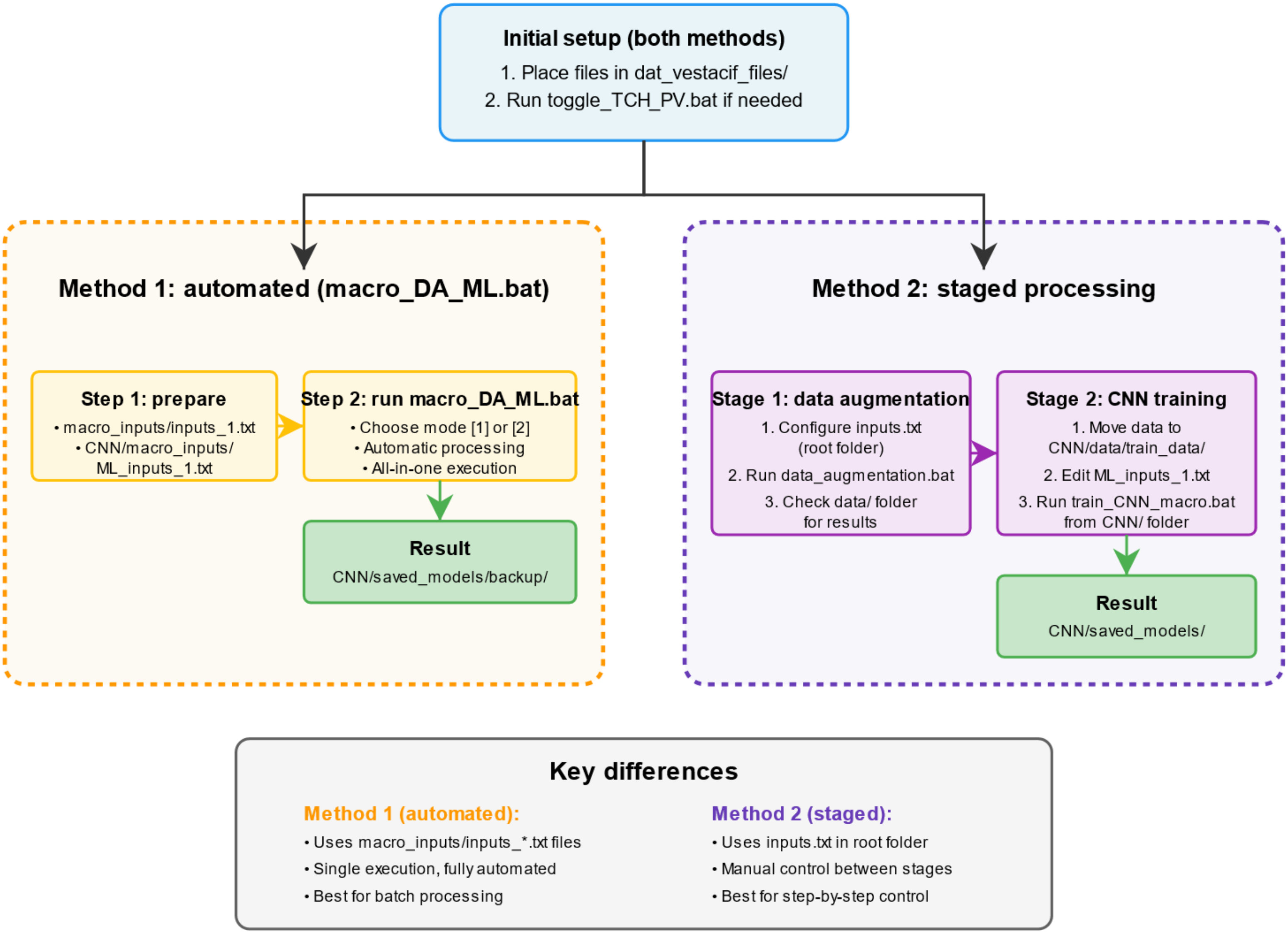
```
██████╗  █████╗ ██████╗ ██╗██████╗ 
██╔══██╗██╔══██╗██╔══██╗██║██╔══██╗
██████╔╝███████║██████╔╝██║██║  ██║
██╔══██╗██╔══██║██╔═══╝ ██║██║  ██║
██║  ██║██║  ██║██║     ██║██████╔╝
╚═╝  ╚═╝╚═╝  ╚═╝╚═╝     ╚═╝╚═════╝ 
```

