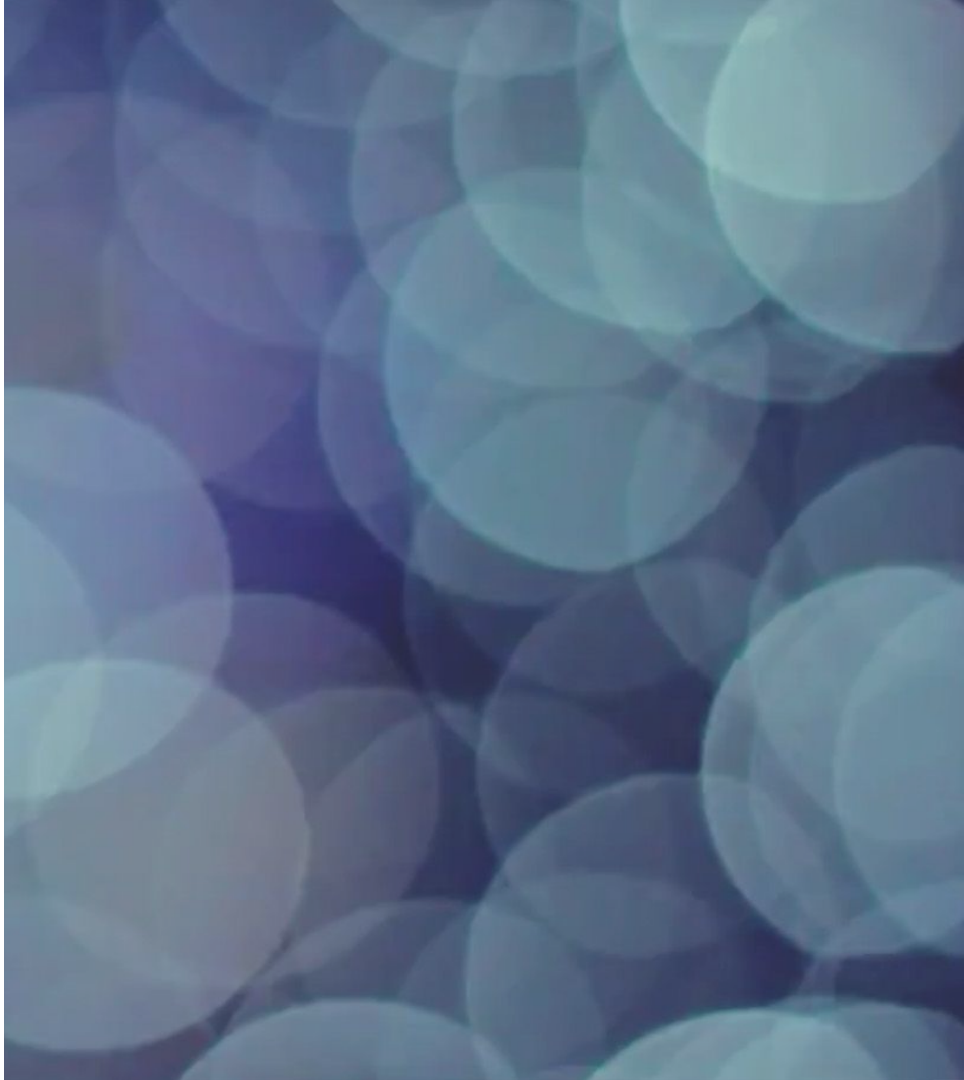
Physicochemical & Pharmacokinetic Dataset Mining and Insights

Taylor Townes

14JUN24



Brief Data Set Descriptions -

- Three related data sets that include physicochemical, pharmacokinetic, gene, and target assay fields:

Compound_Off_Target_Activity

- This dataset provides a detailed view of the off-target activities of various compounds, helping in the assessment of compound specificity and potency.
- It can be used to identify potential off-target effects.

Gene_Drug_Adverse_Event_Relationships

- The dataset provides a comprehensive view of the complex interactions between genes, drugs, and adverse events.
- It can be used to understand which genes are associated with certain adverse events when specific drugs are administered, helping in the assessment of drug safety and the identification of potential side effects based on genetic information.

Project_Level_Data

- This dataset holds the compounds' properties & activities in relation to the primary target.
- It helps to understand the compounds' physicochemical, bioactive, and pharmacokinetic properties.

