Project number: 963845

Project acronym: ONTOX

Project title: ontology-driven and artificial intelligence-based repeated dose toxicity testing of chemicals for next generation risk assessment

01/05/2021 – 30/04/2026

Deliverable number:	D3.3
Deliverable title:	Read-across tools based on the ontologies of the selected systemic repeated dose toxicity effects
Work package:	WP3
Leading partner:	IRFMN
Participating partners:	MN, PROTO
Due date:	30/04/2025
Submission date:	18/04/2025
Dissemination level:	public

Table of contents

1. Introduction and objectives

Within ONTOX, WP3 (chemical domain) is focused on the profiling of chemical properties for the selected project's chemicals with respect to relevant systemic effects.
The present deliverable describes the results achieved as part of Task 3.3, that focused in developing structure-activity relationship (SAR) models to predict molecular initiating events (MIEs) associated with adverse outcome pathways (AOPs) for liver, kidney, and developmental brain toxicities. Complementing the SAR models, a read-across strategy was developed that identifies chemical analogues with known experimental data and extrapolates their activity. Finally, to enhance the reliability of predictions, an integration workflow was designed to combine outputs of SAR and read-across with predictions provided by QSAR and molecular docking models (described in Deliverable 3.2) into a tiered decision framework.

2. Results

2.1. Research statement

The rationale behind the work described in the deliverable stems from the limitations of current animal-based toxicological assessments. These methods are costly, time-consuming, and often do not fully translate to human responses. In this regard, SAR models and the read-across strategy developed here provide, when integrated with other *in silico* models already described in the Deliverable 3.2., a more efficient and mechanistically informed alternative to animal tests, by integrating mechanistic knowledge from AOPs with data-driven methodologies.
While these tasks were planned, the integration strategy was not initially foreseen but was developed in response to feedback from the Scientific Advisory Board and the need to incorporate all computational tools into the project's unified assessment framework. This novel approach goes beyond the state of the art by combining QSAR, SAR, and read-across predictions using a weighted scoring system, improving accuracy and applicability of

predictions. In cases where this scoring approach yields inconclusive or conflicting results, molecular docking models are applied as a complementary tool to provide mechanistic insight and strengthen the final prediction.

2.2. Research and scientific evidence

2.2.1. Dataset compilation

The protein targets associated with the molecular initiating events (MIEs) leading to downstream toxicities were selected from AOP networks developed by WP7-9. These networks focus on cholestasis, steatosis, tubular necrosis (TN), cognitive functional defects (CFD), and neural tube closure (NTC) [1-5]. The inclusion of each target in the final modeling list was based on data availability and the balance between active and inactive chemicals.
Bioactivity data for each target was retrieved from the ChEMBL database (version 34) (https://www.ebi.ac.uk/chembl/) [6].
The dataset underwent curation at both the structural and endpoint levels, following the protocol outlined by Gadaleta et al. [7]. Specifically, only half-maximal response values (e.g., molar EC_{50}, AC_{50}, Ki, Kd, IC_{50}, potency, and ED_{50}) expressed on a negative logarithmic scale were considered. When available, data specific to *Homo sapiens* was prioritized. Binary classification was assigned to each data point based on information from the 'standard_relation,' 'standard_value,' and 'comment' fields in ChEMBL. Records were classified as "active" if the 'Standard Value' was below 10,000 nM and the 'Standard Relation' was '=' or '<'. Conversely, they were classified as "inactive" if the 'Standard Value' was greater than or equal to 10,000 nM and the 'Standard Relation' was '=', '>', or '>='. Additionally, the 'comment' field was reviewed for keywords indicating activity (e.g., "active," "inactive," "inhibitor," "not inhibitor"). Records with conflicting classifications were discarded.
This protocol was applied to each target, and the resulting datasets were then divided into a Training Set (TrS) and a Test Set (TeS) in an 80:20 ratio, ensuring that the original proportion of active and inactive compounds was maintained in both subsets. Table 1 provides an overview of the modeled proteins, the number of chemicals included in the TrS and TeS, and the ratio of active versus inactive chemicals.
The TrS was used to extract structural alerts (SAs), which are substructures commonly found in chemicals exhibiting a specific bioactivity classification. These SAs were formulated as SMARTS (SMiles ARbitrary Target Specification) strings (https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html). Acting as classification rules, SAs enable the assignment of activity labels to chemicals that match the corresponding structural pattern.

