Metabolic markers for the yield of lipophilic indole alkaloids in dried woad leaves (Isatis tinctoria L.)

Thi-Kieu-Oanh Nguyen, Paulo Marcelo, Eric Gontier, Rébecca Dauwe

Title:

Metabolic markers for the yield of lipophilic indole alkaloids in dried woad leaves (*Isatis tinctoria* L.)

Authors (4):

Thi-Kieu-Oanh Nguyen

Department of Pharmacological Medical and Agronomical Biotechnology, University of Science and Technology of Hanoi

Vietnam Academy of Science and Technology

18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam

nguyen-thi-kieu.oanh@usth.edu.vn

Paulo Marcelo

Plateforme ICAP, Centre Universitaire de Recherche en Santé, Université de Picardie Jules Verne

Avenue Laënnec, 80054 Amiens cedex 1, Amiens, France

paulo.marcelo@u-picardie.fr

Eric Gontier

EA3900 BioPI, UFR Sciences, Université de Picardie Jules Verne

33 rue Saint Leu, 80039 Amiens cedex, France

eric.gontier@u-picardie.fr

Corresponding author:

Rebecca Dauwe

EA3900 BioPI, UFR Sciences, Université de Picardie Jules Verne

33 rue Saint Leu, 80039 Amiens cedex, France

rebecca.dauwe@u-picardie.fr

Mobile phone: (+33) (0)6 73 80 95 29

Phone: (+33) (0)3 22 82 76 81

Phone & FAX (assistant): (+33) (0)3 22 82 76 48

ABSTRACT

The pharmacologically active dichloromethane extracts of dried woad leaves (*Isatis tinctoria* L.), and the methanol extracts of comparable fresh leaves of the same plants, were analyzed by LC-MS^n. The fresh leaf metabolite profile revealed a complex pattern of indolic compounds. Besides the known indigo precursors, isatan A, isatan B and indican, seven previously unreported indole derivatives were characterized: acetylindican, malonylindican, two dioxindole glucosides, dioxindole malonylglucoside, 6-hydroxyindole-3-carboxylic acid 6-O-glucoside and 6-hydroxyindole-3-carboxylic acid glucose ester. The integration of 122 compounds in fresh leaves and of five selected compounds (indoxyl, isatin, indigo, indirubin, and tryptanthrin) in dried leaves, formed the input data for a stepwise modelling procedure generating five predictive linear models. The structure of the predictive models and a cross validation provide evidence that the models could predict well or moderately well the accumulation of the selected lipophilic compounds, and were simple enough to be used in a woad cultivation program. PLS regression models relating each of the five selected dry leaf indolics to the fresh leaf metabolome were then fitted in order to deduct potential precursors and mechanisms leading to the formation of these lipophilic indolics in drying woad leaves. The models suggested glucobrassicin, isatan A and isatan B as the main candidate precursors of these compounds, besides a minor contribution of other fresh leaf indolics, including malonylindican, actylindican and dioxindole malonylglucoside. Dioxindole malonylglucoside was identified here as isatan C. The models further suggested that the accumulation of phenylpropanoid antioxidants in woad leaves has a negative impact on the formation of indoxyl, isatin, indigo, indirubin and tryptanthrin.

KEYWORDS

Isatis tinctoria L., Brassicaceae, LC-MS, indigo, tryptanthrin, isatin, stepwise variable selection, PLS regression, metabolic markers

Introduction

Isatis tinctoria L. (woad, Brassicaceae) is a biennial herbaceous plant that was extensively cultivated in Europe from the 12th till the 17th century for the production of the indigo dye (Clark et al., 1993; Hurry, 1930). In a lengthy process, the woad leaves were crushed into a pulpy paste, kneaded into "woad balls" and dried. Spontaneous aerobic fermentation of the dried paste under the addition of water and urine led then to the "finished woad", which was used for dying. The import of cheaper indigo from *Indigofera spp* in the late 17th century initiated the decline of the woad industry in Europe and finally, in the late 1890's, the production of synthetic indigo completely replaced natural indigo production. During the period when woad cultures were flourishing in Europe, the plant was also indicated for medicinal use. Renaissance herbals, which are 16th century treatises on the use of medicinal plants, recommend woad for the treatment of wounds, ulcers and tumors, hemorrhoids, snake bites and various inflammatory diseases (Hamburger, 2002; Hurry, 1930). Woad has an equally rich history as dye and medicinal herb in China and, until now, root and leaf extracts from the taxonomically closely related *Isatis indigotica* Fort. remain popular herbal drugs for the treatment of inflammatory diseases in traditional Chinese medicine (Kang et al., 2014; Mohn et al., 2009). Recently, the rising demand for naturally sourced dyes and the medicinal potential of woad have driven a renewed interest in the plant. Since the year 2000, numerous pharmacological studies have been dedicated to woad and its bio-active constituents. A comparative pharmacological profiling of lipophilic and polar extracts from fresh and dried *Isatis tinctoria* roots and leaves, in various animal models, showed that, of all extracts tested, the lipophilic extracts from dry leaves display the strongest anti-inflammatory potential (Hamburger, 2002). Among the best characterized bio-active constituents of *I. tinctoria* are the indole alkaloids isatin, tryptanthrin, and indirubin. Isatin is a useful chemical scaffold for a variety of chemical transformations and this synthetic versatility makes isatin a widely applicable pharmacological molecule. Isatin derivatives exhibit antiviral, anti-inflammatory, anticonvulsant and antitumor activities, among others (Silva, 2013; Vine et al., 2009). Tryptantrin has potent inhibitory activity on prostaglandin and leukotriene synthesis and on inducible NO synthase (Danz et al., 2002a; Danz et al., 2002b; Danz et al., 2001; Ishihara et al., 2000; Oberthur et al., 2005; Ruster et al., 2004). Indirubin is a red indigoid pigment and inhibits inflammatory reactions in delayed-type hypersensitivity and is a potent inhibitor of cyclin-dependent kinase 5 (CDK5) and glycogen synthase kinase 3β (GSK3B) (Hoessel et al., 1999; Leclerc et al., 2001).

The alkaloids indigo, indirubin, isatin, and tryptanthrin are not true secondary metabolites resulting from biosynthetic pathways. They are undetectable or only found at trace levels in rapidly frozen fresh *I. tinctoria* leaves and their appearance depends on the post-harvest treatment. Indigo and indirubin, for example, are indigo dyes that appear under the fermentative conditions that were used in the ancient natural indigo production. Their formation is also induced by a drying process at 40 °C. Tryptanthrin, on the other hand, is formed by this same drying process at 40 °C, but does not appear under the fermentative conditions used in the ancient indigo production. Isatin is considered to be a "by-product" that appears when indigo dyes are formed (Maugard et al., 2001). Questions regarding the nature of putative precursors and the mechanisms leading to the formation of these alkaloids have attracted the interest of chemists since the 19th century but remain largely unresolved today. These alkaloids all contain at least one indole moiety, and *I. tinctoria* indeed contains high amounts of indole metabolites, mainly the indole-type glucosinolates glucobrassicin and neoglucobrassicin, but also the indoxyl derivatives indoxyl-β-D-glucoside (indican), indoxyl-β-3'-keto-D-glucoside (Isatan B), and 6'O-malonyl-isatan B (Isatan A). Previous studies have focused on reactions under mild acid or alkaline conditions of

