Reproduced with permission from Chemical Research in Toxicology, Sleight, T.; Sexton, C.; Mpourmpakis, I.; Gilbertson, L.; Ng, C. A Classification Model to Identify Direct-Acting Mutagenic Polycyclic Aromatic Hydrocarbon Transformation Products. Chem. Res. Toxicol. Rev. 2021. Copyright 2021, American Chemical Society (in review)

A Classification Model to Identify Direct-Acting Mutagenic Polycyclic Aromatic Hydrocarbon Transformation Products

Supporting Information

Trevor W. Sleight1, Caitlin N. Sexton2, Giannis Mpourmpakis2, Leanne M. Gilbertson1,2\*, Carla A. Ng1,3\*

1. Civil and Environmental Engineering, University of Pittsburgh, Pittsburgh, PA 15261

2. Chemical and Petroleum Engineering, University of Pittsburgh, Pittsburgh, PA 15261

3. Environmental and Occupational Health, University of Pittsburgh, Pittsburgh, PA 15261

\*Corresponding authors: [carla.ng@pitt.edu](mailto:carla.ng@pitt.edu), (412) 383-4075; [leanne.gilbertson@pitt.edu](mailto:Leanne.gilbertson@pitt.edu), (412) 624-1683

Contents

[S-1 Principal Component Analysis Based Clustering 8](#_Toc70948078)

[S-2 Descriptor Selection and Regression 13](#_Toc70948079)

[S-2.1 Initial Dataset 14](#_Toc70948080)

[S-2.2 Large Molecules Cluster 17](#_Toc70948081)

[S-2.3 Small Molecules 19](#_Toc70948082)

[S-2.4 Descriptor Priority for Closely Correlated Descriptors 22](#_Toc70948083)

[S-3 Swarm Plots of Differences Between Clusters 24](#_Toc70948084)

[S-4 PaDEL-Descriptor Weightings 29](#_Toc70948085)

[S-5 PaDEL-Descriptor Equations 31](#_Toc70948086)

[S-5.1 Barysz Matrix 31](#_Toc70948087)

[S-5.2 Burden Modified Matrix 31](#_Toc70948088)

[S-5.3 Broto-Moreau Autocorrelation 32](#_Toc70948089)

[S-5.4 Moran Autocorrelation 32](#_Toc70948090)

[S-5.5 Geary Autocorrelation 32](#_Toc70948091)

[S-6 Representative Structures 33](#_Toc70948092)

**Figures and Tables**

[Figure S1.1: % Variance described by each principal component of the descriptors 5](#_Toc75768357)

[Figure S1.2: Elbow plot of within-cluster sum of squares. 6](#_Toc75768358)

[Figure S1.3: Silhouette Plot. 7](#_Toc75768359)

[Figure S1.4: Davies-Bouldin Plot. 8](#_Toc75768360)

[Figure S1.5: PC Plot of data used in this study. 9](#_Toc75768361)

[Figure S1.6: Silhouette values using three clusters. 10](#_Toc75768362)

[Figure S2.1: F1 plots from recursive feature elimination for the full dataset. 11](#_Toc75768363)

[Figure S2.2: Initial dataset ROC plot. 11](#_Toc75768364)

[Table S2.3: Confusion matrix for 10 rounds of cross validation for the entire dataset with 72 descriptors 12](#_Toc75768365)

[Figure S2.4: Initial dataset Spearman correlation coefficients. 12](#_Toc75768366)

[Table S2.5: Removal of correlated descriptors for the full dataset.. 13](#_Toc75768367)

[Figure S2.6: Large molecules F1 plots from recursive feature elimination. 14](#_Toc75768368)

[Table S2.8: Confusion matrix for 10 rounds of cross validation for the large molecules cluster. 15](#_Toc75768369)

[Figure S2.9: Large molecules spearman correlation coefficients. 16](#_Toc75768370)

[Table S2.10: Removal of correlated descriptors for the large molecule cluster. 16](#_Toc75768371)

[Figure S2.11 Small molecule cluster F1 plots from recursive feature elimination.. 17](#_Toc75768372)

[Figure S2.12: Small molecules cluster ROC plot. 18](#_Toc75768373)

[Table S2.13: Confusion matrix for 10 rounds of cross validation for the small molecules cluster. 18](#_Toc75768374)

[Figure S2.14: Small molecules cluster Spearman correlation coefficients. 19](#_Toc75768375)

[Table S2.15: Removal of correlated descriptors in the small molecules cluster. The same methods were used as described for the entire dataset and the large molecules cluster. 19](#_Toc75768376)

[Table S2.16: Assigned preference for selecting descriptors with VIF < 0.5. 20](#_Toc75768377)

[Figure S3.1: Swarm plot for HOMO-LUMO gap, full dataset. Each molecule is indicated by a circle. In the full dataset the HOMO-LUMO gap data is similar in both positive and negative molecules. 22](#_Toc75768378)

[Figure S3.2: Swarm plot for HOMO-LUMO gap, large molecules cluster. 23](#_Toc75768379)

[Figure S3.3: Swarm plot for HOMO-LUMO gap, small molecules cluster.. 24](#_Toc75768380)

[Figure S3.4: Full dataset swarm plot for radial distribution function, 5.5 Å. 25](#_Toc75768381)

[Figure S3.5: Large molecules cluster swarm plot for radial distribution function, 5.5 Å, 26](#_Toc75768382)

[Figure S3.6: Small molecules swarm plot for radial distribution function 5.5 Å. 27](#_Toc75768383)

[Table S4.1: Static weighting schemes (table values). 28](#_Toc75768384)

[Table S4.2: Intrinsic state of common structures. 28](#_Toc75768385)

[Table S5.1: Atom electronegativity terms for the modified Burden matrix 29](#_Toc75768386)

[Figure S6.1: Representative mutagenic structures in the large molecules cluster. 31](#_Toc75768387)

[Figure S6.2: Representative non-mutagenic structures in the large molecules cluster. 32](#_Toc75768388)

[Figure S6.3: Representative mutagenic structures in the small molecules cluster. 33](#_Toc75768389)

[Figure S6.4: Representative non-mutagenic structures in the small molecules cluster. 34](#_Toc75768390)

[Figure S6.5: Other molecules (third cluster) representative structures. Red indicates a mutagen. 35](#_Toc75768391)

# S-1 Principal Component Analysis Based Clustering

Principle Component Analysis (PCA) is used to improve the clarity of the k-means clustering by focusing on the principal components that explain the largest amount of variance. The first 5 PCs explain about 60% of the variance, which is adequate for clustering in this application. K-means clustering is then performed on these 5 PCs. Selecting the appropriate number of clusters for the dataset is critical to obtaining accurate results using the regression within each cluster.

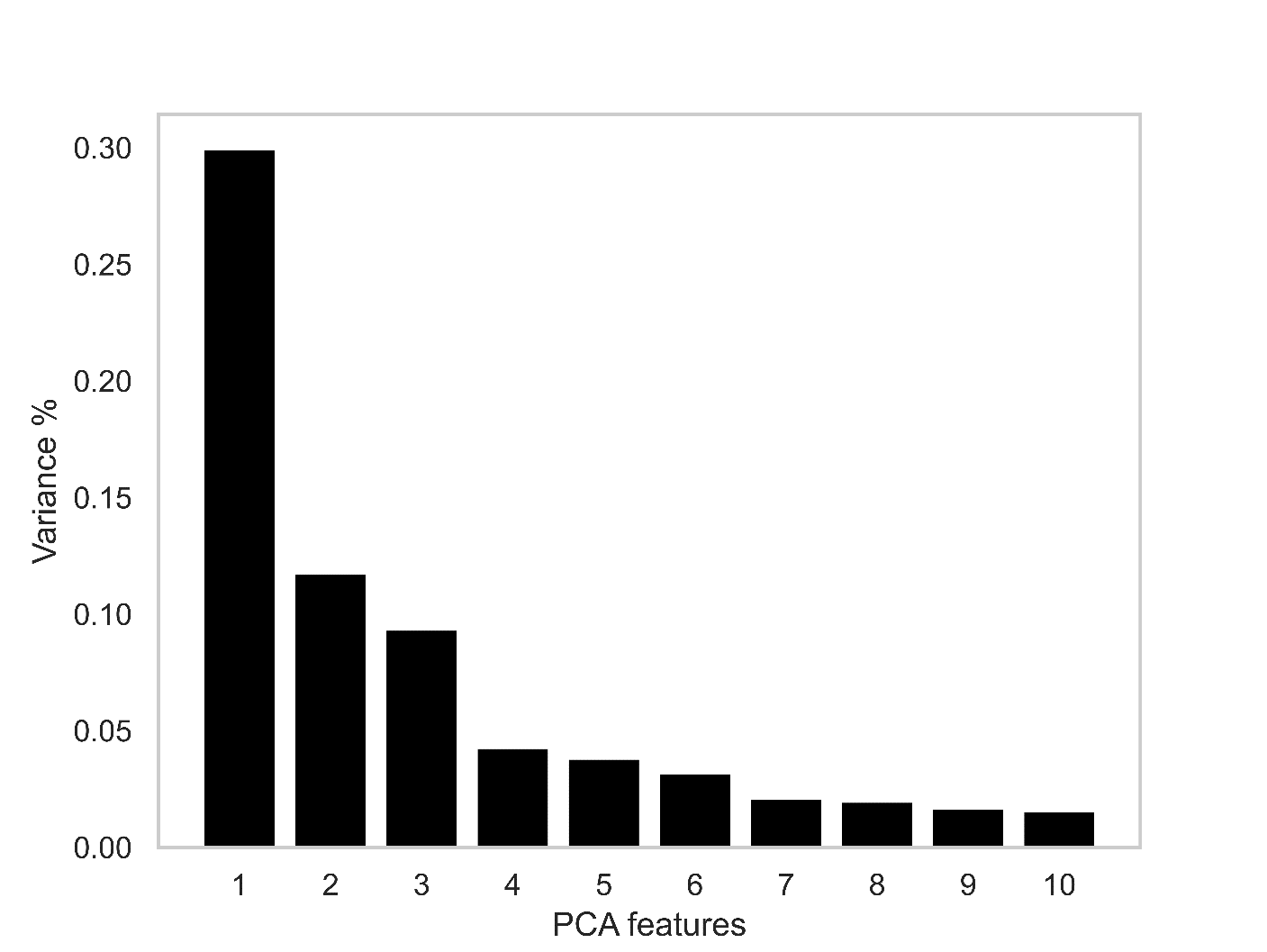


Figure S1.1: % Variance described by each principal component of the descriptors for the full test/train dataset of 557 molecules.

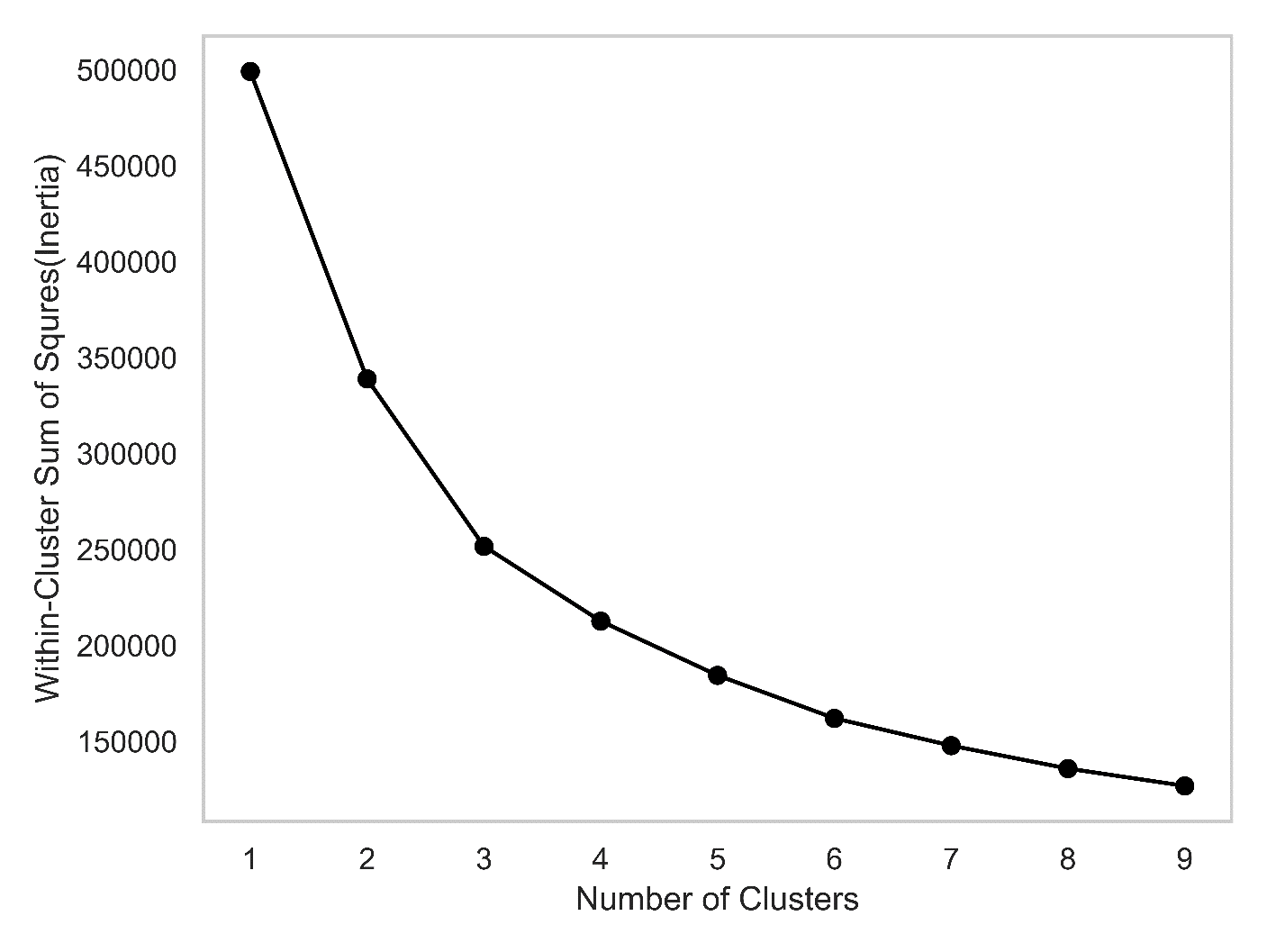


Figure S1.2: Elbow plot of within-cluster sum of squares (also termed inertia) for different numbers of clusters. Less improvement in the within-cluster sum of squares is seen above 3 clusters than below 3 clusters.

The elbow method1 is a basic method of evaluating how well each number of clusters describes the dataset. Adding more clusters will almost always reduce the within-cluster sum of squares, but after a certain point the reduction in the within cluster sum of squares score begins to level off. This point is referred to as the “elbow”.

