Cell Painting morphological profiles can complement QSAR models for rat acute oral toxicity prediction

Fabrice Camilleri[1,2], Joanna M. Wenda[3], Claire Pecoraro-Mercier[3], Jean-Paul Comet[2], David Rouquié[1]

[1] Toxicology Data Science, Bayer SAS Crop Science Division, 355 rue Dostoïevski, CS 90153 Valbonne, 06906 Sophia Antipolis Cedex, France

[2] I3S UMR 7271 du CNRS, Université Côte d'Azur, Bâtiment Algorithme-Euclide-B, 2000 Route des Lucioles, B.P. 121, 06903 Sophia Antipolis, France

[3] Early Toxicology, Bayer SAS Crop Science Division, 355 rue Dostoïevski, CS 90153 Valbonne, 06906 Sophia Antipolis Cedex, France

Keywords: Rat acute oral toxicity, Cell Painting, QSAR, NAMs, Morphological profiles, LD50, NGRA, KNN, Applicability Domain, compound derisking.

ABSTRACT

Small molecule discovery entails the identifying and developing of chemical compounds with optimized safety properties in different target species including laboratory animals, human and environmental species. Early de-risking supports the identification of candidates with improved safety profiles in the development of new chemical compounds. In vivo studies are performed in the early assessment of acute oral toxicity, an important endpoint in the development of crop protection products. Because in vivo studies are long, costly, and raise ethical concerns, non-animal alternatives are needed. Several models were analyzed to classify compounds as highly acutely toxic (LD50 ≤ 60 mg/kg) or not. First, the publicly available QSAR model, CATMoS, was used to classify 630 Bayer Crop Science compounds. This model did not show good results (balanced accuracy of 0.52) as the compounds were in gaps of the model applicability domain. Interestingly, training a K nearest neighbor's model (equivalent to read-across) using specifically the compound structure information, good classifications were obtained (balanced accuracy of 0.81). Then, this study explored whether morphological profiles obtained using Cell Painting *in vitro* assay on U2OS cells could assist in the prediction of rat acute oral toxicity and how those predictions could complement those made by QSAR models using chemical structures. Compounds with known acute oral toxicity were selected and a Cell Painting campaign were conducted on 226 compounds at 10 µM, 31.6 µM, and 100 µM. Binary classifiers based on K nearest neighbors, were developed to categorize compounds as highly acutely toxic (LD50 ≤ 60 mg/kg) or not. These classifiers were built, either using the compound chemical structure information (Morgan Fingerprint) or morphological profiles obtained from Cell Painting. Our results showed that the classification of compounds, using a read-across approach, as very acutely oral toxic or not, was possible using chemical structure information, U2OS cell morphological profiles or the combination of both. When classifying compounds structurally similar to those used to train the classifier, the chemical structure information was more predictive (mean balanced accuracy of 0.82). Conversely, when compounds to classify were structurally different from compounds used to train the classifier, the U2OS cell

morphological profiles were more predictive (mean balanced accuracy of 0.72). The combination of both models allowed, when classifying compounds structurally similar to those used to train the classifiers, to slightly enhance the predictions (mean balanced accuracy of 0.85).

Introduction

Small molecule discovery entails the identifying and developing of chemical compounds with optimized safety properties in different target species including laboratory animals, human and environmental species. This complex process requires the integration of chemical and biological exploration. Chemists design diverse compounds to modulate specific biological targets or pathways while biological assays assess efficacy and safety, guiding further optimization. Balancing potency with selectivity and minimizing off-target effects presents a challenge. Additionally, the complexity of biological systems poses hurdles, as the interplay of various factors influences a molecule's properties. Success in small molecule discovery hinges on a multidisciplinary approach, where chemists, biologists and data scientists collaborate to navigate the intricate landscape of chemical and biological interactions, ultimately advancing the development of new active substances. In agrochemical discovery, early de-risking is crucial involving systematic assessment of compounds for potential safety issues to be addressed as early as possible. Addressing genotoxicity and acute oral toxicity is essential due to established cut-off criteria.

Traditionally, early genotoxicity evaluations are performed with in vitro test methods (Ames and Micronucleus assays) whereas acute oral toxicity profile is assessed using the LD50 which represent the single dose at which mortality is induced in at least in 50 % of the tested animals. Acute oral toxicity assessment requires in vivo studies to have an early estimation of the LD50 range. Conducting in vivo studies for acute oral toxicity assessment is time-consuming, expensive and does not allow testing large number of chemicals. Animal studies also raise ethical concerns and must be reduced or if possible, eliminated. Faster and cheaper alternatives are necessary in a higher throughput manner for the de-risking of more chemical compounds enabling prioritization of small molecules candidates with acceptable toxicological profiles. Non-animal alternatives are available to replace in vivo studies including in vitro approaches and in silico models for early estimation of the LD50. For instance, the in vitro 3T3 neutral red uptake assay (NRU) (Erhirhie, Ihekwereme and Ilodigwe, 2018) can serve to categorize

compounds with LD50 greater or lower than 2000 mg/kg. However, this LD50 threshold is high and may not be fully informative for cases where the LD50 can be smaller.

On the other hand, in silico models, such as quantitative structure activity relationship (QSAR) models, rely on machine learning models based on structural information of chemical compounds. One notable example is the Collaborative Acute Toxicity Modeling Suite (CATMoS), a public QSAR model, built in a collaboration of several research groups. CATMos model was trained on more than 10k compounds and demonstrated high performances in the prediction of acute oral toxicity in the rat (Mansouri *et al.*, 2021). CATMoS can indeed, as a regression model, predict the LD50, along with the classification of chemicals into the five GHS (Global Harmonized System) categories.

Combining structural and in vitro high biologically dense information could help enhancing the prediction of acutely toxic compounds, while providing potentially mechanistic insights related to adverse outcome pathways (AOP) leading to acute toxicity (Becker *et al.*, 2020; Edwards *et al.*, 2022).

The Carpenter–Singh Lab at the Broad institute has developed an in vitro assay, Cell Painting, capturing the morphological information of cells perturbed by chemicals (Bray *et al.*, 2016). The main advantage of Cell Painting lies in its untargeted nature, allowing it to capture in theory any bioactivity inducing a cell morphological change. This assay has already been used successfully, in 'hit' discovery and in MoA (Mode of Action) prediction. In toxicology, the US Environmental Protection Agency (EPA) explored it to screen bioactive compounds in the context of risk assessment (Nyffeler *et al.*, 2020). Moreover, Cell Painting has also been utilized for the prediction of mitochondrial toxicity (Seal et al., 2022), and liver toxicity (Lejal et al., 2023).

Our work explores the use of chemical structure information and morphological profiles obtained from Cell Painting to predict rat acute oral toxicity via a simple read-across approach. We present the performances of two k nearest neighbor's models trained to classify chemicals as very acute oral toxic (LD50 ≤ 60 mg/kg) or not (LD 50 > 60 mg/kg). Two types of classifiers were built utilizing either the compound structure information, or the compound morphological

effects on U2OS cells obtained through the Cell Painting assay. Results indicated that classifying compounds similar to those in the training set, QSAR models (mean balanced accuracy of 0.82) outperformed morphological profile-based models (mean balanced accuracy of 0.75). However, for the classifying compounds structurally different from those in the training set, morphological profile-based models (mean balanced accuracy of 0.72) outperformed QSAR models (mean balanced accuracy of 0.60). Additionally, we propose a combined approach to support decision-making when the two types of classifiers yield contradictory predictions. Overall, our results showed that using a read-across approach, the classification of compounds, as very acutely oral toxic or not, was possible using chemical structure information, U2OS cell morphological profiles or the combination of both. When classifying compounds structurally similar to those used to train the classifier, the chemical structure information was more predictive. Conversely, when compounds to classify were structurally different from compounds used to train the classifier, the U2OS cell morphological profiles were more predictive.

Materials and methods

Acute oral toxicity compound classes

Compounds were categorized into two classes. The class 'Very acutely oral toxic', abbreviated VAOT referred to compounds having a LD50 less than or equal to 60 mg/kg. The class 'Non very acutely oral toxic', abbreviated NVAOT, referred to compounds having a LD50 greater than 60 mg/kg (Table 1).

Acute oral toxicity classes	
Very acutely oral toxic – VAOT	Rat oral LD50 ≤ 60 mg/kg
Non very acutely oral toxic - NVAOT	Rat oral LD50 > 60 mg/kg

Table 1. Definition of the two oral acute toxicity classes VAOT and NVAOT

Compound selection

Compounds with acute oral toxicity results in rats

To select compounds, Bayer internal databases were queried. A total of 765 compounds with rat acute oral toxicity results were found for two doses: 60 mg/kg and 300 mg/kg. Out of them, 109 compounds were acute toxic at the dose of 60 mg/kg, meaning compounds belonging to the VAOT class, and 521 compounds were not toxic at 300 mg/kg, meaning compounds belonging to the NVAOT class. Compounds being acute toxic at 300 mg/kg and not acute toxic at 60 mg/kg were voluntary excluded, hoping to have a good contrast of morphological profiles between the VAOT and NVAOT classes.

This made the first dataset of 630 compounds that were used to test the acute toxicity prediction of the collaborative Acute Toxicity Modeling Suite (CATMoS) (Mansouri *et al.*, 2021) and to train a bespoke chemical structure based classifier.

For the Cell Painting campaign, we checked which of the previous dataset compounds were available in Bayer compound logistics. 81 VAOT compounds were found. To complete the list of VAOT compounds, we queried the chemIDplus public database ('ChemIDplus', 2023), and selected 29 compounds that were available in Bayer compound logistics, making a total of 110 VAOT compounds. To have a balanced dataset, 116 NVAOT compounds were selected. To have a good chemical structure diversity among them, the Butina algorithm (Butina, 1999) was used to cluster the 521 compounds of the previous dataset, based on the Tanimoto similarity of their Morgan fingerprint: a maximum of clusters was selected and the number of compounds coming from the same cluster were minimized. This selection resulted in having a total of 116 NVAOT compounds and 110 VAOT.

To summarize, two sets of compounds were defined. The first one, called 'QSAR only compound set', was a set of 630 compounds; 109 were VAOT and 521 were NVAOT. The second one, called 'Cell Painting compound set', was a set of 226 compounds; 110 were VAOT, and 116 were NVAOT (Figure 1).

