Characterization of molecular belts' shape in the context of porous molecular materials

Ismael Gómez García, Marta Chaves-Salado and Maciej Haranczyk

AUTHOR ADDRESS

IMDEA Materials, Tecnogetafe, C/ Eric Kandel, 2, 28906, Getafe, Madrid

KEYWORDS

Porous molecular materials, molecular belts, molecular shape, material discovery.

ABSTRACT

Porous molecular belts are a common type of porous molecules that can be assembled into nanotube porous crystals for various applications. The inherited nature of crystal porosity allows exploiting molecular properties in order to fine-tune the nanotube crystals for specific applications. However, molecular features determining nanotube formation remain unclear. Molecular shape has been suggested as a potential aspect determining packing at crystal level, but this hypothesis has not yet been tested. In this work, we define the first set of methods to characterize the shape of molecular belts, demonstrating their application to discover nanotube crystals by screening large datasets. Moreover, we introduce and characterize (in terms of porosity, shape and chemistry) the largest available repository of molecular belts, with 4412 molecules mined from the PubChem dataset. With this study, we show that molecular shape can

play a major role in solid-phase assembly of porous molecular materials, opening new avenues in molecular characterization and material discovery.

INTRODUCTION

Molecular belts are defined as macrocycles with permanent cavities that can be accessed from two windows. These molecules have been of interest due to their capacity to capture chemical species within the opening of their structures. Crown ethers are a prime example of such species, and have been investigated, designed and used to capture ions from the solution for decades.[1,2] Interestingly, there are new application opportunities for molecular belts in the context of porous molecular materials, in which molecules with permanent internal voids assemble into solids with exploitable porosity in applications such as molecular separations, storage, ion transport and sensing. Typically, porous organic cage materials, i.e. based on molecules with internal cavity accessible by at least 3 windows, have been considered in this context, [3,4] whereas molecular belts have been less explored. There have been reports of their window-to-window assemblies forming nanotube-like solid structures. Numerous experimental examples of such structures exist in the literature.[5–8] For instance, Sakamoto et. al. assembled molecular belts formed by benzene rings into crystals with high adsorption capacity[7]. In another work, Kameta et. al. introduced artificial chaperones based on hydrogel nanotubes. [8]

The ability of molecular belts to assemble into tubular structures relies on weak interactions, and sometimes on the presence of solvent.[9–11] In a recent work, we discussed a set of three known crystals that showed nanotube-like crystal structures.[12] All three presented noria-like shapes (i.e. molecular belts with high symmetry and three wide corners) and showed nanotube structure in presence of solvent. In that study, we proposed a modified version of these molecules that retained the noria-like shape forming a nanotube crystal structure in the absence of solvent. This study

illustrated how molecular shape affects the macromolecular arrangements with exploitable porosity.

Thus far, the characterization of molecules with voids has been mostly focused on the internal cavity due to its importance in the context of porous solids. Previous works in this direction include efforts from Miklitz et. al., that characterized molecular pore and windows for cage molecules by combining information about the center of mass of the molecule and a Monte Carlo approach to scan its vicinity to detect windows.[13] Another approach, taken by our group, incorporated the Voronoi tessellation to improve detection of void spaces within the molecular framework, and the definition of Pore Exposure Ratio, allowing for discrimination between porous and non-porous molecules.[14] These efforts focused on detection of cavity and characterization of size and its accessibility to guest molecules. In a recent article, Sturluson and coworkers introduced a method to characterize shape of cavities of molecular cages.[15] They utilized singular value decomposition to construct the latent space of the screenshots of molecular cavities, for a total of 74 cage molecules, obtaining a clean space where similar cage molecules appeared together. The rather small set of molecules used to train the space is the main limitation of their approach.

Molecular cavity in belts heavily determines their application. Thus, there have been numerous efforts to combine different chemistries in order to design it. For instance, Sung Kuk et. al. introduced a method to tune calix[4]arene cavities using different ion pairs.[16] In another work, McCann et. al. introduced computational methods to tune bis-phosphine oxide based molecular belts for extraction of lanthanides.[17]

In this article, we introduce a geometry-based analysis of the shape of molecular belts. Our approach generates a set of four descriptors that encode key aspects of the

molecules' shape. The goal is to produce molecular descriptions that allow characterization of large set of molecular belts such as those deposited in PubChem, a repository c.a. 94M molecules, being a representation of all molecules considered by chemists up to date. [18] In previous works, our group screened this dataset, with focus on porous organic molecules (including both cages, belts and other molecules such as bowls and molecules with inaccessible pores, e.g. fullerenes). From those studies, we extracted a set of porous molecules,[19] out of which 4412 are molecular belts. This is the largest available repository of molecular belts. Previous efforts to construct repositories porous molecules led to repositories containing, respectively, 41 porous molecules,[13] and 481 porous crystals,[20] out of which only 50 (~10%) were formed by porous molecules, and only 15 (~3%) were molecular belts. Thus, our repository increases by two orders of magnitude the size of similar existing repositories in current literature. We demonstrate our tools through the analysis of our repository of molecular belts, including: (1) a description the set of molecular belts in terms of their molecular shape, studying the distribution of the different descriptors introduced; (2) study the capacity of novel descriptors to identify categories of molecules within molecular belts; (3) observe the relationships between molecular porosity and molecular shape in molecular belts using classical statistics; and (4) analysis of the chemistry of the molecules deposited. We further, demonstrate the application of our descriptors via screening of molecular candidates with given shape properties, shared with noria-like molecules previously discussed (Bernabei et. al. [12]), to discover nanotube crystals in the Cambridge Structural Database (CSD).[21]

Materials and Methods

By observing multiple examples of molecular belts, we have identified typical geometrical features that both allow classifying the shapes and are natural to the way in which chemists view and describe these molecules.

1) Number of corners: This corresponds with the number of vertices of the geometry (e.g. a triangular molecule will have three corners).
2) Wideness of corners: Our observations revealed that these corners may not be vertex-like, but rather wide. We want to distinguish, for instance, noria-like molecules from triangular molecules.
3) Elongation: Some molecules present a more elongated shape (e.g. rectangular molecules) and we want to consider this feature.
4) Regularity: The shape of molecular belts is not always identifiable with a particular geometrical object geometry.

In the following sections, we introduce a set of descriptors that capture these features, and algorithms to compute them. To validate our algorithms, we analyze several molecular examples, comparing descriptor values with visual inspection.

Molecular shape analysis

The underlying geometry of molecular belts is 2-dimensional; thus, we can obtain information about the shape by analyzing a projection of their top view. Additionally, we can approximate the overall geometry of the molecules by constructing a minimal error ellipse w.r.t. the atoms' projections. This ellipse will leave all corners outside, facilitating their identification.

