Artificial intelligence model for the prediction of cleansing foam formulations with excellent make-up removability
～Is an "in silico formulator" superior to a human formulator?～

Masugu Hamaguchi[1*]; Hideki Miwake[2]; Ryouichi Nakatake[2]; and **Noriyoshi Arai**[3];
[1] Kirin Central Research Institute, Kirin Holdings, 26-1, Muraoka-Higashi 2-chome. Fujisawa, Kanagawa 251-8555, Japan;
[2] Research Institute, Fancl Corporation, 12-13 Kamishinano, Totsuka-ku, Yokohama, Kanagawa 244-0806, Japan;
[3] Department of Mechanical Engineering, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

* Masugu Hamaguchi, Kirin Central Research Institute, Kirin Holdings, 26-1, Muraoka-Higashi 2-chome. Fujisawa, Kanagawa 251-8555, Japan, +81-80-1123-3846, masugu_hamaguchi@kirin.co.jp

Abstract (Maximum of 250 words)

Cleansing foams can contain ingredients in an infinite number of combinations, which renders formulation optimization difficult. In this study, we used artificial intelligence (AI) with machine learning to build a cleansing capability prediction system that incorporates the effects of surfactant self-assembly and chemical characteristics of the ingredients. On one hand, more than 500 cleansing foam samples were prepared and tested. On the other hand, we applied molecular descriptors and Hansen solubility index for estimation of cleansing capability for each formulation set. Five machine learning models were applied to predict the cleansing capability. We also used an in-silico formulation, which produces formulations virtually with PC and predicts the cleansing capabilities with the established AI model.

An accuracy of R^2=0.765 was obtained. We observed that mixtures of cosmetic ingredients demonstrated interactions among each other, and this type of non-linear behavior increased the difficulty of predicting the cleansing performance. However, the use of descriptors of chemical characteristics ensured high accuracy, and with the assistant of the in-silico formulation, we identified a cleansing foam formulation consisting of eicosaglycerol hexacaprylate and PPG-9 diglyceryl ether or cyclohexylglycerin that exhibited a high cleansing capability of >85% for the removal of waterproof eyeliner. We expect that our

system will help significantly reduce the effort required for the development of new and effective cosmetics.

Keywords: AI; machine learning; cleansing capability; formulation; cleansing foam

Introduction.

Cleansing foam is used to wash excess sebum and dirt from skin and make-up remover is used to remove make-up cosmetics. Recently, from the viewpoint of shortening time (reducing procedures) and eco-friendliness (saving water and reducing the release of chemical substances into environment), there is an increasing need for a single product to serve as cleansing foam and make-up remover within one product[1].

Solvent-based cleansing agents such as make-up remover oils exhibit high solubility to the makeup products, which themselves are made of oil and pigments, resulting in excellent removability. However, problems are associated with solvent-based cleansing agents such as high environmental loads and material costs, in addition to the feeling of residual oiliness after rinsing[2]. In contrast, surfactant-based cleansing agents such as cleansing foams have excellent rinsing properties but weak oil removability, because they are mainly water-based. In this study, the latter approach was adopted to improve the cleansing performance of cleansing foams.

Cleansing foams are composed of numerous ingredients, which makes formulation optimization difficult as an infinite number of ingredient combinations are possible. Therefore, artificial intelligence (AI) using machine learning was introduced into the formulation design to construct a cleansing capability prediction system that considers the effects of surfactant self-assembly and chemical characteristics of ingredients. Moreover, in-silico simulations was introduced in order to assist human formulators and obtain desirable products in a product development process.

Materials and Methods.

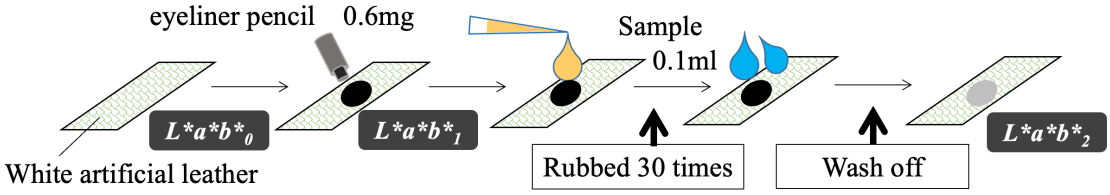
1. Evaluation of cleasing capability

Cleansing foam samples (more than 500 formulations), consisting of ionic surfactants, amphoteric surfactants, nonionic surfactants, polyols, a pH adjuster, and water, were prepared by mixing thoroughly by heating while mixing thoroughly and stirring. Some examples of ingredients are listed in Table 1. To test the prepared samples, first, a waterproof eyeliner pencil on a piece of white artificial leather, which was dried for 30 min. Then, 0.1 mL of the corresponding cleansing foam sample was added on the dried eyeliner, which was rubbed 30 times, and rinsed, and dried. A schematic of all procedures is shown in Fig. 1.

