Multitask Classification of Antimicrobial Peptides for Simultaneous Assessment of Antimicrobial Property and Structural Fold

Michaela Areti Zervou[1,2], Effrosyni Doutsi[2], Yannis Pantazis[3], Panagiotis Tsakalides[1,2]

[1]Computer Science Department, University of Crete, Greece

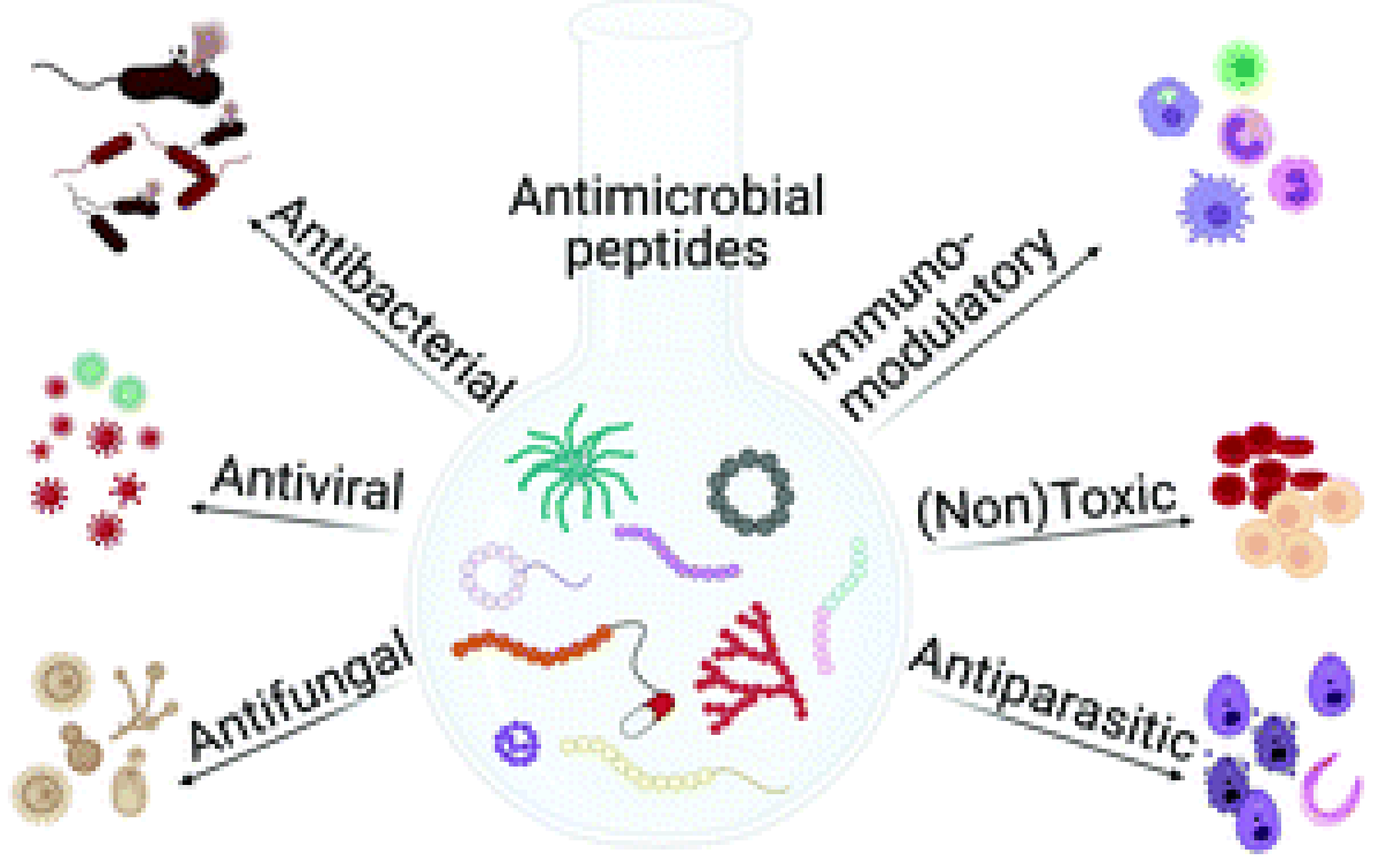
[2]Institute of Computer Science - FORTH, Crete, Greece

