SUPPLEMENTARY INFORMATION

Self-Organizing Maps as a data-driven approach to reveals the variety of packing motives of PDI derivatives

Francesco Marin[a], Alessandro Zappi*[,a], Dora Melucci[a], Lucia Maini*[,a]

[a]Dipartimento di Chimica "G. Ciamician", via Selmi 2, Università di Bologna, 40126 Bologna, Italy

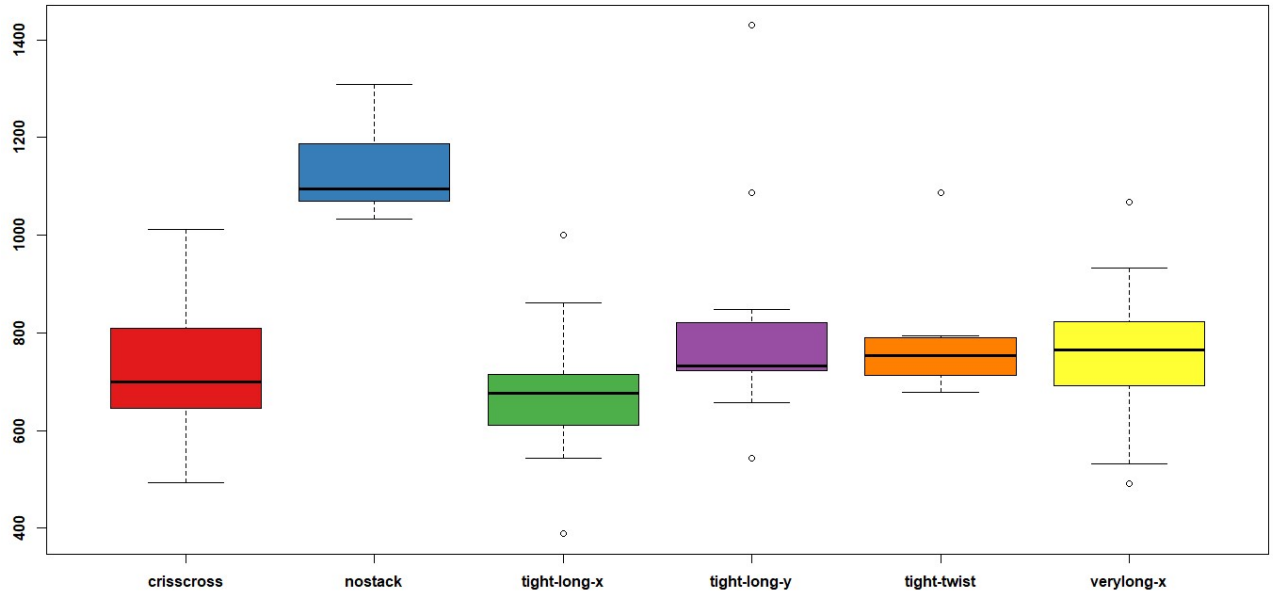

Figure S1. Boxplot of the molecular volume (V_{mol}) distribution among the identified families.

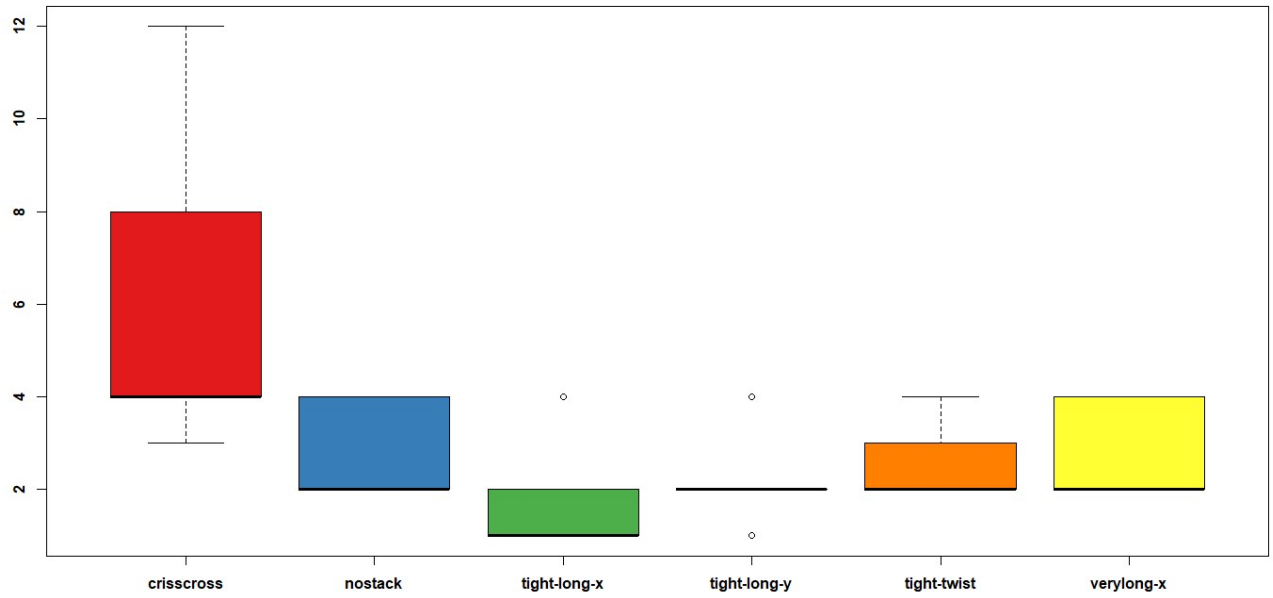

Figure S2. Boxplot of the number of molecules in the unit cell (Z) distribution among the identified families.

