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:	D5.4
Deliverable title:	Report on Chemical Hazard and Potency Prediction by the AI System
Work package:	WP5
Leading partner:	JHSPH
Participating partners:	IRFMN, UM, MN, PROTO, BAYER, TT
Due date:	30/04/2025
Submission date:	28/04/2025
Dissemination level:	public

Table of contents

1. Introduction and objectives

This report outlines the application and integration of artificial intelligence (AI) methodologies within the ONTOX project framework at month 48 of the project, specifically targeting chemical hazard and potency predictions to facilitate next-generation risk assessment. ONTOX, an EU Horizon 2020 initiative (grant no. 963845), focuses on ontology-driven and AI-based toxicity testing, seeking to eliminate reliance on animal testing methods. The primary objective is to enhance accuracy, efficiency, and ethical standards in chemical risk assessments.

2. Results

2.1. Research statement

WP5 is among the most critical parts of ONTOX, as it envisages data integration as well as the establishment of the 6 ontologies and AI systems. WP5 generated the systematic review tools necessary for data mining, especially to be used in WP1-4. Furthermore, WP5 established a big data platform and perform data gap filling for integration of data collected in WP1-4. As a key task, WP5 developed the 6 ontologies for the selected systemic repeated dose toxicity effects. As described in this report, most importantly, machine learning, including deep learning, approaches were set up and applied to hazard prediction. This was achieved by relying on the RASAR approach, which was expanded by inclusion of physico-chemical, biological and toxicological properties. The unpublished model "ToxTransformer" described here has no known competition; the enormous interest by the regulatory community is evidence of its expected value. This report summarizes the progress according to plan.

2.2. Research and scientific evidence

1. Overall Approach and AI Integration

The ONTOX project employs a structured and multi-tiered development and validation approach within its innovative OPRA (ONTOX AI-supported probabilistic risk assessment) framework. This sophisticated approach integrates several advanced methodologies to predict chemical hazards and assess potency accurately. Central to the OPRA model are elements such as ontology-based structures, the Read-Across-Based Structure Activity Relationship (RASAR), probabilistic hazard assessments, probabilistic exposure assessments, and advanced artificial intelligence models including deep learning and large language models (LLMs). We started off by mapping current AI use in toxicology and the emerging opportunities (Hartung, 2023; Kleinstreuer and Hartung, 2024).

At its core, the ontology element serves as the foundational framework for standardizing and harmonizing terminologies, facilitating seamless integration and interpretation of complex toxicological data across diverse platforms and databases. Ontologies help capture relationships among chemical structures, biological activities, and toxicological endpoints, ensuring that information remains coherent, accessible, and actionable throughout the project's lifespan. They are used to illustrate biological perturbations leading to hazards, aka adverse outcomes.

RASAR methodologies (Luechtefeld *et al.*, 2018) within OPRA further enhance predictive capabilities by utilizing structural analogs of chemicals to infer potential toxicities. By leveraging historical and experimental data, RASAR aids in predicting toxicity outcomes for chemicals lacking extensive testing data, significantly improving risk assessments and reducing uncertainties typically associated with limited data scenarios. In its latest iteration, ONTOX has built a chemical property transformer model, tentatively named ToxTransformer.

A significant component of the OPRA model involves probabilistic hazard and exposure assessments; these concepts have been developed in a white paper and three workshops (Maertens *et al.*, 2022, 2024, 2025, in preparation). These assessments use sophisticated statistical and mathematical techniques to estimate the likelihood of adverse effects and levels of exposure within populations. Probabilistic methods provide a nuanced understanding of risk, accounting for variability and uncertainties inherent in real-world scenarios, thereby improving decision-making and regulatory outcomes. Notably, AI methods deliver the most probable result and based on pattern learning can express the probability of this suggested result.

The integration of advanced AI models represents a critical advancement within ONTOX. Notably, the project utilizes deep learning techniques, known for their superior ability to manage and interpret vast datasets, identifying intricate patterns and relationships beyond traditional analytical capabilities. To access the necessary Big Data, we developed a database import concept of BioBricks (Gao, submitted), which allowed to create a very large database Chemharmony, then analyzed by machine learning. These deep learning algorithms are complemented by large language models (LLMs), which enhance the interpretability and usability of complex datasets by translating technical data into comprehensible formats for stakeholders, including regulators, industry experts, and researchers (Corradi *et al.*, 2022).

We are currently exploring the possibility to leverage LLMs as information retrieval instruments for the large collection of physiological maps and combined AOP networks. This will facilitate question-answering for mode of action and collecting and aggregating information from the ontologies in a human-readable way (Hackathon in prep.)

A prime example of an AI-driven tool within the ONTOX framework is the ToxTransformer model that's MIT licensed. This advanced encoder-decoder transformer model excels in multi-task predictions across diverse chemical endpoints, providing comprehensive toxicological profiling of chemicals. By leveraging its powerful architecture, the ToxTransformer captures complex chemical interactions, achieving high predictive accuracy across multiple benchmarks. This AI model has shown substantial promise in facilitating regulatory toxicology decisions by significantly reducing reliance on traditional animal testing methods and offering more robust, rapid, and accurate hazard predictions (Gao *et al.*, in preparation).

Overall, the integration of ontology structures, RASAR methodologies, probabilistic assessments, and advanced AI models like deep learning and LLMs positions the OPRA framework as a pioneering tool in modern chemical risk assessment. By systematically addressing various challenges associated with traditional toxicological methods, OPRA not only enhances predictive accuracy but also ensures a more ethical and scientifically robust approach to chemical hazard assessment and potency prediction.

2. Probabilistic Risk Assessment within the ONTOX project framework

The ONTOX project's Probabilistic Risk Assessment (PRA) framework (Figure 1) represents a significant shift from traditional deterministic risk assessment approaches, responding to evolving regulatory needs and advancements in toxicological sciences. PRA embraces uncertainty and variability inherent to biological systems and chemical exposures, integrating these into sophisticated computational models for hazard and risk predictions. At its core, PRA within the ONTOX project leverages probabilistic methodologies to evaluate chemical risks more realistically. Unlike deterministic approaches that depend on fixed safety thresholds and conservative point estimates, PRA utilizes statistical models to estimate the likelihood of adverse outcomes, thereby offering a nuanced perspective on chemical safety. This approach accounts explicitly for biological variability, dose-response relationships, and uncertainties in diverse data sources, enabling more precise and individualized risk characterizations.

