

- Setup

```
In [1]: import sys, os
current_directory = os.getcwd()
root_directory = os.path.abspath(os.path.join(current_directory, os.pardir))
sys.path.append(root_directory)
```

- Import utils

```
In [9]: from pretrained.predictor import SingleKModel, MultiKModel, OneTestKModel, BatchSingleKModel
import metrics
```

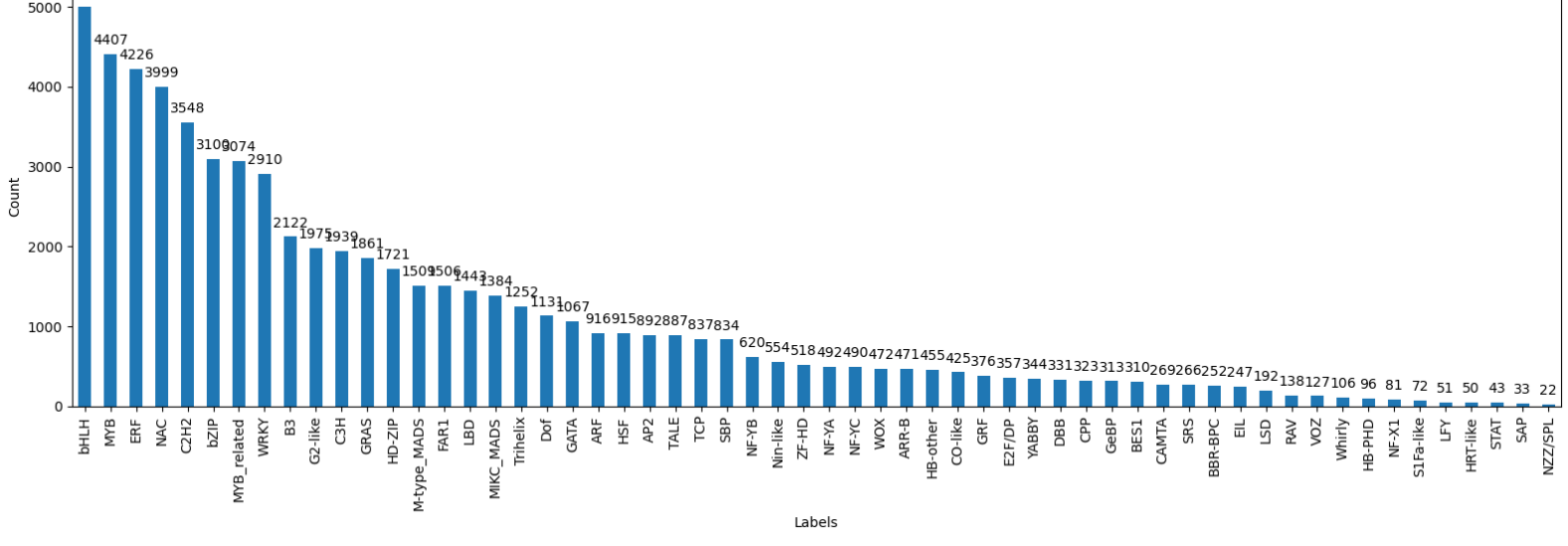
- Predict

1- Using single K Model

```
In [3]: kmodel = BatchSingleKModel(kmer_size=4, batch_size=2000)
kmodel.set_load_config("../data/testset-full/k4/testset.csv", format="csv", type='kmer_file')
genboard = kmodel.predict()
```

