Feature Assignment, Feature Engineering, and Monte Cat component manual

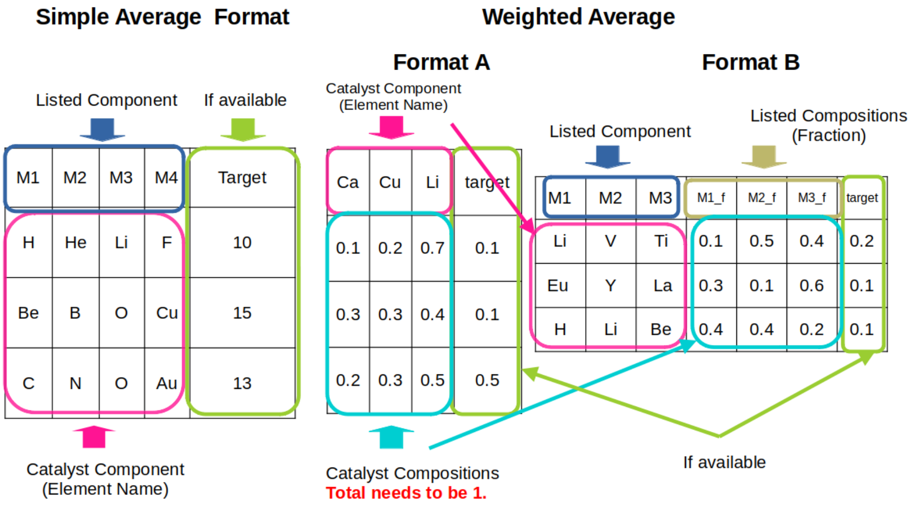
1 Introduction

These three component(Feature Assignment, Feature Engineering, and Monte Cat) serve as pre-processing for machine learning and other applications. Therefore, they output the processing results as a csv file for user's input file.

2 Component Manual

2.1 Feature Assignment

This component convert user's catalyst and its composition information into physical property information based on XenonPy. Figure1 shows the required data structure of the data in the csv file which user uploads. If the user does not comply, an error may occur. User can get simple average or weighted average of dataset.

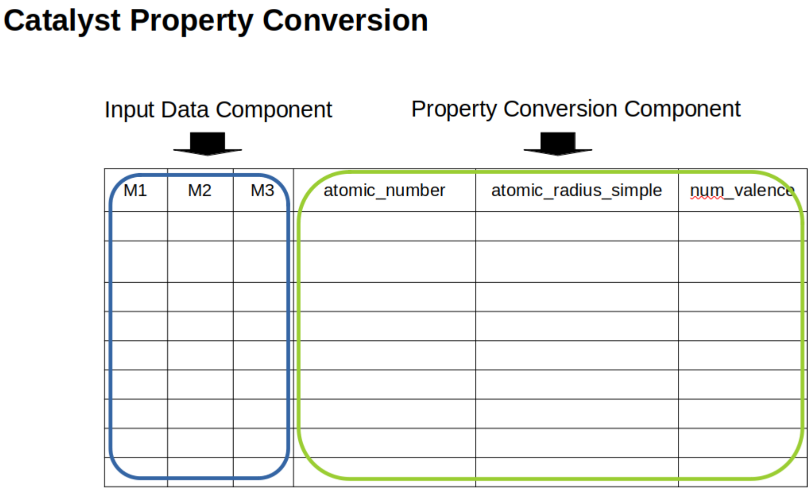


M1	M2	M3	M4	Target
H	He	Li	F	10
Be	B	O	Cu	15
C	N	O	Au	13

Ca	Cu	Li	target
0.1	0.2	0.7	0.1
0.3	0.3	0.4	0.1
0.2	0.3	0.5	0.5

M1	M2	M3	M1_f	M2_f	M3_f	target
Li	V	Ti	0.1	0.5	0.4	0.2
Eu	Y	La	0.3	0.1	0.6	0.1
H	Li	Be	0.4	0.4	0.2	0.1

Figure 1: Input CSV File Data Requirements Format.

