2018

A Statistical Framework for the Development of Prediction and Clustering Models in the Hazard Assessment of Nanomaterials

Ying Pei

A Statistical Framework for the Development of Prediction and Clustering Models in the Hazard Assessment of Nanomaterials

Ying Pei

A dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Industrial Engineering

Feng Yang, Ph.D., Chair
Majid Jaridi, Ph.D.
Bhaskaran Gopalakrishnan, Ph.D.
Xinjian He, Ph.D.
Xi Chen, Ph.D.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia
2018

Keywords: Risk Assessment, Variable Selection, Design of Experiments, Prediction Model, Kriging, Shape Clustering, Nanotoxicology

ABSTRACT

A Statistical Framework for the Development of Prediction and Clustering Models in the Hazard Assessment of Nanomaterials

Ying Pei

Compared to conventional materials or chemicals, it remains challenging to fully assess the hazards of various nanomaterials (NMs) stemming from their physicochemical properties. Tremendously large variety of NMs calls for high-throughput screening methods, and the ultimate goal of nanotoxicology is to develop prediction models, also referred to as QNAR (quantitative nanostructure-activity relationships), which relate the adverse bioactivity effects of NMs to their physicochemical properties. Such models enable the prediction of a new NM's toxicity without performing additional biological experiments, which leads to substantial savings in time and money. For the efficient development of prediction models, a statistical framework is provided and demonstrated in this dissertation.

There are four stages of analysis and modeling in this framework: variable selection, design of experiments, quantitative modeling and shape clustering. Variable selection is first performed on existing nanotoxicology data to identify the important predictors, most of which are materials' physicochemical properties, for NMs' toxicity. Then design of experiments is carried out in the space of identified predictors for efficient data collection. Third, stochastic kriging with qualitative factors (SKQ) is employed to model the relationship between predictors and toxicity responses for the development of prediction models. Lastly, shape clustering methods are adapted to cluster NMs based on their toxicity profiles.

This framework has been illustrated by a simulation case derived from a nanotoxicology database including 25 in-vivo studies for 1899 rodent animals.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Feng Yang for her great guidance and continuous support throughout my study at West Virginia University. It has been a great honor to be her Ph.D. student. Without her inspiration, it wouldn't be possible for me to finish the probably most challenging thing, completion of Ph.D., of the first 26 years of my life. I am also thankful to Dr. Majid Jaridi, Dr. Bhaskaran Gopalakrishnan, Dr. Xinjian He and Dr. Xi Chen for serving on my committee, and for their insightful ideas and assistance in preparing this dissertation.

Contents

List of Figures

List of Tables

List of Acronyms

NMs Nanomaterials

QNAR Quantitative Nanostructure-Activity Relationships

ASM All Subsets Models

BMD Benchmark Dose

ED Estimation Data

VD Validation Data

DBCAU Density-Based Clustering Algorithm with Uncertainty

KSCAU K-Shape Clustering Algorithm with Uncertainty

DOE Design of Experiments

GA Genetic Algorithm

Inh Inhalation

IT Instillation

LASSO Least Absolute Shrinkage and Selection Operator

MEM Mixed Effects Modeling

NM Nanomaterial

PA Pharyngeal Aspiration

PMN Polymorphonuclear Leukocyte

SKQ Stochastic Kriging with Qualitative Factors

SLHD Sliced Latin Hypercube Design

SW StepWise

Chapter 1

Introduction

1.1 Motivation of the Research

Commercialization of nanotechnology is giving rise to extensive applications of nanomaterials (NMs) in areas such as environment [1, 2, 3, 4], energy [5, 6, 7, 8] and biomedicine [9, 10, 11, 12]. The large introduction of engineered NMs to the market [13] will inevitably lead to increasing exposure of humans and environment to NMs, which presents social concerns about the potential hazard of NMs [14]. Therefore, developing appropriate hazard assessment approaches for NMs is under urgent and rising need to ensure the sustainable exploitation of NMs [15]. Hazard assessment methods and tools applied to conventional chemicals and materials may not be readily and suitably adaptable for NMs due to insufficient capacity to fully assess risks associated with all NMs [16, 17, 18, 19]. The rapid proliferation of different types of NMs means that efforts to test toxicity of these variants through traditionally biological experimentation will be burdensome, even infeasible.

How to reduce the work of testing the large and growing number of NMs, and eventually achieve the transition out of animal testing paradigm for huge variety of newly emerging NMs? To address this question, a statistical framework is proposed which gives birth to the prediction and clustering models. The prediction model is a mathematical description of how the adverse bioactivity effects are functionally related to the identifiers of NM. Herein, the scope of identifiers covers physicochemical properties of NMs, experimental conditions and study scenarios [20, 21]. The adverse bioactivity effects refer to the toxic endpoints which may vary in different context [22]. The prediction model enables the quantitative assessment

of a new NM's bioactivity behavior without performing biological experiments on it. The clustering model is able to cluster NMs based on their entire bioactivity profiles.

1.2 Plan and Objective of the Research

The proposed framework consists of four portions given as follows.

Variable Selection. Based on the real nanotoxicology database constructed from 25 in-vivo studies, variable selection is performed to identify the important factors which are mainly responsible for variations in the toxicity response.

Design of Experiments. In the space formed by the important variables identified above, design of experiments (DOE) is carried out for efficient data collection.

Quantitative Modeling. Based on the sample data collected following the experimental design, a kriging model named stochastic kriging with qualitative factors (SKQ) is employed to model these data and generate the prediction model quantifying NMs' bioactivity profiles.

Shape Clustering. Cluster the different types of NMs based on their bioactivity profiles, the estimation and inference of which are rendered by SKQ modeling.

The objective of this work is to develop a statistical framework for the establishment of prediction and clustering models. Given a new material's physicochemical properties, the prediction model is able to estimate its bioactivity profile, based on which the clustering model is then employed to assign the material to a potency group (Figure 1.1). The outputs of this framework are threefold. First, a list of important variables contributing to the NM toxicity is generated which potentially provides guidance for the future design and production of safer NMs [23, 24, 25]. Second, a quantitative prediction model is built, which fulfills the goal of predicting the biological effects (toxicity) of a NM from its physicochemical properties without performing biological experiments on it. Third, the clustering model is

able to classify new materials based on their entire predicted bioactivity profiles [26, 27, 28].

1.3 Challenge of the Research

A number of challenges are involved in developing such a framework. First, the nanotoxicology data collected from biological experiments are relatively scarce [29, 30, 31], highly variable and subject to variance heterogeneity [32, 33, 34]. Second, the target quantitative relationship is high-dimensional, complex and nonlinear [35, 14, 36] . As mentioned in Section 1.1, NM toxicity is potentially affected by a large number of factors including various physicochemical properties. These factors include both quantitative and qualitative ones. Third, the estimated bioactivity profiles of NMs are subject to uncertainty (inherited from the randomness of biological data), and the clustering methods need to accommodate that estimation uncertainty.

1.4 Contribution of the Research

The QNAR (quantitative nanostructure-activity relationships [37]) model investigated by the framework and methods in this work is of the largest scale and most powerful prediction capability compared to those existed in the nantoxicology literature. Six material physicochemical properties (e.g., diameter, surface area and zeta potential), one animal-related factor (i.e., gender) and three exposure conditions (e.g., dose and post-exposure time) are identified from the NIOSH/CIIT/ENPRA database as having significant impacts on NM's bioactivity. These ten factors serve as predictors in the QNAR model fitted by a kriging method from simulation data mimicking the NIOSH/CIIT/ENPRA data.

Based on the QNAR estimation and inference for an new NM, shape clustering methods were adapted to classify this material into a potency category according to its entire 5-dimensional bioactivity profile rendered by the QNAR while factoring into account the

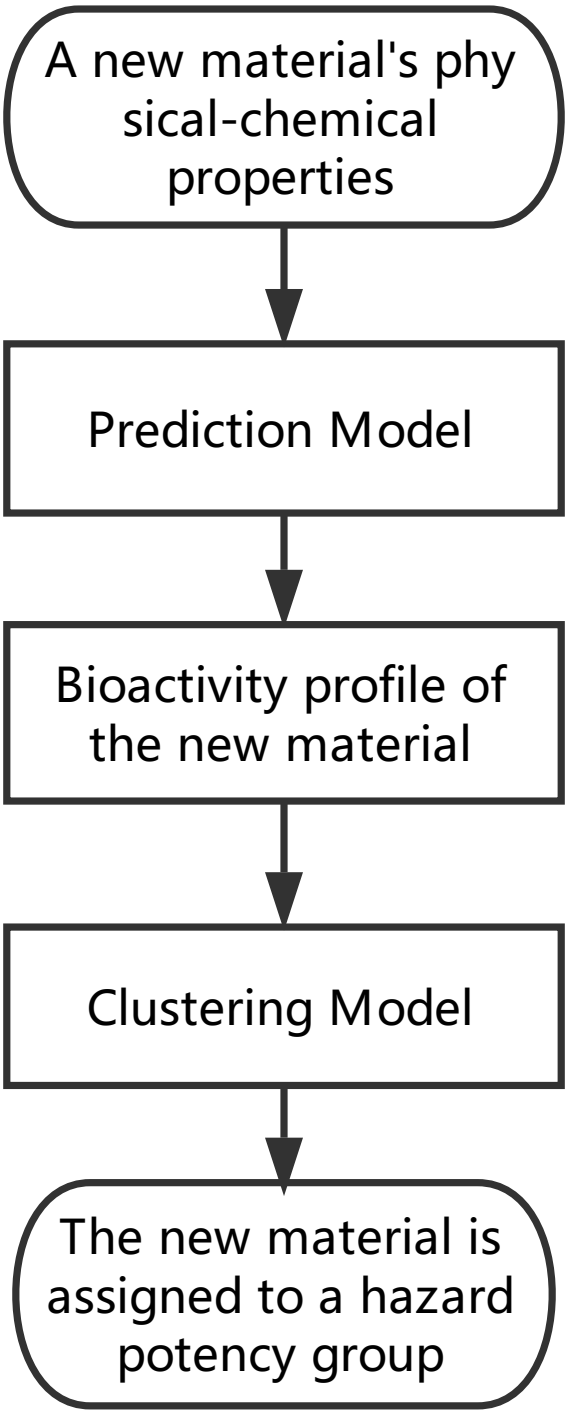

Figure 1.1: The Objective of the Framework

estimation uncertainty. To the best of our knowledge, this is the first effort to cluster NM based on their predicted high-dimensional profiles subject to uncertainty.

In addition, this work illustrates the importance of experimental design for modeling high-dimensional QNAR, the quantitative relationship between the bioactivity response and multiple (10 in our case) predictors. Design of experiments methods were developed to provide a good coverage of the 10-dimensional predictor space as well as to achieve a complex and nonlinear QNAR of high quality. The experimental design for fitting the target QNAR sheds light on the real experimental effort (samples sizes) needed to achieve a high-quality QNAR model of high dimension.

Chapter 2

Data Description

The data involved in this work is derived from the NIOSH/CIIT/ENPRA database of nanotoxicology [38]. It includes 25 in-vivo rodent studies comprising of data for 1899 unique animals, which were provided by researchers from NIOSH (National Institute for Occupational Safety and Health), CIIT (renamed Hamner Institute), and ENPRA (the European Framework 7 Program on Engineered Nano-Particle Risk Assessment). The toxicity endpoint considered is the PMNs (polymorphonuclear leukocyte cells), which is a popular measure of pulmonary inflammation. The endpoint indicator was measured in BALF (bronchoalveolar lavage fluid) extracted from the lungs of rodents and reported as counts per subject animal. To address the differences in BALF and PMNs counting methods across laboratory, the PMN percentage is used as the response of interest, which is calculated as the number of PMNs counted in the cell sample divided by the total number of cells counted.

In the complied NIOSH/CIIT/ENPRA database, there are a total of 21 input variables which can be classified as material-related variables (physicochemical properties), animal-related variables and exposure condition variables.

Material-Related Factors:

- **Material:** The database includes the study of six materials: titanium dioxide (TiO_2), iron oxide (Fe_3O_4), silver (Ag), multi-walled carbon nanotubes (MWCNT), zinc oxide (ZnO), and crystalline silica. Each material is characterized by several physicochemical properties.

- **Material Category:** It describes the type of materials according to their chemical structure. In this data set, all the materials belong to one of the three material categories: metal, metal oxide and carbon.
- **Material Manufacturer:** This factor indicates the source of materials. All NMs in this database come from eight companies or labs, e.g. Purest Colloids, Inc., Dr. Nianqiang Wu's lab, etc.
- **Structure Form:** It describes the physical structure of a material: belt, particle and tube.
- **Crystal Structure:** This is a binary factor representing whether the material has crystal structure or not.
- **Crystal Type:** It gives the specific crystal form if a material has crystal structure. Four crystal types are involved in this data set: anatase, crystalline quartz, rutile and zincite.
- **Diameter:** The median diameter of the individual material either measured or stated by manufacturer's specification. It ranges over $[7, 300]$ nm.
- **Length:** The median length of the individual material either measured or stated by manufacturer's specification. It ranges over $[0.13, 20]$ μm. The values of this factor are missing for half of the observations in data set.
- **Aerodynamic Diameter GDS and Median Aerodynamic Diameter:** The description of particle size. Only 4 out of 25 studies reported the values of them as $\{1.84, 1.94, 1.71, 2.6\}$ for aerodynamic diameter GDS and $\{1.62, 1.8, 1.44, 1.44\}$ for median aerodynamic diameter in the unit of μm.
- **Surface Area:** The specific surface area of materials measured by the BET gas adsorption method in the range of $[5.1, 322]$ m^2/g. Almost 26% of the observations of this factor are missing.

