Toward quantitative structure activity relationship (QSAR) models for nanoparticles

Kate Odziomek[1,2], Daniela Ushizima[2], Tomasz Puzyn[1], Maciej Haranczyk[2]

[1]Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland;
[2]Computational Research Division, Lawrence Berkeley National Lab, Berkeley, CA.

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

This work was partially supported by the Office of Energy Research, U.S. Department of Energy, under Contract Number DE-AC02-05CH11231.

Towards Quantitative Structure-Activity Relationship (QSAR) Models for Nanoparticles

Katarzyna Odziomek[1,2], Daniela Ushizima[2], Tomasz Puzyn[1], Maciej Haranczyk[2]

[1]*Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland*
[2]*Computational Research Division, Lawrence Berkeley National Laboratory, One Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720-8139, USA*

1. Introduction

Quantitative structure-activity relationship (QSAR) methods employ linear and non-linear combinations of structural descriptors in prediction of physical-chemical and/or biological properties of chemical substances. Calculation of descriptors requires precise definition of the considered molecule, e.g. a molecular graph and/or a 3D structure.

Nanoparticles (NPs) are another common form of chemicals, ubiquitous in pharmaceuticals, food products, cosmetics, and other technologies. There is much interest in the ability to quantitatively predict their toxicity and other properties related to risk assessment. Nanoparticles exhibit different characteristics than bulk materials or isolated molecules; their samples are typically non-uniform and present various shapes and sizes, which gives rise to their properties. Using conventional, "molecular" QSAR techniques is therefore not possible. We aim to incorporate new nanoparticle descriptors into QSAR-proven statistical modeling methods to provide capability to estimate toxicity or biological impact of nanoparticles without expensive experimental testing.

One of the biggest challenges related to modeling nanoparticle properties is the scarcity and diversity of the structural data. Imaging techniques such as scanning electron microscopy (SEM) offer valuable insights into the morphology of NP samples. Our goal is to investigate SEM images as the source of morphological information for the statistical modeling of NP properties. We outline the first steps towards NP characterization by applying computer vision algorithms to quantify the morphological and topographical information contained in the SEM images, and convert them into numerical descriptors of NPs. Furthermore, we ask a fundamental question: how to select the optimal images one needs to properly capture both diversity and statistics of the nanoparticles present in a sample set? The motivation for this question is depicted in Figure 1, which illustrates the variation between images of the same dataset.

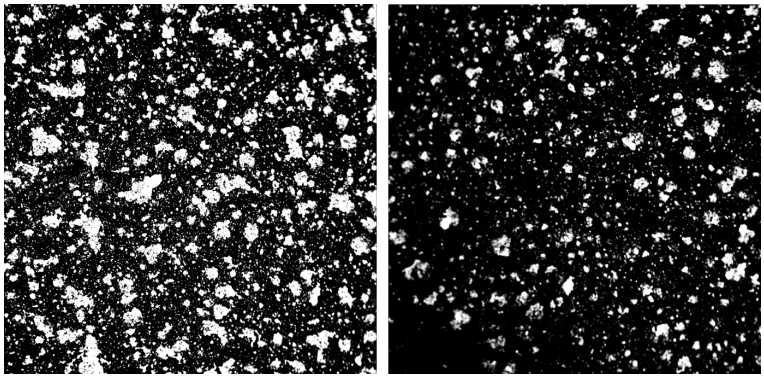

Figure 1. Two SEM images of tricalcium phosphate (TCP). Visible variation in particle shape , size and density (number) between the two micrographs.

2. Methods

2.1 Images

Tricalcium phosphate (TCP), $Ca_3(PO_4)_2$, a member of the calcium orthophosphate family, is a biocompatible and biodegradable compound used both in bulk and nanoparticle form e.g. as a component in composite biomaterials. In order to characterize sub-macro orthophosphate particles, we investigate and analyze micrographs of TCP.

Using Phenom ProX Desktop Scanning Electron Microscope (accelerating voltage: 5 000 – 15 000 V), we obtained 15 grayscale (8-bit) .tiff images of TCP grains. At x400 magnification, the 2048 by 2048 pixels (px) micrographs were scaled to 3.061 px/µm (0.3267 µm/px). The particle number, density and shape varied between SEM images (Figure 1**Error! Reference source not found.**).

