Université de Montréal

Modélisation de la relation quantitative de structure-activité (QSAR) du passage placentaire des contaminants environnementaux

Par

Laura Lévêque

Département de santé environnementale et santé au travail,
École de santé publique

Mémoire présenté en vue de l'obtention du grade de Maîtrise en santé environnementale et santé au travail, Option générale

Mai 2021

Université de Montréal

Unité académique : Département de santé environnementale et santé au travail, École de Santé Publique

---

Ce mémoire intitulé

Modélisation de la relation quantitative de structure-activité (QSAR) du passage placentaire des contaminants environnementaux

Présenté par
Laura Lévêque

A été évalué(e) par un jury composé des personnes suivantes
Ludwig Vinches
Président-rapporteur

Marc-André Verner
Directeur de recherche

Jérôme Baudry
Membre du jury

Résumé

La diversité croissante dans l'environnement de composés potentiellement fœtotoxiques est une préoccupation de santé publique. L'objectif de ce travail était de contribuer à l'élaboration de méthodes rapides et efficaces pour en évaluer l'exposition prénatale. La modélisation de la relation quantitative structure à activité (QSAR) est apparue comme une méthode de choix dans l'élaboration d'un modèle prédictif pour le passage placentaire des contaminants. Les ratios fœto-maternels de concentrations sanguines pour 105 contaminants ont été compilés à partir de la littérature, et 214 descripteurs moléculaires ont été générés. Dix modèles prédictifs ont été élaborés à l'aide du logiciel Molecular Operating Environnement (MOE) et des langages de programmation Python et R. Les jeux de données d'entrainement et de test ont été utilisés, respectivement, pour élaborer et valider les modèles. L'outil Applicability Domain v1.0 a été utilisé pour déterminer le domaine d'applicabilité (DA). Les modèles élaborés avec les méthodes de régression des moindres carrés partiels dans MOE et SuperLearner dans R, ont montré les meilleures valeurs de précision et de prédictivité avec des coefficients de détermination internes (R^2) de 0,88 et 0,82, des R^2 de validation croisée de 0,72 et 0,57, et des R^2 externes de 0,73 et 0,74, respectivement. Le recouvrement de toutes les molécules du jeu de test par le domaine d'applicabilité a permis de démontrer la fiabilité et la pertinence des prédictions des modèles. Les résultats obtenus démontrent que les modèles élaborés peuvent aider à quantifier l'exposition fœtale aux composés toxiques de l'environnement à partir des concentrations sanguines de la mère.

Mots-clés : QSAR, contaminants, passage placentaire, modélisation, *in silico*.

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. Fetal 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. The 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. The inclusion of all test chemicals by the domain of applicability demonstrated the reliability and relevance of the model predictions. The results obtained demonstrate that QSAR modeling can help quantify placental transfer of environmental chemicals.

Keywords: QSAR, contaminants, placental transfer, modeling, *in silico*.

Table des matières

Liste des tableaux

Article scientifique

Liste des figures

Introduction

Méthodologie

Article scientifique

Liste des sigles et abréviations

ANN	Réseaux de neurones artificiels (*Artificial Neural Network*)
BT	Bootstrap (*Bootstrapping*)
DDE	Dichlorodiphényldichloroéthylène
DDT	Dichlorodiphényltrichloroéthane
DOHaD	Origines développementales de la santé et des maladies (*Developmental Origins of Health and adult Diseases*)
DT	Arbres de décisions (*Decision Tree*)
GA-MLR	Algorithme génétique-Régression Linéaire Multiple (*Genetic Algorithm-Multiple Linear Regression*)
HCH	Hexachlorocyclohexane
HOP	Halogènes Organiques Persistants
kNN	Méthode des k plus proches voisins (*k Nearest Neighbors*)
KRR	Régression Kernel Ridge (*Kernel Ridge Regression*)
LMOCV	Leave-Many-Out Cross-Validation
LOOCV	Leave-One-Out Cross-Validation
LR	Régression Lasso (*Lasso Regression*)
MLR	Régression Linéaire Multiple (*Multiple Linear Regression*)
MOE	Molecular Operating Environment
OCDE	Organisation de coopération et de développement économiques
PAH	Hydrocarbure aromatique polycyclique (*Polycyclic aromatic hydrocarbon*)
PBDE	Polybromodiphényléthers
PBPD	Pharmacodynamique à base physiologique (*Physiologically Based Pharmacodynamic*)
PBPK	Pharmacocinétique à base physiologique (*Physiologically Based Phamarcokinetic*)

PCB	Polychlorobiphényles
PCDD	Polychlorodibenzo-p-dioxines
PCDF	Polychlorodibenzofuranes
PCR	Régression sur principales composantes (*Principal Component Regression*)
PFAS	Substances per- et polyfluoroalkyliques (*Per and Polyfluoroalkyl Substances*)
PFOA	Acide perfluorooctanoïque (*Perfluorooctanoic acid*)
PFOS	Acide perfluorooctanesulfanique (*Perfluorooctanesulfonic acid)*
PLS	Régression des moindres carrés partiels (*Partial Least Squares*)
POP	Polluants Organiques Persistants
QSAR	Relation quantitative de structure à activité (*Quantitative Structure-Activity Relationship*)
RF	Forêts aléatoires (*Random Forest*)
RMSE	Erreur quadratique moyenne (*Root Mean Square Error*)
SAR	Relation Structure-activité (*Structure-Activity Relationship*)
SMILES	Simplified Molecular Input Line Entry Specification
SVM	Machine à vecteurs de support (*Support Vector Machine*)
VIP	Importance des variables par projection (*Variable importance in projection*)

Remerciements

Ce mémoire signe la fin de ma maîtrise et d'un long processus d'apprentissage et d'acquisition de nouvelles compétences et connaissances. Par-là, je souhaite adresser mes sincères remerciements à toutes les personnes qui m'ont permis de mener à terme ce projet, et qui ont participé au développement de la personne que je suis aujourd'hui.

Je tiens à adresser tout d'abord un immense remerciement à mon directeur de recherche, le professeur Marc-André Verner, qui a toujours cru en moi et m'a épaulée tout au long de mon programme de maîtrise. Il a fait preuve à mon égard d'une grande patience, d'une excellente écoute et d'un incroyable soutien. J'ai beaucoup appris sous son aile et j'aimerais ainsi lui témoigner mon entière gratitude et ma plus grande reconnaissance.

Je tiens à remercier grandement ma collègue de laboratoire et amie Nadia Tahiri pour son implication au sein de mon projet et les longues heures de travail partagées. Je la remercie sincèrement pour les conseils et le soutien qu'elle m'a apportée tout au long de mon parcours.

Je remercie également le chercheur Rocky Goldsmith pour son implication et sa participation à mon projet de recherche, et pour toutes les connaissances qu'il m'a apporté sur la modélisation QSAR.

Je remercie du plus profond de mon être mon conjoint Peter qui n'a cessé de m'encourager, de me soutenir et de m'aider tout au long de cette maîtrise, particulièrement en ce contexte de pandémie.

Je tiens à remercier mes parents qui ont toujours cru en moi et qui m'ont toujours soutenue dans mes études, me poussant à constamment me surpasser, même à des kilomètres de chez moi.

Enfin je remercie mes collègues de laboratoire Sherri, Élyse, Lucie, Antoine, Lilit, et Emmanuel, ainsi que mes amis, pour leur aide et leur soutien tout au long de ma maîtrise.

1 Introduction

L'exposition quotidienne aux contaminants toxiques de l'environnement dans la population est un problème majeur de santé publique, tout particulièrement quand elle touche à la santé des femmes enceintes. En effet, plusieurs contaminants sont susceptibles de traverser la barrière placentaire et se rendre à l'organisme en développement (Aylward et al., 2014), mettant en danger la santé du fœtus et du nouveau-né. Plusieurs études expérimentales et épidémiologiques ont montré que le fœtus est particulièrement vulnérable aux atteintes toxiques des contaminants et de leurs effets néfastes à court et long termes (Perera et al., 2004). Étant donné le vaste éventail de composés chimiques sur le marché (plus de 85 000) et de contaminants dans l'environnement, il est nécessaire d'élaborer des outils pour évaluer rapidement la capacité de ces substances à traverser le placenta. La modélisation toxicologique prédictive apparaît alors comme un outil de choix pour étudier ce phénomène. Parmi les outils de toxicologie prédictive, les relations quantitatives de structure à activité (QSAR) se démarquent, permettant de lier la structure d'une molécule à son activité (Tropsha, 2010). Dans le cadre de ce mémoire, la pertinence des modèles QSAR réside dans leur habilité à prédire la capacité d'une molécule de passer au travers du placenta. Les études publiées traitant de la modélisation QSAR pour le passage placentaire des molécules sont limitées et se concentrent principalement sur les médicaments ou sur des familles spécifiques de composés, présentant donc des domaines d'applicabilités restreints à ces molécules. Ainsi, il n'existe pour l'instant aucun modèle QSAR permettant d'estimer le passage d'un large spectre de contaminants de l'environnement au travers du placenta, ce qui gêne considérablement l'identification des contaminants les plus susceptibles de se rendre à l'organisme en développement durant la grossesse. L'élaboration d'un modèle QSAR basé sur des données de contaminants de l'environnement pourrait permettre d'étendre ce domaine d'applicabilité et de fournir un modèle plus adapté aux besoins en santé environnementale.

1.1 Contaminants environnementaux et exposition prénatale

1.1.1 Les contaminants dans l'environnement

Les contaminants environnementaux désignent l'ensemble des substances présentes dans l'environnement de sources naturelles ou anthropiques, et qui sont susceptibles de nuire à la santé des êtres vivants (Boddington, 1990). Parmi ces contaminants, on compte les radionucléides, les métaux lourds (par exemple le mercure, le plomb, l'arsenic ou le cadmium), les composés organiques volatils, et les polluants organiques persistants. Cette section vise à introduire succinctement certains groupes de contaminants d'intérêt dans le mémoire.

Les contaminants sont des sujets d'intérêt majeurs puisqu'ils représentent une diversité grandissante de substances chimiques capables de causer des effets néfastes à court et long termes sur la santé de l'Homme (Boddington, 1990). Certains de ces contaminants, notamment les polluants organiques persistants, ont la capacité de persister dans l'environnement sur de très longues périodes et de résister aux dégradations chimique, photolytique et biologique (Ashraf, 2017). L'exposition quotidienne aux contaminants persistants résulte en une bioaccumulation et bioamplification des molécules toxiques dans l'organisme par le réseau trophique. Une exposition importante à court terme ou plus faible sur le long terme, peut engendrer des effets toxiques aigus ou chroniques respectivement. Dans le but de réduire et réglementer les polluants organiques persistants dans l'environnement, la Convention de Stockholm de 2001 a établi une liste de 12 substances chimiques dangereuses pour les organismes vivants, à laquelle 9 autres substances sont venues s'ajouter lors de la Conférence des Parties en 2009.

Parmi les substances dangereuses et toxiques pour l'Homme se trouvent le chlordane, l'hexachlorocyclohexane (HCH ou lindane), ou encore le dichlorodiphényltrichloroéthane (DDT) (Lallas, 2001). Ces substances sont des pesticides organochlorés dont l'utilisation a été majoritairement interdite, mais est encore pratiquée dans certaines parties du monde. Le DDT par exemple est un insecticide que l'on retrouve encore dans certaines régions d'Afrique pour lutter contre le vecteur de la malaria (Van den Berg et al., 2017). Les métabolites des pesticides organochlorés sont également étroitement contrôlés, car ils s'avèrent parfois plus toxiques que le composé parent (Rani et al., 2017). C'est le cas de l'oxychlordane, le métabolite du chlordane, ou de certaines fumées toxiques et irritantes résultant des processus de combustion.

Les polluants incluent également les polychlorobiphényles (PCB), les polychlorodibenzo-p-dioxines (PCDD) et les polychlorodibenzofuranes (PCDF). Alors que les dioxines et les furanes sont des sous-produits de la combustion de matière organique, les polychlorobiphényles ont été synthétisés pour être utilisés comme isolants thermiques et électriques dans de nombreux matériaux. Les dioxines, furanes et polychlorobiphényles sont persistants et se bioaccumulent dans les graisses, et sont nocifs notamment pour la reproduction, le développement, et les systèmes endocrinien, immunitaire et nerveux (Carrier, 1995).

Les contaminants incluent également des substances per- et polyfluoroalkyliques (PFAS). Les substances per- et polyfluoroalkyliques (PFAS) sont des composés synthétiques très stables et persistants dans l'environnement, qui sont utilisés dans de nombreux produits industriels et de consommation en tant qu'enduits de surface antitaches et imperméables. L'acide perfluorooctanoïque (PFOA) et l'acide perfluorooctanesulfanique (PFOS), dont la production et l'importation est réglementée au Canada et aux États-Unis, ont été utilisés comme revêtement antitache, dans la mousse anti-incendie et dans les produits antiadhésifs. Ces substances ont été associées à plusieurs effets néfastes, notamment la perturbation endocrinienne, le cancer, et les altérations du système immunitaire (Jensen & Leffers, 2008).

Les hydrocarbures aromatiques polycycliques (PAH) sont quant à eux des sous-produits de combustion dont la structure de base est le cycle benzénique. Contrairement aux contaminants cités plus haut, ces derniers ne sont pas considérés comme persistants. Le benzène est le plus simple des hydrocarbures aromatiques et agit comme précurseur dans la synthèse de multiples composés organiques (Lundstedt et al., 2007). Volatil et inflammable, il est notamment retrouvé dans les pesticides, l'essence, le plastique, le caoutchouc, les médicaments, les détergents et les solvants. Les composés possédants plus de deux cycles benzéniques constituent les hydrocarbures aromatiques polycycliques et peuvent être retrouvés naturellement dans le charbon et le pétrole, ou peuvent résulter de la combustion fossile et de l'incinération des déchets. La contamination par ces composés se fait à travers l'alimentation, l'air, l'eau et le sol et leur exposition peut causer à court terme des lésions cutanées, des irritations et des altérations du foie et des reins, ou des effets cancérogènes, et génotoxiques à long terme (Lundstedt et al., 2007).

Les polluants environnementaux sont donc des substances que l'on retrouve partout dans l'environnement et auxquelles l'Homme est exposé principalement par l'alimentation, l'eau, l'air

et le sol. Les effets de cette exposition dépendent de la dose, de la durée, de la voie et de l'exposition simultanée à d'autres contaminants, ainsi que des caractéristiques individuelles et socio-culturelles (INSPQ, 2019). La bioaccumulation de ces substances dans les tissus vivants et les effets nocifs à long terme sur les systèmes endocrinien, nerveux, digestif et reproducteur rendent primordiale l'évaluation de l'exposition populationnelle aux contaminants. L'évaluation de l'exposition de la femme enceinte et du fœtus aux contaminants constitue notamment un enjeu majeur, l'organisme en développement étant particulièrement vulnérable au passage placentaire des contaminants (Aylward et al., 2014).

1.1.2 **La période critique de l'exposition prénatale**

L'exposition du fœtus aux contaminants environnementaux pendant la grossesse est susceptible de provoquer des effets néfastes à court terme sur le développement et à long terme sur la santé. En effet, les contaminants peuvent causer des effets tératogènes sur l'embryon et perturber la croissance de l'organisme, créant ainsi des prédispositions à certaines maladies chroniques et psychiatriques de l'adulte (Lejarraga, 2019). L'évaluation de l'exposition fœtale s'est intensifiée au cours des dernières années, avec une attention toute particulière sur les effets des facteurs environnementaux lors des périodes critiques du développement prénatal. Des études toxicologiques et épidémiologiques ont notamment mis en évidence la présence de substances xénobiotiques dans le sang au cordon (Aylward et al., 2014). D'autres modèles expérimentaux et épidémiologiques réalisés chez l'Homme et l'animal ont démontré la capacité de certains composés à franchir la barrière placentaire (Needham et al., 2011). Enfin des associations ont été établies entre les perturbations du développement fœtal et l'apparition de maladies chez l'individu à l'âge adulte (Almond & Curry, 2011).

1.1.2.1 ***Sensibilité de la période prénatale***

L'exposition du fœtus est déterminée par plusieurs facteurs, notamment l'exposition de la femme avant et tout au long de la grossesse, la demi-vie biologique et le passage au travers du placenta (Wan et al., 2010). L'exposition prénatale aux substances toxiques est particulièrement préoccupante en raison de la susceptibilité et de la vulnérabilité des événements qui se produisent

au cours de cette fenêtre développementale. Le premier trimestre qui comprend l'embryogenèse, est particulièrement sensible aux atteintes toxiques des substances exogènes physiques, chimiques et biologiques en raison des périodes critiques de la morphogénèse, de l'organogenèse, et de l'histogenèse (Pryor et al., 2000). Une toxicité plus tardive au cours du développement fœtal vient impacter la croissance des organes, les processus de différenciation cellulaire, le développement des complexes enzymatiques et immunitaires, des systèmes protéiques membranaires, et du système nerveux (Magnarelli et Guiñazú, 2012).

1.1.2.2 *Hypothèse de Barker et origines développementales de la santé et des maladies (DOHaD)*

L'influence de l'environnement aux stades précoces du développement et ses conséquences dans l'apparition des maladies de l'adulte, dite origine fœtale des maladies de l'adulte, est un concept initialement élaboré par le docteur David Barker en Angleterre dans les années 1980 (Lejarraga, 2019). Ce concept fait suite aux observations faites entre un petit poids à la naissance lié à une sous-nutrition pendant la grossesse et l'augmentation du taux de mortalité des maladies cardiovasculaires à l'âge adulte dans une même zone géographique. Barker en déduit que les carences nutritionnelles aux périodes précoces du développement conduisent à des changements permanents structurels, physiologiques, métaboliques, et hormonaux (Barker, 1995). Ainsi les stimuli environnementaux peuvent entraîner des modifications de l'expression des gènes via des modifications des mécanismes épigénétiques, menant à une reprogrammation du phénotype biologique (Perera et Herbstman, 2011). Le nouveau phénotype dit « économe » est une réponse adaptative visant la survie de l'individu dans un environnement intra-utérin perturbé ou réduit, et le préparant à un environnement extra-utérin similaire. Dans le cas où l'environnement extra-utérin est différent de celui dans lequel s'est développé l'individu, une mauvaise adaptation des capacités fonctionnelles de ses différents systèmes vitaux est à l'origine de prédispositions aux maladies chroniques telles l'obésité, l'hypertension et le diabète (Sartori, 2007). L'adaptation de l'organisme aux effets des facteurs environnementaux auxquels il est soumis lors de son développement témoigne de sa capacité de plasticité phénotypique (Tian & Marsit, 2018). Les résultats obtenus par Barker et ses collègues ont permis d'élargir la recherche épidémiologique chez les enfants, se concentrant sur les perturbations prénatales aux différentes étapes du développement fœtal. Le

concept des origines fœtales des maladies de l'adulte a ainsi graduellement évolué pour celui des origines développementales de la santé et des maladies ou DOHaD (Developmental origins of health and adult diseases).

La recherche en lien avec la DOHaD s'intéresse à l'identification, la quantification et l'évaluation des effets des facteurs environnementaux sur les mécanismes épigénétiques et les conséquences de leurs altérations (Mandy & Nyirenda, 2008). La malnutrition, les composés xénobiotiques, les toxines, les stress comportementaux et psychologiques sont tous autant de facteurs susceptibles d'influencer la régulation épigénétique et la plasticité développementale, et d'entraîner un phénotype économe irréversible. Les marques épigénétiques sont adaptatives, transmissibles, réversibles, et n'altèrent pas la structure de l'ADN (Grova et al., 2019). Elles régulent notamment les mécanismes cellulaires de prolifération, de différenciation et de maturation fonctionnelle mis en jeu dans le développement de l'embryon et du fœtus. Un des principaux facteurs de régulation épigénétique est la méthylation de l'ADN, c'est-à-dire l'ajout d'un groupement méthyl sur une portion de la double hélice d'ADN et qui permet de réprimer l'expression du gène codé par cette portion (Tian & Marsit, 2018). Ce mécanisme est notamment employé pour réprimer des gènes qui ne sont pas nécessaires à un moment donné du développement, permettant de limiter les risques de mutagenèse, et d'assurer la stabilité du génome et le bon déroulement de la croissance. Lors d'un stress induit par un agent environnemental, le processus de méthylation est temporairement désactivé au profit de l'activation de protéines de défense (Baccarelli & Bollati, 2009). Toutefois, un stress environnemental tel que l'exposition à des substances toxiques, peut causer la production en continu de ces protéines de défense et entraîner l'irréversibilité de certaines modifications épigénétiques. Une "nouvelle" mémoire épigénétique est alors forgée, et est susceptible de se retrouver dans le patrimoine génétique de l'individu et de ses descendants (Baccarelli & Bollati, 2009).

1.1.2.3 *Effets de l'exposition aux contaminants environnementaux*

L'exposition prénatale aux contaminants et aux substances toxiques de l'environnement peut affecter les différents systèmes de l'organisme en développement, impactant non seulement le développement des organes, mais également les systèmes endocrinien, reproducteur, immunitaire, métabolique et cérébral (Baccarelli et Bollati, 2009). Les effets toxiques incluent des

défauts de naissance, des morts fœtales, de petits poids à la naissance, des déficiences neurologiques et comportementales, et le développement de maladies métaboliques, cardiovasculaires et mentales (Griffiths et Campbell, 2015).

