

1-1-2022


## Shape investigations of structures formed by the self-assembly of aromaticamino acids using the density-based spatial clustering of applications with noise algorithm

MEHMET GÖKHAN HABİBOĞLU

HELEN W. HERNANDEZ

ŞAHİN UYAYER

Follow this and additional works at: <https://dctubitak.researchcommons.org/elektrik>

 Part of the [Computer Engineering Commons](#), [Computer Sciences Commons](#), and the [Electrical and Computer Engineering Commons](#)

---

### Recommended Citation

HABİBOĞLU, MEHMET GÖKHAN; HERNANDEZ, HELEN W.; and UYAYER, ŞAHİN (2022) "Shape investigations of structures formed by the self-assembly of aromaticamino acids using the density-based spatial clustering of applications with noise algorithm," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 30: No. 1, Article 14. <https://doi.org/10.3906/elk-2003-144>  
Available at: <https://dctubitak.researchcommons.org/elektrik/vol30/iss1/14>

This Article is brought to you for free and open access by TÜBİTAK Academic Journals. It has been accepted for inclusion in Turkish Journal of Electrical Engineering and Computer Sciences by an authorized editor of TÜBİTAK Academic Journals.

## Shape investigations of structures formed by the self-assembly of aromatic amino acids using the density-based spatial clustering of applications with noise algorithm

M. Gökhan HABİBOĞLU<sup>1,\*</sup>, Helen W. HERNANDEZ<sup>2</sup>, Şahin UYAYER<sup>3</sup>

<sup>1</sup>Department of Electrical and Electronics Engineering, Faculty of Engineering, Turkish-German University, İstanbul, Turkey

<sup>2</sup>KAL Research Initiatives LLC, Houston, TX, United States

<sup>3</sup>Department of Energy Science and Technology, Faculty of Science, Turkish-German University, İstanbul, Turkey

Received: 24.03.2020

Accepted/Published Online: 24.10.2020

Final Version: 19.01.2022

**Abstract:** Tyrosine, tryptophan, and phenylalanine are important aromatic amino acids for human health. If they are not properly metabolized, severe rare mental or metabolic diseases can emerge, many of which are not researched enough due to economic priorities. In our previous simulations, all three of these amino acids are discovered to be self-organizing and to have complex aggregations at different temperatures. Two of these essential stable formations are observed during our simulations: tubular-like and spherical-like structures. In this study, we develop and implement a clustering analyzing algorithm using density-based spatial clustering of applications with noise (DBSCAN) to measure the shapes of the formed structures by the self-assembly processes of these amino acids. We present the results in quantitative and qualitative ways. To the best of our knowledge, the geometric shapes of the formed structures by the self-assembly processes of these amino acids are not measured quantitatively in the literature. Analytical measurements and comparisons of these aggregations might help us to identify the self-aggregations quickly at early stages in our simulations and hence provide us with more opportunity to experiment with different parameters of the molecular simulations (like temperature, mixture rates, and density). We first implement the DBSCAN method to identify the main self-aggregation cluster and then we develop and implement two algorithms to measure the shapes of the formed structures by the self-assembly processes of these amino acids. The measurements, which are completely in line with our simulation results, are presented in quantitative and qualitative ways.

**Key words:** Clustering analysis, density-based spatial clustering of applications with noise, self-assembly, amino acids, sphericity, cylindricity

### 1. Introduction

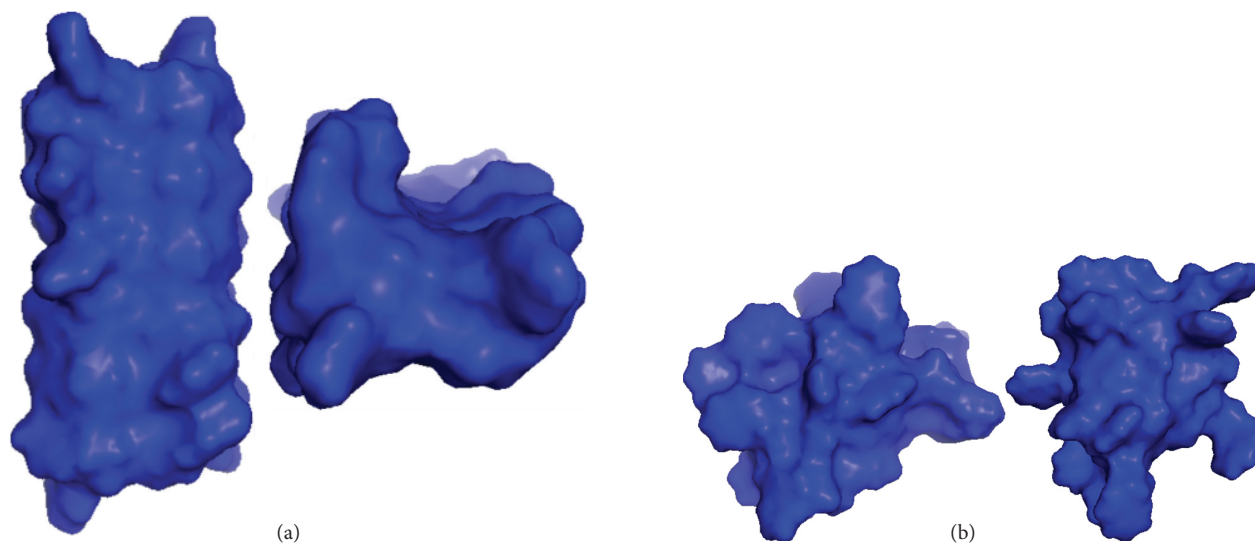
An amino acid is a molecule which has a carbon atom, called  $\alpha$ -carbon, with an amine group, a carboxyl group, a hydrogen, and an R group, or side chain, which determines the identity of the molecule. Some side chains are hydrophobic, while some are hydrophilic, some are amphipathic. The 20 common R groups give rise to the 20 common amino acids present in the human body. When two amino acids are connected to each other covalently, they are joined through a peptide bond, where the carboxyl group of one amino acid connects to the amine group of the next amino acid as the result of a dehydration reaction. When multiple of these amino acid units are connected to each other, they form a polymer more commonly termed a protein. Thousands