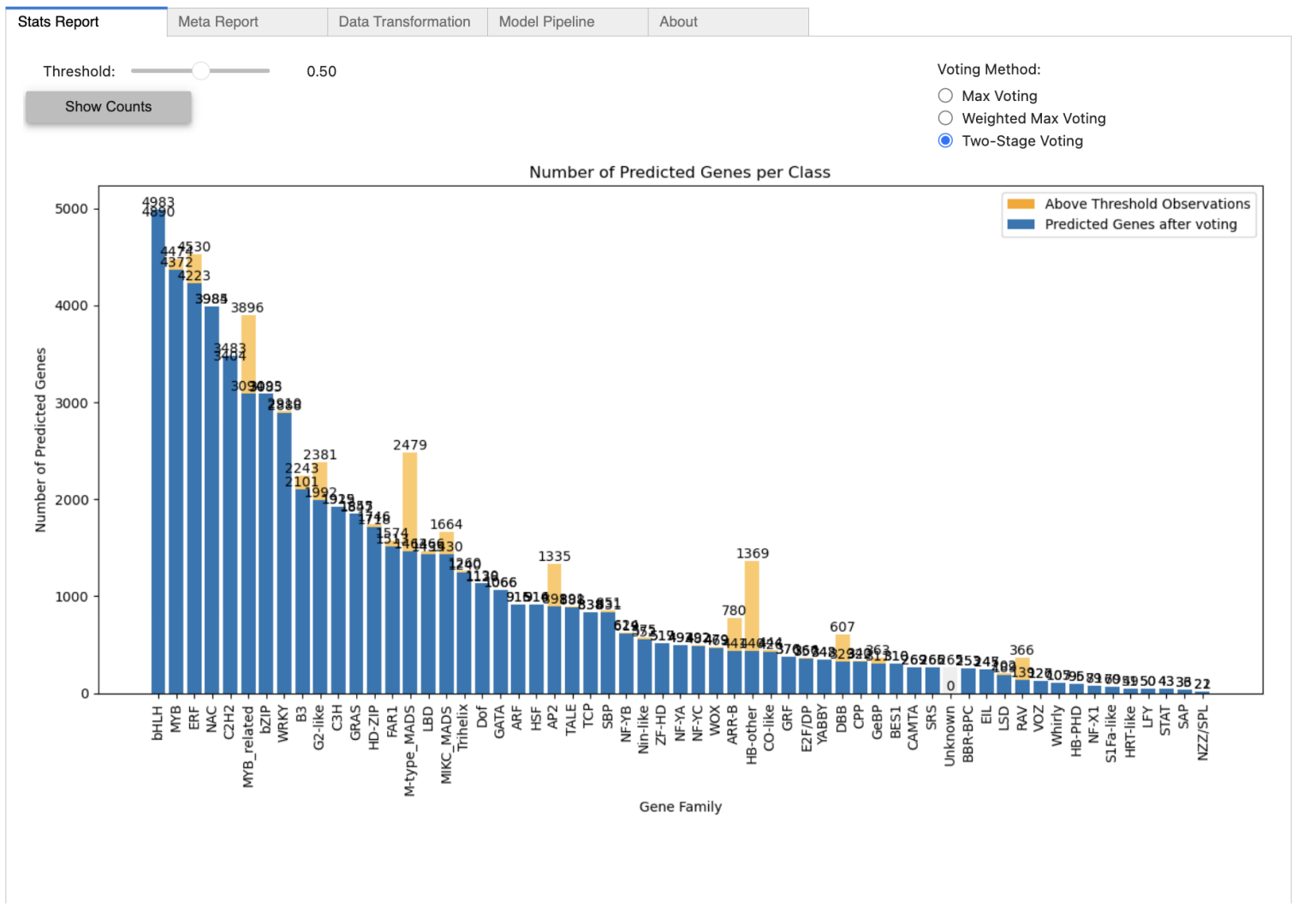
Batch Predictions: 32it [31:35, 59.23s/it]

```
In [4]: metrics.plot_testset(true_label_df_path="../data/testset-full/k4/true_labels.csv", class_mapping_df
```



```
In [5]: genboard.display()
```

Tab(children=(VBox(children=(HBox(children=(VBox(children=(FloatSlider(value=0.5, continuous_update=False, des...



```
In [6]: genboard.prediction.shape
```

Out[6]: (63359, 58)

