Electronic Supplementary Material (ESI) for RSC Advances.

SUPPORTING INFORMATION

Activity landscape sweeping: Insights into the mechanism of inhibition and optimization of DNMT1 inhibitors

J. Jesús Naveja and José L. Medina-Franco

Table of Contents

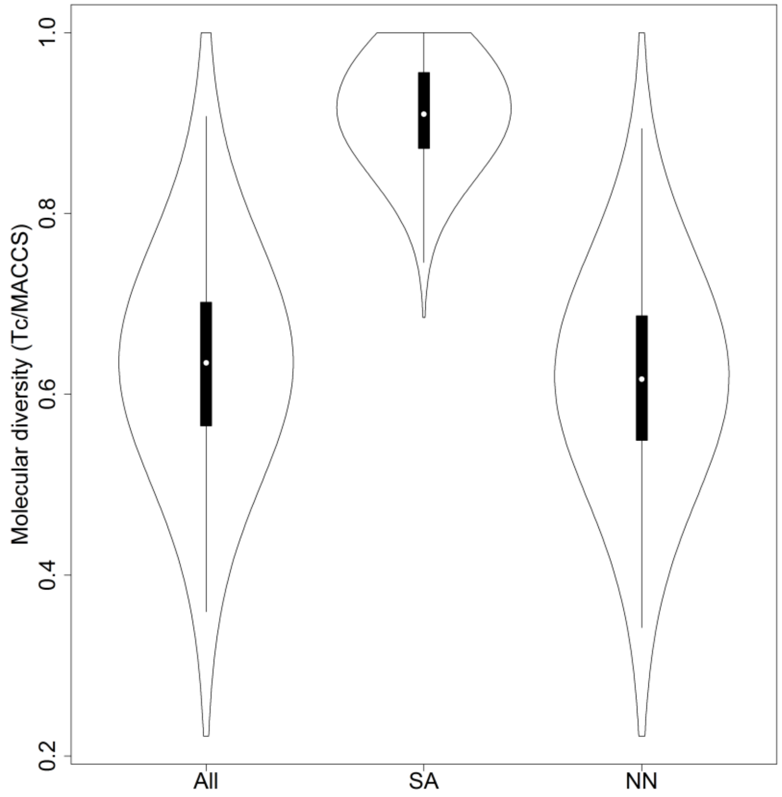


Figure S1. Distribution of the similarity values of the entire data set (All), SAM-analogues (SA) and non-nucleosides (NN).

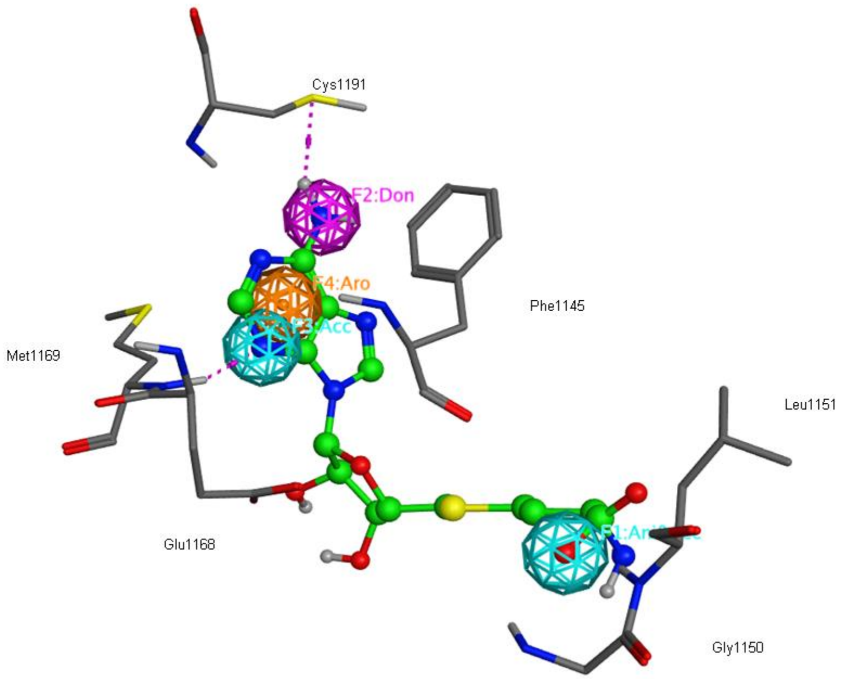


Figure S2. Pharmacophoric constraints considered for the docking of SAM-analogues. These include two hydrogen bond acceptors, a hydrogen bond donor, and an aromatic ring. SAH carbon atoms are rendered in green.

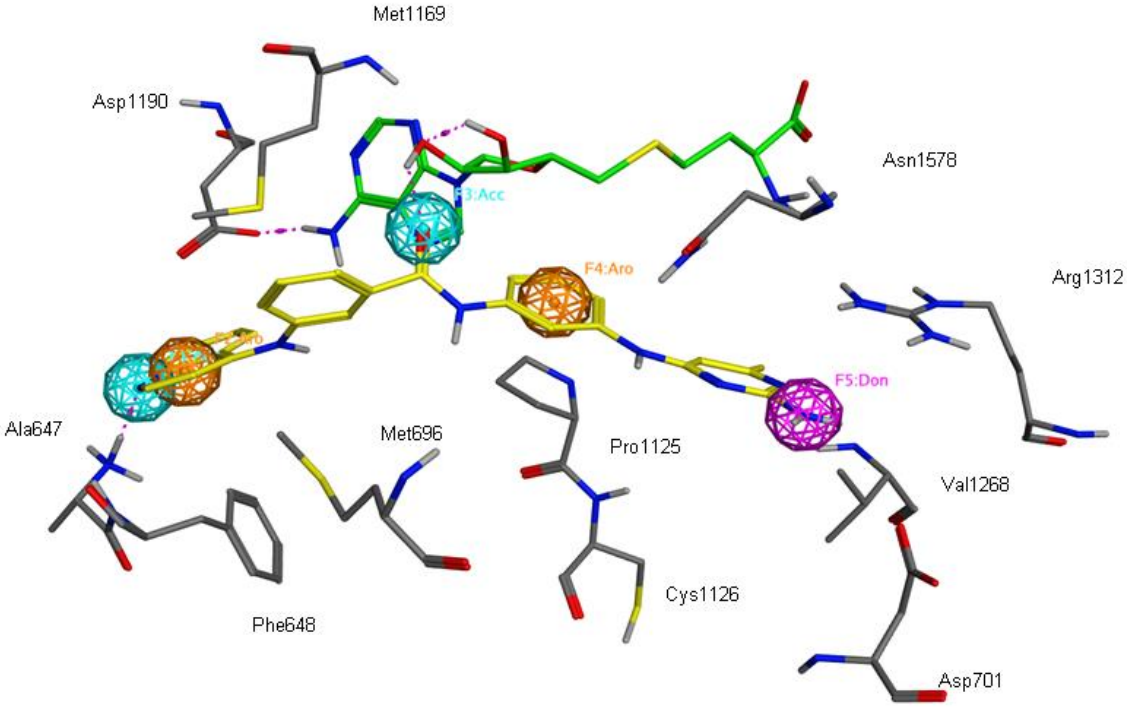


Figure S3. Pharmacophoric constraints considered for docking of the non-nucleosides with DNMT1. These include two aromatic rings, an hydrogen bond donor, and two hydrogen bond acceptors. Carbon atoms of SAH and most potent regioisomer of SGI-1027 are rendered in green and yellow, respectively.

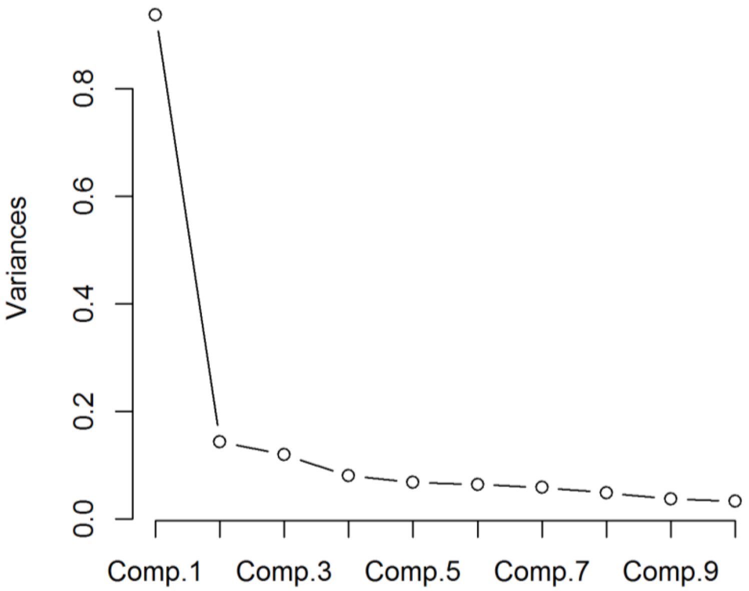


Figure S4. Scree plot depicting the variance explained by each component (in absolute terms). PC #1 explains at least 40% of the variance.

