The datasets are modified to exclude the missing values before being passed into the algorithms to recognize patterns between those representations and their properties. This project attempts to reproduce and benchmark different methods across different datasets. The field is relatively new and there is still a need for good benchmarks and comparisons during the drug development process. The goal of the study is to reduce the need to synthesize a large number of compounds in a laboratory. The datasets contain properties such as water solubility, membrane permeability, and ability to inhibit HIV replication. To translate molecules into mathematical representations and apply different ML models, the parameters are optimized by cross-validation approaches. ROC-AUC is the metric used. The higher the y-axis, the better the model performs.

### What’s Next?
- If we can continue to optimize the performance of these models, we will be able to profile molecules and efficiently obtain their properties in the future without having to physically synthesize them in a laboratory.
- This work provides a good benchmark for future work.

### References
1. Shen et al., Drug Discovery Today Technologies, 12:30, 29–36
2. Wu et al., Journal of Cheminformatics 7(1), 530–538

**Dataset Preprocessing**
- Some rows of the original datasets contain missing values.
- The datasets are modified to exclude the missing values before being passed into the ML models.

**Exhaustive Search for Optimized Parameters**
- Hyperparameter tuning to determine the optimal values for the ML models.
- The parameters are optimized by cross-validated grid-search over a parameter grid.

**Performance Evaluation Metrics**
- **RMSE**: Root Mean Square Error
  \[ \text{RMSE} = \sqrt{\frac{1}{N} \sum (y - \hat{y})^2} \]
- **AUC**: Area under the curve of receiver operating characteristic curve (ROC-AUC)

**Classification and Regression Algorithms**
- **Classification**: DecisionTreeClassifier, SVC, RandomForestClassifier
- **Regression**: DecisionTreeRegressor, SVR, RandomForestRegressor

**Translation of SMILES to Fingerprints**
- Create fingerprints using SMILES
- Append fingerprints column to dataset

**Datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prediction Tasks</th>
<th>Tasks</th>
<th>Compounds</th>
<th>Properties</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSV</td>
<td>Virtual screening</td>
<td>17</td>
<td>10,127</td>
<td>Biophysics</td>
<td>classification</td>
</tr>
<tr>
<td>HIV</td>
<td>Ability to inhibit HIV replication</td>
<td>1</td>
<td>41,913</td>
<td>Biophysics</td>
<td>classification</td>
</tr>
<tr>
<td>BBBP</td>
<td>Permeability</td>
<td>1</td>
<td>2,053</td>
<td>Physiology</td>
<td>classification</td>
</tr>
<tr>
<td>Tox21</td>
<td>Toxicology measurements</td>
<td>12</td>
<td>8,014</td>
<td>Physiology</td>
<td>classification</td>
</tr>
<tr>
<td>MOE</td>
<td>Molecular and advanced drug reactions (MADR)</td>
<td>17</td>
<td>1,037</td>
<td>Physiology</td>
<td>classification</td>
</tr>
<tr>
<td>QM8</td>
<td>Electronic spectra and excited state energy</td>
<td>12</td>
<td>21,786</td>
<td>Quantum Mechanics</td>
<td>regression</td>
</tr>
<tr>
<td>DDL</td>
<td>Water solubility</td>
<td>1</td>
<td>1,328</td>
<td>Physical Chemistry</td>
<td>regression</td>
</tr>
<tr>
<td>Lipinski</td>
<td>Membrane permeability and solubility</td>
<td>1</td>
<td>8,200</td>
<td>Physical Chemistry</td>
<td>regression</td>
</tr>
</tbody>
</table>

**Preventative Measure Against Overfitting**
- Overfitting occurs when the model is trained so much to the existing data that it loses its ability to generalize and cannot perform accurately against unseen data.
- Cross-validation can be used to assess the performance of the model with an unknown dataset.
- The process is repeated for each fold in the dataset.

**Predictive Performance Comparison of Machine Learning Models**
- The predicted performance of machine-learning approaches. ROC-AUC is the metric used. The higher the y-axis, the better the model performs.
- Regression errors of machine-learning approaches. RMSE is the metric used. The lower the y-axis, the better the model performs.