purified woad metabolites or on TLC separated methanolic woad extracts. These experiments showed for example that isatans A and B, as well as indican are, at least *in vitro*, indigo precursors (Oberthür et al., 2004; Oberthur et al., 2004). Similar to the GUS staining reaction that is widely used in molecular biology, under the acid or alkaline conditions tested, the indoxyl moiety is released from the precursors, dimerizes and is oxidized to form indigo. Indirubin is formed when isatin is incubated, in alkaline solution, with the indoxyl derivative isatan B or with an as yet uncompletely characterized dioxindole derivative named isatan C (Maugard et al., 2001). Isatin, in turn, is thought to be an oxidation product of indoxyl, produced in oxygen rich environment (Maugard et al., 2001). In accordance with a role as indigo precursor, isatan A and B totally disappear during the drying process of woad leaves. Nevertheless, it cannot be deduced from the *in vitro* reactions what their contribution is to the formation of indole alkaloids in plant leaves. Because most studies have analyzed reactions involving pure reagents, the contribution of the different indolic metabolites, their potential interactions and also the influence of the overall metabolite composition onto the formation of the post-harvest indolics in woad leaves remain unclear. A potential contribution of the strongly abundant indole glucosinolates has for example not been investigated, although these metabolites also disappear during the drying process. Furthermore, an increase of indican levels when woad leaves are dried has suggested the presence of higher molecular weight acylated indican derivatives, and on TLC separated methanolic woad extracts, several as yet unknown blue or red pigment forming compounds are apparent (Oberthur et al., 2004). It seems thus that the formation of pigments and other indolic oligomers in drying woad leaves may be a rather complex process.

In the context of the renewed interest in woad as a herbal drug or for the production of natural dyes, breeding programs targeting the post-harvest production of indole alkaloids could be considered. The use of genetic markers such as single nucleotide polymorphisms (SNPs) in selection programs is an approach amenable to high throughput and is therefore becoming standard for breeders. The production of post-harvest oligomeric indole alkaloids is most probably a complex trait, influenced by many genes, but on the other hand directly related to the leaf metabolome at the moment of harvest. In such case, metabolic markers are a logic alternative to genetic markers (Fernandez et al., 2016). It has been shown that, for complex traits in maize, 130 metabolites were almost as good predictors as 38000 SNPs (Riedelsheimer et al., 2012). The cost for the use of metabolic markers is much higher, per sample, as compared to the cost for use of a molecular marker, but depends strongly on the number of metabolic markers that are measured. If a small set of metabolic markers with relatively high predictive power can be found, LC-MS targeted profiles should enable high numbers of samples to be processed at low costs, enabling screens of large populations or complex experimental setups (Fernandez et al., 2016). Moreover, metabolic markers have the additional advantage that they may open the way to mechanistic insights and in that way be highly relevant.

Here, we hypothesized that, for the prediction of post-harvest indolic alkaloid formation in woad leaves, it should be possible to construct simple models based on a limited number of polar metabolites, and still explaining a relatively high amount of the variance. We tested this hypothesis based on a large scale metabolite profiling of the rosette leaves of a large number of *I. tinctoria* individuals (27 individuals). Half of the rosette leaves of each plant was quenched in liquid nitrogen directly at harvest and was subjected to untargeted metabolite profiling. The other half of the rosette leaves was air-dried and the accumulation of a selected number of indole alkaloids that are formed during the drying process were

relatively quantified. We then analyzed the predictive power of small sets of polar fresh leaf metabolites for the post-harvest formation of the lipophilic indole alkaloids.

Results and discussion

Metabolite profiling

In a targeted analysis by HPLC-Hybrid Ion Trap-Orbitrap Mass Spectrometry of the dried rosette leaves, we measured the relative content of five apolar indole derivatives that have been previously described in *I. tinctoria*: indoxyl, isatin, indigo, indirubin, and tryptanthrin (Figure 1). These compounds were identified here based on high resolution-MS, MS^n, and literature data on their fragmentation and chromatographic retention (Mohn et al., 2009). A strong variation in the intensities of the five dry leaf indolics was observed among the 27 biological replicates, with coefficients of variation going from 52% for isatin to 97% for indoxyl. A relatively strong positive correlation was observed mutually among the "end products" indigo, indirubin and tryptanthrin, but not between indoxyl and isatin and their presumed oxidation products (Table 1).

The polar extracts of the rapidly frozen rosette leaves were subjected to large scale untargeted metabolite profiling by HPLC-Hybrid Ion Trap-Orbitrap Mass Spectrometry. Based on accurate masses obtained in the orbitrap and a nested fragmentation approach obtained in the LTQ linear ion trap, in which MS^2 first product ions were further fragmented into MS^3 second product ions and these second product ions were, in turn, fragmented into MS^4 third product ions, we characterized and relatively quantified a total of 122 compounds of different metabolic classes (supplemental table 1). These compounds comprised three primary metabolites, 90 phenylpropanoids (flavonoids, hydroxycinnamic acid derivatives, and mono- and oligolignols), two benzenoids, seven glucosinolates of which three contained an indole moiety, ten other indole derivatives, one anthranilic acid derivative, and nine apocarotenoids. Most of these compounds have been described earlier as *I. tinctoria* leaf metabolites (Mohn et al., 2009; Nguyen et al., 2017; Oberthür et al., 2004) but seven of the indole derivatives are described here for the first time in *I. tinctoria* leaves. We have recently described the MS^n fragmentation pattern of the phenylpropanoids, benzenoids and glucosinolates in *I. tinctoria* leaves (Nguyen et al., 2017). For the indole derivatives, anthranilic acid derivative, apocarotenoids and primary metabolites, detailed information on the structural characterization is given in the supplemental data, online. The indole derviatives are depicted in Figure 1. They comprised the three known indigo precursors (isatans A and B and indican) (Oberthür et al., 2004; Oberthur et al., 2004). Besides, indican was also present under acylated form as acetylindican and malonylindican. The presence of acylated indican derivatives has been predicted earlier based on TLC analyses of extracts of rapidly frozen *I. tinctoria* leaves (Oberthur et al., 2004). The malonyl derivative of indican shows a strong resemblance to isatan A, which is also a malonylester of isatan B. Further, we characterized two dioxindole glucosides and a dioxindole malonylglucoside. These compounds are candidates for the partially characterized dioxindole derivative that was reported in young *I. tinctoria* leaves and was named "isatan C" (Maugard et al., 2001). Finally, a 6-hydroxyindole-3-carboxylic acid 6-O-glucoside (6-GlcO-ICOOH) and a 6-hydroxyindole-3-carboxylic acid glucose ester (6-HO-ICOOGlc) were characterized here in the fresh leaf extracts. These compounds have not been described in *I. tinctoria* before but are known as pathogen inducible indolic metabolites in Arabidopsis leaves (Böttcher et al., 2014). Interpretation of the mass spectral fragmentation of the indolic compounds led to three general observations. First, fragmentation spectra of the dioxindole

derivatives were characterized by a strong base peak at m/z 160, which was attributed to the formation of a highly stable 2,3-methylenedioxyindole. Secondly, all malonyl esters (isatan A, malonylindican, and dioxindole malonylglucoside) easily lost a carbon dioxide moiety, both by in source fragmentation leading to an $[M\text{-}CO2\text{-}H]^-$ ion in the MS^1 spectrum, and by collision induced dissociation (CID) leading to a neutral loss of 44 Da in the MS^2 spectrum. Further fragmentation of the first product ion $[M\text{-}CO2\text{-}H]^-$ lead then to the neutral loss of an acetic acid moiety (CH3COOH, 60 Da), resulting in a peak representing the dehydrated demalonylated compound $[M\text{-}C3H4O4\text{-}H]^-$ in the MS^3 spectrum. Finally, malonyl derivatives (isatan A, malonylindican, and dioxindole malonylglucoside) had retention times on a reversed phase column that were about one minute longer than their unmalonylated counterparts (isatan B, indican, dioxindole glucoside).

