Constructing Xenobiotic Maps of Metabolism to Predict the Role of Enzymes in DNA Adduct Formation

Maël Conan
Institut de Recherche en Santé, Environnement et Travail

Nathalie Théret
Institut de Recherche en Santé, Environnement et Travail

Sophie Langouet
Institut de Recherche en Santé, Environnement et Travail

Anne Siegel (✉️ anne.siegel@irisa.fr )
Institut de Recherche en Informatique et Systèmes Aléatoires

Research Article

Keywords: Metabolism, Heterocyclic aromatic amines, xenobiotics, DNA binding ability, Site of metabolism

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-157802/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Constructing xenobiotic maps of metabolism to predict the role of enzymes in DNA adduct formation

Mael Conan¹,², Nathalie Théret¹,², Sophie Langouet¹*† and Anne Siegel²*†

Abstract

Background: The liver plays a major role in the metabolic activation of xenobiotics (drugs, chemicals such as pollutants, pesticides, food additives...). Among environmental contaminants of concern, heterocyclic aromatic amines (HAA) are xenobiotics classified as possible or probable carcinogens (2A or 2B) by IARC for which low information exist in humans. While HAA is a family of more than thirty identified chemicals, the metabolism activation and DNA adduct formation have been fully characterized in human liver for few of them (MeIQx, PhIP, AαC).

Results: We developed a modeling approach in order to predict all the possible metabolite derivatives of a xenobiotic. Our approach relies on the construction of an enriched and annotated map of derivative metabolites from an input metabolite. The pipeline assembles reaction prediction tools (SyGMa), sites of metabolism prediction tools (Way2Drug, SOMP and Fame 3), a tool to estimate the ability of a xenobiotics to form DNA adducts (XenoSite Reactivity V1), and a filtering procedure based on Bayesian framework. This prediction pipeline was evaluated using caffeine and then applied to HAAs. The method was applied to determine enzyme profiles associated with the maximization of DNA adducts formation derived from each HAA. These profiles could be very different depending on the chemicals allowing to classify HAAs which have been grouped by their associated profiles.

Conclusions: Overall, such a predictive toxicological model based on an in silico systems biology approach open perspectives to estimate genotoxicity of various chemical classes of environmental contaminants. Moreover, our approach based on enzymes profile determination open the perspective to predict various xenobiotics derived metabolites susceptible to bind DNA adducts in both normal and physiopathological situations.

Keywords: Metabolism; Heterocyclic aromatic amines; xenobiotics; DNA binding ability; Site of metabolism

Background

Heterocyclic Aromatic Amines (HAA) and their metabolic derivatives The liver plays a major role in the metabolic activation of xenobiotics (drugs, pollutants, pesticides, food additives...). HAA are environmental contaminants formed during the cooking of meat or fish, in cigarette smoke or exhaust gas [1, 2, 3]. HAA are contaminants of concern because previous studies have shown that they are mutagenic in bacteria, carcinogen in animals and due to a lack of epidemiological studies there
are classified as possible and probable carcinogens by the International Agency for Research on Cancer [4].

Among 30 HAAs have been identified so far. They are divided in two classes depending on their formation reaction. First there is the pyrolytic ones which are formed by a pyrolysis reaction of amino acids at temperature as high as 250°C. As an example amine of this group there is AoC (2-Amino-9H-pyrido[2,3-b]indole). Amino Imidazo Arene correspond to the second class of HAA which are produced by Maillard reaction between hexose and amino acids at a temperature greater than 150°C. MeIQx (2-amino-3,8-dimethylimidazo[4,5-f]quinoline), PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine) and IQ (2-amino-3-methylimidazo[4,5-f]quinoline) belong to this class with quinoline, pyridine and quinoline core, respectively.

In human, HAA metabolism includes two transformation steps as illustrated in Figure 1. The first one is catalyzed by phase I metabolism enzymes which consist of an oxidation mainly catalyzed by cytochromes P450 (CYPs). The oxidative metabolite is then conjugated with phase II xenobiotic metabolism enzymes such as UDP glucuronyl transferase (UGTs) but also by glutathione S transferase (GSTs), N-acetyltransferase (NATs) and sulfotransferase (SULTs). Conjugate metabolites can be either excreted or in some case they can form aryl nitrenium ion through heterolytic cleavage; this ion is very nucleophile, can bind DNA and form DNA adducts [4, 2].

![Figure 1: Representation of Heterocyclic Aromatic Amine Metabolism.](image)

In order to predict the genotoxicity of a metabolite, different tools infer the possibility for a compound to bind DNA (potential DNA adduct). A first strategy is to search for specific chemical structure assumed to bind DNA because a known compound, with the similar structures, has been shown to form DNA-adducts [5, 6, 7].
Another strategy to determine if a compound can form DNA adducts is based on a Quantitative Structure-Toxicity Relationship (QSAR) score that models toxicity according to molecular descriptors of compounds [5, 6, 8]. More recent tools use deep learning to infer from the descriptors of each atom if it can bind DNA [9, 10]. The tool predicts site of reactivity (SOR) associated with a SOR score representing the probability to bind DNA.

The main cornerstone to use toxicity prediction tools on HAA is that compounds which may bind to DNA result from one or several metabolic transformations, which are unstable and cannot be experimentally characterized. As a consequence, the bioactivation metabolism and DNA adduct formation are fully characterized for only three HAAs i.e., AcC, MeIQx and PhIP in human liver [11, 12, 13, 14]. This advocates for the use of in silico methods to predict HAA derivatives and potential DNA adducts derived from HAAs bioactivation in order to drive research about toxicity of HAAs family.

**Prediction of metabolites** To overcome the lack of information about metabolic bioactivation of HAAs and potential formation of DNA adducts, tools for metabolism prediction have been developed allowing identification of potential biomarkers of exposure in humans. Methods for the prediction of metabolites and reactions use biochemical transformation rules describing chemical reactions and linking an input chemical structure to an output chemical structure. For a given compound, the prediction tool searches for chemical structures matching with such input structures and when they are found, the rule is applied and the resulting derivatives are predicted as metabolites. Several tools implement such methods of prediction including MetaSite, METEOR, META, PROXIMAL, TIMES, UM-PPS BioTransformer or SyGMa [15, 16, 17, 18, 19, 20, 21, 22]. The predictions are often represented as a metabolism map containing the predicted metabolites and the reactions that link them. The main drawback of these approaches is that the use of a high number of transformation rules can lead to a great number of predictions with a high number of unknown metabolites [23].

**Prediction of sites of metabolism (SOM)** Another method for predicting metabolite structures uses prediction of site of metabolism (SOM) that can reduce the high number of unknown predicted metabolites in metabolism map. SOM-based tools predict the reaction of an atom by using a set of specific reactions. This set is generated by associating reactions catalysed by the same enzyme. It results in models that predict the probability for an atom to interact with specific enzymes or isoforms. These methods use molecular descriptors which describe different parameters of each atom of a compound. Some tools such as QMBO, CypScore, SMARTCyp or MetaSite [24, 25, 26, 27, 15, 28] rely on the hydrogen abstraction reaction which is the energy necessary to remove an hydrogen linked to the atom. Other tools such as Way2Drug SOMP, FAst MEtabolizer (FAME) or XenoSite Metabolism 1.0 [29, 30, 31, 9, 10] use structure parameters such as the atom nature and the nearest neighbour atoms. In these tools, machine learning methods are used to determine a score based on atom molecular descriptors, which represents the probability of an atom to be a SOM. Others SOM predictors such as IDSite, IMPACTS or MLite
[32, 33, 34] use docking methods and similarities between ligand structure and structure of the compound of interest.

The literature highlights two strategies for using SOM prediction to predict metabolic maps. A first strategy classifies and evaluates the confidence of different predicted pathways by interpreting SOM as the probability of a reaction to occur in the map. Ranking pathways with these probabilities permits to analyse the predicted metabolites and reactions. To the state of our knowledge, this recent method was only applied to the metabolism of Terbinafine (TBF), permitting to detect a new pathway that can explain the formation of TBF-A from TBF [35].

