QSAR EDA

Edward Francisco Mendez-Otalvaro, Daniel Alberto Barragan, Isaias Lans-Vargas

2021

Exploratory dataset analysis

The dataset of the 2,6-disubstituted glucosamines consists of 27 molecules with a common **scaffold** (glucosamine, an amino sugar). These molecules has inhibitory activity against HKII and are experimentally validated in a range of μM (EC50). For more details about the dataset, please refer to (Lin et al. 2016).

The molecules were modelled in *Spartan 14'* according to the topology given by (Lin et al. 2016). After that, a conformer distribution of 100 conformes were carried out for each molecule, and the molecule with the minimun potential energy was selected. The protonation state was establish with *Marvin software* at physiological conditions (T=310 K and pH= 7.0) using the theoretical titration curve for each ligand. The tautomer distribution was carried out with *Dimorphite-DL software*. After all the modeling steps, each molecule was manually inspected in order to get reliable chemical models.

The 27 ligand models were uploaded into *OCHEM web platform* in order to calculate mechanistic interpretable 2D and 3D molecular descriptors. Namely:

From **AlvaDesc V.1.0.22** package

- Constitutional descriptors (48)
- Atom-centred fragments (115)
- Ring descriptors (32)
- Pharmacophore descriptors (165)
- P_VSA-like descriptors (58)
- 2D atom pairs (1596)
- Geometrical descriptors (38)
- Charge descriptors (15)
- Functional group counts (154)
- Drug-like indices (28)

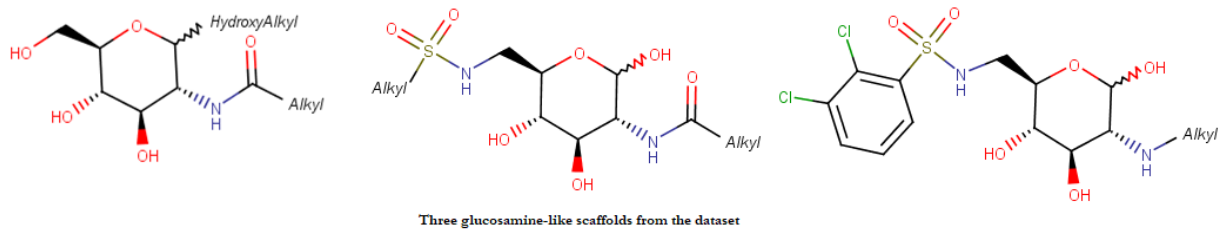
From **ALogPS** package:

- Octanol/water partition coefficient (1)
- Solubility in water (1)

From **MOPAC2016** package:

- Electronic molecular descriptors (16)

For more details about the process, please refer to the Master Thesis Manuscript.



Three glucosamine-like scaffolds from the dataset

Importing dataset

Now, the dataset previously prepared is upload (we concatenate the molecular descriptors with the biological activity):

```
Descriptores_I <- read.csv("Descriptores_Alvrunner.csv", stringsAsFactors = FALSE)

Descriptores_II <- read.csv("Descriptores_MOPAC.csv", stringsAsFactors = FALSE)

Descriptores <- cbind(Descriptores_I, Descriptores_II)


## Biological activity

Actividad <- read.csv2("Actividad.csv", stringsAsFactors = FALSE)

## Joining molecular descriptors with activity

Dataset <- cbind(Descriptores, Actividad)

## Cleaning labels

Dataset$ID <- NULL
Dataset$IC50..uM. <- NULL

## Renaming biological activity

names(Dataset)[names(Dataset) == "IC50..M."] <- "pIC50"
```

Converting biological activity (EC50) into logarithmic scale, and then, calculating dimension of the dataframe

```
Dataset[, 2284] <- -log10(Dataset[, 2284])

dim(Dataset)
```

```
## [1]   27 2284
```

So, there are 27 molecules with 2283 molecular descriptors and a biological activity response.

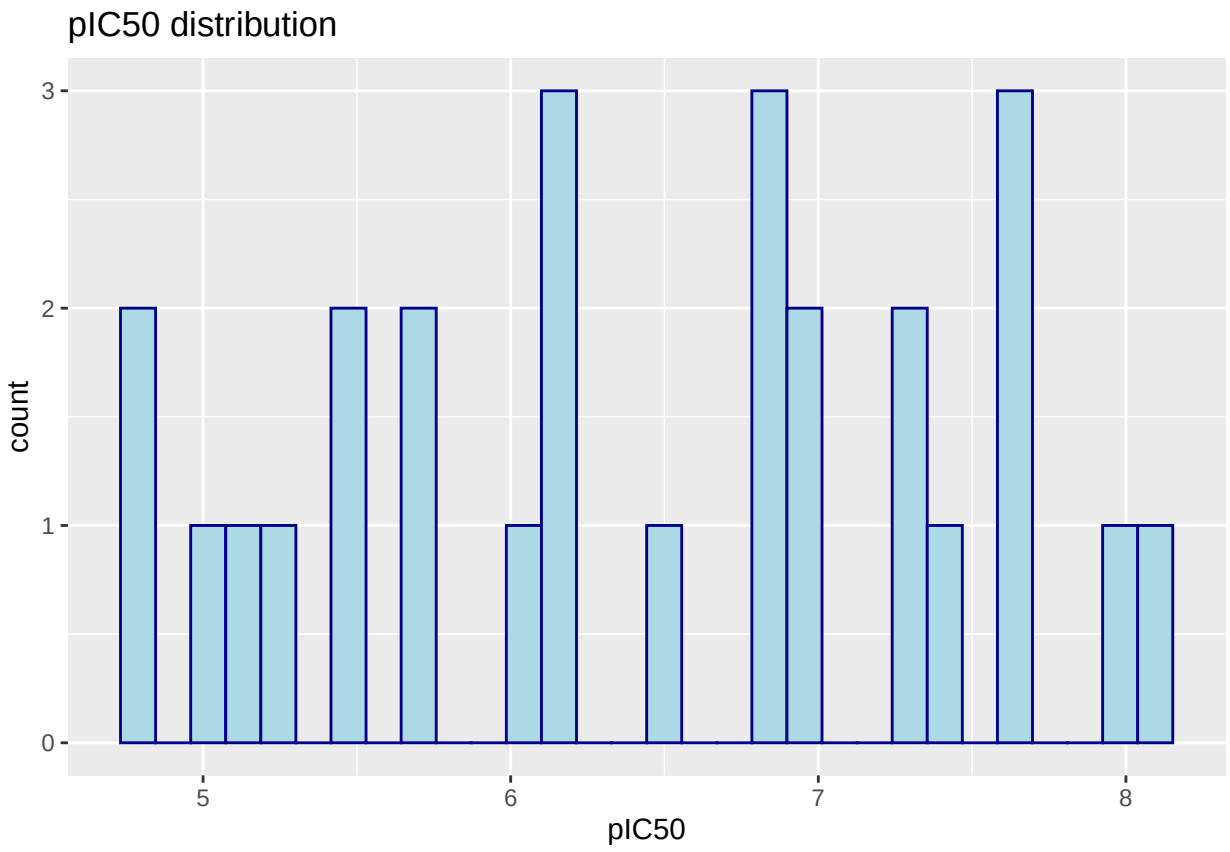
Behaviour of the response variable (pIC50):

In order to get insight about the distribution of the biological activity. Some exploratory analysis of the data will be carried out.