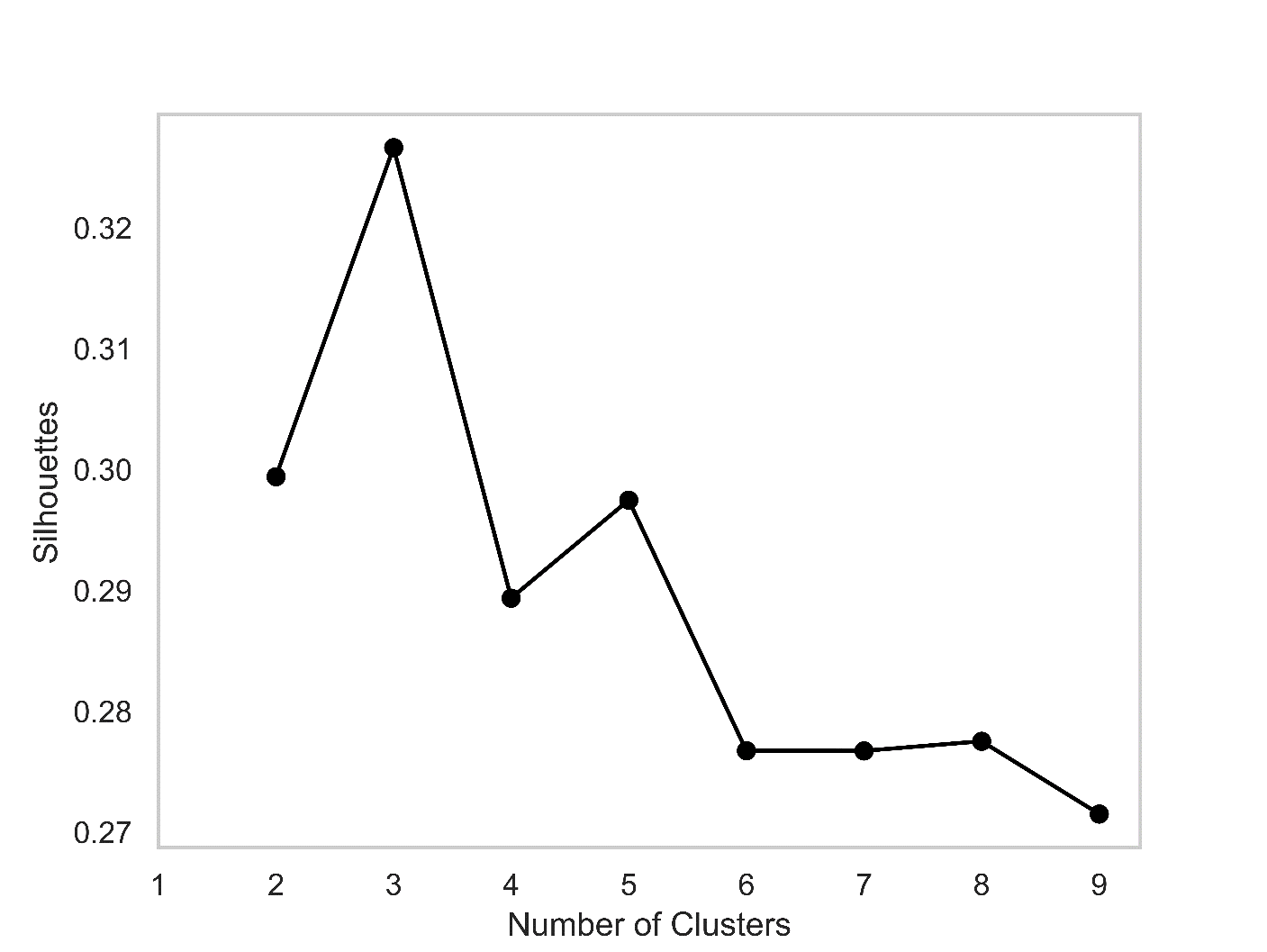


Figure S1.3: Silhouette Plot. Higher values indicate greater separation between the clusters, and therefore better clustering. The average silhouette score is maximized with 3 clusters.

The silhouette score2 (S3) is calculated as the difference between a point’s distance from other points within its own cluster and its distance from other clusters. The silhouette score requires at least 2 clusters to compute, so this plot begins with two clusters.

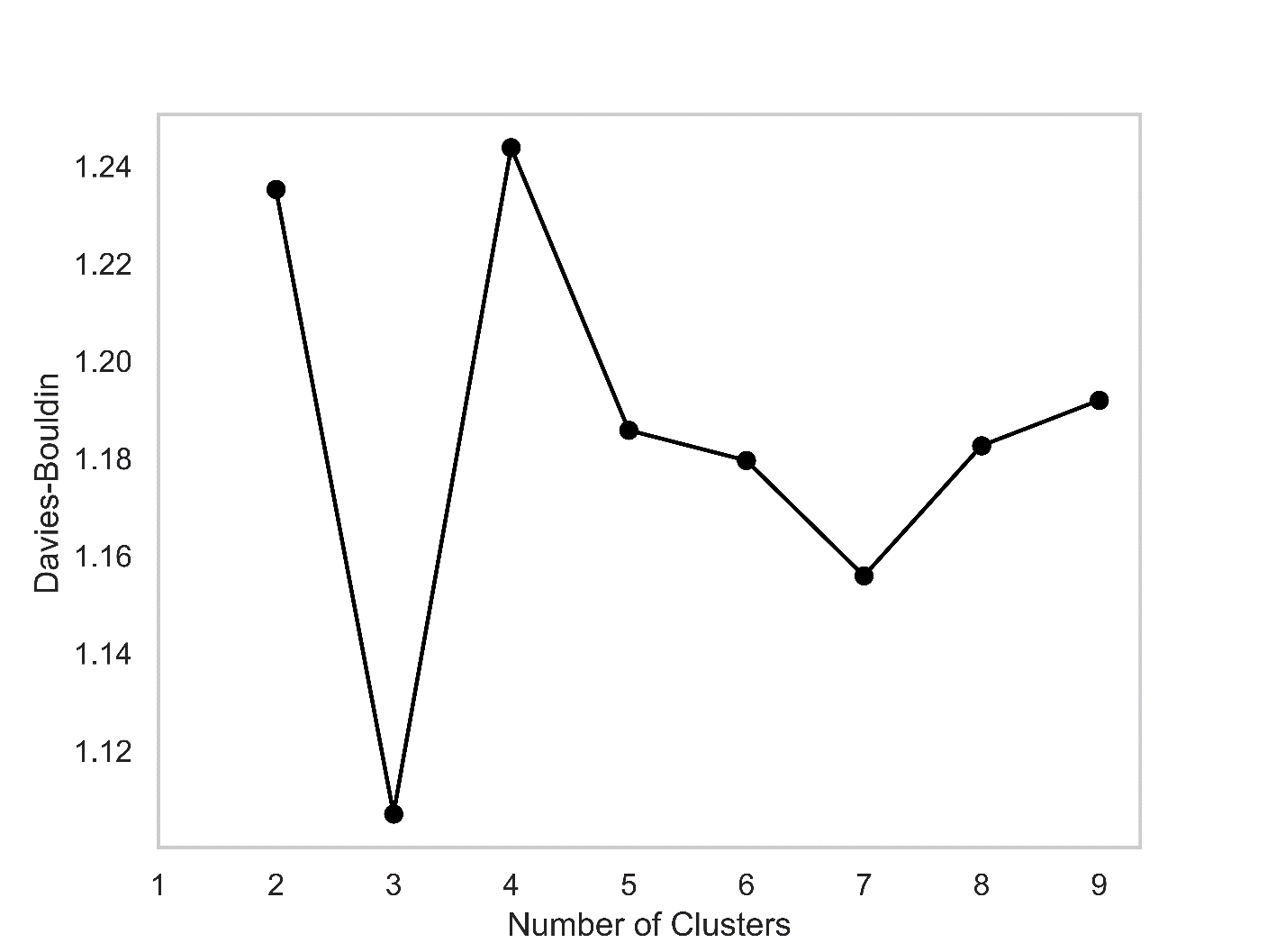


Figure S1.4: Davies-Bouldin Plot. The Davies-Bouldin plot3 is an additional method of measuring the quality of clustering, defined as the ratio between the scatter within the cluster and the cluster’s separation from other clusters. Lower values indicate better clustering.

All 3 measures of the quality of clustering indicate that 3 clusters is the best description of this dataset. The elbow plot begins to level off significantly at 3 clusters. The silhouette plot is maximized at 3 clusters. The Davies-Bouldin plot is minimized at 3 clusters.

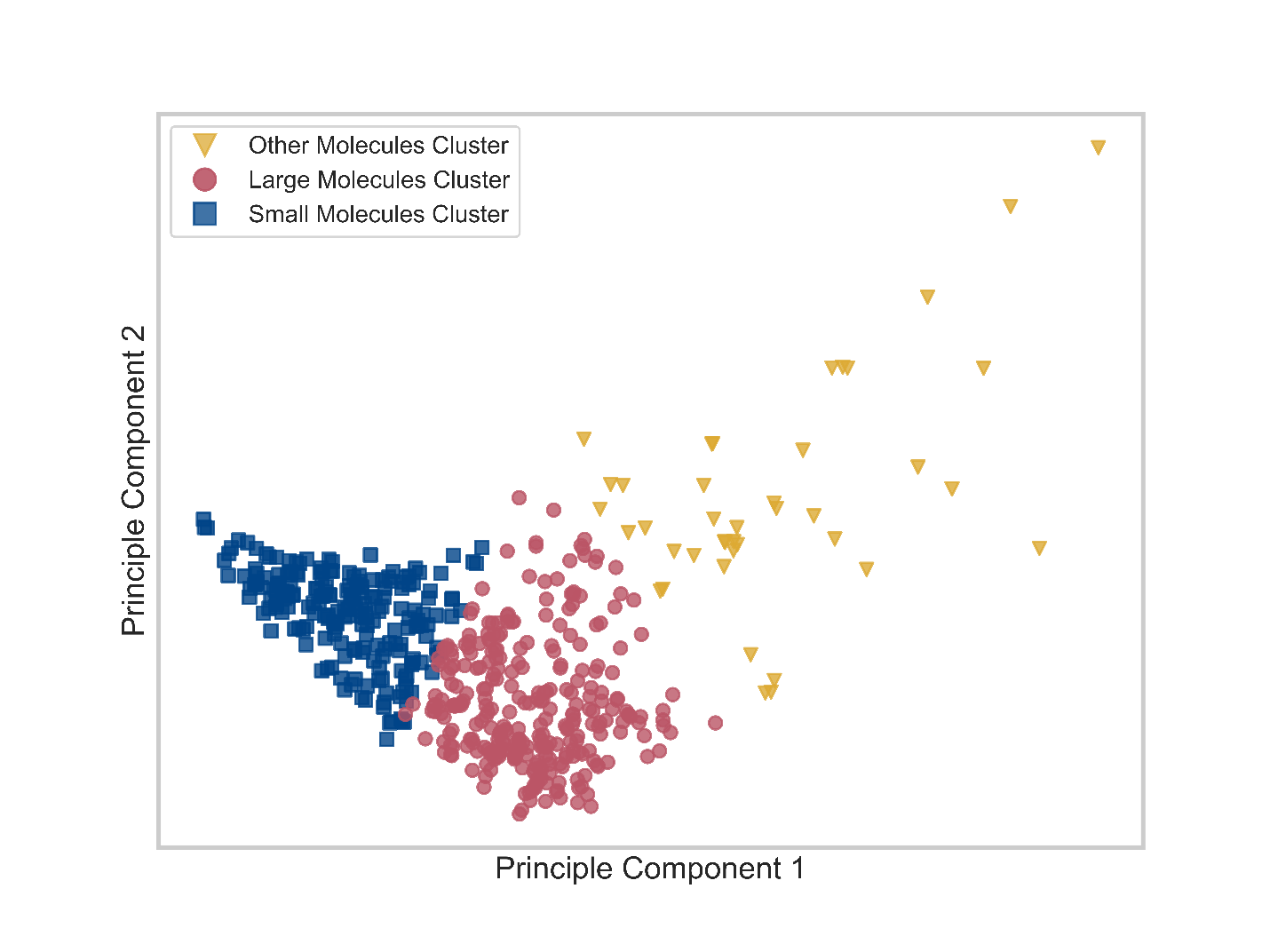


Figure S1.5: PC Plot of data used in this study. This is the same plot as Figure 2 in the main text, included here for context to show the results of the k-means clustering based off of three clusters.

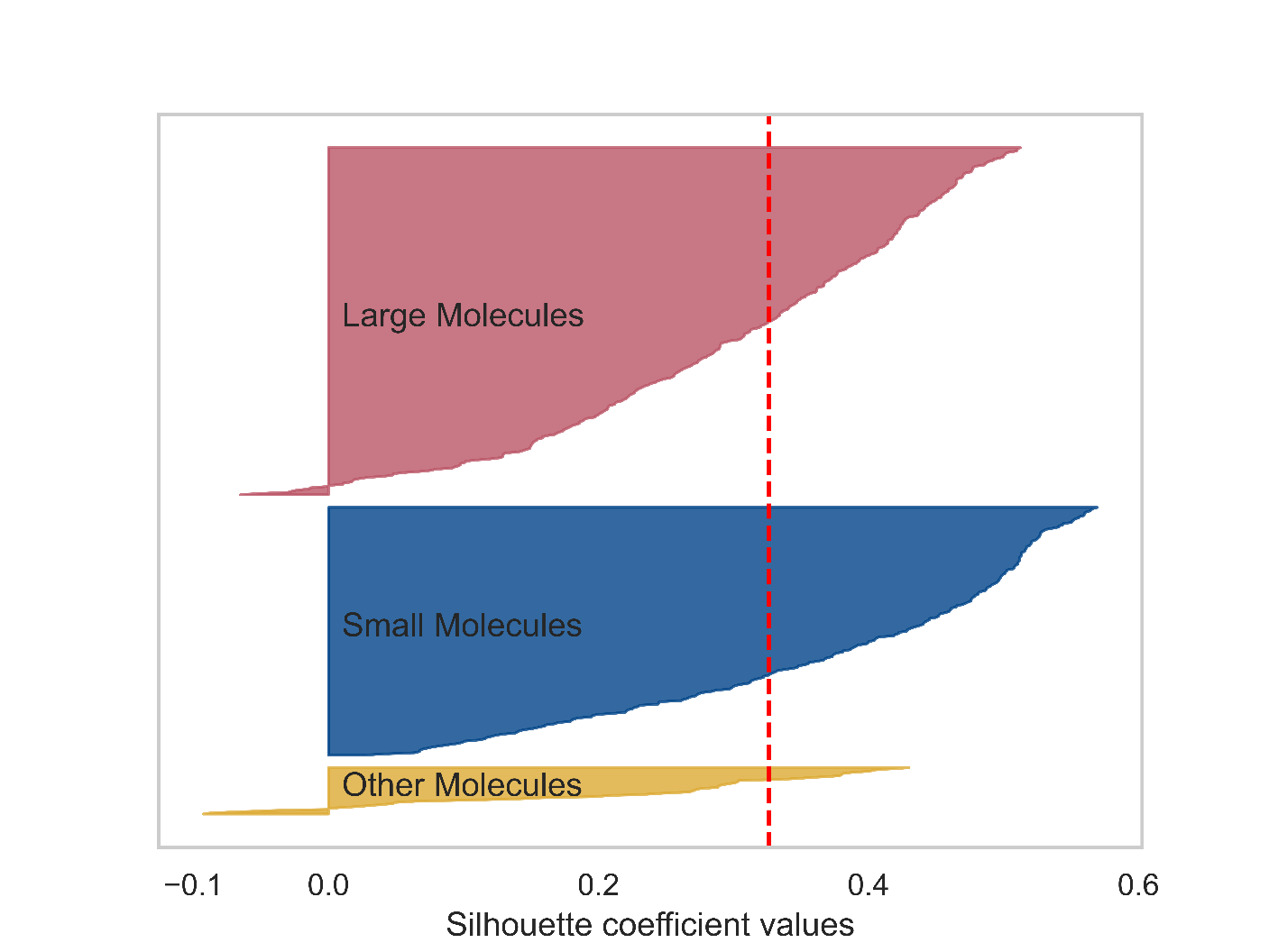


Figure S1.6: Silhouette values using three clusters. Each individual data point has a silhouette value. The dashed red line indicates the average silhouette score across all 3 clusters, and the distance horizontally of the shaded areas indicate the silhouette distance of each point within the cluster from the nearest neighboring cluster.