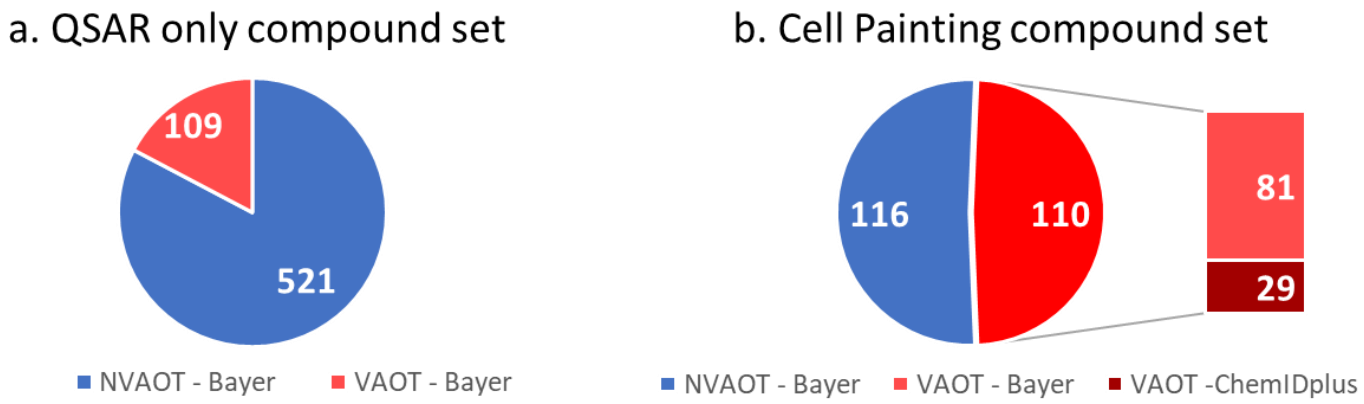

Figure 1.

Composition of the two datasets by compound class: VAOT (Very Acutely Oral Toxic) and NVAOT (Non Very Acutely Oral Toxic); and by source: Bayer and ChemIDplus.
a: QSAR only compound set. Set used for QSAR models only.
b: Cell Painting compound set, Set used for QSAR and morphological profile models.

Negative and positive controls

For the Cell Painting assay, DMSO (dimethyl sulfoxide) only (0.1%) were used as the negative control.

To monitor the Cell Painting assay's performance and assess the quality of experiment's replicates, a set of positive controls was used – compounds inducing reproducible and distinct morphological profiles in U2OS cells. The selection was based on published literature and pilot tests in our lab (Supplementary information).

CATMoS QSAR model

As a first attempt, the acute oral toxicity class of the compounds were predicted with the collaborative Acute Toxicity Modeling Suite (CATMoS) QSAR model, implemented in the OPERA (version 2.9) (Mansouri *et al.*, 2018, 2021) QSAR suite. To match the two classes of this paper (VAOT and NVAOT), 3 CATMoS predictions were needed: the EPA classification, the GHS classification and the LD50 range prediction. All compounds being classified by CATMoS as EPA category 1 (LD50 ≤50 mg/kg), or GHS category 1 (LD50 ≤ 5 mg/kg), or GHS category 2 (5 mg/kg < LD50 ≤ 50 mg/kg) were assigned as VAOT. For compounds being classified as EPA category 2 (50 mg/kg < LD50 ≤ 500 mg/kg), or GHS category 3 (50 mg/kg < LD50 ≤ 300 mg/kg), the inferior limit of the LD50 range predictions were considered: if the inferior limits were smaller than 60, the compounds were classified as VAOT. For all other predictions, we classified the compounds as NVOAT.

The Opera implementation of CATMoS provides three prediction reliability metrics (Mansouri *et al.*, 2018) that were used to understand the predictions made by the model. A global Applicability Domain Boolean value is calculated, indicating if a compound falls within the

training set chemical space. Additionally, an Applicability Domain index is calculated, ranging from zero to one, revealing the close (high value) or distant (low value) vicinity of the queried compound. Finally, a confidence index is computed, informing on the accuracy of the prediction of the neighbors of the queried compound.

Cell painting campaign

A Cell Painting campaign was performed in our laboratory to obtain the morphological profiles of our 'Cell Painting compound set' (set of 226 compounds). We used the Cell Painting Protocol v3 of the Broad institute on U2OS human osteosarcoma cells with 4 biological replicates (Cimini *et al.*, 2023).

A previous Cell Painting pilot done at the unique dose of 10 µM showed that few agrochemical compounds had morphological changes compared to the negative control, for this reason, in this campaign, to increase the chance of capturing a morphological response, the compounds were screened at 3 concentrations: 10µM, 31.6µM and 100µM.

Cell Culture and Seeding

Human osteosarcoma cells U2OS have been purchased from ATCC (ref.: HTB-96, lot: 70025046). The McCoy's 5A Modified Medium with GlutaMAX™ Supplement (Thermo Fisher, ref: 36600021) supplemented with 10% Fetal Bovine Serum (Gibco, ref.: 16000044) and penicillin/streptomycin mix (Sigma Aldrich, ref: P4458) was used for culturing cells in T75 or T175 flasks in a standard humidified incubator (37°C, 5% CO2). The passages were performed when the culture achieved about 80% confluency. Trypsin (Thermo Fisher, ref. 25200056) was used to detach the cells during passage and the number of live cells was calculated with an automatic cell counter (Countess II, Thermo Fisher) after staining the cells with trypan blue (Sigma, ref.: T8154). For the creation of a cell bank, the vial with frozen cells received from the supplier was thawed and expanded until internal passage no. 3 (P3). At this stage the cells were cryopreserved in complete media supplemented with 10% DMSO in an ultra-low temperature freezer (-150°C) creating a master bank. One vial of master bank was then thawed, expanded until internal passage no. 6 (P6), and cryopreserved as before to create a working bank. Vials of

the working bank were then directly used for seeding the microplates. One vial of cells (containing 4 million cells) was removed from -150°C freezer and thawed in the water bath. The contents of the vial were immediately added to 10 ml of pre-heated complete media and centrifuged (5 min, 120xg). After removing the supernatant, the cell pellet was resuspended in 10 ml of complete medium through thorough pipetting. The cell suspension was then added to 150 ml of medium in a round bottle with a magnetic stirrer and immediately used for seeding the 384-well microplates (Greiner BioONE CELLSTAR µCLEAR®; ref: 781091). Multidrop (Thermo Fisher) was used to automatically distribute 36 µl of cell suspension per well, resulting in a seeding density of around 900 cells/well. The cells were then incubated at 37°C, in an atmosphere of 5% CO2 in an automatised incubator (Cytomat 2, Thermo Fisher). All experimental replicates were performed on a different day, using a separate cell vial originating from the same working bank (P6).

Chemical treatment

The test compounds were received in powder form in 96-well deep well plates. They were then dissolved in DMSO (dimethyl sulfoxide) to create 100 mM stock solutions, aliquoted in 96-well V-bottom plates (V96 PP Plate, Thermo Fisher) and frozen at -20°C until the day of the treatment. Every biological replicate of the experiment originates from a separate aliquot of the stock solution plate, so that the compounds undergo only 1 freeze-thaw cycle. On the day of the treatment (24 h post-seeding), the plates containing stock solutions were thawed and the compounds were diluted in DMSO to create dose plates containing 3 concentrations per compound: 100 mM, 31.6 mM, and 10 mM. The dilutions were performed with the use of Viper liquid handler (Synchron). The compound solutions form the dose plates were then administered to the cell plates in a two-step process. Firstly, an intermediate dilution was prepared: 1 µl of the compound solution was diluted in 100 µl of complete cell medium (1:100 dilution), next 4 µl of the resulting intermediate solution was administered to the cell plate (4 µl of the diluted compound into 36 µl of cell media, 1:10 dilution). The final concentrations of compounds that the cells were exposed to were therefore: 10 µM, 31.6 µM and 100 µM, the

final vehicle (DMSO) concentration was 0.1%. The treated cell plates were subsequently incubated with the compounds for 48 h.

Staining

The staining and fixation were performed following the published protocol (Cimini *et al.*, 2023) with the use of PhenoVue JUMP kit (Perkin Elmer, ref.: PING23). Briefly, 20 µl/well of the Mitotracker solution were distributed to the cell plates with Multidrop (final concentration: 500 nM). After 30 minutes of incubation at 37°C, 20 µl/well of 16% PFA solution (Thermo Fisher, ref.: 28908) were added. The fixation was performed at room temperature (25°C). Two washes with HBSS buffer (Gibco, ref.: 14065-056) were performed with the aid of Mutlifo washer (BioTek). 20 µl/well of the staining solution (HBSS, 1% BSA, 0.1% Triton X-100, 43.7 nM PhenoVue Fluor 555 – WGA; 48 nM PhenoVue Fluor 488 - Concanavalin A; 8.25 nM PhenoVue Fluor 568 – Phalloidin; 1.62 µM PhenoVue Hoechst 33342 Nuclear Stain; 6 µM PhenoVue 512 Nucleic Acid Stain) were added and the plates were incubated for 30 min at room temperature before being washed again three times with HBSS. The plates were then sealed with aluminium foil and images were recorded directly.

Morphological profile generation

Image acquisition

ImageXpress Micro 4 epifluorescent microscope (Molecular Devices) with 20x air objective was used for recording the fluorescent images (16-bit). The camera binning was set to 2x2. The total imaged area per well spanned 2163 µm x 2163 µm and consisted of 3 by 3 adjacent fields of view placed in the centre of the well. For each field of view images were recorded in 5 channels. The following filter sets were used: DAPI, GFP, Cy3, Texas Red, Cy5. The Z-offset and exposure times were set separately for each channel. A total of 207,360 images were acquired in this campaign.

Feature extraction

Morphological features were extracted using CellProfiler (version 4.2.1), the cell analysis software developed by the Broad Institute (Stirling *et al.*, 2021). Two CellProfiler pipelines were used: one pipeline for image illumination correction, and one pipeline for the image analysis. The image illumination correction works at plate level and averages the intensity of the images of each channel. With the image analysis pipeline, objects were segmented on each image, the objects were labelled using the channel they were segmented on, and thousands of measurements were made on those objects at cell levels. Measures were also taken at image level. A total of 4761 features were measured and formed the morphological profile of a given cell.

Aggregation and normalization

After the extraction of the cell morphological profiles with CellProfiler, features were aggregated at well levels, by taking the means of each feature.

The features were then normalized using the "mad robustized" method of the Python pycytominer package provided by the Broad Institute (Serrano *et al.*, 2023). The normalization was made relative to all wells of a plate; for each feature, the median of the wells of a plate was subtracted, divided by the median absolute deviation (MAD) of the wells of a plate and multiplied by 1.4826 to have an unbiased estimator (Park, Kim and Wang, 2022). A value of 1e-18 was added to the MAD to avoid having a null denominator when the MAD was null.

Quality check

To check the quality of the Cell Painting experiment, several metrics were computed. First, the number of cells of the negative control treatment (DMSO) were monitored, which was an output of the CellProfiler segmentation: the cell numbers should lay within the range of [1800; 3000] cells per well. The coefficient of variation of the number of cells for the negative control treatment should not be above 15% per plate. All plates passed this initial quality control.

To identify other potential technical issues with the experiment, the Pearson correlations of positive controls across plates were calculated. These positive controls were selected to elicit very distinct and reproducible morphological profiles and were included in each plate. In an experiment of good quality, the correlation between replicates of these treatments should be above 0.8. No outlier plates were identified at this stage, all plates passed this quality control step.

