Recognizing Molecular Structural Features by Pattern Recognition Techniques

Qing Lu (qinglu@iccas.ac.cn)  
ICCAS: Institute of Chemistry Chinese Academy of Sciences  
https://orcid.org/0000-0003-4063-9100

Wensheng Bian  
ICCAS: Institute of Chemistry Chinese Academy of Sciences  
https://orcid.org/0000-0003-2125-8054

Methodology

Keywords: Pattern recognition, cluster analysis, convex hull, molecular structural recognition

Posted Date: September 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-860064/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Recognition of molecular structural features is one of the most attractive fields in chemistry, especially when combining with machine learning techniques. Pattern recognition techniques are straightforward in recognizing graphic features, but little attention was given to recognize molecular structural features. In this work, we propose a new method taking advantage of pattern recognition techniques to analyze structural features and obtain novel chemical insights. Specifically, the cluster analysis is presented to recognize structural features, which provides an alternative to the most widely used root mean square deviation (RMSD) method and the recently proposed blob detection method. Based on this, the convex hull of the molecule is constructed. The convex hull of molecules is highly appealing in the sense that one can introduce established theorems and properties from other disciplines into chemistry. Novel molecular descriptors based on convex hulls can be defined and show encouraging results, especially in providing new insights in understanding non-covalent interactions, adsorption processes, etc.

Introduction

Machine learning techniques have prevailed across many disciplines in recent years. In chemistry, it has exhibited great strengths in different fields, such as conformation exploration, catalysis design, reaction optimization, etc. Given the fast development of machine learning techniques and the increasingly complex molecular systems, one would expect machine learning techniques would become more critical in understanding chemical behaviors.

For successful supervised or unsupervised learning, a large amount of input data is critical. Accordingly, comparing and categorizing different samples are of paramount importance to avoid redundancy or bias during the learning process. In chemistry, this identification process could refer to differentiating molecular structures, such as comparing atomic coordinates between theoretical and experimental structures. Such a comparison is one of the most fundamental applications in computational chemistry, as it is often the starting point for various sophisticated computational studies. In addition, there are studies to combine existing benchmark sets by generating a larger benchmark set. The construction of such a super benchmark set needs the attention of removing duplicated samples from individual sets. Accordingly, it is necessary to recognize unique molecules for building a non-redundant super set.

To examine structural similarities, the root-mean-square-deviation (RMSD) calculation is probably the most commonly used method. It calculates the square sum of distances between corresponding atoms \( d_i \) in the two structures, and takes the division by the total number of atoms (N), followed by a square root operation.

\[
RMSD = \sqrt{\frac{\sum d_i^2}{N}}
\]
However, the same molecule in different benchmark sets may have totally different XYZ coordinates, though one may translate and rotate the molecule to align molecular orientation. Despite this, the RMSD measurement also suffers other limitations such as lack of normalization, being difficult for interpretation, and diminishing ability to distinguish conformers with increasing system size(10-12). Some improvements upon RMSD have been proposed to remedy these problems, such as introducing weighting functions into the calculation of RMSD(12), or taking advantage of the graph theory(13) or symmetry(14). Other alternatives include configuration fingerprint vector(15), global and local descriptors(16), geometric hashing algorithm(17), and several different score functions(10, 18-23).

On another aspect, pattern recognition techniques have received significant successes in recent years. Notably, there are mathematically proved theorems, which can be brought in chemistry for structural analysis. However, very few studies were carried out in this respect. Previously, the blob detection technique was used to recognize molecules(24). Although it achieves considerable success, there are still some limitations remaining. Firstly, during the blob detection, the graphic color was converted to grayscale to boost efficiency. However, such a trick sacrifices the ability to differentiate isotopes or elements in the same family (although the original blob detection study was designed for conformation analysis). Secondly, the blob detection uses a Gaussian function-based kernel for the convolution calculation. Yet this noise-filter step is not necessary as long as the image is not transformed.