\*Correspondence: uyayer@tau.edu.tr

of different proteins are built from these 20 amino acids and are essential to normal human function. The noncovalent interactions (hydrogen bonding, electrostatic interaction, pi-stacking, etc.) between amino acids and surrounding solvent molecules play a very important role in the protein folding process, but not all protein folding, because some may require some folding moderators [1]. Since proteins are polymers of amino acids, amino acids are the core units in the protein folding process. In the case of protein folding, the amino acids interact with one another but are restricted in how they can interact by their relative location on the polymer strand which governs their proximity, as well as limitations in rotations around the protein backbone which limits the angles of the approach of one amino acid to another in the same protein. When single, isolated amino acids are considered, the same noncovalent interactions are at play, but with much more freedom. Many of the fundamental amino acids are known to form self-assembled structures [2–5]. The understanding of this self-assembly process is of high importance in the study of metabolic diseases and also in material science. For instance, phenylalanine (Phe) molecules assemble into different nanostructures like 4-fold, fibrillar, or zig-zag structures. The accumulation of Phe molecules due to a genetic disorder occurs in the disease known as phenylketonuria (PKU). PKU is fairly rare worldwide with the highest degree of prevalence in Turkey and some of the lowest degrees of prevalence in Finland and Japan [6]. The management of this disease requires constant attention to what is being consumed in the affected individual's diet since phenylalanine naturally occurs in many food sources. There are low adherence rates to the strict diet needed to maintain recommended blood phenylalanine levels, especially in adolescence [7]. If this disease is left unmanaged, severe neuro-degeneration occurs. Different degenerative diseases are connected to other isolated amino acids [8]. The study of the self-assembly of single amino acids is still less worked out compared to the cases of many peptides. Recently, in our molecular dynamics simulations, we have investigated the self-assembly processes of the amino acids with aromatic rings, namely, phenylalanine (Phe), tyrosine (Tyr), and tryptophan (Trp) [9–12].

Investigating these formed structures is important to elucidate the aspects of the self-assembly processes and to understand the resulting structures. Since the self-assembly process can be seen as making clusters or groups based on the chemical attributes between the molecules, clustering analysis is mostly done on the obtained data. Clustering analysis is the task of grouping a set of objects, in the case of the self-assembly of individual amino acids, it is a set of molecules, in such a way that the objects in the same set or cluster are seen as members of a structure even though they are not linked by covalent bonds. The algorithm of cluster analysis is not specific, but it may require an extensive effort. Popular methods to determine a cluster are evaluating small distances between objects, dense areas of the data space, intervals, or particular statistical distributions [13–16]. The clustering is not an automated task, but an iterative calculation that is needed to figure out the clusters on the data. It is a commonly used technique in various areas from biology to social science.

This work is devoted to the structural analysis of the self-assemblies of aromatic amino acids, namely Phe, Tyr, and Trp. We take the data obtained in our previous molecular dynamics simulations [10] and focus on describing them mathematically as cylindrical and spherical structures (see Figure 1). In this figure, we display the snapshots of one typical cylindrical structure and one globular or spherical one. We investigate the structures of the molecules of the amino acids implementing the DBSCAN algorithm. We analyze several of these formed structures in terms of how far they deviate from a perfect cylindrical or spherical shape. Any other remaining structure is not included in the paper. The paper is outlined as follows: In the next section, after giving the details of the system and method, we describe the mathematical aspects underlying in the analyses. Next, we present the results and provide our discussion.



**Figure 1.** Some of the structures obtained in the simulations of aromatic amino acids. (a) Cylindrical “tubelike” structure (Tyr molecules at 350 K in the picture). The cross-sectional view of this structure is given on the right. (b) Globular spherical structures (Trp molecules at 275 K in the picture). The data were taken from [10].

## 2. Materials and methods

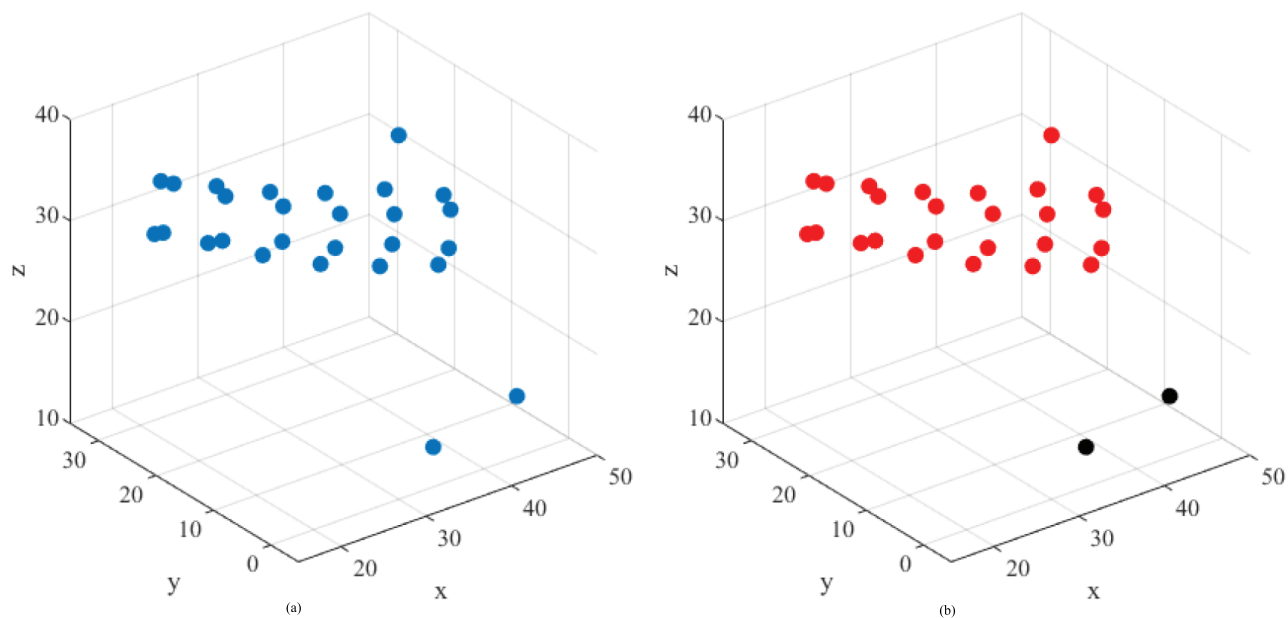
We use the data from our previous work [10], where the aromatic amino acids, Phe, Tyr, and Trp, were simulated using Gromacs software at four different temperatures, namely  $T = 275$  K, 300 K, 325 K and 350 K, and were shown to form various nano-structures including 4-fold (or tube-like) structures or just aggregation (globules). In the current work we have made our analyses over 30,000 frames per each amino acid per temperature. The self-assembly of aromatic amino acids at different temperatures seems to take the shape of either a cylinder-like tube or sphere-like aggregation. To have a better analytical understanding about the formation of these structures, two measures are used in this study. We apply one cylindricity measure for cylindrical aggregations and use the well-known sphericity measure for spherical formations [12]. We have also taken the ratio of the largest cluster to the total number of molecules into account to reflect the contribution of the degree of clustering.

