# PyRegX: A Multivariate Regression Plug-in for PyMol

***PyRegX*** is a GUI-based plugin designed to perform robust ***Multiple Linear Regression (MLR)*** analysis on structured datasets directly from PyMol. Built with Customtkinter, it allows users to load datasets, define regression targets, validate models using statistical metrics, and export results-all without writing code.

## 🔍 Key Features

- 📂 Load training and test datasets in ***.CSV*** format
- 🧹 Automatic removal of null-valued columns
- ✍️ Select dependent (***Y***) variable *via* dialog
- 📈 Run multivariate regression using ***statsmodels.OLS***
- ✅ **Validate models with**:
  - ***Internal Metrics:*** *$R^2$, Adjusted $R^2$, MAE, RMSE, VIF, PRESS, SEE*
  - ***External Metrics:*** *$Q^2$ ($f_1$, $f_2$), MAE, RMSE*
  - ***Cross-Validation:*** Leave-One-Out *$Q^2$* and *MAE*
- 📝 **Export**:
  - Statistical summaries (txt)
  - Prediction tables (_train.csv, _test.csv)
- 📊 **Visualization**:
  - Correlation Heatmap
  - Observed vs Predicted Scatter Plot
- 🧩 **Easy integration as a PyMol Plug-in** (Menu -> Plugin -> PyRegX gui)

***NOTE: The Index column must be at the first column of the input files.***

## 📦 Dependencies

Ensure the following packages are installed in your Python environment:

- os, sys, subprocess, warnings
- NumPy, Pandas
- Scikit-Learn, statsmodels
- Seaborn, Matplotlib
- Tkinter
- Customtkinter (for modern GUI components)
- PyMol (for integration)

# ⚙ Installation

## Option 1: Install in PyMol environment (manual)

1. Place the plugin .py file in your PyMol plugins folder (e.g. ~/.pymol/startup/ or via Plugin Manager).
2. Launch PyMol and navigate to **Plugin > Manage > Install** and select the .py file.
3. PyRegX will appear in the **Plugin** menu.

## Option 2: Install required Python dependencies

If you're not using a pre-configured Python environment, install dependencies with pip:

```
$ pip install pandas numpy scikit-learn statsmodels matplotlib seaborn customtkinter
```

# 🚀 Usage Workflow

1. **Launch PyRegX GUI** from PyMol:
   *Plugin → PyRegX gui (v0.23)*
2. **Load Training Dataset**
   - Click *"Load Training Data"* and select a ***`.csv`*** file
   - The file must contain both independent and dependent variables
3. **Load Test Dataset**
   - Click *"Load Test Data"* and select a `.csv` file with the same structure
4. **Enter Dependent Variable (*Y*)**
   - Click *"Enter Dependent Column"*
   - Provide the exact column name to be predicted
5. **Select Output File**
   - Click *"Select Output File"* to specify where to save results
6. **Run Analysis**
   - Click *"Run Analysis"*
   - Generates regression summary, validation metrics, and prediction output
7. **Visual Output**
   - Correlation heatmap and scatter plot will pop up post-analysis

***NOTE: The Index column must be at the first column of the input files.***

# 📊 Statistical Metrics Used

## ✅ Internal Validation

- ***$R^2$ (Coefficient of Determination/Squared correlation coefficient)***
  Measures how well the regression model explains the variance.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- ***Adjusted $R^2$***
  Accounts for number of predictors:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1}\right]$$

- ***SEE (Standard Error of Estimate)***

$$SEE = \sqrt{\frac{\sum(y-\hat{y})^2}{n-p-1}}$$

- ***$Q^2_{LOO}$ (Internal Leave-One-Out cross-validated coefficient of determination)***
  Parameter indicates model fit under cross-validation (*See later*).
- ***PRESS (Predicted Residual Error Sum of Squares)***
  Parameter indicates model fit under cross-validation.
- ***MAE (Mean Absolute Error), RMSE (Root Mean Square Error)***
  Parameter to quantify average prediction error.
- ***VIF (Variance Inflation Factor)***
  Parameter for multicollinearity assessment

$$VIF = \frac{1}{1-R^2}$$

## 📤 External Validation

- ***$Q^2$ ($f_1$) / $R^2_{Pred}$***

$$Q^2_{f1} = 1 - \frac{\sum(y_{test} - \hat{y}_{test})^2}{\sum(y_{test} - \bar{y}_{train})^2}$$