Another strategy, detailed in [36] uses SOM predictions to filter metabolic maps by removing predicted reactions which are not supported by an accurate SOM prediction. In this study, the SOM-filter threshold is determined by using a training set of analog chemicals to the chemicals of interest. The method was applied to predict HAA metabolism using SOM predictors of CYPs and UGTs enzymes. Metabolites predictions and DNA reactivity prediction were then used to predict potential DNA adducts derived from each HAA. The potential of each HAA to form DNA adducts was finally characterized by the ratio between the number of metabolites predicted to bind DNA and the number of total predicted metabolites. The main limitation of this approach is that the filtration of the metabolic maps relied on SOM scores associated with the reaction producing the putative DNA binding metabolites, getting rid of both the predecessor reactions which are required to produce intermediary metabolites and the possible multiple pathways that produce the same metabolite, as evidenced in [35].

**Contribution** To get further in the prediction of formation of DNA adducts by HAA, we introduce a new method which combines the concept of filtered metabolic map introduced in [36] and the concept of ranked pathways introduced in [35]. Instead of filtering metabolic maps according to individual reaction SOM scores, we introduce a *production probability score* which describes the probability for a metabolite to be produced according to one or several chains of reactions weighted by SOM scores.

Our method consists in a three steps pipeline: first step is the prediction of metabolite derivatives of the compound of interest, second step is the annotation of the resulted *metabolic map* using SOM scores and the third step is the computation of the production probability score for each metabolite with Bayesian networks in order to rank and filter metabolite maps.

We used caffeine as a matter of validation of our modeling approach based on SOM predictions of phase I and phase II xenobiotic metabolism enzymes. Indeed, caffeine metabolism is well described and shares enzymes with HAAs metabolism such as phase I enzyme especially CYP1A2, the main enzyme of caffeine metabolism, but also CYP3A4, CYP2E1, CYP2D6 and phase II enzymes including NATs. In addition, some caffeine metabolites can be produced through distinct pathways similarly to HAAs predicted metabolites. After validation of the method using caffeine, the method was applied to HAAs to predict DNA adducts formation and to identify associated enzymes signature.
Results

Definition and construction of enriched metabolic maps

*Map of metabolism* We define the concept of *enriched maps of metabolism* to be oriented graphs where nodes represent chemicals compounds and edges represent reactions that model the transformation of the input compound into the output compound. In these maps, different information is added as labels of reactions and nodes in order to enable the exploration of predicted metabolism results. As detailed below, in enriched metabolic maps, nodes are labeled by *smile formula DNA reactivity label* and *production probability score* and edges are labeled by *rule name, atom number, rank label, enzyme name* and *enzyme family*. Consequently, two edges with the same enzyme family but different enzymes are associated with different edges. An example is shown in Fig. 2.

More precisely, nodes of enriched maps of metabolism represent chemical compounds. They are associated with a SMILE formula, which is interpreted in a 2D structure allowing to label atoms with numbers according to the International Union of Pure and Applied Chemistry (IUPAC) standard conventions [37]. An other label of the node is its *DNA reactivity label*, an information provided by site of reactivity (SOR) predictors.

Edges are first labeled by a unique identifier and by a *rule name* that refers to a SMIRKS rule which encodes the transformation [38]. Each reaction is also associated with an *atom number of reaction*, the label of the atom in 2D structure of the input compound which is transformed by the reaction to form the output compound. Finally, each edge is labeled by an *enzyme name*, which catalyzes the reaction. The *production probability score* of the edge is determined by using site of metabolism (SOM) predictors (see methods for details) and the atom number of reaction.

A specific node is identified in the graph, named the *original compound*. It is defined to be the source of the map of metabolism, and a compound is described by its SMILES canonical formula available in PubChem database [39]. We introduce a *rank* label for each edge, which describes the position of the reaction in the graph with respect to the original compound: *first rank reactions* correspond to edges having the original compound as input, *second rank reactions* corresponds to edges whose input is the output of a first rank reaction, etc...

*Pipeline for building a map of metabolism* Maps of metabolism were built by combining several tools, which are precisely described in the Methods section.

The pipeline starts with the selection of an original compound described by its SMILES canonical formula available in PubChem database. In this paper, the method was applied to 31 original compounds: caffeine (for the sake of validation of the method), and 30 HAAs, see Results below.

For each original compound, the SyGMA python package [22] is applied to compute nodes (e.g, metabolic derivatives of the original compound) and edges (e.g., transformations between metabolites) of the associated map of metabolism. In our studies, SyGMA was iterated twice, in order to predict first-rank and second-rank reactions with respect to the original compound, e.g, all possible derivatives of the original compound with at most two transformations.
Figure 2: **Definition and construction of enriched maps of metabolism**

(a) **Labels of a map of metabolism.** Metabolites are depicted by their 2D structure in black squares. Arrows between metabolites represent reactions, which consume a metabolite to produce another one. Left side: labels are shown in dashed squares (green for metabolites and purple for reactions). Right side: examples of values of labels for a part of the left side network. (b) **Pipeline to obtain labels for metabolites and reactions.** Labels are depicted in green (metabolites) and purple (reactions) squares. They are computed thanks to different tools and methods (grey squares) which require as input the information provided by other labels. SMIRKS labels dictionary is detailed in Additional file 1.

The **SMILES label** (and its associated 2D structure) of each node is generated by the RDKit package implemented in SyGMA. To define the **DNA reactivity label** of each node, we consider that a node is reactive with DNA if at least one of the atom of the metabolite has a score of reactivity computed by XenoSite Reactivity [9] greater than 0.85.

For each edge, a manual curation procedure is undergone to determine the **atom number of reaction.** This label represents the number of the atom, according to IUPAC numbering [37], of the input metabolite of the edge, on which the reaction occurs to produce the output metabolite of the edge. These atom numbers of reactions are obtained by manually comparing the structure of the input and output metabolites of each reaction provided by the MarvinView tool [40], in order to identify the IUPAC numbering of the transformed atom.
The next step of the pipeline consists in annotating edges with rank, rule name and enzyme labels. For each edge, the rank label is defined to be the number of iterations of SyGMa from the original compound allowing to predict the reaction. The rule name label is also provided by SyGMa, according to a catalogue of 176 SMIRKS rules (149 for phase I xenobiotic metabolism reactions and 27 for phase II xenobiotic metabolism reactions).

In order to label each edge with an enzyme, we created a dictionary mapping every SMIRKS rule label to an enzyme family label (see Additional file 1). The 149 SMIRKS rules corresponding to phase I reactions are mapped to the CYPs enzyme family. Among the 27 SMIRKS rules corresponding to phase II reactions, 25 label rules are associated with the UGTs (13 SMIRKS labels), NATs (5 SMIRKS labels), SULTs (6 SMIRKS labels) and GSTs enzyme family (1 SMIRKS labels). The two remaining SMIRKS rule labels are not related to CYPs, UGTs, NATs, GSTs or SULTs and are out of the scope of the method. The corresponding edges are removed from the metabolic map. In addition, nodes appearing to be isolated in the map after this curation are also removed from the map.

The pipeline continues with a procedure used to annotate each edge with a site of metabolism (SOM) prediction score. This procedure depends on its rank, enzyme label and atom number of reaction. (a) Based on knowledge about xenobiotics metabolism, we assume that reactions of first-rank can be considered mostly as phase I reactions, catalyzed by different isoforms of CYPs. Therefore, the tool Way2Drug SOMP [29] is used to compute SOM scores for edges of first-rank, because it provides refine annotation of CYP isoforms (CYP1A2, CYP3A4, CYP2D6, CYP2C9 and CYP2C19), involved in phase I metabolism. As the tool also provides annotations for reactions catalyzed by UGTs, the predicted scores for such reactions are also conserved. (b) Assuming that reactions of second-rank can be considered mostly as reactions of phase II (catalyzed by SULT, UGT, NAT, GST), the tool FAME3 [31] is used to annotate reactions of second rank, because it is associated with the largest family of phase II enzymes. Note that reactions of second-rank catalyzed by UGT are annotated with a different score than reaction of first-rank catalyzed by UGTs, in order to have homogeneous and comparable scores for reactions with the same rank. (c) Notably, when an edge can be annotated with SOMs associated with different enzymes (especially for isoform predictions), the edge is duplicated for each enzyme to avoid confusion. All reactions which could not be annotated with a SOM score are removed from the metabolic map, as well as the resulting isolated nodes.