2.2.2. SAR models

Two different software programs were evaluated for extracting structural alerts from the datasets:

Table 1. List of MIE targets, the associated ChEMBL target ID, the size of the TrS and the TeS and the ratio between active and inactive records.

Adverse outcome	Protein Acronym	Protein full name	ChEMBL ID[1]	TrS	TeS	Ratio

Cholestasis	BSEP	Bile Salt Export Pump	6020	234	60	1:5
	MRP2	Multidrug Resistance-associated Protein 2	5748	62	17	1:4
	MRP4	Multidrug Resistance-associated Protein 4	1743128	98	26	1:2
Steatosis	AHR	Aryl Hydrocarbon Receptor	3201	218	56	4:1
	LXR	Liver X Receptor	3706564	36	10	11:1
	NRF2	Nuclear Factor Erythroid 2-related Factor 2	1075094	2882	721	31:1
	PPARα	Peroxisome Proliferator-Activated Receptor Alpha	239	2443	612	2:1
	PPARγ	Peroxisome Proliferator-Activated Receptor Gamma	235	3338	835	4:1
	PXR	Pregnane X Receptor	3401	617	156	1:1
TN	COX-1	Cyclooxygenase-1	221	1919	481	1:2
	OAT-1	Organic Anion Transporter 1	3539896	42	11	1:1
CFD	ACHE	Acetylcholinesterase	220	4353	1092	2:1
	NMDAR	N-Methyl-D-Aspartate Receptor	2094124	360	91	2:1
	THRα	Thyroid Hormone Receptor Alpha	1860	396	101	2:1
	THRß	Thyroid Hormone Receptor Beta	1947	1254	314	1:1
	TTR	Transthyretin	3194	184	47	1:1
	VGSC	Voltage-Gated Sodium Channels	1845, 4187, 5163, 5202	281	71	1:1
NTC	BMP	Bone Morphogenetic Proteins	3898, 1926496, 5350, 3286078	630	159	20:1
	CYP26	Cytochrome P450 Family 26	5141	141	36	8:1
	FGF	Fibroblast Growth Factors	4739699, 2362983, 3713913, 3286071, 2120, 3107	90	24	4:1
	HistDeac	Histone Deacetylases	2093865	1904	477	5:1
	WNT	Wingless-related Integration Site Proteins	1255132, 6079, 1255137	98	25	3:1
	FGFR-1	Fibroblast Growth Factor Receptor 1	3650	2818	705	3:1
	FGFR-2	Fibroblast Growth Factor Receptor 2	4142	1055	266	5:1
	FGFR-3	Fibroblast Growth Factor Receptor 3	2742	1490	374	7:1
	FGFR-4	Fibroblast Growth Factor Receptor 4	3973	1081	272	9:1
	SMO	Smoothened Homolog	5971	594	150	34:1

[1]*Some protein acronyms correspond to families or groups of proteins, which are associated with multiple ChEMBL IDs.*

1. **SARpy (SAR in Python, v1.0)** (https://sarpy.sourceforge.net/) is a software tool that identifies and collects SAs by fragmenting SMILES structures to generate all possible TrS fragments, which are then evaluated as potential SAs by comparing the activity of

TrS chemicals matching these fragments with their assigned activity flag. Ultimately, SARpy compiles a 'ruleset' containing SAs, their assigned activity flag (active/inactive), and an associated likelihood ratio (LR), a statistical parameter describing SA precision [8]. The settings were configured as follows:

- Structural Alert Options (SAO): Default settings were maintained (minimum occurrences = 3, minimum number of atoms = 2, maximum number of atoms = 18).
- Structural Alert Precision (CSAP): This parameter, which determines the lowest accepted SA precision, was varied between "automatic" (minimum or maximum) and "manual" (where the lowest accepted LR was iteratively increased from 2 to 4).

2. **MoSS (Molecular Substructure Miner)** (https://borgelt.net/moss.html) employs traversal search tree and basic search tree pruning to identify structural fragments frequently present in a dataset. The algorithm ensures that the proportion of chemicals matched by a SA belonging to one activity class (focus class) is significantly higher than that of the other class (support class). The settings used were minimum focus support: 10%; maximum complement support: 2%; generated fragment size: 2 to 15 atoms; ring size range: 3 to 8 atoms [9].

