Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2017, 51, 12528-12536

pubs.acs.org/est

Downloaded via UNIV OF SASKATCHEWAN on August 1, 2018 at 15:42:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Extended Virtual Screening Strategies To Link Antiandrogenic Activities and Detected Organic Contaminants in Soils

Jing Guo,[†,‡,§] Wei Shi,[*,†,‡,§] Qinchang Chen,[†,‡,§] Dongyang Deng,[†,‡,§] Xiaowei Zhang,[†,‡,§] Si Wei,[†,‡,§] Nanyang Yu,[†,‡,§] John P. Giesy,[†,∥,⊥,#] and Hongxia Yu[†,‡,§]

[†]State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, [‡]Jiangsu Environmental Monitoring Center, and [§]Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, Nanjing University, Nanjing, Jiangsu 210023, China

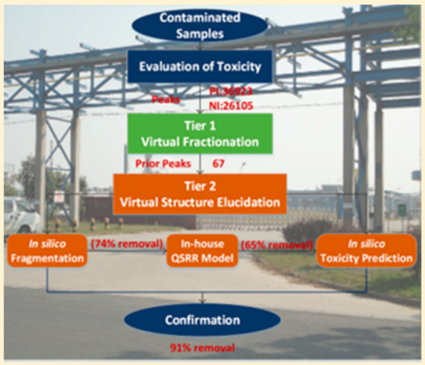
[∥]Department of Veterinary Biomedical Sciences and Toxicology Centre, University of Saskatchewan, Saskatoon, Saskatchewan S7N5B3, Canada

[⊥]Department of Zoology and Center for Integrative Toxicology, Michigan State University, East Lansing, Michigan 48824, United States

[#]School of Biological Sciences, University of Hong Kong, Hong Kong, SAR China

Supporting Information

ABSTRACT: A tiered screening strategy based on extensive virtual fractionation and elucidation was developed to simplify identification of toxicants in complex environments. In tier1-virtual fractionation, multivariate analysis (MVA) was set up as an alternative of physical fractionation. In tier2-virtual structure elucidation, in-house quantitative structure−retention relationship (QSRR) models and toxicity simulation methods were developed to simplify nontarget identification. The efficiency of the tiered virtual strategy was tentatively verified by soil samples from a chemical park contaminated by antiandrogenic substances. Eight out of 18 sites were detected as antiandrogenic, while none of them exhibited androgenic agonist potencies. Sixty-seven peaks were selected for further identification by MVA, among which over 90% were verified in androgenic fractions in traditional effect-directed analysis (EDA). With 579 tentative structures generated by in silico fragmentation, 74% were elucidated by QSRR and 65% were elucidated by in silico toxicity prediction. All prior peaks were identified at different confidence levels with over 40% of the identified peaks above confidence level 2b, which has been increased over 40% with less than half of the time spent compared to traditional EDA. Such a combination of tiered virtual screening methods provides more efficient and rapid identifications of key toxicants at contaminated sites.



INTRODUCTION

Due to rapid development of industries in the past few decades, large numbers and amounts of chemicals have been released into the environment and have accumulated, especially in sediments and soils. Therefore, hot spots with intensive contaminations emerged, including pesticide application areas, chemical industrial parks, and landfills.[1] Simultaneously, such hot spots were widely characterized to reveal estrogenic and androgenic effects,[2,3] which results in endocrine-disrupting effects on aquatic organisms and especially results in feminization of fishes.[4,5] However, key androgenic compounds contributing most to observed effects remain unknown. Therefore, identification and quantification of the androgenic compounds in these complex environments tend to be challenging

The current strategy to identify and quantify pollutants in a complex environment is focused primarily on priority pollutants, and unknown toxicants can be missed. Thus, an alternative method, effect-directed analysis (EDA)[6] that combines chemical and biological characterization has been established to fill the gap in cause−effect relationships between the compounds present and biological effects.[7] EDA has been widely used to identify key toxicants in various environmental matrices including waters,[8] sediments,[9−11] soils,[12] and drinking water.[13] Key processes in EDA usually include toxicity evaluation of environmental samples, stepwise fractionation of toxic samples, and structure elucidation of toxicants. However, several issues emerged in the key steps and limited its applications. Fractionation is a key step to reduce complexity and remove non- or less toxic fractions. However, if requested, multiple fractionations need to be performed, which is time-consuming and can result in poor recoveries. Structure elucidation is another key step to identify the individual

Received: June 30, 2017
Revised: September 26, 2017
Accepted: September 28, 2017
Published: September 28, 2017

© 2017 American Chemical Society

DOI: 10.1021/acs.est.7b03324

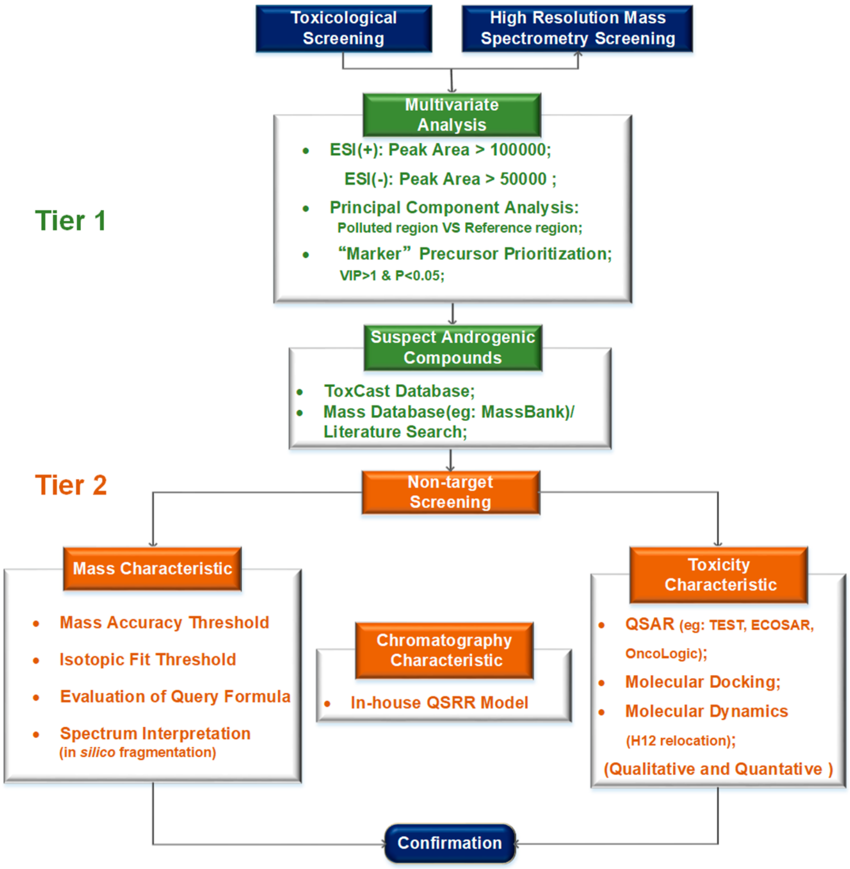

Toxicological Screening
High Resolution Mass Spectrometry Screening
Tier 1
Multivariate Analysis
ESI(+): Peak Area > 100000;
ESI(-): Peak Area > 50000 ;
Principal Component Analysis:
Polluted region VS Reference region;
"Marker" Precursor Prioritization;
VIP>1 & P<0.05;
Suspect Androgenic Compounds
ToxCast Database;
Mass Database(eg: MassBank)/ Literature Search;
Tier 2
Non-target Screening
Mass Characteristic
Mass Accuracy Threshold
Isotopic Fit Threshold
Evaluation of Query Formula
Spectrum Interpretation
(in silico fragmentation)
Chromatography Characteristic
In-house QSRR Model
Toxicity Characteristic
QSAR (eg: TEST, ECOSAR, OncoLogic);
Molecular Docking;
Molecular Dynamics
(H12 relocation);
(Qualitative and Quantative)
Confirmation

Figure 1. Workflow of the extensive virtual screening strategy. The virtual screening strategy contains two tiers. Tier 1 is virtual fractionation by use of multivariate analysis. Tier 2 is virtual elucidation of structures by combining suspect screening and nontarget screening. Note: VIP = variable importance in the project. QSRR = quantitative structure−retention relationship. QSAR = quantitative structure−activity relationship.

structures of potential toxicants from thousands of chemical features in toxic fractions. However, limited strategies for spectral interpretation and structure elucidation result in difficulty in identifying related formulas and structures, which makes the identification time-consuming, low-throughput, and with a tendency for false positives.

With increased resolution and sensitivity of instruments, virtual strategies have been[14,15] widely used to increase identification throughput, reduce cost, and increase confidence, especially in prediction of toxicity[16−18] and identification of biomarkers, which can be employed to improve the key steps in identification of individual toxicants in complex environmental samples. First, multivariate analysis (MVA), a method widely used to identify the most differential ones from thousands of variables, is recommended as an alternative of fractionation to reduce the complexity of the environmental samples and elute chemical features correlated with the observed effects.[19] Second, high-throughput, virtual-screening strategies including suspects and nontarget screening have been developed to simplify elucidation of structures. For screening of suspect toxicants, databases such as ToxCast and various lists of chemicals of various categories (pesticides[20] and pharmaceuticals[21]) provide various information, which enables rapid, reliable screening of a number of suspect compounds. For nontarget screening with limited prior information, some virtual tools also help simplify identification of putative, causative agents. Databases of MS/MS spectra (M^2 data), such as MassBank[22] and fragmentation prediction platforms including MetFrag[23,24] and MetFusion,[25] provide more rapid elucidation of structures relative to manual interpretation of fragment spectra. In-house retention time (RT) prediction models[26] provide additional information to predict retention time of individual structures, which enables high-throughput removal of false positive structures. What's more, in silicoprediction of toxic potency,[27,28] based on simulations of molecular dynamics (MD), provides a more efficient method for confirmation. Therefore, the tiered virtual steps mentioned above provide a new view of virtual screening in EDA, which likely results in more rapid, easy, and efficient identification of agents causing particular responses.

The objectives of the present study were to (i) develop a combined virtual screening strategy to simplify identification of toxicants in complex environmental mixtures; (ii) identify key antiandrogenic compounds in contaminated soils by applying the virtual screening strategy; and (iii) compare results obtained by use of the new virtual screening strategy with results of more traditional EDA methods then to further evaluate its advantages and limitations.

■ MATERIALS AND METHODS

Chemicals, Sampling, and Analysis. Seventy-five chemicals with wide range of physicochemical properties were made as artificial suspect mixtures for validation of the suspect screening approach (Table S1). Detailed information on these reference standards, reagents, and solvents used in this study are given in the Supporting Information (SI). Eighteen samples of soil were collected (Figure S1) in or around a chemical

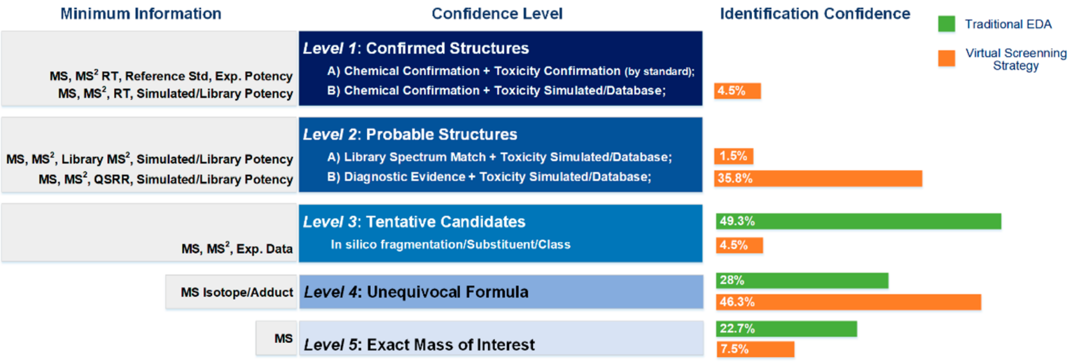


Figure 2. Comparison of optimized confidence levels, which combines structure confidence and toxicity confidence, of identifications by the virtual screening strategy and traditional EDA method. Note: exp potency is the effect potency obtained from biotests of purchased standards; exp data is the spectral information by chemical analysis.

industrial park in Jiangsu Province, China. Details of sample preparation and chemical analysis are described in the SI. Briefly, 20 g samples of soil were freeze-dried and extracted by use of accelerated solvent extraction (ASE) for further evaluation of androgenic potency and chemical characterization. Evaluations of androgenic potencies were applied by reporter gene assays, which are further described in the SI. Chemical analyses were conducted by use of a high-performance liquid chromatography (HPLC) system (Agilent 1260, Agilent, USA) coupled with a QTOF-MS system (TripleTOF 5600, AB Sciex, USA). Chemical profiles were collected in full-scan mode from m/z 100−1250.

New Virtual Screening Strategy. To further simplify identification of toxicants in complex environments, as an alternative of traditional EDA, a tiered, virtual screening strategy including virtual fractionation and virtual structure elucidation was developed, which combined progressive virtual methods for step-by-step elucidation of suspect structures (Figure 1). No more steps will be applied as long as the previous step was able to meet the demands of identification.

Tier 1: Virtual Fractionation. Virtual fractionation was conducted by MVA. Raw mass data were further processed by MS-DIAL[29] based on a data-independent MS/MS acquisition method for peak alignment, picking, and deconvolution. After removing missing values in each sample class by use of the "80% Rule",[30] sample information and normalized peak areas of individual ions with spectral region of blank water and methanol excluded, were introduced to software SIMCA-P 13.0.3 (Umetrics, Umea, Sweden). Principle-component analysis (PCA) and Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) were conducted to visualize differences between various locations and identify the most significant chemicals which were sufficiently unique to discriminate more contaminated regions from reference regions. All variables were mean-centered by sample and pareto-scaled prior to modeling. Variable importance in the projection (VIP) was calculated for each variable to show the contribution in classification, and those variables with VIP > 1.0 were considered to be most relevant for classification.[31] Also, significant differences between classes of samples were identified by use of unpaired Student's t tests for which the P value threshold was set to 0.05 for statistical significance. Therefore, variables with VIP greater than 1.0 and P values less than 0.05 were selected as potential marker precursors for further identification.

