Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Optimization-based Cosmetic Formulation: Integration of Mechanistic Model, Surrogate Model, and Heuristics

Xiang Zhang[1], Teng Zhou[2], and Ka Ng[3]

[1]Affiliation not available
[2]Max Planck Institute for Dynamics of Complex Technical Systems
[3]Hong Kong University of Science and Technology

May 14, 2020

Abstract

Multiple functional and hard-to-quantify sensorial product attributes that can be satisfied by a large number of cosmetic ingredients are required in the design of cosmetics. To overcome this challenge, a new optimization-based approach for expediting cosmetic formulation is presented. It exploits the use of a hierarchy of models in an iterative manner to refine the search for creating the highest-quality cosmetic product. First, a systematic procedure is proposed for optimization problem formulation, where the cosmetic formulation problem is defined, design variables are specified, and a set of models for sensorial perception and desired product properties are identified. Then, a solution strategy that involves iterative model adoption and two numerical techniques (i.e., generalized disjunctive programming reformulation and model substitution) is applied to improve the efficiency of solving the optimization problem and to find better solutions. The applicability of the proposed procedure and solution strategy is illustrated with a perfume formulation example.

Introduction

The chemical industry designs and produces a vast number of chemical products to serve the society. Chemical products are classified as molecular products, formulated products, functional products, and devices.[1,2] Among them, formulated products such as cosmetic and paint are formed by mixing selected ingredients in a formula, which may possess certain microstructures of their own (e.g., powder and emulsion). The formula (i.e., ingredient selection and composition) has a significant impact on formulated product quality. Thus, the major aim of formulated product design is to find a formula that exhibits consumer-desired properties.[2,3]

As a major component of formulated products, cosmetics are applied to the human body for cleansing, beautifying, promoting attractiveness, or altering appearance. They are sold in many forms. Table 1 lists the commonly used cosmetic product forms such as cream and gel. The global cosmetic market valued at $532 billion in 2017 is large, but competitive and dynamic.[4] Many cosmetic products exist on the market but they tend to have short product life. To succeed in this environment, rapid formulation of new and improved cosmetics is crucial. The quality of cosmetics can be broadly represented by two types of attributes. One is sensorial attributes (e.g., smell and sight) perceived by five human senses during and after the application of cosmetics. The other is functional attributes (e.g., stability and safety), which ensures that cosmetics can be assuredly used with the desired functions. Table 2 lists the relevant sensorial and functional attributes of four cosmetic products with different product forms. For instance, the senses of how lipstick is felt by the lips and how the lips look after application are part of the lipstick quality. Meanwhile, lipsticks should be stable, safe, and not broken in use. It is known that the sensorial quality is the dominant consideration for consumers to choose one cosmetic product over another.[5,6] Thus, it needs to be explicitly considered in cosmetic formulation.

[insert Table 1 here] and **[insert Table 2 here]**

Creating a qualified cosmetic formula with desirable attributes is challenging because there exist a large number of cosmetic ingredients leading to numerous possible recipes. Many attributes involve complex physicochemical phenomena, some of which are not yet fully understood. More importantly, it is hard to quantify or predict consumers' sensations since they are elusive, subjective, and affected by consumer status.[7] In this case, the use of any single model or tool cannot capture the cosmetic formulation problem in its totality.[8] The design of related personal care products such as shampoo and toothpaste faces similar issues. Currently, new cosmetics are usually developed by experimental trial-and-error. This is expensive and time-consuming. The search space is limited and there is no guarantee that an optimal formula is found.[5] For expediting new cosmetic formulation, it is highly desirable to develop an effective model-based optimization approach to complement the efforts of experienced cosmetic formulators.

Model-based computer-aided mixture/blend design (CAM$^{\text{b}}$D) methods have been applied extensively. The ingredients are generated using the group contribution (GC) approach. Linear and simple nonlinear mixing rules are applied to predict mixture properties. The CAM$^{\text{b}}$D methods are usually applied to mixtures with less than six ingredients such as solvent mixtures[9–12] and blended fuels.[13–17] This is because much greater computational effort is needed as the number of ingredients increases.[12,13] Since the number of ingredients in cosmetic products is typically larger than 15 and can be up to 50,[18] it is highly desirable to develop alternative methods for cosmetic formulation.

From a product design perspective, several model-based methods have been proposed and applied to cosmetics and the highly related personal care products. Omidbakhsh et al.[19] built statistical models to design disinfectant. Disinfection effect is first correlated with ingredient composition and then composition is optimized to design a new disinfectant with maximal disinfection effect. Smith and Ierapetritou[20] optimized the formula of an under-eye cream using regressed polynomial functions that correlate product attributes with ingredient composition and operating conditions. Bagajewicz et al.[21] started with a base-case formula of skin lotion and optimized its composition for maximum profit in a competitive market. In these studies, cosmetic formulation is treated as a nonlinear programming problem. Only the composition of pre-selected ingredients is optimized without considering the selection of other ingredient alternatives. Obviously, a more superior formula can be easily missed without considering all the available ingredients. Conte et al.[3,22] combined computer-aided modeling with experimental testing to formulate sunscreen spray. Ingredients are selected using databases, knowledge-base, and GC methods. Kontogeorgis et al.[5] extended this integrated modeling-experimental approach to formulate emulsified products. Zhang et al.[23] proposed an integrated framework for formulated product design considering the optimal identification of ingredients, composition, microstructure, etc. Arrieta-Escobar et al.[24] incorporated heuristics and mixed-integer nonlinear programming (MINLP) to identify the optimal ingredients and composition of hair conditioner. By integrating different methods and tools, the above studies can properly select ingredients from a set of candidates with optimized composition. However, only certain mixture properties (e.g., color and greasiness) related to sensorial attributes have been considered.[5,24] Only recently has machine learning been used to predict sensorial perceptions.[25–27] Even less is available on how sensorial satisfaction can be explicitly quantified, modelled, and incorporated into product formulation.[25] By integrating machine learning models, a grey-box optimization problem was formulated and solved using genetic algorithm (GA) for food product design.[25] This method is applied with simplified property models involving a limited number of equations, because GA is inefficient in handling a large number of complex constraints that are common in cosmetic formulation problem.[23] In addition, GA cannot guarantee ε-optimality.

To fill this gap, a novel optimization-based approach is developed for the formulation of cosmetics. Figure 1 illustrates the overall methodology. For given consumer needs and a set of potential chemical ingredients, an MINLP problem is formed by integrating (rigorous and short-cut) mechanistic models, data-driven surrogate models, and mathematical equations derived from heuristics. The objective is to maximize the sensorial

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

perception. Then, a novel solution strategy that involves an iterative adoption of a hierarchy of models and different numerical techniques is applied to solve the optimization problem efficiently. Then, the optimal formula can be verified by experiments. The paper is organized as follows. A systematic procedure is first introduced for problem formulation. Then, the iterative procedure for model adoption and optimization solution strategy are described. Finally, a perfume example is discussed to illustrate the applicability of the proposed approach.

[**insert Figure 1 here**]

Systematic Procedure for Optimization Problem Formulation

Figure 2 shows the 3-step procedure. In Step 1, the cosmetic formulation problem is defined where the desired product form, product attributes, and product specifications and properties are determined based on market study and product knowledge. In Step 2, potential ingredient candidates and relevant microstructural descriptors (if applicable) are generated to specify the design variables. Step 3 identifies a set of mechanistic models and surrogate models and converts heuristics into mathematical equations. For each step, the sources of the input are shown on the left and the outputs are on the right. The details are described below.

[**insert Figure 2 here**]

Step 1: Problem Definition

When a new cosmetic design project is launched, the product type and form such as a facial powder or lipstick are first decided by the marketing team based on the target market, potential consumers, competing products, etc. The quality of the new cosmetic product depends on its sensorial and functional attributes. Usually, the consumer-desired attributes can be specified through interview and survey with potential consumers. The sensorial perceptions given by a cosmetic are the most essential for its satisfaction and repeated use by consumers.[6] In practice, sensorial perception is assessed through sensorial evaluation. A number of panelists assess various cosmetic samples using well-defined protocols and their perceptions are quantified using sensorial ratings. Then, an overall sensorial rating can be obtained to represent the degree of satisfaction of the cosmetic.[7] Note that in addition to perception, other factors such as packaging and price affect consumer's purchase decision. These factors are not considered in this work and the objective function is to maximize the overall sensorial rating (q).

max q (1)

In addition, the functional attributes are also needed to be satisfied. Each cosmetic has its unique functional attributes. For instance, a hair spray should dry rapidly and perfume should be transparent (Table 2).

