# Sequence Slider GUI Documentation

Roberto Carlos Navarro Quiroz
rnqcri@ibmb.csic.es
Rafael Borges
rb.unesp@gmail.com

May 27, 2025

## Contents

# 1 Overview

The **Sequence Slider GUI** is a Python-based application designed to facilitate protein structure analysis and amino acid sequence prediction. It integrates the SEQUENCE SLIDER tool with machine learning (ML) models to process crystallographic data from PDB and MTZ files, predict amino acid sequences, and visualize results through sequence logos and interactive HTML visualizations. The application provides a user-friendly graphical interface built with Tkinter, making it accessible to researchers and scientists in structural biology.

The tool supports two ML model types:

- **Model 1 (without Consurf)**: Uses structural and energetic features for predictions.
- **Model 2 (with Consurf)**: Incorporates conservation scores from Consurf for enhanced predictions.

# 2 Features

The Sequence Slider GUI offers the following features:

- **File Input**: Load PDB and MTZ files for crystallographic analysis.
- **SEQUENCE SLIDER Execution**: Run the SEQUENCE SLIDER script to process input files.
- **Data Processing**: Transform and merge data from multiple sources (clashes, interactions, residue depth, rotamers, statistics).
- **ML Predictions**: Apply pre-trained XGBoost models to predict amino acid sequences with probability scores.
- **Visualization**:
  - Generate sequence logos for predicted sequences using Logomaker.
  - Create interactive HTML visualizations with probability-based font sizing and tooltips.
- **Output**:
  - FASTA files for predicted sequences per protein chain.
  - CSV files with prediction results.
  - JPG files for sequence logos.
  - Interactive HTML files for sequence visualization.
- **GUI Features**:
  - Progress bar for tracking operations.
  - Tooltips for user guidance.

  - Tabbed interface for displaying multiple sequence logo chunks.
- **Logging**: Comprehensive logging to a file (`sequence_slider.log`) and GUI display for tracking operations.

# 3 Dependencies

To run the Sequence Slider GUI, ensure the following Python packages are installed:

```
pandas
numpy
matplotlib
tkinter
subprocess
xgboost
logomaker
scikit-learn
```

Install dependencies using pip:

```
pip install pandas numpy matplotlib xgboost logomaker scikit-learn
```

**Note**: Tkinter is typically included with Python installations. If not, install it via your package manager (e.g., `python3-tk` on Ubuntu).

Additionally, the `SLIDER_VENOM.py` script from the `slider-master` directory is required to run the SEQUENCE SLIDER tool.

# 4 Installation

1. **Clone or Download the Repository**:
   - Obtain the script (`gui_vertion2025Amayo_run_model2.py`) and ensure it is placed in a directory with the required model files and `slider-master` folder.
2. **Model Files**:
   - Ensure the following model files are present in the specified directories:
     - **Without Consurf**:
       * `model_without_consurf_C/modelo_xgb_multiclase_2025_only_C_2`
       * `model_without_consurf_C/label_encoder_model_modelo_xgb_mult`
     - **With Consurf**:
       * `model_consurf_C/modelo_xgb_multiclase_consurf_2025.pkl`
       * `model_consurf_C/label_encoder_modelo_xgb_multiclase_consurf_`
3. **Directory Structure**:

```
project_directory/├──
 gui_vertion2025Amayo_run_model2.py├──
 slider-master/│
    └── SLIDER_VENOM.py├──
 model_without_consurf_C/│
    ├── modelo_xgb_multiclase_2025_only_C_271_sequence_slider.pkl│
    └──
    label_encoder_model_modelo_xgb_multiclas2025_only_C_271_sequence_slider
    .pkl├──
 model_consurf_C/│
    ├── modelo_xgb_multiclase_consurf_2025.pkl│
    └── label_encoder_modelo_xgb_multiclase_consurf_2025.pkl

```

4. **Run the Application**:

```
python gui_vertion2025Amayo_run_model2.py

```

# 5 Usage

1. **Launch the GUI**:
   - Run the script to open the Tkinter-based GUI.
2. **Load Input Files**:
   - Click **Load PDB** to select a `.pdb` file.
   - Click **Load MTZ** to select a `.mtz` file.
3. **Run SEQUENCE SLIDER**:
   - Once both files are loaded, click **Run SEQUENCE SLIDER** to execute the analysis. Outputs are saved in an `output-<pdb_name>` directory.
4. **Process Data**:
   - Click **Prepare Data for ML Model** to process the SEQUENCE SLIDER output. Select the output directory when prompted.
5. **Load ML Model**:
   - Click **Load ML Model** and choose between **Model 1 (without Consurf)**, **Model 2 (with Consurf)**, or load a custom data file.
6. **Load Consurf File** (Optional):
   - For Model 2, click **Load Consurf Result** to select a Consurf CSV file.
7. **Run ML Predictions**:
   - Click **Run ML Model** to generate predictions. Results are saved as CSV and FASTA files, and sequence logos are generated.
8. **Visualize Results**:

- Click **Plot Result** to generate an interactive HTML visualization, which opens in your default browser.

9. **Output Files**:
    - **FASTA Files**: `<pdb_id>_chain_<chain>_predicted_sequence.fasta`
    - **Sequence Logos**: `<pdb_id>_logo_chunk_<number>.jpg`
    - **Prediction Results**: `<pdb_name>_result_df_[without_consurf|consurf].cs`
    - **Interactive Visualization**: `<pdb_id>_sequence_visualization.html`

# 6 Output Description

- **FASTA Files**: Contain predicted amino acid sequences for each protein chain.
- **Sequence Logos**: Visualize amino acid probabilities in chunks (default: 25 positions per chunk). Logos maintain the original sequence order and include:
    - Amino acid probabilities (stacked bars).
    - Position labels (Chain:ResidueNumber).
    - Predicted sequence and confidence statistics.
- **CSV Files**: Contain detailed prediction results, including:
    - `PDBid`, `Chain`, `ResN` (residue number), `ResT` (residue type).
    - Probabilities for each amino acid (A, C, D, ..., Y).
    - `Predicted_AA`: The predicted amino acid with the highest probability.
    - `Probability`: The confidence score for the predicted amino acid.
- **HTML Visualization**:
    - Displays the predicted sequence with amino acids colored by type (acidic, basic, polar, hydrophobic, glycine, cysteine).
    - Font size reflects prediction probability.
    - Hover tooltips show position, amino acid, and probability.
    - Controls to toggle position numbers and adjust font size.

# 7 Color Scheme

The sequence logos and HTML visualization use the following color scheme for amino acids:

- **Acidic (D, E)**: Red (`#FF0000`)

- **Basic (K, R, H)**: Blue (`#0000FF`, `#0066CC` for H)
- **Polar (S, T, N, Q, Y)**: Green (`#00AA00`, `#00CC66` for Y)
- **Hydrophobic (A, V, I, L, M, F, W, P)**: Dark Blue (`#1f77b4`)
- **Glycine (G)**: Orange (`#FFAA00`)
- **Cysteine (C)**: Yellow (`#FFFF00`)

# 8 Logging

- Logs are saved to `sequence_slider.log` in the working directory.
- The GUI displays real-time logs in the output text area.

# 9 Contact

For support or inquiries, contact:


- **Roberto Carlos Navarro Quiroz**: rnqcri@ibmb.csic.es
- **Rafael Borges**: rb.unesp@gmail.com


# 10 License