The cleansing capability was evaluated using the eyeliner-pencil residual ratio, calculated by color differences as follows:

$$\text{Cleansing capability (\%)} = \frac{\sqrt{(L_1^*-L_2^*)^2+(a_1^*-a_2^*)^2+(b_1^*-b_2^*)^2}}{\sqrt{(L_1^*-L_0^*)^2+(a_1^*-a_0^*)^2+(b_1^*-b_0^*)^2}} \times 100$$

where L^* indicates lightness, and both a^* and b^* indicate chromaticity, and the $L^*a^*b^*$ is the color space measured by a colorimeter (CM-2600d, Konica Minolta, Inc.). $L^*a^*b^*_0$, $L^*a^*b^*_1$, and $L^*a^*b^*_2$ represent the color space value of the white artificial leather before applying the eyeliner-penci, after applying it, and after cleansing it respectively[3].

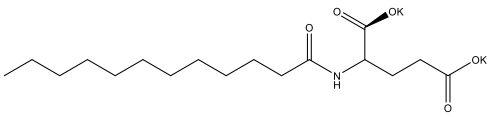
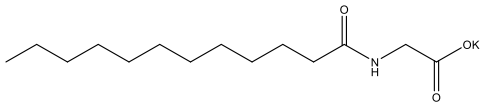
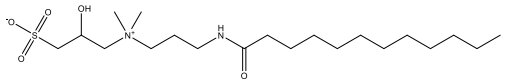
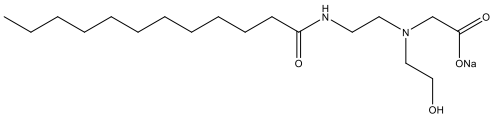
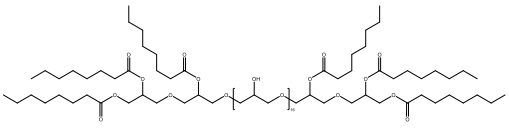
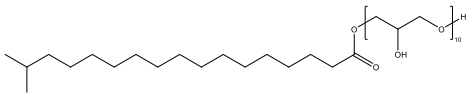
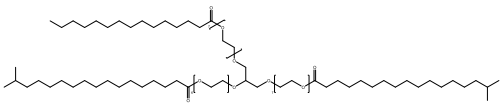
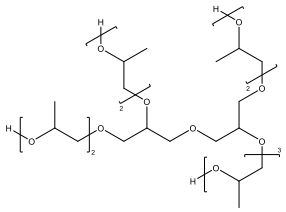
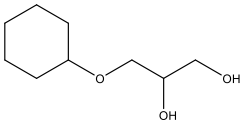
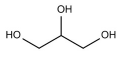
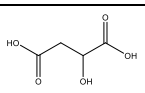


$$\text{Cleansing capability (\%)} = \frac{\sqrt{(L*_1-L*_2)^2+(a*_1-a*_2)^2+(b*_1-b*_2)^2}}{\sqrt{(L*_1-L*_0)^2+(a*_1-a*_0)^2+(b*_1-b*_0)^2}} \times 100$$

$L*a*b*$ value were measured by a colorimeter (CM-2600d, Konica Minolta, Inc.)

Fig. 1 Schematic of the evaluation test to determine the cleansing capability of the prepared formulations in this study

Table 1 Examples of ingredients used in the formulations prepared in this study

Categoly	The # of Ingredients	Examples	
		Material Name	Structure
anionic surfactant	8	potassium cocoyl glutamate	
		potassium cocoyl glycinate	
amphoteric surfactant	4	lauramidopropyl hydroxysultaine	
		sodium cocoamphoacetate	
nonionic surfactant	24	eicosaglycerol hexacaprylate	
		decaglycerol isostearate	
		PEG-20 glyceryl triisostearate	
polyols	33	PPG-9 diglyceryl ether	
		cyclohexylglycerin	
		glycerin	
pH adjuster	1	citric acid	
base	1	water	H_2O

2. Modeling of AI

Fig. 2 shows the date processig flowchart.