- ***$Q^2$ ($f_2$) / $R^2_{Test}$***

$$Q^2_{f2} = 1 - \frac{\sum(y_{test} - \hat{y}_{test})^2}{\sum(y_{test} - \bar{y}_{test})^2}$$

## 🔁 Cross Validation

- ***Leave-One-Out $Q^2$ (Internal Leave-One-Out cross-validated coefficient of determination)***
  Validates the model with each observation removed once. ***$Q^2_{LOO} = R^2$***
  (on predicted *vs* actual in *LOO* setting)

# References

## *Statistical References*


- Banerjee, Adhikari, Amin and Jha (2020). *Structural exploration of tetrahydroisoquinoline derivatives as HDAC8 inhibitors through multi-QSAR modeling study*, Journal of Biomolecular Structure and Dynamics, 38(5):1551-1564. **DOI:** 10.1080/07391102.2019.1617782.
- Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). **DOI:**10.1002/9781118625590.
- Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to Linear Regression Analysis.* **DOI:**10.1111/insr.12020_10
- Golbraikh, A. and Tropsha, A. (2002) *Beware of $q^2$!* Journal of Molecular Graphics and Modelling, 20, 269. **DOI:** 10.1016/s1093-3263(01)00123-1.
- Roy, Kar, and Das (2015). *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. **DOI:** 10.1016/C2014-0-00286-9.


## *Software, Tools and Modules*

- https://pymol.org/
- https://github.com/schrodinger/pymol-open-source
- https://numpy.org/
- https://pandas.pydata.org/
- https://scikit-learn.org/
- https://www.statsmodels.org/
- https://github.com/TomSchimansky/CustomTkinter
- https://docs.python.org/3/library/tkinter.html
- https://matplotlib.org/
- https://seaborn.pydata.org/

---

# Author


**SUVANKAR BANERJEE**

**Email:** suvankarbanerjee1995@gmail.com


**Version:** 0.23

# ✅ Custom Non-Commercial License v1.0



***1. Non-Commercial Use Only***

The Software may *only be used for non-commercial purposes*.

This means:

- You may not sell, license, sublicense, or distribute this Software or any derivative works for a fee or other commercial benefit.

- You may not use the Software as part of a commercial service, product, SaaS platform, or other revenue-generating activity.

- You may not use the Software in any context where you, your company, or any third party derives a commercial advantage or financial compensation.

***2. Attribution Required***

You must give appropriate credit to the original author(s), provide a link to this license, and indicate if changes were made.

Attribution must not suggest endorsement by the original author(s).

***3. No Warranty***

The Software is provided *"as is", without warranty of any kind*, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose, and noninfringement.

In no event shall the authors or copyright holders be liable for any claim, damages, or other liability arising from the use of the Software.

***4. Redistribution***

You may redistribute this Software only if:

- It is accompanied by this license in full.

- It is clearly marked as a derivative (if modified).

- It is not sold or used commercially.

***5. Commercial Licensing Option***

To obtain a commercial license for use beyond these terms, including resale or integration into commercial offerings, please contact the author(s) at: suvankarbanerjee1995@gmail.com

***6. Termination***

Any violation of these terms automatically terminates your rights under this license.

By using the Software, you agree to the terms of this License.

***This license shall be governed by Indian and international copyright laws, and any disputes arising under this license shall be resolved under the principles of fairness, good faith, and mutual respect, without limiting it to any single jurisdiction***.

#  Available at