Suspect Screening. Prior precursors were further identified by comparison to a database containing 3472 suspected, antiandrogenic compounds (Table S2), which were collected based on physicochemical and toxicological information, especially the extract mass, provided by ToxCast. After the mass accuracy threshold was limited to 5 ppm, a signal-to-noise ratio (S/N) greater than 10 and isotopic differences less than 20%, query precursors were identified by XIC Manager (PeakView, Foster City, CA). Precursors with clear MS/MS spectra, containing at least two product ions with intensities greater than 100, were selected for further evaluation. Query structures were further evaluated by matching at least two product ions and relative peak intensity with spectral information provided by databases of spectra, including MassBank, Spectral Database for Organic Compounds (SDBS), and published spectral information. Furthermore, those structures for which no reference spectral information was available were further elucidated by evaluation of fragmentation of suspect structures by use of the Fragment Pane (PeakView). After evaluation, those structures with diagnostic confidence needed further confirmation by comparison to reference standards including comparisons of RT and MS/MS spectra. Validation of suspect screening was performed using 75 standards of suspect chemicals at a concentration of 20 μg/L in a blank matrix. Chemicals with formulas detected were further confirmed by comparison of MS/MS spectra with standards. Rates of detection were further calculated on the basis of confirmed structures.

Tier 2: Virtual Structure Elucidation To Simplify Nontarget Screening. Additional precursors were identified by use of a nontarget strategy of identification. The mass accuracy threshold was set as 2mDa/5 ppm, S/N was greater than 10, and isotope difference was less than 20%. Numbers of elements were further eliminated as $C_{50}H_{200}N_{10}O_{10}P_5S_5Cl_{10}Br_{10}F_{10}$. Formulas were further calculated by use of Formula Finder (PeakView) by linking to the chemicals database (ChemSpider). Peaks with clear MS/MS spectra, which contained more than two product ions with intensities greater than 100, were selected for further elucidation of structures. Experimental MS/MS spectra were first compared with spectra of standards in databases such as MassBank,[22] SDBS, and published spectral information. For the other precursors, structures were generated by in silico fragment simulation platforms including MetFusion[25] and MetFrag.[23] After inputting the accurate mass, adduct ions, chemical database, and spectral information, structures with scores less than 0.7 were eliminated. With acceptable false negative and lower false positive, query structures were further elucidated by limiting the retention time predicted by the in-house QSRR model with relative error less than 20%. Probable structures were predicted to be androgenic or not by use of in silicoMD simulation. Structures with H12 relocation stably during simulations were further confirmed with reference standards.

In-House QSRR Model. Forty-four chemicals (Table S3) with a wide range of physicochemical properties were used to develop the QSRR model. Retention times were predicted by the multifactor linear regression based on the most relevant molecular descriptors (Table S4) selected by MVA. Details of the development of the model and the application domain are described extensively in the SI.

In Silico Prediction of Toxic Potency. Prediction of toxic potency, based on previously described MD simulations,[32] was used for elucidation of androgenic structures. Briefly, Chemoffice was used to build and optimize 3D structures for both ligands and receptors. Suspect structures were bound to active positions of androgen receptor ligand binding domain (AR-LBD) by use of the Surfle-Dock model in Software Sybyl 7.3 (Tripos Inc., St. Louis, MO). MD modeling data were also processed by use of Gromacs 4.5.[32−34] Query structures were classified as androgen-active or not by monitoring whether H12 relocation was stable within 20 ns. Details for structure preparation, docking, and MD simulations are described extensively in the SI.

Communicating Confidence. Since it was impossible to identify and confirm all structures by use of reference standards, communicating confidence (Figure 2) combining structure confidence and toxicity confidence, which had been developed based on the previous study,[35] was applied. Level 1a represents structures confirmed by comparisons to reference standards with measurement of MS, RT, MS/MS, and effect potency by biotests. Level 1b represents structures with chemical confirmation as Level 1a as well as prediction of toxic potency by MD simulation or toxicants databases. Level 2a represents probable structures confirmed by use of spectral information provided by databases of spectra, published literature, as well as prediction of toxic potency by MD simulation or toxicants databases. Level 2b represents probable structures with diagnostic chemical confidence, including in silico fragmentation combined with RT prediction by QSRR as well as toxicity prediction by MD simulation or toxicants databases. Level 3 represents possible structures assessed by simple interpretation with less confidence, such as structures assessed only by in silicofragmentation. Level 4 represents unequivocal formulas with no structures available, including precursors with bad MS/MS spectra. Level 5 represents the exact mass of interest while lacking information to assign even formulas.

Verification by Use of the Classical EDA Method. The efficiency of the new method was verified by the traditional EDA method that combines physical fractionation, chemical characterization, and biotests. Consequently, the strategy included the following:

Fractionation. Prefractionation was performed on Oasis HLB SPE columns. Raw extracts were fractionated into four fractions by eluting sequentially with solvents: methanol, mixed solvent of methanol, and dichloromethane (DCM) (v/v, 1/1), mixed solvent DCM and *N*-hexane (Hex) (v/v, 4/1), and Hex, respectively. Preparative fractionation of androgenic prefractions was performed on reversed-phase (RP) semipreparative HPLC. The collection time of each fraction was determined on the basis of specific peaks and water content, which is shown in detail (Table S5). Fractionation as well as identification and quantification were evaluated by use of nine typical androgenic substances and nine antiandrogenic substances (Table S6), which were added at the level in environmental concentration at 5 μg/L.

Identification and Confirmation. Peaks with intensity greater than 1250/500 (in positive-/negative-ion mode) and signal-to-noise (S/N) greater than 10 were selected for further identification. Formulas were calculated after limiting the isotopic difference to less than 20%, numbers of elements, and comparison with ChemSpider by Formula Finder in PeakView. Structures of precursors with clear MS/MS spectra and at least two product ions were further generated by in silico fragmentation conducted on MetFrag. Query structures with MetFrag scores less than 0.7 were eliminated. With no more elucidation strategy available, query structures were finally confirmed by comparing RT and MS/MS spectra with reference standards under the same conditions of chemical analysis and evaluating AR antagonist potencies by a series of biotests.

Data Processing. Antiandrogenic equivalents (anti-AR EQ) were calculated as concentration of flutamide divided by the dilution factors of individual samples that exhibited 20% inhibition of the response, which was selected to avoid underestimates or overestimates,[13] to 1.0 × 10^{-9} M DHT. Consequently, anti-AR EQ of samples were calculated (eq 1). All exposures were conducted in triplicate. One-way ANOVA followed by Dunnett's multiple comparisons tests were used to evaluate acquired data, with *P* values less than 0.05 considered significant. Dose−response curves were fitted in nonlinear regressions by use of Graphpad 5.4 (San Diego, CA)

$$\text{anti-AREQ}_{\text{s}} = \frac{\text{concentration of known flutamide}}{\text{enrichment ratio of tested samples}} \tag{1}$$

where daily intake of androgenic equivalent (DIAR EQ) was estimated (eq 2)

$$\text{DIAREQ}_{\text{S}} = \text{anti-AREQ}_{\text{s}} \times A \tag{2}$$

where DIAR EQs is the DIAR EQ of individual sampling sites, anti-AR EQs is the anti-AR equivalent of the sampling site, and *A* is a constant which represents 200 mg soil/d inhaled estimated by EPA.[36]

RESULTS AND DISCUSSION

Evaluation of the Virtual Screening Approach. Fractionation, identification, and quantification were evaluated by use of 18 model androgenic chemicals spiked at 5 μg/L. Procedural recoveries ranged from 80 to 107% (Table S7). None of these substances were found in procedural or solvent blanks, except for dibutyl phthalate and butyl benzyl phthalate. Reliability and sensitivity of biotests were also evaluated. The dose−response curves of known AR agonist and antagonist ligand, DHT and flutamide, obtained from the reporter gene assay are shown (Figure S2). No AR agonist or AR antagonist potency was detectable in the blank.

The strategy for screening suspect chemicals was further validated by use of 75 reference chemicals, and a mixture of artificial suspects was added to the blank each at a concentration of 20 μg/L by use of XIC Manager (PeakView). With accurate mass ([M + H] in positive, [M − H] in negative) and calculated formulas the only information available, 65 precursors were found in the positive mode and 34 were found in the negative mode. These were further confirmed by use of MS/MS spectra. During the entire suspect screening strategy, 93% of the 75 suspect chemicals were detected.

The QSRR model, developed in-house, was also evaluated. On the basis of the PCA score plot (Figure S3), the training set and the test set were scattered evenly in the 2D plot, which shows that there was no significant difference between training and test sets. log Kow and log D (pH = 6), were selected as the most significant descriptors evaluated by VIP (Table S4) acquired after partial least-squares (PLS) regression. After multifactor regression with 34 training chemicals, the retention time was calculated (eq 3)

$$\mathrm{RT} = 4.147 \log \mathrm{Kow} + 2.81 \log D\ (\mathrm{pH} = 6) + 6.355 \tag{3}$$

where predicted RT and the experimental RT showed a significant, linear relationship with r^2 = 0.9585, which further demonstrated good efficiency of the in-house QSRR model (Figure S4).

Identification by Use of Tiered Virtual Screening Strategy. *Evaluation of AR Antagonist Potency.* Potencies of androgen antagonist were detectable in 8 of 18 sites, while none of the 18 samples exhibited AR agonist potencies. AR antagonist equivalents (AntEQ) of samples (Table S8) ranged from 23.23 μg of flutamide (FLU)/g soil to 143.90 μg of FLU/g soil. Samples with detectable AR antagonist potency were all located in the downwind region, while the rest of the samples were mainly located around the edge. Therefore, locations S4, S6, S8, S9, S11, S16, S17, and S18 were considered to be in the more contaminated region, while other locations (S1, S2, S3, S5, S7, S10, S12, S13, S14, and S15) were considered to be less contaminated (Figure S1). Although little information is available for concentrations of AR antagonists in soils, the results observed during this study were similar to previously reported results[37] where no agonist potency was detected and the AR AntEQ varied from not detected (N.D.) to 178.05 FLU μg/g in soils collected along the Songhua River. When compared to the proposed rate of inhalation of 200 mg/d soil suggested by the US EPA to represent that inhaled by children, AR AntEQ of these samples of soils ranged from 4.65 to 28.78 μg FLU/g/d with two sampling sites exceeding the threshold of 24.87 μg FLU/g/d, suggested by US EPA for protection of people living nearby. Thus, so that they could be controlled, key androgenic substances which contributed most of the AR antagonist potencies, needed to be further identified.

Tier 1: Virtual Fractionation by Multivariate Analysis. After applying toxicity evaluation and high-resolution mass spectrometry screening, on average, 70221 (52515−96288) peaks in positive-ion mode (PI) and 64845 (31948−95282) peaks in negative-ion mode (NI) were observed in soils from the more contaminated region, while 40798 (26757−45916) peaks (PI) and 22908 (17284−27044) peaks (NI) were observed in soils from the less contaminated region by MS DIAL defined by retention time and exact mass. A total of 36923 chemical features (PI) and 26105 chemical features (NI) were further extracted from each soil sample for further multivariate analysis. Peaks were mainly found to have retention times between 20 and 25 min and between 25 and 30 min. (Figure S5), with related log Kow predicted by use of the in-house QSRR ranging from 2.16 to 3.78, which indicated that these compounds are mostly medium hydrophilic.

To determine whether compounds were sufficiently unique to distinguish the more contaminated region from the reference area, PCA score plots were applied on the basis of the 36923 chemical features (PI) and 26105 chemical features (NI) extracted. On the basis of the results of quantile−quantile plots of the first two principle components of the PCA, no sample was identified as being a potential outlier (Figure 3A,B).

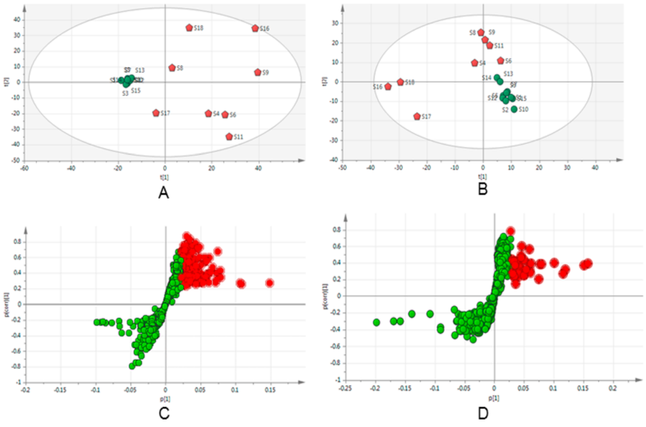


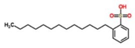
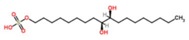
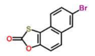
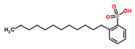
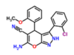
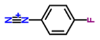
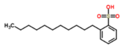
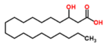
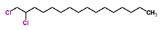
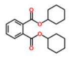
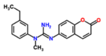
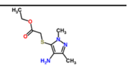
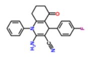
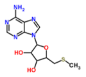
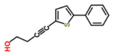
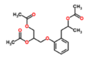
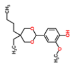
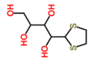
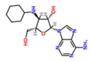
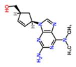
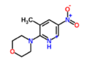
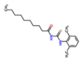
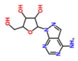
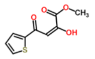
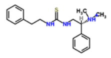
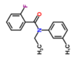
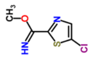
Figure 3. PCA score and S-plot in positive- and negative-ion mode. The dots in the plot all corresponded to the features in the raw LC−MS data. (A, B) PCA score plot of 18 soil samples in positive and negative mode, respectively. Red star dots were samples with AR antagonist potency, while the green dots were noneffective. (C, D) S-plot of 18 soil samples in positive and negative mode, respectively. Potential markers with VIP > 1.0 and P < 0.05 were labeled in red on plots.