# S-2 Descriptor Selection and Regression

Because PaDEL can predict approximately 1500 descriptors, and not all are helpful or relevant, a method of reducing the dataset to only the most relevant descriptors for mutagenicity was needed. A two-stage method of selecting the most relevant features was used.

Sets of 10 to 90 descriptors were selected by recursive feature elimination and the weighted F1 score of each set was calculated for the entire dataset and also within each major cluster, at each of the different numbers of descriptors. This process was conducted with 10 iterations and a 2/3 to 1/3 test/train split, as with the final analysis. The average F1 scores were smoothing with a moving average window of 7 datapoints and the first derivative of this curve was taken to determine the point at which the F1 score stopped improving with more descriptors. The first time that the 1st derivative of the moving average was negative was used as the starting point for classification analysis, and all descriptors selected in at least 3 of the 10 iterations of recursive feature elimination were included.

## S-2.1 Initial Dataset

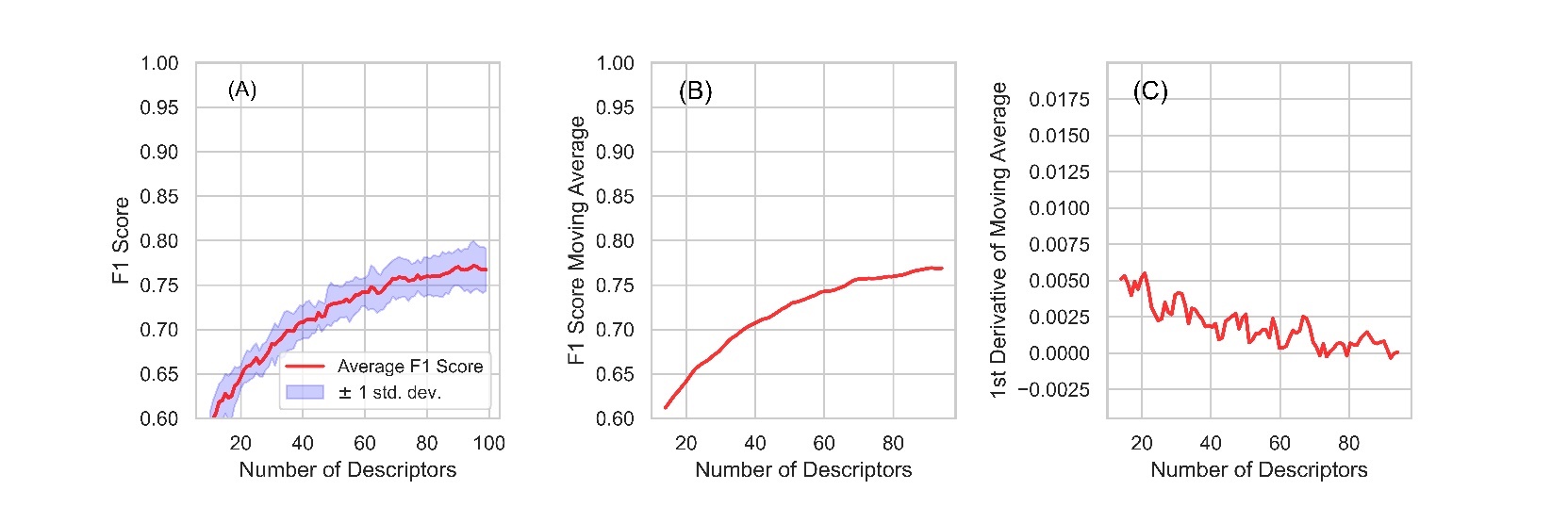


Figure S2.1: F1 plots from recursive feature elimination for the full dataset. (A) Average and standard deviation of the F1 score from 10 rounds of 3-fold cross validation of different numbers of descriptors, selected by recursive feature elimination. (B) 7-point moving average of the line in panel A. (C) The first derivative of the line in panel B. Based on the point at which the curve in C became negative, indicating that additional descriptors did not improve the F1 score, 72 descriptors were used as a starting point.

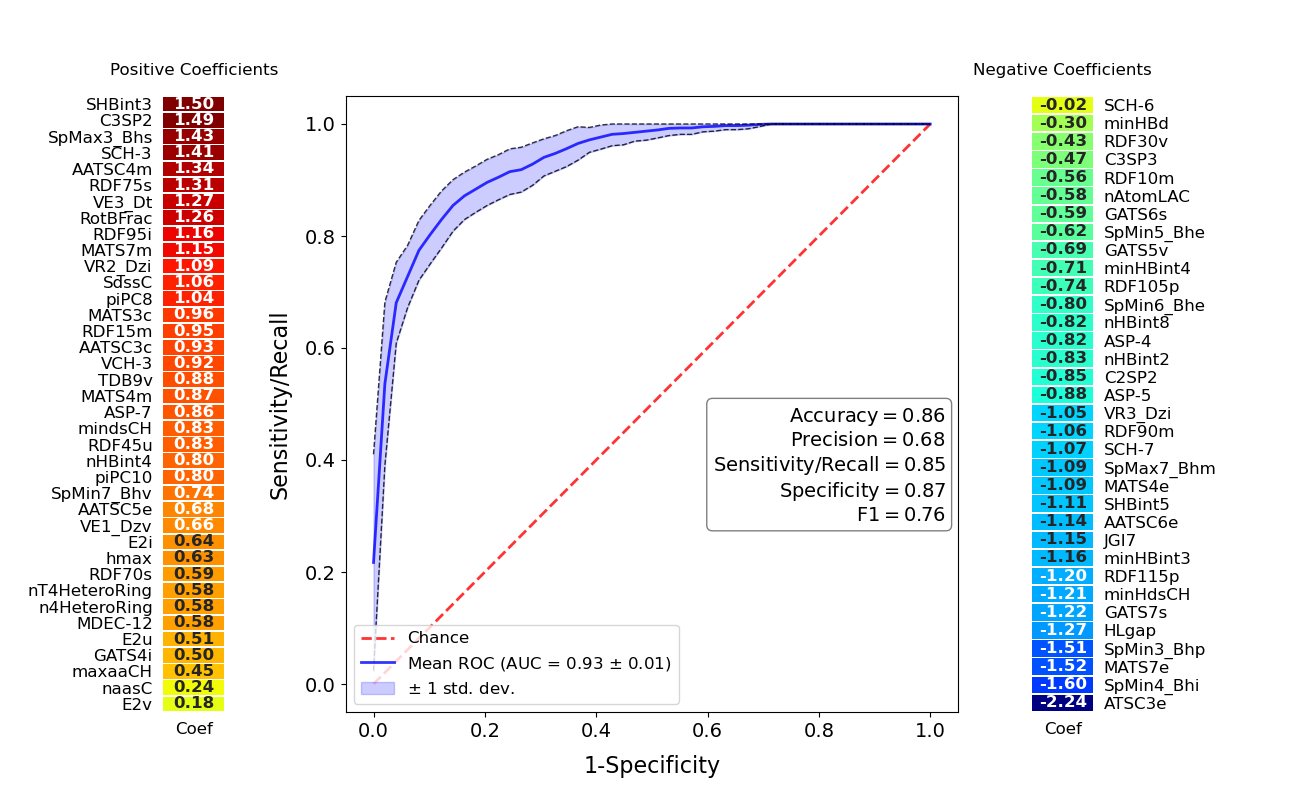


Figure S2.2: Initial dataset ROC plot. The results of Logistic Regression using 72 descriptors with 10 rounds of 3-fold cross validation. Descriptors with positive coefficients for mutagenicity are shown on the left of the plot and negative coefficients on the right. The receiver operating characteristic plot and classifier metrics are identical to those in the main text.

Table S2.3: Confusion matrix for 10 rounds of cross validation for the entire dataset with 72 descriptors

|  |  |  |
| --- | --- | --- |
|  | True Mutagens | True Non-Mutagens |
| Predicted Mutagens | (True Positives)  1197 | (False Positives)  560 |
| Predicted Non-Mutagens | (False Negatives)  213 | (True Negatives)  3600 |

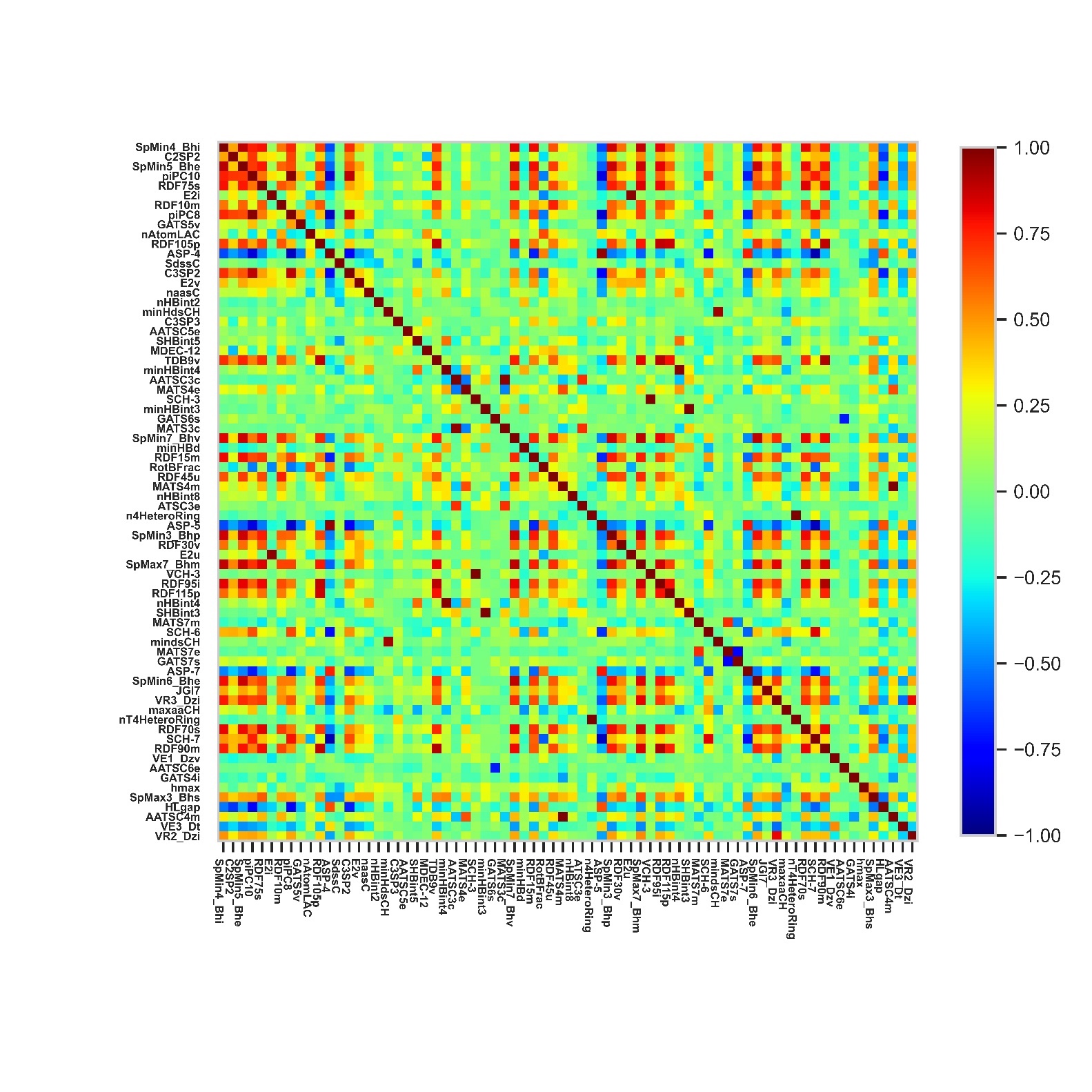


Figure S2.4: Initial dataset Spearman correlation coefficients. Red indicates a positive correlation. Blue indicates a negative correlation.

As shown in Figure S7.1, Recursive feature elimination selected the most relevant features and the F1 score was used to inform how many features to start with. However, several highly correlated descriptors were selected by this method, (Fig S7.3) so the variance inflation factor (VIF) was used to eliminate the highly correlated descriptors. The descriptor with the highest VIF was eliminated and the VIF was recalculated until all descriptors had a VIF less than 5. If the VIF of the highest VIF descriptor and the second highest were within 0.5 of each other, an assigned priority was used. The assigned priority was chosen to select descriptors that were simpler to interpret.

A VIF of less than 5 shows that redundant descriptors have been removed to a level acceptable for our purposes. In cases where the second highest VIF was within 0.5 of the highest VIF, the priority assigned in Table S10 was used. This allowed us to select descriptors that were easier to interpret from a mechanistic perspective over less comprehensible descriptors in cases where the impact was negligible.