The product attributes can be translated into relevant physicochemical properties (e.g., melting point for lipstick) and product specifications (e.g., sun protection factor for sunscreen product) using engineering know-how. For the four cosmetics in Table 2, the last column lists various properties related to their sensorial and functional attributes. How a lipstick is sensed by the lips is affected by its viscosity. The pH of a skin cream affects its safety. Then, a set of design targets (i.e., lower and upper bounds) on the properties can be identified based on the engineering know-how and product in-house data. These bounds serve as constraints in the optimization problem.

$PL^k \leq P^k \leq PU^k, \quad k \in K$ (2)

where P^k is the k -th desired property. K is the set of properties. PL^k and PU^k are the lower and upper bounds, respectively. Note that the nomenclature is presented in Supporting Information.

Step 2: Ingredient Candidate Generation

To provide multiple desired attributes, many chemical ingredients are needed. Cosmetic ingredients are classified into different types based on their functionalities. Table S1 lists the ingredient types that are widely used in various cosmetics and their functions.[7,28] For instance, an abrasive in a facial cleanser is made up of solid particles used for physically cleaning hard surface such as epidermis. Three types of moisturizers (i.e., emollient, humectant, and occlusive) can be used to provide hydration effect. Emollient can improve the skin's water-oil balance, humectant inhibits water evaporation, and occlusive can form a water-repellent layer to reduce water loss. For a cosmetic, the needed ingredient types can be identified based on the fundamental formulation science and the desired product attributes.

For each ingredient type, a set of ingredient candidates can be generated using databases[29,30] and computer-aided tools.[14,31] Regarding each of the ingredient types in Table S1, the last column lists two commonly used ingredient candidates. For instance, lactic acid and triethanolamine are often used as an acidic and alkaline pH buffers, respectively. With the years of development in the cosmetic industry, hundreds of ingredient candidates exist for each ingredient type. To reduce the search space, the candidates can be pre-screened using ingredient screening tools based on cost, regulations, availability, etc. to generate a more organized pool of ingredient candidates

$$\begin{matrix} I_A = \{I_{A,1}, I_{A,2}, \ldots, I_{A,a}\} \\ I_B = \{I_{B,1}, I_{B,2}, \ldots, I_{B,b}\} \\ \cdots \\ I_Z = \{I_{Z,1}, I_{Z,2}, \ldots, I_{Z,z}\} \end{matrix} \quad (3)$$

where I_A, I_B,. . . , I_Z are ingredient types.$I_{A,1}$, $I_{A,2}$,. . . , $I_{A,a}$, etc. represent the generated ingredient candidates. Here, the subscripts a, b, and$z$ denote the number of candidates in each ingredient type. Each candidate has different properties (e.g., density, solubility and pH) which can be collected from the literature, database, and experiment. The selection of ingredients is intuitively a discrete-continuous optimization problem. Each ingredient candidate can be assigned a binary variable S_i to control ingredient selection and a continuous variable (e.g., volume fraction V_i) to denote its composition. If the i -th candidate is selected, S_i is equal to 1 andV_i is constrained by its lower (VL_i) and upper (VU_i) bounds. Otherwise, S_i and V_i are equal to 0.

$\sum_i V_i = 1$ (4)

$VL_i \bullet S_i \leq V_i \leq VU_i \bullet S_i$,$i \in \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\}$ (5)

In addition to ingredients, microstructure can affect the properties when certain product forms are used. Typically, the major microstructural features can be characterized by some geometric descriptors that can be correlated with the mixture properties by experiment and multi-scale modeling to account for the microstructure-property relationship.[32] The last column of Table 1 lists the relevant microstructure descriptors for various commonly used cosmetic product forms. For example, the oil droplet size affects the viscosity and texture of a moisturizing lotion in the form of an oil-in-water emulsion.[33] The emulsion type and particle shape can be decided using heuristics.[34] Geometric descriptors ms such as particle size are continuous variables

$msL \leq ms \leq msU$ (6)

where msL and msU are the lower and upper bounds, respectively. The microstructure is decided by both the formulation and manufacturing process design.[35]

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Step 3. Model Identification

Model for Sensorial Perception

Surrogate model that captures the input-output data is built to predict the sensorial rating. After a surrogate model is trained, its analytical form can be used for optimization.

$$q = f\left(V_{I_{A,1}}, I_{I_{A,2}} \ldots, V_{I_{Z,z}}, ms\right) \ (7)$$

The first task is to collect training data. The input data can be the cosmetic recipes and the microstructures, namely $(V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}, ms)$. The output data is the corresponding sensorial rating (q). Here, the historical data of sensorial evaluations can be utilized. When the historical data is scarce, additional data sampling is required. By far, many efficient sampling approaches have been used in the cosmetic industry such as Latin-hypercube sampling, Plackett-Burman, full-fractional, etc. Referring to the "one in ten" rule, the number of data samples is preferably ten times more than the number of ingredient candidates. The second task is to build an accurate surrogate model. Currently, multiple types of surrogate models can be utilized such as linear regression, kriging, artificial neural network (ANN), radial basis function, etc. Among them, some surrogate models (e.g., random forest) cannot provide available derivative information while the derivatives of many other surrogate models are symbolically available such as linear regression, ANN with tansig kernel function, etc.[36] Here, a surrogate model with available derivative information is preferred because solving a discrete-continuous optimization problem with no derivative information is very challenging. The hyperparameters of the surrogate model structure should be carefully tuned. The heuristics and experience reported in the literature can be consulted.[36,37] Afterward, model accuracy needs to be validated. The widely used validation methods include K-fold cross validation and holdout method. If the model is not sufficiently accurate, the type of surrogate model and the hyperparameters should be re-selected.

Models for Target Properties

Three types of models can be applied for predicting the target properties: rigorous mechanistic model, short-cut model, and surrogate model. Typically, the formulation and application of cosmetics involve various phenomena (e.g., kinetics, thermodynamics, and transport). For any property, the associated phenomena should be first identified based on the basic engineering sciences and domain knowledge, followed by the identification of the relevant mechanistic models. Generally, rigorous models are the most accurate but more complex and sometimes with unknown parameters. The perfume diffusion model[38] and ingredient percutaneous absorption model[39] are examples. Instead of accounting fully the physical phenomena, simple short-cut model captures the property's dependence on the most influential factors. Usually, short-cut model is sufficiently accurate within pre-specified conditions. Note that both rigorous and short-cut models involve many intermediate variables for describing the relevant phenomenon. The rigorous or short model for k -th desired property (P^k) can be represented as

$$P^k = G^k(IM^k_{I_{A,1}}, IM^k_{I_{A,2}}, \ldots, IM^k_{I_{Z,z}}), k \in K \ (8)$$

$$\mathrm{IM}^k_i = \mathrm{IMG}^k\,(V_i, ms), i \in \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\} \ (9)$$

where IM^m_i denotes the intermediate variable related to the i -th ingredient candidate (e.g., vapor pressure and activity coefficient). If there are no suitable mechanistic models but data are available, surrogate models can be adopted,[40] although the model validity is often limited to the range of available data. The input data should be the sampled cosmetic recipes and microstructure. The output data are the target properties. For the k -th property (P^k), its surrogate model is

$$P^k = g^k(V_{I_{A,1}}, I_{I_{A,2}} \ldots, V_{I_{Z,z}}, ms), k \in K \ (10)$$

Accordingly, for any desired property, a set of models (rigorous, short-cut, and surrogate) should be identified for use in the optimization.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

The use of heuristics is often inevitable in cosmetic formulation.[24,41] The reason is that some phenomena have not been identified or are poorly understood. For instance, a hydrocolloid thickener with a weak gel network structure is preferred for use in emulsion-based product to generate thixotropic behavior, although no formal justification has been given.[34] In addition, heuristics can effectively help reduce the search space. Many heuristics, although not all, can be transformed into mathematical design constraints for use in the optimization. Table 3 shows the widely used forms of heuristics and the associated equations for formulated product design. For instance, if the number of ingredients for certain type of ingredient is suggested, an inequality constraint$TL \le \sum_i S_i \le TU, \quad i \in I_X$ can be generated.