RAPID - User Manual

(Version 2025.08.25)

%Table of Contents

%Overview


Initial setup (both methods)
1. Place files in dat_vestacif_files/
2. Run toggle_TCH_PV.bat if needed
Method 1: automated (macro_DA_ML.bat)
Step 1: prepare
• macro_inputs/inputs_1.txt
• CNN/macro_inputs/ ML_inputs_1.txt
Step 2: run macro_DA_ML.bat
• Choose mode [1] or [2]
• Automatic processing
• All-in-one execution
Result
CNN/saved_models/backup/
Method 2: staged processing
Stage 1: data augmentation
1. Configure inputs.txt (root folder)
2. Run data_augmentation.bat
3. Check data/ folder for results
Stage 2: CNN training
1. Move data to CNN/data/train_data/
2. Edit ML_inputs_1.txt
3. Run train_CNN_macro.bat from CNN/ folder
Result
CNN/saved_models/
Key differences
Method 1 (automated):
• Uses macro_inputs/inputs_*.txt files
• Single execution, fully automated
• Best for batch processing
Method 2 (staged):
• Uses inputs.txt in root folder
• Manual control between stages
• Best for step-by-step control

The RAPID Pipeline **(Windows only)** provides two methods for processing XRD data:

1. Automated Method (macro_DA_ML.bat): Combines data augmentation and CNN training in one automated workflow
2. Staged Method: Separate execution of data augmentation (data_augmentation.bat) and CNN training (train_CNN_macro.bat)

Both methods require the same initial file preparation and produce similar results, but differ in workflow control and configuration file locations.

%Initial Setup and File Preparation

Required Files:

1. Crystal structure file: Must be named [material]_vesta.cif
   - Example: CeO2_vesta.cif
   - This file contains atomic positions and unit cell parameters
   - Must have "_vesta" in the filename
   - Values extracted: space group, lattice parameters (a, b, c, α, β, γ), atom types and occupancies, thermal parameters (Uiso/Uani converted to Biso)
2. XRD pattern file: Must be named [material].dat
   - Example: CeO2.dat
   - This is your experimental XRD data
   - Contains 2-theta values and intensities
3. Reference PCR file: [material].pcr
   - Example: CeO2.pcr
   - This is a FullProf control file with refinement parameters
   - If provided, the system will extract parameters from it
   - Values extracted: Profile parameters (U, V, W), peak asymmetry parameters (Asy1-4), background coefficients, zero shift, and scale factor

Place ALL three files in this file path:

```
root_directory/
    dat_vestacif_files/         <-- PUT _vesta.cif AND .dat FILES HERE
        CeO2_vesta.cif
        CeO2.dat
        reference_pcr_format/   <-- PUT reference .pcr FILE HERE
          CeO2.pcr
```

Ensure you have two conda environments:

- py27: Python 2.7 environment for data augmentation
- base: Python 3.x environment for CNN training

%Profile Function Configuration

Understanding Profile Functions

- Npr=5 (Pseudo-Voigt): Simple combination of Gaussian and Lorentzian functions (tbbaco)
- Npr=7 (Thompson-Cox-Hastings): More sophisticated profile function with better instrumental/sample separation

Checking and Setting Profile Function

1. Run toggle_TCH_PV.bat to toggle between profile functions
2. The script will display which function is currently active
3. Materials like tbbaco often require Npr=5 for better fitting

Important: Check this setting before running any data processing

%Method 1: Automated End-to-End Processing

>Step 1: Prepare configuration files

You need to create TWO configuration files in DIFFERENT locations:

File 1: Data Augmentation Configuration

Location: macro_inputs/inputs_1.txt

```
root_directory/
      macro_inputs/                <-- CREATE THIS FOLDER
             inputs_1.txt           <-- PUT DATA AUGMENTATION CONFIG HERE
             inputs_2.txt           (optional - for multiple materials/fine tuning)
```