### 2.1. Clustering

DBSCAN is a deterministic and density-based clustering algorithm, where data points are searched for the regions of different data densities. In this algorithm, density regions are clustered based on data densities [17, 18]. This algorithm is less irritable to occupants and able to find clusters of improper shapes. If a density criterion, minimum number of points within a radius, is fulfilled, these points are connected and an impulsive shape is drawn to consist of these points. The algorithm starts with an arbitrary occurrence (p) in a data set (D) and acquires all distances of D with respect to the radius of the neighborhood of a point based on a distance (Eps) and to the minimum number of points required to be in a cluster (MinPts). Eps can be selected as Euclidean, Manhattan, or Minkowski metric. DBSCAN locates data points within Eps distance from the centers of the clusters [19]. It is widely implemented in the research of satellite images, X-ray crystallography, anomaly detection in temperation and so on [20]. For example, Abdolzadegan et al. [21] benefited from the DBSCAN algorithm to detect the diagnosis of autism spectrum disorder (ASD) from EEG (electroencephalogram) signal.

The algorithm and its usage are still of high interest. The authors in [22] utilized the DBSCAN as minimizing the search area in digital images for single and multiple copy-move forgery detection and localization. Zhang et al. [23] applied the DBSCAN method in another very common research on belief-rule-based system, where they suggested an approach to reduce extended belief rule base. An improved DBSCAN algorithm was implemented at visible light communication systems to improve the signal-to-noise ratio and weaken the damage of noise to the communication quality [24]. A few more studies utilizing the DBSCAN algorithm in different fields of research can be found in very recent studies [25–28]. On the other hand, importantly, in a recent article, the authors [29] searched a way to determine the correct values of the DBSCAN parameters, by detecting the sharp increase in distance. The approach in this study may reduce the difficulty of estimating the DBSCAN parameters. A modification of the DBSCAN algorithm was proposed in [30], in which the clusters with various data densities in a given set of data points are recognized, but in an efficient way in clustering a set of data. In a separate work, the same authors proposed an efficient method for solving the multiple generalized circle detection method [31], in which a combination of the k-means and DBSCAN algorithms is achieved.

To measure sphericity or cylindricity of the formed structures, we first need to identify the isolated monomers, if any exists. After removing these isolated monomers, we obtain a large cluster, “major cluster”, the shape of which will be studied in the following sections both in terms of sphericity and cylindricity. C- $\alpha$  atoms at each molecule are considered to be the center of mass, representing that molecule, which is used for the clustering process instead of all atoms (see Figure 2).



**Figure 2.** (a) Aggregation of 27 Tyr molecules. (b) Isolated monomers are identified (black dots) and removed to obtain the major cluster (red).

The clustering process is achieved by using the DBSCAN algorithm as characterized in [17, 32, 33]. All core and noncore molecules within a radius of Eps form a cluster. All other molecules, which are further from any molecule at that cluster than the radius of Eps, form either another cluster or outliers (in our case they are called isolated molecules). For sphericity measurement, after manually tuning, the parameters needed for

the algorithm are selected as  $Eps = 8$  and  $MinPts = 4$ , whereas for cylindricity measurements they are selected as  $Eps = 7$  and  $MinPts = 3$ . The parameters for these two cases are given in Table 1. Euclidean metric is used for the  $Eps$  parameter. DBSCAN has mainly three strong properties for our purpose as compared to other clustering algorithms: (i) DBSCAN does not need prior knowledge of the specific number of clusters. Our data consists of 30,000-time frames for each amino acid, where the major cluster formed due to self-aggregation should be identified, (ii) the self-aggregation process ends up in a large major cluster and a few outliers, in our case the monomers, nearby. Since the clustering of self-aggregation can be mainly seen as a density-based spatial clustering problem, the outlier molecules correspond to the noise in the clustering problem, which can be easily handled by the DBSCAN algorithm (as its name suggests), (iii) the self-aggregation observed in three amino acids are mostly in spherical or cylindrical shapes. DBSCAN is known to identify clusters in arbitrary shapes as opposed to some other algorithms like the k-mean, k-medoids, Gaussian mixture modeling which usually cluster well enough a specific class of shapes. Thus, this algorithm is less irritable to occupants and able to find clusters of improper shapes. On the other hand, the well-known algorithm DBSCAN may have limitations if the data set is too sparse and the densities vary considerably. Moreover, if the system consists of a huge number of particles, one may need to make a special effort for the performance of the calculation. If the data contains multiple regions with different densities, DBSCAN might not cluster properly. It is important to emphasize that our data at each time frame is reduced to 27 points ( $C-\alpha$  atoms). Therefore, there is not much room to have multiple clusters with different densities. At each time frame, we have usually either a large major cluster and a few outliers, or a medium-sized major cluster and one or two minor clusters. The success of DBSCAN clustering depends highly on the parameters of  $MinPts$  and  $Eps$ . Inappropriate selection of these parameters might end up with less suitable clustering. We have manually selected and tested different parameters, and observed that the parameters given in Table 1 provide us with satisfactory results for our purpose. The DBSCAN algorithm has difficulties to cluster high-dimensional data; however, our data set contains only the 3-dimensional spatial coordinates of the atoms. Thus, we do not suffer from the so-called ‘‘curse of dimensionality’’. For these reasons, DBSCAN is preferred over other clustering algorithms to distinguish the monomers from the major cluster.

**Table 1.** The parameters of DBSCAN at estimating sphericity and cylindricity.

	Sphericity	Cylindricity
$Eps$	8	7
$MinPts$	4	3

After the DBSCAN algorithm is implemented, if there is a major cluster with an aggregation of 4 molecules or below for cylindricity and 12 molecules or below for sphericity, then these aggregations are counted as ‘‘nonsignificant’’ and those time frames are excluded from our quantitative analysis. This incident occurs at the early stages of the simulation, where the molecules still need time for proper aggregation.