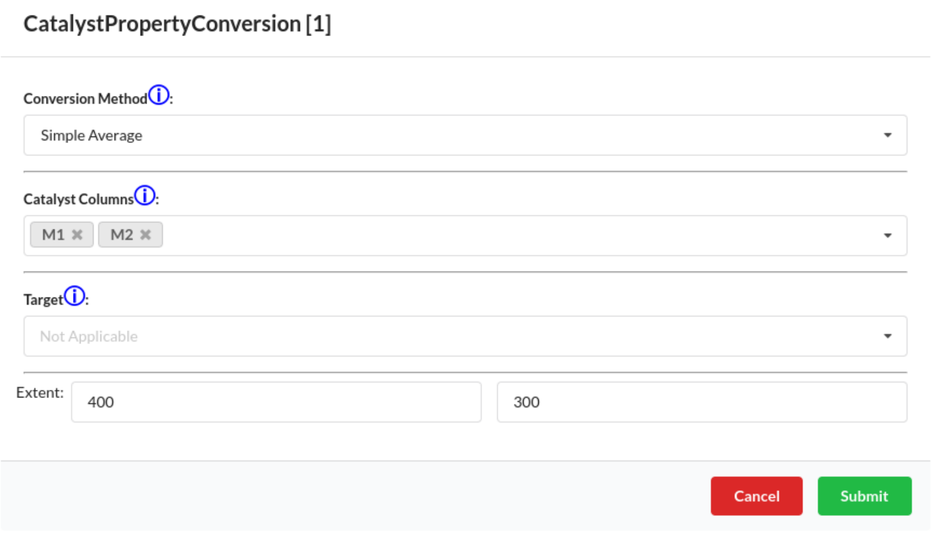


M1	M2	M3	atomic_number	atomic_radius_simple	num_valence

Figure 2: Output CSV File Format.

2.1.1 Form Fields

- **Form**
  Conversion Method: Select which method to use to convert catalyst and its composition into physical property.
  Catalyst Columns: This is the columns of catalysts and means Listed(Catalyst) Component in Figure1.
  Catalyst Composition Columns: This is the columns of catalyst compositions and means Listed Compositions of Format B in Figure1. **Target**: Not necessarily to be entered. After feature assgnment, if user wants to do feature engineering and monte cat, it will be useful to enter target.



CatalystPropertyConversion [1]

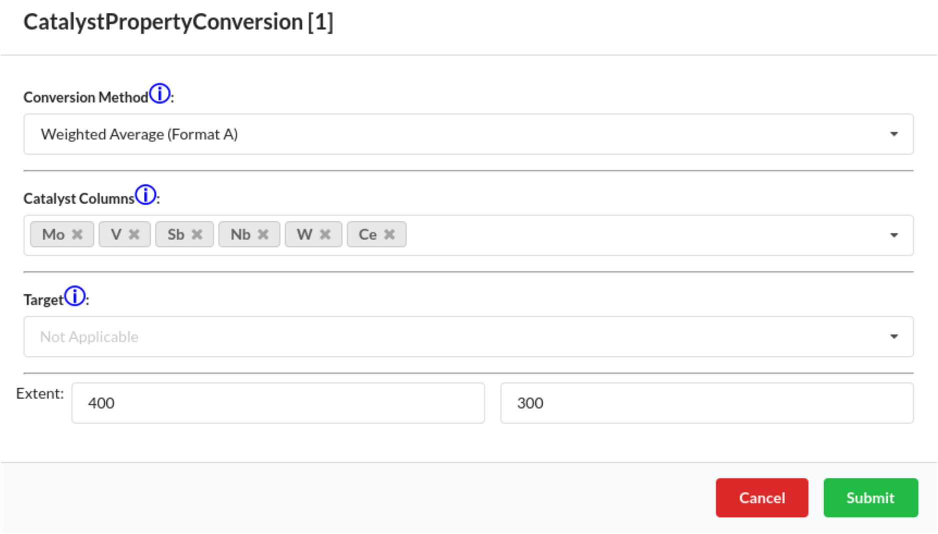
Conversion Method: Simple Average

Catalyst Columns: M1 ✖ M2 ✖

Target: Not Applicable

Extent: 400 300

Figure 3: **Form Fields of Simpale Average. Simple Average method converts a dataset consisting of information of catalyst components without composition information to return the simple average of the properties considering the numbers in each individual catalyst.**



CatalystPropertyConversion [1]