2D projection and minimal error ellipse. To obtain a projection that properly resembles the shape of the object, we rotate the molecule using its largest entry path (an edge connecting the cavity with the surroundings, while keeping maximum distance with atoms; it is computed with Molipor[19]) to determine the rotation. Both the molecule and the entry path are rotated together so that the entry path gets aligned with the Z-axis. This rotation produces a top-view of the molecule's atoms. These are then projected on the XY-plane, i.e. an atom with coordinates (x, y, z), is projected into (x, y). This set of 2D coordinates is used to compute the minimal error ellipse. This done with help of a numerically stable method based on direct least squares. [22] The minimal error ellipse is defined by an equation with of the following form:

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0 \quad (1)$$

External elliptical sharp regions (EESR). Using equation (1), we define the following function:

$$G(x, y) = Ax^2 + By^2 + Cxy + Dx + Ey + F \quad (2)$$

We want to identify the clusters of atoms external to the ellipse, each capturing a corner of the molecule. To do so, first, for any atom projection point p_a, it is classified as being inside the ellipse as follows:

$$p_a \subset \epsilon \Leftrightarrow \text{sign}(\text{G}(p_a)) = \text{sign}(\text{G}(\text{c})) \quad (3)$$

where c is the center of the ellipse. Now, we apply DBSCAN for all $p_a \not\subset \epsilon$ (i.e. for every external point). The distance function used is the polar angle, φ, with two points being clustered together if $\varphi_{diff} < \frac{\pi}{8}$, where φ_{diff} is the difference in polar angle between these two points (module 2π). The minimum number of points to

define a cluster is one, as a single atom can define a corner of a molecule. The number of clusters defines our first descriptor, the External Elliptical Sharp Regions (EESR).

$$EESR(M) = \#\{Clusters\} \quad (4)$$

The EESR captures the number of corners the molecule has, capturing an important aspect of its geometry. Thus, molecules with triangular shapes will have EESR=3, whereas squared and rectangular molecules will have EESR=4, and so on. The process of computing EESR is depicted in Figure 1.

Arc proportion. We introduce a new descriptor named arc proportion. This descriptor is a number between 0 and 1 that reflects how wide corners are (the closer to 1, the wider). Formally, it is the proportion of the 2π arc surrounding the center of the molecule covered by corners. To obtain the arc proportion, we first compute the arc covered by each corner, i.e. the arc difference between the leftmost (clockwise) and the rightmost (clockwise) points of the corner. Formally, this is defined as:

$$CA = \min(2\pi - \max(\varphi) + \min(\varphi), \min(\varphi) - \max(\varphi)) \quad (7)$$

The CA value is the difference between the maximum and minimum angles within the points of the cluster, corrected if the cluster is in the region of periodicity of polar coordinates. Let CA_i be the arc covered by the i-th cluster. Then, the arc proportion is then defined as:

$$AP = \frac{1}{2\pi \sum_i CA_i} \quad (8)$$

Ellipse ratio. This descriptor accounts for the elongation of the molecule. Let R_M be the major radius of the ellipse ϵ, and R_m be its minor radius. The ellipse ratio is computed as the ratio between the major and the minor radii of ϵ, i.e:

$$ER = \frac{R_M}{R_m} \quad (9)$$

Algorithm implementation and datasets

We compute descriptors of molecular porosity using Molipor[19]. In particular: (1) the molecular largest cavity diameter (mLCD), i.e. the largest sphere that can lie within the molecule; (2) the pore exposure ratio (PER), a number determining how exposed the cavity is (0 totally surrounded, 1 totally exposed); (3) the number of windows of the molecule; (4) the entry paths to the molecule (i.e. the minimal collision trajectories for molecular probes that enter the molecule); (5) the internal surface area, i.e. the surface of the atoms placed inside the molecule; and (6) the maximum window size, i.e. the diameter of the largest window of the molecule. We introduce a repository of 4412 molecular belts. These are extracted from our repository of 6020 porous molecules mined from the PubChem database. The default output from PubChem are the lowest energy conformers available, and these are the candidates deposited in our repository and used for descriptor calculation. Our repository of molecular belts consists of molecules selected via chemical filters (e.g. molecules with >50 atoms, organic, etc.) and further characterized molecular porosity using Molipor. We deposit all molecules presenting PER < 0.45, mLCD > 1.0 A and exactly 2 windows. Our molecules are presented as SDF files, alongside tabularized data files containing molecular descriptors. We implement a GitHub repository in which, with help of version control, older files are kept after changes in the records (due to, e.g. refinement of molecular structure via more accurate estimation or experimental measure, algorithm refinement, etc.).

We demonstrate the applications of the introduced tools by characterizing the repository of molecular belts, comparing with the information provided by porosity

descriptors. We apply both classical statistics and unsupervised learning techniques, namely principal component analysis (PCA)[23] and density-based cluster analysis (DBSCAN).[24]

All the statistical analyses done in this work are performed with R (version 3.6.1) and RStudio.[25] We implement an R library that computes the shape descriptors for molecular belts, based on output description provided by our tool Molipor.

Results & Discussion

Shape descriptors validation

We validate the proposed shape descriptors over a diverse set of molecular belts. In Figure 2, molecules with different shapes are shown, alongside their descriptors. From the Figure, it can be seen that triangular molecules (e.g. Fig. 2a) have EESR = 3, whereas squared and rectangular molecules (e.g. Fig. 2f, 2g) have EESR = 4, and so on. Thus, the EESR captures the number of corners of the molecules. Certain molecules present "wide corners" (e.g. Fig. 2b, 2d), placed outside the ellipse and clearly separated among each other. The arc proportion allows discrimination between molecules with sharp and wide corners (e.g. Fig 2a and 2b present molecules that are very different but have similar EESR and ellipse ratio). Finally, the ellipse ratio identifies molecules that are more elongated. For instance, Fig. 2f and 2g present, respectively, a rectangular and squared molecule, with equal EESR and similar arc proportion. The ellipse ratio successfully differentiates these molecules. In our inspection of the dataset, we identified some issues on the algorithms. Having a poor projection of the molecule on the XY plane due to an anomalous entry path is one of the main. These are rare cases that cannot be considered molecular belts (e.g.

molecules with 2 windows but presence of side groups that break the belt shape) and can be detected and filtered out by combining Molipor descriptors and information about the minimal error ellipse.

