A Novel Fuzzy Approach towards *in silico* B-cell Epitope Identification Inducing Antigen-Specific Immune Response for Vaccine Design

Aviral Chharia [1, 2] and Apurva Narayan [3]

[1]Mechanical Engineering Department, Thapar Institute of Engineering & Technology, India
[2]Computer Science & Engineering Department, Thapar Institute of Engineering & Technology, India
[3]Computer Science & Engineering Department, The University of British Columbia, Canada

This work is selected as Oral Presentation at the "21st IEEE International Conference on BioInformatics & Bioengineering, Kragujevac, Serbia"

IEEE BIBE

Mitacs Globalink

IDSL

UBC THE UNIVERSITY OF BRITISH COLUMBIA

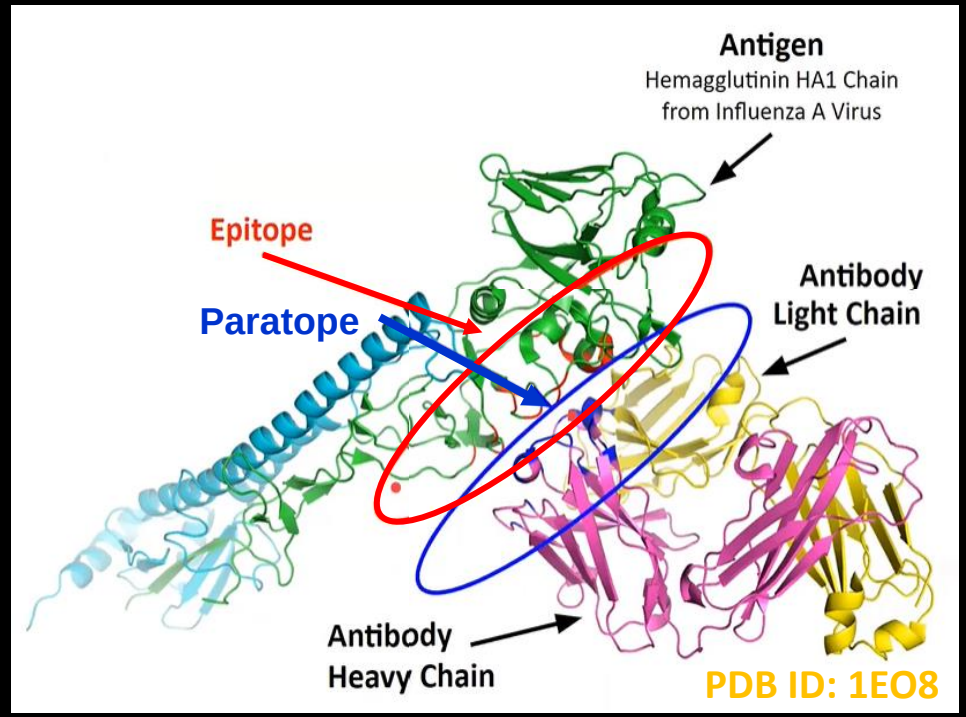
THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY (Deemed to be University)

Introduction

Importance

B-cell epitopes are present on surface of pathogen which is recognized by the human antibody. But, it is found that only certain regions on the antigen surface induce immune response rather than whole antigen.

This makes determination of these specific regions (i.e., ‘B-cell epitopes’) that elicit antigen-specific immune response crucial for immuno-detection & immuno-therapeutic applications, including the development of safe & high efficacy vaccines.