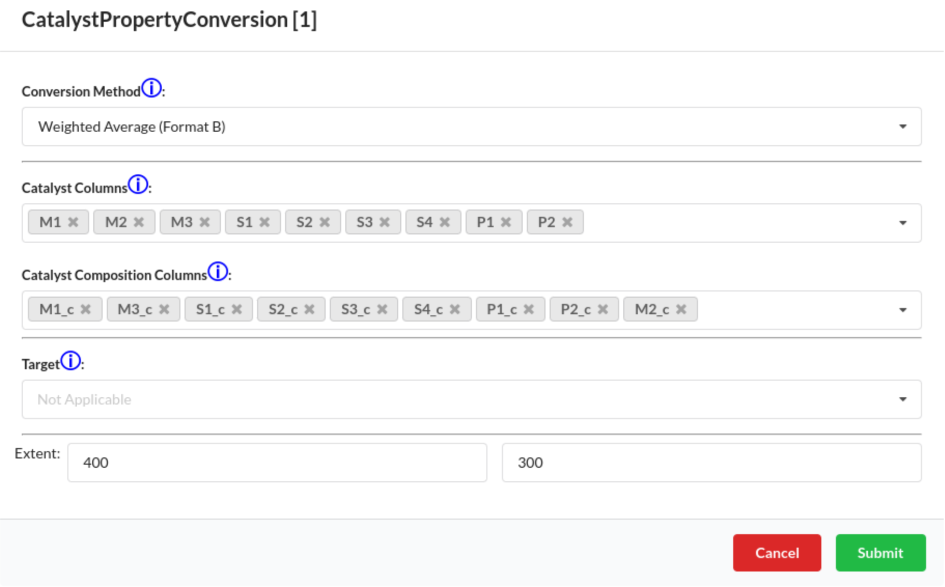
Conversion Method: Weighted Average (Format A)

Catalyst Columns: Mo ✖ V ✖ Sb ✖ Nb ✖ W ✖ Ce ✖

Target: Not Applicable

Extent: 400 300

Figure 4: **Form Fields of Weighted Average FormatA. This method converts a dataset consisting of information of catalyst components and composition in the shape of one 2D matrix with the catalysts' components as column headers, and all the compositions within the matrix. Each catalyst comprises a row. The difference of FormatA and FormatB is that the number of catalysts in dataset is one or multiple. While formatA is the dataset of the composition of one type of catalyst in Catalyst Component, formatB is the dataset of multiple catalysts included in catalyst components, and the composition information is contained in the same row.**



CatalystPropertyConversion [1]

Conversion Method:

Weighted Average (Format B)

Catalyst Columns:

M1 ✖ M2 ✖ M3 ✖ S1 ✖ S2 ✖ S3 ✖ S4 ✖ P1 ✖ P2 ✖

Catalyst Composition Columns:

M1_c ✖ M3_c ✖ S1_c ✖ S2_c ✖ S3_c ✖ S4_c ✖ P1_c ✖ P2_c ✖ M2_c ✖

Target:

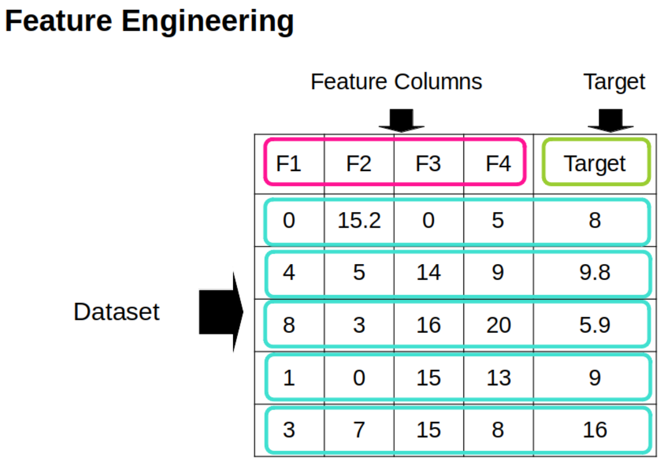
Not Applicable

Extent: 400 300

Figure 5: **Form Fields of Weighted Average FormatB. This method converts a dataset consisting of information of catalyst components and composition in the shape of two subsections next to each other: one comprised of columns depicting catalyst components, and the other depicting catalyst compositions.**

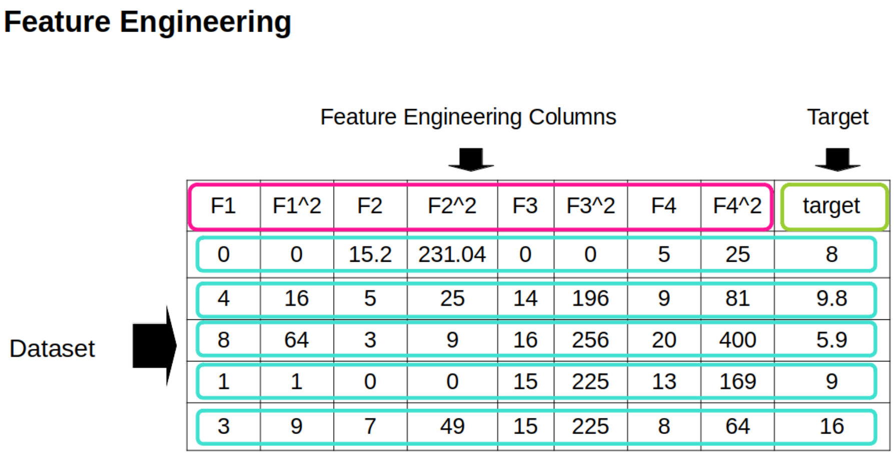
2.2 Feature Engineering

This component creates several First Order Feature analogues from existing numerical features in a dataset. The script trims the calculated features that are invariant or present NaN or infinite values toward the end. The data in the rows with blank are removed before feature engineering.