- **Zeta Potential:** It describes the degree of repulsion b/w surface charge of materials. Half of the observations in the database have zeta potential values which ranges over $[-39.5, -9.35]$ mV.
- **Density:** It is a physical property of material. This factor was not reported quantitatively. Only 5 out of 25 studies provided density of materials.
- **Contaminants:** This factor indicates whether a material contains impurity and the corresponding type of impurities. Among all the 25 studies, only one used the ZnO with impurity of Fe.
- **Modification Type:** This factor describes what modification was done for materials. There are three types of modification reported in 5 studies: purified, functionalized and coated.

Animal-Related Factors:

- **Species:** This factor indicates the species of rodent animals used in the studies: rat and mouse.
- **Gender:** The sex of animals: female and male.
- **Strain:** This factor indicates the specific strains of animals which includes Sprague-Dawley, F344, C57BL/6N, C57BL/6J, and C57BL/6-Apoetm1 across the studies.

Exposure Condition Factors:

- **Dose Metric:** A normalized deposited dose was used as the dose metric. In order to account for the differences in animal size, the deposited dose mass was normalized by the wet lung weight of the species. The dose metric ranges over $[0, 2677.78]$ in the unit of $\mu g/g$.

- **Post-Exposure Duration:** The number of days between the final exposure to NMs and the measurement of the toxicity status of the animal subject, also referred to as recovery period. It is within the range of $[0, 364]$ days.
- **Exposure Route:** The mode of exposure utilized in studies: inhalation (Inh), pharyngeal aspiration (PA) and instillation (IT).
- **Exposure Days:** The time period in days between the first day of exposure and the last day of exposure by the animal subject to NMs.

Herein, the NIOSH/CIIT/ENPRA database has 1899 observations in rows, a total of 21 input variables and one dependent variable in columns.

Chapter 3

Variable Selection

Variable selection was performed first to identify, among the 21 input variables, the most important ones (e.g., diameter, zeta potential, etc.) that have a significant impact on an NM's toxicity.

3.1 Literature Review

In this section, a brief overview of general variable selection methods is given, followed by discussions regarding suitable methods for nanotoxicology data in particular.

Among the numerous variable selection methods, the All Subsets Models (ASM) method is one of the most simple and straight-forward ones [39]. It consists in the generation of models including all the possible combinations of the entire variable candidates, from size 1 to p, with p being the total number of variables. The drawbacks of this method are apparent. First, this method may well be infeasible for cases with a large number of variable candidates due to the extremely heavy computational burden: the number of combination subsets is $2^p - 1$. Second, this method is dependent on selection criteria such as SSE, R^2, Mallows' C^p and AIC. Different criteria may result in different selections, causing confusion and extra work.

StepWise methods (SW) are also commonly known and widely used, which employ two different schemes: forward selection and backward elimination [40, 41]. The forward selection starts with a model with no variable and proceeds by adding one variable at a time until the stop criterion is met. Backward elimination proceeds in the opposite direction: It

starts from a model including all the p variables and eliminates variables step by step. The main problem of SW lies in that it is a greedy algorithm, making locally optimal choices at each step while may well lead to suboptimal results in the end. It is also computationally expensive for large-p cases due to the iterations involved [42].

Least squares regression shrinkage is a variable selection method which is more computationally intensive [43]. It includes a number of efficient algorithms: LASSO, Ridge regression and Elastic Net. These methods are the constrained versions of ordinary least squares, each of which penalizes the l_1 norm of weights in different ways. LASSO minimizes residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant [44, 45]. The additional constraint of ridge regression is regarding the sum of the squared values of the coefficients [46, 47]. Elastic Net combines the penalty terms of LASSO and ridge regression [48, 49]. Due to such constraints, it tends to force (or shrink) some coefficients to be exactly zero and hence realizes variable selection. The shortcoming of these methods is that the degree of sparsity in the solution (shrinkage level) is dependent on the tuning parameter λ [42, 44, 50]. The higher λ is, the more coefficients are shrunk to 0.

As noted earlier, the NIOSH/CIIT/ENPRA data have substantial missing values, and the responses are subject to heterogeneous errors. The variable selection methods above are not ready to be applied on this dataset. First, these methods are not able to cope with missing values present in the input variables. Second, these regression-based methods reply on the assumption of homogeneous errors and don't work well with the drastic heterogeneity in the nanotoxicology database.

Considering the features of the NIOSH/CIIT/ENPRA database, regression tree (RT) has been identified as a feasible method of variable selection. RT successively partitions the whole data set into binary groups corresponding to subregions of the input space, and within each subregion the homogeneity assumption approximately holds. The binary search scheme

Table 3.1: List of the 10 Predictor Variables

Category	Selected Variable	Units	Variable Type
Exposure Condition	Dose Metric	$\mu g/g$	Quantitative
	Post-Exposure Duration	day	Quantitative
	Exposure Route	instillation, inhalation, aspiration	Qualitative
Animal-related	Gender	female, male	Qualitative
Material-related	Diameter	nm	Quantitative
	Length	um	Quantitative
	Zeta Potential	mV	Quantitative
	Surface Area	m^2/g	Quantitative
	Material Category	carbon, metal, metal oxide	Qualitative
	Structure Form	belt,particle,tube	Qualitative

of RT allows it to accommodate dataset with missing values while utilizing all the available data.

3.2 Results of Variable Selection

Applying RT to the NIOSH/CIIT/ENPRA database, 10 out of 21 independent variables are chosen as important and thus used to build the prediction model. These 10 predictors include six quantitative and four qualitative variables (Table 3.1).

3.3 Discussion

From table 3.1, the selected predictors consist of 3 exposure-condition variables, 1 animal-related variable, and 6 material-related variables. These factors have also been reported

in the nanotoxicology literature to have impacts on animals' bioactivity responses to NM exposure.

Dose level and post-exposure duration are long known to affect NM's toxicity. The limits of silver nanoparticles used for medicinal purposes were suggested in Tiwari et al. [51] by exploring the effect of various doses of silver nanoparticles in rats. Pauluhn et al. [52] provided strong evidence that pulmonary toxicity was dependent on the volume-based cumulative lung dose of NMs. Particle volume dose was used to predict the OELs (occupational exposure levels) of low-toxicity isometric biopersistent particles [53]. Schmid et al. [54] showed that surface area dose was the best predictor of acute pulmonary inflammation in mice and rates exposed to various types of NMs by instillation. Post-exposure duration provides information on the persistence of pulmonary inflammation after the end of exposure for long-term effects of the material [55]. Exposure route was shown to influence the pulmonary inflammatory response, with statistically significant increase in BALF neutrophils after TiO_2 instillation compared to TiO_2 inhalation which resulted in a modest increase in BALF neutrophils[56].

Different in-vivo toxicity profiles were reported in many papers. A gender-related difference in the accumulation of silver nanoparticles was noted in the kidneys of rats, with a twofold higher concentration in the female kidneys than that in the male kidneys after inhalation exposure [57, 58]. Gender-related differences were also reported for mice exposed to silver nanoparticles in blood and distribution in lungs and kidneys [59].

As selected material-related variables, diameter, length and surface area play a significant role in the toxicity behavior of NMs. Particle sizes and surface areas of NMs were reported to be influential in dictating their toxicity [60, 61, 62, 63]. Researchers also showed a significant correlation between zeta potential and pulmonary inflammation [64]. The impacts of material category and structure form have been studied as well. Porter et al. [65] investigated the effect of structure form on lung toxicity in rodents by three shapes of titanium dioxide NMs. It was found that the severity of pulmonary response for particle type

of NM is less than that of belt type. A comparison between metal oxide nanoparticles and carbon nanotubes was made in Karlsson et al. [66], and the toxicity effects of these two types of NMs were different. In the study of Studer et al. [67], two forms of copper (copper oxide and carbon-coated copper) showed distinctly different responses, with copper oxide being more toxic compared to copper.

3.4 Simulation Model

The NIOSH/CIIT/ENPRA database includes a large amount of data collected from 25 in-vivo studies for 1899 rodent animals. However, these data points are far from well designed to provide a fair coverage of the 10-dimensional space spanned by the 10 identified important predictors. Thus, the database is not adequate to quantify the dependence of toxicity responses upon the 10 predictors, much less to serve as a source for both estimation (estimating the relationship of toxicity vs. the 10 predictors) and validation (validating the prediction capability of the estimation model by using a different dataset) data.

Due to the inadequacy of the NIOSH/CIIT/ENPRA database, we developed a simulation model to demonstrate and assess the statistical framework and methods. The simulation model was derived from the NIOSH/CIIT/ENPRA database, and designed to generate data that have the major features of a biological dataset such as error heterogeneity. In comparison to real experiments, simulation-based experiments render the following advantages in illustrating statistical methods. First, simulation experiments can be designed by a statistician and carried out on a computer. As will be seen in Chapter 4, an efficient design in the 10-dimensional input space is critical to achieving a quality model of the target relationship. Second, a simulation model provides the true benchmark to evaluate the estimated model at any point in the input space.

The simulation model consists of a number of neural network (NN) models and normally-distributed random errors with heterogeneous variances. The NN models are fitted from the NIOSH/CIIT/ENPRA data representing the input-output relationships, after

missing values are filled in by the tree-based imputation method [68]. There are three types of approaches to handle missing values in the literature: deletion, disregarding and imputation. When the missing percentage is more than 5% of the total number of observations, which is the case with the NIOSH/CIIT/ENPRA data, deletion is not recommended. Disregarding means using the modeling methods which can be applied with the presence of missing values [68, 69, 70], and the resulting models from these methods may well provide the same response predication for a wide range of inputs. Imputation seeks to fully utilize the available information in the data and fill in the missing values.

It is worth noting that a setting of 4 qualitative variables corresponds to a combination category. For each category, there is a seven-dimensional response surface which is modeled by neural network and heterogeneous variance errors are introduced based on the features of existing nanotoxicology data.

Chapter 4

Design of Experiments

4.1 Literature Review

Currently, the adopted designs in nanotoxicology studies are typically generated by experimenters based on their empirical experiences [71, 72]. Since most of the studies only focus on dose-response or dose-time-response relationships [73, 74, 72, 75], the traditional designs usually involve equally-spaced levels in dose and/or time range for each NMs of interest. If such designs are applied to the 10-dimensional input space, it will lead to a tremendously large number of design points, which are not affordable with limited resources.

How to allocate limited samples to a high-dimensional input space, with the target response surfaces being complex and nonlinear? Sequential experimental design procedures involving multiple (greater than or equal 2) stages of experimentation are usually employed to enable a learning process [76, 77]. In the initial stage, the input space is well defined with each input factor having its specified ranges or categories, but no information is available regarding the target response surfaces. To achieve a fair coverage of the input space for initial exploration, a model-independent design [78] such as a fractional factorial design [79] or space-filling design [80, 81] is usually adopted here. On the initial data, features of the target response surfaces (e.g., surface nonlinearity and variance heterogeneity) are explored, and the information obtained is used to guide the follow-up stage of experimentation. In a follow-up stage, the design of additional samples are determined aiming at optimizing the quality of the resulting model fitted from all the data collected so far plus the samples to be allocated by this stage. A model-based design [82] is typically adopted for design

augmentation factoring into account the particular features of target response surfaces. A sequential design is terminated once the desired model quality has been achieved or the limited budget has been exhausted. In the context of biological experimentation, a limited budget (sample size) is imposed and a two-stage procedure is adopted.

4.2 Two-Stage Optimum Design of Experiments

The two-stage procedure developed by [78] was adopted to perform the design of experiments in the 10-dimensional input space for the modeling of the target response surfaces quantifying the dependence of the PMN toxicity response upon the 10 predictors.

In Table 4.1, the ranges for each of the 6 quantitative factors and levels involved in the initial design are given.

4.2.1 Initial Design

Two types of designs are commonly used for initial exploration: space-filling design and fractional factorial design. In this case, there are both quantitative and qualitative factors.

To implement a factorial design, certain levels need to be selected for the quantitative factors. As shown in Table 4.1, 3 discrete levels were selected for each quantitative factor: lowest, middle, and highest values of the specified range, leading to one factor (Gender) with 2 levels, and nine factors with 3 levels. A full factorial design requires the complete combinations of all the factor levels, that is, $2 \times 3^9 = 39366$ distinct design points to be sampled, which is unrealistically large. Thus, a mixed fractional factorial design was employed here: $d_1 = 2 \times 3^{9-4} = 486$ design points were generated to provide a good coverage in the 10-dimensional space.

If the initial budget only allows $d_1 < 486$ design points, then a space-filling design can be considered. In the subspace of the 4 qualitative factors, a fractional factorial design such as $2 \times 3^{3-1} = 18$ can be first generated. Then, across the 18 slices (categories formed by the 4 qualitative factors), the sliced Latin hypercube design [83] can be performed to

Table 4.1: Ranges of Predictors and Levels Involved

Predictor Variables	Variable Type	Feasible Range	Levels in Initial Design	Levels for Check Points
Dose Metric	Quantitative	[0, 2677.78]	0,1338.89, 2677.78	669.445, 2008.34
Post-Exposure Duration	Quantitative	[0, 364]	0, 182, 364	91, 273
Exposure Route	Qualitative	Inh, IT, PA	NA	NA
Gender	Qualitative	Male, Female	NA	NA
Diameter	Quantitative	[7, 300]	7, 154, 300	80.25, 226.75
Length	Quantitative	[0.13, 20]	0.13, 10, 20	5, 15
Surface Area	Quantitative	[5.1, 322]	5.1, 163, 322	84, 243
Zeta Potential	Quantitative	[-39.5, -9.35]	-39.5, -24.35, -9.35	-32, -17
Material Category	Qualitative	Carbon, Metal, Metal Oxide	NA	NA
Structure Form	Qualitative	Belt, Particle, Tube	NA	NA

generate $d_1/18$ points in each slice, leading to an initial design of d_1 distinct points in total. The resulting sliced design provides an even coverage of each slice, and all the design points have the maximum stratification in any one-dimensional projection when collapsed across the slices.

