APPLIED ML PROJECT

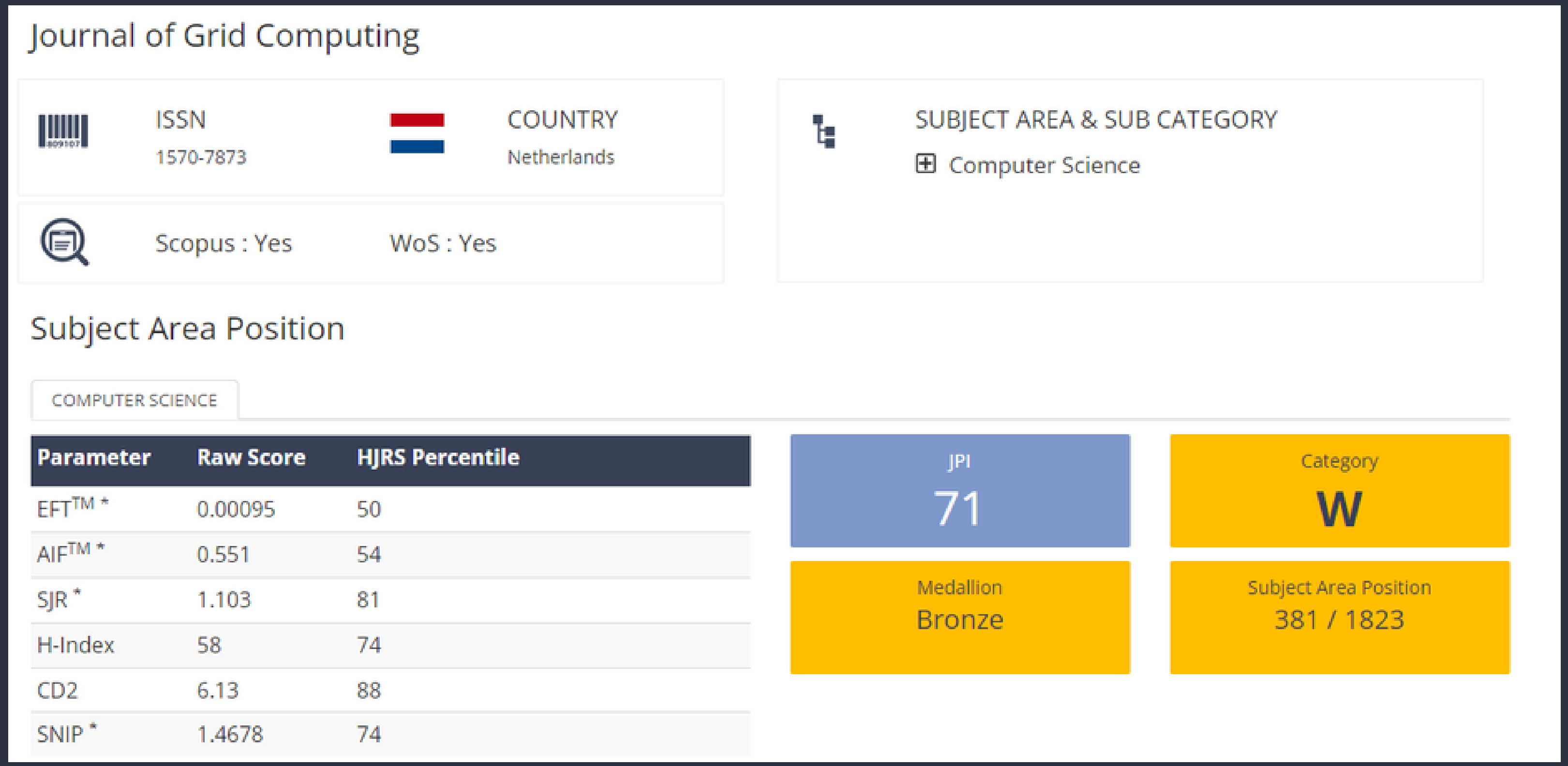
Presented by: Bilal Ahmed Khan | K200183

ABOUT THE RESEARCH PAPER

- Name: **Beyond Homology Transfer: Deep Learning for Automated Annotation of Proteins**
- Published in: **Journal of Grid Computing**
- Authors:
  - Mohammad Nauman
  - Hafeez Ur Rehman
  - Gianfranco Politano
  - Alfredo Benso

MEDALLION OF THE RESEARCH PAPER

- The Journal of Grid Computing is a top-rated journal with a **‘W Category’** and **‘Bronze Medallion’**



Journal of Grid Computing

ISSN
1570-7873

COUNTRY
Netherlands

SUBJECT AREA & SUB CATEGORY
⊞ Computer Science

Scopus : Yes

WoS : Yes

Subject Area Position

COMPUTER SCIENCE

Parameter	Raw Score	HJRS Percentile
EFT^{TM} *	0.00095	50
AIF^{TM} *	0.551	54
SJR *	1.103	81
H-Index	58	74
CD2	6.13	88
SNIP *	1.4678	74