In this work, we propose a new method to recognize molecular structural features taking advantage of pattern recognition techniques. The blob detection was circumvented by applying cluster analysis to the image matrices, which successfully detected all atomic positions. The new method is fast and accurate for molecular structural comparisons. Based on this, the convex hull, which sets up a polyhedron to enclose the molecule, is constructed. Accordingly, the established theorems and properties of convex hulls from other disciplines can be introduced to chemistry to analyze structural features. Specifically, by creating the convex hulls, the molecular volume and surface area can be defined. One can therefore explore new chemistry with these new molecular descriptors. A few applications are bought up to exhibit some applications, which show promising results in providing novel chemical understandings.

**Methods**

In this work, the proposed method consists of three steps to recognize structural features: pre-treatment of the molecules and images, feature extraction, and post-treatment with convex hull constructions, as shown in Fig. 1.

### 1. Pre-treatment of molecules and images

To remove redundancy, the chemical bonds are eliminated from molecular structure images, since the chemical bonds are determined solely by the atom positions. As a result, the problem of recognizing
molecules is equivalent to identify a set of scattered atoms/dots.

The molecular image is the basis for pattern recognition. Providing the atoms have been well aligned(25), the molecular 3D image is generated from its XYZ coordinates (Fig. 2), taking the C60 system as an example. The 60 carbon atoms are scattered after removing all chemical bonds. For easier visualization and treatment in the latter stage, the azimuthal angle and the elevation angle were set as 90 degrees for exhibiting the image (along Z axis). Unless otherwise stated, the following discussion is based on this projection angle.

For complicated molecules, it is increasingly difficult to find a projection angle that all atoms can be projected on a plane without overlaps. One may recognize the molecular features by analyzing its facet of profile. However, the core structure cannot be recognized by this way. To circumvent this problem, we sliced the whole molecule into layers, and took snapshots for each layer to extract features (Fig. 2). Nonetheless, it is not a trivial work to slice the molecule, as the double counting of atoms may take place. Eventually, we sliced the molecule along the projection angle, and set the distance between layers to be 0.7 angstroms. This value is close to a H-H bond distance. For any reasonably determined structure, it is not possible to have two atoms with a distance smaller than 0.7 angstroms. Therefore, the layers separated by 0.7 angstroms can well slice the whole molecule into different layers.

As any other pattern recognition applications, the quality of the picture is essential. Following the parameters given by the blob detection study, we set the picture height and width of 10 * 10 inches with 80 dots per inch (dpi). Accordingly, the final resolution of the figure is 800 * 800 pixels.

2. Feature extraction

To recognize the atoms on each layer, we took advantage of cluster analysis to filter the image matrices. The image matrices are non-diagonal sparse matrices, with dimensions equal to the resolution. In the previous blob detection study, the colored image was first converted to gray scale. Consequently, if atoms were assigned with close color codes, the grayscale conversion would mistakenly consider the different atoms as identical ones.

In this work, the colored image matrix was first separated into three primary-color matrices, namely the R(ed) matrix, G(reen) matrix and B(lue) matrix. As a result, if an atom in the molecule is substituted by its isotope or its family member, the color matrices can reveal its trace. And the atoms with close color codes can be distinguished.

By converting a graph into an image matrix, an atom in the graph is represented by a group of pixel coordinates. Ideally, the atom size determines the number of pixel coordinates. However, such a number is not unambiguously determined, as the boundary of an atom may be blurred especially if the resolution is low. The number of pixel coordinates is thus subject to the round-off error. As a consequence, it is generally not helpful to directly compare the image matrices.
Instead, the K-means algorithm (26) of cluster analysis was used in this work. For a given primary-color matrix, the local extreme values were first filtered out. The local extremes represent pixel occupation for each atom in the image matrix. Whether the local extreme is a maximum or a minimum depends on the background color being black or white. Supposing the background color is black, the non-zero elements were first filtered out as the basis for cluster analysis.