Shape description of molecular belts

We characterize the set of 4412 molecular belts and further examine their shape descriptors from a statistical perspective to gain an insight on the information they add (e.g. compared to porosity descriptors). Our results are summarized in both Fig. 3, and in Table 1. In Fig. 3a, a bar plot representing the number of molecules within each value of EESR is presented. It can be seen that molecules with 4 elliptical sharp regions predominate, followed by those with 3, 5 and 6 regions. In other words, molecules that are (approximately) squared/rectangular, triangular and pentagonal are more common in the dataset. The plot resembles a normal distribution, thus we run the Shapiro-Wilk test for normality, obtaining a p-value of 0.00059, which confirms this observation. In Table 1, we present the main statistics for the three continuous shape descriptors: arc proportion, ellipse ratio and regularity. The results inform about the distribution of the data: (1) 75% of the molecules have an ellipse ratio value under 1.29, reflecting that non-elongated shapes are much more common among molecular belts. (2) 75% of the molecules have an arc proportion under 0.38, indicating that molecules with "sharp" corners (as in Fig 2a) predominate over "wide" corners (as in Fig 2b) in the population. For reference, see the rectangular molecule presented in Fig 2f, which has an ellipse ratio of 1.91. (3) 25% of the molecules regularity over 0.79, 50% of them are over 0.66, and 75% of them are over 0.56. Typically, regularity values over 0.7 indicate regular shape and good prediction of EESR corners and arc proportion. Regularity values between 0.5 and 0.7 seem to correspond with small

variations in one of the corners, whereas regularities below 0.5 typically indicate that at least one corner that is very different (typically much wider) than the others, which may correspond with a molecule that has no clear underlying geometry. However, in some cases, especially when the corners are too close to the ellipse, small perturbations on the shape may lead two corners to get merged, lowering regularity even though visual inspection suggests a clear underlying geometry (more details in SI). Interestingly, in such cases, the EESR tends to show small variations (typically of only 1 point). Also, the arc proportion can increase, but not significantly, and the ellipse ratio remains equal. Overall, regularity serves as an indicator of either absence of underlying geometry, or of it being fragile (i.e. sensitivity to perturbations in molecular structure). Even in this case, the EESR remains close to intuition. These variations are not expected to heavily affect the distribution of EESR presented above, as all categories (except EESR=0) may loose and gain few candidates. Thus, despite the limitations of these methods (coming from the unsupervised design they rely on), they provide an insight on the molecular shapes existing on our repository and can be used to mine molecules with desired shapes from large datasets.

We analyzed the pairwise correlations among shape and porosity descriptors. We considered PER, LCD, ISA and number of entries, being the least self-correlated porosity descriptors. [19] The pairwise Spearman correlations are presented in Fig. 3b. The correlations between porosity and shape descriptors are small to non-existent, confirming that shape descriptors add information w.r.t. porosity descriptors. The largest shape-porosity correlation is the negative association between ellipse ratio and LCD (-0.3 Spearman correlation), indicating that elongation disfavors porosity. We found a weak positive association between LCD and regularity (0.24 Spearman correlation), suggesting that larger molecules tend to be more regular. Regarding the

correlations among shape descriptors with themselves, they are also weak. The strongest correlation found is a negative association between EESR and regularity (-0.39 Spearman correlation) i.e. more corners tend to imply less regularity. This aligns with our observation that molecules with lower regularity may be rounded, with many corners close to the minimal error ellipse. Another negative correlation can be found between arc proportion and EESR (-0.29 Spearman correlation) indicating that molecules with "wide" corners tend to have a smaller number of such corners. Naturally, if the molecule has corners that occupy more space, necessarily needs to have less of them.

Categories of molecular belts

We analyzed the space of molecular belts based on their shape descriptors. The aim is to produce a map of the molecules based on their shape features. The results of this analysis can be seen in Fig. 4. Interestingly, shape descriptors produce a disperse plot on the PCA space, with variance explanation of PC1 and PC2 is of 51.1% and 29.8%, respectively. The plot and variance explanation reflect the high heterogeneity in shapes of molecular belts. DBSCAN over molecular shape successfully identifies numerous clusters, corresponding mostly with EESR values. This is not surprising due to the discrete nature of this descriptor. Clusters are mixed up within the PCA components and the different directions indicate increase in arc proportion (horizontal axis) and ellipse ratio (vertical axis) respectively, which is indicative of these descriptors driving an important amount of variance in the population.

The clustering capacity showed by shape descriptors cannot be achieved strictly with porosity descriptors. We performed a similar analysis over these, without being able to determine shapes nor construct any clusters. More details are given in SI.

Chemistry of molecular belts

To have a more complete description of our repository of molecular belts, we studied their chemistry analyzing the organic chemical groups present in the molecules. These results are shown in Fig. 5, where the number of structures containing each group (or pair of groups) is shown. It can be seen that aromatic groups predominate over the rest, due to the large presence of benzene rings among molecular belts. Some pairs of groups can be found more often together (e.g. aromatic ethers and aromatic groups coexist in high numbers). In general, aromatic groups are those paired more often, which is a natural consequence of them being present in most of the molecules presented. We further studied chemistry correlations with porosity and shape properties of molecules. No relevant correlations were found, suggesting that there's a broad chemical heterogeneity leading to different molecular shapes and pore sizes (more details can be found in SI).

Shape-based discovery of nanotube-like PMM

We aim to demonstrate the potential application of shape analysis for the discovery of porous molecular materials with specific properties. Our goal is to detect porous crystals with nanotube-like channels, with help of shape analysis. To do so, we select three noria-like molecules known to form nanotube crystals and characterize their shapes. In a previous work by our group, we extracted individual molecular candidates from CSD.[26] Here, we extract and characterize molecular belts within CSD dataset to find molecules with shape descriptors similar to the selected noria-like molecules. In particular, these noria-like molecules presented arc proportion values ranging from 0.42 to 0.48. We allowed molecules in the interval [0.3, 0.6], selecting those with EESR = 3, mLCD > 3A and regularity > 0.8. All the structures obtained

are further analyzed with help of Zeo++[27] to see if they presented mPLD/PLD ratios close to 1, an expected property of nanotube crystals. Since this property could also hold coincidentally on non-nanotube structures, visual confirmation is needed, which was done with the Mercury visualization tool provided by CSD.

The shape-based filter resulted in the finding of 11 structures with building blocks that presented shape descriptors within the desired ranges. Posterior analyses confirmed that 6 of these structures were nanotube-like, i.e. a relevant percentage (61.3% if counting the three initial structures) of the structures formed by rounded molecules with three wide corners form the crystals of our interest. Although this percentage suggests that shape can be informative about the properties of crystal structure, this message must be taken with caution, as 39.7% of the molecules found do not have this property. Five of the discovered structures can be found in Fig. 6. All 11 molecules from initial filter, along their shape descriptors, mLCD and PLD, are described in Table 2.

To further extend this analysis, we applied the same filters to our repository of 4412 belts, in order to discover a set of potential candidate molecules that could form nanotube-like structures. This led us to the detection of 52 molecules with similar shape descriptors. We intend to further examine their crystal structures in the future to confirm their material properties. The list of molecules is presented in the Supporting Information file.

