

# Improving Structure Based Models for Predicting Chemical Functions and Weight Fractions in Cosmetic Products using Ensemble Support Vector Machine

Tanja B Vojinovic (✉ [vojinovic.t@ucg.ac.me](mailto:vojinovic.t@ucg.ac.me))

University of Montenegro Faculty of Medicine <https://orcid.org/0000-0002-6593-6315>

Ljiljana Kascelan

University of Montenegro Faculty of Economics <https://orcid.org/0000-0001-9831-7599>

Zorica Potpara

University of Montenegro Faculty of Medicine study program Pharmacy <https://orcid.org/0000-0002-8128-8434>

Natasa Radonjic

University of Montenegro Faculty of Medicine study program Pharmacy <https://orcid.org/0000-0003-3811-9958>

Zorka Kascelan

University of Montenegro Faculty of Medicine study program Pharmacy <https://orcid.org/0000-0002-7786-7484>

---

## Research article

**Keywords:** Support Vector Machine, Chemical functions, Weight fractions, Cosmetic products, Toxic chemicals

**Posted Date:** July 6th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-39565/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

# Abstract

Through usage of a large number of cosmetic products, consumers are very often exposed to toxic chemicals. This paper is aimed at proposing a model for the prediction of chemical functions and weight fractions in these products based on the structural and physical-chemical properties of the chemicals. Due to the imbalance of classes we used Support Vector Machine (SVM) method, which can complement a smaller class with the examples that are most similar to it and identify the examples that are most different. The generality of the SVM method was additionally enhanced by combining it with ensemble Bootstrap Aggregation (Bagging). The research results show that the proposed bagging SVM method can overcome the disadvantages of previously applied methods. Further, it can help address the lack of information needed to assess exposure to risk from the use of cosmetic products containing toxic chemicals in their composition. The proposed models can be applied to predict whether a certain chemical may be a substitute for a function performed by another possibly toxic chemical in a cosmetic product, as well as to determine the quantity proportion of a certain dangerous chemical on the basis of its chemical structure and physical-chemical properties.

## Introduction

Instead of a beneficial effect of a cosmetic product, the undesirable and harmful occurrences often appear at the area of application, and sometimes in distant parts of the skin. This may be due to the active cosmetic substances and also to the components of the vehikulum. Irritant-toxic reactions occur as a result of the use of a harmful chemical or several of those and will occur with anyone who uses such a product. An allergic reaction is caused by a sensitivity to one or several substances. Irritant acts as follows: it chemically reacts onto the cells of the epidermis, dissolves or removes some of the essential metabolites, disrupts the osmotic balance of the cells, reduces the membrane potential and thus changes the cell permeability, irreversibly denatures the proteins. Allergic reactions are more common than the irritative ones. Acute and chronic inflammations of the skin are acute allergic dermatitis. Allergic contact skin inflammation is most commonly caused by low molecular weight substances. These are partial antigens that, by binding to the epidermal and dermal proteins, become full antigens and cause an allergic reaction or specific hypersensitivity. Some low molecular weight substances are converted to primary irritants or allergens only after UV-A and UV-B radiation and short visible spectrum waves (400–600 nm wavelength) [1, 2, 3, 4].

Users of cosmetic products are exposed to the toxic chemicals used in the manufacture of these products on a daily basis. In most common cases, manufacturers provide neither a complete information about their qualitative composition, i.e. the set of chemicals they contain, nor the quantitative composition, i.e. the weight fraction of chemicals. The main goals of green chemistry are to reduce the use of toxic chemicals (cancerous as well as those that affect the reproductive and nerve system) while preserving the functionality of the chemical ingredients and the efficiency of the product. Classification of the chemicals by the chemical function they perform may contribute to finding more easily the less toxic substitutes among chemicals that have the same functionality [5, 6]. The risk of exposure is highly

dependent on the quantity of toxic chemical used, and those information are often unavailable or incomplete due to the lack of adequate regulations and business policies of the manufacturer [7, 8, 9]. A large number of chemicals appear on the market, while no high-quality in vivo toxicity data exist for most of them, and for many there are none toxicity data [10]. The standard method for toxicity assessment is high-throughput screening (HTS) applied by the United States Environmental Protection Agency (EPA). This method involves assessing the potential harmfulness of a chemical in vitro by quantifying its bioactivity. The EPA evaluated about 8,000 chemicals using this method. However, there are approximately 100,000 chemicals in use in the US market, which are difficult to test individually [6]. EPA is currently working to develop methods for the substitution of chemicals that have been identified as hazardous with safe chemicals, aimed at providing formulations that are safe for humans and environment [11, 12]. Such methods involve identifying alternatives from available chemical databases. They take an individual chemical to evaluate and return multiple possible alternatives (low-throughput approach). Such an approach is not efficient enough for many chemicals, therefore, the methods are used to automatically identify groups of chemicals with appropriate function in large databases (high-throughput-HT access) for further more accurate testing. Recently, there have been a number of studies proposing quantitative structure – use relationship (QSUR) methods for classifying and predicting the functionality of chemicals based on their chemical structure and/or physical chemical properties [6, 13].

Isaacs et al., 2016 [13] proposed a model for prediction of functions and weight proportion of chemicals in cosmetic products based on physical-chemical properties and use of chemicals in production, based on the random forest method. Due to the imbalance of classes (the number of examples corresponding to one function is much smaller than the number of examples corresponding to all other functions), they applied undersampling so that the smaller class equals the greater one. Phillips et al., 2017 [6] applied the same balanced random forest method for prediction of chemical function in in a broader set of consumer products. The main drawback of the method used in this studies is random undersampling within the ensemble method, which leads to higher number of misclassified examples of negative class i.e. small precision for the positive class. Such a model well classifies known examples i.e. chemicals in a training set, but on an unknown set there is less potential for accurate classification of positive class examples (chemicals that could have a certain function based on their structure and physical chemical properties).

Therefore, the aim of our research is to more efficiently overcome the problem of class imbalance, i.e. to generate the QSUR models with a higher precision of the positive class and thus with a more accurate prediction, i.e. by targeting of chemicals that could be potential functional substitutes for use in cosmetic products.