For the K-means algorithm, it clusters the matrix elements by relocating each point to its new nearest center. In the context of feature extraction, this corresponds to determine the center of each set of pixel coordinates. The metric mean of the member points to corresponding cluster centers was calculated, and such relocating-and-updating process iterated until the desired number of cluster centers was found, which was the number of atoms in each layer.

By clustering the centers, the arrays containing each center position were obtained. The Euclidean norm between centers of two structures was compared. If the norm differed by more than 5 pixels, the corresponding atoms were considered as occupying different locations.

To find out which atoms differ in the two structures, a register table was first established to map 2D pixel coordinates and 3D atomic coordinates. The table was constructed by mapping atomic coordinates and pixel coordinates atom by atom. Next, the cluster analysis was carried out for the second structure. By differentiating the cluster centers out of two structures, the atoms at different positions can be filtered out by mapping with the register table.

3. Post-treatments of constructing convex hulls

The convex hull is the smallest polyhedron that encloses a set of points, where intersections between any points in the polyhedron are still in the polyhedron. Originally, the concept of convex hulls was used in other disciplines such as computational geometry, functional analysis, image processing, etc. It depicts a set of n-dimensional (usually 2-dimensional) data, with many mathematically proved theorems or properties such as the separating hyperplane theorem. Such theorems are very appealing in the context of molecule recognition that if properly used, one may readily know the molecular properties without complicated calculations. Therefore, we are particularly interested in studying the convex hull for molecules, as established theorems and properties of convex hulls can be borrowed from other disciplines to study molecular interactions.

To construct the 3-dimensional convex hull for a molecule, the QHull algorithm was used (27). A polyhedron enclosed the molecule was generated based on atomic coordinates. The molecular surface area was calculated as the total surface area of all facets of the convex hull. The specific molecular area was calculated as the molecular surface area over the molecular mole mass. Similarly, the molecular density was calculated as the molecular mole mass over the total volume of the convex hull.
To recognize molecular geometric features and compare their structures, the C60 molecule and the manually distorted C60 were examined. This comparison resembles the comparison between structures in different databases or between theoretical and experimental structures. The distorted molecule was generated by adding a random displacement between −0.5 to 0.5 Å to the XYZ coordinates of the first 5 carbon atoms. A constraint is given to the random numbers that the displacements (d_x, d_y) should be larger than 0.2 Å. Otherwise, the geometric difference might not be recognized. The d_z displacement is left out from the constraint since by convention the projection angle is along the Z direction. Figure 3a shows the graph of the undistorted C60 molecule, where the first 5 carbon are marked as blue squares. The remaining carbon atoms are plotted as gray circles. As a comparison, Fig. 3b shows the distorted C60 molecule, where the distorted carbon atoms are highlighted with red color. Figure 3c shows the overlap of the two structures. The atoms at different positions are contrasted from the figure, and the atoms at same positions are overlapped. Since the random number is involved to generate the distorted molecule, 100 trials have been carried out for the recognition. The successful hit reaches 100%.

Having extracted graphic features, we construct convex hulls for molecules. The convex hull is the smallest polyhedron that encloses the molecule. Fig. 4 shows some examples of molecules with their convex hulls. For high-symmetry molecules, such as SF_6 (Oh point group) in Fig. 4a, its convex hull is an octahedron. The 6 fluorine atoms locate on the vertices of the octahedron, while the sulfur atom sits in the center. By definition, all atoms are enclosed in the octahedron. And connections between two atoms are still in the octahedron. Fig. 4b and 4c show two other examples with more complicated geometric features and their convex hulls.

