io/projects/conda/en/latest/index.html (accessed October 30, 2019).

[104] kpinto-gil/Thesis: Development and validation of pharmacoinformatic similarity-based tools for safety assessment of chemicals. https://github.com/kpinto-gil/Thesis (accessed November 2, 2019).

[105] Python, version 3.6: A dynamic, open source programming language. Python Software Foundation. https://www.python.org/ (accessed April 20, 2019).

[106] Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. 2011.

[107] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. vol. 12. 2011.

[108] Conformal prediction framework. https://github.com/josecarlosgomezt/nonconformist/ (accessed April 20, 2019).

[109] SciPy: version 1.2.1. https://www.scipy.org/ (accessed November 2, 2019).

[110] NumPy: version 1.16.2.. https://numpy.org/ (accessed November 2, 2019).

[111] F.Atkinson. Standardiser, 2014 2014. https://github.com/flatkinson/standardiser (accessed April 20, 2019).

[112] Swain M. MolVS, version 0.1.1.: Molecule Validation and Standardization 2019. https://molvs.readthedocs.io/en/latest/ (accessed April 20, 2019).

[113] Milletti F, Storchi L, Sforna G, Cruciani G. New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. J Chem Inf Model 2007;47:2172–81. doi:10.1021/ci700018y.

[114] MoKa, eTOX version 3.0. Molecular Discovery, 2017. 2017. http://www.moldiscovery.com/software/moka/ (accessed April 20, 2019).

[115] Pastor M. Flame. https://github.com/phi-grib/flame (accessed April 20, 2019).

[116] Gomez Tamayo JC, Pinto-Gil K. Structure Resolver 2019.

[117] pickle — Python object serialization — Python 3.7.3 documentation. https://docs.python.org/3/library/pickle.html (accessed May 16, 2019).

[118] Anatomical Therapeutic Chemical (ATC) Classification System ontology 2018. https://bioportal.bioontology.org/ontologies/ATC

(accessed April 20, 2019).

[119] WHO Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment, 2019. Oslo, 2018. https://www.whocc.no/atc_ddd_index_and_guidelines/guidelines/ (accessed April 20, 2019).

[120] Side effect Resorce (SIDER), version 4.1. http://sideeffects.embl.de/ (accessed April 20, 2019).

[121] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res 2016;44:D1075–9. doi:10.1093/nar/gkv1075.

[122] Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 2010;6. doi:10.1038/msb.2009.98.

[123] Medical Dictionary for Regulatory Activities Terminology (MedDRA) ontology 2018. https://bioportal.bioontology.org/ontologies/MEDDRA (accessed April 20, 2019).

[124] MedDRA. https://www.meddra.org/ (accessed April 20, 2019).

[125] Greene N, Fisk L, Naven RT, Note RR, Patel ML, Pelletier DJ. Developing Structure−Activity Relationships for the Prediction of Hepatotoxicity. Chem Res Toxicol 2010;23:1215–22. doi:10.1021/tx1000865.

[126] O'Brien PJ, Irwin W, Diaz D, Howard-Cofield E, Krejsa CM, Slaughter MR, et al. High concordance of drug-induced human hepatotoxicity with in vitro cytotoxicity

measured in a novel cell-based model using high content screening. Arch Toxicol 2006;80:580–604. doi:10.1007/s00204-006-0091-3.

[127] E J, E O, P P, Others. SciPy: Open Source Scientific Tools for Python 2001.

[128] Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLOS Comput Biol 2019;15:e1006907. doi:10.1371/journal.pcbi.1006907.

[129] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 2001;58:109–30. doi:10.1016/S0169-7439(01)00155-1.

[130] Shafer G, Vovk V. A Tutorial on Conformal Prediction. vol. 9. 2008.

[131] Mulliner D, Schmidt F, Stolte M, Spirkl H-P, Czich A, Amberg A. Computational Models for Human and Animal Hepatotoxicity with a Global Application Scope. Chem Res Toxicol 2016;29:757–67. doi:10.1021/acs.chemrestox.5b00465.

[132] Chen M, Borlak J, Tong W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. Hepatology 2013;58:388–96. doi:10.1002/hep.26208.

[133] Liu Z, Shi Q, Ding D, Kelly R, Fang H, Tong W. Translating Clinical Findings into Knowledge in Drug Safety Evaluation - Drug Induced Liver Injury Prediction System (DILIps). PLoS Comput Biol 2011;7:e1002310. doi:10.1371/journal.pcbi.1002310.

[134] Russell J, Cohn R. Mathhews Correlation Coefficient. 2012.

[135] Daneshian M, Kamp H, Hengstler J, Leist M, van de Water B. Highlight report: Launch of a large integrated European in vitro toxicology project: EU-ToxRisk. Arch Toxicol 2016;90:1021–4. doi:10.1007/s00204-016-1698-7.

[136] Raies AB, Bajic VB. *In silico* toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 2016;6:147–72. doi:10.1002/wcms.1240.

[137] Kavlock R. The future of toxicity testing—The NRC vision and the EPA's ToxCast program national center for computational toxicology. Neurotoxicol Teratol 2009;31:237. doi:10.1016/J.NTT.2009.04.007.

[138] Hartung T, FitzGerald RE, Jennings P, Mirams GR, Peitsch MC, Rostami-Hodjegan A, et al. Systems Toxicology: Real World Applications and Opportunities. Chem Res Toxicol 2017;30:870–82. doi:10.1021/acs.chemrestox.7b00003.