Le développement cérébral est particulièrement vulnérable à la toxicité des molécules et de leurs métabolites actifs. Qualitativement différent de celui de l'adulte, le cerveau du fœtus présente une barrière hémato-encéphalique qui n'est pas complètement formée et perméable aux xénobiotiques. Les atteintes toxiques altèrent la réplication neuronale et peuvent diminuer de façon irréversible le nombre de neurones (Lejarraga, 2019). Les perturbations au niveau des formations synaptiques et des processus de myélinisation viennent de plus impacter la communication intercellulaire. Les agents neurotoxiques tels que l'alcool, la fumée de cigarette, les polychlorobiphényles et les métaux, ont été associés à des retards de croissance cérébrale in utéro, des perturbations du quotient intellectuel, du comportement, et d'erreurs de processus développementaux menant à des prédispositions aux maladies psychiatriques (Axelrad et al., 2009; Rice & Barone, 2000). La fumée du tabac est particulièrement préoccupante puisqu'elle est responsable d'une diminution de la croissance et résulte en un poids réduit à la naissance et d'une réduction de la circonférence de la tête, impactant le quotient intellectuel et les fonctions cognitives (Perera et al., 2004). Les composants du tabac augmentent également les risques de défauts congénitaux du cœur et jouent un rôle dans l'apparition des maladies cardiovasculaires (asthme, obésité, diabète, etc.) (Perera et Herbstman, 2011). Les atteintes du développement neuronal résultant du passage des contaminants au travers de la barrière placentaire ne sont que très rarement corrigées et ont donc des conséquences permanentes (Grova et al., 2019). En raison de la plasticité cérébrale, les réseaux neuronaux restent sensibles aux expositions environnementales même après la naissance, et les changements épigénétiques peuvent impacter les comportement émotionnel et cognitif (Perera et Herbstman, 2011). Certains agents toxiques tels le méthylmercure, le plomb et les polychlorobiphényles peuvent provoquer une neurotoxicité pouvant se manifester plusieurs années après arrêt de l'exposition, et causer des déficits cognitifs, comportementaux, moteurs, et attentionnels. Les perturbateurs endocriniens, notamment les polybromodiphényléthers (PBDE), impactent également le développement neurologique. Une étude de cohorte menée par Herbstman et son équipe en 2010 (Herbstman et al., 2010) a montré que l'exposition prénatale aux polybromodiphényléthers entraînent des retards de développement. Les enfants chez qui ont été mesurées de plus grandes concentrations de PBDE dans le sang au cordon présentaient, dès les

premières années de vie, des signes mesurables de défauts mentaux et physiques. L'étude de Siddiqui et al. (2003) a de plus observé une association entre les niveaux sanguins dans le cordon ombilical de deux types de pesticides, le dichlorodiphényltrichloroéthane (DDT) et l'hexachlorocyclohexane (HCH), et le retard de croissance intra-utérine. Une association négative significative a également été observée entre les concentrations sanguines maternelles et fœtales de dichlorodiphényldichloroéthylène (DDE), un produit de la dégradation du DDT, et le poids des enfants à la naissance et. Ainsi l'exposition de la mère aux contaminants de l'environnement tout au long de la grossesse peut être à l'origine d'altérations du développement et de la croissance fœtale. Les retards de croissance et les naissances prématurées peuvent notamment résulter d'une maturation hétérogène des tissus placentaires et de perturbations endocriniennes découlant de l'exposition du fœtus aux pesticides (Magnarelli et Guiñazú, 2012).

L'adaptation du phénotype biologique aux stimuli de l'environnement témoigne ainsi de la plasticité de l'organisme à déprogrammer et reprogrammer son génome. La plasticité est notamment visible durant la période intra-utérine du développement au moment où la période d'ancrage du profil génomique est la plus étroite, et où les régulations épigénétiques sont les plus sensibles aux influences exogènes. Le nouveau programme épigénétique produit joue un rôle de protection de l'organisme face aux attaques extrêmes de l'environnement, par le biais de réponses adaptatives physiologiques et métaboliques. Les profils épigénétiques établis sont copiés et amplifiés grâce à la rapidité du développement cellulaire et au haut taux de synthèse de l'ADN, et persistent même après la fin du stress environnemental, voire chez les générations futures (Marsit, 2015) La plasticité surtout présente lors du développement embryonnaire, laisse peu à peu place à l'expression de différents phénotypes (capacités fonctionnelles fixes dans le temps) à partir d'un génotype en particulier.

L'évaluation des risques toxiques pour le fœtus passe l'évaluation des expositions de la femme enceinte et de l'organisme en développement aux contaminants de l'environnement. Les effets néfastes à court et long terme décrits ci-dessus sont en effet variés, et dépendent du type de contaminant et de la dose d'exposition. La diversité grandissante des substances toxiques dans l'environnement rend de plus nécessaire la quantification de l'exposition fœtale et la caractérisation des facteurs influençant le passage placentaire des contaminants (Mattison, 2010).

1.2 Estimation du passage placentaire

La tragédie de la thalidomide dans les années 1960 a mis en doute le statut d'imperméabilité de la membrane placentaire aux substances xénobiotiques (Myren et al., 2007). La découverte d'anomalies congénitales chez des milliers d'enfants exposés in utero à ce médicament a déclenché une série de recherches sur le passage placentaire des médicaments et des contaminants environnementaux. Des modèles animaux et humains *in vivo*, *in vitro*, *ex vivo*, et plus récemment *in silico*, ont été développés afin d'étudier les caractéristiques du transport placentaire et les effets de l'exposition aux substances toxiques chez le fœtus.

1.2.1 **Le placenta : rôle et fonction**

Le placenta est une structure complexe qui joue un rôle indispensable dans la croissance et le développement optimal du fœtus. Cet organe vascularisé, qui relie l'utérus de la mère à l'unité foeto-placentaire, a pour unité de base la villosité choriale (Griffiths et Campbell, 2015). Cette villosité est une projection de tissus fœtaux entourés de chorion qui est localisée dans la chambre intervillaire, un espace séparant les tissus maternels des tissus fœtaux. La double couche de cellules trophoblastiques qui constituent le chorion sert à mettre en contact le sang maternel et le sang fœtal. Les substances provenant de la circulation maternelle et qui baignent dans l'espace intervillaire doivent traverser les cellules syncytiotrophoblastes de la première couche puis les cytotrophoblastes de la seconde, pour accéder à la circulation fœtale et rejoindre le fœtus par le biais du cordon ombilical (Figure 1). Les circulations maternelle et fœtale ne se mélangent donc jamais, bien qu'en contact étroit perpétuel (Griffiths et Campbell, 2015). L'apport des nutriments essentiels au fœtus est assuré par le placenta qui prend en charge de nombreuses fonctions telles que la respiration, la nutrition et l'excrétion. Les échanges métaboliques et gazeux prenant place au niveau de la membrane placentaire fournissent au fœtus l'oxygène et les éléments nutritifs nécessaires à sa croissance, et permettent l'élimination des déchets et du dioxyde de carbone. Le placenta a également pour fonction la production de cellules immunitaires et hormonales nécessaires au bon déroulement de la grossesse et de la maturité optimale de l'organisme (Magnarelli et Guiñazú, 2012. Cette barrière joue enfin un rôle de filtre protecteur contre les composés xénobiotiques potentiellement toxiques pour le développement fœtal, les maladies ou encore les infections (Pemathilaka et al., 2019).

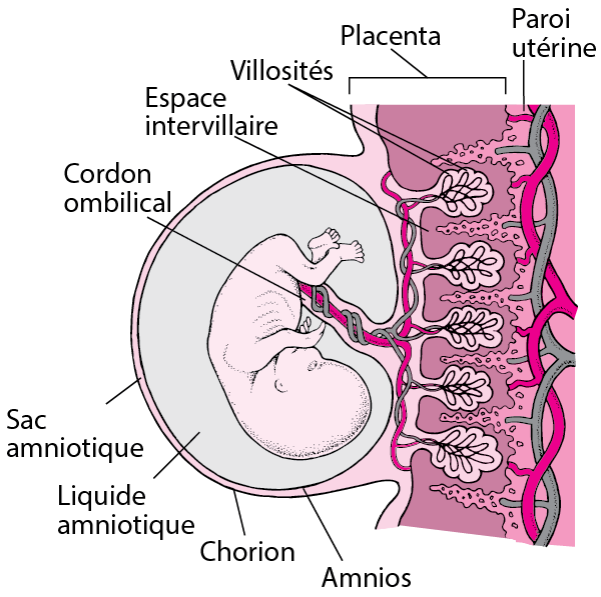

Figure 1. Illustration de la barrière placentaire et du fœtus à 8 semaines de grossesse
Source : https://www.msdmanuals.com/fr/accueil/multimedia/figure/gyn_placenta_embryo_8_weeks_fr

Le sang de la mère permet de nourrir le fœtus tout au long de la grossesse et de lui apporter les éléments énergétiques nécessaires à son développement (eau, glucose, acides aminés, vitamines, électrolytes, acides gras pour la signalisation cellulaire, etc.). Les échanges se font principalement par diffusion passive selon les lois du gradient de concentration et du gradient électrochimique (Griffiths et Campbell, 2015). La constante de diffusion, k, est notamment dépendante des propriétés physicochimiques du composé qui traverse la membrane telles que le poids moléculaire, la solubilité lipidique, le degré d'ionisation, et la capacité de se lier aux protéines. La membrane placentaire étant constituée de cellules avec une bicouche lipidique, elle laisse difficilement traverser les molécules de grosse taille (plus de 500 Da) ou à faible hydrophobicité (Griffiths et Campbell, 2015). De même, les molécules chargées (au pH sanguin maternel), qui sont liées à des protéines ou qui vont à l'encontre de leur gradient, ne peuvent passer par diffusion passive et nécessitent une diffusion active facilitée par des transporteurs membranaires. Les molécules trop grosses pour les diffusions passive et active, sont, elles, soient dégradées en peptides et acides aminés, soit acheminées via des vésicules membranaires (pinocytose), comme le sont les immunoglobulines qui assurent l'immunité passive du fœtus pendant les premiers mois de la grossesse. La régulation du transport placentaire se fait grâce à des signaux hormonaux produits en rétroaction aux besoins du fœtus. La fonction endocrinienne

permet notamment de stimuler la circulation maternelle et de fournir les peptides et hormones stéroïdes indispensables à leur croissance fœtale (Griffiths et Campbell, 2015).

De par son rôle de barrière et sa fonction nourricière, le placenta constitue la principale porte d'entrée des contaminants dans l'unité foeto-placentaire. La régulation du passage des molécules chimiques à travers la barrière placentaire n'est pas uniforme tout au long de la grossesse (Rudge et al., 2009). Le taux de transfert des composés dépend en effet de nombreux facteurs dont la concentration molaire, le mode de transport à travers la membrane, ou encore la surface disponible pour les échanges métaboliques. La surface de contact entre les structures maternelles et fœtales, et le débit sanguin influencent notamment les concentrations en molécules qui vont être retrouvées dans le sang au cordon (Rudge et al., 2009). De nombreuses études ont d'ailleurs démontré la présence de contaminants et leurs métabolites dans les tissus placentaires et au niveau du cordon ombilical (Needham et al., 2011; Morello-Frosch et al., 2016; Zhang et al., 2017). Aylward et al., (2014) ont résumé, dans une revue de la littérature, des ratios de concentrations sanguines de paires mères-enfants pour de nombreux contaminants tels que des pesticides organochlorés, des substances per- et polyfluoroalkyliques, des polychlorobiphényles, ou encore des polybromodiphényléthers. L'étude de Zhang et al., (2021) a montré que la diffusion passive et le transport actif sont les principaux mécanismes du transport transplacentaire des polluants organiques persistants halogénés. Ces mécanismes dépendant respectivement de l'affinité de liaison aux protéines plasmatiques et des transporteurs d'influx et d'efflux.

Les études expérimentales et épidémiologiques chez l'animal et l'humain ont montré que les structures et fonctions placentaires jouent un rôle vital dans le passage placentaire des composés chimiques dans l'évaluation de l'exposition fœtale. Plusieurs méthodes d'estimation ont été développées afin de quantifier l'exposition fœtale et mettre en évidence les mécanismes sous-jacents et caractéristiques du transfert placentaire (Myren et al., 2007).

1.2.2 **Modèles toxicologiques animaux**

Le scandale de la thalidomide a conduit aux premières études animales *in vivo* de toxicité pour tester et mesurer les profils toxicologiques des substances pharmaceutiques et environnementales. Les considérations et limites éthiques dans l'évaluation de la toxicité fœtale sont à l'origine du développement de modèles toxicologiques expérimentaux chez les animaux. Plusieurs espèces ont été considérées pour l'étude du transfert placentaire des contaminants. Parmi

celles-ci, sont retrouvés les primates non humains, les lapins, les cochons d'Inde, les rats, la souris ou encore le mouton, ces deux dernières espèces étant favorisées pour leurs similitudes en termes de barrière et structures placentaires, et de structures vasculaires foeto-placentaires respectivement (Grigsby, 2016). Les modèles animaux ont permis d'étudier les expositions à court, moyen et long-terme, et d'observer les relations dose-réponse en lien avec le développement, la survie et la fertilité, en identifiant les organes cibles de la toxicité. L'identification des paramètres physiologiques et biochimiques entrant en jeu dans le transport et le métabolisme des substances toxiques a notamment permis de déterminer les sujets d'intérêts toxicologiques chez l'Homme (Grigsby, 2016). L'exposition de la femme enceinte aux contaminants est toutefois très variable et dépend de nombreux facteurs individuels et socio-économiques (Edwards, 2017). Il peut être ainsi difficile de reproduire l'exposition du fœtus humain chez certaines espèces malgré les similarités gestationnelles et physiologiques, résultant en des différences qualitatives et quantitatives entre l'animal et l'Homme. De plus, les différences pharmacocinétiques (absorption, distribution, métabolisme, excrétion) entre les espèces peuvent également expliquer les différences de sensibilité à un même agent toxique.

Ainsi, bien que les modèles animaux présentent l'avantage d'étudier le transfert transplacentaire des substances xénobiotiques en conditions *in vivo*, les différences physiologiques inter-espèces sont sources d'erreurs et rendent les résultats difficilement extrapolables à l'Homme (Ehrhardt et Kim, 2007). Cela est d'autant plus vrai que le placenta est une structure spécifique propre à l'espèce. En effet, malgré certaines similitudes entre espèces, les caractéristiques physiologiques et fonctionnelles du placenta humain ne sont retrouvées intégralement chez aucune autre espèce (Grigsby, 2016). L'absence de données concluantes en termes d'efficacité a conduit au délaissement graduel des études animales, cela étant également motivé par des raisons financières (coûts d'installation), temporelles (nombre de contaminants, délai de prise de décision), et éthiques (bien-être animal).

1.2.3 **Modèles toxicologiques humains**

1.2.3.1 ***Modèles humains***

Les incertitudes et difficultés liées à l'extrapolation des données *in vivo* animales à l'Homme inter-espèces ont mené à la quantification du passage placentaire des substances toxiques

par des mesures des concentration sanguines chez la mère et au cordon ombilical. Cette méthode permet de façon éthique et sécuritaire d'estimer le taux de transfert placentaire. Bien que simples à mettre en place, ces mesures demeurent limitées et n'apportent aucune information sur les propriétés du transport placentaire (Bourget et al., 1995). Le métabolisme et les variations pharmacocinétiques des profils de concentrations sanguines ne sont également pas pris en compte (Ehrhardt et Kim, 2007). Enfin il peut être complexe d'étudier l'exposition du fœtus aux contaminants en raison de concentrations parfois en dessous du seuil de détection, du nombre d'échantillons disponibles, de possibles mixtures de composés, ou encore lorsqu'un contaminant n'est pas encore sur le marché (Myren et al., 2007). Les limites de cette approche ont ainsi mené au développement de différentes approches, notamment les modèles *in vitro* et *ex-vivo*.

1.2.3.2 ***Modèles in vitro***

L'estimation du passage placentaire *in vitro* chez l'humain utilise des modèles à base de cellules et de tissus du placenta humain. Ces modèles permettent d'étudier rapidement les effets de médicaments et de composés toxiques sur la barrière placentaire, et d'obtenir des résultats précis au niveau de l'interaction des substances avec les molécules de la membrane. La réplication des interactions entre la mère et le fœtus est ainsi rendue possible grâce aux méthodes de micro-ingénierie telles que le modèle « placenta-on-a-chip », une technique de culture *in vitro* de fragments de tissus ou de lignées cellulaires (Pemathilaka et al., 2019). Cette technique permet d'étudier le transport et le métabolisme des composés au niveau de la barrière, et de comprendre les processus menant aux altérations du développement fœtal. Les cellules placentaires primaires trophoblastiques sont généralement utilisées puisqu'elles sont retrouvées de part et d'autre de la barrière placentaire (Ehrhardt et Kim, 2007), et peuvent servir à étudier les échanges placentaires, la régulation de l'expression génique et les mécanismes toxiques cellulaires et moléculaires (Göhner et al., 2014). Le prélèvement, l'isolement et la culture de trophoblastes à partir placentas jeunes ou matures permettent d'étudier les types de transporteurs impliqués dans le transfert placentaire et leurs interactions avec les substances xénobiotiques (Pemathilaka et al., 2019). Les conditions de culture de ces cellules rendent cependant leur étude difficile et notamment en termes de maintien de leur viabilité (Elefant et Beghin, 2009). Les lignées cellulaires dérivées de choriocarcinome humain (tumeur maligne placentaire) telles que les BeWo, les JAr et les JEG, sont utilisées pour l'étude de la physiologie et du transport placentaire, ces cellules étant comparables

aux trophoblastes en termes de morphologie, de marqueurs biochimiques et de sécrétions hormonales. Les vésicules de membranes villositaires sont également utilisées dans l'étude des mécanismes du transport placentaire pour chaque côté de la membrane placentaire. L'exposition de cellules aux substances toxiques permet d'obtenir des paramètres biologiques essentiels et d'évaluer leur composition en facteurs solubles (hormones, cytokines, facteurs de régulation immunitaire, etc.). Enfin, la culture d'explants placentaires de 1er et 3ème trimestre offre l'avantage supplémentaire d'étudier les effets toxiques des substances sur la structure des tissus (Elefant et Beghin, 2009).

Les méthodes *in vitro* permettent ainsi de répliquer les fonctions et la physiologie du placenta humain, offrant une approche mécanistique que les mesures sanguines dans le sang maternel et fœtal ne permettent pas. Les cellules et fragments de tissus cultivés sont faciles à manipuler, et il est possible d'étudier le passage placentaire à différents moments de la grossesse selon les prélèvements effectués. Certains tests de caractéristiques essentielles étant encore à réaliser, la réplication des conditions physiologiques de la membrane placentaire est limitée au niveau pharmacocinétique (Göhner et al., 2014). La technique placenta-on-a-chip est complexe, longue et coûteuse, et limite le contrôle de la performance et des habiletés de modèles. Ainsi les modèles *in vitro*, bien que réduisant l'utilisation animale, restent difficilement interprétables et extrapolables *in vivo*, cela tout particulièrement lorsque la complexité des fonctions étudiées augmente. Une autre approche, soit la perfusion *ex vivo* de vrais placentas récupérés à la naissance, permet de contourner certains obstacles des tests in vitro.

1.2.3.3 ***Modèles ex vivo***

Les expériences *ex vivo* consistent en l'étude d'organes et de tissus vivants en dehors de l'organisme. Développée en 1967, la perfusion placentaire *ex vivo* permet la recréation de la circulation fœto-maternelle au niveau d'un cotylédon, l'unité fonctionnelle vasculaire du placenta (Bouazza et al., 2019). Cette méthode est celle qui se rapproche le plus des conditions *in vivo* chez l'humain, tout en permettant de contourner les contraintes éthiques et les variations inter-espèces. En effet, l'utilisation d'organes placentaires permet de conserver la physiologie, l'organisation des tissus et la dynamique des échanges *in vivo*. Il est ainsi possible d'étudier le transport et le métabolisme placentaire, le stockage des toxines, le mouvement, la cinétique, et la distribution

(clairance) des composés entre les compartiments maternel et fœtal, les fonctions endocrinienne et enzymatiques, et la prolifération et la différenciation cellulaire (Myren et al., 2007).

La perfusion placentaire peut être à circulation simple pour étudier du mouvement des composés d'un compartiment vers l'autre afin de déterminer l'indice de clairance, ou à double circulation (mouvement de la mère vers le fœtus et du fœtus vers la mère) pour calculer le taux de transfert. Les modèles doubles recirculatoires présentent les conditions les plus similaires au mouvement des substances endogènes et exogènes *in vivo* et sont donc idéaux pour l'étude du transfert transplacentaire et des effets chimiques sur les structures placentaires (Myren et al., 2007). Un tel modèle nécessite ici encore de répliquer de façon optimale les réactions métaboliques se produisant *in vivo*, cela passant par une réplication de l'apport en nutriments et le respect des différences des pH sanguins. Des études des propriétés pharmacocinétiques de la barrière placentaire sont donc nécessaires préalablement à la mise en place de modèles de perfusion et à l'estimation du transfert transplacentaire. L'intégrité et la viabilité des tissus sont assurées tout au long de la perfusion par un suivi de la pression artérielle et de la consommation de glucose. L'évaluation du transfert placentaire des substances se fait en fonction de celui de l'antipyrine, un médicament de référence qui diffuse librement à travers la barrière et permet la normalisation du passage des composés xénobiotiques (Elefant et Beghin, 2009). La perfusion placentaire permet également d'étudier le rôle de transporteurs protéiques de la membrane puisque la polarité du transport est respectée, et d'étudier l'impact de l'inhibition de ces derniers. Elle permet aussi d'étudier les pathologies de la circulation placentaire en mesurant les changements du ratio systolique-diastolique au niveau de l'artère ombilicale par occlusion progressive de la circulation artérielle placentaire (Pemathilaka et al., 2019). La perfusion placentaire a notamment été utilisée pour démontrer et caractériser la sensibilité du placenta au cadmium (Wier et al., 1990). Les paramètres pharmacocinétiques mesurés avec les modèles *ex vivo*, peuvent être par la suite être utilisés pour estimer les paramètres des modèles pharmacocinétiques (PBPK) et pharmacodynamiques (PBPD) à base physiologique (Wier et al., 1990).