Prediction models – stepwise variable selection

We built individual linear models, fitting the abundance of indigo, tryptanthrin, indirubin, isatin, or indoxyl in dry leaves, as a function of the abundance of a limited set of metabolite predictors in the fresh leaves. Generally, adding variables improves the model fit, and, if the added variables are meaningful, also the prediction performance of the model. However, simpler models can be more readily applied. Therefore, simpler, yet effective models are often favored over more complex models even with extreme prediction power. Here, the determination of the polar metabolites showing the strongest predictive power for the lipophilic indole alkaloids was achieved by a step-wise selection procedure minimizing the Bayesian information criterion (BIC). BIC is an estimate of the loss of information in the model, and penalizes for the addition of predictors (i.e. metabolites). The procedure thus selects a model that fits well, but has a minimum number of predictor metabolites. After each step in the step-wise selection procedure, the prediction accuracy of the model was calculated and expressed as the RMSEP and as the Pearson correlation coefficient (cc) between predicted and observed values (see Experimental section). BIC, RMSEP and cc were plotted for the models obtained with increasing numbers of considered steps (Supplemental Figure 1). For indirubin, the step-wise procedure was stopped after adding two predictors to the model (glucobrassicin with positive coefficient and a flavonoid with negative coefficient) based on the dropping BIC value, although no strong prediction accuracy was reached (RMSEP 0.88; cc 0.47). For the other Y-variables, the BIC value only dropped after the addition of more than 25 predictor variables and the RMSEP values continued to decrease steadily until 20 to 25 predictors were added. However, strong correlation coefficients between predicted and observed values were obtained with much simpler models. Because we judged that the measurement of more than twenty predictors can be problematic for a cost-effective approach, and that, on the other hand, a correlation between predicted and observed variables of 0.8 can be highly relevant, we decided to further simplify the models by forcing the selection procedure to stop when this correlation threshold was reached. For indoxyl, isatin, indigo and tryptanthrin, the threshold of prediction accuracy was reached with four, five, three and five predictors, respectively (cc 0.81, 0.82, 0.82 and 0.82 between predicted and measured values). The selected predictors, their corresponding coefficients and the prediction accuracy of the prediction models are presented in table 2. The coefficient of each metabolite in the model indicates the relative influence, with a positive or negative sign reflecting a positive or negative relationship between a metabolite and the accumulation of the lipophilic indolic in the response. The significance of the prediction accuracy was calculated based on a distribution of prediction accuracies obtained with randomly permutated response vectors (Figure 2, see Experimental section) and expressed as a p-value. The simplest and at the same

time one of the most accurate models was thus obtained for indigo. Isatan B and dioxindole malonylglucoside contributed positively to the amounts of indigo in dried leaves, whereas a dilignol hexoside had a relatively strong negative relationship with indigo accumulation. In addition to contributing to predicting indigo content, isatan B was also the strongest predictor for indoxyl (an intermediate in indigo formation), and tryptanthrin. No other indolic compound was selected as tryptanthrin predictor and yet the tryptanthrin prediction model showed a relatively strong prediction accuracy (cc 0.82), with only five predictor metabolites. On the other hand, two flavonoids (vicenin 2 and luteolin-6-C-glucoside-7-O-glucoside) had a relatively strong negative impact on the tryptanthrin prediction model. Remarkably, isatan A, which is by far the most abundant indoxyl derivative in fresh woad leaves (Oberthür et al., 2004), contributed negatively to the prediction model of indoxyl and did not contribute positively to any of the models. Surprisingly, none of the known "indigo precursors" (isatans A and B and indican), was selected as a predictor for isatin or indirubin accumulation. Instead, glucobrassicin (indolylmethyl glucosinolate) was the only predictor with positive coefficient in the indirubin model and glucotropaeolin (benzyl glucosinolate) was the predictor with the strongest positive coefficient in the isatin model. Dihydroascorbigen, which is a breakdown product of glucobrassicin, also contributed positively to the indoxyl model.

Exploring the mechanisms - Partial Least Squares (PLS) regression

We further examined the relationship between the full fresh leaf metabolome and the accumulation of selected indole alkaloid products in dried leaves, in order to deduct potential precursors and mechanisms leading to the formation of the latter compounds. Using leave-one-out cross-validated PLS-regression, we related the fresh leaf metabolome (X matrix) to each of the five dry leaf indole alkaloids separately (Y variable). Models with moderate predictive power could be obtained for indigo ($R^2y = 0.35$; $Q^2 = 0.31$) and tryptanthrin ($R^2y = 0.30$; $Q^2 = 0.32$) and a model with very low predictive power could be obtained for indirubin ($R^2y = 0.16$; $Q^2 = 0.08$). The weak predictive power of both the PLS regression model for indirubin and the model obtained by stepwise variable selection, indicates that another factor, which was not included in the list of 122 fresh leaf metabolites, plays a determining role in the formation of indirubin. Surprisingly, although stepwise variable selection led to simple linear models that fitted indoxyl and isatin relatively well, no well-fitting PLS regression model could be obtained for these compounds ($R^2y = 0.11$ and $R^2y = 0.01$ respectively). The advantage of the use of PLS models for the exploration of the mechanisms, with respect to the stepwise variable selection approach described above, is that PLS efficiently handles the full X matrix including the correlated variables. In correlation circle plots for each of the five Y variables, we plotted the correlations between the variables and the first three components of the PLS regression models (Figure 3). Indirectly, correlations between X-variables, and between X- and Y-variables can be derived from these plots. Highly similar plots were obtained for the indigo, indirubin and tryptanthrin models, indicating that these compounds could be generated from a common set of precursors. When component 1 and component 2 were considered, a strong correlation was apparent between glucobrassicin and the Y-variables indigo, indirubin and tryptanthrin. Isatan A and isatan B were also positively correlated with these Y-variables, although slightly less pronounced than glucobrassicin. Other indolics, including malonylindican, acetylindican, dioxindole malonylglucoside, and 6-hydroxyindole-3-carboxylic acid glucose ester, showed a weaker positive correlation with the dry leaf indole alkaloids. No correlation was observed between any of the Y-variables and indican. The plots of the indoxyl and isatin models were also highly similar to those for

indigo, indirubin and tryptanthrin, but the correlation with glucobrassicin was less pronounced, and no strong correlation with isatan A was apparent. Correlation does not imply a causal link but, if indole alkaloids in dry leaves are derived from indolic precursors in fresh leaves, we expect that this will be reflected in a positive correlation, unless another factor is determining the rate of the conversion. Therefore, the above mentioned positively correlated fresh leaf indolics are the first candidates for being precursors of the studied dry leaf alkaloids. The PLS models thus suggest that the most plausible contributors to indigo, indirubin and tryptanthrin formation in dry leaves are glucobrassicin, isatan A and isatan B, but that a range of other indolics may also contribute. A stronger or more plausible contribution was suggested from the acetylated and malonylated forms of indican than from the underived form. However, if a true precursor shows positive correlation with indigo, indirubin or tryptanthrin, positive correlation of this true precursor with other fresh leaf indolics does not imply a role as precursor for the latter indolics. The authentic contribution of each candidate precursor should therefore be further tested, for example through feeding studies with labeled derivatives. Among the X-variables, a strong positive correlation was observed within the group of mono- and oligolignols, within a large group of hydroxycinnamic acids, to lesser extent within the group of flavonoids, and between these three groups of phenylpropanoids. When component 1 and component 3 were considered, a strong negative correlation was observed between the mono- and oligolignols, hydroxycinnamic acids and flavonoids, on one hand, and all studied dry leaf indolics on the other hand. These phenylpropanoids are potent antioxidants, due to the presence of phenolic hydroxyl groups, and therefore they could negatively impact the oxidative reactions leading to the formation of the post-harvest indolics in dry leaves.