It has been confirmed in the literature that the Support Vector machine method could successfully solve the problem of class imbalance by eliminating data noise that leads to class overlaps, so this method can be used as a pre-processor that refines data [14, 15, 16]. The classifiers applied to SVM output (on the refined dataset) significantly improve their predictive performance [17, 18, 19]. Farquard & Bose, 2012 tested SVM as a pre-processor for highly unbalanced data (94%: 6%) and showed that SVM can balance data better than standard techniques commonly used such as undersampling (taking a subset of a larger

class) or oversampling (supplementing of minor class with new examples). SVM provides more minor class examples by associating the most similar major class examples to the minor class. Also, the authors showed that when SVM refines data and balances classes, other classification methods (such as neural network, logistic regression, and Random Forest), on such a refined dataset, significantly reduce the misclassification of minor classes. Recently, the SVM method has been increasingly used in biochemistry and drug design researches to generate QSAR models that predict the activity of molecules based on their structural and other properties [20–24]. The Ensemble Bootstrap Aggregation (Bagging) method enhances learner predictive performances by generating multiple models using a random sampling (with replacement) of the training set and selecting random subsets of predictors when creating models. Prediction is performed by ensemble average of all models, and the number of models that “vote” for a result determines the probability of accuracy or confidence of the result.

Taking into consideration the prior researches, this paper proposes a bagging SVM method as a pre-processor of data to generate QSUR models aiming at enhancing their predictive performances. Also, to solve the problem of chemical high-fraction class misclassification highlighted by Isaacs et al., (2016) [13], a multi-class classification was proposed using the LibSVM implementation of the SVM method [25, 26] which applies the one-against-one approach for  $k$ -class learning problems by solving  $\frac{k(k-1)}{2}$  binary SVMs. Hsu & Lin (2002), [27] have shown that the one-against-one approach is better than the one-against-all for large datasets and the training time for this method is shorter.

Continuation of work is organized as follows. The second section describes the data and methods used, while the third section presents and discusses the results obtained. Finally, concluding considerations are given.

## Materials And Methods

This section aims at defining the data and their sources, as well as the predictive methods that will be applied.

### Data

This paper uses two publicly available datasets that were formed in the research described in the previous section.

Using a publicly available database of the chemical ingredients of cosmetics with the function have in a product (CosIng), Isaacs et al. (2016) [13] formed the Functional Use (FUse) database. They have reduced many functions by harmonizing similar functions into a common one. Each chemical in this base has its own unique CASRN code, while one chemical can have multiple functions (in different products). Data on the weight fractions of chemicals in individual cosmetic products were obtained from the Consumer Product Chemical Profile Database (CPCPdb) and on-line product information provided by manufacturers

(Material Safety Data Sheets-MSDS sheets). The products are assigned to the appropriate category, and each chemical within a product has an appropriate harmonized function taken from the FUse database and merged based on a unique CASRN code of the chemical. Thus, they obtained a data set containing in each row a chemical as an ingredient in a product of a certain category with a unique harmonized function and a weight fraction that it has in that category (the unique identifier is the CASRN for the chemical, the product category and the function of the chemical). This dataset is coupled via CASRN to the corresponding physical and chemical properties of the chemicals and to their use in manufacturing. The combined data set includes 828 chemicals and 17,103 of their functional uses (35 harmonized functions) in 4,115 cosmetic products, i.e. 66 categories of these products.

Phillips et al., 2017 [6] expanded the FUse database from previous research with new chemicals and their functions in consumer products, using product composition information from the manufacturers' web pages. The functions are harmonized so that each chemical has a unique function. This set of chemicals is coupled to sets of chemicals that have structural and physical-chemical descriptors via a unique CASRN code. Structural descriptors are publicly available and taken from the EPA's Distributed Structure-Searchable Toxicity -DSSTox database, and physical-chemical descriptors (molecular weight, vapour pressure, water solubility, Henry's Law constant, log of the octanol – air partition coefficient - log(Koa), the log of the octanol – water coefficient - log(Kow), half-life of a chemical in soil, sediment, water, and air, and the persistence of a chemical in the environment) were obtained using the US EPA's Estimation Program Interface (EPI) Suite. Thus, a training set of 4,791 unique chemicals was obtained with 729 structural properties and 11 physical-chemical properties. Retaining only harmonized functions that include at least 10 chemicals, 49 remained in in the training set.

Our first dataset was taken from the FUse database created in the research of Isaacs et al. (2016) [13] covering functional use and weight fraction of chemicals in cosmetic product categories (17 103 functional uses). The data were merged (via a unique CASRN) with structural descriptors (729 descriptors in total) and physical-chemical properties of chemicals (11 descriptors in total) taken from supplementary materials in a study of Phillips et al. (2017) [6]. Structural descriptors are binary (dummy) variables that have a value of 1 if a chemical has a structural property defined by that descriptor (for example, atom.element\_main\_group, atom.element\_metal\_group\_I\_II, bond.CC..O.C\_ketone\_alkane\_cyclic\_.C4., etc.), while if it does not have such a property, the value of the variables is 0. Physical descriptors are numerical variables such as molecular weight, vapour pressure, water solubility, Henry's Law constant, the log of the octanol – air partition coefficient - log(Koa), the log of the octanol – water coefficient - log(Kow), half-life of a chemical in soil, sediment, water, and air, and the persistence of a chemical in the environment.

After merging and eliminating the functions that have less than 10 uses, there are 11,240 functional uses remaining in this dataset, thus representing the final training set (Fuse\_Str\_Pc) (Table A1 in Appendix and in supplementary materials). According to Isaacs et al., (2016) [13] weight fractions were generalized to three categories - low (0.0001-0.01), medium (0.01–0.3) and high (0.3- 1) so as to apply the classification predictive method. In this way, the dataset was prepared for prediction of weight fractions of chemicals in

cosmetic products based on structural and physical-chemical descriptors, functional uses and categories of cosmetic products, using the multi-class SVM bagging method.

The second dataset was taken from an expanded FUse database created in a study by Phillips et al., (2017) [6], which contains data on chemicals and their harmonized functions in consumer products. As in the previous case, the data were merged (via a unique CASRN) with structural and physical-chemical descriptors taken from the same work and functions that included less than 10 chemicals were removed. Thus, a final training set (HFunc\_Str\_Pc) was obtained, comprising 4,665 unique chemicals with 729 structural features, 11 physical-chemical properties and 43 harmonized functions (Table A2 in Appendix and in supplementary materials). This set was used to generate the QSUR model (for prediction of functional uses based on structural and physical-chemical descriptors) by the SVM bagging method.

The following sections shall define the SVM method used and the predictive procedures.