Samples in the more contaminated region were separated from samples in the reference region, which is consistent with results of bioassay of AR antagonist for these samples. Also, the PCA score plot also showed that there was no significant difference among the 10 samples in the reference region, which indicates the analytical method was stable. The result also showed that chemical features were quite different between the more contaminated region and the reference region which might be important to explain AR antagonist potencies exhibited by extracts of soils in the more contaminated regions.

OPLS-DA were conducted to reveal and explain differences between the more contaminated region and the reference region. The OPLS-DA score plot (Figure S6A,B) showed that samples from the two regions were separated, which is consistent with results of the PCA score plot. Validations of the OPLS models were evaluated by use of the permutation test with 100 iterations (Figure S7). Variables that were at the edges (VIP > 1 and P < 0.05) of the S-plot (Figure 3C,D) were likely to be the most significant markers relevant to the classification, which efficiently reduced the risk of determining false positives. As a result, 67 peaks (52 peaks in PI/15 peaks in NI) (Table S9) responsible for the separation were considered as potential markers to explain the observed effect, which needed further identification. These results were consistent with those of a previous study, where 78 prior peaks were selected among 3391 chemical features to explain the most significant difference between the oil polluted region and the reference region.[38]

Suspect Screening of Prior Precursors. After multivariate analysis was applied, 67 selected prior precursors were identified by rapid screening of suspect antiandrogenic compounds. Screening the 67 prior precursors in the database

Table 1. Diagnostic Structures with Identification Confidence of Level 2b or Greater

Ion Mode	Precursor m/z	Confidence Level	Structure	Precursor m/z	Confidence Level	Structure	Precursor m/z	Confidence Level	Structure
Negative	339.2003	1b		381.231	2b		278.9142	2b	
	325.1831	1b		377.0839	2b		121.0312	2b	
	311.1687	1b		327.2904	2b		293.1765	2b	
Positive	331.1895	2a		322.1574	2b		230.0952	2b	
	360.149	2b		298.0966	2b		229.0671	2b	
	353.157	2b		293.1712	2b		227.0394	2b	
	349.1968	2b		275.161	2b		225.1081	2b	
	335.2168	2b		268.1032	2b		213.0235	2b	
	329.192	2b		258.1269	2b		176.9885	2b	

of 3472 AR antagonist suspects (Table S2), which was collected from the ToxCast database, yielded 6 hits in PI and 7 hit in NI, when the mass accuracy threshold was set to 5 ppm and isotopic difference less than 20%. By further evaluation by use of the S/N (threshold of 10) and MS/MS spectra (intensity of product ions threshold >100), seven precursors were selected for further identification. After all the steps and comparison with spectral information in published papers and databases such as MassBank and SDBS, 1 suspect in PI and 1 suspect in NI were tentatively identified. The suspect identified in PI, dicyclohexyl phthalate, was assigned an accuracy level of 2a with two main product ions 231 and 249 matching well with the SDBS spectrum (Figure S8). The other identified suspect in NI, dodecylbenzenesulfonic acid, matched well with spectral information in a published paper. Since 15 of the 17 product ions of dodecylbenzenesulfonic acid were interpreted well by in silicofragmentation (Figure S10), the suspect was further confirmed by comparison of the RT and the MS/MS spectrum of a reference standard and was then assigned confidence of identification of 1b. For the other five precursors which failed to match spectra in either databases or published papers, were treated as unidentified targets to be identified by use of nontarget screening.

Tier 2: Virtual Structure Elucidation To Simplify Nontarget Screening. Nontarget screening was conducted for the 65 other prior peaks, which had remained unidentified after applying suspect screening. Since it was not possible to identify each individual precursor by comparisons to MS/MS spectra of reference standards, precursors were identified and confirmed at various levels of confidence (Figure 2). After the mass accuracy threshold, isotopic difference, and elements numbers were limited, five precursors failed to be assigned formulas by Formula Finder, which thus remained at level 5 confidence. Two hundred fifty-six formulas were calculated for another 60 prior precursors by use of Formula Finder then further filtered to 89 formulas by use of their isotopic patterns and limiting relative error of MS/MS spectra when compared to spectra in ChemSpider to less than 10 ppm. Since ionization is not always easy for some precursors and acquisition rate of MS/MS spectra was limited to 1−20, identification of 31 precursors lacking MS/MS spectra remained at level 4 confidence, with only unequivocal formulas. For the other 29 precursors for which spectral information was available, 579 structures (467 in positive mode and 112 in negative mode) were generated after in silico fragmentation conducted by use of MetFrag with a score threshold of 0.7. These query structures were further elucidated to 151 structures (112 in positive mode and 38 in negative mode) after prediction of the RT by the in-house QSRR model with the relative error threshold set at 20%. The rate of removal of false positive structures by QSRR was 74% (76% in PI and 66% in NI). Probable structures were further elucidated via toxicity prediction by MD simulation, with stable H12 relocation in 20 ns during the process (Figure S9A,B). Sixty-five percent more false positive structures with no AR effect potencies were eliminated after prediction of toxic potencies. Thus, these 53 structures related to 26 prior precursors predicted to be androgenic by MD simulation were assigned a confidence level of 2b (Table 1), while the other three precursors were identified and assigned at confidence level of 3. Detailed identification process by use of the nontarget screening strategy are given (Table S10). Overall, 40% of the prior precursors were identified and confirmed over a confidence level of 2b (Figure S9).

As an example of nontarget screening, identification of alkyl benzenesulfonic acids is described. Three alkyl benzenesulfonic acids which lie in the top right corner of the S-plot (Figure 3D) were identified in negative mode. VIPs of these three markers were 1.78, 1.26, and 1.34, respectively, and statistical differences for these three markers between the AR antagonist class and noneffective class were significant. The result showed that intensities of these three chemicals in the more contaminated region were significantly greater than they were in the reference region. After limiting mass accuracy, isotopic fitness, and S/N ratio, these three precursors were found to be homologues with molecular mass to charge ratios (m/z) of 311.1687, 325.1831, and 339.2003, corresponding to formulas $C_{17}H_{28}O_3S$, $C_{18}H_{30}O_3S$, and $C_{19}H_{32}O_3S$. MS/MS spectra of these three chemicals were similar, with 17 product ions in common (Figure S10). By searching ChemSpider and PubChem via MetFusion, three to six structures were retrieved but only two candidates, which are isomers, for each formula were able to explain all of the fragments with the greatest MetFusion and MetFrag score. Among the tentatively identified toxicants, both series of homologues were predicted by in silico MD simulations to have a binding affinity to the AR, with settling time of H12 relocation less than 20 ns. As a result, tentative structures of undecylbenzenesulfonic acid, dodecylbenzenesulfonic acid, and tridecylbenzenesulfonic acid were further confirmed by use of authentic standards (Figure S10).

Verification by Use of Traditional EDA Methods. Identification of antiandrogenic compounds by the novel, virtual strategy was compared to results of traditional EDA methods for five extracts of soils from the same contaminated region, shown to have antiandrogenic potencies. After prefractionation, six of 20 prefractions exhibited AR antagonist potencies (Figure S11A), which contributed from 76 to 101% of the AR antagonist potency of extracts of soils determined by use of bioassays. More polar chemicals and moderately polar chemicals in fractions 0 and 1 contributed most to the AR antagonist potency. This observation is consistent with results of a previous study where medium polar fractions and polar fractions explained 84 to 110% AR antagonist potencies of the raw extracts of soils.[37] Prefractions exhibiting androgenic potencies were further fractionated into 51 preparative fractions by use of preparative LC, with preparative fractions ranked in top 10 accounted for 88% to 119% of AR antagonist potencies of related androgenic prefractions (Figure S11B). On the basis of targeted screening of 14 known antiandrogenic compounds (Table S11, selecting strategy are described extensively in the SI), only nonylphenol (NP), octylphenol (OP), and bisphenol A (BPA) were observed in extracts of soils. Further, nontargeted identification revealed 3872 peaks in PI and 1042 peaks in NI by information dependent acquisition (IDA) for compounds in fractions exhibiting antiandrogenic potency, which was accomplished by limiting the intensity threshold (1250, PI and 500, NI), mass accuracy threshold and S/N ratio. Since nearly 5000 peaks were found, which was deemed to be too many to be further identified, only 75 peaks (56 in positive mode, 19 in negative mode), which exhibited the greatest intensities and observable peaks, were selected for further identification (Table S12). After isotopic distributions and numbers of elements were limited, 92 formulas were calculated by use of Formula Finder. Furthermore, after the score threshold was limited to 0.7, 214 query structures were generated by in silico fragmentation conducted by use of MetFrag. Since no more additional information on predicted retention times or toxic potency was available, these query structures were determined with a confidence level of 3, which would need to be confirmed with authentic standards.

To further compare the results of the novel, virtual screening strategy with the traditional EDA method, of the 67 prior markers eluted by MVA in the extensive virtual screening strategy, more than 97% were found with high intensity in the 6 androgenic prefractions (Table S13). More than 73% of the prior precursors were simultaneously found in more than three of the six prefractions, which demonstrated the significance of selecting these precursors by use of the MVA strategy. More than 88% of the prior precursors were also found in the top five preparative fractions, which exhibited the greatest AR potencies. Since more than 90% of the precursors in the virtual fractionation were verified by traditional EDA method, it was concluded that the two methods gave similar results. However, selecting compounds by use of MVA was more rapid and efficient. In the traditional EDA method, since no additional elucidation strategies are available to remove false positive structures, for over 90% of the peaks identified, confidence in their identifications remained at level 3 or worse. While in the tiered virtual screening strategies, QSRR and MD simulation, efficiently removed false positive structures to a great extent with removal rates of 74% and 65%, which efficiently increased confidence of identification with over 40% being of level 2b or greater. Thus, the proportions of structures identified at confidence of level 2b or greater, were 40% greater by using the virtual screening strategy with less than half of the time spent compared to traditional EDA methods.

The virtual screening strategies efficiently simplified key steps in the traditional EDA process and confidence in identification was greater. First, selecting peaks by virtual fractionation is a more rapid, less labor intensive, and efficient strategy. More than 97% of the prior precursors selected by MVA were detected with greater intensity in androgenic fractions during the EDA. Second, combined virtual screening strategies, including QSRR and MD simulation, efficiently removed false positive structures to a great extent with rates of 74% and 65%. Thus, time and cost of confirmation of false positive query structures were reduced, which also efficiently simplified nontarget identification. Last, the virtual screening strategy efficiently increased confidence of identification, with over 40% of precursors identified and confirmed with confidence levels equal to or greater than 2b (Figure 2). The proportion achieving this level of confidence was increased more than 40% compared to the traditional EDA method. Although these virtual steps were applied to reduce uncertainties that occur in traditional screening processes, such a combination of virtual methods also generated some uncertainties, including false positives of QSRR and potential bias of prediction of toxicity, while the uncertainties could be minimized with enough validations and combination of more prediction methods for multiple validations. Therefore, the new virtual strategy is a good alternative for the classical EDA method to identify key toxicants in intensively polluted regions rapidly and efficiently.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.7b03324.

Supporting materials and addition information for this text, including Tables S1−S9 and S11−S13 and Figures S1−S11 (PDF)

Detailed identification process of non-target screening (Table S10) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: njushiwei@nju.edu.cn.

ORCID

Wei Shi: 0000-0001-9499-818X

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (21577058 and 21307054), the Science Fund for Excellent Young Scholars of Jiangsu Province (BK20170077), and the National Water Pollution Control and Treatment Science and Technology Major Project (2017ZX07602-002). The computational calculations were performed on the IBM Blade cluster system in the High Performance Computing Center (HPCC) of Nanjing University. Prof. Giesy was supported by the program of 2012 “High Level Foreign Experts” (No. GDT20143200016) funded by the State Administration of Foreign Experts Affairs, P.R. China to Nanjing University, and the Einstein Professor Program of the Chinese Academy of Sciences. He was also supported by the Canada Research Chair program and a Distinguished Visiting Professorship in the School of Biological Sciences of the University of Hong Kong.

■ REFERENCES

(1) Coors, A.; Jones, P. D.; Giesy, J. P.; Ratte, H. T. Removal of estrogenic activity from municipal waste landfill leachate assessed with a bioassay based on reporter gene expression. *Environ. Sci. Technol.* **2003**, *37* (15), 3430−3434.

(2) Thomas, K. V.; Hurst, M. R.; Matthiessen, P.; McHugh, M.; Smith, A.; Waldock, M. J. An assessment of in vitro androgenic activity and the identification of environmental androgens in United Kingdom estuaries. *Environ. Toxicol. Chem.* **2002**, *21* (7), 1456−1461.

(3) Dizer, H.; Fischer, B.; Sepulveda, I.; Loffredo, E.; Senesi, N.; Santana, F.; Hansen, P. D. Estrogenic effect of leachates and soil extracts from lysimeters spiked with sewage sludge and reference endocrine disrupters. *Environ. Toxicol.* **2002**, *17* (2), 105−112.

(4) Jobling, S.; Nolan, M.; Tyler, C. R.; Brighty, G.; Sumpter, J. P. Widespread sexual disruption in wild fish. *Environ. Sci. Technol.* **1998**, *32* (17), 2498−2506.

(5) Kabir, E. R.; Rahman, M. S.; Rahman, I. A review on endocrine disruptors and their possible impacts on human health. *Environ. Toxicol. Pharmacol.* **2015**, *40* (1), 241−258.

(6) Brack, W. Effect-directed analysis: a promising tool for the identification of organic toxicants in complex mixtures? *Anal. Bioanal. Chem.* **2003**, *377* (3), 397−407.

(7) Brack, W.; Ait-Aissa, S.; Burgess, R. M.; Busch, W.; Creusot, N.; Di Paolo, C.; Escher, B. I.; Hewitt, L. M.; Hilscherova, K.; Hollender, J. Effect-directed analysis supporting monitoring of aquatic environments—An in-depth overview. *Sci. Total Environ.* **2016**, *544*, 1073−1118.

(8) Grung, M.; Lichtenthaler, R.; Ahel, M.; Tollefsen, K.-E.; Langford, K.; Thomas, K. V. Effects-directed analysis of organic toxicants in wastewater effluent from Zagreb, Croatia. *Chemosphere* **2007**, *67* (1), 108−120.