JPI
71

Category
W

Medallion
Bronze

Subject Area Position
381 / 1823

PAPER ABSTRACT

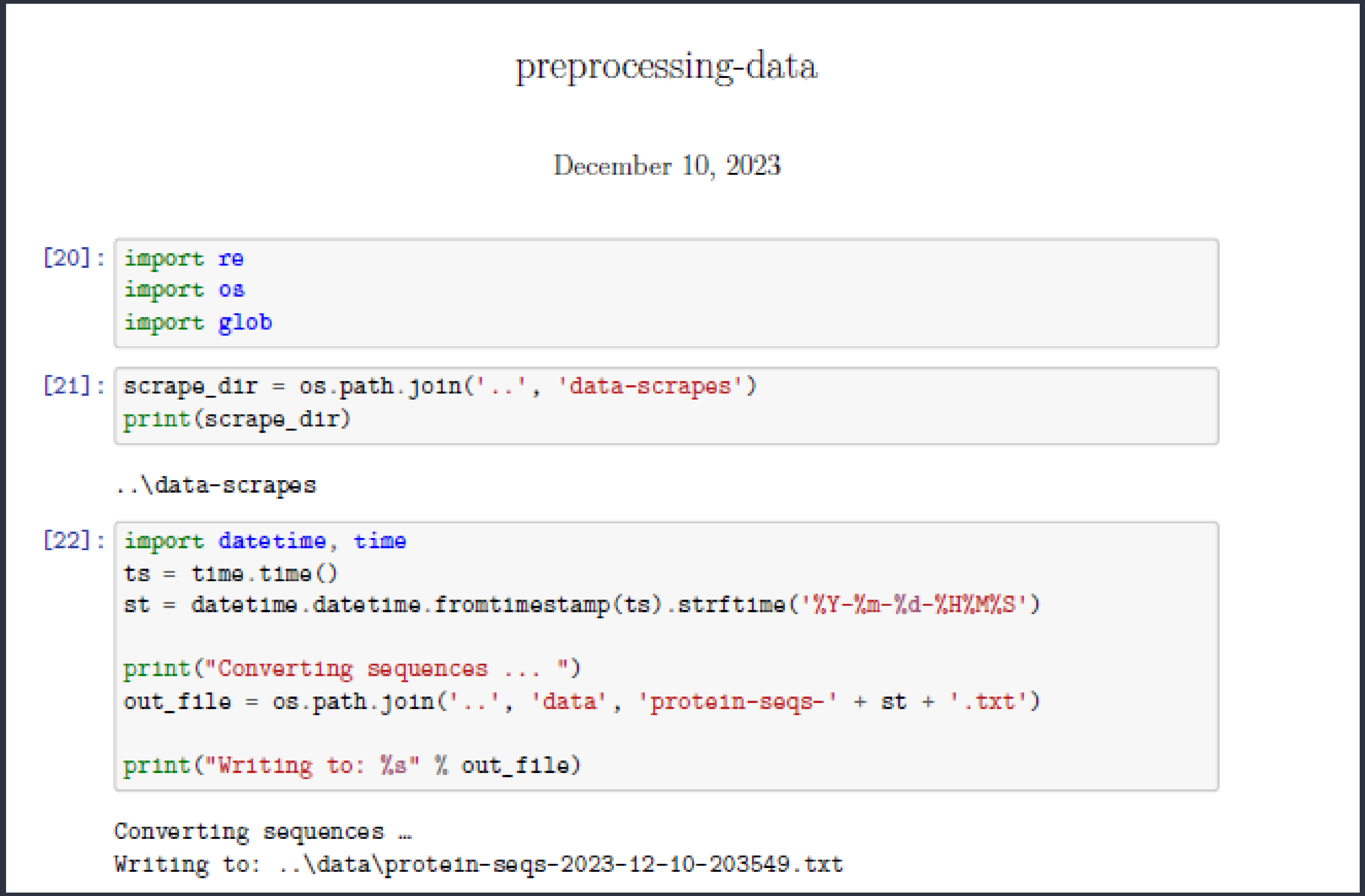
The research addresses the challenge of accurately annotating protein functions, particularly for uncharacterized proteins with limited supporting information beyond their amino acid sequences. We introduce DeepSeq, a novel deep learning architecture that relies solely on protein sequence data to predict associated functions.

Unlike traditional methods, DeepSeq does not require handcrafted features, automatically extracting representations from input sequences. Experimental results demonstrate a significant improvement in prediction accuracy compared to other sequence-based approaches, achieving an impressive 86.72% validation accuracy and a 71.13% F1 score.

Remarkably, DeepSeq's automatically learned features enable successful resolution of related tasks, such as protein function localization, without human intervention. Our findings suggest the potential for applying DeepSeq to more complex challenges, including predicting 2D and 3D protein structures and protein-protein interactions.

