Towards uncoding hepatotoxicity of approved drugs through navigation of multiverse and consensus chemical spaces

Edgar López-López[1,2*] José L. Medina-Franco[2*]

[1] *DIFACQUIM research group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City 04510, Mexico*

[2] *Department of Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), Mexico City 07360, Mexico*

**Contact authors: E-mail:elopez.lopez@cinvestav.mx; medinajl@unam.mx*

Abstract: Drug-induced liver injury (DILI) is the principal reason for failure in developing drug candidates. It is the most common reason to withdraw from the market after a drug has been approved for clinical use. Therefore, a current challenge is enhancing the accuracy of DILI events' predictive models. In this context, data from animal models, liver function tests, and chemical properties could complement each other to understand DILI events better and prevent them. Since the chemical space concept improves decision-making drug design related to the prediction of structure-property relationships, side effects, and polypharmacology drug activity (uniquely mentioning the most recent advances), it is an attractive approach to combining different phenomena influencing DILI events (e.g., individual “chemical spaces”) and exploring all events simultaneously in an integrated analysis of the DILI-relevant chemical space. However, currently, no systematic methods allow the fusion of a collection of different chemical spaces to collect different types of data on a unique chemical space representation, namely "consensus chemical space." This study is the first report that uses the data fusion concept to combine different chemical spaces to facilitate the analysis and prediction of DILI-related events. The present work remarks the importance of analyzing together *in vitro* and chemical data (e.g., topology, bond order, atom types, presence of rings, ring sizes, and aromaticity of compounds encoded on RDKIT fingerprints). These properties could be aimed at improving the understanding of DILI events.

Keywords: clustering, chemoinformatics, consensus chemical space, data fusion, drug design, drug-induced liver injury, multi-objective optimization, unsupervised learning.

Introduction

Drug-induced liver injury (DILI) is one of the most frequent reasons to stop the drug candidate optimization process (around 67% of these optimizations have been stopped for this issue), and it is the most common feature related to post-marketing withdrawals [1]. For this reason, a current challenge is to enhance the understanding of DILI events. In this context, the current non-multidisciplinary approaches to studying hepatotoxic activity have not been exploiting and combining the large diversity of information (in silico, in vitro, in vivo, and clinical data) available to study this endpoint [2,3].

Recent studies have demonstrated that combining different data types increased the description and prediction of DILI events. For example, He et al. demonstrated that the combination of physicochemical and topological descriptors improved the accuracy of predictive DILI models [4]. Thakkar et al. remarked that the compounds associated with DILI events could be classified using mainly by anatomical (e.g. drugs used against the nervous system, anti-infectives for systemic use, antineoplastic immunomodulating agents, alimentary tract, and metabolism agents) and therapeutical features (e.g. drugs that act as antidepressants, anti-inflammatory, antirheumatic and antivirals products) [5]. Furthermore, a recent review by Vall et al. described the potential of artificial intelligence (AI) methods to predict liver injuries emphasizing that the combination of chemical structures, gene expression, *in vitro*, *in vivo*, and imaging assays could be used to decode the side effects of drugs [6]. These recent findings encourage the development of novel methodologies to simultaneously study a large diversity of data to predict DILI events. The next logical question is, “what kind of data and what type of data combinations could help to improve the classification and description of DILI-associated compounds?

In drug design and development, visualization methods are key resources in data mining and information extraction from constantly increasing data sets. Indeed, visualization methods are important tools for rationalizing and interpreting experimental and calculated data [7]. A key example is the chemical space concept, defined as “an M-dimensional cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors” [8]. Thus, chemical space allows the simultaneous study of different data types, such as structural, chemical, physicochemical, biological, clinical, and/or post-market data, to name a few examples. Since the chemical space depends directly on the descriptors used to define the M-dimensional cartesian space, it is important to mention that it is possible for the coexistence of parallel (or alternative) chemical spaces for the same set of molecules, namely, a multiverse chemical space. Also, it is possible to combine the alternative

chemical spaces to create a single “consensus” chemical space [9]. The chemical space application has demonstrated improvement in drug design, making decisions related to the prediction of structure-properties relationships (SPR), side effects, and polypharmacology drug activity, to mention a few of the most recent advances [10].

In this regard, data fusion methods allow putting multiple data observations or calculations (descriptors) together to increase the consistency and confidence of the information derived from the data [11]. Data fusion was developed initially to improve similarity searching. It is a crucial concept that demonstrates its utility to increase the predictability of drug design models against different endpoints (e.g., properties, bioactivity, biological pathways, -omics relationships, etc.) from a huge data diversity such as structural, physicochemical, spectrometry, bioactivity, transcriptomic, imaging, histological data, etc. [12-16].

The present work aims to improve understanding of DILI events through a novel integration of data fusion concepts using chemical, physicochemical, and biological data, consensus chemical space, and chemical multiverses.

Methodology

Data set construction and curation

The dataset was constructed by data deposited on public databases (DrugBank [17] and ChEMBL v.30 [18]) and bibliographic data collected by X. Liu et al. [19] and S.Thakkar et al. [5]. A total of 2,309 drugs approved for clinical use were classified according to their DILI-associated events: if each compound has been associated with fatal hepatic adverse drug reaction, liver failure, liver transplantation, jaundice, bilirubin, and liver enzymes increase, hepatomegaly, hepatitis, and/or hepatotoxicity. Only 186 (~ 8%) of the approved drugs were associated with DILI events.

In parallel, a total of 190,068 compounds tested against the hepatic cell lines HepG2 and Huh7 (ChEMBL ID: 3307718 and 3307515, respectively) and/or the clinically important cytochromes CYP1A2, CYP2A6, CYP2C9, CYP2D6, CYP3A4 (ChEMBL ID: CHEMBL3356, CHEMBL5282, CHEMBL3397, CHEMBL289, CHEMBL340, respectively) were retrieved from ChEMBL v.30.

