PregPred: an *In-Silico* Alternative to Animal Testing for Predicting Developmental Toxicity Potential.

Ricardo Scheufen Tieghi[#a], Marielle Rath[#a], José Teófilo Moreira-Filho[#b], James Wellnitz[a], Holli-Joi Martin[a], Kathleen Gates[c], Helena T. Hogberg-Durdock[b], Nicole Kleinstreuer*[b], Alexander Tropsha*[a], Eugene N. Muratov*[a].

[#] These authors contributed equally.

[a] *UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA.*

[b] *National Toxicology Program Interagency Center for Evaluation of Alternative Toxicological Methods (NICEATM), Research Triangle Park, NC, 27711, USA.*

[c] *Department of Psychology & Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA*

*To whom correspondence and materials requests should be addressed: 100K Beard Hall, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA; Telephone: (919) 966-2955; FAX: (919) 966-0204; E-mail: nicole.kleinsreuer@nih.gov, alex_tropsha@unc.edu, murik@email.unc.edu,

Competing Interests

AT and ENM are co-founders of Predictive, LLC, which develops novel alternative methodologies and software for toxicity prediction. All the other authors declare no conflicts.

ABSTRACT

Background: Understanding potential prenatal and development toxicity hazard associated with the use of pharmaceutical and cosmetic products is an important component of women health. This hazard can be estimated from chemical structure of respective agents using Quantitative Structure-Activity Relationship (QSAR) models; however, the development of reliable models is challenging due to the complex nature of this endpoint.

Methods: Aggregating and curating data from the Food and Drug Administration (FDA), Teratogen Information System (TERIS) database, and select independent studies, we have created, to the best of our knowledge, the largest publicly available dataset comprising compounds annotated as developmental toxicants or not toxicants.

Results: We built several binary classification QSAR models exhibiting a correct classification rate of 62-72%, a sensitivity of 66-75%, a specificity of 59-82%, and high coverage of 70-90% assessed using five-fold external validation protocol. We developed a publicly accessible web portal PregPred for developmental toxicity prediction of both overall toxicity and trimester-specific toxicity predictions.

Conclusions: Due to high accuracy and coverage as well as public accessibility of the respective web portal, our models can be employed as a computational tool to support regulatory assessment of pharmaceutical and cosmetic products in alignment with the 3Rs (refining, reducing, and replacing) of animal testing. This *in silico* model holds the potential to substantially influence the field of developmental toxicology, steering regulatory practices toward safer drug development for pregnant women. The first-of-its-kind curated dataset of developmental toxicants and all developed models implemented as a user-friendly web tool, PregPred, are freely available at https://pregpred.mml.unc.edu/).

Introduction

In high-income countries, approximately 80% of pregnant women take prescription medications, and an even higher proportion take over-the-counter medications.[1] Despite the high usage of medication among pregnant women, drug safety in this population is grossly understudied. For example, clinical trials rarely include pregnant women, leaving knowledge gaps as to the potential toxicity to the fetus and pharmacokinetic differences between non-pregnant and pregnant women.[2] Additionally, there remains a lack of sizeable epidemiological cohort studies following children who were exposed to medications *in utero*.[3] Therefore, pregnant women and their physicians are often faced with the difficult decision of whether to continue the medication throughout pregnancy based on limited data. Further investigation into medication safety for pregnant women and the developing fetus is thus warranted.

Developmental toxicity is broadly defined as the potential for a compound, whether it be a medication, environmental, cosmetic, or other chemical, to cause abnormalities in the developing fetus upon a mother's exposure.[4] Manifestations of developmental toxicity include, but are not limited to, spontaneous abortions, physical abnormalities of organs and bones, low fetal birthweights, jaundice, deafness, and intellectual deficits.[5,6] Given the broad nature of this endpoint, there are many possible implicated mechanisms. Developmental toxicants can interfere with cells in various ways, including inappropriately upregulating or downregulating molecular pathways, binding to DNA and proteins, and oxidatively damaging macromolecules. These changes may be widespread, occurring in many parts of the developing fetus, or localized by damaging specific organs or the neural tube – the latter contributed to the developmental toxicity of thalidomide, the most well-known human teratogen.[7]

The 1960s thalidomide tragedy inspired the incorporation of pregnancy information into drug labels and the implementation of regulatory developmental toxicity guidelines.[8] In brief, animal studies for developmental toxicity typically involve administering the chemical to the pregnant animal through the intended route of administration, removing the fetus, and evaluating structural abnormalities.[9] These are described in further detail in the Organization for Economic Co-operation and Development (OECD) guidelines 414 for developmental toxicity testing and 426 for developmental neurotoxicity, as well as in the Environmental Protection Agency (EPA) guidelines OPPTS 870.3700 for prenatal toxicity and 870.6300 for developmental toxicity.[10,11] Important considerations include the relevance of the species used, the stage of pregnancy during which the chemical is administered, and the dosage.

Adverse Outcome Pathways (AOPs) are a series of linked events at different levels of biological organization (e.g., cell, tissue, organ) that lead to an adverse health effect in an organism following exposure to a stressor.[12] Various AOPs exist for developmental toxicity; however, zebrafish, mice, and rabbit models are currently the best *in vivo* models for screening.[13–16] It is also worthwhile to consider Clinical Outcome Pathways (COPs); while conceptually similar to AOPs, COPs consider a series of critical molecular and cellular events that underlie the therapeutic effects of drug molecules.[17]

Currently, animal studies are the major way to determine developmental toxicity experimentally; however, they are associated with several pitfalls. Thus, animal toxicity studies are time-consuming and expensive. In 2018, the EPA reported that a single developmental toxicity rodent study costs around $128,000, and developmental neurotoxicity study costs around $750,000 per compound.[18] They also require that many animals be used, raising ethical concerns. For example, both EPA and OECD guidelines require that at least 20 animals be used in the control

and test-dose groups. Another potential issue is that the indicated disease state is often not induced in the maternal animal model. Consequently, a medication that may restore homeostasis in an indicated individual may disrupt functioning in a healthy individual. This can result in false positives, wherein medications that would not induce developmental toxicity *in utero* of a mother with the indicated condition may induce toxicity in a healthy mother's fetus.[19] Another downside is that these studies do not track neurobehavioral endpoints over time, an essential manifestation of developmental neurotoxicity.[20] Despite these limitations, animal studies are still a valuable data source in toxicology research. Luckily, this data has become more widely accessible.[21]

Over the past several decades, expanding publicly available biomedical databases has supported the development of computational models for developmental toxicity, furthering the 3Rs in toxicology: reducing, refining, and replacing animal testing.[22] One powerful *in silico* approach is quantitative structure-activity relationship (QSAR) modeling. Several groups have argued against the use of QSAR modeling due to its lack of mechanistic insight; however, we have previously reported QSAR models with high externally validated accuracy for complex endpoints such as skin sensitization, cardiotoxicity, and pharmacokinetics.[23–25] QSAR modeling entails the collection of chemicals and their respective biological activity for the desired endpoint; these data may come from animal studies, *in vitro* studies, or, more favorably, human studies such as clinical trials or epidemiological cohort studies. The next critical step is to perform biological and chemical curation of the dataset; we emphasize that this is critical to developing reliable QSAR models but, unfortunately, is neglected in many studies.[26]

Developing reliable computational models to predict developmental toxicity remains a significant challenge.[29,30] Herein, we have collected developmental toxicity data from large publicly available databases, such as DailyMed (includes Food and Drug Administration (FDA)

labels) and Teratogen Information System (TERIS), as well as from smaller toxicity studies.[31,32] We meticulously validated all the compounds classified as nontoxic by searching the literature for regulatory-approved animal studies and epidemiological cohort studies, and removed compounds with ambiguous toxicity results. Other groups have modeled developmental toxicity but did not rigorously validate the negative compounds as we have.[33,34] Compounds were also validated for overall toxicity during pregnancy and individual risk per trimester. In addition to this extensive biological curation, we followed the best practices in the field for chemical data curation, which we have previously shown is necessary to develop reliable and accurate QSAR models.[35–37] Finally, we used this curated data to build and rigorously validate a QSAR model for developmental toxicity and enabled its use via a user-friendly web tool (PregPred), publicly available to the research community at https://pregpred.mml.unc.edu/.

