New Journal and we have not received input yet 21 (2022) 100211

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.sciencedirect.com/journal/computational-toxicology

Quantitative Structure-Activity Relationship (QSAR) modeling to predict the transfer of environmental chemicals across the placenta

Laura Lévêque [a,b,1], Nadia Tahiri [a,b,1], Michael-Rock Goldsmith [c], Marc-André Verner [a,b,*]

[a] Centre de recherche en santé publique, Université de Montréal and CIUSSS Du Centre-Sud-de-l'île-de-Montréal, Pavillon 7101 avenue du Parc, C.P. 6128, Succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada
[b] Department of Occupational and Environmental Health, School of Public Health, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada
[c] Congruence Therapeutics, 1011 South Hamilton Road, Suite 124, Chapel Hill, NC 27517, USA

ARTICLE INFO

Keywords:
QSAR modeling
Environmental contaminants
Placental transfer
Risk assessment

ABSTRACT

The increasing diversity of environmental chemicals in the environment, some of which may be developmental toxicants, is a public health concern. The aim of this work was to contribute to the development of rapid and effective methods to assess prenatal exposure. Quantitative structure–activity relationships (QSAR) modeling has emerged as a promising method in the development of a predictive model for the placental transfer of contaminants. Cord to maternal plasma or serum concentration ratios for 105 chemicals were extracted from the literature, and 214 molecular descriptors were generated for each of these chemicals. Ten predictive models were built using Molecular Operating Environment (MOE) software, and the Python and R programming languages. Training and test datasets were used, respectively, to build and validate the models. The Applicability Domain Tool v1.0 was used to determine the applicability domain. Models developed with the partial least squares regression method in MOE and SuperLearner in R showed the best precision and predictivity, with internal coefficients of determination (R^2) of 0.88 and 0.82, cross-validated R^2s of 0.72 and 0.57, and external R^2s of 0.73 and 0.74, respectively. All test chemicals were within the domain of applicability. The results obtained in this study suggest that QSAR modeling can help estimate the placental transfer of environmental chemicals.

1. Introduction

Epidemiological and toxicological studies have demonstrated that several contaminants are able to cross the placental barrier and reach the developing organism [2], potentially leading to adverse health effects. However, assessing the health risks of fetal exposure to environmental chemicals is challenging considering the wide range of chemicals currently on the market and new chemicals entering every year [19] (Krimsky, 2017).

Multiple experimental approaches have been used to characterize the ability of chemicals to cross the placenta. Studies have used maternal and cord blood samples collected at or around delivery from volunteers to determine placental transfer, namely through the calculation of cord-to-mother serum or plasma concentration ratios (which we will refer to as CS:MS concentration ratios in this manuscript) [2]. Also, toxicological experiments have been conducted in animals to study the characteristics of *in vivo* placental transfer for toxic compounds and pharmaceuticals. Non-human primates, and animals such as sheep, guinea pigs, and rodents have been used due to either their placentation similarities to the human placenta, or the ability to more easily and rapidly study their pregnancy events [14]. *Ex vivo* placental perfusion models have been used to study the transfer of chemicals between the maternal and fetal compartments using real and intact tissues. Those models, combined with *in vitro* studies on placental cell lines and isolated trophoblastic cells, offer an effective way to study the physiology of the placenta organ and the metabolism of xenobiotics [27]. Although all the approaches presented above provide information on placental transfer, they all have limitations including duration of experiments, interspecies extrapolation, and costs. The increasing diversity of chemicals in the environment demands the development of new ways to quantify fetal exposure rapidly and precisely to such compounds, and their ability to cross the

* Corresponding author at: Department of Occupational and Environmental Health, School of Public Health, Université de Montréal, 2375 chemin de la Côte-Sainte-Catherine, office 4105, Montreal, QC H3T 1A8, Canada.
E-mail address: marc-andre.verner.1@umontreal.ca (M.-A. Verner).
[1] Contributed equally.

https://doi.org/10.1016/j.comtox.2021.100211
Received 13 July 2021; Received in revised form 1 November 2021; Accepted 15 December 2021
Available online 20 December 2021
2468-1113/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Nomenclature

Abbreviations

AD	Applicability Domain
AFP	Alpha FetoProtein
AHA	AracHidonic Acid
CS:MS	CCord-to-maternal serum concentration ratio
DHA	DocosaHexaneoic Acid
GA	Genetic Algorithm
GA-MLR	Genetic Algorithm-Multiple Linear Regression
HSA	Human Serum Albumin
LFA	Long-chain Fatty Acid
LOO	Leave-One-Out
LOO-CV	Leave-One-Out-Cross-Validation
MAE	Mean Absolute Error
MLR	Multiple Linear Regression
MOE	Molecular Operating Environment
OCP	Organochlorine Pesticides
OECD	Organization for Economic Co-operation and Development
OH-PBDE	Hydroxylated PolyBrominated Diphenyl Ether
OH-PCB	Hydroxylated PolyChlorinated Biphenyl
PAH	Polycyclic Aromatic Hydrocarbon
PBB	PolyBrominated Biphenyl
PBDE	PolyBrominated Diphenyl Ether
PCA	Principal Component Analysis
PCB	PolyChlorinated Biphenyl
PCDD	PolyChlorinated Dibenzo-p-Dioxin
PCDF	PolyChlorinated DibenzoFuran
PCR	Principal Component Regression
PFAS	PolyFluoroAlkyl Substance
PLS	Partial Least Squares
PUFA	Poly-Unsaturated Fatty Acid
QSAR	Quantitative Structure-Activity Relationship
R^2	Coefficient of determination
R^2_{cv}	Coefficient of determination - cross-validation
R^2_{ext}	Coefficient of determination - external dataset
RF	Random Forest
RMSE	Root Mean Square Error
$RMSE_{cv}$	Root Mean Square Error - cross-validation
$RMSE_{ext}$	Root Mean Square Error - external dataset
SMILES	Simplified Molecular-Input Line-Entry System
SVM	Support Vector Machine
SVR	Support Vector Regression

placental barrier.

In silico predictive toxicology appears to be a promising alternative to the experimental approaches described above; new approach methodologies (NAMs) have been gaining traction for the assessment of environmental chemicals by regulatory agencies like the U. S. Environmental Protection Agency [36]. Among the major approaches, quantitative structure–activity relationships (QSAR) modeling emerges as a useful tool for the prediction of the biological activity or property of a compound by providing a mathematical association with its structural features [39]. QSAR models have been widely used in the fields of drug design and environmental toxicology and have become central for the molecular interpretation of biological properties. The biological activity of a compound can be described by spatial, hydrophobic, electronic, and steric parameters, as well as quantum chemistry, encoded into a set of descriptors. The molecular descriptors required for QSAR can be obtained by experiment (i.e., empirically derived) or calculated by relevant software packages and platforms, such as SwissADME web tool [6], Molecular Operating Environment [24], PaDEL-descriptor [50], and E-Dragon [23] based on Simplified molecular-input line-entry system (SMILES) [1,45,46].

The prediction of placental transfer of pharmaceuticals and environmental contaminants using QSAR modeling has been investigated in few published studies, using *in vivo* or *ex vivo* transplacental transfer data as training/testing data. Hewitt et al. [15] developed a QSAR model based on the clearance index of 86 heterogeneous pharmaceuticals compiled from literature. *Ex vivo* data were used by Giaginis et al. [12] to build a QSAR model for pharmaceuticals transport across the placental barrier, with 88 clearance indices found in the literature. The same database was used later by Zhang et al. [53] to develop a predictive model using a partial least squares regression (PLS) procedure [38]. Takaku et al. [35] created a multiple linear regression (MLR) model [4] based on 55 CS:MS concentration ratios compiled or calculated from published studies, for a variety of pharmaceuticals and a few organochlorine pesticides . Later, Wang et al. [44] used the same database to build predictive multiple linear regression models, improving the general predictive ability of the model with a four-step feature selection method. Eguchi et al. [7] proposed the first QSAR model developed exclusively with environmental contaminants, using 31 fetal-maternal concentrations ratios of polychlorinated biphenyls (PCBs), polybrominated diphenyl ethers (PBDEs), organochlorine pesticides, and dioxin-like compounds for the prediction of placental transfer. These QSAR modeling efforts allowed identifying some descriptors that are associated with placental transfer, including polarity, lipophilicity and molecular weight. Although these QSAR models have been helpful to quantify and characterize the molecular underpinnings of placental transfer, they were either developed using relatively small datasets or mostly included pharmaceuticals. Therefore, the applicability of these models to predict placental transfer for the diverse range of environmental contaminants is questionable. Also, many of these QSAR models were developed using data from placenta tissue perfused *ex vivo*, which may not fully recapitulate extra-placental factors like protein binding and clearance in the fetus. The development of a QSAR model with a larger environmental chemical database using *in vivo* measurements (i.e., concentrations in maternal and cord blood) could extend this domain of applicability and provide a more suitable model for risk assessment of potential developmental toxicants.

In the current study, our objectives were to create a database of CS:MS concentration ratios for a variety of environmental chemicals based on published articles, and to develop a predictive QSAR model.