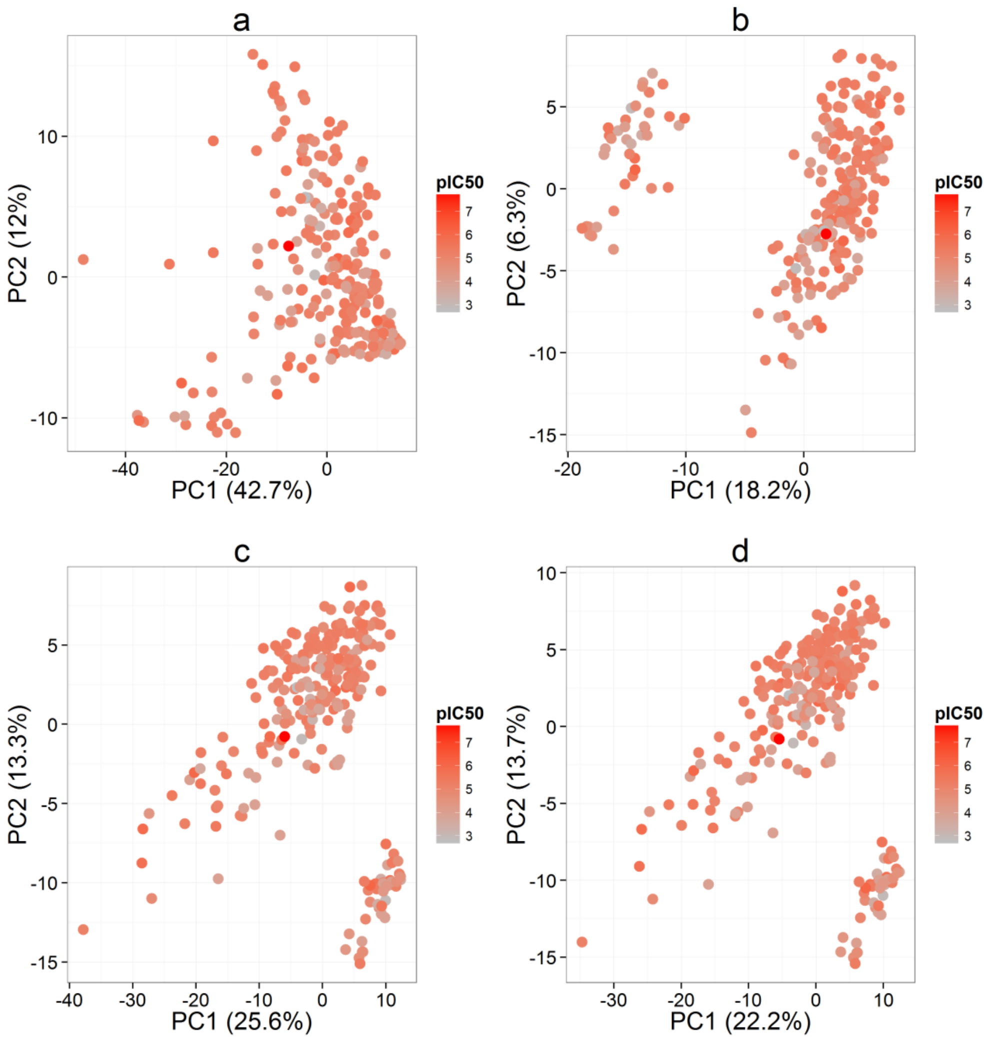


Figure S5. Visual representation of the chemical space generated with different chemical representations: MACCS, ECFP, fusion mean and fusion-Z (a-d, respectively). Points are colored by the pIC_{50} values. *X* and *y* axes refer to the principal components employed to plot the chemical space, PC1 and PC2, with the variance explained by them between parenthesis.

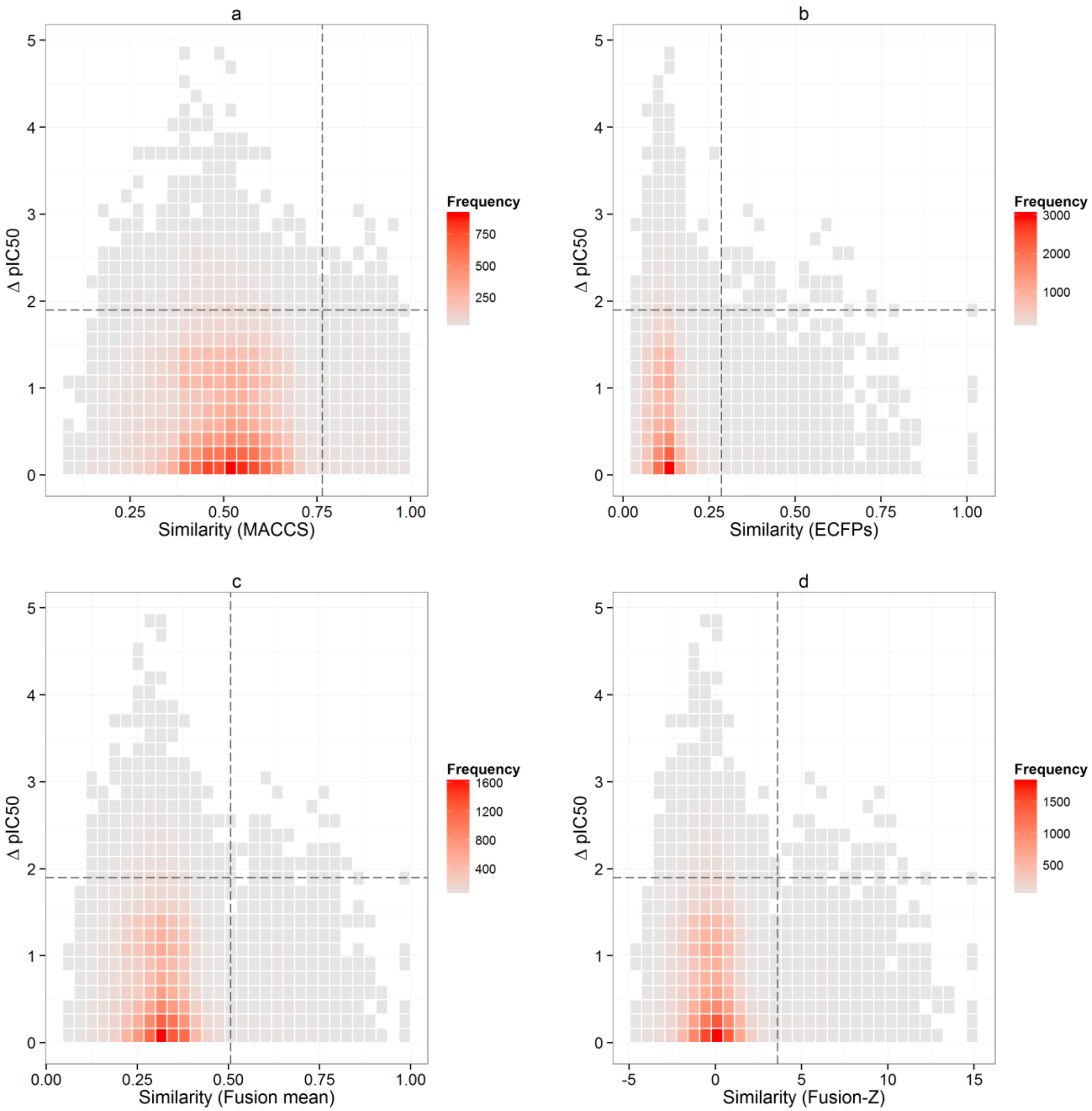


Figure S6. Density SAS maps obtained with different similarity assessing approaches as labeled in the *x*-axis. The dashed lines subdivide the SAS map in the classical four quadrants as explained in the main text. Thresholds are defined as two standard deviations above the mean of the respective variable.