Conclusions

We presented a set of molecular shape descriptors dedicated to molecular belts, which are based on a projection of the molecules into according planes and further characterization of the projection with a minimal error ellipse. Additionally, we introduced the largest known repository of molecular belts, mined from PubChem, with a total of 4412 molecules. We analyzed this repository in terms of molecular shape, chemistry and porosity, gaining an insight of the properties of the deposited molecules. Our analyses on the repository revealed that: (1) the set of molecular belts is dominated by non-elongated molecules, and the main shape (characterized through EESR) follows a normal distribution centered in 4-corners (i.e. squared or rectangular molecules); (2) there are weak correlations between molecular porosity and shape, indicating that molecular shape adds relevant information about molecular belts, whereas molecular chemistry shows no correlation with porosity or shape; and (5) molecular shape provides a form of classification for molecular belts that wasn't possible with previously existing descriptors (e.g. porosity descriptors). We further demonstrated the introduced methods with an application in the field of materials science, focusing on the discovery of porous molecular materials. Our efforts led us to the identification of five porous molecular crystals with nanotube structure based on similarity analysis with molecules with the same known properties from CSD database. We extended this analysis to our repository, identifying 52 belt candidates to be assembled into nanotube crystals.

FIGURES

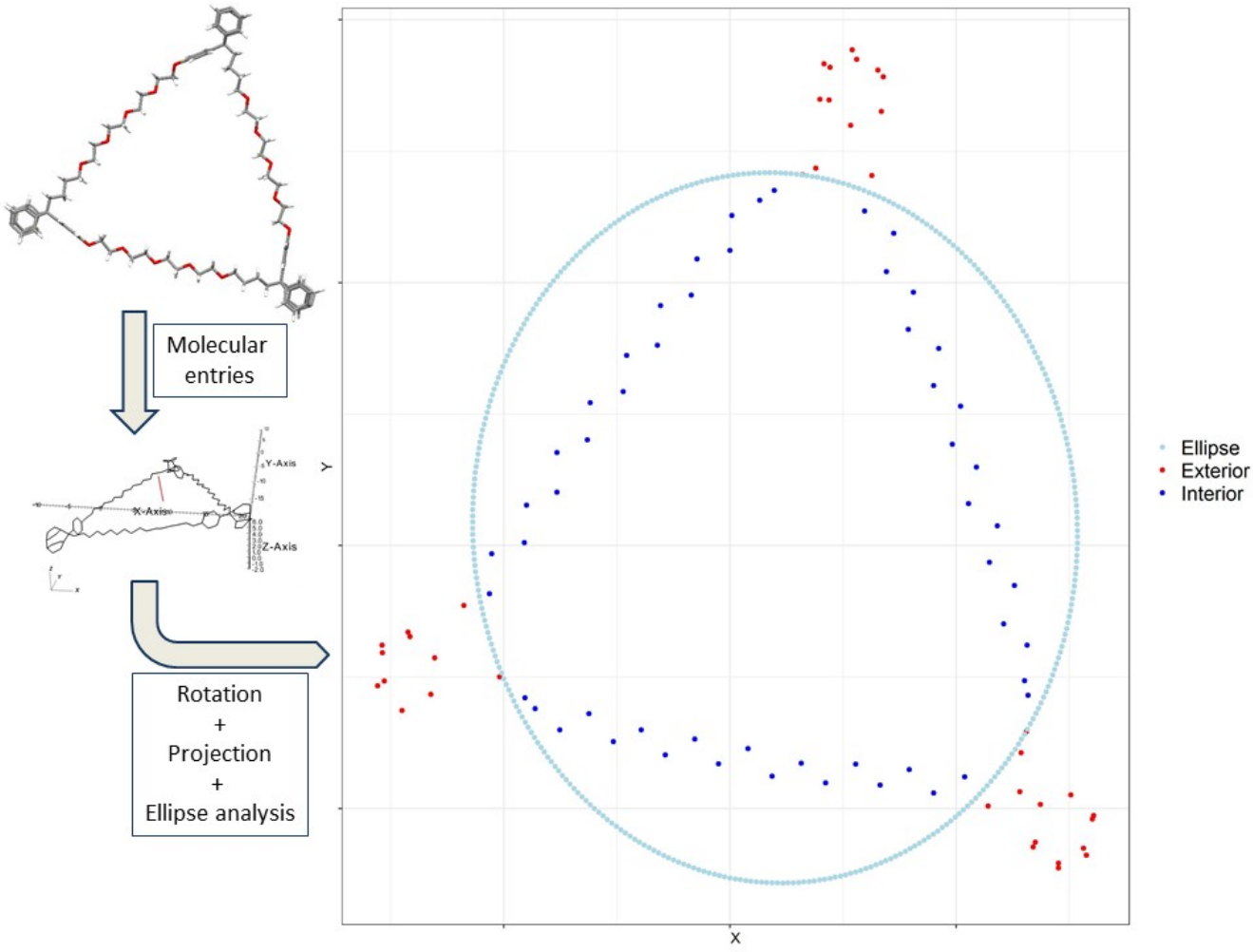

Figure 1. Shape analysis depiction. To determine the shape of a molecular belt, we first rotate it so that the largest molecular entry (previously computed with Molipor) is aligned with the Z-axis. Then we project the atoms on the XY-plane. By doing this, we obtain a 2-dimensional representation of molecular atoms that resembles molecular shape. Then, we compute the minimal error ellipse, and determine the external regions. With help of density-based clustering (DBSCAN) we determine the number of corners as the number of clusters in the DBSCAN algorithm. This is the EESR number. Other parameters are determined using both the external points (arc proportion and regularity) and the ellipse major and minor radii (ellipse ratio).