[3]Institute of Applied and Computational Mathematics - FORTH, Crete, Greece

E-mails: zervou@ics.forth.gr, edoutsi@ics.forth.gr, pantazis@iacm.forth.gr, tsakalid@ics.forth.gr

1. Antimicrobial Peptides

- ***Proteins*** are macromolecules that serve the ***biological functions*** of any ***living organism.***
- The ***function of proteins*** is highly associated with their ***three-dimensional structure***.
- An important category of proteins is the ***antimicrobial peptides (AMPs)*** that *play a significant role in guiding drug design, advancing targeted therapies, and cancer treatment research*.



- AMPs particularly favor ***alpha-helical structures***, or alpha-folds, due to their ability to disrupt the protective layers that surround cells effectively and their structural stability.

alpha-helix

2. Antimicrobial Peptide Classification

- ***State-of-the-art AMP classifiers*** focus on AMP property detection***, overlooking the structural fold's*** valuable insights into AMP function.

Disadvantages

- These classifiers ***rely on a multitude of biological attributes*** for accurate classification resulting in ***increased computational complexity***, potentially ***hindering efficiency and scalability***.

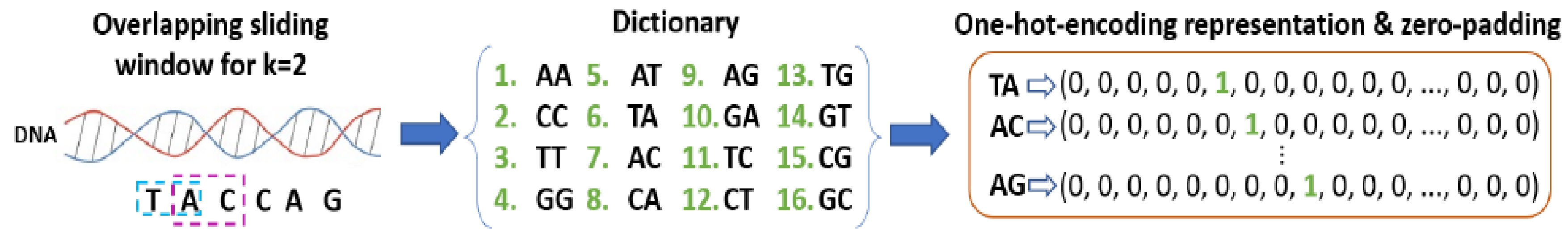
Solution: Multitask classification on the primary sequence.

3. Multitask Classification

- A ***multitask classifier*** operates by addressing ***multiple tasks simultaneously***, eliminating redundant processes, and ***optimizing computational usage***.
- In this case, it efficiently assesses ***two critical aspects***:
  - The presence of the ***AMP property*** within a given sequence,
  - The decision is whether the sequence exhibits an ***alpha-helical fold***.

4. Sequence Representation via k-mers Technique

- The ***k-mers*** technique is a Natural Language Process (NLP) tool that represents a sequence of characters as ***subsequences of length k***. It is extensively used in bioinformatics too as it can capture ***local structural and functional properties.***
- By considering both sequence and structural information, this representation can provide a ***more comprehensive understanding of AMPs***, enhancing the prediction accuracy and functional insight.



Overlapping sliding window for k=2
DNA
T A C C A G
Dictionary
One-hot-encoding representation & zero-padding