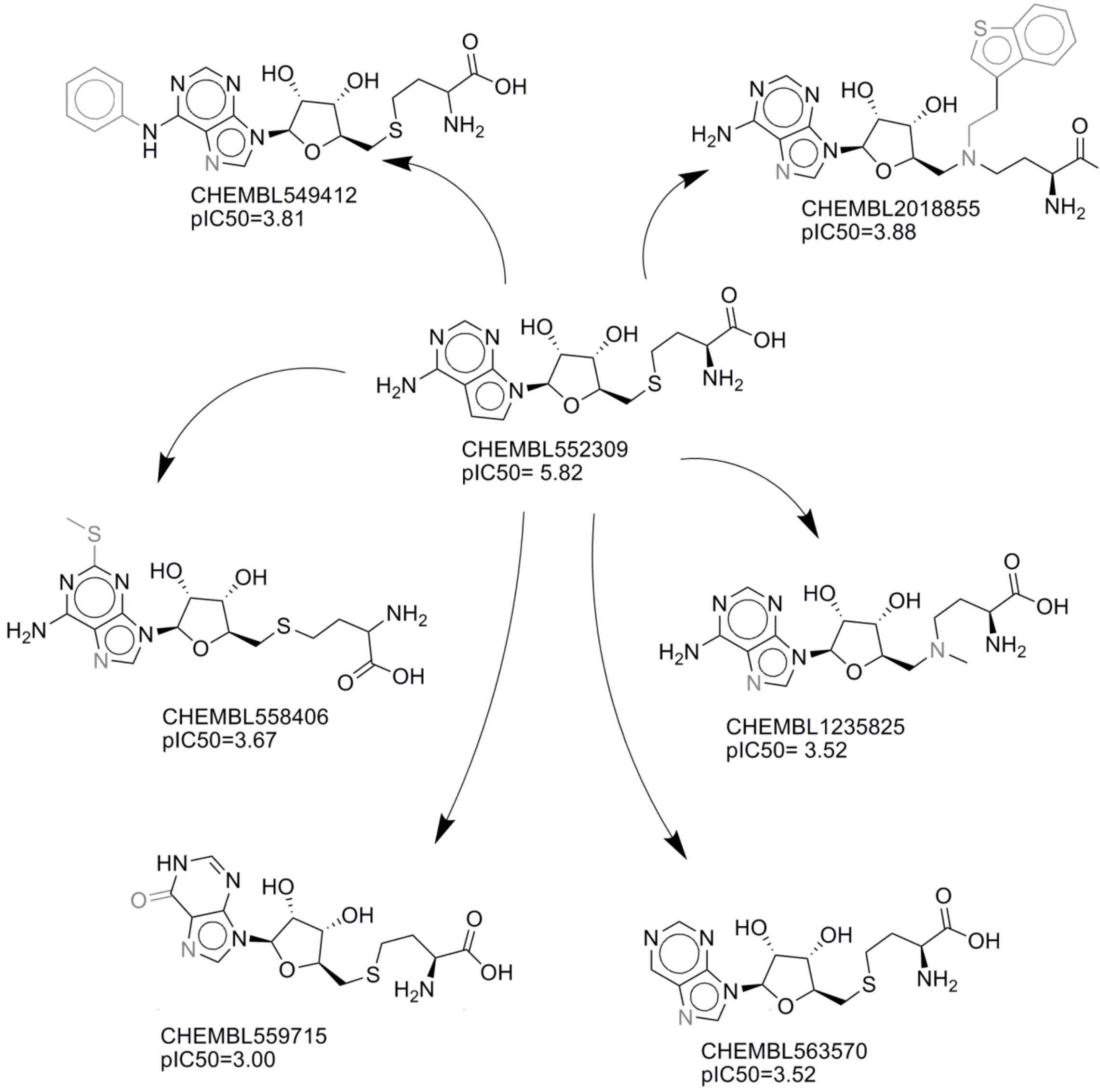


Figure S7. Compounds with which the activity cliffs generator CHEMBL552309 forms activity cliffs.

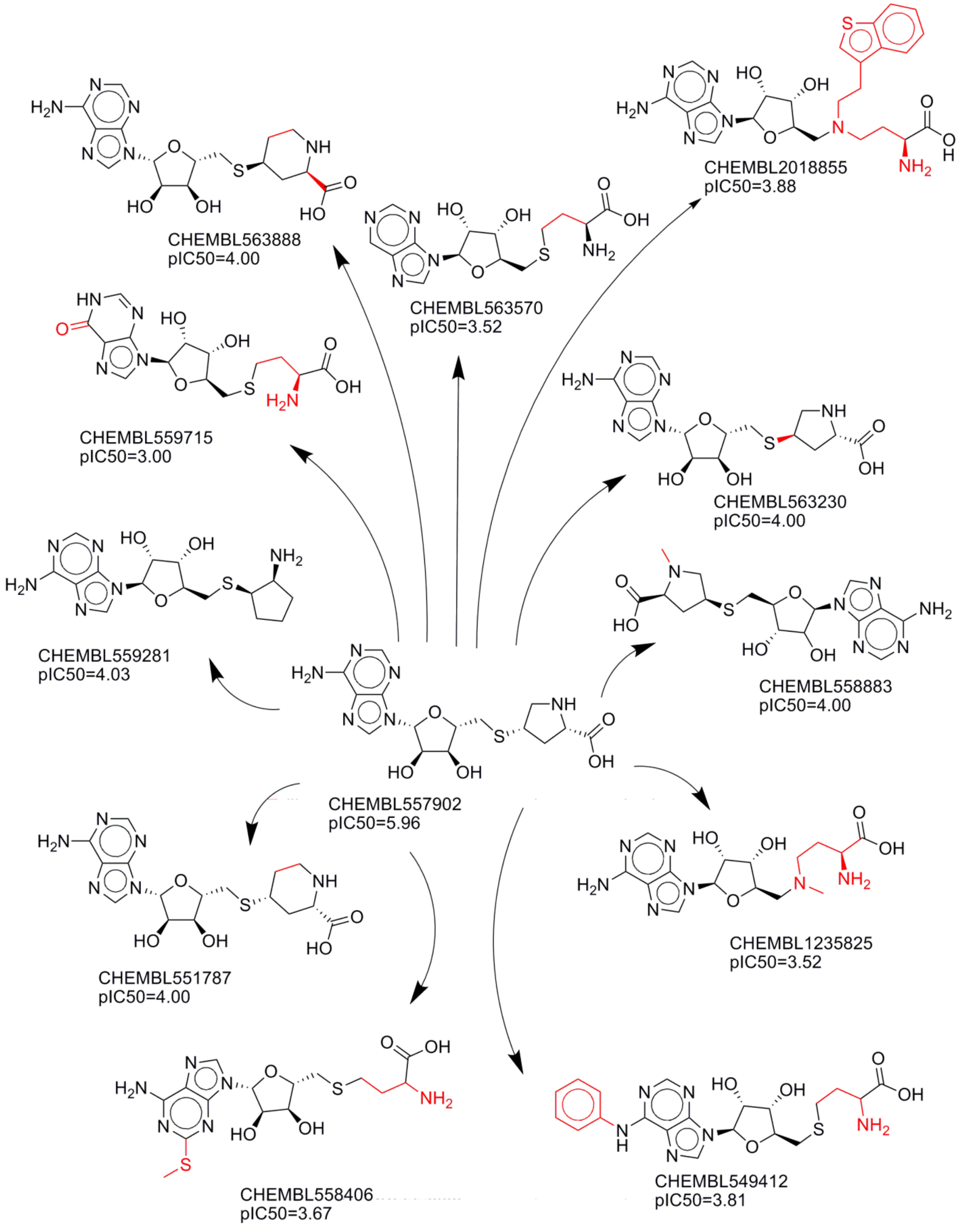


Figure S8. Compounds with which the activity cliffs generator CHEMBL557902 forms activity cliffs.