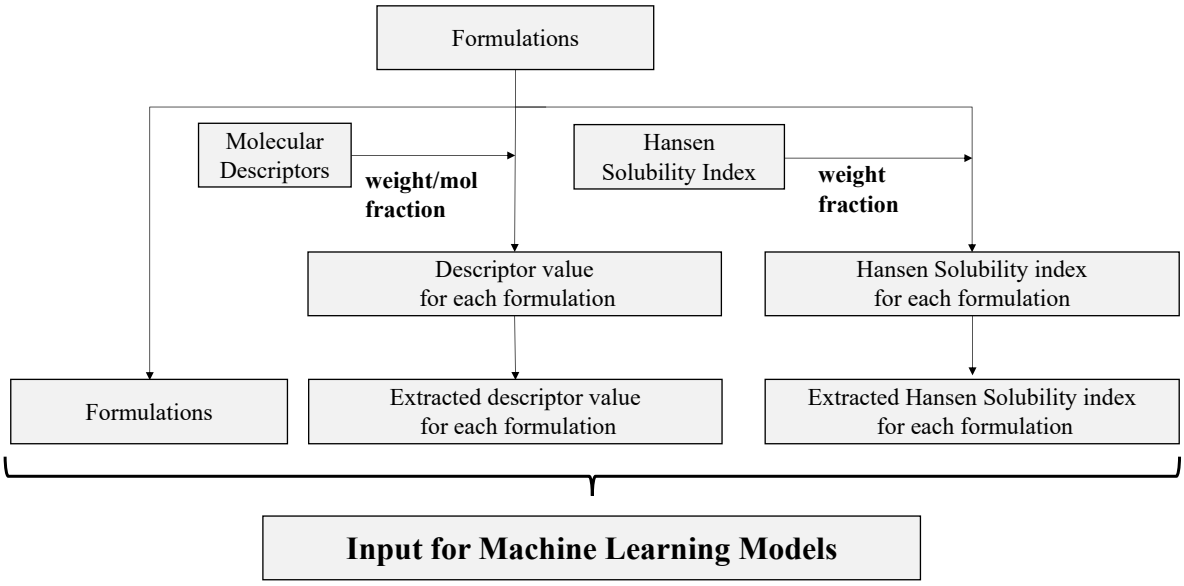


Fig. 2 Data processing flow

2.1 Molecular Descriptors

A molecular descriptor is defined as a numerical basic molecular property extracted from the chemical structure. Each type of molecular descriptors is related to a specific type of interaction between chemical groups in a particular molecule. Descriptors are utilized for prediction of chemical properties of not only single chemical, but also of chemical mixtures [4-7]. Also descriptors are applied for prediction of CMC of gemini surfactants[8]. Therefore, we extracted information from ingredients and predicted cleansing capabilities of the prepared formulations using molecular descriptors. These values were calculated from each ingredient's chemical structure formula with chemoinformatic tools, rdkit[9] and PaDEL-descriptors[10]. Entries with infinite or only one values were removed, and a *k*-NN imputer was applied to predict missing values. Then, the weighted average of an ingredient was calculated using mol or weight fraction to estimate the descriptor values of mixtured ingredients.

2.2 Hansen Solubility Parameters (HSPs)

We applied Hansen Solubility Parameters (HSPs) for the prediction of cleansing capability. HSPs were developed by Charles M. Hansen to predict if a material's ability to dissolve in another material and form a solution. HSPs are usually used to estimate whether solute dissolve in solvent by calculating HSPs distance between solute and solvent. There are some reports in which HSPs are utilized for prediction of properties of surfactants [11, 12]. In this study, instead of solute and solvent, we calculated the distance between each sample and obtained cleansing form samples with highest cleansing capability. We adopted this procedure because solute, an eyeliner in this study, was made of many ingredients and difficult to identify its structural formula. The HSPs distance is defined as $\{4*(dD1-dD2)^2+(dP1-dP2)^2+(dH1-dH2)^2\}^{0.5}$, where dD1, dP1 and dH1 are values of each sample —weighted average by weight fraction for mixture— and dD2, dP2 and dH2 are average values of three highest cleansing capability. We expected that HSPs can estimate the effects of the interactions between the ingredients in a formulation better than the descriptor method, in which the non-linear effect of the interaction of the ingredients was difficult to measure. HSPs values were calculated by the HSPiP software. Because some HSPs cannot be calculated by HSPiP for molecules with high molecular weight, missing HSPs values were imputed by a *k*-NN imputer as the descriptor caluculation.