Table S2.5: Removal of correlated descriptors for the full dataset. The VIF for each descriptor was calculated and then the highest VIF is removed and the VIF is calculated again until all descriptors have a VIF of less than 5.

|  |  |  |
| --- | --- | --- |
| **Difference in VIF Between Descriptors with the Two Highest VIF Values** | **Descriptor Kept** | **Descriptor Removed** |
| Perfect correlation | n4HeteroRing | n4THeteroRing |
| 0.526 | AATSC4m | MATS4m |
| 0.799 | E2i | E2u |
| 2.150 | SCH-3 | VCH-3 |
| 8.944 | SCH-6 | SCH-7 |
| 6.188 | SpMax7\_Bhm | ASP-4 |
| 3.873 | piPC10 | SpMax7\_Bhm |
| 0.923 | RotBFrac | piPC10 |
| 0.651 | SpMin7\_Bhv | RotBFrac |
| 0.464 | RDF105p | SpMin7\_Bhv |
| 2.178 | SpMin4\_Bhi | RDF105p |
| 0.158 | piPC8 | SpMin4\_Bhi |
| 0.706 | SpMin5\_Bhe | piPC8 |
| 0.067 | AATSC3c | SpMin5\_Bhe |
| 0.193 | RDF95i | AATSC3c |
| 0.199 | RDF95i | SpMax3\_Bhs |
| 0.449 | C3SP2 | RDF95i |
| 1.163 | VR3\_Dzi | C3SP2 |
| 0.130 | TDB9v | VR3\_Dzi |
| 0.539 | RDF45u | TDB9v |
| 0.074 | RDF15m | RDF45u |
| 0.061 | mindsCH | RDF15m |
| 0.094 | mindsCH | AATSC4m |
| 0.228 | minHdsCH | mindsCH |

Eliminating the descriptors as described and re-computing the ROC plot yields Figure 2 in the main text. The same procedure was conducted within each cluster.

## S-2.2 Large Molecules Cluster

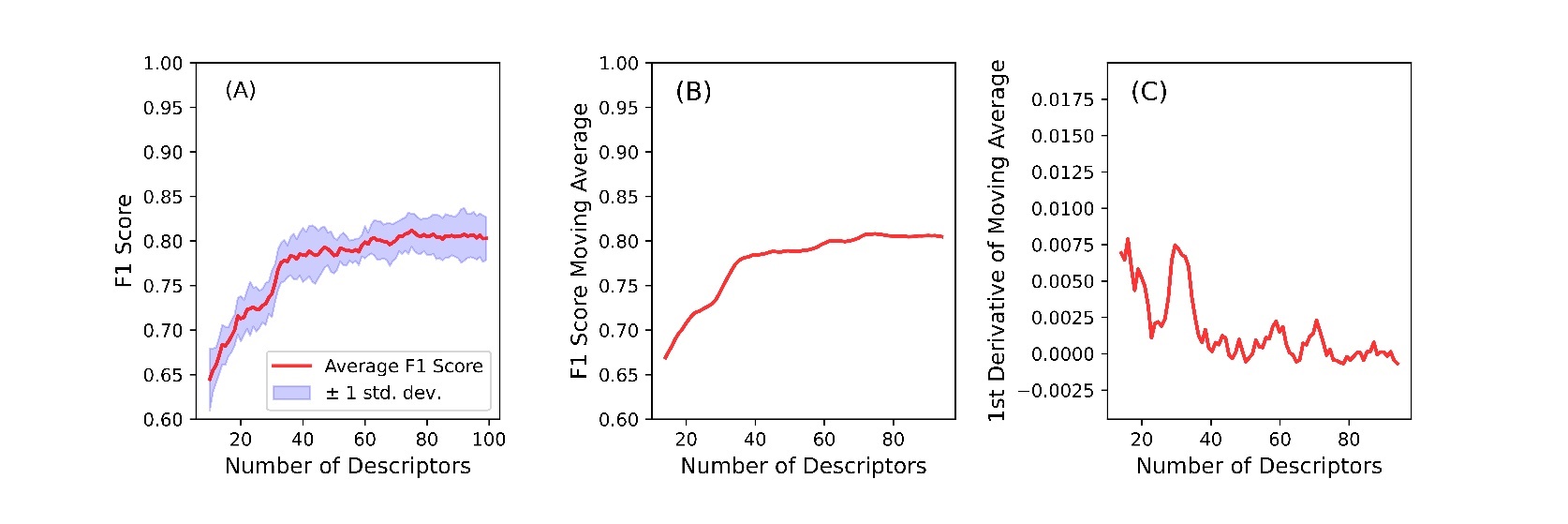


Figure S2.6: Large molecules F1 plots from recursive feature elimination. (A) Average and standard deviation of the F1 score from 10 rounds of 3-fold cross validation of different numbers of descriptors, selected by recursive feature elimination. (B) 7-point moving average of the line in panel A. (C) The first derivative of the line in panel B. Based on the point at which the curve in C became negative, indicating that additional descriptors did not improve the F1 score, 45 descriptors were used as a starting point. 46 descriptors were ultimately used in order to consistently select descriptors that were found by at least 3 of the 10 rounds of analysis.

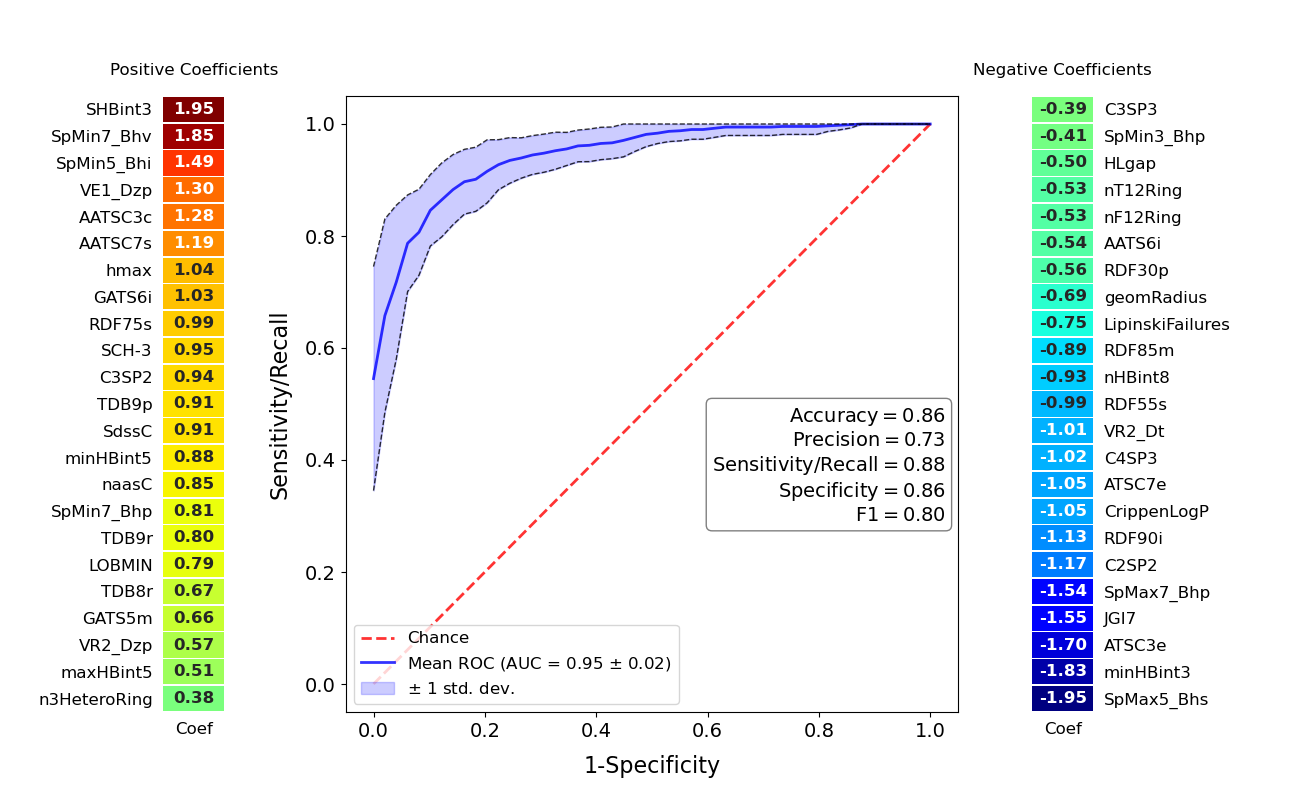


Figure S2.7: Large molecules cluster ROC plot. The results of Logistic Regression using all 46 descriptors with 10 rounds of 3-fold cross validation. Descriptors with positive coefficients for mutagenicity are shown on the left of the plot and negative coefficients on the right. The receiver operating characteristic plot and classifier metrics are identical to those in the main text.

Table S2.8: Confusion matrix for 10 rounds of cross validation for the large molecules cluster.

|  |  |  |
| --- | --- | --- |
|  | True Mutagens | True Non-Mutagens |
| Predicted Mutagens | (True Positives)  806 | (False Positives)  298 |
| Predicted Non-Mutagens | (False Negatives)  114 | (True Negatives)  1792 |

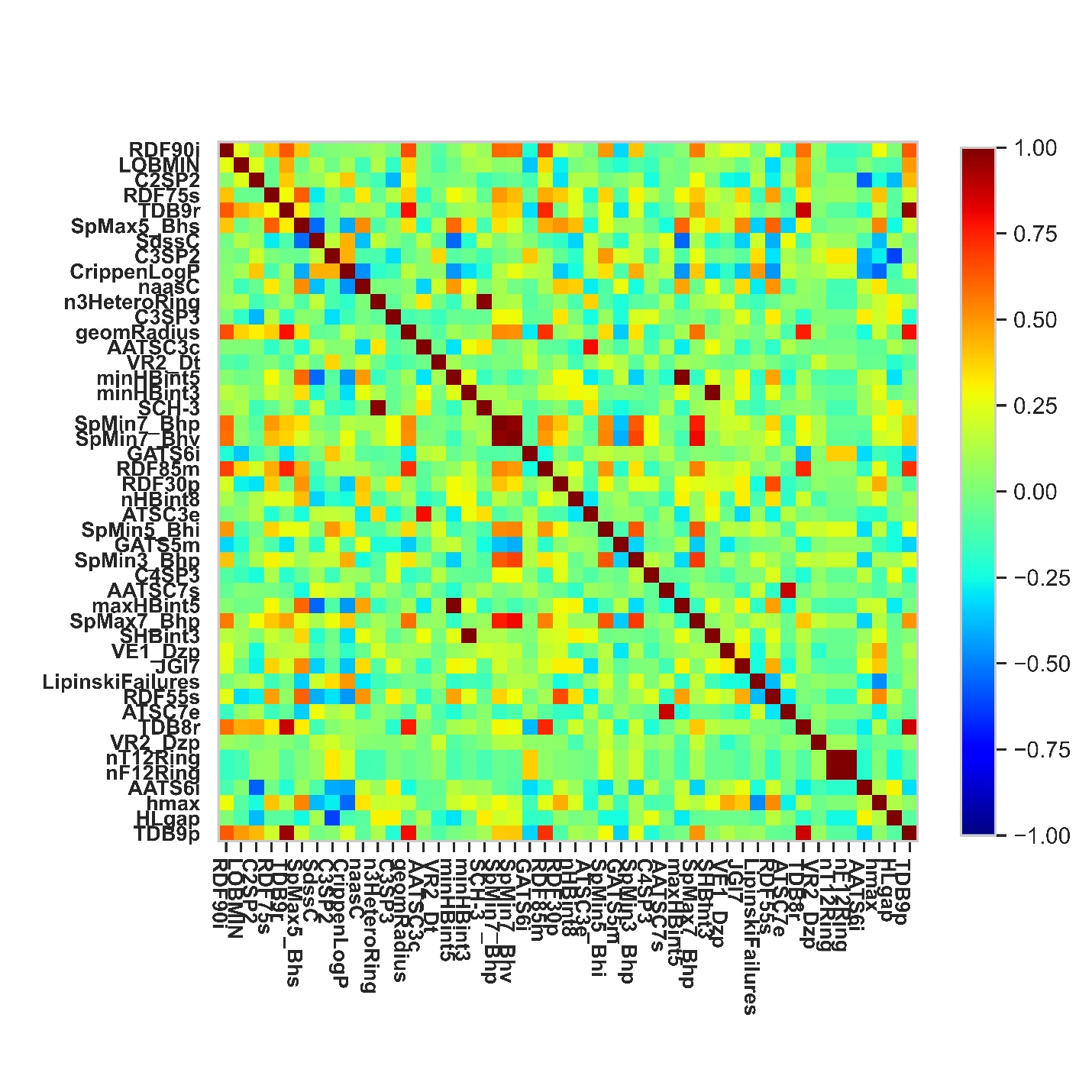
****

Figure S2.9: Large molecules spearman correlation coefficients.

Table S2.10: Removal of correlated descriptors for the large molecule cluster. The VIF for each descriptor was calculated and then the highest VIF is removed and the VIF is calculated again until all descriptors have a VIF of less than 5.

|  |  |  |
| --- | --- | --- |
| **Difference in VIF between the two highest VIF variables** | **Descriptor Kept** | **Descriptor Removed** |
| Perfect correlation | nT12Ring | nF12Ring |
| 10.045 | SpMin7\_Bhp | SpMin7\_Bhv |
| 0.465 | n3HeteroRing | SCH-3 |
| 1.689 | TDB9p | maxHBint5 |
| 1.923 | TDB9r | TDB9p |
| 0.777 | TDB8r | TDB9r |
| 0.303 | SpMin7\_Bhp | SpMax5\_Bhs |
| 0.002 | SpMin7\_Bhp | SpMax7\_Bhp |
| 0.687 | ATSC7e | AATSC7s |
| 0.0478 | SHBint3 | C2SP2 |
| 0.539 | SpMin5\_Bhi | SHBint3 |

## S-2.3 Small Molecules

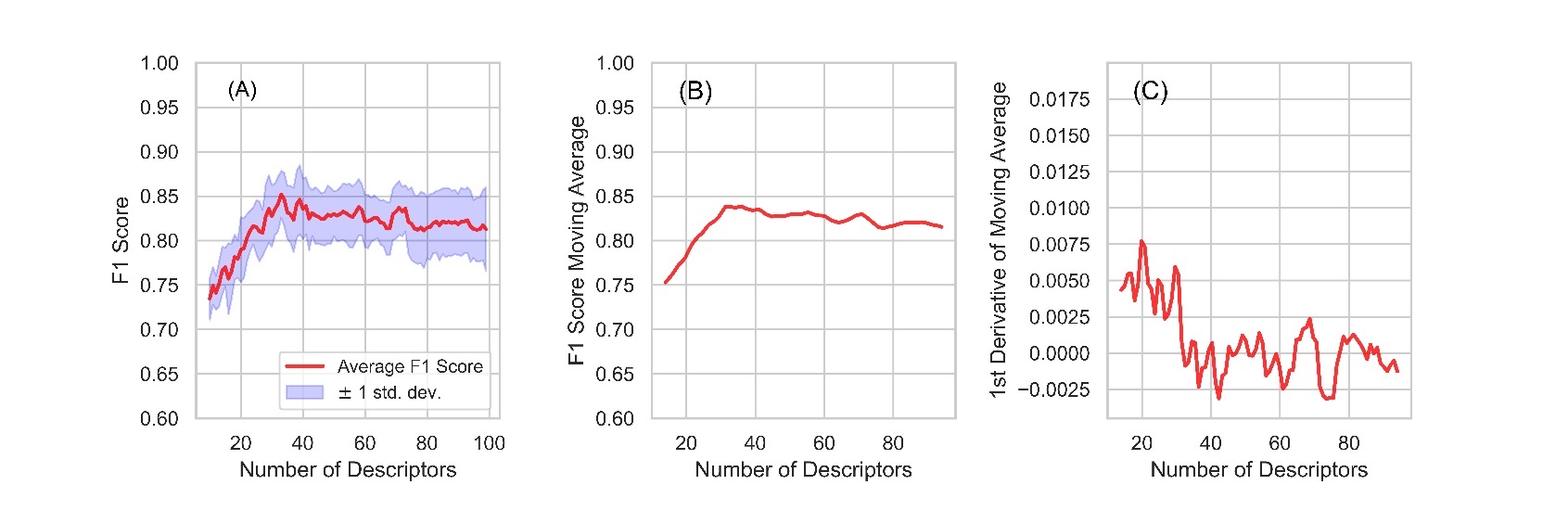


Figure S2.11 Small molecule cluster F1 plots from recursive feature elimination. (A) Average and standard deviation of the F1 score from 10 rounds of 3-fold cross validation of different numbers of descriptors, selected by recursive feature elimination. (B) 7-point moving average of the line in panel A. (C) The first derivative of the line in panel B. Based on the point at which the curve in C became negative, indicating that additional descriptors did not improve the F1 score, 32 descriptors were used as a starting point. 33 descriptors were ultimately used in order to consistently select descriptors that were found by at least 3 of the 10 rounds of analysis.