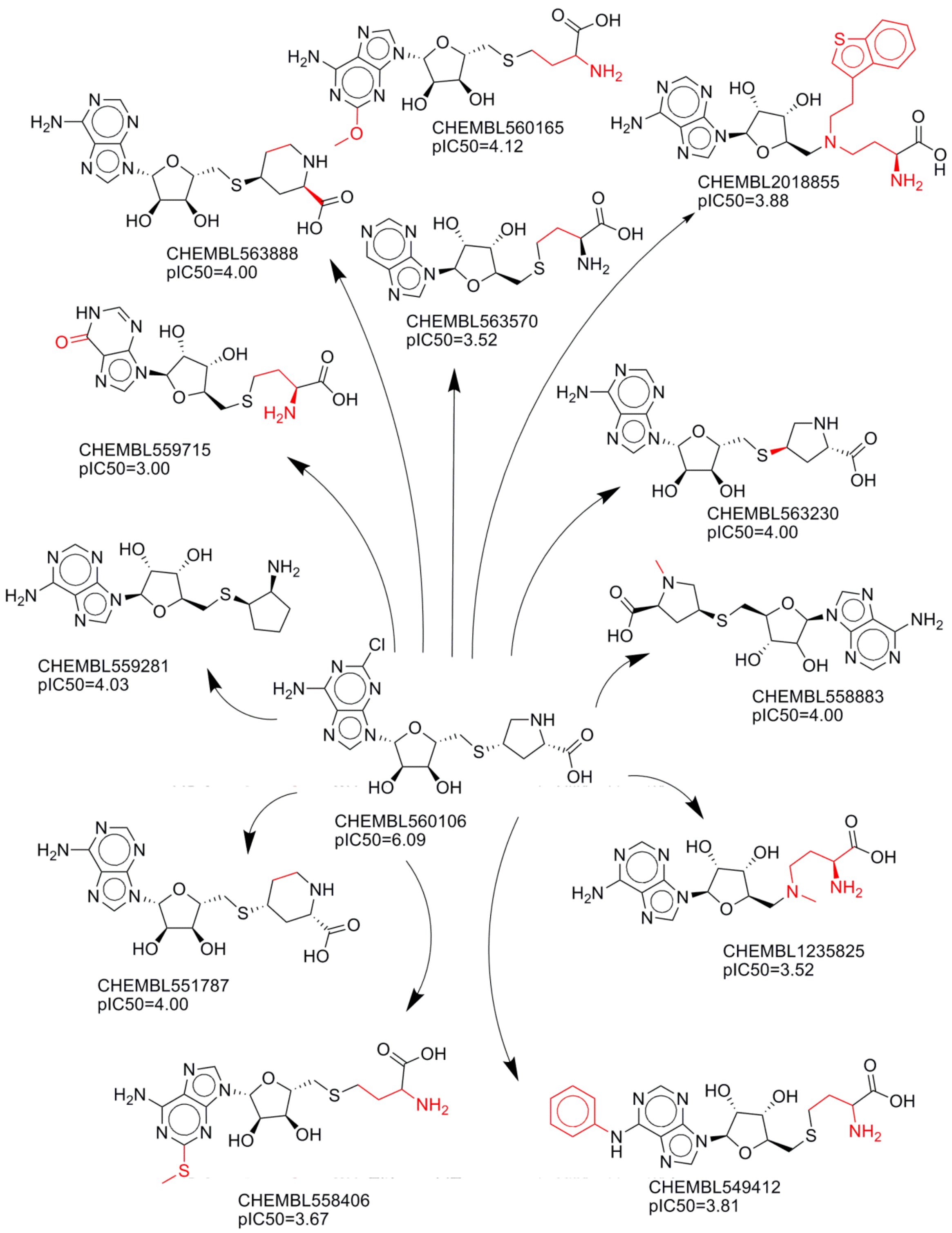


Figure S9. Compounds with which the activity cliffs generator CHEMBL560106 forms activity cliffs.

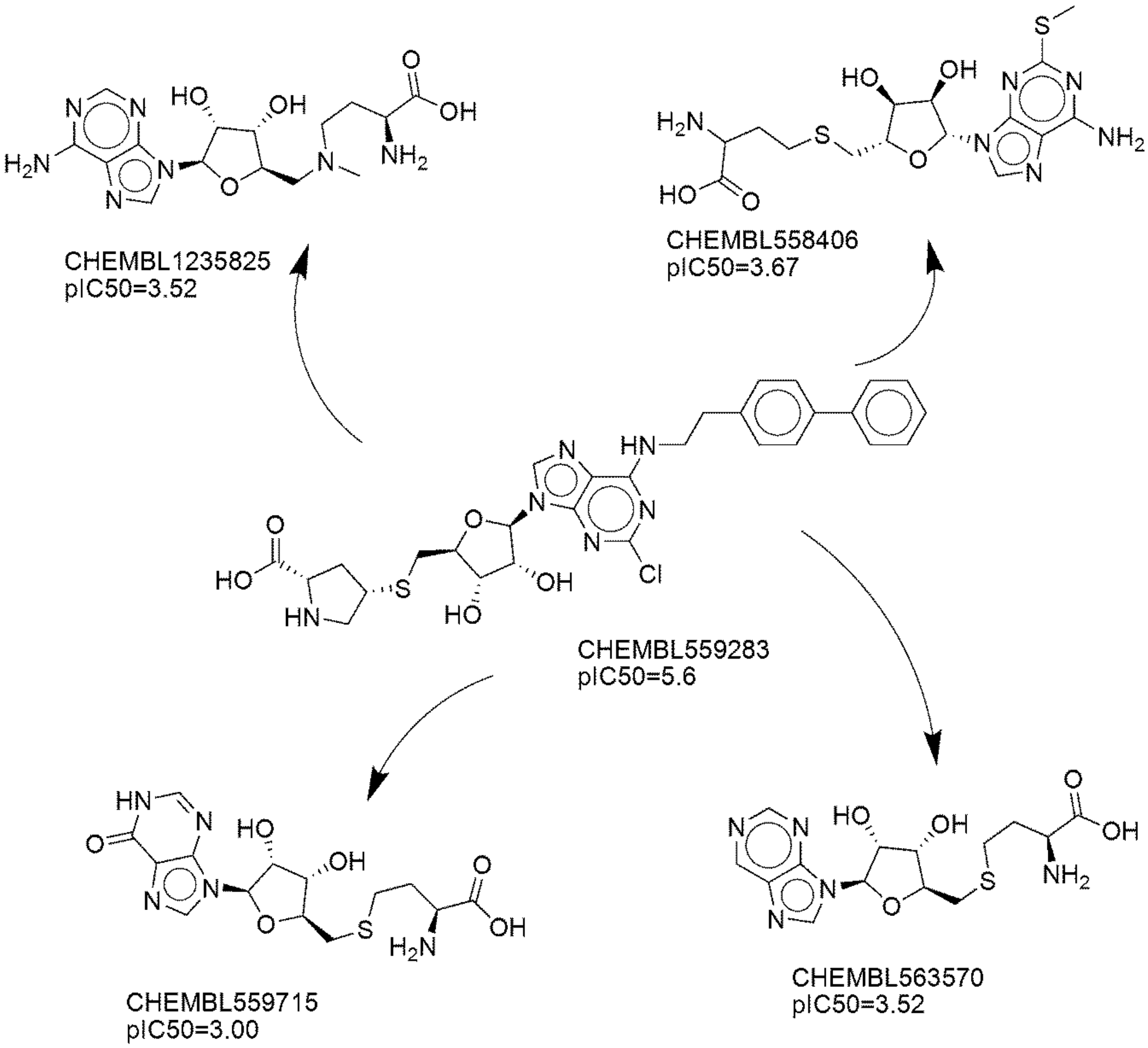


Figure S10. Compounds with which the activity cliffs generator CHEMBL559283 forms activity cliffs.