La perfusion placentaire constitue ainsi à ce jour la méthode idéale pour l'étude du transfert placentaire des substances toxiques chez le fœtus car les similarités au milieu in vivo augmente la fiabilité de prédiction des mesures (Ehrhardt et Kim, 2007). Comme les autres modèles de toxicité placentaire chez l'Homme, la perfusion placentaire est limitée par plusieurs facteurs. Tout d'abord, les variations interindividuelles des organes ne permettent pas d'extrapolation les données obtenues

à partir d'un placenta en particulier (Hutson et al., 2011). De plus les placentas en santé récupérés au terme de la grossesse ont une maturité maximale et ne permettent pas l'étude des caractéristiques du transfert placentaire des composés xénobiotiques aux stades précoces du développement (Hutson et al., 2011). Ces stades constituent les moments du développement fœtal, là où la vulnérabilité du fœtus est maximale. Le placenta mature présente également une plus courte distance entre les vaisseaux sanguins maternels et fœtaux correspondant à une plus grande activité d'échanges métaboliques, ce qui ne permet pas d'extrapoler les résultats aux stades antérieurs de la grossesse. Le modèle de perfusion est également statique au niveau physiologique et métabolique et n'est pas représentatif de la barrière placentaire tout au long de la grossesse, notamment en termes de changements de la composition en protéines membranaires des deux côtés du placenta. Les composés métaboliques des xénobiotiques ne sont également pas pris en compte de ce modèle (Myren et al., 2007). Au niveau matériel, la perfusion placentaire est limitée par la disponibilité des organes et leur viabilité qui ne dépasse généralement pas 48 heures (Myren et al., 2007). Ces limites restreignent ainsi le nombre de composés pouvant être étudiés. La nature de l'exposition et les variations pharmacocinétiques influent donc sur les différences entre les résultats *in vivo* et *ex vivo*. Finalement, l'absence de standardisation et de modèle générique de la perfusion placentaire diminuent la validité des modèles *ex vivo* développés (Göhner et al., 2014).

Les modèles *in vivo*, *in vitro* et *ex vivo* chez l'Homme et l'animal aident à l'étude du transfert placentaire des composés xénobiotiques. La combinaison des forces de chaque modèle permet notamment une meilleure compréhension des mécanismes du transport placentaire (Myren et al., 2007). La complexité de l'établissement des conditions *in vivo* d'exposition de la femme enceinte et du fœtus dans ces modèles limite considérablement l'étude de composés variés. La diversité grandissante de nouveaux contaminants environnementaux rend difficile l'estimation du passage placentaire pour chaque contaminant, et a mené au développement de modèles toxicologiques prédictifs *in silico*. En permettant une estimation à grande échelle des taux de transfert placentaire, les modèles prédictifs permettent de quantifier l'exposition fœtale aux contaminants de l'environnement et d'agir rapidement sur les potentiels risques toxiques.

1.3 Modélisation QSAR

1.3.1 La modélisation *in* silico

Les modèles *in silico* désignent les modèles élaborés à l'aide de méthodes informatiques. Initialement développés pour le domaine pharmaceutique, notamment pour le design de médicaments, les modèles *in silico* se sont rapidement étendus à de plus vastes applications biologiques et toxicologiques. Ce type de modélisation à base de calculs complexes, permet de stocker, analyser, explorer, et curer de nombreuses données provenant d'essais chimiques, afin de modéliser de nouvelles substances ou d'évaluer des substances déjà existantes (Knudsen et al., 2015). L'objectif pour la recherche pharmaceutique est de développer de nouvelles molécules capables d'interagir avec des cibles moléculaires spécifiques. Pour la recherche toxicologique, le développement de modèles mathématiques est particulièrement intéressant dans l'étude de la toxicité des substances environnementales à partir de données obtenues avec les modèles *in vivo*, *in vitro* et *ex vivo* (Modi et al., 2012). En effet, de tels modèles peuvent permettre d'identifier des substances potentiellement dangereuses, de déterminer leur degré de toxicité, ou encore de caractériser leur toxicocinétique (Knudsen et al., 2015). Le développement de modèles fiables, robustes et scientifiquement interprétables permet ainsi d'orienter la recherche pour le développement de substances non dangereuses ou de fournir des renseignements sur certains composés pour lesquels les données de toxicocinétique sont limitées (Cronin et Madden, 2010).

Le nombre croissant de contaminants dans l'environnement et qui se retrouvent aujourd'hui sur le marché (plus de 85 000), sur lesquels très peu d'information est disponible relativement à leur toxicité et toxicocinétique, a mené au développement de méthodes à haut débit d'identification et d'évaluation du risque toxicologique (Knudsen et al., 2015). Les contaminants environnementaux n'étant effectivement pas conçus initialement pour interagir avec l'Homme et ses propriétés biologiques et cinétiques, leurs niveaux d'exposition, leurs structures et leurs modes d'action ne sont pas toujours très bien connus (Hartung et Hoffman, 2009). L'élaboration de modèles toxicologiques prédictifs à partir de résultats de modèles expérimentaux, permet de prédire des réponses ou propriétés biologiques pour des molécules dont la toxicité n'a pas encore été étudiée *in vivo* ou *in vitro*. Dans l'estimation du transfert transplacentaire des contaminants, les modèles *in silico* offrent ainsi l'avantage d'estimer les taux de transfert placentaire de nouvelles molécules de manière non-invasive. Ces modèles sont d'autre part relativement accessibles et

faciles à implémenter, ne nécessitant pas d'animaux ou de tissus vivants, et permettent ainsi d'étudier rapidement un très grand nombre composés toxiques (Cronin et Madden, 2010). Cela permet notamment de contourner les difficultés liées à la réplication exacte du système placentaire et des caractéristiques physiologiques et métaboliques *in vivo*. Les méthodes *in silico* permettent ainsi de modéliser les processus pharmacocinétiques et dynamiques, offrant une étude plus mécanique et quantitative des processus en jeu au niveau de la barrière placentaire, et l'identification de substances capables d'agir sur les propriétés du placenta (Knudsen et al., 2015).

Les outils de modélisation *in silico* permettent d'obtenir des résultats variés et adaptés aux effets ou propriétés toxicologiques étudiés. Il est ainsi possible de développer des modèles de classifications à l'aide de méthodes de relation structure-activité (SAR) qui permettent une compréhension mécaniste d'effets observés. Parmi ces méthodes se trouve les alertes structurelles qui consistent en l'identification de pharmacophores (ou toxicophores), des sous-structures de composés dont l'activité est associée à une toxicité ou propriété, et qui peuvent par exemple perturber la fonction de certaines macromolécules comme les protéines de transports, les enzymes, ou l'ADN (Modi et al., 2012). Ces alertes sont notamment utilisées pour la classification binaire de composés potentiellement dangereux. Les systèmes experts (ou références croisées) sont d'autres méthodes *in silico* qui utilisent l'expérience ou les connaissances d'experts afin d'établir des règles de prise de décision. Ce type de méthode permet de déduire ou d'expliquer les mécanismes de composés toxiques à partir de données de composés similaires (Hartung et Hoffman, 2009). La modélisation *in silico* permet également l'élaboration de procédures d'analyses de données afin de guider le flux de travail d'analyses *in vivo* et *in vitro* et d'améliorer le design des expérimentations (Hartung et Hoffman, 2009).

Les modèles *in silico* peuvent aussi être des modèles prédictifs qui utilisent des données *in vivo*, *in vitro*, ou *ex vivo* obtenues avec des méthodes de référence, et des algorithmes mathématiques, pour prédire de nouveaux résultats pour une même famille de composés. Parmi ces modèles prédictifs se trouvent les modèles pharmacocinétiques à base physiologique (PBPK) dont l'objectif est de prédire le devenir de substances dans l'organisme en fonction du temps selon les paramètres physiologiques, physico-chimiques et biochimiques (Raunio, 2011). Un autre type de modélisation prédictive toxicologique est celle de la relation quantitative structure-activité (QSAR). Branche de l'intelligence artificielle et plus spécifiquement de l'apprentissage machine, la modélisation QSAR est basée sur le concept proposé par Alexander Crum-Brown et Thomas R.

Fraser en 1868, de l'existence d'une relation mathématique entre la structure et l'activité d'un composé (Hemmerich et Ecker, 2020). Ainsi des composés de structures similaires auront des activités similaires, et la description des structures chimiques corrélées aux activités ou propriétés biologiques d'intérêts permet de faire de classification ou des prédictions continues par régression (Hemmerich et Ecker, 2020).

Les différentes méthodes de modélisation permettent ainsi d'apporter des informations qualitatives pour l'identification du risque toxique, d'évaluer et de quantifier la toxicité, d'orienter la prise de décision pour la priorisation de développements ou de tests de composés, ou encore de prédire des données *in vivo*.

1.3.2 **Relation quantitative de structure-activité (QSAR)**

1.3.2.1 ***Généralités***

La modélisation QSAR est un modèle mathématique qui associe une mesure quantitative de la structure d'une molécule à son activité biologique (processus chimique ou cellulaire) ou à ses propriétés (parfois appelée QSPR dans ce cas) (Hartung et Hoffman, 2009). Les étapes de développement d'un modèle prédictif QSAR restent relativement similaires peu importe le domaine d'application envisagé. La première étape est le choix de l'effet biologique, de l'activité ou de la propriété à prédire. Cela permet de déterminer le mécanisme ou le processus à utiliser, et la nature des données nécessaires au développement du modèle. Le sujet d'étude doit être précis et peut faire partie d'un système biologique plus complexe, mais est en général précisé par la disponibilité des données (Hartung et Hoffman, 2009). La seconde étape consiste en la collecte des données. Celles-ci doivent être appropriées à l'objet de recherche puisque la qualité des données contribue à la qualité du modèle (Yang et al., 2018). Les données peuvent provenir d'une revue de la littérature ou d'expérimentations *in vivo, in vitro* et *ex vivo* réalisées dans le cadre de l'étude. Une base de données construite à partir de cette dernière technique présente plusieurs avantages à la revue de littérature, notamment la possibilité de vérifier les données, d'adapter le protocole expérimental le cas échéant, de définir le domaine d'applicabilité d'intérêt, ou encore d'investiguer les anomalies (Combes, 2012). La fiabilité du modèle est dépendante de la fiabilité des données, celle-ci pouvant être assurée par des critères de sélection plus précis assurant la validité et limitant la diversité des sources de données. La taille du jeu de données doit être raisonnable et présenter

une diversité adéquate. La qualité des données est reflétée par leur corrélation avec l'effet recherché et la bonne construction de la base de données (nomenclature correcte, pureté des composés, etc.) (Combes, 2012). Les molécules sont ensuite converties au format SMILES (Simplified Molecular Input Line Entry Specification), un langage de description de la structure chimique moléculaire sous forme de chaînes de caractères (Combes, 2012). C'est à partir des SMILES que sont générés les descripteurs moléculaires, des critères numériques discrets ou continus qui décrivent la structure chimique, les propriétés physico-chimiques et la topologie des composés chimiques (Yang et al., 2018). Les descripteurs peuvent décrire des propriétés structurelles (descripteurs 2D) ou des propriétés conformationnelles (descripteurs 3D) et traiter en autres de la stéréochimie, de la topologie, de la configuration électronique, ou de l'hydrophobicité. Les empreintes moléculaires sont d'autres types de descripteurs qui portent sur les formes et les patrons retrouvés dans la structure moléculaire. Les empreintes sont généralement plus interprétables car elles réfèrent à une sous-structure définie (Yang et al., 2018). La sélection des descripteurs doit être logique et être en lien avec l'effet étudié, certains types de descripteurs pouvant ne pas être corrélés ou l'être de manière superficielle. Notamment la sélection doit tenir compte de si l'on s'intéresse à l'entièreté de la molécule ou à un groupe fonctionnel spécifique dans l'étude de la toxicité ou de la propriété (Combes, 2012). Une fois les descripteurs calculés, le jeu de données est divisé en un jeu d'entraînement qui servira à la phase d'apprentissage du modèle, et en un jeu de test qui servira à tester et valider le modèle développé (Tropsha, 2010). L'élaboration des modèles se fait selon l'approche envisagée, en utilisant des outils statistiques appropriés à la nature de l'effet étudié, au type de données (mesures discrètes, continues ou de catégories), aux descripteurs disponibles, à la taille de la base de données, et à l'utilité du modèle (Yang et al., 2018). Différents outils de classification et de régression peuvent être utilisés tels que la machine à vecteur de support (SVM), les forêts aléatoires (RF), le bootstrap (BT), la méthode des k plus proches voisins (kNN), les réseaux de neurones artificiels (ANN), ou encore les analyses de régression linéaire multiple (MLR) fréquemment utilisés pour la transparence de leur processus (Yang et al., 2018). Les modèles QSAR peuvent être locaux et portés sur un espace chimique réduit ou des composés spécifiques, ou globaux, c'est-à-dire étant développés avec des données variés et qui sont applicables plus généralement (Raunio, 2011).

1.3.2.2 *Validation*

La validation d'un modèle QSAR est une étape primordiale pour son acceptation à des fins réglementaires. L'organisation de coopération et de développement économiques (OCDE) encourage depuis les années 1990 le développement des méthodes SAR et QSAR, et a défini en 2004 cinq principes de validation des modèles QSAR (Raunio, 2011). Ces principes fournissent les lignes directrices pour le développement de modèles fiables et robustes et sont les suivants :
i) Un effet défini : l'effet toxicologique et le protocole d'expérimentation doivent être identifiés, la base de données doit être définie.

ii) Un algorithme non ambigu : l'algorithme mis au point dans l'élaboration du modèle ne doit pas être trop complexe au risque de ne pas permettre la transparence de l'expérimentation et la reproductibilité des calculs réalisés.

iii) Un domaine d'applicabilité défini : l'espace physico-chimique couvert par le modèle doit être décrit afin de garantir la confiance dans les prédictions obtenues.

iv) Des mesures appropriées de la qualité de l'ajustement, de la robustesse et de la prédictivité : le modèle doit être statistiquement validé, à la fois de façon interne (qualité de l'ajustement et robustesse) et de façon externe (prédictivité).

v) Une interprétation mécanistique, si possible soit elle : les descripteurs ayant permis d'élaborer le modèle doivent pouvoir expliquer les prédictions obtenues.

L'évaluation de la performance du modèle est réalisée grâce aux mesures statistiques des validations interne et externe comme l'énonce le quatrième principe de l'OCDE. Les mesures fréquentes de performance incluent généralement le coefficient de détermination R^2, une mesure de la variance des résultats qui témoigne d'une haute précision (d'une faible variance) lorsque le R^2 se rapproche de 1, et l'erreur quadratique moyenne RMSE, une mesure de la justesse et de la précision du modèle, qui témoigne d'une faible variance lorsque la valeur se rapproche de 0 (Yang et al., 2018). La validation interne permet d'estimer la qualité de l'ajustement aux données et la robustesse du modèle, et donc d'assurer que les prédictions sont stables et sont bien dépendantes des paramètres utilisés. Pour ce faire, et notamment pour éviter que le modèle ne soit trop ajusté aux données sur lesquelles il est développé (surentraînement), une étape de validation croisée (*cross-validation*) peut être réalisée (Modi et al., 2012). La validation croisée consiste à séparer le

jeu d'entrainement en sous-groupes, chaque sous-groupe étant mis de côté tour à tour. À chaque tour, les sous-groupes restants sont utilisés pour développer un modèle qui est alors testé sur le sous-groupe mis de côté. Autant de modèles sont produits qu'il y a de sous-groupes, et un coefficient de détermination Q^2 (ou R^2_{CV}) est déterminé à partir de tous ces modèles (Figure 2) (Chtita et al., 2017). Lorsque le jeu d'entrainement est découpé en *k* sous-groupes, on parle de « leave-many-out cross-validation » (LMOCV) ou de « *k*-fold cross validation ». Lorsqu'il y a autant de sous-groupes qu'il y a d'observations ($k = n$), on parle alors de « leave one out cross-validation » (LOOCV) (Combes, 2012).

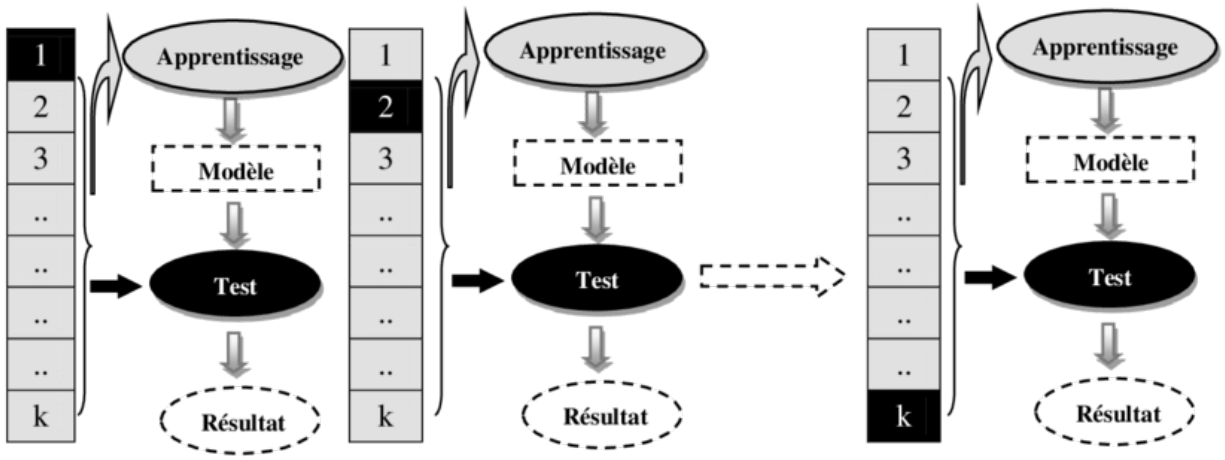

Figure 2. Procédure de validation croisée à k sous-groupes
Source : Chtita et al., 2017

D'autres méthodes de validation croisée existent telles que le *bootstrapping* (ré-échantillonnage de données) (Combes, 2012). La validation croisée permet ainsi d'estimer le pouvoir de prédiction du modèle envers les données du jeu d'entraînement, mais pas d'estimer la prédictivité générale du modèle. Pour déterminer à quel point le modèle est prédictif sur des données qu'il n'a jamais vues, une étape de test est réalisée à l'aide d'un jeu de test externe. Cette étape dite de validation externe, permet d'estimer l'incertitude associée aux prédictions externes (Raunio, 2011).

La performance du modèle dépend avant tout des données et des descripteurs utilisés. En effet, lorsque peu d'informations sont utilisées pour construire le modèle, la confiance du pouvoir de prédiction en est réduite. Cela est également le cas lorsque les variables indépendantes (descripteurs) sont corrélées entre elles, ou lorsque le nombre de descripteurs est trop élevé par rapport au nombre d'observations (Hemmerich et Ecker, 2020). Dans ce dernier cas, un choix

approprié d'un nombre maximum de descripteurs, soit 5 ou 7 observations pour 1 descripteur (Gramatica, 2013), permet de réduire le risque de prédictions non réalistes et hasardeuses.

1.3.2.3 *Domaine d'applicabilité*

Le troisième principe de la modélisation QSAR énoncé par l'OCDE, est que chaque modèle doit avoir un domaine d'applicabilité. Le domaine d'applicabilité est l'espace physico-chimique qui permet de considérer les prédictions obtenues comme fiables et dignes de confiance lorsque les molécules associées se trouvent dans cet espace (Raunio, 2011). Un modèle fiable et robuste développé avec certains types de composés chimiques, n'est en effet pas applicable à tous les composés chimiques qui existent, au risque de faire des prédictions hasardeuses et non précises (Hemmerich et Ecker, 2020).

Ce sont les propriétés structurelles, physico-chimiques ou les mécanismes d'action des molécules du jeu d'entrainement qui déterminent et définissent le domaine d'applicabilité. Les composés chimiques qui sont similaires aux composés du jeu d'entrainement et qui se situent donc dans la couverture chimique du modèle, verront ainsi leurs prédictions considérées comme fiables. À l'inverse les prédictions de composés test dont les structures sont différentes de celles des composés d'entrainement (et qui se situent donc hors du domaine d'applicabilité), ne seront pas considérées fiables, puisqu'il est impossible de déterminer si ces prédictions sont dues au hasard ou à un modèle performant (Hemmerich et Ecker, 2020). Les composés du jeu d'entrainement qui sont situés hors du domaine d'applicabilité sont appelés « outliers » (ou données aberrantes). Pour le jeu de test, les composés sont dits comme n'appartenant pas au domaine d'applicabilité (Roy et al., 2015).

Plusieurs méthodes peuvent être utilisées pour déterminer le domaine d'applicabilité, selon des approches basées sur l'intervalle des descripteurs, sur la distance, sur la géométrie, ou sur la densité de probabilité de distribution (Jaworska et al., 2005). Une méthode géométrique fréquemment employée est le « leverage » (ou effet de levier), une méthode de leviers qui utilise les variances à la variable dépendante (effet ciblé) de résidus standardisés et les moyennes de distances entre les descripteurs (h_i) (Roy et al., 2015). Un seuil est défini selon l'équation $h^*=3p/n$, p étant le nombre de descripteurs plus 1, et n le nombre d'observations. Un composé dont le levier hi est supérieur à ce seuil, est alors considéré comme étant hors du domaine d'applicabilité (Roy et al., 2015).

1.3.2.4 *Interprétation mécanistique*

Le modèle QSAR élaboré doit présenter un équilibre entre l'interprétation du modèle, sa complexité et son pouvoir de prédiction. La relation entre la variable dépendante et les descripteurs n'est pas toujours simple, et doit être retranscrite par l'algorithme développé, sur lequel repose le modèle (Raunio, 2011). Il peut être difficile de déterminer les mécanismes sous-jacents à un modèle et pour cela, le choix de descripteurs interprétables peut aider à la compréhension de ces mécanismes, voire permettre d'expliquer les processus chimiques représentés par le modèle. De par la complexité de certains descripteurs, l'interprétation du modèle n'est parfois pas possible, notamment lorsque celui-ci est développé avec un grand nombre de variables indépendantes. De plus la corrélation entre les descripteurs et l'activité étudiées n'est pas toujours pertinente, notamment lorsqu'il n'existe pas de relation mécanistique entre les deux au niveau biologique, cela pouvant mener à un faible taux de prédiction même pour des composés incluent dans le domaine d'applicabilité (Veerasamy et al., 2011). Une sélection optimale doit ainsi comprendre un nombre adéquat de descripteurs qui sont corrélés avec la variable d'intérêt et avec les observations (pour éviter le surentrainement ou un sous-entrainement), afin de faciliter au maximum l'interprétation du modèle (Combes, 2012).