The convex hulls have been widely used in other disciplines. In chemistry, the probably easiest way of taking advantage of convex hulls is to define the molecular density and the specific surface area. For molecular density, it is calculated as the molecular mole mass over the volume of the convex hull. Although one can also calculate the density by dividing the mole mass over volume of a cubic cell, this cubic cell volume cannot reflect the shape of the molecule (cf. the SF_6 instance). And the volume of the cubic cell would be always larger than that of the convex hull, since the convex hull by definition is the smallest polyhedron enclosing the molecule. Such a difference may lead a significant improvement in the data training, as the molecular density obtained based on convex hulls might be a better molecule descriptor. Fig. 5 shows the molecular density and corresponding convex hulls for different sizes of fullerenes. It is evident that the molecular density decreases as the sphere size increases.

The calculation of surface area is another possible application regarding convex hulls. The specific surface area is an important parameter in studying adsorption processes. Fig. 6 shows the specific surface area for different types of fullerenes. It can be seen that the specific area variates less than the molecular density. If we approximate that the inner surface is equal to the outer surface of the polyhedron, the method can be further used to study adsorption processes of zeolites or nanotubes.

Lastly, the Temozolomide (TMZ)-C60 system was exhibited as a preliminary application to study non-covalent interactions by analyzing convex hulls (Fig. 7). The TMZ-C60 system was theoretically studied...
as a brain anticancer drug. The fullerene loads TMZ and transports the drug across the blood brain barriers. Obviously, the drug adsorbs the molecule by non-covalent interactions. And such interactions are subjected to the contact area. However, it is not quantitatively known about the relationship between the contact area and interaction strength. Therefore, it would be beneficial to study such correlation for better design of drug delivery. Further study is under process in this lab.

**Conclusions**

In this work, pattern recognition techniques are developed for molecular structure recognition. The method provides a new approach to recognize molecular geometrical features, and thus can be used for structural identifications. The cluster analysis of K-Means algorithm was used to determine the pixel centers. This is more straightforward than the previous blob detection technique in the sense that the convolution calculation is saved. A new post-treatment is proposed to construct convex hulls of molecules. Accordingly, the properties of convex hulls can be borrowed into chemistry and provide novel insight. To illustrate some possible applications, the molecular specific surface area and density were calculated based on the total surface and volume of convex hulls for different sizes of fullerenes. It shows that such properties are promising to be used as new molecular descriptors in machine learning studies, and it provides a new dimension to understand molecular interactions. Further study is under development in this lab.

**Abbreviations**

RMSD, root mean square deviation

TMZ, Temozolomide

**Declarations**

**Availability of data and materials**

All data and source code are freely available by the request from the authors.

**Competing interests**

The authors declare no competing interests.

**Funding**

The author gratefully acknowledges the support from the National Natural Science Foundation of China (Nos. 22003068), the Beijing Municipal Natural Science Foundation (Nos. 2214065).
Authors’ contributions

Not applicable

Acknowledgements

Not applicable

References

1. Mater AC, Coote ML (2019) Deep Learning in Chemistry. J Chem Inf Model 59(6):2545–2559
2. Zhang M, Wu H, Yang J, Huang G (2021) A Computational Mechanistic Analysis of Iridium-Catalyzed C(sp3)–H Borylation Reveals a One-Stone–Two-Birds Strategy to Enhance Catalytic Activity. ACS Catalysis 11(8):4833–4847
3. Lu Q, Neese F, Bistoni G (2019) London dispersion effects in the coordination and activation of alkanes in sigma-complexes: a local energy decomposition study. Phys Chem Chem Phys 21:11569–11577
4. Lu Q, Neese F, Bistoni G (2018) Formation of Agostic Structures Driven by London Dispersion. Angew Chem Int Ed Engl 57(17):4760–4764
5. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences. 2020;117(3):1496 – 503
6. Kleywegt GJ. Recognition of spatial motifs in protein structures11Edited by J. Thornton. Journal of Molecular Biology. 1999;285(4):1887-97
7. Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19(13):1644–1649
8. Sylvetsky N, Kesharwani MK, Martin JML. MP2-F12 basis set convergence for the S66 noncovalent interactions benchmark: Transferability of the complementary auxiliary basis set (CABS). AIP Conference Proceedings. 2017;1906(1):030006
9. Schneebeli ST, Bochevarov AD, Friesner RA (2011) Parameterization of a B3LYP specific correction for non-covalent interactions and basis set superposition error on a gigantic dataset of CCSD(T) quality non-covalent interaction energies. J Chem Theory Comput 7(3):658–668
10. Baber JC, Thompson DC, Cross JB, Humblet C (2009) GARD: A Generally Applicable Replacement for RMSD. J Chem Inf Model 49(8):1889–1900
11. Hawkins PCD (2017) Conformation Generation: The State of the Art. J Chem Inf Model 57(8):1747–1756
12. Wagner A, Himmel H-J, aRMSD (2017) A Comprehensive Tool for Structural Analysis. J Chem Inf Model 57(2):428–438
13. Helmich B, Sierka M (2012) Similarity recognition of molecular structures by optimal atomic matching and rotational superposition. J Comput Chem 33(2):134–140