Different SAR models were generated for each dataset by varying the software settings, and their predictive performance was evaluated using the TeS. The optimal solutions for each target were determined based on the best balance between predictivity and coverage.

MoSS exhibited lower performance compared to SARpy and was only able to generate SAs for both activity classes in a limited number of targets, specifically NRF-2, BMP-1, FGFR-4, CYP-26, and SMO (Figure 1). As a result, MoSS models were discarded in favor of SARpy models, which produced rulesets capable of covering a broader range of targets with higher predictive accuracy.

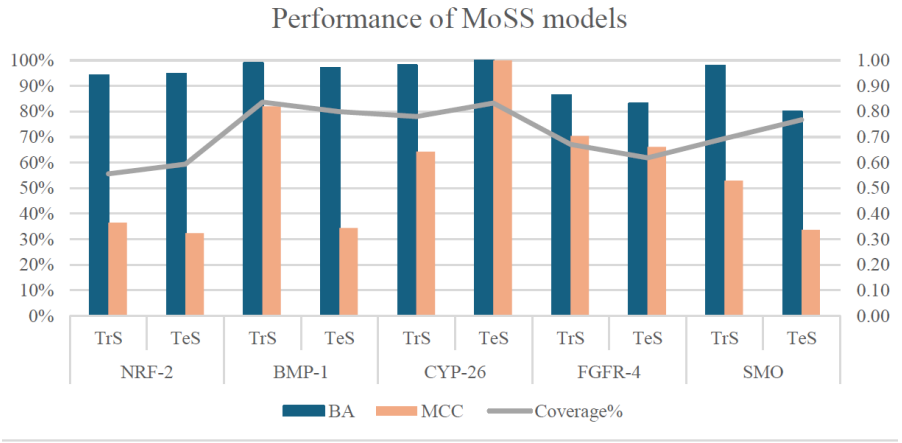

Figure 1. Validation performance of SAR models developed with MoSS. MCC is Matthew's Correlation Coefficient, BA is Balanced Accuracy