It has been reported earlier that a dioxindole derivative extracted from *I. tinctoria* leaves, reacts *in vitro*, in alkaline medium, with isatan B to form indigo (Maugard et al., 2001). The name isatan C was proposed for this compound but the molecular weight and formula were not conclusively resolved. The reported mass spectral peak of this compound at m/z 398 corresponds however to the m/z of protonated dioxindole malonyl glucoside. Therefore, and because the present models are in accordance with a possible contribution of dioxindole malonylglucoside to the formation of indigo in woad leaves, we propose to assign the name “isatan C” to dioxindole malonylglucoside. A possible reaction mechanism leading to the formation of indigo from isatan B and isatan C is presented in Figure 4. When the leaves dry, indoxyl and dioxindole are liberated from respectively isatan B and C, and radicals may be formed through oxidation of their hydroxyl function, after which the radical may delocalize to the C2 position. Radical-radical coupling, followed by two consecutive hydrogen rearrangements can then lead to the formation of indigo.

In conclusion, an LC-ESI-MS^n analysis led here to the characterization of seven previously unreported indolic metabolites in *I. tinctoria* leaves. One of these compounds was dioxindole malonylglucoside, which we named isatan C. From an applied point of view, we propose simple statistical models, based on a small number of predictor metabolites, that predict moderately well the accumulation of indoxyl, isatin, indigo and tryptanthrin in dried woad leaves. The present models are based on a relatively low number of observations obtained in a homogeneous set of plants. If, in subsequent studies, the models can be generalized to various cultivars, developmental stages and environmental conditions, a cost-efficient strategy targeting the predictor metabolites could be developed to screen high numbers of plants for selection purposes. From a more fundamental point of view, the simple predictive models, in combination with PLS regression models, led to a number of hypotheses regarding the precursors and mechanisms leading to the formation of indigo and related alkaloids in dry leaves. These hypotheses to

be tested are that, in drying woad leaves, (i) glucobrassicin, isatan A, and isatan B are important precursors of indigo, indirubin and tryptanthrin, (ii) indigo is formed partially as the oxidative coupling product between isatans B and C, and (iii) the formation of lipophilic indolic products depends on oxidative conditions and is negatively influenced by the accumulation of phenylpropanoid antioxidants.

Experimental

Growth conditions and plant material

Seeds of *Isatis tinctoria* L. (strain French Woad) were kindly provided by the botanical garden of Amiens (Jardin des Plantes), France, where specimens are kept under accession number 061003-3. Seeds were sown in soil and plantlets were grown in a greenhouse under the following conditions: 24 °C, 70% moisture and photoperiod 16h. Rosette leaves were harvested from 11-week-old woad plants. For each of 27 plants, half of the leaves were immediately frozen in liquid nitrogen, and lyophilized. The comparable other half of the leaves of each plant was dried in a thermostated oven at 40 °C. Temperature, relative humidity and weight loss were monitored during the drying process. Constant weight was achieved after 3-4 days. The dried leaf samples were stored at room temperature for a few days in aluminum paper to avoid exposure to light. Immediately before extraction, dried leaf material was ground with a ZM 1 ultra-centrifugal mill (Retsch, Haan, Germany), with 0.75 mm Conidur ring sieve. The frozen leaf material represented the metabolite pattern at the time of harvest (fresh leaf). The oven-dried leaf material represented the medicinal leaf material (dried leaf).

Metabolite extraction

Metabolite extraction was performed by pressurized liquid extraction (PLE) of 1.0 g of frozen and powdered leaf sample, using methanol or dichloromethane as extraction solvents for fresh and dried leaves respectively, following the previously described instruments and conditions (Nguyen et al., 2017). After extraction, the solvents were evaporated in the speed vac. The dried methanol extracts were redissolved in a solution of trimethoxycinnamic acid (internal standard) in water/MeOH/cyclohexane (1/1/2, v/v/v) (10 g/L), of which the polar phase was analyzed by LC-MS. The dried dichloromethane extracts were dissolved in a solution of diazepam (internal standard) in dichloromethan (1 mg/l) for analysis by LC-MS. Internal standards were purchased from Sigma-Aldrich, Saint-Quentin Fallavier, France.

Metabolite profiling

Methanol and dichloromethane extracts were analyzed with an Ultimate module 3000 LC system (Dionex), hyphenated to an LTQ Orbitrap XL hybrid FTMS (Thermo Electron) via an electrospray ionization source. The separation was performed on a reversed phase Sunfire C18 column (150 x 3 mm, 3.5 mm; Waters, Milford, MA), coupled to a guard column (20 x 3.0 mm; Waters, Milford, MA) with 0.1% aqueous formic acid as solvent A and acetonitrile as solvent B. The flow rate was 300 ml per min and the column temperature was 40 °C. For the methanol extracts of fresh leaves, the following gradient was applied: 0 min, 5% B; 2 min, 5% B; 32 min, 18% B; 60 min, 24% B; 65 min, 100% B. For the dichloromethane extracts of dried leaves, the following gradient was applied: 0 min, 5% B; 2 min, 5% B; 32 min, 93% B; 49 min, 93% B; 55 min, 100% B. Ionization was in the negative mode for the methanol extracts (spray voltage 4 kV, capillary temperature 275 °C, sheath gas 20 (arb), aux gas 50 (arb)) and in

the positive mode for dichloromethane extracts (spray voltage 4.2 kV, capillary temperature 250 °C, sheath gas 20 (arb), aux gas 5 (arb)). Full Fourier transform-MS spectra were recorded between 120 and 1400 m/z. The most abundant ion in the MS^1 spectrum, with a signal threshold of 5000 counts, was fragmented by collision-induced dissociation (CID, 35% collision energy) and the resulting MS^2 spectrum was recorded in the LTQ linear ion trap. In order to obtain additional information for the elucidation of the structures, a few samples were subjected to further fragmentation. For these, four data-dependent MS^n spectra were obtained with 35% collision energy and recorded on the LTQ linear ion trap: an MS^2 spectrum of the base peak of the MS^1 spectrum, followed by MS^3 scans of the base peak and second most abundant ion in the MS^2 spectrum and finally an MS^4 scan of the most abundant second product ion obtained from the base peak in the MS^2 spectrum. Peak detection, integration and chromatogram alignment was performed using XCMS version 1.26.1 (Smith et al., 2006) essentially as described earlier (Nguyen et al., 2017). For peak detection, a matched filtration Gaussian model peak with full with at half maximum of 15 seconds was used and the signal to noise ratio threshold was set at 5.