Un bon modèle QSAR est donc un modèle développé en transparence, avec des données et des descripteurs appropriés, un domaine d'applicabilité défini, et qui peut être interprétable.

Les méthodes *in silico* sont donc des méthodes de développement rapide, qui possèdent une grande capacité de traitement, et qui peuvent être constamment améliorées. Les modèles élaborés offrent une plus grande reproductibilité à très faible coûts, et l'intégration de méthodologies plus complexes (Hartung et Hoffman, 2009). La modélisation QSAR permet d'investiguer les effets nocifs de composés, d'estimer les paramètres toxicocinétiques, de prioriser les composés pour l'évaluation toxique, et ce de façon plus rapide et abordable, réduisant notamment les besoins en expérimentations animales (Cronin et Madden, 2010). Les modèles QSAR sont particulièrement intéressants lorsque les données toxiques sont insuffisantes, que l'expérimentation n'est pas possible, ou que seules les structures des molécules sont connues (Raunio, 2011). Les méthodes in silico viennent ainsi compléter et non remplacer les méthodes expérimentales *in vivo*, *in vitro* et *ex vivo* dans la compréhension et la prédiction des mécanismes de toxicité in vivo. Elles sont d'ailleurs rarement utilisées seules et servent souvent à appuyer poids de la preuve dans les décisions

scientifiques (Raunio, 2011). Intégrées aux approches in vivo et in vitro, les méthodes in silico permettent de standardiser l'analyse et la transparence des résultats.

1.3.3 **Études QSAR du transfert placentaire**

Peu d'études se sont intéressées à la modélisation QSAR pour le transfert placentaire des contaminants. Trois équipes ont développé des modèles QSAR exclusivement à partir de jeux de données de médicaments. En 2007, Hewitt et al., ont développé trois modèles QSAR à partir de trois jeux de données différents et d'analyses de régression linéaire multiple. Le premier jeu incluant 86 indices de clairance de médicaments obtenus par perfusion placentaire *ex vivo*, le deuxième et le troisième incluant 58 et 21 indices de transfert par perfusion placentaire *ex vivo* respectivement. La validité interne de ces trois modèles a été évaluée avec une procédure de validation croisée leave-on-out et les coefficients de détermination pour l'entrainement et la validation croisée suivants ont été obtenus, respectivement : $R^2 = 0.635$ et $Q^2 = 0.576$ pour le jeu 1, $R^2 = 0.620$ et $Q^2 = 0.586$ pour le jeu 2 ($n = 58$), et $R^2 = 0.763$ et $Q^2 = 0.692$ pour le jeu 3 ($n = 21$). Aucune validation externe n'a été réalisée dans cette étude. En 2009, Giaginis et al., et en 2015, Zhang et al., ont développé des modèles QSAR en utilisant un même jeu de données *ex vivo* pour une variété de médicaments, les indices de clairance ayant été obtenus par perfusion placentaire et compilés à partir de la littérature. Ces deux études ont utilisé l'approche par régression des moindres carrés partiels, mais différaient au niveau de la sélection des descripteurs. Giaginis et al. (2009) ont effectué une sélection de 16 descripteurs optimaux à partir de 82 descripteurs initiaux, utilisant l'importance des variables par projection (VIP), le poids de chaque variable dans l'analyse des principaux composants, et la taille des coefficients des descripteurs. Zhang et son équipe (2015) ont réalisé une sélection optimale de 48 descripteurs suivant l'importance des variables par projection (VIP) de la méthode de sélection de variables de la régression des moindres carrés partiels. La validité interne a été évaluée à l'aide des procédures de validation croisée LMO (Giaginis et al., 2009) et LOO (Zhang et al., 2015) et des phases de test ont permis d'évaluer les validités externes. Giaginis et son équipe ont obtenu des coefficients de détermination de $R^2 = 0.73$ (entraînement), $Q^2 = 0.71$ (LMO), et une erreur quadratique moyenne de $RMSE_{ext} = 0.15$ pour la validation externe. Aucune donnée n'est fournie pour le coefficient de détermination pour la phase de test. Zhang et al. (2015) ont quant à eux obtenu des coefficients de détermination de $R^2 = 0.9064$ (entrainement), $Q^2 = 0.7323$ (LOO), et $R^2_{ext} = 0.7656$ (validation

externe). L'erreur quadratique moyenne pour la validation externe n'est pas précisée pour cette étude. Enfin Giaginis et al. (2009) n'ont pas défini clairement le domaine d'applicabilité du modèle développé contrairement à Zhang et al. (2009) qui ont utilisé la méthode d'effet de levier (*leverage*).

Les études de Takaku et al. (2015) et Wang et al. (2020) ont étudié le passage placentaire de médicaments et d'une famille de contaminants à partir d'un même jeu de données et d'une analyse de régression linéaire. Ainsi 55 ratios de concentrations sanguines fœto-maternelles *in vivo* compilées de la littérature ont été utilisés, avec 48 ratios de médicaments, et 7 ratios de pesticides organochlorés. L'étude de Wang et al. (2020) a utilisé une procédure de quatre étapes pour faire la sélection des descripteurs. Les validités interne par validation croisée LOO et externe des modèles ont donné des coefficients de détermination respectifs de $R^2 = 0.73$, $Q^2 = 0.71$, $R^2_{ext} = 0.51$, et de $R^2 = 0.875$, $Q^2 = 0.850$, $R^2_{ext} = 0.847$, respectivement.

Le dernier modèle présenté ici et le plus pertinent à ce mémoire est celui de Eguchi et al. (2018). Ce modèle est le premier ayant été exclusivement élaboré avec des ratios de concentrations sanguines fœto-maternelles *in vivo* de contaminants. Les ratios de 31 contaminants environnementaux de quatre familles de contaminants (dioxines, pesticides organochlorés, polychlorobiphényles et polybromodiphényléthers) ont été extraits de la littérature. Trois méthodes statistiques (forêts aléatoires, régression des moindres carrés partiels et la régression linéaire multiple) et 10 descripteurs moléculaires ont été utilisés pour élaborer trois modèles QSAR. Les résultats de validation interne avec une validation croisée 10-fold et de validité externe sont décrits par les coefficients de détermination et les erreurs quadratiques moyennes respectifs suivants: $Q^2 = 0.566$, $RMSE_{CV} = 0.0648$, $R^2_{ext} = 0.519$ et $RMSE_{ext} = 0.0514$ pour le modèle de forêts aléatoires, $Q^2 = 0.492$, $RMSE_{CV} = 0.0699$, $R^2_{ext} = 0.123$ et $RMSE_{ext} = 0.112$ pour la régression des moindres carrés partiels , et $Q^2 = 0.425$, $RMSE_{CV} = 0.0740$, $R^2_{ext} = 0.129$ et $RMSE_{ext} = 0.0897$ pour la régression linéaire multiple. Une sélection de 10 descripteurs a été réalisée à partir des valeurs des coefficients de corrélation de Spearman, où une variable est retirée de chaque couple de variables dont le coefficient de corrélation est supérieur à 0.7. Cette sélection finale, utilisée dans le développement des trois modèles, incluait deux descripteurs physico-chimiques huit descripteurs de chimie quantiques. Le poids moléculaire et la polarité ont été déterminés comme les deux paramètres les plus importants dans la prédiction du transfert transplacentaire des molécules.

1.4 Problématique

La diversité grandissante des polluants et contaminants environnementaux en lien avec l'exposition populationnelle met à l'épreuve les méthodes d'évaluation de la toxicité. L'exposition de la femme enceinte en particulier et les effets néfastes potentiels sur le fœtus rendent nécessaire l'élaboration d'outils prédictifs permettant d'évaluer rapidement la capacité des substances toxiques à traverser le placenta. La modélisation toxicologie prédictive apparaît comme un outil de choix pour étudier ce phénomène. Parmi les outils de toxicologie prédictive disponibles, les relations quantitatives de structure à activité (QSAR), permettre de lier la structure d'une molécule à ses propriétés. Les études QSAR disponibles ont été développées jusqu'à présent exclusivement sur des médicaments, ou sur des jeux mixtes n'incluant que peu de contaminants (13% du jeu de données). La seule étude traitant exclusivement de contaminants utilisait un petit jeu de données et une diversité restreinte à quatre familles de xénobiotiques toxiques, limitant l'applicabilité du modèle à plus de contaminants environnementaux. Certaines de ces études n'ont également pas respecté les principes de validation de l'OCDE en termes de modélisation QSAR, limitant leur validité et leur fiabilité. L'évaluation de l'exposition fœtale aux différents contaminants présents dans l'environnement nécessite ainsi l'élaboration d'un modèle possédant un domaine d'applicabilité plus large, en utilisant notamment des données plus diversifiées et représentatives de l'exposition réelle du fœtus.

1.5 Objectifs

Les travaux présentés dans ce mémoire visaient l'élaboration d'un modèle QSAR pour le passage placentaire de contaminants environnementaux. Plus précisément, l'article présenté visait :
i) à établir une base de données à partir des données publiées sur les concentrations de contaminants de l'environnement mesurées dans des échantillons de sang de la mère et du cordon ombilical au moment de l'accouchement;
ii) et à élaborer un modèle QSAR pour le transfert placentaire, c'est-à-dire un modèle permettant d'estimer la capacité d'une molécule à traverser la barrière placentaire (exprimée en termes de ratio entre les concentrations dans le sang au cordon ombilical et dans le sang maternel) à partir de sa structure.

2 Méthodologie

2.1 Revue de littérature et construction de la base de données

Une revue de littérature a été réalisée à partir des moteurs de recherche PubMed, ScienceDirect, Scopus et Google Scholar à l'aide des mots-clés suivants : « fetus », « fetal exposure », « toxic compound », « placental transfer », « environmental contaminants », « pesticides », et « maternal-fetal exposure ». Seules les études publiées en français et en anglais, et réalisées chez l'humain ont été retenues. La recherche bibliographique s'est concentrée sur la diversité des contaminants à l'étude, le moment de la collecte des mesures, et le type de mesures (concentrations sanguines non ajustées). Au final, 14 études ont été sélectionnées pour une diversité de contaminants (furanes, dioxines, pesticides), et pour lesquelles la collecte des sangs maternel et au cordon a été effectuée au moment de la naissance, à l'exception de trois polluants organiques persistants (PCB163, BDE209 et BDE154) dont les échantillons maternels ont été collectés au cours du premier trimestre de la grossesse. Plusieurs études ont toutefois démontré que la concentration des polluants organiques persistants restent stables tout au long de la grossesse (Vizcaino et al., 2014). Sur ces 14 études, 10 présentaient directement les ratios médians ou moyens fœto-maternels de concentrations sanguines calculés à partir de paires mère-enfants, pour un total de 84 contaminants; 2 études ne présentaient que les concentrations moyennes ou médianes dans le sang des mères et des enfants, nous permettant de calculer les ratios fœto-maternels pour 11 composés à partir de ces valeurs; enfin pour 2 études, la disponibilité des données brutes de concentrations sanguines pour chaque paire mère-enfant nous a permis de calculer les ratios médians pour 7 contaminants. Dans le cas où plusieurs ratios étaient disponibles pour le même composé, la priorité a été mise sur les radios médians, puis sur le nombre de paires mère-enfants.

2.2 Génération des descripteurs

La génération des descripteurs moléculaires pour les 105 contaminants de la base de données a été réalisée avec le logiciel Molecular Operating Environment (MOE). Les composés ionisables ont été mis sous leurs formes dominantes normalement retrouvée dans le sang au cordon où le pH est de 7.3. Une étape de minimisation d'énergie a ensuite été effectuée pour transformer

les structures 2D des molécules au format SMILES en leurs structures 3D. Le calcul des descripteurs a inclus 193 descripteurs 2D relatives aux propriétés structurelles et topologiques, et 117 descripteurs i3D, des descripteurs 3D qui dépendent des coordonnées internes et des propriétés conformationnelles des molécules. Une sélection optimale a été réalisée à partir des 310 descripteurs initiaux avec le module « QuaSAR Contingency » implémenté dans MOE. Ce module analyse statistiquement à l'aide de quatre scores d'évaluation (coefficient de contingence, test V de Cramer, incertitude entropique, coefficient de corrélation linéaire) les descripteurs les plus significativement corrélés à la variable dépendante (le ratio de transfert transplacentaire dans ce cas). La sélection finale obtenue était de 214 descripteurs 2D et i3D.

2.3 Développement des modèles QSAR

2.3.1 **Séparation du jeu de données**

Le jeu de données de 105 composés chimiques a été séparé en un jeu d'entraînement et un jeu de test selon un ratio 80 % (n = 84) – 20 % (n = 21) respectivement. La séparation a été réalisée sur la base des descripteurs afin de former le sous-groupe des composés les plus divers. La méthode du « Diverse Subset » implémentée directement dans MOE, a permis de calculer les distances euclidiennes au sein de chaque cluster de molécules et de trouver les 84 molécules les plus diverses à inclure au jeu d'entraînement. Cette technique de séparation permet ainsi d'obtenir le domaine d'applicabilité le plus large possible, et de s'assurer que les molécules du jeu de test sont similaires à celle du jeu d'entraînement afin d'augmenter la confiance dans les prédictions des modèles. Les analyses statistiques ont été réalisées à l'aide de trois outils différents : MOE, le langage de programmation Python, et le logiciel R Studio (Figure 3).

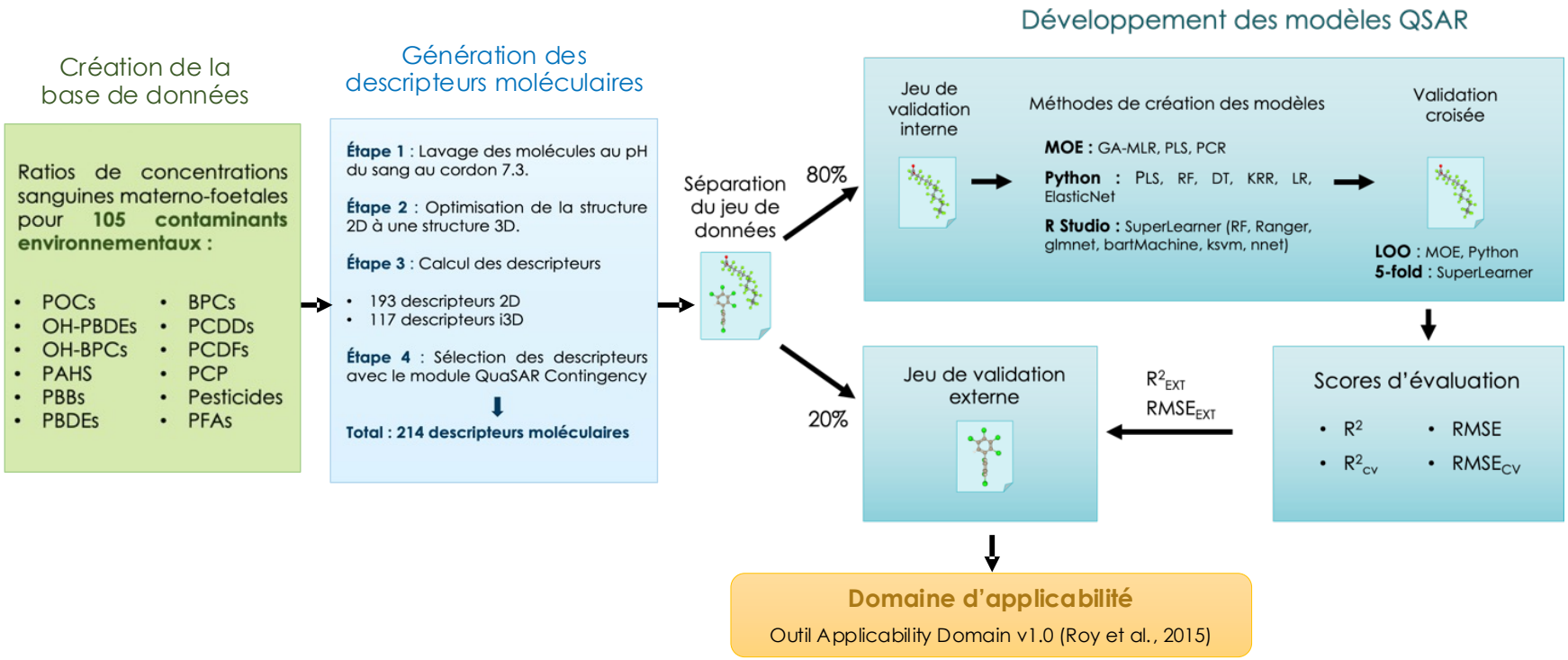

Figure 3. Schéma de la méthodologie et du flux de travail pour le développement de modèles QSAR du passage placentaire des contaminants environnementaux

2.3.2 **Molecular Operating Environment**

Le logiciel MOE a été utilisé pour développer trois modèles QSAR à l'aide d'analyses par régression des moindres carrés partiels (PLS), par algorithme génétique-régression linéaire multiple (GA-MLR) et par régression sur principales composantes (PCR).

2.3.3 **Python**

Le langage de programmation Python et la librairie Scikit-Learn ont été utilisés pour développer six modèles statistiques incluant l'analyse de régression des moindres carrés partiels (PLS), les forêts aléatoires (RF), les arbres de décisions (DT), ElasticNet, la régression Lasso (LR) et la régression Kernel Ridge (KRR).

2.3.4 **R Studio**

Le dernier modèle a été développé sur R Studio avec le package SuperLearner. L'algorithme SuperLearner permet de construire un modèle optimal à partir d'une combinaison de modèles statistiques individuels (Van Der Laan et Dudoit, 2003). Une sélection de six modèles a été réalisée sur la base de la pertinence de chaque modèle dans la prédiction finale, et inclut l'analyse par forêts aléatoires, l'analyse Ranger (une implémentation rapide des forêts aléatoires), la méthode glmnet (des modèles Lasso et ElasticNet linéaires généralisés et régularisés), l'analyse bartMachine (un support des arbres de régression additive bayésienne), l'algorithme de machine à vecteurs de support de Kernlab (ksvm) et l'analyse par réseaux neuronaux adaptés (nnet). Les valeurs par défauts ont été utilisées pour tous les modèles sélectionnés.

2.3.5 **Validation**

La validation des modèles a été effectuée dans le respect des principes de l'OCDE pour la modélisation QSAR. Dans un premier temps, une validation interne a été faite pour évaluer la qualité de l'ajustement et la robustesse du modèle. Dans un second temps une validation externe a été réalisée avec une phase de test pour évaluer la prédictivité du modèle.

La validation interne permet de s'assurer que le modèle a bien appris et qu'il pas trop ajusté aux données, c'est-à-dire qu'il n'est pas surentraîné. Le surentraînement ou surapprentissage (*overfitting* en anglais), est un concept d'intelligence artificielle et en particulier du domaine de

l'apprentissage machine. Ce phénomène apparaît lorsqu'un modèle apprend trop bien les données qui lui sont fournies lors de la phase d'entraînement. Lorsqu'il est évalué sur ces mêmes données, les prédictions sont excellentes voire parfaites. Toutefois, la performance du modèle diminue considérablement lorsqu'il doit prédire des données qu'il n'a jamais vues (Modi et al., 2012). Un bon modèle est donc un modèle qui présente un équilibre entre sa performance et son pouvoir de prédiction afin de pouvoir représenter adéquatement la relation entre les descripteurs moléculaires et l'effet étudié. À l'inverse du surentrainement, il existe également le sousentrainement (*underfitting* en anglais) qui survient lorsque les paramètres utilisés pour construire le modèle ne sont pas adaptés ou en nombre insuffisant, et qui ne permet pas au modèle d'apprendre (Figure 4) (Modi et al, 2012).

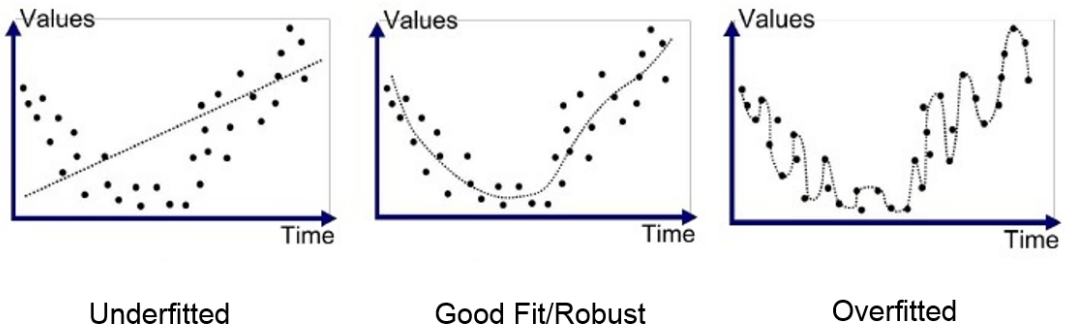

Figure 4. Illustration des phénomènes de sur- et sousapprentissage
Source : Bhande, 2018

Pour éviter le surentraînement des modèles développés, une validation croisée a été réalisée pour chaque modèle. Les 9 modèles développés avec MOE et le langage Python ont été soumis à une validation croisée leave-one-out donc avec un k = 84, tandis que le modèle SuperLearner développé avec R Studio a été soumis à une validation croisée leave-many-out avec un k = 5 (donc une division du jeu d'entrainement en 5 sous-groupes). La qualité de l'ajustement et la robustesse des modèles ont été évaluées avec deux mesures statistiques : le coefficient de détermination R^2 pour la précision, et l'erreur quadratique moyenne RMSE pour la justesse et la précision.