In the morphological profiles, at well level, 10 values were missing. Most of them were coming from the feature 'Cells_AreaShape_FormFactor'. This feature was removed, along with 2 wells, to remove all missing values.

Additional outliers were detected based on the number of cells within a group of replicates. Some wells had an extreme difference of cell counts, more than 1800 cells, compared to other replicates of the same treatment. 22 wells were identified as outliers and removed.

Unsupervised feature selection

To reduce the dimensionality of the normalized morphological profiles, an unsupervised feature selection was performed with the "feature_select" function of the pycytominer Python package (Serrano *et al.*, 2023).

This function performed several steps to select the features. First, highly correlated features were removed: for a pair of features having a Pearson correlation greater or equal to 0.9, the feature having the smallest sum of correlations with other features were removed. Second, features with low variances were taken out; for a given feature, if the count of the second most common feature value divided by the count of the most common feature value was lower than 0.05, the feature was removed. Moreover, features having the ratio defined as the number of unique feature values divided by the number of samples, less than 0.01, were excluded. Third, a list of features (contained in the package), that are known to be noisy and generally unreliable have been removed. Fourth, features with at least one absolute value greater than 500, values

considered as outliers, were not kept. Fifth and finally, within one treatment group, noisy features were removed; they are features with a standard deviation within the same group of replicates greater than 1.2.

Eventually, a total of 644 features were kept, and were used for downstream analysis.

Consensus profiles

For a given treatment, in our case a chemical compound at a given concentration, the consensus profiles were obtained by aggregating the replicates, taking for each of the remaining features after the unsupervised feature selection, the median values of the replicates.

Morphological change signal measure

We used the grit score (Serrano *et al.*, 2023), a metric developed by the Broad Institute, to measure the morphological changes, with regards to the negative control treatment (DMSO), of a treatment replicate. To calculate this metric, different Pearson correlation coefficients were calculated. First, the correlations between the morphological profiles of a given treatment replicate and each of the negative control treatments were calculated. The distribution of those correlations was defined by its mean and its standard deviation. Then, the correlations between the morphological profiles of a given treatment replicate and other replicates of this treatment were computed. Each of the previous correlation coefficients were z-transformed using the distribution of the correlations with the negative controls: the mean was subtracted, and the results were divided by the standard deviation. The grit scores were then obtained by taking the mean of the transformed values.

Thus, the grit score informed on how much a given replicate profile deviated from the negative control profiles. A negative grit score indicated a problem of correlations between the

replicates. A high grit score indicated a high deviation of the profile from the negative control profiles.

Median grit score for a given treatment were also calculated, taking the median of the treatment replicate grit scores. This median grit score value allowed measuring how much a treatment impacts the morphology of U2OS cells, compared to the negative controls (DMSO).

We set a threshold of 1 from which a treatment is considered to induce a morphological change compared to the negative control. Indeed, a grit score of 1 means that the correlation of the morphological profile of a treatment to its replicates, is one standard deviation away from the mean of its correlation with the negative control profiles.

Molecular fingerprints

The compound structures were extracted from Bayer database as SMILES (Simplified molecular-input line-entry system). We used the Morgan fingerprints on 1024 bits in this analysis(Morgan, 1965; Rogers and Hahn, 2010). To obtain them from the SMILES, we performed the following steps using the RDKIT Python package (Landrum *et al.*, 2023).

First, the SMILES were cleaned using the MolStandardize module of RDKIT: hydrogens were removed, metal atoms were disconnected, the molecule were normalized and ionized again. When several fragments of a compound existed, the parent fragments were kept. The molecule was then neutralized and, the canonical tautomer was returned. Finally, the cleaned smiles were used to compute the Morgan fingerprints on 1024 bits, with a radius of three.

Chemical compounds clustering with Butina

The Butina clustering algorithm groups molecules based on their structural similarity (Butina, 1999). The RDKit implementation of the Butina algorithm was used to cluster the chemical

compounds (Landrum *et al.*, 2023). The clustering was based on the Tanimoto similarity of the Morgan fingerprints of the molecules, with a cut-off value of 0.7.

Dataset splits

To assess the performances of the binary classifiers, the dataset was split several times into training and testing sets. Two kinds of split were performed: a random one, without considering the chemical similarities of the compounds, and another split trying to create sets of structurally different chemicals, to produce cases where compounds to classify were outside the Applicability Domain of the chemical space (Figure 2).

For the random split, called 'easy case', a stratified 10 cross validation was performed to split the dataset into 10 different training and testing sets. The scikit-learn python package (Pedregosa *et al.*, 2011) was used to perform those splits with the StratifiedKFold function. The dataset was split 10 times with a 10-fold cross-validation having for each cross-validation a different random state, making a total of 100 different splits. Each testing set contained 22 or 23 compounds.

For the splits based on the chemical structures of the molecules, called 'difficult case', the compound structures were clustered using the Butina clustering algorithm (Butina, 1999). A cluster number was then assigned to each compound. The StratifiedGroupK-Fold function of scikit-learn was used to make a 10-fold cross-validation based on the cluster number (Pedregosa *et al.*, 2011). Indeed, this function assigned in the testing sets cluster numbers different from the cluster numbers in the training set and tried to keep in each set the same class (VAOT and NVAOT) ratio. The dataset was split in this manner 10 times, with different random state, making a total of 100 different splits. If some of those splits would not have compounds from the two classes in the test sets, they would be discarded.

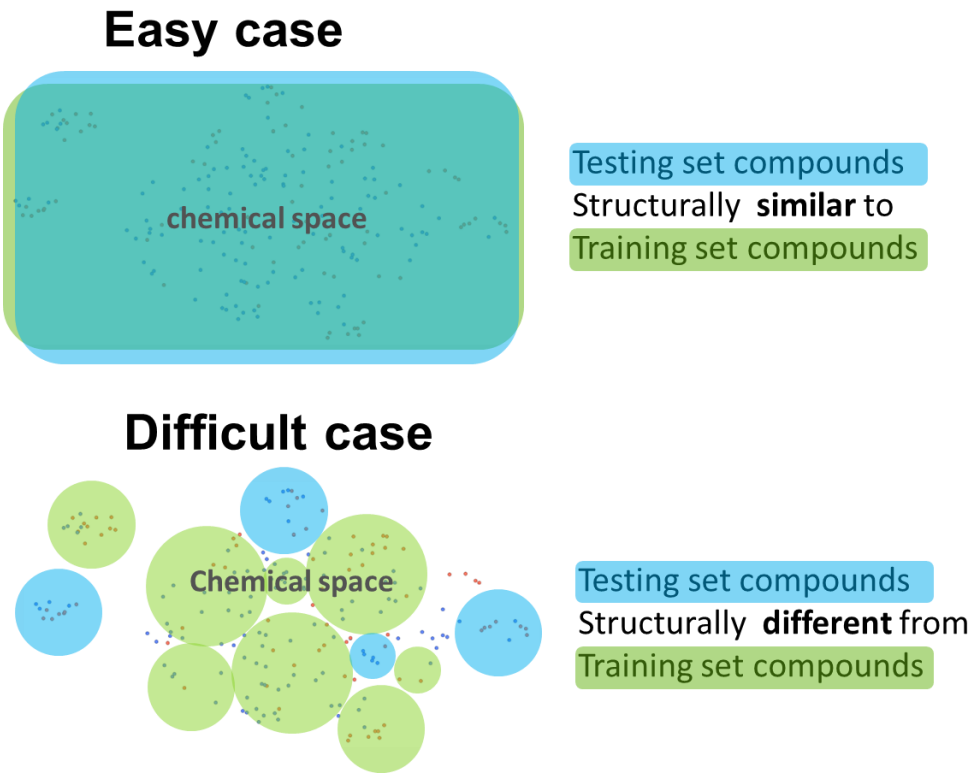

Figure 2.
Two data holdout strategies are used to train and test models: the easy case and the difficult case.

Binary classification model

To classify compounds as VAOT or NVOAT, several algorithms were tested. For this analysis, we decided to use a K Nearest Neighbors (KNN) algorithm (Kowalski and Bender, 1972), as it showed good performances (supplementary data). The KNN algorithm has also the advantage of being explainable and functions like a read-across, technique commonly used for toxicity prediction (Escher and Bitsch, 2021).

We used the scikit-learn (Pedregosa *et al.*, 2011) implementation of the K Nearest Neighbors (KNN) classification algorithm. Several models were built depending on the data that were used as input. When using the chemical Morgan fingerprints, the Tanimoto (Jaccard) distance was used and when using the morphological profiles, the correlation distance was used. For all

models, we set the number of neighbors to one. The choice of the distances and the number of neighbors were the results of benchmarking done on both datasets (supplementary data).

Decision support model

To aid the decision when the two types of classifiers (Morgan fingerprint and Cell Painting morphological profile models) did not predict the same class, a model was built, similar to the Similarity-based merger model (Seal *et al.*, 2023). This ensemble model takes as input the predictions of the two KNN classifiers, along with the distances of the nearest neighbors of each prediction. A classifier was trained in each training set, for the cases where the two KNN models did not agree on the predicted class. In the test sets, we used this model only when the two KNNs did not predict the same classes, otherwise the consensual predicted classes of the two model were set as the final class (Figure 3).

For this decision support model we used a SVM classifier (Cristianini and Ricci, 2008) implementation of Scikit learn (Pedregosa *et al.*, 2011).

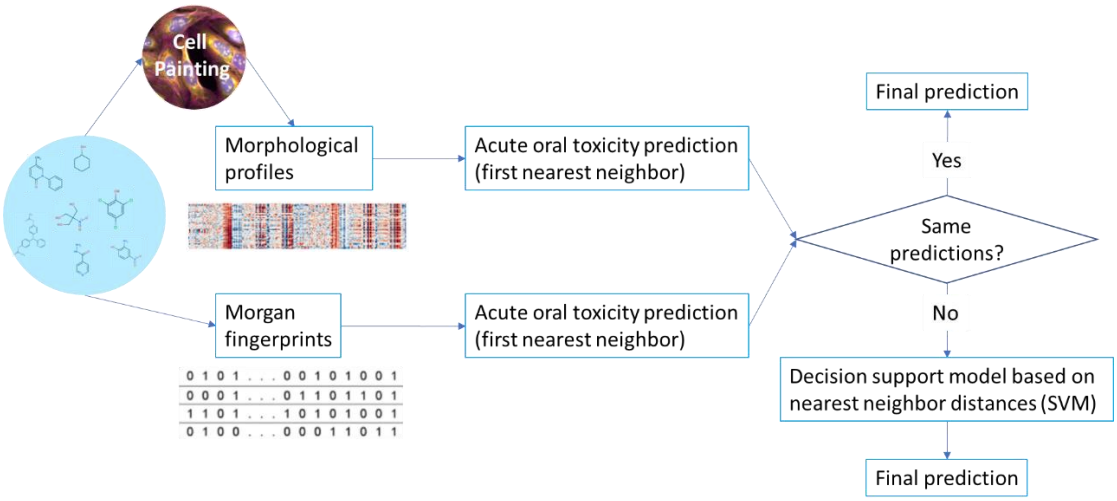

Figure 3.
Decision support model

Model performance evaluation

To evaluate the performance of the classifiers, we used different metrics. They were all based on the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN), which were the results of the model classification of a given testing set in a confusion matrix.