The approved drugs dataset associated with DILI events and the dataset with cell-based and cytochrome activity data were merged based on their canonical SMILES. Only 471 compounds (~20% of 2309 approved drugs) are associated with cell-hepatotoxicity activity (HepG2 and/or Huh7) and/or

cytochrome inhibition (CYP1A2, CYP2A6, CYP2C9, CYP2D6, and/or CYP3A4). The KNIME software [20] was used to assemble, merge, and curate the datasets. The KNIME workflow is available in the Supplementary material section (file Multitiverse_DataFusion.knwf).

Descriptor calculation

Based on the published findings that suggest that the combination of chemical, physicochemical, and structural/topological descriptors improves the classification of DILI-related compounds [4,6], these types of descriptors were calculated in this work.

To describe the chemical and physicochemical context of the data set DataWarrior v. 5.5.0 software [21] was used to calculate the number of H-donor bonds, number of H-acceptor bonds, number of rotatable bonds, molecular weight, cLogP, and topological surface area (TPSA) for each compound on the data set. Additionally, three types of structural/topological descriptors e.g Molecular ACCes System (MACCS) Keys, RDKit, and ECFP4 fingerprints were computed using the RDKit [22] module implemented by python programming language.

Chemical space construction

From the dataset with 471 compounds associated with DILI reports (available in the Supplementary material: “DB_ConsensusChemSpace_DILI.csv”), hepatotoxicity cell activity and cytochrome inhibition data were constructed and analyzed in their different chemical space representations based on chemical, physicochemical, structural, and *in vitro* (bioactivity) profile: cytochrome and hepatotoxic cell activity. The implementation of different chemical representations to analyze chemical spaces has been recently termed multiverse chemical space analysis [9].

Before combining all bi-dimensional representations of chemical spaces, each representation was constructed using KNIME software v. 4.3.4 and the module “t-SNE” which is widely used to reduce high-dimensional data to two dimensions [23]. In t-SNE, the parameters were: 1,000 iterations, 0.5 theta value, and 30 perplexity values to generate t-SNE 1 and t-SNE 2 coordinates.

Assignment of weights to each chemical space

Previous to data fusion it is important to establish the relative importance (weights) of each variable (chemical space coordinates, i.e., t-SNE coordinates) to describe the studied data (chemical structures associated with DILI reports). For this reason, we propose a simple metric, Quadrant weight (QW) -

Equation (**1**), that allows uncovering specific regions on the chemical spaces (2D plot coordinates) that are enriched with compounds associated with DILI events:

$$(\mathbf{1})\ QW = (A * 100) / n) + ((NA * 100/n)) / 2$$

where “A” and “NA” represent the number of compounds associated or not with DILI events in a specific quadrant of the chemical space plot, respectively; “n” is the total number of compounds contained in the database. A positive QW value suggests that a region of the chemical space (2D plot coordinates) is enriched with positive DILI compounds.

For this work, we define nine regions of each chemical space representation using the minimum and maximum values of the t-SNE coordinates that contain positive DILI compounds (this step is schematically explained in Figure S1 in the Supplementary material). The criteria to delimit each region are available in the Supplementary material (MetricOfDataFusion.xlsx). Finally, each weight peer quadrant was multiplied by the coordinate (t-SNE 1 or 2) of each compound contained in each chemical space representation.

Data fusion

Normalized value of weighted t-SNE coordinate (NWtSNE) was calculated to compare directly the representation of the chemical spaces i.e., based on *in vitro* data, chemical and physicochemical properties, and fingerprints. Each of the two dimensional coordinates t-SNE 1 and t-SNE 2 were calculated using Equation (**2**):

$$(\mathbf{2})\ NWtSNE = ((WtSNE) - (MIN (WtSNE))) / (MAX (WtSNE) - MIN(WtSNE))$$

where “WtSNE” is the weighted t-SNE coordinate, “MIN” and “MAX” are the minimum and maximum t-SNE value, respectively.

Finally, the consensus t-SNE coordinates were generated by summing the normalized coordinates of each chemical space representation of each compound. The automatic workflow of this method was implemented in KNIME and it is available in the Supplementary material (Multitiverse_DataFusion.knwf). The interactive visualizations of the chemical spaces were generated with DataWarrior software v.5.5.0., and are available in the Supplementary material (DB_ConsensusChemSpace_DILI.dwar) [21,24].

Distance calculations

A strategy to evaluate if the clustering of associated and non associated DILI compounds is efficient is calculating the distance between each compound in each chemical space representation. Namely, the shortest distances between DILI-associated compounds indicate that the clustering method is more efficient. The largest distance in the clustering between DILI-associated compounds indicates that the method is not capable of clustering them. To this end, the Manhattan distances were calculated by each pair of compounds on the dataset [25]. The distances were calculated using the "distance matrix calculate" node in KNIME. The protocol is available in the Supplementary material (Multitiverse_DataFusion.knwf). The mean distance between associated (or non-associated) DILI compounds and their standard deviation were calculated and plotted.

Results

In this section, we discuss the chemical multiverse of compounds associated with DILI reports, and a methodology to integrate chemical space data. Figure 1 shows chemical structures of representative compounds associated with DILI events. Interestingly, these compounds exhibit a notable structural diversity with different chemical scaffolds, and also present different types of atoms (e.g., O, N, S, Cl, F, P, etc.) that confer different kinds of properties.