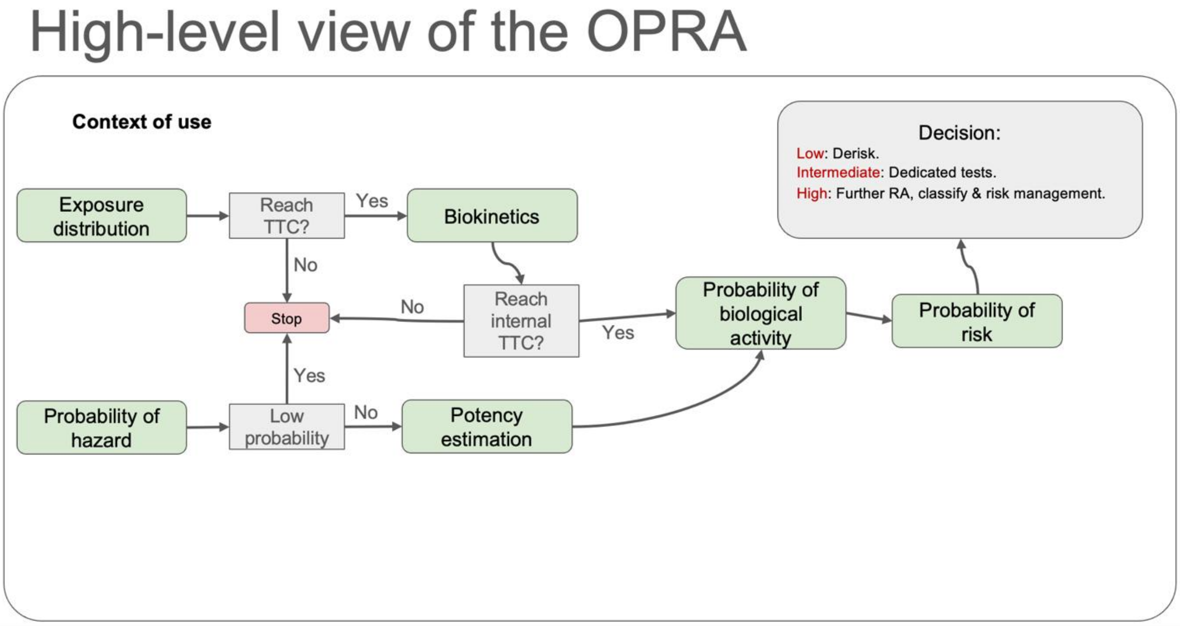


Figure 1. The probability of risk is ultimately driven by the probability of hazard, probability of exposure, and probability of biological effect. Each of these can be predicted or tested with different data streams, all of which should be considered for predicting risk.

The conceptual model underpinning PRA remains rooted in the fundamental risk assessment paradigm—risk as a function of hazard and exposure. However, it refines this relationship by quantifying both hazard and exposure as probabilities rather than binary or simplistic classifications. For instance, hazard identification is no longer merely categorizing a chemical as hazardous or non-hazardous; instead, it involves quantifying the potency and establishing confidence intervals using points-of-departure or benchmark doses. Similarly, exposure assessment is refined using sophisticated modeling techniques, such as physiologically-based pharmacokinetic (PBPK) models and probabilistic exposure models, capturing real-world variability and complexity more effectively.

A key strength of PRA lies in its ability to integrate complex and diverse data types, including high-throughput screening data, omics technologies, computational predictions, and *in vitro* test results. The integration is facilitated by advanced computational infrastructure and artificial intelligence (AI) tools, such as machine learning algorithms and natural language processing, to identify patterns, predict outcomes, and quantify uncertainties. These technological innovations allow PRA to not only handle but also leverage large-scale, multi-dimensional datasets, enhancing the reliability and accuracy of hazard predictions and regulatory decision-making.

One critical application of PRA in the ONTOX project is its focus on biological perturbations as the foundation for assessing hazards. This involves characterizing chemical interactions at the molecular level (molecular initiating events or MIEs), subsequent key events in biological pathways, and ultimately adverse outcomes. By adopting a quantitative Adverse Outcome Pathway (qAOP) approach, PRA models mathematically define internal concentrations triggering biological tipping points and assess the probabilities of these events occurring. This mechanistic grounding provides a robust basis for predicting chemical hazards and understanding individual susceptibility differences across populations.

Despite its advantages, PRA implementation within the ONTOX framework faces challenges. The complexity of probabilistic models requires extensive computational resources, specialized expertise, and robust software capable of handling intricate analyses and large data volumes. Additionally, validation remains a significant hurdle, with ongoing efforts needed to establish standardized good practices and benchmarks for comparison against probabilistic outcomes. Effective communication of probabilistic results to regulators, industry, and the public, accustomed to simpler deterministic conclusions, further complicates its widespread adoption.

Nevertheless, the ONTOX project's PRA methodology signifies a critical advancement in toxicological risk assessment. By incorporating probabilistic and AI-driven approaches, it aligns toxicology with modern scientific capabilities and regulatory demands, providing a comprehensive, nuanced, and realistic appraisal of chemical risks. Ultimately, this probabilistic approach offers enhanced transparency, efficiency, and accuracy, moving toward more informed and evidence-based regulatory decisions that better protect human health and the environment. By currently integrating the PRA approach into a software package with an user-interface as OPRA, a type of toxicology co-pilot, the approach becomes accessible for practitioners. The remaining time of the ONTOX project will be used to integrate OPRA in an agentic AI systems to expand and finetune its predictive capabilities and evaluate / pre-validate its functionality with domain experts in ONTOX for liver, kidney and the developing brain hazards.

3. The ONTOX BioBricks Approach

The BioBricks approach within the ONTOX project leverages an innovative data management and integration system designed to streamline and accelerate chemical and biological research through enhanced accessibility, reproducibility, and scalability of life sciences data. The BioBricks.ai platform, central to this methodology, addresses the significant challenges researchers face in biomedical and toxicological sciences, particularly regarding the integration and use of diverse, complex datasets from numerous disparate sources (examples in Table 1). Documentation for building, installing, and configuring BioBricks.ai is available at docs.biobricks.ai.

Table 1 - Examples of Databases in BioBricks.ai

Repository	Description
SMRT	Small Molecule Retention Time Dataset.
dictrank	Drug-Induced Cardiotoxicity Rank Dataset.
ice	Integrated Chemical Environment - High quality in vitro and in vivo toxicology data.
biogrid	Data from BioGRID.
ctgov	Data from ClinicalTrials.gov.
mirbase	Data from miRBase.
skinsensdb	Skin sensitization database.
ctdbase	Data from Comparative Toxicogenomics Database.
tox21	Tox21 quantitative high throughput screening (qHTS) 10K library data.
targetscan	Data from TargetScan.
USPTO_ChemReaction	Data from USPTO Chemical Reaction Database.
moleculenet	Molecular datasets for machine learning.
pubchem	PubChem data.
toxvaldb	Toxicity endpoint data.
dbgap	Genotype-phenotype interaction data.
zinc	ZINC purchasable compound database.
toxcast	EPA in vitro toxicity data.
pdb	Protein Data Bank 3D structure data.
geneontology	Gene Ontology knowledgebase.
cpdat	Consumer Product Data.
cpcat	Chemical Product Categories.
chembl	Bioactive molecule data.