Methods

Data collection and curation

We collected chemical data from human and animal developmental toxicity studies, including records from the FDA, TERIS, and independent studies. Drug, environmental, and cosmetic chemicals were included. We focused on acquiring the most credible literature, using widely recognized and cited sources. As we have shown previously, it is generally acceptable to include drug, cosmetic, and environmental compounds in the same dataset for QSAR modeling, especially when employing the applicability domain, given the overlap in chemical space.[38] Additionally, while human data is preferred over animal data, they are often scarce for toxicity endpoints. Therefore, we have also included results from animal studies. Since many chemicals lacked standard identifiers (International Chemical Identifier (InChI), Chemical Abstracts Service

Registry Number (CASRN), or Simplified molecular-input line-entry system (SMILES)), having instead only the compound name, we retrieved SMILES for each molecule in KNIME using the "Chemical Identifier Resolver" node.[39] All outcomes were binary (*i.e.*, toxic or nontoxic). Datasets were biologically and chemically curated according to the best practices in the field.[35–37] As for chemical curation, we removed mixtures, inorganics, and large organic compounds, removed counterions, cleaned and neutralized salts, and normalized chemotypes using the ChemAxon Standardizer software.[40] We followed one of the two appropriate procedures for handling duplicates: (i) if the outcomes of all duplicates were concordant, one record was kept with the respective outcome; (ii) if any outcomes disagreed, they were further investigated, as described in the Section "Merging the datasets and verification of nontoxic compounds**."** We developed four binary QSAR models to predict developmental toxicity risk in (i) any trimester, (ii) first trimester, (iii) second trimester, and (iv) third trimester. Dataset-specific biological curation is described in the following sections.

FDA database

In 1979, the FDA implemented a policy that pregnancy safety information be included on FDA-approved drug labels, if available.[41] Drugs may be classified into one of the five following classes: A, B, C, D, and X (see Table 1). In 2015, the FDA began shifting from these classes to a new labeling system called the Pregnancy and Lactation Labeling Rule (PLLR). While the PLLR provides a more holistic view of compound-mediated pregnancy risk, the old classification system (A, B, C, D, and X) is more practical for use in our study for the following reasons: (i) there is greater data availability, given that many drug labels have not been updated with PLLR data, and (ii) they fit better into the schema of binary classification (*i.e.*, toxic vs. nontoxic). A plethora of

information on FDA-approved prescription drugs and other biological products (such as cosmetics) is contained in the DailyMed database; for example, their indications, contraindications, dosage, route of administration, and use in specific populations (for example, pregnant women). [31]

As of March 2023, there were 46,943 records in DailyMed.[31] We collected these records and curated the data. After removing compounds with missing SMILES, 42,075 records remained. We excluded the 25,245 non-classified records, leaving 16,830 records. Of these, 9,455 were category C, in which there are no satisfactory studies in pregnant women, but animal studies demonstrated a risk to the fetus. Due to this, we excluded class C compounds, leaving 7,375 records. After removing inorganics, organometallics, and mixtures and normalizing chemotypes, 4,023 compound entries remained. After removing duplicate entries for the same compound, 221 remained (102 nontoxic categories A and B, 119 toxic categories D and X, and 3 non-concordant). The non-concordant compounds were investigated further once the datasets were merged.

Table 1. FDA Pregnancy-Risk Categories and their respective definitions.

Category	Definition
A	No risk in human studies (studies in pregnant women have not demonstrated a risk to the fetus during the first trimester).
B	No risk in animal studies (there are no adequate studies in humans, but animal studies did not demonstrate a risk to the fetus).

C	Risk cannot be ruled out. There are no satisfactory studies in pregnant women, but animal studies demonstrated a risk to the fetus; the potential benefits of the drug may outweigh the risks.
D	Evidence of risk (studies in pregnant women have demonstrated a risk to the fetus; potential benefits of the drug may outweigh the risks).
X	Contraindicated (studies in pregnant women have demonstrated a risk to the fetus, and/or human or animal studies have shown fetal abnormalities; risks of the drug outweigh the potential benefits).

*Definitions from "Pregnancy Medications" by Leek and Arif.[42]

Teratogen Information System (TERIS)

The Teratogen Information System (TERIS) database comprises over 1,700 compounds paired with in-depth summaries of their teratogenic risk.[32] Unfortunately, we did not have access to this database. However, a dataset consisting of 293 compounds from either/both the FDA and TERIS was published by Arena et al.[43] We collected these compounds. We removed the compound, "Azatguiorube," for which no structural information was available. After removing inorganics, organometallics, and mixtures and normalizing chemotypes, 275 compounds remained. We removed one duplicate, leaving 274 compounds (160 nontoxic and 114 toxic).

Aschner et al. (2017)

Aschner et al. compiled a list of 75 positive and negative compounds for developmental neurotoxicity.[44] It should be noted that this endpoint is more specific than "developmental

toxicity"; for example, a compound that may not induce neurotoxicity may exhibit developmental toxicity. As described in the section, "**Merging the datasets and verification of nontoxic compounds,"** we verified all negative compounds after merging the datasets, so this difference does not invalidate the use of these chemicals. We removed records for compounds with no structural information, leaving 73 compounds remaining. After removing inorganics, organometallics, and mixtures and normalizing chemotypes, 62 compounds remained. There was only one duplicate compound. The final dataset was comprised of 61 compounds (35 nontoxic and 26 toxic).

Grandjean (2006; 2014)

Grandjean and Landrigan compiled one list of six developmentally neurotoxic compounds in 2006 and another set of six in 2014.[45,46] Two entries were not specific compounds but classes of compounds (for example, polychlorinated biphenyls) and were therefore excluded. After removing inorganics, organometallics, and mixtures and normalizing chemotypes, 5 compounds remained (all toxic).

Abortion medications

There are two medication abortion compounds in the United States, specifically, mifepristone and misoprostol. These structures were cleaned and standardized as described above.

Web search for non-developmentally toxic compounds

To increase the number of non-toxicants in our dataset, we performed a web search for medications that are widely accepted as safe for the pregnant population. We required that two or

more medical websites or peer-reviewed studies supported the safety of these compounds. The web search was performed with adapted web scraping code, which enabled a quick compilation and filtering of relevant articles.[47] PubMed articles were searched with combinations of: "Compound selected," and "Teratogenic," and "Trimester Risk."

Martin (2022) dataset

It is challenging to validate compounds as being developmentally nontoxic. Martin et al. sought to address this issue. Specifically, a panel of experts from the Center for Computational Toxicology and Exposure of the U.S. Environmental Protection Agency performed a thorough literature review on 39 compounds suggested by previous studies to be non-developmentally neurotoxic. They found that 29 chemicals did not have sufficient evidence to be categorized as nontoxic. After merging the datasets, according to this study, we removed all compounds that did not have sufficient evidence to classify them as nontoxic.[48] We verified the remaining ten compounds in the study that were classified as having sufficient evidence to be categorized as nontoxic, as described below.

Merging the datasets and verification of nontoxic compounds

After curating each dataset individually, we merged them. Using KNIME, we removed overlapping compounds (identical InChiKey) between datasets, leaving 482 unique compounds. Then, we exported this list of compounds to Comma-Separated Values (CSV) so that we could manually investigate all compounds.

Two separate searches were conducted: overall toxicity and trimester toxicity. Each of the 482 compounds was evaluated for overall toxicity based on the available literature. If toxicity was

identified at any stage, the compound was labeled as toxic. Rigorous criteria were used for dataset creation in compliance with the US EPA guidelines for Developmental Toxicity Risk Assessment[49,50], wherein only stringent and validated data would enter the dataset. Human studies were only included if they followed the criteria stated in the US EPA guidelines with the change that the study was included if the sample size in the study was more significant than or equal to 80, as studies with larger sample sizes testing developmental toxicity were scarce. Animal studies were only included if the ratio of the tested dose was stated and if the tests were performed on rats, rabbits, or mice in compliance with OECD guidelines.[14–16] Compounds that did not contain reliable studies were removed. We removed these compounds if animal studies demonstrated a compound's developmental toxicity but were not up to regulatory standards. If epidemiological studies showed developmental toxicity, we replaced the outcome with "toxic," as human data is preferred over animal data. We also removed compounds for which reliable studies were not available. After the literature search and careful curation, removing compounds without reliable sources, 144 compounds remained (59 toxic and 85 nontoxic).

Then, a separate literature search was performed to classify compounds based on their trimester toxicity. This search was more rigorous, wherein a study was included only if it labeled the trimester at which the toxicity occurred. In this case, various drugs ended up being labeled as "toxic" or "nontoxic" for a particular trimester or removed for trimesters where articles were unavailable or not up to regulatory standards. In this study, data labeling was performed using a binary classification system within KNIME, where "toxic" compounds were encoded as "1" and "nontoxic" as "0" to facilitate streamlined computational processing and analysis. After excluding irrelevant compounds, only 156 remained for the first trimester, 65 for the second, and 60 for the

third. As an additional measure of validation, we manually inspected the compounds in both datasets before proceeding with QSAR modeling.