## 2.2. Sphericity

Classical compactness of a 3D object can be interpreted as the relation between the object’s volume and its enclosing surface area. The measure is given as  $A^3/V^2$ , where  $A$  and  $V$  represent the area and the volume of the object respectively; hence, it is dimensionless [34]. The essential property of this measure for our study is that the sphere minimizes the measure by  $36\pi$ . The well-known sphericity [35, 36] is specifically defined to

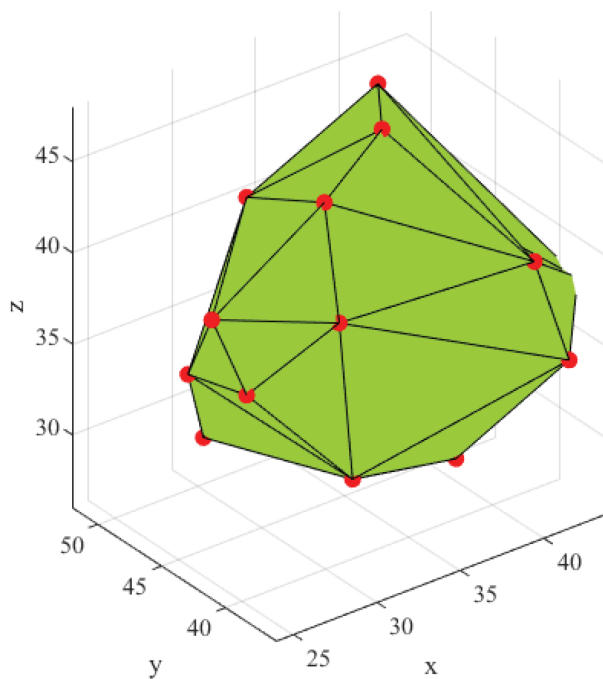
measure the compactness in sphere-like objects and can be defined as follows:

$$\Psi = \frac{(36V^2\pi)^{1/3}}{A}, \quad (1)$$

where  $V$  and  $A$  represent the volume and area of the shape, respectively. This measure is maximized for a perfect sphere at value 1.0 and decreases as the shape diverges from a perfect sphere. In our study, we first use the DBSCAN [17,18] algorithm to cluster the molecules as defined above. Any cluster consisting of less than 12 molecules is considered to be an ignorable cluster and is therefore disregarded. Then, we created a 3-dimensional minimum envelope around the major cluster, which encloses all C- $\alpha$  atoms and hence provides us a 3-dimensional solid shape for the amino acid assembly (see Figure 3). After that, the volume and the area of this shape are measured to calculate the sphericity of the shape. Finally, the effect of the ratio of the number of molecules forming the spherical shape to the total number of molecules is also considered. Since a sphere-like structure in a crowded aggregation is more significant for our purpose compared to a perfect sphere formation in a tiny cluster, we have taken this effect into consideration by multiplying sphericity with the ratio of the number of the molecules in the major cluster to the total number of molecules and thus defined weighted sphericity measurement:

$$\Psi_R = R \times \Psi, \quad (2)$$

where  $R$  is the ratio of the number of molecules in the major cluster  $n_c$  to the number of all molecules  $n_{all}$ . If all molecules form a single aggregation (i.e. there are no isolated monomers or small clusters), then  $R$  reaches its maximum value 1. Hence,  $\Psi_R$  is maximal at  $R = 1$  if all molecules form a perfect sphere and thus can take the values between [0,1].



**Figure 3.** Sketch of the aggregation of 27 Trp molecules at 275 K and 300 ns.



### 2.3. Cylindricity

Cylindricity measures are of high interest in the efficiency and cost-effectiveness of systems, like liquid injection systems. There are many ways of defining the cylindricity of an object. One way is, for example, the minimum zone method of analysis [37, 38], in which an ideal error value is approached and International Organization for Standardization standard is verified [39, 40]. Murthy conducted the measurement and evaluation of cylindrical objects through the algorithm based on the orthographic projection of the axis of the cylinder and normal least squares fit [41]. Considering that zone construction is a very complex geometric problem, and considering that even the zone hyperboloid technique does not fit to the solutions of many problems, Lao et al. proposed a method in which axis estimation and hyperboloid technique provide an integrated methodology for cylindricity evaluation [42]. On the other hand, the limitations of the measurement instruments and techniques for cylindricity were analyzed in [43], where possible improvements were also studied.

To measure the degree of a cylindrical structure formed, we have developed a simple application of cylindricity. After the major cluster is identified, we first calculate the 3-dimensional line of best fit, which passes through all molecules (Figure 4). The standard deviation of all molecules to this line gives us an analytical understanding of how uniformly the molecules are distributed around this axis. In a perfect cylinder, the standard deviation would be 0, since all points on the surface of this hypothetical cylinder have the same distance to this axis. Therefore, as this deviation increases, we can state that the similarity of the shape to a perfect cylinder decreases. The number of molecules in the major cluster has also a significant role in determining how strong the formation of molecular assembly is. We have taken this effect into consideration by dividing the standard deviation explained above by the ratio of the number of all molecules in the major cluster to the number of all molecules:

$$\varphi = \frac{\sigma(d_i)}{R}, \quad (3)$$

where  $d_i$  is the distance of the  $i^{th}$  molecule to the best line of fit,  $\sigma_i$  is the standard deviation operator,  $R$  is the ratio of the number of molecules in the major cluster to the number of all molecules, and  $\varphi$  is the weighted cylindricity, which can take the values between  $[0, \infty]$ . The reason  $R$  contributes as a fractional parameter is that  $\varphi$ , on the contrary of sphericity, decreases as the aggregation is more cylinder-like.