As a final step, all the metabolites of the map are associated with a SOM-based pathway production probability score (production probability score). This score depicts the probability for each node to be formed for each metabolites according to all the annotations of the reactions of the metabolic map. The approach relies on the formalism of Bayesian networks [41], a relevant framework to ensure that all possible production pathways are contributing factor to a probability of metabolite production (See methods for details).
Validation of the method: construction and analysis of a map of metabolism for caffeine

As a matter of validation of our modeling approach based on SOM predictions of phase I and phase II xenobiotics metabolisms enzymes, we applied the pipeline to caffeine. Indeed, caffeine metabolism is well described and share enzymes with HAA metabolism, linked to xenobiotics metabolism enzymes, such as CYP1A2, the main enzyme of caffeine metabolism, but also CYP3A4, CYP2E1 and CYP2D6[42]. Caffeine metabolism is also known to involve metabolites produced by NATs, the phase II enzymes of xenobiotic metabolism.

Fig. 3 shows maps of metabolism obtained as several steps of the pipeline applied the molecule of caffeine, modelled by its SMILES formula extracted from pubchem[39]. The first step of the pipeline consisted in prediction of caffeine metabolites according to two transformation steps using SyGMa[22]. SyGMa can make chemical structures predictions using two parameters that define a scenario: (i) The first parameter is a group of SMIRKS reactions to use in the predictions. We choose to use reactions related to xenobiotics metabolism available in SyGMa as two set of reactions named “phase I” and ”phase II”[22]. (ii) The second parameter is a maximum number of transformations that can occur between the original chemical and a predicted metabolite. We choose to set this parameter at 2 due to the fact that the first two reactions of xenobiotics metabolism[2] (phase I and phase II) are the main biotransformation steps which are included in SyGMa SMIRKS reaction sets. The resulting map contained 23 metabolites and 31 reactions shown in Fig. 3.

Figure 3: Map of metabolism of caffeine predicted by the tools SyGMA and Way2Drug Among the 23 metabolites, the caffeine node is shown in red and metabolites which have been experimentally observed in previous studies are shown in green. Among the 31 reactions depicted by arrows, the 13 black arrows depict transformations with a literature-based evidence. Site of metabolism (SOM) score annotating reactions are shown in orange bolt on arrows. They are provided by SOM prediction tools (Way2Drug SOMP, FAME 3).

*Predicted metabolites* Our results are consistent with the literature knowledge [43, 44, 42, 45, 46] since 11 of the 16 known derivatives of caffeine, including caf-
feine itself, are effectively recovered according to two steps of reactions associated with phase I and phase II reactions. This metabolites are shown as green in Fig 3, except for caffeine which is red in the figure. Only two metabolite 5-acetylamino-6-formylamino-3-methyluracil, i.e., AFMU, and 6-amino-5-(N-formylmethylamino)-1,3-dimethyluracil, i.e., 137-TAU are not identified and are both associated to a NAT catalysed reaction. The other known metabolites of caffeine, such as 3-methyluric acid (i.e., 3MU), 7-methyluric acid (i.e., 7MU), and 1-methyluric acid (i.e., 1MU), are out of the scope of the method because they correspond to other conditions including for instance three steps of reactions.

The method predicted 16 reactions between known metabolites (green nodes), including the 13 reactions supported by the literature (black arrows) [45, 46, 42]. Therefore, our method predicted three new possible transformations from 137U (1,3,7-trimethyluric acid, node 4 in the figure) to 13U (node 8), 17U (node 17) and 37U (node 13), through a demethylation. The latter is similar to other reactions of the model. This suggests that the metabolite 137U has the capability to be transformed into several metabolites although this hypothesis has not been tested because of the low quantities produced in human metabolism and its elimination [42].

In addition to 11 known metabolites of caffeine, the method predicted that 12 other compounds are potential caffeine derivatives, which are called caffeine metabolites (blue nodes). Indeed the metabolite node 5 (Cn1c(=O)c2c(n(C)c1=O)[n+]- (O)cn2C) is predicted to be derived from caffeine after one step reaction. Ten other compounds are predicted to be derived from caffeine after two step reactions. Among these ten metabolites, three (nodes 9, 14, 18) are also predicted to be derivatives of node 5. A single metabolite (node 22) is predicted to be derived only from the newly metabolite node 5, which has the specificity to lose the aromatic structure of the imidazole, a chemical structure which is improbable according to our knowledge.

Production probability scores of predicted metabolites In order to estimate the confidence of the predicted metabolites, we computed production probability scores for each of them. As detailed in the Methods section, the production probability score first takes into account scores associated with reactions, which are computed from several site of metabolism (SOM) scores predicted by dedicated tools (Way2Drug SOMP, FAME3), e.g. the chance that a transformation occurs on a given atom. In addition, the production probability score for the metabolites also take into account the different pathways, e.g. chains of reaction, from caffeine to the considered metabolite, combined according to a Bayesian framework.

The SOM score associated to reactions are indicated as orange label on arrows in Fig. 3. We notice that five reactions have a null score. The production probability score for metabolites are shown in Fig. 4a, where metabolites are ordered from top-score metabolites to lowest scores. The 11 known metabolites (green nodes) have the largest scores. The 12 unknown predicted metabolites are represented with a blue bar.

We observe that these predicted metabolites can be gathered into three groups, each of them is shown by an ellipse with a specific color. The first group (black ellipse) describes metabolites with a score greater than 0.70: it contains all known
metabolites predicted by our method and two unknown metabolites, node 9 (Cn1c(=O)c2[nH]c[n+]O)c2n(C)c1=O) and node 11 (Cn1c(=O)c2c(ncn2C2OC-(C(=O)O)C(=O)C(O)C2O)n(C)c1=O). The second group (violet ellipse) contains three derivatives with a medium score (between 0.20 and 0.70) corresponding to the unknown metabolites node 14, node 5 and node 18. The last group of seven unknown metabolites (yellow ellipse) with low score (< 0.20) corresponding to the nodes 15, 20, 22, 10, 16, 19 and 21. The remaining nodes 10, 16, 19 and 21 have a score of 0.00, which is explained by the fact that all the pathways which produce them contain at least a reaction with a null score.

![Production probability scores](image)

(a) Selection of metabolites according to Bayesian model

(b) Filtered predicted caffeine map of metabolism

Figure 4: Filtered map of metabolism for caffeine. (a) Production probability score of predicted metabolites The production probability score of each metabolite is computing according to a Bayesian model based on the reaction production probability scores. Red, green and blue bars are associated to caffeine, known and unknown metabolites, respectively. Metabolites are grouped into three groups according to their range of scores (black, red and yellow ellipses). (b) Filtered map of metabolism of caffeine This map is the filtration of the previous predicted map in order to keep metabolites with a high production probability score (black ellipse, score > 0.70). The map contains 13 metabolites and 17 reactions. Two unknown metabolites (nodes 9 and 11) are predicted to be equivalently producible than the known metabolites according to the production probability scores.

Filtered caffeine metabolic map The production probability scores were used to filter the caffeine map of metabolism as follows. The metabolites of the filtered map are all metabolites of the first group (black ellipse) in Fig. 4a, and the reactions
in the map are those of Fig. 3 transforming nodes in the black ellipse group. The filtered map is shown in Fig. 4b. We observe that all the known 11 metabolites and 13 reactions belong to this map and therefore were conserved by the filtration procedure.