QSAR Modeling:

Calculation of Molecular Fingerprints

We used the RDKit node in KNIME to calculate Extended-Connectivity Fingerprints, with a diameter 4 (ECFP4) with 2,048 bits[51] and Molecular ACCess System (MACCS) Fingerprints.[52]

Chemical Space Analysis

Our dataset comprised toxic and nontoxic compounds as defined by testing for the overall developmental toxicity collected from the FDA[53], TERIS[32], and select articles.[45,48] The 144 compounds were investigated by plotting a similarity map generated using OSIRIS DataWarrior Software v.05.02.01.[54] The similarity map utilizes a Rubberbanding Forcefield approach, which translates similarity (vertices) between compounds (nodes). The similarity map approach entails the following steps: (*i*) all compounds are randomly positioned in a 2D space; (*ii)* Calculation of similarity matrix between all compounds using Tanimoto coefficients (Tc) and Datawarrior's default substructure-based binary fingerprint (FragFP)[55]; (*iii)* determination of most similar neighbors (Tc>0.8), considered for every compound; and (*iv)* stepwise relocation of all compounds to ensure similar molecules are located in proximity to each other.[56]

Model Development and Performance Assessment

QSAR models were developed and validated according to the best practices in the field.[57] The models were developed using the RF algorithm, wherein trees were decorrelated via

bootstrapping with replacement, and LightGBM, a gradient-boosting algorithm that optimizes model performance through a leaf-wise tree construction approach. Models used LightGBM[58] and RF[27] were implemented through Scikit-learn v.1.4.0.[59]. The consensus among trees defined in each model was used to ascertain the confidence of a binary prediction. Since all compounds were manually verified during data curation, models were built with 4 different criteria: models for overall toxicity, where toxicity in any trimester flags the compound as toxic, and models for each the first, second, or third trimester toxicities where toxicity would only be flagged if predicted to be toxic during the given trimester.

Because all data were binarized (*i.e.*, toxic vs. nontoxic), the following statistical metrics were used to assess different aspects of the performance of classification models (Equations 1-6):

Sensitivity (SE):

$$SE = \frac{N_{TP}}{N_{TP} + N_{FN}} \qquad [\text{Eq. } 1]$$

Specificity (SP):

$$SP = \frac{N_{TN}}{N_{TN} + N_{FP}} \qquad [\text{Eq. } 2]$$

Correct Classification Rate (CCR):

$$CCR = \frac{SE + SP}{2} \qquad [\text{Eq. } 3]$$

Positive Predictive Value (PPV):

$$PPV = \frac{N_{TP}}{N_{TP} + N_{FP}} \qquad [\text{Eq. } 4]$$

Negative Predictive Value (NPV):

$$NPV = \frac{N_{TN}}{N_{TN} + N_{FN}} \qquad [\text{Eq. } 5]$$

Area Under the Receiver Operating Characteristic Curve (AUC):

$$AUC = \sum_i [(SE_{i+1})((SP_{i+1} - (SP_i] \qquad [\text{Eq. } 6]$$

N represents the number of compounds, N_{TP} and N_{TN} represent the number of true positives and true negatives, and N_{FP} and N_{FN} represent the number of false positives and false negatives, respectively.

Compounds known to cause developmental toxicity were classified as positive (class 1), and nontoxic compounds were classified as negative (class 0).

Hyperparameter Optimization

Taking into consideration that the performance of machine learning (ML) is closely related to its hyperparameters, the models were optimized using a Bayesian approach, implemented in Optuna[60] v. 3.5.0. Optuna's framework can perform Bayesian hyperparameter optimization for a given set of descriptors and ML algorithms. The best hyper-parameters were then used to fine-tune the model using the entire training set of compounds and tested during the 5-fold cross-validation step.

Dimensionality Reduction

The dimensionality reduction method implemented was the filter of low-variance descriptors using the "Low Variance Filter" node in KNIME. Molecular fingerprints with a variance less than an established threshold were removed from the data set because they did not provide relevant information for the model. In this study, a threshold of 0.01 was utilized. We employed a threshold of 0.01 for dimensionality reduction to efficiently filter out noise and retain

only the most predictive features, thereby enhancing the model's ability to accurately predict developmental toxicity.

Dataset Split and 5-Fold Cross Validation

We employed 5-fold external cross-validation.[61] For this, the dataset is split into five equal parts, wherein one subset (20%) is used as the test set, and the remaining compounds (80%) compose the training set. This procedure is repeated five times, and each subset is used as the validation set exactly once. Models are built using the training set only, and compounds in the test set must not be present in the training set.

Threshold-Moving

We tried threshold-moving calibration of probability estimates to increase prediction confidence without losing data, i.e., without needing to balance the data. QSAR models probability thresholds were adjusted using a threshold-moving approach, incorporated into Scikit-learn Version 1.4.0.[56,59]. Threshold-moving was used to select the binary classification probability threshold for the model that produced the highest geometric-mean values on these test sets. The geometric mean was chosen as it better assesses the performance of models when predicting imbalanced data.[56,62–64]

Applicability domain

We have previously established the importance of the applicability domain when analyzing predictions from developed QSAR models. The AD must be addressed for the given chemical space of predictive models to identify "reliable" and "unreliable" regions for

predictions.[56,65] Thus, users should only consider the model's predictions if their set of compounds is within the AD.

We employed the "Applicability Domain" meta-node to assess the AD of our models. Within this meta node, the "Domain-Similarity" node uses Euclidean distances to measure chemical similarity between a compound from the test set and its nearest neighbor in the training set. The prediction may be unreliable if the distance of a compound not present in the test set to its nearest neighbor is higher than an arbitrary parameter (Z=0.5) that controls the significance level.[66]

Model Interpretation

Contribution maps [67,68] were generated from QSAR models to visualize atoms and fragments contributing to developmental toxicity. An atom's "weight" was considered a predicted probability difference obtained when bits in the fingerprints corresponding to the atom were removed. Then, the normalized weights were used to color atoms in a topography-like map in which green indicates the contribution to toxicity (i.e., predicted probability decreases when bits are removed) and red indicates a negative contribution to toxicity (i.e., predicted probability increases when bits are removed).[68]

Model Implementation – The PregPred Web Application

The QSAR models developed in this study have been implemented as a web application, PregPred, which runs on an Ubuntu server. The PregPred application is encoded using Flask[69], uWSGI[70], Nginx[71], Python 3.8[72], RDKit[52], scikit-learn[59], and Javascript[73]. PregPred also includes the JSME molecule editor[74], which is written in JavaScript and supported by most popular web

browsers. The server takes input chemicals and produces the developmental toxicity predictions for the user.

Results and Discussion

In the present study, we integrated, curated, and carefully verified the most extensive collection of developmental toxicants for overall trimester risk and risk for each trimester (Supplemental File S2). Data curation represents a quintessential step for the construction of QSAR models. However, given the inconsistencies we found between FDA pregnancy category labeling and the most up-to-date literature – many compounds were labeled as nontoxic by the FDA, but there were more recent epidemiological and animal studies – we emphasize the importance of our rigorous biological curation. Specifically, we searched the literature for evidence for or against developmental toxicity for each compound in the dataset and removed compounds for which tests were not up to regulatory standards. OECD testing guidelines were utilized when verifying compound activity data with current literature.[15,75] We do not intend to criticize other groups for using the FDA and TERIS developmental toxicity classifications at the face value; it seems reasonable to expect that the FDA data, especially, would be up to date. Instead, this is an issue that needs to be addressed by drug regulatory agencies to ensure that drug-specific pregnancy information is current. After merging the datasets, the same compound may contain multiple entries in the modeling and external sets. QSAR models developed with duplicate models will have low accuracy if toxicity outcomes are dissimilar or over-optimistic performance if outcomes are identical.[35,36] Nevertheless, we took extensive measures in this study to ensure our developmental toxicity model was built on the most reliable and accurate data possible. Table 2 demonstrates the compounds obtained after each trimester's thorough, up-to-date literature search.

Table 2. Distribution of chemicals in each dataset for overall toxicity model and first, second, and third-trimester toxicity models.