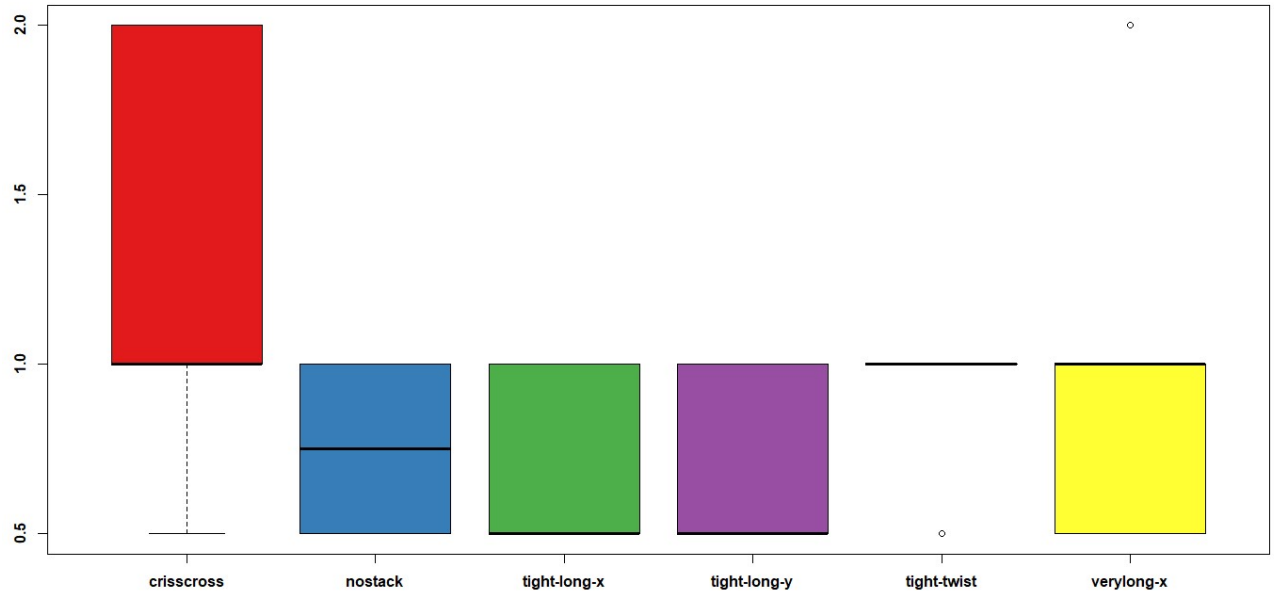

Figure S3. Boxplot of the number of molecules in the asymmetric unit (Z') distribution among the identified families.

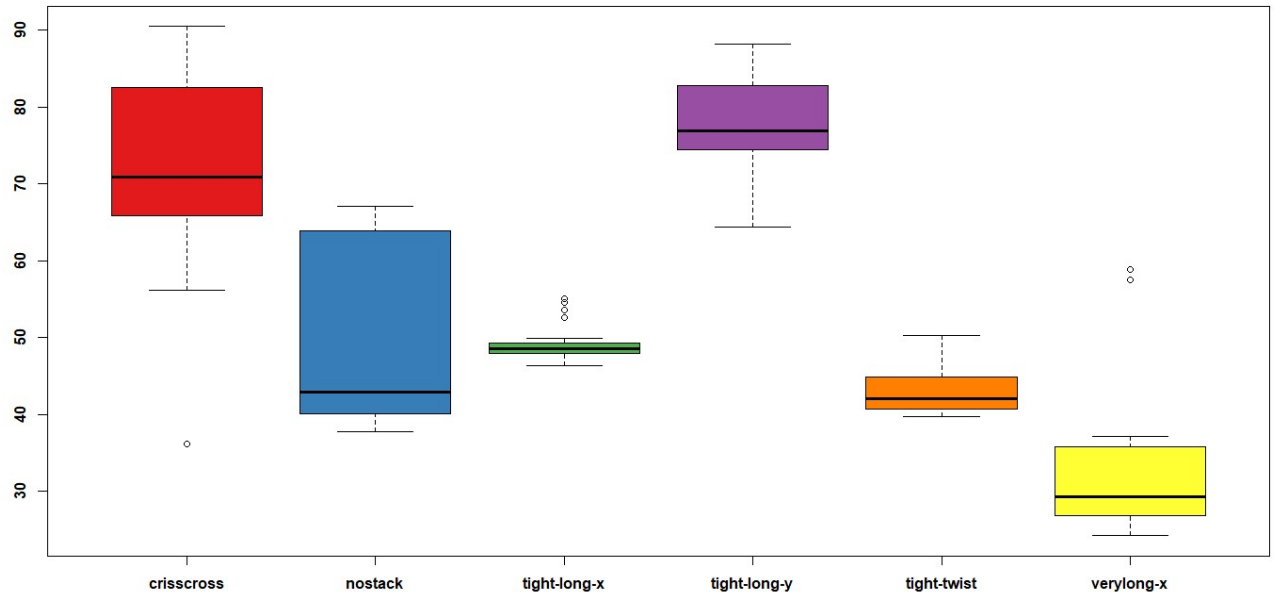

Figure S4. Boxplot of the angles of the direction cosines of the SV with the perylene x-axis (χ) distribution among the identified families.