## SVM

To classify linearly non-separable classes, i.e. in case of class overlaps and noise in the data, Vapnik (2010) [28] proposed the SVM method that maps data (it views as  $n$ -dimensional vectors) from original space to feature space, where the classes can be separated using hyper-alignment. Finding such a hyper-alignment minimizes the distance between its end position (so that the gap between classes, i.e. the margins, is as large as possible) and the closest points (support vectors). Instead of an explicit mapping function to a greater-dimension space, it is used a kernel function that allows the calculation of the scalar product of the vector in the original space (kernel trick). Maximizing the margin in a greater-dimension space is reduced to the quadratic programming optimization problem in the original space, using a kernel function. Different kernel functions can be used, but it is often the most efficient and the most widely used Radial Basis Function (RBF) [29]. Training of SVM classifiers is being realized by choosing the optimal values of the gamma parameter for the RBF kernel, and the parameter  $C$  that represents the boundary for the margin i.e. the empty space between the classes. Choosing smaller values for parameter  $C$  reduces over-fitting and increases the generality of the SVM model i.e. its predictive performance.

## Grid-search and k-fold cross-validation

A grid-search technique combined with  $k$ -fold cross-validation is used to select the optimal SVM parameters ( $C$  and  $\gamma$ ). Grid-search sets appropriate rankings and steps for parameters (i.e. defines grid) and then tests their combinations so that the best predictive learner performances are reached. The  $k$ -fold cross-validation process is implemented by dividing the training set into  $k$  subsets, of which  $k-1$  is used to train the model and the one remaining to test the predictive performances of the model using unknown examples. The procedure is iteratively repeated so that each of the  $k$  sub-sets serves as a test set. The final predictive performances are the averages of the model performances obtained in these  $k$  iterations.

## Results And Discussion

In order to predict the functions of chemicals in cosmetic products based on their structural and physical-chemical properties, i.e. to generate QSUR models, on the training set HFunc\_Str\_Pc (numerical data are normalized 0–1 rank transformation), the bagging SVM learners were trained. Model training involves optimizing the C and  $\gamma$  parameters using 5-fold cross-validation so that maximum predictive learner performances are achieved. Using the grid-search technique i.e. the setting appropriate rankings for the parameters, optimal combinations of these parameters were found for the 43 bagging SVM models to predict each of the 43 harmonized functions. Table 1 shows the optimal parameters and predictive performances of the generated models.

Table 1

Predictive performances of the bagging SVM models (5- fold cross-validation, positive class: 1)

Function	optimal parameters	Accur	cl_error	precis	sensit	specif
skin conditioner	SVM.C = 960.4 SVM.gamma = 0.01	95.99%	4.01%	38.08%	29.81%	98.25%
hair dye	SVM.C = 1280.2 SVM.gamma = 0.082	98.78%	1.22%	49.62%	51.06%	99.37%
Buffer	SVM.C = 1280.2 SVM.gamma = 0.082	99.04%	0.96%	39.19%	47.62%	99.42%
antimicrobial	SVM.C = 1600.0 SVM.gamma = 0.046	96.63%	3.37%	56.23%	46.29%	98.60%
hair conditioner	SVM.C = 1600.0 SVM.gamma = 0.064	98.48%	1.52%	54.46%	50.67%	99.28%
catalyst	SVM.C = 320.8 SVM.gamma = 0.064	98.29%	1.71%	72.41%	66.35%	99.23%
preservative	SVM.C = 320.8 SVM.gamma = 0.082	99.40%	0.60%	50.00%	26.00%	99.85%
skin protectant	SVM.C = 640.6 SVM.gamma = 0.01	99.36%	0.64%	15.38%	10.00%	99.76%
flavorant	SVM.C = 320.8 SVM.gamma = 0.046	87.40%	12.60%	50.07%	44.65%	93.57%
flame retardant	SVM.C = 640.6 SVM.gamma = 0.01	99.16%	0.84%	73.32%	71.90%	99.59%
colorant	SVM.C = 320.8 SVM.gamma = 0.064	95.33%	4.67%	71.86%	64.16%	97.91%
masking agent	SVM.C = 1600.0 SVM.gamma = 0.1	97.02%	2.98%	29.34%	22.19%	98.73%
antioxidant	SVM.C = 960.4 SVM.gamma = 0.082	97.68%	2.32%	37.13%	29.61%	98.99%
fragrance	SVM.C = 320.8 SVM.gamma = 0.01	81.63%	18.37%	69.58%	75.13%	84.66%
additive	SVM.C = 960.4 SVM.gamma = 0.064	99.38%	0.62%	13.33%	15.00%	99.68%
surfactant	SVM.C = 640.6 SVM.gamma = 0.028	98.26%	1.74%	73.49%	71.22%	99.14%
heat stabilizer	SVM.C = 1600.0 SVM.gamma = 0.01	99.38%	0.62%	18.75%	18.33%	99.72%
reducer	SVM.C = 320.8 SVM.gamma = 0.01	99.83%	0.17%	85.33%	68.33%	99.94%



Function	optimal parameters	Accur	cl_error	precis	sensit	specif
emulsion stabilizer	SVM.C = 1280.2 SVM.gamma = 0.1	99.64%	0.36%	36.67%	40.00%	99.83%
ubiquitous	SVM.C = 960.4 SVM.gamma = 0.046	96.46%	3.54%	22.68%	18.66%	98.40%
chelator	SVM.C = 320.8 SVM.gamma = 0.064	99.55%	0.45%	47.86%	33.33%	99.76%
plasticizer	SVM.C = 1600.0 SVM.gamma = 0.064	99.19%	0.81%	34.67%	42.00%	99.55%
solvent	SVM.C = 1280.2 SVM.gamma = 0.046	98.52%	1.48%	27.27%	23.11%	99.31%
Vinyl	SVM.C = 960.4 SVM.gamma = 0.046	99.98%	0.02%	96.00%	100.00%	99.98%
monomer	SVM.C = 640.6 SVM.gamma = 0.046	98.80%	1.20%	70.24%	69.15%	99.37%
viscosity controlling agent	SVM.C = 1600.0 SVM.gamma = 0.046	99.27%	0.73%	20.00%	12.00%	99.74%
Emollient	SVM.C = 1280.2 SVM.gamma = 0.01	97.73%	2.27%	59.32%	57.14%	98.85%
Foamer	SVM.C = 320.8 SVM.gamma = 0.01	99.72%	0.28%	28.57%	20.00%	99.89%
Crosslinker	SVM.C = 1280.2 SVM.gamma = 0.01	97.11%	2.89%	51.34%	46.06%	98.65%
film forming agent	SVM.C = 640.6 SVM.gamma = 0.1	98.91%	1.09%	47.59%	45.56%	99.44%
UV absorber	SVM.C = 640.6 SVM.gamma = 0.01	98.71%	1.29%	59.42%	53.63%	99.39%
Perfumer	SVM.C = 1600.0 SVM.gamma = 0.046	96.40%	3.60%	15.17%	12.57%	98.29%
additive for liquid system	SVM.C = 320.8 SVM.gamma = 0.046	99.72%	0.28%	20.00%	10.00%	99.91%
wetting agent	SVM.C = 960.4 SVM.gamma = 0.028	99.51%	0.49%	50.00%	53.00%	99.74%
rheology modifier	SVM.C = 960.4 SVM.gamma = 0.046	99.61%	0.39%	14.29%	10.00%	99.87%
adhesion promoter	SVM.C = 640.6 SVM.gamma = 0.1	99.38%	0.62%	36.67%	24.00%	99.72%
organic pigment	SVM.C = 320.8 SVM.gamma = 0.028	99.01%	0.99%	77.79%	74.74%	99.52%