2. Methods

2.1. Data curation and annotation

A literature review was conducted to identify studies on fetal exposure to contaminants and placental transfer rate of toxic compounds. Four scientific databases including PubMed, ScienceDirect, Scopus and Google Scholar, were searched with the following keywords: “fetus”, “fetal exposure”, “toxic compound”, “placental transfer”, “environmental contaminants”, “pesticides”, “maternal-fetal exposure”, “fetal blood”, and ”cord blood”. The search was limited to publications written in French or English, and to studies conducted in humans. A database was created compiling CS:MS concentration ratios of environmental contaminants from published epidemiologic/biomonitoring studies ranging from 2002 to 2020. Chemicals included 18 pesticides, 2 hydroxylated polybrominated diphenyl ethers (OH-PBDEs), 4 polybrominated diphenyl ethers (PBDEs), 19 polycyclic aromatic hydrocarbons (PAHs), 1 polybrominated biphenyl (PBB), 6 hydroxylated polychlorinated biphenyls (OH-PCBs), 24 polychlorinated biphenyls (PCBs), 4 polychlorinated dibenzo-p-dioxins (PCDDs), 7

polychlorinated dibenzofurans (PCDFs), and 17 per- and polyfluoroalkyl substances (PFAS). In addition, ratios of 4 combined measures of PCB congeners (TriCBs, TetraCBs, OctaCBs, NonaCBs) were included to represent PCBs that were measured but not reported individually. Similar to the approach presented in Eguchi et al. [7], we used specific isomers to represent these groups: 2,3,4-trichlorobiphenyl (TriCBs), 2,4,4',5-tetrachlorobiphenyl (TetraCBs), 2,2',3,3',4,4',5,5'-octachlorobiphenyl (OctaCBs), and 2,2',3,3',4,4',5,5',6-nonachlorobiphenyl (NonaCBs). All CS:MS concentration ratios have been calculated using maternal and cord blood samples from mother-infant pairs collected at or around delivery and expressed on a wet weight basis. Exceptions were made for the compounds BDE154, BDE209, and PCB163, for which maternal blood was collected during the first trimester, but the concentrations of these compounds have been shown to be relatively stable throughout pregnancy [43]. When median or mean CS:MS concentration ratios were reported, these values were used (87 chemicals). Where CS:MS concentration ratios were not reported in the publication, but mean or median maternal and cord serum levels were, we used these values to calculate ratios (11 chemicals). Finally, when CS:MS concentration ratios were not reported in the publication but the raw data on maternal and cord serum levels were provided, we used these concentrations to calculate median ratios (7 chemicals). If more than one CS:MS concentration ratio was available for a given chemical (e.g., two studies reporting a CS:MS ratio), priority was set on the median ratio over the mean ratio (to reduce the potential influence of outliers), then on the number of mother-infant pairs when several median ratios were available. Chemicals included in the database are shown in Table 1. CS:MS concentration ratios were ln-transformed prior to model development and testing. Chemicals were represented using their SMILES canonicalized structures, retrieved either from the publications or interest or from PubChem using their chemical name or CASRN.

2.2. Descriptors calculation

The Molecular Operating Environment (MOE) software version 2019–0102 (Chemical Computing Group) was used to generate molecular descriptors of chemicals. For ionizable molecules, we used their dominant form at the placental pH, i.e., 7.3. Energy minimization was then applied to all compounds to generate the 3D structure from the 2D SMILES. Molecular descriptors calculated included 193 2D descriptors, (i.e., topological properties), and 117 i3D descriptors (3D descriptors that are internal coordinate dependent). The statistical application "QuaSAR-Contingency module" implemented in MOE was subsequently applied to the 310 descriptors that had been calculated to select the optimal descriptors for the QSAR model development; the contingency module is a bivariate analysis which provided the optimal selection of 2D and 3D descriptors to find the independent variables that most significantly correlated to the dependent variable and important to the model's development. A total of 214 2D and 3D descriptors were retained based on the calculations of four contingency scores (contingency coefficient, Cramer's V, entropic uncertainty, and linear correlation).

2.3. Model development

The dataset of 105 chemicals was separated into training (80%; 84 chemicals) and test sets (20%; 21 chemicals) using the diverse subset method based on the calculated descriptors implemented in MOE. All 214 descriptors were used to calculate the Euclidean distance within each cluster and gave the 84 most diverse chemicals corresponding to the training set. Statistical analyses were performed with three different tools: MOE, Python programming language version 3.7 (Oliphant) [28], and RStudio software version 1.3.887 (RStudio Team) [32] (Fig. 1).

2.4. AutoQSAR in molecular Operating Environment (MOE)

Three models were built using MOE's AutoQSAR & QSAR-Evolution Scientific Vector Language (SVL) codes [18]; (1) Partial Least Square (PLS), (2) Genetic Algorithm - Multiple Linear Regression (GA-MLR), and (3) Principal Component Regression (PCR) analysis. PLS is a method frequently used in QSAR modelling to predict a variable explained by large numbers of factors. The algorithm searches for manifest factors as well as latent factors, the latter being mostly responsible for variations of the predicted variable [38]. Although models developed with PLS have a high predictive power, their mechanistic and biological interpretations can be hard to understand [33]. QSAR studies have demonstrated that with a large number of descriptors compared to the observations, overfitting can occur and substantially reduce the predictive power of the model. A powerful hybrid method combining genetic algorithms (GA) and multiple linear regression (MLR) has been proposed to solve variable selection and facilitate the resolution of optimization problems [33]. Mimicking the natural evolution concept, the GA-MLR method searches for approximate solutions by genetic operations (mutation, selection, crossover), accelerating and facilitating the optimization process [17,21]. The third model was developed with PCR analysis, a method based on principal component analysis (PCA) that aims to minimize the correlation between variables by reducing the size of the dataset, and to maximize the variance.

2.5. Exploring QSAR parametrization with a variety of machine learning approaches implemented in Python

The dataset also was subjected to six statistical models implemented within the Python language version 3.7: Support Vector Regression (SVR), Random Forest (RF), Ridge Regression (RR), Lasso regression (LR), ElasticNet (EN), and PLS.

Support Vector Regression (SVR) is an approach derived from the Support Vector Machine (SVM) developed by Vapnik [41]. SVR regression is based on kernel functions (e.g., linear, polynomial, Gaussian - rbf, sigmoid) thus belonging to nonparametric techniques. Learning the hyperplane in SVR is done by transforming the problem using linear algebra to minimize the best line within a threshold of values (i.e., epsilon-insensitive tube). The points of error inside the tube are qualified as the least important. However, points outside the tube count towards penalties. For all parameters, the default values have been used.

The aim of the random forest (RF) algorithm [3] is to retain most of the strengths of decision trees while eliminating their drawbacks, in particular their vulnerability to overfitting and the complexity of pruning operations. It is a non-parametric regression algorithm which is proving to be very flexible and robust. To avoid overfitting, a few parameters have been adjusted. The number of trees in the forest was fixed at 1000, the maximum depth of the tree was set at 10, and the function to measure the quality of a split was calibrated to be mean absolute error (MAE).

The Ridge Regression [16,22] is a technique for analyzing multiple regression variables that suffer from multicollinearity. In the case of multicollinearity, the estimates of least squares are not biased but have large variances. By adding a degree of bias to the regression estimates, the Ridge regression creates a net effect that reduces the standard errors in order to give more reliable estimates. The alpha value was increased to 200 to reduce the variance of the estimates and specify stronger regularization.

Lasso Regression or least absolute shrinkage and selection operator [37] is a technique that uses descriptors and regularization process to select the most accurate and interpretable model. The parameter alpha value was set to 0.005.

The Elastic Net [55] is a regularized regression method that linearly combines the penalties of the lasso and ridge methods. The alpha value was set here to 1e-7.

Table 1
Dataset used to develop and validate QSAR models of cord-to-maternal serum (CS:MS) concentration ratios.