[139] Carrió P, Sanz F, Pastor M. Toward a unifying strategy for the structure-based prediction of toxicological endpoints. Arch Toxicol 2015. doi:10.1007/s00204-015-1618-2.

[140] Abboud G, Kaplowitz N. Drug-Induced Liver Injury. Drug Saf 2007;30:277–94. doi:10.2165/00002018-200730040-00001.

[141] Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W. FDA-approved drug labeling for the study of drug-induced liver injury. Drug Discov Today 2011;16:697–703. doi:10.1016/j.drudis.2011.05.007.

[142] Kotsampasakou E, Montanari F, Ecker GF. Predicting drug-induced liver injury: The importance of data curation. Toxicology 2017;389:139–45. doi:10.1016/j.tox.2017.06.003.

ANNEX

Annex Publications

Generating modelling data from repeat-dose toxicity reports

Oriol López-Massaguer, **Kevin Pinto-Gil**, Ferran Sanz, Alexander Amberg, Lennart Anger, Manuela Stolte, Carlo Ravagli, Philippe Marc and Manuel Pastor. Toxicol Sci. 2018 Mar 1;162(1):287-300.

doi: https://doi.org/10.1093/toxsci/kfx254

Contributions:

- Dataset structure normalization, Pre-Clinical liver toxicity predictive models were built: 1) Degeneration, 2) Inflammation and 3) Non-neoplasic proliferative lesions. The statistical analysis. Manuscript revision.

Annex Figures

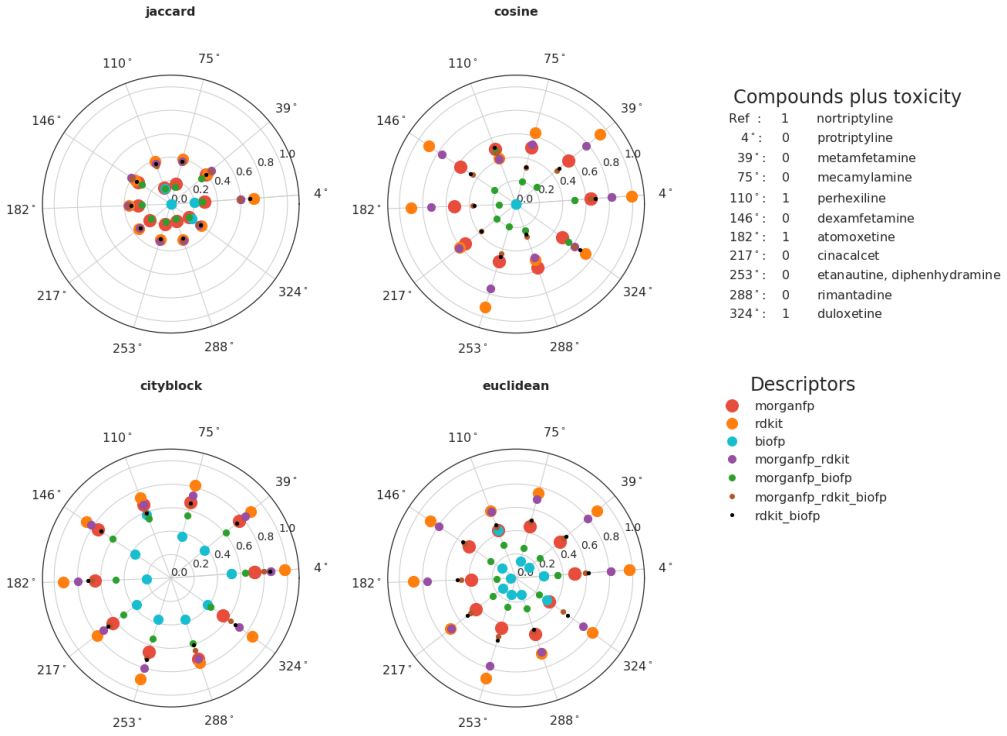


Fig. S1 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using RDKit Descriptors. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

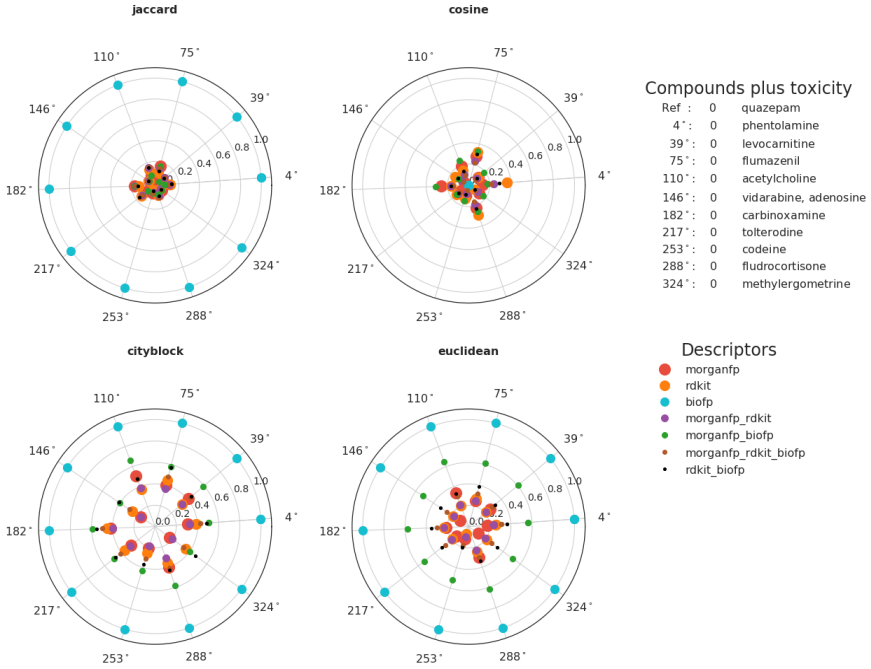


Fig. S2 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using Biofingerprints. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