BioBricks.ai functions akin to a "package manager for data," providing a standardized and efficient framework for researchers to manage dependencies on various scientific datasets. This system significantly reduces the redundant effort typically expended on data collection, curation, and cleaning, which traditionally consumes substantial research resources. By employing a centralized Data Version Control (DVC) registry, BioBricks.ai ensures consistent, reproducible access to up-to-date datasets, thereby enabling researchers to focus more on analysis and scientific innovation rather than data preparation (example figure 1).

```
import sqlite3, pandas as pd
import biobricks as bb

# Install and load the ToxRefDB asset
bb.install("toxrefdb")
assets = bb.assets("toxrefdb")
conn = sqlite3.connect(assets.toxrefdb_sqlite)

# Query to get LOAEL data
data = pd.read_sql_query("""
SELECT c.preferred_name AS chemical_name,
       g.name AS guideline_name,
       p.dose_level < p.max_dose_level AS adverse
FROM pod p
JOIN chemical c ON p.chemical_id = c.chemical_id
JOIN study s ON p.study_id = s.study_id
JOIN guideline g ON s.guideline_id = g.guideline_id
WHERE p.pod_type = 'loael'
""", conn)

# Count hazardous compounds for each guideline
data.groupby(['guideline_name','adverse']).size()
```

ToxRefDB Table Output

Guideline	NonHazard	Hazard
90-day Oral Toxicity in Rodents	4213	1680
Combined Chronic Toxicity/Carcinogenicity	3877	1228
Reproduction and Fertility Effects	2764	1480
Prenatal Developmental Toxicity Study	2680	1316
Chronic Toxicity	2423	1075
Carcinogenicty	2004	764

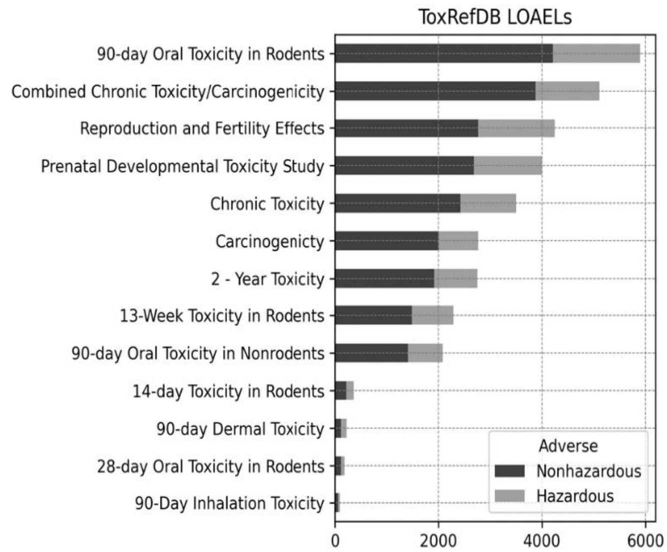


Figure 2. Example of BioBrick implementation allowing analysis of a database showcasing the number of studies that indicate hazardous status on

Top left - A code example to install, load, and analyze ToxRefDB data. Bottom left - the result of running the code example. Right - tabular data in bar chart form.

In practice, BioBricks organizes datasets into modular units known as 'bricks,' each represented as a git repository containing structured data, code, and metadata for easy installation, updating, and integration. These bricks can be primary data sources, such as genetic databases, chemical toxicology repositories, or regulatory guideline datasets. Alternatively, bricks can also be composite structures, combining multiple sources to create more complex, integrated datasets. A key example is ChemHarmony (Figure 3), a BioBrick specifically designed to integrate chemical safety data from numerous databases into a unified schema, thus simplifying complex heterogeneities across datasets and enhancing usability for downstream predictive modeling.

ChemHarmony Dataset: All model training and evaluation draws on the ChemHarmony integrated database of chemical activities (BioBricks.ai: A Package Manager for Public Health Data). Each data entry in ChemHarmony links a chemical substance to a property with a binary activity value (1 or 0) indicating presence/absence of that property or activity (GitHub - biobricks-ai/chemharmony: integrated chemical-property-values from many source databases.). For example, a property could be "acute oral toxicity" and an activity value of 1 means the compound is classified as acutely toxic in that context. For continuous endpoints (e.g., LD_{50} values or binding affinities), the data were binarized by thresholding (often using median or regulatory thresholds) to fit this framework. ChemHarmony consolidates data from numerous sources covering physico-chemical properties, bioactivity, toxicity, and regulatory classifications. Major sources include ChEMBL, PubChem, Tox21, ToxCast, eChemPortal, EPA CompTox, BindingDB, and more, collectively providing a broad chemical activity landscape. The current version of ChemHarmony is based on 12 BioBricks with 117 million chemicals, for which 254 million chemical activities are availability as triple chemical / property /result. Notably, ChemHarmony contains 4,026 distinct properties with at least 1,000 known compound activities each – a scale that enables training a single model on thousands of tasks.

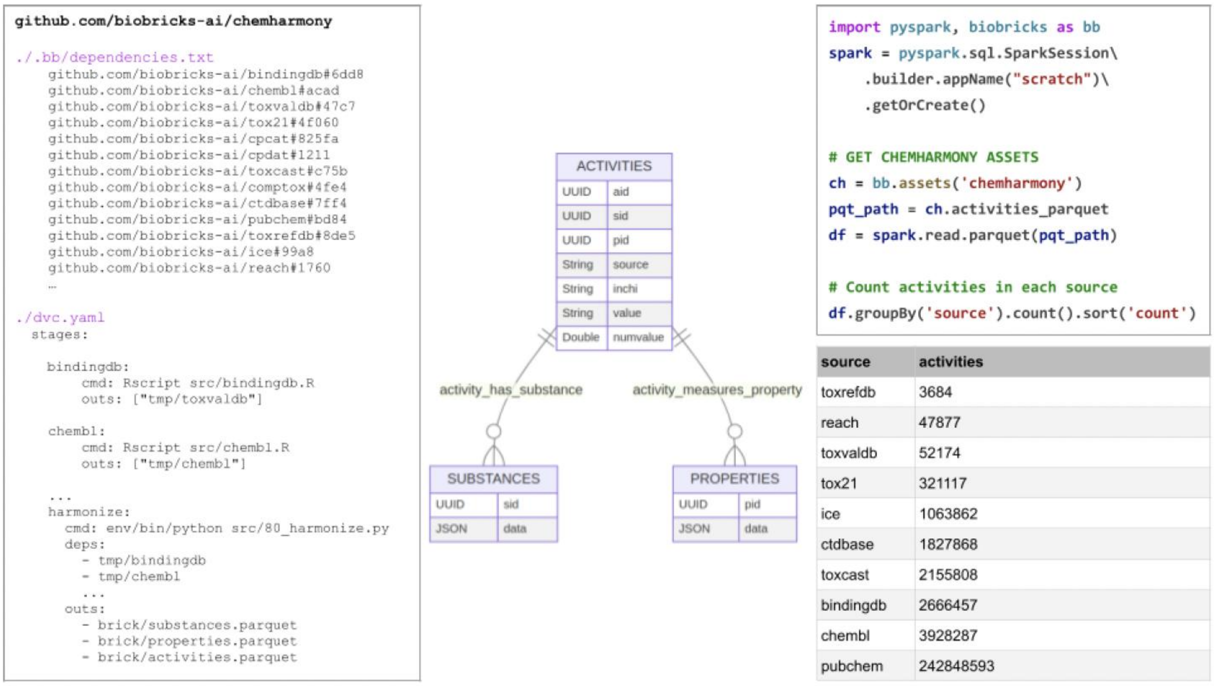


Figure 3 ChemHarmony created from 12 BioBricks

Left - truncated versions of the (1) .bb/dependencies.txt and (2) dvc.yaml file in the ChemHarmony BioBrick. Center, the 3-table schema of ChemHarmony, a simple chemical activities dataset with a substances, properties, and activities table. Right shows how to count activities by source by installing the ChemHarmony.