2.3 Modeling and Feature Selection

With the use of descriptors and HSPs, the number of explanatory variables were greater than 1,000. Therefore, we applied machine learning to obtain lows for predicting cleansing performance from these numerous features. There are three types of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. Because our intention was to predict results within a continuous output, we selected supervised learning (regression) in this study. The input dataset was described in sections 2.1 and 2.2, and the output was the cleansing capability. We adopted two tree-based models (Random Forest Regressor and Extra Tree Regressor), two linear-based models (Lasso and Partial Lease Square), and one support-vector based model (SVR). Hyperparameters are shown in Table 2. The hyperparameters were optimized based on a grid-search method. All explanatory features were standarized with a mean of zero and standard deviation of one.

Because numerous features will cause a noise in the modeling, we also adopted Boruta method[13] to extract important features.

Table 2 Hyperparameter set for modeling

Model Name	Hyperparameter	Minimum	Maximum	Interval
ExtraTreeRegressor	n_estimators	25	125	25
	max_features	50	300	50
	min_sample_leaf	20	50	10
	min_sample_split	15	45	15
RandomForrestRegressor	max_depth	5	25	5
	n_estimators	10	25	5
SVR	$\log_{10}$(C)	-4	2	1
	kernel		Linear or RBF	
PLS	n_components	1	20	1
Lasso	$\log_{10}$(Alpha)	-4	2	1

2.4 Model Evaluation

The prediction performance was evaluated based on 10-fold cross validation with the indices of validated R^2, which represents the proportion of the variance for a dependent variable that is explained by independent variables.

3. In-silico formulation and actual cleansing capabilities

To evaluate whether the AI models could support human formulators, we made formulations virtually with a computer by the rules described below. We call this procedure the 'in-silico formulation'.

- All ingredients were assigned into one of six categories (same categories described in Table 1), which are anionic surfactants, amphoteric surfactants, nonionic surfactants, polyols, a pH adjuster (only citric acid), and a base (only water).
- In order to compare predicted and actual cleansing capabilities, the selection of anionic and amphoteric surfactant was restricted to one kind respectively —an anionic surfactant was restricted to potassium cocoyl glutamate, and an amphoteric surfactant was restricted to lauramidopropyl hydroxysultaine.
- Only one kind of ingredient was selected from each category, e.g., two nonionic surfactants could not be selected in one formulation.

- The addition rates of each ingredient except for pH adjuster (citric acid) and water were randomized for each category within the predefined range described in Table 3. The addition rate of citric acid was fixed with the value of 0.8 weight%, and the addition rate of water was calculated so that the sum of all ingredients became 100%.
- 10^5 of formulations were made in the procedure, and these formulations were predicted with the best model describe the Section 2.

In order to validate the predictions made by the in-silico formulation, actual cleansing capabilities of some formulations were measured (the formulations of measured samples will be shown in the Results section).

Table 3 Condition of in silico formulation
(water content was excluded in the weight% expression)

Categoly	Material Name	Randomize the addition rate	Minimum weight%	Maximum weight%
anionic surfactant	potassium cocoyl glutamate	Yes	3	20
amphoteric surfactant	lauramidopropyl hydroxysultaine	Yes	3	20
nonionic surfactant	eicosaglycerol hexacaplyrate PPG-20 glyceryl triisosterarate etc. (24 kinds in total)	Yes	0	10
polyols	PPG-9 diglyceryl ether glycerin etc. (33 kinds in total)	Yse	0	30
pH adjuster	citric acid	No	0.8 % (fixed ratio)	
base	water	No	100-Σ(other material amount)	

Results.

1. Modeling of AI

An AI model was established to predict cleansing capability. The prediction accuracy for each model was shown in Table 4. The best prediction accuracy was obtained with R^2=0.765. The prediction accuracy increased significantly with the use of descriptors. The results of the best model, Random Forest Regressor with a use of molecular descriptors, Hansen Solubility Index and feature extraction are shown in Fig. 3.