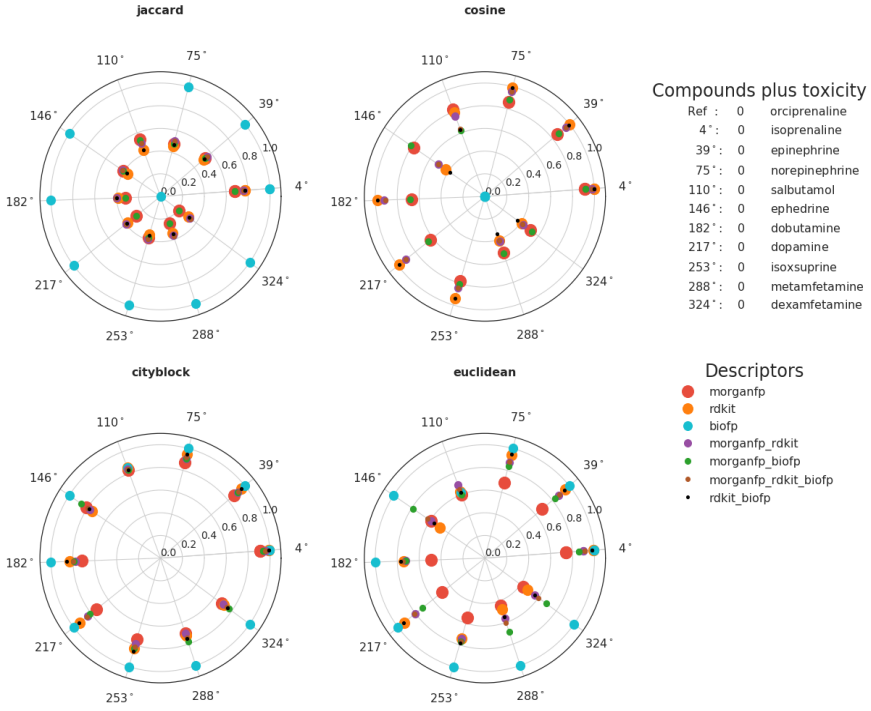


Fig. S3 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using morgan fingerprints plus RDkit descriptors combination. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

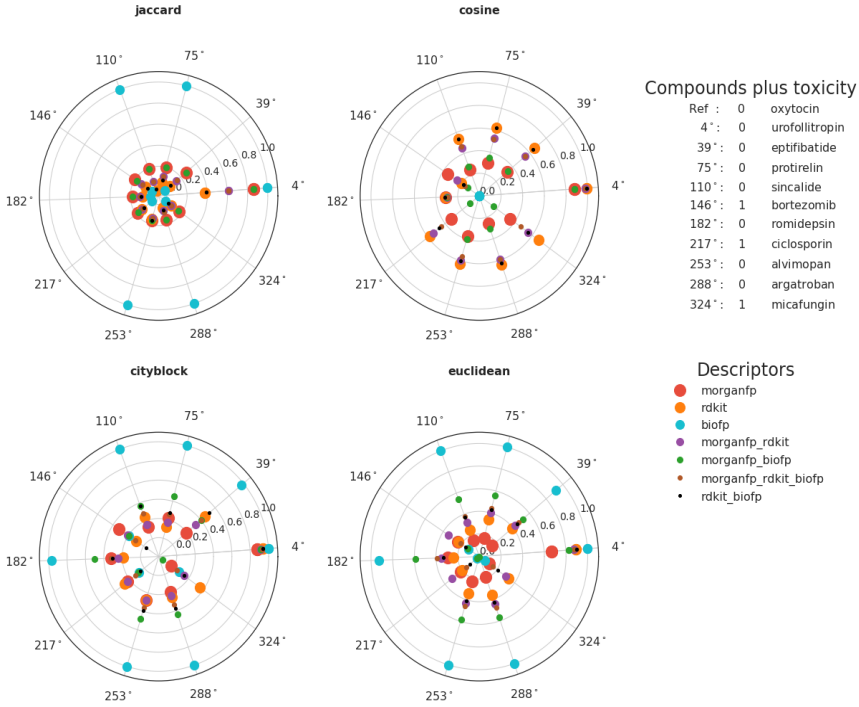


Fig. S4 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using morgan fingerprints plus biofingerprints combination. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