La validation externe a été évaluée lors de la phase de test des modèles élaborés pour évaluer leurs pouvoirs de prédiction sur des données inconnues (n = 21) qui n'ont pas participé à leurs développements. Les deux mêmes mesures statistiques (R^2 et RMSE) ont été utilisées pour évaluer la prédictivité des modèles.

2.3.6 **Domaine d'applicabilité**

Le domaine d'applicabilité permet définir l'espace physico-chimique dans lequel les prédictions de molécules s'y trouvant, sont considérées comme fiables (Raunio, 2011). Le domaine d'applicabilité étant défini par les données du jeu d'entrainement, ce domaine est le même pour tous les modèles développés. Le domaine d'applicabilité de notre jeu de données a été déterminé avec l'outil Applicability Domain v1.0 développé par Roy et al., (2015), un algorithme implémenté en Java qui est basé sur l'approche par standardisation. Cette approche veut que dans le cas d'une distribution normale des données du jeu d'entrainement, 99.7% de celles-ci se trouvent dans un espace dont les limites sont définies par la moyenne de la population ± 3 écarts-type. Toute molécule se trouvant statistiquement en dehors de cet espace est donc différente de la majorité des autres molécules, et est qualifiée de valeur aberrante ou de molécule en dehors du domaine d'applicabilité dépendamment qu'elle appartienne au jeu d'entrainement ou au jeu de test respectivement (Roy et al., 2015). L'approche par standardisation consiste en la standardisation de chaque colonne de descripteurs sur la base de la moyenne et de l'écart-type. Les valeurs standardisées de chaque descripteur i d'une observations k sont nommées S_{ki}. Lorsqu'une valeur S_{ki} est supérieure à 3, l'observation k est alors hors de l'espace moyenne ± 3 écarts-type, basé sur ce descripteur. Lorsqu'une observation a plusieurs descripteurs, si son S_i maximal est inférieur à 3, il est facile placer cette observation dans le domaine d'applicabilité. À l'inverse, une observation dont le S_i minimum est supérieur à 3 sera considérée hors du domaine d'applicabilité (Roy et al., 2015). Le classement d'une observation ayant un S_i minimal inférieur à 3 mais un S_i maximal supérieur à 3 n'est pas toujours évident. Roy et son équipe ont ainsi élaboré un algorithme utilisant le score standard (Z) qui dans le cas d'une distribution normale standardisée est de 1.28 et signifie que 90% des données ont une fréquence d'occurrence relative inférieure à 1.28 fois l'écart-type. Pour une observation k données, si la valeur (S_{new}) de la moyenne des S_i de ses descripteurs plus 1.28 fois l'écart-type est inférieure à 3, alors 90% des valeurs S_i de cette observation sont supposées également inférieures à 3 et l'observation est considérée être dans le domaine d'applicabilité. Ainsi l'outil Applicability Domain v1.0 a permis dans un premier temps de calculer les S_i des 214 descripteurs pour les 84 molécules du jeu d'entrainement et les 21 molécules du jeu de test. Dans un second temps, les valeurs des S_i ont été analysées pour chaque observation k suivant les trois cas suivants (Figure 5) :

a) Si le S_i maximal est inférieur ou égal à 3, le composé se trouve dans le domaine d'applicabilité.
b) Si le S_i maximal est supérieur à 3, et si le S_i minimal est également supérieur à 3, le composé est une valeur aberrante ou se trouve hors du domaine d'applicabilité.
c) Enfin si le S_i maximal est supérieur à 3 mais que le S_i minimal est inférieur à 3, alors le S_{new} est calculé. Si le S_{new} est inférieur ou égal à 3, alors le composé est dans le domaine d'applicabilité.

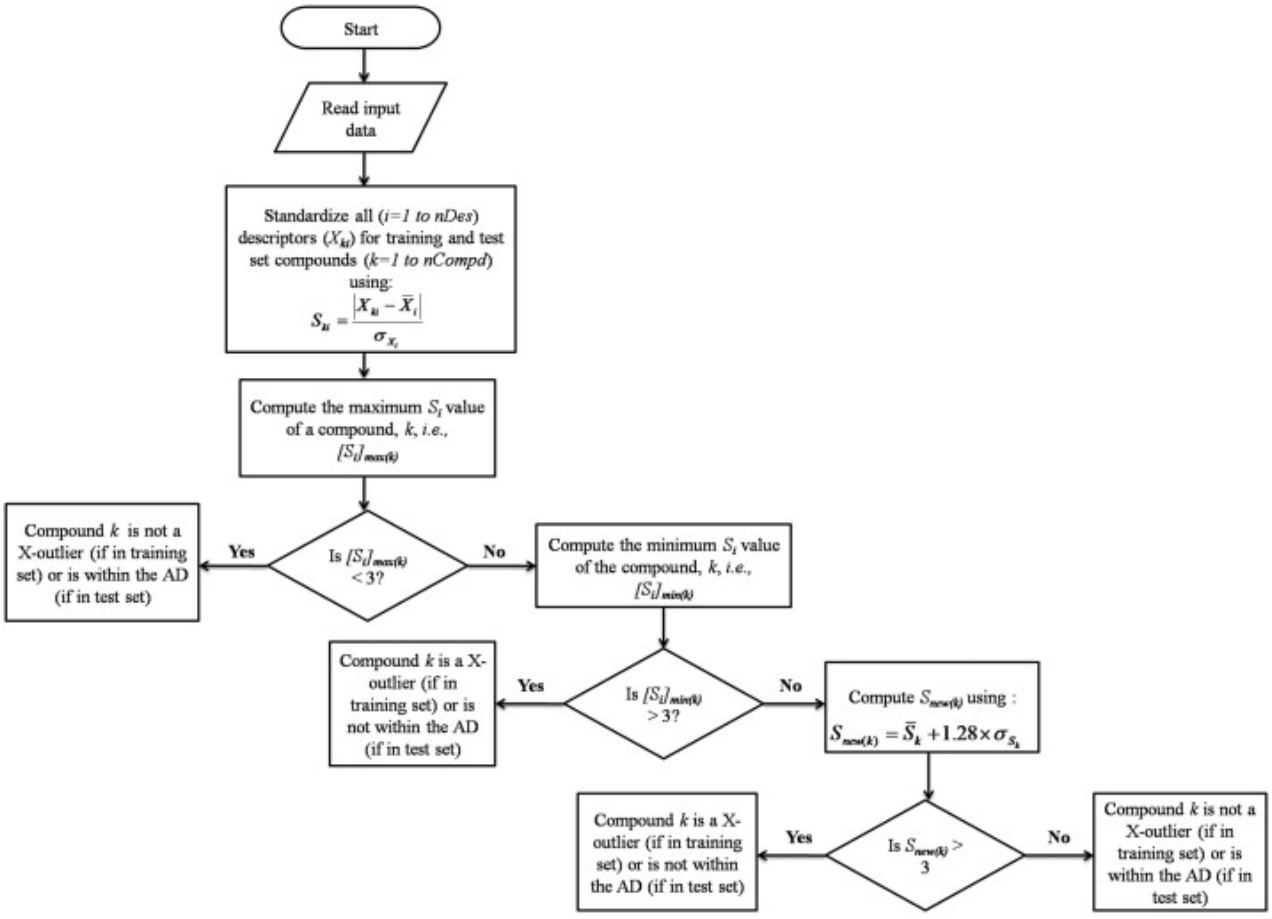

Figure 5. Schéma de l'algorithme développé par Roy et al. (2015) utilisant l'approche par standardisation pour la détermination du domaine d'applicabilité.

3 Article scientifique

CONTRIBUTION DES AUTEURS

Laura Lévêque

- A créé la base de données
- A développé les modèles QSAR
- A analysé et interprété les résultats
- A participé à la rédaction de l'article

Nadia Tahiri

- A créé les scripts des algorithmes pour les modèles QSAR
- A analysé et interprété les résultats
- A participé à la rédaction de l'article

Michael-Rock Goldsmith

- A analysé et interprété les résultats
- A participé à la rédaction de l'article

Marc-André Verner

- A élaboré le protocole et supervisé la recherche
- A analysé et interprété les résultats
- A participé à la rédaction de l'article

Quantitative Structure-Activity Relationship (QSAR) Modeling to Predict the Transfer of Environmental Chemicals across the Placenta

Laura Lévêque[1,2], Nadia Tahiri[1,2], Michael-Rock Goldsmith[3], Marc-André Verner[1,2,*].

1. Center for Public Health Research, Montreal, QC

2. Department of Occupational and Environmental Health, School of Public Health, Université de Montréal, Montreal, QC

3. Congruence Therapeutics, Chapel Hill NC, 27514 (rgoldsmith@congruencetx.com)

*. Corresponding author: marc-andre.verner.1@umontreal.ca

3.1 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. Fetal 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. The 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. The inclusion of all test chemicals by the domain of applicability demonstrated the reliability and relevance of the model predictions. The results obtained demonstrate that QSAR modeling can help quantify placental transfer of environmental chemicals.

Keywords: QSAR, contaminants, placental transfer, modeling, *in silico*.

3.2 Introduction

Epidemiological and toxicological studies have demonstrated that several contaminants are able to cross the placental barrier and reach the developing organism (Aylward et al. 2014), potentially leading to adverse health effects. 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 (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:mother concentration ratios (Aylward et al. 2014). Toxicological experiments have been conducted in animals to study the characteristics of *in vivo* placental transfer for toxic compounds and drugs. Non-human primates, and animals such as sheep, guinea pigs, and rodents have been used due to either their placentation similarities to human placenta, or the ability to study easily and rapidly their pregnancy events (Grigsby, 2016). *Ex vivo* placental perfusion models have been developed 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 properties of xenobiotics (Myren et al. 2007). Although all of the approaches presented above provide information on placental transfer, they all have limitations including duration of experiments, interspecific extrapolation, and costs. The increasing diversity of chemicals in the environment demands the development of new ways to rapidly and precisely quantify fetal exposure to such compounds, and their ability to cross the placental barrier.

In silico predictive toxicology appears to be a promising alternative to the experimental approaches described above. 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 correlation with its structural features (Tropsha, 2010). 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. A large number of molecular descriptors can be obtained in experiments or calculated by relevant

softwares like SwissADME web tool (Daina et al. 2017), Molecular Operating Environment (MOE) (MOE, 2019), PaDEL-descriptor (Yap, 2011), and E-Dragon (Mauri et al. 2006) based on Simplified molecular-input line-entry system (SMILES) (Weininger, 1988).

The prediction of placental transfer of drugs 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. In 2007, Hewitt et al. developed a QSAR model based on the clearance index of 86 heterogeneous drugs compiled from literature. *Ex vivo* data were used by Giaginis et al. (2009) to build a QSAR model for drugs transport across the placental barrier, with 88 clearance indices found in the literature. The same database was used later by Zhang et al. in 2015 to develop a predictive model using a partial least squares regression (PLS) procedure (Tobias, 1995). Takaku et al. (2015) created a multiple linear regression (MLR) model (Chakraborty and Goswami, 2017) based on 55 fetal-maternal ratios compiled or calculated from published studies, for a variety of drugs and a few organochlorine pesticides (OCPs). In 2020, Wang et al. (2020) used the same database to build predictive multiple linear regression models, improving the general predictive ability of the model with four steps feature selection method. In 2018, Eguchi et al. 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 maternal-fetal transfer rates.

While 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 drugs. Therefore, the applicability of these models to predict placental transfer for the diverse range of environmental contaminants is questionable. The development of a QSAR model with a larger environmental chemical database should extend this domain of applicability and provide a more suitable model to answer the need/demand for fetal exposure risk assessment.

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

3.3 Materials and methods

3.3.1 **Data**

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 », et « maternal-fetal exposure ». The search was limited to publications written in French or English, and to studies conducted in humans. A database was created compiling maternal-fetal blood concentrations ratios of 105 environmental contaminants from published epidemiologic/biomonitoring studies ranging from 2002 to 2020. Chemicals included 11 organochlorine pesticides, 7 pesticides, 2 hydroxylated polybrominated diphenyl ethers (OH-PBDEs), 4 polybrominated diphenyl ethers (PBDEs), 19 polycyclic aromatic hydrocarbons (PAHs), one polybrominated biphenyl (PBB), 6 hydroxylated polychlorinated biphenyls (OH-PCBs), 29 polychlorinated biphenyls (PCBs), 5 polychlorinated dibenzo-p-dioxins (PCDDs), 7 polychlorinated dibenzofurans (PCDFs), and 17 per- and polyfluoroalkyl substances (PFAS). All ratios have been calculated on maternal and cord blood samples from mother-infants pairs collected at or around deliver 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 whose concentrations have been shown to be relatively stable throughout pregnancy (Vizcaino et al. 2014). When median/mean concentration ratios were reported, these values were used (n = 87 chemicals). Where ratios were not reported in the publication, mean/median maternal and cord blood levels were used to calculate ratios (n = 11 chemicals). Finally, when ratios were not reported in the publication but matched concentrations of maternal/cord serum of each pair mother-infant were available, those data were used to calculate the median ratios (n = 7 chemicals). If more than one ratio was available for the same chemical, priority was set on the median ratio first, then on the number of pairs mother-infant when several median ratios were available. Chemicals included in the database are shown in Table 1.

Table 1. Maternal/fetal blood concentrations ratios, ratios Ln, compound names, and SMILES for 105 environmental contaminants.

SMILES	Compound name	Ratio	Ratio Ln	Reference
OCPs				
C1=CC(=CC=C1C(=C(Cl)Cl)C2=CC=C(C=C2)Cl)Cl	4,4'-DDE	0.4	-0.92	Morello-Frosch et al., 2016
C1(C(C(C(C(C1Cl)Cl)Cl)Cl)Cl)Cl	b-HCH	0.3	-1.20	Morello-Frosch et al., 2016
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	HCB	0.6	-0.51	Morello-Frosch et al., 2016
C12C(C(C3(C1O3)Cl)Cl)C4(C(=C(C2(C4(Cl)Cl)Cl)Cl)Cl)Cl	Oxychlordane	0.1	-2.30	Morello-Frosch et al., 2016
C12C(C(C(C1Cl)Cl)Cl)C3(C(=C(C2(C3(Cl)Cl)Cl)Cl)Cl)Cl	Transnonachlor	0.3	-1.20	Morello-Frosch et al., 2016
C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)C(Cl)Cl)Cl	o,p'-DDD	0.34	-1.08	Yin et al., 2019
C1=CC=C(C(=C1)C(=C(Cl)Cl)C2=CC=C(C=C2)Cl)Cl	o,p'-DDE	0.37	-0.99	Yin et al., 2019
C1=CC=C(C(=C1)C(C2=CC=C(C=C2)Cl)C(Cl)(Cl)Cl)Cl	o,p'-DDT	0.41	-0.89	Yin et al., 2019
C1=CC(=CC=C1C(C2=CC=C(C=C2)Cl)C(Cl)Cl)Cl	p,p'-DDD	0.37	-0.99	Yin et al., 2019
C1=CC(=CC=C1C(C2=CC=C(C=C2)Cl)C(Cl)(Cl)Cl)Cl	p,p'-DDT	0.39	-0.94	Yin et al., 2019
OH-PBDEs				
C1=CC(=C(C=C1Br)Br)OC2=C(C=C(C(=C2)Br)O)Br	4'-OH-BDE-49	0.4	-0.92	Morello-Frosch et al., 2016
C1=CC(=C(C=C1Br)Br)OC2=C(C=C(C(=C2)O)Br)Br	5-OH-BDE-47	1.1	0.10	Morello-Frosch et al., 2016
OH-PCBs				
C1=C(C(=CC(=C1Cl)Cl)Cl)C2=CC(=C(C(=C2Cl)O)Cl)Cl	3-OH-CB153	0.68	-0.39	Park et al., 2008
C1=CC(=C(C(=C1C2=CC(=C(C(=C2Cl)O)Cl)Cl)Cl)Cl)Cl	3'-OH-CB138	1.01	0.01	Park et al., 2008
C1=CC(=C(C=C1C2=CC(=C(C(=C2Cl)Cl)O)Cl)Cl)Cl	4-OH-CB107	0.57	-0.56	Park et al., 2008
C1=C(C(=CC(=C1Cl)Cl)Cl)C2=CC(=C(C(=C2Cl)Cl)O)Cl	4-OH-CB146	0.78	-0.25	Park et al., 2008
C1=C(C(=CC(=C1Cl)Cl)Cl)C2=C(C(=C(C(=C2Cl)Cl)O)Cl)Cl	4-OH-CB187	0.68	-0.39	Park et al., 2008
C1=C(C(=C(C(=C1Cl)O)Cl)Cl)C2=CC(=C(C(=C2Cl)Cl)Cl)Cl	4'-OH-CB172	1.03	0.03	Park et al., 2008
PAHs				

SMILES	Compound name	Ratio	Ratio Ln	Reference
CC1=CC=CC2=CC=CC=C12	1-MNAP	1.3	0.26	Sexton et al., 2011
CC1=C2C=CC3=CC=CC=C3C2=CC=C1	1-MPA	2	0.69	Sexton et al., 2011
CC1=C2C=C(C(=CC2=CC=C1)C)C	1,6,7-TMNAP	1.5	0.41	Sexton et al., 2011
CC1=CC2=CC=CC=C2C=C1	2-MNAP	1.4	0.34	Sexton et al., 2011
CC1=CC2=C(C=C1)C=C(C=C2)C	2,6-DMNAP	2	0.69	Sexton et al., 2011
C1=CC=C(C=C1)C2=CC=CC=C2	Biphenyl	1.4	0.34	Sexton et al., 2011
C1=CC=C2C(=C1)C3=CC=CC=C3O2	Dibenzofuran	1.2	0.18	Sexton et al., 2011
C1=CC=C2C(=C1)C3=CC=CC=C3S2	Dibenzothiophene	1.4	0.34	Sexton et al., 2011
C1=CC=C2C=CC=CC2=C1	Naphtalene	3.1	1.13	Sexton et al., 2011
C1CC2=CC=CC3=C2C1=CC=C3	Acenaphtene	0.26	-1.35	Zhang et al., 2017
C1=CC2=C3C(=C1)C=CC3=CC=C2	Acenaphtylene	0.39	-0.94	Zhang et al., 2017
C1=CC=C2C=C3C=CC=CC3=CC2=C1	Anthracene	0.42	-0.87	Zhang et al., 2017
C1=CC=C2C(=C1)C=CC3=CC4=CC=CC=C4C=C32	Ben(a)anthracene	0.35	-1.05	Zhang et al., 2017
C1=CC=C2C=C3C4=CC=CC5=C4C(=CC=C5)C3=CC2=C1	Benzo(b+k)fluoranthene	0.4	-0.92	Zhang et al., 2017
C1=CC=C2C(=C1)C=CC3=C2C=CC4=CC=CC=C43	Chrysene	0.35	-1.05	Zhang et al., 2017
C1=CC=C2C(=C1)C3=CC=CC4=C3C2=CC=C4	Fluoranthene	0.34	-1.08	Zhang et al., 2017
C1C2=CC=CC=C2C3=CC=CC=C31	Fluorene	0.29	-1.24	Zhang et al., 2017
C1=CC=C2C(=C1)C=CC3=CC=CC=C32	Phenanthrene	0.36	-1.02	Zhang et al., 2017
C1=CC2=C3C(=C1)C=CC4=CC=CC(=C43)C=C2	Pyrene	0.43	-0.84	Zhang et al., 2017
PBBs				
C1=C(C(=CC(=C1Br)Br)Br)C2=CC(=C(C=C2Br)Br)Br	BB153	0.18	-1.71	Frederiksen et al., 2010
PBDEs				
C1=CC(=CC=C1OC2=C(C=C(C=C2)Br)Br)Br	BDE28	0.45	-0.80	Frederiksen et al., 2010
C1=CC(=C(C=C1Br)Br)OC2=C(C=C(C=C2Br)Br)Br	BDE100	0.3	-1.20	Morello-Frosch et al., 2016
C1=C(C(=CC(=C1Br)Br)Br)OC2=CC(=C(C=C2Br)Br)Br	BDE153	0.2	-1.61	Morello-Frosch et al., 2016