Dataset	Classification models		
	Toxic	**Nontoxic**	**Total**
Overall Toxicity	59	85	144
First Trimester Toxicity	50	106	156
Second Trimester Toxicity	18	47	65
Third Trimester Toxicity	15	45	60

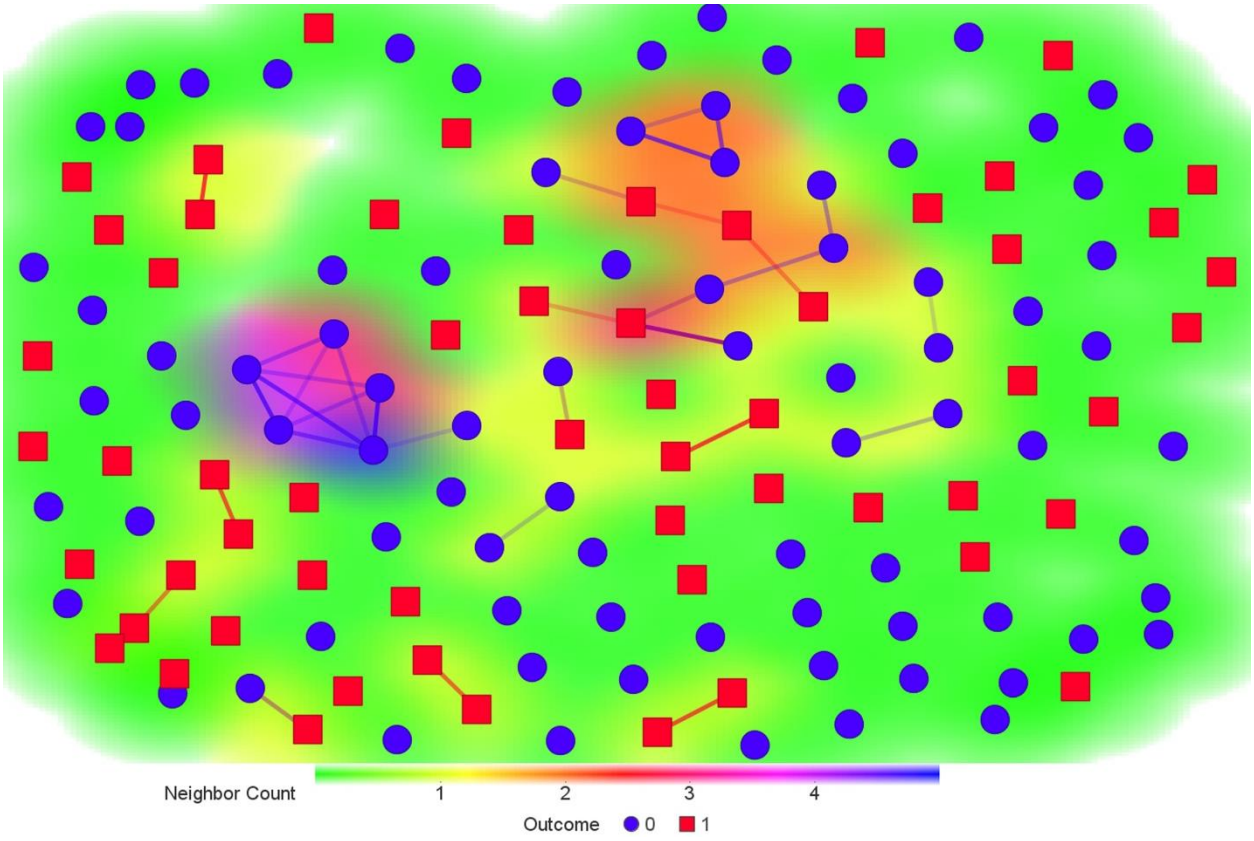

Figure 1. Structural distribution of toxic (1) and nontoxic (0) compounds from the Overall Toxicity dataset. Clusters of highly similar compounds are connected. Blue circles represent nontoxic compounds, and red squares represent toxic compounds. Compounds with a Tanimoto coefficient >0.8 are connected by vertices. The color scheme in the background represents the number of neighbors.

The chemical space analysis was performed using the overall toxicity dataset. The analysis has been performed by plotting the dataset using similarity maps. [54] As shown in Fig. 1, most toxic and nontoxic compounds do not share the same clusters or are not connected and the dataset contains few toxicity cliffs (i.e., structurally similar compounds with a significant difference in toxicity), which improves the effective discrimination between toxic and nontoxic entities, and paves the way for more reliable predictions with our models.[76–78]

As described in the *QSAR Modeling* section, 16 models were built with various combinations of fingerprints and ML methods. The RF and LightGBM models were built in KNIME and validated with 5-fold cross-validation. The statistical characteristics of the model are shown in Tables 3 and S1. All cross-validated models for the overall toxicity, first, second, and third-trimester toxicities are present in Table 3 after threshold moving. Threshold moving increases prediction confidence without losing data (i.e., we tried threshold-moving calibration of probability estimates without the need to balance the data).[79,80] All cross-validated developmental toxicity endpoint models showed high predictive accuracy on 5-fold external cross-validation based on several metrics, including CCR, SE, SP, PPV, and NPV.

Briefly, overall toxicity models showed reasonable CCR (65-68%), SE (62-72%), and SP (62-82%). The models also displayed a good coverage of 75%-82% with varying calibrations. The

ECFP4 + RF model showed excellent coverage and consistent metrics, so it was implemented into the web tool. The first trimester models showed reasonable CCR (60-68%), SP (56-80%) and PPV (73-81%), with a good coverage of (72.7-90%). The ECFP4 + RF model was also selected for the web tool to represent the first trimester, as it showed the highest coverage and accuracy for first trimester models. The second-trimester models had adequate CCR (59-73%), PPV (84-87%), and coverage of 54.5%-90%. The ECFP4 + LightGBM model showed the most promising performance, so it was selected for the web tool. Lastly, third-trimester models showed superior SP (73-80%), PPV (83-88%), and coverage (70-90%). Given its performance, the ECFP4 + LightGBM model was selected for the web tool.

Table 3. Statistical characteristics of the developmental toxicity calibrated QSAR models.

Fingerprint	**Method**	**CCR**	**SE**	**SP**	**PPV**	**NPV**	**Coverage (%)**	**PT**
Overall Toxicity								
ECFP4	RF	0.65	0.66	0.63	0.55	0.73	82.1	0.35
MACCS	RF	0.67	0.72	0.62	0.57	0.76	75.0	0.38
ECFP4	LightGBM	0.65	0.48	0.82	0.65	0.70	82.1	0.46
MACCS	LightGBM	0.62	0.60	0.64	0.54	0.70	75.0	0.41
First Trimester Toxicity								
ECFP4	RF	0.66	0.75	0.56	0.75	0.56	90.9	0.72
MACCS	RF	0.63	0.59	0.66	0.75	0.47	72.7	0.67
ECFP4	LightGBM	0.63	0.71	0.54	0.73	0.51	90.9	0.72
MACCS	LightGBM	0.64	0.48	0.80	0.81	0.46	72.7	0.73
Second Trimester Toxicity								

ECFP4	RF	0.70	0.46	0.94	0.95	0.44	90.9	0.88
MACCS	RF	0.65	0.51	0.78	0.84	0.41	54.5	0.73
ECFP4	LightGBM	0.72	0.76	0.67	0.84	0.55	90.9	0.79
MACCS	LightGBM	0.66	0.49	0.83	0.87	0.42	54.5	0.75
Third Trimester Toxicity								
ECFP4	RF	0.73	0.59	0.87	0.92	0.45	90.0	0.89
MACCS	RF	0.62	0.51	0.73	0.83	0.37	70.0	0.78
ECFP4	LightGBM	0.70	0.67	0.73	0.87	0.46	90.0	0.87
MACCS	LightGBM	0.62	0.44	0.8	0.85	0.35	70.0	0.79

RF, Random Forest; ECFP4, extended connectivity fingerprints with diameter 4; LightGBM, Light Gradient-boosting machine; MACCS, Molecular Access Systems keys fingerprint; CCR, Correct Classification Rate; SE, Sensitivity; SP, Specificity; PPV, Positive predictive value; NPV, Negative predictive value; Coverage, a ratio of the test set or external set compounds within the applicability domain; Probability Threshold, PT; Statistical results all obtained after threshold-moving calibration. Statistical results obtained from the default probability threshold available in Supplementary Material.

In the developed RF and LightGBM models, a continuous value represents the probability of a given compound belonging to a specific class. In this study, independent training of classification models using developmental toxicity in overall or individual trimesters aims to differentiate toxic and non-toxic compounds. Due to the nature of the endpoint and the lack of available literature, binary classification models were utilized instead of continuous ones to

categorically differentiate between compounds as toxic or non-toxic, simplifying the prediction of developmental toxicity risks.