The filtering procedure removed metabolites associated with nodes 5, 14, 18 and 22, which have a specific configuration regarding nitrogen atoms on the imidazole part. These four metabolites are the only ones that are oxidized derivatives of caffeine with a pattern N=C-N or N-C-N such that one of the nitrogens is methylated and the other nitrogen is linked to an oxygen atom.

The other metabolites that were eliminated by the filtration procedures are nodes 10, 15, 16, 19, 20, and 21. They are all glucuronyl-conjugates of caffeine and caffeine metabolites whose glucuronyl group is not associated with the methylated nitrogen atom of the caffeine imidazole part. The fact that most of the glucuronyl conjugates are filtered is consistent with literature because there is no glucuronyl-conjugate metabolite of caffeine described in human yet. The only glucuronyl-conjugate appearing in the final map is node 11 (with SMILES formula: Cn1c(=O)c2c(ncn2C2OC(C(=O)O)C(O)C(O)C2O)n(C)c1=O, or IUPAC name: 6-(1,3-dimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-7-yl)-3,4,5-trihydroxyoxoxane-2-carboxylic acid), whose glucuronyl group is linked to the imidazole part.

This suggests that node 1 (theophylline or 13x), which is already known to be metabolized into nodes 6 (3-methylxanthine or 3MX), 7 (1-methylxanthine or 1MX) and 8 (1,3-dimethyluric acid or 13U) could have also the potential to be biotransformed into another new metabolite. However the 13U and 1MX which are the most known produced metabolites from 13X [42], could compete with this new metabolite and make it undetectable.

Apart to node 11, the only unknown metabolite of the initial map conserved after the filtration is node 9 (Cn1c(=O)c2c[nH]c[n+]|O)c2n(C)c1=O or 9-hydroxy-1,3-dimethyl-2,6-dioxo-2,3,6,7-tetrahydro-1H-purin-9-imin). According to our method, this metabolite is a derivative of node 1, e.g. 1,3-dimethylxanthine (theophylline or 13x), with a SOM score predicted by Way2Drug of 0.714 for enzyme CYP3A4. This reaction from node 1 is an oxidation on a nitrogen atom of the imidazole part. Contrary to the oxidation of the carbon between nitrogen atom in the imidazole part which produced node 8 (13U) from node 1 with a predicted SOM score of 0.606 for enzyme CYP2D6, with the same order of magnitude, this reaction occurs on an un-methylated nitrogen atom of the imidazole part. This suggests than oxidation on nitrogen atoms of caffeine could theoretically occurs although it has never been experimentally observed.

Application to Heterocyclic Aromatic Amines (HAA) and DNA reactivity predictions

The caffeine example suggests that there is an added-value to build map of metabolism by combining several approaches such as assembling reaction prediction tools, predicting sites of metabolism and filtering the map according to a production probability score. Based on this validation, we further investigated how this method may facilitate the prediction of DNA adducts formation derived from xenobiotics. To that matter, we first constructed the predicted maps of metabolism of
the 30 human HAAs (see Additional file 2 for details). Then we annotated maps of metabolism for six HAAs of interest in order to study DNA adducts formation prediction.

**Unfiltered maps of metabolism of HAAs** The pipeline was applied to predict maps of metabolism of the 30 identified HAAs. The characteristics of the maps predicted by the tool SygMa are described in Table 1. HAAs are ordered according to the number of metabolites in the maps predicted by the SyGMa tool. HAAs associated with the largest map are 4-CH2OH-8-MeIQx and 4,7,8-TriMeIQx (194 predicted metabolites). The smallest maps correspond to the HAA Harman (70 metabolites) and NorHarman (50 metabolites). The size of these maps could be explained by the chemical structure of both HAAs that have few sub-structures on which SyGMA transformation rules can be applied. The maps predicted for two of the three well characterized HAAs in primary human hepatocytes [11, 12, 14], MeIQx - 155 metabolites - and PhIP - 128 metabolites, have a medium size. On the contrary, AcC, the third one, is associated with one of the smallest metabolic map.

We also noticed that the maps associated with pairs of HAAs such as 7,8-DiMeIQx and 7,9-DiMeIQGx, MeIQx and 7,MeIQGx or IQ and IQ[4,5-b] are associated with maps with similar characteristics: they have the same number of metabolites, reactions and metabolites reactive to DNA. The closeness of the chemical structure of isomers could explain such a similarity between maps: as plane isomers are composed of the same chemical substructures, the transformation rules contained in the SyGMA database have a high probability to occur equivalently on the derivatives of both HAAs.

For each HAAs, we estimated the number of metabolites reactive to DNA by assuming that each metabolite with a XenoSite Reactivity score greater than 0.85 is reactive to DNA, following the criteria introduced in [36]. The four HAAs, 4-CH2OH-8-MeIQx, 7,8-DiMeIQx, 7,9-DiMeIQGx and 4,7,8-TriMeIQx have both the largest map and the greatest amount of metabolites reactive to DNA (91, 80, 81 and 81). NorHarman HAA characterized by the smallest map contains no metabolite reactive to DNA. Harman, AMPNH and APNH are also associated with a very small ratio of metabolites reactive to DNA (12.7%, 10.1% and 8.1%). The ratio for the other HAAs ranges from 38.8% to 55.8%. MeIQx and PhIP have a relatively high ratio of metabolites reactive to DNA (45.2% and 44.5%) while AcC has a lower ratio (38.8%) in spite of its known higher reactivity towards DNA compared with MeIQx or PhIP [14]. This observation might be related to its smallest map of metabolism (average and median of numbers of metabolites and reactions in the map). Several hypotheses can be made about the variability of the sizes of the maps of metabolism: (a) the metabolites reactive to DNA do not have the same importance *in vitro*, (b) the reactions predictions performed by SyGMa may be incomplete, as we observed it for NAT2 reactions in the case of caffeine, (c) the reactions performed by SyGMA may not be homogeneous. As detailed below, the analysis of the production probability scores of the maps suggests that the two last hypotheses are highly probable.