## 3. Results

### 3.1. Sphericity

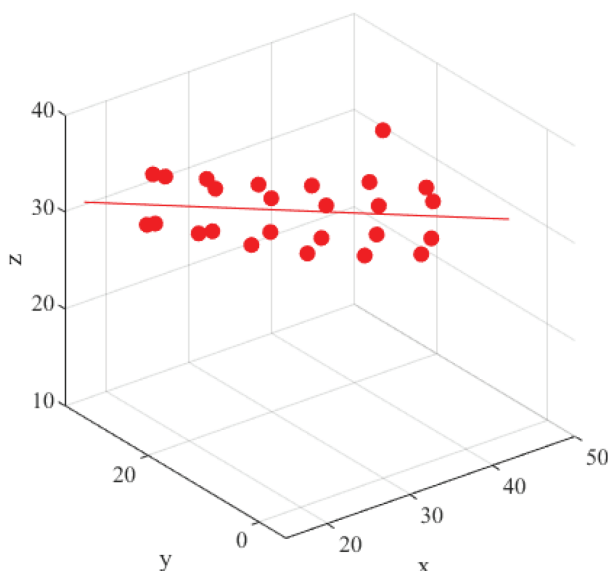
The ratio of molecules in the major cluster to all molecules might have a significant role in the sphericity measure. To have a better understanding of  $\Psi_R$ , Table 2 summarizes the average  $R$  over 250-3-00 ns of simulation results.

**Table 2.** The time averages of  $R$  between 250 and 300 ns for all 3 amino acids at 4 different temperatures when clustering for sphericity.

Amino acid	275 K	300 K	325 K	350 K
Tyr	0.8005	0.4972	0.9134	0.9056
Phe	0.5758	0.2552	0.5352	0.5868
Trp	0.8469	0.6911	0.3912	0.6916

For Phe at 300 K, for example, a very low value of  $R$  indicates that for those 50 ns the molecules are not very well assembled and while ordered assemblies may be observed, they are relatively unstable in time. In





**Figure 4.** Line of best fit through all the molecules.

contrast, Tyr assemblies at high temperatures give very high  $R$  values indicating that most of the molecules are found in the major cluster throughout the final 50 ns. The weighted sphericity measurement  $\Psi_R$  for all 3 amino acids at different time intervals are given in Tables 3–5.

**Table 3.** Time averages of  $\Psi_R$  for Tyr at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	0.3714	0.3382	0.4054	0.4307
100–200	0.5452	0.5054	0.4874	0.5663
200–250	0.5405	0.4891	0.5655	0.5639
250–300	0.5444	0.4583	0.5748	0.5606

**Table 4.** Time averages of  $\Psi_R$  for Phe at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	0.3470	0.3513	0.3927	0.3476
100–200	0.3986	0.4170	0.3905	0.3748
200–250	0.4927	0.3754	0.3574	0.3744
250–300	0.4369	0.3582	0.3886	0.4107

Table 3 supports the qualitative observation that Tyr molecules equilibrate relatively quickly at all temperatures compared to the other two aromatic amino acids. However, at 325 K, Tyr molecules continue to change their shape roughly up to 200 ns and only then start to stabilize. Tyr molecules assemble at 325 K in slightly more sphere-like structures compared to other temperatures. At the time interval of 250–300 ns the deviations of the formed structures from being a perfect sphere ( $\Psi_R = 1$ ) is calculated as 0.4252%, 0.4394 %, 0.4556%, and 0.5417% for the temperatures 325 K, 350 K, 275 K, and 300 K, respectively. Thus, the degrees

**Table 5.** Time averages of  $\Psi_R$  for Trp at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	0.5418	0.5394	0.4732	0.4328
100–200	0.5757	0.4839	0.4385	0.5490
200–250	0.6896	0.4908	0.3766	0.5499
250–300	0.6754	0.4733	0.3903	0.4645

of sphericity are

$$\Psi_R^{325K} > \Psi_R^{350K} > \Psi_R^{275K} > \Psi_R^{300K}.$$

Phe molecules exhibit low  $\Psi_R$  at all temperatures. One important reason for this is the relatively low average values of  $R$ . This fact is in agreement with [10], where Phe molecules are found to be in more fluctuating structures compared to Try and Trp molecules. In the last 50 ns, the deviations from being a perfect sphere are 0.5631%, 0.5896%, 0.6114%, and 0.6418% for the temperatures 275 K, 350 K, 325 K, and 300 K. Thus,

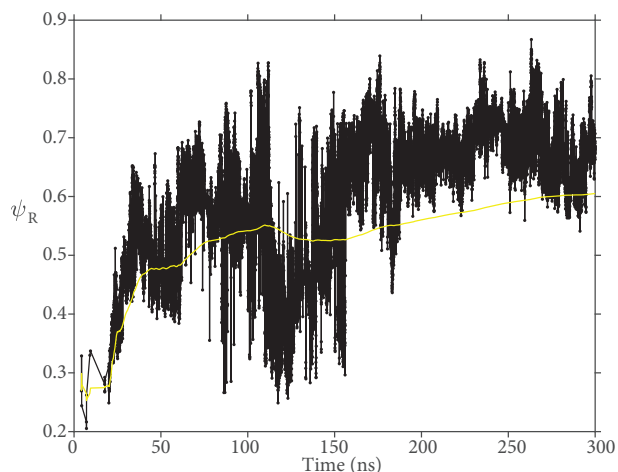
$$\Psi_R^{275K} > \Psi_R^{350K} > \Psi_R^{325K} > \Psi_R^{300K}.$$

The structure formation rates of Phe molecules at all temperatures are similar, where the molecules reach close to their final form approximately after 100 ns.

When we compare the equilibrium states in the last 50 ns, Trp molecules at 275 K have the strongest sphericity among all 3 amino acids. As calculated for Tyr and Phe molecules above, the deviations of the structures from being a perfect sphere in the last 50 ns are 0.3246%, 0.5267%, 0.5355%, and 0.6097% for the temperatures 275 K, 300 K, 350 K, and 325 K, respectively. Thus, the degrees of sphericity are:

$$\Psi_R^{275K} > \Psi_R^{300K} > \Psi_R^{350K} > \Psi_R^{325K}.$$

At 300 K and 325 K, the decline in sphericity after 100 ns is especially notable. The evolution of  $\Psi_R$  at 275 K and its moving average is depicted in Figure 5.

**Figure 5.** Evaluation of  $\Psi_R$  and its moving average (yellow line) for Trp at  $T = 275$  K.

Consequently, when all 3 amino acid types are compared at the temperatures they reach their highest weighted sphericity average  $\Psi_R$  between 250 and 300 ns, Trp exhibits the highest sphere-like formation:

$$\Psi_R^{Trp,275K} > \Psi_R^{Tyr,325K} > \Psi_R^{Phe,275K}.$$

### 3.2. Cylindricity

The ratio of molecules in the major cluster to all molecules also has an important impact on the cylindricity measure. We give the average of  $R$  between 250 and 300 ns for all 4 amino acids in Table 6.