F1	F2	F3	F4	Target
0	15.2	0	5	8
4	5	14	9	9.8
8	3	16	20	5.9
1	0	15	13	9
3	7	15	8	16

Figure 6: Input CSV File Data Requirements Format.



F1	F1^2	F2	F2^2	F3	F3^2	F4	F4^2	target
0	0	15.2	231.04	0	0	5	25	8
4	16	5	25	14	196	9	81	9.8
8	64	3	9	16	256	20	400	5.9
1	1	0	0	15	225	13	169	9
3	9	7	49	15	225	8	64	16

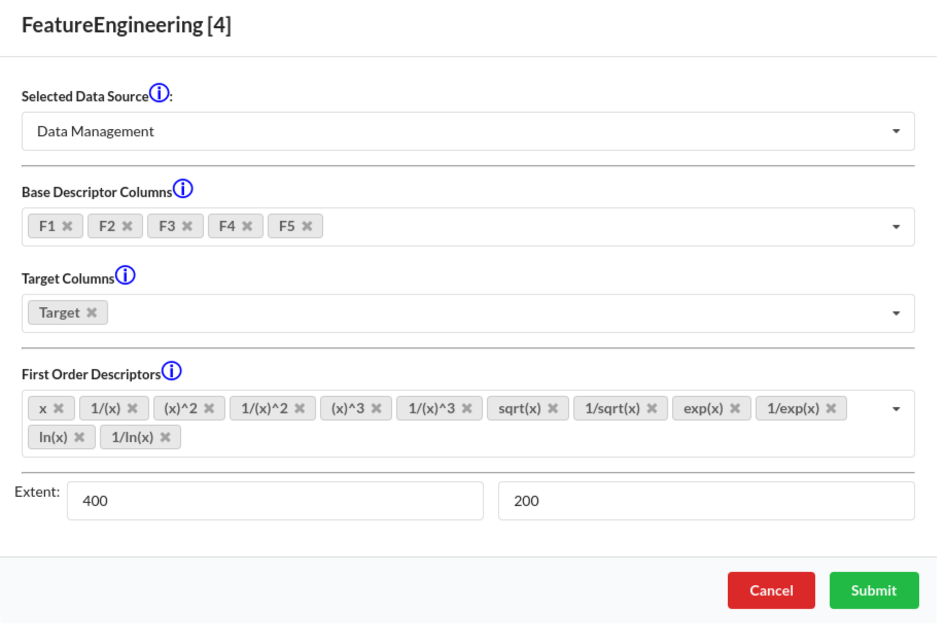
Figure 7: Output CSV File Data Format.

2.2.1 Form Fields

if user want to use dataset which upload to Datamangement, firstly user selects "Data Management" in "Selected Data Source" form.(See section 3 about how to use "Feature Assignment Component" in "Selected Data Source")

- **Form**
  Selected Data Source: User can choose to use the data source of Data Management or Feature Assignment Component.
  Base Descriptor Columns: This is existing numerical features which user has.
  Target Columns: This is target for the dataset.
  First Order Descriptors: This is the list of first order descriptors which user want to generate. User can select 12 types of first order descriptors: simple, inverse, square, inverse square, cube,

inverse cube, sqrt, inverse sqrt, exponential, inverse exponential, natural logarithm, inverse natural logarithm.