Once the locations of the initial design points have been determined, then 5 replications (which is the typical practice in biological experiments) were assigned to each point. In the fractional factorial design adopted for this work, a total of $486 \times 5 = 2430$ samples were used.

4.2.2 Second-Stage Design

Following the second-stage design method in Pei et al. [78], information was derived from the initial-stage data and used to find the augmented design by solving an IMSE (integrated mean squared error) minimization problem. The additional design points were restricted to the 18 categories formed by the 4 qualitative factors from the initial stage, and the optimization search was performed over the 18 slices.

In this work, $d_2 = 162$ additional design points ($810 = 162 \times 5$ samples) were added in the second stage. In both stages, a total of $d = d_1 + d_2 = 648$ distinct design points were included with 3240 samples. Considering the fact that the input space is 10-dimensional, this experimental budget is reasonable and low, and can be lowered if space-filling designs are used in the initial stage. Recall that with only 3 levels selected for each of the six quantitative factors, the naive full factorial design calls for $2 \times 3^9 = 39366$ design points (factor combinations) and 39366×5 samples, which serves as a conservative benchmark budget under the current design practice in nanotoxicology .

4.2.3 Estimation and Validation Data

The estimation data (ED), to which the prediction model was fitted, were generated by simulation experiments at the $d = 648$ design points from the two-stage procedure. The ED

are spread across the 18 categories out of the total $2 \times 3^3 = 54$ combination settings of the 4 qualitative factors.

To evaluate the fitted model, a validation data (VD) set was generated separately at 2304 check points. These check points are allocated in the remaining $36 = 54 - 18$ categories of the qualitative factor settings, and constitute a full combination of the quantitative factor levels in each category. The levels selected for each quantitative factor in check points are given in the last column of Table 4.1.

Chapter 5

Estimation and Inference of the Prediction Model

In this chapter, the prediction model, which enables the quantitative assessments of a new NM's bioactivity behavior without performing additional biological experiments, is estimated by a statistical modeling method referred to as SKQ (stochastic kriging with qualitative factors) [84].

SKQ is a statistical modeling method allowing its predictors to be high-dimensional quantitative and qualitative factors [84]. SKQ is flexible and general. In particular, it employs an adaptive mechanism (Gaussian correlation) to capture the inherent similarity (more or less or none) across categories, which could be different study types, different material types/shapes, etc. By synergistically modeling the data across different categories, SKQ pools information together and results in fitted models of improved quality.

Compared to other powerful statistical models such as support vector machine [85, 86, 87] and artificial neural network [88, 89, 90], the main advantage of SKQ lies in its ability to capture heterogeneous errors, which is a known feature of biological data, and enable valid statistical inference: SKQ not only provides a fitted bioactivity profile, but also quantifies the uncertainty of the estimated profile.

The comparison of SKQ and its closest parametric counterpart, mixed effects modeling (MEM) [91, 92, 93], was performed in Wang et al. [85] and briefly summarized here. As a parametric regression method, MEM is subject to restrictive assumptions such as a prior-assumed functional form (e.g., logistic function) for the target relationships. Moreover, MEM builds its information-pooling ability upon the assumed model commonality: A

common functional form has to be used for bioactivity profiles across all the categories, and such a commonality assumption is also required for the error variance structure. These assumptions can be easily violated in biological data. Through an empirical case, Wang et al. shows that SKQ has superior performance over MEM even when all the MEM assumptions are satisfied.

In the literature of nanotoxicology, the majority of existing work for QNAR development are restricted to one type of NMs. For instance, Mu et al. [94, 95, 96] developed prediction models based on the data for metal oxide nanoparticles alone, and their models cannot be used for prediction of other types of NMs (e.g., metal or carbon NMs). In addition, the predictors in a lot of papers [97, 98, 35] involve quantitative factors only, while qualitative factors such as gender are disregarded. The 10 predictors identified from the NIOSH/CIIT/ENPRA database, which includes 25 individual studies, provide a broader coverage of material properties across different types of NMs and involve both quantitative and qualitative factors.

5.1 Model Estimation and Inference

Following the design in Chapter 4, the locations of a total of $d = 648$ distinct design points denoted as $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ were determined in the 10-dimensional input space. The estimation data (ED) were generated via simulation experiments (Section 3.4) at these design points with 5 replications at each point. Denote the estimation data as $\{\mathbf{w}_i, y_j(\mathbf{w}_i)\}$ for $i = 1, 2, \ldots, d$ and $j = 1, 2, \ldots, 5$ and the sample variances obtained from the 5 replications at design points as $\widehat{\mathrm{Var}}[\epsilon(\mathbf{w}_i))]$.

On the ED, the SKQ fitting procedure [84] was applied, and the resulting prediction model allows for the estimation and inference of the expected toxicity response at an arbitrary point/setting $\mathbf{w}$ of the input space. That is, the estimated expected response $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variance $\widehat{\mathrm{Var}}[\epsilon(\mathbf{w}_i))]$ are both available analytically.

5.2 Model Validation

As explained in Section 4.2.3, a separate validation data (VD) set was generated to evaluate the goodness of the prediction model fitted from the ED.

There are 4 qualitative predictors leading to a total of 54 combination categories, which are listed in Table 5.1. The 648 design points in the ED spread across 18 categories, as shown in the first column of the table. The 2304 check points included in the VD are allocated in the remaining 36 categories. The quantitative levels of the check points also differ from those of the design points, as pointed out previously in Section 4.2.3.

At each check point, the deviation between the estimated expected response and its true value (available from the simulation model) is calculated. The deviations at check points included in each one of the 36 validation categories are plotted in one histogram, leading to 36 histograms in Figure 5.1-5.2. From these deviation histograms, it can be seen that the SKQ prediction model is able to provide accurate estimates for the toxicity responses throughout the input space. Note that the histograms in Figure 5.1-5.2 are obtained from one macro-replication: Generate an ED following the design procedure, fit SKQ to the ED and evaluate the fitted SKQ at all the check points. One hundred macro-replications have been performed in our work, and each one leads to similar histograms in terms of the ranges of deviations.

This case is built on the simulation model (Section 3.4) whose output data mimic those from real biological experiments. As explained earlier, simulation is typically employed for the illustration and evaluation of statistical methods [99], because the design and implementation of simulation experiments can be easily done and the true benchmarks are always available to evaluate the fitted model.

This case demonstrates the critical role of experimental design and SKQ's prediction capability in the development of QNAR.

Table 5.1: Combination Categories of Estimation and Validation Data.

Data Type	Combination Category	The Setting of 4 Qualitative Variables			
		Gender	Exposure Route	Material Category	Structure Form
Estimation	Category1	female	inhalation	carbon	belt
Validation	Category2	female	instillation	carbon	belt
Validation	Category3	female	aspiration	carbon	belt
Estimation	Category4	male	inhalation	carbon	belt
Validation	Category5	male	instillation	carbon	belt
Validation	Category6	male	aspiration	carbon	belt
Validation	Category7	female	inhalation	metal	belt
Validation	Category8	female	instillation	metal	belt
Estimation	Category9	female	aspiration	metal	belt
Validation	Category10	male	inhalation	metal	belt
Validation	Category11	male	instillation	metal	belt
Estimation	Category12	male	aspiration	metal	belt
Validation	Category13	female	inhalation	metal oxide	belt
Estimation	Category14	female	instillation	metal oxide	belt
Validation	Category15	female	aspiration	metal oxide	belt
Validation	Category16	male	inhalation	metal oxide	belt
Estimation	Category17	male	instillation	metal oxide	belt
Validation	Category18	male	aspiration	metal oxide	belt
Validation	Category19	female	inhalation	carbon	particle
Estimation	Category20	female	instillation	carbon	particle
Validation	Category21	female	aspiration	carbon	particle
Validation	Category22	male	inhalation	carbon	particle
Estimation	Category23	male	instillation	carbon	particle
Validation	Category24	male	aspiration	carbon	particle
Estimation	Category25	female	inhalation	metal	particle
Validation	Category26	female	instillation	metal	particle
Validation	Category27	female	aspiration	metal	particle
Estimation	Category28	male	inhalation	metal	particle
Validation	Category29	male	instillation	metal	particle
Validation	Category30	male	aspiration	metal	particle
Validation	Category31	female	inhalation	metal oxide	particle
Validation	Category32	female	instillation	metal oxide	particle
Estimation	Category33	female	aspiration	metal oxide	particle
Validation	Category34	male	inhalation	metal oxide	particle
Validation	Category35	male	instillation	metal oxide	particle
Estimation	Category36	male	aspiration	metal oxide	particle
Validation	Category37	female	inhalation	carbon	tube
Validation	Category38	female	instillation	carbon	tube
Estimation	Category39	female	aspiration	carbon	tube
Validation	Category40	male	inhalation	carbon	tube
Validation	Category41	male	instillation	carbon	tube
Estimation	Category42	male	aspiration	carbon	tube
Estimation	Category43	female	inhalation	metal	tube
Validation	Category44	female	instillation	metal	tube
Validation	Category45	female	aspiration	metal	tube
Validation	Category46	male	inhalation	metal	tube
Estimation	Category47	male	instillation	metal	tube
Validation	Category48	male	aspiration	metal	tube
Estimation	Category49	female	inhalation	metal oxide	tube
Validation	Category50	female	instillation	metal oxide	tube
Validation	Category51	female	aspiration	metal oxide	tube
Estimation	Category52	male	inhalation	metal oxide	tube
Validation	Category53	male	instillation	metal oxide	tube
Validation	Category54	male	aspiration	metal oxide	tube

Experimental design plays a critical role in the quantification of high-dimensional relationships such as the one investigated here. An efficient design allows for the development of a high-quality prediction model with a reasonable amount of experimental effort. In this case, 648 distinct design points are employed in the 10-dimensional input space to provide a fair coverage of the space as well as to precisely capture the complex and nonlinear response surfaces. (Please refer to Section 4.2 for the experiments/samples required by a naive full-combination design.)

To a well-designed data set, SKQ is able to fit a high-quality prediction model. The 648 design points in the ED involve material-factor settings corresponding to the NMs on which biological experiments have been performed. To the ED collected on these old NMs, the prediction model is fitted. The 2304 check points in the VD involve material-factor settings corresponding to new NMs that have not been investigated. The prediction model is able to render accurate estimates for these new NMs' expected toxicity behaviors.

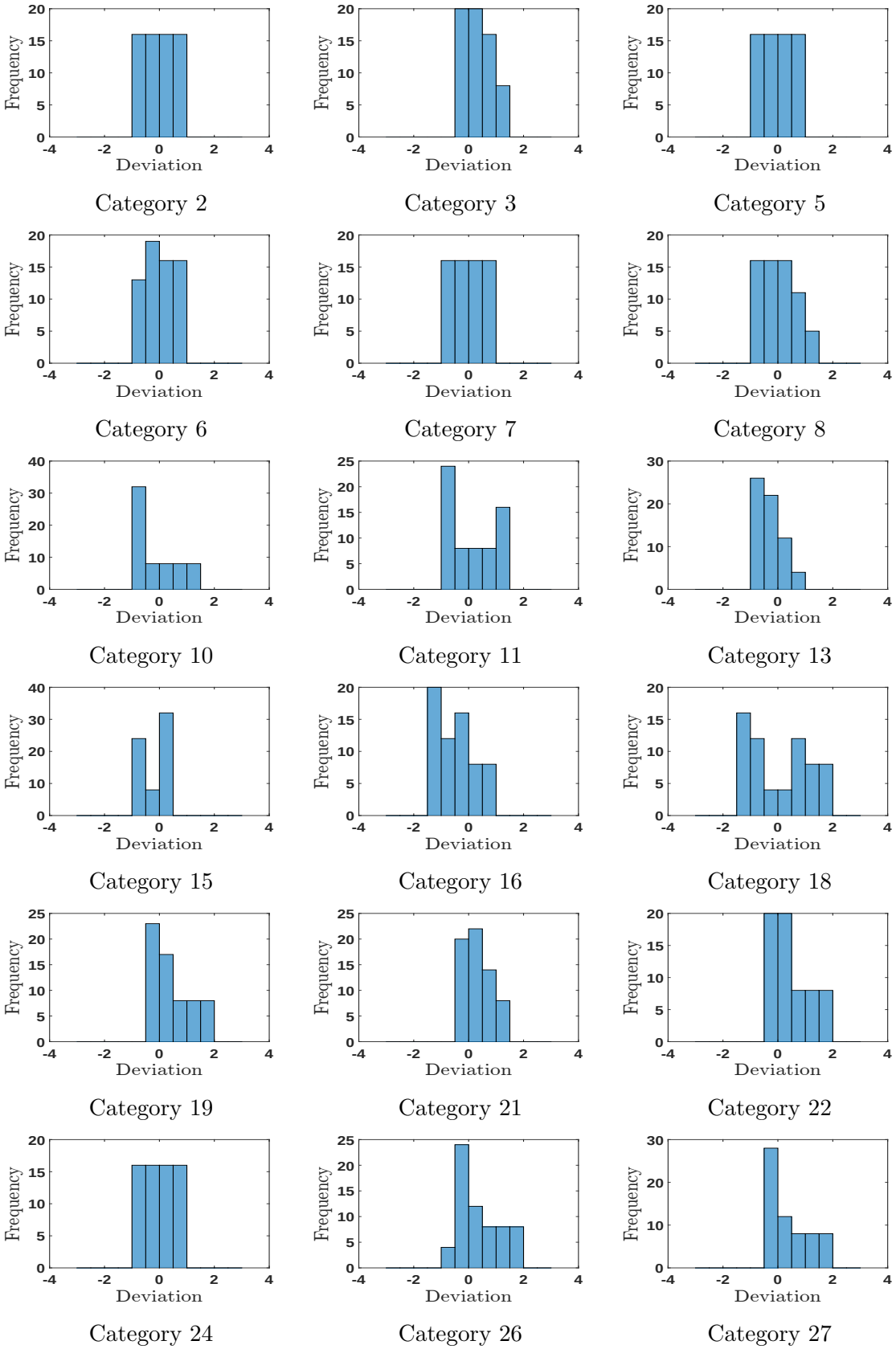

Figure 5.1: Histograms of deviations for the first group of 18 categories in VD.