14. Allen WJ, Rizzo RC (2014) Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-Based Design. J Chem Inf Model 54(2):518–529

15. Sadeghi A, Ghasemi SA, Schaefer B, Mohr S, Lill MA, Goedecker S (2013) Metrics for measuring distances in configuration spaces. J Chem Phys 139(18):184118

16. Ramirez-Manzanares A, Peña J, Azpiroz JM, Merino G (2015) A hierarchical algorithm for molecular similarity (H-FORMS). J Comput Chem 36(19):1456–1466

17. Wallace AC, Borkakoti N, Thornton JM. Tess (1997) A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 6(11):2308–2323

18. Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins: Struct Funct Bioinf 57(4):702–710

19. Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

20. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A (2001) A study of quality measures for protein threading models. BMC Bioinformatics 2(1):5

21. Rychlewski L, Fischer D, Elofsson A (2003) LiveBench-6: Large-scale automated evaluation of protein structure prediction servers. Proteins: Struct Funct Bioinf 53(S6):542–547

22. Zemla A, Venclovas Č, Moul J, Fidelis K (1999) Processing and analysis of CASP3 protein structure predictions. Proteins: Struct Funct Bioinf 37(S3):22–29

23. Siew N, Elofsson A, Rycklewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16(9):776–785

24. Lu Q, Molecular structure recognition by blob detection, in review

25. Temelso B, Mabey JM, Kubota T, Appiah-Padi N, Shields GC (2017) ArbAlign: A Tool for Optimal Alignment of Arbitrarily Ordered Isomers Using the Kuhn–Munkres Algorithm. J Chem Inf Model 57(5):1045–1054

26. Jin X, Han J (2010) K-Means Clustering. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer US, Boston, pp 563–564

27. Barber CB, Dobkin DP, Huhdanpaa H (1996) The quickhull algorithm for convex hulls. ACM Trans Math Softw 22(4):469–483

28. Samanta PN, Das KK (2017) Noncovalent interaction assisted fullerene for the transportation of some brain anticancer drugs: A theoretical study. J Mol Graph Model 72:187–200

Figures
Figure 1

General flowchart of molecule structural feature recognition.
Figure 2

Illustration of the molecule pre-treatment.

Figure 3

a) The bird view of a normal C60 molecule. The blue atoms are referenced for comparison; b) The bird view of a distorted C60 molecule. The red atoms are displaced by a random number; C) The overlap of distorted and non-distorted C60 molecule.

Figure 4

Examples of convex hulls constructed for a) SF6 molecule; b) AcOH dimer system; c) L-lysine;
Figure 5

The molecular density and convex hulls for different type of fullerenes.
Figure 6

The molecular specific surface area for different type of fullerenes.

Figure 7

Illustration of interaction area defined by convex hulls. a) The chemical structure of Temozolomide (TMZ); b) The molecular structure of TMZ; c) The convex hull of TMZ interacting with C60; d) The top
view of the TMZ convex hull; e) The side view of TMZ convex hull.