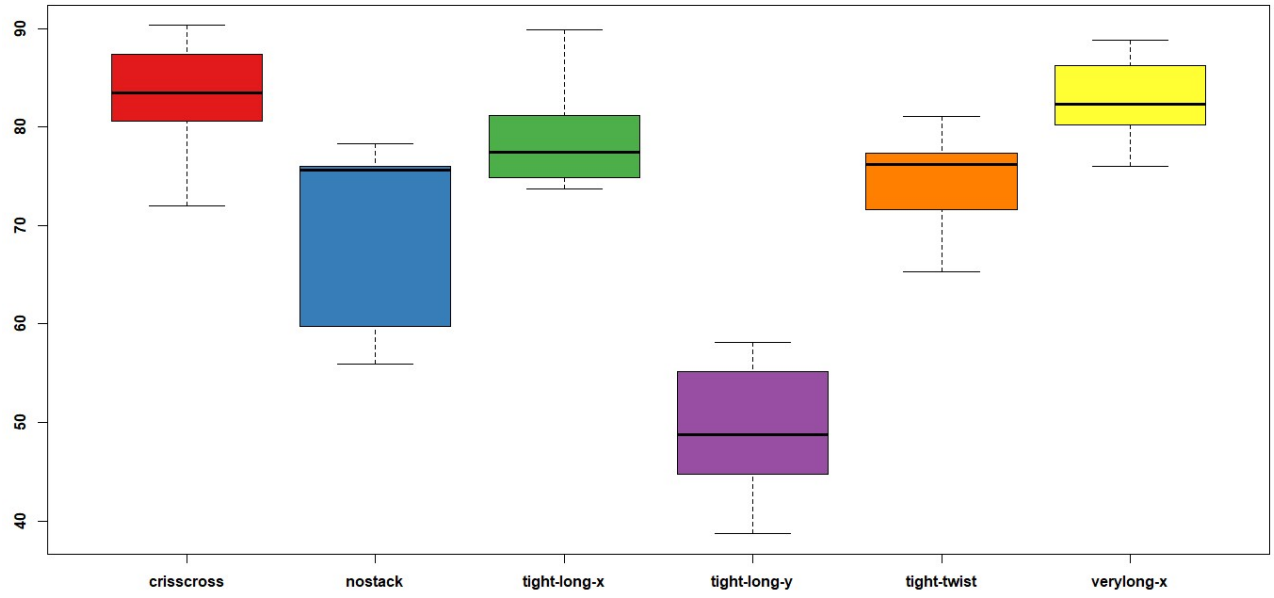

Figure S5. Boxplot of the angles of the direction cosines of the SV with the perylene y-axis (ψ) distribution among the identified families.

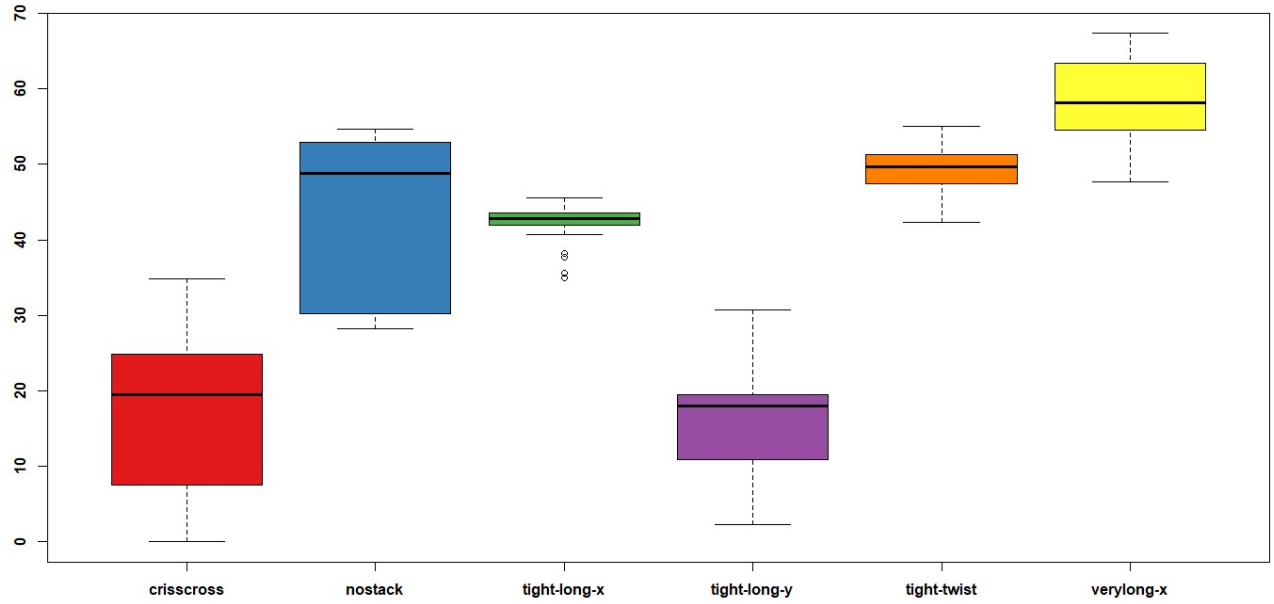

Figure S6. Boxplot of the pitch angles (P) distribution among the identified families.

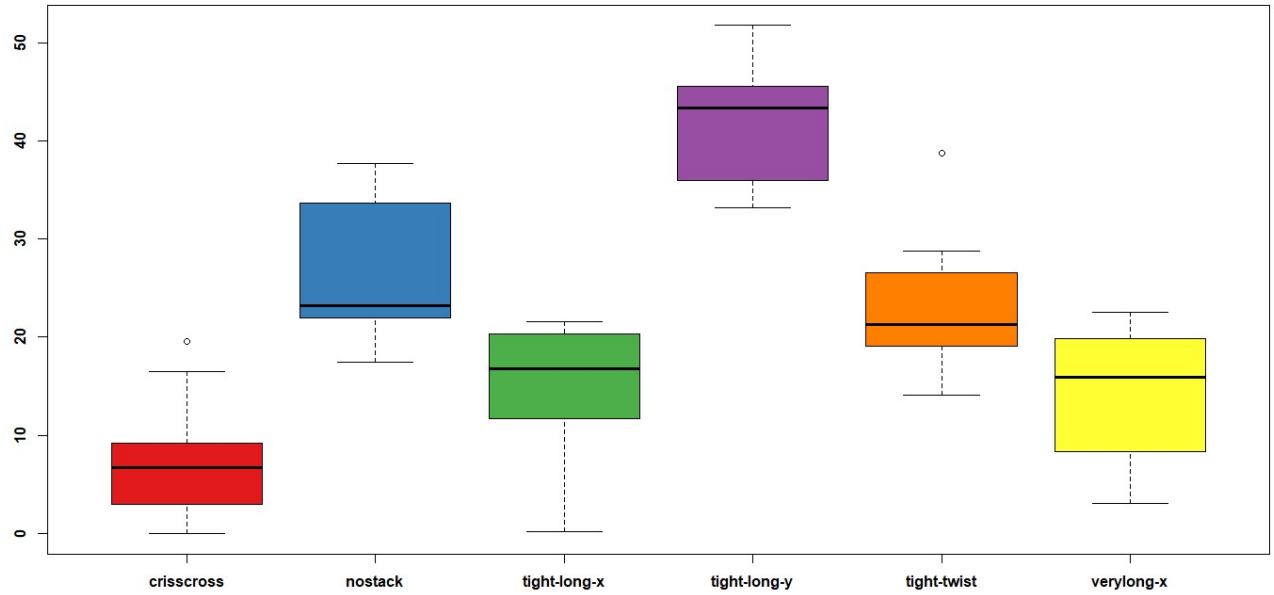

Figure S7. Boxplot of the roll angles (R) distribution among the identified families.