Table 4 Prediction accuracy for each model

#	Mol/Weight Fraction	Descriptors	Hansen Solubitiy Index	Feature Extraction	ExtraTree Regressor	RandomForest Regressor	SVR	Lasso	PLS
1	weight	Not Use	Not Use	Not Use	0.425	0.654	0.494	0.531	0.527
2	mol	Not Use	Not Use	Not Use	0.408	0.654	0.485	0.509	0.502
1 – 2	-	-	-	-	0.017	-0.001	0.009	0.022	0.025
3	weight	**Use**	Not Use	Not Use	0.581	0.724	0.528	0.571	0.557
4	mol	**Use**	Not Use	Not Use	0.598	0.757	0.477	0.535	0.535
3 – 4	-	-	-	-	-0.018	-0.032	0.051	0.037	0.022
5	weight	**Use**	**Use**	Not Use	0.579	0.722	0.545	0.597	0.559
6	mol	**Use**	**Use**	Not Use	0.598	0.757	0.503	0.564	0.541
5 – 6	-	-	-	-	-0.019	-0.035	0.042	0.033	0.018
7	weight	**Use**	**Use**	**Use**	0.610	0.755	0.544	0.581	0.549
8	mol	**Use**	**Use**	**Use**	0.657	**0.765**	0.511	0.535	0.518
7 – 8	-	-	-	-	-0.047	-0.010	0.033	0.046	0.032

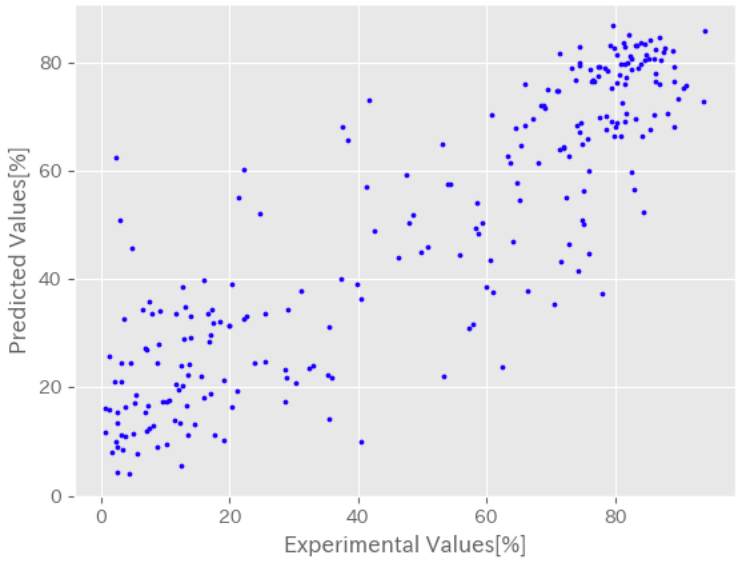


Fig. 3 Experimental vs predicted values of the cleansing capabilities of the cleansing form formulations

2. In-silico formulation and actual cleansing capabilities

The cleansing capabilities of the in-silico formulations are shown in Fig. 4 and Fig. 5. A box with light ang dark gray color in these figures indicate the middle 50 percent of the data (that is, the middle two quartiles of the data's distribution), and horizontal bars display all points within 1.5 times the interquartile range (in other words, all points within 1.5 times the width of the adjoining box), or all points at the maximum or minimum extent of the data. In Fig. 4,

we found eicosaglycerol hexacaprylate exhibited the highest cleansing capabilities in the point of both the median value (middle horizontal line in each box) and the best value (top horizontal line).

In Fig. 5, we stratified these formulation data into three categories, those not using nonioic surfactants, those using nonionic surfactants other than eicosaglycerol hexacaprylate, and those using eicosaglycerol hexacaprylate. Next, we stratified each category into subcategories with a kind of polyols in order to estimate the interactions with nonionic surfactants and polyols. We found the nonionic surfactants increased the cleansing capabilities, and hydrophobic PPG-9 diglyceryl ether or cyclohexylglycerin with lower IOB (Inorganic and Organic Balance) values boosted the cleansing capability more than glycerin. Eicosaglycerol hexacaprylate, PPG-9 diglyceryl ether, and cyclohexylglycerin have similar IOB values, suggesting that hydrophobic polyols inhibit an aggregation of eicosaglycerol hexacaprylate molecules, and promote efficient adsorption of eicosaglycerol hexacaprylate on makeup dirt.