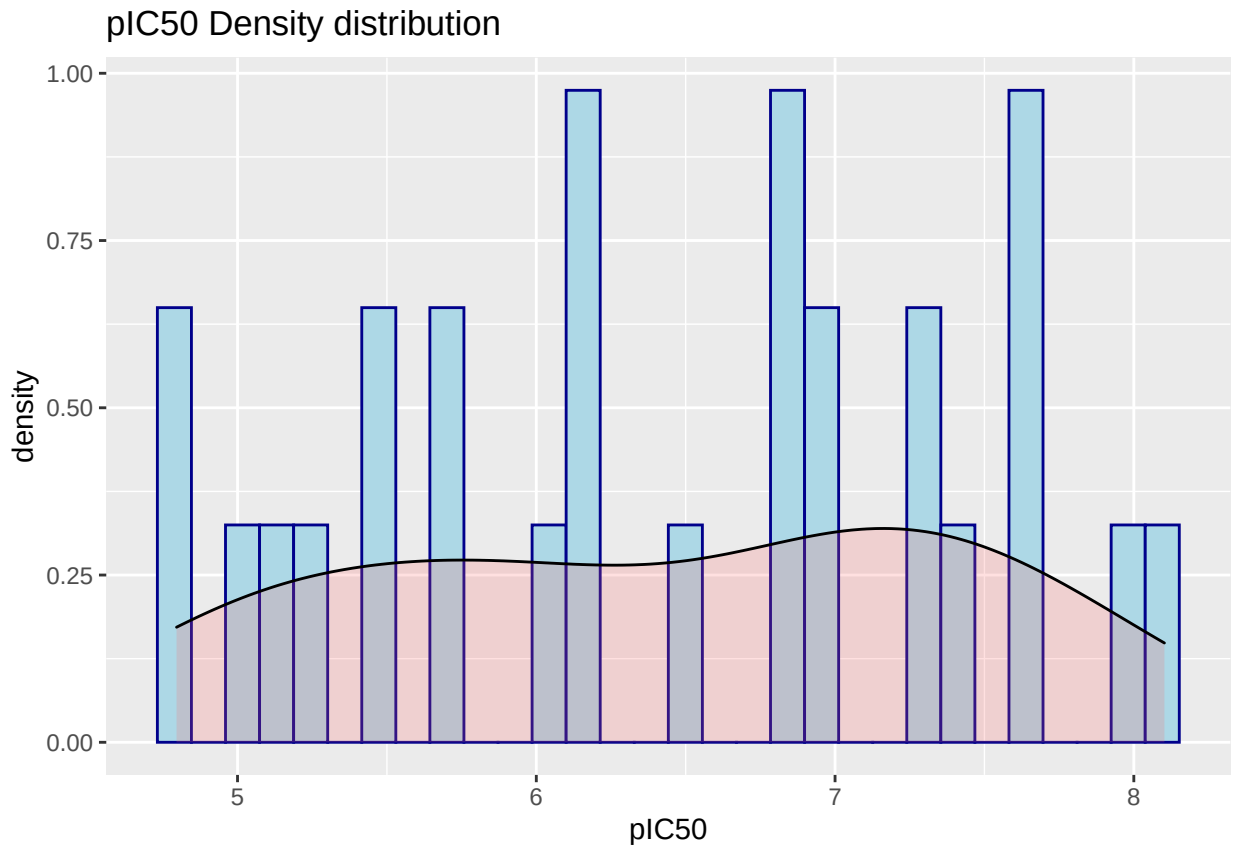
First, distribution of the response variable

```
library(ggplot2)
```

```
## Basic histogram

gg_bas <- ggplot(Dataset, aes(x = pIC50)) + geom_histogram(color = "darkblue", fill = "lightblue") +
    ggtitle("pIC50 distribution")
gg_bas
```




```
## Density histogram

gg_dens <- ggplot(Dataset, aes(x = pIC50)) + geom_histogram(aes(y = ..density..),
    color = "darkblue", fill = "lightblue") + geom_density(alpha = 0.2, fill = "#FF6666") +
    ggtitle("pIC50 Density distribution")
gg_dens
```



Apparently, the biological activity of the molecules is normally-flattened distributed. A Shapiro-Wilk statistical test will be carried out to get insight about the normality of this property. The null hypothesis states that the data come from a normal distribution. The alternative hypothesis states that the data do not come from a normal distribution. The test will be carried out with a significance level of 95% (p-value = 0.05)

```
shapiro.test(Dataset$pIC50)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Dataset$pIC50
## W = 0.94585, p-value = 0.1698
```

Since `p-value>0.05`, then the null hypothesis is accepted, so the data follow a normal distribution with a statistical significance level of 95%.

Molecular similarity of the set of ligands:

According to the structure-activity principle, there will be structura similarities between the molecules, since are related with a biological activity spectrum against HKII. Let's calculate a similarity index between molecules, specifically Tanimoto index based on atom-pairs distance, and after that, let's apply a clustering algorithm to validate the structural molecular space.

```
library(rcdk)
```

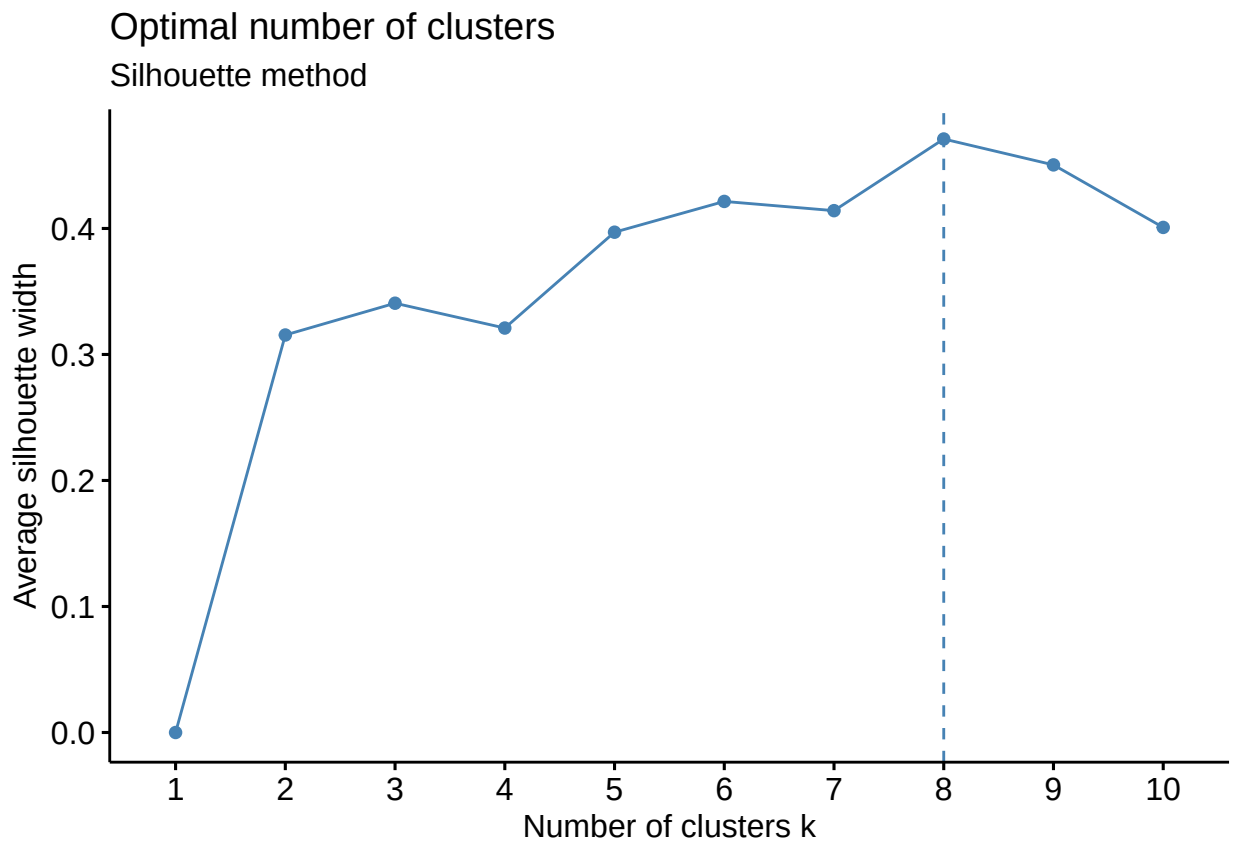
```
## Loading required package: rcdklibs
```

```
## Loading required package: rJava
```

```r
## Reading dataset in SMILES
Dataset_smi <- read.table("Dataset_smi.smiles", allowEscapes = FALSE)

## Parsing into CDK object
Dataset_smi <- as.vector(Dataset_smi$V1)
gluc_amine_ens <- parse.smiles(Dataset_smi)

## Estimating PubChem fingerprints
fps_pubch <- lapply(gluc_amine_ens, get.fingerprint, type = "pubchem")

## Calculating pairwise similarity matrix between ligands according to Tanimoto
## index, and then, the distance matrix.
fp.sim <- fingerprint::fp.sim.matrix(fps_pubch, method = "tanimoto")
fp.dist <- 1 - fp.sim
```