SMILES	Compound name	Ratio	Ratio Ln	Reference
C1=CC(=C(C=C1Br)Br)OC2=C(C=C(C=C2)Br)Br	BDE47	0.4	-0.92	Morello-Frosch et al., 2016
C1=CC(=C(C=C1Br)Br)OC2=CC(=C(C=C2Br)Br)Br	BDE99	0.3	-1.20	Morello-Frosch et al., 2016
C1=C(C=C(C(=C1Br)OC2=CC(=C(C=C2Br)Br)Br)Br)Br	BDE154	0.46	-0.78	Vizcaino et al., 2014
C1(=C(C(=C(C(=C1Br)Br)Br)Br)Br)OC2=C(C(=C(C(=C2Br)Br)Br)Br)Br	BDE209	0.8	-0.22	Vizcaino et al., 2014
PCBs				
C1=CC(=C(C=C1Cl)Cl)C2=CC(=C(C=C2Cl)Cl)Cl	PCB99	0.33	-1.12	Covaci et al., 2002
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)C2=C(C(=C(C(=C2Cl)Cl)Cl)Cl)Cl	DecaCBs	0.09	-2.41	Mori et al., 2014
C1=C(C(=C(C(=C1Cl)Cl)Cl)Cl)C2=C(C(=C(C(=C2Cl)Cl)Cl)Cl)Cl	NonaCBs	0.11	-2.21	Mori et al., 2014
C1=C(C(=C(C(=C1Cl)Cl)Cl)Cl)C2=CC(=C(C(=C2Cl)Cl)Cl)Cl	OctaCBs	0.12	-2.12	Mori et al., 2014
C1=CC(=CC=C1C2=CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl	PCB114	0.2	-1.61	Mori et al., 2014
C1=CC(=C(C=C1Cl)Cl)C2=CC(=C(C(=C2)Cl)Cl)Cl	PCB123	0.22	-1.51	Mori et al., 2014
C1=CC(=C(C=C1C2=CC(=C(C(=C2)Cl)Cl)Cl)Cl)Cl	PCB126	0.18	-1.71	Mori et al., 2014
C1=CC(=C(C=C1C2=CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl)Cl	PCB156	0.16	-1.83	Mori et al., 2014
C1=CC(=C(C(=C1C2=CC(=C(C(=C2)Cl)Cl)Cl)Cl)Cl)Cl	PCB157	0.18	-1.71	Mori et al., 2014
C1=C(C=C(C(=C1Cl)Cl)Cl)C2=CC(=C(C=C2Cl)Cl)Cl	PCB167	0.18	-1.71	Mori et al., 2014
C1=C(C=C(C(=C1Cl)Cl)Cl)C2=CC(=C(C(=C2)Cl)Cl)Cl	PCB169	0.15	-1.90	Mori et al., 2014
C1=C(C=C(C(=C1Cl)Cl)Cl)C2=CC(=C(C(=C2Cl)Cl)Cl)Cl	PCB189	0.13	-2.04	Mori et al., 2014
C1=CC(=C(C=C1C2=CC(=C(C=C2)Cl)Cl)Cl)Cl	PCB77	0.33	-1.11	Mori et al., 2014
C1=CC(=CC=C1C2=CC(=C(C(=C2)Cl)Cl)Cl)Cl	PCB81	0.27	-1.31	Mori et al., 2014
C1=CC(=CC=C1C2=CC(=C(C=C2Cl)Cl)Cl)Cl	TetraCBs	0.24	-1.43	Mori et al., 2014
C1=CC=C(C=C1)C2=C(C(=C(C=C2)Cl)Cl)Cl	TriCBs	0.59	-0.53	Mori et al., 2014
C1=CC(=C(C=C1C2=C(C(=CC(=C2Cl)Cl)Cl)Cl)Cl)Cl	PCB163	0.23	-1.47	Fisher et al., 2016
C1=CC(=C(C=C1C2=C(C(=C(C=C2)Cl)Cl)Cl)Cl)Cl	PCB105	0.16	-1.83	Park et al., 2008
C1=CC(=C(C=C1C2=CC(=C(C=C2Cl)Cl)Cl)Cl)Cl	PCB118	0.12	-2.12	Park et al., 2008
C1=CC(=C(C(=C1C2=CC(=C(C=C2Cl)Cl)Cl)Cl)Cl)Cl	PCB138	0.21	-1.56	Park et al., 2008

SMILES	Compound name	Ratio	Ratio Ln	Reference
C1=C(C(=CC(=C1Cl)Cl)Cl)C2=CC(=C(C=C2Cl)Cl)Cl	PCB153	0.19	-1.66	Park et al., 2008
C1=CC(=C(C(=C1C2=CC(=C(C(=C2Cl)Cl)Cl)Cl)Cl)Cl)Cl)Cl	PCB170	0.16	-1.83	Park et al., 2008
C1=C(C(=CC(=C1Cl)Cl)Cl)C2=CC(=C(C(=C2Cl)Cl)Cl)Cl	PCB180	0.18	-1.71	Park et al., 2008
PCDDs				
C1=C2C(=C(C(=C1Cl)Cl)Cl)OC3=C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	1,2,3,4,6,7,8-HeptaCDD	0.19	-1.66	Mori et al., 2014
C1=C2C(=CC(=C1Cl)Cl)OC3=C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	1,2,3,4,7,8-HexaCDD	0.22	-1.51	Mori et al., 2014
C1=C2C(=CC(=C1Cl)Cl)OC3=C(C(=C(C=C3O2)Cl)Cl)Cl	1,2,3,7,8-PentaCDD	0.23	-1.47	Mori et al., 2014
C1=C2C(=CC(=C1Cl)Cl)OC3=CC(=C(C=C3O2)Cl)Cl	2,3,7,8-TetraCDD	0.28	-1.27	Mori et al., 2014
C12=C(C(=C(C(=C1Cl)Cl)Cl)Cl)OC3=C(O2)C(=C(C(=C3Cl)Cl)Cl)Cl	OctaCDD	0.11	-2.21	Mori et al., 2014
PCDFs				
C1=C2C3=C(C(=C(C(=C3Cl)Cl)Cl)Cl)OC2=C(C(=C1Cl)Cl)Cl	1,2,3,4,6,7,8-HeptaCDF	0.33	-1.11	Mori et al., 2014
C1=C2C(=CC(=C1Cl)Cl)OC3=C2C(=C(C(=C3Cl)Cl)Cl)Cl	1,2,3,4,7,8-HexaCDF	0.25	-1.39	Mori et al., 2014
C1=C2C3=C(C(=C(C=C3OC2=C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	1,2,3,6,7,8-HexaCDF	0.3	-1.20	Mori et al., 2014
C1=C2C(=CC(=C1Cl)Cl)OC3=CC(=C(C(=C23)Cl)Cl)Cl	1,2,3,7,8-PentaCDF	0.33	-1.11	Mori et al., 2014
C1=C2C3=CC(=C(C(=C3OC2=C(C(=C1Cl)Cl)Cl)Cl)Cl)Cl	2,3,4,6,7,8-HexaCDF	0.27	-1.31	Mori et al., 2014
C1=C2C3=CC(=C(C(=C3OC2=CC(=C1Cl)Cl)Cl)Cl)Cl	2,3,4,7,8-PentaCDF	0.22	-1.51	Mori et al., 2014
C1=C2C3=CC(=C(C=C3OC2=CC(=C1Cl)Cl)Cl)Cl	2,3,7,8-TetraCDF	0.42	-0.87	Mori et al., 2014
PCP				
C1(=C(C(=C(C(=C1Cl)Cl)Cl)Cl)Cl)O	PCP	1.1	0.10	Park et al., 2008
Pesticides				
CC(C)OC1=CC=CC=C1O	2-Isopropoxyphenol	1.13	0.12	Whyatt et al., 2003
CC1(OC2=C(O1)C(=CC=C2)OC(=O)NC)C	Bendiocarb	0.84	-0.17	Whyatt et al., 2003
CCOP(=S)(OCC)OC1=NC(=C(C=C1Cl)Cl)Cl	Chlorpyrifos	0.98	-0.02	Whyatt et al., 2003
CCOP(=S)(OCC)OC1=NC(=NC(=C1)C)C(C)C	Diazinon	0.85	-0.16	Whyatt et al., 2003

SMILES	Compound name	Ratio	Ratio Ln	Reference
C1=C(C=C(C(=C1Cl)N)Cl)[N+](=O)[O-]	Dicloran	1.06	0.06	Whyatt et al., 2003
C1=CC=C2C(=C1)C(=O)NC2=O	Phthalimide	0.87	-0.14	Whyatt et al., 2003
C1C=CCC2C1C(=O)NC2=O	Tetrahydrophthalimide	0.91	-0.09	Whyatt et al., 2003
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	N-EtFOSAA	1.2	0.18	Morello-Frosch et al., 2016
CN(CC(=O)O)S(=O)(=O)C(C(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	N-MeFOSAA	0.9	-0.11	Morello-Frosch et al., 2016
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	PFOSA	1.1	0.10	Morello-Frosch et al., 2016
C1=CC(=CC=C1C(=O)O)F	PFBA	1.67	0.51	Li et al., 2020
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	PFHpS	0.8	-0.22	Li et al., 2020
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)(F)F)(F)F)(F)F)O	PFTeDA	4	1.39	Li et al., 2020
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)(F)F)(F)F)(F)F)O	PFTrDA	1.38	0.32	Li et al., 2020
C(CS(=O)(=O)O)C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F	6:2FTS	1.66	0.51	Yang et al., 2016
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)(F)F)O	PFDA	0.25	-1.39	Yang et al., 2016
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)(F)F)(F)F)O	PFDoA	0.51	-0.67	Yang et al., 2016
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	PFHxS	0.35	-1.05	Yang et al., 2016
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)(F)F)O	PFNA	0.43	-0.84	Yang et al., 2016
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	PFOA	0.65	-0.43	Yang et al., 2016
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	PFOS	0.29	-1.24	Yang et al., 2016
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)(F)F)(F)F)O	PFUnA	0.27	-1.31	Yang et al., 2016
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)O	PFHpA	1.20	0.18	Zhang et al., 2013
C(=O)(C(C(C(C(C(F)(F)F)(F)F)(F)F)(F)F)(F)F)(F)F)O	PFHxA	1.07	0.07	Zhang et al., 2013

3.3.2 **Descriptors calculation**

The Molecular Operating Environment (MOE) software 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 then applied to the 310 descriptors calculated to select the optimal descriptors for the QSAR model 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).

3.3.3 **Model development**

The dataset of 105 chemicals was separated into training (80%; $n = 84$) and test sets (20%; $n = 21$) using the diverse subset method based on the calculated descriptors (MOE, 2019). 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, and R Studio software (Figure 1).

3.3.3.1 ***Molecular Operating Environment***

Three models were built in MOE by Partial Least Square (PLS), Genetic Algorithm - Multiple Linear Regression (GA-MLR), and Principal Component Regression (PCR) analysis. PLS is a method frequently used in QSAR modelling to predict a variable explained by a large number of factors. The algorithm searches for manifest factors as well as latent factors, the latter being mostly responsible for variations of the predicted variable (Tobias, 1995). Although models developed with PLS have a high predictive power, their mechanistic and biological interpretations can be hard to understand (Saxena and Prathipati, 2003). PLS method is similar to MLR method, but often preferred when there is a high number of independent variables or factors. Indeed, the MLR technique consists of predicting the relationship between a dependent variable and two or more independent variables. However, QSAR studies have demonstrated that with a large number

of descriptors compared to the observations, an overfitting phenomenon can occur and significantly reduce the predictive power of the model. Hence a powerful hybrid method combining genetic algorithms (GA) and MLR has been proposed to solve variable selection and facilitate the resolution of optimization problems (Saxena and Prathipati, 2003). Genetic algorithms are optimization techniques created by John Holland in 1992 (Katoch et al. 2020). Mimicking natural evolution concept, the GA-MLR method searches for approximate solutions by genetic operations (mutation, selection, crossover), accelerating and facilitating the optimization process (McCall, 2005). The third model was developed with PCR analysis, a method based on principal component analysis (PCA) that aims to minimize correlation between variables by reducing the size of the dataset, and to maximize the variance.

3.3.3.2 *Python language*

The dataset also was subjected to six statistical models implemented within a Python language, including Decision tree (DT), Random Forest (RF), Ridge Regression, Lasso regression, ElasticNet, and PLS.

Decision trees (DT) are highly flexible, supervised, and non-parametric machine learning models that are readily interpretable and well-suited to regression problems. These models suffer from a well-known sensitivity to the data used in their construction. Finding an optimal tree that minimizes the number of segmentation criteria is a difficult problem in terms of complexity theory and enhances therefore the need of a heuristic method. To reduce overfitting during testing, the smallest tree depth needs to be used.

The aim of the random forest (RF) algorithm (Breiman, 2001) is to retain most of the strengths of DT 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 optimize the final model and avoid the 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 (Hoerl and Kennard, 1970; McDonald, 2009) 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 (Tibshirani, 1996) 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 (Zou and Hastie, 2005) 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.

3.3.3.3 *SuperLearner with R Studio*

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 (Van Der Laan et Dudoit, 2003). 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 discussed QSAR prediction algorithm with the dataset used, are freely available at: https://github.com/TahiriNadia/QSAR_PTR.

3.3.3.4 *Validation and applicability domain*

Validation is a crucial step in acceptance of a QSAR model for regulatory purposes. In 2004, the Organization for Economic Co-operation and Development (OECD) established 5

Principles for (Q)SAR validation that any QSAR model should follow: i) a define 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 (Worth et al. 2005). 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 correlation coefficient 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, in order to predict the variable 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, they cannot account for the predictivity of the model, i.e., its ability to predict new data not included in its development. External validation techniques consist in determining the predictive power of the model by comparing predicted and observed data for a test set of compounds. Predictivity of all models was assessed with a coefficient of determination R^2_{ext}, a relative measure of fit, and $RMSE_{ext}$, an absolute measure of fit (Figure 1).

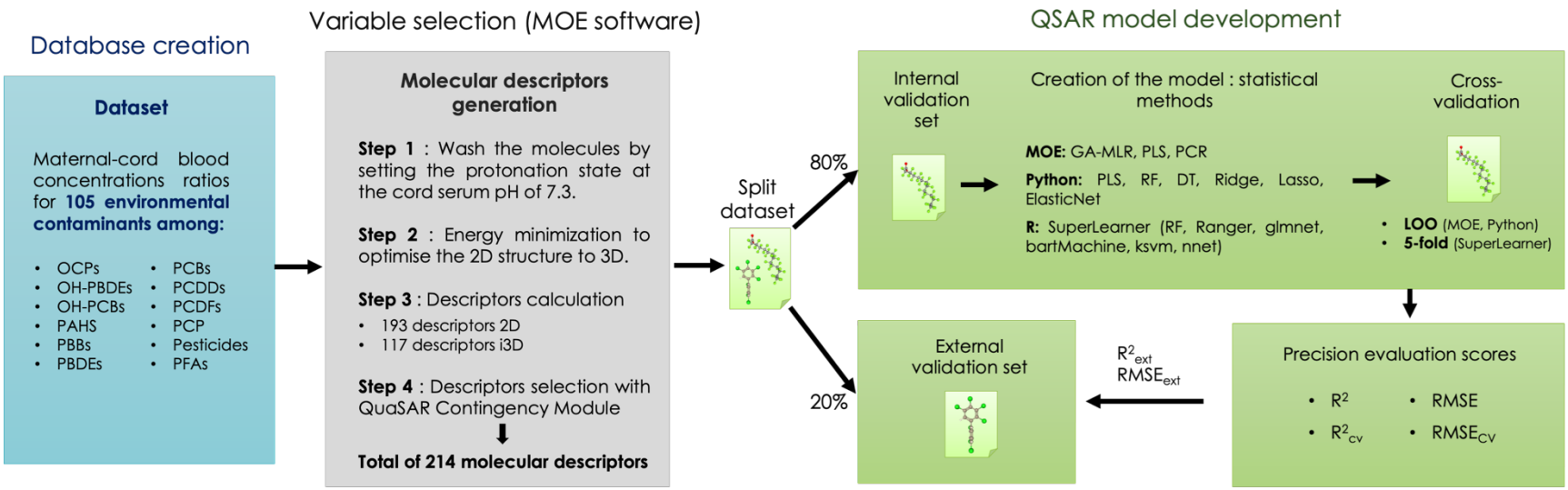

Figure 1. QSAR model workflow for the prediction of the placental transfer of environmental chemicals.

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 R Studio 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 and 5-fold cross validation techniques, and external validation was computed with the test set (green section). Models goodness-of-fit, robustness and predictivity were assessed with correlation of determination (R^2) and root mean square error (RMSE).

Validation techniques are fundamental to assess the performance of a model and are closely related to its applicability domain as a good model should not only give accurate but also reliable predictions. The third principle of OECD for QSAR modelling recommends that each model have a defined applicability domain (AD), i.e., physical-chemical space where predictions should fall within (Roy et al. 2015). 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 (Roy et al. 2015). 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 (Veerasamy et al. 2011). 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. (2015). 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.

3.4 Results

An optimal selection of 2D and 3D 214 descriptors was performed with the MOE module QuaSAR contingency using bivariate analysis in order to find the independent variables the most significantly correlated to the dependent variable and important to the model's development.

Molecular Operating Environment was first used to calculate 310 2D et 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 bound 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.

Ratios were Ln-transformed prior to model development and testing. Overall, we developed ten models with Molecular Operating Environment (MOE) software, Python and R programming languages. Statistical results for internal and external validation for all the models are summarized in Table 2. Evaluation scores included coefficient 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. The next sections present models developed, and results obtained with each method.

3.4.1 **Molecular Operating Environment**

The chemical compounds set was used to develop 3 models using the MOE software (PLS, GA-MLR, and PCR) (Figure 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 through the LOO-CV and external validation step. Among the three models, the PLS analysis obtained the highest predictive performance through the external validation test ($R^2_{ext} = 0.73$ and $RMSE_{ext} = 0.32$).

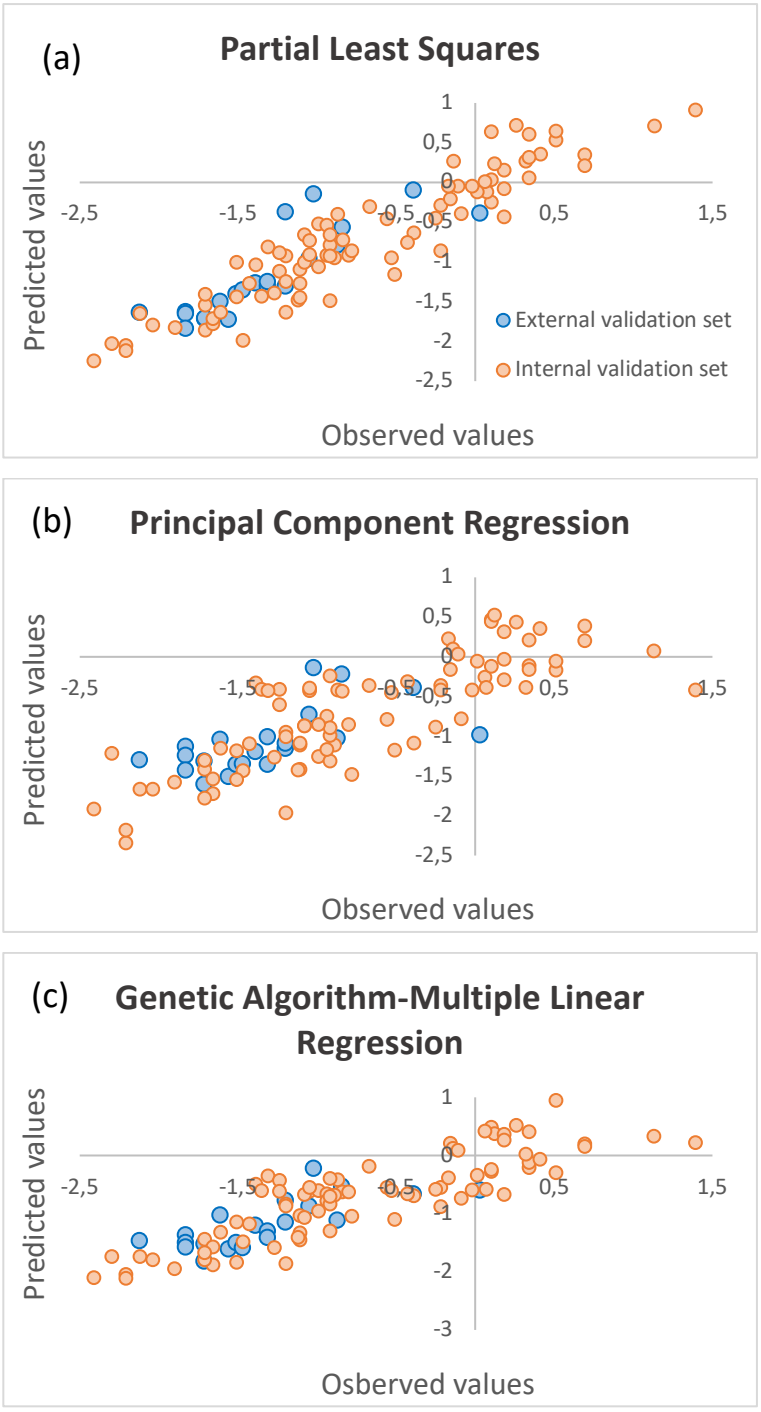

Figure 2. Comparison of predicted and observed ratios for the internal validation set (orange circle) and the external validation set (blue circle) using the MOE software. Three models have been developed. The results of the Partial Least Squares model are indicated in (a), Principal Component Regression in (b), and Genetic Algorithm-Multiple Linear Regression in (c).

3.4.2 **Python language**

We developed six models in Python version 3.6 (Random Forest, Lasso Regression, Partial Least Squares, ElasticNet, Kernel Ridge Regression) (Figure 3). The PLS model provides a point of comparison between the 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 performance for cross-validation and external validation. However, the results of the training set show an opposite performance, which suggests that the PLS model developed with the Python script is overfitting. The selection of attributes was performed with the *selectkbest* method provided with the *sklearn package* in Python. The method relies on the Pearson correlation coefficient between one attribute and the predictor (Ratio Ln). The results indicate that among the 214 initial descriptors 20 are the most meaningful and best interpreted the descriptor (see supplementary materials). The decision tree was built with a maximum depth fixed at 11, given that beyond this threshold the scores of R^2 stabilize, and under, results are deteriorated in the test stage. This algorithm showed good performance in the test set ($R^2_{ext} = 0.70$). However, the R^2 score was equal to 1 for the training set, which means that this model over-learned during its training phase and may be of lower quality for unknown data. The selection of parameters was made on all the initial data, which ensured that the attributes are representative of all the data and not only the training data or the test data. In the supplementary materials, Figure S1 shows a comparison of the selection only based on the training sets or only based on the initial full data set.