SMILES	DTXSIDs	Compound name	CS:MS ratio	ln(CS:MS ratio)	Reference
	Pesticides				
C1 = CC(=CC = C1C(=C(Cl)Cl)C2 = CC = C(C = C2)Cl)Cl	DTXSID9020374	4,4′-DDE	0.4	−0.92	[25]
C1(C(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl	DTXSID7020687	b-HCH	0.3	−1.20	[25]
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	DTXSID2020682	HCB	0.6	−0.51	[25]
C12C(C(C3(C1O3)Cl)Cl)C4(C(=C(C2(C4(Cl)Cl)Cl)Cl)Cl)Cl	DTXSID9044166	Oxychlordane	0.1	−2.30	[25]
C12C(C(C(C1Cl)Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	DTXSID8042214	Transnonachlor	0.3	−1.20	[25]
C1 = CC = C(C(=C1)C(C2 = CC = C(C = C2)Cl)C(Cl)Cl)Cl	DTXSID4020373	o,p'-DDD	0.34	−1.08	[51]
C1 = CC = C(C(=C1)C(=C(Cl)Cl)C2 = CC = C(C = C2)Cl)Cl	DTXSID4022313	o,p'-DDE	0.37	−0.99	[51]
C1 = CC = C(C(=C1)C(C2 = CC = C(C = C2)Cl)C(Cl)(Cl)Cl)Cl	DTXSID6022345	o,p'-DDT	0.41	−0.89	[51]
C1 = CC(=CC = C1C(C2 = CC = C(C = C2)Cl)C(Cl)Cl)Cl	DTXSID4020373	p,p'-DDD	0.37	−0.99	[51]
C1 = CC(=CC = C1C(C2 = CC = C(C = C2)Cl)C(Cl)(Cl)Cl)Cl	DTXSID4020375	p,p'-DDT	0.39	−0.94	[51]
CC(C)OC1 = CC = CC = C1O	DTXSID5041431	2-Isopropoxyphenol	1.13	0.12	[47]
CC1(OC2 = C(O1)C(=CC = C2)OC(=O)NC)C	DTXSID9032327	Bendiocarb	0.84	−0.17	[47]
CCOP(=S)(OCC)OC1 = NC(=C(C = C1Cl)Cl)Cl	DTXSID4020458	Chlorpyrifos	0.98	−0.02	[47]
CCOP(=S)(OCC)OC1 = NC(=NC(=C1)C)C(C)C	DTXSID9020407	Diazinon	0.85	−0.16	[47]
C1 = C(C = C(C(=C1Cl)N)Cl)[N +](=O)[O-]	DTXSID2020426	Dicloran	1.06	0.06	[47]
C1 = CC = C2C(=C1)C(=O)NC2 = O	DTXSID3026514	Phthalimide	0.87	−0.14	[47]
C1C = CCC2C1C(=O)NC2 = O	DTXSID4052849	Tetrahydrophthalimide	0.91	−0.09	[47]
	OH-PBDEs				
C1 = CC(=C(C = C1Br)Br)OC2 = C(C = C(C(=C2)Br)O)Br	DTXSID90873927	4′-OH-BDE-49	0.4	−0.92	[25]
C1 = CC(=C(C = C1Br)Br)OC2 = C(C = C(C(=C2)O)Br)Br	DTXSID3030056	5-OH-BDE-47	1.1	0.10	[25]
	OH-PCBs				
C1 = C(C(=CC(=C1Cl)Cl)Cl)C2 = CC(=C(C(=C2Cl)O)Cl)Cl	DTXSID90202637	3-OH-CB153	0.68	−0.39	[29]
C1 = CC(=C(C(=C1C2 = CC(=C(C(=C2Cl)O)Cl)Cl)Cl)Cl)Cl	DTXSID50904125	3′-OH-CB138	1.01	0.01	[29]
C1 = CC(=C(C = C1C2 = CC(=C(C(=C2Cl)Cl)O)Cl)Cl)Cl	DTXSID70165175	4-OH-CB107	0.57	−0.56	[29]
C1 = C(C(=CC(=C1Cl)Cl)Cl)C2 = CC(=C(C(=C2Cl)Cl)O)Cl	DTXSID60163004	4-OH-CB146	0.78	−0.25	[29]
C1 = C(C(=CC(=C1Cl)Cl)Cl)C2 = C(C(=C(C(=C2Cl)Cl)O)Cl)Cl	DTXSID40166371	4-OH-CB187	0.68	−0.39	[29]
C1 = C(C(=C(C(=C1Cl)O)Cl)Cl)C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl	DTXSID80166370	4′-OH-CB172	1.03	0.03	[29]
	PAHs				
CC1 = CC = CC2 = CC = CC = C12	DTXSID9020877	1-MNAP	1.3	0.26	[34]
CC1 = C2C = CC3 = CC = CC = C3C2 = CC = C1	DTXSID6025648	1-MPA	2	0.69	[34]
CC1 = C2C = C(C(=CC2 = CC = C1)C)C	DTXSID8061189	1,6,7-TMNAP	1.5	0.41	[34]
CC1 = CC2 = CC = CC = C2C = C1	DTXSID4020878	2-MNAP	1.4	0.34	[34]
CC1 = CC2 = C(C = C1)C = C(C = C2)C	DTXSID0029187	2,6-DMNAP	2	0.69	[34]
C1 = CC = C(C = C1)C2 = CC = CC = C2	DTXSID4020161	Biphenyl	1.4	0.34	[34]
C1 = CC = C2C(=C1)C3 = CC = CC = C3O2	DTXSID2021993	Dibenzofuran	1.2	0.18	[34]
C1 = CC = C2C(=C1)C3 = CC = CC = C3S2	DTXSID0047741	Dibenzothiophene	1.4	0.34	[34]
C1 = CC = C2C = CC = CC2 = C1	DTXSID8020913	Naphtalene	3.1	1.13	[34]
C1CC2 = CC = CC3 = C2C1 = CC = C3	DTXSID3021774	Acenaphtene	0.26	−1.35	[54]
C1 = CC2 = C3C(=C1)C = CC3 = CC = C2	DTXSID3023845	Acenaphtylene	0.39	−0.94	[54]
C1 = CC = C2C = C3C = CC = CC3 = CC2 = C1	DTXSID0023878	Anthracene	0.42	−0.87	[54]
C1 = CC = C2C(=C1)C = CC3 = CC4 = CC = CC = C4C = C32	DTXSID5023902	Ben(a)anthracene	0.35	−1.05	[54]
C1 = CC = C2C = C3C4 = CC = CC5 = C4C(=CC = C5)C3 = CC2 = C1	DTXSID4075455	Benzo(b + k) fluoranthene	0.4	−0.92	[54]
C1 = CC = C2C(=C1)C = CC3 = C2C = CC4 = CC = CC = C43	DTXSID0022432	Chrysene	0.35	−1.05	[54]
C1 = CC = C2C(=C1)C3 = CC = CC4 = C3C2 = CC = C4	DTXSID3024104	Fluoranthene	0.34	−1.08	[54]
C1C2 = CC = CC = C2C3 = CC = CC = C31	DTXSID8024105	Fluorene	0.29	−1.24	[54]
C1 = CC = C2C(=C1)C = CC3 = CC = CC = C32	DTXSID6024254	Phenanthrene	0.36	−1.02	[54]
C1 = CC2 = C3C(=C1)C = CC4 = CC = CC(=C43)C = C2	DTXSID3024289	Pyrene	0.43	−0.84	[54]
	PBBs				
C1 = C(C(=CC(=C1Br)Br)Br)C2 = CC(=C(C = C2Br)Br)Br	DTXSID70858838	BB153	0.18	−1.71	[11]
	PBDEs				
C1 = CC(=CC = C1OC2 = C(C = C(C = C2)Br)Br)Br	DTXSID4052710	BDE28	0.45	−0.80	[11]
C1 = CC(=C(C = C1Br)Br)OC2 = C(C = C(C = C2Br)Br)Br	DTXSID4052689	BDE100	0.3	−1.20	[25]
C1 = C(C(=CC(=C1Br)Br)Br)OC2 = CC(=C(C = C2Br)Br)Br	DTXSID4030047	BDE153	0.2	−1.61	[25]
C1 = CC(=C(C = C1Br)Br)OC2 = C(C = C(C = C2)Br)Br	DTXSID3030056	BDE47	0.4	−0.92	[25]
C1 = CC(=C(C = C1Br)Br)OC2 = CC(=C(C = C2Br)Br)Br	DTXSID9030048	BDE99	0.3	−1.20	[25]
C1 = C(C = C(C(=C1Br)OC2 = CC(=C(C = C2Br)Br)Br)Br)Br	DTXSID3052692	BDE154	0.46	−0.78	[43]
C1(=C(C(=C(C(=C1Br)Br)Br)Br)Br)OC2 = C(C(=C(C(=C2Br)Br)Br)Br)Br	DTXSID9020376	BDE209	0.8	−0.22	[43]
	PCBs				
C1 = CC(=C(C = C1Cl)Cl)C2 = CC(=C(C = C2Cl)Cl)Cl	DTXSID1073496	PCB99	0.33	−1.12	[5]
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)C2 = C(C(=C(C(=C2Cl)Cl)Cl)Cl)Cl	DTXSID4047541	DecaCBs	0.09	−2.41	[26]
C1 = C(C(=C(C(=C1Cl)Cl)Cl)Cl)C2 = C(C(=C(C(=C2Cl)Cl)Cl)Cl)Cl	DTXSID0022513	NonaCBs	0.11	−2.21	[26]
C1 = C(C(=C(C(=C1Cl)Cl)Cl)Cl)C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl	DTXSID101019427	OctaCBs	0.12	−2.12	[26]
C1 = CC(=CC = C1C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl	DTXSID9074226	PCB114	0.2	−1.61	[26]
C1 = CC(=C(C = C1Cl)Cl)C2 = CC(=C(C(=C2)Cl)Cl)Cl	DTXSID50867160	PCB123	0.22	−1.51	[26]
C1 = CC(=C(C = C1C2 = CC(=C(C(=C2)Cl)Cl)Cl)Cl)Cl	DTXSID3032179	PCB126	0.18	−1.71	[26]
C1 = CC(=C(C = C1C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl)Cl	DTXSID0052706	PCB156	0.16	−1.83	[26]
C1 = CC(=C(C(=C1C2 = CC(=C(C(=C2)Cl)Cl)Cl)Cl)Cl)Cl	DTXSID6074205	PCB157	0.18	−1.71	[26]
C1 = C(C = C(C(=C1Cl)Cl)Cl)C2 = CC(=C(C = C2Cl)Cl)Cl	DTXSID7074165	PCB167	0.18	−1.71	[26]
C1 = C(C = C(C(=C1Cl)Cl)Cl)C2 = CC(=C(C(=C2)Cl)Cl)Cl	DTXSID2038314	PCB169	0.15	−1.90	[26]
C1 = C(C = C(C(=C1Cl)Cl)Cl)C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl	DTXSID4074144	PCB189	0.13	−2.04	[26]
C1 = CC(=C(C = C1C2 = CC(=C(C = C2)Cl)Cl)Cl)Cl	DTXSID5022514	PCB77	0.33	−1.11	[26]

(*continued on next page*)

Table 1 (*continued*)