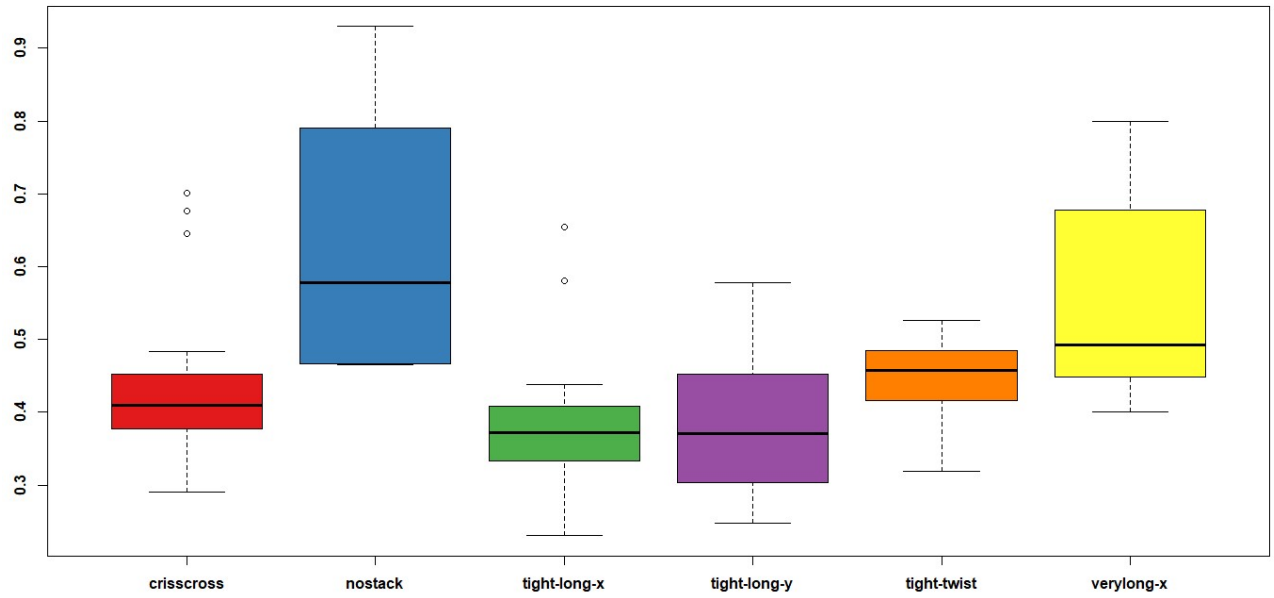

Figure S8. Boxplot of the aspect ratio of the molecule described as enclosed in a rectangular box, with here showed the ratio of the medium and long box axes (M/L) distribution among the identified families.

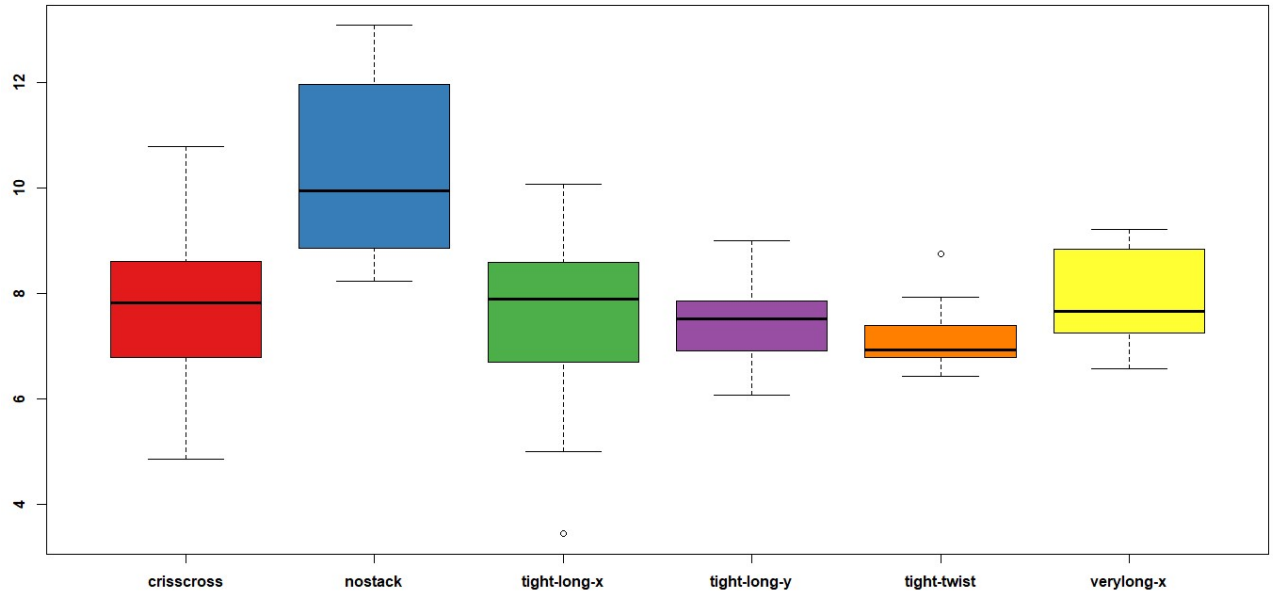

Figure S9. Boxplot of the aspect ratio of the molecule described as enclosed in a rectangular box, with here showed the length of the short box axis (S) distribution among the identified families.