Technologically, BioBricks.ai employs robust data formats such as Parquet, SQLite, and HDT (Header, Dictionary, Triples). Parquet files optimize data storage and retrieval through effective compression and partitioning strategies, making them particularly suitable for handling large datasets efficiently. SQLite databases support sophisticated indexing and querying, enabling rapid data access and manipulation. HDT files manage semantic knowledge graphs, facilitating the storage and querying of complex relational data. Noteworthy, there are 140 BioBricks, many of them already available on Github.com and the others to follow the publication (Gao *et al.*, submitted). The approach has considerable advantages over traditional methods (Table 2).

Feature	BioBricks.ai	Federated Data Approaches (semantic web)	Harmonization Assets	Traditional Data Repositories
Data Integration	Supports multiple data types and sources	When standards are adopted	Within specific domains	Often siloed
Standardization	Standardized access, flexible data formats	Requires coordination on ontologies	Within asset scope	Varies widely
Ease of Use	Package manager-like system	Requires specialized knowledge	Depends on asset	Often requires manual navigation
Reproducibility	Version control built-in	Depends on implementation	Within asset scope	Often lacks versioning
Scalability	Designed for large datasets	Depends on infrastructure	Asset-dependent	Often limited by original design
Community Contribution	Open-source model	Depends on governance	Often centrally managed	Usually closed systems
Data Update Frequency	Can be real-time	Depends on participants	Often periodic releases	Varies widely
Interoperability	Common access method for diverse data	When standards are adopted	Within asset scope	Often requires custom integration
Learning Curve	New system, but designed for ease of use	Requires understanding of complex standards	Asset-specific knowledge needed	Varies - Often high for each new source
Cost Efficiency	Reduces redundant work	High initial investment	Reduces some redundancy	Often leads to redundant work
Flexibility for New Data Sources	Can easily add new 'bricks'	Requires adherence to existing standards	Often limited to predefined scope	Can add, but often in isolation
Support for AI/ML Applications	Designed with AI/ML needs in mind	Depends on data quality and format	Often designed for specific AI/ML tasks	Often requires significant preprocessing

Table 2. Comparison of BioBricks with other approaches

The ONTOX project specifically leverages BioBricks.ai to improve the development and validation of probabilistic risk assessment (PRA) models, which are fundamental to ONTOX's mission of predictive toxicology without animal testing. By integrating diverse datasets, including omics, chemical toxicity endpoints, mechanistic pathways, and regulatory data, BioBricks.ai significantly enhances the predictive capability and mechanistic understanding of chemical hazards.

Furthermore, BioBricks.ai supports advanced AI and machine learning applications central to the ONTOX project's goals. Its streamlined data infrastructure enables high-throughput data analysis, facilitating the rapid development and validation of sophisticated AI models such as ToxTransformer and RASAR (Read-Across Structure-Activity Relationships). These models benefit significantly from the harmonized and integrated datasets that BioBricks provides, ultimately improving their accuracy and reliability.

In conclusion, the BioBricks approach in ONTOX not only represents a significant advancement in data integration and management but also actively supports the broader goals of ethical and efficient chemical safety assessment. By significantly reducing redundancy and enhancing reproducibility in data management, BioBricks.ai positions ONTOX at the forefront of innovative toxicological research, paving the way for next-generation chemical risk assessments.

4. Chemical Property Transformer Approach

The Chemical Property Transformer, commonly referred to as ToxTransformer, represents a cutting-edge development in computational toxicology, meticulously designed to enhance the precision, efficiency, and comprehensiveness of chemical hazard predictions and property assessments. Its robust capabilities have wide-reaching implications across regulatory frameworks, academic research, and industry applications. ToxTransformer integrates state-of-the-art deep learning technologies, especially the powerful transformer-based neural network architectures, which have originally gained prominence through their remarkable success in natural language processing (NLP). These technologies have since been ingeniously adapted to the chemical informatics domain.

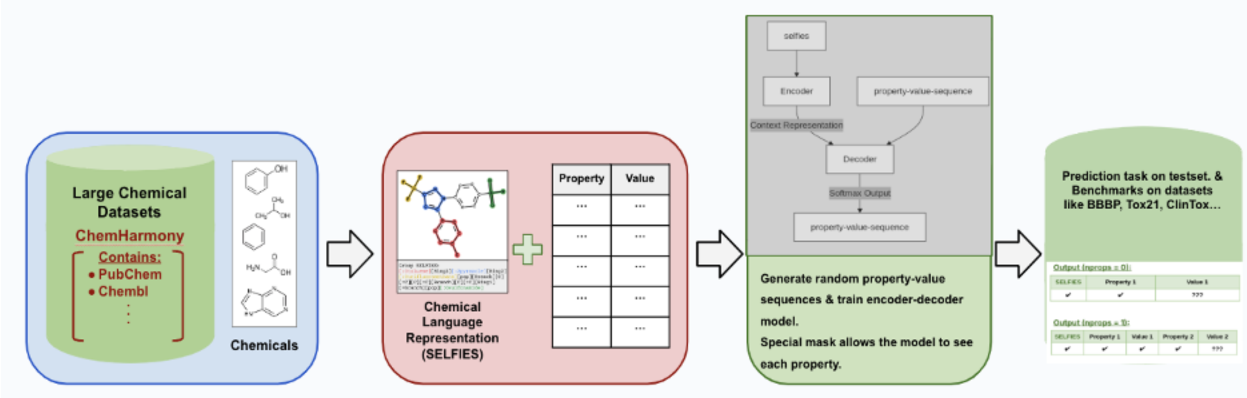


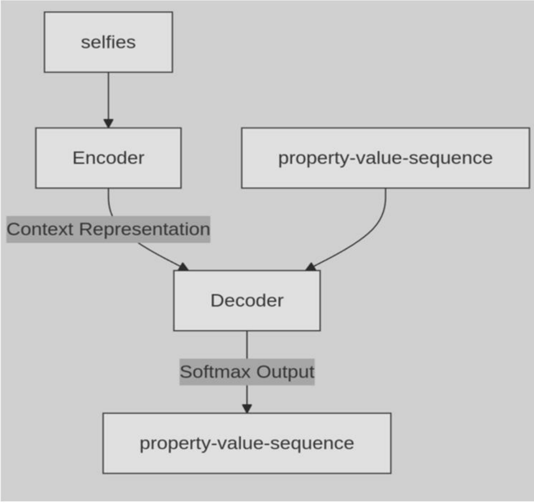
Figure 4. Overview of ToxTransformer pipeline

The transformer based model is trained on SELFIES sequences corresponding to a large collection of chemicals from a centralized database ChemHarmony, including multiple public databases, like PubChem, Chembl, Tox21, etc. The model is designed with mixture of experts system, learning not only the chemical structures, but also tokenized extracted chemical properties and values(binary values demonstrating toxic or not). We allow model to "see" different numbers of prior properties (as nprops, e.g. nprops =0 is no prior property is known and only predict using SELFIES, nprops =1 is given one random known property) to predict the target property value.