Stepwise variable selection

All statistical analyses were performed using R version 3.5.3. Linear models fitting the abundance of individual compounds in dried leaves (tryptanthrin, indigo, 5-hydroxyoxindole, indirubin, indoxyl, or isatin) in terms of a selection of fresh leaf metabolite intensities, were generated using the R lm and step functions. To allow for direct comparison between metabolites that exhibited unequal range distributions, the data of each metabolite were centered and scaled using the R function scale. The initial model in the stepwise search was a null model. The R step function selected important predictors among the centered and scaled intensity values for all fresh leaf metabolites by a stepwise procedure which minimized the BIC. The relative influence of each predictor metabolite in the model was expressed as the coefficient of each metabolite in the model equation, which was of the form: μ = (estimate 1) x (selected fresh leaf metabolite 1) + … + (estimate n) x (selected fresh leaf metabolite n); where μ is the expected value of the response compound and n is maximally 5.

Cross validation

After each step in the stepwise procedure, an assessment of the predictive accuracy of the models was performed by cross-validation as described in Robinson et al. (2009) (Robinson et al., 2009). During this cross validation, each sample in the experiment was left out one by one, and the models were refitted for each of the 27 reduced datasets (R update function) and used to predict the responses for the excluded sample (R predict function). The predictive accuracy of the models was expressed as the RMSEP and as the Pearson correlation between the complete sets of measured and predicted intensity values. If the Pearson correlation coefficient reached 0.8, the stepwise procedure was stopped and the model was accepted. If the Pearson correlation was lower than 0.8, the stepwise procedure was continued and another predictor was added or removed if this resulted in a reduced BIC value. To evaluate the significance of the predictive accuracy, a distribution of accuracies was generated by repeating the cross-validation procedure using 500 test datasets consisting of the original dataset in which the response variable was randomly permutated. We then calculated the p-value of the accuracy of a model as the probability of having, in a distribution of accuracies obtained with test datasets, an accuracy that was equal to or higher than the accuracy obtained with the real dataset (R pt function).

PLS regression

The R package mixOmics (Rohart et al., 2017) was used to fit PLS (Partial Least Squares) regression models (pls function), to evaluate their performance using "leave-one-out" validation (perf function), and to generate circular correlation circle plots (plotVar function). The number of components to include in the models for which R^2y and Q^2 values are given (four, three, three, six, and six components for indoxyl, isatin, indigo, indirubin and tryptanthrin models respectively), was chosen based on RMSEP curves (Supplemental Figure 2).

Acknowledgements

This work was developed with the financial support of the Hanoi Science and Technology University (USTH French and Vietnamese consortium), the French Ministry of Higher Education and the Picardie Region Research Council (program Métabo-Typage-Végétal). We thank the botanical garden of Amiens (Jardin des Plantes), France, for kindly providing *Isatis tinctoria* seeds.

References

Böttcher, C., Chapman, A., Fellermeier, F., Choudhary, M., Scheel, D., Glawischnig, E., 2014. The Biosynthetic Pathway of Indole-3-Carbaldehyde and Indole-3-Carboxylic Acid Derivatives in Arabidopsis. Plant Physiology 165, 841-853.
Clark, R. J. H., Cooksey, C. J., Daniels, M. A. M., Withnall, R., 1993. Indigo, woad, and Tyrian Purple: important vat dyes from antiquity to the present. Endeavour 17, 191-199.
Danz, H., Baumann, D., Hamburger, M., 2002a. Quantitative Determination of the Dual COX-2/5-LOX Inhibitor Tryptanthrin in Isatis tinctoria by ESI-LC-MS. Planta Med 68, 152-157.
Danz, H., Stoyanova, S., Thomet, O. A., Simon, H.-U., Dannhardt, G., Ulbrich, H., Hamburger, M., 2002b. Inhibitory activity of tryptanthrin on prostaglandin and leukotriene synthesis. Planta medica 68, 875-880.
Danz, H., Stoyanova, S., Wippich, P., Brattström, A., Hamburger, M., 2001. Identification and isolation of the cyclooxygenase-2 inhibitory principle in Isatis tinctoria. Planta medica 67, 411-416.
Fernandez, O., Urrutia, M., Bernillon, S., Giauffret, C., Tardieu, F., Le Gouis, J., Langlade, N., Charcosset, A., Moing, A., Gibon, Y., 2016. Fortune telling: metabolic markers of plant performance. Metabolomics 12, 158.
Hamburger, M., 2002. Isatis tinctoria – From the rediscovery of an ancient medicinal plant towards a novel anti-inflammatory phytopharmaceutical. Phytochemistry Reviews 1, 333.
Hoessel, R., Leclerc, S., Endicott, J. A., Nobel, M. E., Lawrie, A., Tunnah, P., Leost, M., Damiens, E., Marie, D., Marko, D., Niederberger, E., Tang, W., Eisenbrand, G., Meijer, L., 1999. Indirubin, the active constituent of a Chinese antileukaemia medicine, inhibits cyclin-dependent kinases. Nat Cell Biol. 1(1), 60-67.
Hurry, J. B., 1930. The woad plant and its dye. Oxford University Press, London.
Ishihara, T., Kohno, K., Ushio, S., Iwaki, K., Ikeda, M., Kurimoto, M., 2000. Tryptanthrin inhibits nitric oxide and prostaglandin E(2) synthesis by murine macrophages. European journal of pharmacology 407, 197-204.
Kang, L., Du, X., Zhou, Y., Zhu, B., Ge, X., Li, Z., 2014. Development of a complete set of monosomic alien addition lines between Brassica napus and Isatis indigotica (Chinese woad). Plant Cell Reports 33, 1355-1364.
Leclerc, S., Garnier, M., Hoessel, R., Marko, D., Bibb, J. A., Snyder, G. L., Greengard, P., Biernat, J., Wu, Y. Z., Mandelkow, E. M., Eisenbrand, G., Meijer, L., 2001. Indirubins inhibit glycogen synthase kinase-3 beta