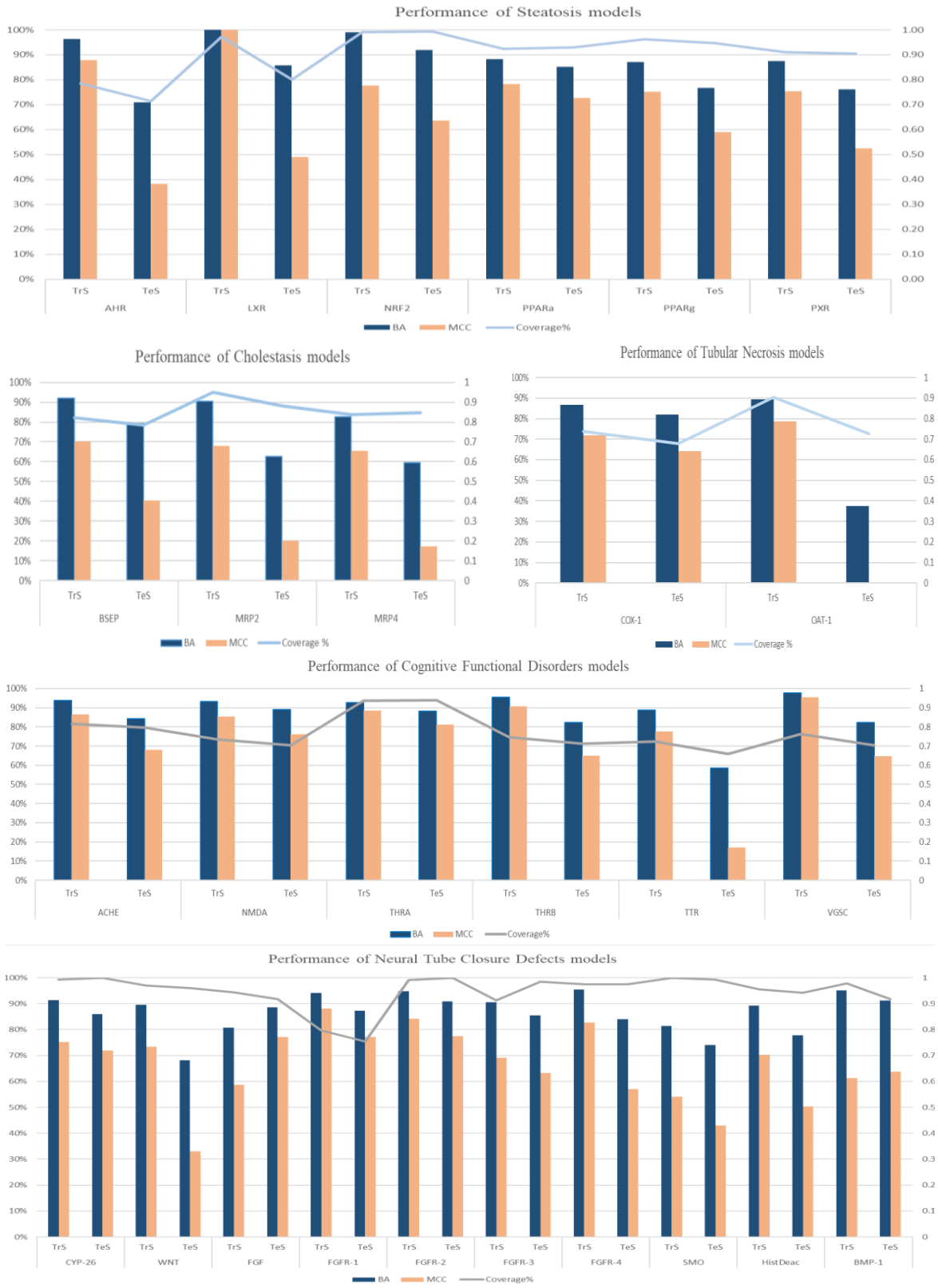

Figure 2. Validation performance of SAR models developed with SARpy. MCC is Matthew's Correlation Coefficient, BA is Balanced Accuracy

Figure 2 provides a detailed overview of the statistical performance of the SARpy models on the TeS. Notably, dataset size and precision settings played crucial roles in determining model

performance. Specifically, 22 out of 27 target models achieved balanced accuracy (BA) greater than 70%, with those based on larger datasets exceeding 85%. Setting CSAP to 'max' resulted in rulesets with high predictivity but unacceptably low coverage (i.e., the percentage of TeS chemicals covered by SAs), while setting CSAP to 'min' led to models with high coverage but lower predictive accuracy. In this context, the 'manual' CSAP setting was the most frequently selected among top-performing models, whereas the 'automatic' setting was preferred in only seven out of 27 models.

2.2.3. Read-across strategy

A read-across strategy was developed to support MIE predictions generated by computational models developed by WP3 by incorporating information from chemical analogues.

A set of 17 different molecular fingerprints (FPs) was calculated for a large reference dataset (ChEMBL v.34) [9]. Sample distributions of Tanimoto score values for each FP were determined by comparing random molecule pairs (Figure 3). Redundant FPs (Pearson correlation >0.80) were removed after verifying inter-FP correlation by comparing Tanimoto distributions. For the remaining FPs, Tanimoto thresholds defining the top 5% of each distribution were established.

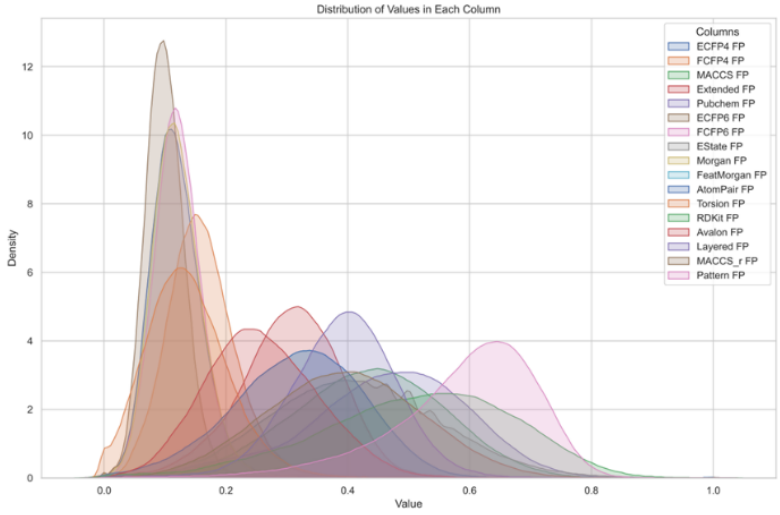

Figure 3. Example of distribution of TS calculated with different FP.