**Manual annotation and filtering of six maps of metabolism** As the pipeline for the study of maps of metabolism encompasses a part of manual annotations for atom
<table>
<thead>
<tr>
<th>HAA</th>
<th>Metabolites</th>
<th>Reactions</th>
<th>Metabolites which are reactive to DNA</th>
<th>Ratio of metabolites reactive to DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,7,8-TriMeIQx</td>
<td>194</td>
<td>282</td>
<td>91</td>
<td>46,9</td>
</tr>
<tr>
<td>4-CH2OH-8-MeIQx</td>
<td>189</td>
<td>266</td>
<td>80</td>
<td>42,3</td>
</tr>
<tr>
<td>7,8-DiMeIQx</td>
<td>174</td>
<td>250</td>
<td>81</td>
<td>46,6</td>
</tr>
<tr>
<td>7,9-DiMelgQx</td>
<td>174</td>
<td>250</td>
<td>81</td>
<td>46,6</td>
</tr>
<tr>
<td>4,8-DiMeIQx</td>
<td>174</td>
<td>250</td>
<td>79</td>
<td>45,4</td>
</tr>
<tr>
<td>6,7-DiMelgQx</td>
<td>169</td>
<td>245</td>
<td>76</td>
<td>46,0</td>
</tr>
<tr>
<td>AMPNH</td>
<td>157</td>
<td>225</td>
<td>16</td>
<td>10,1</td>
</tr>
<tr>
<td>7-MelgQx</td>
<td>155</td>
<td>220</td>
<td>70</td>
<td>45,2</td>
</tr>
<tr>
<td>MeIQx</td>
<td>155</td>
<td>220</td>
<td>70</td>
<td>45,2</td>
</tr>
<tr>
<td>GluP1</td>
<td>142</td>
<td>202</td>
<td>64</td>
<td>45,1</td>
</tr>
<tr>
<td>IQx</td>
<td>137</td>
<td>192</td>
<td>62</td>
<td>45,3</td>
</tr>
<tr>
<td>IgQx</td>
<td>133</td>
<td>188</td>
<td>56</td>
<td>42,1</td>
</tr>
<tr>
<td>TrP1</td>
<td>129</td>
<td>179</td>
<td>61</td>
<td>47,3</td>
</tr>
<tr>
<td>PIP</td>
<td>128</td>
<td>177</td>
<td>57</td>
<td>44,5</td>
</tr>
<tr>
<td>MeIQ</td>
<td>125</td>
<td>175</td>
<td>64</td>
<td>51,2</td>
</tr>
<tr>
<td>4'-OH-PhIP</td>
<td>123</td>
<td>166</td>
<td>49</td>
<td>39,8</td>
</tr>
<tr>
<td>3,5,6-TMIP</td>
<td>122</td>
<td>172</td>
<td>57</td>
<td>46,7</td>
</tr>
<tr>
<td>APNH</td>
<td>120</td>
<td>165</td>
<td>10</td>
<td>8,3</td>
</tr>
<tr>
<td>MeαC</td>
<td>113</td>
<td>154</td>
<td>48</td>
<td>42,5</td>
</tr>
<tr>
<td>TrP2</td>
<td>113</td>
<td>154</td>
<td>49</td>
<td>43,4</td>
</tr>
<tr>
<td>GluP2</td>
<td>110</td>
<td>151</td>
<td>46</td>
<td>41,8</td>
</tr>
<tr>
<td>IQ</td>
<td>109</td>
<td>150</td>
<td>59</td>
<td>54,1</td>
</tr>
<tr>
<td>IQ[4,5-b]</td>
<td>109</td>
<td>150</td>
<td>59</td>
<td>54,1</td>
</tr>
<tr>
<td>1,5,6-TMIP</td>
<td>107</td>
<td>153</td>
<td>59</td>
<td>55,1</td>
</tr>
<tr>
<td>1,6-DMIP</td>
<td>95</td>
<td>132</td>
<td>53</td>
<td>55,8</td>
</tr>
<tr>
<td>IFP</td>
<td>90</td>
<td>123</td>
<td>37</td>
<td>41,1</td>
</tr>
<tr>
<td>ααC</td>
<td>85</td>
<td>111</td>
<td>33</td>
<td>38,8</td>
</tr>
<tr>
<td>PheP1</td>
<td>76</td>
<td>97</td>
<td>35</td>
<td>46,1</td>
</tr>
<tr>
<td>Harman</td>
<td>70</td>
<td>98</td>
<td>9</td>
<td>12,7</td>
</tr>
<tr>
<td>NorHarman</td>
<td>50</td>
<td>65</td>
<td>0</td>
<td>0,0</td>
</tr>
<tr>
<td>Average</td>
<td>127,6</td>
<td>178,7</td>
<td>53,7</td>
<td>41,0</td>
</tr>
<tr>
<td>Median</td>
<td>124</td>
<td>173,5</td>
<td>58</td>
<td>45,2</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of the maps of metabolism predicted according to biotransformations rules by the SyGMA tool for 30 HAA.
number of reactions, we applied the pipeline to a selection of six HAAs among
the 30 HAAs. We first selected the three well described HAAs in human hepa-
tocytes AαC, PhIP and MeIQx [11, 12, 14, 13]. We complemented this list with the
two HAAs with the largest map (4-CH2OH-8-MeIQx and 4,7,8-TriMeIQx). We fi-
nally selected 7,8-DiMeIQx which represents the pair of isomers (7,8-DIMeIQx and
7,9-DiMeIQx) having large maps. Note that the three selected HAAs (4-CH2OH-8-
MeIQx, 4,7,8-TriMeIQx and 7,8-DIMeIQx) have also the largest amount of metabo-
lites reactive with DNA. We applied the prediction pipeline and the annotated maps
of metabolism obtained were explored to identify enzyme families and isoforms as-
associated with reactions. We confirmed that most enzymes families (SULTs, NATs,
CYPs, UGTs) and CYPs isoforms (CYP1A2, CYP2C19, CYP2C9, CYP2D6 and
CYP3A4) annotate at least one reaction in each map with the exception of GSTs
that are not found in annotated map.

The annotated maps of metabolism obtained from AαC, PhIP and MeIQx were
explored in order to identify the metabolites corresponding to the metabolites de-
scribed in humans of these three AHAs. Upon the 11, 10 and 9 derivatives experi-
mentally shown for AαC, MeIQx and PhIP, 9 and 7 are found in the annotated
maps of metabolism. Among the 8 known derivatives not present in annotated maps
of metabolism, three are N-sulfonyl derivatives of each HAA and three are N-acetoxy
derivatives of each HAA. This suggests that N-sulfonyl and N-acetoxy derivatives
may not be predicted using SyGMA’s SMIRKS rules, supporting the hypothesis (b)
above. The two last known derivatives of MeIQx, 7-oxo-MeIQx and N-desmethyl-
7-oxo-MeIQx are not present in annotated maps of metabolism but we identi-
fied two metabolites, with SMILES formula Cc1nc2c(ccc3c2nc(N)n3C)n1O and
Cc1nc2c(ccc3[nH]c(N)nc32)nc1O, which are close to these known missing deriva-
tives. The main structural difference between those metabolites and 7-oxo-MeIQx
and N-desmethyl-7-oxo-MeIQx is that the ketone group of the 7-oxo-MeIQx part
is replaced by an hydroxyl group. This result suggests also that SyGMA’s SMIRKS
rules are not able to predict ketone group linked to a carbon of an heterocycle, still
supporting the hypothesis (b) above.

Figs. 5a, 5b and 5c, show that the distribution of the production probability scores
of the known metabolites of each AHA is rather scattered with scores ranging from
0.9 to 0.2. Surprisingly, the metabolite MeIQx-N2-SO3H, in the map of metabolism
of MeIQx, is the only experimentally identified metabolite associated with a pro-
duction probability score of 0.0. A null score suggests that the metabolic pathways
leading to the metabolite contains at least one reaction that has not been annotated
either by Way2Drug for a reaction of rank 1, or by FAME 3 for a reaction of rank
2.

More generally, we notice that PhIP and MeIQx have a high amount of metabolites
associated with a low production probability score, compared to AαC. 83 metabo-
lites (53.5% of the map) of the MeIQx map have a score lower than 0.1, including
36 metabolites with a null score (23.2% map). Similarly, 65 metabolites (50.8% of
the map) of the PhIP map have a score lower than 0.1 including 18 metabolites
with a null score (14.1% map) and no known metabolite. On the contrary, only 26
metabolites (30.6% of the map) of the AαC map have a score lower than 0.1, in-
cluding 13 metabolites with a null score (15.3% map). Based on this remark, and
ratio of metabolites predicted to be reactive to DNA (between 40% and 50%). This tends to homogenize the size of the final maps of metabolites, and a similar trend also homogenizes the size of the final maps of metabolites with little support according to our model.

After comparing the effect of different thresholds, we used the value 0.10 to filter metabolites with little support according to our model. The origin compounds of each map is shown with a red bar. Metabolites known as metabolites with little support according to our model. Details about metabolites contained in these maps are provided in Table 2. Table 2 describes the characteristics of the six selected filtered maps of metabolism. We observe that the filtering procedure based on prediction probability scores favors the detection of metabolites with lower reaction rates with DNA. However, this filtering step is not entirely based on prediction probability scores, but also takes into account the threshold value used.

**Figure 2: Distribution of prediction probability scores in annotated maps of PhIP.**

- (a) Distribution of the prediction probability score of each metabolite of the annotated map of PhIP.
- (b) Distribution of the prediction probability score of each metabolite of the annotated map of 7,8-DiMeIQx.
- (c) Distribution of the prediction probability score of each metabolite of the annotated map of 4',7,8-TrDiMeIQx.
suggests that there is a strong interest in using SOM scores (as included in the production probability scores) to homogenise maps of metabolism and eliminate unsupported DNA reactive metabolites.