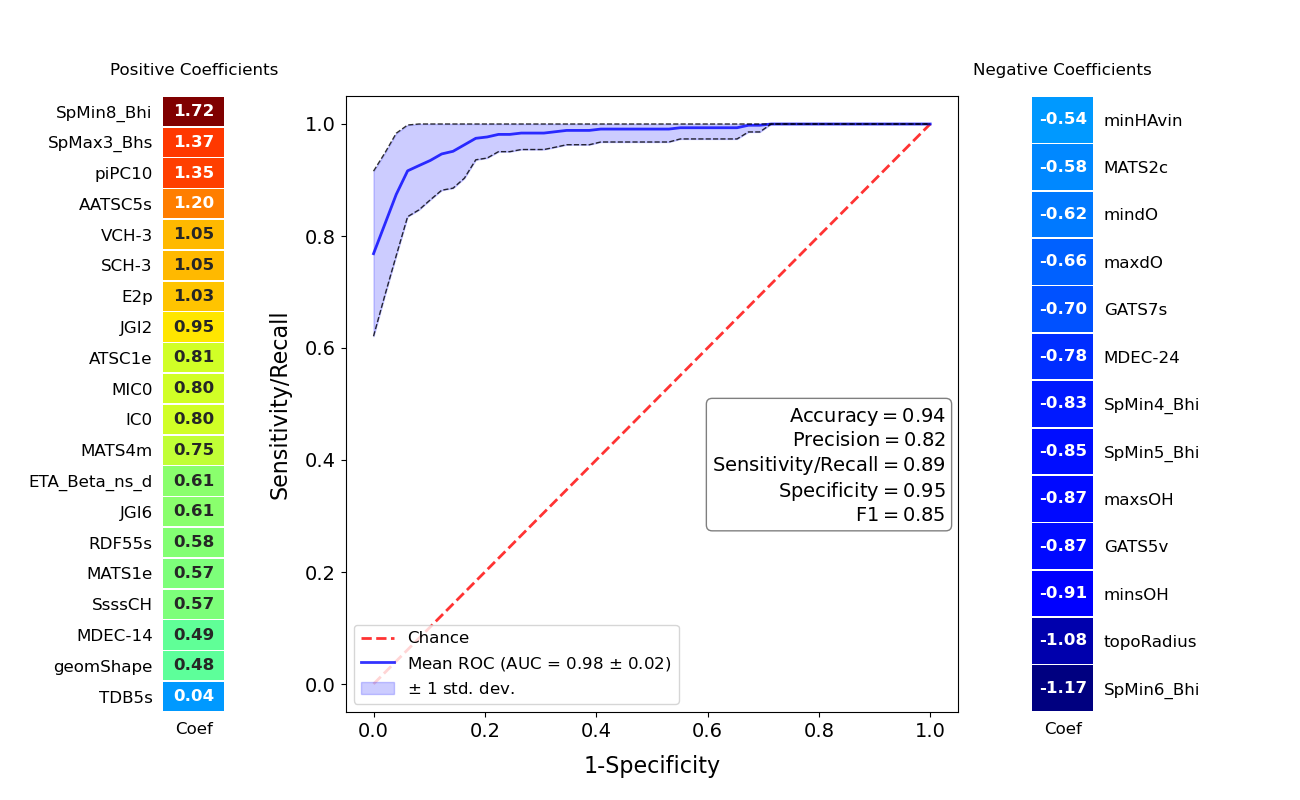


Figure S2.12: Small molecules cluster ROC plot. The results of Logistic Regression using all 33 descriptors with 10 rounds of 3-fold cross validation. Descriptors with positive coefficients for mutagenicity are shown on the left of the plot and negative coefficients on the right. The receiver operating characteristic plot and classifier metrics are identical to those in the main text.

Table S2.13: Confusion matrix for 10 rounds of cross validation for the small molecules cluster.

|  |  |  |
| --- | --- | --- |
|  | True Mutagens | True Non-Mutagens |
| Predicted Mutagens | (True Positives)  382 | (False Positives)  89 |
| Predicted Non-Mutagens | (False Negatives)  48 | (True Negatives)  1631 |

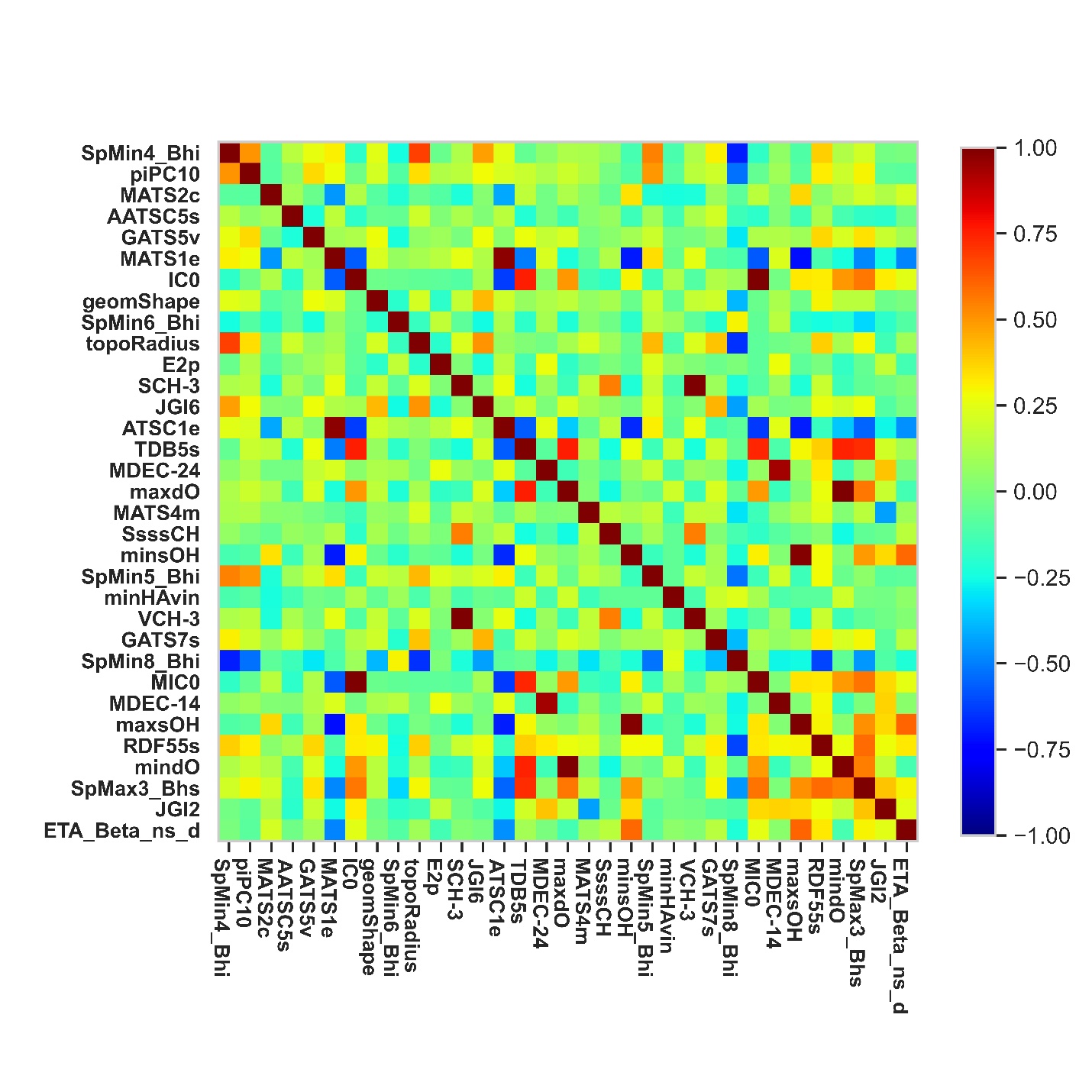


Figure S2.14: Small molecules cluster Spearman correlation coefficients.

Table S2.15: Removal of correlated descriptors in the small molecules cluster. The same methods were used as described for the entire dataset and the large molecules cluster.

|  |  |  |
| --- | --- | --- |
| **Difference in VIF between the two highest VIF variables** | **Descriptor Kept** | **Descriptor Removed** |
| Perfect correlation | SCH-3 | VCH-3 |
| 95.583 | mindO | maxdO |
| 34.703 | minsOH | maxsOH |
| 1.544 | IC0 | MIC0 |
| 1.357 | ATSC1e | MATS1e |
| 1.608 | TDB5s | SpMax3\_Bhs |
| 0.395 | TDB5s | SpMin8\_Bhi |
| 1.553 | ATSC1e | TDB5s |

## S-2.4 Descriptor Priority for Closely Correlated Descriptors

Table S2.16: Assigned preference for selecting descriptors with VIF < 0.5. Abbreviations for each type of descriptor are in accompanying excel document.

|  |  |
| --- | --- |
| **Descriptor Type** | **Rank** |
| Atom type electrotopological state | 1 |
| Molecular distance edge | 2 |
| Walk counts | 3 |
| Acidic group count | 4 |
| Basic group count | 5 |
| Largest chain | 6 |
| Carbon types | 7 |
| Weight | 8 |
| Aromatic atoms count | 9 |
| Aromatic bonds count | 10 |
| Atom count | 11 |
| Bond count | 12 |
| Hbond acceptor count | 13 |
| Hbond donor count | 14 |
| Ring count | 15 |
| Rotatable bonds count | 16 |
| Crippen LogP and MR | 17 |
| ALOGP | 18 |
| APol | 19 |
| Mannhold LogP | 20 |
| XLogP | 21 |
| Chi chain 1 | 22 |
| Chi chain 2 | 23 |
| Chi cluster | 24 |
| Chi path cluster | 25 |
| Chi path | 26 |
| RDF | 27 |
| Path counts | 28 |
| Longest aliphatic chain | 29 |
| Extended topochemical atom | 30 |
| Charged partial surface area | 31 |
| Information content | 32 |
| Zagreb index | 33 |
| Gravitational index | 34 |
| Eccentric connectivity index | 35 |
| Vertex adjacency information (magnitude) | 36 |
| Hybridization ratio | 37 |
| Weighted path | 38 |
| Wiener numbers | 39 |
| Kappa shape indices | 40 |
| Length over breadth | 41 |
| Moment of inertia | 42 |
| Petitjean shape index | 43 |
| BCUT | 44 |
| BPol | 45 |
| Constitutional | 46 |
| FMFDescriptor | 47 |
| Fragment complexity | 48 |
| Largest Pi system | 49 |
| McGowan volume | 50 |
| Molecular linear free energy relation | 51 |
| Petitjean number | 52 |
| Rule of five | 53 |
| Topological | 54 |
| Topological charge | 55 |
| Topological distance matrix | 56 |
| Topological polar surface area | 57 |
| Van der Waals volume | 58 |
| 3D autocorrelation | 59 |
| Autocorrelation | 60 |
| Barysz matrix | 61 |
| Detour matrix | 62 |
| Burden modified eigenvalues | 63 |
| WHIM | 64 |
| Gaussian Descriptors | 65 |

# S-3 Swarm Plots of Differences Between Clusters

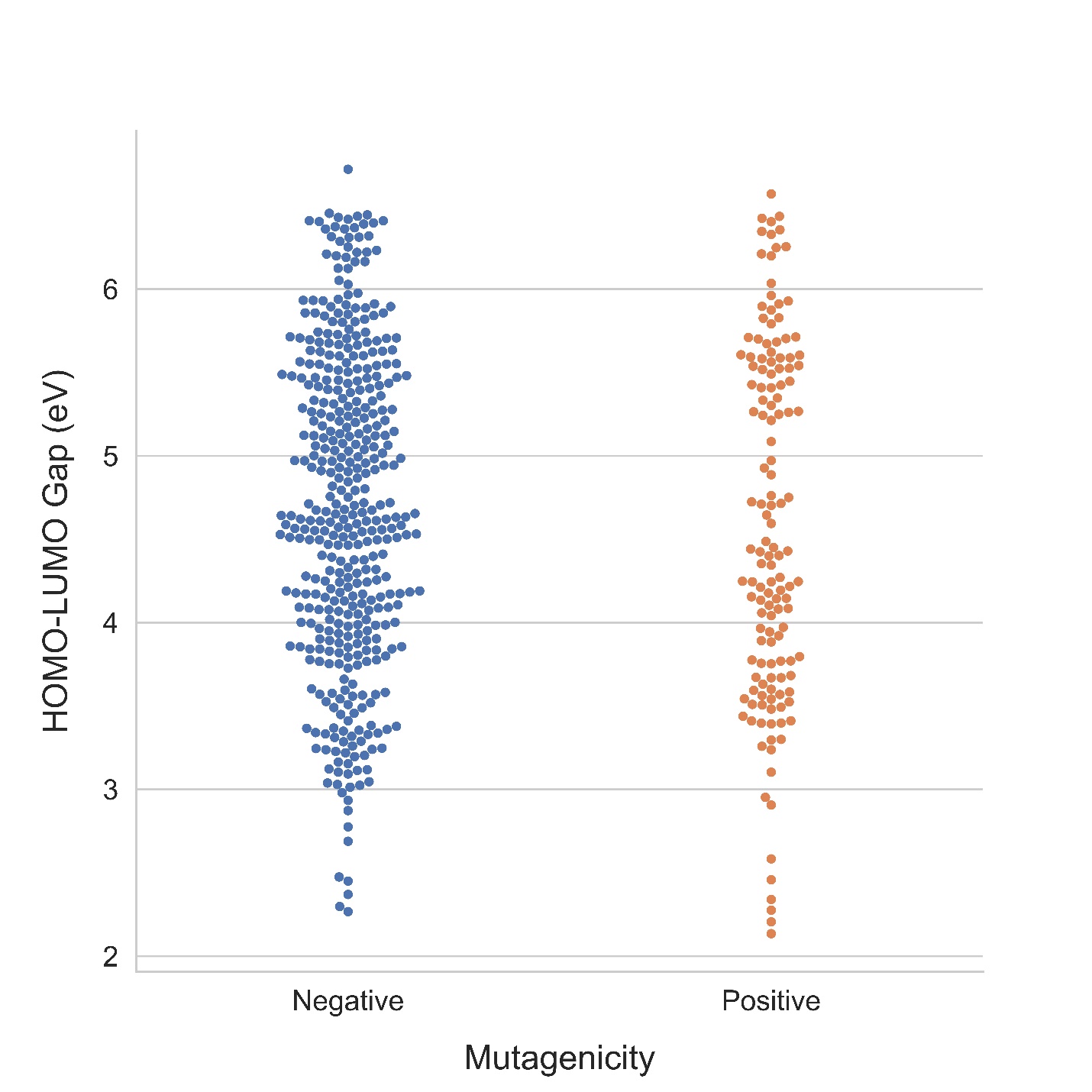


Figure S3.1: Swarm plot for HOMO-LUMO gap, full dataset. Each molecule is indicated by a circle. In the full dataset the HOMO-LUMO gap data is similar in both positive and negative molecules.