Data Preprocessing

Data Preprocessing -

1) Combined columns with same values - MW & Molecular_Weight.
2) Checked the proportion of null values for each field in each dataset.

Proportion of null values in Compound_Off_Target_Activity:

CompoundID	0.0
Gene Target	0.0
pIC50	0.0

Proportion of null values in Gene_Drug_Adverse_Event_Relationships:

GeneSymbol	0.006119
Ensembl ID	0.338218
EntrezGene	0.281146
ae	0.000000
thresholdset	0.000000
gene_count	0.000000
drug_count	0.000000
ae_count	0.000000
drugs_with_ae	0.000000
bioactive_drugs	0.000000

Proportion of null values in Project_Level_Data:

CompoundID	0.000000
Primary_Target_Assay	0.000000
Primary_Target_Assay_BioActivity	0.010196
TPSA	0.000000
ClogP	0.000000
LogD	0.010196
Num_H_Donors	0.000000
Num_H_Acceptors	0.000000
Num_AromaticRings	0.000408
F_SP3	0.000408
Drug_Class	0.076672
Cell Permeability	0.219005
Cmpd Solubility (uM)	0.175775
fafg (Rat)	0.983279
Bioavailability (Rat)	0.970636
Clint,mic (L/hr/kg) (Rat)	0.210848
Molecular_Weight (amu)	0.000000

3) Mapped Gene-Ensembl ID-Entrez Gene ID relationships to fill missing data & pulled further missing data from web sources.
4) Removed duplicate rows from each dataset.

Compound_Off_Target_Activity:
Initial rows: 26358
Final rows: 25624
Rows removed: 734

Gene_Drug_Adverse_Event_Relationships:
Initial rows: 19449
Final rows: 19449
Rows removed: 0

Project_Level_Data:
Initial rows: 2452
Final rows: 2411
Rows removed: 41

Physicochemical Properties & Bioavailability Relationship Exploration

Physicochemical Properties & Bioavailability Relationship -

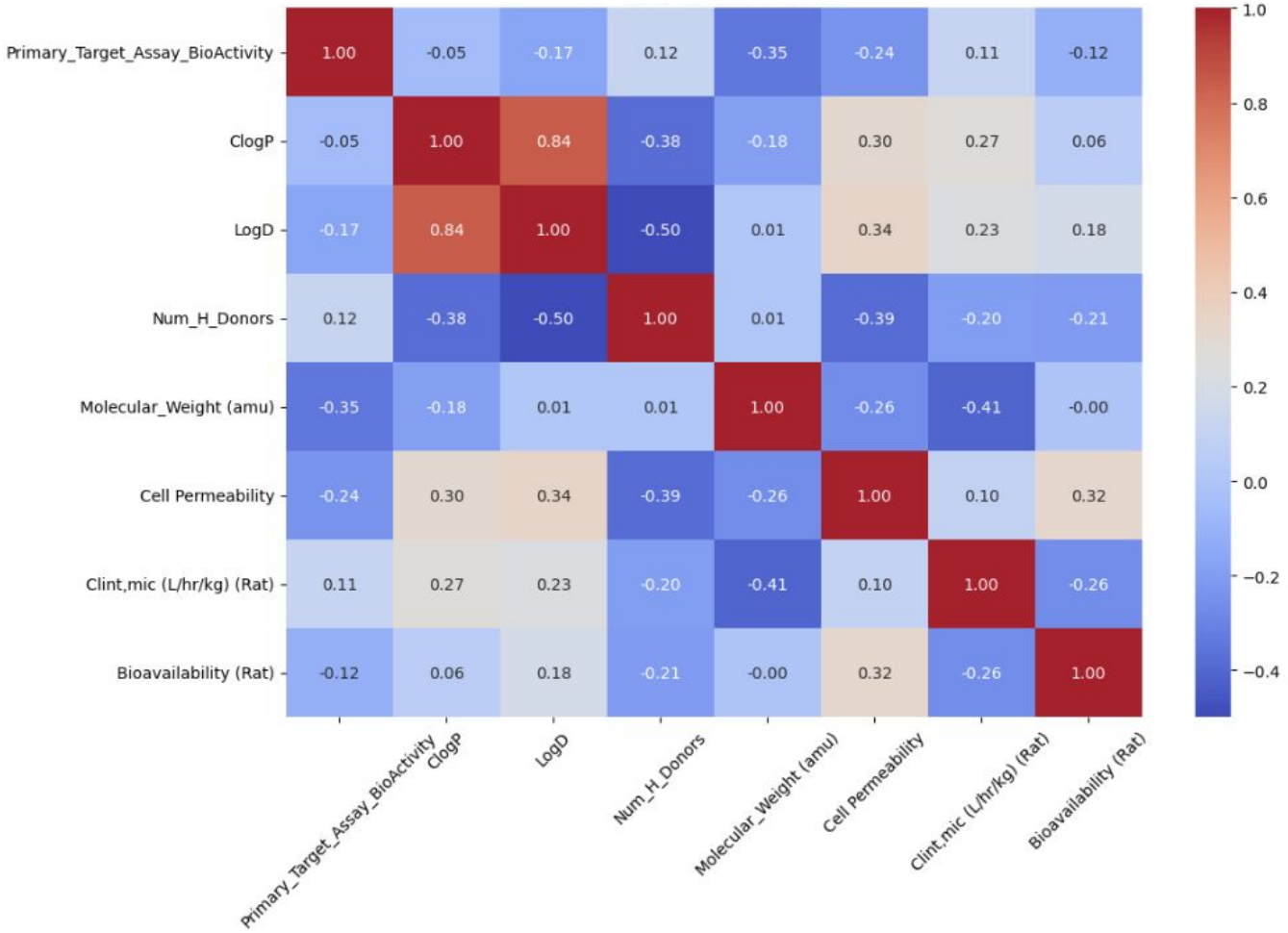
Expected:

- Positive Correlations:
  - Cell Permeability & Compound Solubility
- Negative Correlations:
  - TPSA, Number of H Donors, Number of H Acceptors, Molecular Weight, CLint

Observed:

Cell Permeability	0.225295
LogD	0.185580
Molecular_Weight (amu)	0.154249
Primary_Target_Assay_BioActivity	0.110425
ClogP	0.110376
Num_H_Acceptors	0.056352
Cmpd Solubility (uM)	0.039484
Num_AromaticRings	0.010859
TPSA	-0.039900
Num_H_Donors	-0.219144
Clint,mic (L/hr/kg) (Rat)	-0.262636

Physicochemical & Bioavailability Correlation Matrix



Predicting Bioavailability Using Highest Correlated Features -

- Used Linear Regression Model to Predict Bioavailability based on features.
- The R^2 value, 0.1029, shows a weak ability to accurately predict bioavailability based on these features.

```python
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error

# Initialize and train the model
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)

# Make predictions
y_pred = linear_model.predict(X_test)

# Evaluate the model
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)

print(f'R^2 Score: {r2}')
print(f'Mean Squared Error: {mse}')
```

```
R^2 Score: 0.10288663090476668
Mean Squared Error: 644.5064617898353
```

Predicting Bioavailability Using Highest Correlated Features -

- Used Gradient Boosting, Lasso & Ridge Regression, & Hyperparameter Tuning to achieve a more accurate prediction model.

```
Gradient Boosting R^2 Score: -0.5073564540307123
Gradient Boosting Mean Squared Error: 1082.91884650342

Best Ridge Parameters: {'alpha': 10}
Best Lasso Parameters: {'alpha': 0.01}
Ridge Regression R^2 Score: 0.09620549983309667
Ridge Regression Mean Squared Error: 649.3063369183254
Lasso Regression R^2 Score: 0.10283045181077644
Lasso Regression Mean Squared Error: 644.5468221170153
```

- The R^2 values did not improve from the original linear regression.
- This indicates that there ***is not*** a strong linear relationship between the bioavailability of the drugs and the features chosen, and this limited bioavailability data cannot easily be used to predict other compounds' bioavailability based on the presented physicochemical properties.