(9) Schwab, K.; Altenburger, R.; Varel, U. L. v.; Streck, G.; Brack, W. Effect-directed analysis of sediment-associated algal toxicants at selected hot spots in the River Elbe basin with a special focus on bioaccessibility. *Environ. Toxicol. Chem.* **2009**, *28* (7), 1506−1517.

(10) Qi, H.; Li, H.; Wei, Y.; Mehler, W. T.; Zeng, E. Y.; You, J. Effect-Directed Analysis of Toxicants in Sediment with Combined Passive Dosing and in Vivo Toxicity Testing. *Environ. Sci. Technol.* **2017**, *51* (11), 6414−6421.

(11) Hecker, M.; Giesy, J. P. Effect-directed analysis of Ah-receptor mediated toxicants, mutagens, and endocrine disruptors in sediments and biota. In *Effect-directed analysis of complex environmental contamination*;Springer, 2011; pp 285−313

(12) Legler, J.; van Velzen, M.; Cenijn, P. H.; Houtman, C. J.; Lamoree, M. H.; Wegener, J. W. Effect-directed analysis of municipal landfill soil reveals novel developmental toxicants in the zebrafish Danio rerio. *Environ. Sci. Technol.* **2011**, *45* (19), 8552−8558.

(13) Shi, W.; Hu, X.; Zhang, F.; Hu, G.; Hao, Y.; Zhang, X.; Liu, H.; Wei, S.; Wang, X.; Giesy, J. P. Occurrence of thyroid hormone activities in drinking water from eastern China: contributions of phthalate esters. *Environ. Sci. Technol.* **2012**, *46* (3), 1811−1818.

(14) Eide, I.; Neverdal, G.; Thorvaldsen, B.; Grung, B.; Kvalheim, O. M. Toxicological evaluation of complex mixtures by pattern recognition: correlating chemical fingerprints to mutagenicity. *Environ. Health Perspect.* **2002**, *110* (Suppl 6), 985.

(15) Nash, M. S.; Chaloud, D. J. Partial least square analyses of landscape and surface water biota associations in the Savannah River Basin. *ISRN Ecol.* **2011**, *2011*, 1−11.

(16) Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. Computational ecotoxicology: Simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environ. Int.* **2014**, *73*, 288−294.

(17) Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environ. Sci. Technol.* **2014**, *48* (24), 14686−14694.

(18) Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **2014**, *6* (18), 10623−10630.

(19) Hug, C.; Sievers, M.; Ottermanns, R.; Hollert, H.; Brack, W.; Krauss, M. Linking mutagenic activity to micropollutant concentrations in wastewater samples by partial least square regression and subsequent identification of variables. *Chemosphere* **2015**, *138*, 176−182.

(20) Moschet, C.; Piazzoli, A.; Singer, H.; Hollender, J. Alleviating the reference standard dilemma using a systematic exact mass suspect screening approach with liquid chromatography-high resolution mass spectrometry. *Anal. Chem.* **2013**, *85* (21), 10312−10320.

(21) Vergeynst, L.; Van Langenhove, H.; Joos, P.; Demeestere, K. Suspect screening and target quantification of multi-class pharmaceuticals in surface water based on large-volume injection liquid chromatography and time-of-flight mass spectrometry. *Anal. Bioanal. Chem.* **2014**, *406* (11), 2533−2547.

(22) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45* (7), 703−714.

(23) Wolf, S.; Neumann, S. MetFrag—Match Predicted Fragments with Mass Spectra. *German Conference on Bioinformatics*; Citeseer, 2009; p 120.

(24) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.* **2016**, *8* (1), 1.

(25) Gerlich, M.; Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **2013**, *48* (3), 291−298.

(26) Aalizadeh, R.; Thomaidis, N. S.; Bletsou, A. A.; Gago-Ferrero, P. Quantitative Structure Retention Relationship Models To Support Nontarget High-Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples. *J. Chem. Inf. Model.* **2016**, *56* (7), 1384−1398.

(27) Wang, X.; Yang, H.; Hu, X.; Zhang, X.; Zhang, Q.; Jiang, H.; Shi, W.; Yu, H. Effects of HO-/MeO-PBDEs on androgen receptor: in vitro investigation and helix 12-involved MD simulation. *Environ. Sci. Technol.* **2013**, *47* (20), 11802−11809.

(28) Wang, X.; Zhang, X.; Xia, P.; Zhang, J.; Wang, Y.; Zhang, R.; Giesy, J. P.; Shi, W.; Yu, H. A high-throughput, computational system to predict if environmental contaminants can bind to human nuclear receptors. *Sci. Total Environ.* **2017**, *576*, 609−616.

(29) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12* (6), 523−526.

(30) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J.; Jellema, R. H. Fusion of mass spectrometry-based metabolomics data. *Anal. Chem.* **2005**, *77* (20), 6729−6736.

(31) Jansson, J.; Willing, B.; Lucio, M.; Fekete, A.; Dicksved, J.; Halfvarson, J.; Tysk, C.; Schmitt-Kopplin, P. Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* **2009**, *4* (7), e6386.

(32) Chen, Q.; Wang, X.; Shi, W.; Yu, H.; Zhang, X.; Giesy, J. P. Identification of Thyroid Hormone Disruptors among HO-PBDEs: In Vitro Investigations and Coregulator Involved Simulations. *Environ. Sci. Technol.* **2016**, *50* (22), 12429−12438.

(33) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435−447.

(34) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **1995**, *91* (1), 43−56.

(35) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097−2098.

(36) EPA, U. *Exposure factors handbook 2011 ed. (Final)*; Washington, DC, 2011.

(37) Li, J.; Wang, Y.; Kong, D.; Wang, J.; Teng, Y.; Li, N. Evaluation and characterization of anti-estrogenic and anti-androgenic activities in soil samples along the Second Songhua River, China. *Environ. Monit. Assess.* **2015**, *187* (11), 1−10.

(38) Wang, B.; Wan, Y.; Zheng, G.; Hu, J. Evaluating a tap water contamination incident attributed to oil contamination by nontargeted screening strategies. *Environ. Sci. Technol.* **2016**, *50* (6), 2956−2963.

Extended virtual screening strategies to link anti-androgenic activities and detected organic contaminants in soils

Jing Guo[†,‡,§], Wei Shi[†,‡,§,]*, Qinchang Chen[†,‡,§], Dongyang Deng[†,‡,§], Xiaowei Zhang[†,‡,§], Si Wei[†,‡,§], Nanyang Yu[†,‡,§], John P. Giesy [†,‖,⊥, #], Hongxia Yu[†,‡,§]

[†] State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, Jiangsu 210023, China

[‡] Jiangsu Environmental Monitoring Center, Nanjing University, Nanjing, Jiangsu 210023, China

[§] Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, Nanjing University, Nanjing 210023, China

[‖] Department of Veterinary Biomedical Sciences and Toxicology Centre, University of Saskatchewan, Saskatoon, Saskatchewan S7N5B3, Canada

[⊥] Department of Zoology and Center for Integrative Toxicology, Michigan State University, East Lansing, Michigan 48824, United States

[#] School of Biological Sciences, University of Hong Kong, Hong Kong, SAR China

[*]Corresponding author: Dr. Wei Shi, School of the Environment, Nanjing University, Nanjing, Jiangsu 210023, China

E-mail: njushiwei@nju.edu.cn

Supporting Information

This file includes:

Pages: S1-S38

Tables: S1-S13 (Table S2 and Table S10 were shown in SI Tables.xls)

Figures: S1-S11

Sample Preparation

Eighteen samples of soil were collected from in or around a chemical industrial park in Jiangsu province, China. Surface soil with a depth of 2 cm were firstly removed. 20 g soil samples were collected by wooden spoons and were kept in brown glass bottles which were protected from light. Samples of soil were freeze-dried within 48 h of collection, sieved through 60 mesh stainless steel sieves and were saved in brown bottles. A 20 g aliquot of each soil was uniformly mixed with 0.5 g celite, which had been washed with hexane, DCM and acetone. Soils were further extracted by accelerated solvent extraction (ASE). Extracts were concentrated to 2 ml by rotary vacuum evaporation (type TVE-1000, EYELA, Tokyo, Japan). 1 ml of the concentrate were solvent-changed to dimethyl sulfoxide (DMSO) and was saved at -20 ˚C. Identification and quantification of organic constituents were carried out on an HPLC system (Agilent 1260, Agilent, USA) couple with QTOF-MS system (TripleTOF 5600, AB Sciex, USA). Chromatographic separation was carried out by use of an Agilent HPLC C18 column (2.1 × 100 mm^2, 1.7μm particle size) and with a mobile phase consisting of (A) water with 5% (volume percentage) acetonitrile and (B) methanol with a flow rate of 0.4 mL min^{-1} to acquire enough response of all potential chemicals in the sample extracts.

Accelerate solvent extraction (ASE)

ASE extractions of soil samples were conducted on Dionex ASE350 (Dionex, German). Samples of spoils were extracted three times with mixed solvent of dichloromethane and N-hexane (volume ratio 1:1) followed by extractions in triplicate with pure methanol. The extraction process started with static duration for 5 min, followed by 5-min extraction three times with extraction pressure set at 1500 psi, temperature of extraction cells set at 100 °C, volume was reduced by blowing with nitrogen for 120 seconds.

Fractionation

Extracts of ASE in mixed solvent of DCM and Hex were cleaned-up and enriched on HLB (Oasis, Waters) solid phase extraction (SPE) columns which were pre-conditioned with 10 mL n-hexane, 10 mL of dichloromethane, and 10 mL of methanol with loading rate of 1-2 drops/second. HLB columns were eluted successively with 10 mL of mixed solvents including methanol: Dichloromethane = 1:1 (volume ratio), 10 mL of mixed solvents including dichloromethane: N-hexane = 4:1 (volume ratio), and 10 mL of hexane. Thus, extracts of soil samples were separated into four primary fractions by elution with solvents with different hydrophobicity. Elutes were further concentrated by rotary evaporation and nitrogen blowing.

Preparation and separation were conducted on preparative high performance liquid chromatography (HPLC) Waters AutoPurification (Waters, USA) for the primary fractions with detectable anti-androgenic effect. Preparative chromatography techniques are used to effectively perform high-throughput separation of the primary fractions based on hydrophobicity and retention time of fractions. Waters XBridge C18 preparative columns（19 mm×150 mm, a particle diameter of 5 pm) was used for the fractionation. The mobile phases were water and methanol, and the flow rate was controlled at 5 mL/min. 15 mL of glasses are used to collect elutes according to the time period. Detail collection method was described extensively in Table S1.

Reporter gene assay

The MDA-Kb2 cell line (ATCC CRL-2713, American Tissue Culture Collection, Manassas, VA, USA), which is stably transformed with murine mammalian tumor virus (MMTV)-luciferase, was cultured as recommended. The MDA-Kb2 cell line was cultured in L-15 medium (Sigma, St, Louis, MO, USA) with 10% fetal bovine serum (FBS, Gibco, Invitrogen Corparation, Carlsbad, CA, USA) at 37 °C in ambient atmosphere without CO2. Background potencies were minimized by replace 10% FBS to 10% charcoal-dextran—stripped FBS (CDS-FBS, Biological Industries Ltd. Israel) in the bioassays. Cells were plated on 384-well plate (Corning Inc. Corning, NY, USA) in 80μL assay medium at a concentration of 1×10^{5} cell per Ml after cultured in assay medium for at least 24 hours. After incubation for 24 h, solvent-control, tested extracts and 1.0×10^{-9} M DHT were added into the wells. The cells were then exposed to seven dilutions of tested samples with or without DHT (1 nM). After exposure for 24 hours, exposure medium was removed, and 10 μL of 1×lysis buffer (Promega, Madison, WI, USA) per well was added. A blank control and a solvent control were presented in each plate. After cell lysis for 10 min, 25 μL of luciferase was added per well, and luminescence was quantified immediately in a Synergy H4 hybrid microplate reader (BioTek Instruments Inc., Winooski, VT, USA). During the assay, DMSO was diluted to less than 0.1% to avoid cytotoxicity. Flutamide and DHT were chosen as positive control for anti-androgenic and androgenic potencies, respectively. Each sample was assayed independently at least 3 times (3 replicate assays). Detailed information of cell culture and in *vitro* assays were further described in the SI.

80% Rules

For the LC/MS data set, 80% of the data contained a zero. To reduce the number of zeros present, the following procedure was applied, which will be referred to as the "80% rule". Every sample can be assigned to a certain experiment (experiments 1-10, with the exception of experiment 8, which is not present). For both data sets mentioned, a variable is kept if the variable has a nonzero value for at least 80% of all samples for at least one experiment.

Sample size of OPLS-DA

OPLS-DA was applied in the strategy to select key toxicants from thousands of chemical features, especially efficient when chemical features were far more than sample size. Generally, for valid OPLS-DA, the minimum sample size in each group is 6. Recently, OPLS-DA has been widely applied for identification of metabolites and biomarkers with sample sizes generally around 20, which also varies according to sampling types and sampling areas.

In-house Quantitative Structure-Retention Relationship (QSRR) Model

Fourty four (44) chemicals were ranked in increasing order of log Kow and then used to develop the QSRR model. 4 chemicals were selected every 5 chemicals, 34 chemicals in total served as the training set to develop the model, and the remaining 10 chemicals were used for validation of the model. 1096 molecular descriptors were calculated by PaDEL-Descriptor for each chemical, then normalized for further MVA. PCA was conducted by use of SIMCA-P 13.0.3 (Umetrics, Umea, Sweden) to compare the division of the dataset into the training set and the test set. Partial least squares regression (PLS) was also conducted to find the most important descriptors related to the RT of individual chemicals. VIP were calculated for each descriptor (Table S4) to reveal their contributions to the regression. Log Kow and Log D (pH=6), with the greatest VIP values, were chosen to develop the relationship. Descriptors of the 34 training chemicals were used to develop the model to predict related retention time by multi factor linear regression method in SPSS 23.0, which were further validated by comparison of predicted and observed RTs of the remaining 10 chemicals.

