

```
#####  
##  
##### INSIGHT REPORT  
#####  
#####  
##
```

==== Endpoint Distribution Warning =====

Data sources with a significantly different logCL distribution:

1) Astrazeneca --> Mean difference: 0.724

***Figure reference: logCL Distribution Across Data Sources (Violin Plots) [Fig 3.2].

==== Value Range Warning =====

=====

No data source shows an inconsistent logCL value range (> 1 order of magnitude).

==== Skewed Distribution Warning =====

Data sources with a skewed logCL distribution:

1) GombardHall --> Skewness: -1.08 (Moderate Left Skewed)

2) Varma2010 --> Skewness: -1.02 (Moderate Left Skewed)

***Figure reference: logCL Distribution Across Data Sources (Overlaid Histogram) [Fig 3.1].

==== Outliers Identification Warning =====

Data sources with a significant proportion of outliers ($\geq 1\%$):

1) GombardHall --> Outliers Proportion: 1.52% (8 outliers)

2) Varma2010 --> Outliers Proportion: 1.31% (4 outliers)

3) Varma2009 --> Outliers Proportion: 1.29% (5 outliers)

4) Lombardo --> Outliers Proportion: 1.27% (17 outliers)

5) Obach --> Outliers Proportion: 1.21% (8 outliers)

6) Iwata --> Outliers Proportion: 1.08% (8 outliers)

***Figure reference: Visualization of logCL Outliers [Fig 1].

==== Abnormal Data Identification Warning =====

Data sources with abnormal values:

1) Iwata --> Abnormals Proportion: 0.13% (1 data points)

2) GombardHall --> Abnormals Proportion: 0.19% (1 data points)

***Figure reference: Visualization of logCL Outliers [Fig 1].

==== Dissimilar Datasets Warning =====

No data source is dissimilar in terms of the molecular feature profile.

==== Conflicting Datasets Warning =====

No data source pair has conflicting logCL value annotations for shared molecules.

==== Divergent Datasets Warning =====

Data sources which are divergent due to a low overlap with the rest of sources (<10% shared molecules):

1) Astrazeneca --> Percentage of common molecules: 6.45%

***Figure reference: Between-Source Discrepancies [Fig 9].

===== Redundant Datasets Warning

=====

Data source pairs which are redundant due to significant proportion of shared molecules (>80% molecules in common):

1) Varma2009 vs. Obach --> Percentage of common molecules (from first source): 98.97%

2) Varma2010 vs. Obach --> Percentage of common molecules (from first source): 95.74%

3) GombarHall vs. Obach --> Percentage of common molecules (from first source): 93.9%

4) Varma2010 vs. Varma2009 --> Percentage of common molecules (from first source): 85.9%

5) Varma2010 vs. GombarHall --> Percentage of common molecules (from first source): 82.62%

6) Varma2009 vs. GombarHall --> Percentage of common molecules (from first source): 81.65%

7) Iwata vs. Lombardo --> Percentage of common molecules (from first source): 81.11%

***Figure reference: Between-Source Discrepancies [Fig 9].

===== Reference Molecules Underrepresentation Warning =====

Data sources with a low proportion of reference molecules (<15%):

1) Astrazeneca --> Percentage of reference molecules: 9.68% (72 molecules)

***Figure reference: Comparison of logCL Distribution for Reference vs. Non-Reference Molecules Across Data Sources [Fig EXTRA].