[insert Table 3 here]

Iterative Model Adoption and Optimization Solution Strategy

Figure 3 presents an iterative model adoption framework to generate an optimization problem that can be solved efficiently. The strategy is to first employ the most accurate rigorous mechanistic model for property prediction. This is expected to provide a reliable solution. In case a rigorous model is not available, the relatively simple but less accurate short-cut model can be adopted. The surrogate model is used when there is no suitable mechanistic model. Through this strategy, the cosmetic formulation problem can be explicitly expressed as an MINLP optimization problem below.

$q = f(V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}, ms)$Sensorial rating (11)

s.t. $PL^k \le P^k \le PU^k$, $k \in K = MM \cup SM$ Design target

$P^m = G^m(IM^m_{I_{A,1}}, IM^m_{I_{A,2}} \ldots, IM^m_{I_{Z,z}})$,IM$^m_i = IMG^m(V_i, ms)$, $m \in MM$ Mechanistic model

$P^s = g^s(V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}, ms)$,$s \in SM$ Surrogate model

$H\left(S_{I_{A,1}}, S_{I_{A,2}}, \ldots, S_{I_{Z,z}}, V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}\right) \le 0$Heuristics in Table 3

$msL \le ms \le msU$ Design variables

$S_i \in \{0,1\}^i$, $\sum_i V_i = 1$,$VL_i \bullet S_i \le V_i \le VU_i \bullet S_i$,$i \in \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\}$

where P^m is the m -th property predicted using a (rigorous or short-cut) mechanistic model. MM is the set of properties predicted using mechanistic-based models. P^s is the s -th target property (P^s) predicted using a surrogate model.SM is the set of properties predicted using surrogate models.

[insert Figure 3 here]

The computational difficulty of the MINLP problem depends on the number of ingredient candidates and the complexity of the adopted models. A large number of ingredient candidates often create a combinational problem. Many rigorous models (e.g., thermodynamic and transport phenomena models) involve nonlinear and nonconvex equations. Along with complex surrogate models (e.g., neural network), the optimization problem is prone to convergence failure if the problem is directly solved using standard MINLP solvers. Some of these problems can be handled with advanced algorithms. For instance, Schweidtmann and Mitsos[42,43] recently developed and applied an efficient global solver for ANN embedded MINLP problems. Alternatively, the problem can be resolved by reformulating the optimization problem.[44] Two techniques are proposed for enhancing optimization convergence and finding better solutions (Figure 3). If the problem can be directly solved, the solution is sent for experimental validation. Otherwise, generalized disjunctive programming (GDP) can be used because of the need to calculate the multiple intermediate variables in the mechanistic models. If the GDP problem still cannot be solved or better solutions are needed, the model(s) in use are replaced with alternative model(s) and repeat the calculations.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

GDP reformulation

As can be seen in Eq. 11, even if i -th ingredient candidate is not selected, its intermediate variables (IM_i^m) must be calculated. Also, forcing V_i to 0 may lead to singularity at$V_i = 0$ for some models (e.g., logarithmic function). Thus, when mechanistic models are employed and multiple intermediate variables are calculated through complex equations, the redundant constraints and singularities can lead to convergence failure. Similar to the tray selection problem in distillation column design,[44]cosmetic formulation problem can be formulated using GDP. As an alternative way to program discrete-continuous problem, GDP is a logic-based method containing Boolean and continuous variables. Constraints are expressed as disjunctions, algebraic equations, and logic propositions.[44] The following disjunction can be used to express the bounds and intermediate variables for Eq. 11.

$$\begin{bmatrix} Y_i \\ VL_i \le V_i \le VU_i \\ IM_i^m = IMG^m(V_i, ms) \end{bmatrix} \bigvee$$

$$\begin{bmatrix} \neg Y_i \\ V_i = 0 \\ IM_i^m = 0 \end{bmatrix}, i \in \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\} \ (12)$$

where Y_i is the Boolean variable for ingredient selection. If the$i$ -th ingredient is selected, the bounds on V_i are fulfilled and its intermediate variables IM_i^m are calculated. Otherwise, they are not calculated and simply set as 0.

To solve a GDP problem, it is often transformed back into MINLP using big-M or convex-hull relaxation to take advantage of standard MINLP solvers. It is found that the big-M method is more appropriate in solving mixture design problem since singularity issue can still occur in the convex-hull relaxation.[12] After transforming the above disjunction using big-M approach, the cosmetic formulation problem is reformulated below. S_i has a one-to-one correspondence with Y_i. bm is a sufficiently large parameter.

$q = f(V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}, ms)(13)$

s.t. $PL^k \le P^k \le PU^k$, $k \in K = MM \cup SM$

{

$$\begin{array}{c} VL_i - bm \bullet (1 - S_i) \le V_i \le VU_i + bm \bullet (1 - S_i) \\ -bm \bullet S_i \le V_i \le bm \bullet S_i \\ \mathrm{IMG}^m\,(V_i, ms) - bm\,(1 - S_i) \le IM_i^m \le IMG^m\,(V_i, ms) + bm(1 - S_i) \\ -bm \bullet S_i \le IM_i^m \le bm \bullet S_i \end{array} \quad \text{Big-M constraint}$$

$P^m = G^m(IM_{I_{A,1}}^m, IM_{I_{A,2}}^m, \ldots, IM_{I_{Z,z}}^m)$,$m \in MM$

$P^s = g^s(V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}, ms)$,$s \in SM$

$H\left(S_{I_{A,1}}, S_{I_{A,2}}, \ldots, S_{I_{Z,z}}, V_{I_{A,1}}, V_{I_{A,2}}, \ldots, V_{I_{Z,z}}\right) \le 0$

$msL \le ms \le msU$, $S_i \in \{0,1\}^i$,$\sum_i V_i = 1$,$i \in \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\}$

Model substitution

Some rigorous mechanistic models are too complicated to be directly used for optimization even if they are programmed using GDP. There is always a trade-off between model accuracy and traceability. In this case, the complicated but accurate rigorous models can be replaced by simple short-cut model or surrogate model to reduce the computational effort and to seek out even better solutions. A surrogate model is a good choice when it is relatively easy to generate simulation data as training data from the rigorous model. Although

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

the model accuracy is reduced, it is easier to solve and to obtain the global solution.[42,45–47] After model substitution, the newly generated optimization problem should be solved and the optimal solution obtained can be denoted as V_i^*. This solution must be validated using the original rigorous mechanistic models. If the validation fails, the newly generated optimization problem should be re-solved by adding the equation below to remove this solution (V_i^*) that fails validation. Otherwise, the solution can be sent for experimental verification.

$\sum_i \left(V_i^{**} - V_i^*\right)^2 \geq tol$ (14)

Here, V_i^{**} is the solution of a new round optimization. The parameter tol is a small tolerance.

Case Study: Liquid Perfume

As a popular cosmetic, perfume is a liquid mixture releasing pleasant scents. The global perfume market is valued at $31.4 billion in 2018. Based on the volume fraction of fragrant compounds, perfume can be classified into several types. Extrait contains 15-30%, Eau de parfum 10-20%, and Eau de cologne 3-5%. For each type, thousands of products exist on the market. Most perfumes are made from various synthetic fragrances for easy quality control. The experienced perfumers create new recipes by trial-and-error. Here, the proposed framework and solution strategy are applied to formulate a new Eau de parfum.

Step 1: Problem definition

Table 2 shows that the most critical sensorial attribute of perfume is the smell. After applying the perfume, the fragrant compounds begin to evaporate and are detected by an observer away from the location of release. The scents change over time because each constituent is released at a different rate. This process can last several hours. Based on the order in which the odors appear, the released scents are classified as: top note, middle note, and base note. Top note is comprised of the scents perceived immediately after perfume application and generally lasts 5-15 minutes. The scents in the middle note emerge after the top note dissipates and remain for around an hour. The base note appears close to the end of middle note and can last several hours. During the sensorial evaluation of a perfume, each note is assessed and rated. An average rating can then be obtained to represent consumer preference.[48] Thus, the objective function is to maximize the overall sensorial rating on the smell of perfume (q_s).

q_s (15)

A perfume can be formulated to provide any specific scent with certain intensity. In this study, it is assumed that the marketing team decides that a lemon-like odor should dominate in the top note of the new Eau de parfum. There are no specific odors required for the middle and base notes as long as the overall sensorial rating is maximized. Thus, the odor type with the highest intensity in the top note (OTTN) is

$OTTN = \text{lemon} - \text{like}$ (16)

Moreover, since homogeneous liquid solution is transparent, all the perfume ingredients must be completely miscible with each other. Perfume safety is related to its toxicity and flammability. Toxicity can be measured by the median lethal dose (LD_{50}). The larger theLD_{50} is, the safer the perfume is. Since a solution withLD_{50} larger than 5000 mg/kg can be regarded as non-toxic, this is chosen as the design target as defined in Eq. 17. Flammability depends on the flash point (T_{fp}), which is the lowest temperature at which liquid vapor ignites given an ignition source. A higher flash point indicates lower flammability. Here, the flash point is required to exceed 15 °C which is roughly the value of existing perfume products.