FeatureEngineering [4]

Selected Data Source ⓘ:

Data Management

Base Descriptor Columns ⓘ

F1 ✖ F2 ✖ F3 ✖ F4 ✖ F5 ✖

Target Columns ⓘ

Target ✖

First Order Descriptors ⓘ

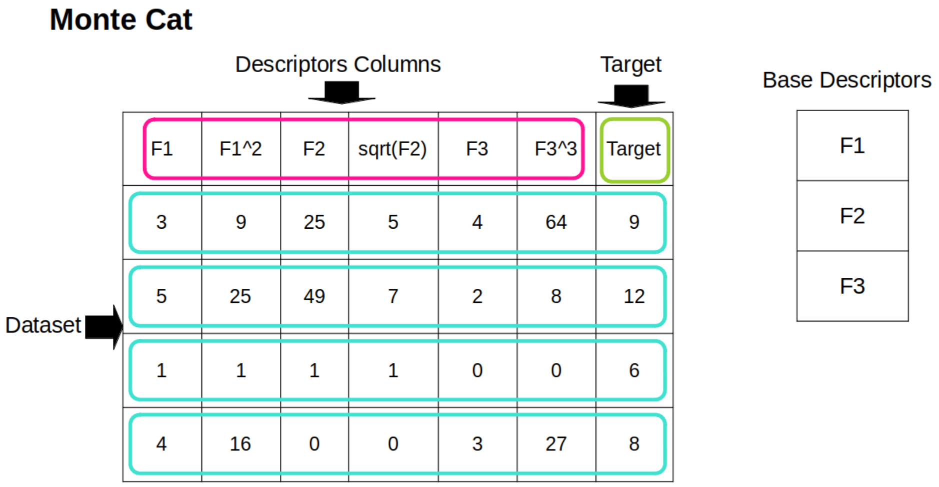
x ✖ 1/(x) ✖ (x)^2 ✖ 1/(x)^2 ✖ (x)^3 ✖ 1/(x)^3 ✖ sqrt(x) ✖ 1/sqrt(x) ✖ exp(x) ✖ 1/exp(x) ✖ ln(x) ✖ 1/ln(x) ✖

Extent: 400 200

Figure 8: Form Fields of Feature Engineering(selected Data Source: Data Management)

2.3 Monte Cat

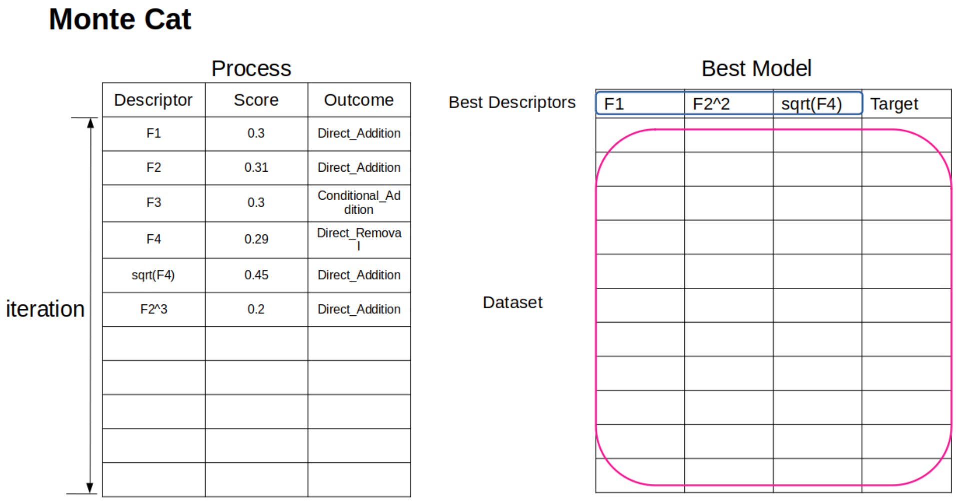
This component follows a forward descriptor addition procedure coupled with a Acceptance Criterion value to search for suitable variables in a regresion model from a big pool of engineered Descriptors. The number of Steps (Iterations) in the Descriptor search and the Temperature modulating the acceptance tolarance are determined by the user.



F1	F1^2	F2	sqrt(F2)	F3	F3^3	Target
3	9	25	5	4	64	9
5	25	49	7	2	8	12
1	1	1	1	0	0	6
4	16	0	0	3	27	8

Base Descriptors
F1
F2
F3

Figure 9: Input CSV File Data Requirements Format.



Descriptor	Score	Outcome
F1	0.3	Direct_Addition
F2	0.31	Direct_Addition
F3	0.3	Conditional_Addition
F4	0.29	Direct_Removal
sqrt(F4)	0.45	Direct_Addition
F2^3	0.2	Direct_Addition