Function	optimal parameters	Accur	cl_error	precis	sensit	specif
soluble dye	SVM.C = 320.8 SVM.gamma = 0.046	99.66%	0.34%	66.67%	46.67%	99.87%
Photoinitiator	SVM.C = 1600.0 SVM.gamma = 0.01	99.74%	0.26%	61.54%	53.33%	99.89%
foam boosting agent	SVM.C = 640.6 SVM.gamma = 0.01	99.87%	0.13%	83.00%	90.00%	99.91%
Whitener	SVM.C = 320.8 SVM.gamma = 0.1	99.91%	0.09%	100.00%	60.00%	100.00%
antistatic agent	SVM.C = 960.4 SVM.gamma = 0.082	99.72%	0.28%	67.33%	65.00%	99.87%
Emulsifier	SVM.C = 1280.2 SVM.gamma = 0.028	99.68%	0.32%	30.00%	23.33%	99.85%
Average		98.07%	1.93%	49.34%	43.99%	98.94%

As can be seen from the Table 1, the accuracy of the model based on 5-fold cross-validation ranges from 81.63% to 99.98%. All models are valid in contrast to the results obtained by Phillips et al., (2017) [6] where 8 of the 46 models had a balanced accuracy of less than 75%. For these very important functions such as masking agent, solvent, viscosity, controlling agent and perfumer, models with accuracy greater than 96.40% were obtained. However, it should be taken into consideration that this is not a balanced accuracy, so much of the accuracy falls on the major (negative) class (true negative rate i.e. sensitivity averages 98.94%, while true positive rate i.e. specificity averages 43.99%). The lower sensitivity of the models indicates that the models are not overfitted, i.e. too dependent on the training set for the positive class, which gives them the potential for better performance on an unknown dataset. A great deal of model specificity is present because the negative class is much larger than the positive class. The small precision of the positive class, which averages 13.73%, indicates that a number of examples of the negative class (chemicals that have no function) are the least distant from the positive class (chemicals most similar in structure and physical-chemical properties to chemicals that have this function) by an SVM qualifier associated with a positive (minor) class.

Applying the resulting bagging SVM models to the training set, a prediction was generated for 43 functions. Each bagging prediction generates multiple SVM models and accordingly each bagging prediction has its own probability. Predictions with high probability mean that the greatest number of SVM models thus generated voted for a chemical to have/does not have a certain function.

The goal now is to increase the precision and sensitivity of the model (precision of a positive class and true positive rate) by taking from a large number of negative class representatives only those chemicals that are farthest from the chemicals belonging to the positive class (those having the structure and properties differing to the greatest extent). Those members of the negative class to be selected were

determined by bagging SVM through assigning them the highest probability of belonging to the negative class.

Therefore, the next step is to determine the optimal threshold (minimum likelihood) of  $Pr$ , for predictions that will be accepted. For this purpose, the DT method with the gini\_index measure for partitioning [28] and 5-fold cross validation were used. Specifically, the optimal threshold for the probability of an SVM prediction was determined for each of the 43 models as follows. A number of chemicals whose major class prediction was less than the  $Pr$  value were excluded from the training set and 5-fold DT predictive performances were tested. The parameter  $Pr$  is determined to obtain the maximum predictive performance of the DT learner. Table 2 shows the values of this parameter as well as the predictive performances of the QSUR models thus obtained for each of the 43 functions.

Table 2  
 Predictive performance of DT models on bagging SVM output (5- fold cross-validation, positive class: 1)

Function	accur	cl_err	precis	sensit	specif	Pr
skin conditioner	92.13%	7.87%	93.89%	92.30%	91.84%	> 0.98
hair dye	99.57%	0.43%	92.18%	80.27%	99.88%	> 0.98
Buffer	99.25%	0.75%	90.95%	77.62%	99.77%	> 0.98
antimicrobial	89.57%	10.43%	90.75%	91.03%	87.69%	> 0.98
hair conditioner	91.93%	8.07%	80.30%	73.71%	95.99%	> 0.94
catalyst	92.10%	7.90%	86.37%	87.19%	94.01%	> 0.96
preservative	92.75%	7.25%	100.00%	78.00%	100.00%	> 0.95
skin protectant	90.29%	9.71%	81.00%	71.67%	94.48%	> 0.99
flavorant	90.80%	9.20%	84.67%	88.78%	91.82%	> 0.95
flame retardant	99.25%	0.75%	88.66%	86.92%	99.64%	> 0.95
colorant	92.37%	7.63%	93.18%	92.85%	91.77%	> 0.95
masking agent	91.69%	8.31%	78.00%	73.07%	95.54%	> 0.97
antioxidant	96.07%	3.93%	88.79%	75.59%	98.73%	> 0.97
fragrance	94.06%	5.94%	91.94%	95.41%	92.92%	> 0.8
additive	99.72%	0.28%	100.00%	73.33%	100.00%	> 0.98
surfactant	99.02%	0.98%	96.42%	92.06%	99.67%	> 0.98
heat stabilizer	99.24%	0.76%	83.33%	76.67%	99.61%	> 0.98
reducer	99.90%	0.10%	90.00%	76.67%	99.97%	> 0.95
emulsion stabilizer	95.46%	4.54%	88.33%	66.67%	98.57%	> 0.95
ubiquitous	89.01%	10.99%	86.50%	78.71%	93.49%	> 0.99
chelator	98.19%	1.81%	80.00%	70.00%	99.25%	> 0.97
plasticizer	97.97%	2.03%	100.00%	71.00%	100.00%	> 0.98
solvent	97.04%	2.96%	80.17%	84.56%	98.08%	> 0.995
Vinyl	99.98%	0.02%	96.00%	100.00%	99.98%	> 0.5
monomer	99.25%	0.75%	92.08%	88.73%	99.66%	> 0.95
viscosity controlling agent	91.77%	8.23%	83.50%	72.00%	95.49%	> 0.99