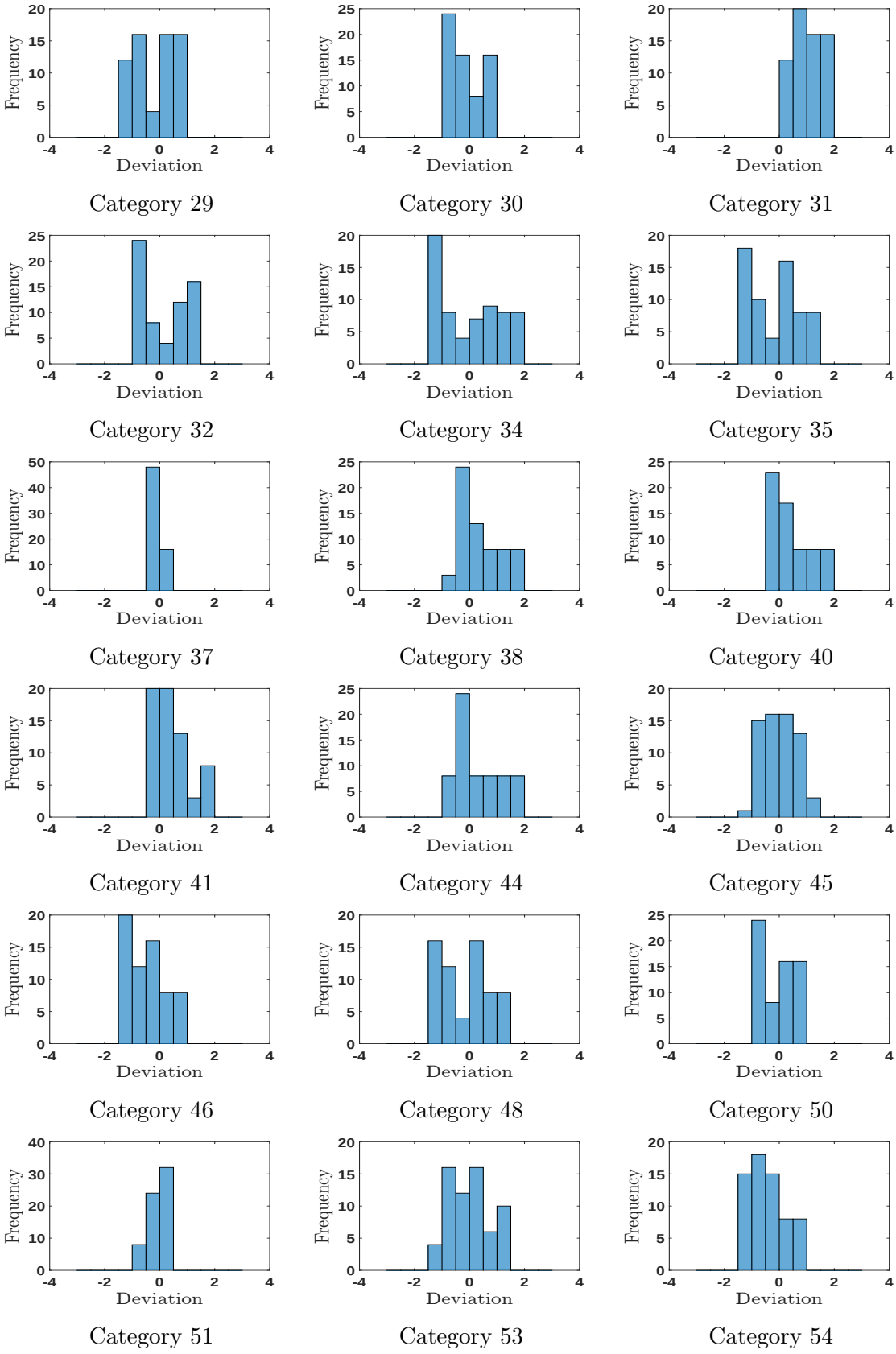

Figure 5.2: Histograms of deviations for the second group of 18 categories in VD.

Chapter 6

Shape Clustering

Clustering of NMs based on their potency is an important step in hazard assessment [16]. With the SKQ prediction model developed in the previous chapter, a new NM's 5-dimensional toxicity profile (the toxicity response vs. 4 non-material factors) can be estimated without performing additional experiments on this material. In this chapter, clustering methods will be developed/adapted to group NMs based on their estimated toxicity profiles along with the estimation uncertainty.

6.1 Literature Review

In the literature of nanotoxicology, the majority of clustering methods directly used material properties [26, 100, 27] to group NMs. In these work, clustering was performed in the space formed by one, two, or several material properties, with no effort to quantify how these properties affect NMs' bioactivity. Some researchers [38, 101, 102] sought to derive potency information from exposure-response profiles for NM clustering. For instance, Drew et al. [38] used BMD (benchmark dose) estimates as a measure of potency to group materials. However, BMD is a single point on the 5-dimensional toxicity profile for a specified NM, and can at best provide a snapshot of the material's toxicity behavior. Thus, in this chapter, shape clustering methods are adapted to cluster NMs by utilizing their complete toxicity profiles while taking into account the profile estimation uncertainty.

6.2 Statement of Shape Clustering Problem

For convenience of discussions, notations for shape clustering are provided as follows.

- n: the number of objects (or shapes) to be clustered.
- $\mathbf{O}_j$: the j^{th} object, with $j = 1, 2, \ldots, n$.
- k: the number of clusters with $k < n$.
- $\{C_i; i = 1, 2, \ldots, k\}$: the collection of k clusters formed from n objects.
- n_i: the size of (number of objects included in) the i^{th} cluster C_i with $i = 1, 2, \ldots, k$.
- $\boldsymbol{\mu}_i$: the centroid of the i^{th} cluster C_i.
- $d(\mathbf{O}, \mathbf{O}')$: the distance between the two objects $\mathbf{O}$ and $\mathbf{O}'$.

Given the n objects $\{\mathbf{O}_j; j = 1, 2, \ldots, n\}$ and the desired number of clusters k, shape clustering groups the objects into k disjoint clusters $\mathbf{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$ which seeks to maximize the homogeneity within a cluster and/or the separation across clusters.

The optimization objective of shape clustering can be formulated in different ways [103], and herein, three most commonly-used formulations are breifly reviewed.

- Objective I: Minimize within-cluster sum of squared distance [104, 105].

$$\min \sum_{i=1}^{k} \sum_{\mathbf{O} \in C_i} d(\mathbf{O}, \boldsymbol{\mu}_i)^2 \tag{6.1}$$

  where $\boldsymbol{\mu}_i$ is a centroid obtained from

$$\boldsymbol{\mu}_i = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{\mathbf{O} \in C_i} d(\mathbf{w}, \mathbf{O})^2 \tag{6.2}$$

  with $\mathbf{w}$ being a vector of the same dimension of an object.

- Objective II: Minimize within-cluster variance [106].

$$\min \sum_{i=1}^{k} \frac{2}{n_i} (\sum_{\mathbf{O}_a \in C_i} \sum_{b<a, \mathbf{O}_b \in C_i} d(\mathbf{O}_a, \mathbf{O}_b)^2) \tag{6.3}$$

- Objective III: Minimize within-cluster distances and maximize distances between clusters [107].

$$\min \sum_{\mathbf{O}_a \in C_i} \sum_{b<a, \mathbf{O}_b \in C_i} d(\mathbf{O}_a, \mathbf{O}_b)^2 \tag{6.4}$$

and

$$Max_{\mathbf{O}_a \in C_i} \underset{\mathbf{O}_b \in C_j, C_i \neq C_j}{Min} d(\mathbf{O}_a, \mathbf{O}_b)^2 \tag{6.5}$$

In a shape clustering objective, the distance $d(\mathbf{O}, \mathbf{O}')$ can be defined by different metrics as follows.

- Euclidean distance. Let $\mathbf{X} = (x_1, x_2, \ldots, x_p)$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_p)$ be two points in the p-dimensional Euclidean space. The Euclidean distance is defined as:

$$d_E(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}. \tag{6.6}$$

- Hausdorff distance[108]. Let $\mathbf{X}$ and $\mathbf{Y}$ be two non-empty subsets. We define their Hausdorff distance $d_H(\mathbf{X}, \mathbf{Y})$ by

$$d_H(\mathbf{X}, \mathbf{Y}) = \max\{\sup_{x \in \mathbf{X}} \inf_{y \in \mathbf{Y}} d(x, y), \sup_{y \in \mathbf{Y}} \inf_{x \in \mathbf{X}} d(y, x)\} \tag{6.7}$$

where *sup* represents the supremum and inf the infimum. The Hausdorff distance is very sensitive to noise: an outlier can substantially affect the distance value.

- Frechet Distance [108]. Let X and Y be two parameterized curves $X(\alpha(t))$ and $Y(\beta(t))$. Then, the Frechet distance is defined as the following:

$$d_F(X,Y) = \inf_{\alpha,\beta} \max_{t\in[0,1]} \{d(X(\alpha(t)), Y(\beta(t)))\} \tag{6.8}$$

How to solve the optimization problem of clustering? The algorithms can be roughly divided into two categories: optimal and heuristic methods. Examples of the optimal methods include Wu et al. [109], which developed a graph theoretic approach to solve the image segmentation problem, and Gath et al. [110], which used maximum-likelihood estimation to obtain the optimal solutions. Clustering heuristics mainly include K-means [111, 112], hierarchical clustering [113] and density-based clustering methods [114].

6.3 Shape Clustering of Nanomaterials

In light of the fact that NMs' toxicity profiles (objects) are estimated from the SKQ prediction model and subject to uncertainty, we chose to adapt the following two algorithms for NM clustering: K-shape clustering algorithm with uncertainty (KSCAU) [105], and density-based clustering algorithm with uncertainty (DBCAU) [114]. These algorithms were adapted to accommodate the estimation uncertainty of the prediction model.

6.3.1 K-Shape Clustering Algorithm with Uncertainty

KSCAU [105] is a heuristic clustering method, which solves the optimization problem (6.1) through a series of refinement iterations. The KSCAU algorithm adapted for NM clustering is given in Algorithm 1.

The inputs of KSCAU are listed and explained as follows.

- $\{\mathbf{O}_j; j = 1, 2, \ldots, n\}$: n objects to be clustered. In our case, each object is a response vector representing the toxicity profile of an NM, which is estimated from the SKQ

Algorithm 1: KSCAU

Input: (a) The n toxicity profiles estimated by the SKQ prediction model $\{\mathbf{O}_1,\mathbf{O}_2,...,\mathbf{O}_n\}$; (b) The preliminary number of clusters k; (c) The fitted SKQ prediction model $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$; (d) The locations of design points $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ and the sample variances at the design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

Output: $\{l_j; j = 1, 2, \ldots, n\}$: the cluster label of each object $\mathbf{O}_j$, with $l_j \in \{1, 2, \ldots, k\}$

.

Initial Step:

Randomly assign n profiles to k clusters and compute the initial k centroids by Equation (6.2).

Iterative Refinement Procedure:

while *!Stop* **do**

  for $j = 1$ *to* n **do**

    Compute the expected distance from each object to the k centroids by applying Algorithm 2 with Inputs (c) and (d) and the current cluster labels of the objects.

    Update the cluster label of each object as follows:

$$l_j = \underset{i}{argmin}(Expected\ Distance\ (\mathbf{O}_j, \boldsymbol{\mu}_i)). \tag{6.9}$$

  end

  for $i = 1$ *to* k **do**

    Check if C_i is empty, and delete empty clusters to update k.

  end

  for $i = 1$ *to* k **do**

    Based on the current cluster labels of each object, compute the k centroids by Equation (6.2).

  end

  if $\mathbf{L}^{(Iteration)} = \mathbf{L}^{(Iteration-1)}$ **then**

    Stop = 1

  end

  $Iteration = Iteration + 1$

end

prediction model:

$$\mathbf{O}_j = (\widehat{\mathrm{Y}}_{j1}, \widehat{\mathrm{Y}}_{j2}, \ldots, \widehat{\mathrm{Y}}_{jg}).$$

A total of g points in the 5-dimensional profile (evenly-spaced across the 4 non-material factors) for a NM selected to form its numerical object.