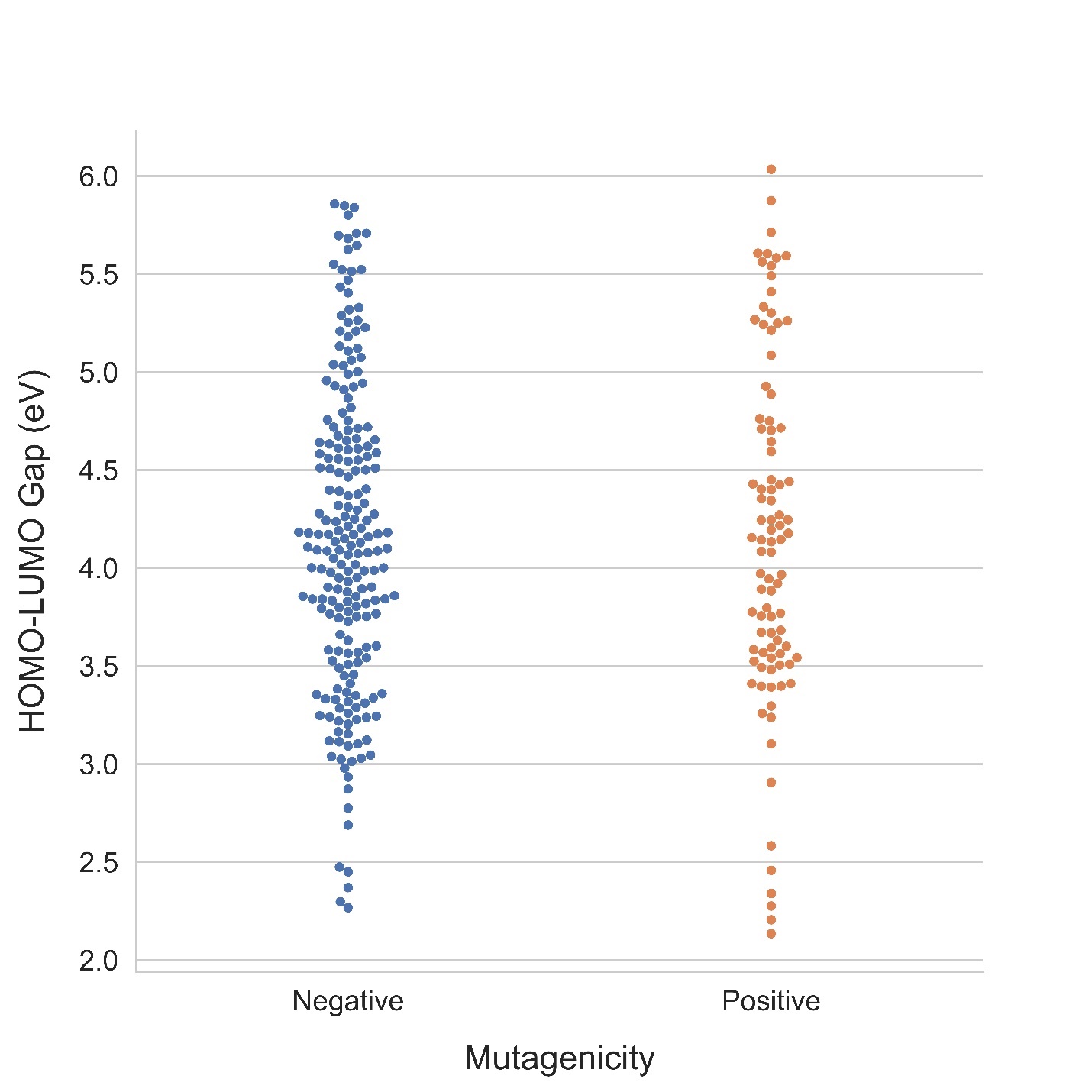


Figure S3.2: Swarm plot for HOMO-LUMO gap, large molecules cluster. In the large molecules cluster, there are more mutagenic (positive) molecules with smaller HOMO-LUMO gaps, particularly between 3 eV and 4.5 eV, whereas the negative molecules have a greater proportion of molecules with values above 5 eV.

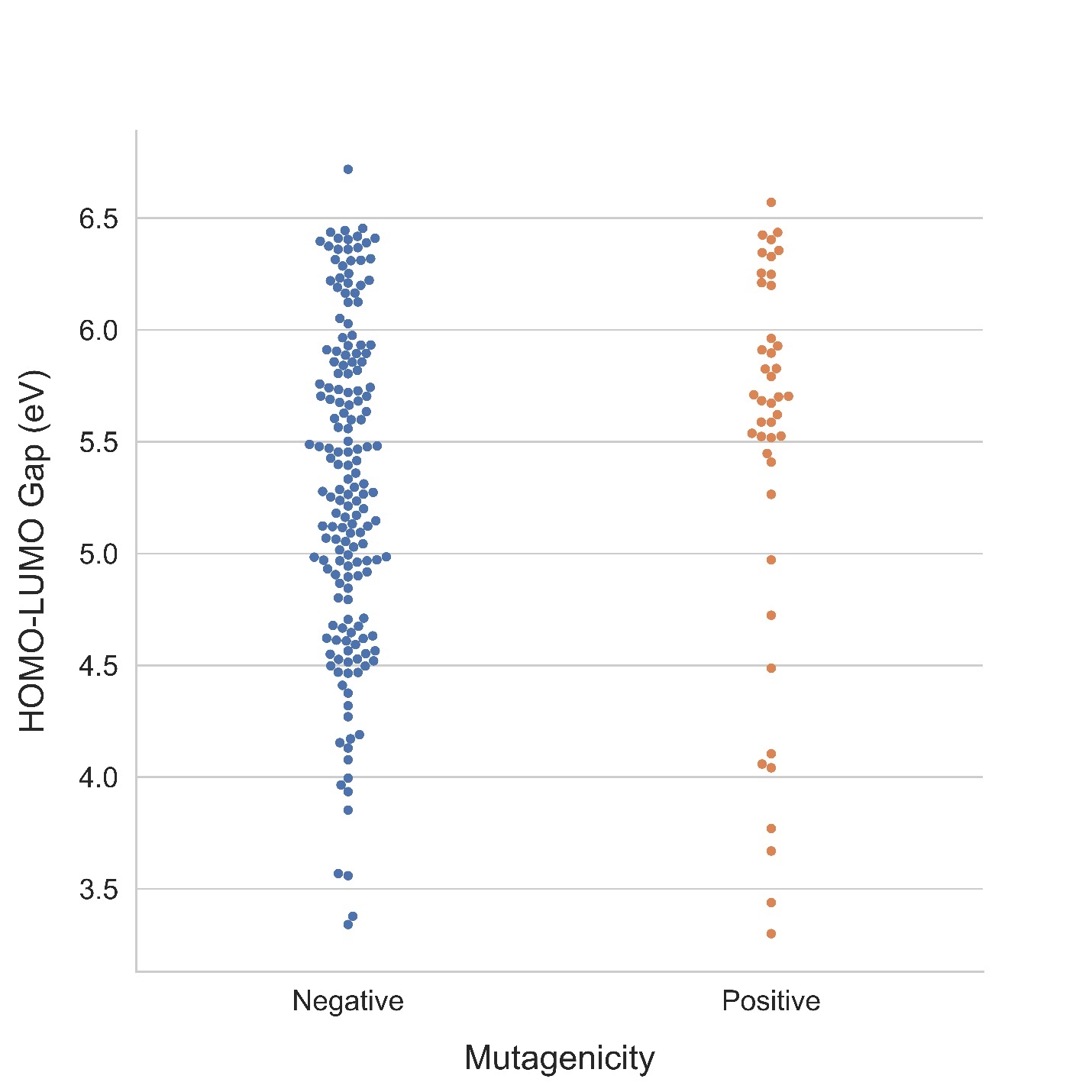


Figure S3.3: Swarm plot for HOMO-LUMO gap, small molecules cluster. In the small molecules cluster, many mutagens have HOMO-LUMO gaps between 5.0 and 6.5 eV.

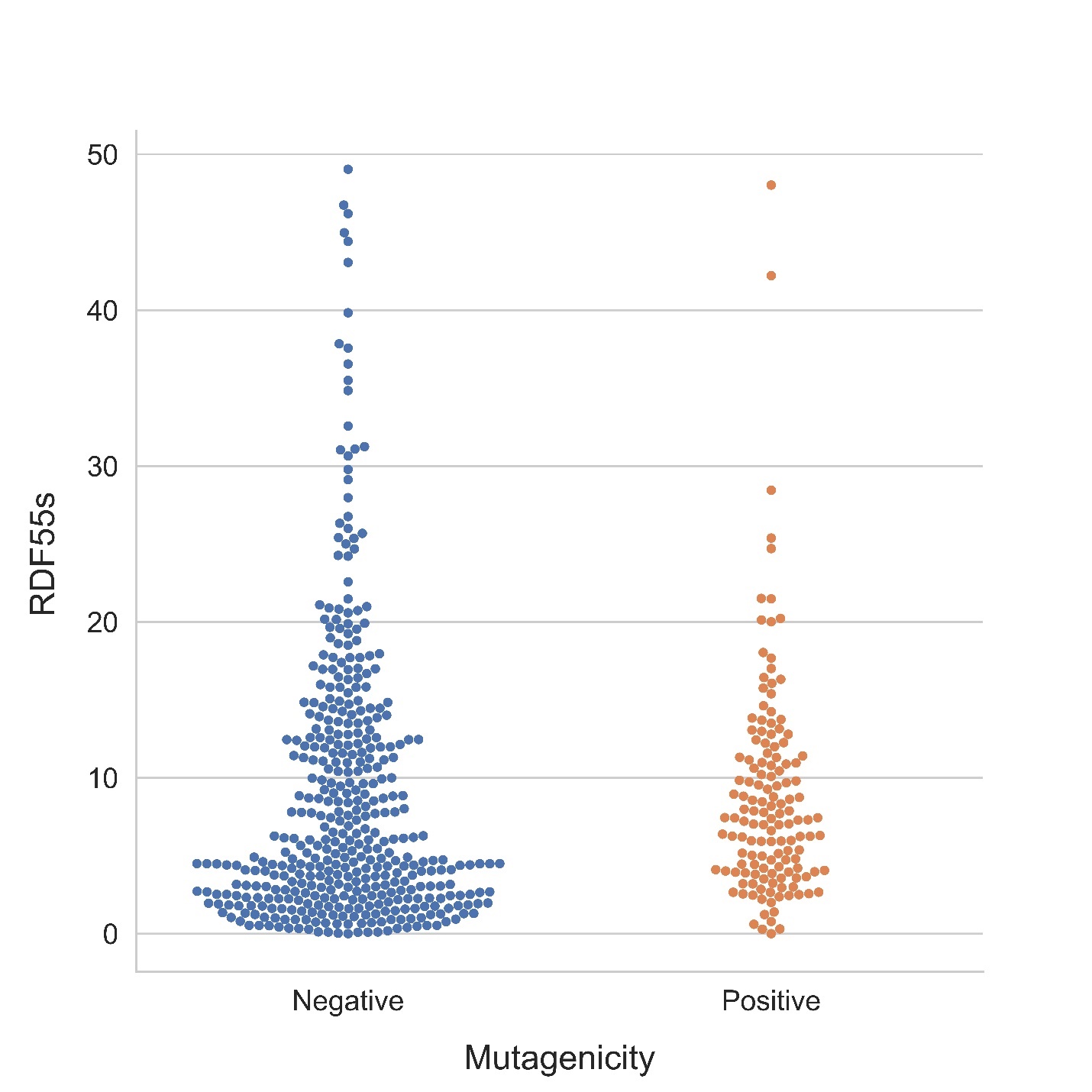


Figure S3.4: Full dataset swarm plot for radial distribution function, 5.5 Å, weighted by intrinsic state (RDF55s). In the full dataset, the RDF55s data distribution is similar in both positive and negative molecules.