In order to validate prediction data obtained in the in-silico formulation, we selected some formulations and measured the cleansing capabilities of them. The formulations and results were shown in Table 5. Nonionic surfactant eicosaglycerol hexacaprylate and cyclohexylglycerin/PPG-9 diglyceryl ether —(A), (B), and (C) in Table 5— showed the highest cleansing capabilities in actual formulations. Formulations with other nonionic surfactants and cyclohexylglycerin/PPG-9 —(D) to (H) in Table 5— showed lower cleansing capabilities. Formulations with glycerin —(I) and (J) in Table 5— showed much lower cleansing capabilities regardless of the kind of nonionic surfactant. These tendencies corresponded to results in Fig. 4 and Fig. 5. The prediction accuracy was better in the range of cleansing capabilities with more than 70% and those with less than 15%, which also corresponded to results in Fig. 3 where predictions of both high and low cleansing capabilities were more accurate.

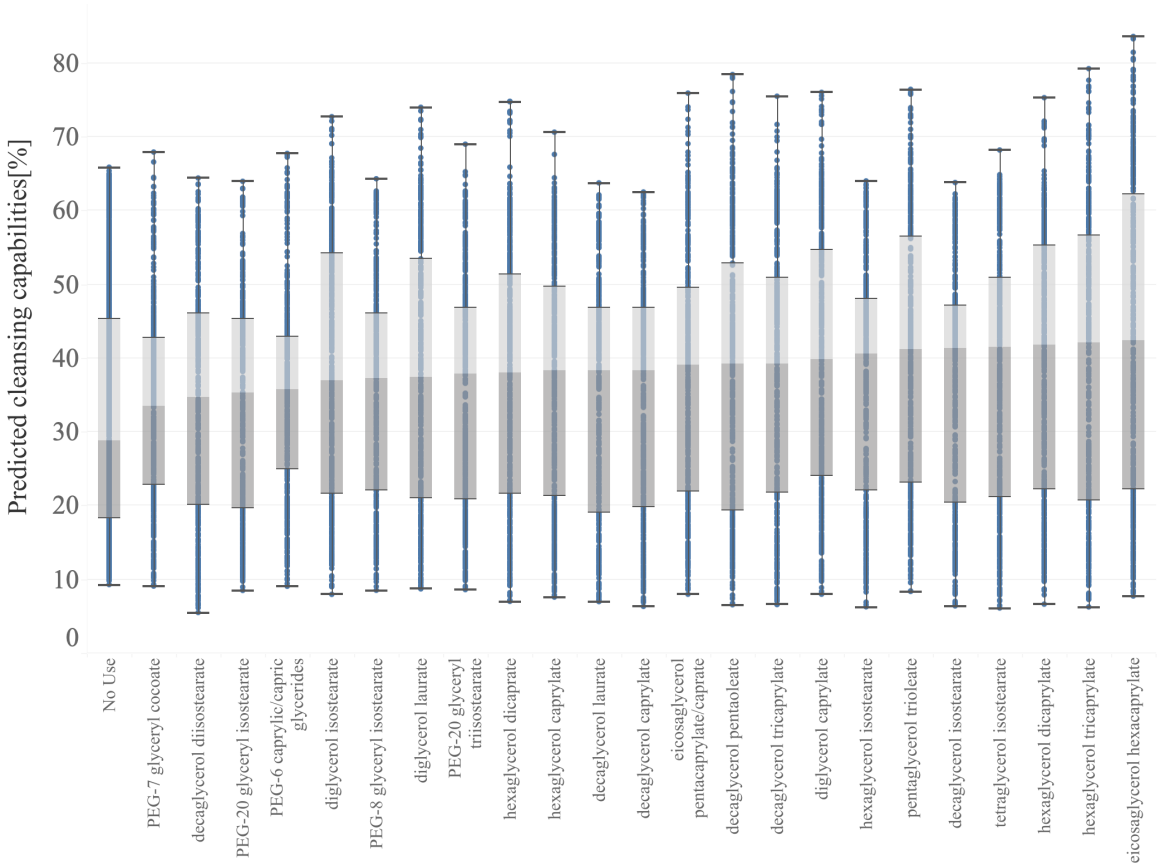


Fig. 4 Box plots of cleansing capabilities with the in-silico formulation, stratified with nonionic surfactants