File 2: CNN Training Configuration

Location: CNN/macro_inputs/ML_inputs_1.txt

```
root_directory/
      CNN/
             macro_inputs/          <-- THIS FOLDER SHOULD EXIST
                   ML_inputs_1.txt    <-- PUT CNN CONFIG HERE
```

>Step 2: run the automated pipeline

1. Double-click macro_DA_ML.bat in the root directory
2. Select processing mode:

   ```
   Choose mode:
   [1] Process multiple datasets
   [2] Fine-tune a compound
   ```

3. For option 1 (single or multiple datasets):
   - The script automatically finds all inputs_*.txt files in macro_inputs/
   - Generates datasets for each configuration
   - Trains CNN models on each dataset
   - Organizes results automatically
4. For option 2 (fine-tuning):
   - Review the displayed inputs.txt configuration
   - Select parameters to fine-tune (1-7)
   - Specify number of iterations
   - Enter shift values for each parameter
   - The script creates variations and processes them

>Step 3: review results

Results are organized in:

```
CNN/saved_models/
      backup/
            models_[dataset_name]/
                  [model_folders]/
```

%Method 2: Staged Processing

Stage 1: Data Augmentation

>Step 1: Configure inputs.txt

Location: Edit inputs.txt in the ROOT directory (NOT in macro_inputs/)

```
root_directory/
        inputs.txt                 <-- EDIT THIS FILE
        data_augmentation.bat
        ...
```

>Step 2: Run Data Augmentation

1. Double-click data_augmentation.bat in the root directory
2. The script will:
   - Create a timestamped folder in data/
   - Generate modified PCR files
   - Run Rietveld refinements
   - Create combined dataset files

>Step 3: Check Generated Data

Look for output in:

```
data/
        [material]_YYYYMMDD_HHMMSS/
                [material]_simulated_data_row_param.dat     # Main dataset for CNN
                classification_report.dat                   # Quality metrics
                output_figures/                             # Visualizations
```

Stage 2: CNN Training

>Step 1: Move Dataset to CNN Folder

MANUALLY COPY the entire timestamped folder. If we were processing CeO2 data:

FROM:

```
data/
      CeO2_20250417_062558/    <-- COPY THIS ENTIRE FOLDER
```

TO:

```
CNN/
      data/
            train_data/
                  CeO2_20250417_062558/    <-- PASTE HERE
```

>Step 2: Configure ML_inputs Files

Location: CNN/macro_inputs/ML_inputs_1.txt

```
CNN/
      macro_inputs/
            ML_inputs_1.txt         <-- EDIT THIS FILE
```

Edit the file to match your dataset name EXACTLY.

>Step 3: Create Multiple Configs (Optional)

To train multiple models:

1. Navigate to CNN/ folder
2. Double-click create_ml_inputs.bat
3. Enter the total number of models you want

>Step 4: Run CNN Training

1. Navigate to CNN folder
2. Double-click train_CNN_macro.bat
3. The script will process all ML_inputs_*.txt files

>Step 5: Review Results

Find results in:

```
CNN/saved_models/
      model_[name]/
             model_[name].pth           # Trained model file
             refinement_result/         # Refinement outputs
             training_result/           # Training metrics
             feature_importance/        # Correlation analysis
```

%Configuration Files Explained

%inputs.txt/inputs_x.txt - Complete Line-by-Line Guide (46 lines)

→This file controls data augmentation. Each line has a specific purpose:

File Identification (Lines 1-5)

```
1>Name of .CIF file
CeO2_vesta.cif
```

- Purpose: Identifies your crystal structure file
- What to input: Exact filename including "_vesta.cif"
- Notes: Must match the file in dat_vestacif_files/ folder exactly

```
2>Name of .DAT file
CeO2.dat
```

- Purpose: Identifies your experimental XRD pattern
- What to input: Exact filename with .dat extension
- Notes: Material name must match the CIF file

```
3>Name of .PCR file to be generated
CeO2.pcr
```