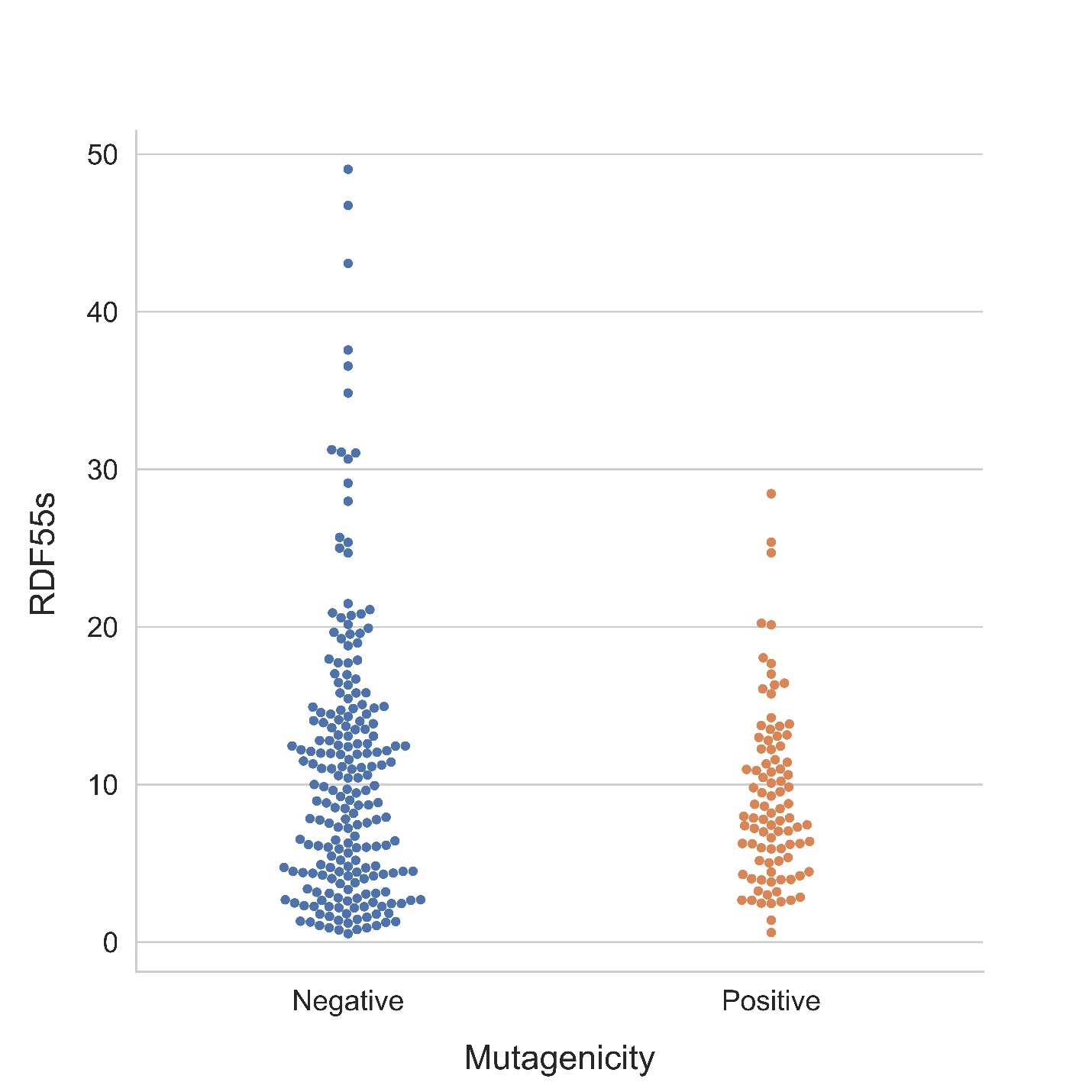


Figure S3.5: Large molecules cluster swarm plot for radial distribution function, 5.5 Å, weighted by intrinsic state (RDF55s). The RDF55s values of the positive molecules are slightly shifted towards smaller values.

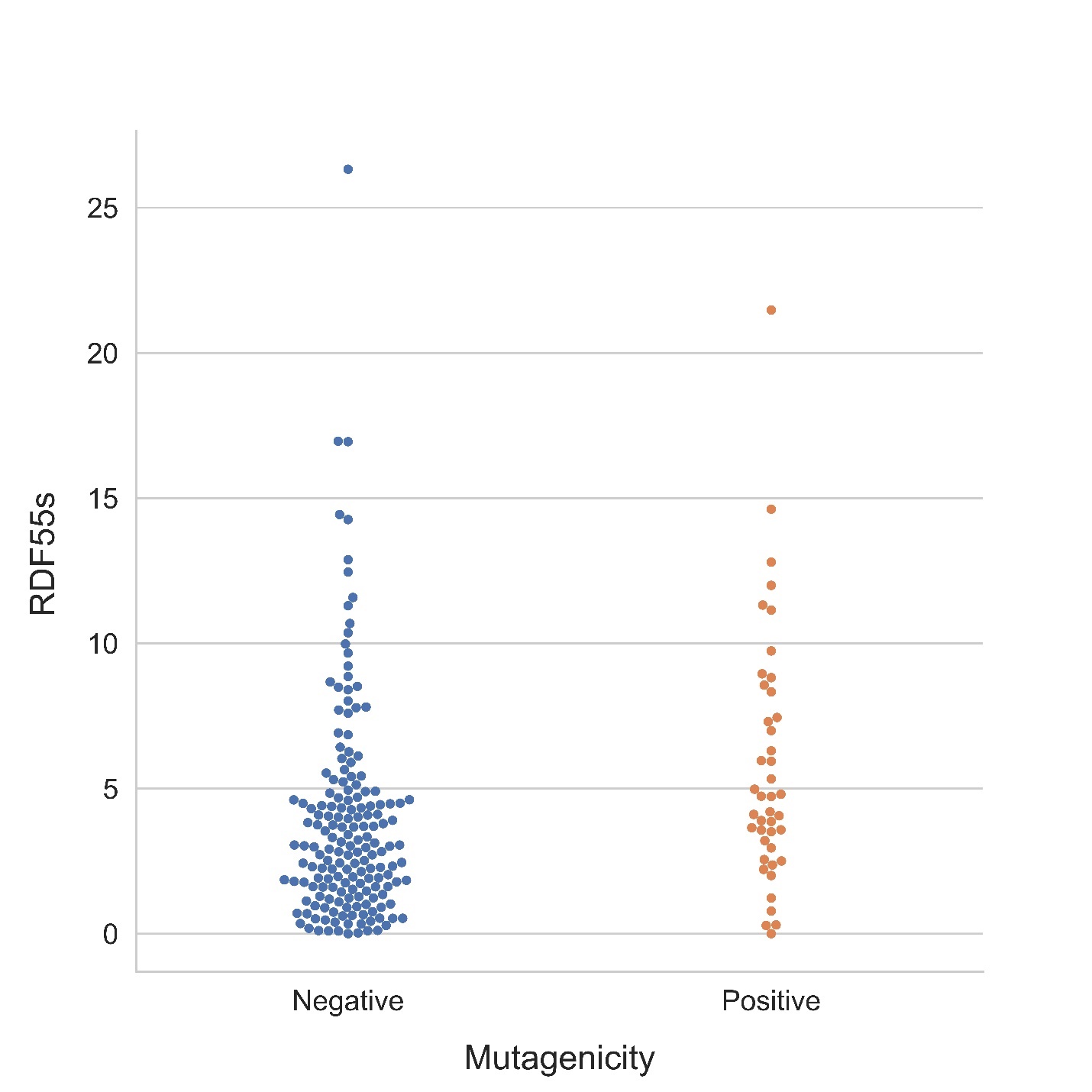


Figure S3.6: Small molecules swarm plot for radial distribution function 5.5 Å, weighted by intrinsic state (RDF55s). The RDF55s values of the positive molecules are slightly shifted towards the higher range.

# S-4 PaDEL-Descriptor Weightings

The following is a summary of the PaDEL descriptors that were most relevant to our analysis, with illustrations and examples. For full details the reader is referred to PADEL-Descriptor’s website [www.yapcwsoft.com/dd/padeldescriptor/](http://www.yapcwsoft.com/dd/padeldescriptor/) as well as to our github twsleight/Environmental\_PAH\_Mutagenicity

The book Molecular Descriptors for Chemoinformatics by Todeschini and Consonni is the most widely cited reference by the PADEL-Descriptor documentation and is an excellent reference for background information.4

Table S4.1: Static weighting schemes (table values).

|  |  |  |
| --- | --- | --- |
| **Weighting Scheme** | **Carbon** | **Oxygen** |
| Ionization Potential (i) | 11.2603 | 13.6181 |
| Relative Polarizability (p) | 1.76 | 0.8 |
| Mass (m) | 12.01 | 16 |
| Sanderson Electronegativity (e) | 2.75 | 3.65 |
| Van Der Waals Volume (v) | 20.58 | 14.71 |
| Covalent Radius (r) | 76 pm | 66 pm |

A useful table to weighting schemes for DRAGON™ software, which uses many of the same calculations can be found here [Dragon 6 user's manual (talete.mi.it)](http://www.talete.mi.it/help/dragon_help/index.html?weighting_schemes.htm)

The definition of intrinsic state (abbreviation s) is

Where Li is the principal quantum number, δiv is the number of valance electrons, and δi is the number of sigma electrons. Since our dataset contains only carbon and oxygen, which have principal quantum numbers of 2, this can be simplified to

Table S4.2: Intrinsic state of common structures.

|  |  |
| --- | --- |
| **Atom Group** | **Intrinsic State Value** |
| =O | 7.0 |
| -OH | 6.0 |
| -O- | 3.5 |
| aOa | 3.5 |
| -CH3 | 2.0 |
| =CH2 | 3.0 |
| -CH2- | 1.5 |
| =CH- | 2.0 |
| aCHa | 2.0 |
| =C= | 2.5 |

# S-5 PaDEL-Descriptor Equations

## S-5.1 Barysz Matrix

The Barysz matrix is a vertex-distance matrix. A simple vertex-distance matrix is computed by counting the number of vertices (atoms). The Barysz matrix uses the following formula:

Off Diagonal Terms:

Diagonal Terms:

This results in the carbon atoms having a 0 value on the diagonal, and oxygen atoms having the value 1-Zc/Zi

Since there are only two non-hydrogen atoms in our dataset. The following table can be developed for the possible values used for Zc

## S-5.2 Burden Modified Matrix

The Burden Modified Matrix is a molecular connectivity matrix which accounts for electronegativity terms in the diagonal. The off-diagonal terms are the square root of the conventional bond order.

Table S5.1: Atom electronegativity terms for the modified Burden matrix

|  |  |
| --- | --- |
| **Atom** | **Value** |
| C | 0.00 |
| H | 0.15 |
| N | 0.90 |
| O | 0.90 |
| F | 2.30 |
| Cl | 0.90 |
| Br | 0.80 |
| I | 0.50 |
| S | 0.50 |
| P | 0.50 |

## S-5.3 Broto-Moreau Autocorrelation

Where w is an atomic property, such as mass, van er Waals volumes, or Sanderson electronegativities.

## S-5.4 Moran Autocorrelation

The Moran autocorrelation is based on the difference between the atomic property value for each atom and the average value across the entire molecule. In this equation, ∆*k* is the lag (distance between molecules in the molecules graph) and A is total number of atoms

## S-5.5 Geary Autocorrelation

The Geary Autocorrelation is similar to the Moran autocorrelation in that the correlation is computed based on the differences between the atomic properties. However in the Geary autocorrelation the difference is between different atoms rather than between each atom and the average value.

# S-6 Representative Structures

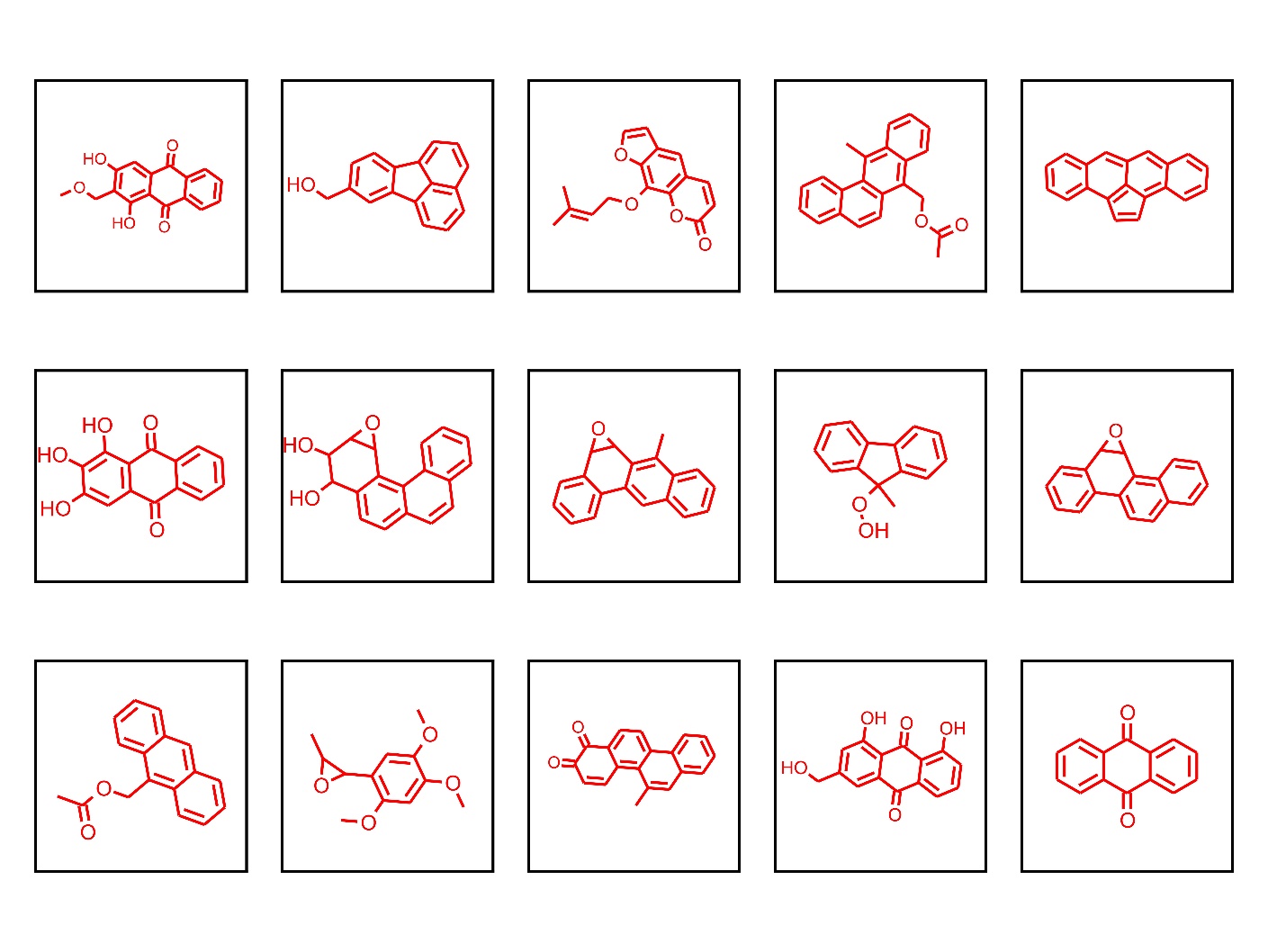


Figure S6.1: Representative mutagenic structures in the large molecules cluster.