- k: the preliminary number of clusters. k can be set in a somewhat arbitrary manner as long as it is less than n. The algorithm iterates to refine the value of k. k is recommended to be set as a large number.
- The fitted SKQ prediction model, denoted as $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances of the expected responses at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$.
- The locations of design points in the estimation data $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ to which the SKQ model was fit and the sample variances obtained from the 5 replications at design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

These inputs will be passed onto Algorithm 2 to compute the expected distance from an object to a centroid by bootstrapping resampling methods.

In the initial step, the n objects are randomly assigned to k clusters. Then, the centroid of each cluster is computed by Equation (6.2). Once the initial centroids are obtained, an iterative refinement procedure is performed with two steps involved in each iteration.

Step 1: For each object, apply the bootstrapping algorithm 2 to estimate its expected distance to each of the k centroids, and update its cluster label by assigning it to the closest centroid cluster. In contrast to the traditional distance metrics which are deterministic, the expected distance is a probabilistic measure of the object-to-centroid distances subject to estimation uncertainty. Herein, the Euclidean distance metric (6.6) is adopted. How to obtain the expected Euclidean distance between two estimated vectors (one object and one centroid)? SKQ enables the analytical approximations for such expectation estimates by providing analytical expressions for the mean, variance and covariance estimates of the estimated responses $\widehat{\mathrm{Y}}(\cdot)$ (vector components). However,

since the distance metric (6.6) is a nonlinear function of the two vectors, analytical approximations may well fall short of accuracy compared to numerical methods, which is adopted for the estimation of expected distances in this work. Algorithm 2 provides the bootstrap resampling procedure for this purpose: For each resampled estimation data set, a SKQ model is fitted, and provides a collection of estimated objects and centroids leading to the estimates of object-to-centroid distances; Across the resampled data sets, the distance expectation can be estimated based on the distances obtained from each resampled data.

Step 2: After the cluster label of each object is updated, the clusters with no affiliation will be deleted. Accordingly, the value of k is updated and new centroids are formed based on the current cluster memberships.

These two steps are repeated until there is no change in cluster labels from the previous to current iteration.

KSCAU is a shape clustering algorithm which accommodates uncertain objects. The resulting number of clusters is obtained from the iterative refinement procedure, and does not need to be specified in advance as traditional K-means algorithm. However, if the initial value of k is set to be a very small number, it can not increase with iterations. In addition, different initial partition of objects in the initial step may result in different clustering results.

6.3.2 Density-based Clustering Algorithm with Uncertainty

Density-based clustering algorithm with uncertainty (DBCAU) (Algorithm 3) is an adapted density-based clustering algorithm. Clusters are considered as regions in which the objects are dense. The key idea of DBCAU lies in identifying core objects and their reachable neighbors by means of calculating the respective probabilities.

Some definitions of regarding DBCAU are provided below.

Definition 1 Radius r and threshold m:

r is a distance threshold and m the minimum number of objects contained in a cluster.

Algorithm 2: Estimating the expected distance via bootstrapping

Input: (a) The cluster labels of the n objects; (b) The resample size B; (c) The fitted SKQ prediction model $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$; (d) The locations of design points $\{\mathbf{w}_i; i = 1, 2, \ldots, d\}$ and the sample variances at the design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

Output: The expected distances between objects and centroids.

for $i = 1$ *to* B **do**

(i) Resampling: At each design point $\mathbf{w}_i$ $(i = 1, 2, \ldots, d)$, generate t random errors $e_j^b(\mathbf{w}_i)$ from the normal distribution $N(0, \widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\mathbf{w}_i)] + \widehat{\mathrm{Var}}[\epsilon(\mathbf{w}_i)])$, and the resampled observations are represented as:
$y_j^b(\mathbf{w}_i) = \widehat{\mathrm{Y}}(\mathbf{w}_i) + e_j^b(\mathbf{w}_i); j = 1, 2, ..., t.$

(ii) To the resampling data $\{\mathbf{w}_i, y_j^b(\mathbf{w}_i)\}$, generated at the design points , fit the SKQ model and denote the resulting model as $\widehat{\mathrm{Y}}^b(\cdot)$.

(iii) Based on the fitted SKQ model $\widehat{\mathrm{Y}}^b(\cdot)$, generate the toxicity profiles of the n NMs denoted as $\{\mathbf{O}_j^b; j = 1, 2, \ldots, n\}$;

(iV) Based on the cluster labels of the n objects, compute the centroid $\boldsymbol{\mu}_i^b$ $(i = 1, 2, \ldots, k)$ for each cluster of estimated profiles;

(V) For $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$, calculate the Euclidean distance between $\mathbf{O}_j^b$ and $\boldsymbol{\mu}_i^b$ denoted as $d_E^b(\mathbf{O}_j^b, \boldsymbol{\mu}_i^b)$.

end

The expected distance between $\mathbf{O}_j$ and $\boldsymbol{\mu}_i$ is obtained by the following equation:

$$d_E(\mathbf{O}_j, \boldsymbol{\mu}_i) = \frac{\sum_{b=1}^{B} d_E^b(\mathbf{O}_j^b, \boldsymbol{\mu}_i^b)}{B}; \quad i = 1, 2, \ldots, k; j = 1, 2, \ldots, n \tag{6.10}$$

Definition 2 Core object:

At a given r and m, $\mathbf{O}$ is called a core object if the number of its neighbors is not less than m. Object $\mathbf{O}'$ is called the neighbor of $\mathbf{O}$ if the distance between $\mathbf{O}'$ and $\mathbf{O}$ is not greater than r.

Definition 3 Core object probability:

Since DBCAU considers the uncertainty of objects, the core object definition needs to be enhanced by involving the likelihood that $\mathbf{O}$ is a core object. Let $\mathbf{D}$ be the set of all objects. Then, the core object probability of an object $\mathbf{O}$ is defined as:

$$P^{core}(\mathbf{O}) = \sum_{A \subseteq \mathbf{D}, |A| \geq m} \prod_{\mathbf{O}' \in A} P_d(\mathbf{O}', \mathbf{O})(r) \prod_{\mathbf{O}'' \in \mathbf{D} \setminus A} (1 - P_d(\mathbf{O}'', \mathbf{O})(r)) \tag{6.11}$$

where

$$P_d(\mathbf{O}', \mathbf{O})(r) = P(d(\mathbf{O}', \mathbf{O}) \leq r) \tag{6.12}$$

Definition 4 Directly density-reachable

An object $\mathbf{O}'$ is directly density-reachable from an object $\mathbf{O}$ w.r.t r and m if two requirements are met: (i) $\mathbf{O}$ is a core object; (ii) $\mathbf{O}'$ is the neighborhood of $\mathbf{O}$.

Definition 5 Reachability Probability

Similarly, definition 4 needs to be adapted to the probability of object $\mathbf{O}'$ to be directly density-reachable to $\mathbf{O}$. The reachability probability of $\mathbf{O}'$ is defined as follows:

$$P^{reach}(\mathbf{O}', \mathbf{O}) = P^{core}(\mathbf{O}) \cdot P_d(\mathbf{O}', \mathbf{O})(r) \tag{6.13}$$

The inputs of DBCAU are as follows.

- r: The radius which is a distance threshold;
- m: A threshold representing the minimum number of objects contained in one cluster.

Algorithm 3: DBCAU

Input: (a) The radius r; (b)The threshold m; (c) The fitted SKQ prediction model $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$; (d) The locations of design points $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ and the sample variances at the design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

Output: $\{l_j; j = 1, 2, \ldots, n\}$, the cluster labels of each object $\mathbf{O}_j$ and $l_j \in \{1, 2, \ldots, k\}$

for $j = 1$ *to* n **do**
  Compute the core probability of each object by Equation (6.11) and Algorithm 4.
  if *The probabality is greater than 0.5* **then**
    $I(j) = 1$
  else
    $I(j) = 0$
  end
end
for $j = 1$ *to* n **do**
  Compute reachability probabilities of $\mathbf{O}_j$ to all core objects. Get the cluster label based on the maximum reacheability probability

$$l_j = \underset{i}{argmax}\{P^{reach}(\mathbf{O}_j, \mathbf{O}_i) \cdot I(i)\} \tag{6.14}$$

end
for $i = 1$ *to sum(I)* **do**
  Merge clusters to refine l_j
end

- The fitted SKQ prediction model, denoted as $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances of the expected responses at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$.
- The locations of design points in the estimation data $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ to which the SKQ model was fit and the sample variances obtained from the 5 replications at design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

The DBCAU is based on the fact that a cluster is equivalent to the set of objects which are reachable from an arbitrary core object. The retrieval of these reachable objects is performed by a scheme including three steps.

Step 1: Check the r-neighborhood of every object in the set and compute the probabilities of each object to be a core by Equation (6.11). If the core object probability is greater than 0.5, then this object is identified as a core object. In our case, Euclidean distance is used as the distance metric. How to obtain the value of $P_d(\mathbf{O}, \mathbf{O}')(r)$ involved in Equation (6.11)? Due to the same reasons mentioned in Section 6.3.1, a numeric method was adopted to estimate the value of $P_d(\mathbf{O}, \mathbf{O}')(r)$ (Algorithm 4). Algorithm 4 used the bootstrap resampling procedure to generate a certain number of resampled estimation data sets, to which the SKQ models were fitted. Thus, a collection of estimated profiles of the desired two NMs $\mathbf{O}$ and $\mathbf{O}'$ was produced, leading to the estimates of distances between $\mathbf{O}$ and $\mathbf{O}'$. Then the value of $P_d(\mathbf{O}, \mathbf{O}')(r)$ is estimated by the ratio of the number of distances not greater than r to the total number of distances.

Step 2: Go through every object again. For a certain object, compute the reachability probabilities of this object to all identified core objects; Assign it to the cluster of that core object with the maximum reachability probability.

Step 3: Go through every core object and check their neighbors. For one core object, if there are any (or only one) core objects contained in its neighborhood, then merge the affiliation of those (or one) core objects into this neighborhood as one cluster.

DBCAU inherits the advantage of density-based clustering algorithm in that the resulting number of clusters does not need to be predetermined, and that it is robust to outliers. However, the clustering results are sensitive to the setting of radius r and threshold m. A big r and small m tend to produce less clusters, and a small r and large m may result in more clusters.

Algorithm 4: $P_d(\mathbf{O}, \mathbf{O}')(r)$ Calculation

Input: (a) The resample size B; (b) The radius r; (c) The fitted SKQ prediction model $\widehat{\mathrm{Y}}(\cdot)$ and the estimated variances at the design points $\widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\cdot)]$; (d) The locations of design points $\{\mathbf{w}_i; i = 1, 2, \ldots d\}$ and the sample variances at the design points $\widehat{\mathrm{Var}}[\epsilon(\cdot)]$.

Output: $P_d(\mathbf{O}, \mathbf{O}')(r)$

for $i = 1$ *to* B **do**

(i) Resampling: At each design point $\mathbf{w}_i$ $(i = 1, 2, \ldots, d)$, generate t random errors $e_j^b(\mathbf{w}_i)$ from the normal distribution $N(0, \widehat{\mathrm{Var}}[\widehat{\mathrm{Y}}(\mathbf{w}_i)] + \widehat{\mathrm{Var}}[\epsilon(\mathbf{w}_i)])$, and the resampled observations are represented as: $y_j^b(\mathbf{w}_i) = \widehat{\mathrm{Y}}(\cdot) + e_j^b(\mathbf{w}_i); j = 1, 2, ..., t.$

(ii) To the resampling data set $\{\mathbf{w}_i, y_j^b(\mathbf{w}_i)\}$, generated at the design points , fit the SKQ model and denote the resulting model as $\widehat{\mathrm{Y}}^b(\cdot)$.

(iii) Based on the SKQ model $\widehat{\mathrm{Y}}^b(\cdot)$, generate the estimated profiles of the desired two NMs and denote the obtained profiles as $\mathbf{O}^b$ and $\mathbf{O}'^b$.