- Purpose: Names the PCR file that will be created
- What to input: Material name + .pcr
- Notes: This file will be generated automatically, don't create it beforehand

```
4>Value of INS
10
```

- Purpose: Specifies instrument resolution file number
- What to input: Usually "0" for no resolution file, or 1-10 for specific instruments
- Notes: 0 is standard for most cases unless you have instrument-specific files

```
5>Enter the name of the subfolder to be created in the data directory (the purpose
of the subfolder is to organize and store related data files and refinement results
for analysis):
```

```
CeO2
```

- Purpose: Names the output folder in data/ directory
- What to input: Simple material name (no spaces, no special characters)
- Notes: A timestamp will be added automatically (e.g., CeO2_20250417_062558)

Parameter Variation Settings (Lines 6-13)

```
6>Input value for zero parameter (input 'skip' if you wish to skip to next row and
input maximum absolute shift range instead):
```

```
skip
```

- Purpose: Sets the zero-shift correction value
- What to input: A number (e.g., 0.02) for fixed value, or "skip" for random variation
- Notes: "skip" is recommended to explore parameter space

```
7>Input the maximum absolute shift range for generating a random number around the
zero parameter value (example: 0.03). You may ignore this part if you did not skip
part 6>:
```

```
0.05
```

- Purpose: Defines random variation range for zero parameter
- What to input: Maximum shift value (e.g., 0.05 means ±0.05)
- Notes: Only used if line 6 is "skip". Typical values: 0.01-0.1

```
8>Input value for background parameter (input 'skip' if you wish to skip to next row
and input maximum absolute shift range instead):
```

```
skip
```

- Purpose: Sets background polynomial coefficient
- What to input: A number for fixed background, or "skip" for variation
- Notes: Usually use "skip" for realistic data augmentation

```
9>Input the maximum absolute shift range for generating a random number around the background parameter value (example: 0.03). You may ignore this part if you did not skip part 8>:

0
```

- Purpose: Background variation range
- What to input: 0 for no variation, or a positive number (e.g., 50)
- Notes: Set to 0 if background is stable in your data

```
10>Input values (example: 10, 10, 10, 90, 90, 90) for lattice parameters a, b, c, alpha, beta, gamma (input 'skip' if you wish to input maximum percentage scaling range instead):

skip
```

- Purpose: Set unit cell parameters directly
- What to input: Six numbers "a, b, c, alpha, beta, gamma" or "skip"
- Notes: "skip" uses CIF values with percentage variation from line 12. Recommended to skip.

```
11>Enter the lattice type (cubic, tetragonal, orthorhombic, hexagonal, monoclinic, triclinic, trigonal):

cubic
```

- Purpose: Specifies crystal system for proper constraints
- What to input: One of: cubic, tetragonal, orthorhombic, hexagonal, monoclinic, triclinic, trigonal
- Notes: Must match your actual crystal structure

```
12>Input the percentage range for generating a random number around the lattice parameters with degrees of freedom (example: 3%). You may ignore this part if you did not skip part 10>:

0.01%
```

- Purpose: How much to vary lattice parameters
- What to input: Percentage with % symbol
- Notes: Only used if line 10 is "skip"

```
13>Input the number of datasets to be generated: (The first half will be uniformly
distributed points on the grid defined by the ranges in 7>, 9>, and 12>, and the
second half will be randomly distributed points on this grid)
```

```
10000
```

- Purpose: How many modified PCR files to create
- What to input: Integer number (typically 10000)
- Notes: First half uses uniform grid sampling, second half uses random sampling

Data Collection Settings (Lines 14-22)

```
14>Use the default two theta sampling range? (Y/N); If 'Y', the range is determined
automatically. If 'N', specify the range as: initial, step, final (example:
N;5,0.1,110):
```

```
N;25,0.2,100
```

- Purpose: Define 2-theta range and step size
- What to input: "Y" for auto-detection, or "N;start,step,end"
- Notes: Format example: N;25,0.2,100 means 25° to 100° with 0.2° steps. The range you specify must match or be within the actual range of your experimental .dat file. Check your .dat file first to confirm its 2-theta range and use those values here.

```
15>Display AutoFP GUI interface (Y/N)?
```

```
N
```