and CDK5/p25, two protein kinases involved in abnormal tau phosphorylation in Alzheimer's disease. A property common to most cyclin-dependent kinase inhibitors? J Biol Chem 276, 251-260.
Maugard, T., Enaud, E., Choisy, P., Legoy, M. D., 2001. Identification of an indigo precursor from leaves of Isatis tinctoria (Woad). Phytochemistry 58, 897-904.
Mohn, T., Plitzko, I., Hamburger, M., 2009. A comprehensive metabolite profiling of Isatis tinctoria leaf extracts. Phytochemistry 70, 924-934.
Morreel, K., Dima, O., Kim, H., Lu, F., Niculaes, C., Vanholme, R., Dauwe, R., Goeminne, G., Inzé, D., Messens, E., Ralph, J., Boerjan, W., 2010a. Mass Spectrometry-Based Sequencing of Lignin Oligomers. Plant Physiology 153, 1464-1478.
Morreel, K., Kim, H., Lu, F., Dima, O., Akiyama, T., Vanholme, R., Niculaes, C., Goeminne, G., Inzé, D., Messens, E., Ralph, J., Boerjan, W., 2010b. Mass Spectrometry-Based Fragmentation as an Identification Tool in Lignomics. Analytical Chemistry 82, 8095-8105.
Nguyen, T. K., Jamali, A., Grand, E., Morreel, K., Marcelo, P., Gontier, E., Dauwe, R., 2017. Phenylpropanoid profiling reveals a class of hydroxycinnamoyl glucaric acid conjugates in Isatis tinctoria leaves. Phytochemistry 144, 127-140.
Oberthür, C., Graf, H., Hamburger, M., 2004. The content of indigo precursors in Isatis tinctoria leaves—a comparative study of selected accessions and post-harvest treatments. Phytochemistry 65, 3261-3268.
Oberthur, C., Jaggi, R., Hamburger, M., 2005. HPLC based activity profiling for 5-lipoxygenase inhibitory activity in Isatis tinctoria leaf extracts. Fitoterapia 76, 324-332.
Oberthur, C., Schneider, B., Graf, H., Hamburger, M., 2004. The elusive indigo precursors in woad (Isatis tinctoria L.) - Identification of the major indigo precursor, isatan A, and a structure revision of isatan B. Chemistry & Biodiversity 1.
Riedelsheimer, C., Czedik-Eysenberg, A., Grieder, C., Lisec, J., Technow, F., Sulpice, R., Altmann, T., Stitt, M., Willmitzer, L., Melchinger, A. E., 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature genetics 44, 217-220.
Robinson, A. R., Dauwe, R., Ukrainetz, N. K., Cullis, I. F., White, R., Mansfield, S. D., 2009. Predicting the regenerative capacity of conifer somatic embryogenic cultures by metabolomics. Plant biotechnology journal 7, 952-963.
Rohart, F., Gautier, B., Singh, A., Lê Cao, K.-A., 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS computational biology 13, e1005752-e1005752.
Ruster, G. U., Hoffmann, B., Hamburger, M., 2004. Inhibitory activity of indolin-2-one derivatives on compound 48/80-induced histamine release from mast cells. Die Pharmazie 59, 236-237.
Silva, B. V., 2013. Isatin, a versatile molecule: studies in Brazil. Journal of the Brazilian Chemical Society 24, 707-720.
Vine, K. L., Matesic, L., Locke, J. M., Ranson, M., Skropeta, D., 2009. Cytotoxic and Anticancer Activities of Isatin and Its Derivatives: A Comprehensive Review from 2000-2008. Anti-Cancer Agents in Medicinal Chemistry- Anti-Cancer Agents) 9, 397-414.

Figure 1 Structures of alkaloids detected in fresh (A) and dried (B) *I. tinctoria* rosette leaves

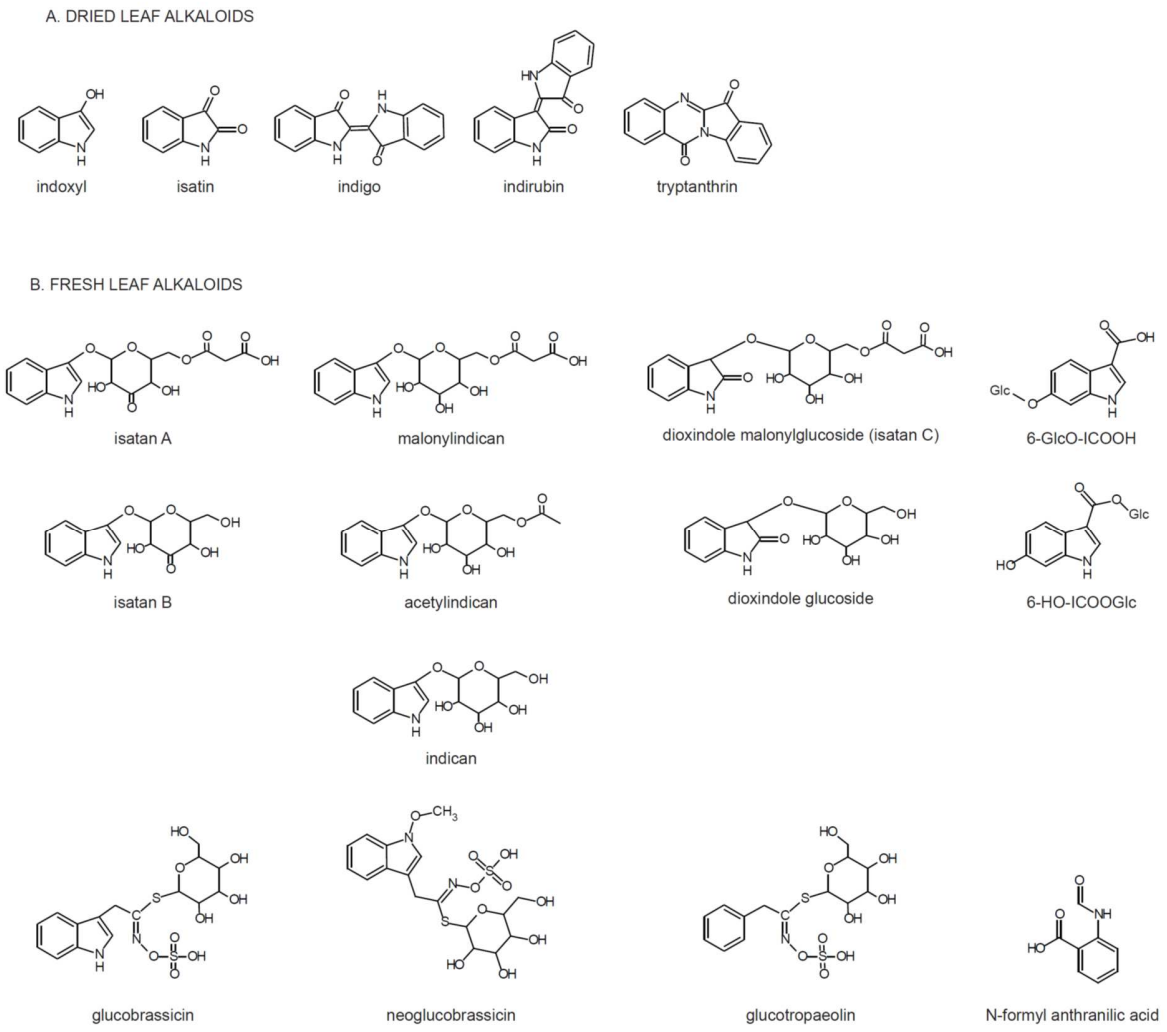
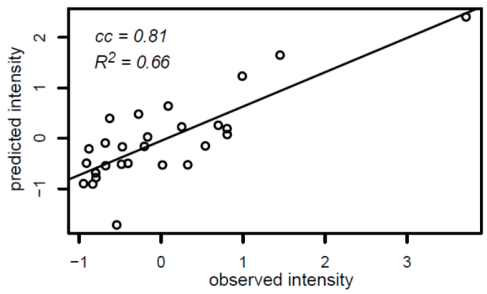


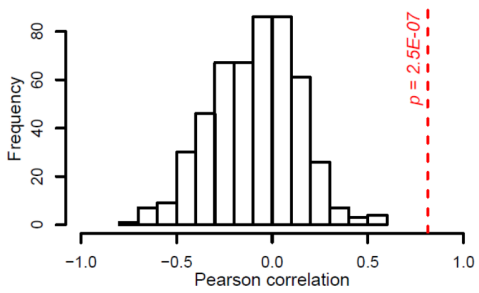
Figure 2 cross validation of prediction models

Left: Plot of measured versus predicted production of each of five indole alkaloids in dried *I. tinctoria* leaves, as determined by complete cross-validation of the prediction model. “Intensity” values on x and y axes are the, respectively observed and predicted, standardized (by internal standard and dry weight) and scaled peak integration values of an extracted ion current peak corresponding to the pseudomolecular ion $[M+H]^+$ of the compound of interest in an LC-MS chromatogram. Each dot represents a single plant. The line represents the fit of least squares regression. The prediction accuracy (Pearson correlation coefficient, cc) and percentage of variance that was predicted (coefficient of determination, R^2) are provided.