SMILES	DTXSIDs	Compound name	CS:MS ratio	ln(CS:MS ratio)	Reference
C1 = CC(=CC = C1C2 = CC(=C(C(=C2)Cl)Cl)Cl)Cl	DTXSID6074209	PCB81	0.27	−1.31	[26]
C1 = CC(=CC = C1C2 = CC(=C(C = C2Cl)Cl)Cl)Cl	DTXSID8073473	TetraCBs	0.24	−1.43	[26]
C1 = CC = C(C = C1)C2 = C(C(=C(C = C2)Cl)Cl)Cl	DTXSID0074180	TriCBs	0.59	−0.53	[26]
C1 = CC(=C(C = C1C2 = C(C(=CC(=C2Cl)Cl)Cl)Cl)Cl)Cl	DTXSID8074233	PCB163	0.23	−1.47	[10]
C1 = CC(=C(C = C1C2 = C(C(=C(C = C2)Cl)Cl)Cl)Cl)Cl	DTXSID8038306	PCB105	0.16	−1.83	[29]
C1 = CC(=C(C = C1C2 = CC(=C(C = C2Cl)Cl)Cl)Cl)Cl	DTXSID4032116	PCB118	0.12	−2.12	[29]
C1 = CC(=C(C(=C1C2 = CC(=C(C = C2Cl)Cl)Cl)Cl)Cl)Cl	DTXSID8038300	PCB138	0.21	−1.56	[29]
C1 = C(C(=CC(=C1Cl)Cl)Cl)C2 = CC(=C(C = C2Cl)Cl)Cl	DTXSID2032180	PCB153	0.19	−1.66	[29]
C1 = CC(=C(C(=C1C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl)Cl)Cl	DTXSID2073481	PCB170	0.16	−1.83	[29]
C1 = C(C(=CC(=C1Cl)Cl)Cl)C2 = CC(=C(C(=C2Cl)Cl)Cl)Cl	DTXSID6038299	PCB180	0.18	−1.71	[29]
	PCDDs				
C1 = C2C(=C(C(=C1Cl)Cl)Cl)OC3 = C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	DTXSID1052034	1,2,3,4,6,7,8-HeptaCDD	0.19	−1.66	[26]
C1 = C2C(=CC(=C1Cl)Cl)OC3 = C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	DTXSID8052067	1,2,3,4,7,8-HexaCDD	0.22	−1.51	[26]
C1 = C2C(=CC(=C1Cl)Cl)OC3 = C(C(=C(C = C3O2)Cl)Cl)Cl	DTXSID7052078	1,2,3,7,8-PentaCDD	0.23	−1.47	[26]
C1 = C2C(=CC(=C1Cl)Cl)OC3 = CC(=C(C = C3O2)Cl)Cl	DTXSID2021315	2,3,7,8-TetraCDD	0.28	−1.27	[26]
C12 = C(C(=C(C(=C1Cl)Cl)Cl)Cl)OC3 = C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	DTXSID4025799	OctaCDD	0.11	−2.21	[26]
	PCDFs				
C1 = C2C3 = C(C(=C(C(=C3Cl)Cl)Cl)Cl)OC2 = C(C(=C1Cl)Cl)Cl	DTXSID8052350	1,2,3,4,6,7,8-HeptaCDF	0.33	−1.11	[26]
C1 = C2C(=CC(=C1Cl)Cl)OC3 = C2C(=C(C(=C3Cl)Cl)Cl)Cl	DTXSID6029915	1,2,3,4,7,8-HexaCDF	0.25	−1.39	[26]
C1 = C2C3 = C(C(=C(C = C3OC2 = C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	DTXSID2069155	1,2,3,6,7,8-HexaCDF	0.3	−1.20	[26]
C1 = C2C(=CC(=C1Cl)Cl)OC3 = CC(=C(C(=C23)Cl)Cl)Cl	DTXSID7052234	1,2,3,7,8-PentaCDF	0.33	−1.11	[26]
C1 = C2C3 = CC(=C(C(=C3OC2 = C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	DTXSID3052276	2,3,4,6,7,8-HexaCDF	0.27	−1.31	[26]
C1 = C2C3 = CC(=C(C(=C3OC2 = CC(=C1Cl)Cl)Cl)Cl)Cl	DTXSID7030066	2,3,4,7,8-PentaCDF	0.22	−1.51	[26]
C1 = C2C3 = CC(=C(C = C3OC2 = CC(=C1Cl)Cl)Cl)Cl	DTXSID3052147	2,3,7,8-TetraCDF	0.42	−0.87	[26]
	PCP				
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)O	DTXSID7021106	PCP	1.1	0.10	[29]
	PFAS				
CCN(CC(=O)O)S(=O)(=O)C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	DTXSID60892504	N-EtFOSAA	1.2	0.18	[25]
CN(CC(=O)O)S(=O)(=O)C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	DTXSID10624392	N-MeFOSAA	0.9	−0.11	[25]
C(C(C(C(C(F)(F)S(=O)(=O)N)(F)F)(F)F)(F)F)(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F	DTXSID3038939	PFOSA	1.1	0.10	[25]
C1 = CC(=CC = C1C(=O)O)F	DTXSID4059916	PFBA	1.67	0.51	[20]
C(C(C(C(F)(F)F)(F)F)(F)F)(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(F)F	DTXSID8059920	PFHpS	0.8	−0.22	[20]
C(=O)(C(C(C(C(C(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID3059921	PFTeDA	4	1.39	[20]
C(=O)(C(C(C(C(C(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID90868151	PFTrDA	1.38	0.32	[20]
C(CS(=O)(=O)O)C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	DTXSID6067331	6:2FTS	1.66	0.51	[49]
C(=O)(C(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID3031860	PFDA	0.25	−1.39	[49]
C(=O)(C(C(C(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID8031861	PFDoA	0.51	−0.67	[49]
C(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(C(C(F)(F)F)(F)F)(F)F)(F)F	DTXSID80873012	PFHxS	0.35	−1.05	[49]
C(=O)(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID8031863	PFNA	0.43	−0.84	[49]
C(=O)(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID8031865	PFOA	0.65	−0.43	[49]
C(C(C(C(C(F)(F)S(=O)(=O)O)(F)F)(F)F)(F)F)(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F	DTXSID3031864	PFOS	0.29	−1.24	[49]
C(=O)(C(C(C(C(C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID8047553	PFUnA	0.27	−1.31	[49]
C(=O)(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID1037303	PFHpA	1.20	0.18	[52]
C(=O)(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)O	DTXSID3031862	PFHxA	1.07	0.07	[52]

The selection of attributes was performed with the *SequentialFeatureSelector* method provided with the *scikit-learn package* in Python using version 0.23.2 [9]. The method relies on the coefficient of determination between one subset of attributes and the predictor (ln-transformed CS:MS concentration ratio). The optimal combination of the *k* best descriptors was achieved based on the Sequential Forward Selection (SFS) procedure [30]. This technique involves finding the best model for an attribute, then finding the second best by trying to add any possible descriptor (except the previous one) to the previously selected one, and continuing in this mode until no improvement is achieved in the process. This algorithm, which scales quadratically, is time-consuming but with a large number of descriptors (214 in this study), the choice of this approach turns out to be the best, thus making it possible to reach the target number of descriptors (i.e., less than the number of chemicals divided by 5) the most efficiently.

2.6. SuperLearner with RStudio

Finally, a model was developed using R programming language and the SuperLearner algorithm. SuperLearner is a stacking generalization algorithm that combines predictions from several machine learning techniques to create an accurate and high prediction final model [40]. This stacked generalization method creates an ensemble, i.e., a machine learning model obtained by weighting performances and evaluating contributions of each machine learning technique on the same dataset. Out-of-fold predictions obtained from the *k*-fold cross-validation of each model considered (*k* being the same for all models) are used as input data to develop the SuperLearner model. In this study, the optimal combination was estimated using out-of-fold predictions of six different machine learning models, listed as follows: 1) Random Forest, 2) Ranger (fast implementation of Random Forest), 3) glmnet (Lasso and Elastic-Net Regularized Generalized Linear Models), 4) bartMachine (Support Bayesian additive regression trees), 5) ksvm (Kernlab's support vector machine algorithm) and 6) nnet (Neural network). For all parameters, the default values have been used.

Python and R programs, called QSAR_PTR (QSAR placental transfer ratios), implementing the QSAR prediction algorithms along with the dataset are freely available at: https://github.com/TahiriNadia

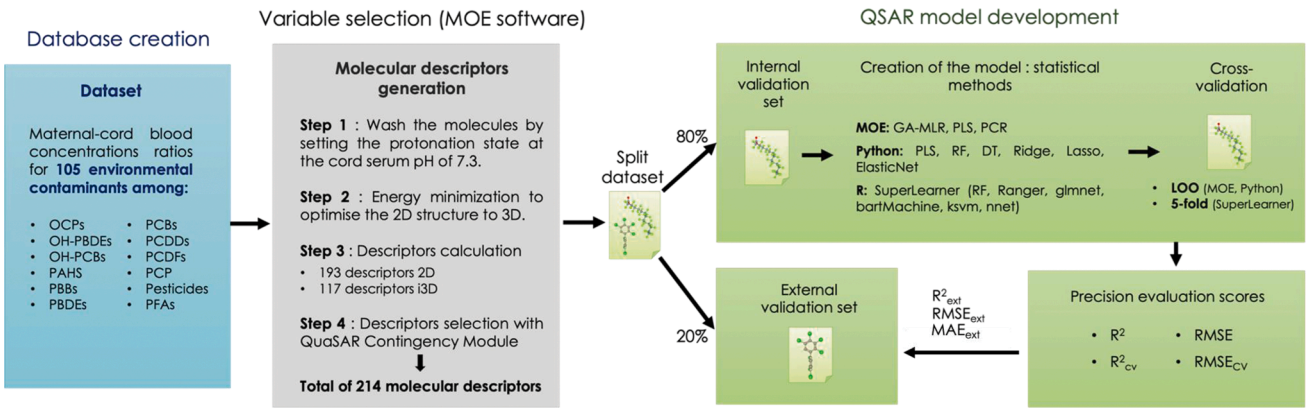

Fig. 1. QSAR model workflow for the prediction of the placental transfer of environmental chemicals. A 105 molecules database of diverse environmental contaminants (blue section) was used to generate a pool of 214 2D and i3D descriptors (gray section). MOE, Python language, and RStudio software were used to develop ten QSAR models based on an 80–20% training-test split with the most diverse subset method implemented in MOE. Internal validation was performed with LOO (MOE and Python) and 5-fold cross-validation (RStudio) techniques, and external validation was performed with the test set (green section). Models goodness-of-fit, robustness and predictivity were assessed with coefficients of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

/QSAR_PTR.

2.7. Validation and applicability domain

Validation is a crucial step for the acceptance of a QSAR model for regulatory purposes. In 2004, the Organization for Economic Co-operation and Development (OECD) established five principles for (Q)SAR validation that any QSAR model should follow: i) a defined endpoint; ii) an unambiguous algorithm; iii) a defined domain of applicability; iv) appropriate measures of goodness-of-fit, robustness and predictivity; v) a mechanistic interpretation, if possible [48]. The formulation of the fourth principle refers to two aspects of the model performance evaluation: an internal one expressed as goodness-of-fit and robustness, and an external one expressed as predictivity.