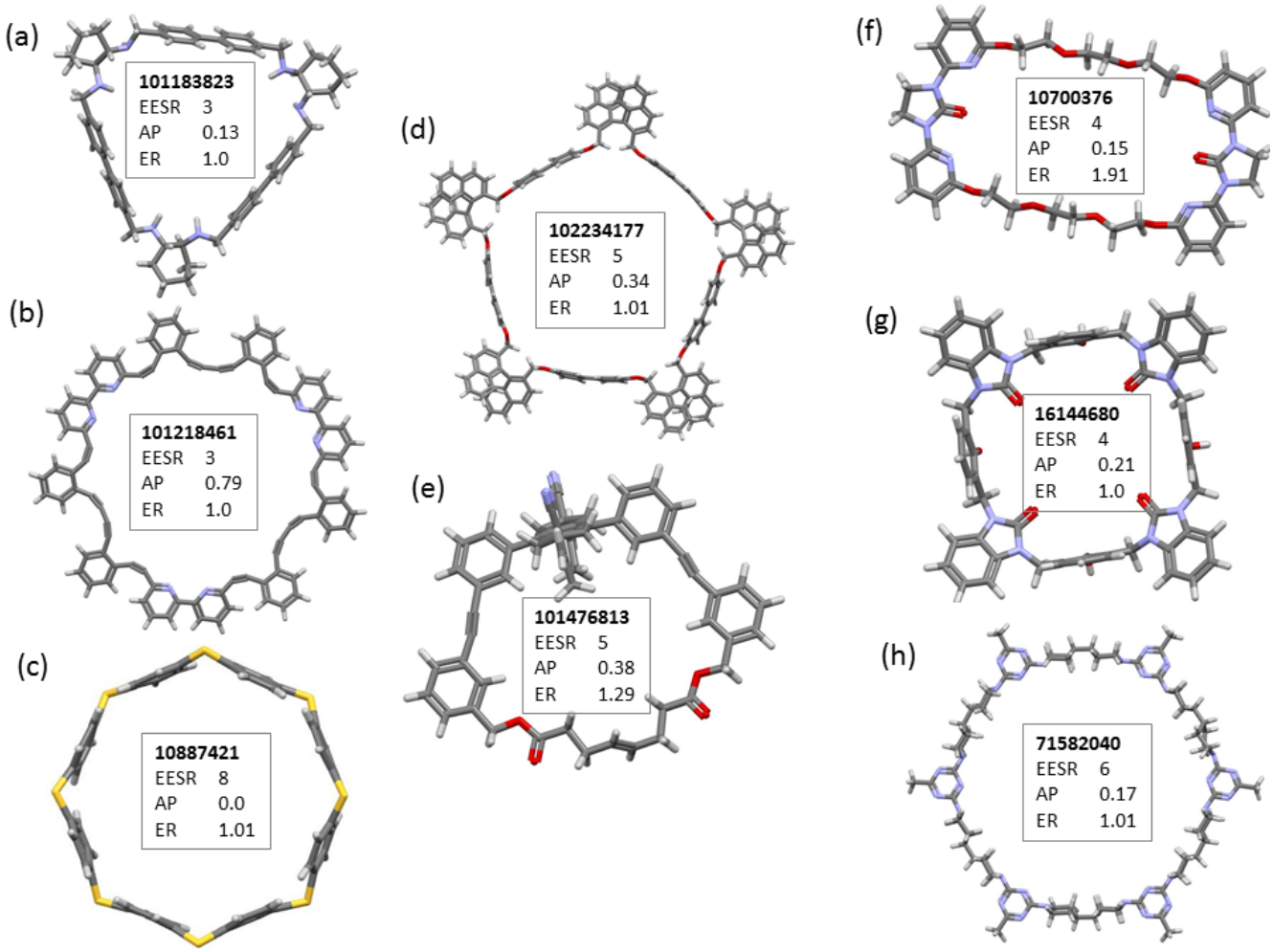

Figure 2. Shape analysis examples. Eight molecules, taken from our repository, are shown to demonstrate the methodology. Each molecule is presented with EESR, arc proportion (AP) and ellipse ratio (ER). In all cases, EESR coincides with the intuition of what would be the number of vertices of the equivalent polygon. Arc proportion is demonstrated through examples (b), (d) and (e): these have wide corners and larger AP values. Ellipse ratio and its capability to identify rectangular/elongated shapes is demonstrated through examples (e) and (f).

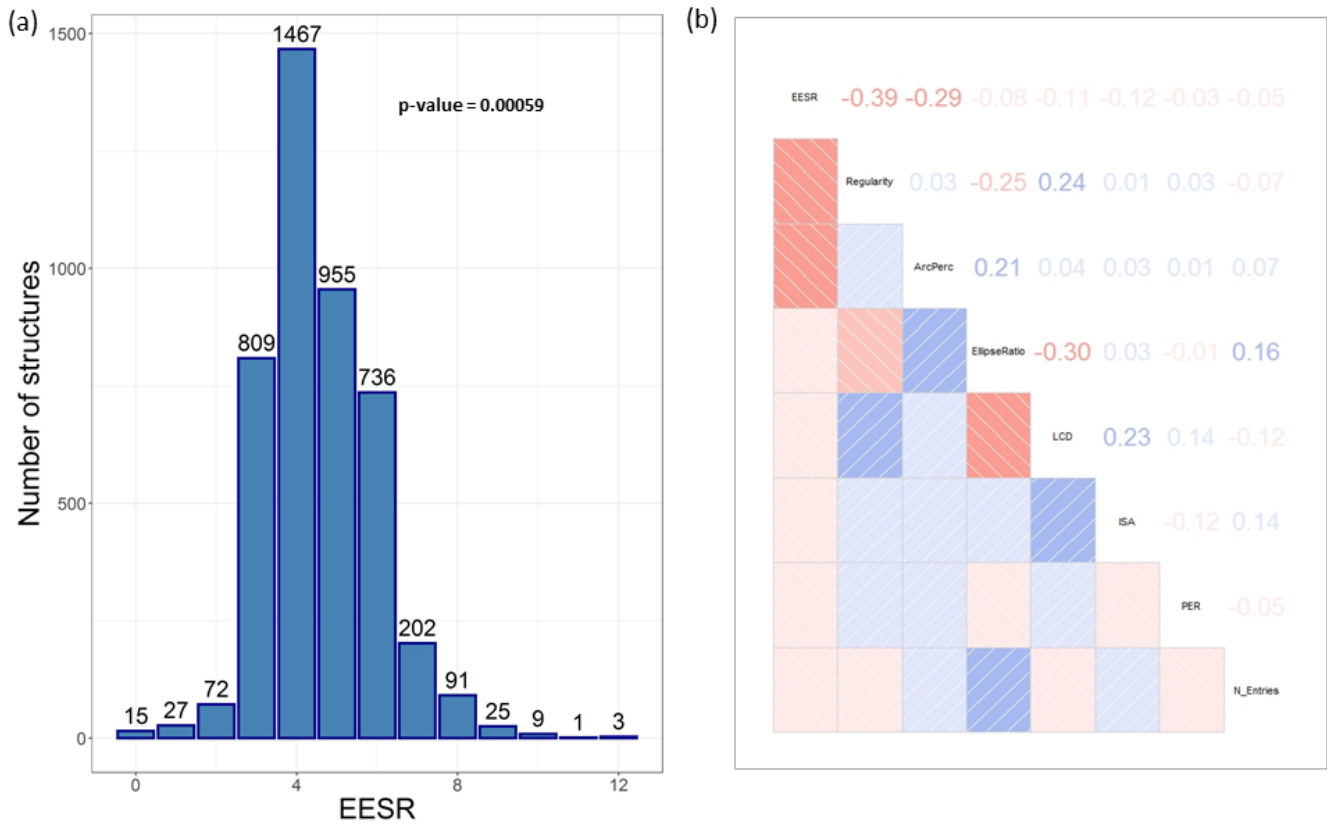

Figure 3. Shape descriptive and descriptor correlations. (a) Distribution of EESR values within the population of belt-like molecules, alongside the p-value from Shapiro-Wilk test for normality. (b) Correlation diagram reflecting Spearman correlations between pairs of shape and porosity descriptors for the molecular belts of the repository. Upper triangular panel shows actual correlation values, whereas lower triangular panel shows color coding corresponding to the values shown in the upper panel. Red and blue colors indicate negative and positive correlation, respectively.