The presence of analogues with experimental values in ChEMBL datasets for each MIE target was verified by assessing chemical similarity relative to the queried compound. Analogues were pre-selected if their Tanimoto scores exceeded the predetermined top 5% thresholds for at least 80% of the FPs. The use of multiple FPs enabled a comprehensive assessment of structural similarity, mitigating biases inherent to single-FP evaluations. Pre-selected analogues were further filtered based on their similarity to the target compound in terms of lipophilicity ($\Delta \log K_o w \leq 1.0$) to account for physicochemical similarity in addition to the structural one.

The target was classified based on the most frequent activity among selected analogues, with a score expressed as the percentage of analogues exhibiting the assigned activity. To ensure reliability, at least two analogues meeting the above conditions were required for an acceptable assessment.

2.2.4. Integration workflow for computational predictions

A strategy was developed to integrate the diverse *in silico* prediction methods from WP3 into a unified test battery, ensuring an accurate and reliable assessment of chemical interactions with protein targets relevant to the project's MIEs. The workflow combines outputs from QSAR and docking methods (described in Deliverable 3.2) with SAR models and read-across approaches (detailed in this deliverable). The integration framework consists of three hierarchical tiers, with each tier considered only if the previous one fails to provide a definitive chemical classification regarding MIE binding (Figure 4).

- **Tier 1:** The ChEMBL database is queried for experimental data on the chemical's interaction with the target MIE-associated protein. If reliable and conclusive experimental data are available, no further predictions are necessary.
- **Tier 2:** A consensus approach combines QSAR, SAR, and read-across predictions using a weighted scoring system to derive a final classification for each chemical-protein pair. Each method contributes to the final weighted assessment only if its predictions are available and within its applicability domain. Weighting is assigned based on prediction reliability as follows:
    - **QSAR models:** Binary classification, with a score based on the probability of prediction. If multiple valid QSAR models exist for a single protein, individual predictions are scored and combined using a weighted average before integration with other methods.
    - **SAR models:** Binary classification, with a score proportional to model accuracy, expressed as the positive predictive value (PPV) or negative predictive value (NPV) of the relevant SA matching the predicted chemical.
    - **Read-across:** Binary classification based on the most frequent activity among valid selected analogues. The score reflects the percentage of analogues with the assigned activity.

  Predictions from all methods are aggregated into a Weighted Prediction Score (WPS) for each chemical-protein pair using the following formula:

  $$\textit{Weighted Prediction Score (WPS)} = (\sum wi \times \textit{positive prediction scores} - \sum wi \times \textit{negative prediction scores}) / \textit{number of predictions}$$

  Threshold values for WPS are used to classify chemicals as follows: binder (WPS ≥ 0.75), likely binder (0.75 > WPS ≥ 0.50), uncertain (0.50 > WPS > -0.50), likely non-binder (-0.50 ≥ WPS > -0.75), non-binder (WPS ≤ -0.75).

  Chemicals classified as 'uncertain', 'likely active' or 'likely non-active' require further evaluation in the next tier.

- **Tier 3:** Docking parameters, such as binding affinity and interaction fraction, are computed using the DockTox platform (https://chemopredictionsuite.com/DockTox). These values are compared with reference ligands to calculate a binding energy score (BES) and an interaction fraction score (IFS) as explained in Table 2. A docking score (DS) is then derived as the average of BES and IFS. Chemicals with a DS > 0.5 are classified as "possible binders" whereas those with a lower DS are classified as "possible non-binders". This tier provides an additional layer of refinement for cases where chemical classification remains ambiguous after the previous steps.