(iV) Calculate the Euclidean distance of these two objects denoted as $d_E^b(\mathbf{O}^b, \mathbf{O}'^b)$.

end

Idx=0;

for $i = 1$ *to* B **do**

if $d_E^b(\mathbf{O}^b, \mathbf{O}'^b)$ *is not greater than* r **then**

Idx=Idx+1

end

end

Compute $P_d(\mathbf{O}, \mathbf{O}')(r)$ by the following equation:

$$P_d(\mathbf{O}, \mathbf{O}')(r) = \frac{\mathrm{Idx}}{B} \tag{6.15}$$

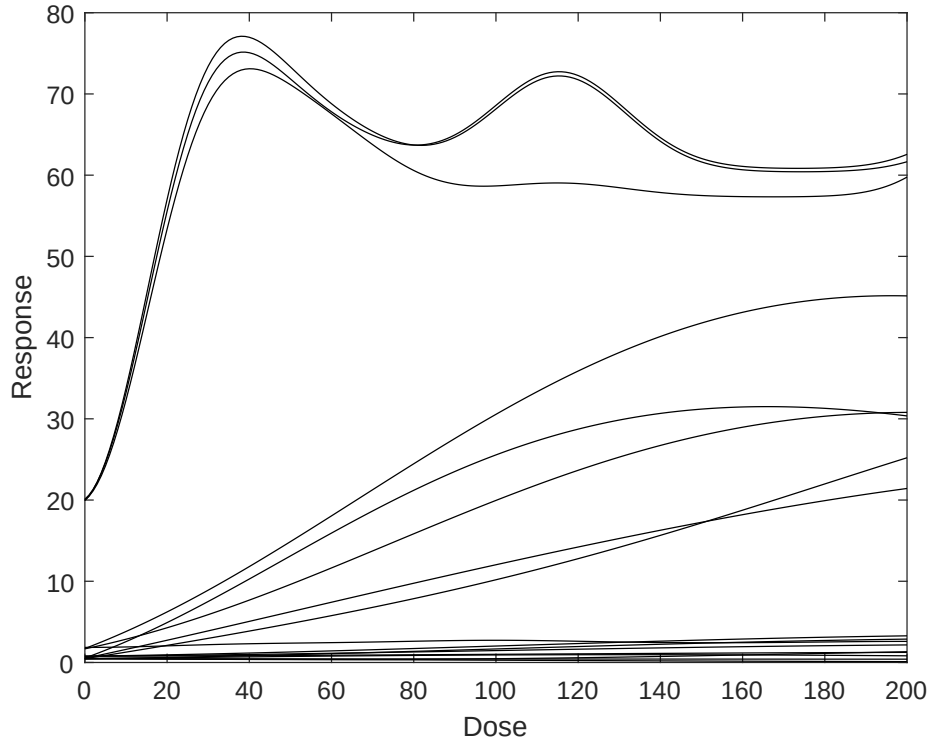

Figure 6.1: The plot of 17 fitted dose-response curves

6.3.3 Empirical Studies

Empirical case studies were designed and performed to illustrate the two clustering algorithms.

Case 1: A simple case is developed whose clustering results can be displayed in 2-D graphs, with objects being 17 dose-response curves.

Case 2: A large case is developed with objects being 81 five-dimensional profiles.

Case 1: Clustering 17 dose-response curves

Based on the estimated SKQ prediction model, 17 dose-response curves are extracted to be the target objects. Figure 6.1 shows the plot of 17 fitted dose-response curves.

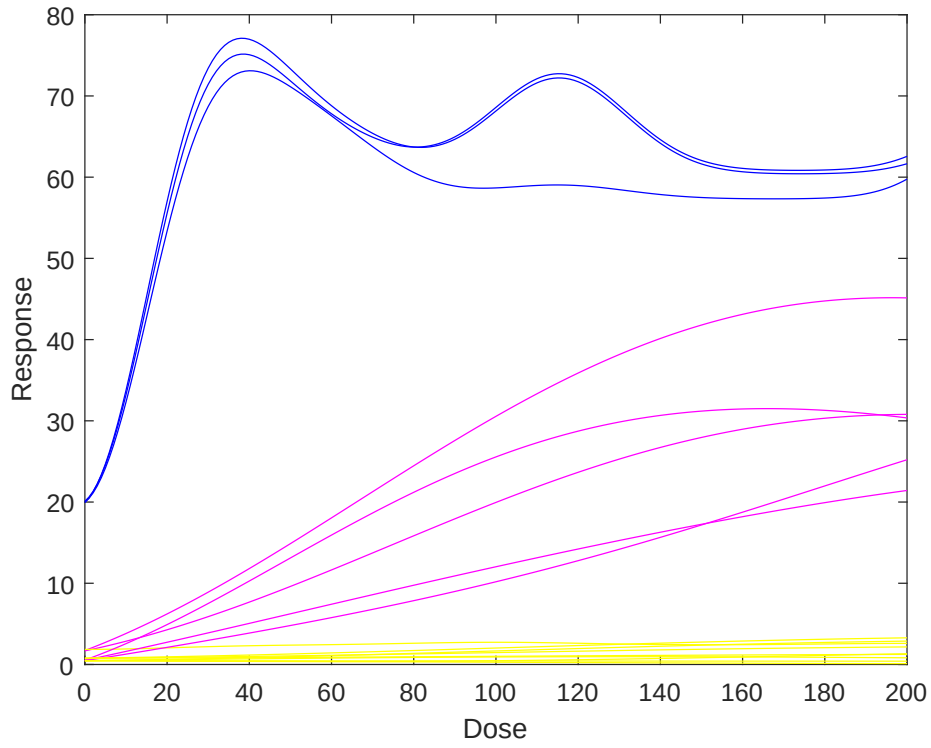

Figure 6.2: KSCAU clustering results of Case 1

Clustering Results Figure 6.2 and 6.3 displayed the clustering results by KSCAU and DBCAU separately. In Figure 6.2, three clusters colored as blue, pink and yellow are generated among the 17 profiles. In Figure 6.3, 17 dose-response profiles are divided into 4 clusters colored as light blue, red, yellow and pink.

Although the clustering results from two algorithms are a little bit different, they are both reasonable as can be seen from the figures. DBCAU results are sensitive to the two pre-specified parameters: r and m. With suitable values of r and m, the clustering result of DBCAU may be the same as that of KSCAU. For KSCAU, the resulting number of clusters k is generated by the iterative procedure. But the initial value of k is better set as a relatively large number smaller than n, since the value of k cannot be increased with iterations.

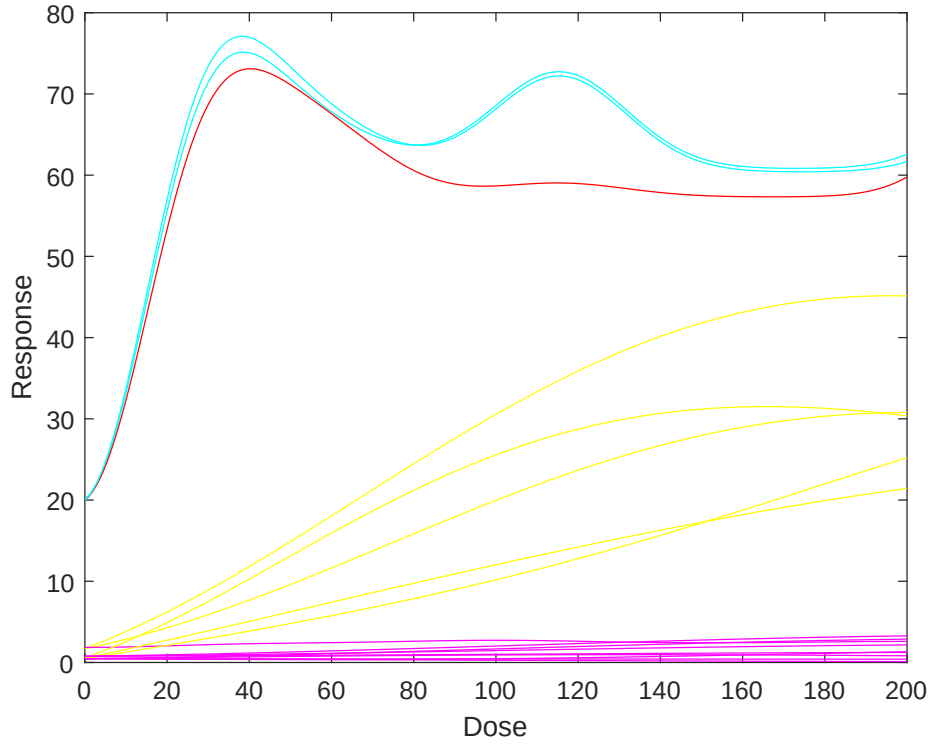

Figure 6.3: DBCAU clustering results of Case 1 with $r = 1.5$ and $m = 2$

Case 2: Clustering 81 five-dimensional profiles

Based on the estimated SKQ prediction model, five-dimensional toxicity profiles of 81 NMs are derived to be clustered.

Generating the 81 five-dimensional Profiles There are six material-related variables in the prediction model obtained in Chapter 5: diameter, length, surface area, zeta potential, material category and structure form. Since diameter, length and surface area determine the size of a NM together, the values of these three variables are bundling. That is, taking the maximum, middle and minimum values of theses three variables simultaneously to represent the large, median and small size of NMs. For zeta potential, three discrete levels are taken as well. Since there are three levels for the qualitative variables "material category" and "structure form", 81 different kinds of NMs characterized by the six material-related variables

are defined. On the basis of the prediction model , 81 five-dimensional toxicity profiles of NMs were estimated.

Clustering Results Table 6.1 displays the resulting clusters by two methods and the values of material-related factors of the corresponding kinds of NMs. Similar to the situation occurring in the first example, the clustering results of two algorithms differ with each other a little bit. DBCAU produced 8 clusters while KSCAU produced 7 clusters. As I mentioned before, the difference is due to the determination of values of the two parameters involved in DBCAU.

The clustering results are consistent with the general knowledge in the nanotoxicology literature. NMs of belt or tube shape are generally more toxic than those shaped in particles. NMs of small size are generally more toxic than those of large size.

Table 6.1: Clustering Results of the 81 NMs

NM	KSCAU cluster label	DBCAU cluster label	Material Category	Structure Form	Diameter	Length	Surface Area	Zeta Potential
NM1	1	1	Metal	Tube	300	20	322	-9.35
NM2	1	1	Metal Oxide	Tube	300	20	322	-9.35
NM3	1	1	Carbon	Tube	300	20	322	-9.35
NM10	1	1	Metal	Tube	154	10	163	-9.35
NM11	1	1	Metal Oxide	Tube	154	10	163	-9.35
NM12	1	1	Carbon	Tube	154	10	163	-9.35
NM19	1	1	Metal	Tube	7	0.13	5.1	-9.35
NM20	1	1	Metal Oxide	Tube	7	0.13	5.1	-9.35
NM21	1	1	Carbon	Tube	7	0.13	5.1	-9.35
NM4	2	2	Metal	Belt	300	20	322	-9.35
NM5	2	2	Metal Oxide	Belt	300	20	322	-9.35
NM6	2	2	Carbon	Belt	300	20	322	-9.35
NM13	2	2	Metal	Belt	154	10	163	-9.35
NM14	2	2	Metal Oxide	Belt	154	10	163	-9.35
NM15	2	2	Carbon	Belt	154	10	163	-9.35
NM22	2	2	Metal	Belt	7	0.13	5.1	-9.35
NM23	2	2	Metal Oxide	Belt	7	0.13	5.1	-9.35
NM24	2	2	Carbon	Belt	7	0.13	5.1	-9.35
NM7	3	3	Metal	Tube	300	20	322	-9.35
NM8	3	3	Metal Oxide	Tube	300	20	322	-9.35
NM9	3	3	Carbon	Tube	300	20	322	-9.35
NM16	3	3	Metal	Tube	154	10	163	-9.35
NM17	3	3	Metal Oxide	Tube	154	10	163	-9.35
NM18	3	3	Carbon	Tube	154	10	163	-9.35
NM61	3	8	Metal	Tube	300	20	322	-39.5
NM62	3	8	Metal Oxide	Tube	300	20	322	-39.5
NM63	3	8	Carbon	Tube	300	20	322	-39.5
NM70	3	8	Metal	Tube	154	10	163	-39.5
NM71	3	8	Metal Oxide	Tube	154	10	163	-39.5
NM72	3	8	Carbon	Tube	154	10	163	-39.5
NM79	3	8	Metal	Tube	7	0.13	5.1	-39.5
NM80	3	8	Metal Oxide	Tube	7	0.13	5.1	-39.5
NM81	3	8	Carbon	Tube	7	0.13	5.1	-39.5
NM25	4	4	Metal	Tube	7	0.13	5.1	-9.35
NM26	4	4	Metal Oxide	Tube	7	0.13	5.1	-9.35
NM27	4	4	Carbon	Tube	7	0.13	5.1	-9.35
NM28	5	5	Metal	Particle	300	20	322	-24.35
NM29	5	5	Metal Oxide	Particle	300	20	322	-24.35
NM30	5	5	Carbon	Particle	300	20	322	-24.35
NM34	5	5	Metal	Particle	300	20	322	-24.35
NM35	5	5	Metal Oxide	Particle	300	20	322	-24.35
NM36	5	5	Carbon	Particle	300	20	322	-24.35
NM37	5	5	Metal	Particle	154	10	163	-24.35
NM38	5	5	Metal Oxide	Particle	154	10	163	-24.35
NM39	5	5	Carbon	Particle	154	10	163	-24.35
NM43	5	5	Metal	Particle	154	10	163	-24.35
NM44	5	5	Metal Oxide	Particle	154	10	163	-24.35
NM45	5	5	Carbon	Particle	154	10	163	-24.35
NM46	5	5	Metal	Particle	7	0.13	5.1	-24.35
NM47	5	5	Metal Oxide	Particle	7	0.13	5.1	-24.35
NM48	5	5	Carbon	Particle	7	0.13	5.1	-24.35
NM52	5	5	Metal	Particle	7	0.13	5.1	-24.35
NM53	5	5	Metal Oxide	Particle	7	0.13	5.1	-24.35
NM54	5	5	Carbon	Particle	7	0.13	5.1	-24.35
NM55	5	5	Metal	Particle	300	20	322	-39.5
NM56	5	5	Metal Oxide	Particle	300	20	322	-39.5
NM57	5	5	Carbon	Particle	300	20	322	-39.5
NM64	5	5	Metal	Particle	154	10	163	-39.5
NM65	5	5	Metal Oxide	Particle	154	10	163	-39.5
NM66	5	5	Carbon	Particle	154	10	163	-39.5
NM73	5	5	Metal	Particle	7	0.13	5.1	-39.5
NM74	5	5	Metal Oxide	Particle	7	0.13	5.1	-39.5
NM75	5	5	Carbon	Particle	7	0.13	5.1	-39.5
NM31	6	6	Metal	Belt	300	20	322	-24.35

Table 6.2: Continued

NM	KSCAU	DBCAU	Material Category	Structure Form	Diameter	Length	Surface Area	Zeta Potential
NM32	6	6	Metal Oxide	Belt	300	20	322	-24.35
NM33	6	6	Carbon	Belt	300	20	322	-24.35
NM40	6	6	Metal	Belt	154	10	163	-24.35
NM41	6	6	Metal Oxide	Belt	154	10	163	-24.35
NM42	6	6	Carbon	Belt	154	10	163	-24.35
NM49	6	6	Metal	Belt	7	0.13	5.1	-24.35
NM50	6	6	Metal Oxide	Belt	7	0.13	5.1	-24.35
NM51	6	6	Carbon	Belt	7	0.13	5.1	-24.35
NM58	7	7	Metal	Belt	300	20	322	-39.5
NM59	7	7	Metal Oxide	Belt	300	20	322	-39.5
NM60	7	7	Carbon	Belt	300	20	322	-39.5
NM67	7	7	Metal	Belt	154	10	163	-39.5
NM68	7	7	Metal Oxide	Belt	154	10	163	-39.5
NM69	7	7	Carbon	Belt	154	10	163	-39.5
NM76	7	7	Metal	Belt	7	0.13	5.1	-39.5
NM77	7	7	Metal Oxide	Belt	7	0.13	5.1	-39.5
NM78	7	7	Carbon	Belt	7	0.13	5.1	-39.5

Chapter 7

Summary

This work developed a statistical framework including four stages.