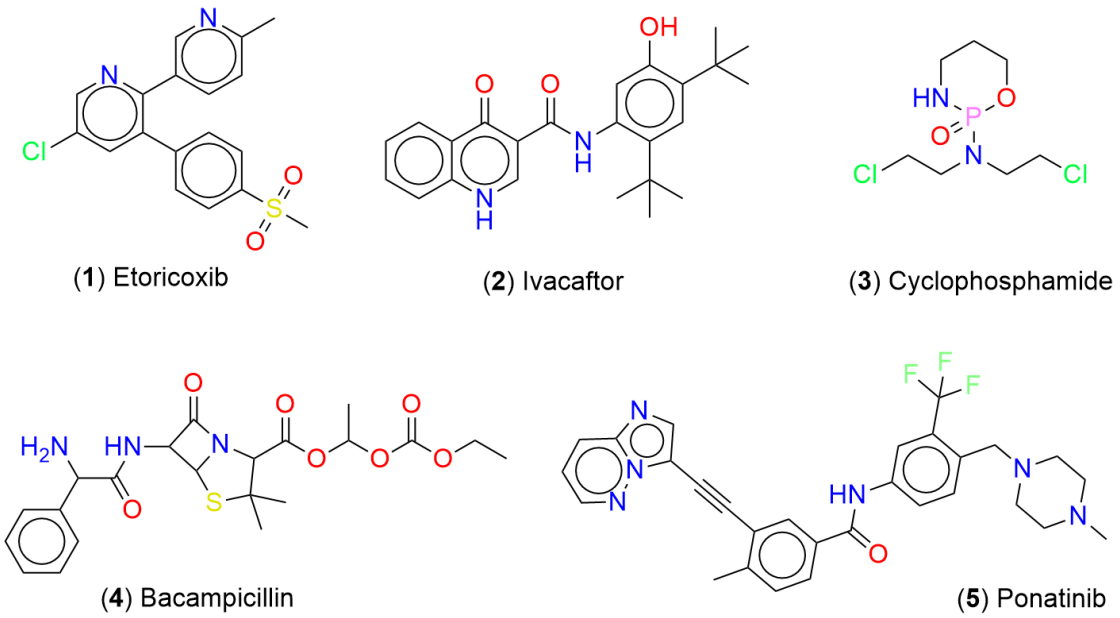

Figure 1. Chemical structures of representative compounds associated with DILI events.

Figure 2A-E shows the multiverse chemical space (i.e., different chemical space representations to the same data set) of 471 compounds associated with DILI reports. Each chemical space

representation illustrates structural (e.g., MACCS keys), topological (e.g., RDKIT, and ECFP4), chemical and physicochemical (e.g., drug-like properties), or *in vitro* data of this dataset. The data points colored in red represent compounds associated with DILI events (i.e., compounds associated with hepatotoxic signatures), in contrast with the compounds represented with data points in blue (that have not been related to DILI issues). Figure 2 illustrates an overview of the impact of each kind of descriptor on the clustering of compounds associated with DILI events. For example, the poor clustering generated by data from bidimensional structural descriptors (MACCS fingerprint - Figure 2-A) suggests that this information is not enough to cluster the compounds according to their DILI events. In contrast, topological (tridimensional) descriptors (like RDKIT) offer a better clustering of compounds associated with DILI events (red dots). Interestingly, the poor clustering based on drug-like properties (Figure 2-D) and *in vitro* data (Figure 2-E) suggests that these features (independently) do not guarantee the correct description of DILI events.