Internal validation was performed for all MOE and Python models by a Leave-One-Out Cross-Validation (LOO-CV) and assessed by two statistical measures: 1) a coefficient of determination (R^2_{CV}), and 2) root-mean square error ($RMSE_{CV}$). The cross-validation method LOO consists of removing from the training set each molecule once, to predict the activity/property of the molecule left out with a model developed based on the other training molecules. The operation is repeated as many times as there are molecules in the training set and R^2 is determined as the mean of the external coefficient of determination of the *k* models. Internal validation for the SuperLearner model was performed by a 5-fold Cross-validation, a resampling procedure that divides the training set into 5 groups and uses each group once to test the model developed based on the 4 other groups. The two same statistical measures were used to assess the cross-validation performance. Though internal validation techniques help to prevent overfitting in the model, it does not quantify the ability of the model to predict data that was not part of the development dataset. External validation entails determining the predictive power of the model by comparing predicted and observed data for a test set of compounds that were not used in model development. Predictivity of all models was assessed with a coefficient of determination (R^2_{ext}), a relative measure of fit, and the root mean squared error ($RMSE_{ext}$) and mean absolute error (MAE_{ext}), two absolute measures of fit (Fig. 1).

Both model validation, and applicability domain assessment are fundamental QSAR model analyses to instill confidence in a model (i.e. not only give accurate but also reliable predictions). The third principle of the OECD for QSAR modeling recommends that each model have a defined applicability domain (AD), i.e., physical–chemical, descriptor or end-point space or basis of training chemistries where test predictions should fall within [31]. Also known as the interpolation space, the AD sets the boundaries of a model due to its restrictions in terms of types of molecules and molecular features. Determining this interpolation space will then allow the user to estimate the reliability of the predictions for both the training and the test set [31]. A model developed without a defined AD could predict all sorts of chemicals for which the model was not built for and lead to inaccurate extrapolation and predictions [42]. The definition of the AD makes it possible to determine the types of molecules covered by the model based on the chemicals and descriptors used to develop the model. Several statistical methods can be employed to characterize the interpolation space of the model and have been summarized by Roy et al. [31]. In the same article, the authors proposed a new method to determine the AD using the standardization approach. Implemented in Java and available at https://dtclab.webs.com/software-tools as an open access tool "Applicability domain using Standardization approach", the method offers an easy way to detect outliers in the training set, and molecules that fall outside of the AD in the test set. We applied this tool to our dataset using the training-test split and descriptors described above, to determine the applicability domain of all our models.

3. Results

As a first analysis we used MOE to calculate 310 2D and 3D molecular descriptors for all 105 contaminants of the database. A sub-analysis to find the most important descriptors to build the model was done with the QuaSAR contingency module: ratios were used by the model as the activity field to perform a bivariate analysis to the 310 descriptors individually. The four coefficient scores calculated for each descriptor, contingency coefficient (descriptor useful when value was above 0.6), Cramer's V, entropic uncertainty and linear correlation R^2 (a value above 0.2 being useful) gave the optimal 214 final descriptors significantly correlated to the dependent variable and important to do QSAR modelling. Among the 2D descriptors selected: 29 were partial charge descriptors; 4 were pharmacophore feature descriptors; 27 were adjacency and distance matrix descriptors; 15 were Kier & Hall Connectivity and Kappa Shape index; 19 were atom count and bond count descriptors; 10 were subdivided surface area; and 12 were physical properties. Among the i3D descriptors selected: 65 were surface area, volume and shape descriptors, depending on the connectivity and the conformation of the structure; and 14 were conformation dependant charge descriptors.

Overall, we developed ten models with MOE, Python and R programming languages. Statistical results for internal and external validation for all the models are summarized in Table 2. Evaluation scores

Table 2
Evaluation scores of the QSAR models developed with MOE, Python and RStudio, with coefficient of determination (R^2) and root mean square error (RMSE) for the training phase (R^2 and RMSE), cross-validation (R^2_{CV} and $RMSE_{CV}$), and testing phase (R^2_{ext}, $RMSE_{ext}$, and MAE_{ext}).

Models	R^2	RMSE	R^2_{CV}	$RMSE_{CV}$	R^2_{ext}	$RMSE_{ext}$	MAE_{ext}
MOE							
Genetic Algorithm-Multiple Linear Regression	0.76	0.41	0.66	0.49	0.56	0.36	0.28
Partial Least Squares	0.88	0.29	0.72	0.45	0.73	0.32	0.22
Principal Component Regression	0.66	0.49	0.50	0.60	0.38	0.47	0.36
Python Language							
Random Forest	0.96	0.17	0.62	0.38	0.64	0.30	0.23
Lasso Regression	0.61	0.52	0.48	0.61	0.53	0.34	0.22
Partial Least Squares	0.63	0.51	0.5	0.45	0.51	0.35	0.23
Support Vector Regression	0.66	0.49	0.39	0.39	0.39	0.39	0.29
ElasticNet	0.64	0.50	0.50	0.59	0.37	0.4	0.24
Kernel ridge regression	0.82	0.36	0.56	0.54	0.54	0.34	0.24
RStudio							
SuperLearnerCombination of: SL.bartMachine, SL.randomForest, SL.ranger, SL.glmnet, SL.ksvm, SL.nnet	0.82	0.36	0.57	0.55	0.74	0.29	0.20

include the coefficient of determination (R^2) and root mean square error (RMSE) for the internal training set, for the cross-validation procedure and for the external testing set. We also calculated the mean absolute error (MAE) for the external testing set. The next sections present models developed, and results obtained with each method.

3.1. AutoQSAR in molecular operating environment (MOE)

The chemical compounds set was used to develop 3 models using the AutoQSAR Scientific Vector Language (SVL) module developed for MOE software (PLS, GA-MLR, and PCR) (Fig. 2). The first model was developed using partial least squares (PLS) analysis with a selection of 35 descriptors and 19 principal components. The second was elaborated with principal component regression analysis (PCR) and 11 descriptors as well as 11 components were used. Finally, the third model was developed with genetic algorithm-multiple linear regression (GA-MLR) approach and an optimal selection of 10 descriptors. PLS and GA-MLR models provided high robustness and goodness-of-fit performances with good statistical internal validation scores of leave-one-out cross-validation coefficient of determination and root mean square error (See Table 2). The PCR model showed low predictive precision in cross-validation and external validation steps. Among the three models, the model developed using PLS obtained the highest predictive performance in the external validation set ($R^2_{ext} = 0.73$, $RMSE_{ext} = 0.32$, $MAE_{ext} = 0.22$).

3.2. Python language

We developed six models in Python (Random Forest, Lasso Regression, Partial Least Squares, ElasticNet, Kernel Ridge Regression, Support Vector Regression) (Fig. 3). The PLS model provides a point of comparison between two tools (MOE and Python). Table 2 shows the performance of PLS of MOE and Python. The PLS model developed using MOE was more robust in terms of internal performance for cross-validation but degraded more rapidly relative to its Python counterpart in external validation; The decreased performance of the MOE-build AutoQSAR's PLS implementation, when going from internal to external dataset validation, may arise as a result of over-parametrization. Models included 8 (Ridge Regression), 6 (Lasso Regression), 9 (ElasticNet), 10 (Partial Least Squares), 9 (Random Forest) and 9 (Support Vector Regression) descriptors (see supplementary materials S3 and S4). Among the six models, the model developed using Random Forest obtained the highest predictive performance in the external validation test with a $R^2_{ext} = 0.64$ and a $RMSE_{ext} = 0.30$, although Lasso Regression had a slightly lower MAE_{ext} (see Table 2).

3.3. SuperLearner

Finally, a QSAR model was developed using the SuperLearner package in RStudio (Fig. 4). Internal and external validation scores, obtained respectively with a 5-fold cross-validation process and a test phase, are presented in Table 2. The probability predictions of the six selected machine models (SL.randomForest, SL.ranger, SL.glmnet, SL.bartMachine, SL.ksvm, and SL.nnet) are combined by averaging and weighting each model. The model that contributed the most to the model was SL.ksvm with a weight of 0.8, followed by the model SL.glmnet with a weight of 0.1, and finally the other four models contribute with a total weight of 0.1. Of note, the lower cross-validated R^2 observed with the SuperLearner compared to some other algorithms may be related to the cross-validation technique (5-fold cross-validation vs. leave-one-out for other algorithms).

3.4. Applicability domain

The applicability domain was defined with the tool Applicability Domain v1.0 developed by Roy et al., [31]. Results showed one outlier in the training set (PFTeDA, CS:MS ratio = 4), but no observation outside the applicability domain in the test set. When we evaluated absolute and relative errors for the chemicals included in the test set, we did not observe a pattern of larger errors for any particular class of chemicals.

4. Discussion

A total of ten QSAR models were developed with three different tools including the Molecular Operating Environment (MOE) software, the Python programming language, and RStudio. All models have been evaluated internally through a cross-validation process, and externally through a test phase. Best models for each tool have been selected based on goodness-of-fit, robustness and predictivity performances. Two statistical measures were used to measure precision and deviation of model's predictions to the actual data set, including the coefficient of determination (R^2) and the root mean square error (RMSE). The partial least squares analysis was developed with 35 descriptors and gave the best results for MOE with the smallest RMSE and the best R^2 for the external validation. Among the six models developed with the Python programming language, the ElasticNet and the Random Forest models showed the best evaluation scores for external validation with similar R^2s and RMSEs. Finally, the SuperLearner developed in RStudio showed good results during external validation, with a high R^2 ($R^2_{ext} = 0.74$) and a low RMSE ($RMSE_{ext} = 0.29$). The applicability domain showed that all test compounds were included in the interpolation space.