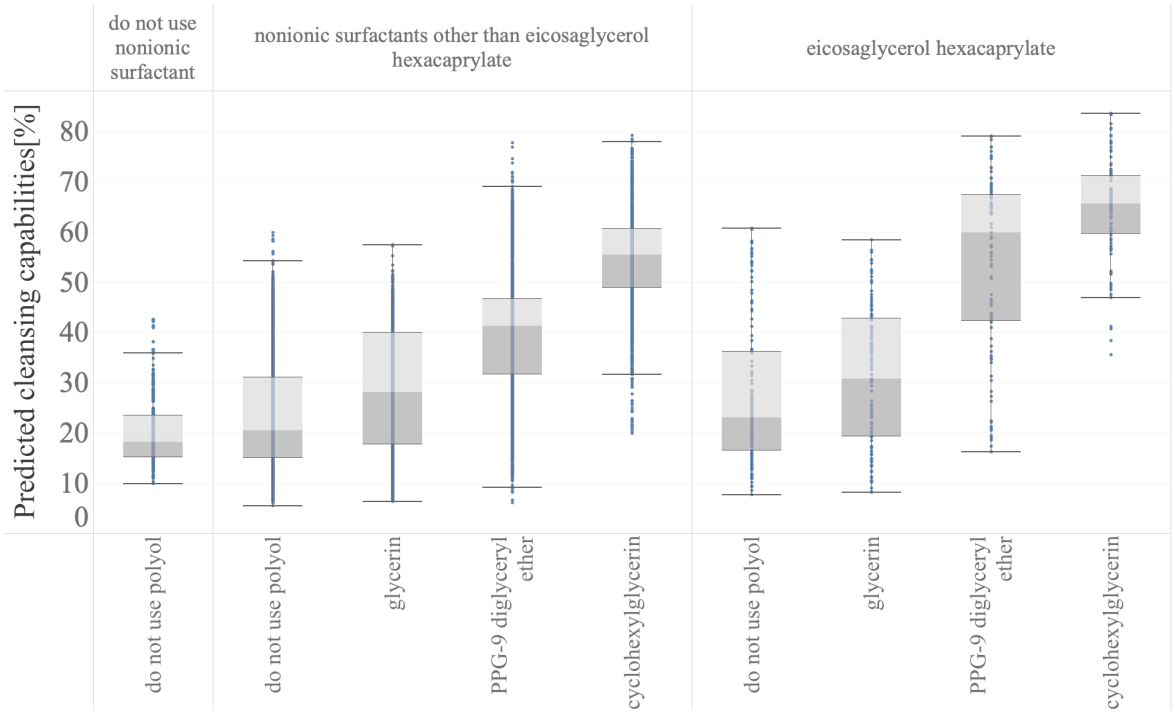


Fig. 5 Box plots of cleansing capabilities with the in-silico formulation, stratified with nonioic surfactants and polyols

Table 5 Formulations for comparison of predicted and actual cleansing capabilities

Category	Material Name	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)
anionic surfactant	potassium cocoyl glutamate	6	7	6	7	13	9	13	8	8	7
amphoteric surfactant	lauramidopropyl hydroxysultaine	5	9	7	8	4	6	6	9	7	7
nonionic surfactants	eicosaglycerol hexacaprylate	10	10	7						10	
	hexaglycerol caprylate						10				
	decaglycerol isostearate					10					
	decaglycerol laurate							10			
	PEG-20 glyceryl triisostearate								10		10
	PEG-8 glyceryl isostearate				10						
polyols	PPG-9 diglyceryl ether		11				14	13	11		
	cyclohexylglycerin	9		6	11	14					
	glycerin									11	13
pH adjuster	citric acid	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
base	water	69.2	62.2	73.2	63.2	58.2	60.2	57.2	61.2	63.2	62.2
Cleansing capability test	prediction / %	79.0	78.8	75.6	59.6	58.5	40.6	38.1	34.7	10.6	9.8
	actual / %	85.8	85.9	79.1	37.3	39.8	15.4	29.7	46.3	8.3	3.6

Discussion.

The use of descriptors increased the prediction accuracy increased in all models, which indicates that the chemical properties expressed as molecular descriptors successfully

enabled the prediction of cleansing capabilities. However, HSPs did not improve the prediction accuracy.

Furthermore, for the calculation of weighted average calculation, mol% was suitable for tree-based models, whereas wt% was suitable for linear-based models. The mol% of the weighted average is considered potentially more accurate based on stoichiometry, because water constitutes >97 mol% on average in formulations. Therefore, the influence of water is more dominant. Linear-based models are affected more by this influence than tree-based models. For the prediction of cleansing capability, nonlinear behavior should also be considered owing to the interactions between surfactants and water molecules, and their self-assembly. Tree-based models are usually more suitable for non-linear prediction; therefore, their prediction accuracies are higher than those of linear-based models.