2.2 Computer vision

We use an open source, Java-based program called ImageJ to construct our computer vision methodology. ImageJ offers scripting capability useful to analyze multiple images. Our analysis algorithm was divided into two sections: image processing (preparation) and image analysis (numerical transformation).

2.2.1 Image processing

Prior to any analysis, the SEM images require appropriate preprocessing, such as border improvement and contrast enhancement. In this study, we implemented a workflow with three main processing steps: filtering, thresholding and segmentation. The graphical output (result) at each stage of the procedure is presented in Figure 2, with a brief description given below:

1. FILTERING. Reducing noise, (e.g. dust, dirt, artifacts), through smoothing local variations in the image. *Method:* ***anisotropic diffusion*** *- a non-linear filter that blurs areas with similar intensity while preserving the edges.*

2. THRESHOLDING. Separating the background from the objects based on their pixel intensity (0-255). *Method:* ***intermodes****, which iteratively smoothes out the intensity histogram until there are only two maxima, and then uses their average as the limit value.*

3. SEGMENTATION. Identifying and selecting regions of interest which are most likely to correspond to NP. *Method:* ***discarding*** *objects on edges or with size* ***smaller*** *than* ***10 µm²***

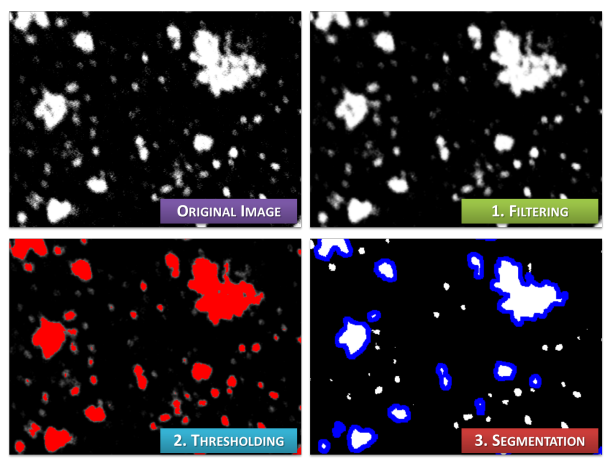

Figure 2. Output of each image processing step implemented in NP characterization.

2.2.2 Image analysis

After identifying and selecting the objects (particles) of interest, our algorithm proceeds in extracting NP features (in the form of numerical descriptors). We obtained ten shape and size descriptors, briefly described in Table 1.

Table 1. Calculated morphological NP features: shape (purple) and size (green) descriptors.

Feature	Description	Unit
Area	area of selection	**μm^{-2}**
Perimeter	length of the outside boundary of the selection	**µm**
Major Axis	primary axis of the best fitting ellipse; $S_{ellipse} = S_{particle}$	**µm**
Minor Axis	secondary axis of the best fitting ellipse; $S_{ellipse} = S_{particle}$	**µm**
Aspect Ratio	the ratio of Major Axis to Minor Axis	—
Feret's diameter MAX	maximum distance between the two parallel lines restricting the object perpendicular to a specific direction; maximum caliper	**µm**
Feret's diameter MIN	minimum distance between the two parallel lines restricting the object perpendicular to a specific direction; minimum caliper	**µm**
Circularity	comparison of the surface area of a particle to that of a circle with a perimeter of an equal length; $l_{particle} = l_{circle}$	—
Roundness	comparison of the surface area of a particle to that of a circle with a major axis (diameter) of an equal length; $MjrAx_{particle} = MjrAx_{circle}$	—
Solidity	ratio of the particle Area and Convex area; compactness	—

2.2.2.1 Representativeness

After establishing a protocol to extract useful NP descriptors from SEM images, we stood before the issue of image selection. Before an image can be used as a source of descriptors for QSAR/QSPR modeling, we must first establish its representativeness, i.e. the level of (morphological) feature diversity. A micrograph containing fewer but more diverse particles is more valuable than one with a large number of particles differing very little from each other. In some cases, it might be necessary to use more than one SEM image in order to incorporate all the possible variations of a descriptor value (particle feature).

We developed a means of assessing the significances of information carried in SEM images by employing NP statistics. We designed a series of analytical steps, and implemented an algorithm using the R statistical software:

STEP 1: Calculating the overall probability density for each descriptor. The probability density function (PDF), describing the relative likelihood for a variable to take on a specific value, is the ideal tool for assessing the range and frequency of descriptor values. The probability density was calculated using a kernel density estimation (KDE) method implemented in R's *stats* package. These kernels assume a Gaussian function to approximate the data distribution locally, following the equation below:

$$G(x;\ \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma^2}}$$

where G is the Gaussian kernel, x is a data point and σ stands for standard deviation of the data.