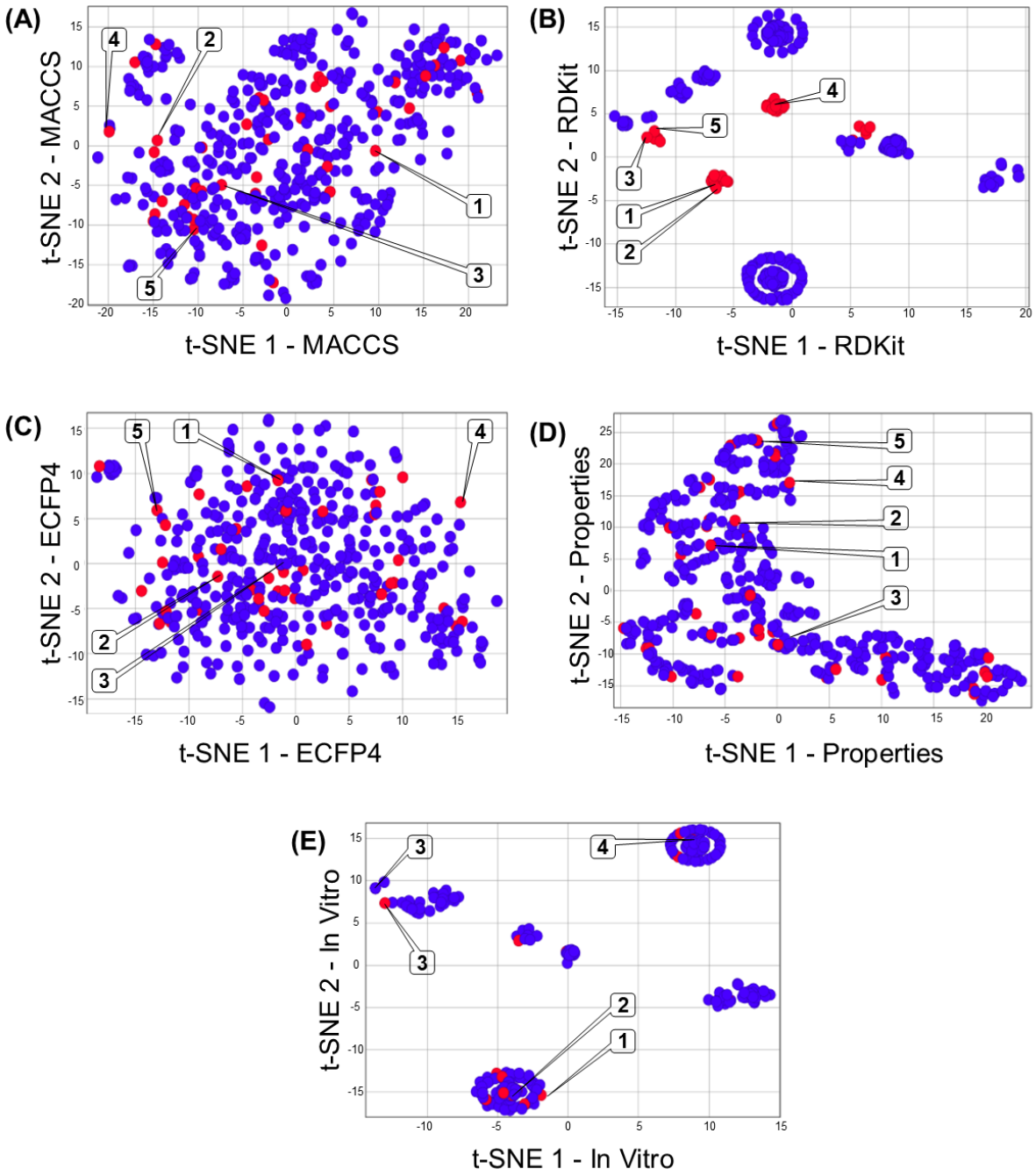

Figure 2. Representation of the multiverse chemical space of 471 compounds associated with DILI. Each chemical space visualization was constructed by dimensional reduction (t-SNE coordinates) of fingerprints (**A**) MACCS keys, (**B**) RDkit, (**C**) ECFP4, (**D**) chemical and physicochemical properties, and (**E**) *in vitro* data. Each data point in the graph represents a chemical structure, and the color of these points indicates if the chemical structure has been associated (red) or not (blue) with DILI events. Representative compounds are labeled with the compound numbers as in Figure 1.

Figure 3 shows the consensus chemical space representation. This new chemical space representation improves the visual identification of positive DILI compounds (red data points). Each region of each consensus chemical space representation is constructed, as per equations **1** and **2**, to improve the separation of the positive and negative DILI compound cases. Figure 3-A the new t-SNE coordinates generated from the fusion of multiverse chemical space data (e.g., structural, topological,

chemical, physicochemical, and *in vitro* data). Figure 3-B shows the new coordinates generated from the fusion of structural (RDKIT fingerprint) and *in vitro* data.