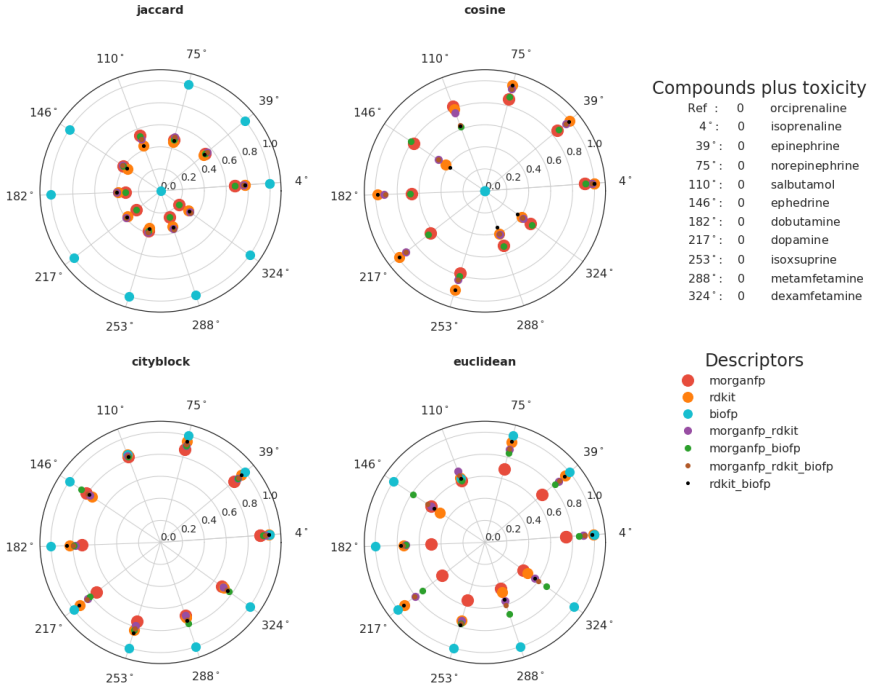


Fig. S5 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using morgan fingerprints + RDkit descriptors + biofingerprints combination. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

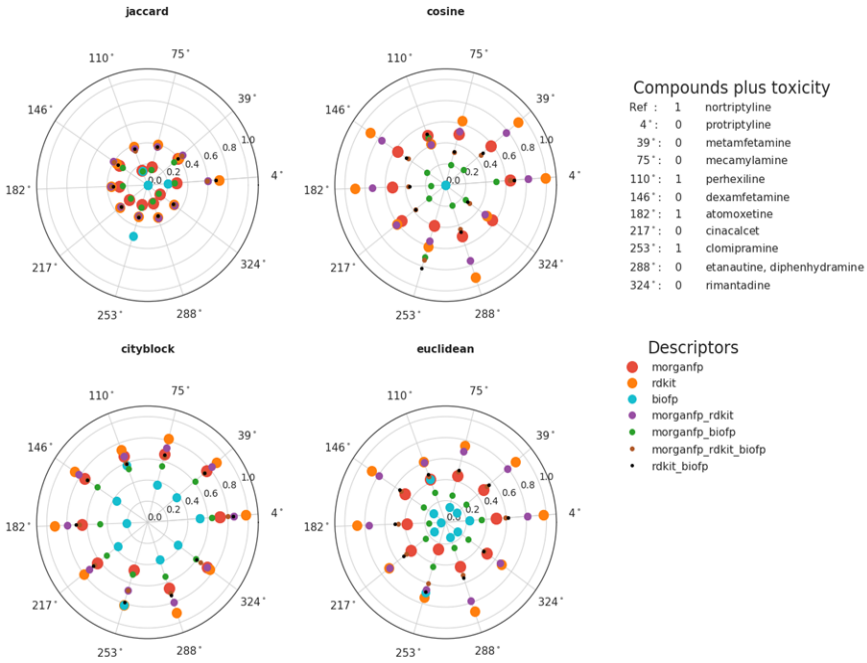


Fig. S6 **| Radial plot Biased RA using combination of descriptors from Table 7.** These compounds are ordered by Jaccard metric using RDkit descriptors + biofingerprints combination. Morgan fingerprints (morganfp) in red, rdkit in orange, Biofingerprints (biofp) in blue, morganfp plus rdkit in purple, morganfp plus biofp in green, morganfp plus rdkit plus biofp in grey and rdkit plus biofp in black.

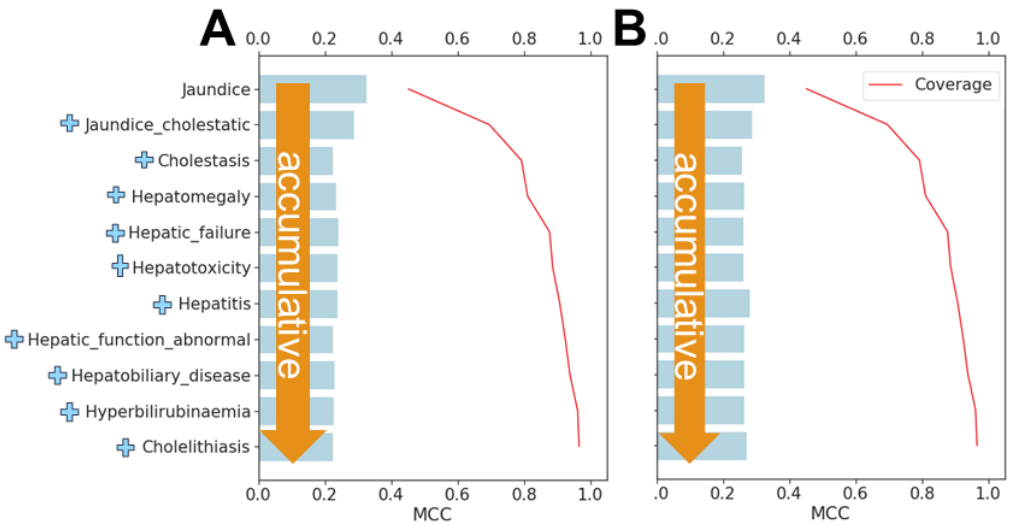


Fig. S7 **| Optimization AE models Performance with Mulliner Dataset.** A) General Rule. B) Majority Voting rule. MCC is shown in X axis and Adverse effects optimization models. Red line means the coverage molecules taken into accounts in the model building.