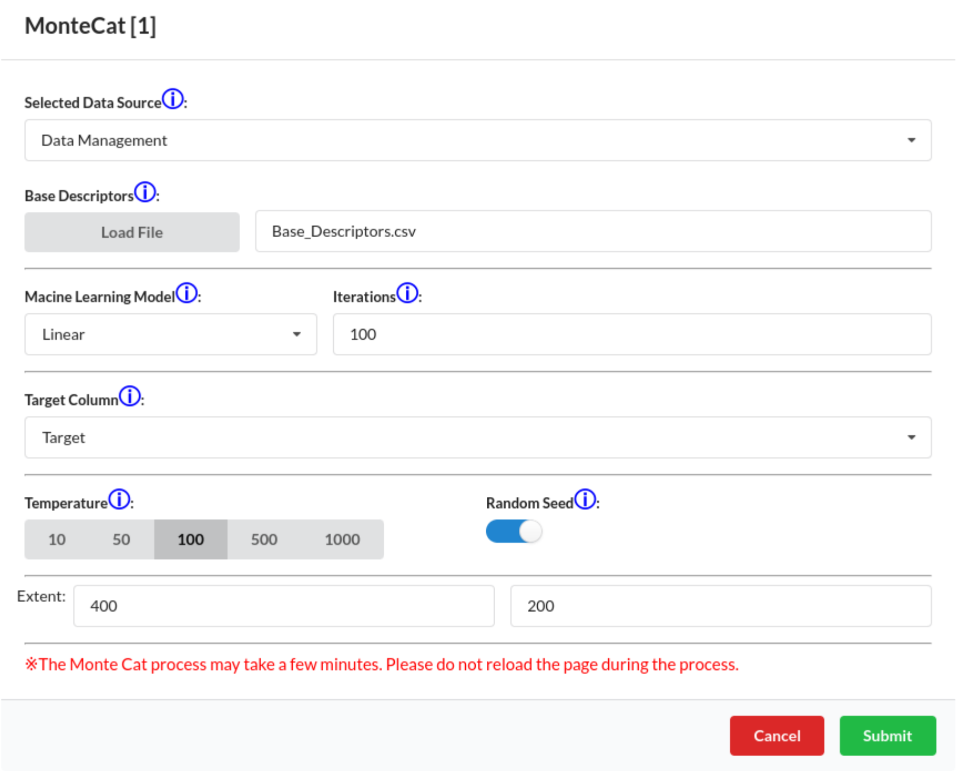
F1	F2^2	sqrt(F4)	Target

Figure 10: Output CSV File Data Format.

2.3.1 Select data set from "Data Management" as data source

if user want to use dataset which upload to Datamangement, firstly use selects "Data Management" in "Selected Data Source" form.(See section 3 about how to use "Feature Engineering Component" in "Selected Data Source")

- **Form Fields**
  Selected Data Source: User can choose to use the data source of Data Management or Feature Engineering Component.
  Base Descriptors: This file contains a list of the base Descriptor names prior to engineering the different analogues.CSV format is only allowed. This List is the same as Base Descriptor Columns of Feature Engineering Component .
  Macine Learning Model: machine learning model(Linear Model, Support Vector Regressio, Random Forest) to run MonteCat.
  Iterations:the number of iterations for the model, limited to 50-100 for RandomForest, 100-1000 for Linear and Support Vector Regression.
  Target Column: Objective variable.
  Temperature: Temperature parameter used to tune the Acceptance Probability curve behavior
  Random Seed: A specific random seed value can be selected by the user if reproducibility is desired. If not, the outcome will be randomized.



MonteCat [1]

Selected Data Source: Data Management

Base Descriptors: Base_Descriptors.csv

Macine Learning Model: Linear

Iterations: 100

Target Column: Target

Temperature: 10 50 100 500 1000

Random Seed:

Extent: 400 200

※The Monte Cat process may take a few minutes. Please do not reload the page during the process.

Figure 11: Form Fields of Monte Cat(selected Data Source: Data Management)

3 Manual for Combining Three Components

The three components can be combined by selecting the "Feature Assignment Component" and "Feature Engineering Component" in "selected Data Source" form of "Feature Engineering Component" and "Monte Cat Component", respectibely.(Figure14, 15) This is user-friendly because it allows three components to run simultaneously in a workspace.