The underlying architecture of the ToxTransformer employs an encoder-decoder model that is foundational to transformer systems. This configuration enables the model to proficiently process, analyze, and interpret chemical data structures with great efficacy. The encoder aspect of this model captures and represents chemical information into structured data embeddings, which the decoder then utilizes to generate highly accurate predictions of chemical properties and potential toxicological outcomes. This transformative architecture has proven exceptionally adept at handling complex, structured, relational data that characterize chemical compounds.

Unlike traditional quantitative structure-activity relationship (QSAR) models, which typically depend on manually selected molecular descriptors, ToxTransformer relies on molecular SELFIES as input data for the graph convolutional networks. We take a SELFIES input, which is another way of encoding the chemical graph, but it's parsed as language tokens by the transformer. These molecular graphs offer a detailed representation of chemical compounds, encapsulating both atomic-level data and sophisticated relational information. This representation allows the model to process molecular structures in a way analogous to language models interpreting textual data—understanding chemical relationships and interactions in depth, similarly to how advanced language models grasp semantic and syntactic details within language. The model architecture is shown in Figure 5.

Model Architecture



Inputs:

SELFIES	Property 1	Value 1	Property 2	Value 2	...
[C][C][O]	LD50 injection	6.6g/kg	Fatal if swallowed	Inactive	...
[1, 32, 32,...]	2493	6798	1075	6799	...

Output (nprops = 0):

SELFIES	Property 1	Value 1
✓	✓	???

Output (nprops = 1):

SELFIES	Property 1	Value 1	Property 2	Value 2
✓	✓	✓	✓	???

Figure 5. ToxTransformer Architecture

The model employs a sequence-to-sequence transformer architecture comprising an encoder and a decoder. The encoder processes tokenized SELFIES representations of molecules, translating them into latent features via multiple layers of self-attention and feed-forward networks with positional encoding. The decoder, structured similarly, generates sequences of property and value tokens, using masked self-attention for causal sequence generation and cross-attention to integrate encoder-derived molecular information. A customized causal mask ensures correct dependencies between properties and their corresponding values. Additionally, the decoder incorporates Mixture-of-Experts (MoE) layers to efficiently handle numerous prediction tasks, allowing specialized routing of tasks to different experts.

The use of molecular graph-based representations equips the ToxTransformer with superior interpretative capabilities. By efficiently parsing structural motifs and relational networks, it can identify subtle molecular interactions and complex relationships influencing chemical behavior. This depth of interpretative insight significantly enhances its predictive power, providing highly accurate and nuanced forecasts about chemical toxicity and other relevant properties.

A defining strength of the ToxTransformer is its strategic deployment of multi-task learning methodologies. Multi-task learning involves concurrently training the model to predict multiple chemical endpoints, such as toxicity, bioavailability, environmental persistence, and bioaccumulation potential, among others. The synergy achieved through this approach facilitates enhanced learning efficiency and predictive accuracy across all tasks. By sharing common parameters and transferring knowledge between tasks, the model not only reduces computational demands but also achieves improved generalization, making the predictions more reliable and robust across diverse chemical properties.

Extensive benchmarking exercises consistently demonstrate the exceptional performance of ToxTransformer across diverse and complex chemical datasets. These benchmark studies reveal that ToxTransformer reliably outperforms traditional QSAR models and other conventional computational approaches, especially in critical toxicological endpoints that have profound regulatory implications. This high predictive accuracy is crucial in facilitating sound and evidence-based decision-making in regulatory toxicology, helping authorities make informed judgments regarding chemical safety and regulatory compliance.

Another significant advantage of the ToxTransformer lies in its integration of explainable AI (xAI) methodologies. Explainability is increasingly emphasized within regulatory frameworks, as stakeholders, including regulatory agencies, industry partners, and researchers, require transparency and understandability of predictions to build trust and effectively implement computational predictions in decision-making processes. The model's ability to deliver interpretable outcomes, where it clearly elucidates how particular predictions are derived, substantially enhances its utility and acceptance within regulatory and industrial environments.
In practical scenarios, ToxTransformer has shown substantial utility in ONTOX, where its predictive prowess significantly contributes to reducing the reliance on animal-based testing methodologies. By providing reliable, high-quality, and interpretable chemical predictions, ToxTransformer supports the broader scientific and ethical objectives of advancing alternative testing methods. This alignment with global trends toward more ethical, humane, and sustainable research practices underscores its relevance and applicability in contemporary chemical safety assessments.
The versatility of the ToxTransformer also extends to its adaptability in handling varied chemical data, making it well-suited for emerging and complex chemical safety evaluation contexts. Its design is flexible enough to accommodate continuous integration of new chemical data and evolving endpoints, ensuring its sustained relevance and utility in rapidly advancing scientific and regulatory landscapes.
Moreover, the computational efficiency and scalability of the ToxTransformer make it ideally suited for high-throughput screening and large-scale chemical assessments. Its capability to quickly process extensive chemical libraries and provide rapid, accurate predictions dramatically accelerates the process of chemical risk assessment, fostering quicker decision-making and significantly reducing research and development timelines.
In conclusion, the ToxTransformer is a groundbreaking tool in computational toxicology, leveraging advanced deep learning architectures and robust interpretability techniques to provide accurate, reliable, and comprehensive predictions. Its sophisticated approach, integrating molecular graph-based inputs, transformer architectures, and multi-task learning, positions it as an innovative and transformative solution for chemical safety evaluation. Ultimately, the ToxTransformer substantially enhances predictive accuracy, facilitates transparency in regulatory processes, reduces reliance on animal testing, and significantly accelerates the progress of chemical risk assessment methodologies.

5. From Physiological Maps to Hazard Assessment

Our article (Staumont *et al.*, 2025) introduces the concept of Physiological Maps (PMs) within the ONTOX project, significantly enhancing the development and refinement of alternative methodologies for chemical safety assessments. These maps serve as comprehensive graphical representations of biochemical processes and molecular interactions associated with specific organ functions, standardized using Systems Biology Graphical Notation (SBGN) for consistency and clarity. Physiological Maps aim to bridge a critical gap in the existing framework of New Approach Methodologies (NAMs), particularly addressing the need for detailed mechanistic insights at the gene and cellular levels. By complementing the widely utilized Adverse Outcome Pathways (AOPs), PMs provide an enriched layer of physiological detail, integrating pathology and normal physiological processes. This integration enables the identification and validation of new molecular initiating events and key biological events, thus filling essential gaps in current AOP models.
We outline the structured workflow employed to develop PMs, encompassing four main phases: planning, curation, updating, and application. This workflow highlights the necessity for close collaboration between domain experts and curators, leveraging data from extensive

scientific literature and authoritative databases such as Reactome, KEGG, and WikiPathways. Notably, PMs are continuously refined using advanced AI-driven tools for data extraction and integration, ensuring that the maps remain current and scientifically robust.