We used the default smoothing bandwidth estimation procedure from the *stats* package, which is based on Silverman's rule of thumb. The complete 15-SEM image set was used as the reference curve, such that the probability density for each descriptor was estimated on all the NPs from all 15 SEM images combined. The respective minimum and maximum descriptor values were used as lower and upper value

range limits. Within those set boundaries, 512 evenly spaced points compose the density estimation. Thus obtained, the descriptor population PDFs would serve as reference curve in convergence analysis.

STEP 2: Determining all the possible image combinations (subsets). In order to investigate potential procedures for optimal image selection, we took all possible image combinations (subsets) under consideration. When listing all possibilities, we started with one-image subsets, taking into account only one image at a time. Following that, we found all two-image combinations, then all three-image combinations (subsets) and so on – up to 14-image subsets. In total, we found 32,766 possible combinations (subsets) of 15 images.

STEP 3: Calculating subset probability density function for each descriptor. Implementing the method described in **STEP 1**, we calculated the PDF of each descriptor for each subset. We used the previously established descriptor boundaries (lower and upper range limits) in the subset probability density estimations, thereby ensuring the 512 sampling point intervals were identical each time. Additionally, the smoothing bandwidths selected during the overall density estimations were implemented here as default.

STEP 4: Calculating the dissimilarity score. In order to compare the overall (y_i) and subset (f_i) descriptor probability distributions at the *i-th* sample point, we calculated the mean absolute error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

We obtained 10 MAE values per subset, each for the respective descriptor. Since the mean absolute error value for two probability densities is *de facto* difference between them expressed in numerical form, we call it the *dissimilarity score (DS).*

3. Results

3.1 Particle features and corresponding dissimilarity scores

The average particle number per image oscillates around 900-1100, with the exception of images 1, 2, 3, and 12 (Figure 3 A). The majority of the particles are small, under 100 µm (Figure 3 B: Area) and quite compact (Figure 3 B: Solidity).

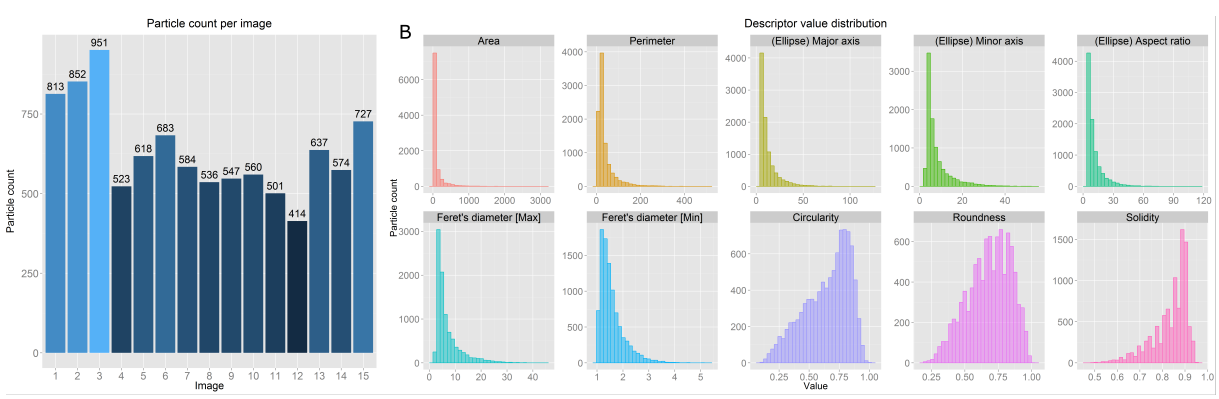

Figure 3. Particle count per image (A) and overall descriptor value distributions (B)

For all descriptors we observed a decrease of dissimilarity score values with increasing subset size (Figure 4). It is readily noticeable in the form of a "shift" from highest DS values for 1-element subsets

(bottom, peach-colored bar) to lowest values for 14-element subsets (top, pink closed bar) for each feature.