$LD_{50} \geq 5000$ mg/kg (17)

$T_{\text{fp}} \geq 15$ (18)

Accordingly, four design targets are specified: constraints on LD_{50} and flash point, a homogeneous solution, and a dominant lemon-like odor in the top note.

Step 2: Ingredient candidate generation

Table 4 lists the four required ingredients types and their functions.[49] Various fragrances are used to provide different scents. Based on the volatility, fragrance compounds can be classified into three types in accordance with the top note, middle note, and base note. For instance, the top note fragrances are most volatile with a vapor pressure typically larger than 0.1 mmHg. The vapor pressure of middle note and base note fragrances are 0.001–0.1 mmHg and less than 0.001 mmHg, respectively.[50]

Referring to the perfume manual,[51] 48 common perfume ingredients are generated in the four ingredient types (see Table 4). 17 candidates are top note fragrances, 16 candidates are middle note fragrances, and 13 candidates are base note fragrances. Each candidate has a different odor. For instance, as a top note fragrance, limonene occurs naturally in the oil of citrus peels and offers a lemon-like odor. Coumarin is the source of tonka bean's distinctive aroma and is often added as a base note fragrance. An ethanol and water mixture is by far the most common solvent in perfume.[52] Ingredient selection is controlled by the binary variable S_i and ingredient composition is represented by volume fraction V_i. If the$i$ -th ingredient is not selected, S_i and V_i are set to be 0. Otherwise, S_i is equal to 1 and V_i is constrained by its lower and upper bounds (VL_i and VU_i).

$\sum_i V_i = 1$ (19)

$VL_i \bullet S_i \leq V_i \leq VU_i \bullet S_i$ (20)

Their values as well as the properties of 48 candidates (e.g., density, toxicity, etc.) are given in Table S2 in Supporting Information. These are used as parameters in the optimization. Since perfume is a liquid solution, no microstructural descriptors are considered.

[insert Table 4 here]

Step 3: Model identification

The third step is to identify the models for the average sensorial rating on perfume smell (q_s) and the four target properties. The models are elaborated below.

ANN-based surrogate model for sensorial rating

A surrogate model is developed for predicting q_s. Perfume sensorial data are generated by matching the general consumers' preferences reflected in various perfume review websites. Here, the data is used to represent consumers' satisfaction. A total of 761 data samples are uploaded in https://github.com/zx2012flying/Perfume-Case-Study. These data samples only involve the 48 ingredient candidates in Table 4. For each data sample, the input data includes the selected ingredients and their volume fractions. The output data is the overall sensorial rating. For consistency, the ratings are scaled to [0, 100] with 100 denoting the best smell. The minimum and maximum ratings for these samples are 50.2 and 89.7, respectively. Based on these data, several surrogate models such as linear regression, artificial neuron network (ANN), and support vector regression are built using the Surrogate Modeling Toolbox, Pyrenn, and Scikit-learn packages in Python 3.6. The hyperparameters are tuned manually and the model accuracy is evaluated through 10-fold cross validation. A three-layer ANN model (i.e., one input layer, one hidden layer, and one output layer) was found to offer the highest accuracy. Figure S1 shows the schematic structure of the ANN model. The tansig and purelin functions are applied in the hidden and output layer, respectively. The number of neurons in the hidden layer is tuned to be 8. Figure 4 presents the histogram of the absolute errors between the true values and predicted values ($q_s^{\text{true}} - q_s^{\text{pre}}$). 90% of the deviations are less than 10.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

The mean average error (MAE) and mean average percentage error (MAPE) are equal to 4.8 and 6.9%, respectively. This ANN model provides an accurate prediction of q_s, which is explicitly expressed as

$q_s = \sum_{l=1}^{8} wo_l \bullet \quad f_h(ah_l) + bo$ (21)

$f_h\left(ah_l\right) = 1 - \frac{2}{1+e^{2\times ah_l}}, \quad l = 1, \ldots, 8$(22)

$ah_l = \sum_{i=1}^{48} \mathrm{wh}_{l,\ i} \bullet V_i + bh_l, \quad l = 1, \ldots, 8$(23)

where wo_l and bo are the weights and bias in the output layer, respectively. f_h is the tansig function in the hidden layer. ah_l is the intermediate variable in the hidden layer. $\mathrm{wh}_{l,\ i}$ and bh_l are the weights and biases in the hidden layer, respectively. These model parameters are provided in the Github platform mentioned above.

[insert Figure 4 here]

$\mathbf{LD_{50}}$

LD_{50} of the perfume solution is calculated by Eq. 24. It depends on the toxicity of ingredients ($LD_{50,i}$) and the mass fraction (m_i) converted from the volume fraction V_i.

$LD_{50} = \frac{1}{\sum_{i=1}^{48} \frac{m_i}{LD_{50,i}}}$ (24)

$m_i = \frac{V_i \bullet \rho_i}{\sum_{j=1}^{48} V_j \bullet \rho_j}$(25)

where $LD_{50,i}$ and the density (ρ_i) for the 48 ingredient candidates are given in Table S2.

Flash point

The flash point (T_{fp}) of a flammable liquid mixture can be theoretically determined based on the Le Chatelier's mixing rule.[53]

$\sum_{i=1}^{48} \frac{\mathrm{FP}P_i}{\mathrm{FPLF}L_i} = 1$ (26)

where $\mathrm{FP}P_i$ and $\mathrm{FPL}FL_i$ are the partial pressure and lower flammability limit of the i -th ingredient candidate at the flash point, respectively. $\mathrm{FPL}FL_i$ is calculated by

$\mathrm{FPL}FL_i = LFL_i^* - \frac{0.182\times(T_{\mathrm{fp}}-298)}{Hc_i}$(27)

where Hc_i and $\mathrm{LF}L_i^*$ are the heat of combustion and lower flammability limit at 298 K (see Table S2), respectively.$\mathrm{FP}P_i$ is calculated via the vapor-liquid equilibrium in Eq. 28. The UNIFAC model is used to calculate the activity coefficient$\Phi\Pi\Upsilon_i$ at the flash point. The mole fraction x_i is converted from mass fraction m_i. FPPsat_i is the saturated vapor pressure at flash point, which is calculated using the Antoine equation in Eq. 31.

$\mathrm{FP}P_i \ = \Phi\Pi\Upsilon_i \bullet x_i \bullet FPP\mathrm{sat}_i$(28)

$\Phi\Pi\Upsilon_i = f_{\mathrm{unifac}}\left(x_i, T_{\mathrm{fp}}\right)$(29)

$x_i = \frac{m_i}{MW_i \bullet \sum_j \frac{m_j}{MW_j}}$(30)

$\mathrm{FPPsat}_i = A_i - \frac{B_i}{C_i + T_{\mathrm{fp}}}$(31)

The molecular weight MW_i, UNIFAC parameters, and Antoine coefficients A_i, B_i, and C_i for the 48 ingredient candidates are given in Table S2.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Homogeneous solution

To ensure a homogeneous solution, the volume of selected organic fragrances must be less than their volume solubility ($SV_{i,ew}$) in the ethanol-water solvent system.

$$\frac{V_i}{(V_{47}+V_{48})} \leq SV_{i,ew}, \quad i = 1, \ldots, 46(32)$$

It is found that it is quite hard to calculate $SV_{i,ew}$ using rigorous thermodynamic models due to the many missing parameters. In the literature, several short-cut models have been developed to predict$SV_{i,ew}$. The log-linear mixture rule below is widely used.[54]

$$\log \mathrm{SV}_{i,ew} = \log SV_{i,w} + \beta \bullet \log \frac{SV_{i,e}}{SV_{i,w}}, \quad i = 1, \ldots, 46(33)$$

$$\beta = \frac{V_{47}}{V_{47}+V_{48}} \ (34)$$

$$\log \frac{SV_{i,e}}{SV_{i,w}} = M \bullet \log K_{ow,i} + N, \quad i = 1, \ldots, 46(35)$$

where $SV_{i,e}$ and $SV_{i,w}$ are the volume solubility in ethanol and water, respectively. $K_{ow,i}$ is the n-octanol/water partition coefficient of the i -th candidate. M and N are the cosolvent constants. Based on experimental data, their values have been regressed as 0.81 and 0.85, respectively.