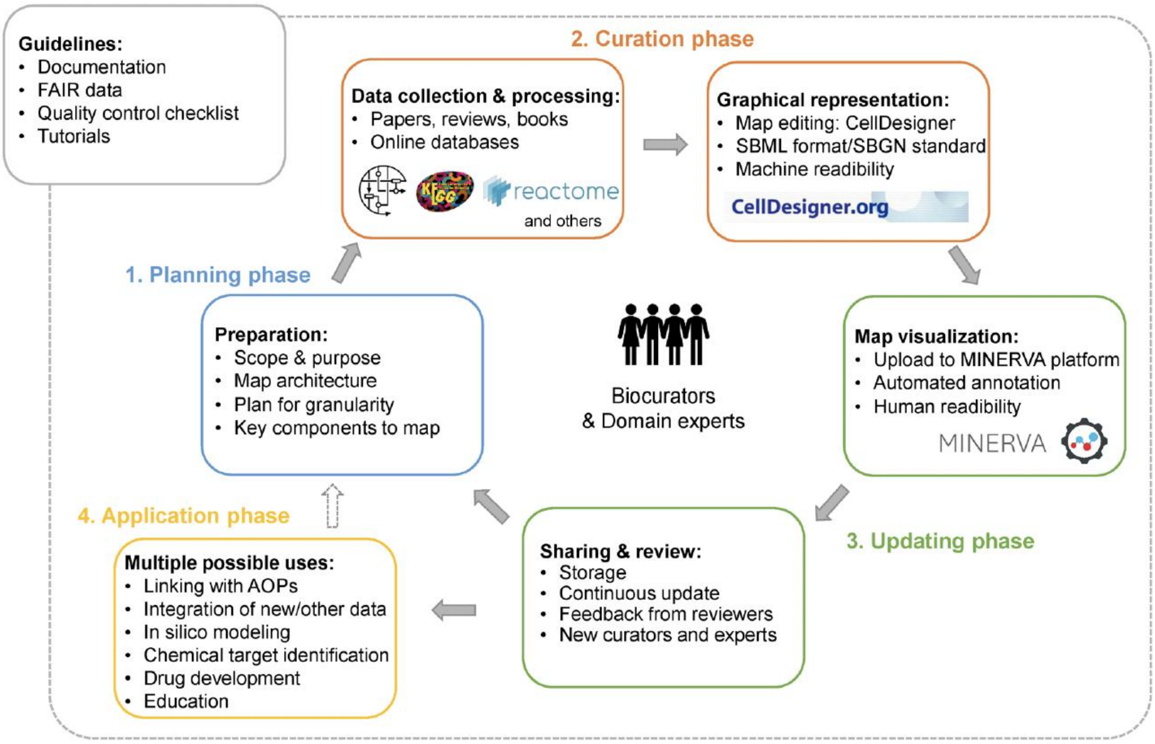


Figure 6. Workflow to construct ONTOX physiological maps

The workflow consists of four main phases: (1) planning, (2) curation, (3) update and (4) application.

PMs also offer practical utilities beyond their research-oriented applications. They provide user-friendly, standardized platforms for visualizing and analyzing omics data, which can significantly enhance chemical safety assessments and the development of targeted *in vitro* assays. Furthermore, these maps serve educational purposes, making complex biological processes accessible and understandable to broader audiences, thereby fostering interdisciplinary collaboration and knowledge dissemination.
We advocate for continued development and maintenance of PMs through collaborative community efforts, emphasizing the importance of ongoing contributions from the toxicology and systems biology communities. By promoting standardized, interoperable, and comprehensive data frameworks, the PMs developed within ONTOX contribute substantially to advancing human-relevant chemical safety assessments, aligning closely with regulatory demands and ethical considerations in toxicology.
Overall, the Physiological Maps initiative represents a substantial advancement in integrating detailed mechanistic biological understanding into predictive toxicology frameworks, thus significantly contributing to the overall goals and effectiveness of the ONTOX project.

6. Tiered Validation Approach

The ONTOX project has established a systematic and comprehensive validation strategy (consensus report as deliverable D6.3 recently submitted), structured meticulously into three primary tiers to ensure robust, reliable, and scientifically defensible outputs from the OPRA

(ONTOX AI-supported probabilistic risk assessment) framework. Each validation tier progressively builds upon the previous one, increasing in complexity, comprehensiveness, and regulatory relevance.

Notably, ONTOX participates in the global development of new validation approaches and standards for NAMS with a number of publications (Hartung, 2024; Hartung *et al.*, 2024a,b, Hartung and Kleinstruer, 2025; Hartung *et al.*, 2025).

Tier 1 (Module-level Validation):

Tier 1 validation concentrates on the foundational elements of the OPRA framework. This initial validation step is crucial for ensuring each individual module functions correctly and reliably before integration into the broader system. Key activities in this tier involve developing standardized operating protocols (SOPs), conducting internal validations by ONTOX project partners, and performing targeted benchmarking through specific case studies.

Protocol development and internal validation ensure methodological consistency and operational clarity across the diverse range of tools and predictive models employed by ONTOX. Each module is thoroughly tested internally to verify accuracy, reproducibility, and robustness in isolation. Benchmarking is particularly emphasized to assess the modules against well-characterized substances with known toxicological profiles. A notable example within Tier 1 involves the detailed examination of substances such as 3-monochloropropanediol (3-MCPD) and 2-monochloropropanediol (2-MCPD), which are industrially relevant due to their prevalence in food processing. Collaborations with major industrial stakeholders, such as Nestlé, enable practical testing scenarios and validation under realistic operational conditions, enhancing the reliability and industrial applicability of the outcomes.

Tier 2 (Probabilistic Risk Assessment Validation):

Building upon the robust module-level validations conducted in Tier 1, Tier 2 introduces probabilistic elements, significantly expanding the validation scope. At this stage, advanced computational and artificial intelligence-driven methods, including e-validation techniques, are employed to rigorously test and validate probabilistic risk assessment (PRA) methodologies.

The use of AI-enhanced approaches in Tier 2 involves sophisticated simulation models and AI-driven chemical selection processes, optimizing validation procedures by effectively capturing and managing uncertainties and variabilities in chemical hazard and exposure scenarios. These techniques enhance the realism and accuracy of validation outcomes, making the probabilistic risk assessments more reflective of real-world scenarios.

Tier 2 employs detailed case studies, carefully selected for their complexity and relevance to regulatory toxicology. A prominent example is the probabilistic risk assessment of perfluorooctanoic acid (PFOA), a chemical known for its environmental persistence and toxicological significance. This case study integrates hazard and exposure assessments and ensures the coherence and interoperability among different tools and modules developed within the ONTOX framework. Through such rigorous evaluations, Tier 2 validation aims to refine the PRA methodologies, ensuring they deliver coherent, reliable, and transparent predictions that can support regulatory decision-making processes.

Tier 3 (Full OPRA System Validation):

Tier 3 validation represents the culmination of the ONTOX validation strategy, involving a comprehensive evaluation of the fully integrated OPRA decision tree. At this stage, the entire framework—comprising ontology-driven hazard identification, probabilistic risk assessment, and potency estimation based on integrated *in vitro* and *in silico* predictions—is systematically validated. This tier ensures the seamless integration and functional coherence of all modules, providing robust evidence to support the reliability and regulatory acceptability of the full OPRA system.

This extensive validation process in Tier 3 encompasses specific toxicological endpoints related to organ toxicity, addressing critical regulatory concerns. The targeted toxicities include hepatic outcomes such as steatosis and cholestasis, renal outcomes like tubular necrosis and crystallopathy, and neurological endpoints such as neural tube closure defects and cognitive function impairments. Each of these endpoints involves complex interactions and pathophysiological mechanisms, requiring the integration of diverse datasets, computational models, and probabilistic assessments.