Application domain of the in-house QSRR

Since the application domains are mainly limited to the selection of the training set, varied phase systems and detection of outliers, to minimize the uncertainties, the QSRR model applied during the process was developed with chemicals with wide range of physicochemical properties, including chemicals with Log Kow ranged from -1.74 to 5.28, pKa ranged from -1.48 to 15.06 and Log D(pH=7.4) ranged from -1.37 to 5.19 (added in Table S3), which indicated chemicals in the training set were sufficiently diverse to represent a wide range of chemicals exist in the environment, especially for more polar compounds. Thus, additional attention should be paid when physicochemical properties of suspect structures lie out of the ranged of the training set, which need further validation by additional methods including spectral interpretation and MD simulation. What's more, since successful model depends on selection of suitable molecular descriptors, multivariate analyses (MVA), including principle component analysis (PCA) and partial least-squares (PLS) regression, were applied to select the most representative molecular descriptors to develop the model. For liquid systems, the model built in-house was valid for predictions in the same system, while further corrections should be made by artificial standards when applied to other systems. Furthermore, due to the uncertainty of the model, the threshold of the relative error was set as 20% to minimize the false negative with acceptable false positive.

Structure preparation and in silico simulation

Structures of ligands and receptors were constructed and optimized by Chemoffice (PerkinElmer, USA). A Surflex-Dock module of Sybyl software was linked to an AR-LBD active site of a small molecule for a test, then Grimaces 4.0 molecular modeling software was used for MD simulations adopting for field processing CHARMM27 protein receptors and ligand molecules. TIP3P based spherical layers of water molecules were added to every composite system. A minimum spacing edge of a solute and solvent was 10 Å. Sodium or chloride ion was added so that the system was in equilibrium charge state. All systems were used a steepest-descent method to optimize energy and then to restrict ligand positions, within 40 picoseconds (PS) time the temperature rises from 0 K to 300 K, in the condition of one atmosphere and 300 K, balancing 1 nanosecond (ns), and molecular dynamics simulations were followed, wherein the electrical interaction was applied with a particle mesh Ewald (PME) method to calculate, Van der Waals cutoff was set to 10 Å, all simulations for 22 ns, a step was set to 2 femtoseconds (fs), saving every 2ps. MD simulation data obtained were also processed using GROMACS 4.0, monitoring whether H12 relocation is stable within 20 ns. Query structures were estimated as androgen active or not by monitoring whether H12 relocation was stable within 22 ns.

Uncertainties of the screening process

First, OPLS-DA was applied to select chemicals specific enough to discriminate the polluted region and the reference region, which was evaluated by permutation test for 100 times. Then, application domain of QSRR was mainly limited by the selection of the training set, varied phase systems and detection of the outliers. To minimize the uncertainty, the threshold of the elucidation was extended to a relative error of 20% and the application domains and conditions was further claimed. Moreover, toxicity prediction of suspect chemicals will certainly generate additional uncertainties. However, MD simulation improved the accuracy of the prediction to a great extent compared to simply molecular docking. Nevertheless, for more accurate predictions, multiple validations including molecular docking, MD simulation and binding free energy should be combined.

Table S1. List of artificial suspect compounds.

Chemical Name	CAS No.	Formula	Chemical Name	CAS No.	Formula
Imidacloprid	138261-41-3	C9H10ClN5O2	Pirimiphos-methyl	29232-93-7	C11H20N3O3PS
Fluometuron	2164-17-2	C10H11F3N2O	Prosulfocrab	52888-80-9	C14H21NOS
Buprofezin	69327-76-0	C16H23N3OS	Terbutryn	886-50-0	C10H19N5S
Vitavax	5234-68-4	C12H13NO2S	Triadimefon	43121-43-3	C14H16ClN3O2
Cloquintocet-mexyl	99607-70-2	C18H22ClNO3	Cimetidine	51481-61-9	C10H16N6S
Cyproconazole	94361-06-5	C15H18ClN3O	Bensulfuron methyl	83055-99-6	C16H18N4O7S
Diazinon	333-41-5	C12H21N2O3PS	Isoproturon	34123-59-6	C12H18N2O
Difenoconazole	119446-68-3	C19H17Cl2N3O3	Pyrimethanil	53112-28-0	C12H13N3
Diuron	330-54-1	C9H10Cl2N2O	Lidocaine	137-58-6	C14H22N2O
Linuron	330-55-2	C9H10Cl2N2O2	Nicotine	54-11-5	C10H14N2
Malathion	121-75-5	C10H19O6PS2	Sulpiride	15676-16-1	C15H23N3O4S
Metazachlor	67129-08-2	C14H16ClN3O	Telmisartan	144701-48-4	C33H30N4O2
Praziquantel	55268-74-1	C19H24N2O2	Terbutylazine	5915-41-3	C9H16ClN5
Prometon	1610-18-0	C10H19N5O	Fluconazole	86386-73-4	C13H12F2N6O
Propamocarb	24579-73-5	C9H20N2O2	Irbesartan	138402-11-6	C25H28N6O
Thiabendazole	148-79-8	C10H7N3S	Ornidazole	16773-42-5	C7H10ClN3O3
Thiamethoxam	153719-23-4	C8H10ClN5O3S	Melamine	108-78-1	C3H6N6
2-Benzoylacetanilide	103-84-4	C8H9NO	Erythromycin	114-07-8	C37H67NO13
Triisopropanolamine	122-20-3	C9H21NO3	Venlafaxine	93413-69-5	C17H27NO2
Levamisole	14769-73-4	C11H12N2S	Nicosulfuron	111991-09-4	C15H18N6O6S
Ketoconazole	65277-42-1	C26H28Cl2N4O4	2,4-Dichlorophenol	120-83-2	C6H4Cl2O
Dimethomorph	110488-70-5	C21H22ClNO4	Chloramphenicol	56-75-7	C11H12Cl2N2O5
Imazalil	35554-44-0	C14H14Cl2N2O	N-Lauroylsarcosine	97-78-9	C15H29NO3

Metalaxyl	57837-19-1	C15H21NO4	Fluroxypyr	69377-81-7	C7H5Cl2FN2O3
Prochloraz	67747-09-5	C15H16Cl3N3O2	2-Methyl-4-chlorophenoxyacetic acid	94-74-6	C9H9ClO3
Propiconazole	60207-90-1	C15H17Cl2N3O2	2-Amino-4,6-dimethoxypyrimidine	36315-01-2	C6H9N3O2
Tebuconazole	107534-96-3	C16H22ClN3O	Benzimidazole	51-17-2	C7H6N2
Atrazine	1912-24-9	C8H14ClN5	Imidacloprid-urea	120868-66-8	C9H10ClN3O
Azoxystrobin	131860-33-8	C22H17N3O5	Atrazine-2-hydroxy	2163-68-0	C8H15N5O
Carbendazim	10605-21-7	C9H9N3O2	Propazine-2-hydroxy	7374-53-0	C9H17N5O
(S)-Metolachlor	87392-12-9	C15H22ClNO2	Clothianidin	210880-92-5	C6H8ClN5O2S
Prometryn	7287-19-6	C10H19N5S	Cloquintocet	88349-88-6	C11H8ClNO3
Tricyclazole	41814-78-2	C9H7N3S	4,6-Dimethoxypyrimidine	5270-94-0	C6H8N2O2
Amantadine	768-94-5	C10H17N	2,4,6-Trichlorophenol	88-06-2	C6H3Cl3O
Flutriafol	76674-21-0	C16H13F2N3O	2,4-Dichlorophenoxyacetic acid	94-75-7	C8H6Cl2O3
Acetamiprid	135410-20-7	C10H11ClN4	Roxithromycin	80214-83-1	C41H76N2O15
Clomazone	81777-89-1	C12H14ClNO2	Bis(4-fluorophenyl)-methanone	345-92-6	C13H8F2O
Isoprothiolane	50512-35-1	C12H18O4S2			

Table S3. Reference standards for development of in-house QSRR.

Classification	Chemical Name	CAS No.	Formula	RT (min)	Predict RT(min)
Training	Pirimiphos-methyl	29232-93-7	C11H20N3O3PS	29.56	26
	Terbutryn	886-50-0	C10H19N5S	27.83	26.72
	Prometryn	7287-19-6	C10H19N5S	27.83	26.77
	Buprofezin	69327-76-0	C16H23N3OS	30.6	29.85
	Prometon	1610-18-0	C10H19N5O	25.92	25.45
	Diazinon	333-41-5	C12H21N2O3PS	29.19	28.94
	Atrazine	1912-24-9	C8H14ClN5	24.08	23.21
	Pyrimethanil	53112-28-0	C12H13N3	26.26	24.62
	Cloquintocet-mexyl	99607-70-2	C18H22ClNO3	30.63	34.05
	Ketoconazole	65277-42-1	C26H28Cl2N4O4	29.21	30.27
	Prosulfocrab	52888-80-9	C14H21NOS	30.17	30.03
	Thiabendazole	148-79-8	C10H7N3S	18.99	20.89
	Difenoconazole	119446-68-3	C19H17Cl2N3O3	29.74	33.85
	Tebuconazole	107534-96-3	C16H22ClN3O	28.98	27.96
	Flutriafol	76674-21-0	C16H13F2N3O	24.56	22.82
	Cyproconazole	94361-06-5	C15H18ClN3O	27.7	25.35
	(S)-Metolachlor	87392-12-9	C15H22ClNO2	28.27	26.11
	Prochloraz	67747-09-5	C15H16Cl3N3O2	29.43	29.01
	Propazine-2-hydroxy	7374-53-0	C9H17N5O	21.95	19.93
	Fluometuron	2164-17-2	C10H11F3N2O	23.45	21.79
	Vitavax	5234-68-4	C12H13NO2S	22.56	18.3

Test	Carbendazim	10605-21-7	C9H9N3O2	16.4	18.84
	Tricyclazole	41814-78-2	C9H7N3S	18.23	21.23
	Sulpiride	15676-16-1	C15H23N3O4S	11.13	11.14
	Atrazine-2-hydroxy	2163-68-0	C8H15N5O	18.1	17.61
	Clomazone	81777-89-1	C12H14ClNO2	25.69	24.28
	Cimetidine	51481-61-9	C10H16N6S	11.88	12.47
	Venlafaxine	93413-69-5	C17H27NO2	23.64	21.62
	2-Amino-4,6-dimethoxypyrimidine	36315-01-2	C6H9N3O2	12.33	15.68
	Metazachlor	67129-08-2	C14H16ClN3O	24.67	22.9
	N-Lauroylsarcosine	97-78-9	C15H29NO3	27.46	26.43
	Diuron	330-54-1	C9H10Cl2N2O	24.64	23.19
	Imazalil	35554-44-0	C14H14Cl2N2O	28.96	27.75
	Terbutylazine	5915-41-3	C9H16ClN5	26.6	24.74
	Propiconazole	60207-90-1	C15H17Cl2N3O2	29.15	31.96
	Isoproturon	34123-59-6	C12H18N2O	24.61	23.71
	Triadimefon	43121-43-3	C14H16ClN3O2	27.36	25.25
	Linuron	330-55-2	C9H10Cl2N2O2	26.14	23.99
	Isoprothiolane	50512-35-1	C12H18O4S2	27.17	24.17
	Amantadine	768-94-5	C10H17N	17.65	18.85
	Chloramphenicol	56-75-7	C11H12Cl2N2O5	17.34	17.15
	Ornidazole	16773-42-5	C7H10ClN3O3	13.09	7.32
	Triisopropanolamine	122-20-3	C9H21NO3	7.04	3.32
	Imidacloprid	138261-41-3	C9H10ClN5O2	14.3	8.5

QSRR: Quantitative structure-retention relationship.

Table S4. VIP value for 20 molecular descriptors.

Var ID (Primary)	M1.VIP[3]	1.89456 * M1.VIP[3]cvSE
LogD (pH=6)	2.38661	0.518198
Log Kow	2.23073	0.647711
CrippenLogP	1.84141	0.682962
SpMin5_Bhs	1.69664	0.69485
nHBint7	1.68477	1.72477
SpMin4_Bhm	1.65558	0.577209
SpMin5_Bhm	1.64881	0.579996
SpMax8_Bhp	1.63827	0.298686
SpMax8_Bhv	1.63602	0.312633
SpMin4_Bhs	1.6185	0.625441
ATSC0v	1.61478	0.241709
SpMin6_Bhm	1.60712	0.499958
SpMax5_Bhv	1.60252	0.482636
SpMax8_Bhm	1.60053	0.423572
SpMax7_Bhp	1.59866	0.493937
SpMin6_Bhs	1.58905	0.585074
SpMax8_Bhe	1.58699	0.323708
SpMin6_Bhv	1.58343	0.499536
SpMax5_Bhp	1.58272	0.514182
SpMin6_Bhe	1.57835	0.580445

VIP: Variable importance in the projection. Top 20 molecular descriptors were selected according to the order of VIP value.

Table S5. Collection time of preparative fractionation.

Fraction	Collection Time (min)	Fraction	Collection Time (min)	Fraction	Collection Time (min)
1	0.5	18	9.2	35	25
2	1	19	9.7	36	26
3	1.5	20	10.2	37	27
4	2	21	11.4	38	28
5	2.5	22	12.6	39	29
6	3	23	13	40	30
7	3.5	24	14	41	31
8	4	25	15	42	32
9	4.5	26	16	43	33
10	5	27	17	44	34
11	5.7	28	18	45	36
12	6.2	29	19	46	38
13	6.7	30	20	47	40
14	7.2	31	21	48	42
15	7.7	32	22	49	45
16	8.2	33	23	50	50
17	8.7	34	24	51	65

Table S6. Detail information of typical androgenic compounds.