Typically, probabilities less than 0.5 are assigned to the nontoxic class, while values greater than or equal to 0.5 are assigned to the tox class. However, when dealing with imbalanced data, QSAR models for classification often yield poor probability estimates (<0.5) for the minority class.[56] We explored various probability thresholds ranging from 0 to 1 to identify the optimal threshold for model performance. The statistical performance of these QSAR models is detailed in Table S1. Overall, the statistical performances of the calibrated models for developmental toxicity resulted in significant enhancements in the statistical performance of these QSAR models (Fig 2A-D). Consequently, the ideal thresholds in the table were retained as the final model for predicting the developmental toxicity of new compounds. The calibrated models outperformed the uncalibrated counterparts or scored similarly, rarely underperforming models without calibration. (File S1)

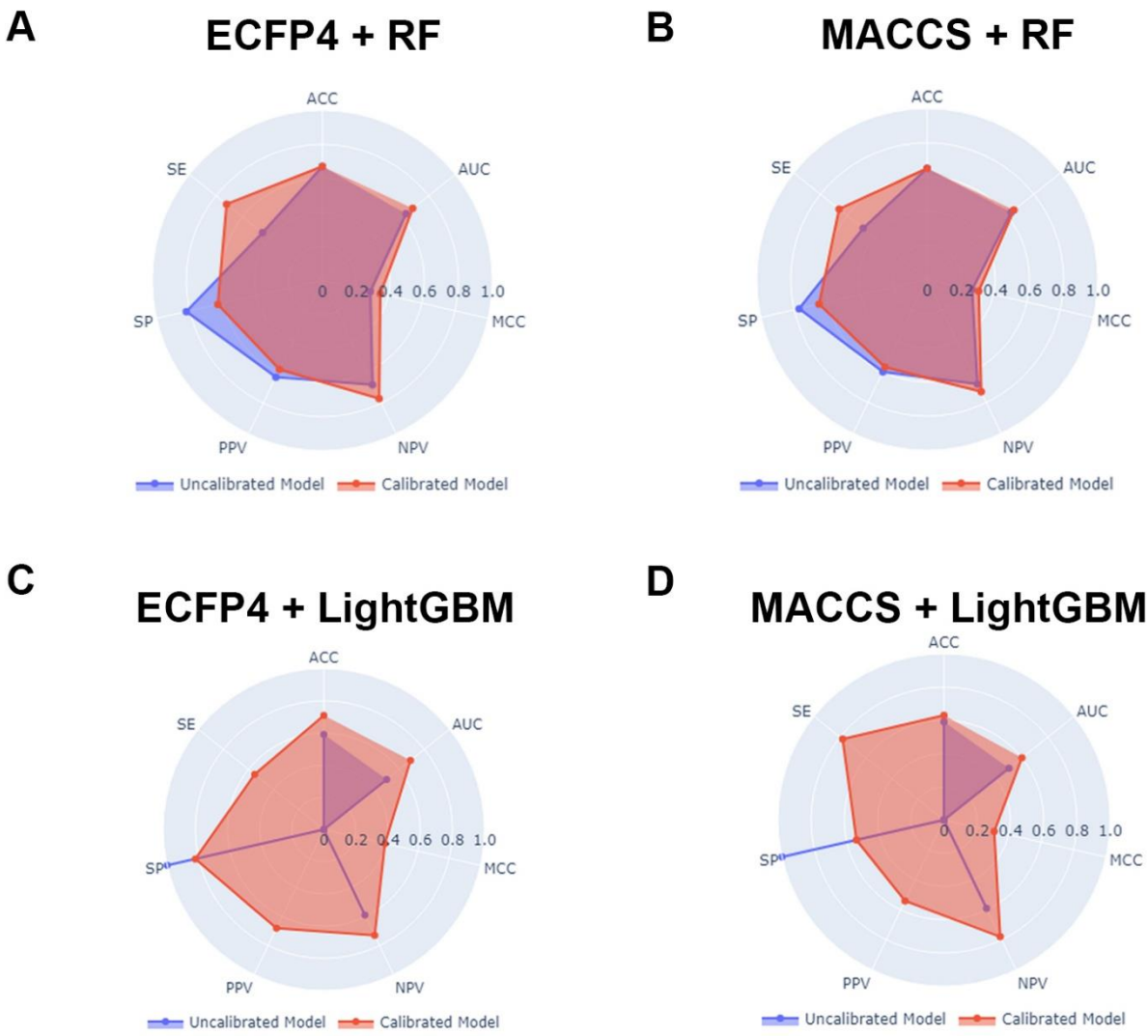

Figure 2. Radar chart for Overall Toxicity Models Predictions with Uncalibrated and Calibrated Models. **1A)** Uncalibrated and Calibrated model statistics for ECPF4+ RF model **1B)** Uncalibrated and Calibrated model statistics for MACCS + RF models **1C)** Calibrated and Uncalibrated models for ECFP4 and LightGBM models. **1D)** Calibrated and Uncalibrated models for MACCS+LightGBM models. The calibrated models outperformed the uncalibrated counterparts or scored similarly, rarely underperforming models without calibration.

The most predictive classification model for the developmental toxicity category was implemented in the PregPred web application (https://pregpred.mml.unc.edu/). The PregPred web tool has an intuitive user interface (Figure 3.), in which the user may draw a compound of interest in the "molecular editor" box or directly paste the SMILES string of the queried chemical structure. The user will be prompted to select the models they wish to use in a checkbox format (e.g., overall toxicity, first trimester toxicity). After hitting the "Get Properties" button, the user will receive the classification outcomes (e.g., toxic, nontoxic) using the best classification model for each of the selected models. The user will also be shown the predicted probability values, which are helpful for estimating the confidence of classification outcomes. [81] All predictions also contain the AD estimates and mechanistic interpretation using color-coded maps of fragment contribution. [67,68] For the fragment contribution maps, atoms or fragments promoting positive toxicity are highlighted in green, while those decreasing the toxicity are highlighted in purple. The models developed in this study are available within the PregPred web application (https://pregpred.mml.unc.edu/).