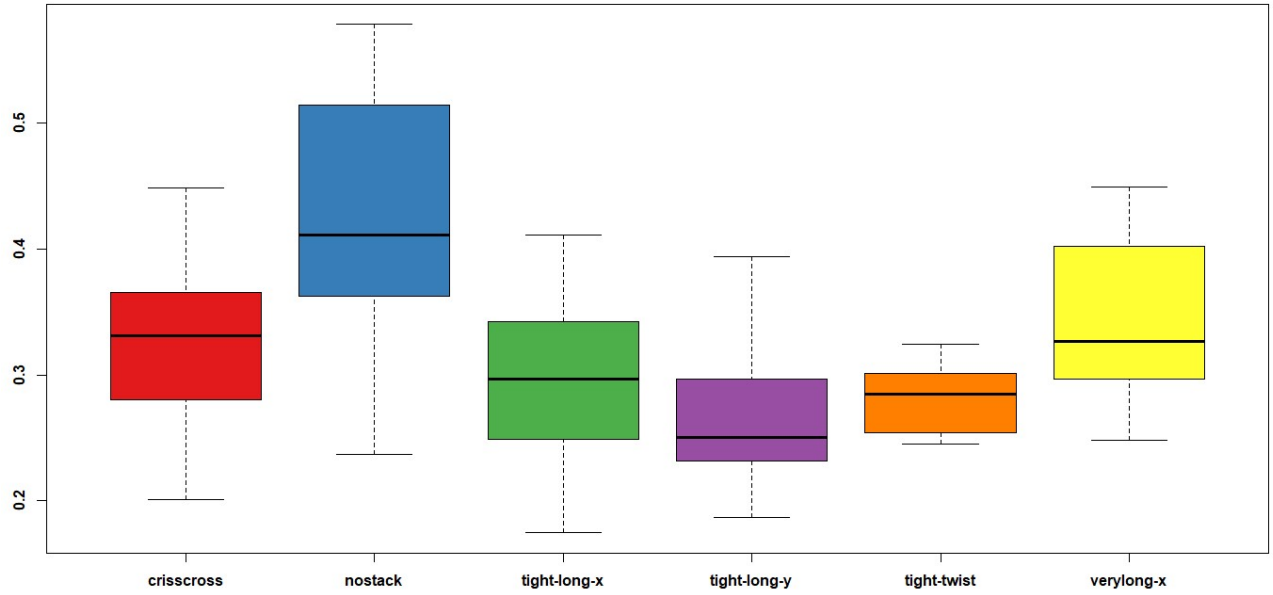

Figure S10. Boxplot of the aspect ratio of the molecule described as enclosed in a rectangular box, with here showed the ratio of the short and long box axes (S/L) distribution among the identified families.

The result of SOM grouping visualized using the PCA scores plot are shown, with the *extra* divided into the 5 small groups. The scores of PC1 vs PC2 are shown, the colour of the scores correspond to the group in which each object was categorized.

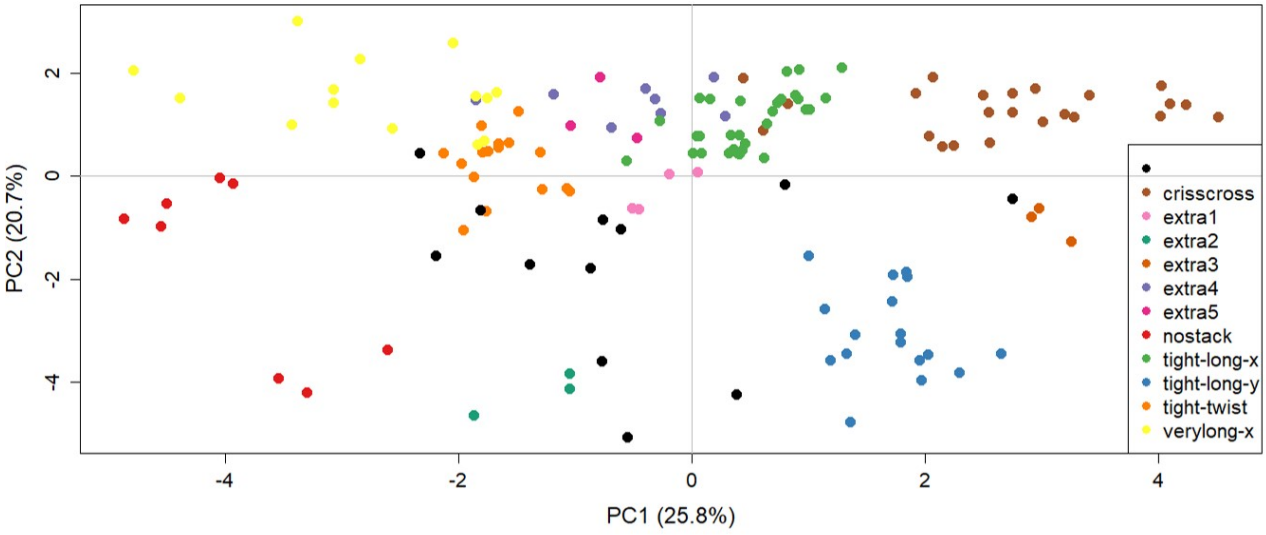

Figure S11. PC1-PC2 scores plot obtained by PCA, with the scores categorized by the group they belong to. In particular, the 6 major families are shown, and the extra is divided into the respective 5 small groups. With extra1 in pink; extra2 in dark green; extra3 in dark orange; extra4 in purple; extra5 in violet.

The packing of YIWMEY, which is one of the two objects displaying dipole-π interaction between the perylene core and the carbonyl oxygen of the closest molecule, is shown as an example.

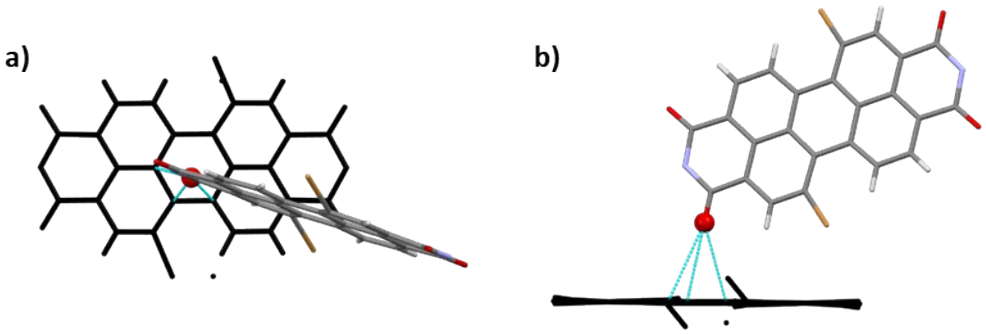

Figure S12. Dipole-π interaction in YIWMEY packing; a) viewed along z-direction; b) viewed along x-direction.