Function	accur	cl_err	precis	sensit	specif	Pr
emollient	98.29%	1.71%	91.20%	82.92%	99.41%	> 0.985
foamer	97.22%	2.78%	83.33%	70.00%	99.26%	> 0.982
crosslinker	91.44%	8.56%	89.31%	85.88%	94.32%	> 0.99
film forming agent	94.27%	5.73%	87.14%	76.39%	97.72%	> 0.99
UV absorber	92.69%	7.31%	91.67%	79.23%	97.07%	> 0.99
perfumer	92.14%	7.86%	81.43%	79.16%	95.39%	> 0.99
additive for liquid system	95.29%	4.71%	76.67%	90.00%	96.25%	> 0.98
wetting agent	99.64%	0.36%	96.00%	69.00%	99.95%	> 0.98
rheology modifier	99.15%	0.85%	85.71%	53.33%	99.83%	> 0.98
adhesion promoter	97.65%	2.35%	93.33%	81.67%	98.97%	> 0.93
organic pigment	99.52%	0.48%	91.79%	89.30%	99.78%	> 0.9
soluble dye	96.77%	3.23%	79.33%	70.00%	98.15%	> 0.95
photoinitiator	99.42%	0.58%	78.33%	80.00%	99.71%	> 0.99
foam boosting agent	99.95%	0.05%	100.00%	86.67%	100.00%	> 0.85
whitener	99.18%	0.82%	93.33%	90.00%	99.57%	> 0.85
antistatic agent	94.85%	5.15%	89.47%	93.33%	95.00%	> 0.93
emulsifier	99.86%	0.14%	80.00%	72.73%	99.94%	> 0.99
Average	95.95%	4.05%	88.49%	80.57%	97.40%	

It can be seen from the Table 2 that the average accuracy of the final QSUR models is 95.95%, the precision averages 88.49%, while the average sensitivity and specificity are 80.57% and 97.40%, respectively. Thus, after removing the example of a negative class whose probability of belonging to the class is less than the Pr threshold from the training set, precision increased on average by 74.76%, sensitivity on average by 36.58%, while the sensitivity decreased on average by 1.54%.

In the study of Phillips et al., (2017) [6] for 49 harmonized function the 49 balanced random forest models were generated, of which 41 were valid (with balanced accuracy of >75%). For 8 functions (which include some of the most important ones such as perfumer and solvent) no valid models were obtained i.e. random balanced undersampling was not a method effective enough to predict significant differences in the structure and physical-chemical properties of these chemicals compared to the others. Most models have well recognized the chemicals that make up the positive class in the training set (sensitivity

models average about 85%), however the average precision (positive classes) is only about 14%, which means that the predictive power of the model is weak. This is due to the large number of false positives (chemicals that do not have a specific function but are misclassified by the model as having them). To identify the chemicals that could be functional substitutes, the generated models were applied to 6,356 chemicals in the Tox21 library for which they are structural and physical-chemical descriptors available. Consistent with the small precision of the positive class of obtained models, about 88% of the predictions were invalid (with a probability of less than 50%).

Comparing the performances of the QSUR models thus obtained with the results from the confusion matrix obtained by Phillips et al., (2017) [6] (Table A3 below) [6], it can be seen that their accuracy averaged 91.81%, sensitivity 84.62%, specificity 91.83%, and precision 13.73%. The precision of our QSUR models is significantly improved over theirs, while the other 3 indicators were similar. The foregoing could lead to a conclusion that the predictive potential of our QSUR models for the positive class (chemicals having some function) is increased.

Bagging SVM prediction did the pre-processing of data for the DT learner and significantly improved its predictive performances even in the case of strong class imbalance. Instead of random undersampling used in the study by Phillips et al., (2017) [6] to ensure class balancing, an undersampling of the major class was made here based on bagging SVM prediction, i.e. only prediction with a higher probability of belonging to the major class was included in the sample. Thus, valid models were obtained for all 43 features with high predictive performances.

Based on the generated DT models, it can be concluded which structural and physical-chemical properties are most responsible for distinguishing the chemicals belonging to the positive class from the others. Thus, chemicals that have these properties can be identified and tested for potential substitutes. For example, the rules for the positive class derived from the DT model for the fragmentation function in Table 3 show the structural and physicochemical properties that should satisfy the potential substitutes for this function in cosmetic products.

Table 3  
Positive class rules derived from DT models for the *fragrance* function

DT rules	DT rules
Rule 1: 1 {0 = 2, 1 = 14}, Accuracy = 87.50%	Rule 6: 1 {0 = 0, 1 = 20} Accuracy = 100%
logKoa_unitless > 0.199	logKoa_unitless ≤ 0.199
logKoa_unitless > 0.229	logP_unitless > 0.375
bond.C..O.O_carboxylicEster_aromatic = true	bond.X.any._halide = false
logKoa_unitless ≤ 0.286	atom.element_metal_metalloid = false
Rule 2: 1 {0 = 0, 1 = 6}, Accuracy = 100%	bond.OZ_oxide_peroxy = false
logKoa_unitless > 0.199	bond.CN_amine_aliphatic_generic = false
logKoa_unitless ≤ 0.229	bond.C..O.O_carboxylicEster_alkenyl = true
chain.aromaticAlkane_Ph.C1_acyclic_generic = false	chain.alkeneLinear_diene_1_2.butene = true
bond.COH_alcohol_ter.alkyl = false	Rule 7: 1 {0 = 2, 1 = 10} Accuracy = 83.33%
chain.alkaneCyclic_ethyl_C2_connect_noZ. = true	logKoa_unitless ≤ 0.199
ring.hetero_.5._N_pyrrole_generic = false	logP_unitless ≤ 0.375
persistence_units(hr) ≤ 0.084	bond.C..O.O_carboxylicEster_alkyl = false
Rule 3: 1 {0 = 0, 1 = 8}, Accuracy = 100%	bond.CC..O.C_ketone_generic = false
logKoa_unitless > 0.199	ring.hetero_.5._O_oxolane = false
logKoa_unitless ≤ 0.229	logP_unitless > 0.359
chain.aromaticAlkane_Ph.C1_acyclic_generic = false	persistence_units(hr) ≤ 0.020
bond.COH_alcohol_ter.alkyl = true	chain.alkeneLinear_mono.ene_ehtylene_terminal = false
Rule 4: 1 {0 = 4, 1 = 45}, Accuracy = 91.83%	vapor_pressure_units(Pa) > 0.000
logKoa_unitless > 0.199	Rule 8: 1 {0 = 2, 1 = 25}, Accuracy = 92.59%
logKoa_unitless ≤ 0.229	logKoa_unitless ≤ 0.199
chain.aromaticAlkane_Ph.C1_acyclic_generic = true	logP_unitless ≤ 0.375
bond.C..O.O_carboxylicAcid_generic = false	bond.C..O.O_carboxylicEster_alkyl = false
water_solubility_units(mg/L) ≤ 0.129	bond.CC..O.C_ketone_generic = true

