# SuccKD\_MultiSpecies

**SuccKD\_MultiSpecies:** A novel approach exploiting Multi-Teacher Knowledge Distillation and Word Embedding for Cross-Species Prediction of Protein Succinylation Sites.

# Requirement

* Goolge Colab pro
* Keras
* Numpy
* Sklearn
* Pandas
* Tensorflow

**Dataset**

We have developed an architecture for predicting succinylation sites in both generic and cross-species contexts. The dataset, sourced from SuccinSite2.0 [[12](#_heading=h.1fob9te)] and GPSuc [[10](#_heading=h.3znysh7)], mirrors their training and test sets to ensure consistency in model development and comparison. It includes data from nine species: *Homo sapiens (H.sapiens), Mus musculus (M.musculus), Escherichia coli (E.coli), Mycobacterium tuberculosis (M.tuberculosis), Saccharomyces cerevisiae (S.cerevisiae), Toxoplasma gondii (T.gondii), Solanum lycopersicum (S.lycopersicum)*, and others.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Species** | **Dataset** | **Succinylated proteins** | **Succinylation sites** | **Nonsuccinylation sites** |
| Generic | Training | 2,198 | 4,750 | 9,500 |
| Testing | 124 | 254 | 2,977 |
| H.sapiens | Training | 500 | 1,351 | 2,701 |
| Testing | 50 | 54 | 2,004 |
| H.capsulatum | Training | 150 | 334 | 665 |
| Testing | 33 | 50 | 591 |
| M.musculus | Training | 240 | 414 | 828 |
| Testing | 24 | 24 | 679 |
| E.coli | Training | 786 | 1942 | 3,884 |
| Testing | 79 | 289 | 1,381 |
| M.tuberculosis | Training | 369 | 699 | 1,398 |
| Testing | 36 | 61 | 242 |
| S.cerevisiae | Training | 364 | 961 | 1,922 |
| Testing | 36 | 90 | 1,423 |
| T.gondii | Training | 98 | 282 | 564 |
| Testing | 10 | 26 | 261 |
| S.lycopersicum | Training | 150 | 242 | 484 |
| Testing | 16 | 33 | 274 |
| T.aestivum | Training | 53 | 113 | 226 |
| Testing | 20 | 31 | 310 |

**Model Learning:**

We use Google colab pro buid this model

* Cross validation: CV\_Generics\_KD2\_Succi.ipynb
* Train model: Generic\_ KD2\_Succi.ipynb.

Model saved and named: KD2\_Succi.h5. You can use KD2\_Succi.h5 in the Model folder for predicting and independent testing.

* Independent test and predict files are located in the Code\_KD2\_Succi \_Generic\_predict\_Species folder:

|  |  |  |
| --- | --- | --- |
| **No.** | **Codes using 2teacher\_1student.h5 model to predict nine species** | **Species** |
| 1 | predict\_H.Sapien\_from\_Generic.ipynb | H.sapiens |
| 2 | predict\_H.capsulatum\_from\_Generic.ipynb | H.capsulatum |
| 3 | predict\_M.musculus\_from\_Generic.ipynb | M.musculus |
| 4 | predict\_E.coli\_from\_Generic.ipynb | E.coli |
| 5 | predict\_M.tuberculosis\_from\_Generic.ipynb | M.tuberculosis |
| 6 | predict\_S. cerevisiae\_from\_Generic.ipynb | S.cerevisiae |
| 7 | predict\_T.gondii\_from\_Generic.ipynb | T.gondii |
| 8 | predict\_S.lycopersicum\_from\_Generic.ipynb | S.lycopersicum |
| 9 | predict\_T.eastivum\_from\_Generic.ipynb | T.aestivum |

# Contact

Please feel free to contact us if you need any help: [nvnui@ictu.edu.vn](mailto:nvnui@ictu.edu.vn)