Tier 3 employs a variety of validation techniques, including sensitivity analyses, uncertainty quantification, and cross-validation against existing regulatory data. This comprehensive approach ensures not only the internal coherence of predictions but also their external validity and applicability to real-world regulatory contexts. Regulatory agencies and industry stakeholders are extensively involved at this stage to provide feedback, enhance the regulatory relevance of outcomes, and facilitate broader acceptance.

Through this structured tiered validation approach, the ONTOX project will deliver a scientifically rigorous, transparent, and robust OPRA framework capable of significantly advancing chemical risk assessments. By systematically addressing methodological accuracy, predictive coherence, and regulatory applicability, this validation strategy ensures that the OPRA system stands as a reliable and innovative tool for modern toxicological risk assessment, ultimately aiding in the transition towards more ethical, efficient, and animal-free testing paradigms.

2.3. Impact

The integration of artificial intelligence (AI), specifically advanced models like the ToxTransformer, significantly enhances the predictive accuracy and efficiency of chemical hazard and potency assessments. These transformer-based models capitalize on sophisticated deep learning architectures to provide precise, reliable predictions essential for modern toxicological evaluations. The ToxTransformer, in particular, represents a notable advancement, employing a mixture-of-experts transformer architecture that leverages extensive datasets to predict a wide array of chemical properties and toxicological endpoints.

One of the primary capabilities offered by models such as the ToxTransformer is their facilitation of high-throughput screening (HTS) for chemical toxicity. Traditional toxicity testing methods are labor-intensive, time-consuming, expensive, and ethically controversial due to their reliance on animal models. The high-throughput capacity of AI-driven models drastically reduces the dependency on animal experiments by rapidly evaluating extensive chemical libraries. For example, the ToxTransformer utilizes the comprehensive ChemHarmony dataset, integrating chemical properties from diverse sources such as PubChem, Tox21, and ChEMBL. This extensive integration allows for swift screening and prioritization of hazardous substances, significantly accelerating chemical risk evaluations and regulatory decision-making.

Another critical benefit of these advanced AI models is their ability to handle uncertainty through robust probabilistic outputs. Traditional deterministic approaches are often limited by their inability to fully capture biological variability and chemical interaction complexities. Conversely, AI-driven probabilistic predictions effectively encapsulate these uncertainties, offering a nuanced and realistic representation of risks. The ToxTransformer model demonstrates strong predictive accuracy, with benchmark performance showing high ROC AUC scores (0.70–0.94) across various chemical toxicity endpoints. Such probabilistic modeling thus significantly enhances regulatory confidence by providing more precise and realistic risk characterizations.

The mechanistic insights offered through explainable AI (XAI) methodologies further bolster the credibility and interpretability of AI-derived predictions. ToxTransformer facilitate transparent and interpretable predictions, which are critical for stakeholder acceptance, including regulatory bodies and industry partners. Explainable outputs elucidate the underlying rationale of predictions, bridging the gap between complex computational outputs and practical regulatory application. This transparency not only enhances trust but also deepens the scientific community's mechanistic understanding of chemical-biological interactions, contributing substantially to improved regulatory toxicology practices.

Additionally, the capability of ToxTransformer to condition predictions on known property outcomes represents a significant enhancement. This conditional inference capability markedly improves prediction accuracy when experimental data on some assays are available, offering substantial advantages in practical applications. Such an approach enables strategic testing, optimizing resource allocation by guiding further experimental assays based on initial computational predictions, thus reducing redundant or unnecessary tests.

In conclusion, the integrated capabilities of high-throughput screening, probabilistic risk characterization, mechanistic interpretability, and conditional inference provided by advanced transformer-based models such as ToxTransformer markedly enhance the accuracy, efficiency, and ethical responsibility of chemical risk assessments. These robust predictive tools position ONTOX at the forefront of modern toxicology, setting new benchmarks for chemical safety evaluation and regulatory toxicology, ultimately supporting the transition toward more humane, sustainable, and scientifically rigorous risk assessment methodologies.

Stakeholder Discussion

Documented in Diemar *et al.* (2024) we provide comprehensive insights from stakeholder engagement, particularly emphasizing the implementation of NAMs and probabilistic risk assessment (PRA). This report, derived from the first Stakeholder Network Meeting of the ONTOX project, identifies critical barriers, drivers, and specific challenges faced by stakeholders, including regulatory authorities, industries, academia, and non-governmental organizations. A major contribution of this paper is its detailed analysis of stakeholder perceptions regarding NAMs and PRA. It identifies key topics of agreement and contention, including capacity building, platform sustainability, regulatory acceptance, and AI acceptance. Importantly, the stakeholders emphasized the necessity of robust education and training programs to foster understanding and facilitate the integration of advanced methods like PRA into regulatory frameworks. These training initiatives would address current knowledge gaps among stakeholders, particularly regulators and Contract Research Organizations (CROs).

The stakeholders also underscored the significance of clear and harmonized validation processes to increase regulatory confidence and acceptance of NAMs. The discussions highlight the importance of developing targeted case studies, which demonstrate the practical applicability and benefits of these methodologies, thereby enhancing stakeholder confidence and facilitating regulatory uptake.

Moreover, Diemar *et al.* underline the strategic importance of transparent communication between industry, academia, and regulatory agencies to overcome barriers related to regulatory acceptance. They advocate for a gradual approach with incremental steps, ensuring smooth transitions from traditional practices to NAMs and PRA. Additionally, the report proposes organizing hackathons to creatively address persistent "wicked problems" that could not be immediately resolved during initial stakeholder meetings.

Finally, the paper provides a clear roadmap and actionable strategies aimed at overcoming identified barriers, thereby supporting the successful long-term implementation of NAMs and PRA methodologies within the ONTOX framework. This contribution aligns with the project's goals to enhance regulatory toxicology, promoting ethical, efficient, and scientifically rigorous chemical risk assessments.

3. Conclusions and follow-up

Within its operational timeframe, the ONTOX project has achieved substantial advancements, establishing robust proof-of-concept validations for its innovative AI-driven OPRA framework. These accomplishments underscore the significant potential of AI methodologies in revolutionizing chemical risk assessment paradigms. However, it is acknowledged that a full external validation, covering the complete spectrum of the OPRA system's capabilities, remains beyond the immediate scope of the current project duration. Nevertheless, the foundational work accomplished by ONTOX provides a solid platform upon which future advancements can be effectively built.

To ensure sustained progress and broader regulatory acceptance beyond the project's lifespan, it is essential to leverage AI-enhanced validation methods continually. Alignment with evolving international regulatory standards, notably those outlined by organizations such as the OECD, will be instrumental in securing long-term success and credibility of the OPRA framework. Adherence to these internationally recognized validation guidelines ensures that the methodologies developed by ONTOX remain compatible with regulatory expectations and standards, facilitating their integration into formal regulatory processes.