1. AA 5. AT 9. AG 13. TG
2. CC 6. TA 10. GA 14. GT
3. TT 7. AC 11. TC 15. CG
4. GG 8. CA 12. CT 16. GC

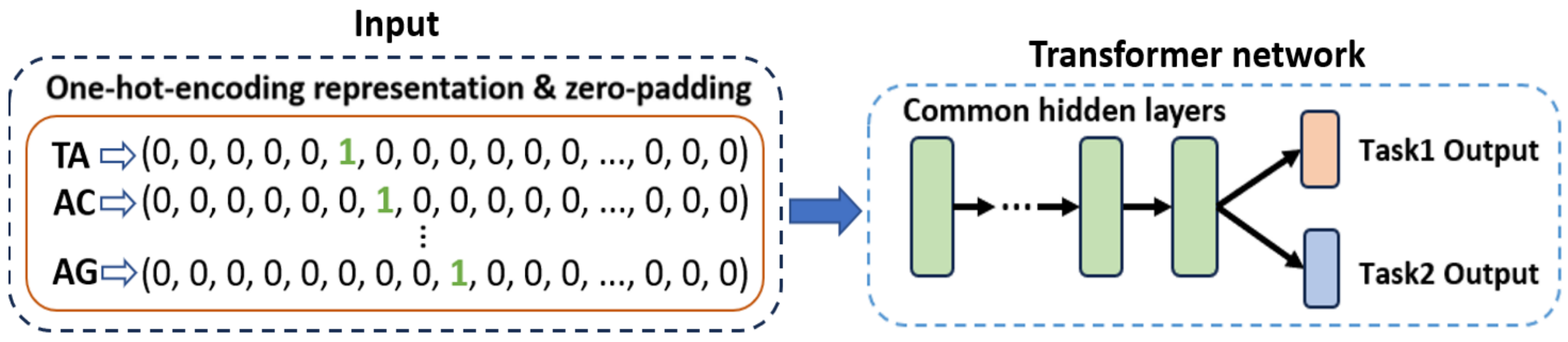
TA ⇨ (0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ..., 0, 0, 0)
AC ⇨ (0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0, 0, 0)
⋮
AG ⇨ (0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0, 0, 0)

5. Multitask Transformer Classifier

- For multitask learning, a ***combined loss function*** is employed to simultaneously optimize predictions for two classification tasks.

$$\mathcal{L}_{\text{combined}} = \lambda \cdot \mathcal{L}_1 + (1 - \lambda) \cdot \mathcal{L}_2$$