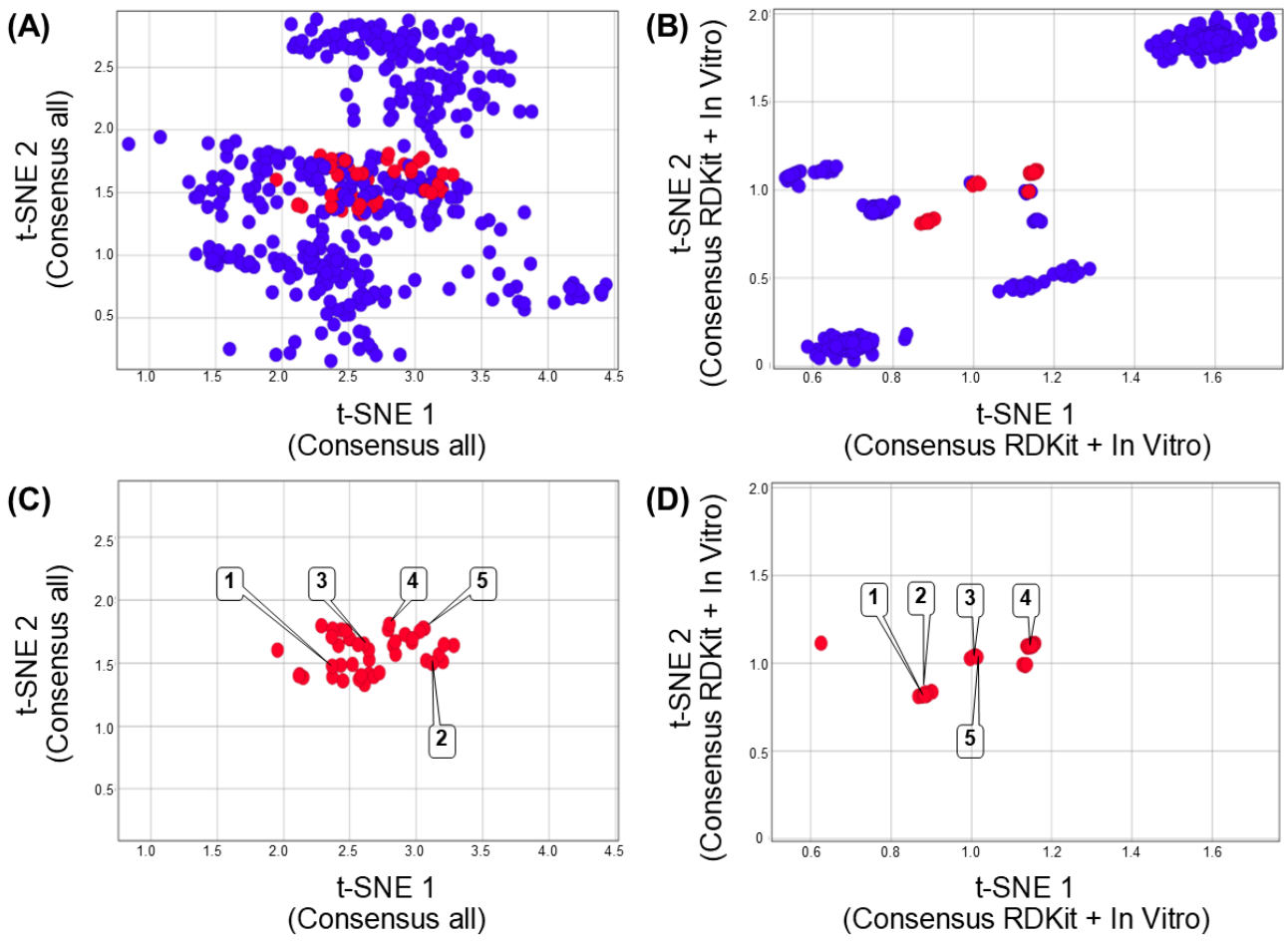

Figure 3. Consensus chemical space of 471 compounds associated with DILI reports. Each chemical space was constructed using the assignment and normalization of the weights by each region of the chemical space. (**A**) Consensus chemical space representation of reduced dimensions generated from fingerprints, chemical/physicochemical properties, and *in vitro* data related to each compound associated with DILI reports. (**B**) Consensus chemical space representation using the reduced dimensions generated from RDKit fingerprint and *in vitro* data. (**C** and **D**) Consensus chemical space representations showing only compounds associated with DILI events. Each point in the chemical spaces represents a chemical structure. Data points are colored by if the chemical structure has been associated with DILI events (red) or not (blue). Representative compounds are labeled with the compound numbers as in Figure 1.

It is remarkable the clustering difference observed in the visualization of the chemical spaces generated by only a type of data (Figure 2) as compared to the combined data (Figure 3).

To remark on the improved clustering of the combined descriptors, Figure 4 shows the mean pairwise distance of associated (red) and non-associated (blue) compounds with DILI events generated by each chemical space representation. Figure 4 indicates that the use of a single data type generates a higher average pairwise distance (low clustering efficiency) of positive DILI compounds (from 12.7 to

25.7), and paired negative DILI compounds (from 17.9 to 25.1). This is in contrast with the consensus chemical space representation (fused data) that exhibits lower mean pairwise distance (high clustering efficiency) between positive DILI compounds (from 1.0 to 1.1) and negative DILI compounds (from 1.7 to 3.2).

Interestingly, using single or fused data, the distance between the non-associated DILI compounds continues to be higher than the distance between associated-DILI compounds. This fact suggests that the non-associated DILI compounds exhibit a higher intrinsically data diversity.

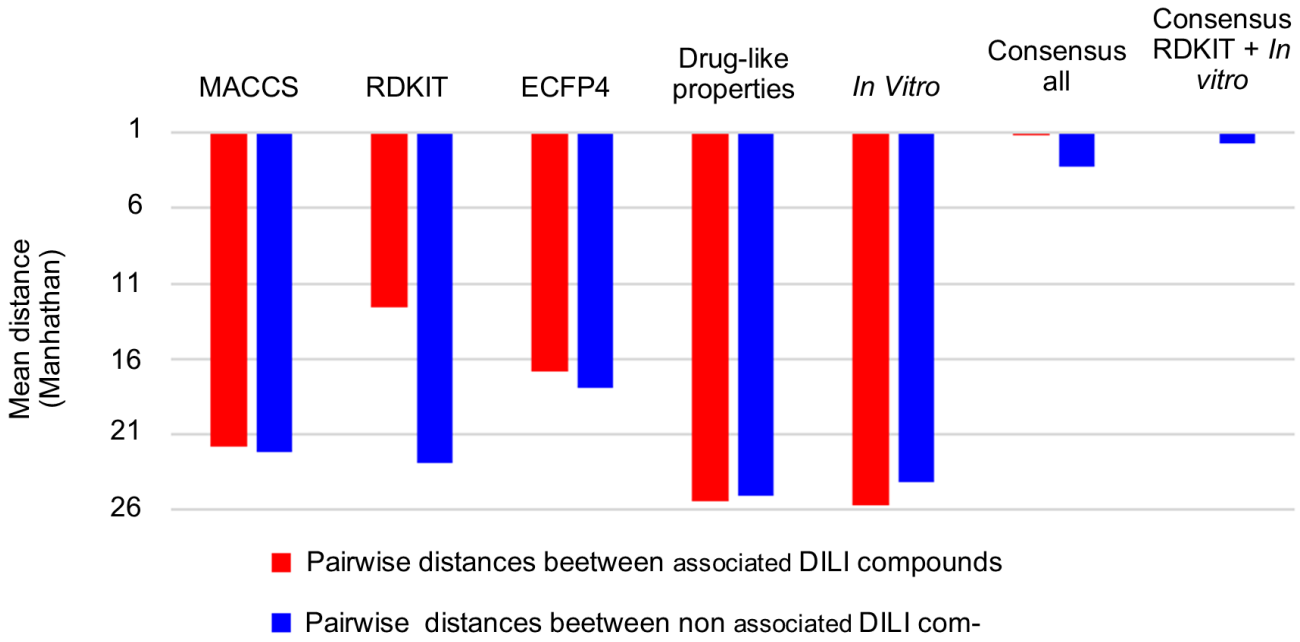

Figure 4. Pairwise Manhattan distances of positive and negative DILI compounds in chemical space representations obtained with different descriptors. The 2D plots of the chemical space visualizations are in Figures 2 and 3.

Each representation offers a unique form to cluster each chemical structure (Figures 2 and 3). However, consensus methods provide a mathematical framework to establish a weight to each region on the different chemical space representations (generating a semi-unsupervised approach to construct enriched chemical space representations, Figure 3). From a pharmacological view, these results remark on the importance of multidisciplinary approaches, using chemical and biological data, to develop methodologies capable of efficiently DILI events.