Variable selection: To identify important predictors for an NM's toxicity, variable selection is first performed.

Design of experiments: In the high-dimensional space of the important predictors, design of biological experiments is performed for modeling efficiency.

Modeling and inference: To the well-designed biological data, kriging-based method is employed to quantify the dependence of an NM's toxicity upon the important predictors.

Shape clustering: Based on the toxicity profiles and estimation uncertainty rendered by the quantitative prediction model, shape clustering is carried out for potency grouping of NMs.

This framework intends to provide a statistical roadmap for the generation of QNAR (quantitative nanostructure-activity relationships) prediction model and high-throughput toxicity grouping of NMs. In particular, through simulation-based studies, the importance of experimental design in generating a high-quality QNAR is demonstrated: With limited sample size, biological experiments need to be designed in such a way that the samples at least provide a fair coverage of the high-dimensional predictor space; the efficient assignment of samples also depends on the shape of target QNAR. Without a good experimental design, the collected biological data will not contain the necessary information needed to map a comprehensive QNAR in the specified predictor (design) space.

References

[1] Ming Li, Qiaoyi Wang, Xiaodong Shi, Lawrence A Hornak, and Nianqiang Wu. Detection of mercury (ii) by quantum dot/dna/gold nanoparticle ensemble based nanosensor via nanometal surface energy transfer. *Analytical chemistry*, 83(18):7061–7065, 2011.

[2] Nianqiang Wu, Minhua Zhao, Jian-Guo Zheng, Chuanbin Jiang, Ben Myers, Shuoxin Li, Minking Chyu, and Scott X Mao. Porous cuo–zno nanocomposite for sensing electrode of high-temperature co solid-state electrochemical sensor. *Nanotechnology*, 16(12):2878, 2005.

[3] Nianqiang Wu, Zheng Chen, Jianhui Xu, Minking Chyu, and Scott X Mao. Impedancemetric pt/ysz/au–ga 2 o 3 sensor for co detection at high temperature. *Sensors and Actuators B: Chemical*, 110(1):49–53, 2005.

[4] Marinella Farré, Krisztina Gajda-Schrantz, Lina Kantiani, and Damià Barceló. Ecotoxicity and analysis of nanomaterials in the aquatic environment. *Analytical and Bioanalytical Chemistry*, 393(1):81–95, 2009.

[5] Mingjia Zhi, Ayyakkannu Manivannan, Fanke Meng, and Nianqiang Wu. Highly conductive electrospun carbon nanofiber/mno 2 coaxial nano-cables for high energy and power density supercapacitors. *Journal of Power Sources*, 208:345–353, 2012.

[6] Liming Dai, Dong Wook Chang, Jong-Beom Baek, and Wen Lu. Carbon nanomaterials for advanced energy conversion and storage. *Small*, 8(8):1130–1166, 2012.

[7] RJ1 Aitken, MQ Chaudhry, ABA Boxall, and M Hull. Manufacture and use of nanomaterials: current status in the uk and global trends. *Occupational medicine*, 56(5):300–306, 2006.

[8] Xiaobo Chen and Samuel S Mao. Titanium dioxide nanomaterials: synthesis, properties, modifications, and applications. *Chem. Rev*, 107(7):2891–2959, 2007.

[9] Ming Li, Jianming Zhang, Savan Suri, Letha J Sooter, Dongling Ma, and Nianqiang Wu. Detection of adenosine triphosphate with an aptamer biosensor based on surface-enhanced raman scattering. *Analytical chemistry*, 84(6):2837–2842, 2012.

[10] Na Ren, Rui Li, Limei Chen, Guancong Wang, Duo Liu, Yingjun Wang, Lin Zheng, Wei Tang, Xiaoqiang Yu, Huaidong Jiang, et al. In situ construction of a titanate–silver nanoparticle–titanate sandwich nanostructure on a metallic titanium surface for bacteriostatic and biocompatible implants. *Journal of Materials Chemistry*, 22(36):19151–19160, 2012.

[11] Michael Giersig and Gennady B Khomutov. *Nanomaterials for application in medicine and biology*. Springer, 2008.

[12] R Jayakumar, Deepthy Menon, K Manzoor, SV Nair, and H Tamura. Biomedical applications of chitin and chitosan based nanomaterialsa short review. *Carbohydrate Polymers*, 82(2):227–232, 2010.

[13] Marina E Vance, Todd Kuiken, Eric P Vejerano, Sean P McGinnis, Michael F Hochella Jr, David Rejeski, and Matthew S Hull. Nanotechnology in the real world: Redeveloping the nanomaterial consumer products inventory. *Beilstein journal of nanotechnology*, 6:1769, 2015.

[14] Günter Oberdörster, Eva Oberdörster, and Jan Oberdörster. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental health perspectives*, 113(7):823, 2005.

[15] Karluss Thomas and Philip Sayre. Research strategies for safety evaluation of nanomaterials, part i: evaluating the human health implications of exposure to nanoscale materials. *Toxicological Sciences*, 87(2):316–321, 2005.

[16] Vicki Stone, Stefano Pozzi-Mucelli, Karin Schberger, Stefania Sabella, Ulla B Vogel, Dominique Balharry, Teresa Fernandes, Stefania Gottardo, Steve Hankin, Mark Hartl, et al. Research prioritisation to deliver an intelligent testing strategy for the human and environmental safety of nanomaterials. 2017.

[17] Josje HE Arts, Muhammad-Adeel Irfan, Athena M Keene, Reinhard Kreiling, Delina Lyon, Monika Maier, Karin Michel, Nicole Neubauer, Thomas Petry, Ursula G Sauer, et al. Case studies putting the decision-making framework for the grouping and testing of nanomaterials (df4nanogrouping) into practice. *Regulatory Toxicology and Pharmacology*, 76:234–261, 2016.

[18] Neus Feliu and Bengt Fadeel. Nanotoxicology: no small matter. *Nanoscale*, 2(12):2514–2520, 2010.

[19] Andrew D Maynard, Robert J Aitken, Tilman Butz, Vicki Colvin, Ken Donaldson, Günter Oberdörster, Martin A Philbert, John Ryan, Anthony Seaton, Vicki Stone, et al. Safe handling of nanotechnology. *Nature*, 444(7117):267, 2006.

[20] Xiong Liu, Kaizhi Tang, Stacey Harper, Bryan Harper, Jeffery A Steevens, and Roger Xu. Predictive modeling of nanomaterial exposure effects in biological systems. *International journal of nanomedicine*, 8(Suppl 1):31, 2013.

[21] Frank Bretz, Jason Hsu, Jose Pinheiro, and Yi Liu. Dose finding–a challenge in statistics. *Biometrical Journal*, 50(4):480–504, 2008.

[22] Jeremy M Gernand and Elizabeth A Casman. A meta-analysis of carbon nanotube pulmonary toxicity studieshow physical dimensions and impurities affect the toxicity of carbon nanotubes. *Risk Analysis*, 34(3):583–597, 2014.

[23] Kunwar P Singh and Shikha Gupta. Nano-qsar modeling for predicting biological activity of diverse nanomaterials. *RSC Advances*, 4(26):13215–13230, 2014.

[24] Jürgen Pauluhn. Subchronic 13-week inhalation exposure of rats to multiwalled carbon nanotubes: toxic effects are determined by density of agglomerate structures, not fibrillar structures. *Toxicological Sciences*, 113(1):226–242, 2009.

[25] David B Warheit, Thomas R Webb, Vicki L Colvin, Kenneth L Reed, and Christie M Sayes. Pulmonary bioassay studies with nanoscale and fine-quartz particles in rats: toxicity is not dependent upon particle size but on surface characteristics. *Toxicological sciences*, 95(1):270–280, 2006.

[26] Josje HE Arts, Mackenzie Hadi, Athena M Keene, Reinhard Kreiling, Delina Lyon, Monika Maier, Karin Michel, Thomas Petry, Ursula G Sauer, David Warheit, et al. A critical appraisal of existing concepts for the grouping of nanomaterials. *Regulatory Toxicology and Pharmacology*, 70(2):492–506, 2014.

[27] Hedwig M Braakhuis, Agnes G Oomen, and Flemming R Cassee. Grouping nanomaterials to predict their potential to induce pulmonary inflammation. *Toxicology and applied pharmacology*, 299:3–7, 2016.

[28] Zhimin Tao, Bonnie B Toms, Jerry Goodisman, and Tewodros Asefa. Mesoporosity and functional group dependent endocytosis and cytotoxicity of silica nanomaterials. *Chemical research in toxicology*, 22(11):1869–1880, 2009.

[29] Tian Xia, Raymond F Hamilton, James C Bonner, Edward D Crandall, Alison Elder, Farnoosh Fazlollahi, Teri A Girtsman, Kwang Kim, Somenath Mitra, Susana A Ntim, et al. Interlaboratory evaluation of in vitro cytotoxicity and inflammatory responses to engineered nanomaterials: the niehs nano go consortium. *Environmental health perspectives*, 121(6):683–690, 2013.

[30] James C Bonner, Rona M Silva, Alexia J Taylor, Jared M Brown, Susana C Hilderbrand, Vincent Castranova, Dale Porter, Alison Elder, Günter Oberdörster, Jack R Harkema, et al. Interlaboratory evaluation of rodent pulmonary responses to engineered nanomaterials: the niehs nano go consortium. *Environmental health perspectives*, 121(6):676, 2013.

[31] Tina M Sager, Michael W Wolfarth, Michael Andrew, Ann Hubbs, Sherri Friend, Teh-hsun Chen, Dale W Porter, Nianqiang Wu, Feng Yang, Raymond F Hamilton, et al. Effect of multi-walled carbon nanotube surface modification on bioactivity in the c57bl/6 mouse model. *Nanotoxicology*, 8(3):317–327, 2014.

[32] G Dunn. optimaldesigns for drug, neurotransmitter and hormone receptor assays. *Statistics in Medicine*, 7(7):805–815, 1988.

[33] JK Lindsey, WD Byrom, Jihui Wang, P Jarvis, and Bradley Jones. Generalized nonlinear models for pharmacokinetic data. *Biometrics*, 56(1):81–88, 2000.

[34] Enrico Burello and Andrew P Worth. Qsar modeling of nanomaterials. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, 3(3):298–306, 2011.

[35] Yi Ting Chau and Chun Wei Yap. Quantitative nanostructure–activity relationship modelling of nanoparticles. *Rsc Advances*, 2(22):8489–8496, 2012.

[36] George M Woodall, Jeffrey S Gift, and Gary L Foureman. Empirical methods and default approaches in consideration of exposure duration in dose–response relationships. *General, Applied and Systems Toxicology*, 1999.

[37] Denis Fourches, Dongqiuye Pu, Carlos Tassa, Ralph Weissleder, Stanley Y Shaw, Russell J Mumper, and Alexander Tropsha. Quantitative nanostructure- activity relationship modeling. *ACS nano*, 4(10):5703–5712, 2010.

[38] Nathan M Drew, Eileen D Kuempel, Ying Pei, and Feng Yang. A quantitative framework to group nanoscale and microscale particles by hazard potency to derive occupational exposure limits: Proof of concept evaluation. *Regulatory Toxicology and Pharmacology*, 89:253–267, 2017.

[39] Matteo Cassotti and Francesca Grisoni. Variable selection methods: an introduction. 2012.

[40] Michael H Kutner, Chris Nachtsheim, and John Neter. *Applied linear regression models*. McGraw-Hill/Irwin, 2004.

[41] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[42] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[43] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[44] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[45] Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

[46] Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

[47] Donald W Marquardt and Ronald D Snee. Ridge regression in practice. *The American Statistician*, 29(1):3–20, 1975.

[48] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[49] Chris Hans. Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, 106(496):1383–1393, 2011.

[50] Tong Tong Wu, Kenneth Lange, et al. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[51] Dhermendra K Tiwari, Takashi Jin, and J Behari. Dose-dependent in-vivo toxicity assessment of silver nanoparticle in wistar rats. *Toxicology mechanisms and methods*, 21(1):13–24, 2011.