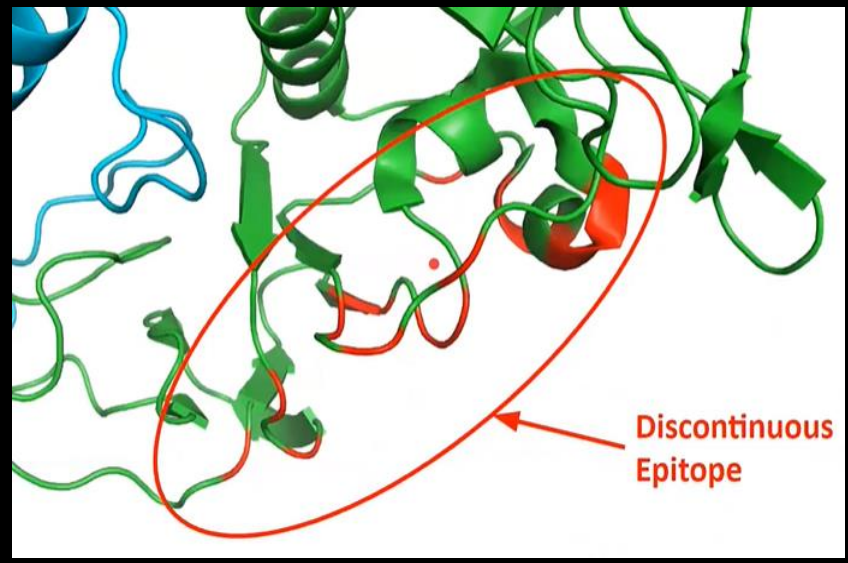
Ahmad, A. et al., “B-cell epitope mapping for the design of vaccines and effective diagnostics.” (2016).
Caoili, S.E.C. et al., “Benchmarking B-cell epitope prediction for the design of peptide-based vaccines: problems and prospects.” (2010).
Barlow, D. J. et al., “Continuous and discontinuous protein antigenic determinants.” (1986).

Introduction

Current Challenges

Identifying diagnostically useful epitopes is a difficult, time-consuming, & resource-intensive procedure. Thus, *in silico* prediction has gained immense attention recently due to its low cost, fast results, & less labor-intensive method compared to NMR spectroscopy & 3D X-ray structural analysis of antibody-antigen complexes.

However, one of the major problems that most established models confront is gathering huge volumes of data. Moreover, most models do not achieve high levels of accuracy. The underlying high complexity & noisy nature of the data further makes the task challenging.



More than 90% of B-cell epitopes in Protein antigens are estimated to be discontinuous/ conformational ***(Barlow et al., Nature)*** making it challenging to predict them.

Ahmad, A. et al., "B-cell epitope mapping for the design of vaccines and effective diagnostics." (2016).
Caoili, S.E.C. et al., "Benchmarking B-cell epitope prediction for the design of peptide-based vaccines: problems and prospects." (2010).
Barlow, D. J. et al., "Continuous and discontinuous protein antigenic determinants." (1986).

IEEE

Mitacs Globalink

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Previous methods and limitations

- Structure based prediction using antigen structure, propensity scales, geometric attributes

(-) Gets very complicated for 3D structure

- Mimotope-based method that combines mimotope sequences from phage display experiments with 3D antigen structure. Locates the best alignment sequences and predicts possible epitopic regions by mapping mimotopes back to parent antigen surface.

(-) Not found effective in actual practice to a great extent.

- Sequence-based (feature matrix formation by scoring amino acids as input antigen chain (Sweredoski et al., Ansari et al., Jespersen et al.)

(-) These attained low AUC scores & required large quantities of data to attain low positive false rate.

- Various ML & deep learning models have been developed including attention-based LSTM networks, deep ensemble learning, etc. (Kavitha et al., Noumi et al., Sun et al.)