Discussion

There are multiple representations available to describe compounds and study the SPR of a data set. The large variety of molecular descriptors is linked to the subjectivity of the “molecular similarity” that

is dependent on the molecular representation [26]. Namely, the similarity of a pair of compounds depends on the features used to compare them. In fact, a pair of compounds could be considered similar if we use structural descriptors, but this does not guarantee that both compounds have similar *in vitro* activity [27]. For this reason, it is crucial to evaluate the similarity of the compounds and, in general, the SPR of data sets using different descriptors and similarity metrics. The combined analysis of alternative representations (aka data fusion) could reduce the information gap between the chemical structures vital in drug development and biological knowledge. However, one of the most important issues in data fusion is assigning adequate weights to each variable that is being combined (e.g., dimensions that define the compound's chemical space) because different mathematical approximations could be used to generate them [28]. In fact, there is no unique and “best” manner to generate consensus chemical spaces. Namely, it is necessary to adapt the data fusion approach to consider each data set. This important point could lead to the feature selection for prospective studies, generating a good starting point for exploring large datasets.

There is a crescent interest in developing protocols capable of predicting DILI events. However, these side effects are complicated to predict because they are associated with (parallelly) multiple pharmacological events and become a typical problem to address with multiple-parameter optimization. For example, existing reports demonstrate the relationships between the chemical structures and physicochemical properties of the DILI events, but at the same time, other authors show that ADME properties, cell-based data, and other *in vitro* assays lead to the identification/prediction of DILI events. Namely, the DILI events are a complex case study that requires using all available data to rationalize (almost in part) and predict their occurrence during the pre-clinical and clinical interventions. Fortunately, the current multi-objective optimization methods could help address this issue briefly [29].

Consensus chemical spaces are an approach to fuse and use different kinds of data (e.g., descriptors that define the multidimensional vector space) to improve predicting a specific, desired property. To this end, the main challenge is to choose from the several methods available to combine high dimensionality of data using a robust mathematical scheme.

Additionally, and as happens in any other predictive methodology, another major issue to address is the limited access to data [30], considering that several results that are regarded as of “no interest” for a particular study (at some point in time) are rarely published. This fact creates a crescent gap in the available information related to compounds associated with poor activity or side effects like DILI events.

For prospective studies, it will be necessary to assess multiple methods to fuse data [31] and use other high-dimensional reduction methods such as principal component analysis and support vector machine, etc. [32].

The DILI understanding is relevant to elucidating molecular mechanisms, identifying novel biomarkers, and preventing drug side effects previously to pre-clinical and clinical interventions. The multiverse chemical space and the consensus chemical space representations (using fused data) enrich the information that could generate useful knowledge. For example, the drug design methods based on fused data could improve the next generation of toxicological and post-marketing decision-making approaches.

The results illustrated in Figure 4 show that the RDKIT fingerprint allows more efficient clustering in contrast with other kinds of fingerprints and descriptors explored in this work. This observation highlights the importance of the intrinsic descriptor encoded by the RDKIT fingerprint (e.g., topology, bond order, atom types, presence of rings, ring sizes, and aromaticity of each compound) that could be used to improve the understanding of DILI events.

Figure 5 shows a classification of the 471 compounds associated with DILI according to the type of chemical taxonomy. The analysis shows that major types of compounds exhibit around 10% of chemical structures associated with DILI events. However, organohalogens, phenylpropanoids, polyketides, organic acids, organosulfur, alkaloids, and organophosphorus compounds exhibit a rate higher than 10% of associated DILI compounds.

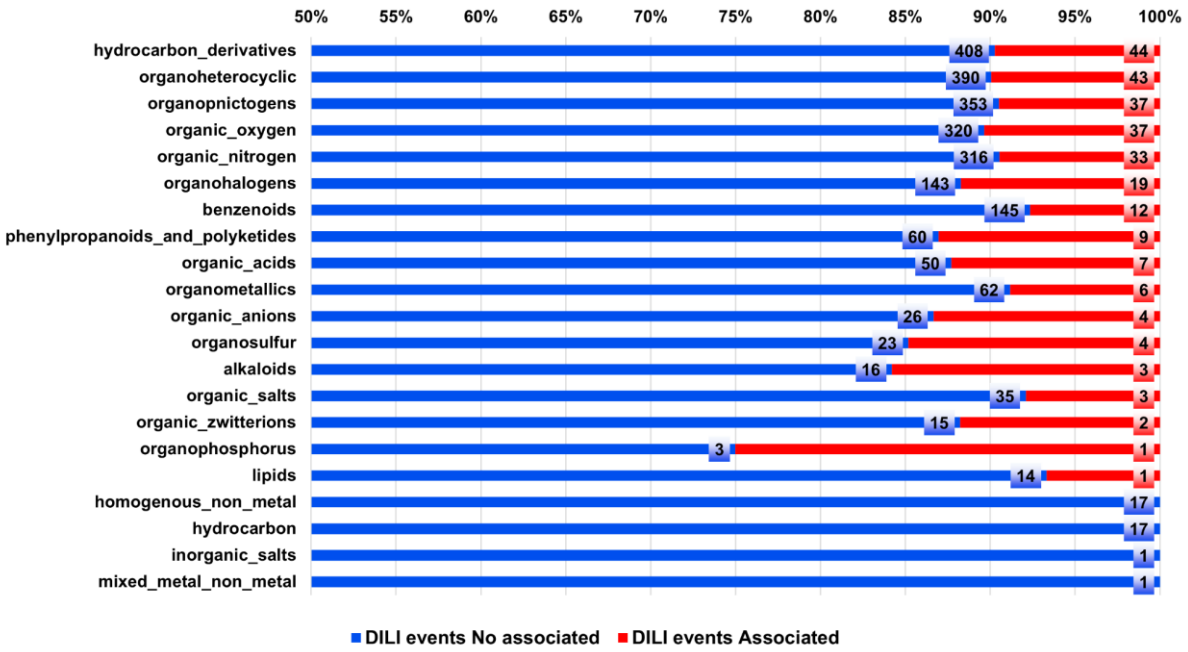

Figure 5. Types of compounds and their association with DILI events. 471 compounds associated with DILI reports were classified [33] according to their chemical taxonomy, and each chemical taxonomy was associated with the number of cases associated (red) and no associated (blue) with DILI events.