Moreover, the dissimilarity score values depend on descriptor type: size descriptors have smaller DS values (10^{-6} – 10^{-3}), whereas shape descriptors tend to reach higher DS values (10^{-3} – 10^{-1}).

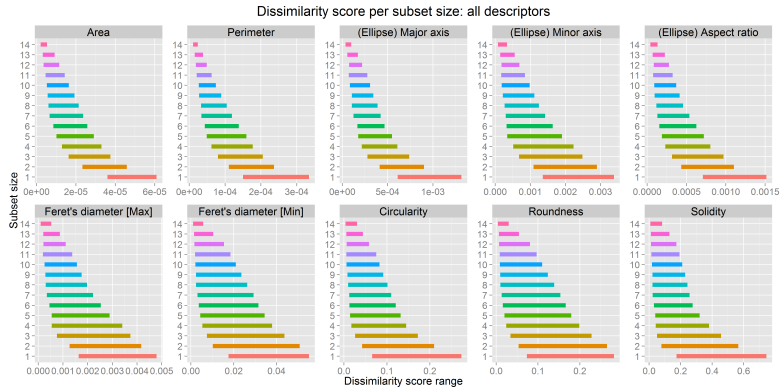

Figure 4. Dissimilarity score per subset size for each descriptor. Differently colored bars represent combined DS values for all image subsets of a given size.

Focusing on one of the NP descriptors, e.g. circularity, we notice that some of the subsets have lower DS values than others, despite being of equal size (Figure 5 A). This difference arises from the fact that image subsets with lower DS contain NPs with more diverse values for circularity. Moreover, certain subsets of different sizes have the same DS (Figure 5 A), signaling not only the number but also the choice of images dictates a subset's representativeness. This phenomenon occurs in all descriptors.

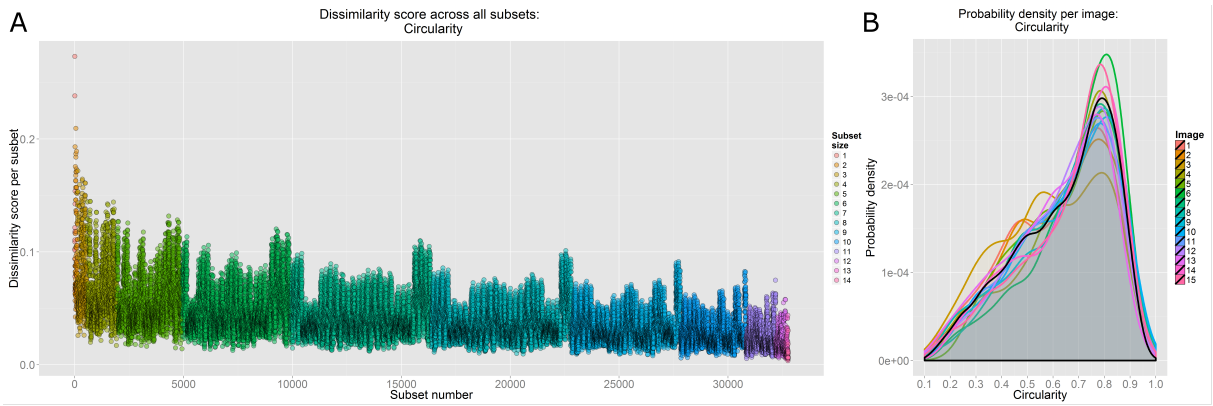

Figure 5. Dissimilarity score of all subsets for Circularity; each dot represents a subset, colored by its size (A), and probability density of Circularity values per image (B).

As illustrated in Figure 5 B, the overall probability density (black line), representing all images, and the probability density estimates for each image (various colors) vary significantly. The PDF curve of Image 11 (blue polygon) almost overlaps with the overall PDF curve, meaning Image 11 is the most representative in terms of circularity. Conversely, the PDF curve for Image 3 (yellow polygon) differs significantly from the overall, making it the least representative of all 15 SEM images.

3.2 Image selection method

We developed an effective way of selecting the most representative images using the lowest dissimilarity score (Tables 2 and 3) as a benchmark. For a given number of images (specific subset size), the lowest DS value will indicate the most representative image combination.

Interestingly, despite containing the largest number of particles, Image 3 is not the most representative according to any descriptor. It is not present in any of the 1-image subsets with the lowest DS values,

proving that image particle count is not indicative of feature diversity and should not be used as a factor during the image selection process.