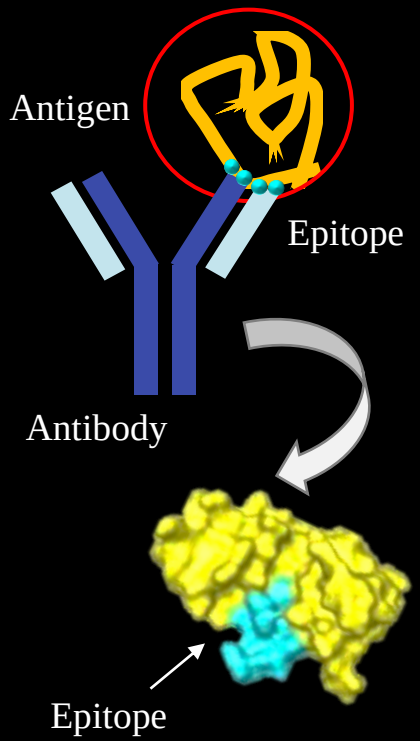
(-) Training data being highly limited and skewed, most deep learning models may be prone to overfitting.

Sweredoski, M. J. et al.., “COBEpro: a novel system for predicting continuous B-cell epitopes.” (2009).
Ansari, H. R. et al., “Identification of conformational B-cell Epitopes in an antigen from its primary sequence.” (2010).
Jespersen, M. C. et al., “BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes.” (2017).
Kavitha, K. V., “Computational prediction of continuous B-cell epitopes using random forest classifier.” (2013).
Noumi, T. et al., “Epitope prediction of antigen protein using attention-based LSTM network.” (2020).
Sun, P. et al., “B-cell Epitope prediction method based on deep ensemble architecture and sequences.” (2019).

Proposed Methodology

The current work is the *first* to propose a 'Fuzzy' approach to *in silico* B-cell epitope prediction. The effectiveness of the proposed approach is demonstrated on severely imbalanced and limited datasets through several experiments. The results show that using the proposed method enhances both accuracy and precision when compared to existing approaches.

Further, the model is tested on the SARS-CoV-1 antigen-antibody PDB complex. The proposed approach outperforms state-of-the-art ML models trained on the same data. Results obtained indicate that applying the proposed method improves the prediction compared to the other approaches.

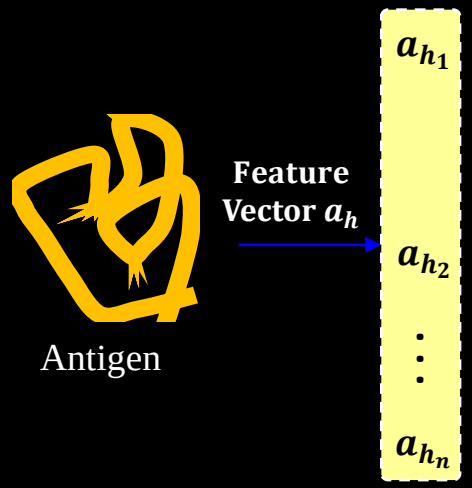


Model Architecture

Antigen

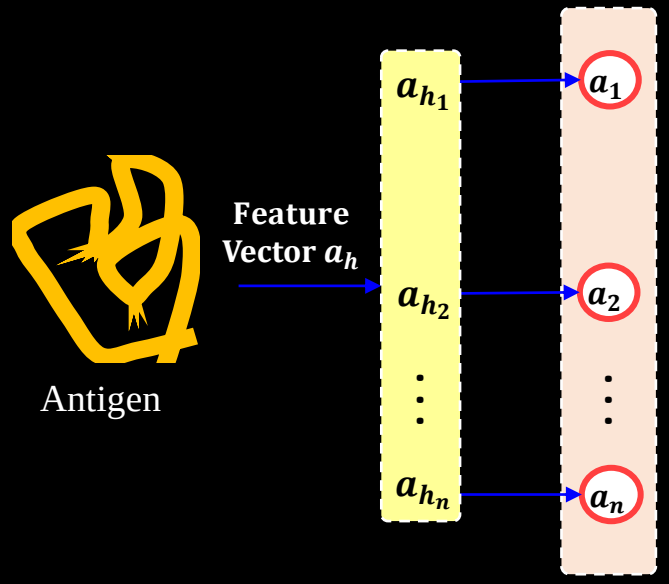
Simpson, P.K., “Fuzzy min-max neural networks. I. Classification.” (1992).
Alpern, B. et al., “The hyperbox.” (2002).
Gabrys, B. et al., “General fuzzy min-max neural network for clustering and classification.” (2000).
Nandedkar, A.V. et al. “A General Reflex Fuzzy Min-Max Neural Network.” (2007).