Compound name	CAS No.	Formula	Purity	Information
Boldenone	846-48-0	C19H26O2	99.90%	J&K
Nandrolone	434-22-0	C18H26O2	99.90%	J&K
Androstenedione	63-05-8	C19H26O2	99.90%	J&K
Formestane	22259-30-9	C19H26O3	99.90%	J&K
Testosterone	58-22-0	C19H28O2	99.90%	J&K
17-Methyltestosterone	58-18-4	C20H30O2	99.90%	J&K
Epiandrosterone	481-29-8	C19H30O2	99.90%	J&K
Stanolone	521-18-6	C19H30O2	99.90%	J&K
Androsterone	53-41-8	C19H30O2	99.90%	J&K
Octylphenol	27193-28-8	C14H22O	99.80%	Sigma
Nonylphenol	25154-52-3	C15H24O	99.80%	Sigma
Diphenylolpropane	80-05-7	C15H16O2	99.90%	Sigma
Dimethyl phthalate	131-11-3	C10H10O4	99.90%	Sigma
Diethyl phthalate	84-66-2	C12H14O4	99.90%	Sigma
Dibutyl phthalate	84-74-2	C16H22O4	99.90%	Sigma
Butyl Benzyl phthalate	85-68-7	C19H20O4	99.90%	Sigma
Bis(2-ethylhexyl) phthalate	117-81-7	C24H38O4	99.90%	Sigma
Diisooctyl phthalate	27554-26-3	C24H38O4	99.90%	Sigma

Table S7. Limit of quantification (LOQ) and recovery of typical androgenic compounds.

Class	Chemicals	LOQ	Procedural Recovery	
		(μg/L)	Recovery	RSD
Androgenic	Androstenedione	0.12	102.30%	2.70%
	Nandrolone	0.31	94.40%	3.70%
	Formestane	0.37	79.70%	9.20%
	17-Methyltestosterone	0.16	97.80%	4.50%
	Stanolone	0.56	97.20%	4.60%
	Testosterone	0.11	96.90%	2.70%
	Epiandrosterone	0.32	98.20%	4.80%
	Androsterone	0.15	82.10%	6.30%
	Boldenone	0.23	98.20%	2.60%
Anti-androgenic	Dimethyl phthalate	0.13	94.10%	3.30%
	Diethyl phthalate	0.43	94.90%	1.30%
	Dibutyl phthalate	1.51	95.20%	4.20%
	Butyl Benzyl phthalate	0.81	92.30%	3.20%
	Bis(2-ethylhexyl) phthalate	1.60	107.40%	11.40%
	Diisooctyl phthalate	0.70	91.80%	8.70%
	Octylphenol	0.43	92.50%	6.70%
	Nonylphenol	0.73	91.80%	5.20%
	Bisphenol A	0.45	82.50%	4.10%

Table S8. Anti-AR EQ of soil samples.

Sample ID	S4	S6	S8	S9	S11	S16	S17	S18
Anti-AR EQ (μg Flutamide/ g Soil)	44.53	24.47	101.20	126.16	143.90	23.23	101.86	33.66

Anti-AR EQ: Androgen antagonist potency equivalent.

Table S9. Prior peaks eluted by MVA and related confidence level.

Ion Mode	Retention Time (min)	Precursor m/z	VIP	Confidence Level	Ion Mode	Retention Time (min)	Precursor m/z	VIP	Confidence Level
Positive	31.18	331.1895	1.363	2a	Positive	29.24	267.1555	1.507	4
	27.75	360.1490	1.084	2b		24.67	265.1039	1.379	4
	25.53	353.1570	2.146	2b		20.29	264.8989	1.542	4
	26.4	349.1968	1.050	2b		28.53	255.1344	1.060	4
	28.86	335.2168	1.212	2b		29.63	254.0949	2.153	4
	24.34	329.1920	1.061	2b		22.49	251.1613	1.414	4
	31.16	322.1574	1.347	2b		2.3	240.0835	1.471	4
	10.82	298.0966	1.743	2b		26.41	239.1253	1.234	4
	23.02	293.1712	1.028	2b		24.39	237.1451	1.145	4
	24.84	275.1610	1.221	2b		6.98	223.0921	1.119	4
	2.51	268.1032	1.545	2b		24.09	209.1142	1.325	4
	31.42	258.1269	1.525	2b		22.82	203.1026	1.287	4
	28.77	230.0952	1.071	2b		10.97	194.9076	1.054	4
	25.72	229.0671	1.896	2b		36.96	188.9168	1.854	4
	22.19	227.0394	1.236	2b		26.36	313.1760	1.272	5
	24.49	225.1081	1.027	2b		19.47	214.9154	1.499	5
	15.92	213.0235	1.155	2b		29.8	205.0831	1.277	5
	3.9	176.9885	1.130	2b		35.71	200.9370	1.298	5
	28.68	391.2076	1.157	3	Negative	31.3	339.2003	1.342	1b
	35.53	318.8741	1.022	3		33.31	325.1831	1.262	1b

7.47	198.9392	1.557	3	27.88	311.1687	1.778	1b
23.7	399.1609	1.340	4	31.83	381.2310	1.469	2b
26.14	351.2125	1.136	4	0.98	377.0839	1.752	2b
29.07	337.2350	1.298	4	31.77	327.2904	1.441	2b
31.44	336.3094	1.130	4	23.04	278.9142	1.003	2b
23.34	321.1668	1.082	4	9.57	121.0312	1.126	2b
27.73	309.2025	1.329	4	30.25	293.1765	1.229	2b
24.76	307.1872	1.129	4	31.9	351.2209	1.810	4
28.95	305.1078	1.631	4	31.86	253.2179	1.024	4
28.58	299.1613	2.052	4	12.3	174.9572	1.186	4
31.22	295.2244	1.147	4	35.36	132.9245	1.223	4
27.64	291.1920	1.819	4	24.13	110.9776	1.244	4
29.6	289.1770	1.099	4	35.08	115.9209	1.720	5
27.21	277.1755	1.702	4				

MVA: Multivariate analysis.

Table S11. 14 selected target androgenic compounds.

Source	Chemical Name	CAS No.
Ministry of the Environment, Government of Japan	Nonylphenol	25154-52-3
	Octylphenol	1806-26-4
	Bisphenol A	80-05-7
	o,p'-DDT	789-02-6
Tier 2 chemicals in EDSP	Chlorothalonil	1897-45-6
	Tebuconazole	107534-96-3
	2-Phenylphenol	90-43-7
	Pentachloronitrobenzene	82-68-8
	Propiconazole	60207-90-1
	Folpet	133-07-3
	Myclobutanil	88671-89-0
	Flutolanil	66332-96-5
	Linuron	330-55-2
	Propargite	2312-35-8

Selecting strategy: Target androgenic compounds were selected from 18 EDCs for further Tier 2 evaluation by EPA, 2 confirmed EDCs by European Community Rolling Action Plan (CoRAP) and 4 EDCs with enough province reported by the Ministry of the Environment, Government of Japan.

Table S12. Structures identified by traditional EDA method.

Precursor	Formula	MassBank ID	MetFrag Score	Log Kow
399.2489	C19H29F3N6	5757977	0.75	1.82
399.2489	C20H34N2O6	3356741	0.82	0.63
274.9743	C10H11Br1O2S1	45038420	0.88	2.19
274.9743	C10H11Br1O2S1	45038421	0.81	2.19
274.9743	C10H11Br1O2S1	40435033	0.78	2.19
287.2687	C16H34N2O2	9646542	0.87	2.61
287.2687	C16H34N2O2	9441173	0.87	2.31
287.2687	C16H34N2O2	35829539	0.87	2.18
287.2687	C16H34N2O2	36197127	0.87	2.1
315.2995	C18H38N2O2	18725135	0.83	2.88
315.2995	C18H38N2O2	23177696	0.83	2.88
342.0758	C15H17Cl2N3O2	30400258	0.89	2.69
342.0758	C15H17Cl2N3O2	35162318	0.89	2.69
342.0758	C15H17Cl2N3O2	30400410	0.89	2.54
342.0758	C15H17Cl2N3O2	30400437	0.89	2.7
342.0758	C15H17Cl2N3O2	12031830	0.75	2.74
409.1064	C20H20N6O4	21840309	0.96	2.48
409.1064	C20H20N6O4	21008081	0.96	3.04
409.1064	C20H23Cl1F2N4O1	9970437	0.88	2.69
409.1064	C20H23Cl1F2N4O1	23289250	0.88	2.59
186.2211	C12H27N1	7340	0.78	4.46
186.2211	C12H27N1	43624453	0.79	4.52
186.2211	C12H27N1	43624264	0.79	4.52
186.2211	C12H27N1	38197322	0.76	4.52
186.2211	C12H27N1	38591762	0.76	4.52
393.2488	C21H36N4O3	17241295	0.84	1.57
315.2997	C18H38N2O2	5310585	0.73	1.17
315.2992	C18H38N2O2	5310585	0.73	1.17
274.9741	C10H11Br1O2S1	45038420	0.88	2.19
274.9741	C10H11Br1O2S1	40435033	0.81	2.19
274.9741	C10H11Br1O2S1	45038421	0.8	2.19
279.1585	C16H22O4	4277003	0.91	4.55
279.1585	C16H22O4	160567	0.91	4.73
279.1585	C16H22O4	19208	0.91	4.73
279.1585	C16H22O4	82082	0.91	4.73
279.1585	C16H22O4	142002	0.91	4.73
279.1585	C16H22O4	98244	0.91	4.73
279.1585	C16H22O4	71681	0.91	4.8
279.1585	C16H22O4	280978	0.91	4.69
279.1585	C16H22O4	136267	0.91	5.07

279.1585	C16H22O4	285481	0.78	4.36
315.3002	C18H38N2O2	5310585	0.74	1.17
315.3002	C18H38N2O2	5310585	0.72	1.17
315.2998	C18H38N2O2	5310585	0.73	1.17
451.1631	C21H27N2O7P1	10152405	0.8	1.48
287.2687	C16H34N2O2	35829539	0.87	2.18
287.2687	C16H34N2O2	36197127	0.87	2.1
302.1095	C16H13F2N3O1	82827	0.91	2.52
302.1095	C16H13F2N3O1	116051	0.91	2.52
302.1095	C16H13F2N3O1	35196717	0.91	2.44
302.1095	C16H13F2N3O1	32644721	0.71	2.26
342.0759	C15H17Cl2N3O2	29477383	0.88	1.91
342.0758	C16H12Cl1N5O2	12031745	0.96	3.86
342.0758	C16H12Cl1N5O2	2346155	0.96	3.86
342.0754	C16H12Cl1N5O2	12031830	0.96	2.74
302.1088	C16H13F2N3O1	82827	0.93	2.52
302.1088	C16H13F2N3O1	116051	0.93	2.52
302.1088	C16H13F2N3O1	35196717	0.9	2.44
308.1518	C16H22Cl1N3O1	37095859	0.9	1.7
308.1518	C16H22Cl1N3O1	37095950	0.9	1.63
308.1518	C16H22Cl1N3O1	37095952	0.9	1.91
399.2489	C19H29F3N6	5757977	0.75	1.82
399.2489	C19H29F3N6	33006171	0.85	2.44
165.0475	C7H10Cl1F1O1	32994015	0.89	2.04
165.0475	C7H10Cl1F1O1	45691398	0.72	1.66
165.0475	C7H10Cl1F1O1	45691405	0.72	1.66
165.0475	C7H10Cl1F1O1	45691416	0.72	1.66
165.0475	C7H10Cl1F1O1	49541520	0.72	1.66
165.0475	C7H10Cl1F1O1	49541521	0.72	1.66
287.2685	C16H34N2O2	9646542	0.87	2.61
287.2685	C16H34N2O2	9441173	0.87	2.31
287.2685	C16H34N2O2	35829539	0.87	2.18
287.2685	C16H34N2O2	36197127	0.87	2.1
342.0757	C15H17Cl2N3O2	30400258	0.87	2.69
342.0757	C15H17Cl2N3O2	35162318	0.87	2.69
342.0757	C15H17Cl2N3O2	30400410	0.87	2.54
342.0757	C15H17Cl2N3O2	30400437	0.87	2.7
342.0757	C15H17Cl2N3O2	29477383	0.74	1.91
209.2003	C13H24N2	9096964	0.88	4.72
209.2003	C13H24N2	37325762	0.88	5.39
318.2451	C16H28N6O2	1184923	0.82	0.25
273.1379	C13H21Cl1N2O2Cl	35743648	0.91	1.1
287.1532	C20H18N2	8801623	0.79	1.17

Table S13. Intensity of prior precursors setected in androgenic fractions.