**Optimal enzymatic signature in terms of DNA reactivity** As described previously, the pipeline relies on the computation for SOM scores on reactions involving manually annotated metabolites to reduce the maps of metabolism according to a Bayesian prediction probability. As SOM scores are directly related to enzymes, the production probability score is influenced by enzymes availability. We define enzymatic contexts to be tables describing all the possible combinations of enzymes that may be considered as available. In our study, there are 512 such different enzymatic context. Each enzymatic context is associated with a specific distribution of production probability scores. Indeed, when an enzyme is described as unavailable in an enzymatic context, the reactions annotated with this enzyme cannot be taken in account for the calculation of the production probability scores.

Based on this assumption, for each of the six filtered maps of metabolism described above, and for each of the 512 enzymatic context, we computed the production probability scores of all metabolites of the map. This allowed us to determine the global reactivity score of a HAA in a given context that we defined as the sum of the production probability score (in the considered context) of all metabolites reactive to DNA in the considered map of metabolism.

In this framework, it becomes possible to define optimal enzymatic signatures in terms of reactivity, which are all the enzymatic contexts where the global reactivity score is maximized while they contain the smallest number of activated enzymes. Intuitively, optimal enzymatic signatures therefore correspond to enzymatic contexts where the chance to obtain at least one metabolite reactive to DNA is maximal according to our models.

**Impact of enzymatic context to the production probability score and application on DNA adduct formation of HAAs** Fig. 6 shows all optimal enzymatic signatures for the six HAA.
Figure 6: **Optimal enzymatic signatures in terms of reactivity.** A blue cell corresponds to an available enzyme. A grey cell corresponds to an unavailable enzyme. The threshold to determine if a metabolite is reactive to DNA was 0.85.

We observe that the enzymes UGTs, SULTs and CYP1A2 are present in all optimal signatures of the six HAA. This is consistent with the literature which describes the implication of SULTs and CYP1A2 in the formation of HAA DNA adducts [2]. In addition, it has been recently shown that UGTs an also be involved in a pathway leading to DNA adduct formation for \( \alpha \)C[14].

Conversely, the enzymes NATs and GSTs are absent in all optimal signatures. While the absence of GSTs is explained by the fact that annotated maps of metabolism do not contain any GSTs, the absence of NATs suggest that the resulting metabolites are not reactive to DNA since NATs are present in all maps of metabolism. However, NAT2 has been previously involved in formation of DNA adducts derived from HAAs [2]. We hypothesise that NATs reaction implicates other isoforms in the maps of metabolism. In accordance with this hypothesis, NAT2 enzymes did not appear in the map of metabolism predicted for caffeine since it is involved in the only known missing metabolite (AFMU).

Fig. 6 suggests that HAA is characterized by a specific enzymatic profile. The profiles according to the availability of the CYP isoforms other than CYP1A2: CYP3A4 is available in four of the six optimal signatures and correspond to all HAAs with a MeIQx chemical structure. This suggests a different involvement of CYP isoforms, other than CYP1A2, in the formation of metabolites that are highly regarded as reactive and depending on the structure of the HAA.

**Impact of the XenoSite Reactivity threshold** In order to test the impact of the reactivity threshold chosen to characterize all the compounds of a DNA reactive map, we decided to explore all the thresholds from 0 to 1.0. For the 101 threshold values considered (step of 0.01), we recalculated all the metabolites considered as "reactive" (i.e. associated with a XenoSite Reactivity score greater than or equal to the threshold), and then calculated the optimal signatures for reactivity to DNA. According with our previous results we did not consider NATs and GSTs. As shown in Fig. 7, the enzymes CYP1A2, UGTs and SULTs are present in the optimal signature whatever the threshold thereby suggesting that the result described in
Fig. 6 for the reactivity threshold value 0.85 are robust. The Fig. 7 also confirms the specificity of CYP3A4 in optimal signatures of HAAs with an MeIQx chemical structure.

We further observed a low variability between HAAs since the largest optimal signature is reached at a reactivity threshold of 0.89 for AαC, 0.91 for 4-CH2OH-8-MeIQx, 0.92 for MeIQx and PhIP, 0.93 for 4,7,8-TriMeIQx and 0.95 for 7,8-DiMeIQx. In addition, the use of any reactivity threshold lower than 0.89 returns the same optimal signature for each HAA. This suggests that enzymes involved in metabolic pathways leading to the most reactive metabolites (given by XenoSite Reactivity scores), are sufficient to activate all the pathways leading to the less reactive metabolites.

The Fig. 8 is a counter-part of Fig. 7 to compare the values of XenoSite reactivity associated with the apparition of each enzyme in an optimal signature. When considering high XenoSite Reactivity thresholds, we observed that cytochromes P450 isoforms are the only enzymes present in all optimal signatures (thresholds from 0.96 to 0.97). This suggests that CYPs are responsible of the production of most of DNA-reactive metabolites, especially CYP1A2 found in all HAAs. Therefore, reactive metabolites derived from CYPs-annotated reactions may form DNA-adduct more easily than reactive metabolites derived from phase II enzymes-annotated reaction. In addition, we observed that the enzymes UGTs and SULTs were present in the optimal signatures for reactivity thresholds between 0.95 and 0.89. For MeIQx, 4-CH2OH-8-MeiQx, 7,8-DiMeIQx and 4,7,8-TriMeIQx, we noted that the reactivity thresholds for which SULTs is present in the optimal signature is greater than the one for UGTs. This suggests that SULTs-associated metabolites are more likely to form DNA adducts than UGTs-associated metabolites when the chemical structure of the HAA is close to the MeIQx structure.

**Discussion and conclusion**

In this study, we introduced a pipeline for predicting the metabolism of xenobiotics in humans. The pipeline was applied to six HAA of interest and caffeine. Our predic-
Figure 8: **Optimal signatures in terms of reactivity for the highest reactivity thresholds.** A grey cell means that the enzyme is not a part of the optimal signature and a colored cell mean the enzyme is available in the optimal signature. Each lines represent the optimal signature at a specific reactivity score threshold. Thresholds marking the integration of new enzyme in the optimal signature are annotated on each bar.

The pipeline is based on the construction of so-called *enriched maps of metabolism*. The main specificity of the pipeline is to use *production probability score* to sort metabolites according both to prediction of site of metabolisms and the topology of the maps predicted by bio-transformation rules. This score allows comparing metabolite production in different physiopathological conditions that permit to explore the role of enzymes in the production of specific metabolites. In this study we focus on metabolites and their DNA adduct formation capacity.

The pipeline was used to reconstruct the maps of caffeine metabolism and of six HAAs. Among the four xenobiotics for which the metabolism was known i.e., MeIQx, PhIP, AoC and caffeine, the majority of their metabolites described in humans were found. Among these known metabolites no one were removed from the maps after the filtration on the production probability scores, with the exception of MeIQx-N2-SO3H.

In the maps of metabolism predicted by our pipeline, most of experimentally identified caffeine and HAA metabolites are associated with a high prediction probability score. On the contrary a large number of predicted metabolites, which can be considered as over-predicted metabolites, are not supported by site of metabolism predictions.

This allowed us to use prediction probability scores as filters that importantly reduced the size of the maps. In the case of HAAs, half of the predicted metabolites were filtered out. Thus map of metabolism produced by other tools could be filtered by this production probability score.
The prediction of the map of caffeine and HAA metabolism highlighted areas for improvement in our pipeline. First, the study of the caffeine map of metabolism evidenced that a metabolite (AFMU), resulting from a NAT2-directed reaction, is not predicted by the SyGMa tool, suggesting that the tool is incomplete in predicting reactions catalyzed by N-acetyl transferases. This assumption is also supported by the analysis of the optimal HAAs signatures as detailed in figure 6. The prediction of HAA maps of metabolism also suggested lacks of prediction of N-Sulfonyl and N-Acetoxy derivatives of 1αC, PhIP and MeIQx, which were not predicted because there are no SMIRKS rules in SyGMa adapted to the prediction of these metabolites. These different unpredicted metabolites advocate for relying on the capability of the SyGMa tool to add new SMIRKS biotransformation rules to its prediction rules in order to complete the maps of metabolism with relevant enzymes.