Model Architecture



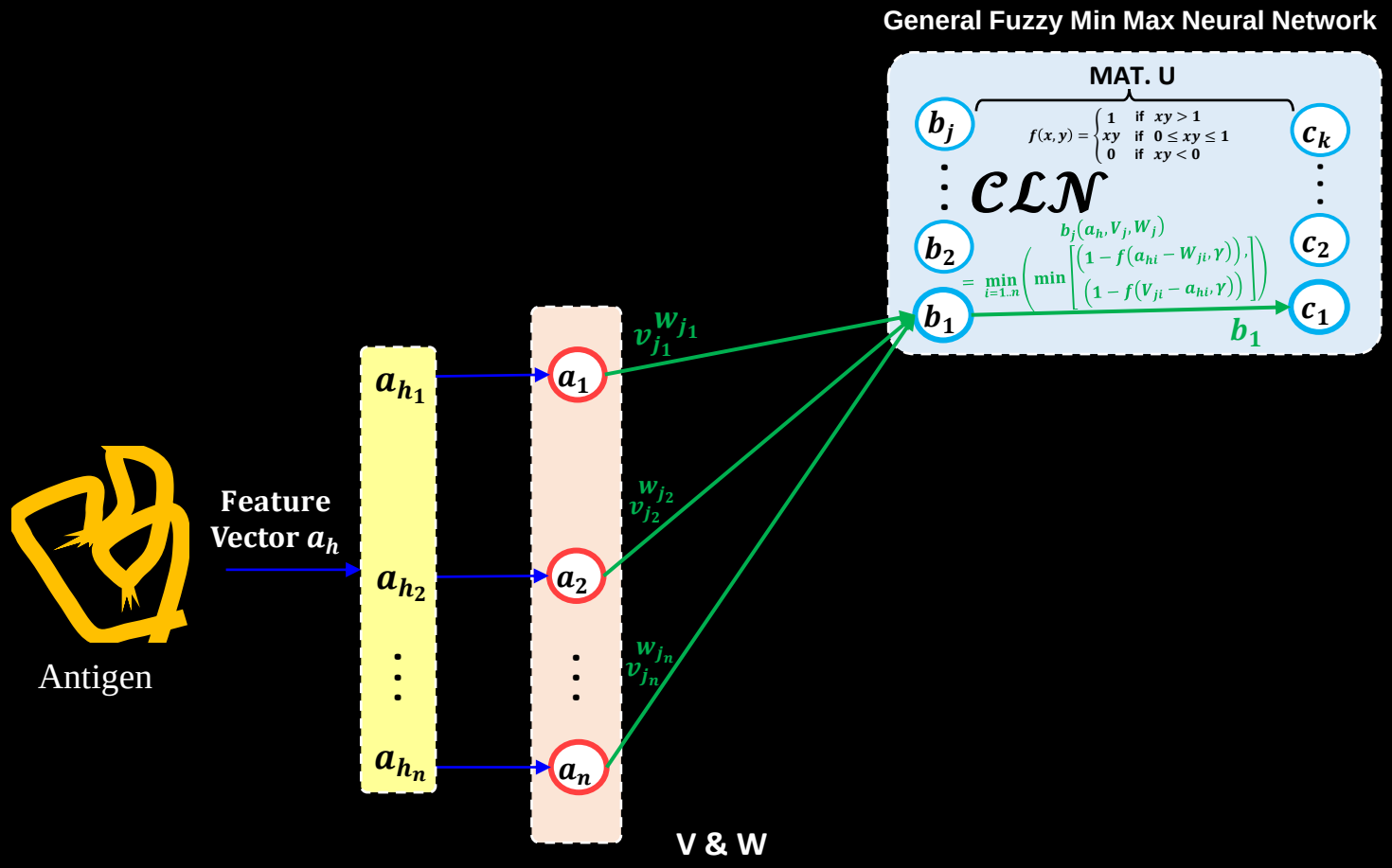
Simpson, P.K., “Fuzzy min-max neural networks. I. Classification.” (1992).
Alpern, B. et al., “The hyperbox.” (2002).
Gabrys, B. et al., “General fuzzy min-max neural network for clustering and classification.” (2000).
Nandedkar, A.V. et al. “A General Reflex Fuzzy Min-Max Neural Network.” (2007).