The cylindricity measures for all 3 amino acids at different time intervals are given in Tables 7–9. Tyr molecules exhibit a cylindrical shape at 350 K, which is in line with our simulation observations.

**Table 6.** The time averages of  $R$  between 250 and 300 ns for all 3 amino acids at 4 different temperatures when clustering for cylindricity.

Amino acid	275 K	300 K	325 K	350 K
Tyr	0.7861	0.5137	0.9071	0.8999
Phe	0.4623	0.4215	0.5138	0.5630
Trp	0.8034	0.6401	0.4577	0.6754

**Table 7.** Time averages of  $\varphi$  for Tyr at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	3.8435	4.2351	3.0616	2.8542
100–200	2.5535	3.0970	2.6313	0.7709
200–250	1.8967	2.6158	2.8321	0.6293
250–300	1.7229	3.4249	2.8342	0.6198

**Table 8.** Time averages of  $\varphi$  for Phe at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	3.6790	4.1312	3.0770	3.7167
100–200	3.2665	2.5203	1.8767	3.5574
200–250	2.8318	2.4996	2.1288	2.4459
250–300	2.8670	2.5545	1.9667	1.5129

**Table 9.** Time averages of  $\varphi$  for Trp at 4 different temperatures.

Time interval (ns)	275 K	300 K	325 K	350 K
0–100	4.0080	3.4203	3.1520	2.3958
100–200	3.1043	3.2516	3.0593	2.1426
200–250	2.9322	2.8113	2.3792	2.2987
250–300	3.0268	2.5574	2.2756	1.4113

Furthermore, the aggregation is achieved after approximately 100 ns. Thus, a stronger and quicker formation is obtained at this temperature compared to other temperatures. It is also interesting to observe that although  $R$  at 325 K is sufficiently high for a complex structure, its cylindricity measure indicates that a good cylinder-like shape is not formed. When simulation results are investigated, one can easily observe the reason behind these values: A crystal-like shape is formed at 325 K.

$$\varphi^{350K} < \varphi^{275K} < \varphi^{325K} < \varphi^{300K}$$

Phe molecules indicate the most cylinder-like shape at 350 K. This temperature is followed by 325 K, 300 K, and 275 K, respectively:

$$\varphi^{350K} < \varphi^{325K} < \varphi^{300K} < \varphi^{275K}$$

At temperatures 300 K and 325 K, the aggregation process appears to be fast, since the average  $\varphi$  does not change a significant amount after 100 ns. Furthermore, by looking at Table 7, we can easily deduce that the  $R$  of Phe molecules are clearly lower than the other two; hence, they aggregate weaker among all amino acids.

Trp molecules have the most cylinder-like shape at 350 K, which is followed by temperatures 325 K, 300 K, and 275 K. It is important to note that at 275 K, the average in the first 100 ns does not improve in the last 50 ns. Considering that Trp molecules aggregate to a sphere-like structure at this temperature, this value is in agreement with our sphericity measurement and thus can be given as the reason for the slight decrease and then increase between 200 and 300 ns. From Table 8, we deduce:

$$\varphi^{350K} < \varphi^{325K} < \varphi^{300K} < \varphi^{275K}$$

Consequently, all 3 amino acids reach their lowest cylindricity average between 250 and 300 ns at 350 K. At this temperature, Tyr reaches the most cylinder-like formation among all 3 amino acids, which is also on par with our simulation results:

$$\varphi^{Tyr,350K} < \varphi^{Trp,350K} < \varphi^{Phe,350K}$$

The time evolution of  $\varphi$  at 350 K and its moving average is shown in Figure 6. We observe that the system reaches its equilibrium structure around 130 ns, which is in full agreement with the result of [10].

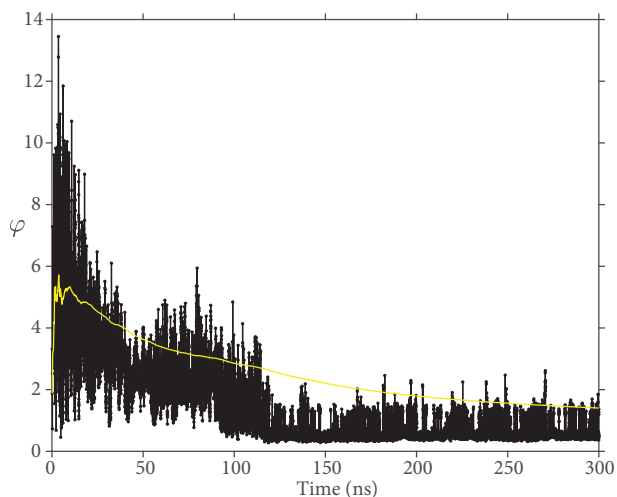
### 3.3. Accuracy and performance

The detection results of the algorithm have been compared with ground truth, which is the process of tagging the result of the algorithm for our data set and it is depicted in Table 10, where we investigate the number of molecules out of the major cluster, namely the outliers, and or member of nonsignificant clusters. The outputs in the table are in a perfect agreement. Clustering using DBSCAN for each time frame lasts less than 0.1 s on average. The computations are conducted on a laptop with Intel i5-8250 and 8 GB RAM. The experimental results show that the accuracy and the performance of the model are high.

The results of the clustering is also in agreement with the results of [9, 10], but here we focus on measuring the shapes of the self-assembled structures quantitatively.

## 4. Conclusions and outlook

In recent years it has been established that many single amino acids self assemble into organized structures. The fact that these assemblies occur from isolated monomers present unique challenges over the more well-studied proteins which assemble from covalently linked amino acids. For example, within the field of protein



**Figure 6.** Evaluation of  $\varphi$  and its moving average (yellow line) for Tyr at  $T = 350$  K.

science, several decades of research has led to numerous and well-understood characterization methods for the proteins themselves, as well as for the computational simulations of proteins. The number and availability of tools to describe the single molecule amino acid assemblies and simulations thereof has not been as thoroughly established. The current approach provides a novel tool to better characterize the simulation results of single amino acids.