[52] Jürgen Pauluhn. Poorly soluble particulates: searching for a unifying denominator of nanoparticles and fine particles for dnel estimation. *Toxicology*, 279(1-3):176–188, 2011.

[53] Jürgen Pauluhn. Derivation of occupational exposure levels (oels) of low-toxicity isometric biopersistent particles: how can the kinetic lung overload paradigm be used for improved inhalation toxicity study design and oel-derivation? *Particle and fibre toxicology*, 11(1):72, 2014.

[54] Otmar Schmid and Tobias Stoeger. Surface area is the biologically most effective dose metric for acute nanoparticle toxicity in the lung. *Journal of Aerosol Science*, 99:133–143, 2016.

[55] Carl J Johnston, Kevin E Driscoll, Jacob N Finkelstein, R Baggs, Michael A O'Reilly, Janet Carter, Robert Gelein, and Günter Oberdörster. Pulmonary chemokine and mutagenic responses in rats after subchronic inhalation of amorphous and crystalline silica. *Toxicological Sciences*, 56(2):405–413, 2000.

[56] Brittany L Baisch, Nancy M Corson, Pamela Wade-Mercer, Robert Gelein, Andrea J Kennell, Günter Oberdörster, and Alison Elder. Equivalent titanium dioxide nanoparticle deposition by intratracheal instillation and whole body inhalation: the effect of dose rate on acute respiratory tract inflammation. *Particle and fibre toxicology*, 11(1):5, 2014.

[57] Yong Soon Kim, Jin Sik Kim, Hyun Sun Cho, Dae Sik Rha, Jae Min Kim, Jung Duck Park, Byung Sun Choi, Ruth Lim, Hee Kyung Chang, Yong Hyun Chung, et al. Twenty-eight-day oral toxicity, genotoxicity, and gender-related tissue distribution of silver nanoparticles in sprague-dawley rats. *Inhalation toxicology*, 20(6):575–583, 2008.

[58] Wan-Young Kim, Jin Kim, Jung Duck Park, Hyeon Yeol Ryu, and Il Je Yu. Histological study of gender differences in accumulation of silver nanoparticles in kidneys of fischer 344 rats. *Journal of Toxicology and Environmental Health, Part A*, 72(21-22):1279–1284, 2009.

[59] Yuying Xue, Shanshan Zhang, Yanmei Huang, Ting Zhang, Xiaorun Liu, Yuanyuan Hu, Zhiyong Zhang, and Meng Tang. Acute toxic effects and gender-related biokinetics of silver nanoparticles following an intravenous injection in mice. *Journal of Applied Toxicology*, 32(11):890–899, 2012.

[60] Helinor J Johnston, Gary Hutchison, Frans M Christensen, Sheona Peters, Steve Hankin, and Vicki Stone. A review of the in vivo and in vitro toxicity of silver and gold particulates: particle attributes and biological mechanisms responsible for the observed toxicity. *Critical reviews in toxicology*, 40(4):328–346, 2010.

[61] Fumio Watari, Noriyuki Takashi, Atsuro Yokoyama, Motohiro Uo, Tsukasa Akasaka, Yoshinori Sato, Shigeaki Abe, Yasunori Totsuka, and Kazuyuki Tohji. Material nanosizing effect on living organisms: non-specific, biointeractive, physical size effects. *Journal of the Royal Society Interface*, pages rsif–2008, 2009.

[62] Margriet VDZ Park, Arianne M Neigh, Jolanda P Vermeulen, Liset JJ de la Fonteyne, Henny W Verharen, Jacob J Briedé, Henk van Loveren, and Wim H de Jong. The effect of particle size on the cytotoxicity, inflammation, developmental toxicity and genotoxicity of silver nanoparticles. *Biomaterials*, 32(36):9810–9817, 2011.

[63] Jun Ho Ji, Jae Hee Jung, Sang Soo Kim, Jin-Uk Yoon, Jung Duck Park, Byung Sun Choi, Yong Hyun Chung, Il Hoon Kwon, Jayoung Jeong, Beom Seok Han, et al. Twenty-eight-day inhalation toxicity study of silver nanoparticles in sprague-dawley rats. *Inhalation toxicology*, 19(10):857–871, 2007.

[64] Wan-Seob Cho, Rodger Duffin, Frank Thielbeer, Mark Bradley, Ian L Megson, William MacNee, Craig A Poland, C Lang Tran, and Ken Donaldson. Zeta potential and solubility to toxic ions as mechanisms of lung inflammation caused by metal/metal oxide nanoparticles. *Toxicological Sciences*, 126(2):469–477, 2012.

[65] Dale W Porter, Nianqiang Wu, Ann Hubbs, Robert Mercer, Kathleen Funk, Fanke Meng, Jiangtian Li, Michael Wolfarth, Lori Battelli, Sherri Friend, et al. Differential mouse pulmonary dose-and time course-responses to titanium dioxide nanospheres and nanobelts. *Toxicological Sciences*, page kfs261, 2012.

[66] Hanna L Karlsson, Pontus Cronholm, Johanna Gustafsson, and Lennart Moller. Copper oxide nanoparticles are highly toxic: a comparison between metal oxide nanoparticles and carbon nanotubes. *Chemical research in toxicology*, 21(9):1726–1732, 2008.

[67] Andreas M Studer, Ludwig K Limbach, Luu Van Duc, Frank Krumeich, Evagelos K Athanassiou, Lukas C Gerber, Holger Moch, and Wendelin J Stark. Nanoparticle cytotoxicity depends on intracellular solubility: comparison of stabilized copper metal and degradable copper oxide nanoparticles. *Toxicology letters*, 197(3):169–174, 2010.

[68] Derrick A Bennett. How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, 25(5):464–469, 2001.

[69] Philip L Roth. Missing data: A conceptual review for applied psychologists. *Personnel psychology*, 47(3):537–560, 1994.

[70] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[71] Raymond F Hamilton, Nianqiang Wu, Dale Porter, Mary Buford, Michael Wolfarth, and Andrij Holian. Particle length-dependent titanium dioxide nanomaterials toxicity and bioactivity. *Particle and fibre toxicology*, 6(1):35, 2009.

[72] Dale W Porter, Ann F Hubbs, Robert R Mercer, Nianqiang Wu, Michael G Wolfarth, Krishnan Sriram, Stephen Leonard, Lori Battelli, Diane Schwegler-Berry, Sherry Friend, et al. Mouse pulmonary dose-and time course-responses induced by exposure to multi-walled carbon nanotubes. *Toxicology*, 269(2-3):136–147, 2010.

[73] Wout Slob. Dose-response modeling of continuous endpoints. *Toxicological Sciences*, 66(2):298–312, 2002.

[74] Hee Chul Park, Jinsil Seong, Kwang Hyub Han, Chae Yoon Chon, Young Myoung Moon, and Chang Ok Suh. Dose-response relationship in local radiotherapy for hepatocellular carcinoma. *International Journal of Radiation Oncology Biology Physics*, 54(1):150–155, 2002.

[75] Agneta Falk Filipsson, Salomon Sand, John Nilsson, and Katarina Victorin. The benchmark dose methodreview of available models, and recommendations for application in health risk assessment. *Critical Reviews in Toxicology*, 33(5):505–542, 2003.

[76] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[77] Brian J Williams, Thomas J Santner, and William I Notz. Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, pages 1133–1152, 2000.

[78] Ying Pei, Feng Yang, Xi Chen, Nianqiang Wu, and Kai Wang. Kriging-based design of experiments for multi-source exposure–response studies in nanotoxicology. *ACS Sustainable Chemistry & Engineering*, 5(4):3223–3232, 2017.

[79] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.

[80] V Roshan Joseph. Space-filling designs for computer experiments: A review. *Quality Engineering*, 28(1):28–35, 2016.

[81] Thomas J Santner, Brian J Williams, and William I Notz. Space-filling designs for computer experiments. In *The Design and Analysis of Computer Experiments*, pages 121–161. Springer, 2003.

[82] Wim CM Van Beers and Jack PC Kleijnen. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European journal of operational research*, 186(3):1099–1113, 2008.

[83] Peter ZG Qian. Sliced latin hypercube designs. *Journal of the American Statistical Association*, 107(497):393–399, 2012.

[84] Kai Wang, Xi Chen, Feng Yang, Dale W Porter, and Nianqiang Wu. A new stochastic kriging method for modeling multi-source exposure–response data in toxicology studies. *ACS sustainable chemistry & engineering*, 2(7):1581–1591, 2014.

[85] Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

[86] Kyung-Shik Shin, Taik Soo Lee, and Hyun-jung Kim. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, 28(1):127–135, 2005.

[87] Weiwu Yan, Huihe Shao, and Xiaofan Wang. Soft sensing modeling based on support vector machine and bayesian model selection. *Computers & chemical engineering*, 28(8):1489–1498, 2004.

[88] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.

[89] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.

[90] Robert J Schalkoff. *Artificial neural networks*, volume 1. McGraw-Hill New York, 1997.

[91] David M Giltinan and Marie Davidian. Assays for recombinant proteins: A problem in non-linear calibration. *Statistics in Medicine*, 13(11):1165–1179, 1994.

[92] Marie Davidian and David M Giltinan. *Nonlinear models for repeated measurement data*, volume 62. CRC press, 1995.

[93] Jean-Louis Steimer, Alain Mallet, Jean-Louis Golmard, and Jean-François Boisvieux. Alternative approaches to estimation of population pharmacokinetic parameters: comparison with the nonlinear mixed-effect model. *Drug metabolism reviews*, 15(1-2):265–292, 1984.

[94] Yunsong Mu, Fengchang Wu, Qing Zhao, Rong Ji, Yu Qie, Yue Zhou, Yan Hu, Chengfang Pang, Danail Hristozov, John P Giesy, et al. Predicting toxic potencies of metal oxide nanoparticles by means of nano-qsars. *Nanotoxicology*, 10(9):1207–1214, 2016.

[95] Christie Sayes and Ivan Ivanov. Comparative study of predictive computational models for nanoparticle-induced cytotoxicity. *Risk Analysis*, 30(11):1723–1734, 2010.

[96] Tomasz Puzyn, Bakhtiyor Rasulev, Agnieszka Gajewicz, Xiaoke Hu, Thabitha P Dasari, Andrea Michalkova, Huey-Min Hwang, Andrey Toropov, Danuta Leszczynska, and Jerzy Leszczynski. Using nano-qsar to predict the cytotoxicity of metal oxide nanoparticles. *Nature nanotechnology*, 6(3):175, 2011.

[97] V Chandana Epa, Frank R Burden, Carlos Tassa, Ralph Weissleder, Stanley Shaw, and David A Winkler. Modeling biological activities of nanoparticles. *Nano letters*, 12(11):5808–5812, 2012.

[98] Supratik Kar, Agnieszka Gajewicz, Tomasz Puzyn, and Kunal Roy. Nano-quantitative structure–activity relationship modeling using easily computable and interpretable descriptors for uptake of magnetofluorescent engineered nanoparticles in pancreatic cancer cells. *Toxicology in Vitro*, 28(4):600–606, 2014.

[99] Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*, volume 2. McGraw-Hill New York, 1991.

[100] Josje HE Arts, Mackenzie Hadi, Muhammad-Adeel Irfan, Athena M Keene, Reinhard Kreiling, Delina Lyon, Monika Maier, Karin Michel, Thomas Petry, Ursula G Sauer, et al. A decision-making framework for the grouping and testing of nanomaterials (df4nanogrouping). *Regulatory Toxicology and Pharmacology*, 71(2):S1–S27, 2015.

[101] Rong Liu, Robert Rallo, Saji George, Zhaoxia Ji, Sumitra Nair, André E Nel, and Yoram Cohen. Classification nanosar development for cytotoxicity of metal oxide nanoparticles. *Small*, 7(8):1118–1126, 2011.

[102] Tommi Tervonen, Igor Linkov, José Rui Figueira, Jeffery Steevens, Mark Chappell, and Myriam Merad. Risk-based classification system of nanomaterials. *Journal of Nanoparticle Research*, 11(4):757–766, 2009.

[103] Pierre Hansen and Brigitte Jaumard. Cluster analysis and mathematical programming. *Mathematical programming*, 79(1-3):191–215, 1997.

[104] MR Rao. Cluster analysis and mathematical programming. *Journal of the American statistical association*, 66(335):622–626, 1971.

[105] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870. ACM, 2015.

[106] Anuj Srivastava, Shantanu H Joshi, Washington Mio, and Xiuwen Liu. Statistical shape analysis: Clustering, learning, and testing. *IEEE Transactions on pattern analysis and machine intelligence*, 27(4):590–602, 2005.

[107] Lin Yu Tseng and Shiueng Bien Yang. A genetic clustering algorithm for data with non-spherical-shape clusters. *Pattern Recognition*, 33(7):1251–1259, 2000.

[108] Remco C Veltkamp. Shape matching: Similarity measures and algorithms. In *Shape Modeling and Applications, SMI 2001 International Conference on.*, pages 188–197. IEEE, 2001.

[109] Zhenyu Wu and Richard Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 15(11):1101–1113, 1993.

[110] Isak Gath and Amir B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 11(7):773–780, 1989.

[111] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[112] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. An efficient k-means clustering algorithm. 1997.

[113] H-P Kriegel and Martin Pfeifle. Hierarchical density-based clustering of uncertain data. In *Data Mining, Fifth IEEE International Conference on*, pages 4–pp. IEEE, 2005.

[114] Hans-Peter Kriegel and Martin Pfeifle. Density-based clustering of uncertain data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 672–677. ACM, 2005.