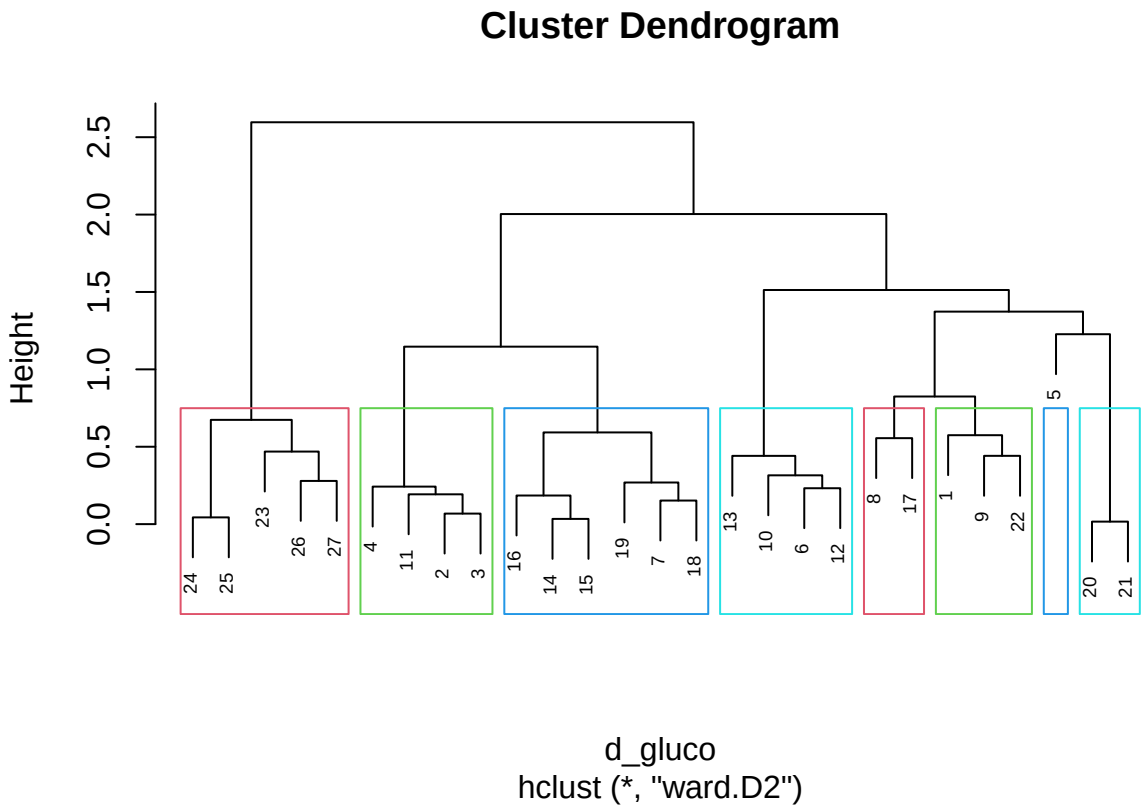
After getting the distance matrix between all the ligands according to Tanimoto index (that was calculated using PubChem fingerprints), let's do a clustering analysis in order to get insight about molecular similarity. First, one needs to estimate the optimal number of clusters.

```r
## Optimal number of clusters
library(factoextra)
fviz_nbclust(fp.dist, kmeans, method = "silhouette") + labs(subtitle = "Silhouette method")
```