Among the four models with best external validation performances selected in this study, two were developed using partial least squares and random forest analysis. The same statistical approaches were used in the study by Eguchi et al. [7]. In their study, the partial least squares and the random forest models yielded lower coefficients of determination for external validation ($R^2_{ext} = 0.123$ and $R^2_{ext} = 0.519$, respectively). Additionally, a third model was developed in the same study using

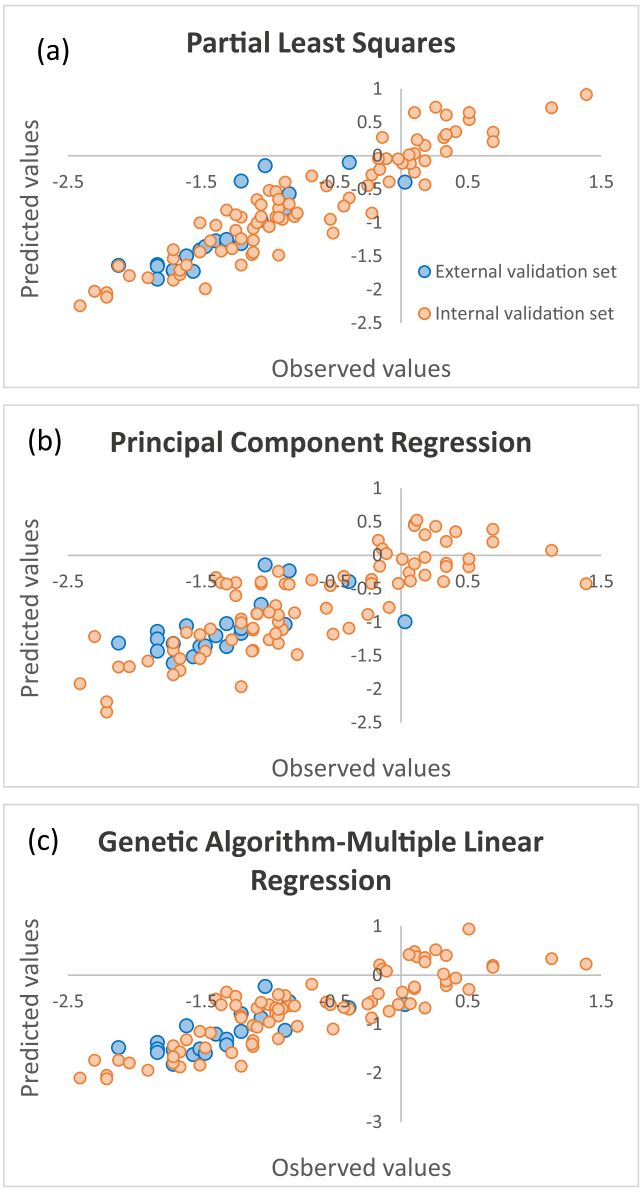

Fig. 2. Comparison of predicted and observed ratios for the internal validation set (orange circles) and the external validation set (blue circles) using the MOE software. Three models have been developed. The three models were tested and based on: (a) Partial Least Squares, (b) Principal Component Regression, and (c), Genetic Algorithm-Multiple Linear Regression. Of note, CS:MS concentration ratios are ln-transformed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

multiple linear regression. Results showed a low external precision (R^2_{ext} = 0.129) in comparison to the multiple linear regression combined with a genetic algorithm developed in our study with MOE (R^2_{ext} = 0.56). The three models elaborated by Eguchi et al. [7] are the only ones that used exclusively environmental contaminants. Their models were developed and tested with PBDE, OCP, PCBs and dioxin-like compounds. However, the dataset included only 31 compounds of which 24 (80%) were used to train the models, and 7 (20%) were used to test the predictive power of the models; Tropsha [39] recommends a minimum of 10 observations in the test set for continuous response variables. A principle of a good QSAR model validation is the adequate ratio between the descriptors and the molecules used to train the model. The poor predictive power observed for the PLS, the RF and the MLR analysis performed by Eguchi et al. [7] could be explained by the high number of descriptors (10) used in the model development, and the small number of observations included in the training set (24), whereas the rule of thumb recommends five chemicals for one descriptor ratio [13]. That being said, we did not develop models using only the same 31 chemicals as Eguchi et al. [7], so we cannot rule out that parameters other than the dataset (e.g., algorithms, selection of training/test sets) could explain, at least partially, the discrepancy between models.

All models developed in this study were elaborated in accordance with the OECD principles, providing transparency of data, an unambiguous algorithm, a specified endpoint, good measures of statistical internal and external validation to assess the performance of the model, a defined applicability domain. The fifth principle mechanistic interpretation refers to "the assignment of physical/chemical/biological meaning to the descriptors after modelling (*a posteriori*)" [48]. Most relevant descriptors on which the models are based should provide insight into the correlation between the chemical structure and their activity or biological properties. A selection of the most important descriptors based on the initial pool of 214 descriptors was performed for the PLS, the GA-MLR and the PCR analysis in MOE, and those models selected 35, 10 and 11 descriptors, respectively. Equations and relative importance of selected descriptors for MOE models are available in the Supplemental material section (Table S1 and S2, respectively). Whereas the PLS model showed the best performance during external validation in MOE, the large number of descriptors make interpretation difficult. Perhaps one of the reasons for the larger number of parameters in the PLS mode is not adhering to the 5 molecules/descriptor paradigm, resulting in over-parameterizing. On the other hand, the smaller set of descriptors selected with the GA-MLR approach (10) provides information on the amplitude and direction of the influence of molecular descriptors on placental transfer. Model parameters included in the equation were related to hydrophobicity (e.g., accessible hydrophobic surface area, hydrophobic volume), refractivity (e.g., molar refractivity), connectivity (e.g., connectivity of atoms on their contributions to logP and molar refractivity), molecular size (e.g., vertex adjacency information in terms of magnitude) and atomic charge (e.g., relative negative partial charge, partial charge based on Van der Waals surface area of atoms). Relative importance of descriptors of the GA-MLR model indicated that molar refractivity, hydrophobic volume, and water accessible surface area are the most relevant variables to explain the transplacental transfer rate. The equation of the model showed a negative contribution of the aqueous surface area and of the hydrophobic volume, but a positive contribution of the molar refractivity. The more hydrophobic a chemical becomes (both surface area and volume) and the less diffuse it becomes (polarizability related to molar refractivity) a likely reduction in efficiently crossing the placental barrier will result. Conversely, molecules with lower hydrophobicity (both surface area and volume) and greater polarizability should cross the barrier more efficiently. Furthermore, a molecule with high molar refractivity, i.e., a higher polarizability, or characteristic capability of a molecule's electronic system to be distorted by an external field, is more likely to penetrate the fetal unit. Depending on the chemical, transfer from maternal circulation to fetal circulation through the placenta can occur through diffusion or protein-mediated facilitated and active transport. Multiple factors can influence placental transfer, including differential lipid and water blood composition, ability to cross bilipid layers, and affinity for cell wall binding proteins [8]. Additionally, the predominant plasma proteins on the fetal versus maternal compartment, alpha fetoprotein (AFP) and human serum albumin (HSA) differ in their payload and affinity for many chemicals, the former having a higher surface charge and a specific affinity for a variety of developmental poly-unsaturated fatty acids (PUFAs) required for neural development, whereas HSA carries small molecules and many saturated long-chain fatty acids (LFAs). One should note that this conceptually also agrees with the fact that many developmental poly-unsaturated fatty acids (such as docosahexaneoic acid (DHA) or arachidonic acid (AHA) with molar refractivity of 10.5 and 9.6 respectively) have a higher affinity for AFP and are also more polarizable than their saturated counterparts