Sensitivity: SN = $T\frac{TP}{TP+FN}$

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

Balanced accuracy: BA = $\frac{SN+SP}{2}$

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

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

When using cross validation, those metrics were averaged over all testing sets, and their standard deviations calculated.

We performed a corrected t-test (Nadeau and Bengio, 2003) to compare the balanced accuracy values over the splits of classification models.

Visualization of the chemical and biological spaces

To visualize on two-dimensional scatter plots the chemical and the biological spaces, UMAP embeddings were generated using the Umap Python package (Sainburg, McInnes and Gentner, 2021). For the chemical space, the chemical compound structure similarities embeddings were calculated with the Tanimoto distances of the compound Morgan fingerprints. As for the biological space, the morphological profile similarities embeddings were computed using their correlation distances. The plots were visualized in TIBCO Spotfire software.

Results

The goal of our analysis was to compare the predictive efficacy or read-across approaches for acute oral toxicity in rats between chemical structural information, in vitro biological data derived from the Cell Painting profiling assay on U2OS cells or the combination of both. Two distinct types of inputs were utilized to construct KNN models classifying compounds as VAOT or NVAOT.

Initially, we categorized 630 Bayer Crop Science (BCS) agrochemical compounds with known acute rat toxicity results using the public QSAR model CATMoS (Mansouri *et al.*, 2021). Additionally, we present the results of custom QSAR model, trained on this specific unbalanced set of 630 compounds. Subsequently, KNN models were trained on a reduced but balanced set of 226 compounds using either the chemical structures or their morphological profiles in U2OS cells. A comprehensive analysis of both chemical space (made by the chemical structures) and biological space (revealed by the U2OS morphological profiles) was conducted to enhance our comprehension of the classifier results. Finally, we explored if combining the predictions of the two models could enhance the accuracy of the predictions.

Results of the QSAR models

Initially, CATMoS was employed to classify BCS compounds as either VAOT or NVAOT. We used the Opera CATMoS implementation of the model on the 'QSAR only compounds set' (Figure 1a), with 630 compounds as external test set. Predictions were mapped to the two classes, resulting in most compounds being classified as NVAOT. Specifically, 5 of the 109 VAOT compounds, and 514 of the 521 NVAOT compounds were correctly predicted, resulting in a low sensitivity of 0.05, a high specificity of 0.99, a balanced accuracy of 0.52 and a MCC of 0.09 (table 2a, table S5). This outcome could be due to the fact that CATMoS QSAR model was not trained on Bayer chemistry, but on mostly publicly available industrial chemical compounds, indicating a possible mismatch in the Applicability Domain of the model for BCS chemistry.

or the predictions of our set of 630 compounds, we could figure out that most compounds, 628 (99,6%) were inside the CATMoS Applicability Domain (Table 2c). Most of them, 448 (71%), had an Applicability Domain index below 0.6, suggesting that the predictions should be considered with caution (Table 2c). The remaining predictions, 180 (29%) having an Applicability Domain index above 0.6, displayed an average confidence level of 0.57, suggesting a relatively low level of confidence in the predictions (Table 2c).

Subsequently, we developed a model, based on our 630 compound set, using a KNN classifier on Morgan fingerprints of chemical compounds. The model's performance was assessed through cross-validation with two data holdout strategies: the "easy case", where compounds from the test sets resemble those from the training sets, and the "difficult case", where compounds from the test sets differ structurally from those in the training sets

For the "easy case", the classifiers exhibited an average balanced accuracy of 0.81, a sensitivity of 0.70, a specificity of 0.92 and a MCC of 0.61 (table 2b).

In the "difficult case", out of 100 theoretical cross-validation splits, only 44 included the two classes (VAOT and NVAOT) in both training and testing sets and hence were valid. On average, the model demonstrated a balanced accuracy of 0.60, a sensitivity of 0.33, a specificity of 0.87 and a MCC of 0.19 (table 2b).

The "difficult case" reiterated the diminished performance of chemical structure-based models when tasked with classifying compounds structurally distant from those used to train the model. In summary, the QSAR models demonstrated good performance in handling known chemistry but as expected by the design of the "difficult case", exhibited a decrease of performance when confronted with unfamiliar chemical structures. This limitation becomes apparent when exploring new areas of chemical space. To address this constraint, we leveraged

the biological effects of chemical compounds for predictions. The subsequent section details our approach, employing Cell Painting assay on U2OS cells to capture the biological effects of the chemical compounds.

QSAR only compounds set (set of 630 compounds)

a. Classification using CATMoS

		Predicted class	
		VAOT	**NVAOT**
True class	**VAOT**	5	104
	NVAOT	7	514

Balanced accuracy	**MCC**	**Sensitivity**	**Specificity**
0.52	0.09	0.05	0.99

b. Prediction reliabilities

Within Applicability Domain	**Applicability Domain Index**	**Number of compounds (percentage)**	**Average confidence index**
No	all	2 (0.3%)	NA
Yes	< 0.6	448 (71.1%)	0.5
Yes	≥ 0.6	180 (28.6%)	0.57

c. Performance of the KNN classifiers trained on 630

Holdout strategy	Balanced accuracy	MCC	Sensitivity	Specificity
Easy case	0.81	0.61	0.70	0.92
Difficult case	0.60	0.19	0.33	0.87

Table 2.

a. Confusion matrix and metrics for the classification of CATMoS over 630 compounds from Bayer Crop Science.

b. Reliability of the predictions: Number of compounds outside the CATMoS applicability domain, number of compounds and average confidence index for compounds within the CATMoS applicability domain and having an Applicability Domain index below or above 0.6.

c. Mean of 4 metrics assessing the performance of the KNN binary classifiers built out 630 Bayer CropScience agrochemical candidates, over the 100 splits of the "easy case" where training and testing sets are split randomly, not taking into account chemical structure similarities, and over the 44 valid splits of the "difficult case" where training and testing sets are split in order to have structurally different compounds over the two sets.

Comparison of QSAR and Cell Painting morphological based models

Our study aimed at comparing two inputs for predicting acute oral toxicity, utilizing a dataset called 'Cell Painting set', a subset of the 'QSAR only compound set', augmented with additional public chemical compounds. This set had a total of 226 compounds (Figure 1b). KNN classifiers were trained using both types of input and employing two data holdout strategies: easy and difficult cases.

Similar to the previous QSAR model on the QSAR only compound set (630 compounds), KNN models were trained on the Morgan fingerprint of the molecules.

For models based on the morphological profiles obtained from Cell Painting, consensus profiles were utilized after normalization, unsupervised features selection, and replicate profile aggregation at treatment level. As for the QSAR models, we used KNN algorithm. Models were built for each tested concentration (10 µM, 31.6 µM, 100 µM).

Results in the Easy Case

In the easy case, the QSAR model outperformed other models, achieving a mean balanced accuracy of 0.82, followed by the 31.6µM morphological profile model, with a mean balanced accuracy of 0.75 (Table 3a). The two other morphological profile models at 10 and 100 µM had lower performances. (Table 3a).

The distribution of the balance accuracy values over the 100 splits for each input type of input showed a narrow range (Figure 4a) confirmed by low standard deviations ranging from 0.08 to 0.09 (Table 3a).

Statistical analysis using the Nadeau and Bengio's corrected t-test to compare the balance accuracy values over the 100 splits of the two top models, indicated the QSAR model significantly outperformed the 31.6 µM morphological profile model (p-value= 0.05).

Results in the Difficult Case

In the difficult case, the 31.6 µM morphological profile model exhibited superior performance achieving a mean balanced accuracy of 0.72, followed by the QSAR model, with a mean balanced accuracy of 0.60 (Table 3b). The two other models had lower performances (Table 3b).

The distributions of the balanced accuracies for each model showed that for some splits, the QSAR model had difficulties making good predictions (Figure 4b). This was also the case, to a lesser extend for the 31.6 µM morphological profile-based models (Figure 4b).

We used The Nadeau and Bengio's corrected t-test suggested the 31.6 µM morphological profile model significantly outperformed the QSAR model (pvalue=0.045).

Summarizing the results, for the QSAR models, we reproduced the pattern of the previous QSAR model that was trained on almost three times more compounds (630 compounds) with good performances in the easy case, and a decrease in performance in the difficult case (the balanced accuracy dropped from 0.82 to 0.60).

Overall, our findings emphasize the superior performance of QSAR model in the easy case while morphological profile-based model remain valuable particularly in the difficult case, where the classifiers based on the 31.6 µM morphological profile demonstrated the highest performance (balanced accuracy of 0.72).

Performances of the classifiers

Types of models	a. Easy case					b. Difficult case				
	ACC	**BA**	**MCC**	**SN**	**SP**	**ACC**	**BA**	**MCC**	**SN**	**SP**
Morgan FP	0.82 ± 0.08	0.82 ± 0.08	0.65 ± 0.16	0.82 ± 0.11	0.82 ± 0.12	0.61 ± 0.15	0.60 ± 0.16	0.20 ± 0.29	0.49 ± 0.28	0.71 ± 0.18
CP 10 µM	0.57 ± 0.08	0.57 ± 0.08	0.14 ± 0.16	0.55 ± 0.14	0.58 ± 0.14	0.50 ± 0.13	0.50 ± 0.14	0.00 ± 0.26	0.46 ± 0.24	0.54 ± 0.15
CP 31.6 µM	0.75 ± 0.09	0.75 ± 0.09	0.51 ± 0.18	0.72 ± 0.13	0.79 ± 0.12	0.71 ± 0.13	0.72 ± 0.13	0.42 ± 0.24	0.68 ± 0.23	0.76 ± 0.13
CP 100 µM	0.63 ± 0.09	0.63 ± 0.09	0.27 ± 0.19	0.61 ± 0.13	0.65 ± 0.14	0.52 ± 0.13	0.53 ± 0.13	0.07 ± 0.25	0.47 ± 0.23	0.60 ± 0.18
DS	0.85 ± 0.07	0.85 ± 0.07	0.71 ± 0.14	0.86 ± 0.11	0.84 ± 0.11	0.71 ± 0.13	0.72 ± 0.13	0.42 ± 0.26	0.67 ± 0.23	0.77 ± 0.16

Table 3.