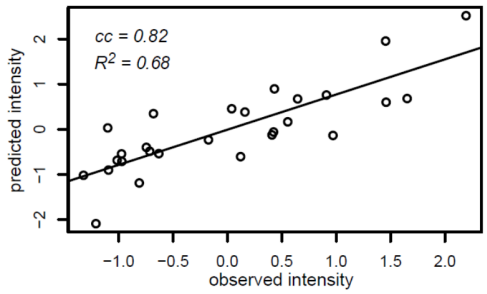
Right: Comparison between the authentic correlation coefficient for observed and predicted intensities (dashed vertical line), and a histogram of the distribution of 500 correlation coefficients generated using a randomly permutated response vector (see materials and methods). The distance between the mean coefficient for the randomly permutated vector and the authentic coefficient is 3.32, 3.26, 3.16, 2.54, and 3.40 standard deviations for indoxyl, isatin, indigo, indirubin and tryptanthrin respectively, and the corresponding upper tail p-values are given on the plots.

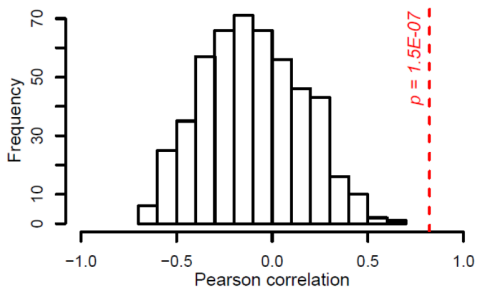
indoxyl



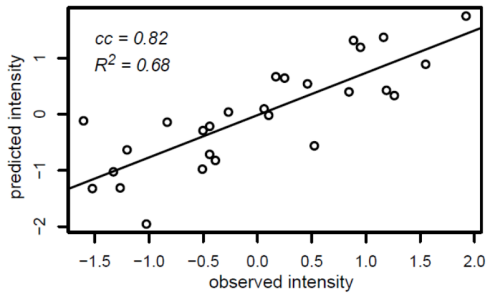


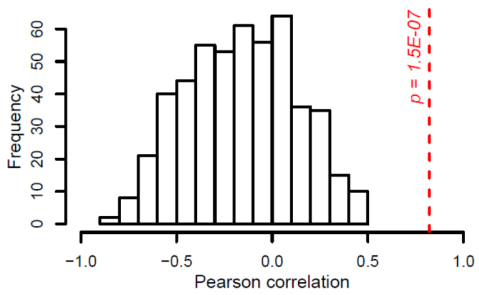
isatin



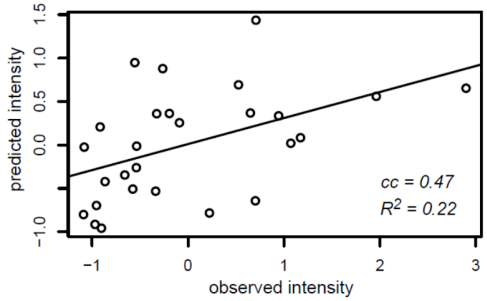


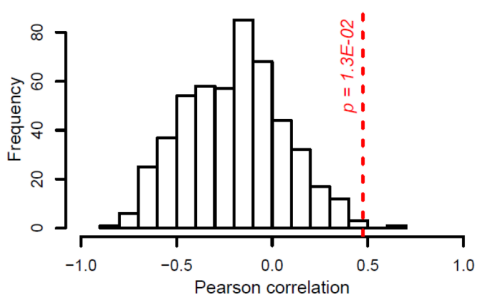
indigo



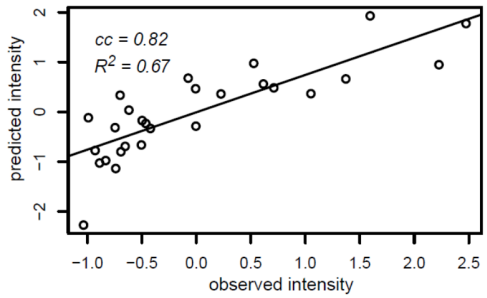


indirubin





tryptanthrin



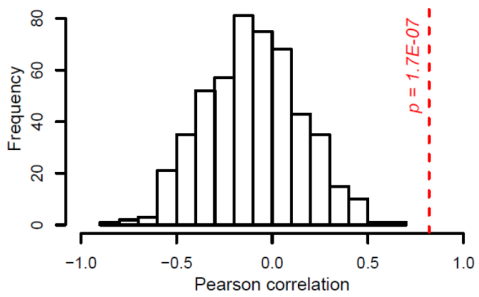
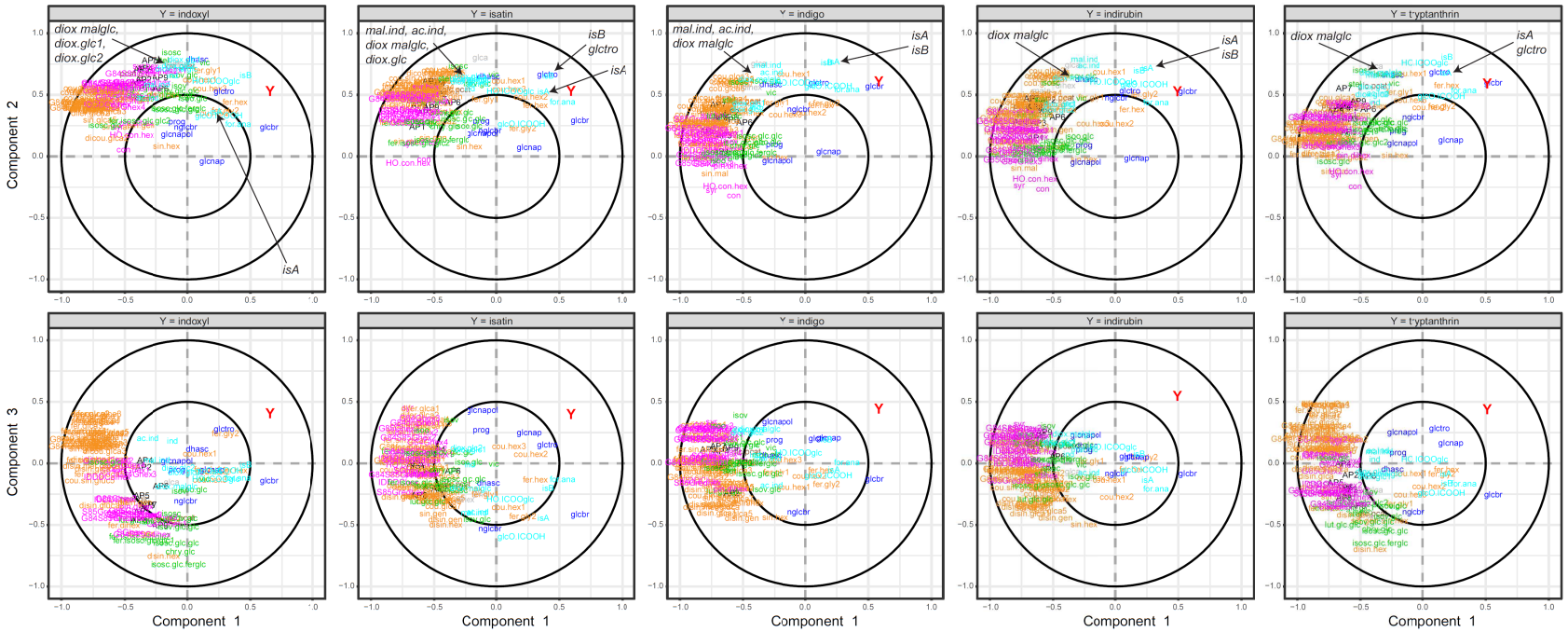


Figure 3. Correlation circles of PLS regression, illustrating the correlations of the variables with the first three PLS components.