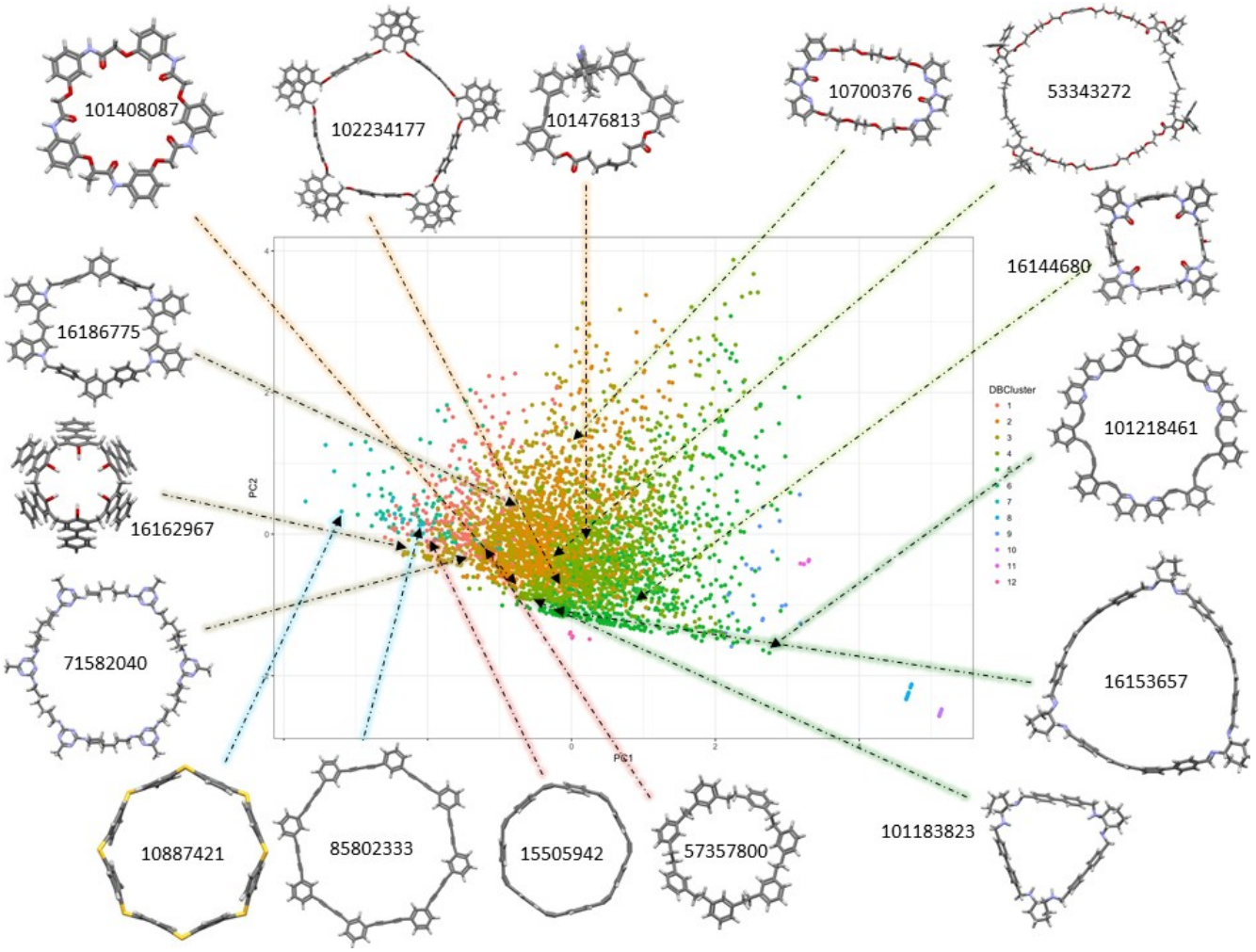

Figure 4. PCA combined with DBSCAN for the set of PubChem molecular belts. The color-coding reflects clustering under DBSCAN. Molecules are clustered mainly by their EESR. The principal components capture combined information between the EESR and the arc proportion descriptors (PC1) and the ellipse ratio (PC2). There's a color gradient across the PC1 axis (i.e. EESR governs this direction). Molecules with higher arc proportion tend to be at the right side of the plot (e.g. 101218461) whereas molecules with smaller arc proportion are placed at middle and left side of the plot (e.g. 101183823). Accordingly, more elongated molecules tend to be at the top of the plot (e.g. 10700376).

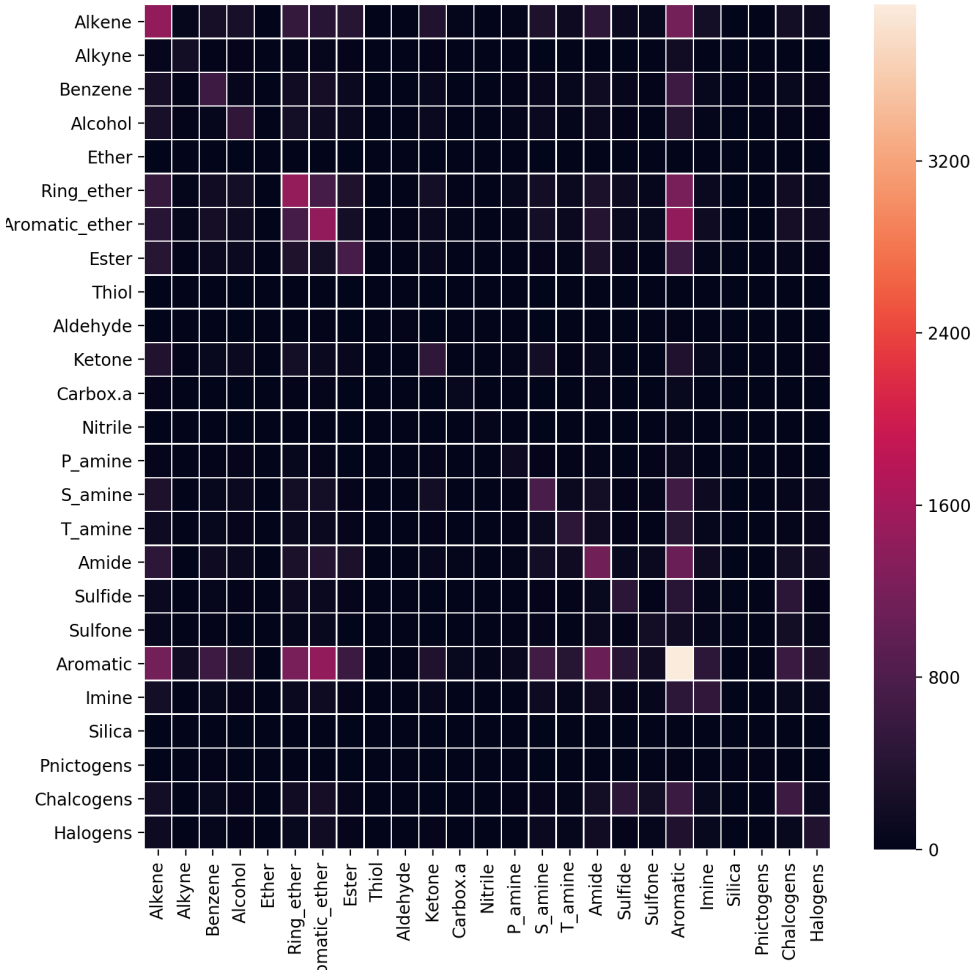

Figure 5. Chemistry of molecular belts. The diagram represents the number of molecules containing a given group (diagonal) and the number of molecules containing two groups (non-diagonal).