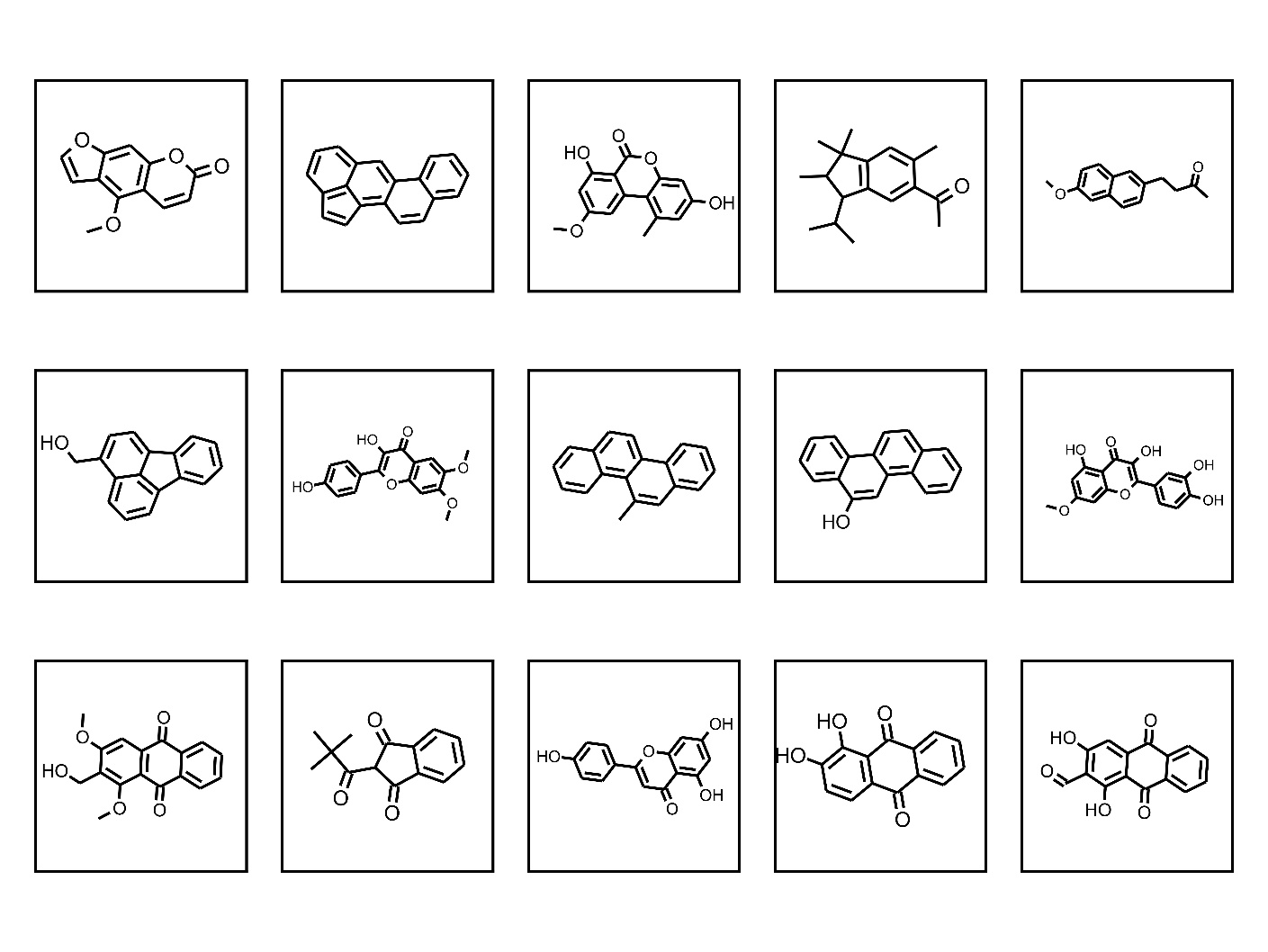
****

Figure S6.2: Representative non-mutagenic structures in the large molecules cluster.

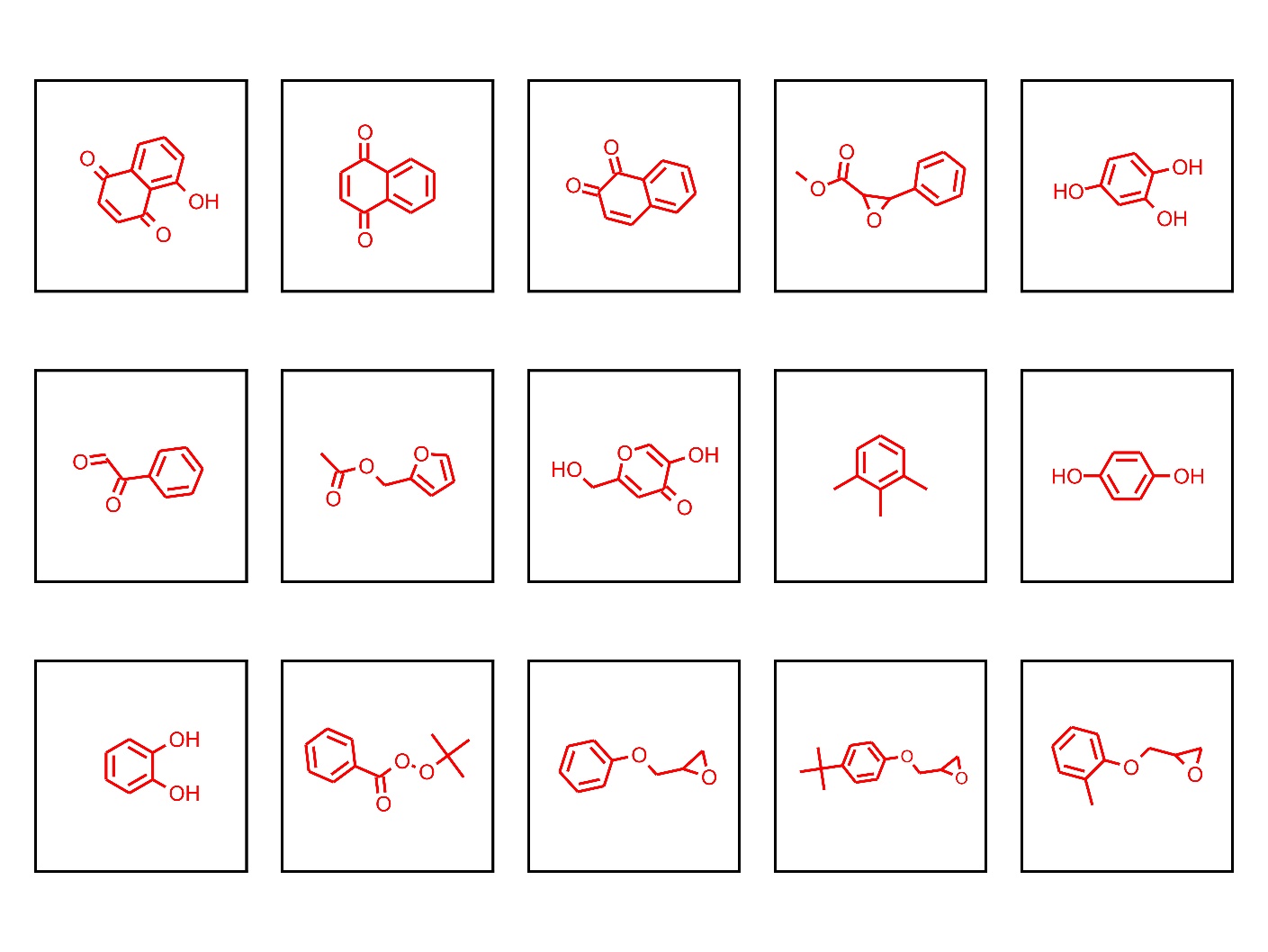


Figure S6.3: Representative mutagenic structures in the small molecules cluster.

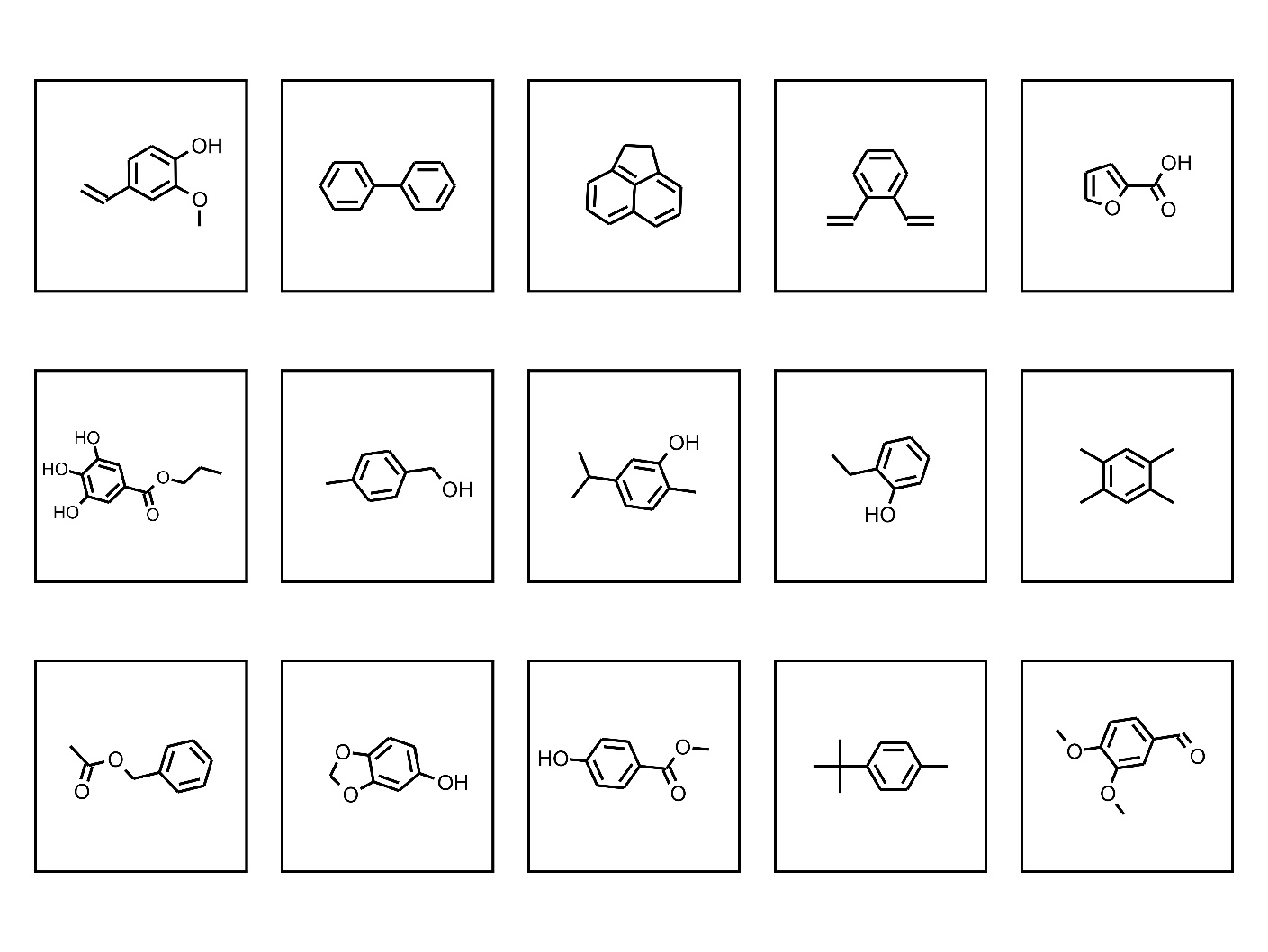
****

Figure S6.4: Representative non-mutagenic structures in the small molecules cluster.

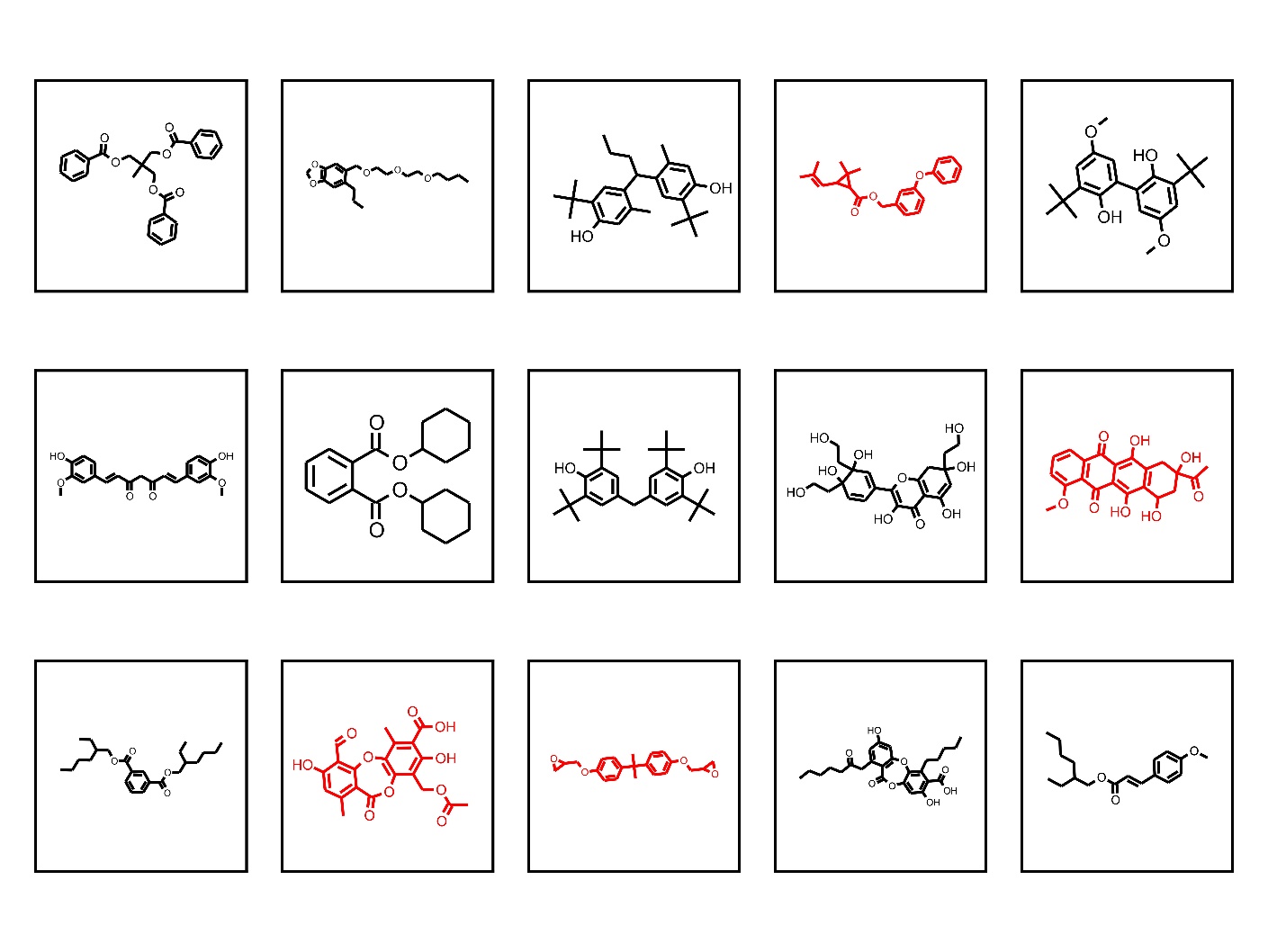
****

Figure S6.5: Other molecules (third cluster) representative structures. Red indicates a mutagen.

**References**

(1) Thorndike, R. L. Who Belongs in the Family? *Psychometrika* **1953**, *18* (4), 267–276. https://doi.org/10.1007/BF02289263.

(2) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

(3) Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1* (2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909.

(4) Consonni, V.; Todeschini, R. Molecular Descriptors. In *Recent Advances in QSAR Studies: Methods and Applications*; Puzyn, T., Leszczynski, J., Cronin, M. T., Eds.; Challenges and Advances in Computational Chemistry and Physics; Springer Netherlands: Dordrecht, 2010; pp 29–102. https://doi.org/10.1007/978-1-4020-9783-6\_3.