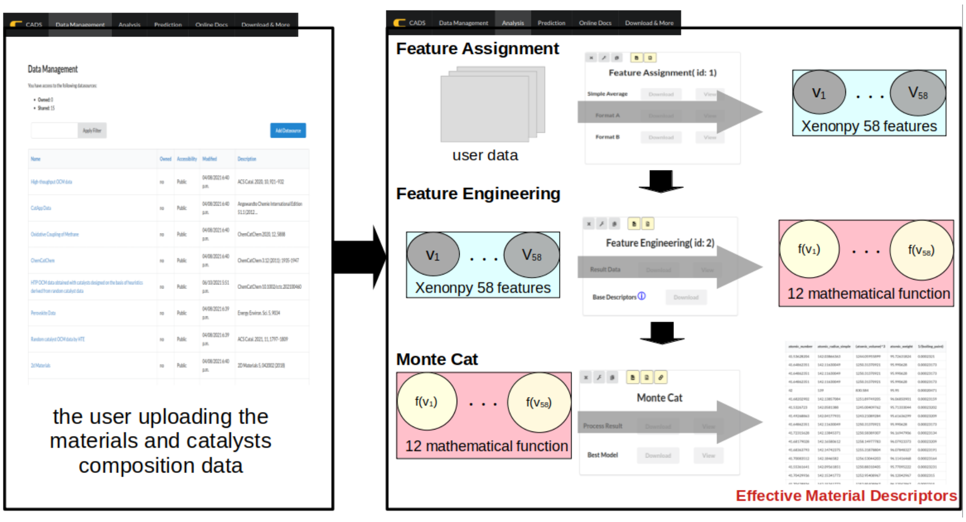

Data Management
the user uploading the materials and catalysts composition data
Feature Assignment
user data
Feature Assignment(id: 1)
V_1 . . . V_{58}
Xenonpy 58 features
Feature Engineering
Feature Engineering(id: 2)
$f(v_1)$ . . . $f(v_{58})$
12 mathematical function
Monte Cat
Monte Cat
Effective Material Descriptors

Figure 12: the architecture for combining three components

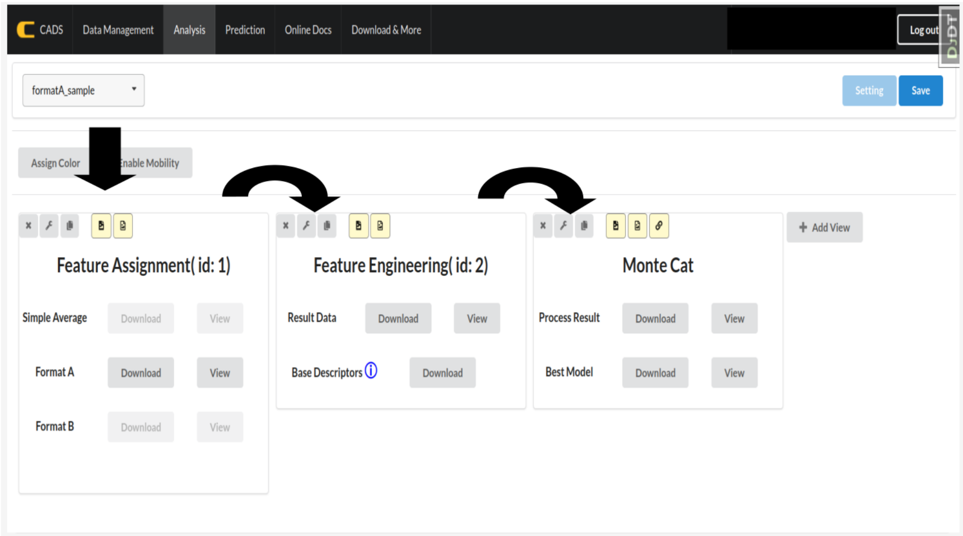
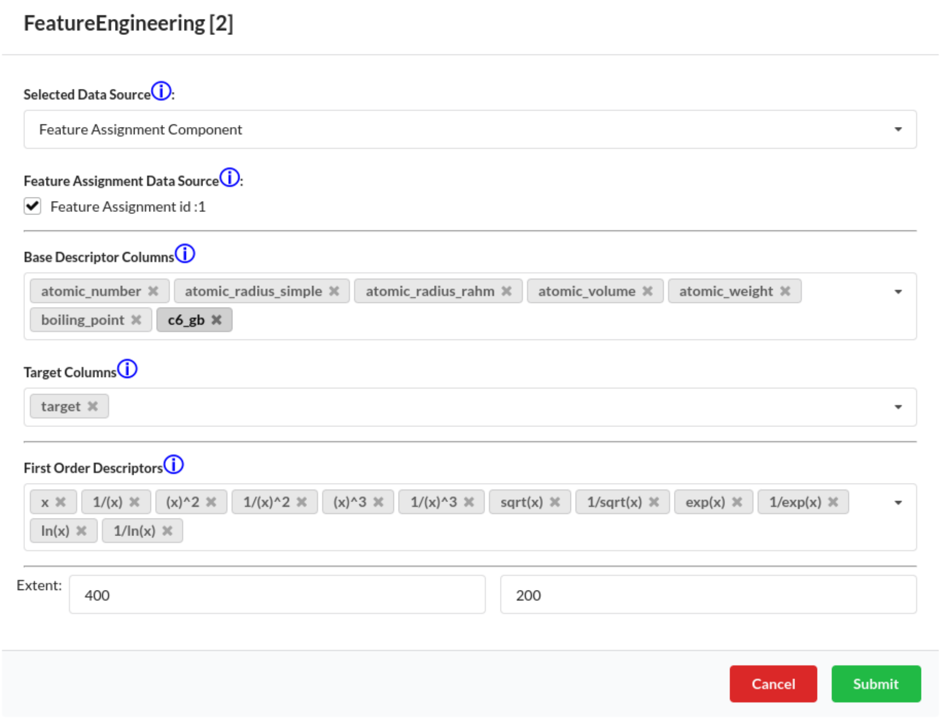