PRE-PROCESSING DATA

The code for pre-processing the data can be inspected in the *‘preprocessing-data.ipynb’* file



preprocessing-data

December 10, 2023

```
[20]: import re
      import os
      import glob
```

```
[21]: scrape_dir = os.path.join('..', 'data-scrapes')
      print(scrape_dir)
```

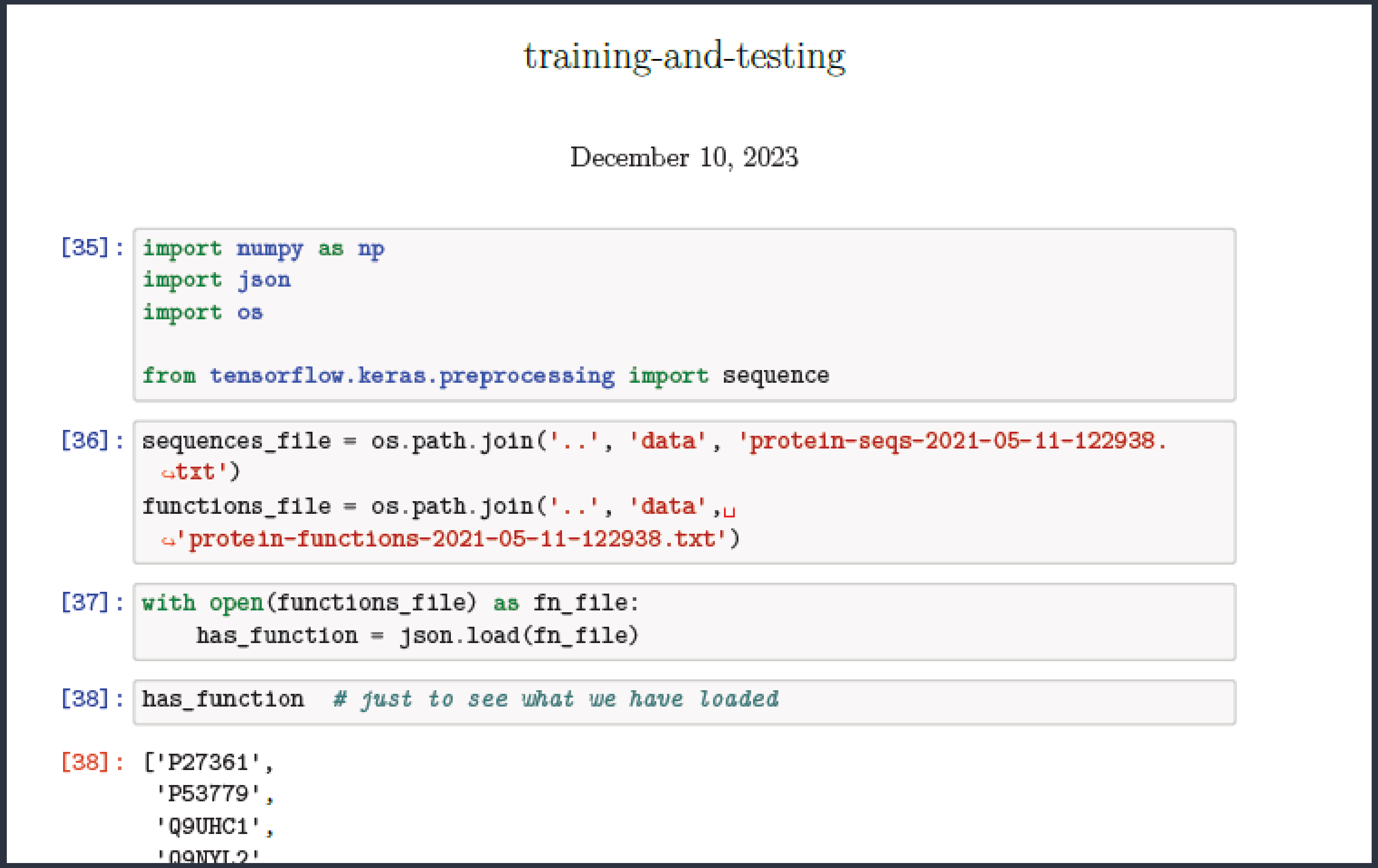
```
..\data-scrapes
```

```
[22]: import datetime, time
      ts = time.time()
      st = datetime.datetime.fromtimestamp(ts).strftime('%Y-%m-%d-%H%M%S')

      print("Converting sequences ... ")
      out_file = os.path.join('..', 'data', 'protein-seqs-' + st + '.txt')

      print("Writing to: %s" % out_file)
```

```
Converting sequences …
Writing to: ..\data\protein-seqs-2023-12-10-203549.txt
```

TRAINING AND TESTING THE MODEL

The code for pre-processing the data can be inspected in the ‘*training-and-testing.ipynb*’ file



training-and-testing

December 10, 2023

```
[35]: import numpy as np
      import json
      import os

      from tensorflow.keras.preprocessing import sequence
```

```
[36]: sequences_file = os.path.join('..', 'data', 'protein-seqs-2021-05-11-122938.
        ↪txt')
      functions_file = os.path.join('..', 'data',␣
        ↪'protein-functions-2021-05-11-122938.txt')
```

```
[37]: with open(functions_file) as fn_file:
          has_function = json.load(fn_file)
```

```
[38]: has_function  # just to see what we have loaded
```

```
[38]: ['P27361',
       'P53779',
       'Q9UHC1',
       'Q9NYL2'
```

THANK YOU!!!