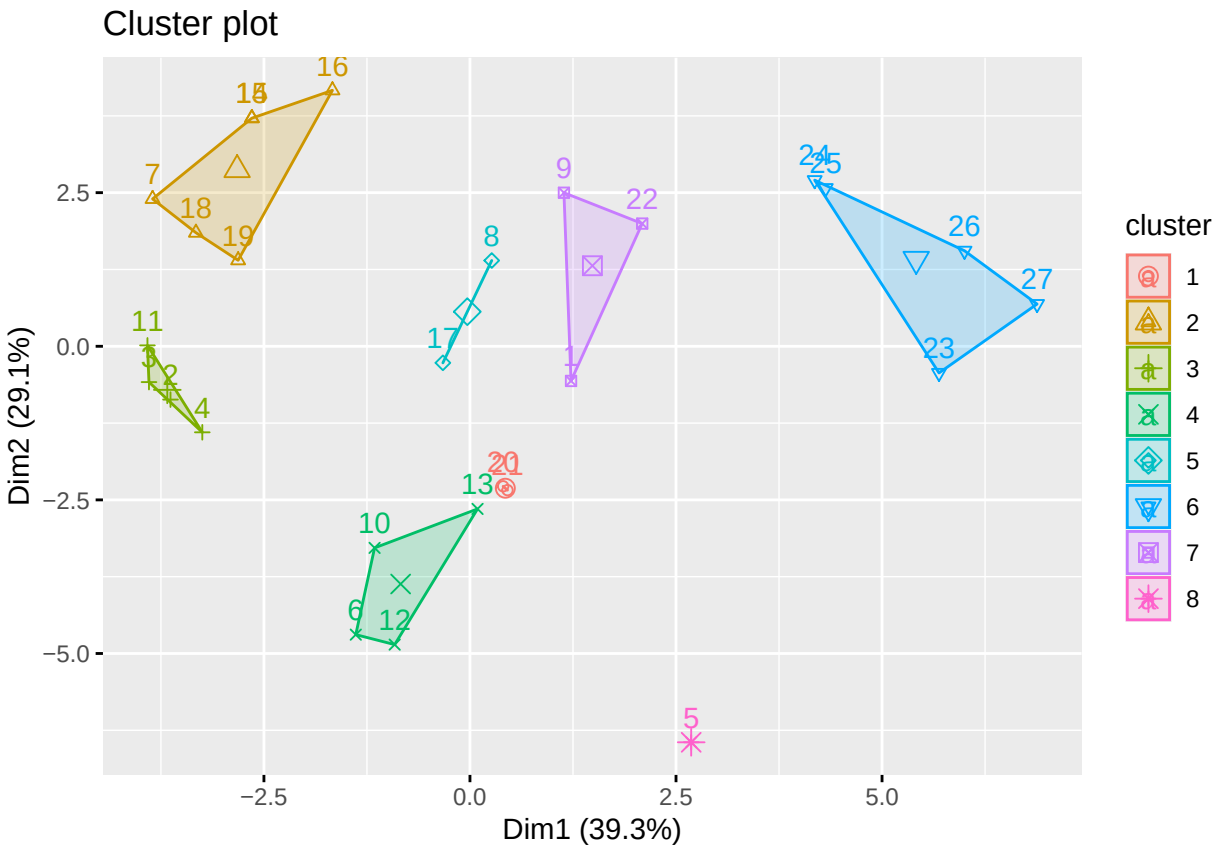
So, with an optimal number of eight clusters, let's do a hierarchical clustering algorithm according to distance matrix.

```
d_gluco <- dist(fp.dist, method = "euclidean")
res.hc_gluco <- hclust(d_gluco, method = "ward.D2")
grp_gluco <- cutree(res.hc_gluco, k = 8)

plot(res.hc_gluco, cex = 0.6)  # plot tree
rect.hclust(res.hc_gluco, k = 8, border = 2:5)
```



Let's compare this molecular distribution with a K-means cluster algorithm.

```
kmeans_gluco <- kmeans(fp.dist, centers = 8, nstart = 25)
fviz_cluster(kmeans_gluco, data = fp.dist)
```



It seems that the biological activity space determined by the structure of the ligands is heterogeneous, so there are subgroups of molecules with similar structural topology according to its inhibitory activity. One can argue that the flattened distribution gives us a wide range of action in the modelling.

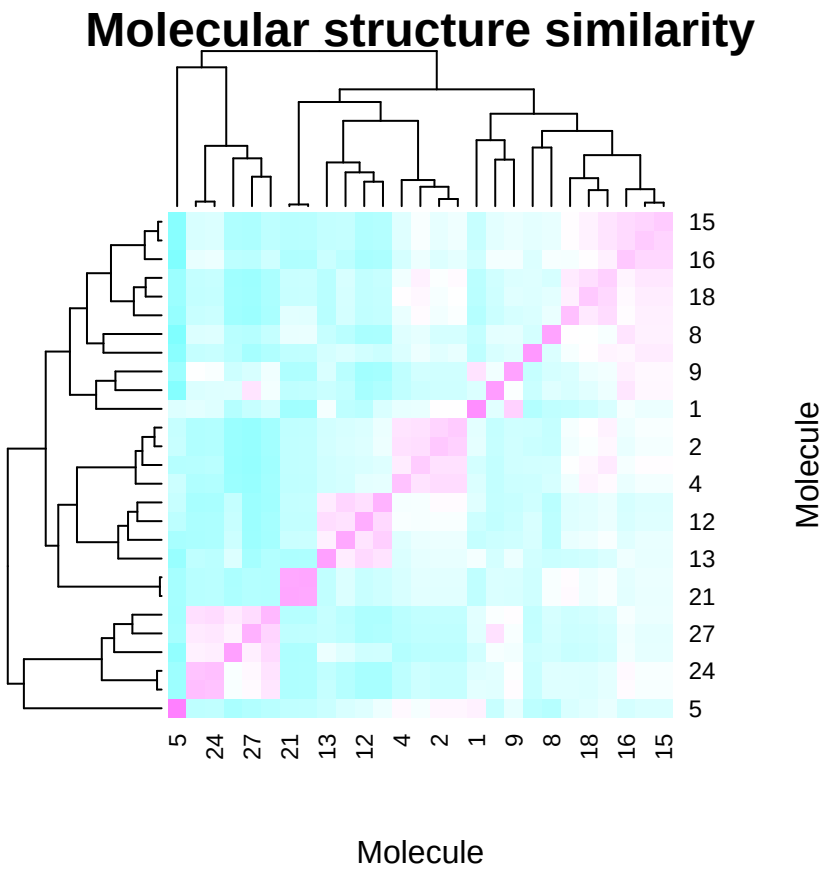
Structure-activity similarity maps:

One last chemoinformatic approach to detect influence from the topology of the molecule in the biological activity is the SAS map. According to [], let's plot a similarity index (In this case, Tanimoto index for each pair of compounds) against the difference between biological activity for each pair of compounds. With this graph, one car argue about the structure activity dependence and activity cliffs in the dataset.

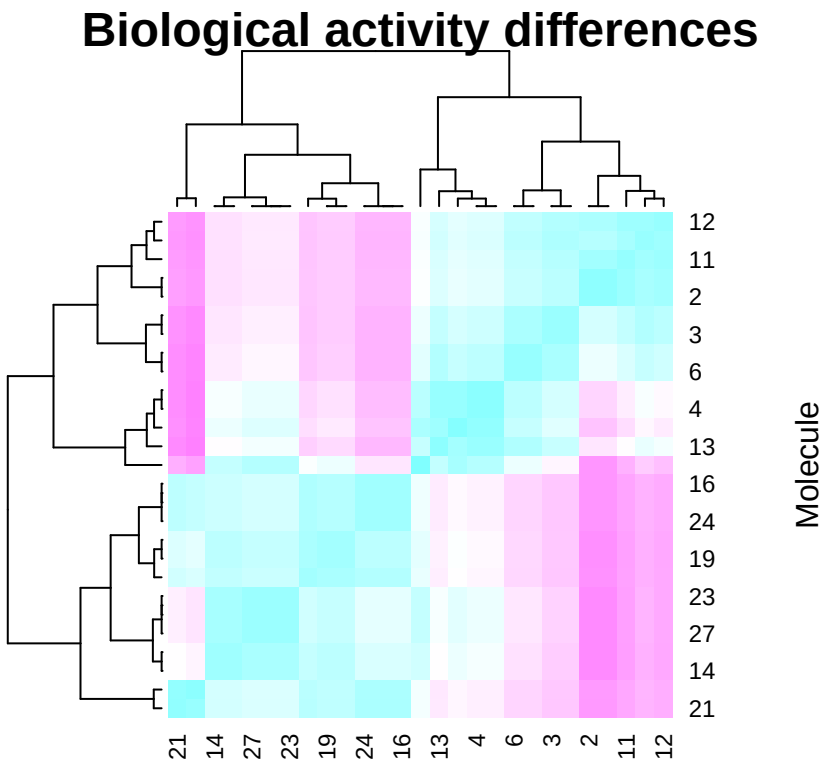
```
## Estimation of difference of biological activity
Activi_lands = t(c(Dataset$pIC50))

delt_act <- list()
for (i in 1:ncol(Activi_lands)) {
    delt_act[[i]] = abs(Activi_lands - Activi_lands[, i])
}

delt_act_mtx <- do.call(rbind, delt_act)
```