where $\mathcal{L}_1, \mathcal{L}_2$, represent the loss calculated for the first and the second classification task respectively, and λ is a hyperparameter that controls the balance between the two losses.



TA ⇨ (0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ..., 0, 0, 0)
AC ⇨ (0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0, 0, 0)
⋮
AG ⇨ (0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ..., 0, 0, 0)

6. Classification Results

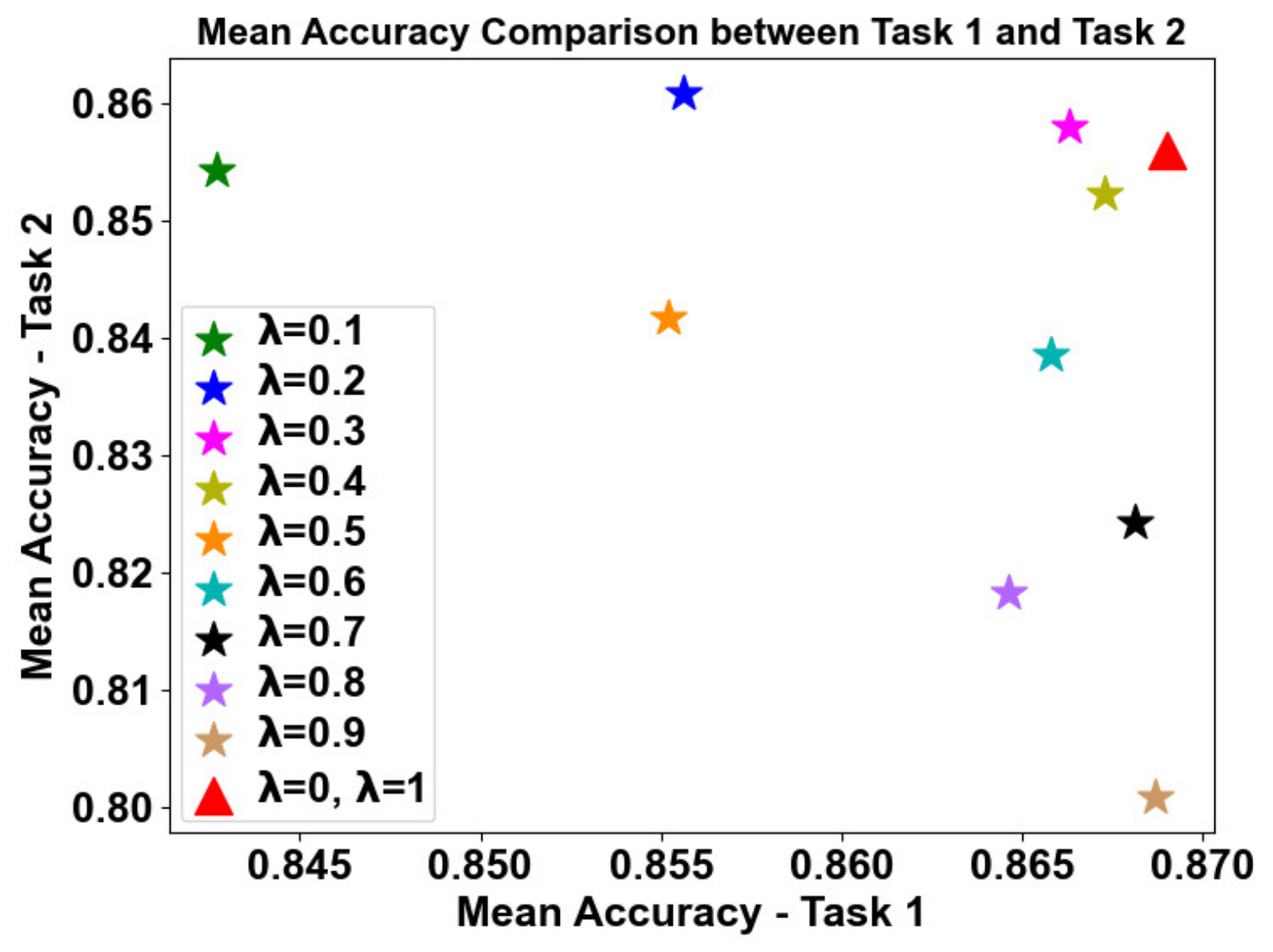
- ***Effect of λ values*** in multitask classification of ***AMP (Task 1)*** and ***alpha-fold (Task 2)***. Standard deviation is reported in parentheses.