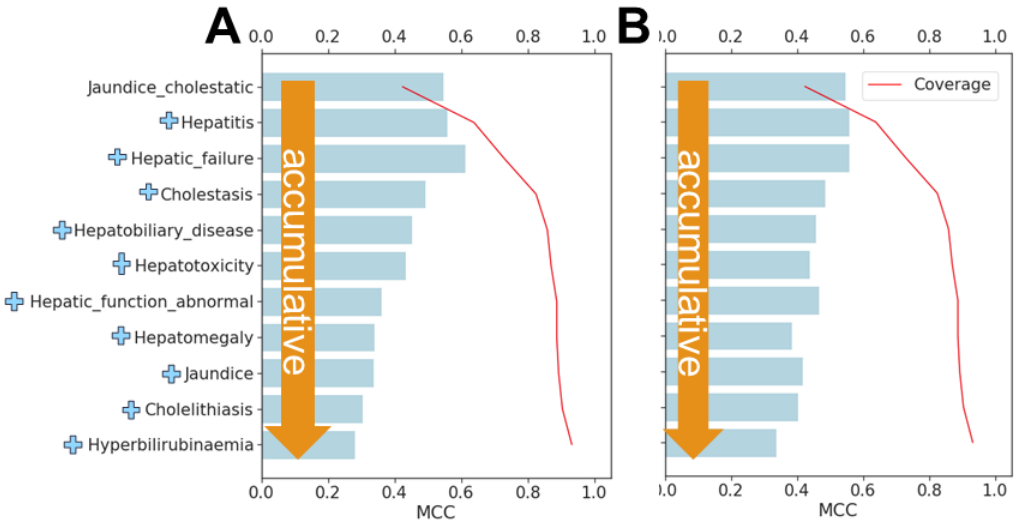


Fig. S8 **| Optimization AE models Performance with MostDILIrank Dataset.** A) General Rule. B) Majority Voting rule. MCC is shown in X axis and Adverse effects optimization models. Red line means the coverage molecules taken into accounts in the model building.

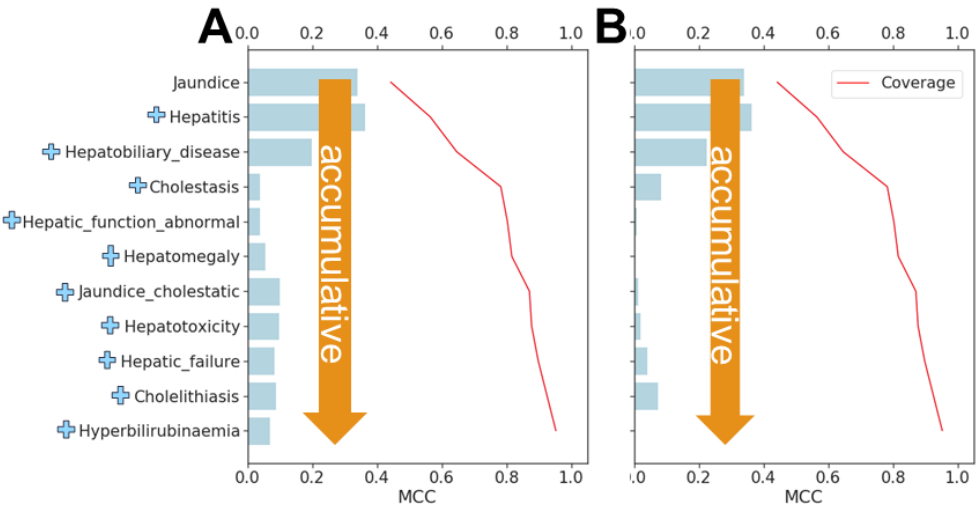


Fig. S9 **| Optimization AE models Performance with Pfizer Dataset.** A) General Rule. B) Majority Voting rule. MCC is shown in X axis and Adverse effects optimization models. Red line means the coverage molecules taken into accounts in the model building.