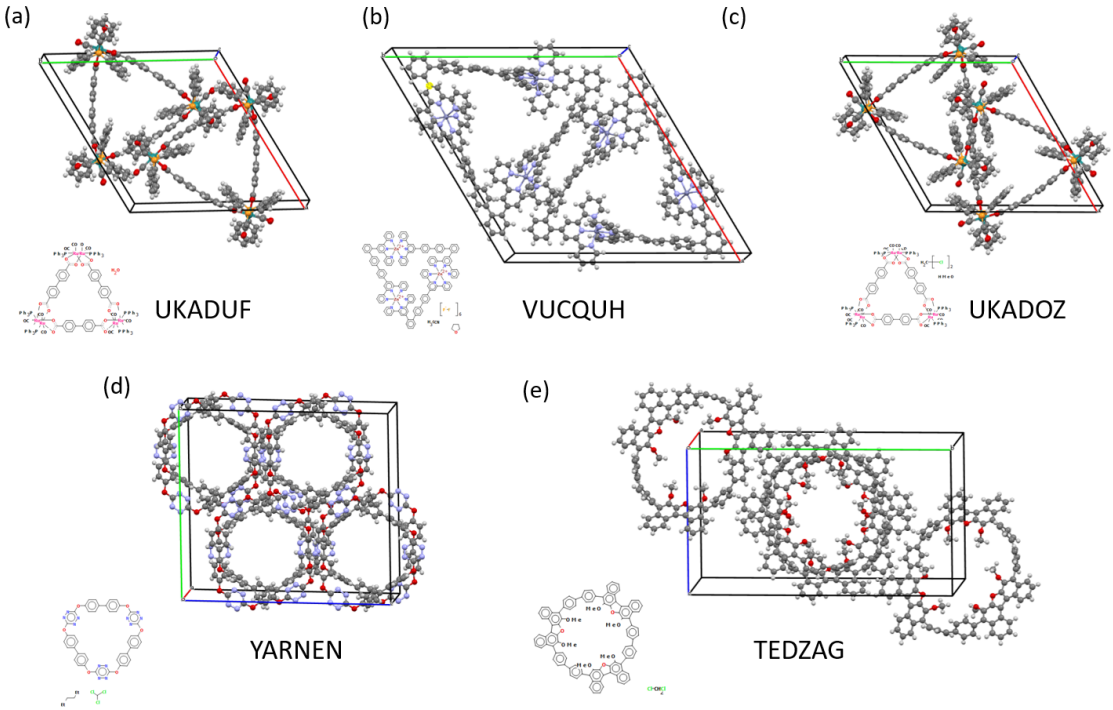

Figure 6. Molecules selected on the screening process (top) and their corresponding nanotube-like crystal structures (bottom). The molecules can be seen to share shape properties with those used as reference (high arc proportion, EESR = 3, ellipse ratio ~ 1).

TABLES

Parameter	*Min.*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max.*
Arc proportion	0	0.28	0.36	0,37	0.45	1
Regularity	0	0.56	0.66	0,67	0.79	1
Ellipse ratio	1	1.11	1.26	1,36	1.48	10.11

Table 1. Summary of the shape descriptors over the set of belt-like molecules. All parameters are presented with minimum, 1st and 3rd quartiles, mean, median and maximum.

CID	*EESR*	*Arc proportion*	*Ellipse ratio*	*Regularity*	*mLCD*	*Material LCD*	**Material PLD**	*Nanotube structure*
CEBCOC	3	0,33	1,01	0,81	6,70	9.93	4.70	No*
KAHQEP	3	0,35	1,05	0,97	4,15	3.17	1,58	No
MEXBEW	3	0,31	1,04	0,96	5,78	6.41	3.73	No
OJIDOB	3	0,4	1,06	0,9	7,88	7.92	5.26	No*
TEDZAG	3	0,39	1,02	0,88	6,21	6.96	4.48	**Yes**
UFANIY	3	0,31	1,05	1	9,29	10.09	4.97	No
UKADOZ	3	0,36	1,01	1	5,84	7.55	4.63	**Yes**
UKADUF	3	0,35	1	1	5,79	7.87	5.10	**Yes**
VUCQUH	3	0,43	1,01	1,01	7,56	8.57	6.53	**Yes**
YAQVEU	3	0,39	1,06	0,98	6,19	10.17	6.47	**Yes***
YARNEN	3	0,5	1,06	0,87	6,86	7.67	6.45	**Yes**
YELKUW	3	0,42	1,09	0,87	6,13	6.49	4.94	Yes
YELLAD	3	0,48	1,02	0,88	6,68	7.87	6.52	Yes

Table 2. Descriptors of mined structures with similar shape to reference. Shape descriptors and mPLD for the molecules with similar shape to those known to form nanotube structures (YELLAD and YELKUW, included in the table) and corresponding material properties of interest. LCD and PLD after solvent removal. Two structures (CEBCOC and OJIDOB) are marked as no nanotube, although they show non-aligned 1-dimensional channels. The YAQVEU structure is a cocrystal

with C60 that shows nanotube channels, but the behavior of the molecule alone is unclear.

ASSOCIATED CONTENT

Supporting Information.

A Supporting information file in PDF is provided alongside this work.

AUTHOR INFORMATION

Corresponding Author

* Maciej Haranczyk

Author Contributions

The development of novel methods, statistical analyses and simulations where conducted by IGG. The analysis of chemistry was conducted by MCS. The conception of this study was done by MH and IGG. All authors contributed in manuscript writing and reviewing.

Funding Sources

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231.

ABBREVIATIONS

LCD, (material) largest cavity diameter; mLCD, molecular largest cavity diameter; PER, pore exposure ratio; ISA, (molecular) internal surface area; MWS, maximum window size.

REFERENCES

(1) Lakshmanan, V. I.; Vijayan, S. A Review on Application of Crown Ethers in Separation of Rare Earths and Precious Metals. *Miner. Met. Mater. Ser.* **2018**. https://doi.org/10.1007/978-3-319-95022-8_159.