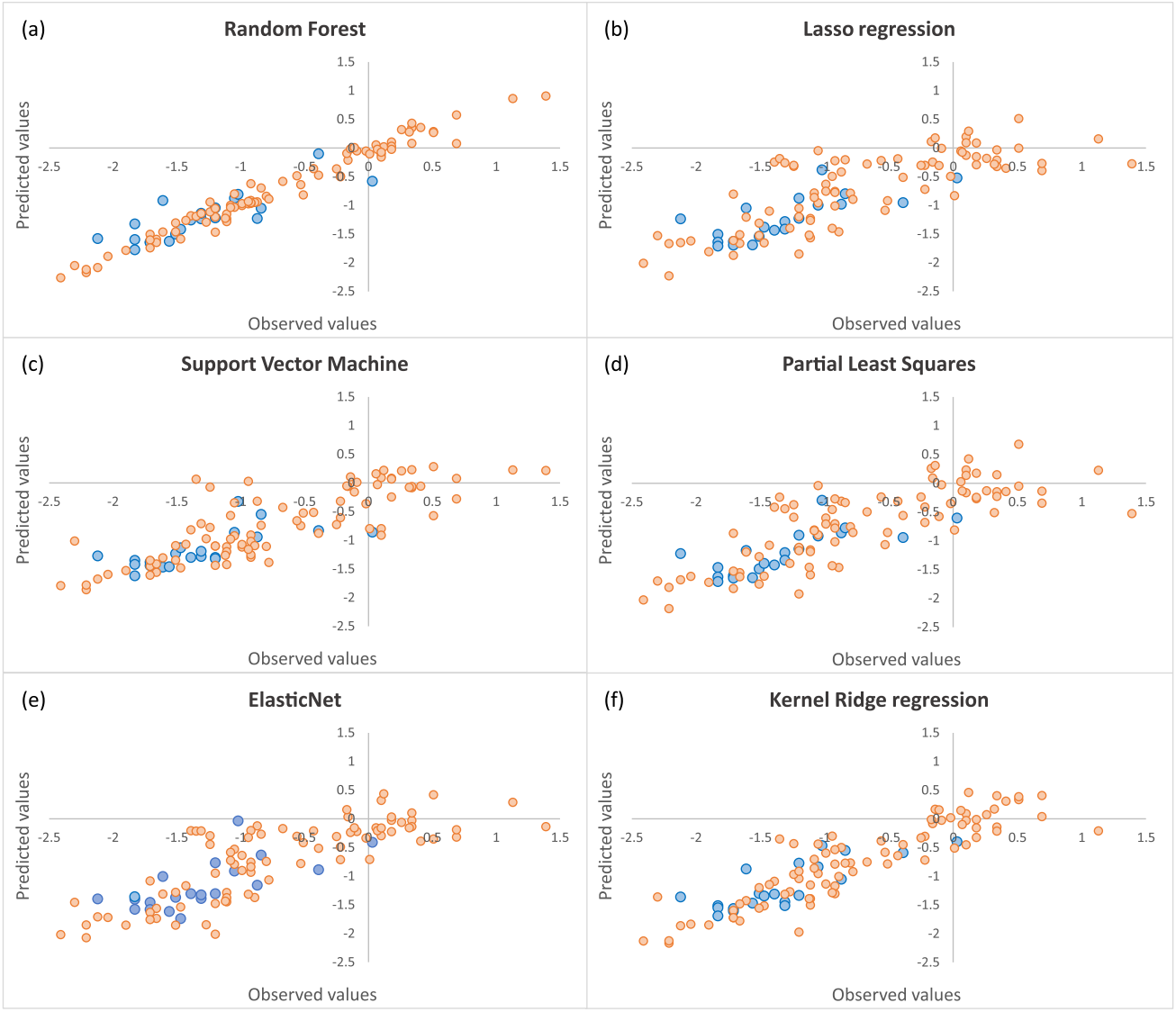

Fig. 3. Comparison of predicted and observed ratios for the internal validation set (orange circles) and the external validation set (blue circles) using the Python programming language and the scikit-learn library. Three models have been developed. The results of the Random Forest model are indicated in (a), Lasso Regression in (b), Support Vector Regression in (c), Partial Least Squares (d), ElasticNet in (e), and Kernel Ridge Regression (f). Of note, CS:MS concentration ratios are ln-transformed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

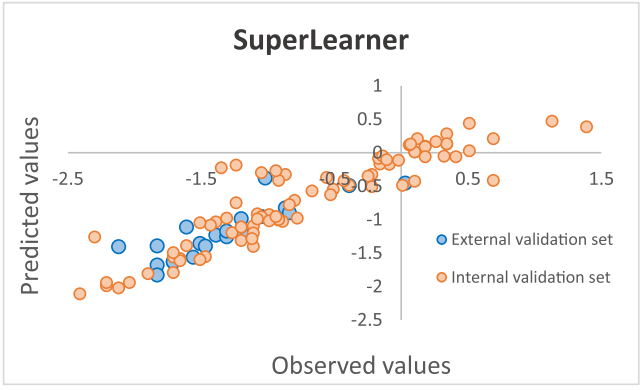

Fig. 4. Comparison of predicted and observed ratios for the internal validation set (orange circles) and the external validation set (blue circles) using the SuperLearner package in RStudio. Of note, CS:MS concentration ratios are ln-transformed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

typically bound to HSA (for instance myristic acid with molar refractivity ~ 7).

Our work has some limitations that need to be discussed. First, our dataset, although it is the largest to date on environmental chemicals, is still relatively small for QSAR model development as it does not allow further separation for the determination of hyperparameters. Determination of hyperparameters (e.g., number of folds in cross-validation, percentage of data used in training/test sets) in a subset of data would have allowed us to optimize QSAR model training, and possibly increase performance. Also, included chemicals were mostly persistent organic pollutants, many within-class analogs that were not particularly diverse, with few chemicals with a shorter biological half-life (which may have different chemical properties). Consequently, the applicability of our QSAR model for non-persistent chemicals should be considered as uncertain. That being said, our model had a better performance and a wider domain of applicability than that reported by Eguchi et al. [7], the only other QSAR effort using only environmental chemicals that we could locate. Another important limitation is the uncertainty and interindividual variability underpinning the concentration ratios that were used for model development and testing. Where more than one

study reported concentration ratios, the values differed (e.g., four studies reported perfluorohexanesulfonate CS:MS concentration ratios of 0.23, 0.29, 0.35 and 0.67), indicating that the calculated ratios can vary for studies in different populations or using different methods for chemical analyses. In studies where measurements were available to calculate CS:MS concentration ratios for all study participants, the values varied substantially from one individual to another; the coefficients of variation for CS:MS concentration ratios calculated using data from Yin et al. [51] on 10 pesticides ranged from 0.29 to 0.59, with *p,p'*-DDE ratios ranging from 0.17 to 0.70. This between-study and between-person variability clearly underline that QSAR predictions are central tendency values and may poorly predict placental transfer in a given individual. Another limitation is that we used the MOE software to generate 2D and 3D descriptors, some of which may not be generated with open platforms: costs associated with software licence may limit future applicability. Finally, because we were interested in the placental transfer of environmental chemicals, we did not include pharmaceuticals in the training and testing sets. Inclusion of pharmaceuticals in subsequent modeling efforts could allow expanding the domain of applicability; where global QSAR models are unable to predict placental transfer adequately, local models could be developed to accommodate specific transfer processes (e.g., transporters). The limitations highlighted in this section likely impacted model performance, domain of applicability, and potential for model application by other researchers.

In conclusion, our study showed that QSAR modeling can be used to estimate fetal plasma concentrations based on maternal plasma concentrations during pregnancy. Models developed herein could be used to parameterize pharmacokinetic models of pregnancy for data-poor chemicals and allow for high-throughput evaluation of fetal exposure, although the predicted CS:MS concentration ratios are estimated at birth and may not adequately represent placental transfer and distribution to the fetus throughout pregnancy. The models could also be used to support the screening/risk assessment of developmental toxicants. Future work could be undertaken to expand the dataset to widen the domain of applicability, and possibly increase the predictivity of QSAR models of placental transfer.

CRediT authorship contribution statement

Laura Lévêque: Conceptualization, Data curation, Methodology, Writing – original draft. **Nadia Tahiri:** Conceptualization, Data curation, Methodology, Writing – original draft. **Michael-Rock Goldsmith:** Conceptualization, Data curation, Methodology, Writing – review & editing. **Marc-André Verner:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors would like to thank Lesa Aylward for providing compiled data on CS:MS concentration ratios. Nadia Tahiri is supported by a postdoctoral training award from the Fonds de recherche en santé – Québec (FRQS). Marc-André Verner is supported by a Research Scholar J2 Award from the FRQS. This study was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2016-06101). Cette étude a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) (RGPIN-2016-06101).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.comtox.2021.100211.

References

[1] E. Anderson, G.D. Veith, D. Weininger, SMILES: A line notation and computerized interpreter for chemical structures. Report No. EPA/600/M-87/021. U.S. Environmental Protection Agency, Environmental Research Laboratory-Duluth, Duluth, MN 55804, 1987.
[2] L.L. Aylward, S.M. Hays, C.R. Kirman, S.A. Marchitti, J.F. Kenneke, C. English, D. R. Mattison, R.A. Becker, Relationships of chemical concentrations in maternal and cord blood: a review of available data, J. Toxicol. Environ. Health, Part B 17 (3) (2014) 175–203.
[3] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[4] A. Chakraborty, D. Goswami, Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN), Arabian J. Geosci. 10 (17) (2017) 1–11.
[5] A. Covaci, P. Jorens, Y. Jacquemyn, P. Schepens, Distribution of PCBs and organochlorine pesticides in umbilical cord and maternal serum, Sci. Total Environ. 298 (1-3) (2002) 45–53.
[6] A. Daina, O. Michielin, V. Zoete, SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules, Sci. Rep. 7 (1) (2017) 1–13.
[7] A. Eguchi, M. Hanazato, N. Suzuki, Y. Matsuno, E. Todaka, C. Mori, Maternal–fetal transfer rates of PCBs, OCPs, PBDEs, and dioxin-like compounds predicted through quantitative structure–activity relationship modeling, Environ. Sci. Pollut. Res. 25 (8) (2018) 7212–7222.
[8] M. Feghali, R. Venkataramanan, S. Caritis, Pharmacokinetics of drugs in pregnancy. In Seminars in perinatology (Vol. 39, No. 7, pp. 512-519). WB Saunders, 2015, November.
[9] M. Feurer, A. Klein, K. Eggensperger, J.T. Springenberg, M. Blum, F. Hutter, Auto-sklearn: efficient and robust automated machine learning. In Automated Machine Learning, Springer, Cham, 2019, pp. 113–134.
[10] M. Fisher, T.E. Arbuckle, C.L. Liang, A. LeBlanc, E. Gaudreau, W.G. Foster, D. Haines, K. Davis, W.D. Fraser, Concentrations of persistent organic pollutants in maternal and cord blood from the maternal-infant research on environmental chemicals (MIREC) cohort study, Environ. Health 15 (1) (2016), https://doi.org/10.1186/s12940-016-0143-y.
[11] M. Frederiksen, C. Thomsen, M. Frøshaug, K. Vorkamp, M. Thomsen, G. Becher, L. E. Knudsen, Polybrominated diphenyl ethers in paired samples of maternal and umbilical cord blood plasma and associations with house dust in a Danish cohort, Int. J. Hyg. Environ. Health 213 (4) (2010) 233–242.
[12] C. Giaginis, A. Zira, S. Theocharis, A. Tsantili-Kakoulidou, Application of quantitative structure–activity relationships for modeling drug and chemical transport across the human placenta barrier: a multivariate data analysis approach, J. Appl. Toxicol. Int. J. 29 (8) (2009) 724–733.
[13] P. Gramatica, On the development and validation of QSAR models. Methods Mol. Biol. (CliftonN.J.) 930 (2013) 499–526, https://doi.org/10.1007/978-1-62703-059-5_21.
[14] P.L. Grigsby, Animal models to study placental development and function throughout normal and dysfunctional human pregnancy. In Seminars in reproductive medicine (Vol. 34, No. 1, p. 11). NIH Public Access. 2016, January.
[15] M. Hewitt, J.C. Madden, P.H. Rowe, M.T.D. Cronin, Structure-based modelling in reproductive toxicology:(Q) SARs for the placental barrier, SAR QSAR Environ. Res. 18 (1-2) (2007) 57–76.
[16] A.E. Hoerl, R.W. Kennard, Ridge regression: applications to nonorthogonal problems, Technometrics 12 (1) (1970) 69–82.
[17] S. Katoch, S.S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future. Multimedia tools and applications, Advance online publication, 2020, pp. 1–36.
[18] Y. Kimura, AutoQSAR & QSAR-Evolution, accessed: 07/09/2021, https://svl.chemcomp.com/#AutoQuaSAR, 2020.
[19] S. Krimsky, L.S. Birnbaum, The unsteady state and inertia of chemical regulation under the US Toxic Substances Control Act, PLoS Biol. 15 (12) (2017) e2002404, https://doi.org/10.1371/journal.pbio.2002404.
[20] J. Li, D. Cai, C. Chu, Q. Li, Y. Zhou, L. Hu, B. Yang, G. Dong, X. Zeng, D.a. Chen, Transplacental transfer of per-and polyfluoroalkyl substances (PFASs): Differences between preterm and full-term deliveries and associations with placental transporter mRNA expression, Environ. Sci. Technol. 54 (8) (2020) 5062–5070.
[21] J. McCall, Genetic algorithms for modelling and optimisation, J. Comput. Appl. Math. 184 (1) (2005) 205–222.
[22] G.C. McDonald, Ridge regression, Wiley Interdiscip. Rev. Comput. Stat. 1 (1) (2009) 93–100.
[23] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: An easy approach to molecular descriptor calculations, Match 56 (2) (2006) 237–248.
[24] Molecular Operating Environment (MOE), Chemical Computing Group ULC, 1010 Sherbrooke St, West, Suite #910, Montreal, QC, Canada, 2019.01;.
[25] R. Morello-Frosch, L.J. Cushing, B.M. Jesdale, J.M. Schwartz, W. Guo, T. Guo, M. Wang, S. Harwani, S.-S. Petropoulou, W. Duong, J.-S. Park, M. Petreas, R. Gajek, J. Alvaran, J. She, D. Dobraca, R. Das, T.J. Woodruff, Environmental chemicals in an urban population of pregnant women and their newborns from San Francisco, Environ. Sci. Technol. 50 (22) (2016) 12464–12472.
[26] C. Mori, N. Nakamura, E. Todaka, T. Fujisaki, Y. Matsuno, H. Nakaoka, M. Hanazato, Correlation between human maternal–fetal placental transfer and molecular weight of PCB and dioxin congeners/isomers, Chemosphere 114 (2014) 262–267.