- Purpose: Show/hide FullProf interface during refinement
- What to input: "Y" or "N"
- Notes: Use "N" for batch processing (faster), "Y" for displaying AutoFP processing. Recommended to input "N"

```
16>Generate all output figures (Y/N)?
```

```
Y
```

- Purpose: Create visualization plots
- What to input: "Y" or "N"
- Notes: "Y" recommended for quality checking

```
17>Display all output figures (Y/N)?

N
```

- Purpose: Show plots on screen during processing
- What to input: "Y" or "N"
- Notes: "N" for batch processing, plots are saved regardless

```
18>Run scripts step 5 and step 6 for R-factor analysis and contour plots (Y/N)?
(this is deprecated for now, suggested to input “N”):

N
```

- Purpose: Additional R-factor statistical analysis
- What to input: "Y" or "N"
- Notes: Adds significant processing time, usually not needed as they are deprecated steps for the moment.

```
19>Make simulated_dataset folder and save its files (sample PCR, PRF files) (Y/N)?:

N
```

- Purpose: Keep individual PCR/PRF files
- What to input: "Y" or "N"
- Notes: Can use lots of disk space, usually not needed, recommended to input "N".

```
20>Make output_figures folder and save its files (analysis plots, distributions)
(Y/N)?:

Y
```

- Purpose: Save all visualization plots
- What to input: "Y" or "N"
- Notes: "Y" recommended for analysis

```
21>Generate .BIN files instead of .DAT files? (Y/N):

N
```

- Purpose: Use binary format for large datasets
- What to input: "Y" or "N"
- Notes: Use "Y" only for >10000 datasets to save space. This setting is deprecated for now.

```
22>Train Convolutional Neural Network model with the generated .DAT file (Y/N)?:

N
```

- Purpose: Immediate CNN training after generation
- What to input: "N" (always)
- Notes: We handle CNN training separately in our pipeline. This setting is deprecated as well.

```
23>Input the percentage range for generating a random number around the Biso
parameter with degrees of freedom (example: 3%):

0.01%
```

- Purpose: Atomic displacement parameter variation
- What to input: Small percentage (e.g., 0.01% to 1%)
- Notes: Controls thermal vibration simulation

```
24>Input the percentage range for generating a random number around the scale factor
parameter with degrees of freedom (example: 5%):

0.1%
```

- Purpose: Overall intensity scaling variation
- What to input: Percentage (e.g., 0.1% to 5%)
- Notes: Simulates different sample amounts/densities

```
25>Input the percentage range for generating a random number around the U, parameter
with degrees of freedom (example: 1%):

0.1%
```

- Purpose: Gaussian peak width variation, U
- What to input: Percentage
- Notes: Affects peak broadening at high angles

```
26>Input the percentage range for generating a random number around the V, parameter
with degrees of freedom (example: 1%):

0.1%
```

- Purpose: Gaussian peak width variation, V
- What to input: Percentage
- Notes: Affects overall peak broadening

```
27>Input the percentage range for generating a random number around the W, parameter
with degrees of freedom (example: 1%):
```

```
0.1%
```

- Purpose: Gaussian peak width variation, W
- What to input: Percentage
- Notes: Affects peak broadening at low angles

```
28>Input the number of epochs for CNN training:
```

```
10
```

- Purpose: CNN training iterations (legacy option)
- What to input: Any integer
- Notes: Ignored in our pipeline, can leave as default

Classification and Quality Control (Lines 29-33)

```
29>Perform Step 3.5 distribution & classification logic (Y/N)?:
```

```
Y
```

- Purpose: Classify results by quality (CLOSE/BOUNDARY/INVALID)
- What to input: "Y" or "N"
- Notes: "Y" strongly recommended for quality control

```
30>Enter Rwp threshold (in %) for a "close" fit classification (or 'default' to use
recommended 80%):
```

```
default
```

- Purpose: Define good fit threshold
- What to input: "default" or a number (e.g., 12)
- Notes: "default" uses automatic thresholds based on material. Recommended to input "default".

```
31>Enter max Rwp (in %) for a "boundary" fit classification (or 'default' to use
recommended 20%):
```

```
default
```