Mean and standard deviations of 5 metrics: Accuracy (ACC), Balanced Accuracy (BA), Matthew's correlation coefficient (MCC), Sensitivity (SN), Specificity 'SP). Four different input data were used to classify compounds as VAOT and NVAOT: chemical structural data (Morgan FP) and Cell Painting (CP) morphological profiles of U2OS cells put in presence of the chemical compounds at 3 concentrations (MP 10 µM, MP 31.6 µM and MP 100 µM). Are highlighted in orange the best average metric. The performances of the decision support (DS) model combining the Morgan Fingerprint and the morphological profile 31.6 µM model predictions are shown in the last row.

a. Assessing the performance of the binary classifiers, over the 100 splits of the easy case where training and testing sets are split randomly, not considering chemical structure similarities.

b. Assessing the performance of the KNN classifiers, over the 99 valid splits of the difficult case where training and testing sets are split to have structurally different compounds over the two sets.

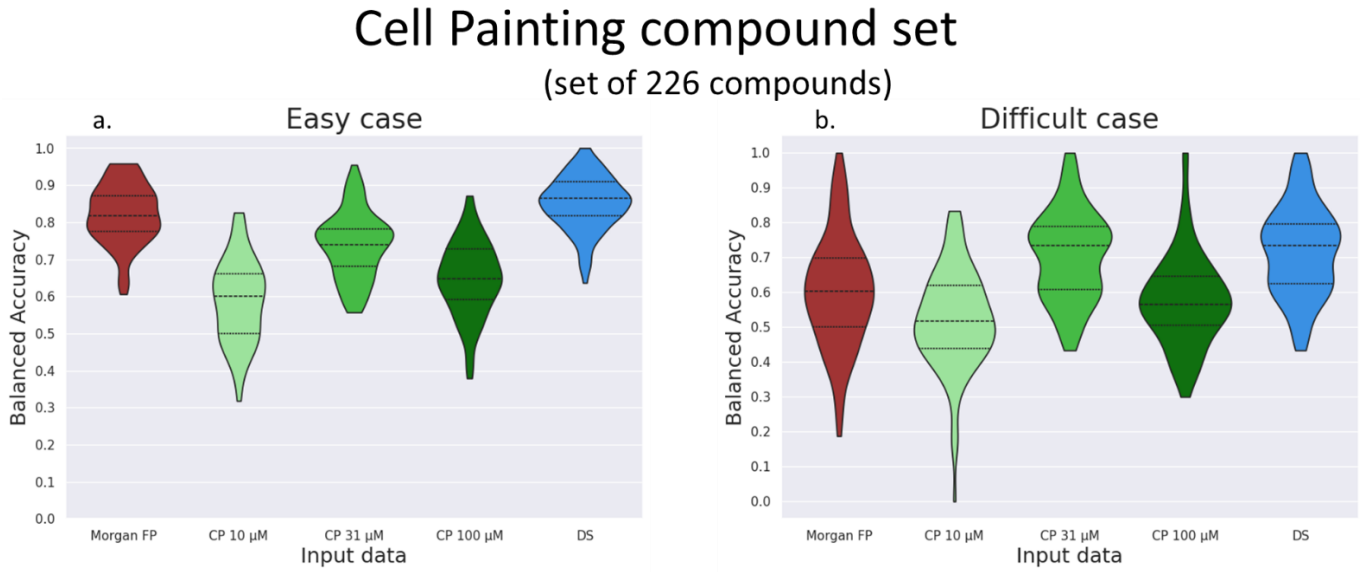

Figure 4.

a. Violin plot representing the balanced accuracies of the binary classifier for the 10 x 10-fold cross validation splits not considering the structure similarities.
b. Violin plot of the balanced accuracies of the KNN binary classifier for 99 valid splits of the 10 x 10-fold cross validation that put in the testing set chemical structurally different from the training set.

Legend: In red, the classifier using the Morgan Fingerprint, in light green the classifier using the morphological profiles at 10 µM, in green the classifier using the morphological profiles at 31.6 µM and in dark green, the classifier using the morphological profiles at 100 µM, in blue, the decision support (DS) model. Inside each violin plot the quartiles are indicated as dash lines.

Comparison of the chemical and biological spaces

To understand the results of the classifiers based on the chemical structures and the morphological profiles, we investigated the chemical and biological spaces, for the Cell painting set of compounds.

Chemical space

Our dataset comprises 226 compounds primarily originating from Bayer Crop Science chemistry and supplemented with 29 public compounds. Employing the Butina algorithm, using the Tanimoto distance and a threshold of 0.7, we identified 91 clusters, 61 of them were represented by a unique compound indicating structural diversity.

Visualizing the similarity of the structures on a UMAP plot (Figure 5), using Morgan fingerprints and Tanimoto distance revealed distinct clusters. Notably, certain clusters exclusively comprised VAOT compounds (e.g., cluster A), others exclusively comprised NVAOT compounds (e.g., cluster B) while several contained a mixture of both (e.g., clusters C and D)

In summary, the chemical space exhibited diversity made of different clusters. Specific areas demonstrated a prevalence of either VAOT or NVAOT compounds, while others presented a combination of both classes.

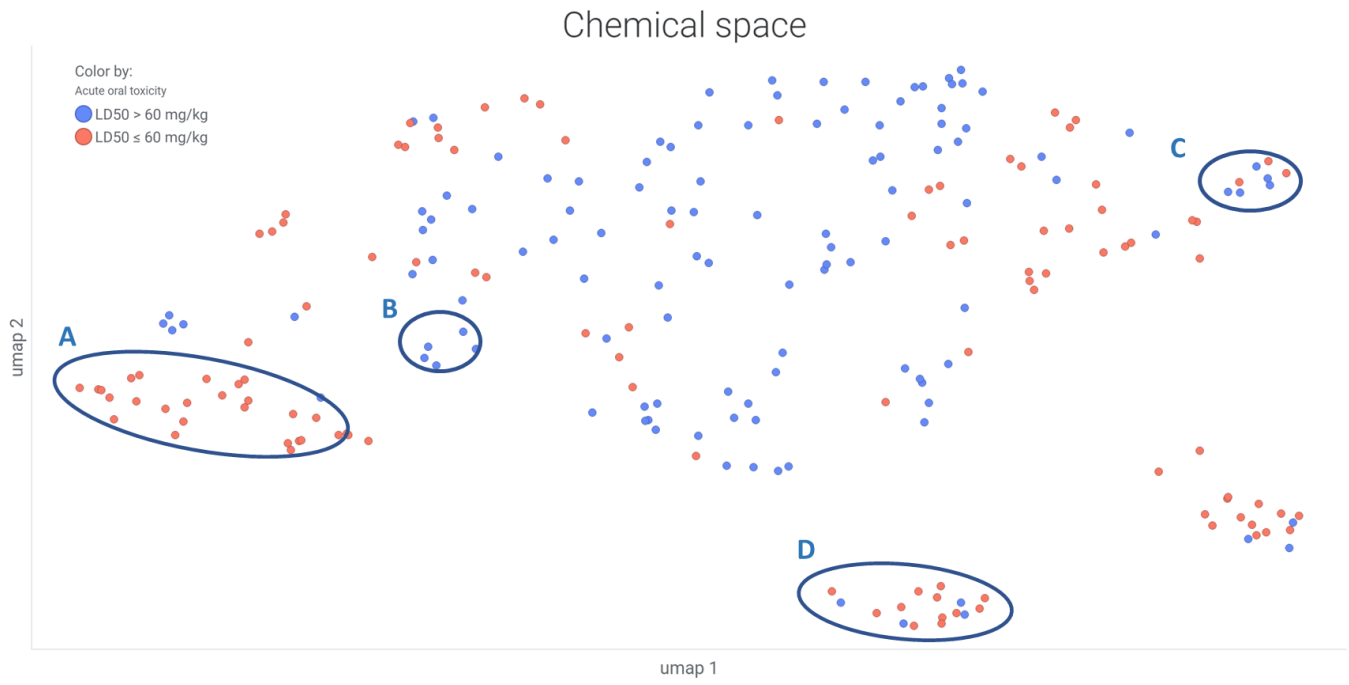

Figure 5.

Scatter plot of the 2-dimensional UMAP embedding of the chemical compound morgan fingerprints. In blue, the chemical compounds that are NVAOT. In red, chemical compounds that are VAOT. Four clusters of compounds are designated by the letter A, B, C and D. Cluster A is an example of a cluster with only VAOT compound. The cluster B is an example of a cluster with only NVAOT compounds. B and C are two examples of clusters with a mix of VAOT and NVAOT compounds.

Biological space

In Figure 6, we depicted the similarity of the biological response of compounds on a UMAP plot, utilizing morphological profiles for the three concentrations and the correlation distance. Unlike the chemical space, a limited number of clusters emerged with only two notable clusters observed. An isolated small cluster (cluster A) was clearly separated from other profiles, and upon inspection these profiles corresponded to instances with notably low cell count.

The second cluster exhibited diverse areas: including regions with a high number of VAOT (e.g., grouping B), NVAOT compounds (e.g., grouping C), and areas with a mix of both classes (e.g., grouping D).

In Figure 7, we focused on the 31.6µM concentration, the concentration yielding optimal performances for the classification model using morphological profiles. Similar observations were made, with an isolated cluster corresponding to profiles with a very low number of cells. Additional distinct areas emerged, showcasing regions with high number of VAOT or NVAOT compounds, as well as areas with a mixed representation of both classes.

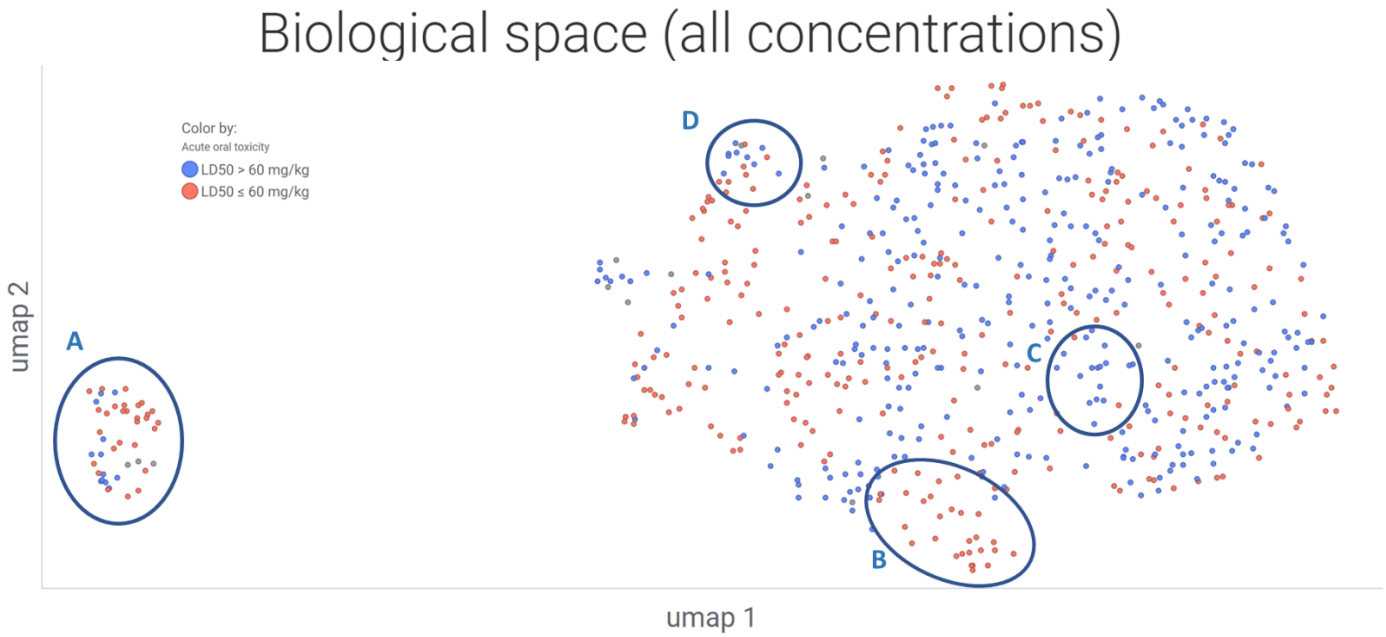

Figure 6.