Ion Mode	Precursor	4-1	6-1	8-0	9-1	9-2	11-2
Positive	399.1609	784	1107	244096	1337	895	19754
	391.2076	0	1975	4022	0	0	546
	360.149	2282	19384	1675	14485	0	679
	353.157	0	0	364692	0	0	10757
	351.2125	776	0	5578	1397	1229	612
	349.1968	0	0	10299	0	0	0
	337.235	325804	128754	80532	914896	940684	128316
	336.3094	0	0	2313	0	0	0
	335.2168	0	6250	6063	4030	0	0
	331.1895	7586	17572	9269	19792	11618	11882
	329.192	5014	15320	6727	4525	7974	9567
	322.1574	0	0	2609	0	0	0
	321.1668	0	620	31724	0	0	508
	318.8741	0	0	0	0	1997	868
	313.176	2370	4783	2807	12970	70678	3738
	309.2025	6776	14186	8805	160788	75029	2178
	307.1872	0	522	11498	2239	1929	822
	305.1078	0	0	154463	859	0	1694
	299.1613	4358	5842	2564	54261	18895	16381
	298.0966	0	0	98337	0	0	808
	295.2244	2262	1466	3616	83285	9010	699
	293.1712	0	2758	21468	0	0	607
	291.192	3045	3360	5314	5530	28725	754
	289.177	805	11088	4559	26046	6329	4316
	277.1755	0	0	11263	0	0	1539
	275.161	0	3875	8945	7824	4171	1315
	268.1032	0	0	0	0	3260	0
	267.1555	15537	3394	9521	78731	105852	693
	265.1039	538	0	58996	0	780	2874
	264.8989	2144	2815	0	4562	3483	1481
	258.1269	0	0	7448	0	0	0
	255.1344	0	1015	11365	1776	0	0
	254.0949	0	0	68315	0	0	1173
	251.1613	520	1429	14711	1754	2474	1529
	240.0835	0	0	37512	0	0	0
	239.1253	991	4145	51891	3439	1679	2987
	237.1451	1725	2086	11018	4482	12367	597
	230.0952	0	1206	70789	2212	0	0
	229.0671	738	2226	37553	0	855	802
	227.0394	0	0	3694	0	0	0

	225.1081	592	1393	7840	1613	1806	815
	223.0921	0	0	11338	4545	0	0
	214.9154	0	0	2629	0	0	0
	213.0235	0	0	0	0	0	0
	209.1142	1160	1472	5338	0	2166	560
	205.0831	0	0	0	0	5960	0
	203.1026	3597	27528	0	26607	45755	8762
	200.937	2001	2312	1423	5118	5466	1477
	198.9392	1642	2776	0	4679	2295	763
	194.9076	785	968	3606	2108	764	0
	188.9168	6426	0	6345	0	0	5433
	176.9885	2168	4622	14463	6978	0	3269
Negative	381.231	25718	48897	6892004	128422	137483	14572
	377.0839	795	0	488041	975	139121	769
	351.2209	15403	48362	6193419	154787	78408	9466
	339.2003	86206	142704	9801307	436283	680005	135555
	327.2904	9538	14811	242446	15771	15899	17354
	325.1831	408901	559718	14678830	1346438	1843822	105493
	311.1687	606984	188290	14413550	704628	1578282	87265
	293.1765	0	882642	0	0	0	0
	278.9142	4234	7040	0	6982	7935	3841
	253.2179	971331	673162	1075150	858247	1081767	693335
	174.9572	26329	38775	191895	35002	40977	17109
	132.9245	97001	87541	67505	70314	68236	63426
	121.0312	2427	6839	84706	7463	14813	2428
	115.9209	0	0	0	0	0	0
	110.9776	0	93167	71031	99908	102292	94255

Figure S1

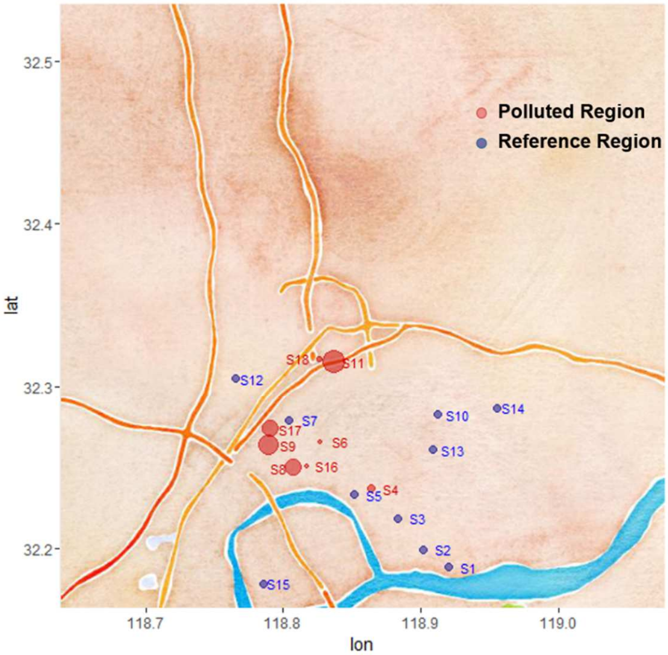
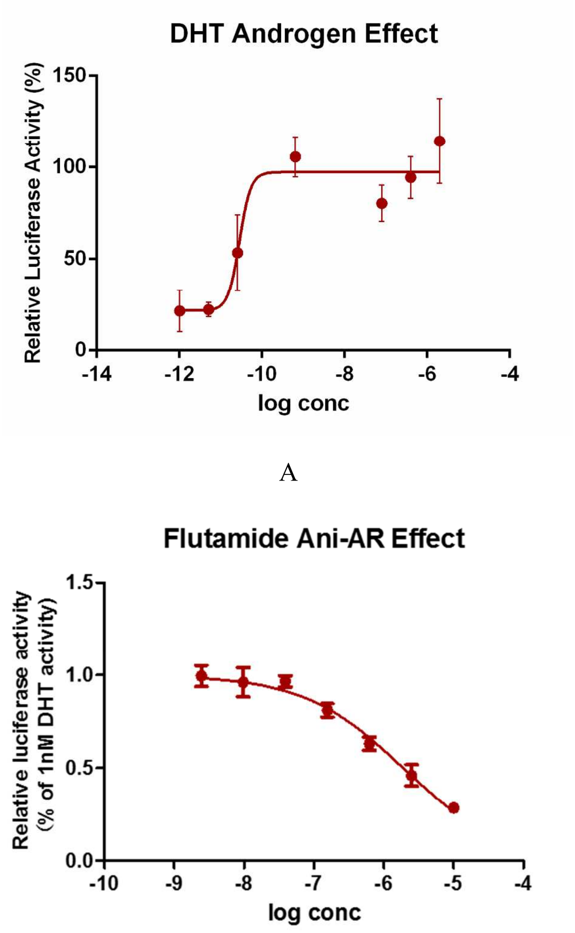


Figure S1. Plot of sample location. Orange dots in the plot represents the polluted sites and the blue dots represents the reference sites. The square of the dots represents related androgenic antagonist equivalents. The greater the dots are, the greater of the androgenic potencies were exhibited.

Figure S2



A

B

Figure S2. Dose-effect curve of positive control. (A) Dose-effect curve of DHT. (B) Dose-effect curve of flutamide.

Figure S3

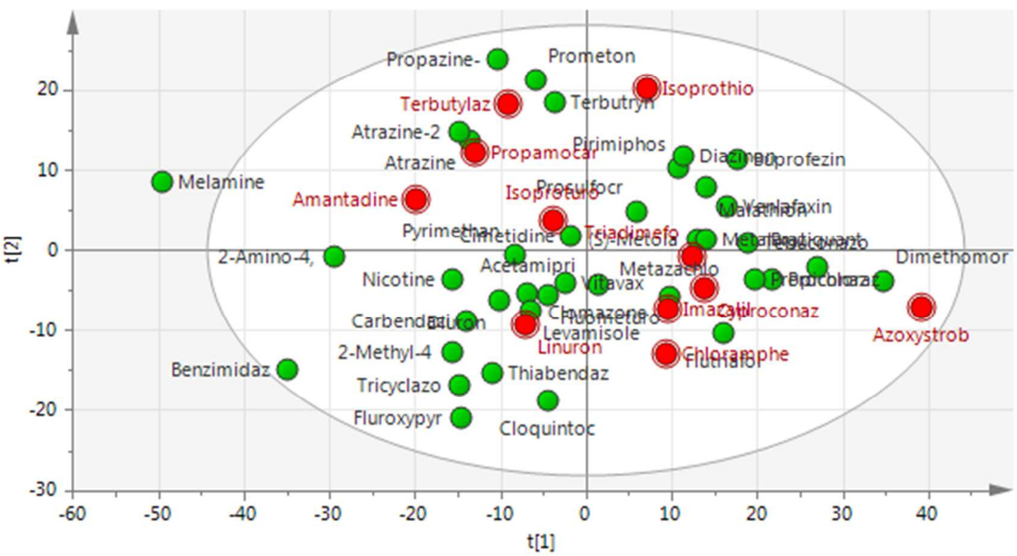


Figure S3. Principle component analysis (PCA) score plot of the training set and the test set in the QSRR model. Green dots are chemicals in the training set and the red dots are chemicals in the test set.

Figure S4

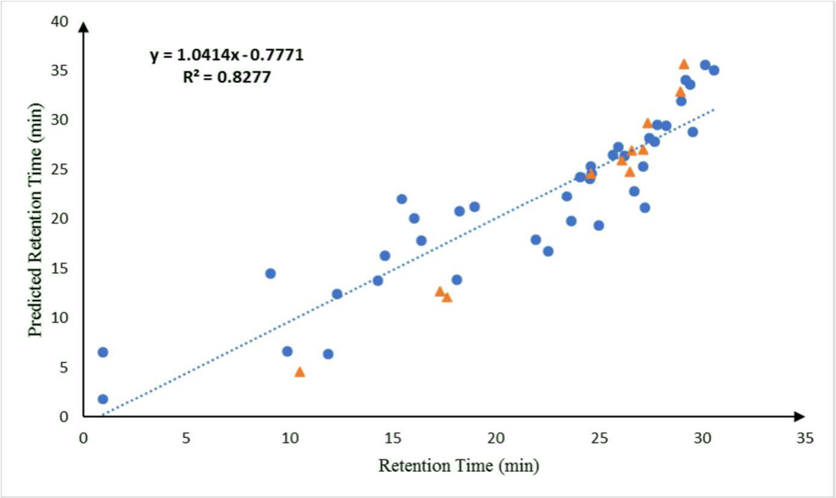


Figure S4. In-house QSRR model.

Figure S5

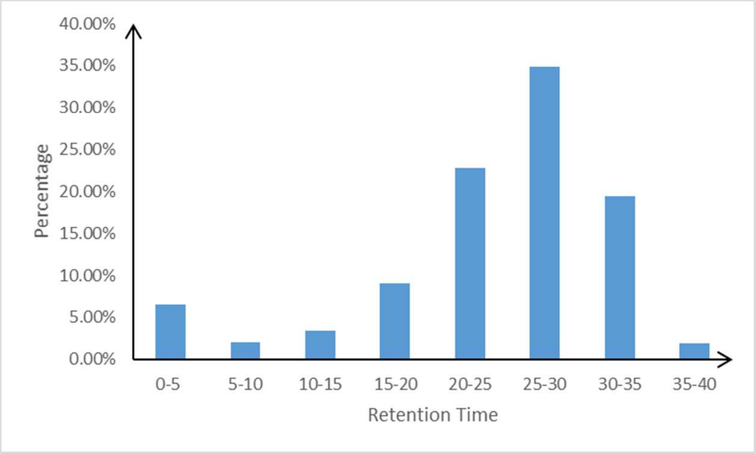
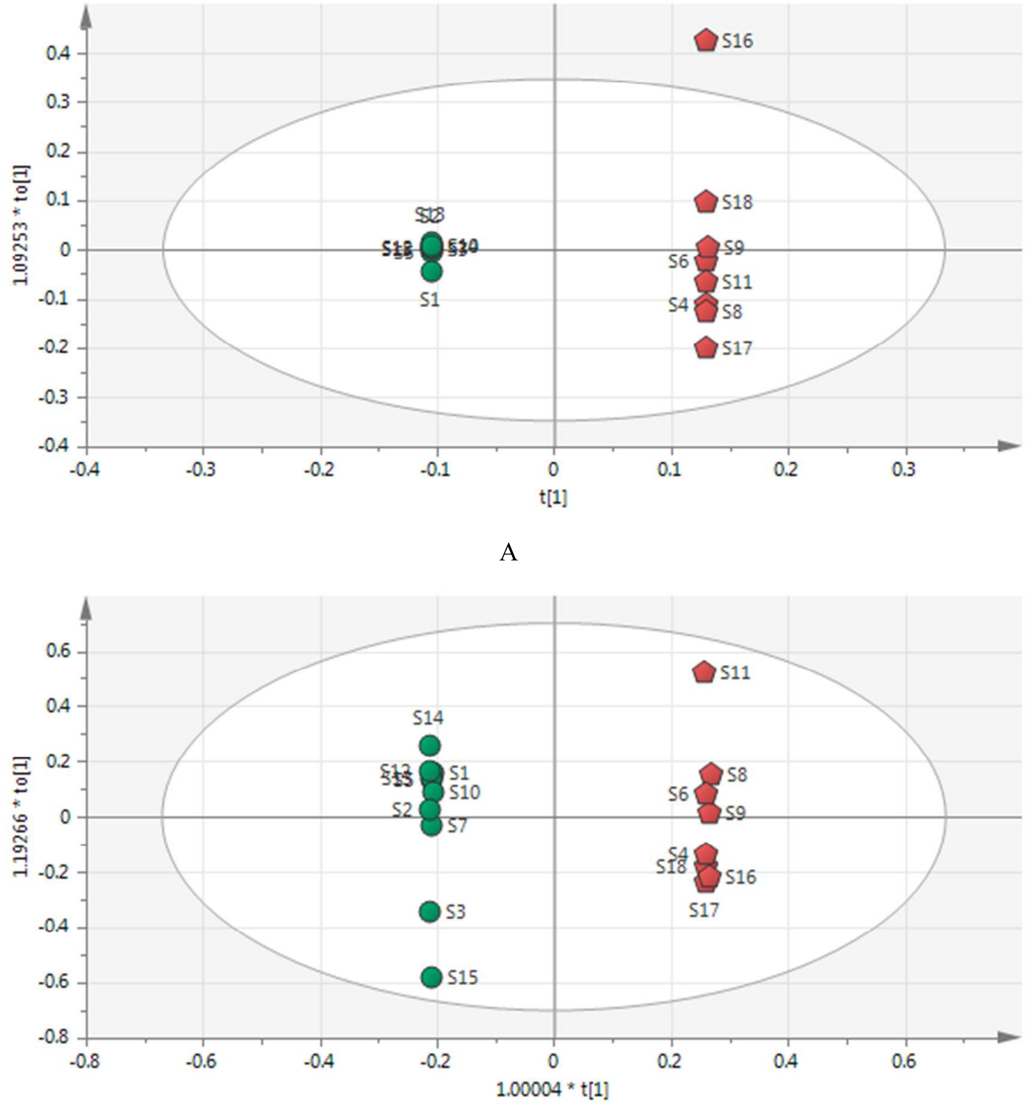


Figure S5. Distribution of retention time of eluted peaks for multivariate analysis.