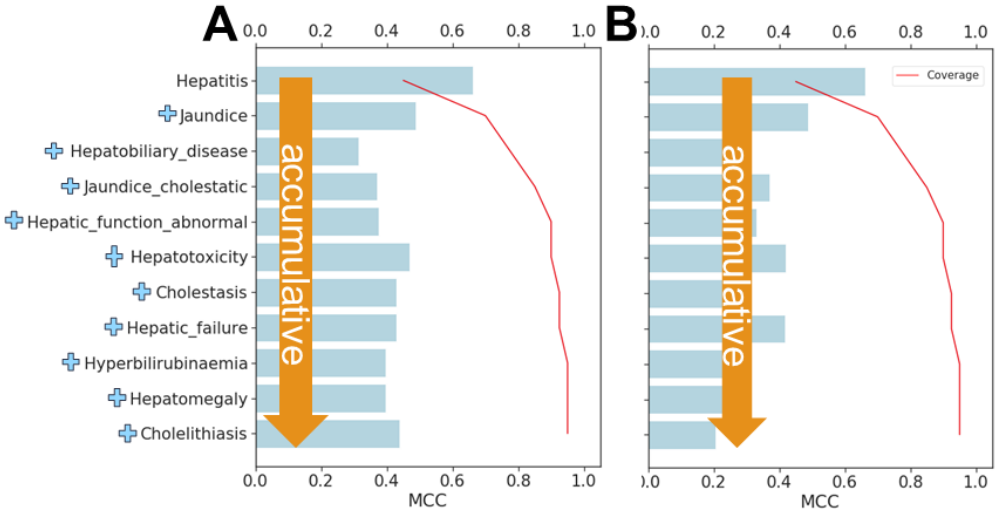


Fig. S10 **| Optimization AE models Performance with O'Brien Dataset.** A) General Rule. B) Majority Voting rule. MCC is shown in X axis and Adverse effects optimization models. Red line means the coverage molecules taken into accounts in the model building.

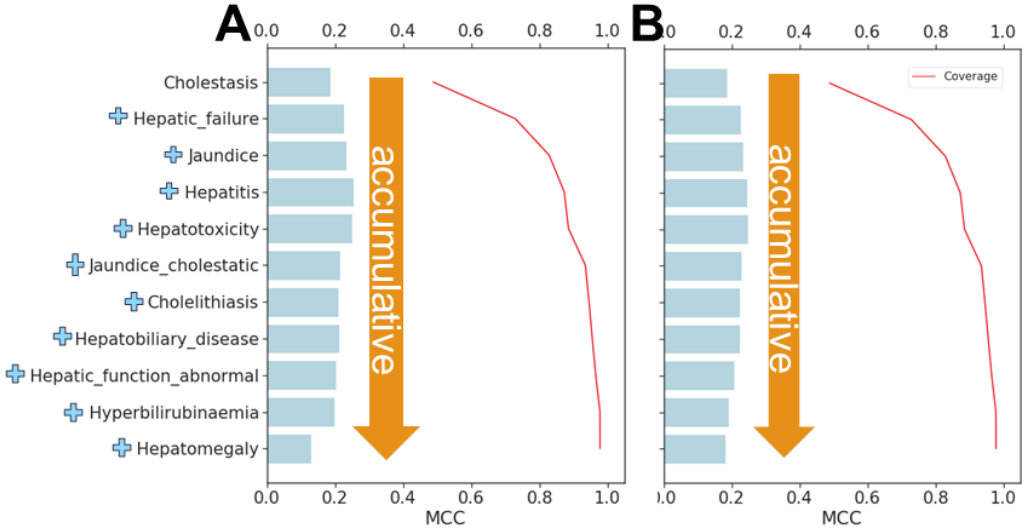


Fig. S11 **| Optimization AE models Performance with DrugBank plus all DILI sets combination.** A) General Rule. B) Majority Voting rule. MCC is shown in X axis and Adverse effects optimization models. Red line means the coverage molecules taken into accounts in the model building.

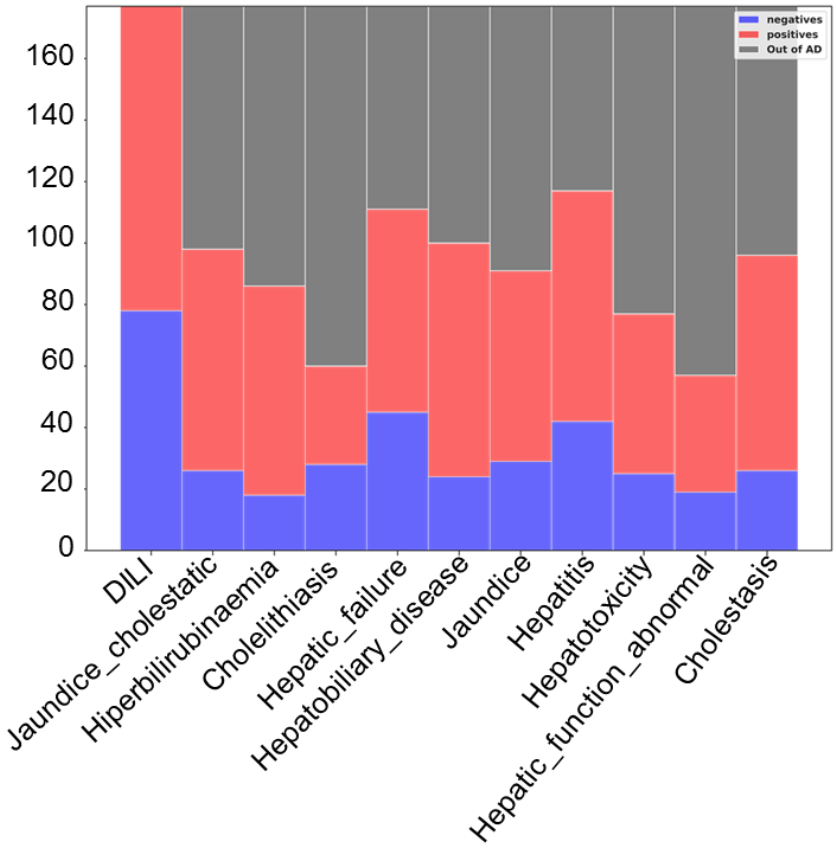


Fig. S12 **| AEs predictions distribution in DILIrank.** X axis Adverse effects expert models. Y axis number of compounds with DILI prediction. Blue colour means DILI negative, Red colour means DILI positive, and Grey colour means DILI could not be predicted because is out of the AD.