Two-dimensional representation of the consensus morphological profile similarities for all treatments and all concentrations, using Uniform Manifold Approximation (UMAP) embedding on 2 components with the Pearson correlation distances. In red, the morphological profiles of U2OS cells perturbated by VAOT compounds. In blue, the morphological profiles of U2OS cells perturbated by NVAOT compounds. Four groups of compounds are designated by the letter A, B, C and D. The group A of compounds corresponds to treatment with very low cell counts. The group B of compounds is an example of grouping with a high number of VAOT compounds. The group C of compounds is an example of grouping with a high number of NVAOT compounds. The group D of compounds is an example of grouping with a mix of VAOT and NVAOT compounds.

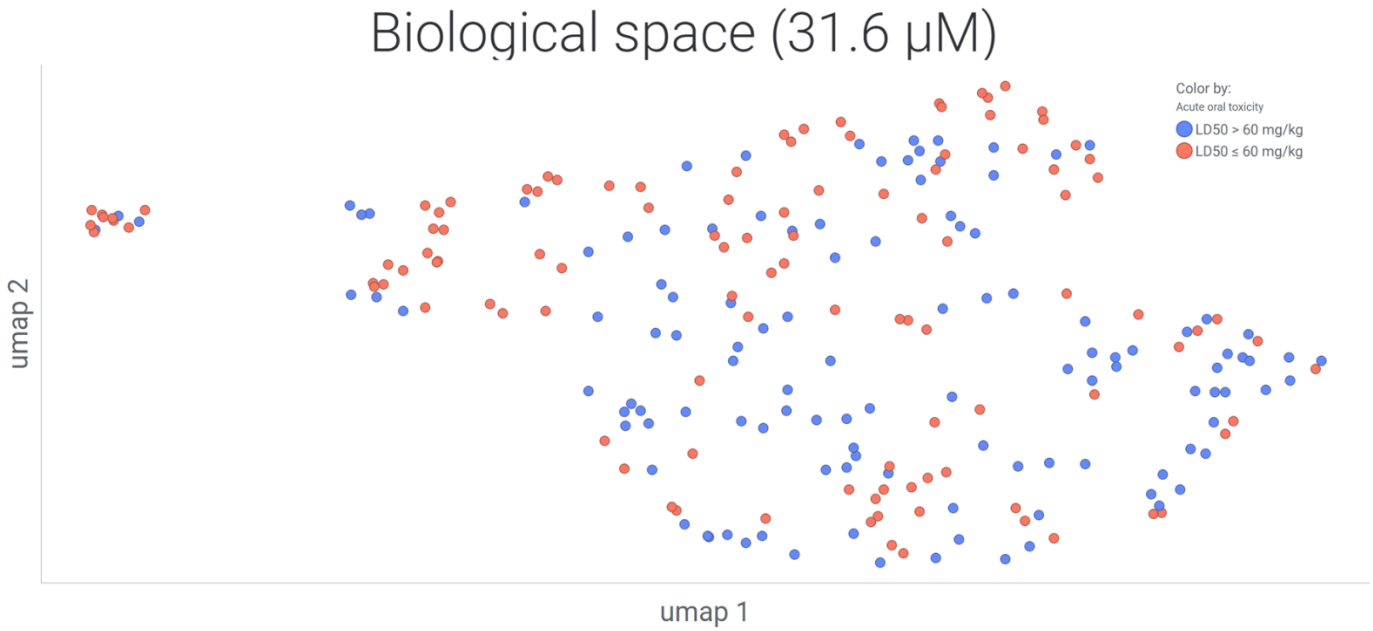

Figure 7.

Two-dimensional representation of the consensus morphological profile similarities for all treatments at 31.6 µM, using Uniform Manifold Approximation (UMAP) embedding on 2 components with the Pearson correlation distances. In red, the morphological profiles of U2OS cells perturbated by VAOT compounds. In blue, the morphological profiles of U2OS cells perturbated by NVAOT compounds.

Comparison of the chemical and biological spaces

We compared different groups of compound structures and groups of morphological profiles to better understand the chemical and biological space interrelationship. Notably, we observed that chemicals clustering together in the chemical space could elicit a diversity of biological responses in the biological space, emphasizing that structurally similar compounds may manifest distinct biological responses (Figure 8, top row). We illustrate a specific case involving a group of chemicals, the carbamates, inducing similar morphologies in U2OS cells (Figure 8, middle row). Interestingly, similar morphological profiles induced by structurally different compounds could also be observed (Figure 8, bottom row). This illustrates that biological

effects of structurally similar molecules may not necessarily be alike. Structurally similar compounds could trigger different biological effects. Conversely, compounds with different structures could cause comparable biological responses.

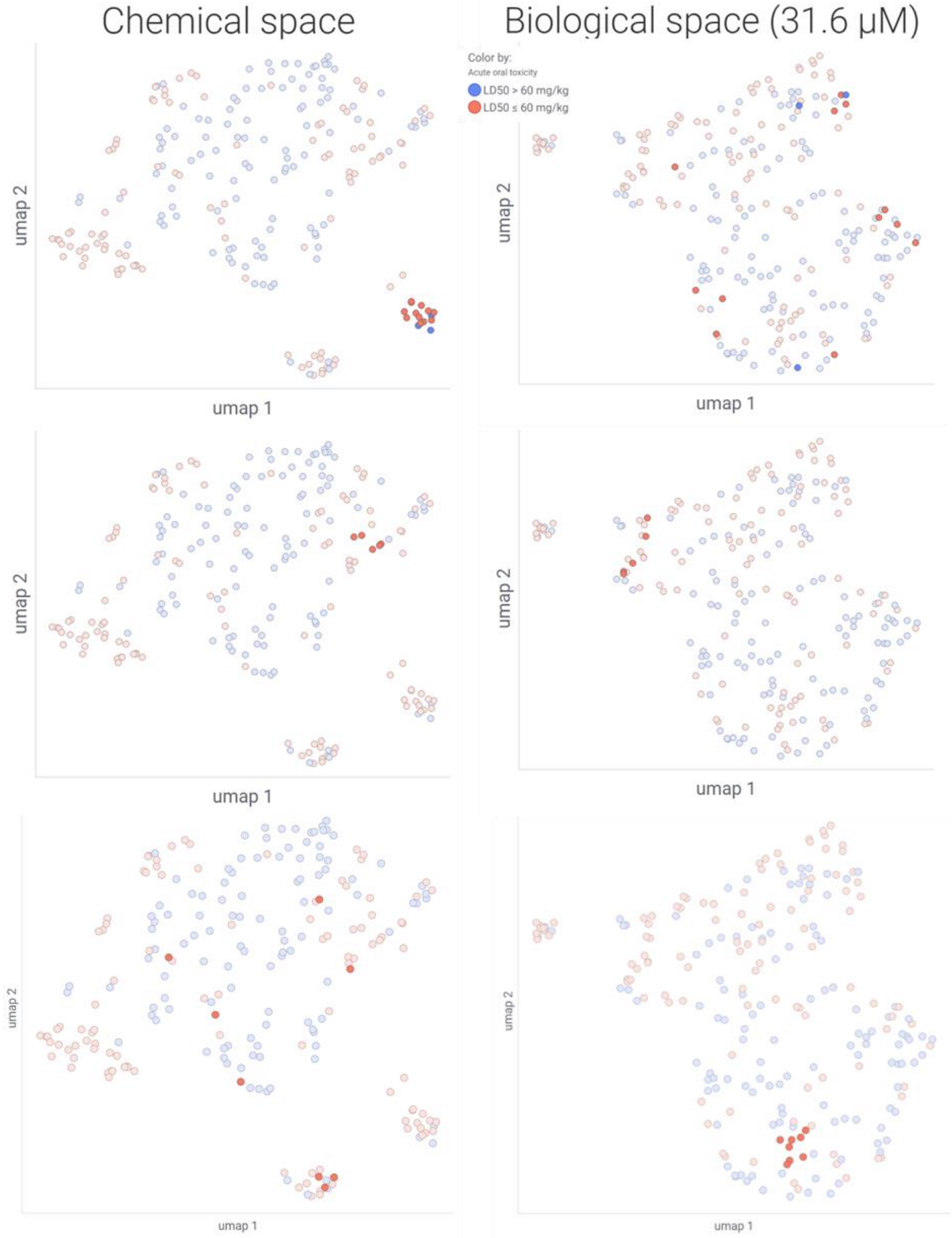
Chemical space
Biological space (31.6 μM)
Color by:
Acute oral toxicity
LD50 > 60 mg/kg
LD50 ≤ 60 mg/kg
umap 2
umap 1

Figure 8.

(left) Chemical space. Scatter plot of the 2-dimensional UMAP embedding of the chemical compound morgan fingerprints.

(right) Biological space. Scatter plot of the 2-dimensional UMAP embedding of the Cell Painting morphological profiles of the chemical compounds at 31.6 µM.

In red, the morphological profiles of U2OS cells perturbated by VAOT compounds.

In blue, the morphological profiles of U2OS cells perturbated by NVAOT compounds.

On each row, the chemical profiles and morphological profiles of the same compounds are selected.

(top) Example of structurally similar compounds, inducing different morphological profiles.

(middle) Example of structurally similar compounds, inducing similar morphological profiles.

(bottom) Example of structurally different compounds, inducing similar morphological profiles.

Biological response

To analyze the biological response of U2OS cells to chemical compound perturbations, we employed morphological profiles, using two metrics: the Grit score (Serrano *et al.*, 2023), and the number of cells. We analyzed how strong the biological response of a given treatment was to better understand how U2OS cells reacted to our set of compounds, hence also helping in understanding the results of the classifiers.

Grit score

The Grit score was an indication of how much the average morphology of U2OS cells perturbed by a treatment deviated from the average negative control morphology of non-perturbed U2OS cells. A high Grit score signifies a more distinct cell morphology from the negative controls. For example, the average grit score of the positive controls was 4.8.