DT rules	DT rules
bond.X.any._halide = false	molecular_weight ≤ 0.113
bond.CN_amine_aliphatic_generic = false	bond.C..O.O_carboxylicAcid_alkyl = false
Rule 5: 1 {0 = 60, 1 = 1347}, Accuracy = 95.73%	chain.alkeneCyclic_ethylene_C_connect_noZ. = false
logKoa_unitless ≤ 0.199	logP_unitless > 0.335
logP_unitless > 0.375	Rule 9: 1 {0 = 0, 1 = 24}, Accuracy = 100%
bond.X.any._halide = false	logKoa_unitless ≤ 0.199
atom.element_metal_metalloid = false	logP_unitless ≤ 0.375
bond.OZ_oxide_peroxy = false	bond.C..O.O_carboxylicEster_alkyl = true
bond.CN_amine_aliphatic_generic = false	air_half_life_units(hr) > 0.000
bond.C..O.O_carboxylicEster_alkenyl = false	
bond.CS_sulfide = false	
molecular_weight ≤ 0.229	

Rule 5 is the most important as it covers the largest number (1347) of positive examples with an accuracy of 95.73%. There are a number of studies/reports stating that 95% of the chemicals used in cosmetic products as fragrance components are of synthetic origin, derived from petroleum. The odour components either of synthetic or of natural origin are the other allergens in frequency of causing reactions, accounting for 12.5% of all reactions [30]. The chemicals can cause symptoms such as: respiratory irritation, increased asthma, allergic reactions, mucosal irritations, migraines, headaches, skin problem, cognitive problem, gastrointestinal problem, contact dermatitis, urticaria, photosensitivity. Some of these chemicals are lylal (synthetic lily scent), nitro and polycyclic musks, amyl cinnamal (usually of synthetic origin, though it may be of natural origin, having a floral jasmine-like scent), etc. These compounds are widely used as fragrances in various personal care products such as cosmetics and perfumery. In recent years, a large number of preparations have been marketed, which are labelled as odourless preparations, with the presence of vegetable ingredients or oils. Concealed allergens include rose oil, vanilla and sweet almond oil. Lilal - Butyl-phenyl-methyl-propanal causes contact dermatitis and is often found as a fragrant ingredient in perfumes, shampoos, bath preparations and lotions [31].

It is most commonly obtained synthetically via cross-aldol condensation between para-terc-butylbenzaldehyde and propanal, followed by hydrogenation of the intermediate alkene. It is the clear, viscous liquid with a strong floral scent. In addition to causing contact dermatitis, it is suspected to have an effect on the endocrine system and estrogen activity [31]. Citronellol is a colourless oily liquid with a floral scent on rose which is used in cleansers, hair care products, lipsticks, perfumes. It is a known skin allergen, causes eczema, and often causes complications in people with psoriasis [32]. Nitro- and



polycyclic musks are two common and important synthetic musks currently in use [33]. In addition, due to their strong photochemical toxicity, [34] carcinogenicity [35] and neurotoxic properties, as well as endocrine dysfunction, nitro-musks (e.g. musk xylene), their use is being monitored in Japan, in the EU they are also under scrutiny, and further research is being conducted on their potential adverse effects on human health.

The process of high-throughput screening of a set of unknown chemicals using the generated QSUR models would consist of the following steps:

1. Prediction of the chemical function using the bagging SVM model
2. Elimination of chemicals whose non-function is determined with probability less than  $P_r$
3. Application of the DT model to predict function on the purified set of chemicals from step 2
4. For each chemical, each of the 41 QSUR models will generate a result
5. The result that has the highest confidence will determine the function for which that chemical can be substitute
6. If two or more results have the same confidence, then the result obtained by the model with the highest precision of positive class will be the winning one

The next step is to generate a model for the prediction of weight fractions, on the training set Fuse\_Str\_Pc, in which the nominal variables are transformed into numerical dummy variables, while the numerical are normalized by a 0–1 rank transformation. By training bagging multi-class SVM learner, i.e. by applying grid-search procedure in combination with 5-fold cross-validation, the optimal combination of parameters SVM.C = 750.0 and SVM.gamma = 0.1 was obtained and the following predictive performance was achieved: accuracy- 88.47%, classification error- 11.53%, mean recall – 83.44% and mean precision – 85.82%.

For the purpose of predicting the fractions of chemicals in the product category, Isaacs et al. (2016) [13] generated a classification model, by generalizing fractions into three categories - low, medium, and high fractions. They used structural and physical-chemical properties of chemicals as well as their functional uses as predictors. Using the random forest method, a model with a 5-fold cross-validation error of 16.7% was generated. The potential for misclassification of the obtained model is highest for high fractions (about 22%), while for low and medium fractions the class precision is about 84%. Accordingly, applying the model to an unknown dataset, less than 1% of chemicals are predicted to have high fractions (30% -100% of total weight), while 35% and 65% of chemicals are predicted to have medium and low fractions. Therefore, the predictive performances of the model for the high-fraction chemicals that make up the majority of cosmetic product composition are not satisfactory.

With our model the precision error for the high class is 15%, which is a better result than the one achieved by Isaacs et al. (2016) [13] having amounted to 22%, and the 5-fold classification error of the model is decreased by about 5%. This indicates that the multi-class bagging SVM model has a better predictive

potential for a class of chemicals with a high participation in cosmetics than the random forest model used in the study mentioned.