Additionally, the most frequent compounds associated with DILI events contain complex ring systems, specific functional groups, and atoms (e.g., double carbon bonds, carboxylic acids, ketones, halogens, sulfur, phosphorus) that *per se* have been associated with hepatic injuries [34-38]. From a chemical perspective, these observations could lead to the early identification of compounds potentially associated with DILI events.

Finally, and from a pharmacological perspective, we remark on the importance of incorporating data that predict the hepatic and microbiota biotransformation [39,40] of xenobiotics to increase the early identification of potential associated DILI compounds. Acetaminophen provides a typical example of the importance of studying biotransformation. This drug is not hepatotoxic but its metabolites generate fulminant liver injuries [41,42].

Conclusions

DILI is the principal reason for failure in developing drug candidates. It is the most common reason to withdraw from the market after a drug has been approved for clinical use. However, the current approaches to predicting DILI have not allowed a complete understanding of chemical and biological alerts to identify early compounds associated with DILI events.

Drug design methodologies based on fused data could be the next generation of tools used in rational design, especially to decode complex pharmacological issues like DILI events. Here, we introduce a combined analysis of DILI-related events using the concept of consensus chemical space and the chemical multiverse, using chemical, physicochemical, structural, biochemical, and biological data to improve the understanding of DILI events. Our results, which suggest that the combination of chemical structural and biological data improves the clustering of associated DILI compounds, pave the way to new opportunities to develop predictive models (like machine and deep learning models) capable of predicting DILI events in an early stage of the drug development process. It was also concluded that organohalogens, phenylpropanoids, polyketides, organic acids, organosulfur, alkaloids, and organophosphorus compounds are associated with a higher rate of DILI events. For this reason, we suggest more exhaustive preliminary studies for these types of compounds with the aim of reducing the cases associated with DILI events.

Data availability statement

The original contributions presented in the study are included in the article and Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

EL-L: Methodology, Investigation, Formal analysis, Writing–original draft, and Writing-review. JLM-F: Conceptualization, Writing-review and editing, Supervision, and Funding acquisition.

Funding

No external funding was received to perform this research. The article processing charges were covered by the authors´s personal resources.

Acknowledgments

E. L.-L. is grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the Ph. D. scholarship number, 762342 (No. CVU: 894234).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary material

The Supplementary Material for this article can be found online at: https://doi.org/10.6084/m9.figshare.21312108

References

1. Babai, S., Auclert, L., and Le-Louët, H. (2021). Safety data and withdrawal of hepatotoxic drugs. *Therapies*, *76*. doi: 10.1016/j.therap.2018.02.004

2. Leeson, P. D. (2018). Impact of Physicochemical Properties on dose and hepatotoxicity of oral drugs. *Chem. Res. Toxicol.*, *31*. doi: 10.1021/acs.chemrestox.8b00044

3. Liu, L., Fu, et al. (2019). Three-Level Hepatotoxicity prediction system based on adverse hepatic effects. *Mol. Pharm.*, *16*. doi: 10.1021/acs.molpharmaceut.8b01048

4. He, S., et al. (2019). An in silico model for predicting drug-Induced hepatotoxicity. *Int. J. Mol. Sci.*, *20*. doi: 10.3390/ijms20081897

5. Thakkar, S., et al. (2020). Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. *Drug Discov. Today*, *25*. doi: 10.1016/j.drudis.2019.09.022

6. Vall, A., et al. (2021). The promise of AI for DILI prediction. *Front. Artif. Intell.*, *4*. doi: 10.3389/frai.2021.638410

7. Medina-Franco, J. L., Naveja, J. J., and López-López, E. (2019). Reaching for the bright StARs in chemical space. *Drug Discov. Today*, *24*. doi: 10.1016/j.drudis.2019.09.013

8. Virshup, A. M., et al. (2013). Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.*, *135*. doi: 10.1021/ja401184g

9. Medina-Franco, J. L., et al. (2022). Chemical multiverse: an expanded view of chemical space. *Mol. Inf. doi:* 10.1002/minf.202200116

10. Medina-Franco, J. L., et al. (2022). Progress on open chemoinformatic tools for expanding and exploring the chemical space. *J. Comput.-Aided Mol. Des.*, *36*. doi:10.1007/s10822-021-00399-1

11. Wang, S., et al. (2021). Advances in data preprocessing for biomedical data fusion: an overview of the methods, challenges, and prospects. *Inf. Fusion*, *76*. doi: 10.1016/j.inffus.2021.07.001

12. Kalliokoski, T., and Sinervo, K. (2019). Predicting pK_a for small molecules on public and in-house datasets using fast prediction methods combined with data fusion. *Mol. Inf.*, *38*. doi: 10.1002/minf.201800163

13. Moreira, R. et al. (2022). Data fusion discovery (DAFdiscovery) pipeline to aid compound annotation and bioactive compound discovery across diverse spectral data. ChemRxiv [Preprint]. Available at: https://doi.org/10.26434/chemrxiv-2022-5fj6v-v2 (Accessed October 8, 2022).