A Mann-Whitney U rank test on the grit scores, for the two compounds groups , VAOT and NVAOT, demonstrated that VAOT compounds elicited a marginal though significant stronger

biological response compared to NVAOT (Grit score respectively 3 and 2.5, p-value of 0.004). (Table 4).

Regarding concentrations, on average, the 10 µM treatments had a grit score of 1.9, the 31.6 µM had a grit score of 2.6 and the 100 µM treatment had a grit score of 3.7. This aligns with our assumption that higher concentrations lead to increased biological responses, a consideration made when designing the Cell painting campaign with 3 concentrations (Table 4).

Identifying compounds with no induced morphological changes, we set a Grit score threshold of 1. Below this threshold, we considered a treatment not inducing any morphological change. Of the 23 compounds falling below this threshold 6 were VAOT.

Number of cells

An additional output of our image analysis was the number of cells per well. For this analysis the number of cells were not normalized, and the median number of cells per well for a given treatment were computed. The average number of cells for the negative controls was 2231. We arbitrary set the number of cells that defines cytotoxicity as a cell count below 50% of the average negative control cell count, meaning a cell count below 1115 defined a cytotoxic treatment.

In total 44 compounds exhibited cytotoxicity: 12 compounds, at 10 µM, 23 compounds, at 31.6 µM and 44 compounds, at 100 µM (Table 4).

Categorizing by class, 28 VAOT compounds (25%) and 16 NVAOT compounds (14%) displayed cytotoxicity for at least one concentration. A chi-square test of independence of variables, with the null hypothesis that the number of cytotoxic compounds is independent of the class (VAOT and NVAOT) gave a p-value of 0.1. We could not conclude that there were a higher percentage of cytotoxicity for VAOT compounds (Table 4).

profiles	Average Grit score	Number of cytotoxic treatments
Negative control	NA	0
VAOT	3	28
NVAOT	2.5	16
10 µM	1.9	12
31.6 µM	2.6	23
100 µM	3.7	44

Table 4.

Average grit score and number of cytotoxic treatments for different groups of profiles: VAOT (very acutely oral toxic compounds), NVAOT (non very acutely oral toxic compounds), 10 µM treatment profiles, 31.6 µM treatment profiles and 100 µM treatment profiles.

Results of the decision-support model

The decision-support model aided the decision when the two KNN classifiers did not predict the same class. The model combined four pieces of information: predictions from the KNN model based on the chemical structure information, predictions from the KNN model based on the morphological profiles and distances to the nearest neighbor in each model.

In the easy case, the model had an average balanced accuracy of 0.84, slightly above the QSAR model's average balanced accuracy of 0.82. In the difficult case, the model had on average a balanced accuracy of 0.65, below the 31.6µM morphological profile model's average balanced accuracy of 0.72.

To understand why in the difficult case, this model did not yield better performances, we computed the mean Morgan fingerprint Tanimoto distances between each chemical compound of the training set and its nearest neighbor in the training set, and the mean distances between each chemical compound of the testing set and its nearest neighbor in the training set.

For the easy case, on average, in the training set, each compound has a distance to its nearest neighbor of 0.50 and in the testing set 0.48. For the difficult case, on average, in the training set, each compound has a distance to its nearest neighbor of 0.49, and in the testing set 0.73.

We could notice that the training set did not have enough examples of distant chemical structures. To help the model, we added synthetic examples of distant chemical structures. To do so, we subset in each training set the cases where the predictions of the QSAR did not match the real class, and we updated the distances of the nearest neighbors with a random number between 0.7 and 0.9 and added those synthetic examples in the dataset used to train the model.

By doing so, in the easy case the model had an average balanced accuracy of 0.85. In the difficult case, the model had an average balanced accuracy of 0.72 (Table 3).

Discussion

Our results showed that the classification of compounds, using a read-across approach, as very acutely oral toxic or not, was possible using chemical structure information, U2OS cell morphological profiles or the combination of both. When classifying compounds structurally similar to those used to train the classifier, the chemical structure information was more predictive. Conversely, when compounds to classify were structurally different from compounds used to train the classifier, the U2OS cell morphological profiles were more predictive.

Initial attempts with the publicly available QSAR model, CATMoS , for the prediction of acute oral toxicity on a set of 630 Bayer compounds did not yield good predictions (Mansouri *et al.*, 2021). CATMoS performance is hindered as Bayer Crop Science chemistry could be considered as locally outside its applicability domain (Table 2b). Although nearly all compounds were globally within the CATMoS applicability domain, most resided in gaps in that applicability domain. It is known that QSAR models excel when compounds being classified fall within their applicability domain of the models, and perform poorly when they do not (Kar, Roy and Leszczynski, 2018). In summary, CATMoS, which is a QSAR model trained on more than 10,000 compounds has very good performances for the prediction of acute oral toxicity but fails to work effectively with Bayer Crop Science chemistry.

To confirm our hypothesis, we trained a simple KNN classifier, resembling a read-acrross approach, on this set of 630 compounds, using their chemical structure information. Working with two data holdout strategy to simulate scenarios within and outside the applicability domain of a QSAR model, we evaluated our models under two conditions: the easy case, simulating scenarios within applicability domain case, and the difficult case, attempting to simulate outside applicability domain case. In the easy case, our model exhibited strong performance, comparable to CATMoS. For example, CATMoS achieved a balanced accuracy of 0. 84, in classifying compounds as very toxic (VT) (LD50 < 50 mg/kg) (Mansouri *et al.*, 2021) whereas our model had a balanced accuracy of 0.82 for the classification of compounds as VAOT (LD50 < 60 mg/kg).

As designed, the performances of the classifier dropped in the difficult case due to the data holdout strategy, which placed Butina compound clusters not present in the training sets into the testing sets . This effectively simulated scenarios outside applicability domain, although the decrease in balanced accuracy (from 0.82 to 0.60) was not as drastic as observed with Bayer CropScience chemistry using CATMoS (from 0.84 to 0.52).

To overcome this chemical applicability domain limitation, we explored whether using the compound biological effects could mitigate this issue. Compound-induced biological effects characterized with transcriptomics has already been used to predict target activities, in association and in comparison, with QSAR models (Baillif *et al.*, 2020; Moshkov *et al.*, 2023). Here we utilized Cell Painting to generate morphological profiles at a more reasonable cost compared to transcriptomics.

Using a smaller but balanced set of 226 compounds (49 % of compounds selected known to have LD50<60 mg/kg), we trained KNN classifiers based on either chemical structure information or U2OS morphological profiles at 3 concentrations.

Similar to the larger set of 630 compounds, we observed similar trends for the classifiers based on the chemical structure information: good performances in the easy case but decreased performances in the difficult case. Morphological profiles at 31.6 µM concentration demonstrated better performances in comparison to the other concentrations, in both the easy and difficult cases. With a balanced accuracy of 0.75 in the easy case and 0.72 in the difficult case, Cell Painting U2OS profiles demonstrated the capability to predict acute oral toxicity classes, interestingly, independently of the structural similarity of the tested compounds.

Cell Painting can indeed identify morphological patterns associated with specific mode of action (MOA) and molecular initiation event (MIE) of compounds (Ljosa *et al.*, 2013; Way *et al.*, 2022). Typically, acute toxicity involves a limited number of MIE (Prieto, 2019). such as narcosis (activity on the lipid bilayer of membrane), acetylcholinesterase inhibition, ion channel modulators and inhibitors of cellular respiration (A. Leblanc, 2004). Cell Painting experiment revealed morphological profiles (initiated by MIE) associated with acute oral toxicity, as

evidenced by the grouping of morphological profiles associated with VAOT compounds. For example, four carbamates (Promecarb, Methiocarb, Propoxur, m-Cumenyl methylcarbamate) known as acetylcholinesterase inhibitors produced similar morphological profiles in U2OS cells (Figure 8, middle row).

This partially explains why morphological profile-based models were capable, of correctly classifying compounds. In the easy case, the performances of the classifiers based on the 31.6 µM morphological profile did not surpass the classifiers based on the chemical structure information but did outperform them classifiers in the difficult case.

Capturing the biological effects of compounds brought limitations: limitation of the cell system to reveal the effects causally associated with acute toxicity, together with technical limitations of the lab experiment itself.

For the limitation of the cell system, we observed, through grit score analysis, that not all the compounds induced a biological response in U2OS cells (10%), regardless of the concentration used. Six VAOT compounds did not elicit any morphological change in U2OS cells compared to the negative controls. Among these compounds, 5 were public compounds and some of them with information on their possible mode of action. The Warfarin, a Vitamin K Antagonist, the Methamidophos, a potent acetylcholinesterase inhibitor, did not produce any biological response in U2OS cells. This suggests that U2OS cells have their own biological applicability domain and may not capture all the bioactivities associated with oral acute toxicity observed in a whole organism like a rat in our case study. Nonetheless, for our set of compounds, Cell Painting on U2OS managed to capture bioactivities for most of the VAOT compounds.

On the contrary, analyzing the number of cells, we could also identify a limitation due to the cytotoxicity of compounds: 44 compounds showed cytotoxicity at least at one concentration, and 12 exhibiting cytotoxicity even at the lowest concentration. Morphological profiles of cytotoxic treatments were not informative as they predominantly consisted of debris and dying cells. It appears that the concentration of 31.6 µM represented a good tradeoff between inducing bioactivity and avoiding cytotoxicity.

For the limitation of the experiments, several quality issues may arise when running an experiment in a laboratory. Experiments are technically demanding and prone to variability and errors. The seeding variability can impact the cell morphologies, hence the morphological profiles. There are other usual problems that could occur in laboratory experiments, such as treatment errors, compounds with low purity and precipitation at high concentration. Those issues could affect the quality of the morphological profiles and hence the performance of a classifier based on morphological profile similarities.

The chemical structural information did not suffer from those limitations, because this information was not subject to quality issue, was not cell system dependent, and not assay design dependent. This information was intrinsic of the description of a given compound. This could explain partly why QSAR models, in the easy case, performed better than biological based models: the full structural information is available, whereas the biological information is partly available and subject to quality issues in particular reproducibility.

In the “difficult case”, morphological profiles-based models did not experience performance drop as much as the QSAR models, suggesting that the biological space did not cluster the same way as the chemical space. This indicates that similar compounds did not consistently induce the same response in U2OS cells (due for example to activity cliffs), and vice versa. The presence of different Butina clusters in the training and test sets, did not necessarily result in different morphological profiles explaining why the morphological profile-based model performances did not drastically drop in the difficult case.

Using biological responses of compounds could be also an advantage with regards to enantiomers. The Morgan fingerprint used in this analysis does not consider chirality. Enantiomers can have different acute oral toxicity, and our QSAR model will not make the difference between them, where morphological profiles could be different.