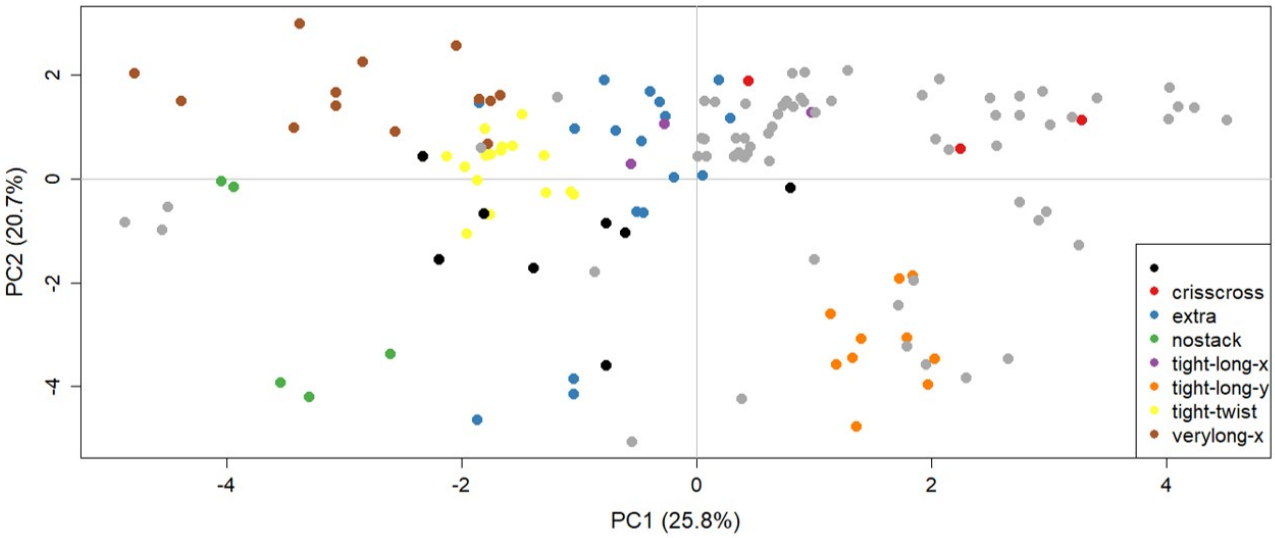

Figure S13. PC1-PC2 scores plot obtained by PCA, with the *bay*-, *ortho*-, and *bay*- and *ortho*-substituted objects categorized by the family they belong to by the colour. The *imide*-substituted objects are coloured in grey. Most of the *core*-substituted objects are at negative PC1 values, where the π-π stacking interactions between perylene cores becomes weaker and other type of interactions becomes dominant.

In the present paragraph, we present the procedure and outcomes of the SOM analysis.[1] As already stated in the main text, the SOM analyses were carried out with the package SOMEnv[2] of the R environment. Here we present the SOM calculated with map dimensions 10 x 6, but the results are analogous for all the other calculated maps.

The dataset is composed of 142 objects (crystal structures) and 17 numeric variables (described in the main text). The computation starts by choosing the map dimensions and shape, in this case, a rectangular map composed of 60 units divided into 10 rows and 6 columns.[3] The first step of the computation produces an output as the one reported in Figure S1 in which the map is shown, and the dimension of the black filling is proportional to the number of objects in each unit. At this stage, it is not important to know how many objects are assigned to each unit, it is most of all interesting to observe that there are regions of the map that are more populated than others and that there are also some empty units.

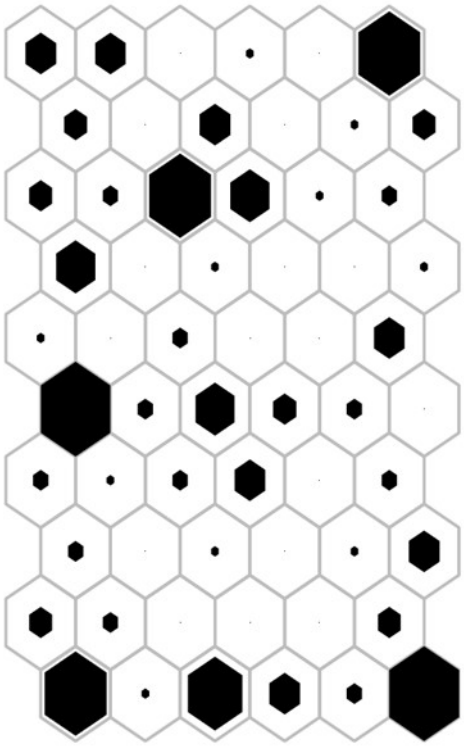

Figure S14. SOM map reporting the number of objects assigned to each unit: black fillings are proportional to the number of objects.

The second important output of this first stage is reported in Figure S2. It concerns the role of variables in the SOM computation. The filling of the hexagons, in this case, is proportional to the basic statistic of each variable: black colour represents higher quartiles in that region for that variable, white colour represents lower quartiles. This means that the objects assigned to a certain region of the map have higher or lower (median) values for that variable, based on the colour reported in Figure S2. From this graph, it is possible to evaluate some features of the variables. In

particular, if two or more variables show similar graphs, they would probably have the same general behaviour, *i.e.* they are strongly correlated. In the case shown in Figure S2, for example, the variables $d_{\pi\text{-}\pi}$ (called d.p.p in Figure S2), M/L (ML) and S have very similar SOM graphs, thus these can be considered correlated.