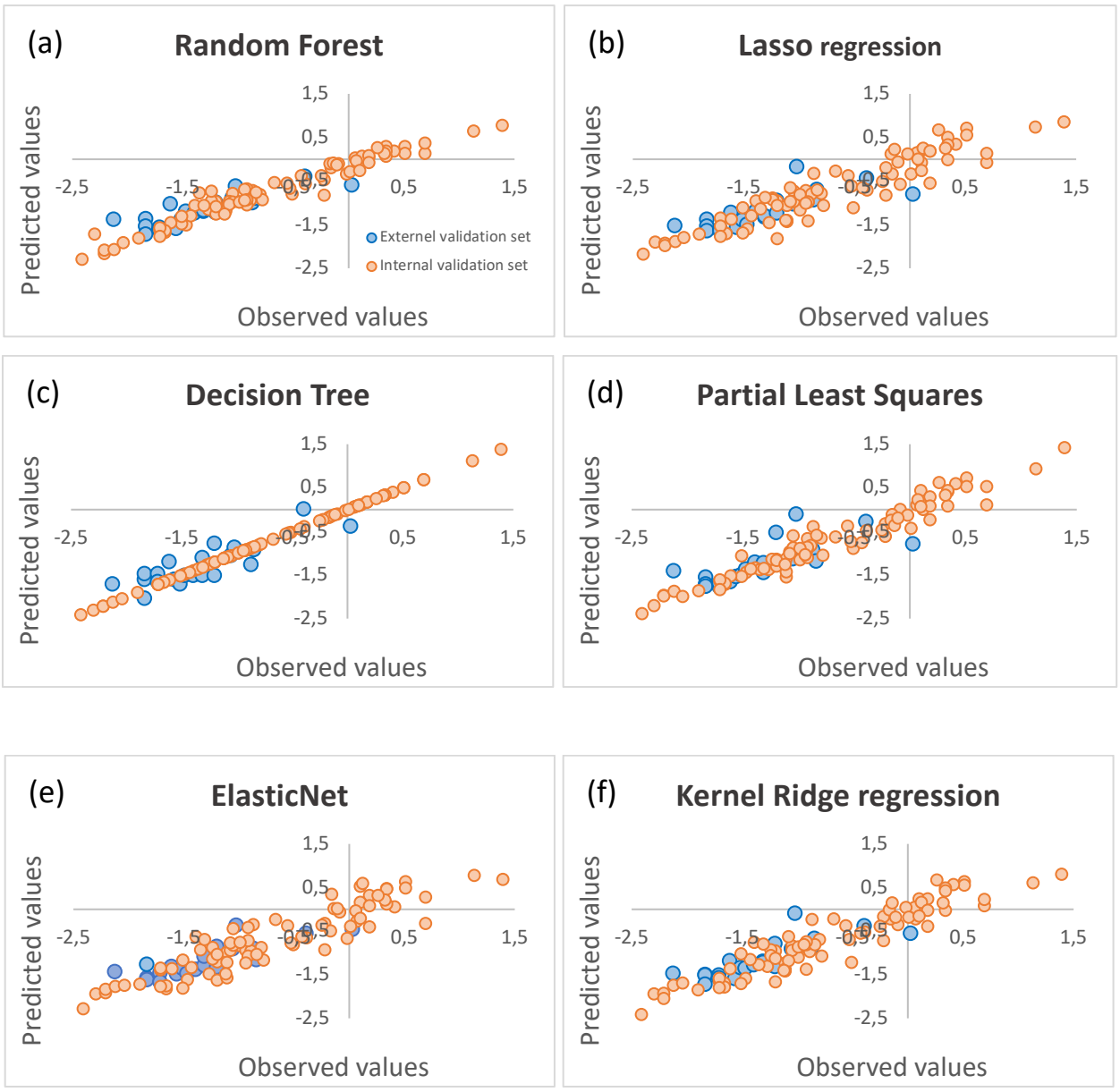

Figure 3. Comparison of predicted and observed ratios for the internal validation set (orange circle) and the external validation set (blue circle) using the Python programming language and the sklearn library. Three models have been developed. The results of the Random Forest model are indicated in (a), Lasso Regression in (b), Decision Tree in (c), Partial Least Squares (d), ElasticNet in (e), and Kernel Ridge Regression (f).

3.4.3 **SuperLearner**

Finally, a QSAR supermodel was developed using the SuperLearner package in R Studio (Figure 4). Internal and external validation scores with coefficient of determination (R^2) and root mean square error (RMSE), 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 contributes the most to the supermodel is 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.

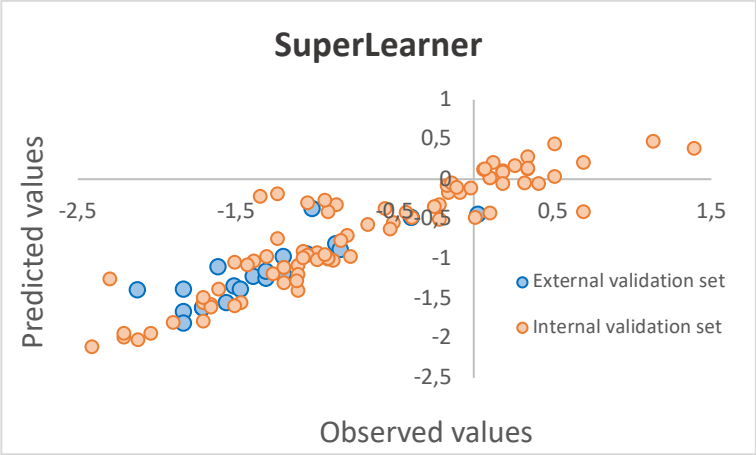

Figure 4. Comparison of predicted and observed ratios for the internal validation set (orange circle) and the external validation set (blue circle) using the SuperLearner package in R Studio.

3.4.4 **Applicability domain**

The applicability domain was defined with the tool Applicability Domain v1.0 developed by Roy et al., (2015). Standardized values of the 214 descriptors were calculated for all the observations and used to determine whether the molecule lays within the applicability domain. Results showed one outlier in the training set, but no observation outside the applicability domain in the test set.

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

Models	R^2	RMSE	R^2_{CV}	$RMSE_{CV}$	R^2_{ext}	$RMSE_{ext}$
MOE						
Genetic Algorithm-Multiple Linear Regression	0.76	0.41	0.66	0.49	0.56	0.36
Partial Least Squares	0.88	0.29	0.72	0.45	0.73	0.32
Principal Component Regression	0.66	0.49	0.50	0.60	0.38	0.47
Python Language						
Random Forest	0.93	0.22	0.48	0.44	0.63	0.31
Lasso Regression	0.87	0.30	0.67	0.48	0.55	0.34
Partial Least Squares	0.94	0.20	0.48	0.44	0.47	0.36
Decision tree	1.00	0.00	0.51	0.32	0.70	0.27
ElasticNet	0.84	0.34	0.67	0.48	0.61	0.31
Kernel ridge regression	0.88	0.29	0.98	0.13	0.56	0.33
R Studio						
SuperLearner (combination 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

3.5 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 R Studio. 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. The genetic algorithm-multiple linear regression and the principal component regression models showed lower R^2 and higher RMSE for external validation with sets of 10 and 11 descriptors, respectively. 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. The Decision Tree analysis showed perfect scores of R^2 and RMSE for the training phase, which considerably decreased for the cross-validation process, demonstrating an overfitting of the model to the training set. Finally, the SuperLearner developed in R Studio 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 has been determined with the Applicability Domain v.1.0, a tool based on the standardisation approach, and showed that all test compounds were included in the interpolation space, thus those external predictions can be considered precise and reliable.

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. (2018). 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 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. (2018) 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 (2010) 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. (2018) could be explained by the high number of descriptors (n = 10) used in the model

development, and the small number of observations included in the training set (n = 24), whereas the rule of thumb recommends five chemicals for one descriptor ratio (Gramatica, 2013).

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*)" (Worth et al. 2005). Most relevant descriptors on which the models are based should provide insight into the correlation between the chemical substances 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 were developed with 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-parametrizing the QSAR. On the other hand, the smaller set of descriptors selected with the GA-MLR approach (n = 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. Thus, a chemical non hydrophobic or with a high hydrophobic surface, is not likely to cross the placental barrier. 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 into 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 (Feghali et al. 2015). 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 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. 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). 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 with respective 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. These limitations likely impacted model performance and domain of applicability.

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. 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.

3.6 Abbreviations

AD	Applicability Domain
AFP	Alpha FetoProtein
AHA	AracHidonic Acid
DHA	DocosaHexaneoic Acid
DT	Decision Tree
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

3.7 References

Aylward, L. L., Hays, S. M., Kirman, C. R., Marchitti, S. A., Kenneke, J. F., English, C., ... & Becker, R. A. (2014). Relationships of chemical concentrations in maternal and cord blood: a review of available data. *Journal of Toxicology and Environmental Health, Part B*, *17*(3), 175-203.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chakraborty, A., & Goswami, D. (2017). Prediction of slope stability using multiple linear regression (MLR) and artificial neural network (ANN). *Arabian Journal of Geosciences*, *10*(17), 1-11.

Covaci, A., Jorens, P., Jacquemyn, Y., & Schepens, P. (2002). Distribution of PCBs and organochlorine pesticides in umbilical cord and maternal serum. *Science of the total environment*, *298*(1-3), 45-53.

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*, *7*(1), 1-13.

Eguchi, A., Hanazato, M., Suzuki, N., Matsuno, Y., Todaka, E., & Mori, C. (2018). Maternal–fetal transfer rates of PCBs, OCPs, PBDEs, and dioxin-like compounds predicted through quantitative structure–activity relationship modeling. *Environmental Science and Pollution Research*, *25*(8), 7212-7222.

Feghali, M., Venkataramanan, R., & Caritis, S. (2015, November). Pharmacokinetics of drugs in pregnancy. In *Seminars in perinatology* (Vol. 39, No. 7, pp. 512-519). WB Saunders.

Fisher, M., Arbuckle, T. E., Liang, C. L., LeBlanc, A., Gaudreau, E., Foster, W. G., ... & Fraser, W. D. (2016). Concentrations of persistent organic pollutants in maternal and cord blood

from the maternal-infant research on environmental chemicals (MIREC) cohort study. *Environmental Health*, *15*(1), 1-14.

Frederiksen, M., Thomsen, C., Frøshaug, M., Vorkamp, K., Thomsen, M., Becher, G., & Knudsen, L. E. (2010). Polybrominated diphenyl ethers in paired samples of maternal and umbilical cord blood plasma and associations with house dust in a Danish cohort. *International journal of hygiene and environmental health*, *213*(4), 233-242.

Giaginis, C., Zira, A., Theocharis, S., & Tsantili-Kakoulidou, A. (2009). Application of quantitative structure–activity relationships for modeling drug and chemical transport across the human placenta barrier: a multivariate data analysis approach. *Journal of Applied Toxicology: An International Journal*, *29*(8), 724-733.

Gramatica P. (2013). On the development and validation of QSAR models. Methods in molecular biology (Clifton, N.J.), 930, 499–526. https://doi.org/10.1007/978-1-62703-059-5_21

Grigsby, P. L. (2016, January). 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.

Hewitt, M., Madden, J. C., Rowe, P. H., & Cronin, M. T. D. (2007). Structure-based modelling in reproductive toxicology:(Q) SARs for the placental barrier. *SAR and QSAR in Environmental Research*, *18*(1-2), 57-76.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, *12*(1), 69-82.

Katoch, S., Chauhan, S. S., & Kumar, V. (2020). A review on genetic algorithm: past, present, and future. Multimedia tools and applications, 1–36. Advance online publication. https://doi.org/10.1007/s11042-020-10139-6.

Krimsky, S. (2017). The unsteady state and inertia of chemical regulation under the US Toxic Substances Control Act. *PLoS biology*, *15*(12), e2002404.

Li, J., Cai, D., Chu, C., Li, Q., Zhou, Y., Hu, L., ... & Chen, D. (2020). Transplacental transfer of per-and polyfluoroalkyl substances (PFASs): Differences between preterm and full-term deliveries and associations with placental transporter mRNA expression. *Environmental science & technology*, *54*(8), 5062-5070.

McCall, J. (2005). Genetic algorithms for modelling and optimisation. *Journal of computational and Applied Mathematics*, *184*(1), 205-222.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(1), 93-100.

Mauri, A., Consonni, V., Pavan, M., & Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations. *Match*, *56*(2), 237-248.

Molecular Operating Environment (MOE), 2019.01; Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

Morello-Frosch, R., Cushing, L. J., Jesdale, B. M., Schwartz, J. M., Guo, W., Guo, T., ... & Woodruff, T. J. (2016). Environmental chemicals in an urban population of pregnant women and their newborns from San Francisco. *Environmental science & technology*, *50*(22), 12464-12472.

Mori, C., Nakamura, N., Todaka, E., Fujisaki, T., Matsuno, Y., Nakaoka, H., & Hanazato, M. (2014). Correlation between human maternal–fetal placental transfer and molecular weight of PCB and dioxin congeners/isomers. Chemosphere, 114, 262-267.

Myren, M., Mose, T., Mathiesen, L., & Knudsen, L. E. (2007). The human placenta–an alternative for studying foetal exposure. *Toxicology in Vitro*, *21*(7), 1332-1340.

Park, J. S., Bergman, Å., Linderholm, L., Athanasiadou, M., Kocan, A., Petrik, J., ... & Hertz-Picciotto, I. (2008). Placental transfer of polychlorinated biphenyls, their hydroxylated metabolites and pentachlorophenol in pregnant women from eastern Slovakia. *Chemosphere*, *70*(9), 1676-1684.

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *145*, 22-29.

Saxena, A. K., & Prathipati, P. (2003). Comparison of mlr, pls and ga-mlr in qsar analysis. *SAR and QSAR in Environmental Research*, *14*(5-6), 433-445.

Sexton, K., Salinas, J. J., McDonald, T. J., Gowen, R. M., Miller, R. P., McCormick, J. B., & Fisher-Hoch, S. P. (2011). Polycyclic aromatic hydrocarbons in maternal and umbilical cord blood from pregnant Hispanic women living in Brownsville, Texas. *International journal of environmental research and public health*, *8*(8), 3365-3379.

Takaku, T., Nagahori, H., Sogame, Y., & Takagi, T. (2015). Quantitative structure–activity relationship model for the fetal–maternal blood concentration ratio of chemicals in humans. *Biological and Pharmaceutical Bulletin*, *38*(6), 930-934.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Tobias, R. D. (1995, April). An introduction to partial least squares regression. In *Proceedings of the twentieth annual SAS users group international conference* (Vol. 20). Cary: SAS Institute Inc.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular informatics*, *29*(6-7), 476-488.

Van Der Laan, M. J., & Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. Int. J. Drug Des. Discov, 3, 511-519.

Vizcaino, E., Grimalt, J. O., Fernández-Somoano, A., & Tardon, A. (2014). Transport of persistent organic pollutants across the human placenta. *Environment international*, *65*, 107-115.

Wang, C. C., Lin, P., Chou, C. Y., Wang, S. S., & Tung, C. W. (2020). Prediction of human fetal–maternal blood concentration ratio of chemicals. *PeerJ*, *8*, e9562.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36.

Whyatt, R. M., Barr, D. B., Camann, D. E., Kinney, P. L., Barr, J. R., Andrews, H. F., ... & Perera, F. P. (2003). Contemporary-use pesticides in personal air samples during pregnancy and blood samples at delivery among urban minority mothers and newborns. *Environmental health perspectives*, *111*(5), 749-756.

Worth, A. P., Bassan, A., Gallegos, A., Netzeva, T. I., Patlewicz, G., Pavan, M., ... & Vračko, M. (2005). *The characterisation of (quantitative) structure-activity relationships: preliminary guidance*. Institute for Health and Consumer Protection, Toxicology and Chemical Substances Unit, European Chemical Bureau.

Yang, L., Li, J., Lai, J., Luan, H., Cai, Z., Wang, Y., ... & Wu, Y. (2016). Placental transfer of perfluoroalkyl substances and associations with thyroid hormones: Beijing Prenatal Exposure Study. *Scientific reports*, *6*(1), 1-9.

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, *32*(7), 1466-1474.

Yin, S., Zhang, J., Guo, F., Zhao, L., Poma, G., Covaci, A., & Liu, W. (2019). Transplacental transfer of organochlorine pesticides: Concentration ratio and chiral properties. *Environment international*, *130*, 104939.

Zhang, T., Sun, H., Lin, Y., Qin, X., Zhang, Y., Geng, X., & Kannan, K. (2013). Distribution of poly-and perfluoroalkyl substances in matched samples from pregnant women and carbon chain length related maternal transfer. *Environmental science & technology*, *47*(14), 7974-7981.

Zhang, Y. H., Xia, Z. N., Yan, L., & Liu, S. S. (2015). Prediction of placental barrier permeability: a model based on partial least squares variable selection procedure. *Molecules*, *20*(5), 8270-8286.

Zhang, X., Li, X., Jing, Y., Fang, X., Zhang, X., Lei, B., & Yu, Y. (2017). Transplacental transfer of polycyclic aromatic hydrocarbons in paired samples of maternal serum, umbilical cord serum, and placenta in Shanghai, China. *Environmental Pollution*, *222*, 267-275.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

3.8 Supplemental material

Table S1. Equations of the partial least squares analysis (PLS), the genetic algorithm-multiple linear regression (GA-MLR) and the principal component regression (PCR) developed with the MOE software.

PLS	GA-MLR	PCR
Ratio Ln =	**Ratio Ln =**	**Ratio Ln =**
4.94465	19.39180	0.38145
+0.01159 * ASA	-0.00778 * ASA_H	+0.09792 * E_vdw
-0.03482 * ASA+	+4.49588 * BCUT_SLOGP_3	+0.03934 * PEOE_VSA_PPOS
+0.00429 * ASA-	-9.12944 * BCUT_SMR_3	-0.00349 * Q_VSA_PNEG
+0.01331 * ASA_H	-2.01493 * GCUT_SMR_0	-0.03744 * SlogP_VSA6
-0.00254 * CASA-	-0.15224 * KierA3	+0.01169 * SlogP_VSA9
-0.01735 * DASA	+0.53636 * PEOE_RPC-	+0.07428 * a_IC
+0.00191 * DCASA	-0.01554 * PEOE_VSA+1	-0.21493 * a_count
-0.03139 * PEOE_VSA+1	-1.87284 * VAdjMa	+0.22336 * a_nH
+0.00626 * PEOE_VSA-1	+1.16785 * mr	+0.18554 * b_ar
+0.02782 * PEOE_VSA_HYD	-0.03036 * vsurf_D8	-0.10873 * bpol
-0.02583 * PEOE_VSA_NEG		-0.03633 * vsurf_D8
+0.00867 * Q_VSA_HYD		
+0.02106 * Q_VSA_PNEG		
+0.00989 * Q_VSA_POL		
+0.00421 * SMR_VSA1		
+0.00594 * SMR_VSA7		
+0.02024 * SlogP_VSA0		
+0.01459 * VSA		
-0.01786 * Weight		
+0.00182 * pmi		
-0.00039 * pmi2		
-0.00098 * pmi3		
+0.01856 * vdw_area		
+0.02939 * vdw_vol		
-0.01921 * vsa_hyd		
-0.00451 * vsurf_D2		
+0.01101 * vsurf_D3		
-0.01191 * vsurf_D4		
-0.02210 * vsurf_D6		
-0.00310 * vsurf_HB1		
-0.01676 * vsurf_HB3		
+0.01566 * vsurf_HB4		
-0.02257 * vsurf_V		
+0.01430 * vsurf_W4		
-0.01859 * zagreb		

Table S2. Relative importance (weight) and name of descriptors selected to build the partial least squares analysis (PLS, n=35), the genetic algorithm-multiple linear regression (GA-MLR, n=10) and the principal component regression (PCR, n=11) with the MOE software.

PLS		**GA-MLR**		**PCR**	
Weight	***Name***	***Weight***	***Name***	***Weight***	***Name***
0.125659	ASA	0.696288	ASA_H	0.374109	E_vdw
0.326923	ASA+	0.232612	BCUT_SLOGP_3	0.241907	PEOE_VSA_PPO
0.073315	ASA-	0.750328	BCUT_SMR_3	0.189426	S
0.339177	ASA_H	0.056361	GCUT_SMR_0	0.204686	Q_VSA_PNEG
0.494355	CASA-	0.104815	KierA3	0.722437	SlogP_VSA6
0.374139	DASA	0.035676	PEOE_RPC-	0.491142	SlogP_VSA9
0.362333	DCASA	0.121027	PEOE_VSA+1	0.684269	a_IC
0.069577	PEOE_VSA+1	0.379170	VAdjMa	0.520870	a_count
0.066013	PEOE_VSA-1	1.000000	mr	0.574263	a_nH
0.260037	PEOE_VSA_HYD	0.731437	vsurf_D8	0.496467	b_ar
0.244535	PEOE_VSA_NEG			1.000000	bpol
0.131074	Q_VSA_HYD				vsurf_D8
0.285087	Q_VSA_PNEG				
0.154437	Q_VSA_POL				
0.059774	SMR_VSA1				
0.096403	SMR_VSA7				
0.045963	SlogP_VSA0				
0.126779	VSA				
0.376834	Weight				
1.000000	pmi				
0.199329	pmi2				
0.503036	pmi3				
0.176409	vdw_area				
0.277359	vdw_vol				
0.185755	vsa_hyd				
0.127897	vsurf_D2				
0.237958	vsurf_D3				
0.232352	vsurf_D4				
0.360279	vsurf_D6				
0.068965	vsurf_HB1				
0.563431	vsurf_HB3				
0.235605	vsurf_HB4				
0.395751	vsurf_V				
0.216094	vsurf_W4				
0.101190	zagreb				

4 Discussion générale

4.1 Rappel rapide des résultats

Plusieurs études toxicologiques et épidémiologiques laissent croire que de nombreux contaminants environnementaux sont capables de traverser la barrière placentaire durant la grossesse, mais que cette capacité varie selon le composé. La caractérisation du passage placentaire des contaminants auxquels sont exposées les femmes enceintes est primordiale, considérant que plusieurs composés chimiques sont susceptibles d'altérer le développement du fœtus. Dans le cadre de ma maîtrise, j'ai compilé 105 ratios de concentrations sanguines fœto-maternels pour une diversité de contaminants à partir d'une revue de la littérature. Ces molécules ont permis de générer une sélection de 214 descripteurs 2D et 3D, lesquels ont été utilisés pour le développement de dix modèles QSAR à l'aide du logiciel Molecular Operating Environment, du langage de programmation Python, et de R Studio. La robustesse, le degré d'ajustement et la prédictivité des modèles ont été évalués lors de phases de validation interne et de validation externe, à l'aide du coefficient de détermination (R^2) et de l'erreur quadratique moyenne (RMSE). Le modèle de régression des moindres carrés partiels développé avec 35 descripteurs a donné les meilleurs résultats de prédiction externe pour l'outil MOE (R^2_{ext} = 0.73 et $RMSE_{ext}$ = 0.32) tandis que l'analyse par algorithme génétique-régression linéaire multiple (GA-MLR) réalisée avec 10 descripteurs a donné le modèle le plus interprétable mécanistiquement. Les analyses ElasticNet et des forêts aléatoires ont montré les meilleurs scores de performance externe pour le langage Python (R^2_{ext} = 0.73 et $RMSE_{ext}$ = 0.32, et R^2_{ext} = 0.73 et $RMSE_{ext}$ = 0.32, respectivement) après l'analyse par arbres de décisions, ce modèle ayant été ignoré puisque présentant un surapprentissage des données à la phase d'entraînement avec un R^2 de 1 et un RMSE de 0. Enfin le dernier modèle développé à l'aide de R Studio et du package SuperLearner a obtenu de bons résultats de performance (R^2_{ext} = 0.74 et $RMSE_{ext}$ = 0.29). La détermination du domaine d'applicabilité basée sur le jeu d'entraînement a permis de montrer la présence des composés du jeu de test dans l'espace physico-chimique d'interpolation et donc la fiabilité des prédictions obtenues. L'analyse des descripteurs du modèle GA-MLR a démontré l'importance de l'hydrophobicité, de la surface accessible à l'eau et de la réfractivité molaire (polarité) dans le passage transplacentaire des molécules chimiques et des contaminants. Ces résultats confirment les résultats de précédentes études, lesquelles ont mis en

évidence l'importance du caractère hydrophobe et de la polarité dans le passage des composés à travers la membrane placentaire (Burton et Fowden, 2015; Pemathilaka et al., 2019; Eguchi et al., 2018).