Model Architecture



Simpson, P.K., “Fuzzy min-max neural networks. I. Classification.” (1992).
Alpern, B. et al., “The hyperbox.” (2002).
Gabrys, B. et al., “General fuzzy min-max neural network for clustering and classification.” (2000).
Nandedkar, A.V. et al. “A General Reflex Fuzzy Min-Max Neural Network.” (2007).

Model Architecture

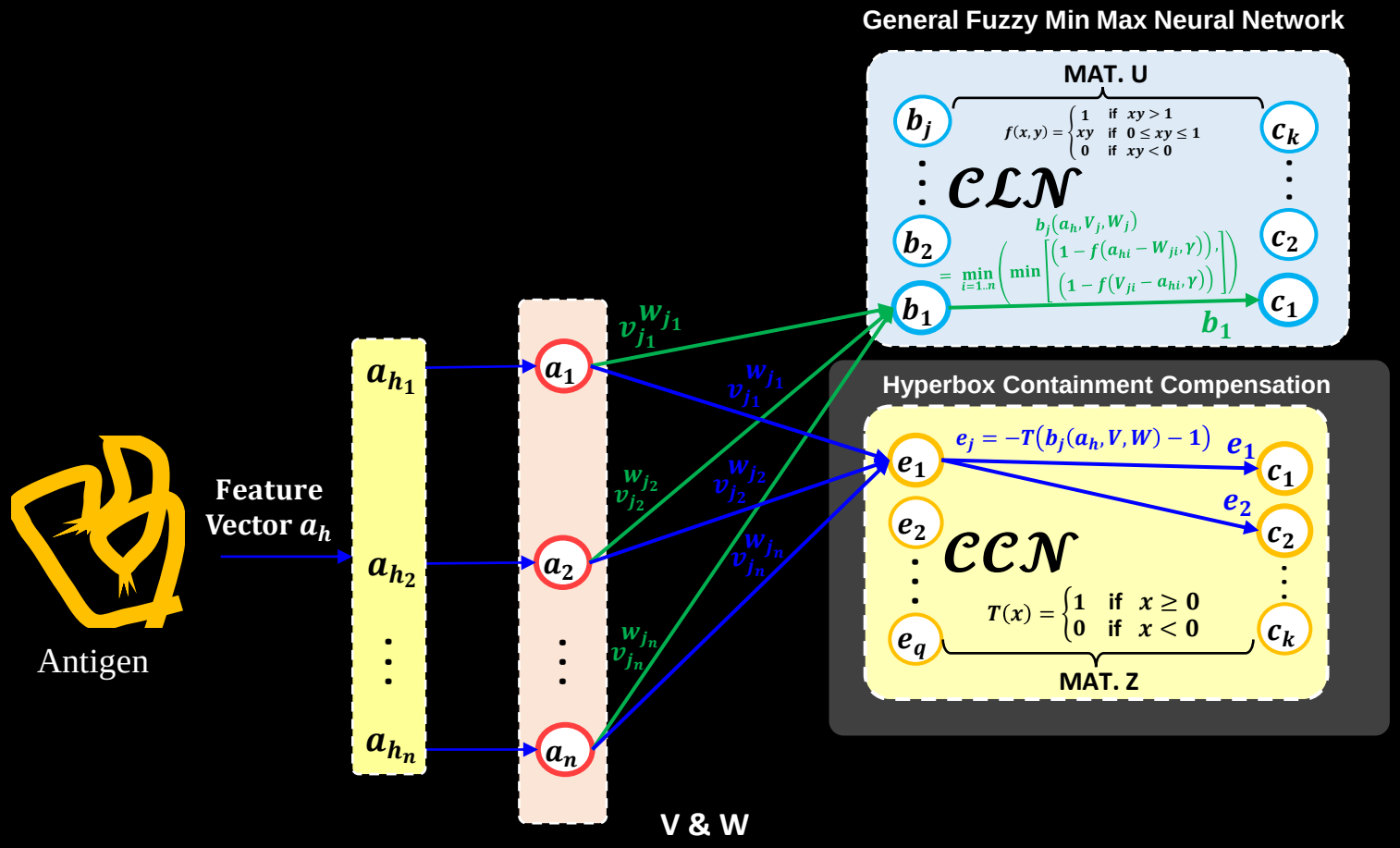


Classifying Layer Neuron (CLN)

$$b_j(a_h, V_j, W_j) = \min_{i=1..n}\left(\min\left[\begin{matrix}(1 - f(a_{hi} - W_{ji}, \gamma)), \\ (1 - f(V_{ji} - a_{hi}, \gamma))\end{matrix}\right]\right)$$

Simpson, P.K., “Fuzzy min-max neural networks. I. Classification.” (1992).
Alpern, B. et al., “The hyperbox.” (2002).
Gabrys, B. et al., “General fuzzy min-max neural network for clustering and classification.” (2000).
Nandedkar, A.V. et al. “A General Reflex Fuzzy Min-Max Neural Network.” (2007).

Model Architecture



Classifying Layer Neuron (CLN)

