```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, multilabel_confusion_matrix, classification_report
```

```
data = pd.read_csv('polymerase_cluster-BI.csv')
data.head()
```

	index	ABS	0	1	2	3	4	5	6	7	...	G1	G2	G3	G4	G
0	1	abstract astrocytes produce granulocytemacroph...	-0.050448	0.017385	-0.039777	-0.067159	-0.029633	0.074573	-0.050444	-0.010799	...	0	0	0	0	
1	2	abstract replication of avian infectious bronc...	-0.128422	-0.084803	0.084813	-0.013748	0.006486	0.128668	0.032655	0.066775	...	0	0	0	0	
2	3	abstract the infectivity of vesicular stomatit...	-0.095019	-0.032279	0.017571	-0.065860	0.001315	0.048199	-0.031072	0.010103	...	0	0	0	0	

```
data.dropna(inplace=True)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2782 entries, 0 to 2782
Data columns (total 42 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   index   2782 non-null   object
 1   ABS     2782 non-null   object
 2   0       2782 non-null   float64
 3   1       2782 non-null   float64
 4   2       2782 non-null   float64
 5   3       2782 non-null   float64
 6   4       2782 non-null   float64
 7   5       2782 non-null   float64
 8   6       2782 non-null   float64
 9   7       2782 non-null   float64
 10  8       2782 non-null   float64
 11  9       2782 non-null   float64
 12  10      2782 non-null   float64
 13  11      2782 non-null   float64
 14  12      2782 non-null   float64
 15  13      2782 non-null   float64
 16  14      2782 non-null   float64
 17  15      2782 non-null   float64
 18  16      2782 non-null   float64
 19  17      2782 non-null   float64
 20  18      2782 non-null   float64
 21  19      2782 non-null   float64
 22  20      2782 non-null   float64
 23  21      2782 non-null   float64
 24  22      2782 non-null   float64
 25  23      2782 non-null   float64
 26  24      2782 non-null   float64
 27  25      2782 non-null   float64
 28  26      2782 non-null   float64
 29  27      2782 non-null   float64
 30  28      2782 non-null   float64
 31  29      2782 non-null   float64
 32  G1      2782 non-null   int64
 33  G2      2782 non-null   int64
 34  G3      2782 non-null   int64
 35  G4      2782 non-null   int64
 36  G5      2782 non-null   int64
 37  G6      2782 non-null   int64
 38  G7      2782 non-null   int64
 39  G8      2782 non-null   int64
 40  G9      2782 non-null   int64
 41  G10     2782 non-null   float64
dtypes: float64(31), int64(9), object(2)
memory usage: 934.6+ KB
```

```
data.describe()
```

	0	1	2	3	4	5	6	7	8	9
count	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000	2782.000000
mean	-0.006256	-0.006805	0.000002	-0.001416	-0.000034	-0.002247	0.000039	0.000487	-0.001165	0.000761
std	0.135629	0.095939	0.082939	0.082707	0.079708	0.077969	0.073826	0.068710	0.066997	0.068602
min	-0.284363	-0.210271	-0.259660	-0.199337	-0.202987	-0.237105	-0.255140	-0.355713	-0.213472	-0.230506
25%	-0.113374	-0.073097	-0.055179	-0.060157	-0.047112	-0.055859	-0.047340	-0.029521	-0.041757	-0.039536
50%	-0.024205	-0.026229	0.001556	-0.015129	-0.012831	-0.007865	-0.008074	-0.000190	-0.004953	0.000924
75%	0.092567	0.038582	0.055740	0.044334	0.033604	0.044956	0.041129	0.029913	0.032115	0.040680
max	0.389576	0.343473	0.269622	0.307179	0.402218	0.358950	0.319398	0.416263	0.437870	0.266771

8 rows × 40 columns

```
data.isna().sum()
```

	0
index	0
ABS	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	0
23	0
24	0
25	0
26	0
27	0
28	0
29	0
G1	0
G2	0
G3	0
G4	0
G5	0
G6	0
G7	0
G8	0
G9	0
G10	0

dtype: int64

```
data.columns
```

```
Index(['index', 'ABS', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10',
       '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22',
       '23', '24', '25', '26', '27', '28', '29', 'G1', 'G2', 'G3', 'G4', 'G5',
       'G6', 'G7', 'G8', 'G9', 'G10'],
      dtype='object')
```

```
X = data[['0','1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17','18','19','20','21','22','23','24','25','26',
y = data[['G1','G2','G3','G4','G5','G6','G7','G8','G9']]
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=42)
```

```
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier()
```

```
y_pred = model.predict(X_test)
```

```
acc = accuracy_score(y_test,y_pred)
print('Accuracy:',acc*100,'%')
```

```
Accuracy: 92.63913824057451 %
```

```
cr = classification_report(y_test,y_pred)
print('Classification Report:',cr)
```

```
Classification Report:               precision    recall  f1-score   support

                0       1.00      1.00      1.00         8
                1       1.00      0.96      0.98        68
                2       1.00      0.94      0.97        77
                3       1.00      0.93      0.97        45
                4       0.97      0.96      0.96        67
```