(2) Wang, J.; Zhuang, S. Cesium Separation from Radioactive Waste by Extraction and Adsorption Based on Crown Ethers and Calixarenes. *Nucl. Eng. Technol.* **2020**, *52* (2), 328–336. https://doi.org/10.1016/j.net.2019.08.001.

(3) Hasell, T.; Cooper, A. I. Porous Organic Cages: Soluble, Modular and Molecular Pores. *Nat. Mater.* **2016**, *1* (9), 16053. https://doi.org/10.1038/natrevmats.2016.53.

(4) Cooper, A. I. Porous Molecular Solids and Liquids. *ACS Cent. Sci.* **2017**, *3* (6), 544–553. https://doi.org/10.1021/acscentsci.7b00146.

(5) Smith, B. W.; Monthioux, M.; Luzzi, D. E. Encapsulated C60 in Carbon Nanotubes. *Nature* **1998**, *396* (6709), 323. https://doi.org/10.1038/24519.

(6) Caira, M. R.; Le Roex, T.; Nassimbeni, L. R. Selective Enclathration of Picoline Isomers by a Resorcinarene Host. *CrystEngComm* **2006**, *8* (3), 275–280. https://doi.org/10.1039/b601560c.

(7) Sakamoto, H.; Fujimori, T.; Li, X.; Kaneko, K.; Kan, K.; Ozaki, N.; Hijikata, Y.; Irle, S.; Itami, K. Cycloparaphenylene as a Molecular Porous Carbon Solid with Uniform Pores Exhibiting Adsorption-Induced Softness. *Chem. Sci.* **2016**, *7* (7), 4204–4210. https://doi.org/10.1039/c6sc00092d.

(8) Kameta, N.; Masuda, M.; Shimizu, T. Soft Nanotube Hydrogels Functioning as

Artificial Chaperones. *ACS Nano* **2012**, *6* (6), 5249–5258. https://doi.org/10.1021/nn301041y.

(9) Tominaga, M.; Takahashi, E.; Ukai, H.; Ohara, K.; Itoh, T.; Yamaguchi, K. Solvent-Dependent Self-Assembly and Crystal Structures of a Salen-Based Macrocycle. *Org. Lett.* **2017**, *19* (7), 1508–1511. https://doi.org/10.1021/acs.orglett.7b00264.

(10) Frontera, A.; Bauzá, A. Concurrent Aerogen Bonding and Lone Pair/anion-π Interactions in the Stability of Organoxenon Derivatives: A Combined CSD and Ab Initio Study. *Phys. Chem. Chem. Phys.* **2017**, *19* (44), 30063–30068. https://doi.org/10.1039/c7cp06685f.

(11) Hisaki, I.; Nakagawa, S.; Tohnai, N.; Miyata, M. A C3-Symmetric Macrocycle-Based, Hydrogen-Bonded, Multiporous Hexagonal Network as a Motif of Porous Molecular Crystals. *Angew. Chemie - Int. Ed.* **2015**, *54* (10), 3008–3012. https://doi.org/10.1002/anie.201411438.

(12) Bernabei, M.; Pérez-Soto, R.; Gómez García, I.; Haranczyk, M. Towards Stable Porous Crystalline Phases of Molecular Belts. *CrystEngComm* **2017**, No. 19, 6932–6935. https://doi.org/10.1039/C7CE01679D.

(13) Miklitz, M.; Jiang, S.; Clowes, R.; Briggs, M. E.; Cooper, A. I.; Jelfs, K. E. Computational Screening of Porous Organic Molecules for Xenon/Krypton Separation. *J. Phys. Chem. C* **2017**, *121* (28), 15211–15222. https://doi.org/10.1021/acs.jpcc.7b03848.

(14) Gómez García, I.; Bernabei, M.; Pérez Soto, R.; Haranczyk, M. Out-of-Oblivion Cage Molecules and Their Porous Crystalline Phases. *Cryst. Growth*

Des. **2017**, *17* (11), 5614–5619. https://doi.org/10.1021/acs.cgd.7b01095.

(15) Sturluson, A.; Huynh, M. T.; York, A. H. P.; Simon, C. M. Eigencages : Learning a Latent Space of Porous Cages. *ArXIV* **2018**, 1–16.

(16) Kim, S. K.; Lynch, V. M.; Hay, B. P.; Kim, J. S.; Sessler, J. L. Ion Pair-Induced Conformational Motion in calix[4]arene-Strapped calix[4]pyrroles. *Chem. Sci*. **2015**, *6* (2), 1404–1413. https://doi.org/10.1039/c4sc03272a.

(17) McCann, B. W.; Silva, N. De; Windus, T. L.; Gordon, M. S.; Moyer, B. A.; Bryantsev, V. S.; Hay, B. P. Computer-Aided Molecular Design of Bis-Phosphine Oxide Lanthanide Extractants. *Inorg. Chem*. **2016**, *55* (12), 5787–5803. https://doi.org/10.1021/acs.inorgchem.5b02995.

(18) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res*. **2016**, *44* (D1), D1202–D1213. https://doi.org/10.1093/nar/gkv951.

(19) Gómez García, I.; Bernabei, M.; Haranczyk, M. Toward Automated Tools for Characterization of Molecular Porosity. *J. Chem. Theory Comput*. **2019**, *15* (1), 787–798. https://doi.org/10.1021/acs.jctc.8b00764.

(20) Evans, J. D.; Huang, D. M.; Haranczyk, M.; Thornton, A. W.; Sumby, C. J.; Doonan, C. J. Computational Identification of Organic Porous Molecular Crystals. *CrystEngComm* **2016**, *18*, 4133–4141. https://doi.org/10.1039/C6CE00064A.

(21) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge

Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72* (2), 171–179. https://doi.org/10.1107/S2052520616003954.

(22) Halir, R.; Flusser, J. Numerically Stable Direct Least Squares Fitting Of Ellipses. *Proc. 6th Int. Conf. Cent. Eur. Comput. Graph. Vis.* **1998**, 125–132.

(23) Jolliffe, I. T. Principal Component Analysis. *Springer-Verlag New Yor* **2002**, 93. https://doi.org/https://doi.org/10.1007/b98835.

(24) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proc. Second Int. Conf. Knowl. Discov. Data Min.* **1996**, *1*, 226–231. https://doi.org/10.1016/B978-044452701-1.00067-3.

(25) Wickham, H. Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **2011**, *3* (2), 180–185. https://doi.org/10.1002/wics.147.

(26) Gómez García, I.; Haranczyk, M. Toward Crystalline Porosity Estimators for Porous Molecules. *CrystEngComm* **2020**, No. 2. https://doi.org/10.1039/c9ce01753d.

(27) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and Tools for High-Throughput Geometry-Based Analysis of Crystalline Porous Materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134–141. https://doi.org/10.1016/j.micromeso.2011.08.020.