A characteristic of the map of metabolism of MeIQx is that MeIQx-N2-SO3H, an experimentally identified metabolite, is associated with a null prediction probability score. This is explained by the fact that the reaction producing MeIQx-N2-SO3H is labeled as a rank 1 reaction, catalyzed by the SULTs. Our pipeline differentiate rank 1 and rank 2 reactions and could not annotate reactions of rank 1 with SO score only available for tools annotating reactions of rank 2. To overcome this issue, we plan to differentiate the SOMs prediction tools according to the enzymes annotating the reactions instead of the rank of the reactions. This however requires to homogenize the level of information about isoforms.

The production probability score that we defined allowed us to analyze the influence of enzymes on the production of DNA reactive metabolites and to propose a specificity of the CYP3A4 enzyme in the production of DNA adducts derived from AHAs close to MeIQx, which will be the subject of further experiments.

In conclusion, our study describes a new method for the construction and analysis of maps of metabolism by combining prediction of biotransformation rules, predictions of site of metabolisms, and prediction of reactivity to DNA. The method was validated and applied to six xenobiotics. The further study will consist in applying the pipeline to the 24 other human HAAs, which requires to automatize the annotation of metabolites predicted by transformation rules. Moreover, our approach based on enzymes profile determination open the perspective to predict various xenobiotics derived metabolites susceptible to bind DNA adducts in both normal and physiopathological situations. The enzymatic contexts extracted from data repositories such as TCGA [47] and GTEX [48] makes this goal achievable.

**Method**

**Tool for the prediction of edges and nodes of maps of metabolism: SyGMa**

The SyGMa python package (Systematic Generation of potential Metabolites) [22] is a rule-based method to predict metabolite (e.g. derivative compounds) from an input chemical compound. In the paper, the input compound was either caffeine or HAA. The method relies on an internal set of metabolic reactions (biotransformation rules) in SMIRKS format which can be applied to the input compound. If the input chemical structure of a SMIRKS reaction is detected in the compound, the reaction is applied and the resulting structure obtained is a predicted metabolite.

The first parameter required by SyGMa to make prediction is the set of SMIRKS reactions to use. In this paper, we used the two sets of SMIRKS reactions, named
"phase I" and "phase II", obtained by data mining the Metabolite Database\cite{22} and corresponding to reactions involved in phase I and II metabolism of xenobiotics.

The second parameter required by SyGMa is the number of reactions (which we call rank) that separates the original compound and a metabolite. If this maximal rank number is greater than 1, the metabolites obtained after a first iteration are used as source for new reactions to obtain second rank metabolites, until the maximal rank is reached. In the paper, the maximal rank number was equal to 2. This allows reproducing the main observed xenobiotic metabolism with a first reaction associated to phase I enzymes of xenobiotic metabolism and a second reaction that conjugates the oxidized metabolite by enzymes of phase II metabolism of xenobiotics.

Tools for the prediction of sites of metabolism (SOM)

We use tools for site of metabolism (SOM) prediction tools to annotate reactions that could be catalysed by phase I or phase II enzymes of xenobiotics metabolism. All methods we used for predicting SOMs consists in inferring models which can be applied to an atom configuration according to a set of valid reactions which is specific to each model, extracted from a database. Using these models, each method can evaluate for each atom of a compound if it can be involved in a reaction similar to those described in the reaction database. This analysis results in a score for each atom of a compound that describes the probability to be transformed by each reaction. Therefore, the methods differ according to the parameters describing the atom configuration and the enzymes involved in the reactions (different enzyme family and/or isoforms).

In order to take advantage of the panel of existing methods, our procedure include different tools to annotate each edge (reaction) with a site of metabolism (SOM) prediction score depending on its rank, enzyme label and atom number of reaction (see results section).

Way2Drug SOMP\cite{29} is used to predict SOMs associated with reactions applied to the original compound as source (first-rank reactions). This SOM predictor has different CYP models and can provide a SOM score for each atom of a compound for five the main cytochrome isoforms: 1A2, 3A4, 2D6, 2C9 and 2C19. It also predicts SOM scores for the UGT enzyme family but does not specify any isoforms.

When input of a reaction is not the original compound, we used FAME 3\cite{31} tool, which provides SOM score for phase II enzymes of xenobiotics metabolism. We use specific models restrained to specific reactions catalyzed by the five enzyme families: N-acetyltransferase (NATs), Sulfotransferases (SULTs) and Glutathione S-transferases (GTSs), UDP-glucuronosyltransferase (UGTs) and cytochromes P450 (CYPs).

Prediction of SOR

XenoSite Reactivity\cite{9, 10} is used to annotate the DNA reactivity of each metabolite. Based on the SOM scores, this tool computes a score of reactivity (SOR) for each atom of the metabolite meaning ability of the atom to bind DNA. It also relies on the atom configuration, described by molecular descriptors, and uses deep-learning to infer a model that predicts the probability for an atom be involved in a
specific set of reactions that are DNA-binding reactions. To determine if a metabolite is considered as reactive to DNA, we apply a threshold on the SOR scores if at least one SOR score of an atom of the metabolite is retained the metabolite is considered as reactive to DNA. We use the same threshold as in [36] where XenoSite Reactivity where a threshold of 0.85 was learned according to metabolites known to be reactive to DNA.

*Scoring metabolites with a SOM-based pathway production probability score* A specific method was designed to compute a probability for each node to be formed for each metabolites according to all the annotations of the reactions of the metabolic map of metabolism. The approach relied on the formalism of Bayesian networks [41] which are probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks were used to predict the the production probability score of each metabolite of the metabolic map, assuming that all possible production pathways are contributing factor to this score.

The graph (DAG) on which probabilities were computed was derived from the map of metabolism based on the principles that (1) when several isoforms of the same enzyme can transform a compound to another, the total affinity of the compound with the enzyme family can be approximated by the maximal isoform enzyme affinity (2) when several members of several enzyme families are competing to produce the same target metabolite from different input metabolites, the recruitment of enzymes in reactions follows an exclusivity principle such that each enzyme family can catalyze the production of the targeted metabolite with at most one reaction.

Consequently, a variable of the Bayesian model was build for each node (e.g. metabolites) of the map of metabolism. It was therefore associated with the event production of the metabolite. The DAG was built according to the following rules, which are illustrated in Fig.9. (a) For reactions which has the same input and output and therefore varied only by the isoform of their enzyme family (in our case, CYPs), we selected the edge with the maximal SOM score and removed the other edges of the graph. (b) For each metabolite which is the output of several edges with different input, we selected the enzyme family (UGTs, NATs, SULTs, CYPs, GSTs) with the largest SOM score and removed all the other edges producing the targeted metabolite with the same enzyme family. Then we selected the remaining enzyme family producing the targeted metabolite with the second largest score and again removed from the graph all the edges leading to the targeted metabolite with an enzyme of the same enzyme family. This was repeated until each enzyme family could produce the targeted metabolite from at most an input compound.

For each edge, the SOM score was interpreted as a conditional probability, which is the probability to get the output compound of the edge assuming the presence of input compound. The structure of the graph, which is both acyclic (because the pipeline for building metabolic maps cannot create cycles) and such that enzymes do not compete for the production of a same metabolite, yields that conditional probabilities are independent. This allowed creating a complete probability table associated with each metabolite production event.
Figure 9: Creating a Bayesian network from an annotated map of metabolism Considering A) An annotated map of metabolism with enzymes named E1, E2, E3 and E4 and a SoM score associated to each enzyme for each reactions. B) Illustrate the application of the rule: ”only one enzyme can catalyze a reaction”. There is a reaction from A to C, divided in three reactions one for each enzyme annotating the reaction. Those reactions need to be reduced in one with a SoM score used as probability for the Bayesian Network. SoM score annotating reactions are indicators of enzymatic affinity so the reaction with maximal SoM score is selected. Here it is the reaction annotated by E3. C) Illustrate a different case where there is reactions from D and from C leading to production of F. This is the application of ”if an enzyme is recruited by a reaction, another enzyme has to be recruited for a second reaction with the same product”. Here the maximal SoM score annotating reactions is 0.77 for the reaction from D to F annotated by E1. Because this reaction is the retain reaction from D to F, E1 is no more allowed to annotate the reduced reaction from C to F. The reaction retain from C to F, is the second reaction with the maximal SoM score which is the reaction annotated by E2. D) Illustrate the final reduced graph used as Bayesian Network to calculate production probability score of metabolites.