Predicting Bioavailability Insight Summary -

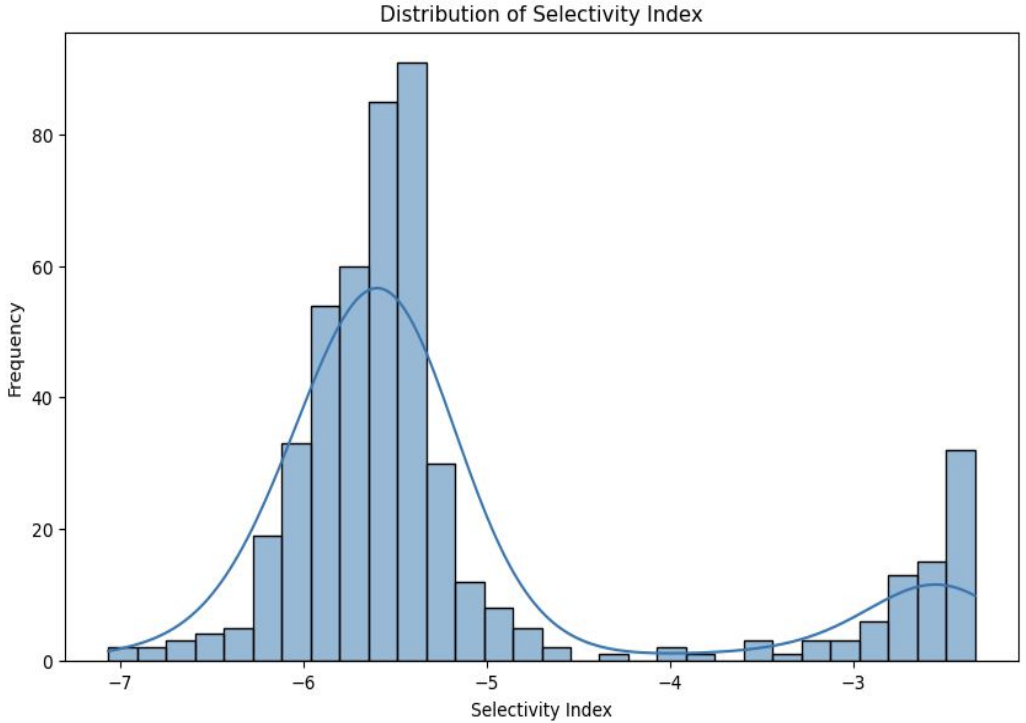
- The data set contains very little data for Bioavailability (~3% of compounds)
- Bioavailability does show weak to moderate correlations to a few physicochemical features.
- The relationship between the physicochemical features and bioavailability **is not** strong enough in this data set to use in predictive models to fill in missing bioavailability data or predict future compounds' bioavailability. More data is needed.

Primary Potency and Selectivity (Off-Target Bioactivity)

Primary Potency and Selectivity (Off-Target Bioactivity) -

By linking off-target activities to adverse events associated with specific genes, we can explore how selectivity (or lack thereof) contributes to adverse events.

- The negative skew of the Selectivity Index indicates that many compounds in the dataset have lower selectivity.
- These compounds may be less potent against their primary target compared to their average potency against off-targets.
- This lower selectivity could lead to more adverse reactions in off-target genes.