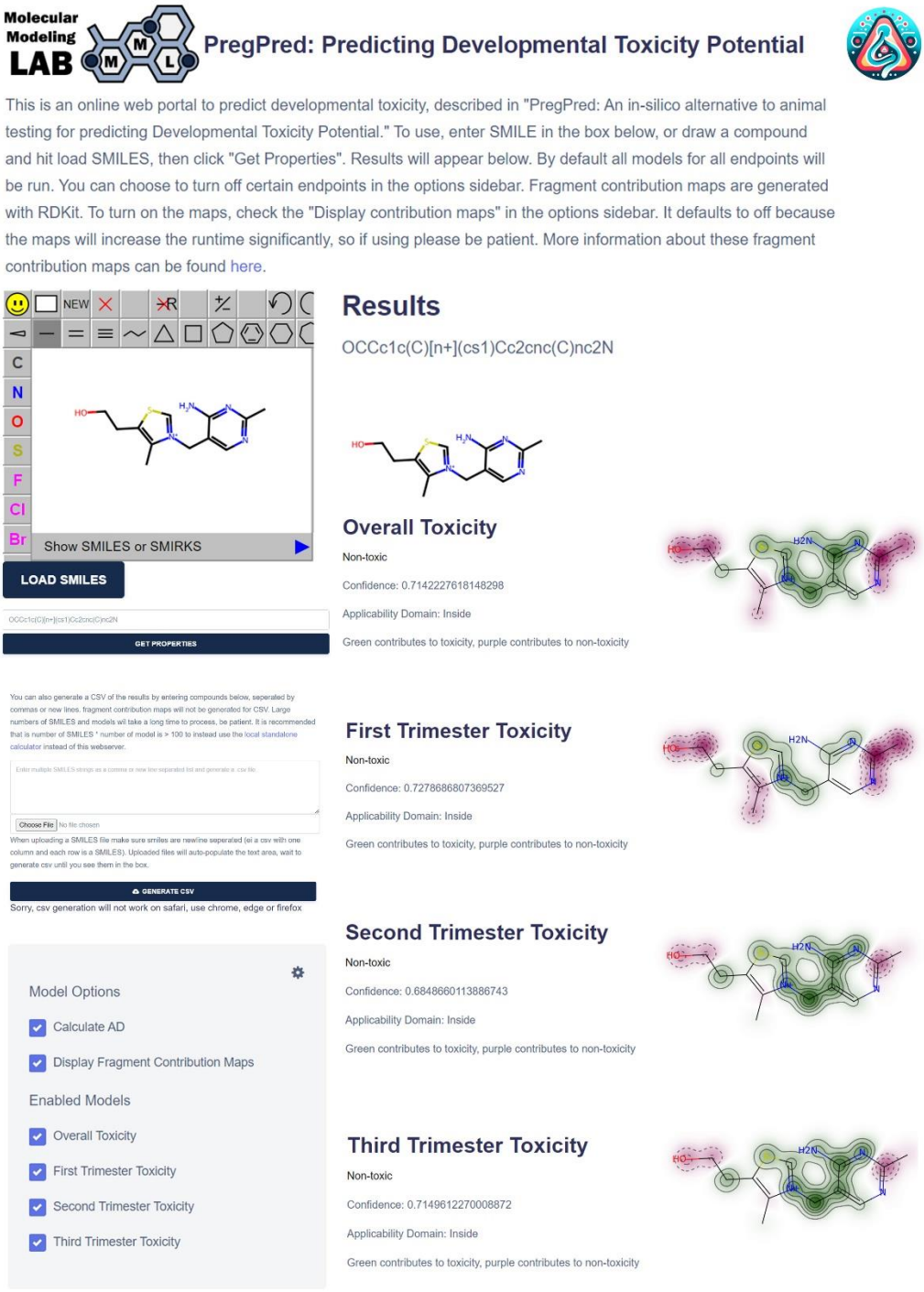

Figure 3. User interface for PregPred. The query chemical can be drawn in the "molecular editor" box or directly inserted by pasting the SMILES strings. After hitting the "get properties" button, the user will receive predicted values for developmental toxicity for all the model options selected

under the box "Model Options" and, if selected, color-coded maps of fragment contributions to toxicity.

In order to carry out an additional statistical validation of the PregPred app, we compiled and prepared a list of 6 additional drugs (not included in any of the QSAR datasets) with developmental toxicity data from various studies.[82–86] These were compounds with known developmental toxicity effects from studies that adhered to the criteria listed under "Merging the datasets and verification of nontoxic compounds." Then, we used PregPred to predict the developmental toxicity potential of these compounds (Fig 4).

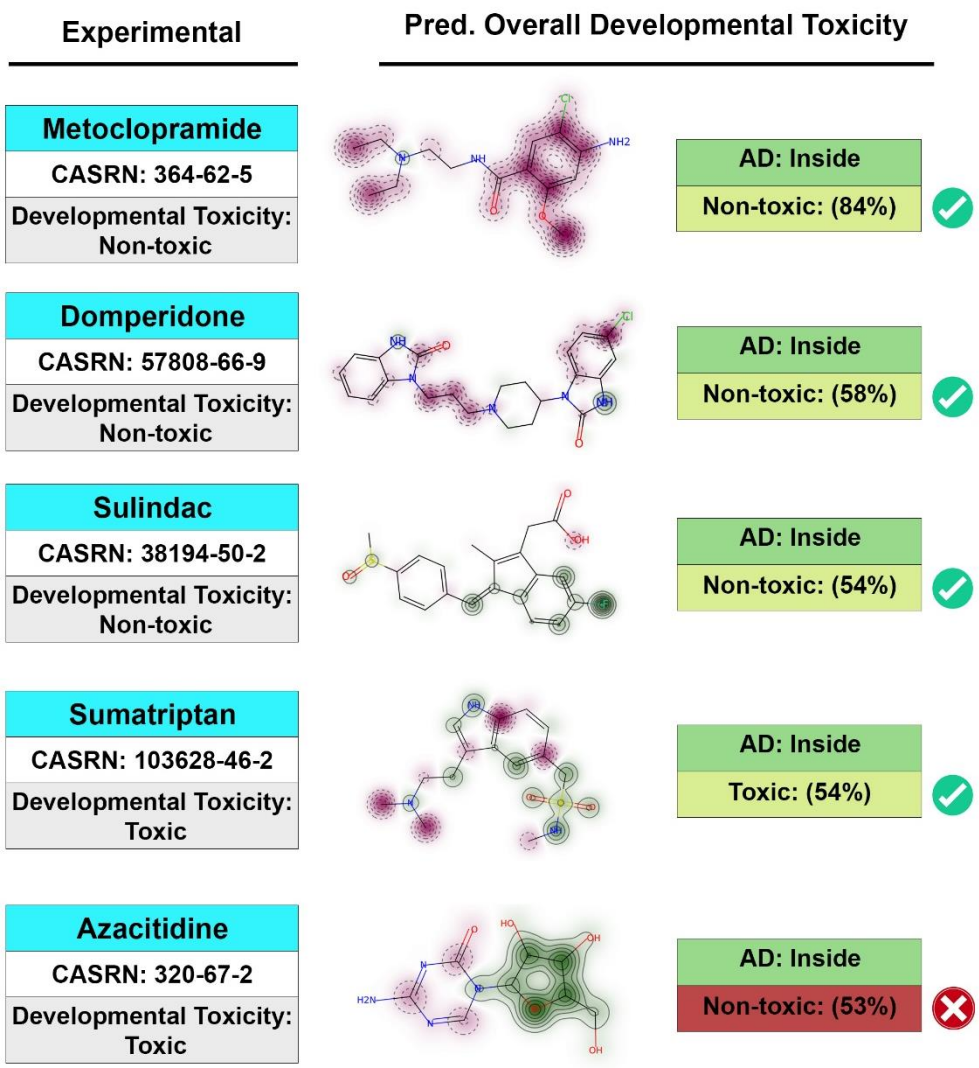

Fig 4. Experimental and predicted toxicity of six developmental toxicants not included in any of the train or test datasets with structural fragments' contribution to toxicity. Fragments contributing to toxicity are highlighted in green, and fragments decreasing toxicity are colored red. Predictions were made using the overall toxicity model from the PregPred web tool. Confidence in the prediction is shown inside the parenthesis.

According to the results of PregPred (Figure 4), the classification models correctly classified 4 out of 5 compounds. These results corroborate the high external predictive power reported above, especially when considering compounds inside the AD. Conversely, one

compound was erroneously predicted by the overall toxicity classification model. The incorrect prediction was inside of the model's applicability; therefore, the analysis of predictions using PregPred should be cautious, as the classification models were trained using small datasets, and the biological mechanism underlying the compound's developmental toxicity might be multifactorial or distinct from those represented in the training data, indicating a gap in the model's ability to generalize across mechanistic pathways.

In toxicity modeling, it is more important for a model to accurately predict toxic than nontoxic compounds; this can reduce animal testing and resource waste by eliminating compounds likely to fail downstream in the development process due to safety concerns. Therefore, the sensitivity of the models (66-76%) indicates their utility, given that toxic compounds were classified as positives or class 1. The specificity (54-83%) indicates that nontoxic compounds are not frequently mislabeled as toxic. Our models yielded PPV ranging from 55% to 87% and an NPV from 42% to 77%, which underscores the models' considerable potential in contributing to the advancement of the 3 Rs (Reduce, Refine, and Replace). The high PPV and NPV scores provide a comprehensive understanding of a model's predictive accuracy, with high values suggesting a precise model in identifying true cases of a condition and reliable in confirming its absence. While the model could be improved by including more diverse true negatives (nontoxic compounds) in the dataset, we still emphasize that sensitivity and specificity are considerable for a model with only 144 compounds.

To our knowledge, the high sensitivity of this developmental toxicity QSAR model outperforms others in the field. Other groups have reported competitive QSAR models for developmental toxicity on similar datasets, such as the FDA and TERIS databases.[24,25] Some groups utilized proper chemical curation protocols (*i.e.*, removing mixtures and salts and

standardizing chemotypes), but not all. In the future, we hope to expand the dataset to include more compounds verified as either developmentally or non-developmentally toxic, especially those with evidence from epidemiological studies. However, it is difficult to definitively classify compounds as being non-developmentally toxic, given the complexity of this endpoint and the many manifestations of developmental toxicity.

Another major challenge for this endpoint is elucidating whether the medication induces developmental toxicity or the mother's underlying illness for which she is being treated. To address this issue, we suggest that regulatory agencies incentivize, for each drug and indication, epidemiological studies following cohorts of babies born to women who took the medication throughout pregnancy and those born to women who did not. This, of course, is outside the scope of this work; however, it would prove a helpful step in furthering maternal and fetal health, which is desperately needed in the US and other countries.

We suggest that the utility of our developmental QSAR model lies within its potential to predict developmental toxicants with high accuracy. Currently, regulatory standards require that animal tests be used to determine developmental toxicity for environmental compounds and FDA-approved drugs and cosmetics. Unfortunately, these studies are time-consuming, expensive, and raise ethical concerns. In contrast, our QSAR model can be easily implemented in the early stages of drug development to reduce animal testing downstream. We hope that this model, as well as the other toxicity models we have developed, progresses further toward regulatory acceptance.