14. Bergenstråhle, L., et al. (2022). Super-resolved spatial transcriptomics by deep data fusion. *Nature Biotech.*, *40*. doi: 10.1038/s41587-021-01075-3

15. Simm, J., et al. (2018). Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chem. Biol.*, *25*. doi: 10.1016/j.chembiol.2018.01.015

16. Bisht, V., Acharjee, A., and Gkoutos, G. V. (2021). NFnetFu: a novel workflow for microbiome data fusion. *Comput. Biol. Med.*, *135*. doi: 10.1016/j.compbiomed.2021.104556

17. Wishart, D. S., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, *46*. doi: 10.1093/nar/gkx1037

18. ChEMBL (2022). CHEMBL v. 30. http://doi.org/10.6019/CHEMBL.database.30 [Accessed October 08, 2022].

19. Liu, X., et al. (2020). Machine-learning prediction of oral drug-induced liver injury (DILI) via multiple features and endpoints. *BioMed Res. Int.*, *2020*. doi: 10.1155/2020/4795140

20. Berthold, M. R., et al. (2008). "*KNIME: the konstanz information miner*", in studies in classification, data analysis, and knowledge organization, ed. V. Rosteck (Springer). doi: 10.1007/978-3-540-78246-9_38

21. Sander, T., Freyss, J., von Korff, M., and Rufener, C. (2015). DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model. 55*. doi: 10.1021/ci500588j

22. Landrum, G. et al. (2022). rdkit/rdkit: 2022_03_5. doi:10.5281/zenodo.6961488.

23. Laurens van der Maaten. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, *9*, 2579–2605.

24. López-López, E., Naveja, J. J., and Medina-Franco, J. L. (2019). DataWarrior: an evaluation of the open-source drug discovery tool. *Exp. Op. Drug Discov.*, *14*. doi: 10.1080/17460441.2019.1581170

25. Kyosev, I., Paun, I., Moshfeghi, Y., and Ntarmos, N. (2020). Measuring distances among graphs en route to graph clustering. *2020 IEEE International Conference on Big Data (Big Data)*. doi: 10.1109/BigData50022.2020.9378333

26. Medina-Franco, J. L., and Maggiora, G. M. (2013). Molecular similarity analysis. In *Chemoinformatics for Drug Discovery*. John Wiley & Sons, Inc. doi: 10.1002/9781118742785.ch15

27. López-López, E., Cerda-García-Rojas, C. M., and Medina-Franco, J. L. (2021). Tubulin inhibitors: a chemoinformatic analysis using cell-Based data. *Molecules*, *26*. doi:10.3390/molecules26092483

28. Azcarate, S. M., Ríos-Reina, R., Amigo, J. M., and Goicoechea, H. C. (2021). Data handling in data fusion: methodologies and applications. *Trends Analyt, Chem,*, *143*. doi: 10.1016/j.trac.2021.116355

29. Nicolaou, C. A., and Brown, N. (2013). Multi-objective optimization methods in drug design. *Drug Discov. Today*, *10*. doi: 10.1016/j.ddtec.2013.02.001

30. López-López, E., Fernández-de Gortari, E., and Medina-Franco, J. L. (2022). Yes SIR! on the structure–inactivity relationships in drug discovery. *Drug Discov. Today*, *27*. doi: 10.1016/j.drudis.2022.05.005

31. Baldwin, E., et al. (2020). On fusion methods for knowledge discovery from multi-omics datasets. *Comput. Struct .Biotech. J.*, *18*. doi: 10.1016/j.csbj.2020.02.011

32. Saldívar-González, F. I., and Medina-Franco, J. L. (2022). Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opin. Drug Discov.*, *17*. doi: 10.1080/17460441.2022.2084608

33. DrugTax (2022). drugtax 1.0.11. https://pypi.org/project/drugtax/ [Accessed October 8, 2022].

34. Wu, J.-P., et al. (2019). Contamination of organohalogen chemicals and hepatic steatosis in common kingfisher (Alcedo atthis) breeding at a nature reserve near e-waste recycling sites in South China. *Sci. Total Environ.*, *659*. doi: 10.1016/j.scitotenv.2018.12.395

35. Oh, H.-A., et al. (2022). Identification of integrative hepatotoxicity induced by lysosomal phospholipase A2 inhibition of cationic amphiphilic drugs via metabolomics. *Biochem. Biophysic. Res. Comm.*, *607*. doi: 10.1016/j.bbrc.2022.03.038

36. Mahomoodally, M. F., Nabee, N., and Baureek, N. (2022). Organosulfur compounds (allyl sulfide, indoles). In *Antioxidants Effects in Health*. Elsevier. doi: 10.1016/B978-0-12-819096-8.00070-7

37. Wang, Z., et al. (2021). Hepatotoxicity of pyrrolizidine alkaloid compound intermedine: comparison with other pyrrolizidine alkaloids and its toxicological mechanism. *Toxins*, *13*. doi: 10.3390/toxins13120849

38. Ramesh, M., et al. (2020). Organophosphorus flame retardant induced hepatotoxicity and brain AChE inhibition on zebrafish (Danio rerio). *Neurotoxicol. Teratol.*, *82*. doi: 10.1016/j.ntt.2020.106919

39. Djoumbou-Feunang, et al. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics*, *11*. doi: 10.1186/s13321-018-0324-5

40. Yang, M., et al. (2019). Intestinal and hepatic biotransformation of pyrrolizidine alkaloid N-oxides to toxic pyrrolizidine alkaloids. Arch. Toxicol., *93*(8). https://doi.org/10.1007/s00204-019-02499-2

41. McClain, et al. (1999). Acetaminophen hepatotoxicity: An update. *Curr. Gastroenterology Rep.*, *1*. doi: 10.1007/s11894-999-0086-3

42. David, A., et al. (2021). Acetaminophen metabolism revisited using non-targeted analyses: Implications for human biomonitoring. *Environ Int.*, *149*. doi: 10.1016/j.envint.2021.106388