Odor type in top note

The fragrance molecules in a perfume solution first evaporate into the air through the liquid-gas interface. Then, the molecules diffuse in the air (assumed to be stagnant) and are detected at certain distance away. The processes of evaporation, diffusion, and detection have been modelled using chemical engineering principles and psychophysics.[38,52,55] Perfume evaporation is simulated using Eq. 36 with an initial condition. The liquid molar changes are equal to the moles of ingredients transported through the interface (i.e., $z = 0$).

$$\frac{dn_{i,t}}{\mathrm{dt}} = C_T \bullet D_i \bullet A_{\mathrm{lg}} \bullet \left.\frac{\partial y_{i,t,z}}{\partial z}\right|_{z=0}(36)$$

Initial condition: $n_{i,t=0} = n_p \bullet x_i$

After discretization, Eq. 37 is obtained.

$$\frac{n_{i,t+t}-n_{i,t}}{t} = C_T \bullet D_i \bullet A_{\mathrm{lg}} \bullet \frac{y_{i,t,z=z_1}-y_{i,t,z=0}}{z_1}(37)$$

where n_p is the initial number of moles of perfume solution.$C_T = P/RT$ is a constant. D_i and A_{lg} are the diffusivity of i -th candidate and interfacial area, respectively.t and $z_1$are the time interval and the first distance interval, respectively. These parameters are given in Table S2.$n_{i,t}$ is the number of moles of the i -th candidate in the liquid at time t. $y_{i,t,z}$ is the molar fraction of i -th ingredient candidate in the air at time t at distance z . It is calculated via vapor-liquid equilibrium.

$$y_{i,t,z=0} = \gamma_{i,t} \bullet x_{i,t} \bullet \frac{\mathrm{Psat}_i}{P}(38)$$

$$\gamma_{i,t} = f_{\mathrm{unifac}}(x_{i,t}, T_r) \ (39)$$

$$x_{i,t} = \frac{n_{i,t}}{\sum_{i=1}^{48} n_{i,t}} \ (40)$$

where $\gamma_{i,t}$ and $x_{i,t}$ are the activity coefficient and mole fraction of i -th ingredient candidate at time t , respectively. Psat_i is the saturated vapor pressure at room temperature $T_r = 298\ K$.

After evaporation, fragrance diffusion is modelled based on Fick's 2nd law of diffusion with one initial condition and two boundary conditions (Eq. 41).

$$\frac{\partial y_{i,t,z}}{\partial t} = D_i \bullet \frac{\partial^2 y_{i,t,z}}{\partial z^2}(41)$$

Initial condition: $y_{i,t=0,z} = 0$

Boundary conditions: Eq. 38, $y_{i,t,z=z_{\max}} = 0$

The initial condition assumes that no fragrances exist in the air before diffusion begins (i.e., $t = 0$). The boundary conditions indicate that vapor-liquid equilibrium is maintained at the interface at any time (i.e., Eq. 38) and no fragrances exist beyond the maximum distance ($z_{\max} = 2m$). This model is discretized using a non-uniform distance grid (Table S2) for reducing the computational difficulty. After discretization, we get

$\frac{y_{i,t+t,z}-y_{i,t,z}}{t} = D_i \bullet \frac{\frac{y_{i,t,z+z_{j+1}}-y_{i,t,z}}{z_{j+1}} - \frac{y_{i,t,z}-y_{i,t,z-z_j}}{z_j}}{0.5\times(z_{j+1}+z_j)}, z \in [0, z_{\max}]$ (42)

where z_j and z_{j+1} are the distance intervals, respectively.

Any fragrance with a different concentration leads to a different intensity. Many theoretical models (e.g., Weber-Fenchner law, power law, and linear law) have been proposed for quantifying odor intensity. The power law is chosen here because it fits experimental data well. The intensity of the i -th odorant is defined as the ratio of its concentration in the air (c_i in g/m^3) to its odor recognition threshold value (ORT_i), raised to a poweroe_i.[52] With this, the odor intensity in the top note is determined based on the mole fraction of fragrances in the air at 5 minutes (t_{tn}) after application at a distance of 0.2 m (z_{tn}).

$\psi_i = \left(\frac{c_i}{\text{OR}T_i}\right)^{oe_i}$ (43)

$c_i = y_{i,t_{\text{tn}},z_{\text{tn}}} \bullet MW_i \bullet C_T$(44)

Given multiple odorants, the one with the highest intensity is more strongly sensed and can be regarded as the major odor type. Thus, the dominant odor type in top note is expressed as

$OTTN = i, \quad if\ \psi_i = \psi_{\max} = \{\psi_i\}$(45)

Heuristics

Following Table 3, constraints for the Eau de parfum formulation are derived from dozens of modern Eau de parfum recipes.[51]It is found that the suggested number of ingredients for each fragrance note can be represented by Eq. 46-48. Eq. 49 shows that Eau de parfum usually contains 10-20% organic fragrances. The suggested volumetric proportions for top note and middle note are 15-25% and 30-40%, respectively (Eq. 50-51). The suggested volume fraction of water is 9-13%.[49,52]

$3 \leq \sum_{i=1}^{17} S_i \leq 6$ (46)

$3 \leq \sum_{i=18}^{33} S_i \leq 6$ (47)

$2 \leq \sum_{i=34}^{46} S_i \leq 5$ (48)

$0.1 \leq \sum_{i=1}^{46} V_i \leq 0.2$ (49)

$0.15 \bullet \sum_{i=1}^{46} V_i \leq \sum_{i=1}^{17} V_i \leq 0.25 \bullet \sum_{i=1}^{46} V_i$(50)

$0.3 \bullet \sum_{i=1}^{46} V_i \leq \sum_{i=18}^{33} V_i \leq 0.4 \bullet \sum_{i=1}^{46} V_i$(51)

$0.09 \leq V_{48} \leq 0.13$ (52)

Iterative Model Adoption and Optimization Solution Strategy

The identified rigorous mechanistic models for LD_{50}, flash point, and odor type, the short-cut model for transparency, the surrogate model for sensorial rating as well as the heuristics in Eq. 46-52 are integrated to form the perfume formulation problem below.

q_s (53)

s.t. Eq. 21-23 ANN-based surrogate model for q_s

Eq. 16-18 Design targets

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Eq. 24-45 Mechanistic models

Eq. 46-52 Heuristics

Eq. 19-20 Design variables

This problem is implemented in GAMS 24.7 on a laptop with Intel 3.30 GHz CPU. The global solver BARON is used first and then the local solver SBB is employed if no optimal solutions are obtained from BARON.

GDP reformulation

Because of the complexity of the identified models and the number of intermediate variables, the problem is directly programmed using GDP. The disjunction is explicitly expressed as

$$\begin{bmatrix} Y_i \\ VL_i \leq V_i \leq VU_i \\ Eq.25 \\ Eq.27-31 \\ Eq.37-44 \end{bmatrix} \bigvee \begin{bmatrix} \neg Y_i \\ V_i = 0 \\ m_i = 0 \\ \text{FPLF}L_i, FPP_i, FP\gamma_i, x_i, FPPsat_i = 0 \\ n_{i,t}, y_{i,t,z}, \gamma_{i,t}, x_{i,t}, c_i, \psi_i = 0 \end{bmatrix} \quad (54)$$

The GDP problem is further reformulated using the big-M approach with the solver JAMS and then solved by SBB. Different initial guesses are utilized. The second column of Table 5 lists the computational statistics. It contains 46 discrete variables, 9783 single variables, and 18230 equations. It takes 3459 seconds to obtain a local optimal solution.