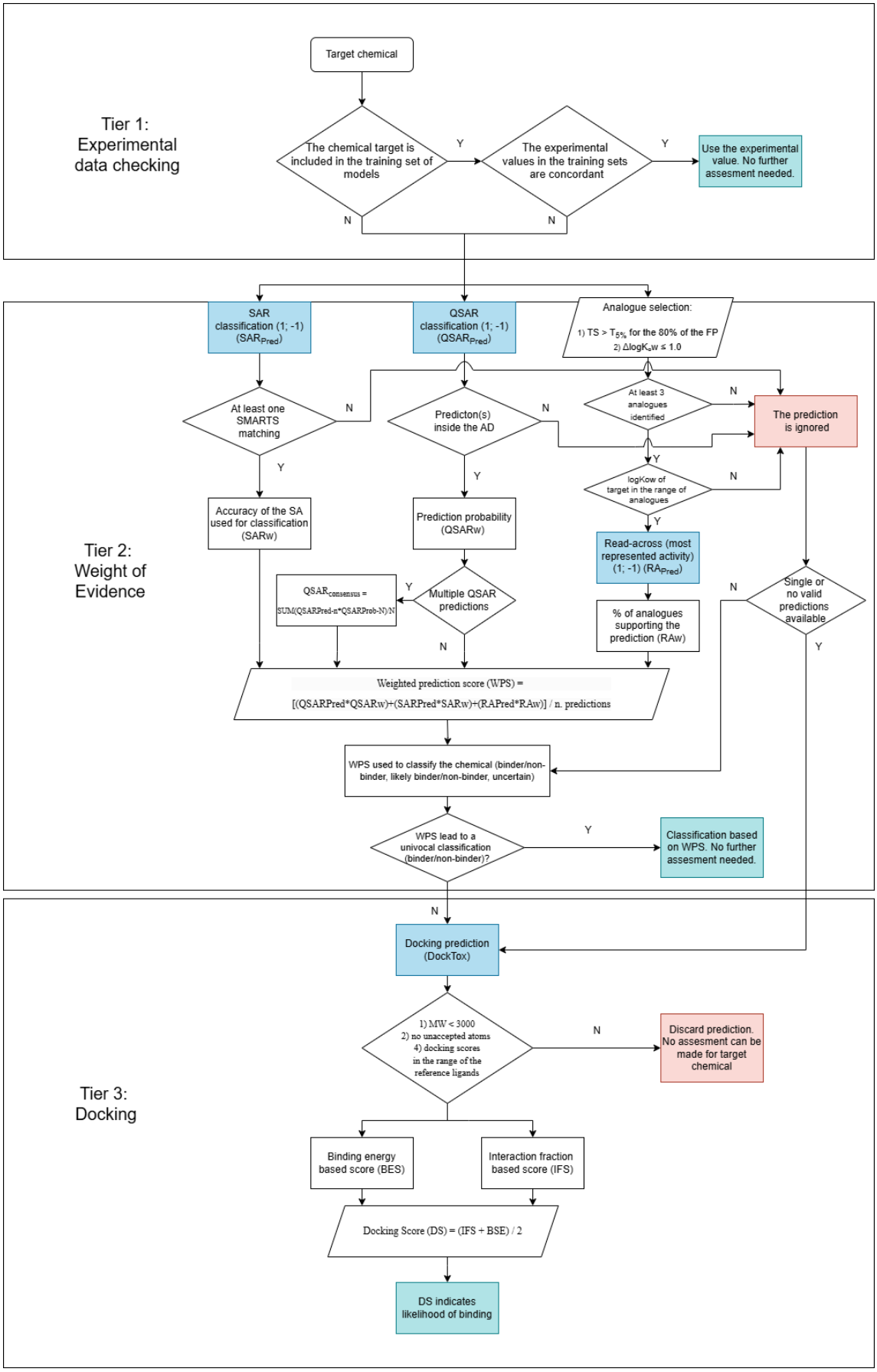

Figure 4: Workflow for the computational integrated strategy

Table 2. Calculation of the energy score (ES) and the Interaction score (IS) finalized at determining the overall docking score in the computational integration workflow.