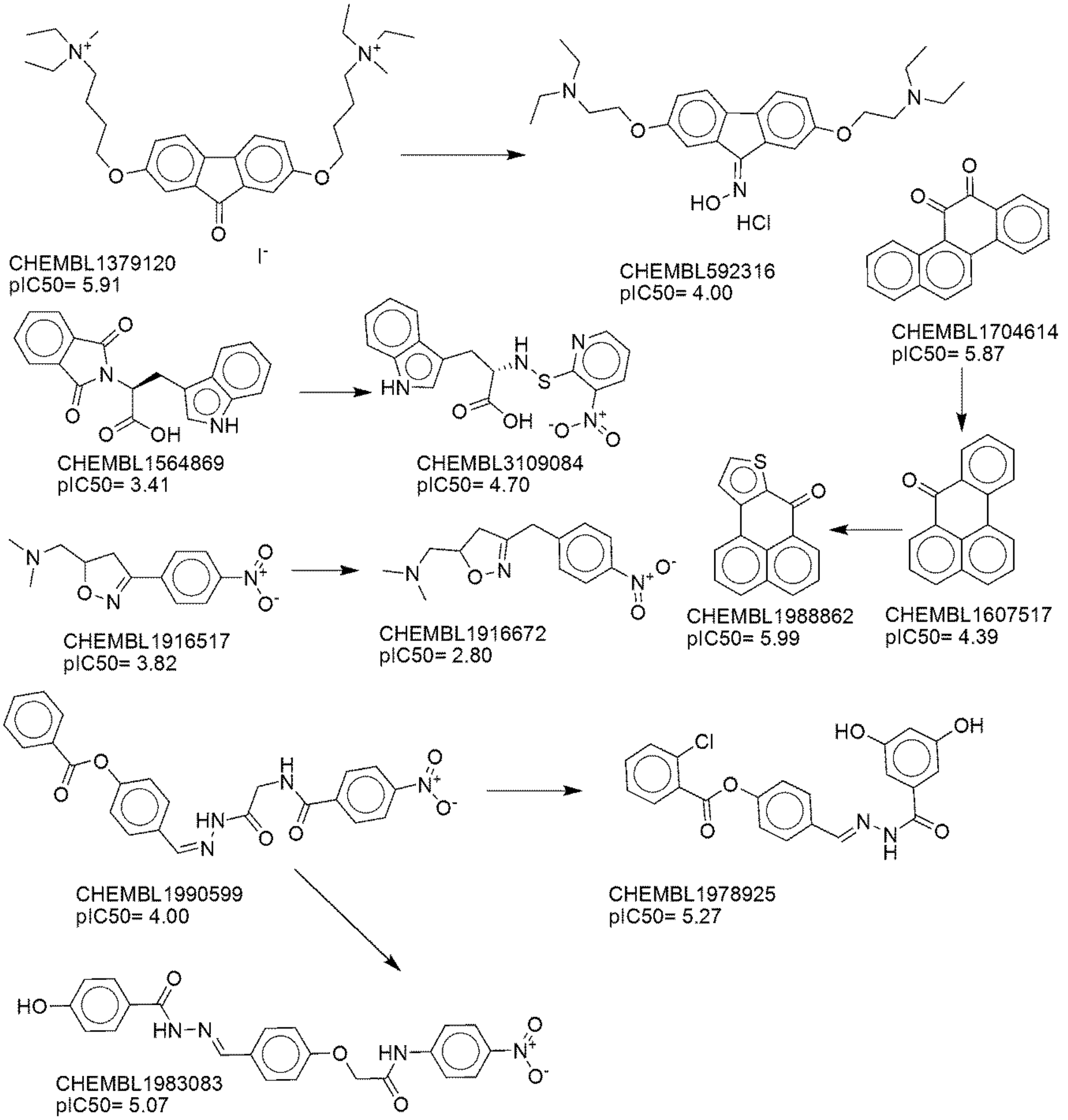


Figure S11. Shallow activity cliffs omitted in the main text. These compounds are less interesting due to either low activities in both compounds forming the pair or potential toxicity by intercalating DNA or reacting unspecifically.

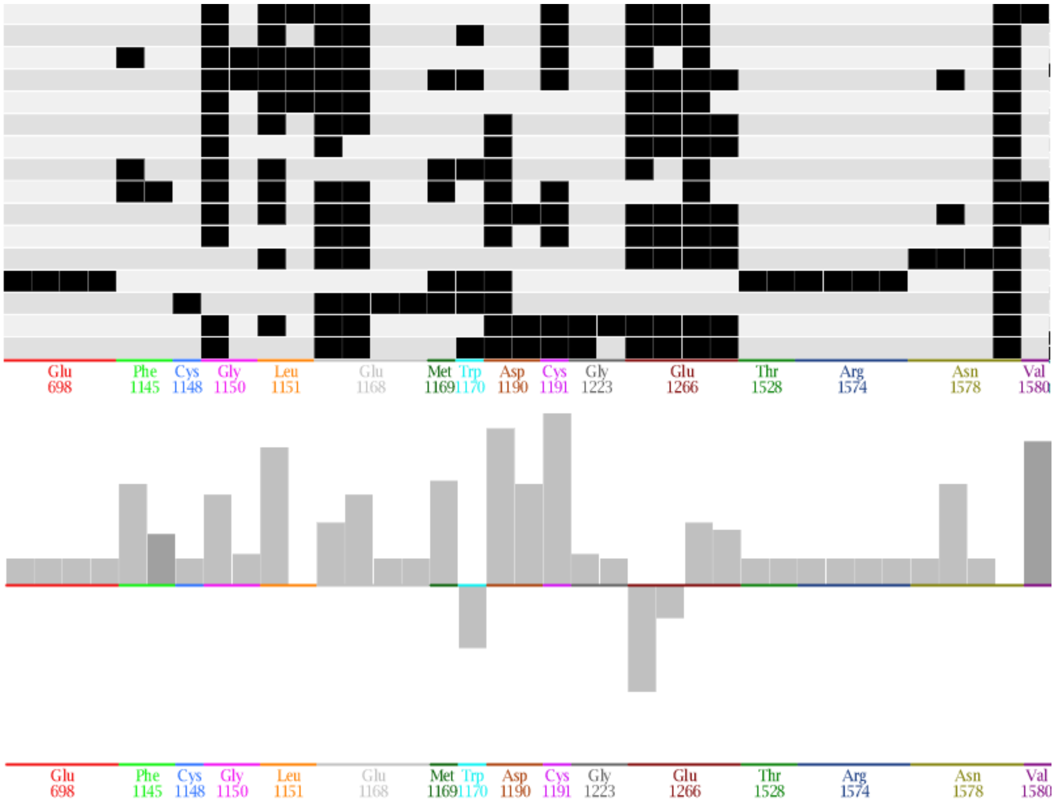


Figure S12. PLIF analysis for the activity cliffs generated by CHEMBL552309, which are depicted in figure S8. Two poses of every molecule were considered. Order is as follows (CHEMBL IDs) 563570; 559715; 558406; **552309**; 549412; 2018855; 1235825. PLIFs are above and significance test below.