The production probability score of each metabolite was defined to be the probability of a metabolite production. These scores were computed by using the conditional probability tables to expand joint probability function (Bayes formula) [41].

Acknowledgements
Not applicable

Funding
The work has supported by the Institut National de la Santé et de la Recherche Médicale (Inserm), University of Rennes 1, Ligue contre le Cancer du Grand Ouest, PNREST Anses cancer TMOI AVIESAN 2013/1/166.

Abbreviations
HAA: heterocyclic aromatic amines
Availability of data and materials
The description of HAAs by SMILE formula were provided in [36]. The SMILES formula of caffeine was extracted from PubChem: https://pubchem.ncbi.nlm.nih.gov/. They are all available in the Supp. data 2.

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
The authors declare that they consent for publication.

Authors' contributions
M.C. developed and implemented the pipeline and produced all the results. N.T. and S.L. analyzed the biological consistency of the results. A.S. contributed to the design of the pipeline, the choice of the different parameters and the data analysis. All authors contributed to the manuscript. A.S. and S.L. contributed equally to the direction of the work.

Author details
1 Irisa, UMR 6074, Univ Rennes, Inria, CNRS, Rennes, FR. 2 Irisa, UMR 6074, Univ Rennes, Inria, CNRS, Rennes, FR.

References


Additional Files

Additional file 1 - Dictionary of SMIRKS rules
The file provides a manually constructed catalogue of SMIRK rules required to map SMIRKS rules to enzyme family labels by taking into account the rank of the reaction in the pipeline for building enriched maps of metabolism.

Additional file 2 - Description of thirty HAAs and Caffeine structure used as pipeline input
This pdf file describes the SMILES formula and 2D-structure of each of the thirty HAAs and the Caffeine studied in the paper.

Additional file 3 - Description of six HAA maps of metabolism
The file provides a detailed description of metabolites for the six HAAs filtered maps of metabolism built in the paper. For each metabolite of each map, the file provides the identifier of the metabolite, its SMILES formula, its production probability score, its reactivity to DNA and the score of XenoSite Reactivity.
Figure 1

Representation of Heterocyclic Aromatic Amine Metabolism. This metabolism is divided in two steps: Phase I (in green) is known as "oxidation reaction" and catalyzed by cytochromes P450 (CYPs) and Phase II (in red) is known as a "conjugation reaction", catalyzed by UDPglucuronyltransferases (UGTs), N-acetyltransferases (NATs), Sulfotransferases (SULTs) or Glutathione S-transferase (GSTs)
Figure 2

Definition and construction of enriched maps of metabolism (a) Labels of a map of metabolism. Metabolites are depicted by their 2D structure in black squares. Arrows between metabolites represent reactions, which consume a metabolite to produce another one. Left side: labels are shown in dashed squares (green for metabolites and purple for reactions). Right side: examples of values of labels for a part of the left side network. (b) Pipeline to obtain labels for metabolites and reactions. Labels are depicted in green (metabolites) and purple (reactions) squares. They are computed thanks to different tools and methods (grey squares) which require as input the information provided by other labels. SMIRKS labels dictionary is detailed in Additional file 1.
Map of metabolism of caffeine predicted by the tools SyGMA and Way2Drug. Among the 23 metabolites, the caffeine node is shown in red and metabolites which have been experimentally observed in previous studies are shown in green. Among the 31 reactions depicted by arrows, the 13 black arrows depict transformations with a literature-based evidence. Site of metabolism (SOM) score annotating reactions are shown in orange bolt on arrows. They are provided by SOM prediction tools (Way2Drug SOMP, FAME 3).
Filtered map of metabolism for caffeine. (a) Production probability score of predicted metabolites. The production probability score of each metabolite is computed according to a Bayesian model based on the reaction production probability scores. Red, green and blue bars are associated to caffeine, known and unknown metabolites, respectively. Metabolites are grouped into three groups according to their range of scores (black, red and yellow ellipses). (b) Filtered predicted caffeine map of metabolism. This map is the filtration of the previous predicted map in order to keep metabolites with a high production probability score (black ellipse, score >0.70). The map contains 13 metabolites and 17 reactions. Two unknown
metabolites (nodes 9 and 11) are predicted to be equivalently producible than the known metabolites according to the production probability scores.

![Figure 5](image)

(a) Distribution of the prediction probability score of each metabolite of the annotated map of metabolism of AαC

(b) Distribution of the prediction probability score on each metabolite of the annotated map of metabolism of MeIQx

(c) Distribution of the prediction probability score on each metabolite of the annotated map of metabolism of PhIP

**Figure 5**

Distribution of prediction probability scores in annotated maps of metabolism of AαC (a), MeIQx (b) and PhIP (c). The origin compound of each map is shown with a red bar. Metabolites known as metabolites experimentally described are associated to a green bar and other metabolite are associated to a blue bar. The names of metabolites shown in X-axis are either the name found for known metabolites or a two
parts name: Unknown + X where X describes the identifier of the node in the map associated with this metabolite.

<table>
<thead>
<tr>
<th></th>
<th>CYP 1A2</th>
<th>CYP 3A4</th>
<th>CYP 2C19</th>
<th>CYP 2C9</th>
<th>CYP 2D6</th>
<th>UGTs</th>
<th>NATs</th>
<th>SULTs</th>
<th>GSTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,8-DiMeIQx</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>4,7,8-TriMeIQx</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>4-CH2OH-8-MeIQx</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>AaC</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>MeIQx</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>PhIP</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

**Figure 6**

Optimal enzymatic signatures in terms of reactivity. A blue cell corresponds to an available enzyme. A grey cell corresponds to an unavailable enzyme. The threshold to determine if a metabolite is reactive to DNA was 0.85.

**Figure 7**

Optimal enzymatic signatures in terms of reactivity for each reactivity score thresholds. Each line describe an optimal enzymatic signature for each HAA for a reactivity score threshold.
Figure 8

Optimal signatures in terms of reactivity for the highest reactivity thresholds. A grey cell means that the enzyme is not a part of the optimal signature and a colored cell means the enzyme is available in the optimal signature. Each line represents the optimal signature at a specific reactivity score threshold. Thresholds marking the integration of new enzyme in the optimal signature are annotated on each bar.
Figure 9

Creating a Bayesian network from an annotated map of metabolism Considering A) An annotated map of metabolism with enzymes named E1, E2, E3 and E4 and a SoM score associated to each enzyme for each reactions. B) Illustrate the application of the rule: "only one enzyme can catalyze a reaction". There is a reaction from A to C, divided in three reactions one for each enzyme annotating the reaction. Those reactions need to be reduced in one with a SoM score used as probability for the Bayesian Network. SoM score annotating reactions are indicators of enzymatic affinity so the reaction with maximal SoM score is selected. Here it is the reaction annotated by E3. C) Illustrate a different case where there is reactions from D and from C leading to production of F. This is the application of "if an enzyme is recruited by a reaction, another enzyme has to be recruited for a second reaction with the same product". Here the maximal SoM score annotating reactions is 0.77 for the reaction from D to F annotated by E1. Because this reaction is the retain reaction from D to F, E1 is no more allowed to annotate the reduced reaction from C to F. The reaction retain from C to F, is the second reaction with the maximal SoM score which is the reaction annotated by E2. D) Illustrate the final reduced graph used as Bayesian Network to calculate production probability score of metabolites.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupFile1.pdf
• Supfile2.pdf
• SupFile3.pdf