The decision support model combined both predictions along with the distances of the nearest neighbor to make final predictions, slightly improving the classification performance in the easy case while decreasing in the difficult case. By adding a few artificial examples in each training

set of higher distances in the chemical spaces allowed increasing the classification accuracy in the easy case not in the difficult case, where the model performed like for the 31.6µM morphological based model. Notably, in the difficult case the classifier favored predictions from the 31.6µM morphological based model predictions over QSAR model predictions.

Further results could expand and refine these findings by employing a broader compound set covering all the molecular initiating event (MIE) linked to acute oral toxicity. Additionally, a wider set of compounds could facilitate the identification of more morphological profiles associated with acute oral toxicity. A larger dataset would also allow isolating a set of compounds as an external dataset to further assess the performance of the model.

The choice of the KNN algorithm in this analysis was deliberate due to its simplicity and resemblance to read-across approach commonly used in toxicology. The quantity of in vivo data being often limited, the read-across approach is often the only analysis that can be performed. For QSAR models, other algorithms yielded similar performances (Supplementary information).

For both types of input data, the optimal number of neighbors was 1 for the KNN algorithm, indicating that few examples of identified profiles leading to high acute oral toxicity or not, were present in the dataset. Having a larger set of compounds, like the CATMoS training set, would help identifying more examples of Cell Painting profiles linked to acute oral toxicity.

To help create public QSAR models with a wider applicability domain, representations of compound structures and results of acute toxicity studies for early candidates which failed to be placed on the markets could be shared by companies and organizations to expand the chemical space coverage.

Additionally, in this analysis, the Morgan fingerprint was the only computed chemical descriptor. Using additional descriptors (such as PaDEL) could help to have better QSAR performances.

We have also seen the limitation of the U2OS cell line, not capturing all the bioactivities of the compounds. Trying different cell lines could allow capturing more bioactivities linked to MIE

leading to acute oral toxicity. Several cell lines have already been used with Cell Painting (Cox et al., 2020; Nyffeler, 2020) and could help defining a set of cell lines capable of capturing a maximum, if not all, MIE leading to acute oral toxicity.

Finally, absorption, distribution, metabolism, and excretion (ADME) properties of compounds were not taken into consideration in this study but incorporating such data could enhance predictive models. We tried to use predicted maximum concentration in plasma and AUC from a predictive model (Schneckener *et al.*, 2019), but this information did not improve our results (data not shown).

In addition, preincubation of the compounds with liver S9 fractions (the 9000g supernatant of a liver homogenate), containing phase I and II metabolic enzymes, to work on the possible metabolites of a parent compounds, could also help when the toxicity is driven by a metabolite, as it is done for example with the Ames assay to test the mutagenic potential of chemical compounds (Hakura *et al.*, 2001; Hopperstad *et al.*, 2022).

In conclusion, a combined approach utilizing QSAR, and Cell Painting morphological profiles-based models based on chemical and biological spaces distances holds promise for predicting acute oral toxicity. Those models could be used in the context of early de-risking and in the future serve in context of Next Generation Risk Assessment (NGRA) aiming at refining if not replacing laboratory animal testing.

Funding Information.

FC holds a doctoral fellowship from the Association Nationale de la Recherche Technique (ANRT CIFRE PhD funding). This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002 through the Affiliated Chair of DR.

Acknowledgment.

The authors thank Dr. Oscar Mendes Lucio for his insightful suggestions to improve the article.

References

A. Leblanc, G. (2004) 'Mechanisms of acute toxicity', in *A Textbook of Modern Toxicology*. Ernest Hodgson.

Baillif, B. *et al.* (2020) 'Exploring the Use of Compound-Induced Transcriptomic Data Generated From Cell Lines to Predict Compound Activity Toward Molecular Targets', *Frontiers in Chemistry*, 8, p. 296. Available at: https://doi.org/10.3389/fchem.2020.00296.

Becker, T. *et al.* (2020) *Predicting compound activity from phenotypic profiles and chemical structures*. preprint. Bioinformatics. Available at: https://doi.org/10.1101/2020.12.15.422887.

Bray, M.-A. *et al.* (2016) 'Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes', *Nature Protocols*, 11(9), pp. 1757–1774. Available at: https://doi.org/10.1038/nprot.2016.105.

Butina, D. (1999) 'Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets', *Journal of Chemical Information and Computer Sciences*, 39(4), pp. 747–750. Available at: https://doi.org/10.1021/ci9803381.

'ChemIDplus' (2023). Available at: https://pubchem.ncbi.nlm.nih.gov/source/ChemIDplus.

Cimini, B.A. *et al.* (2023) 'Optimizing the Cell Painting assay for image-based profiling', *Nature Protocols*, 18(7), pp. 1981–2013. Available at: https://doi.org/10.1038/s41596-023-00840-9.

Cox, M.J. *et al.* (2020) 'Tales of 1,008 small molecules: phenomic profiling through live-cell imaging in a panel of reporter cell lines', *Scientific Reports*, 10(1), p. 13262. Available at: https://doi.org/10.1038/s41598-020-69354-8.

Cristianini, N. and Ricci, E. (2008) 'Support Vector Machines: 1992; Boser, Guyon, Vapnik', in M.-Y. Kao (ed.) *Encyclopedia of Algorithms*. Boston, MA: Springer US, pp. 928–932. Available at: https://doi.org/10.1007/978-0-387-30162-4_415.

Edwards, S.W. *et al.* (2022) 'Mapping Mechanistic Pathways of Acute Oral Systemic Toxicity Using Chemical Structure and Bioactivity Measurements', *Frontiers in Toxicology*, 4, p. 824094. Available at: https://doi.org/10.3389/ftox.2022.824094.

Erhirhie, E.O., Ihekwereme, C.P. and Ilodigwe, E.E. (2018) 'Advances in acute toxicity testing: strengths, weaknesses and regulatory acceptance', *Interdisciplinary Toxicology*, 11(1), pp. 5–12. Available at: https://doi.org/10.2478/intox-2018-0001.

Hakura, A. *et al.* (2001) 'An improvement of the Ames test using a modified human liver S9 preparation', *Journal of Pharmacological and Toxicological Methods*, 46(3), pp. 169–172. Available at: https://doi.org/10.1016/S1056-8719(02)00186-7.

Hopperstad, K. *et al.* (2022) 'Chemical Screening in an Estrogen Receptor Transactivation Assay With Metabolic Competence', *Toxicological Sciences*, 187(1), pp. 112–126. Available at: https://doi.org/10.1093/toxsci/kfac019.

Kar, S., Roy, K. and Leszczynski, J. (2018) 'Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling', in O. Nicolotti (ed.) *Computational Toxicology*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 141–169. Available at: https://doi.org/10.1007/978-1-4939-7899-1_6.

Kowalski, B.R. and Bender, C.F. (1972) 'K-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation', *Analytical Chemistry*, 44(8), pp. 1405–1411. Available at: https://doi.org/10.1021/ac60316a008.

Landrum, G. *et al.* (2023) 'rdkit/rdkit: 2023_03_3 (Q1 2023) Release'. Zenodo. Available at: https://doi.org/10.5281/ZENODO.591637.

Ljosa, V. *et al.* (2013) 'Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment', *Journal of Biomolecular Screening*, 18(10), pp. 1321–1329. Available at: https://doi.org/10.1177/1087057113503553.

Mansouri, K. *et al.* (2018) 'OPERA models for predicting physicochemical properties and environmental fate endpoints', *Journal of Cheminformatics*, 10(1), p. 10. Available at: https://doi.org/10.1186/s13321-018-0263-1.

Mansouri, K. *et al.* (2021) 'CATMoS: Collaborative Acute Toxicity Modeling Suite', *Environmental Health Perspectives*, 129(4), p. 047013. Available at: https://doi.org/10.1289/EHP8495.

Morgan, H.L. (1965) 'The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.', *Journal of Chemical Documentation*, 5(2), pp. 107–113. Available at: https://doi.org/10.1021/c160017a018.

Moshkov, N. *et al.* (2023) 'Predicting compound activity from phenotypic profiles and chemical structures', *Nature Communications*, 14(1), p. 1967. Available at: https://doi.org/10.1038/s41467-023-37570-1.

Nadeau, C. and Bengio, Y. (2003) 'Inference for the Generalization Error', *Machine Learning*, 52(3), pp. 239–281. Available at: https://doi.org/10.1023/A:1024068626366.

Nyffeler, J. *et al.* (2020) 'Bioactivity screening of environmental chemicals using imaging-based high-throughput phenotypic profiling', *Toxicology and Applied Pharmacology*, 389, p. 114876. Available at: https://doi.org/10.1016/j.taap.2019.114876.

Nyffeler, J. (2020) 'Phenotypic profiling for high-throughput chemical bioactivity screening at the U.S. EPA', p. 11457644 Bytes. Available at: https://doi.org/10.23645/EPACOMPTOX.13198346.

Park, C., Kim, H. and Wang, M. (2022) 'Investigation of finite-sample properties of robust location and scale estimators', *Communications in Statistics - Simulation and Computation*, 51(5), pp. 2619–2645. Available at: https://doi.org/10.1080/03610918.2019.1699114.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Prieto, P. (2019) 'Investigating cell type specific mechanisms contributing to acute oral toxicity', *ALTEX*, 36(1), pp. 39–64. Available at: https://doi.org/10.14573/altex.1805181.

Rogers, D. and Hahn, M. (2010) 'Extended-Connectivity Fingerprints', *Journal of Chemical Information and Modeling*, 50(5), pp. 742–754. Available at: https://doi.org/10.1021/ci100050t.

Sainburg, T., McInnes, L. and Gentner, T.Q. (2021) 'Parametric UMAP Embeddings for Representation and Semisupervised Learning', *Neural Computation*, 33(11), pp. 2881–2907.

Schneckener, S. *et al.* (2019) 'Prediction of Oral Bioavailability in Rats: Transferring Insights from in Vitro Correlations to (Deep) Machine Learning Models Using in Silico Model Outputs and Chemical Structure Parameters', *Journal of Chemical Information and Modeling*, 59(11), pp. 4893–4905. Available at: https://doi.org/10.1021/acs.jcim.9b00460.

Seal, S. *et al.* (2023) 'Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data', *Journal of Cheminformatics*, 15(1), p. 56. Available at: https://doi.org/10.1186/s13321-023-00723-x.

Serrano, E. *et al.* (2023) 'Reproducible image-based profiling with Pycytominer'. Available at: https://doi.org/10.48550/ARXIV.2311.13417.

Stirling, D.R. *et al.* (2021) 'CellProfiler 4: improvements in speed, utility and usability', *BMC Bioinformatics*, 22(1), p. 433. Available at: https://doi.org/10.1186/s12859-021-04344-9.

Way, G.P. *et al.* (2022) 'Morphology and gene expression profiling provide complementary information for mapping cell state', *Cell Systems*, 13(11), pp. 911-923.e9. Available at: https://doi.org/10.1016/j.cels.2022.10.001.