4.2 Retombées possibles

L'élaboration de modèles QSAR performants pour la prédiction du taux de transfert des contaminants à travers la barrière placentaire démontre que les concentrations sanguines fœtales peuvent être estimées à partir des concentrations mesurées chez les femmes enceintes durant la grossesse. Ainsi il est possible de quantifier l'exposition du foetus aux composés de l'environnement basé sur l'exposition de la femme enceinte.

L'utilisation des modèles QSAR *in silico* pour l'étude de la toxicité foetale et des mécanismes physiologiques et métaboliques en jeu dans le transport placentaire des composés toxiques est prometteuse pour l'estimation des paramètres pour des modèles pharmacocinétiques à base physiologique (PBPK) de l'exposition des femmes enceintes (Wier et al., 1990). Les modèles PBPK sont des modèles mathématiques *in silico* qui utilisent de façon mécaniste les propriétés physiologiques, physico-chimiques et biologiques entre des compartiments interconnectés afin de décrire l'estimation de la dose ou de l'exposition (Peyret et Krishnan, 2011). Les paramètres des modèles PBPK sont généralement estimés à partir de données *in vivo* et *in vitro*, qui peuvent être en disponibilité limitée, difficilement extrapolables à partir de modèles animaux ou difficilement mesurables. L'estimation de la dose d'exposition du foetus aux contaminants dépend non seulement des propriétés pharmacocinétiques d'absorption, de distribution, de métabolisme et d'excrétion de la mère mais également des propriétés du transport transplacentaire (Codaccioni et al., 2019). L'utilisation de données *in silico* obtenues à partir de modèles QSAR et décrivant le transfert placentaire des contaminants, permettrait ainsi d'estimer les paramètres des modèles PBPK afin de décrire de façon quantitative la relation entre l'exposition, la dose et la réponse (Peyret et Krishnan, 2011). L'incorporation d'un modèle QSAR dans un modèle pharmacocinétique à base physiologique permettra ainsi d'estimer la dose reçue par le fœtus pour plusieurs scénarios d'exposition chez la femme enceinte.

La modélisation QSAR permet d'autre part l'étude d'un grand nombre de composés de façon très rapide et peu coûteuse. Les modèles QSAR permettent de pallier les contraintes des

expérimentations *in vivo*, *in vitro*, et *ex vivo* chez l'Homme et l'animal en termes éthiques et matériels. En effet, le protocole d'expérimentation des modèles QSAR ne nécessite pas d'installation matérielle complexe, ni d'utilisation de tissus ou d'organes placentaires qui peuvent être assez difficiles à obtenir et à en assurer la viabilité. Plusieurs contaminants peuvent ainsi être analysés de façon simultanée, permettant une obtention très rapide des résultats. La modélisation QSAR propose donc une méthode de criblage à haut débit pour l'identification de composés qui peuvent facilement traverser le placenta. À long terme, l'amélioration de la performance et de la précision des modèles QSAR peut entraîner une réduction des tests effectués sur les animaux dans un contexte où les grandes agences réglementaires visent une abolition de ces tests dans les prochaines décennies.

4.3 Avenues de recherche

La nécessité d'élaborer un modèle QSAR général et applicable à une grande diversité de contaminants fut le moteur de cette recherche. Cela est d'autant plus vrai qu'il n'existe à ce jour que peu de modèles sur le transport placentaire des composés chimiques, et en particulier sur le transport des contaminants. En effet, seule l'étude de Eguchi et al. (2018) s'est intéressée au transfert des contaminants entre la circulation de la mère et celle du fœtus, mais ne l'a fait que pour un nombre limité de composés. De plus les trois modèles développés dans cette étude présentent de faibles performances de prédiction, sans compter que la moitié des données du jeu de test se trouvaient en dehors du domaine d'applicabilité. Le jeu de données utilisé dans notre étude a permis d'élargir le domaine d'applicabilité des modèles QSAR pour le passage placentaire des contaminants. Toutefois ce jeu de données reste relativement petit avec de nombreux composés similaires en termes de structure et de famille chimique. L'utilisation d'une plus grande diversité de composés toxiques permettrait d'élargir le domaine d'applicabilité et de créer de futurs modèles QSAR encore plus fiables, robustes et à haute précision pour l'étude de l'exposition foetale aux contaminants.

5 Conclusion

En conclusion, la recherche réalisée dans ce mémoire a permis d'élaborer des modèles QSAR performants à haut pouvoir de prédiction du passage placentaire des composés chimiques de l'environnement. Les résultats obtenus suggèrent que la modélisation QSAR est avantageuse dans l'estimation et l'évaluation de l'exposition du fœtus aux contaminants environnementaux à partir des concentrations sanguines de composés toxiques mesurées chez la femme enceinte tout au long de la grossesse. Les modèles QSAR peuvent notamment être utilisés pour produire des données *in silico* pouvant être incorporées par la suite dans des modèles pharmacocinétiques à base physiologique du placenta qui évaluent le devenir des composés chimiques selon les propriétés d'absorption, de distribution, de métabolisme et d'excrétion. La facilité de développement de tels modèles permet d'évaluer rapidement de nombreux composés, permettant de contourner les obstacles éthiques et matériels rencontrés dans les modèles animaux et humains. L'utilisation d'un plus grand jeu de données ayant une diversité chimique accrue permettra le développement de futurs modèles QSAR possédant de plus grands domaines d'applicabilité et possiblement de meilleures performances prédictives.

6 Références bibliographiques

Almond, D., & Currie, J. (2011). Killing me softly: The fetal origins hypothesis. Journal of economic perspectives, 25(3), 153-72.

Ashraf, M. A. (2017). Persistent organic pollutants (POPs): a global issue, a global challenge.

Axelrad DA, Goodman S, Woodruff TJ. PCB body burdens in US women of childbearing age 2001-2002: An evaluation of alternate summary metrics of NHANES data. Environ Res 2009 May; 109(4):368–78. [PubMed: 19251256]).

Aylward, L. L., Hays, S. M., Kirman, C. R., Marchitti, S. A., Kenneke, J. F., English, C., ... & Becker, R. A. (2014). Relationships of chemical concentrations in maternal and cord blood: a review of available data. *Journal of Toxicology and Environmental Health, Part B*, *17*(3), 175-203.

Baccarelli, A., & Bollati, V. (2009). Epigenetics and environmental chemicals. Current opinion in pediatrics, 21(2), 243.

Barker, D. J. (1995). Fetal origins of coronary heart disease. Bmj, 311(6998), 171-174.

Bhande, A. (2018). What is underfitting and overfitting in machine learning and how to deal with it. GreyAtom. Lien : https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76 [Consulté le 21 avril 2021].

Boddington, M. (1990). Dibenzodioxines polychlorées et dibenzofurannes polychlorés. Ottawa: Environnement Canada.

Bouazza, N., Foissac, F., Hirt, D., Urien, S., Benaboud, S., Lui, G., & Treluyer, J. M. (2019). Methodological approaches to evaluate fetal drug exposure. Current pharmaceutical design, 25(5), 496-504.

Bourget, P., Roulot, C., & Fernandez, H. (1995). Models for placental transfer studies of drugs. Clinical pharmacokinetics, 28(2), 161-180.

Burton, G. J., & Fowden, A. L. (2015). The placenta: a multifaceted, transient organ. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1663), 20140066. https://doi.org/10.1098/rstb.2014.0066.

Carrier, G. (1995). Réponse de l'organisme humain aux BPC, dioxines et furannes et analyse des risques toxiques.

Chtita S. Modélisation de molécules organiques hétérocycliques biologiquement actives par des méthodes QSAR/QSAR. Recherche de nouveaux médicaments [thèse]. [Meknès]: Moulay Ismail; 2017.

Codaccioni, M., Bois, F., & Brochot, C. (2019). Placental transfer of xenobiotics in pregnancy physiologically-based pharmacokinetic models: Structure and data. Computational Toxicology, 12, 100111.

Combes, R. D. (2012). In silico methods for toxicity prediction. *New Technologies for Toxicity Testing*, 96-116.

Cronin, M. T. D., & Madden, J. C. (2010). In Silico Toxicology—An Introduction. In *In Silico Toxicology* (pp. 1-10).

Edwards, M. (2017). The barker hypothesis. Handbook of Famine, Starvation, and Nutrient Deprivation: From Biology to Policy, 191-211.

Eguchi, A., Hanazato, M., Suzuki, N., Matsuno, Y., Todaka, E., & Mori, C. (2018). Maternal–fetal transfer rates of PCBs, OCPs, PBDEs, and dioxin-like compounds predicted through quantitative structure–activity relationship modeling. *Environmental Science and Pollution Research*, *25*(8), 7212-7222.

Ehrhardt, C., & Kim, K. J. (2007). Drug Absorption Studies. Springer Publishing.

Elefant, E., & Beghin, D. (2009). Le passage transplacentaire des médicaments. Bull. Acad. Natle Méd., 193, no 5, 1043-1057, séance du 12 mai 2009.

Jensen, A. A., & Leffers, H. (2008). Emerging endocrine disrupters: perfluoroalkylated substances. International journal of andrology, 31(2), 161-169.

Giaginis, C., Zira, A., Theocharis, S., & Tsantili-Kakoulidou, A. (2009). Application of quantitative structure–activity relationships for modeling drug and chemical transport across the human placenta barrier: a multivariate data analysis approach. *Journal of Applied Toxicology: An International Journal*, *29*(8), 724-733.

Göhner, C., Svensson-Arvelund, J., Pfarrer, C., Häger, J. D., Faas, M., Ernerudh, J., Cline, J. M., Dixon, D., Buse, E., & Markert, U. R. (2014). The placenta in toxicology. Part IV: Battery of toxicological test systems based on human placenta. Toxicologic pathology, 42(2), 345–351.

Gramatica P. (2013). On the development and validation of QSAR models. Methods in molecular biology (Clifton, N.J.), 930, 499–526. https://doi.org/10.1007/978-1-62703-059-5_21

Grigsby, P. L. (2016, January). 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.

Griffiths, S. K., & Campbell, J. P. (2015). Placental structure, function and drug transfer. Continuing Education in Anaesthesia, Critical Care & Pain, 15(2), 84-89.

Grova, N., Schroeder, H., Olivier, J. L., & Turner, J. D. (2019). Epigenetic and Neurological Impairments Associated with Early Life Exposure to Persistent Organic Pollutants. International journal of genomics, 2019, 2085496.

Hartung, T., & Hoffmann, S. (2009). Food for thought... on in silico methods in toxicology. *ALTEX-Alternatives to animal experimentation*, *26*(3), 155-166.

Hemmerich, J., & Ecker, G. F. (2020). In silico toxicology: From structure–activity relationships towards deep learning and adverse outcome pathways. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *10*(4), e1475.

Herbstman, J. B., Sjödin, A., Kurzon, M., Lederman, S. A., Jones, R. S., Rauh, V., Needham, L. L., Tang, D., Niedzwiecki, M., Wang, R. Y., & Perera, F. (2010). Prenatal exposure to PBDEs and neurodevelopment. Environmental health perspectives, 118(5), 712–719.

Hewitt, M., Madden, J. C., Rowe, P. H., & Cronin, M. T. D. (2007). Structure-based modelling in reproductive toxicology:(Q) SARs for the placental barrier. *SAR and QSAR in Environmental Research*, *18*(1-2), 57-76.

Hutson, J. R., Garcia-Bournissen, F., Davis, A., & Koren, G. (2011). The human placental perfusion model: a systematic review and development of a model to predict in vivo transfer of therapeutic drugs. *Clinical pharmacology and therapeutics*, *90*(1), 67–76.

INSPQ, Institut National de Santé Publique du Québec. (Mise à jour : 2 mai 2019). Principaux Contaminants. https://www.inspq.qc.ca/qualite-de-l-air-et-salubrite-intervenir-ensemble-dans-l-habitation-au-quebec/qualite-de-l-air-et-salubrite/principaux-contaminants [Consulté le 24 mars 2020].

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: a review. Alternatives to laboratory animals: ATLA, 33(5), 445–459. https://doi.org/10.1177/026119290503300508

Knudsen, T. B., Keller, D. A., Sander, M., Carney, E. W., Doerrer, N. G., Eaton, D. L., ... & Whelan, M. (2015). FutureTox II: in vitro data and in silico models for predictive toxicology. *Toxicological Sciences*, *143*(2), 256-267.

Lallas, P. L. (2001). The Stockholm Convention on persistent organic pollutants. The American Journal of International Law, 95(3), 692-708.

Lejarraga H. Perinatal origin of adult diseases. Arch Argent Pediatr 2019;117(3):e232-e242.

Lundstedt, S., White, P. A., Lemieux, C. L., Lynes, K. D., Lambert, I. B., Öberg, L., ... & Tysklind, M. (2007). Sources, fate, and toxic hazards of oxygenated polycyclic aromatic hydrocarbons (PAHs) at PAH-contaminated sites. AMBIO: A Journal of the Human Environment, 36(6), 475-485.

Magnarelli, G., & Guiñazú, N. (2012). Placental Toxicology of Pesticides.

Mandy, M., & Nyirenda, M. (2018). Developmental Origins of Health and Disease: the relevance to developing nations. International health, 10(2), 66-70.

Marsit, C. J. (2015). Influence of environmental exposure on human epigenetic regulation. Journal of Experimental Biology, 218(1), 71-79.

Mattison D. R. (2010). Environmental exposures and development. Current opinion in pediatrics, 22(2), 208–218. https://doi.org/10.1097/MOP.0b013e32833779bf.

Modi, S., Hughes, M., Garrow, A., & White, A. (2012). The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug discovery today*, *17*(3-4), 135-142.

Morello-Frosch, R., Cushing, L. J., Jesdale, B. M., Schwartz, J. M., Guo, W., Guo, T., Wang, M., Harwani, S., Petropoulou, S. E., Duong, W., Park, J. S., Petreas, M., Gajek, R., Alvaran, J.,

She, J., Dobraca, D., Das, R., & Woodruff, T. J. (2016). Environmental Chemicals in an Urban Population of Pregnant Women and Their Newborns from San Francisco. Environmental science & technology, 50(22), 12464–12472. https://doi.org/10.1021/acs.est.6b03492

Myren M, Mose T, Mathiesen L, Knudsen LE. The human placenta--an alternative for studying foetal exposure. Toxicol In Vitro. 2007 Oct;21(7):1332-40. doi: 10.1016/j.tiv.2007.05.011. Epub 2007 Jun 7. PMID: 17624715.

Needham, L. L., Grandjean, P., Heinzow, B., Jørgensen, P. J., Nielsen, F., Patterson Jr, D. G., ... & Weihe, P. (2011). Partition of environmental chemicals between maternal and fetal blood and tissues. *Environmental science & technology*, *45*(3), 1121-1126.

Pemathilaka, R. L., Reynolds, D. E., & Hashemi, N. N. (2019). Drug transport across the human placenta: review of placenta-on-a-chip and previous approaches. Interface focus, 9(5), 20190031. https://doi.org/10.1098/rsfs.2019.0031

Perera, F., & Herbstman, J. (2011). Prenatal environmental exposures, epigenetics, and disease. Reproductive toxicology (Elmsford, N.Y.), 31(3), 363–373. https://doi.org/10.1016/j.reprotox.2010.12.055

Perera, F. P., Rauh, V., Whyatt, R. M., Tsai, W. Y., Bernert, J. T., Tu, Y. H., ... & Tang, D. (2004). Molecular evidence of an interaction between prenatal environmental exposures and birth outcomes in a multiethnic population. *Environmental Health Perspectives*, *112*(5), 626-630.

Peyret, T., & Krishnan, K. (2011). QSARs for PBPK modelling of environmental contaminants. SAR and QSAR in environmental research, 22(1-2), 129–169. https://doi.org/10.1080/1062936X.2010.548351

Pryor, J. L., Hughes, C., Foster, W., Hales, B. F., & Robaire, B. (2000). Critical windows of exposure for children's health: the reproductive system in animals and humans. Environmental Health Perspectives, 108(suppl 3), 491-503.

Rani, M., Shanker, U., & Jassal, V. (2017). Recent strategies for removal and degradation of persistent & toxic organochlorine pesticides using nanoparticles: a review. Journal of environmental management, 190, 208-222.

Raunio, H. (2011). In silico toxicology–non-testing methods. *Frontiers in pharmacology*, *2*, 33.

Rice, D., & Barone Jr, S. (2000). Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. Environmental health perspectives, 108(suppl 3), 511-533.Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. Chemometrics and Intelligent Laboratory Systems, 145, 22-29.

Roy, K., Kar, S., & Ambure, P. (2015). On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems*, *145*, 22-29.

Rudge, C. V., Röllin, H. B., Nogueira, C. M., Thomassen, Y., Rudge, M. C., & Odland, J. Ø. (2009). The placenta as a barrier for toxic and essential elements in paired maternal and cord blood samples of South African delivering women. Journal of environmental monitoring : JEM, 11(7), 1322–1330. https://doi.org/10.1039/b903805a.

Sartori, J. B. U. S. C. (2007). Programmation f oetale: un facteur de risque méconnu des maladies cardiovasculaires et métaboliques. Rev Med Suisse, 3, 32638.

Siddiqui, M. K., Srivastava, S., Srivastava, S. P., Mehrotra, P. K., Mathur, N., & Tandon, I. (2003). Persistent chlorinated pesticides and intra-uterine foetal growth retardation: a possible association. International archives of occupational and environmental health, 76(1), 75–80. https://doi.org/10.1007/s00420-002-0393-6.

Takaku, T., Nagahori, H., Sogame, Y., & Takagi, T. (2015). Quantitative structure–activity relationship model for the fetal–maternal blood concentration ratio of chemicals in humans. *Biological and Pharmaceutical Bulletin*, *38*(6), 930-934.

Tian, F. Y., & Marsit, C. J. (2018). Environmentally Induced Epigenetic Plasticity in Development: Epigenetic Toxicity and Epigenetic Adaptation. Current epidemiology reports, 5(4), 450-460.

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. Molecular informatics, 29(6-7), 476-488.

Van Den Berg, H., Manuweera, G., & Konradsen, F. (2017). Global trends in the production and use of DDT for control of malaria and other vector-borne diseases. Malaria journal, 16(1), 1-8.

Van Der Laan, M. J., & Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. Int. J. Drug Des. Discov, 3, 511-519.

Vizcaino, E., Grimalt, J. O., Fernández-Somoano, A., & Tardon, A. (2014). Transport of persistent organic pollutants across the human placenta. *Environment international*, *65*, 107-115.

Wan, Y., Choi, K., Kim, S., Ji, K., Chang, H., Wiseman, S., ... & Giesy, J. P. (2010). Hydroxylated polybrominated diphenyl ethers and bisphenol A in pregnant women and their matching fetuses: placental transfer and potential risks. *Environmental science & technology*, *44*(13), 5233-5239.

Wang, C. C., Lin, P., Chou, C. Y., Wang, S. S., & Tung, C. W. (2020). Prediction of human fetal–maternal blood concentration ratio of chemicals. *PeerJ*, *8*, e9562.

Wier, P. J., Miller, R. K., Maulik, D., & di Sant'Agnese, P. A. (1990). Toxicity of cadmium in the perfused human placenta. Toxicology and applied pharmacology, 105(1), 156-171.

Yang, H., Sun, L., Li, W., Liu, G., & Tang, Y. (2018). In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. *Frontiers in chemistry*, *6*, 30.

Zhang, X., Cheng, X., Lei, B., Zhang, G., Bi, Y., & Yu, Y. (2021). A review of the transplacental transfer of persistent halogenated organic pollutants: Transfer characteristics, influential factors, and mechanisms. Environment international, 146, 106224. https://doi.org/10.1016/j.envint.2020.106.

Zhang, X., Li, X., Jing, Y., Fang, X., Zhang, X., Lei, B., & Yu, Y. (2017). Transplacental transfer of polycyclic aromatic hydrocarbons in paired samples of maternal serum, umbilical cord serum, and placenta in Shanghai, China. *Environmental Pollution*, *222*, 267-275.