Dataset 1	λ=0.2		λ=0.3		λ=0.4		λ=0.5		λ=1	λ=0
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Accuracy	85.6 (2.2)	86.1 (1.0)	86.6 (1.4)	85.8 (0.3)	86.5 (1.1)	85.2 (1.4)	85.5 (1.3)	84.2 (0.9)	86.9 (1.2)	85.6 (0.8)
F1-score	85.7 (2.0)	87.3 (0.9)	86.6 (1.4)	87.2 (0.3)	86.8 (1.2)	86.9 (1.1)	85.3 (1.6)	85.6 (0.9)	87.1 (1.1)	87.1 (0.7)
AUC	93.7 (1.1)	93.5 (0.6)	94.3 (0.9)	93.1 (0.2)	94.7 (0.6)	93.2 (0.9)	94.3 (0.8)	92.0 (0.4)	94.4 (0.8)	93.3 (0.4)

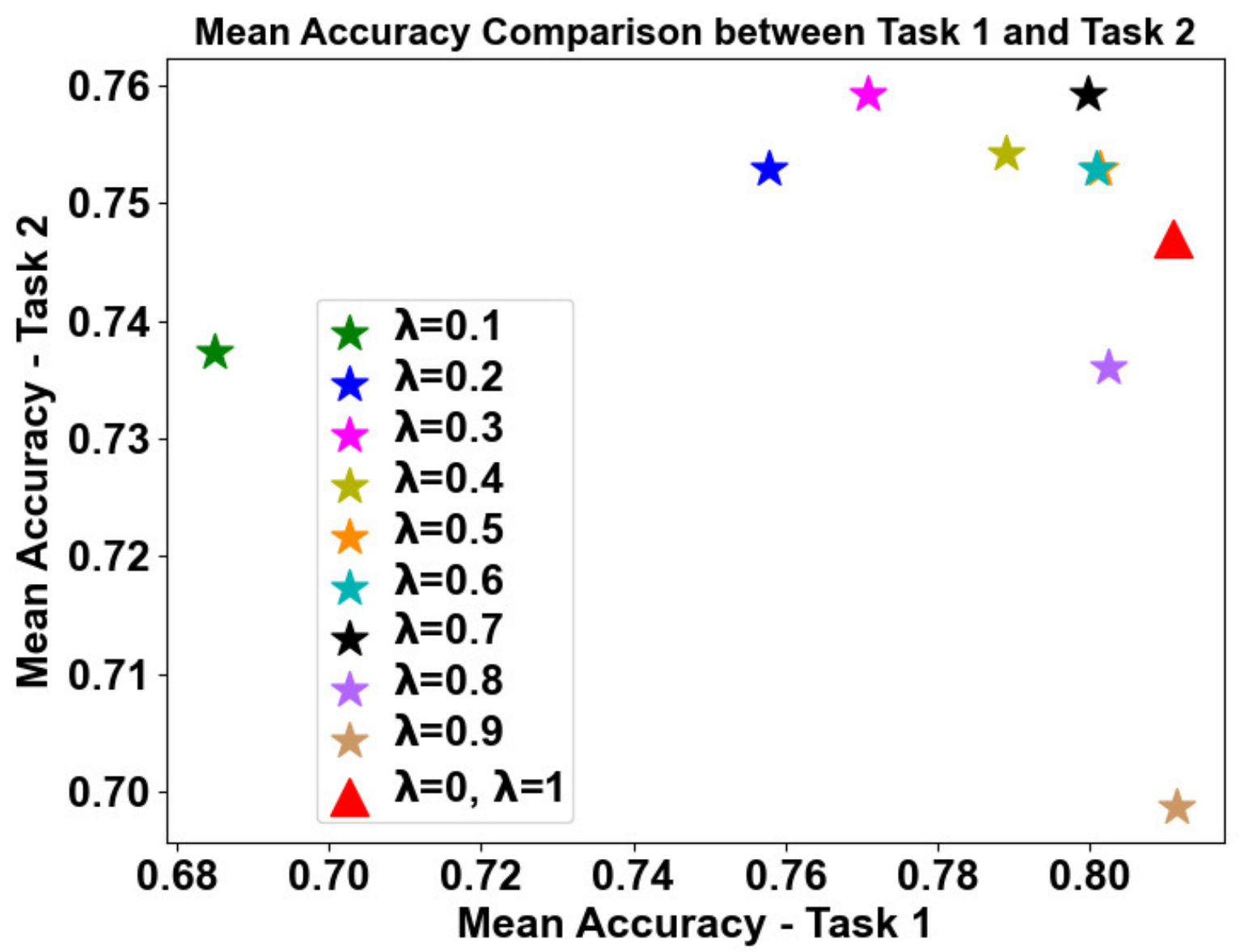
Dataset 2	λ=0.3		λ=0.4		λ=0.5		λ=0.7		λ=1	λ=0
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Accuracy	77.0 (2.7)	75.9 (1.4)	78.9 (0.8)	75.4 (1.5)	80.1 (2.3)	75.3 (1.7)	79.9 (2.1)	75.9 (0.4)	81.1 (1.4)	74.7 (1.2)
F1-score	76.5 (2.9)	81.6 (1.1)	77.3 (1.3)	81.1 (1.0)	79.4 (2.5)	81.3 (2.0)	79.0 (3.5)	82.0 (0.5)	80.9 (1.7)	80.2 (1.1)
AUC	84.6 (3.1)	81.8 (1.7)	86.9 (1.0)	82.7 (1.2)	88.0 (1.7)	82.4 (0.3)	88.3 (0.8)	82.0 (1.7)	88.2 (0.6)	80.7 (1.3)

- Average classification ***accuracy for various λ*** values.

Dataset 1



Dataset 2



7. Conclusions

- Introduction of a novel ***multitask classifier*** leveraging ***k-mers*** representation and ***Transformer*** networks.
- Demonstrating competitive ***performance comparable to single-task*** classification.
- Conducted experimental ***evaluation on real protein data*** demonstrating ***50% reduced training times*** with minimal sacrifice in individual task performance.
- A promising solution for ***resource-constrained*** environments.

8. Acknowledgements

This work was funded by the *TITAN ERA Chair* project (contract no. 101086741 within the *Horizon Europe Framework Program* of the European Commission and the *Hellenic Foundation for Research and Innovation* (Ph.D. Fellowship grant no 5647).