Now, let's plot a heatmap for the similarity matrix between compounds, and another for the difference between biological activities for each compound. Finally the SAS map

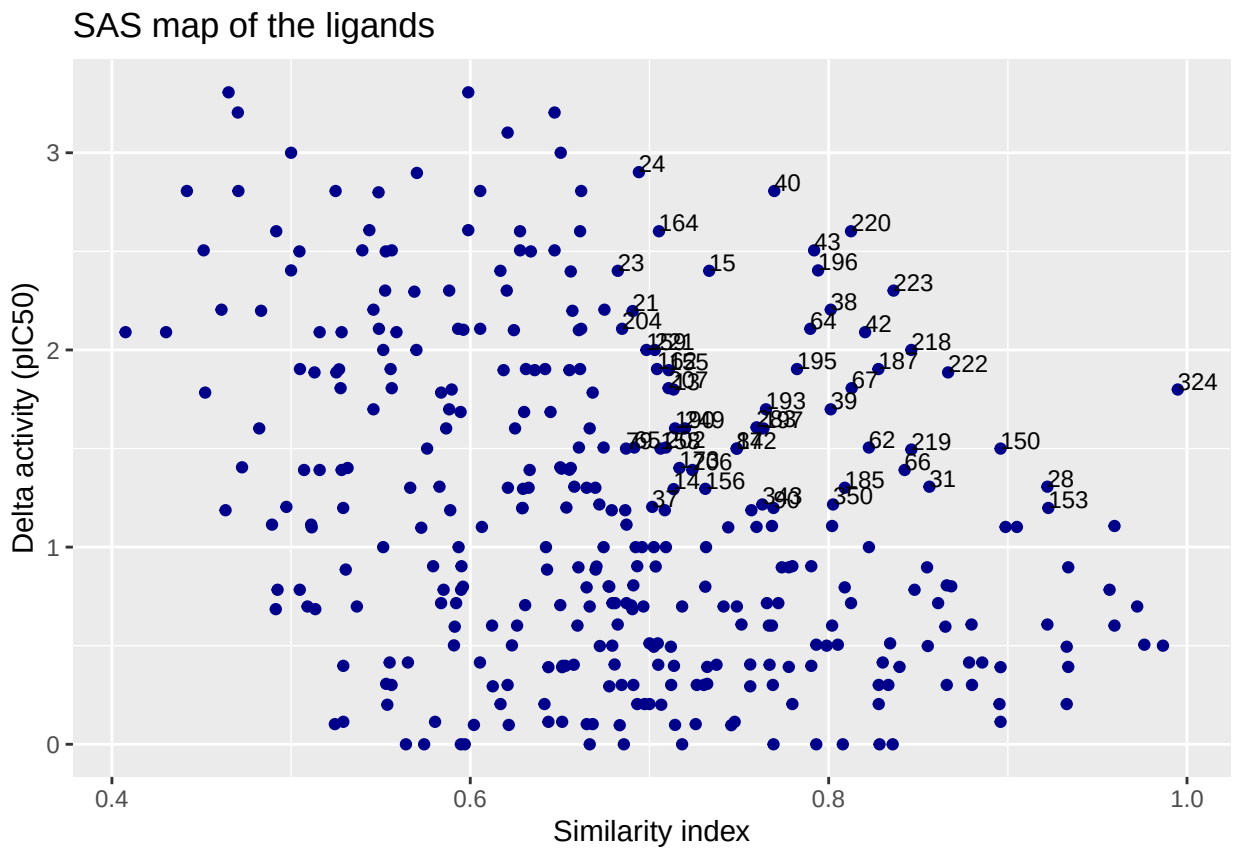
```
heatmap(fp.sim, col = cm.colors(256), main = "Molecular structure similarity", xlab = "Molecule",
    ylab = "Molecule")
```




```
heatmap(delt_act_mtx, col = cm.colors(256), main = "Biological activity differences",
    xlab = "Molecule", ylab = "Molecule")
```




```
## SAS map

Tanimoto_pair <- fp.sim[lower.tri(fp.sim)]
Activity_pair <- delt_act_mtx[lower.tri(delt_act_mtx)]
SAS_pair <- as.data.frame(cbind(Tanimoto_pair, Activity_pair))


## Limits for SAS maps

fp_limit <- mean(Tanimoto_pair)
act_limit <- mean(Activity_pair)

library(ggplot2)

ggplot(SAS_pair, aes(x = Tanimoto_pair, y = Activity_pair)) + geom_point(color = "darkblue") +
    geom_text(aes(label = ifelse(Activity_pair > act_limit & Tanimoto_pair > fp_limit,
        as.character(row.names(SAS_pair)), "")), hjust = 0, vjust = 0, size = 3) +
    ggtitle("SAS map of the ligands") + xlab("Similarity index") + ylab("Delta activity (pIC50)")
```



So, according to the above plots, there are some pair compounds that relies on the activity cliff space, ie, the molecules has a high similarity structure but a dissimilar activity. The most prominent outlier is the pair 324. Let's see which molecule pairs correspond to it.

```
Outlier_SAS <- SAS_pair[324, 1]

Par_molecule_out <- which(fp.sim == Outlier_SAS, arr.ind = TRUE)
Par_molecule_out
```

```
##      row col
## [1,]  21  20
## [2,]  20  21
```