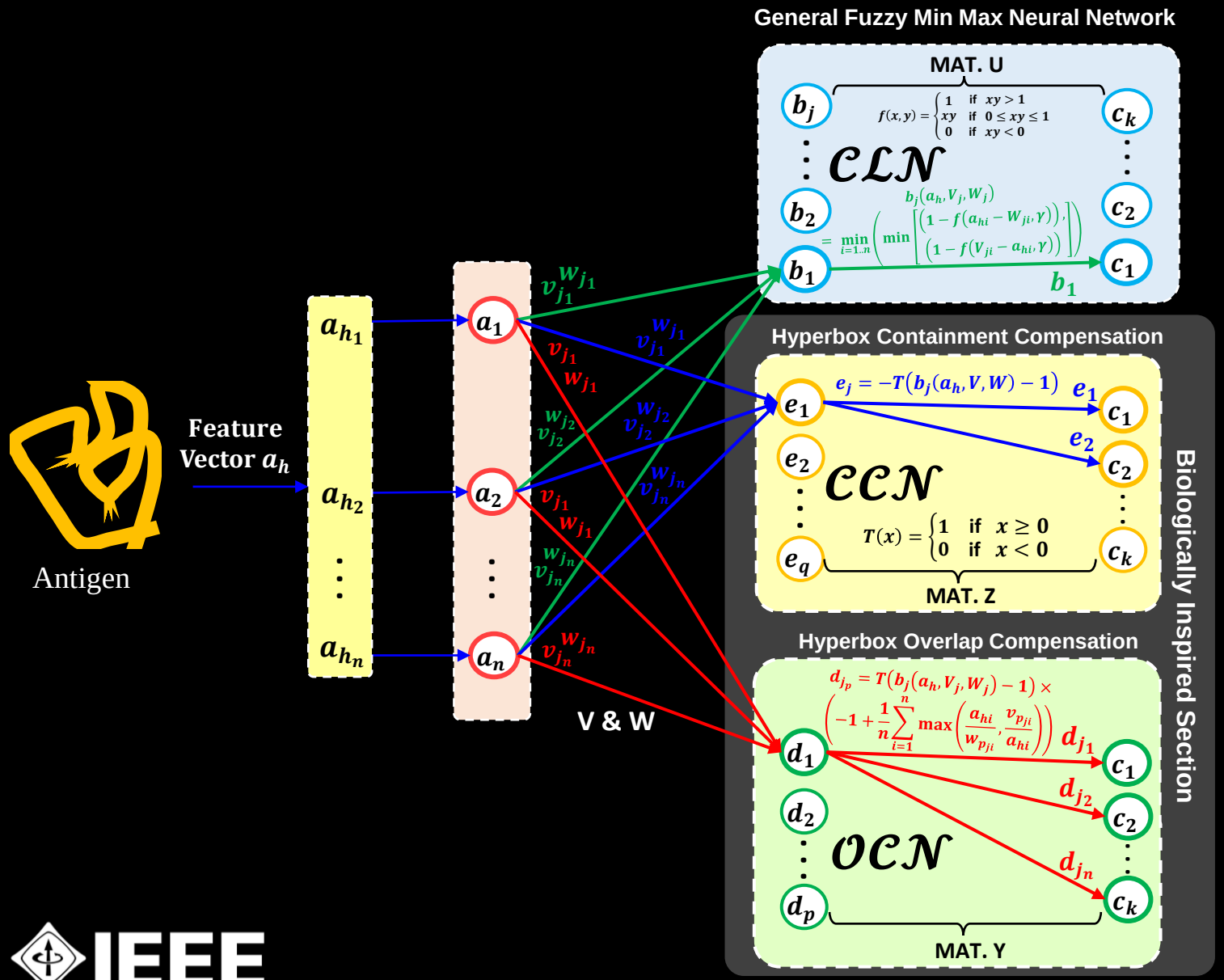
$$b_j(a_h, V_j, W_j) = \min_{i=1..n}\left(\min\left[\begin{matrix}(1 - f(a_{hi} - W_{ji}, \gamma)), \\ (1 - f(V_{ji} - a_{hi}, \gamma))\end{matrix}\right]\right)$$

Containment Compensation Neuron (CCN)

$$e_j = -1 \times T(b_j(a_h, V, W) - 1)$$

Simpson, P.K., “Fuzzy min-max neural networks. I. Classification.” (1992).
Alpern, B. et al., “The hyperbox.” (2002).
Gabrys, B. et al., “General fuzzy min-max neural network for clustering and classification.” (2000).
Nandedkar, A.V. et al. “A General Reflex Fuzzy Min-Max Neural Network.” (2007).

Model Architecture



Classifying Layer Neuron (CLN)

$$b_j(a_h, V_j, W_j) = \min_{i=1..n}\left(\min\left[\begin{matrix}\left(1 - f(a_{hi} - W_{ji}, \gamma)\right), \\ \left(1 - f(V_{ji} - a_{hi}, \gamma)\right)\end{matrix}\right]\right)$$

Containment Compensation Neuron (CCN)

$$e_j = -1 \times T\left(b_j(a_h, V, W) - 1\right)$$

Overlap Compensation Neuron (OCN)

$$d_{j_p} = T\left(b_j(a_h, V_j, W_j) - 1\right) \times \left(-1 + \frac{1}{n}\sum_{i=1}^{n} \max\left(\frac{a_{hi}}{w_{p_{ji}}}, \frac{v_{p_{ji}}}{a_{hi}}\right)\right)$$