The tool includes an algorithm using the DBSCAN algorithm to measure two specific structures coming from the molecular simulations in the field of biophysical science. Insights into these structures are of high interest in biotechnology and health sciences. The described approach in the paper shows that the DBSCAN algorithm can be successfully applied in the computational characterization of the structures of self-assemblies. A measure on cylindricity and sphericity might be used as a gadget to discriminate various important structures.

The code of the DBSCAN algorithm and the measures used on the data set are from our previous molecular dynamic simulations [10] to identify the clusters in the self-assembled structures formed by aromatic amino acids. Then, the cylindrical and spherical structures of the clusters are measured analytically and also depicted graphically. The quantitative results obtained from these measures were in agreement with those of previous qualitative studies.

According to the cylindricity measurements, we observe that all 3 aromatic amino acids, namely Phe, Tyr, and Trp, have a higher degree of fibril-like aggregations at high temperatures, the reasons of which were discussed in [10]. Especially formations of Tyr at 350 K are remarkably ordered and stable. Sphericity measurements for Trp molecules showed that it has the highest spherical aggregation among all studied amino acids. On the other hand, Phe has the lowest aggregation ratio  $R$  compared to the other two amino acids, meaning that it has a relatively less stable or disorganized structure. However, despite having smaller major clusters, Phe exhibits a fairly low value for cylindricity, albeit not the lowest. As described in the previous section, the cylindricity parameter is lower for a more perfect cylinder. Thus, it would not be unfair to claim that Phe also achieves a tubular shape, as temperature increases.

It is also important to recall that this work provides the measurement of the cylindrical and spherical shapes formed by aromatic amino acids under certain conditions after a certain period. These conditions include the temperature of the medium and the density of the distribution of the amino acid. If we observe the dynamics

**Table 10.** Ground truth and detection. The red dots are either the monomers out of the major cluster (outliers) or nonsignificant clusters.

Sample	Time (ns)	Ground truth	Detection	
			Number	Visual
Trp, 275 K	150	3	3	
Trp, 300 K	220	1	1	
Tyr, 325 K	140	5-7	7	
Tyr, 350 K	230	2	2	
Phe, 325 K	275	5	5	
Phe, 350 K	284	9	9	

of amino acids at a relatively early stage or change the temperature and density conditions, then the shapes created by amino acids might be in nonspherical or noncylindrical forms, or even self-aggregation might not occur at all. However, under the conditions mentioned in [10], the quantitative analyses in this study completely match our observations from previous simulations [10] and thus provide us an analytical justification for our naked-eye analysis.

In further work, one may develop this analyzing tool combining all measures within a deep learning algorithm, which may estimate more aggregations types, such as planes, fibrils, and scattered structures, in an

automatic way and give the relevant measures. A visual tool to track specific monomers or groups of monomers may lead to the insight of the self-assembly mechanism. The importance of the self-assembly processes of pure and mixed aromatic amino acids and similar units might become more clear with the use of such improved tools.

### Acknowledgments

This work was supported by the scientific research project 2016BF0018 of Turkish-German University. The simulation data used for this work was fully performed at TÜBİTAK ULAKBİM, High Performance and Grid Computing Center (TRUBA resources).

### Conflict of interest

Helen W. Hernandez is a managing member of KAL Research Initiatives, LLC.

### References

- [1] Goldfarb D. Biophysics Demystified. USA: McGraw-Hill Professional, 2010.
- [2] Perween S, Chandanshive B, Kotamarthi HC, Khushalani D. Single amino acid based self-assembled structure. *Soft Matter* 2013; 9: 10141-10145.
- [3] M'enard-Moyon C, Venkatesh V, Krishna KV, Bonachera F, Verma S et al. Self-assembly of tyrosine into controlled supramolecular nanostructures. *Chemistry A European Journal* 2015; 21: 11681-11686. doi: 10.1002/chem.201502076
- [4] Shaham-Niv S, Adler-abramovich L, Schnaider L, Gazit E. Extension of the generic amyloid hypothesis to nonproteinaceous metabolite assemblies. *Science Advance* 2015; 1: 1-7. doi: 10.1126/sciadv.1500137
- [5] Chakraborty P, Gazit E. Amino acid based self-assembled nanostructures: Complex structures from remarkably simple building blocks. *Chemistry of Nanomaterials for Energy. Biology and More* 2018; 4 (8): 730-740. doi: 10.1002/cnma.201800147
- [6] El-Metwally A, Al-Ahaidib LY, Sunqurah AA, Al-Surimi K et al. The prevalence of phenylketonuria in arab countries, Turkey, and Iran: A systematic review. *BioMed Research International* 2018; 1-12. doi: 10.1155/2018/7697210
- [7] Ford S, O'Driscoll M, MacDonald A. Living with phenylketonuria: Lessons from the PKU community. *Molecular Genetics and Metabolism Reports* 2018; 17:57-63. doi: 10.1016/j.ymgmr.2018.10.002
- [8] Gazit E. Metabolite amyloids: a new paradigm for inborn error of metabolism disorders. *Journal of Inherited Metabolic Disease* 2016; 39: 483-488. doi: 10.1007/s10545-016-9946-9
- [9] German HW, Uyaver S, Hansmann UHE. Self assembly of phenylalanine-based molecules. *The Journal of Physical Chemistry A* 2014; 119 (9): 1609-1615. doi: 10.1021/jp5077388
- [10] Uyaver S, Hernandez HW, Habiboglu MG. Self-assembly of aromatic amino acids: a molecular dynamics study. *Physical Chemistry Chemical Physics* 2018; 20 (48): 30525-30536. doi: 10.1039/C8CP06239K
- [11] German HW, Bharaju M, Uyaver S, Hansmann UHE. Computational Insights into the self-assembly of phenylalanine-based molecules. *Task Quarterly* 2014; 18 (4): 357-363.
- [12] Habiboglu MG, Uyaver S. Shape measurement for cylindrical structures formed by tyrosine molecules. In: *AIP Conference Proceedings* 2183; USA; 2019. 080003:1-4. doi: 10.1063/1.5136196.
- [13] Hartigan JA. Clustering. *Annual Review of Biophysics and Bioengineering* 1973; 2: 81-102. doi: 10.1146/annurev.bb.02.060173.000501
- [14] Leake MC. *Biophysics Tools and Techniques*, USA: CRC Press, 2016.