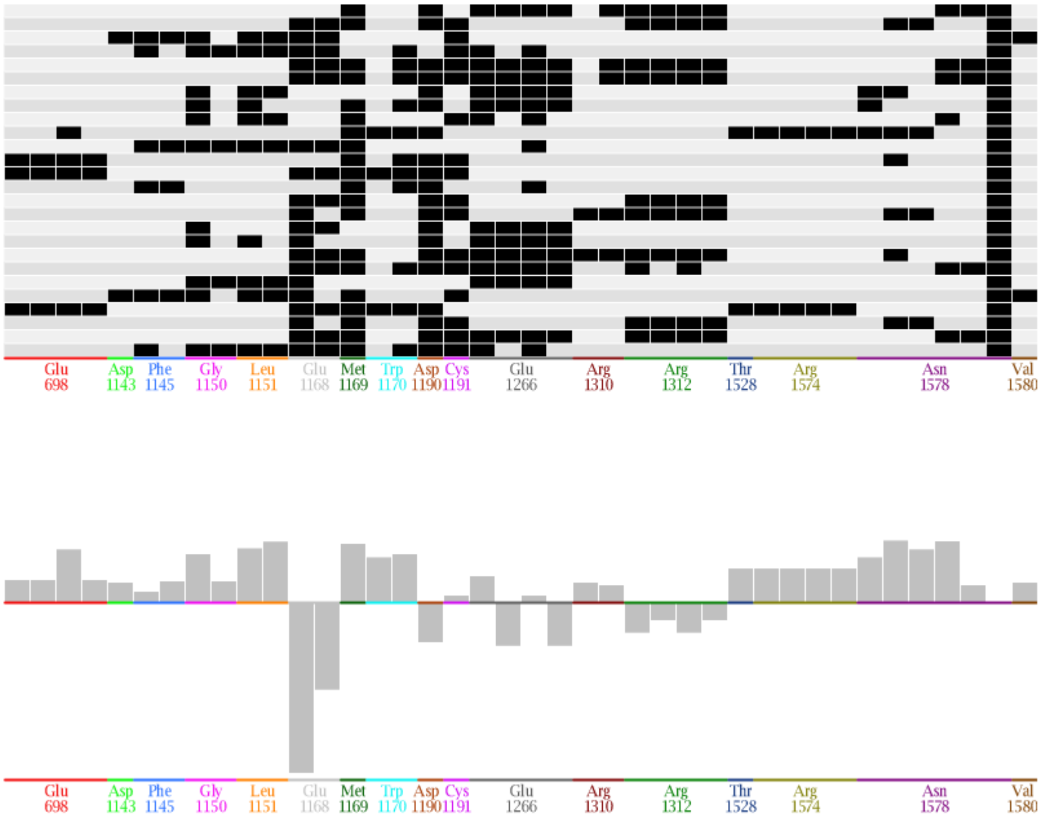
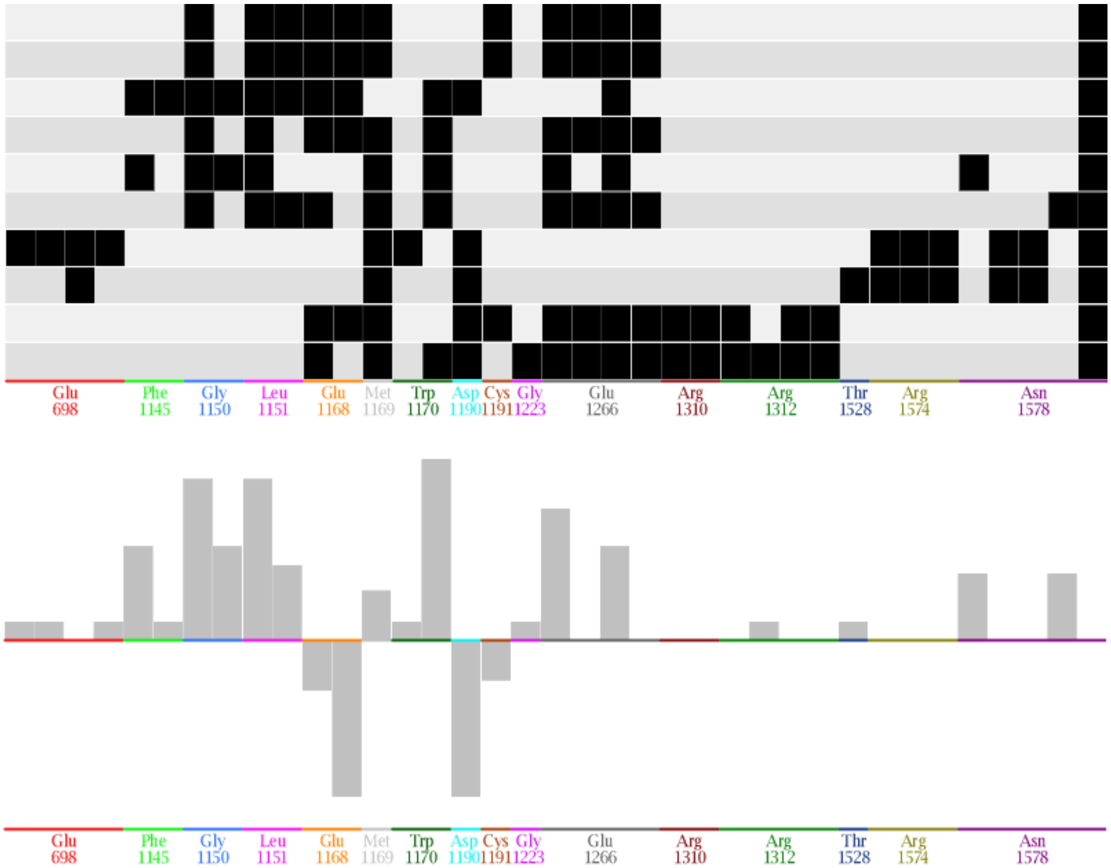


Figure S13. PLIF analysis for the activity cliffs generated by CHEMBL560106, which are depicted in figure S9. Two poses of every molecule were considered. Order is as follows (CHEMBL IDs): 563888; 563570; 563230; 560165; **560106**; 559715; 559281; 558883; 558406; 551787; 549412; 2018855; 1235825. PLIFs are above and significance test below.



.

Figure S14. PLIF analysis for the activity cliffs generated by CHEMBL559283, which are depicted in figure S10. Two poses of every molecule were considered. Order is as follows (CHEMBL IDs): 5636570; 559715; **559283**; 558406; 1235825. PLIFs are above and significance test below.

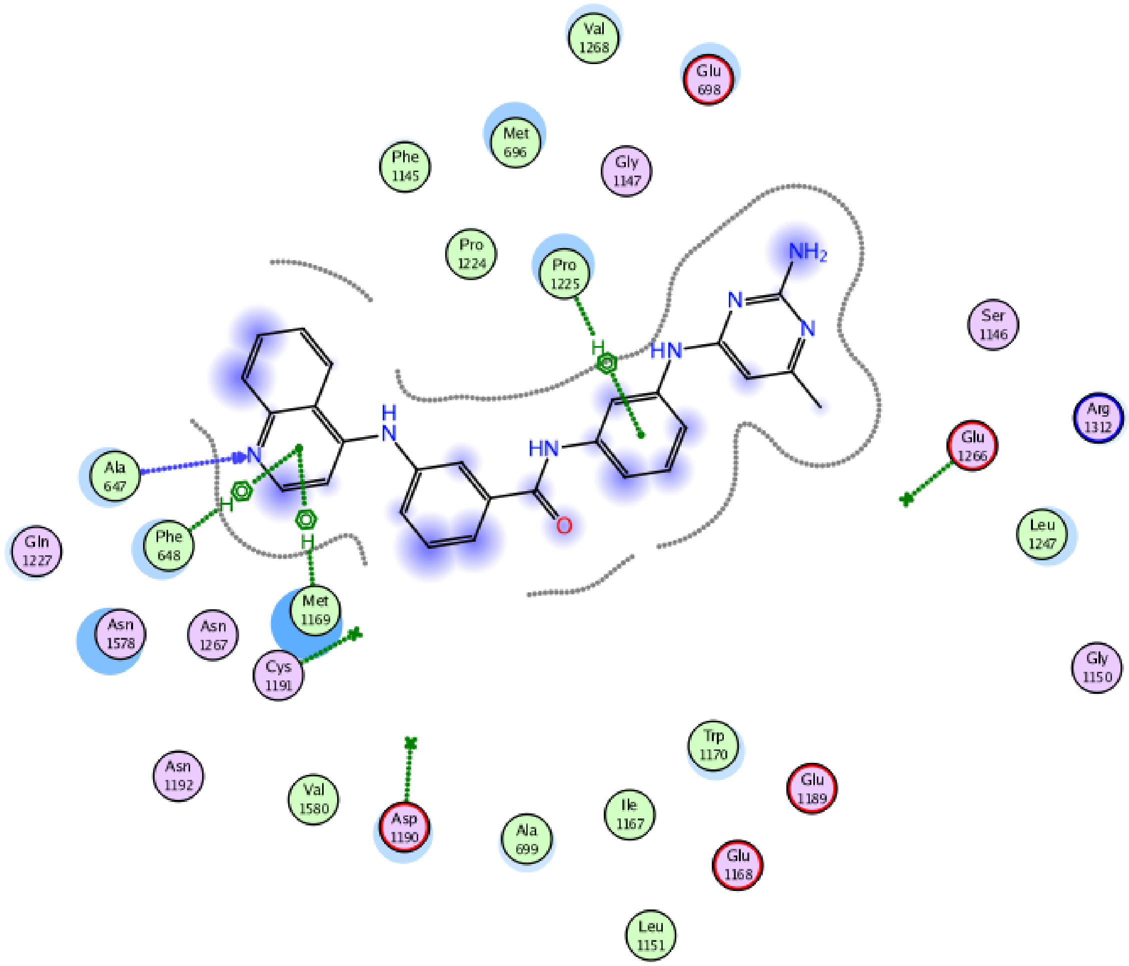


Figure S15. Bidimensional representation of the ligand interactions found for the meta-meta SGI-1027 regioisomer with DNMT1.