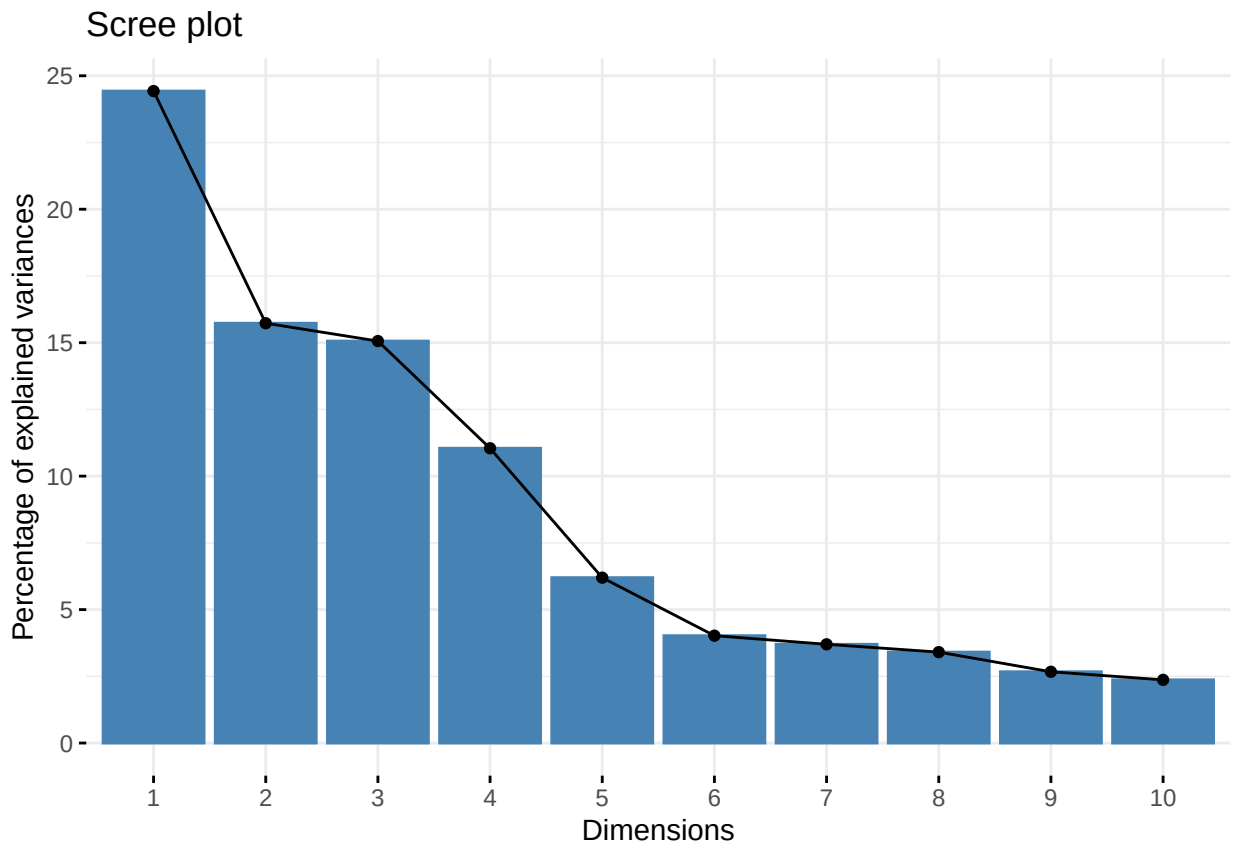
Compound 20 and 21 are apparently activity cliffs. For the sake of QSAR modeling, these kind of phenomena should be avoided, but from a pharmacophore point of view, the activity cliff molecules are valuable information about the chemical fuctional groups that give the biological activity.All the compounds will be keep in the dataset.

Principal component analysis (PCA) over the dataset to determine the chemical space:

Let's carry out a PCA analysis in order to get insights about the chemical space determined by the molecular descriptors in the dataset. With this analysis, one can project the entire dataset (27 x 2284) into two or three latent variables (PC) that explains the majority of the variance.

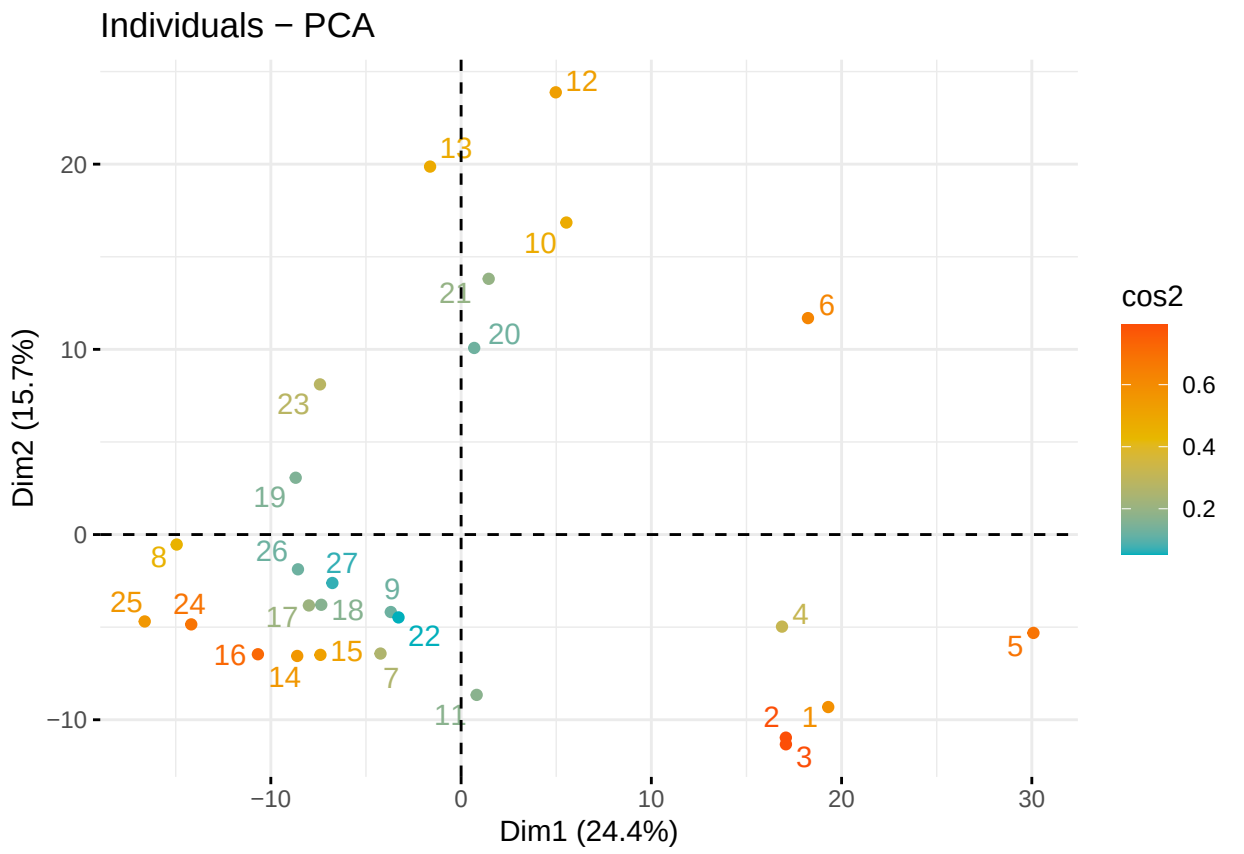
```
library(factoextra)

## Cleaning NA's
Dataset <- Dataset[, !apply(Dataset, 2, function(x) any(is.na(x)))]
```

```r
## Cleaning columns with zero variance
Dataset <- Dataset[, apply(Dataset, 2, var, na.rm = TRUE) != 0]

## Cleaning columns with constant value
Dataset <- Dataset[, !apply(Dataset, 2, function(x) length(unique(x)) == 1)]

## Building PCA model
Dataset_pca <- prcomp(Dataset, scale = TRUE)

## Percentage of variance explained for each PC Eigenvalues (Scree plot)
fviz_eig(Dataset_pca)
```