As with the QSUR model, by generating a DT model on bagging SVM output, taking high Pr results (those voted for by the great number of SVM models obtained by bagging) for each class should further improve the prediction performances. However, since most of the results on the training set had a Pr greater than 0.99 (meaning that 99% of the generated SVM models voted for that result), the DT on the refined dataset had almost the same predictive performance as the SVM.

Nevertheless, the DT model generated on SVM output is useful as it provides explicit classification rules for low, medium and high fractions from which it can be concluded what are the chemical properties that cosmetic products contain to the greatest extent. Table 4 shows some of the most significant rules generated by DT model for all three classes (with the highest accuracy and cover).

Table 4  
Rules derived from DT models for weight fractions

DT rules	DT rules
Rule 1: Medium {Low = 135, Medium = 4193, High = 60}, Acc = 95.55%	Rule 4: High {Low = 0, Medium = 0, High = 114}, Acc = 100%
logKoa_unitless > 0.086	logKoa_unitless > 0.086
bond.S..O.O_sulfonate = false	bond.S..O.O_sulfonate = false
bond.C.O_carbonyl_generic = false	bond.C.O_carbonyl_generic = true
chain.aromaticAlkane_Ar.C_meta = false	logP_unitless ≤ 0.523
logP_unitless > 0.352	bond.C..O.O_carboxylicEster_acyclic = false
chain.alkaneBranch_t.butyl_C4 = false	air_half_life_units(hr) > 0.001
henrys_law_constant_units(atm-m3/mol) ≤ 0.000	Rule 5: Low {Low = 595, Medium = 167, High = 0}, Acc = 78.08%
bond.X.any_halide = false	logKoa_unitless > 0.086
Rule 2: Low {Low = 174, Medium = 22, High = 1}, Acc = 88.32	bond.S..O.O_sulfonate = true
logKoa_unitless > 0.086	Rule 6: High {Low = 12, Medium = 209, High = 830}, Acc = 78.97%
bond.S..O.O_sulfonate = false	logKoa_unitless ≤ 0.086
bond.C.O_carbonyl_generic = false	logP_unitless ≤ 0.466
chain.aromaticAlkane_Ar.C_meta = true	bond.C..O.O_carboxylicAcid_alkyl = false
Rule 3: Medium {Low = 24, Medium = 1301, High = 16}, Acc = 97.01%	
logKoa_unitless > 0.086	
bond.S..O.O_sulfonate = false	
bond.C.O_carbonyl_generic = true	
logP_unitless > 0.523	
bond.CC..O.C_ketone_alkene_cyclic_2.en.1.one_generic = false	
logP_unitless ≤ 0.685	
bond.C.O_carbonyl_1_2.di = false	

Thus, for example, it can be concluded from Table 4, based on Rule 3, that cosmetic formulations can contain up to 30% of chemicals whose octanol-water partitioning coefficient is between 0.523 and 0.685. This means that water pollutants can be significant due to poor water solubility. Also, this rule shows that cosmetic products can contain up to 30% of chemicals with a carbonyl group to which they belong and

some that are dangerous to human health. A substance that has a carbonyl group in it and is a common ingredient in various cosmetic products including liquid soaps, shampoos and shower creams/lotions is formaldehyde [36, 37]. According to the International Agency for Research on Cancer, formaldehyde belongs to a group of human carcinogens because there is enough evidence that it causes cancer in humans. This fact is based on the fact that formaldehyde can lead to nasopharyngeal cancer in humans after inhalation and to squamous cell carcinoma of the nasal passages in rats [38]. This is why formaldehyde and paraformaldehyde are used as preservatives at concentrations of up to 0.1% in products used in cosmetics for oral hygiene (not to be used in aerosol products) and up to 0.2% in other products [39].

The procedure for predicting the unknown fraction of a chemical in a cosmetic product using the generated multi-class SVM model implies that the model input provides information on the product category, the function the chemical has in the product, the structural descriptors of the chemical and its physicochemical properties. Based on this information, the model will predict whether the chemical in the specified category is represented by a low, medium or high weight fraction.

## Conclusions

An assessment of the toxicity exposure of chemicals in consumer products involves knowledge of the qualitative and quantitative composition of these products. Namely, on the basis of knowledge of the structural properties and amount of chemicals used in the product, the negative impact of the product on the consumer and the environment can be assessed.

This paper proposes methods that, based on available information on the functional and quantitative use of chemicals in thousands of real consumer products, generate predictive chemical classification models based on the function and weight fractions that chemicals have in cosmetic products. With these models, the composition of products with unavailable information can be assumed. These methods clearly define the approach by which great libraries of chemicals can be screened to identify potential substitutes for toxic chemicals without impairing the functions that the original chemicals have in the product. Also, a clear procedure is defined on the manner in which the weight fraction of a chemical in a cosmetic product can be estimated using the generated predictive model. Thus, one can implicitly find out the chemical composition of cosmetic products that is inaccurate or completely inaccessible for many products.

The research results show that the proposed bagging SVM method can overcome the disadvantages of previously applied methods, i.e. increase the precision of prediction.

The proposed methods can help address the lack of information needed to assess exposure to risk from the use of cosmetic products containing toxic chemicals in their composition.

## Declarations

## Funding:

This research received no external funding.

## Author Contributions:

"Conceptualization, T.V. and Z.K.; Methodology, Lj.K. and T.V.; Software, Lj.K.; Validation, Z.P. and T.V.; Formal Analysis, T.V.; Investigation, Z.K.; Resources, N.R.; Data Curation, Z.K.; Writing – Original Draft Preparation, T.V.; Writing – Review & Editing, Z.P. and N.R.; Supervision, Z.P.; Project Administration, N.R.

## Conflicts of Interest:

The authors declare no conflict of interest.

The study does not include human participants and / or animals.

No Informed Consent.