[27] M. Myren, T. Mose, L. Mathiesen, L.E. Knudsen, The human placenta–an alternative for studying foetal exposure, Toxicol. In Vitro 21 (7) (2007) 1332–1340.
[28] T.E. Oliphant, Python for scientific computing, Comput. Sci. Eng. 9 (3) (2007) 10–20.
[29] J.-S. Park, Å. Bergman, L. Linderholm, M. Athanasiadou, A. Kocan, J. Petrik, B. Drobna, T. Trnovec, M.J. Charles, I. Hertz-Picciotto, Placental transfer of polychlorinated biphenyls, their hydroxylated metabolites and pentachlorophenol in pregnant women from eastern Slovakia, Chemosphere 70 (9) (2008) 1676–1684.
[30] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (11) (1994) 1119–1125.
[31] K. Roy, S. Kar, P. Ambure, On a simple approach for determining applicability domain of QSAR models, Chemomet. Intellig. Lab. Syst. 145 (2015) 22–29.
[32] RStudio Team, RStudio: Integrated Development for R, RStudio, PBC, Boston, MA, 2020 http://www.rstudio.com/.
[33] A.K. Saxena, P. Prathipati, Comparison of mlr, pls and ga-mlr in qsar analysis, SAR QSAR Environ. Res. 14 (5-6) (2003) 433–445.
[34] K. Sexton, J.J. Salinas, T.J. McDonald, R.M. Gowen, R.P. Miller, J.B. McCormick, S. P. Fisher-Hoch, Polycyclic aromatic hydrocarbons in maternal and umbilical cord blood from pregnant Hispanic women living in Brownsville, Texas, Int. J. Environ. Res. Public Health 8 (8) (2011) 3365–3379.
[35] T. Takaku, H. Nagahori, Y. Sogame, T. Takagi, Quantitative structure–activity relationship model for the fetal–maternal blood concentration ratio of chemicals in humans, Biol. Pharm. Bull. 38 (6) (2015) 930–934.
[36] R.S. Thomas, T. Bahadori, T.J. Buckley, J. Cowden, C. Deisenroth, K.L. Dionisio, J. B. Frithsen, C.M. Grulke, M.R. Gwinn, J.A. Harrill, M. Higuchi, K.A. Houck, M. F. Hughes, E.S. Hunter, K.K. Isaacs, R.S. Judson, T.B. Knudsen, J.C. Lambert, M. Linnenbrink, T.M. Martin, S.R. Newton, S. Padilla, G. Patlewicz, K. Paul-Friedman, K.A. Phillips, A.M. Richard, R. Sams, T.J. Shafer, R.W. Setzer, I. Shah, J. E. Simmons, S.O. Simmons, A. Singh, J.R. Sobus, M. Strynar, A. Swank, R. Tornero-Valez, E.M. Ulrich, D.L. Villeneuve, J.F. Wambaugh, B.A. Wetmore, A.J. Williams, The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency, Toxicol Sci. 169 (2) (2019) 317–332.
[37] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1) (1996) 267–288.
[38] R.D. Tobias, An introduction to partial least squares regression. In Proceedings of the twentieth annual SAS users group international conference (Vol. 20). Cary: SAS Institute Inc., 1995.
[39] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inf. 29 (6-7) (2010) 476–488.
[40] M.J. Van Der Laan, S. Dudoit, Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples, 2003.
[41] V.N. Vapnik, Conclusion: What is Important in Learning Theory?. In The Nature of Statistical Learning Theory, in: V.N. Vapnik (Ed.), The Nature of Statistical Learning Theory, Springer New York, New York, NY, 1995, pp. 167–175, https://doi.org/10.1007/978-1-4757-2440-0_7.
[42] R. Veerasamy, H. Rajak, A. Jain, S. Sivadasan, C.P. Varghese, R.K. Agrawal, Validation of QSAR models-strategies and importance, Int. J. Drug Des. Discov 3 (2011) 511–519.
[43] E. Vizcaino, J.O. Grimalt, A. Fernández-Somoano, A. Tardon, Transport of persistent organic pollutants across the human placenta, Environ. Int. 65 (2014) 107–115.
[44] C.-C. Wang, P. Lin, C.-Y. Chou, S.-S. Wang, C.-W. Tung, Prediction of human fetal–maternal blood concentration ratio of chemicals, PeerJ 8 (2020) e9562.
[45] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.
[46] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci. 29 (2) (1989) 97–101.
[47] R.M. Whyatt, D.B. Barr, D.E. Camann, P.L. Kinney, J.R. Barr, H.F. Andrews, L. A. Hoepner, R. Garfinkel, Y. Hazi, A. Reyes, J. Ramirez, Y. Cosme, F.P. Perera, Contemporary-use pesticides in personal air samples during pregnancy and blood samples at delivery among urban minority mothers and newborns, Environ. Health Perspect. 111 (5) (2003) 749–756.
[48] A.P. Worth, A. Bassan, A. Gallegos, T.I. Netzeva, G. Patlewicz, M. Pavan, M. Vračko, The characterisation of (quantitative) structure-activity relationships: preliminary guidance. Institute for Health and Consumer Protection, Toxicology and Chemical Substances Unit, European Chemical Bureau.
[49] L. Yang, J. Li, J. Lai, H. Luan, Z. Cai, Y. Wang, Y. Zhao, Y. Wu, Placental transfer of perfluoroalkyl substances and associations with thyroid hormones: Beijing Prenatal Exposure Study, Sci. Rep. 6 (1) (2016), https://doi.org/10.1038/srep21699.
[50] C.W. Yap, PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474.
[51] S. Yin, J. Zhang, F. Guo, L.u. Zhao, G. Poma, A. Covaci, W. Liu, Transplacental transfer of organochlorine pesticides: Concentration ratio and chiral properties, Environ. Int. 130 (2019) 104939, https://doi.org/10.1016/j.envint.2019.104939.
[52] T. Zhang, H. Sun, Y. Lin, X. Qin, Y. Zhang, X. Geng, K. Kannan, Distribution of poly- and perfluoroalkyl substances in matched samples from pregnant women and carbon chain length related maternal transfer, Environ. Sci. Technol. 47 (14) (2013) 7974–7981.
[53] Y.H. Zhang, Z.N. Xia, L. Yan, S.S. Liu, Prediction of placental barrier permeability: a model based on partial least squares variable selection procedure, Molecules 20 (5) (2015) 8270–8286.
[54] X. Zhang, X. Li, Y.e. Jing, X. Fang, X. Zhang, B. Lei, Y. Yu, Transplacental transfer of polycyclic aromatic hydrocarbons in paired samples of maternal serum, umbilical cord serum, and placenta in Shanghai, China, Environ. Pollut. 222 (2017) 267–275.
[55] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc.: Ser. B (Statist. Methodol.) 67 (2) (2005) 301–320.