rm(list = ls())

library(ggplot2)

library(ggrepel)

# Original data

x <- c(-1149.5,-680.6,-131.9,-1075,-760.6,-156.2,-226.7,-150,-144.7,-3455.6,-482,-515.6,-456.4,-1290.2,-168,-177.5,-1111,-226.3,-1114.4,-324.3,-232.2,-164.4,-1089.4,-991)

y <- c(718.71, 518.35, 97.27, 1013.12, 643.41, 160.33, 206.26, 72.09, 99.32, 3325.04, 294.18, 328.73, 264.82, 1360.25, 155.36 ,153.54, 1025.21, 137.52, 615.36, 246.17, 173.78, 121.97, 641.70, 947.04)

# Labels for outliers

labels <- c("EAA1 (Canonical)", "A0A7P0Z4R4", "A0A7P0T9Z4", "A0A087X0U3", "A0A7P0T8Q1", "E7EUV6", "E7EUS7", "A0A7P0T9A4", "A0A7P0T807",

"EAA2 (Canonical)", "C9J9N5", "A0A2R8Y4D1", "A0A2R8Y642", "A0A2R8Y4N0", "A0A2R8YHI4", "H0YEB1",

"EAA3 (Canonical)", "H0Y7R2",

"EAA4 (Canonical)", "M0R106", "M0R2V7", "M0QY32",

"EAA5 (Canonical)", "F1T0D4")

# Create a data frame

df <- data.frame(x = x, y = y, label = labels)

# Calculate Pearson correlation

cor\_result <- cor.test(x, y, alternative = "two.sided", method = "pearson", exact=FALSE)

print(cor\_result)

# Function to identify pair outliers using Tukey's fence method

identify\_outliers <- function(x, y, k = 1.5) {

# Calculate Mahalanobis distances

center <- c(mean(x), mean(y))

cov\_matrix <- cov(cbind(x, y))

mahalanobis\_dist <- mahalanobis(cbind(x, y), center, cov\_matrix)

# Calculate Tukey's fences

q1 <- quantile(mahalanobis\_dist, 0.25)

q3 <- quantile(mahalanobis\_dist, 0.75)

iqr <- q3 - q1

lower\_fence <- q1 - k \* iqr

upper\_fence <- q3 + k \* iqr

# Identify outliers

outliers <- mahalanobis\_dist > upper\_fence | mahalanobis\_dist < lower\_fence

return(outliers)

}

# Identify outliers

df$outlier <- identify\_outliers(df$x, df$y)

# Print number of outliers

num\_outliers <- sum(df$outlier)

cat("Number of pair outliers detected:", num\_outliers, "\n")

# Fit linear regression model

lm\_model <- lm(y ~ x, data = df)

# Create ggplot

p <- ggplot(df, aes(x = x, y = y)) +

geom\_point(aes(color = outlier, shape = outlier), size = 4) + # Smaller data points

geom\_smooth(method = "lm", se = FALSE, color = "darkgreen", linetype = "solid", size = 0.5) + # Continuous line

scale\_color\_manual(values = c("steelblue", "red"),

labels = c("Normal", "Outlier"),

name = "Data Points") +

scale\_shape\_manual(values = c(16, 17),

labels = c("Normal", "Outlier"),

name = "Data Points") +

labs(title = " ",

x = " **RNA structure MFE (kcal/mol)**", y = "**Ensemble Diversity**") +

theme\_minimal(base\_size = 13) +

theme(

legend.position = c(1, 1), # Move legend to upper right corner

legend.justification = c(1, 1), # Align legend to top-right

legend.box.just = "right",

legend.margin = margin(2, 2, 5, 5),

legend.background = element\_rect(fill = "white", color = "black", size = 0.5),

panel.grid.major = element\_line(color = "gray80", size = 0.2),

panel.grid.minor = element\_line(color = "gray90", size = 0.05),

axis.line.x = element\_line(size = 0.5), # Thicker x-axis line

axis.line.y = element\_line(size = 0.5), # Thicker x-axis line

axis.ticks.y = element\_line(size = 0.5), # Thicker x-axis ticks

axis.ticks.x = element\_line(size = 0.5), # Thicker x-axis ticks

axis.title.x = element\_text(size = 16), # Enlarged X axis label

axis.title.y = element\_text(size = 16), # Enlarged Y axis label

axis.text.x = element\_text(size = 14), # Enlarged X axis numbers

axis.text.y = element\_text(size = 14), # Enlarged Y axis numbers

legend.text = element\_text(size = 15) # Larger legend text

)

# Add labels to outliers using ggrepel

p <- p + geom\_text\_repel(

data = subset(df, outlier),

aes(label = label),

box.padding = 0.5,

point.padding = 0.5,

force = 2,

max.overlaps = Inf,

size = 4

)

# Add regression equation

eq <- paste0("y = ", round(coef(lm\_model)[2], 4), "x + ", round(coef(lm\_model)[1], 4))

p <- p + annotate("text", x = -Inf, y = Inf, label = eq, hjust = -1.1, vjust = 16, size = 4)

# Extend the plot limits (less than before)

x\_range <- max(df$x) - min(df$x)

y\_range <- max(df$y) - min(df$y)

p <- p + coord\_cartesian(

xlim = c(min(df$x) - 0.01\* x\_range, max(df$x) + 0.01 \* x\_range),

ylim = c(min(df$y) - 0.01 \* y\_range, max(df$y) + 0.01 \* y\_range)

)

# Add Pearson correlation to the plot

cor\_text <- paste("Pearson correlation:", round(cor\_result$estimate, 4))

p <- p + annotate("text", x = -Inf, y = Inf, label = cor\_text, hjust = -0.8, vjust = 12.5, size = 4.5)

# Display the plot

print(p)