- Purpose: Define acceptable fit threshold
- What to input: "default" or a number (e.g., 25)
- Notes: Fits above this are marked as INVALID. Recommended to input "default".

```
32>Allow partial re-run if 80/20 ratio not met? (Y/N)?:

Y
```

- Purpose: Automatic parameter adjustment if too many failures
- What to input: "Y" or "N"
- Notes: "Y" recommended for automatic optimization. Input "N" if you know your parameter configuration can yield the desired 80:20 ratio without adjustment, or when fine-tuning with specific parameter values.

```
33>Generate classification plots after each re-run? (Y/N)?:

N
```

- Purpose: Create plots during re-runs
- What to input: "Y" or "N"
- Notes: "N" for faster processing, final plots always generated

```
34>Reference PCR file to extract parameters from (leave blank or enter "default" to
use default):

CeO2.pcr
```

- Purpose: Use existing PCR as template
- What to input: Filename or "default" for no reference
- Notes: File must be in dat_vestacif_files/ folder. If set to "default" or left blank, the pipeline works fine using a built-in template with standard profile parameters - the reference PCR is just recommended and provides better starting values for faster convergence.

```
35>Select which Bragg reflection peak to zoom into (1=first peak, 2=second, etc., or
'auto' for most intense peak):

auto
```

- Purpose: Focus plots on specific peak
- What to input: Peak number (1, 2, 3...) or "auto"
- Notes: Used in classification plots for detailed view. "auto" actually just shows the first peak, not the most intense peak

```
36>Customize zoom width around the selected peak (in degrees 2θ, or 'auto' for
structure-dependent defaults):

0.3
```

- Purpose: How wide to make zoomed view
- What to input: Degrees 2-theta or "auto"
- Notes: Typical values: 0.25-1.0 degrees

```
37>Enter two-theta range for classification plots (min, max, or 'default' for standard 0-100 range):
```

```
25, 100
```

- Purpose: Plot range for classification
- What to input: "min, max" or "default"
- Notes: Should match your data range. It is also fine to input "25, 100" as shown in the above example.

```
38>Copy asymmetry-shape parameters and preferred orientation parameters from reference PCR (Y/N)?:
```

```
Y
```

- Purpose: Use asymmetry from reference PCR
- What to input: "Y" or "N"
- Notes: "Y" if reference PCR has good asymmetry parameters. Most of the time this will not affect much about the processing.

```
39>Copy scan-range and pattern-selection parameters from reference PCR (Y/N)?:
```

```
Y
```

- Purpose: Use 2-theta range from reference
- What to input: "Y" or "N"
- Notes: Overrides line 14 settings if "Y". Recommended to just input "Y".

Output Control (Lines 40-46)

```
40>Generate classification profile plots (Y/N)?:
```

```
Y
```

- Purpose: Create quality classification plots
- What to input: "Y" or "N"
- Notes: "Y" recommended for visual quality check

```
41>Generate parameter shift distribution plots (Y/N)?:
```

```
Y
```

- Purpose: Visualize parameter sampling
- What to input: "Y" or "N"
- Notes: Useful for confirming proper parameter distribution

```
42>Generate PRF files folder (Y/N)?:
```

```
N
```

- Purpose: Keep all PRF result files
- What to input: "Y" or "N"
- Notes: Takes lots of space, usually not needed

```
43>Generate Rietveld refinement plots (Y/N)?:
```

```
N
```

- Purpose: Individual refinement plots
- What to input: "Y" or "N"
- Notes: Creates one plot per refinement, can be overwhelming, "N" recommended

```
44>Generate Ycal original versus interpolated plots (Y/N)?:
```

```
N
```

- Purpose: Technical validation plots
- What to input: "Y" or "N"
- Notes: Only needed for debugging/checking interpolation issues

```
45>How many ML_inputs files to create for CNN training?
```

```
500
```