[insert Table 5 here]

The perfume formula obtained is shown in the second column of Table 6. The maximum sensorial rating is 92.4. The new perfume consists of 3 fragrances in top note, 4 fragrances in middle note, and 3 fragrances in base note. Their volume fractions vary in the range of 0.3-1.9% and the total volume fraction of fragrances is 10.1%. Furthermore, this recipe fulfills the four design targets. The LD_{50} and flash point are 6815 mg/kg and 15.1 °C, respectively. These are higher than their lowest acceptable design targets (5000 mg/kg and 15 °C in Eq. 17-18). The volume fractions of the 10 fragrances are less than their volume solubility shown in Table S3. Thus, a homogeneous and transparent perfume solution can be obtained. Figure 5a shows the odor profile during the first 350 seconds. The intensity of benzyl acetate (jasmine-like) and octyl acetate (apple-like) are 2.6 and 0.5, respectively. The limonene with a lemon-like odor has the maximum intensity of 2.8 after 5 minutes. Note that since the odor intensities of other fragrances are much less than those of top note fragrances, they are not shown in the figure. Figure 6a shows the simulated diffusion profile of top note fragrances within 2 meters at 5 minutes. Obviously, the maximum intensities are located at $z = 0$. The maximal intensity of limonene can reach 33.5. As the distance increases, the intensity decreases and down to zero beyond 2 meters. The simulated diffusion profiles of 4 middle note fragrances at 1 hour and 3 base note fragrances at 5 hours are illustrated in Figure S2a and S2b, respectively.

[insert Figure 5 here], [insert Figure 6 here], and [insert Table 6 here]

Model substitution

The above result from GDP (92.4 in Table 6) is slightly larger than the maximal sensorial rating (89.7) of the original 761 data samples. Although a local optimal solution has already been obtained, the GDP problem

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

is still challenging to solve. In fact, the choice of the initial values greatly affects whether feasible solutions can be obtained and the quality of local solution. It is found that the major computational difficulties come from the rigorous mechanistic models for perfume evaporation (Eq. 36) and diffusion (Eq. 41), which requires the handling of many highly nonlinear equations. For instance, the vapor-liquid equilibrium and UNIFAC equations must be calculated at every time point (i.e., Eq. 38-40). Thus, in order to solve the formulation problem more efficiently and find better solutions, model substitution is employed here.

Whether the top note of a perfume can be dominated by a lemon-like or non-lemon-like scent is a binary decision. Thus, the prediction of the odor type can be transformed into a classification problem. In other words, the complex mechanistic models (Eq. 36-45) for predicting the odor type in the top note is substituted by a classification-based surrogate model. To do so, random sampling is applied to generate 15000 artificial perfume recipes that account for the heuristic rules in Eq. 46-52. Among them, 7500 recipes consist of 0.25-0.75% limonene (lemon-like), 5000 recipes contain 0.75-1.25%, and 2500 recipes have 1.25-1.75%. These recipes are used as the input data. For each recipe, their odor intensities in the top note are calculated using Eq. 36-45. If a lemon-like odor has the highest intensity, the output is set equal to 1. Otherwise, it is equal to 0. Then, a support vector classification (SVC) model with linear kernel function is trained. Through 10-fold cross validation, the hyperparameter C indicating the regularization strength is tuned to be 10. Figure S3 presents the classification error distribution. For the 7500 data samples containing 0.25-0.75% limonene, the classification accuracy is 93.3%. For the other half samples, the accuracy is 98.9 %. The overall accuracy is 96.1%. These statistics indicate that this SVC model can serve as a relatively simple surrogate for substituting the original complex mechanistic models. The SVC model consists of 2126 support vectors and is expressed as

$OTTN = \sum_{c=1}^{2126} \alpha_c \bullet K_c + bs$ (55)

$K_c = \sum_{i=1}^{48} SV_{c,i} \bullet \mathrm{VN}_i$ (56)

$VN_i = \frac{V_i - V_{i,min}}{V_{i,max} - V_{i,min}}$ (57)

where α_c and bs are the weights for support vector and a constant, respectively. $SV_{c,i}$ is the support vector.$V_{i,max}$ and $V_{i,min}$ are normalization coefficients. These parameters are optimized automatically during the training process and provided in the Github platform mentioned above.

By substituting Eq. 36-45 with Eq. 55-57, the resulting perfume formulation problem (MINLP-SVC) is solved using the global solver BARON. Table 5 shows the computational statistics. It consists of 2860 single variables, 2920 equations, and 2928 nonlinear matrix entries. Clearly, the problem size and nonlinearity are much less than those of the GDP problem. It takes 143 seconds to obtain the global solution given in the last column of Table 6. The maximum sensorial rating is 98.3 which is better than the GDP result. The new perfume formula consists of 13 different fragrances in different volume fractions. The total volume fraction of fragrances is 20%. Moreover, the design targets onLD_{50} and flash point are fulfilled. As listed in Table S4, all the ingredient's volume fractions are less than their volume solubility in the ethanol-water solvent. In addition, the major odor type in the top note is classified as 1 (i.e., lemon-like) by the SVC model. As validated using the original mechanistic models (Eq. 36-45), Figure 5b shows the odor intensity in the first 350 seconds. Again, only the top note fragrances are plotted. It is clear that the lemon-like fragrance limonene has the maximum odor intensity (around 3.5) which is higher than those of other fragrances. This validates the SVC results as well. In addition, Figure 6b shows the diffusion profile of 4 top note fragrances at 5 minutes, which is simulated using the original mechanistic models. Figure S4a and S4b present the simulated diffusion of 5 middle note fragrances at 1 hour and 4 base note fragrances at 5 hours, respectively.

Conclusion

This paper presents a new optimization-based approach for cosmetic formulation. A three-step procedure is proposed to formulate the cosmetic formulation problem as an MINLP problem. For problem definition,

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

the objective function (i.e., sensorial perception) and design targets are identified. Then, a pool of potential ingredient candidates is generated for selection. Design variables include ingredient selection, composition, and microstructure descriptors (not include in the example). Next, models are identified for predicting the sensorial rating and target properties. Meanwhile, common heuristics are translated into mathematical equations which serve as constraints to narrow down the search space. To improve the optimization convergence and to find better solutions, a solution strategy that involves an iterative model adoption and different numerical techniques is proposed. The procedure and solution strategy are illustrated using a perfume case study. Our approach is one of the first attempts to integrate multiple (rigorous, short-cut, surrogate, and heuristic-based) models to account for both sensorial and functional attributes for optimal cosmetic formulation. It can be used for other cosmetics and personal care products provided that the relevant models, data, heuristics, etc. are available.

Product design involves a wide range of issues that include consumer preference, ingredient selection, supply chain analysis, process design, government regulations, economics, corporate social responsibility, sustainability and so on.[56] These issues interact in an exceedingly complex manner as captured in the Grand Product Design Model.[57] While many detailed models exist to describe the separate issues, it is a daunting task to solve the optimization problem for product design when a number of disparate issues are involved. It is interesting to study how the approach described in this paper can be extended to the product design as a whole. Efforts in this direction are underway.

Literature cited

1. Zhang L, Babi DK, Gani R. New vistas in chemical product and process design. *Annual Review of Chemical and Biomolecular Engineering* . 2016;7(1):557-582.

2. Gani R, Ng KM. Product design – Molecules, devices, functional products, and formulated products. *Computers & Chemical Engineering* . 2015;81:70-79.

3. Conte E, Gani R, Ng KM. Design of formulated products: A systematic methodology. *AIChE Journal* . 2011;57(9):2431-2449.

4. *Global Cosmetics Products Market* . 360 research reports. 2018. https://www.360researchreports.com/global-cosmetics-products-market-13100793 (accessed March 2020)

5. Kontogeorgis GM, Mattei M, Ng KM, Gani R. An integrated approach for the design of emulsified products. *AIChE Journal* . 2019;65(1):75-86.

6. Pensé-Lheritier A-M. Recent developments in the sensorial assessment of cosmetic products: a review. *International Journal of Cosmetic Science* . 2015;37(5):465-473.

7. Benson HAE, Roberts MS, Leite-Silva VR, Walters KA. *Cosmetic Formulation Principles and Practice* . 1st ed. Florida, USA: CRC Press; 2019.