- [15] Yeole SD, Gadre S. Molecular cluster building algorithm: Electrostatic guidelines and molecular tailoring approach. *The Journal of Chemical Physics* 2011; 134 (8): 084111. doi: 10.1063/1.3556819
- [16] Li Z, D'Ambro EL, Schobesberger S, Gaston CJ, Lopez-Hilfiker FD et al. A robust clustering algorithm for analysis of composition-dependent organic aerosol thermal desorption measurements. *Atmospheric Chemistry and Physics Discussions* 2019; 20: 2489-2512. doi: 10.5194/acp-2019-733
- [17] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; Menlo Park, CA, USA; 1996. pp. 226-231.
- [18] Lashkov AA, Rubinsky SV, Eistrikh-Heller PA. Application of the DBSCAN algorithm to detect hydrophobic clusters in protein structures. *Crystallography Reports* 2019; 64 (3): 494-502. doi: 10.1134/S1063774519030179
- [19] Xu R, Wunsch II D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 2005; 16 (3): 645-678.
- [20] Ahmad PH, Dand S. Performance evaluation of clustering algorithm using different datasets. *International Journal of Advance Research in Computer Science and Management Studies* 2015; 3 (1): 167-173.
- [21] Abdolzadegan D, Moattar MH, Ghoshuni M. A robust method for early diagnosis of autism spectrum disorder from EEG signals based on feature selection and DBSCAN method. *Biocybernetics and Biomedical Engineering* 2020; 40: 482-493. doi: 10.1016/j.bbe.2020.01.008
- [22] Bilal M, Habib HA, Mehmood Z, Saba T, Rashid M. Single and multiple copy-move forgery detection and localization in digital images based on the sparsely encoded distinctive features and DBSCAN clustering. *Arabian Journal for Science and Engineering* 2020; 45: 2975-2992. doi: 10.1007/s13369-019-04238-2
- [23] Zhang A, Gao F, Yang M, Bi W. A new rule reduction and training method for extended belief rule base based on DBSCAN algorithm. *International Journal of Approximate Reasoning* 2020; 119: 20-39, doi: 10.1016/j.ijar.2019.12.016
- [24] Lu G, Liu H. An effective interference suppression algorithm for visible light communication system based on DBSCAN. *Chinese Optics Letters* 2020; 18 (1): 1-6.
- [25] Li Z, Lu W, Huang J. Crowdsourcing logistics pricing optimization model based on DBSCAN clustering algorithm. *IEEE Access* 2020; 8: 92615 - 92626. doi: 10.1109/ACCESS.2020.2995063
- [26] Li S, Qin N, Huang D, Ke L. Damage localization of stacker's track based on EEMD-EMD and DBSCAN cluster algorithms. *IEEE Transactions on Instrumentation and Measurement* 2020; 69 (5): 1981-1992.
- [27] Chen H, Yu G, Liu F, Cai Z, Liu A et al. Unsupervised anomaly detection via DBSCAN for KPIs jitters in network managements. *Computers, Material & Continua* 2020; 62 (2): 917-927. doi: 10.32604/cmc.2020.05981.
- [28] Zhang P, Wang Y, Liang L, Li X, Duan Q. Short-term wind power prediction using GA-BP neural network based on DBSCAN algorithm outlier identification. *Processes* 2020; 8 (157): 1-15. doi: 10.3390/pr8020157.
- [29] Starczeski A, Goetzen P, Er J. A new method for automatic determining of the DBSCAN parameters. *Journal of Artificial Intelligence and Soft Computing Research* 2020; 10 (3): 209-221. doi: 10.2478/jaiscr-2020-0014.
- [30] Scitovski R, Sabo K. DBSCAN-like clustering method for various data densities. *Pattern Analysis and Applications* 2020; (2): 1-20. doi: 10.1007/s10044-019-00809-z
- [31] Scitovski R, Sabo K. A combination of k-means and DBSCAN algorithm for solving the multiple generalized circle detection problem. *Advances in Data Analysis and Classification* 2020. doi: 10.1007/s11634-020-00385-9
- [32] Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems* 2017; 42 (3): 1-21. doi: 10.1145/3068335
- [33] Ahmad S, Sarai A. Qgrid: clustering tool for detecting charged and hydrophobic regions in proteins. *Nucleic Acids Research* 2004; 32 (2): W104-W107. doi: 10.1093/nar/gkh363
- [34] Wadell H. Sphericity and roundness of rock particles. *The Journal of Geology* 1933; 41 (3): 310-331.

- [35] Bribiesca E. A measure of compactness for 3D shapes. *Computers and Mathematics with Applications* 2000; 40: 1275-1284.
- [36] Wadell H. Volume, shape, and roundness of quartz particles. *The Journal of Geology* 1935; 43 (3): 250-280.
- [37] Tsudaka T, Kanada T. Minimum zone evaluation of cylindricity deviation by some optimization techniques. *Bulletin of the Japan Society of Precision Engineering* 1985; 19 (1): 18-23.
- [38] Lei X, Song H, Xue Y, Li J, Zhou J et al. Method for cylindricity error evaluation using geometry optimization searching algorithm. *Measurement* 2011; 44 (9): 1556- 1563. doi: 10.1016/j.measurement.2011.06.010
- [39] Kanad T. Suzuki S. Evaluation of minimum zone flatness by means of nonlinear optimization techniques and Its verification. *Precision Engineering* 1993; 15: 93-99. doi: 10.1016/0141-6359(93)90343-9
- [40] Zhang K. Research on coaxiality errors evaluation based on ant colony optimization algorithm. In: Fei M, Irwin G, Ma S (editors). *Bio-Inspired Computational Intelligence and Applications*, USA: Springer, 2007, ISBN: 978-3-54074769-7.
- [41] Murthy TSR. A comparison of different algorithms for cylindricity evaluation. *International Journal of Machine Tool Design and Research* 1982; 22 (4): 283-292. doi: 10.1016/0020-7357(82)90006-3
- [42] Lao YZ, Leong HW, Preparata FP, Singh G. Accurate cylindricity evaluation with axis-estimation preprocessing. *Precision Engineering* 2003; 27: 429-437. doi: 10.1016/S0141-6359(03)00044-8
- [43] Dawson DJW. Cylindricity and its measurement. *International Journal of Machine Tools and Manufacture* 1992; 32(1/2): 247-253. doi: 10.1016/0890-6955(92)90085-U