- Purpose: Number of CNN configs to generate, usually for fine tuning purpose
- What to input: Integer (500-2000 typically, for PC, 2000 is recommended)
- Notes: Only relevant for integrated workflow (method 1)

```
46>Atom data format in PCR file (enter "4-line" for standard format with thermal
parameters, or "2-line" for simplified format without thermal parameters - check
your PCR file to see if each atom has 4 lines or 2 lines of data):
```

```
2-line
```

- Purpose: PCR file format specification
- What to input: "4-line" or "2-line"
- Notes: Check your PCR file - count lines per atom to determine

%ML_inputs.txt - Complete Line-by-Line Guide (12 lines)

This file controls CNN training and refinement:

```
# 1> Train from scratch? (y/n).
y;model_CeO2_1
```

- Purpose: Create new model or load existing
- What to input: "y;model_name" for new, or "n" to load existing
- Notes: Model name format: model_[material][description][number]

```
# 2> Name of an existing model folder in 'saved_models' to load.
n
```

- Purpose: Specify model to load if not training from scratch
- What to input: "n" if line 1 is "y", or exact folder name from saved_models/
- Notes: Must match folder name exactly including timestamp

```
# 3>Train the model progressively with more than one dataset?
N;CeO2_20250511_111201
```

- Purpose: Specify training dataset(s)
- What to input: "N;folder" for single, or "Y;folder1,folder2" for multiple
- Notes: Folder name must match exactly what's in CNN/data/train_data/. We no longer use progressive training as this usually leads to unnecessary overfitting.

```
# 4> Experimental .dat file(s) for ML refinement, comma-separated if multiple.
CeO2.dat
```

- Purpose: Files to refine with trained model
- What to input: Single filename, comma-separated list, or "n"
- Notes: Files must exist in dat_vestacif_files/ folder

```
# 5> Do Rietveld Refinement using the refined parameters? (y/n).
y
```

- Purpose: Run refinement after training
- What to input: "y" or "n"
- Notes: Creates refined PCR and analysis plots

```
# 6> Execute feature importance test? (y/n).
y
```

- Purpose: Run correlation analysis for interpretability
- What to input: "y" or "n"
- Notes: Adds 5-10 minutes but provides valuable insights

```
# 7> Omit background parameter from CNN training and prediction? (y/n).
y
```

- Purpose: Exclude background from CNN training
- What to input: "y" or "n"
- Notes: "y" recommended for better stability

```
# 8>Enter the Bragg reflection peak number to zoom in on (or "default"):
default
```

- Purpose: Focus refinement plots on specific peak
- What to input: Peak number or "default"
- Notes: Choose a well-defined, isolated peak

```
# 9>Customize zoom width around selected peak (in degrees 2θ, or 'auto' for
structure-dependent defaults):

1.5
```

- Purpose: Width of zoomed region in degrees
- What to input: Number or "auto"
- Notes: Typical values: 1.0-3.0 degrees

```
# 10> Use digit-based parameter scaling? (y/n) [Normalizes values to similar digit
places for improved consistency]

y
```

- Purpose: Normalize parameters by digit count
- What to input: "y" or "n"
- Notes: Usually "y" – digit-based scaling is stable

```
# 11> Use reference-based adaptive scaling? (y/n) [Scales parameters relative to
reference values from standard materials]

n
```

- Purpose: Scale parameters relative to reference values
- What to input: "y" or "n"
- Notes: "n" recommended for most cases

```
# 12> Specify DAT file format structure (CeO2.dat, pbso4.dat, tbbaco.dat):

CeO2.dat
```

- Purpose: Select parser for DAT file format
- What to input: One of: CeO2.dat, pbso4.dat, tbbaco.dat. Each dat file has a different format.
- Notes: Must match your actual data format structure

%File Naming Conventions

%input Files

- CIF files: [material]_vesta.cif
- DAT files: [material].dat
- PCR files: [material].pcr

%generated Files

- Datasets: `[material]_YYYYMMDD_HHMMSS/`
- Models: `model_[material]_[number]/`
- Data for model training: `[material]_simulated_data_row_param.dat`

%notes

1. Material names must be consistent across all files
2. Avoid spaces in filenames
3. Use lowercase for material names
4. Timestamp format: YYYYMMDD_HHMMSS