Strongly positively correlated variables are projected in the same direction from the origin. Strongly negatively correlated variables are projected at diametrically opposite places on the correlation circle, whereas variables that are not correlated are situated 90° from each other in the circle. The longer the distance to the origin, the stronger the relationship between the variables. Only variables closely located to the circumference of radius 1 can be directly interpreted. For variables closely located to the origin, some information can be carried on other dimensions. Abbreviations: ac.ind, acetylindican; diox.glc, dioxindole glucoside; glcbr, glucobrassicin; glctro, glucotropaeolin; HO.ICOOglc, 6-hydroxyindole-3-carboxylic acid glucose ester; IsA, isatan A; IsB, isatan B; mal.ind, malonylindican. All abbreviations are given in Supplemental Table 1.



Indolics
Flavonoids
Phenylpropanoic acids
Apocarotenoids
Glucosinolates
Mono- and oligolignols
Benzenoids
Primary metabolites

Figure 4. Potential oxidative coupling mechanism between isatan B and isatan C leading to the formation of indigo

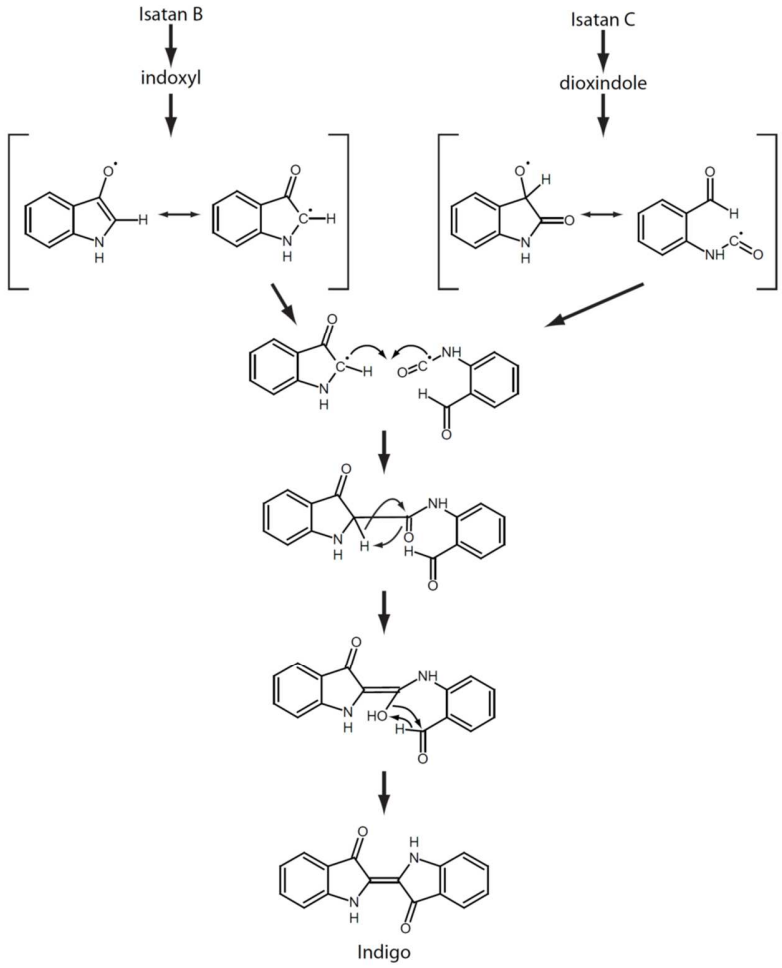


Table 1 Correlation matrix of five indole alkaloids in dried *I. tinctoria* rosette leaves (n=27)

	indoxyl	isatin	indigo	indirubin	tryptantrhin	CV
indoxyl	1					0.97
isatin	0.76	1				0.52
indigo	0.72	0.70	1			0.57
indirubin	0.63	0.54	0.84	1		0.86
tryptantrhin	0.75	0.63	0.81	0.84	1	0.89

CV, coefficient of variation

Table 2 Predictors, regression coefficients, and prediction accuracy of the prediction models for indoxyl, isatin, indigo, indirubin, and tryptanthrin

Feature ID	Compound	Class	indoxyl	Isatin	indigo	indirubin	tryptan-thrin
M408T1402	glucotropaeolin	Indole Glucosinolates		0.72			
M447.1T1654	glucobrassicin					0.61	
M352T1100	dihydroascorbigen	Other indole derivatives	0.27				
M378.1T1805	Isatan A		-0.36				
M292.1T1018	Isatan B		1.03		0.56		1.10
M396.1T1067	dioxindole malonylglucoside				0.38		
M338.1T377	6-GlcO-ICOOH			-0.43			
M593.2T1120	vicenin 2	Flavonoids					-0.54
M785.2T1298	isoscoparin-3''-O-glucoside-7-O-glucoside					-0.33	
M609.1T1115	luteolin-6-C-glucoside-7-O-glucoside						-0.51
M591.2T3456	disinapoyl hex	Phenylpropanoic acids	-0.51				
M385.1T1070	sinapoyl hex			-0.29			
M415.1T764	sinapoylglucaric acid						-0.37
M591.1T2821	feruloyl sinapoyl glucaric acid						0.58
M403.1T483	5-hydroxy-coniferyl alcohol hex	Mono- and oligolignols		-0.64			
M537.2T874	G(8–O–4)G hex			0.52			
M613.2T1520	S(8–O–4)G hex or G(8–O–4)S hex				-0.72		
Pearson correlation (predicted vs observed)			**0.81**	**0.82**	**0.82**	**0.47**	**0.82**
p-value			**3.6E-05**	**1.9E-04**	**3.4E-04**	**4.9E-03**	**1.3E-04**

Metabolites are ordered according to their metabolic class. The coefficient of the strongest positive predictor in each model is underlined. The accuracy of the models is expressed as Pearson correlation coefficient and p-values, as described in materials and methods. Feature ID indicates the LC-MS feature that has been integrated and the corresponding values in each sample are given in supplemental table 1. Shorthand naming of oligolignols is based on Morreel et al. (2010a, 2010b). Units derived from coniferyl alcohol and sinapyl alcohol are annotated as G (guaiacyl unit) and S (syringyl unit), respectively. 8–O–4 indicates a beta-aryl ether linkage. Abbreviations: hex, hexose (ester or ether); 6-GlcO-ICOOH, 6-hydroxyindole-3-carboxylic acid 6-O-glucoside.

Isatis tinctoria
rosette leaves

fresh leaf metabolites

linear prediction models

Drying
40 °C

indigo