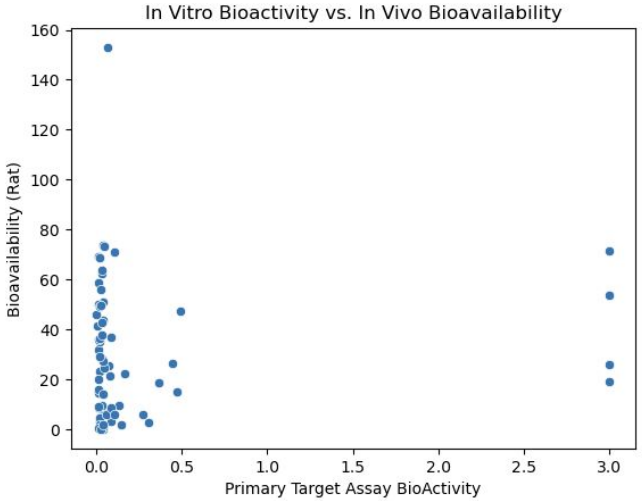


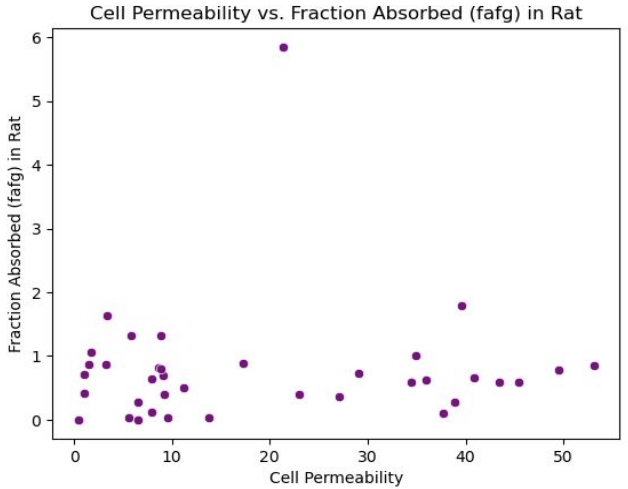
In Vitro to In Vivo Properties

In Vitro to In Vivo Properties -

We can compare in vitro properties like assay bioactivity, cell permeability, and solubility with in vivo properties such as bioavailability and F_aF_g (fraction absorbed).

- The high clustering near low bioactivity suggests that many compounds are potent in vitro but show a wide range of bioavailability in vivo.
- Both plots indicate a lack of strong correlation between the in vitro and in vivo properties analyzed. This suggests that factors other than those plotted might be influencing the in vivo outcomes.
- The presence of outliers in both plots warrants further investigation to understand the unique properties of these compounds that lead to their unusual bioavailability or absorption.





In Vitro to In Vivo Properties -

To gain deeper insights, we perform a multivariate regression analysis to see how multiple in vitro properties jointly predict in vivo outcomes:

- R^2 score for bioavailability was -0.54 & the R^2 score for F_aF_g was -5.4.
- This indicates that the models **do not** fit the data well & that the selected features **do not** explain the variation in bioavailability or F_aF_g .

```
# Prepare the feature set
features = pld_clean_df[['Primary_Target_Assay_BioActivity', 'Cell Permeability', 'TPSA', 'ClogP', 'LogD', 'Num_H_Donors',
                         'Num_H_Acceptors']]

# Prepare the target sets, dropping rows with NaN values
target_bioavailability = pld_clean_df['Bioavailability (Rat)'].dropna()
target_fafg = pld_clean_df['fafg (Rat)'].dropna()

# Ensure that the features and target arrays have the same number of samples
features_bioavailability = features.loc[target_bioavailability.index].dropna()
target_bioavailability = target_bioavailability.loc[features_bioavailability.index]

features_fafg = features.loc[target_fafg.index].dropna()
target_fafg = target_fafg.loc[features_fafg.index]

# Split data for bioavailability model
X_train_bio, X_test_bio, y_train_bio, y_test_bio = train_test_split(features_bioavailability, target_bioavailability,
                                                                    test_size=0.2, random_state=42)

# Split data for fafg model
X_train_fafg, X_test_fafg, y_train_fafg, y_test_fafg = train_test_split(features_fafg, target_fafg, test_size=0.2,
                                                                        random_state=42)

# Train models
model_bio = LinearRegression().fit(X_train_bio, y_train_bio)
model_fafg = LinearRegression().fit(X_train_fafg, y_train_fafg)

# Predict and evaluate
predictions_bio = model_bio.predict(X_test_bio)
predictions_fafg = model_fafg.predict(X_test_fafg)

r2_bio = r2_score(y_test_bio, predictions_bio)
r2_fafg = r2_score(y_test_fafg, predictions_fafg)
```

In Vitro to In Vivo Properties Next Steps -

- Exploring additional features or different sets of features that might have a more direct impact on bioavailability and absorption.

- Consider using more complex models like decision trees, random forests, or neural networks that can capture non-linear relationships.

- Incorporate domain knowledge to select more relevant features. For example, considering metabolic stability, specific enzyme interactions, or other pharmacokinetic properties that could influence bioavailability and absorption.

Next Steps

Data Exploration Next Steps -

1) Create a column with a binary key for neutral/ionizable in relation to the LogD & CLogP values.
2) Use external or other larger domain set that has more bioavailability, absorption, & clearance data to train an ML model to more accurately predict for this dataset.
3) Examine relationship between adverse events, selectivity, & potency of compounds.

Q&A

Appendix -

See GitHub for full code:
https://github.com/tktownes/AbbVie-Interview-Presentation