When comparing subsets in terms of circularity, starting with the smallest one, we observe that Image 11 had the lowest dissimilarity score value (Table 2), meaning it is the most representative of the group. When selecting the two most representative images, we should combine Image 2 and Image 4, as they had the lowest DS value out of all 2-image subsets. When selecting three most representative Images we should combine Images 1,4, and 9, etc.

Table 2. List of Images from the most representative image combinations of particular size based on Circularity. Each Image was assigned a unique color for clarity.

Subset size	Lowest DS [x10^{-2}]	Image list													
1	6.550	11	-	-	-	-	-	-	-	-	-	-	-	-	-
2	4.232	2	4	-	-	-	-	-	-	-	-	-	-	-	-
3	2.594	1	4	9	-	-	-	-	-	-	-	-	-	-	-
4	1.704	9	10	13	14	-	-	-	-	-	-	-	-	-	-
5	1.418	9	10	11	13	14	-	-	-	-	-	-	-	-	-
6	1.276	2	3	6	7	12	15	-	-	-	-	-	-	-	-
7	1.213	2	3	6	7	11	12	15	-	-	-	-	-	-	-
8	0.953	2	3	6	7	10	13	14	15	-	-	-	-	-	-
9	0.804	2	3	6	7	10	12	13	14	15	-	-	-	-	-
10	0.597	1	2	3	4	5	6	7	8	12	15	-	-	-	-
11	0.549	1	2	3	4	5	6	7	8	11	12	15	-	-	-
12	0.607	1	2	3	4	6	7	9	10	11	12	14	15	-	-
13	0.554	1	2	3	4	5	6	7	8	9	11	13	14	15	-
14	0.364	1	2	3	4	5	6	7	8	9	10	12	13	14	15

Table 3. List of Images from the most representative image combinations of particular size based on Circularity. Each Image was assigned a unique color for clarity.

Subset size	Lowest DS [x10^{-5}]	Image list													
1	3.602	4	-	-	-	-	-	-	-	-	-	-	-	-	-
2	2.332	3	13	-	-	-	-	-	-	-	-	-	-	-	-
3	1.628	1	5	9	-	-	-	-	-	-	-	-	-	-	-
4	1.291	3	6	11	15	-	-	-	-	-	-	-	-	-	-
5	1.015	1	3	4	7	15	-	-	-	-	-	-	-	-	-
6	0.854	1	3	4	7	11	15	-	-	-	-	-	-	-	-
7	0.669	1	3	5	6	9	11	15	-	-	-	-	-	-	-
8	0.606	1	3	5	6	9	11	12	15	-	-	-	-	-	-
9	0.567	1	3	5	6	8	9	11	12	15	-	-	-	-	-
10	0.526	1	2	3	4	7	8	11	13	14	15	-	-	-	-
11	0.463	1	2	3	5	6	9	10	11	12	13	15	-	-	-
12	0.360	1	2	3	5	6	8	9	10	11	12	13	15	-	-
13	0.318	1	2	3	5	6	7	8	9	10	11	13	14	15	-
14	0.209	1	2	3	5	6	7	8	9	10	11	12	13	14	15

When choosing a set of SEM images representative in terms of particle area (Table 3), we find that the best combinations, those with the lowest DS value, differ from the subsets for Circularity. Here, Image 4 is most representative of the whole set, whereas for circularity it was Image 11. For two most representative images, we should combine Images 3 and 13, not 2 and 4, as was the case with circularity.

In fact, all descriptors have their own sets of optimal (representative) image combinations, independent from each other.

4. Conclusions

We provided a framework for extracting morphological descriptors of nanoparticles from SEM images. Using this framework, we have looked at 15 SEM images of TCP material, and analyzed their NP distributions in terms of their shape and size.

We have developed a statistical means for image comparison, enabling the selection of most representative images and image sets. We have demonstrated that when choosing a representative set of SEM images, one should make the selection separately for each descriptor. We have proven that information quality (feature diversity) is independent of the number of particles in an image, and the most populated images are not always the optimal choice.

The next stage of our research will focus on devising a measure of information content for SEM images, based on the distribution of particle features. Following that, we will compare the morphological information obtained by means of computer vision for different biomaterials and group them accordingly to those features, looking for natural clusters. Afterwards, we will investigate relationships between particle morphology and their biological properties.