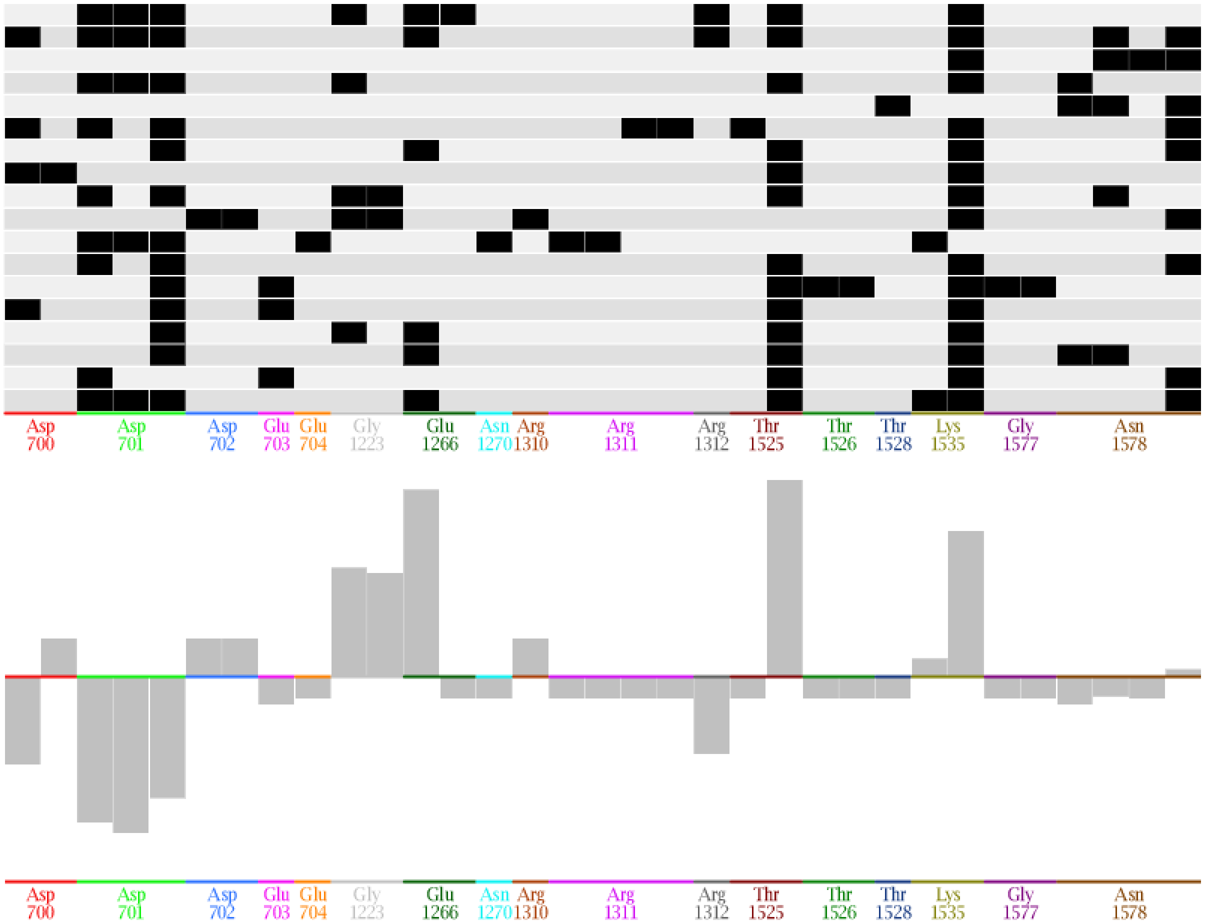


Figure 16. Summary of protein-ligand interaction fingerprint (PLIFs) analysis of the SGI-1027 regioisomers. A) Data matrix. B) Significance test. A darker color means that the interaction is more associated to the active compound. The chemical structures can be found in figure 5. Were defined arbitrarily as active those compounds with pIC50>5. No statistically significant differences were found. Two poses of every molecule were considered. Order is as follows (CHEMBL IDs): 3126649, 3126648, 3126647, 3126646, 3126645, 3126644, *ortho/ortho* (not in ChEMBL), 2336409 and *para/meta* (not in ChEMBL). PLIFs are above and significance test below.

Further details on PCA and K-means in this study

Principal Components Analysis (PCA)

PCA was conducted on the similarity matrices of MACCS and ECFP, and also with the mean and Z-fusion data fusion approaches as described in the main text. It should be noted that the primary output of PCA is 280 eigenvectors (equal to the 280 compounds in the dataset), that are numbered in descending order according to how much variance do they explain (Figure S4). Figure S5 depicts the chemical space with each of these similarity measures.

We will here focus on some important aspects regarding the PCA methodology. For instance, a key feature of PCA is the determination of the number of principal components that would be sufficient to explore the data. As mentioned in the main text, we decided to include only the first two principal components, based on the thumb rule of the curve's "elbow" when plotting the principal components and the variance explained by them. If the alternative thumb rule of choosing the principal components explaining at least 85% of the variance were used, the first 44 principal components would be necessary, and this would evidently difficult the visual representation.

K-means and clustering

Among the statistical methods to perform clustering, hierarchical clustering and K-means are the most popular methods due to their general purpose approach. We preferred K-means over hierarchical clustering in this case provided that we had prior knowledge of the two major types of compounds within the dataset and K-means is a method that requires *a priori* definition of the number of clusters, in contrast with hierarchical clustering. Then, we could easily evaluate (as discussed in the main text) the performance of the clustering with *k=2* (i.e., two clusters defined *a priori*) and finally propose a reasonable number of clusters for the data based on the reduction of the within groups sums of squares, as shown in figure S17. Furthermore, the visualization of the hierarchical clustering results (usually summarized through a dendrogram) may be difficult with such a high number of compounds (280 in this case). For a detailed and updated review on clustering methods, see Jain, A. K. *Pattern Recognition Letters* 31 (2010) 651–666).

As shown in figure S17, a clustering reducing a high amount of the within groups sums of squares could define 6 cluster (see Figure S18), although with 3 clusters a fair quantity of the sums of squares is reduced (see Figure S19). Importantly, the input of the k-means method was the first five principal components obtained with ECF.