## References

1. Fisher AA (1975) Contact dermatitis. Lea and Febinger, Philadelphia
2. Stüttgen G (1981) Benefit and Risk der kosmetischen Mittel. J Soc Cosmet Chem 32, 231
3. Kaminester LH (1981) Alergic reaction to sunscreen products. Arch Dermatol 117, 66
4. Štivić I, Čajkovac M (1972) O nekim pojavama nepoželjnog i štetnog djelovanja kozmetika na kožu. Farm Glas 28, 341
5. Tickner JA, Schifano JN, Blake A, Rudisill C, Mulvihill MJ (2015) Advancing safer alternatives through functional substitution. Environ Sci Technol 49:742-749
6. Phillips KA, Wambaugh JF, Grulke CM, Dionisio KL, Isaacs KK (2017) High-throughput screening of chemicals as functional substitutes using structure-based classification models. Green Chem 19:1063-1074
7. Delmaar C, Bokkers B, Burg W, Schuur G (2015) Validation of an aggregate exposure model for substances in consumer products: a case study of diethylphthalate in personal care products. J Expo Sci Environ Epidemiol 25:317–323
8. Consumer Exposure Model (2015) United States Environmental Protection Agency. [http://www.epa.gov/sites/production/files/2015-09/documents/cemuser\\_guide\\_beta\\_test.pdf](http://www.epa.gov/sites/production/files/2015-09/documents/cemuser_guide_beta_test.pdf)
9. Egeghy PP, Judson R, Gangwal S, Mosher S, Smith D, Vail J, CohenHubal EA (2012) The exposure data landscape for manufactured chemicals. Sci Total Environ 414 :159–166
10. Wambaugh JF, Setzer RW, Reif DM, Gangwal S, Mitchell-Blackwood J, Arnot JA, Judson RS (2013) High-throughput models for exposure-based chemical prioritization in the ExpoCast project. Environ Sci Technol 47: 8479-8488
11. U. S. Environmental Protection Agency. Safer Choice. <https://www.epa.gov/saferchoice>

12. U. S. Environmental Protection Agency. Program for Assisting the Replacement of Industrial Solvents. Apr, 2016
13. Isaacs KK, Goldsmith MR, Egeghy P, Phillips K, Brooks R, Hong T, Wambaugh JF (2016) Characterization and prediction of chemical functions and weight fractions in consumer products. *Toxicol Rep* 3:723-732
14. Farquad M, Bose I (2012) Preprocessing unbalanced data using support vector machine. *Decision Support Systems* 53: 226–233
15. Diederich J (2008) *Rule Extraction from Support Vector Machines: An Introduction*. Springer-Verlag, Berlin Heidelberg
16. Barakat N, Bradley AP (2010) Rule extraction from support vector machines: A review. *Neurocomputing* 74:178–190
17. Martens D, Baesens B, Gestel TV, Vanthienen J (2007) Comprehensible credit scoring models using rule extraction from support vector machines. *Eur J Oper Res* 183:1466-1476
18. Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B (2008) Rule Extraction from Support Vector Machines: An Overview of Issues and Application in Credit Scoring. *Rule Extraction from Support Vector Machines Studies in Computational Intelligence*. Springer-Verlag Berlin Heidelberg
19. Rogic S, Kascelan L (2019) Customer Value Prediction in Direct Marketing Using Hybrid Support Vector Machine Rule Extraction Method. In *European Conference on Advances in Databases and Information Systems*. Springer, Cham 283- 294
20. Byvatov E, Schneider G (2004) SVM-based feature selection for characterization of focused compound collections. *J Chem Inf Comput Sci* 44:993-999
21. Briem H, Günther J (2005) Classifying “kinase inhibitor-likeness” by using machine-learning methods. *Chembiochem* 6:558-566
22. Doucet JP, Barbault F, Xia H, Panaye A, Fan B (2007) Nonlinear SVM approaches to QSPR/QSAR studies and drug design. *Current Computer-Aided Drug Design*.3:263-289
23. Heikamp K, Bajorath J (2014) Support vector machines for drug discovery. *Exprt Opin Drug Discov* 9:93-104
24. Maltarollo VG, Kronenberger T, Espinoza GZ, Oliveira PR, Honorio KM (2019) Advances with support vector machines for novel drug discovery. *Expert Opin Drug Discov* 14:23-33
25. Fan RE, Chen PH, Lin CJ (2005) Working set selection using second order information for training support vector machines. *J Mach Learn Res* 6:1889-1918
26. Chang CC (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2:1-27
27. Hsu CW, Lin CJ (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw*13: 415-425
28. Vapnik VN (2010) *The nature of statistical learning theory*. Springer, New York

29. Regulation (EC) No 1223/2009 of the European Parliament and of the Council. Cosmetic Products Official Journal of the European Union 30 November 2009
30. Sanderson M, Christopher D M, Prabhakar R, Hinrich S (2010) Introduction to Information Retrieval Cambridge University Press. 2008: Natural Language Engineering 16:100-103
31. Breiman L (1984) Classification and regression trees. Wadsworth International Group, Belmont CA
32. Čajkovac M (2005) Kozmetologija. Naklada Slap, Zagreb
33. Usta , Hachem Y, El-Rifai O, Bou-Moughlabey Y, Echtay K, Griffiths D, Nakkash-Chmaisse H, Makki RF (2013) Liral Fragrance chemicals lylal and lial decrease viability of HaCat cells; by increasing free radical production and lowering intracellular ATP level: protection by antioxidants. Toxicol In Vitro 27:339-48
34. Rudbaeck J, Hagvall L, Boerje A, Nilsson U, Karlberg AT (2014) Characterization of skin sensitizers from autoxidized citronellol- impact of the terpene structure on the autoxidation process. Contact Dermat 70:329-339
35. Hopkins ZR, Blaney L (2016) An aggregate analysis of personal care products in the environment: identifying the distribution of environmentally-relevant concentrations. Environ Int 92-93, 301–316
36. Karschuk N, Tepe Y, Gerlach S, Pape W, Wenck H, Schmucker R, Wittern KP, Schepky A, Reuter HA (2010) Novel in vitro method for the detection and characterization of photosensitizers. PLoS One, 5
37. Zhang Y, Huang L, Zhao Y, Hu T (2017) Musk xylene induces malignant transformation of human liver cell line L02 via repressing the TGF-beta signaling pathway. Chemosph 168:1506–1514
38. Lv C, Hou J, Xie W, Cheng H (2015) Investigation on formaldehyde release from preservatives in cosmetics. Int J Cosmet Sci 37:474- 478
39. Halla N, Fernandes IP, Heleno SA, Costa P, Boucherit-Otmani Z, Boucherit K, Rodrigues AE, Ferreira ICFR, Barreiro MF (2018) Cosmetics Preservation: A Review on Present Strategies. Molecules 23: 1571

## Supplementary Files

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

- [Appendix.docx](#)