Figure S6

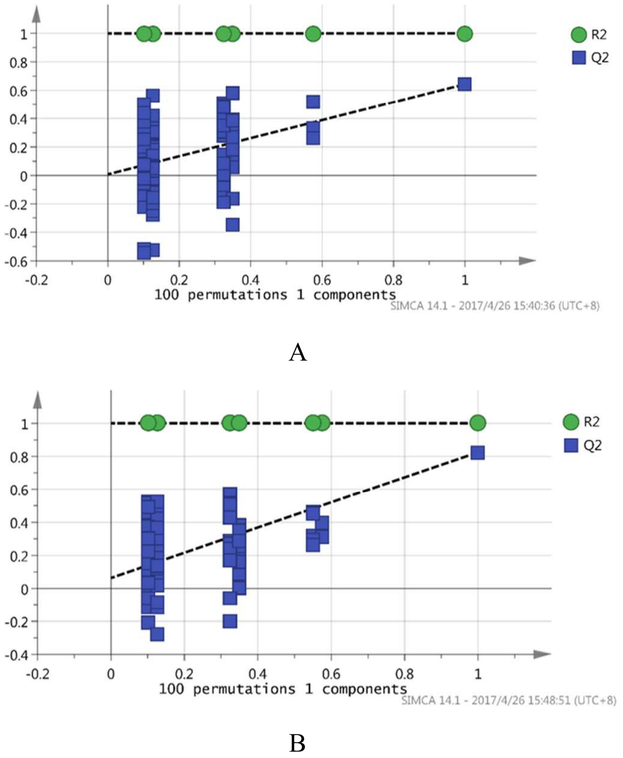


A

B

Figure S6. Orthogonal projection to latent structures-discriminant analysis（OPLS-DA）score plot and S-plot in positive and negative mode. The dots in the plot were all corresponding to the features in the raw LC-MS data. (A) and (B) were OPLS-DA score plot of 18 soil samples in positive and negative mode, respectively; Red dots were samples with AR antagonist potency while the green dots represent non-effective compounds.

Figure S7



A

B

Figure S7. Permuation tests of OPLS models. A and B represent the permutation tests of negative mode (NI) and positive mode (PI), respectively. The developed model is valid if the blue spots (Q^2 values) to the left are lower than the original points to the right.

Figure S8

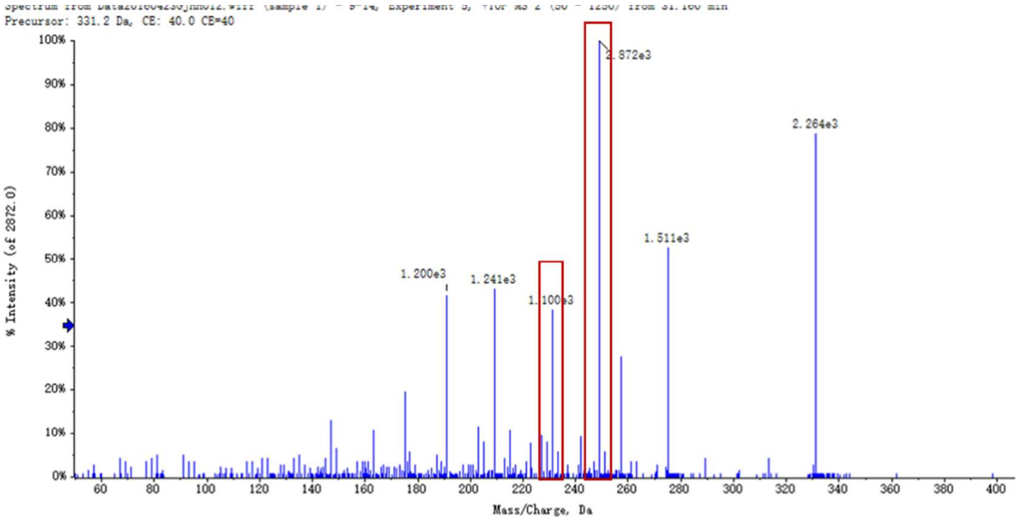
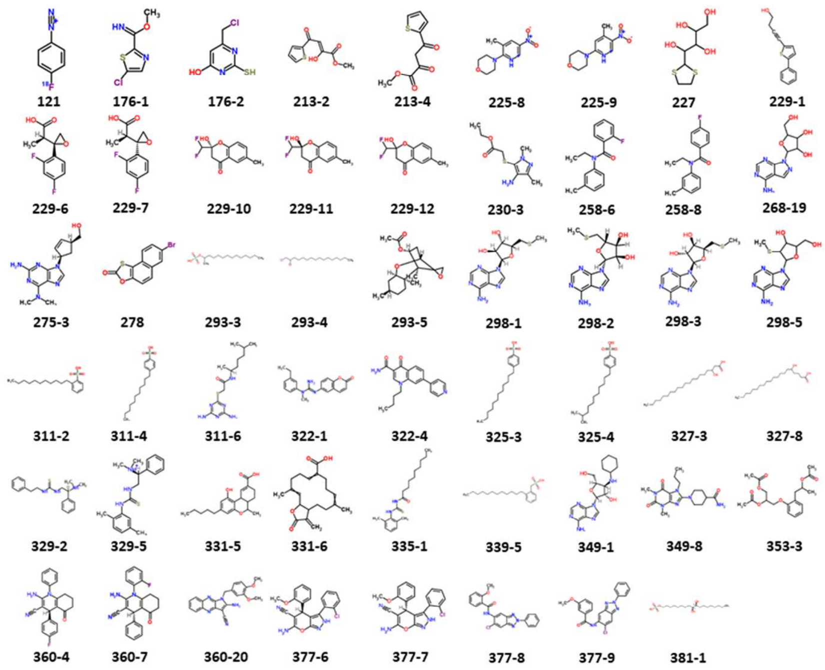
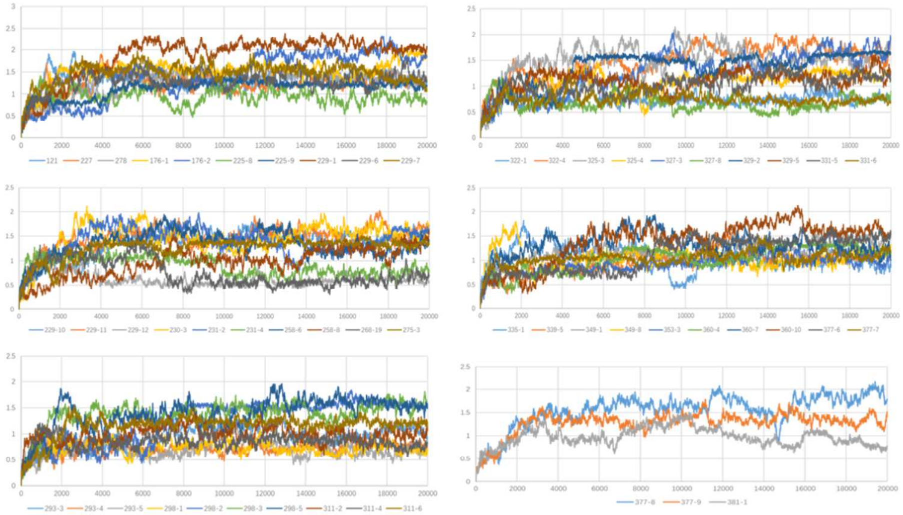


Figure S8. MS/MS spectrum of suspect Dicyclohexyl phthalate.

Figure S9


121
176-1
176-2
213-2
213-4
225-8
225-9
227
229-1
229-6
229-7
229-10
229-11
229-12
230-3
258-6
258-8
268-19
275-3
278
293-3
293-4
293-5
298-1
298-2
298-3
298-5
311-2
311-4
311-6
322-1
322-4
325-3
325-4
327-3
327-8
329-2
329-5
331-5
331-6
335-1
339-5
349-1
349-8
353-3
360-4
360-7
360-20
377-6
377-7
377-8
377-9
381-1

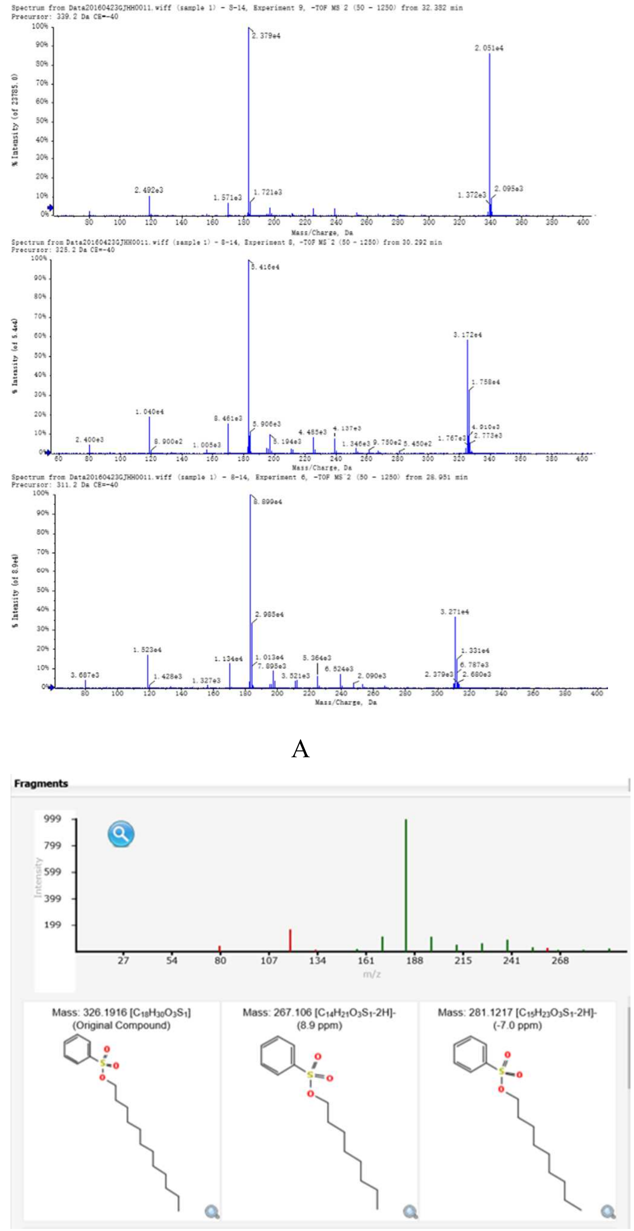
A



B

Figure S9. Diagnostic strucutres with confidence level of 2b after toxicity prediction by molecular dynamics simulation. (A) Individual structure of 53 diagnostic structures wih confidence level 2b. The name of the structure was match with the name in Figure S9B. (B) Toxicity predicition of diagnostic structures by molecular dynamics simulation. Relocations of H12 of all these 53 process were all stable in 20ns, which indicates these 53 structures were potentially anti-androgenic.

Figure S10



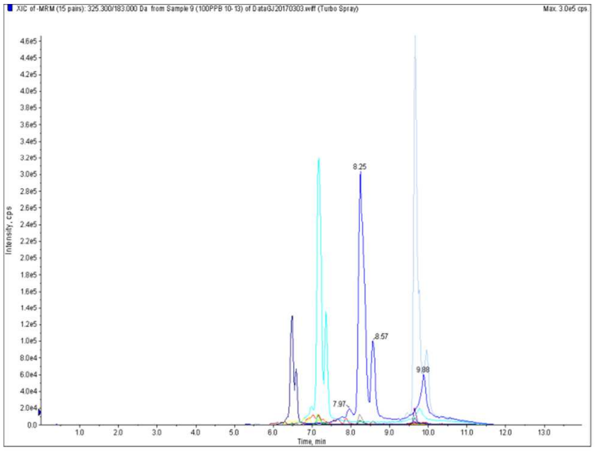
A

Mass: 326.1916 [$C_{18}H_{30}O_3S_1$] (Original Compound)

Mass: 267.106 [$C_{14}H_{21}O_3S_1$-2H]- (8.9 ppm)

Mass: 281.1217 [$C_{15}H_{23}O_3S_1$-2H]- (-7.0 ppm)

B



C

Figure S10. Identification of three alkyl benzenesulfonic acid. (A) MS/MS spectrums of the three homologues. (B) MS/MS peaks interpretation by in silico platform MetFrag. (C) The chemical confirmation with reference standard.

Figure S11

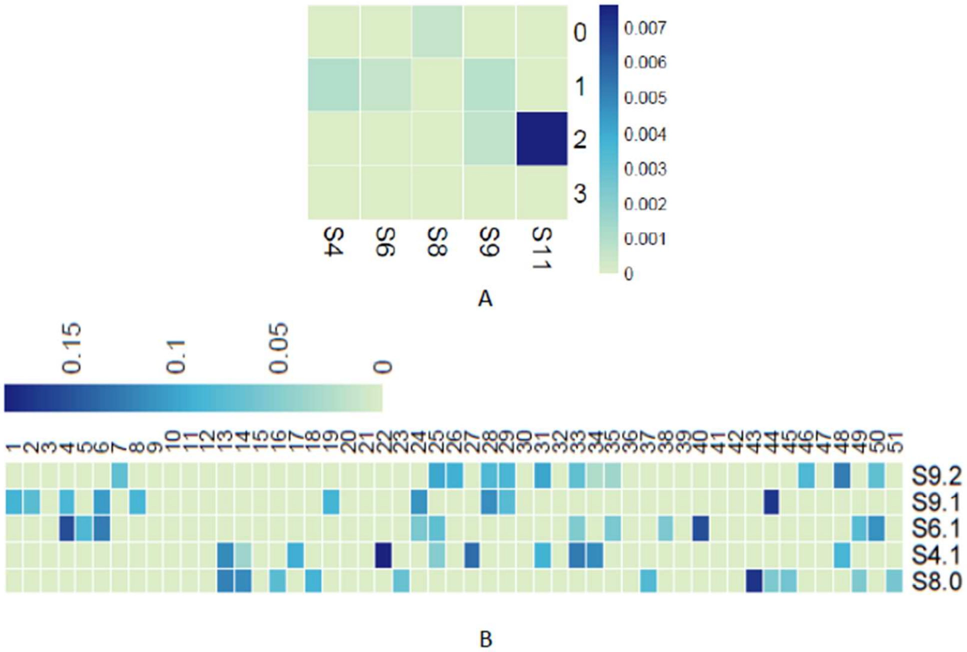


Figure S11. AR antagonist potencies of fractions and preparative fractions. AR antagonist potencies of fractions and preparative fractions were shown in (A) and (B), respectively. The greater the AR antagonist potencies are, the darker the color marked.