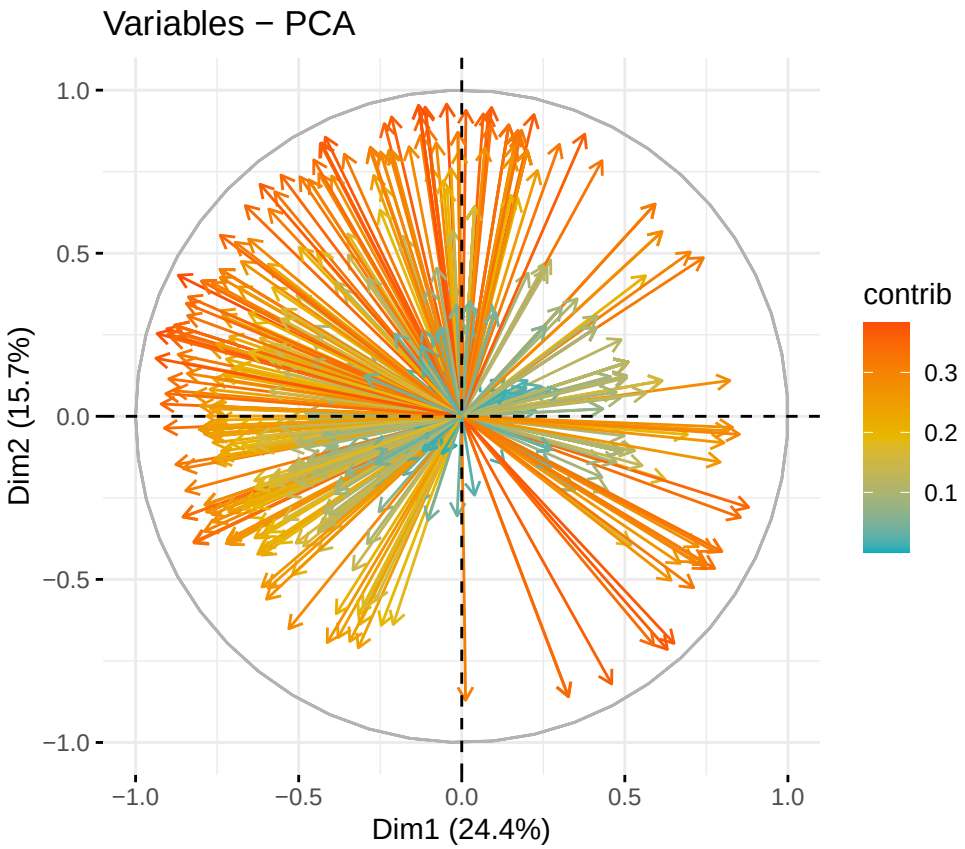
According to the Scree plot, the three PC explained around 55% of the variance of the data.

```r
fviz_pca_ind(Dataset_pca, repel = TRUE, col.ind = "cos2", gradient.cols = c("#00AFBB",
    "#E7B800", "#FC4E07"))
```



The projection of the molecules into the PC space gives us approximately four clusters (a result that is according to the cluster analysis from Tanimoto index). It seems that molecules five and six are some kind of outliers

```
fviz_pca_var(Dataset_pca, col.var = "contrib", gradient.cols = c("#00AFBB", "#E7B800",
    "#FC4E07"), repel = TRUE, label = c("ind"))
```



So apparently the chemical space determined by the 613 molecular descriptors (after cleaning of the data) is very vast and a corrrelation analysis can't be carry out. Let's select the molecular descriptors that contribute to the majority of the PC, since that features should be the most important in the variance

```
## Getting features that have the most weight in the First six principal
## components
Dataset_pca_data <- get_pca_var(Dataset_pca)
contribution_pca <- (Dataset_pca_data$contrib)

print((max.col(contribution_pca))[1:6])
```

```
## [1] 20  1  1  3  1 27
```

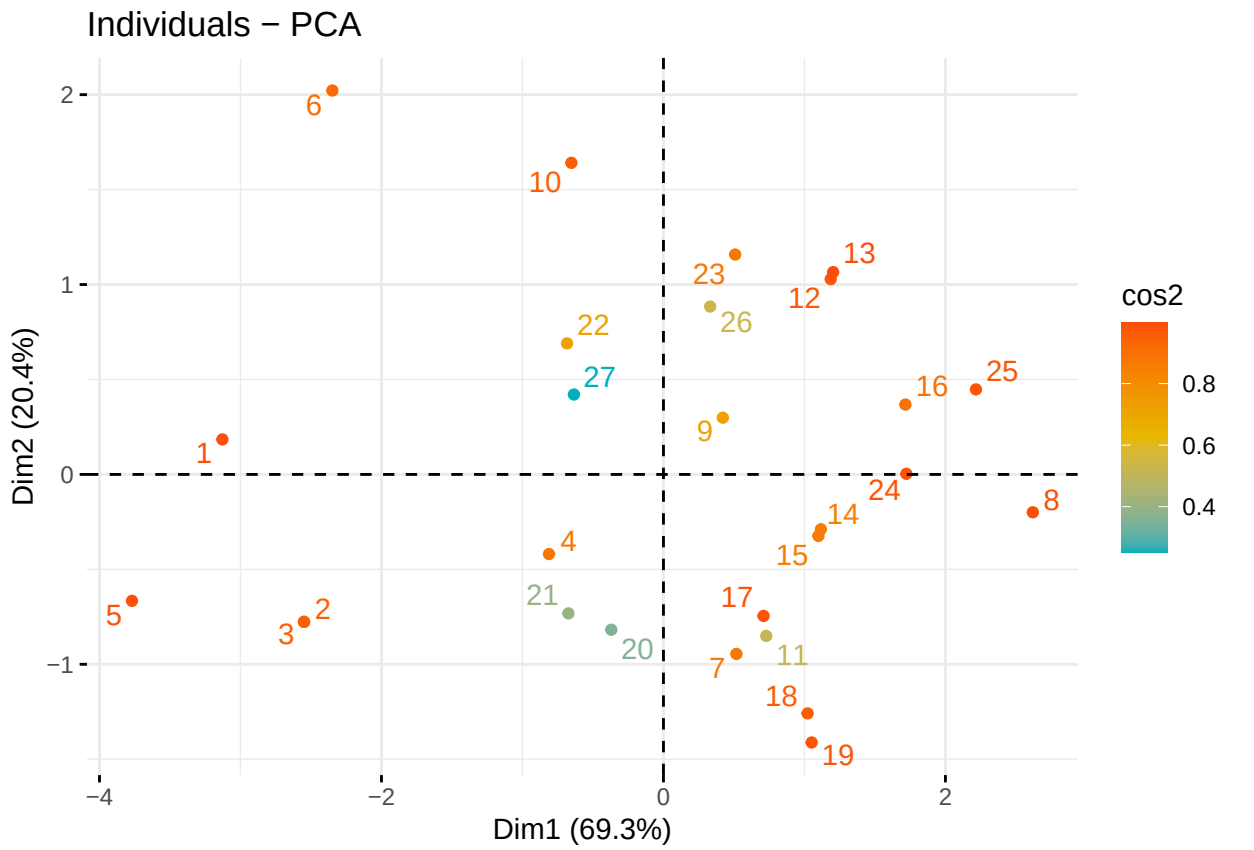
```
## Extracting the features

print(names(Dataset[, c(1, 3, 20, 27)]))
```

```
## [1] "ALogPS_logP" "MW"          "SCBO"        "nC"
```