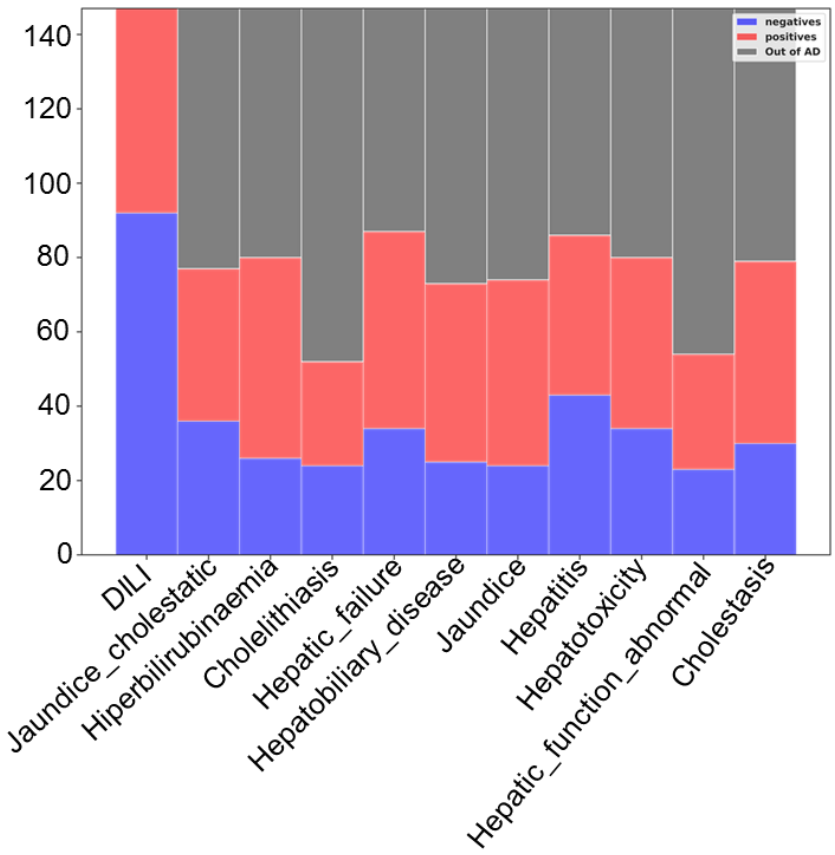


Fig. S13 **| AEs predictions distribution in Pfizer.** X axis Adverse effects expert models. Y axis number of compounds with DILI prediction. Blue colour means DILI negative, Red colour means DILI positive, and Grey colour means DILI could not be predicted because is out of the AD.

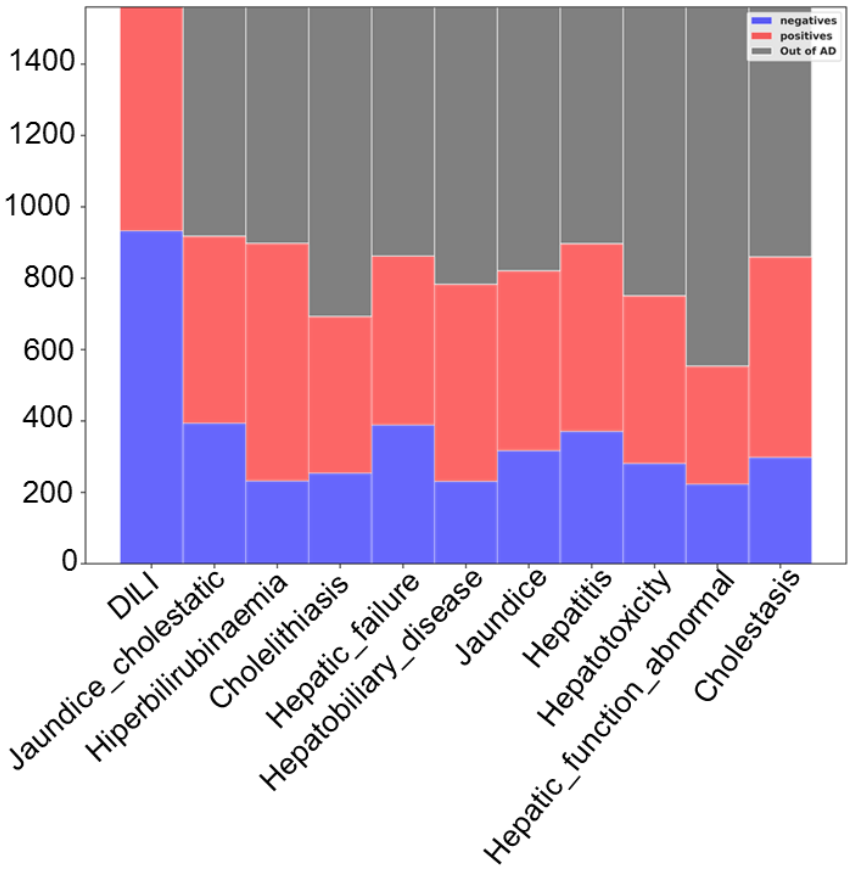


Fig. S14 **| AEs predictions distribution in Mulliner.** X axis Adverse effects expert models. Y axis number of compounds with DILI prediction. Blue colour means DILI negative, Red colour means DILI positive, and Grey colour means DILI could not be predicted because is out of the AD.

Annex Tables

Some tables cannot be shown here because of the extension. In any case, an excel document containing all the excel Tables can be downloaded from [104].