Score	Parameter to check	Condition	Value
Energy Score (ES)	Binding Energy	Below the lower limit of reference ligands	1.00
		Within the range of reference ligands	0.50
		less than 3*st.dev above the higher limit for the reference ligands	0.25
		more than 3*st.dev above the higher limit for the reference ligands	0.00
Interaction Score (IS)	Interaction Fraction	above the average of reference ligands	1.00
		between the average and the lower limit of reference ligands	0.50
		below the lower limit of reference ligands	0.00

2.3. Impact

The results of this deliverable will be disseminated through scientific publications detailing both the SAR models and the integration strategy, ensuring broad visibility within the computational toxicology and regulatory communities. To maximize accessibility and usability, the computational models developed will be implemented in the VEGA QSAR platform (https://www.vegahub.eu/portfolio-item/vega-qsar/), allowing researchers and regulators to apply them in chemical risk assessment. Furthermore, all methods will be evaluated through case studies addressing the different project adversities. Ultimately, the entire framework will be incorporated as a dedicated module within the OPRA assessment framework, ensuring its integration into the broader strategy of the project and enhancing its impact on regulatory decision-making and next-generation risk assessment.

3. Conclusions and follow-up

This deliverable details the work performed in Task 3.3, which successfully developed SAR models and a read-across strategy for predicting chemical interactions with molecular targets associated with key MIEs related to liver, kidney, and developmental brain toxicities. These approaches were integrated with QSAR and docking models from Task 3.2 within a decision workflow designed to reinforce single predictions through weight of evidence.

Future directions include expanding the training dataset to enhance model generalizability and manually refining SAs based on expert judgment to reduce redundancy and improve mechanistic relevance. Additionally, next efforts will focus on exploring how computational predictions can be effectively integrated with other data sources generated within the project, such as *in vitro*, toxicokinetic, clinical, and epidemiological data, to develop a unified assessment framework for chemical risk evaluation.

4. Delays, issues and contingency

No major delays were encountered.

5. References

1. Li, J.; Settivari, R.; LeBaron, M. J.; Marty, M. S. An Industry Perspective: A Streamlined Screening Strategy Using Alternative Models for Chemical Assessment of Developmental Neurotoxicity. Neurotoxicology 2019, 73, 17–30.

2. Barnes, D. A.; Firman, J. W.; Belfield, S. J.; Cronin, M. T. D.; Vinken, M.; Janssen, M. J.; Masereeuw, R. Development of an Adverse Outcome Pathway Network for Nephrotoxicity. Arch. Toxicol. 2024, 1–14.
3. Van Ertvelde, J.; Verhoeven, A.; Maerten, A.; Cooreman, A.; dos Santos Rodrigues, B.; Sanz-Serrano, J.; Vinken, M. Optimization of an Adverse Outcome Pathway Network on Chemical-Induced Cholestasis Using an Artificial Intelligence-Assisted Data Collection and Confidence Level Quantification Approach. J. Biomed. Inform. 2023, 145, 104465.
4. Verhoeven, A.; van Ertvelde, J.; Boeckmans, J.; Gatzios, A.; Jover, R.; Lindeman, B.; Vanhaecke, T. A Quantitative Weight-of-Evidence Method for Confidence Assessment of Adverse Outcome Pathway Networks: A Case Study on Chemical-Induced Liver Steatosis. Toxicology 2024, 505, 153814.
5. Heusinkveld, H. J.; Staal, Y. C.; Baker, N. C.; Daston, G.; Knudsen, T. B.; Piersma, A. An Ontology for Developmental Processes and Toxicities of Neural Tube Closure. Reprod. Toxicol. 2021, 99, 160–167.
6. Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., ... & Leach, A. R. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic acids research, 52(D1), D1180-D1192.
7. Gadaleta, D.; Garcia de Lomana, M.; Serrano-Candelas, E.; Ortega-Vallbona, R.; Gozalbes, R.; Roncaglioni, A.; Benfenati, E. Quantitative Structure–Activity Relationships of Chemical Bioactivity Toward Proteins Associated with Molecular Initiating Events of Organ-Specific Toxicity. J. Chem. Inf. Model. 2024, 16 (1), 122.
8. Ferrari, T., Cattaneo, D., Gini, G., Golbamaki Bakhtyari, N., Manganaro, A., & Benfenati, E. (2013). Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR and QSAR in Environmental Research, 24(5), 365-383.
9. Yang, H., Li, J., Wu, Z., Li, W., Liu, G., & Tang, Y. (2017). Evaluation of different methods for identification of structural alerts using chemical ames mutagenicity data set as a benchmark. Chemical research in toxicology, 30(6), 1355-1364.