- Confusion report for validation

```
In [7]: import json
import pandas as pd

true_label = pd.read_csv('../data/testset-full/k4/true_labels.csv')['true_label'].values
with open('../data/testset-full/k4/class_mapping.json', 'r') as json_file:
    class_mapping = json.load(json_file)
class_mapping['Unknown'] = 0
```

```
In [8]: genboard.show_eval_metric(
    true_label=true_label,
    class_mapping_rules=class_mapping,
    voting_method="Two-Stage Voting",
    voting_threshold=0.5,
    binary_class_threshold=0.5,
    components=['confusion_matrix', 'general_accuracy', 'accuracy_per_family']
)
```

Overall Accuracy

kmer_size=4

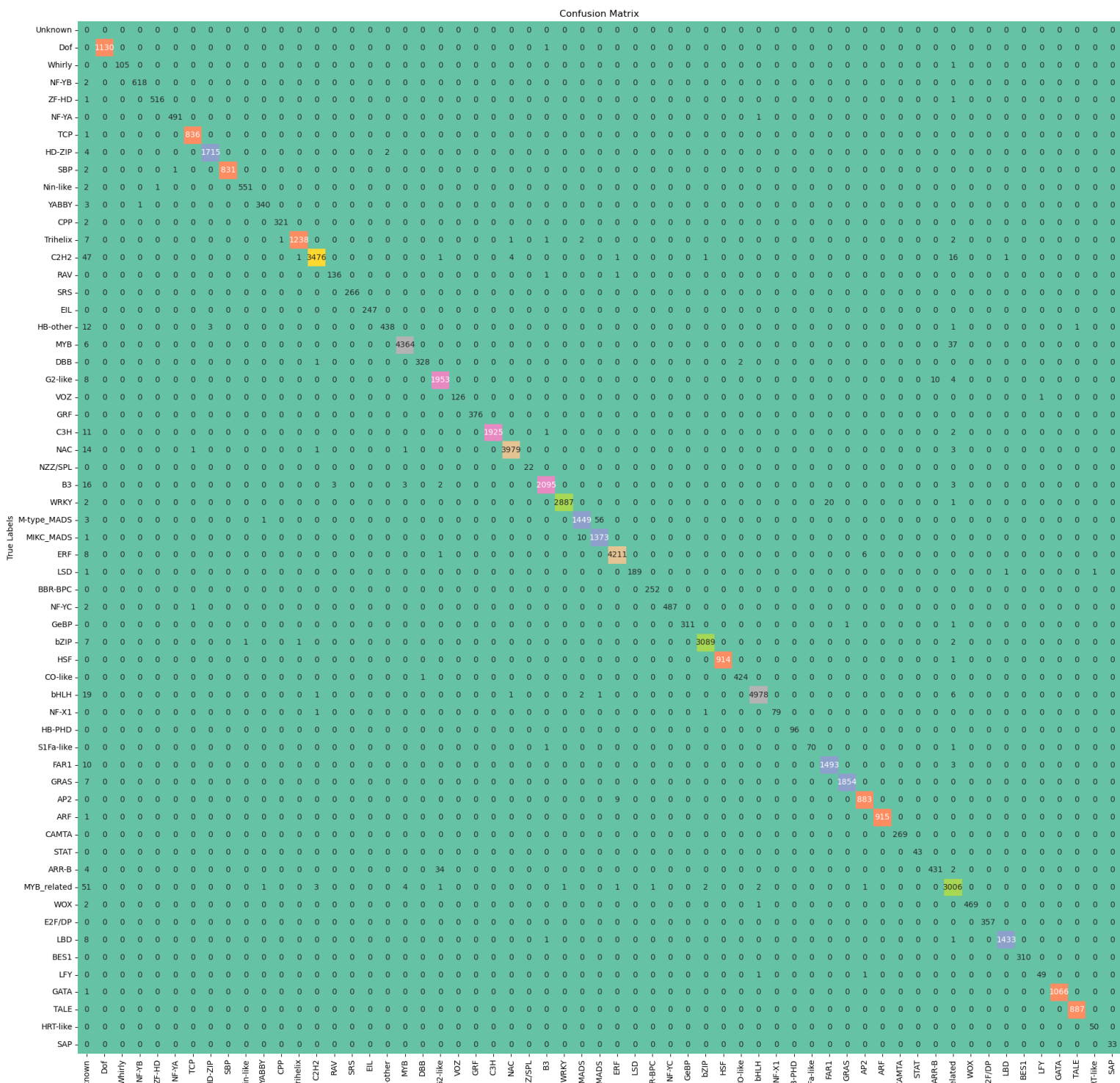
Score	0.99
-------	------

Accuracy per Gene Family

Gene Family	Accuracy	Precision	Recall	F1 Score
AP2	0.99	0.66	0.98	0.79
ARR-B	0.99	0.59	0.99	0.74
BBR-BPC	1.00	1.00	1.00	1.00
C2H2	1.00	0.98	0.94	0.96
CAMTA	1.00	1.00	1.00	1.00
CPP	1.00	0.94	0.99	0.97
Dof	1.00	1.00	1.00	1.00
EIL	1.00	0.99	0.98	0.99
FAR1	1.00	0.93	0.97	0.95
GATA	1.00	1.00	1.00	1.00
GRF	1.00	1.00	1.00	1.00
HB-PHD	1.00	1.00	0.99	0.99
HD-ZIP	1.00	0.98	1.00	0.99
HSF	1.00	1.00	1.00	1.00
LFY	1.00	1.00	0.98	0.99
M-type_MADS	0.98	0.60	0.98	0.75