8. Taifouris M, Martin M, Martinez A, Esquejo N. Challenges in the design of formulated products: multiscale process and product design.*Current Opinion in Chemical Engineering* . 2020;27:1-9.

9. Karunanithi AT, Achenie LEK, Gani R. A new decomposition-based computer-aided molecular/mixture design methodology for the design of optimal solvents and solvent mixtures. *Ind Eng Chem Res* . 2005;44(13):4785-4797.

10. Austin ND, Sahinidis NV, Konstantinov IA, Trahan DW. COSMO-based computer-aided molecular/mixture design: A focus on reaction solvents.*AIChE Journal* . 2018;64(1):104-122.

11. Jonuzaj S, Adjiman CS. Designing optimal mixtures using generalized disjunctive programming: Hull relaxations. *Chemical Engineering Science* . 2017;159:106-130.

12. Jonuzaj S, Akula PT, Kleniati P-M, Adjiman CS. The formulation of optimal mixtures with generalized disjunctive programming: A solvent design case study. *AIChE Journal* . 2016;62(5):1616-1633.

13. Zhang L, Kalakul S, Liu L, Elbashir NO, Du J, Gani R. A computer-aided methodology for mixture-blend design. applications to tailor-made design of surrogate fuels. *Ind Eng Chem Res* . 2018;57(20):7008-7020.

14. Kalakul S, Zhang L, Fang Z, et al. Computer aided chemical product design – ProCAPD and tailor-made blended products. *Computers & Chemical Engineering* . 2018;116:37-55.

15. Yunus NA, Gernaey KV, Woodley JM, Gani R. A systematic methodology for design of tailor-made blended products. *Computers & Chemical Engineering* . 2014;66:201-213.

16. Marvin WA, Rangarajan S, Daoutidis P. Automated generation and optimal selection of biofuel-gasoline blends and their synthesis routes.*Energy Fuels* . 2013;27(6):3585-3594.

17. Liu Q, Zhang L, Liu L, et al. OptCAMD: An optimization-based framework and tool for molecular and mixture product design.*Computers & Chemical Engineering* . 2019;124:285-301.

18. Jones O, Ben Selinger A. The chemistry of cosmetics. https://www.science.org.au/curious/people-medicine/chemistry-cosmetics (access March 2020).

19. Omidbakhsh N, Duever TA, Elkamel A, Reilly PM. A systematic computer-aided product design and development procedure: Case of disinfectant formulations. *Ind Eng Chem Res* . 2012;51(45):14925-14934.

20. Smith BV, Ierapepritou M. Framework for consumer-integrated optimal product design. *Ind Eng Chem Res* . 2009;48(18):8566-8574.

21. Bagajewicz M, Hill S, Robben A, et al. Product design in price-competitive markets: A case study of a skin moisturizing lotion.*AIChE Journal* . 2011;57(1):160-177.

22. Conte E, Gani R, Cheng YS, Ng KM. Design of formulated products: Experimental component. *AIChE Journal* . 2012;58(1):173-189.

23. Zhang L, Fung KY, Zhang X, Fung HK, Ng KM. An integrated framework for designing formulated products. *Computers & Chemical Engineering* . 2017;107:61-76.

24. Arrieta-Escobar JA, Bernardo FP, Orjuela A, Camargo M, Morel L. Incorporation of heuristic knowledge in the optimal design of formulated products: Application to a cosmetic emulsion. *Computers & Chemical Engineering* . 2019;122:265-274.

25. Zhang X, Zhou T, Zhang L, Fung KY, Ng KM. Food product design: A hybrid machine learning and mechanistic modeling approach. *Ind Eng Chem Res* . 2019;58(36):16743-16752.

26. Zhang L, Mao H, Liu L, Du J, Gani R. A machine learning based computer-aided molecular design/screening methodology for fragrance molecules. *Computers & Chemical Engineering* . 2018;115:295-308.

27. Goyal S, Goyal EK. Cascade and feedforward backpropagation artificial neural networks models for prediction of sensory quality of instant coffee flavoured sterilized drink. *Canadian Journal on Artificial Intelligence, Machine Learning and Pattern Recognition* . 2011;2(6):7882.

28. Baki G, Alexander KS. *Introduction to Cosmetic Formulation and Technology* . John Wiley & Sons; 2015.

29. Personal Care Products Council. Cosmetic ingredient dictionary. https://cosmeticsinfo.org/Ingredient-dictionary (accessed March 2020).

30. European Commission. CosIng database. https://ec.europa.eu/growth/sectors/cosmetics/cosing_en (accessed March 2020).

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

31. Gani R, Hytoft G, Jaksland C, Jensen AK. An integrated computer aided system for integrated design of chemical processes.*Computers & Chemical Engineering* . 1997;21(10):1135-1146.

32. Cardona Jaramillo JEC, Achenie LE, Alvarez OA, Carrillo Bautista MP, Gonzalez Barrios AF. The multiscale approach to the design of bio-based emulsions. *Current Opinion in Chemical Engineering* . 2020;27:65-71.

33. Bernardo FP, Saraiva PM. A conceptual model for chemical product design. *AIChE Journal* . 2015;61(3):802-815.

34. Wibowo C, Ng KM. Product-oriented process synthesis and development: Creams and pastes. *AIChE Journal* . 2001;47(12):2746-2767.

35. Wibowo C, Ng KM. Product-centered processing: Manufacture of chemical-based consumer products. *AIChE Journal* . 2002;48(6):1212-1230.

36. Kim SH, Boukouvala F. Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques. *Optim Lett* . May 2019.

37. Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. *Computers & Chemical Engineering* . 2018;108:250-267.

38. Teixeira MA, Rodriguez O, Mata VG, Rodrigues AE. The diffusion of perfume mixtures and the odor performance. *Chemical Engineering Science* . 2009;64(11):2570-2589.

39. Bronaugh RL, Maibach HI. *Percutaneous Absorption: Drugs–Cosmetics–Mechanisms–Methodology* . 3rd ed. New York, USA: Marcel Dekker, Inc.; 1999.

40. Hada S, Herring RH, Eden MR. Mixture formulation through multivariate statistical analysis of process data in property cluster space. *Computers & Chemical Engineering* . 2017;107:26-36.

41. Hill M. Product and process design for structured products.*AIChE Journal* . 2004;50(8):1656-1661.

42. Schweidtmann AM, Mitsos A. Deterministic global optimization with artificial neural networks embedded. *J Optim Theory Appl* . 2019;180(3):925-948.

43. Schweidtmann AM, Huster WR, Luthje JT, Mitsos A. Deterministic global process optimization: Accurate (single-species) properties via artificial neural networks. *Computers & Chemical Engineering* . 2019;121:67-74.

44. Grossmann IE, Trespalacios F. Systematic modeling of discrete-continuous optimization models through generalized disjunctive programming. *AIChE Journal* . 2013;59(9):3276-3295.

45. Beykal B, Boukouvala F, Floudas CA, Pistikopoulos EN. Optimal design of energy systems using constrained grey-box multi-objective optimization. *Computers & Chemical Engineering* . 2018;116:488-502.

46. Boukouvala F, Floudas CA. ARGONAUT: AlgoRithms for Global Optimization of coNstrAined grey-box compUTational problems. *Optim Lett* . 2017;11(5):895-913.

47. Eason JP, Biegler LT. A trust region filter method for glass box/black box optimization. *AIChE Journal* . 2016;62(9):3124-3136.

48. Craig S. How to review fragrance? https://bespokeunit.com/fragrance/formula/ (assessed March 2020).

49. Mata VG, Gomes PB, Rodrigues AE. Engineering perfumes. *AIChE Journal* . 2005;51(10):2834-2852.

50. Shcherbakov D, Massebeuf S, Normand V. Flash-point prediction of fragrances or flavours accounting for non-ideality of the liquid phase.*Flavour and Fragrance Journal* . 2019;34(1):63-69.

51. Poucher WA. *Perfumes, Cosmetics and Soaps: Vol. II, the Production, Manufacture and Application of Perfumes* . 9th ed. Dordrecht: Springer Science; 1993.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

52. Teixeira MA, Rodriguez O, Rodrigues AE. The perception of fragrance mixtures: A comparison of odor intensity models. *AIChE Journal* . 2010;56(4):1090-1106.