S1 Table. Cross-validation statistics of AE models. Detailed cross-validation statistics including other AEs not satisfying quality requirements as well as models built under non-conformal framework.

ML Method	Cross-Validation method	mols	Endpoint	TP	TN	FP	FN	Sensitivity	Specificity	MCC	Conformal coverage	Conformal accuracy	Conformal significance
Random Forest Conformal	5 kfold	937	Jaundice cholestatic	48	42	15	11	0.81	0.74	0.55	0.54	0.78	0.2
			Hyperbilirubinaemia	50	49	10	20	0.71	0.83	0.54	0.6	0.77	
			Hepatic failure	94	83	27	27	0.78	0.75	0.53	0.63	0.77	
			Cholelithiasis	20	20	6	7	0.74	0.77	0.51	0.41	0.75	
			Cholestasis	40	34	14	14	0.74	0.71	0.45	0.55	0.73	

ML Method	Cross-Validation method	mols	Endpoint	TP	TN	FP	FN	Sensitivity	Specificity	MCC	Conformal coverage	Conformal accuracy	Conformal significance
			Hepatitis	135	116	50	48	0.74	0.7	0.44	0.55	0.72	
			Jaundice	116	102	52	36	0.76	0.66	0.43	0.55	0.71	
			Hepatobiliary disease	103	74	39	34	0.75	0.65	0.41	0.48	0.71	
			Hepatomegaly	27	22	11	12	0.69	0.67	0.36	0.44	0.68	
			Hepatotoxicity	37	36	17	19	0.66	0.68	0.34	0.61	0.67	
			Hepatic function abnormal	54	41	24	26	0.68	0.63	0.31	0.46	0.66	
			Ascites	20	16	10	11	0.65	0.62	0.26	0.65	0.63	
			Liver injury	26	15	13	11	0.7	0.54	0.24	0.32	0.63	
			Hepatic necrosis	27	13	11	12	0.69	0.54	0.23	0.44	0.63	

ML Method	Cross-Validation method	mols	Endpoint	TP	TN	FP	FN	Sensitivity	Specificity	MCC	Conformal coverage	Conformal accuracy	Conformal significance
			Hepatocellular injury	32	21	19	20	0.62	0.53	0.14	0.37	0.58	
			Liver disorder	11	8	9	10	0.52	0.47	-0.01	0.25	0.5	
			Cholecystitis	8	15	8	16	0.33	0.65	-0.02	0.39	0.49	
			Hepatitis cholestatic	11	2	4	6	0.65	0.33	-0.02	0.24	0.57	
			Foetor hepaticus	7	5	12	9	0.44	0.29	-0.27	0.28	0.36	
Random Forest	5 kfold	937	Hepatic failure	88	92	90	94	0.48	0.51	-0.01			
			Hyperbilirubinaemia	50	51	56	57	0.47	0.48	-0.06			
			Cholestasis	46	40	53	47	0.49	0.43	-0.08			
			Jaundice cholestatic	46	52	56	62	0.43	0.48	-0.09			
			Hepatitis	131	148	168	185	0.41	0.47	-0.12			

ML Method	Cross-Validation method	mols	Endpoint	TP	TN	FP	FN	Sensitivity	Specificity	MCC	Conformal coverage	Conformal accuracy	Conformal significance
			Ascites	16	22	22	28	0.36	0.5	-0.14			
			Hepatic necrosis	31	28	43	40	0.44	0.39	-0.17			
			Jaundice	106	126	154	174	0.38	0.45	-0.17			
			Cholelithiasis	27	25	40	38	0.42	0.38	-0.2			
			Hepatotoxicity	33	38	52	57	0.37	0.42	-0.21			
			Hepatic function abnormal	62	61	98	97	0.39	0.38	-0.23			
			Hepatomegaly	33	29	52	48	0.41	0.36	-0.23			
			Hepatobiliary disease	90	101	158	169	0.35	0.39	-0.26			
			Liver injury	43	26	77	60	0.42	0.25	-0.33			
			Cholecystitis	17	21	40	44	0.28	0.34	-0.38			

ML Method	Cross-Validation method	mols	Endpoint	TP	TN	FP	FN	Sensitivity	Specificity	MCC	Conformal coverage	Conformal accuracy	Conformal significance
			Hepatocellular injury	42	36	90	84	0.33	0.29	-0.38			
			Hepatitis cholestatic	8	14	33	39	0.17	0.3	-0.54			
			Liver disorder	14	19	56	61	0.19	0.25	-0.56			
			Foetor hepaticus	7	19	40	52	0.12	0.32	-0.57			

S2 Table. PLS AE scores. Higher PLS scores correspond to AEs more related to DILI outcome.

S3 Table. Real AEs and DILI correlation matrix. Correlation between adverse effects and DILI outcome. Compounds present in both SIDER and DILI labelled datasets.
S4 Table. Predicted AEs and DILI correlation matrix. Correlation between predicted adverse effects and DILI outcome. Compounds do not present in SIDER.

S5 Table. QSAR DILI models performance. Includes quality statistics for DILI QSAR models built using the DILI databases considered in this work. The external validation for a given model is calculated by the prediction of DILI in remaining datasets.

S6 Table. Inconsistent DILI labelling. Includes compounds with different assigned DILI outcome in DILIrank and Mulliner datasets.

S7 Table. Model documentation. Includes detailed information on the models built in this work.