MYB	1.00	0.97	0.99	0.98
NAC	1.00	0.99	0.99	0.99
NF-YA	1.00	1.00	1.00	1.00
NF-YC	1.00	0.99	0.99	0.99
Nin-like	1.00	0.94	0.97	0.95
S1Fa-like	1.00	1.00	0.96	0.98
SBP	1.00	0.98	1.00	0.99
STAT	1.00	1.00	1.00	1.00
TCP	1.00	1.00	1.00	1.00
VOZ	1.00	1.00	1.00	1.00
WRKY	1.00	1.00	1.00	1.00
YABBY	1.00	0.97	0.98	0.97
bHLH	1.00	0.99	0.97	0.98

Gene Family	Accuracy	Precision	Recall	F1 Score
ARF	1.00	1.00	1.00	1.00
B3	1.00	0.92	0.97	0.94
BES1	1.00	1.00	1.00	1.00
C3H	1.00	0.98	0.97	0.98
CO-like	1.00	0.95	1.00	0.97
DBB	1.00	0.54	1.00	0.70
E2F/DP	1.00	0.98	1.00	0.99
ERF	0.99	0.93	1.00	0.96
G2-like	0.99	0.82	0.99	0.90
GRAS	1.00	0.99	0.99	0.99
GeBP	1.00	0.83	0.96	0.89
HB-other	0.98	0.29	0.89	0.44
HRT-like	1.00	1.00	0.98	0.99
LBD	1.00	0.97	0.99	0.98
LSD	1.00	0.95	0.99	0.97
MIKC_MADS	1.00	0.83	1.00	0.90
MYB-related	0.98	0.71	0.90	0.79
NF-X1	1.00	1.00	1.00	1.00
NF-YB	1.00	0.99	1.00	0.99
NZZ/SPL	1.00	1.00	0.95	0.98
RAV	1.00	0.37	0.99	0.54
SAP	1.00	0.92	1.00	0.96
SRS	1.00	1.00	1.00	1.00
TALE	1.00	1.00	1.00	1.00
Trihelix	1.00	0.97	0.98	0.98
WOX	1.00	0.98	1.00	0.99
Whirly	1.00	0.99	1.00	1.00
ZF-HD	1.00	0.99	0.99	0.99
bZIP	1.00	0.99	0.99	0.99

Confusion Matrix



Save Report

```
In [ ]:
```