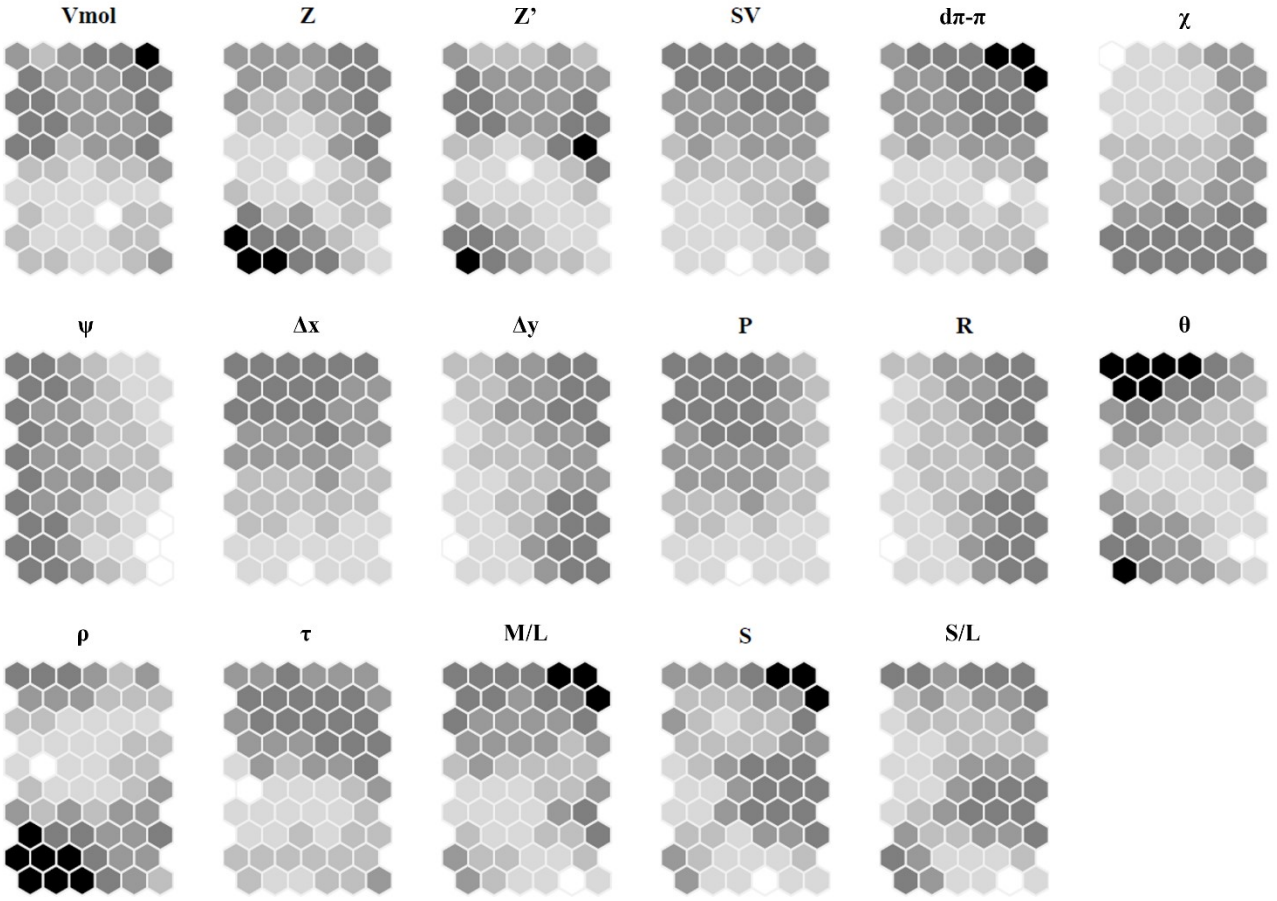

Figure S15. Variable's behaviour in the SOM analysis. Black and white colours represent respectively higher and lower quartiles of each variable in each map region.

From the map shown in Figure S1, a K-means cluster analysis can be computed, using the units as starting data.[4] In this case, the maximum number of possible clusters (k) has to be decided as input data and the algorithm calculates the optimal number of clusters from k=2 to the decided maximum k. The optimization is based on minimization of the DB-index,[5] as explained in the main text. In the present work, k_{max} was always put as 8. For the map presented in this paragraph, the optimal number of clusters resulted to be 5. The graphical output produced by SOMEnv is reported in Figure S3. Figure S3 shows the same map reported in Figure S1, but the colours represent the clusters to which each unit is assigned. Based on these clusters, the computations and the discussion reported in the main text were carried out.

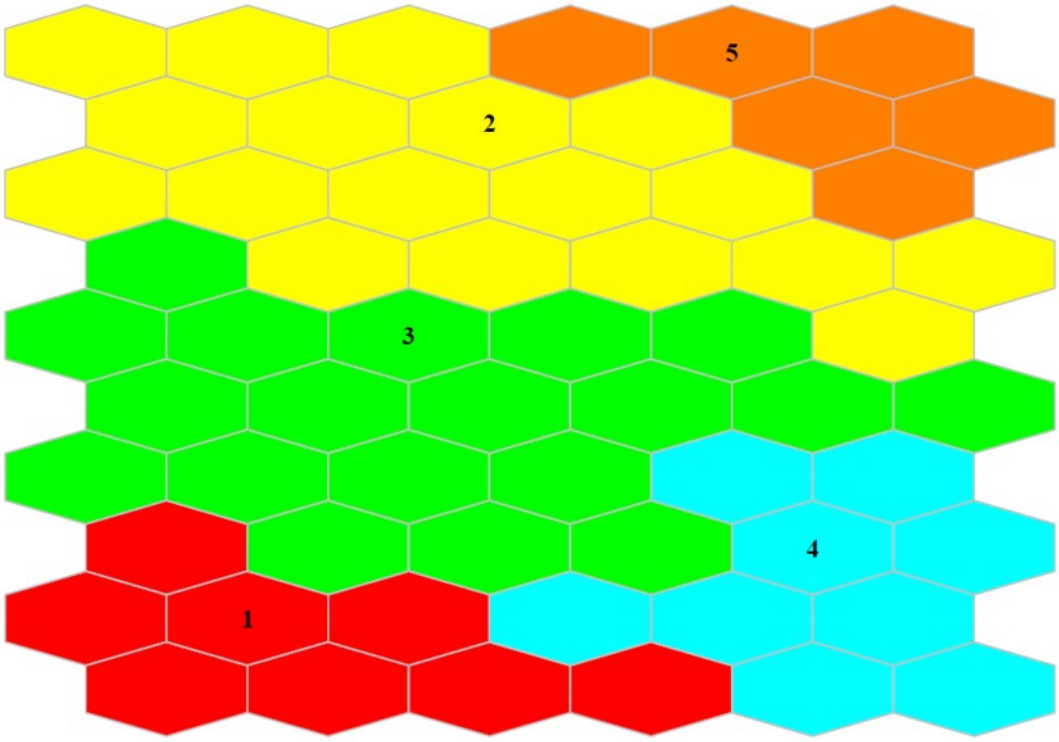

Figure S16. SOM map divided by colours after K-means cluster analysis. The units in which cluster-number is placed indicates the cluster centroids.