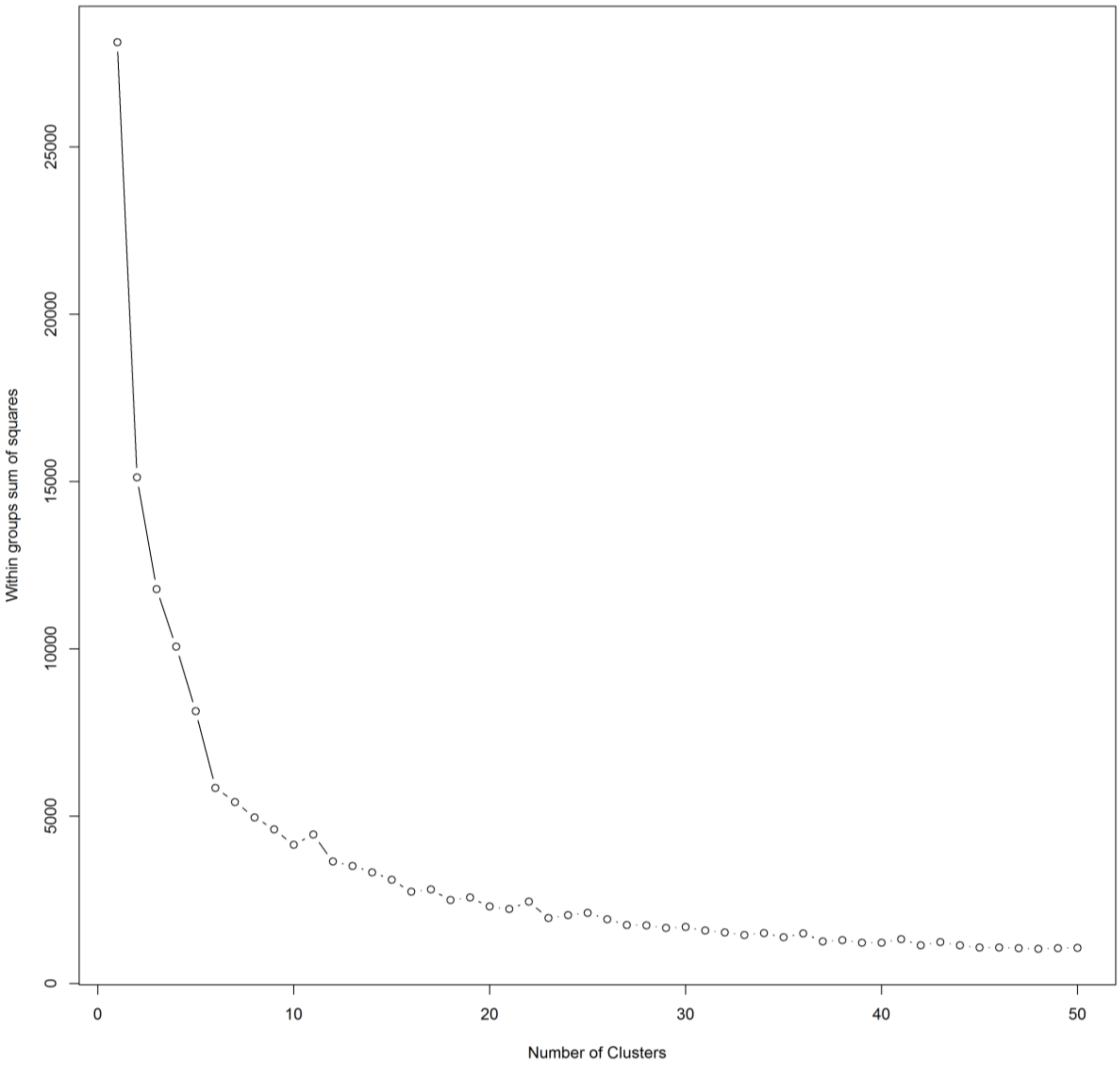


Figure S17. Number of cluster versus within groups sums of squares calculated by the k-means method. Each point represents a different number of clusters. It can be seen than the most abrupt slopes are among the first 6 points. From 3-6 points a considerable diminishing of the within groups sums of squares (approximately 40-70%) can be observed, thus concluding these to be optimal numbers of clusters.

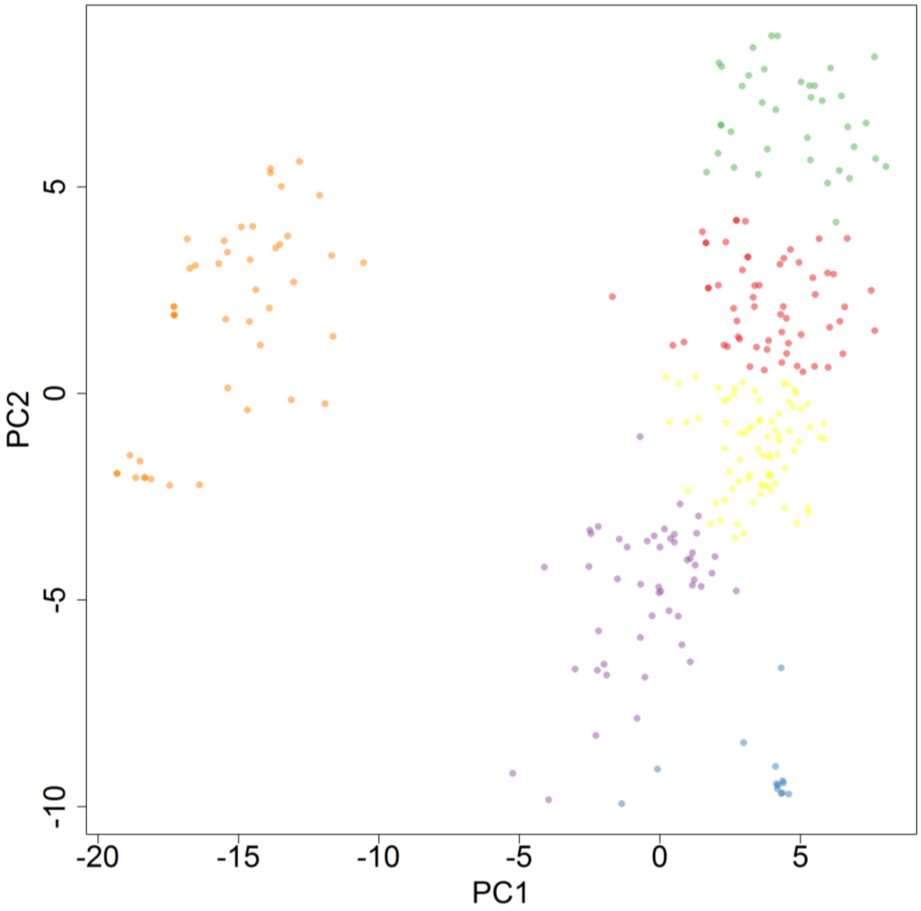


Figure S18. Depiction of the chemical space considering 6 clusters. It will be important to study the local SAR of each cluster separately. PC1 and PC2 account for 18.23% and 6.44%of the variance, respectively.

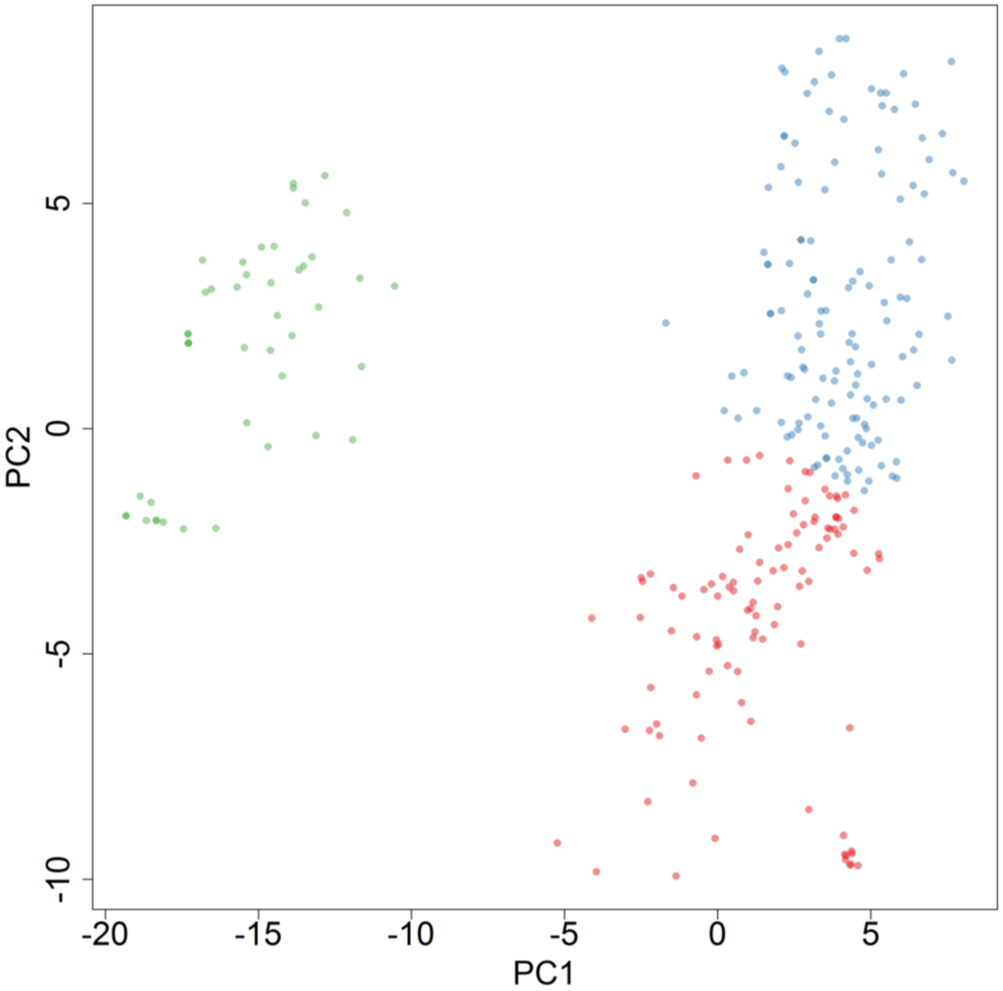


Figure S19. Depiction of the chemical space considering 3 clusters. PC1 and PC2 account for 18.23% and 6.44%of the variance, respectively.