Figure 13: the workspace for combining three components



FeatureEngineering [2]

Selected Data Source: Feature Assignment Component

Feature Assignment Data Source:
☑ Feature Assignment id :1

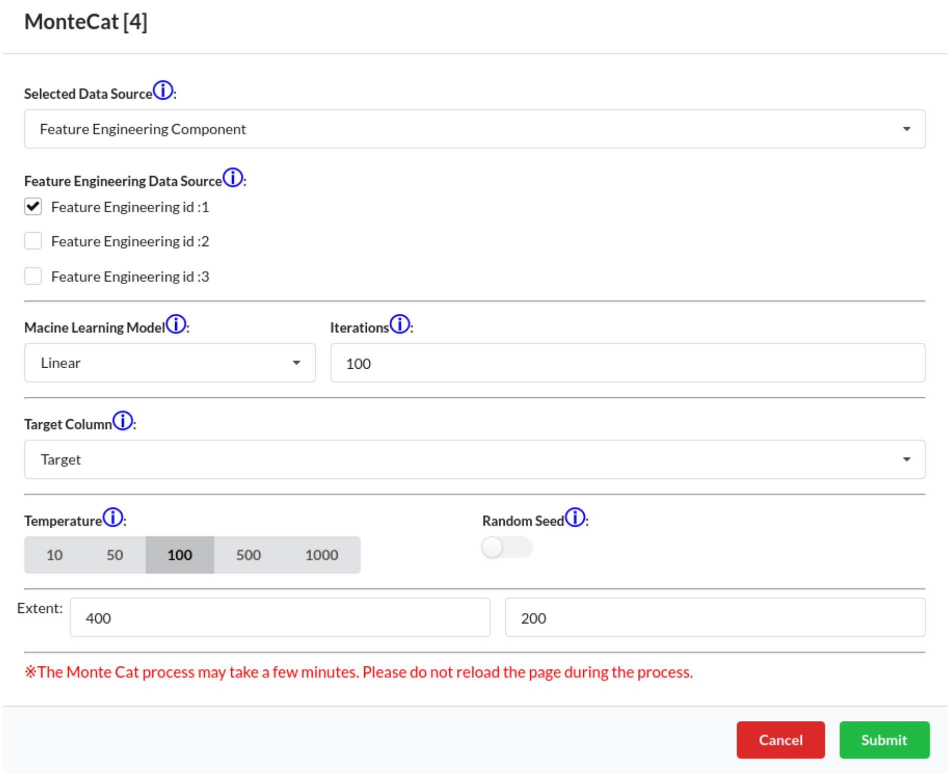
Base Descriptor Columns: atomic_number ✖ atomic_radius_simple ✖ atomic_radius_rahm ✖ atomic_volume ✖ atomic_weight ✖ boiling_point ✖ c6_gb ✖

Target Columns: target ✖

First Order Descriptors: x ✖ 1/(x) ✖ (x)^2 ✖ 1/(x)^2 ✖ (x)^3 ✖ 1/(x)^3 ✖ sqrt(x) ✖ 1/sqrt(x) ✖ exp(x) ✖ 1/exp(x) ✖ ln(x) ✖ 1/ln(x) ✖

Extent: 400 200

Figure 14: Form Fields of Feature Engineering(selected Data Source: Feature Assignment Component)



MonteCat [4]

Selected Data Source: Feature Engineering Component

Feature Engineering Data Source:
☑ Feature Engineering id :1
☐ Feature Engineering id :2
☐ Feature Engineering id :3

Macine Learning Model: Linear

Iterations: 100

Target Column: Target

Temperature: 10 50 **100** 500 1000

Random Seed:

Extent: 400 200

※The Monte Cat process may take a few minutes. Please do not reload the page during the process.

Figure 15: Form Fields of Monte Cat(selected Data Source: Feature Engineering Component)