Future efforts to build upon the foundation established by ONTOX should include comprehensive expansions of external validations and rigorous benchmarking exercises. These validations should involve collaborative engagements with regulatory agencies, industry stakeholders, and independent research organizations. Such collaborative benchmarking will not only verify the reliability and accuracy of the OPRA framework but also significantly enhance its acceptance and integration into practical regulatory environments.

Another critical recommendation is the continuous integration and updating of AI tools and methodologies within the ONTOX Hub. The dynamic nature of AI technologies necessitates an ongoing commitment to refining models and computational approaches, ensuring they remain at the cutting edge of scientific and technological developments. Regular updates, incorporating new datasets, improved computational methods, and advanced AI techniques, will maintain the predictive accuracy, reliability, and relevance of the ONTOX methodologies in rapidly evolving toxicological and regulatory landscapes.

Finally, addressing the current limitations inherent in AI technologies, such as data biases, interpretability challenges, and computational scalability, will require strengthened interdisciplinary collaborations. Collaborative engagements between toxicologists, data scientists, regulatory experts, and ethical specialists can systematically address these challenges, fostering the development of more balanced, transparent, and widely acceptable AI solutions. By actively involving diverse expertise, the ONTOX project can comprehensively tackle the complexities associated with chemical hazard predictions, thereby advancing AI-driven methodologies toward greater robustness, transparency, and regulatory applicability.

In summary, the structured approach adopted by ONTOX positions the project as a pioneering force in AI-supported toxicology, providing a scientifically robust, ethically responsible, and technologically innovative framework for chemical safety assessment. Continued adherence to international validation standards, strategic external validations, ongoing technological updates, and proactive interdisciplinary collaborations will ensure that ONTOX's achievements significantly contribute to long-term advancements in toxicological sciences and regulatory practices.

4. Delays, issues and contingency

This project has seen no delays but enormous acceleration beyond anything imaginable at the time of proposing or the start of ONTOX. Taking full advantage of the general progress in AI, all aspects of work are state of the art. With AI doubling its capabilities every three months, this is a continuing challenge. The dimension of the database with 250+ million entries for 120+ million chemicals is second to none. Predicting more than 4,000 properties with promising accuracies now represents an abundance problem to integrate this into manageable information. With the user interface now available, the planned pre-validation for the domains of expertise in the consortium (liver, kidney and the developing brain) can now proceed. Integration of QSAR, scientific literature, internet information and perturbation of biology as additional input is ongoing. Most noteworthy, the project has been requested to provide briefings by EFSA, OECD test guideline program, FDA and others, which is most promising for its contingency. While the consortium has refrained from claiming IPR in this fast-moving space, a commercial version and services is envisaged by partner ToxTrack and possibly other partners.

5. References

Corradi, M. P. F., de Haan, A. M., Staumont, B., Piersma, A. H., Geris, L., Pieters, R. H. H., Krul, C. A. M., & Teunis, M. A. T. (2022). Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies. Biomaterials and biosystems, 7, 100061. https://doi.org/10.1016/j.bbiosy.2022.100061

Gao Y, Mughal Z, Jaramillo-Villegas JA, Corradi M, Borrel A, Lieberman B, Sharif S, Shaffer J, Fecho K, Chatrath A, Maertens A, Teunis MAT, Kleinstreuer N, Hartung T, and Luechtefeld T. BioBricks.ai: A Versioned Data Registry for Life Sciences Data Assets. https://arxiv.org/abs/2408.17320

Gao et al. ToxTransformer: Advancing Multi-task Chemical Property Prediction for Animal-Free Toxicology. In preparation.

Diemar M, Vinken M, Teunis M, Krul C, Busquet F, Zajac J, Kandarova H, Corvi R, Russo M, Kharina A, Bryndum L, Santillo M, Bloch D, Kucheryavenko O, Panagiotakos D, Rogiers V, Beekhuijzen M, Giusti A, Najjar A, Courage C, Koenig T, Kolle S, Boonen H, Dhalluin S, Boberg J, Müller B, Kukic P, Ritskes-Hoitinga M, Grasselli E, Zietek T, Stoddart G, Heusinkveld H, Castell J, Benfenati E, Yang H, Perera S, Paini A, Kramer N, Hartung T, Janssen M, Fritsche E, Jennen D, Piumatti M, Rathman J, Marusczyk J, Milec L, and Roggen E. Report of the first ONTOX Stakeholder Network meeting: digging under the surface of ONTOX together with the stakeholders. ATLA 2024, 52:117-131. Doi: 10.1177/02611929231225730

Hartung T. ToxAIcology - the Evolving Role of Artificial Intelligence in Advancing Toxicology and Modernizing Regulatory Science. ALTEX 2023, 40: 559-570. doi: 10.14573/altex.2309191.

Hartung T, The validation of regulatory test methods – conceptual, ethical, and philosophical foundations. ALTEX 2024, 41:525–544. doi: 10.14573/altex.2409271.

Hartung T, King N, Kleinstreuer N, Leist M and Tagle D. Leveraging Biomarkers and Translational Medicine for Preclinical Safety - Lessons for Advancing the Validation of Alternatives to Animal Testing. ALTEX 2024, 41: 545-566. doi:10.14573/altex.2410011.

Hartung T, Maertens A and Luechtefeld T. E-validation – Unleashing AI for Validation. ALTEX 2024, 41:567–587. doi: 10.14573/altex.2409211.

Hartung T and Kleinstreuer N. Challenges and Opportunities for Validation of AI-based New Approach Methods. ALTEX 2025, 42:3-21. doi: 10.14573/altex.2412291

Hartung T, Whelan, Tong W, and Califf RM. Is Regulatory Science Ready for Artificial Intelligence? NPJ Digital Medicine 2025, 8:200. https://doi.org/10.1038/s41746-025-01596-0

Hartung T, Hoffmann S and Whaley P. Assessing Risk of Bias in Toxicological Studies in the era of Artificial Intelligence, Archives in Toxicology 2025, in press

Kleinstreuer N and Hartung T. Artificial Intelligence (AI) – it's the end of the tox as we know it (and I feel fine) - AI for Predictive Toxicology. Archives of Toxicology 2024, 98:735–754. Doi: 10.1007/s00204-023-03666-2

Maertens A, Antignac E, Benfenati E, Bloch D, Fritsche E, Hoffmann S, Jaworska J, Loizou G, McNally K, Piechota P, Roggen EL, Teunis M and Hartung T. The probable future of toxicology - probabilistic risk assessment. ALTEX 2024, 41:273–281. doi: 10.14573/altex.2310301.

Maertens A, Kincaid B, Bridgeford E, Brochot C, de Carvalho e Silva A, Dorne J-LCM, Geris L, Husøy T, Kleinstreuer N, Ladeira L, Loizou G, Mcnally K, Middleton A, Reynolds J, Rodriguez B, Roggen EL, Russo G, Thayer K, and Hartung T. From Cellular Perturbation to Probabilistic Risk Assessments. revised.

Staumont B, Ladeira L, Gamba A, Heusinkveld HJ, Piersma A, Fritsche E, Masereeuw R, Vanhaecke T, Teunis M, Luechtefeld TH, Hartung T, Jover R, Vinken M and Geris L. Mapping physiology: a systems biology approach for the development of alternative methods in toxicology. Submitted. ALTEX 2025, 42:301-307. doi:10.14573/altex.2412241