

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

Endpoint Distribution Warning

Data sources with a significantly different logHL distribution:

1) Fan --> Mean difference: -0.923

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

Value Range Warning

Data sources with an inconsistent logHL value range (> 1 order of magnitude):

1) Lombardo --> Source's IQR: 0.82 vs. Other-sources' IQR: 1.04

2) Obach --> Source's IQR: 0.8 vs. Other-sources' IQR: 1.04

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

Skewed Distribution Warning

No data source shows a skewed logHL distribution.

Outliers Identification Warning

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

1) eDrug3D --> Outliers Proportion: 1.32% (17 outliers)

2) DDPD --> Outliers Proportion: 1.22% (15 outliers)

3) Obach --> Outliers Proportion: 1.06% (7 outliers)

4) Fan --> Outliers Proportion: 1.05% (27 outliers)

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

Abnormal Data Identification Warning

Data sources with abnormal values:

1) eDrug3D --> Abnormals Proportion: 0.08% (1 data points)

***Figure reference: Visualization of logHL 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 logHL 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) Fan --> Percentage of common molecules: 4.24%

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

===== Redundant Datasets Warning

=====

No data source pair has a significant percentage of shared molecules (greater than 80%).

===== Reference Molecules Underrepresentation Warning

=====

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

1) Fan --> Percentage of reference molecules: 2.37% (61 molecules)

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