The in-silico formulation helped us understand the effect not only of each material to cleansing capabilities but also of combination by materials with the consequence of molecular interactions. The in-silico formulation also assisted us of making formulations with higher cleansing capabilities.

In this study, by applying the prediction method and the in-silico formulation method, we identified a cleansing foam formulation consisting of eicosaglycerol hexacaprylate and cyclohexylglycerin/PPG-9 that exhibited a high cleansing capability of >85% for the removal of waterproof eyeliner.

Conclusion.

Using an artificial intelligence (AI) with machine learning, we have built a cleansing capability prediction system that incorporates the effects of surfactant self-assembly and chemical characteristics of the ingredients. The accuracy of R^2=0.765 was obtained in the prediction of cleansing performance. Non-linear behavior, i.e. interactions among cosmetic ingredients in formulations, made it more difficult for formulators to predict their performance. However, high accuracy was obtained by incorporating chemical characteristics with descriptors. This AI prediction model based on the molecular structure of the ingredients and surfactant self-assembly showed higher accuracy and was better than conventional approaches such as multiple linear regression. With the use of the in-silico formulation, formulators can obtain information of which ingredients should be selected in

order to obtain highest cleansing capabilities. The prediction model and the in-silico formulation may contribute to a significant reduction in the effort required for the cosmetics development.

Conflict of Interest Statement. NONE.

References.

1. Watanabe K., Sakurai N., Meno T., Yasuda C., Takahashi S., Hori A., Tsuchiya K., Sakai K., (2021) 'Novel Spontaneous Cleansing Feature of Foam — Hybrid Bicontinuous-Microemulsion-Type Foamy Makeup Remover —', Journal of Society of Cosmetic Chemists of Japan, 55(1), pp. 19–27.
2. Watanabe K., Masuda M., Nakamura K., Inaba T., Noda A., Yanagida T., Yanaki T., (2004) 'A new makeup remover prepared with a system comprising dual continuous channels (bicontinuous phase) of silicone oil and water', IFSCC Magazine, 7(4), pp. 310–318.
3. Iwanaga T., Uchida K, Takeuchi N, Abe Y., (2005) 'Development of Oil-Type Make-up Remover Prepared with Polyglycerol Fatty Acid Esters', Journal of Society of Cosmetic Chemists of Japan, 39(3), pp. 186–194.
4. Gaudin, T., Rotureau, P. and Fayet, G. (2015) 'Mixture Descriptors toward the Development of Quantitative Structure–Property Relationship Models for the Flash Points of Organic Mixtures', Industrial & engineering chemistry research, 54(25), pp. 6596–6604.
5. Abbasi, A. and Eslamloueyan, R. (2014) 'Determination of binary diffusion coefficients of hydrocarbon mixtures using MLP and ANFIS networks based on QSPR method', Chemometrics and Intelligent Laboratory Systems, 132, pp. 39–51.
6. Sobati, M.A. et al. (2016) 'A new structure-based model for estimation of true critical volume of multi-component mixtures', Chemometrics and Intelligent Laboratory Systems, 155, pp. 109–119.
7. Gaudin, T., Rotureau, P. and Fayet, G. (2015) 'Mixture Descriptors toward the Development of Quantitative Structure–Property Relationship Models for the Flash

Points of Organic Mixtures', Industrial & engineering chemistry research, 54(25), pp. 6596–6604.

8. Absalan, G. et al. (2004) 'Quantitative structure–micellization relationship study of Gemini surfactants using genetic-PLS and genetic-MLR', QSAR & combinatorial science, 23(6), pp. 416–425.
9. RDKit: Open-Source Cheminformatics Software
10. Yap, C.W. (2011) 'PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints', Journal of computational chemistry, 32(7), pp. 1466–1474.
11. Faasen, D.P. et al. (2020) 'Hansen solubility parameters obtained via molecular dynamics simulations as a route to predict siloxane surfactant adsorption', Journal of colloid and interface science, 575, pp. 326–336.
12. Afzal, O. et al. (2022) 'Hansen solubility parameters and green nanocarrier based removal of trimethoprim from contaminated aqueous solution', Journal of molecular liquids, p. 119657.
13. Kursa, M.B., Jankowski, A. and Rudnicki, W.R. (2010) 'Boruta – A System for Feature Selection', Fundamenta Informaticae, 101(4), pp. 271–285.