53. Liaw H-J, Gerbaud V, Li Y-H. Prediction of miscible mixtures flash-point from UNIFAC group contribution methods. *Fluid Phase Equilibria* . 2011;300(1):70-82.

54. Jouyban A. Review of the cosolvency models for predicting drug solubility in solvent mixtures: An update. *Journal of Pharmacy & Pharmaceutical Sciences* . 2019;22:466-485.

55. Teixeira MA, Rodriguez O, Mota FL, Macedo EA, Rodrigues AE. Evaluation of group-contribution methods to predict VLE and odor intensity of fragrances. *Ind Eng Chem Res* . 2011;50(15):9390-9402.

56. Ng KM, Gani R. Chemical product design: Advances in and proposed directions for research and teaching. *Computers & Chemical Engineering* . 2019;126:147-156.

57. Fung KY, Ng KM, Zhang L, Gani R. A grand model for chemical product design. *Computers & Chemical Engineering* . 2016;91:15-27.

Figure 1. The General Methodology of Optimization-based Cosmetic Formulation

Hosted file

`image1.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

Figure 2. Systematic Procedure for Optimization Problem Formulation

Hosted file

`image2.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

Figure 3. Iterative Model Adoption and Optimization Solution Strategy

Hosted file

`image3.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

Figure 4. Absolute Error Distribution of ANN Model for Predicting Sensorial Rating

Hosted file

`image4.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

Figure 5. Odor Profile of Top Note Fragrance Calculated Using Rigorous Mechanistic Models for Optimal Perfume Recipe from (a) GDP Formulation (b) MINLP-SVC Formulation

Hosted file

`image5.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

(a)

Hosted file

`image6.emf` available at `https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics`

(b)

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Figure 6. Simulated Diffusion of Top Note Fragrances at 5 Minutes for Optimal Perfume Recipe from (a) GDP Formulation (b) MINLP-SVC Formulation

Hosted file

`image7.emf` available at https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics

(a)

Hosted file

`image8.emf` available at https://authorea.com/users/322113/articles/451197-optimization-based-cosmetic-formulation-integration-of-mechanistic-model-surrogate-model-and-heuristics

(b)

Table 1. Dosage Form of Typical Cosmetic Products and the Microstructural Descriptors

Dosage form	Dosage form	Typical cosmetic products	Relevant microstructural descriptors
Solid	Stick	Lipstick, contour stick	Droplet size
	Tablet	Foundation tablet, eyeshadow	Tablet size, porosity, pore size
	Powder/granule	Facial powder, blush	Porosity, pore size, particle size and shape
Semi-solid	Paste	Facial-mask paste, skin paste	Emulsion type, droplet size
	Gel	Eye gel, aftershave gel	/
	Ointment	Hair pomade, facial scrub	Droplet size
	Cream	Hair cream, hand cream	Emulsion type, droplet size
Liquid	Lotion	Body lotion, lip gloss	Droplet size
	Suspension	Nail polish, mascara	Particle size and shape
	Solution	Perfume, makeup remover	/
Gas	Aerosol	Hair spray, shaving foam	Droplet size

Table 2. Sensorial and Functional Attributes of Four Cosmetic Products

	Sensorial attributes	Functional attributes	Relevant properties
Lipstick	sight, touch	no surface defect, hard to break, stable, safe	color, color intensity, viscosity, s
Skin cream	touch, smell, sight	moisturizing, skin protection, ease of use, stable, safe	viscosity, oiliness, odor, color, m
Perfume	smell	transparent, safe	odor intensity, odor type, flash p
Hair spray	sight, smell	effective, rapid drying, easy to remove, stable, safe	color, odor, adhesion, curl retent

Table 3. Typical Heuristics and the Translated Constraints for Formulated Products

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Heuristics
Suggested number of ingredients
Suggested number of ingredients in certain type
Ingredients with certain property is preferred
Certain ingredients cannot be used simultaneously
Certain ingredients should be used simultaneously
Suggested concentration for certain type
Suggested concentration for certain candidate
Total ingredient candidate $TIC = \{I_{A,1}, I_{A,2}, \ldots, I_{Z,z}\}$ Certain ingredient type $I_X = I_A,\ I_B, \ldots,$ or I_Z L, TL, VTL, VL_i, a

Table 4. Ingredient Types, Functions, and Ingredient Candidates for Perfume Example

Ingredient type	**Top note fragrance**	**Top note fragrance**
Function	very volatile, appear immediately to offer the first impression, and last 5-15 minutes	very volatile, appear immediately to offer the first impression, and last 5-15 minutes
Candidates	1. allyl amylglycolate (galbanum-like) 2. alpha-phellandrene (pepper-like) 3. benzylidene acetal (green leaf-like) 4. grapefruit acetal (grapefruit-like) 5. isoamyl propionate (apricot-like) 6. linayl propionate (bergamot-like) 7. methyl 2-octynoate (violet-like) 8. methyl benzoate (blackcurrant-like) 9. propyl octanoate (coconut-like)	10. amyl butyrate (pear-like) 11. benzyl acetate (jasmine-like) 12. limonene (lemon-like) 13. estragole (anise-like) 14. nerol (neroli-like) 15. nonyl aldehyde (rose-like) 16. octanal (orange-like) 17. octyl acetate (apple-like)
Ingredient type	**Middle note fragrance**	**Middle note fragrance**
Function	The body of perfume, dominate after top notes fade, and last up to 1 hour	The body of perfume, dominate after top notes fade, and last up to 1 hour
Candidates	18. amylcinnamaldehyde (jasmine-like) 19. cinnamic alcohol (cinnamon-like) 20. cyclohexylethanol (patchouli-like) 21. ethyl 4-phenylbutyrate (plum-like) 22. ethyl o-anisate (ylang ylang-like) 23. gamma-decalacetone (peach-like) 24. methyl iso-eugenol (carnation-like) 25. phenethyl isobutyrate (rose-like)	26. 1-phenylethanol (gardenia-like) 27. 2-undecanone (orris root-like) 28. amyl phenylacetate (cacao-like) 29. cedryl acetate (woody-like) 30. heliotropin (heliotrope-like) 31. lilyall (lily-like) 32. linalyl salicylate (musk-like) 33. methyl anthranilate (neroli-like)
Ingredient type	**Base note fragrance**	**Base note fragrance**
Function	Lowest volatility, appear close to the end of middle notes to offer the lasting impression, and remain several hours	Lowest volatility, appear close to the end of middle notes to offer the lasting impression, and remain several hours

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Ingredient type	**Top note fragrance**	**Top note fragrance**
Candidates	34. acetyl cedrene (woody-like) 35. alpha-ambrinol (amber-like) 36. amyl-iso-eugenol (incense-like) 37. coumarin (tonke bean-like) 38. patchouli alcohol (patchouli-like) 39. phenethyl phenylacetate (musk-like) 40. sandal hexanol (sandalwood-like)	41. benzoin (benzoin-like) 42. cedrol (cedar-like) 43. ethyl vanillin (vanilla-like) 44. maltol (caramel-like) 45. phenylacetic acid (honey-like) 46. vetiverol (vetiver-like)
Ingredient type	**Solvent**	**Solvent**
Function	Dilute organic fragrant to adjust odor release	Dilute organic fragrant to adjust odor release
Candidates	47. ethanol	48. water

Table 5. Computational Results for the Perfume Case Study

	GDP formulation	MINLP-SVC formulation
Number of discrete variables	46	46
Number of single variables	9783	2860
Number of equations	18230	2920
Number of nonlinear matrix entries	42814	2928
Solver	SBB	BARON
CPU time (s)	3459	143

Table 6. Optimal Perfume Formula Obtained from Two Optimization Formulation

	GDP form
	Recipe
Top note fragrance	benzyl ace limonene octyl aceta
Middle note fragrance	heliotropir 2-undecan 1-phenylet lilyall
Base note fragrance	sandal hex benzoin coumarin
Solvent	ethanol water
LD_{50}	6815 mg/l
Flash point	15.1 °C
Homogeneous	All fragra
OTTN	Lemon-lik
Sensorial rating	92.4

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

Posted on Authorea 14 May 2020 — CC-BY 4.0 — https://doi.org/10.22541/au.158948936.63110621 — This a preprint and has not been peer reviewed. Data may be preliminary.

			GDP form
*Volume fraction in percentage (%)	**Volume solubility data in Table S3	***Volume solubility data in Table S4	*Volume f