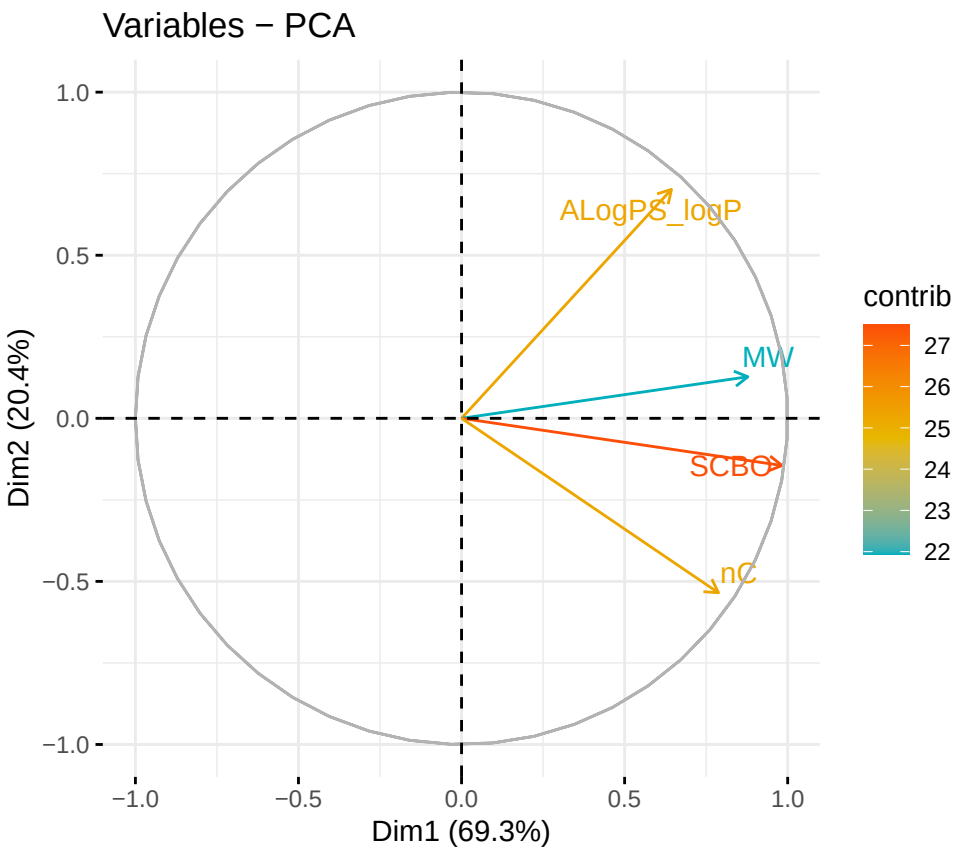
The octanol/water partition coefficient (ALogPS_logP), molecular weight (MW), number of carbons in the structure (nC) and sum of conventional bond orders (H-depleted) (SCBO) are the most weightened features to the PC's. Let's rerun the PCA analysis only with these features in order to get a better insight of the chemical space of the ligands.

```
## Building PCA model with the features
Dataset_features <- prcomp(Dataset[, c(1, 3, 20, 27)], scale = TRUE)

## Scores plot of the compounds in the PC space
```

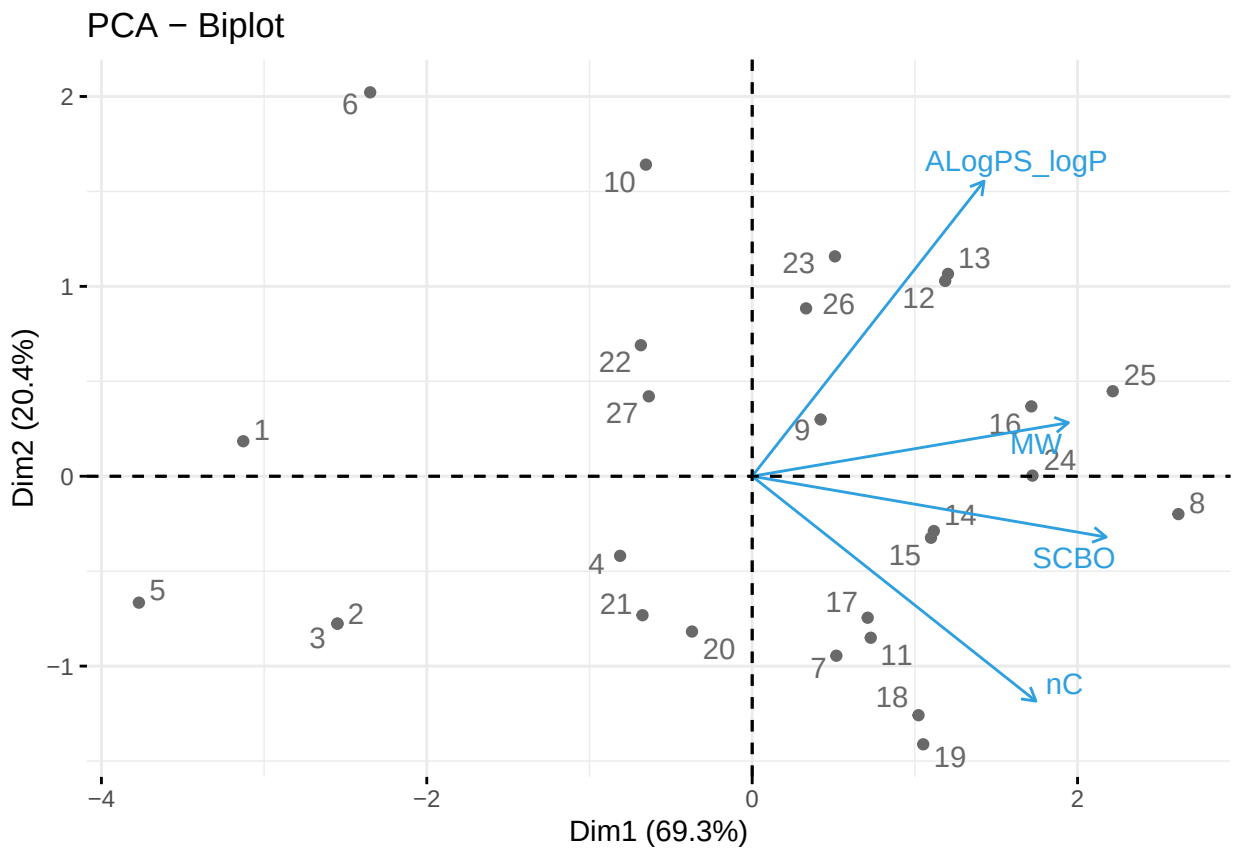
```
fviz_pca_ind(Dataset_features, repel = TRUE, col.ind = "cos2", gradient.cols = c("#00AFBB",
    "#E7B800", "#FC4E07"))
```




```
## Loadings plot of the features selected

fviz_pca_var(Dataset_features, col.var = "contrib", gradient.cols = c("#00AFBB",
    "#E7B800", "#FC4E07"), repel = TRUE)
```




```
## Biplot of the PCA model

fviz_pca_biplot(Dataset_features, repel = TRUE, col.var = "#2E9FDF", col.ind = "#696969")
```



Summary, the four features are positive correlated so there is a total weight to that chemical space. Also, approximately half of the dataset rests in the space determined by these descriptors.

References

Lin, Hong, Jin Zeng, Ren Xie, Mark J. Schulz, Rosanna Tedesco, Junya Qu, Karl F. Erhard, et al. 2016. "Discovery of a Novel 2,6-Disubstituted Glucosamine Series of Potent and Selective Hexokinase 2 Inhibitors." *ACS Medicinal Chemistry Letters* 7 (3): 217–22. https://doi.org/10.1021/acsmedchemlett.5b00214.