Also in this case, the role of variables in the clustering step can be graphically evaluated by the plot reported in Figure S4. In this case, boxplots of auto-scaled variables (each original data is subtracted to its column-mean and the result divided by the column standard deviation) are reported for each cluster. From these graphs, it can be argued which variables are the most characteristic of each cluster. In the example shown, Cluster 3 (in green in Figure S3) have no particular characteristic variables, while Cluster 4 (in light blue) is mainly characterized by variables χ (chi), Δy (Delta_y), and R.

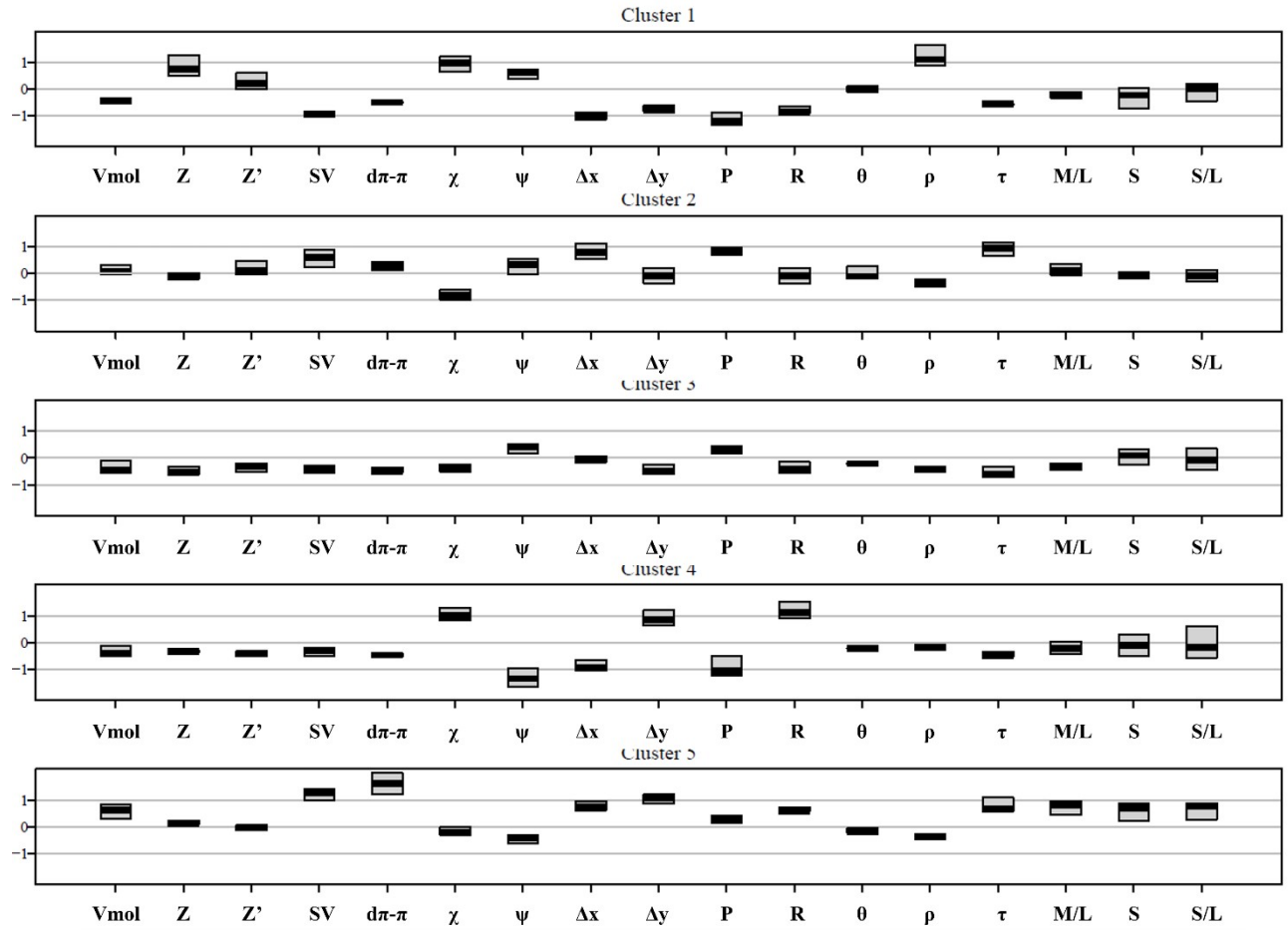

Figure S17. Boxplots of variables for each cluster. Higher median values indicate more characteristic variables for the corresponding cluster.

The boxplots showing the variation of the variables in the six packing families not reported in the main text are reported here. In the boxplots, the variation range of each variable is plotted for each family, and these can give information about the range and distribution of each variable within the family. The use of boxplots is helpful to identify the families' most important characteristics.

By looking only at the objects substituted at the core, we observe that most of them are found at negative PC1 of the scores plot where the π-π stacking interactions between perylene cores becomes weaker and other type of interactions becomes dominant.

REFERENCES

(1) Kohonen, T. The Self-Organizing Map. *Neurocomputing* **1998**, *21* (1), 1–6.

(2) Licen, S.; Franzon, M.; Rodani, T.; Barbieri, P. SOMEnv: An R Package for Mining Environmental Monitoring Datasets by Self-Organizing Map and k-Means Algorithms with a Graphical User Interface. *Microchem. J.* **2021**, *165*, 106181.

(3) Nakagawa, K.; Yu, Z.-Q.; Berndtsson, R.; Hosono, T. Temporal Characteristics of Groundwater Chemistry Affected by the 2016 Kumamoto Earthquake Using Self-Organizing Maps. *J. Hydrol.* **2020**, *582*, 124519.

(4) Everitt, B. S.; Landau, S.; Leese, M. *Cluster Analysis*, 5th ed.; Wiley Publishing, 2011.

(5) Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1* (2), 224–227.