Conclusions

We have created the largest publicly available heavily curated database of developmental toxicity that includes the per-trimester data as well as overall toxicity irrespective of pregnancy

term. We have compiled robust QSAR models for accurately predicting developmental toxicants with a CCR of 62-72%, sensitivity of 54-76%, PPV of 55-87%, and NPV of 42-77%. These models were implemented in the PregPred web app, which is reliable, fast, and user-friendly for the assessment of the developmental toxicity of compounds. Users can make predictions using these rigorous and externally validated computation models that fulfill all the OECD principles for developing and validating QSAR models for regulatory purposes. The web app is intuitive and does not require prior programming or knowledge of computation skills for its utilization. The predictions for a single compound take only a few seconds. Furthermore, the PregPred interface provides users with the following outcomes: (i) toxic/nontoxic classification for overall developmental toxicity and trimester toxicity; (ii) confidence in the predictions; (iii) applicability domain estimation; and (iv) color-coded contribution maps illustrating the relative contribution of chemical fragments for toxicity. Considering the model's accuracy and ease of implementation, we suggest that this be considered a novel alternative approach in light of the 3Rs (refining, reducing, and replacing) for animal testing. Medication safety in pregnant women is vastly understudied, and we hope that our *in silico* model supports the advancement of developmental toxicology.

Supplemental Material

Supplemental Material includes the results for the calibrated and uncalibrated models and curated datasets for the developmental toxicity endpoint in xlsx format.

Author Contributions

Each author has contributed significantly to this work. MR, NK, ENM, HH, and AT conceived and designed the study. RST, MR, HJM, and JTMF curated the data and developed the models. MR, RST, and JTMF analyzed the data. JW and RST implemented the models into the PregPred web application. RST, MR, JTMF, and ENM wrote the first draft of the manuscript. All authors read, edited, and approved the final manuscript.

References:

1. Prescription drug use during pregnancy in developed countries: a systematic review - Daw - 2011 - Pharmacoepidemiology and Drug Safety - Wiley Online Library. Accessed December 19, 2023. https://onlinelibrary.wiley.com/doi/full/10.1002/pds.2184?casa_token=09W67H2ORf0AAAAA%3A6X4tWbqFlHn8FOH3LtlNrb0wbdyS8yjxDg51aMPJIPONGQGhr67FoD7O6IL-DcDr9RZWataw7eXOw5jv

2. Waitt C, Astill D, Zavala E, et al. Clinical trials and pregnancy. *Commun Med*. 2022;2(1):1-5. doi:10.1038/s43856-022-00198-1

3. Gentile S, Galbally M. Prenatal exposure to antidepressant medications and neurodevelopmental outcomes: A systematic review. *J Affect Disord*. 2011;128(1):1-9. doi:10.1016/j.jad.2010.02.125

4. Tyl RW. Toxicity Testing, Developmental. Published online 2014.

5. Gilbert-Barness E. Teratogenic Causes of Malformations. *Ann Clin Lab Sci*. 2010;40(2):99-114.

6. Calado AM, Seixas F, Pires M dos A. Updating an Overview of Teratology. In: Félix L, ed. *Teratogenicity Testing: Methods and Protocols*. Springer US; 2024:1-38. doi:10.1007/978-1-0716-3625-1_1

7. Nishita M, Satake T, Minami Y, Suzuki A. Regulatory mechanisms and cellular functions of non-centrosomal microtubules. *J Biochem (Tokyo)*. 2017;162(1):1-10. doi:10.1093/jb/mvx018

8. Ridings JE. The Thalidomide Disaster, Lessons from the Past. In: Barrow PC, ed. *Teratogenicity Testing: Methods and Protocols*. Humana Press; 2013:575-586. doi:10.1007/978-1-62703-131-8_36

9. Scialli AR, Daston G, Chen C, et al. Rethinking developmental toxicity testing: Evolution or revolution? *Birth Defects Res*. 2018;110(10):840-850.

10. Organisation for Economic Co-operation and Development. *Test No. 414: Prenatal Developmental Toxicity Study*. OECD Publishing; 2018.

11. US Environmental Protection Agency. Health effects test guidelines: OPPTS 870.6300, developmental neurotoxicity study. Published online 1998.

12. Adverse Outcome Pathway (AOP) Development I: Strategies and Principles | Toxicological Sciences | Oxford Academic. Accessed April 2, 2024. https://academic.oup.com/toxsci/article/142/2/312/1621273

13. Vinken M. The adverse outcome pathway concept: A pragmatic tool in toxicology. *Toxicology*. 2013;312:158-165. doi:10.1016/j.tox.2013.08.011

14. No OT. 421: reproduction/developmental toxicity screening test. *OECD Guidel Test Chem Sect*. 2016;4.

15. The underestimated value of OECD 421 and 422 repro screening studies: Putting it in the right perspective - ScienceDirect. Accessed January 21, 2024. https://www.sciencedirect.com/science/article/pii/S0890623814000604?casa_token=RYaLZnq4n74AAAAA:uRxKgIDsj2ggAj9Gu_sUjS5wVapUlms5Uh60_0eUAULKWFkb7PdkR0mcqgKwXuEyj7H4ds4cHQ

16. Test No. 421: Reproduction/Developmental Toxicity Screening Test | en | OECD. Accessed March 20, 2024. https://www.oecd.org/env/test-no-421-reproduction-developmental-toxicity-screening-test-9789264264380-en.htm

17. Korn D, Thieme AJ, Alves VM, et al. Defining clinical outcome pathways. *Drug Discov Today*. 2022;27(6):1671-1678. doi:10.1016/j.drudis.2022.02.008

18. US EPA O. Cost Estimates of Studies Required for Pesticide Registration. Published April 2, 2018. Accessed December 30, 2023. https://www.epa.gov/pesticide-registration/cost-estimates-studies-required-pesticide-registration

19. Van Norman GA. Limitations of animal studies for predicting toxicity in clinical trials: is it time to rethink our current approach? *JACC Basic Transl Sci*. 2019;4(7):845-854.

20. Dubovický M, Kovačovský P, Ujházy E, Navarová J, Brucknerová I, Mach M. Evaluation of developmental neurotoxicity: some important issues focused on neurobehavioral development. *Interdiscip Toxicol*. 2008;1(3-4):206-210.

21. New Approach Methods Work Plan.

22. Ford KA. Refinement, reduction, and replacement of animal toxicity tests by computational methods. *ILAR J*. 2017;57(2):226-233.

23. Ciallella HL, Russo DP, Sharma S, et al. Predicting prenatal developmental toxicity based on the combination of chemical structures and biological data. *Environ Sci Technol*. 2022;56(9):5984-5998.

24. Borba JVB, Braga RC, Alves VM, et al. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem Res Toxicol*. 2021;34(2):258-267. doi:10.1021/acs.chemrestox.0c00186

25. Rath M, Wellnitz J, Martin HJ, et al. Novel Pharmacokinetics Profiler (PhaKinPro): Model Development, Validation, and Implementation as a Web-Tool for Triaging Compounds with Undesired PK Profiles. Published online March 20, 2023. doi:10.26434/chemrxiv-2023-rnc4l

26. Alves VM, Auerbach SS, Kleinstreuer N, et al. Curated data in—trustworthy in silico models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing. *Altern Lab Anim*. 2021;49(3):73-82.

27. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.

28. Hastie T, Tibshirani R, Friedman J, Hastie T, Tibshirani R, Friedman J. Random forests. *Elem Stat Learn Data Min Inference Predict*. Published online 2009:587-604.

29. Julien E, Willhite CC, Richard AM, DeSesso JM, Group IRW. Challenges in constructing statistically based structure-activity relationship models for developmental toxicity. *Birt Defects Res A Clin Mol Teratol*. 2004;70(12):902-911. doi:10.1002/bdra.20087

30. Knudsen TB, Kavlock RJ, Daston GP, Stedman D, Hixon M, Kim JH. Developmental toxicity testing for safety assessment: new approaches and technologies. *Birth Defects Res B Dev Reprod Toxicol*. 2011;92(5):413-420. doi:10.1002/bdrb.20315

31. DailyMed. Accessed December 30, 2023. https://dailymed.nlm.nih.gov/dailymed/

32. Home | TERIS. Accessed December 30, 2023. https://deohs.washington.edu/teris/

33. Cassano A, Manganaro A, Martin T, et al. CAESAR models for developmental toxicity. In: Vol 4. Springer; 2010:1-11.

34. Chen Q, Gan Y, Wang K, Li Q. PregTox: A Resource of Knowledge about Drug Fetal Toxicity. *BioMed Res Int*. 2022;2022.

35. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model*. 2010;50(7):1189-1204. doi:10.1021/ci100176x

36. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model*. 2016;56(7):1243-1252. doi:10.1021/acs.jcim.6b00129

669 37. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol*.
670 2015;11(8):535-535. doi:10.1038/nchembio.1881

671 38. Alves VM, Muratov EN, Zakharov A, Muratov NN, Andrade CH, Tropsha A. Chemical
672 toxicity prediction for major classes of industrial chemicals: Is it possible to develop
673 universal models covering cosmetics, drugs, and pesticides? *Food Chem Toxicol*.
674 2018;112:526-534.

675 39. Berthold MR, Cebron N, Dill F, et al. KNIME-the Konstanz information miner: version 2.0
676 and beyond. *AcM SIGKDD Explor Newsl*. 2009;11(1):26-31.

677 40. Chemaxon. Accessed December 30, 2023. https://chemaxon.com

678 41. Law R, Bozzo P, Koren G, Einarson A. FDA pregnancy risk categories and the CPS: do they
679 help or are they a hindrance? *Can Fam Physician*. 2010;56(3):239-241.

680 42. Leek JC, Arif H. Pregnancy Medications. In: *StatPearls*. StatPearls Publishing; 2023.
681 Accessed December 20, 2023. http://www.ncbi.nlm.nih.gov/books/NBK507858/

682 43. Arena V, Sussman N, Mazumdar S, Yu S, Macina O. The utility of structure–activity
683 relationship (SAR) models for prediction and covariate selection in developmental toxicity:
684 comparative analysis of logistic regression and decision tree models. *SAR QSAR Environ*
685 *Res*. 2004;15(1):1-18.

686 44. Aschner M, Ceccatelli S, Daneshian M, et al. Reference compounds for alternative test
687 methods to indicate developmental neurotoxicity (DNT) potential of chemicals: example lists
688 and criteria for their selection and use. *Altex*. 2017;34(1):49.

689 45. Grandjean P, Landrigan PJ. Developmental neurotoxicity of industrial chemicals. *The*
690 *Lancet*. 2006;368(9553):2167-2178.

691 46. Grandjean P, Landrigan PJ. Neurobehavioural effects of developmental toxicity. *Lancet*
692 *Neurol*. 2014;13(3):330-338.

693 47. PubMed Web Scraper. Devpost. Published March 27, 2020. Accessed December 30, 2023.
694 https://devpost.com/software/pubmed-web-scraper

695 48. Martin MM, Baker NC, Boyes WK, et al. An expert-driven literature review of "negative"
696 chemicals for developmental neurotoxicity (DNT) in vitro assay evaluation. *Neurotoxicol*
697 *Teratol*. Published online 2022:107117.

698 49. Guidelines for Developmental Toxicity Risk Assessment.

699 50. Assessment of the U.S. Environmental Protection Agency methods for identification of
700 hazards to developing organisms, Part II: The developmental toxicity testing guideline -
701 Claudio - 1999 - American Journal of Industrial Medicine - Wiley Online Library. Accessed
702 March 26, 2024. https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-
703 0274(199906)35:6%3C554::AID-AJIM2%3E3.0.CO;2-X

51. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *J Chem Inf Model*. 2010;50(5):742-754. doi:10.1021/ci100050t

52. Landrum G. Rdkit: Open-source cheminformatics software. Published online 2016.

53. Commissioner O of the. U.S. Food and Drug Administration. FDA. Published January 2, 2024. Accessed January 8, 2024. https://www.fda.gov/

54. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis | Journal of Chemical Information and Modeling. Accessed January 21, 2024. https://pubs.acs.org/doi/10.1021/ci500588j

55. Yu T, Nantasenamat C, Anuwongcharoen N, Piacham T. Machine Learning Approaches to Investigate the Structure–Activity Relationship of Angiotensin-Converting Enzyme Inhibitors. *ACS Omega*. 2023;8(46):43500-43510. doi:10.1021/acsomega.3c03225

56. Moreira-Filho JT, Braga RC, Lemos JM, et al. BeeToxAI: An artificial intelligence-based web app to assess acute toxicity of chemicals to honey bees. *Artif Intell Life Sci*. 2021;1:100013.

57. Tropsha A. Best practices for QSAR model development, validation, and exploitation. *Mol Inform*. 2010;29(6-7):476-488.

58. Ke G, Meng Q, Finely T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In: *Advances in Neural Information Processing Systems 30 (NIP 2017)*. ; 2017. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

59. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Mach Learn PYTHON*.

60. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: ; 2019:2623-2631.

61. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22(1):69-77.

62. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J Chem Inf Model*. 2014;54(3):705-712. doi:10.1021/ci400737s

63. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. In: Vol 97. Citeseer; 1997:179.

64. Barandela R, Sánchez JS, Garcıa V, Rangel E. Strategies for learning in class imbalance problems. *Pattern Recognit*. 2003;36(3):849-851.

65. Trisciuzzi D, Alberga D, Mansouri K, et al. Predictive structure-based toxicology approaches to assess the androgenic potential of chemicals. *J Chem Inf Model*. 2017;57(11):2874-2884. doi:10.1021/acs.jcim.7b00420

66. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des*. 2007;13(34):3494-3504.

67. Neves BJ, Braga RC, Alves VM, et al. Deep Learning-driven research for drug discovery: Tackling Malaria. *PLoS Comput Biol*. 2020;16(2):e1007025.

68. Riniker S, Landrum GA. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *J Cheminformatics*. 2013;5:1-7.

69. Grinberg M. *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media, Inc.; 2018.

70. The uWSGI project — uWSGI 2.0 documentation. Accessed March 3, 2024. https://uwsgi-docs.readthedocs.io/en/latest/

71. nginx news. Accessed March 3, 2024. https://nginx.org/

72. Welcome to Python.org. Python.org. Published February 29, 2024. Accessed March 3, 2024. https://www.python.org/

73. Home. Ecma International. Accessed March 3, 2024. https://ecma-international.org/home/

74. Bienfait B, Ertl P. JSME: a free molecule editor in JavaScript. *J Cheminformatics*. 2013;5(1):1-6.

75. Makris SL, Raffaele K, Allen S, et al. A Retrospective Performance Assessment of the Developmental Neurotoxicity Study in Support of OECD Test Guideline 426. *Environ Health Perspect*. 2009;117(1):17-25. doi:10.1289/ehp.11447

76. Stumpfe D, Bajorath J. Exploring Activity Cliffs in Medicinal Chemistry. *J Med Chem*. 2012;55(7):2932-2942. doi:10.1021/jm201706b

77. Stumpfe D, Hu Y, Dimova D, Bajorath J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J Med Chem*. 2014;57(1):18-28. doi:10.1021/jm401120g

78. Wassermann AM, Wawer M, Bajorath J. Activity Landscape Representations for Structure−Activity Relationship Analysis. *J Med Chem*. 2010;53(23):8209-8223. doi:10.1021/jm100933w

79. Collell G, Prelec D, Patil KR. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*. 2018;275:330-340. doi:10.1016/j.neucom.2017.08.035

80. On threshold moving-average models - Gooijer - 1998 - Journal of Time Series Analysis - Wiley Online Library. Accessed April 2, 2024. https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9892.00074?casa_token=ZTtNfOygV6MAAAAA:sKkuIZay-dUK1zKDPYi65YSQ2xw6y_lvQNmcUl3E_j91KEVJtb17Jj9affdcxysTABteVWjWzgJuI93H

81. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des*. 2003;17:241-253.

82. Hishinuma K, Yamane R, Yokoo I, et al. Pregnancy outcome after first trimester exposure to domperidone-An observational cohort study. *J Obstet Gynaecol Res*. 2021;47(5):1704-1710. doi:10.1111/jog.14709

83. Zheng X, Zhu Y, Zhao Y, Feng S, Zheng C. Taxanes in combination with platinum derivatives for the treatment of ovarian cancer during pregnancy: A literature review. *Int J Clin Pharmacol Ther*. 2017;55(9):753-760. doi:10.5414/CP202995

84. Ostesen M. Optimisation of antirheumatic drug treatment in pregnancy. *Clin Pharmacokinet*. 1994;27(6):486-503. doi:10.2165/00003088-199427060-00006

85. Hilaire ML, Cross LB, Eichner SF. Treatment of migraine headaches with sumatriptan in pregnancy. *Ann Pharmacother*. 2004;38(10):1726-1730. doi:10.1345/aph.1D586

86. Fabijanovic D, Serman A, Jezic M, et al. Impact of 5-azacytidine on rat decidual cell proliferation. *Int J Exp Pathol*. 2014;95(4):238-243. doi:10.1111/iep.12088