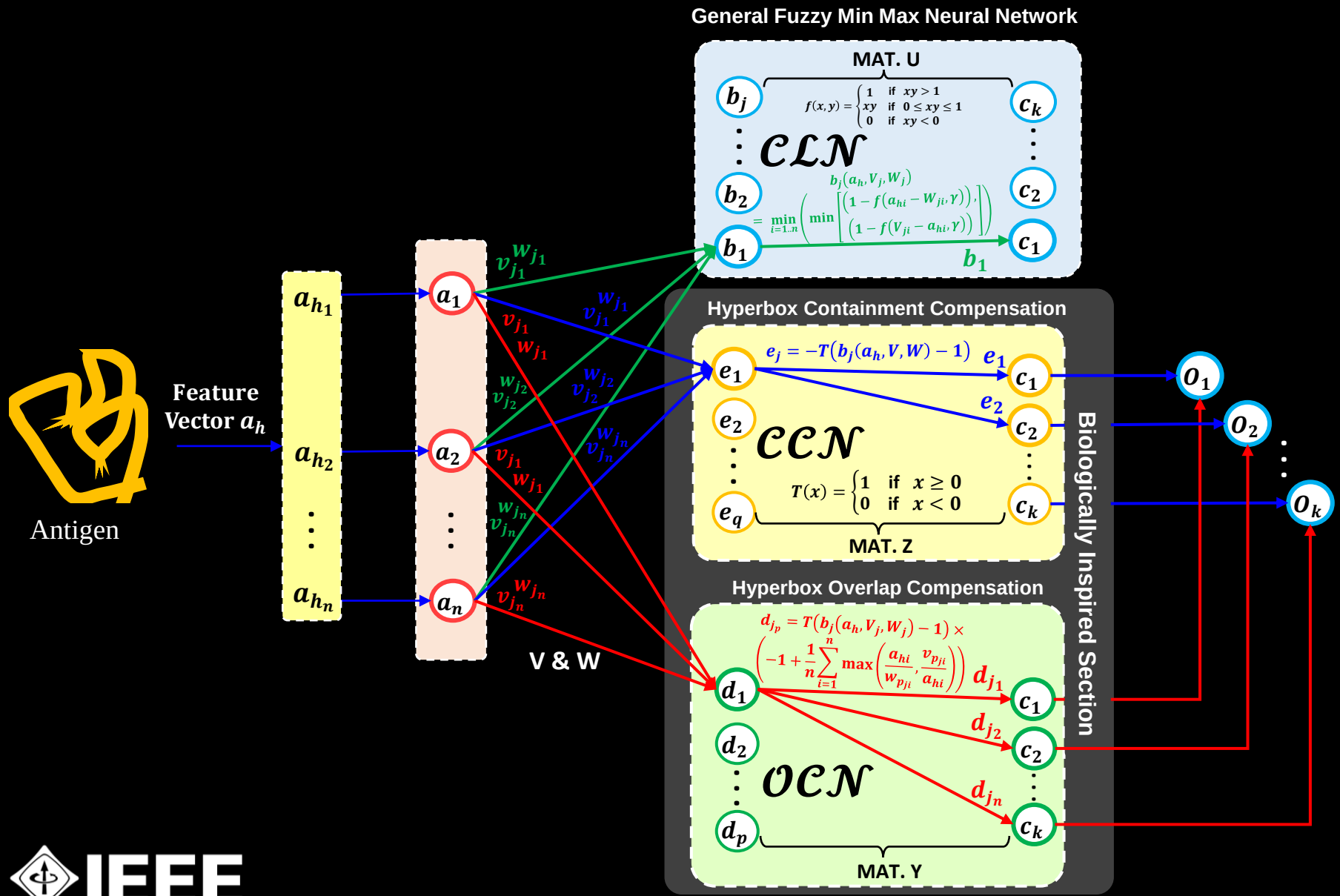
here, $f(x,y) = \begin{cases} 1 & \text{if } xy > 1 \\ xy & \text{if } 0 \le xy \le 1 \\ 0 & \text{if } xy < 0 \end{cases}$ $\quad T(x) = \begin{cases} 1 & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases}$

IEEE

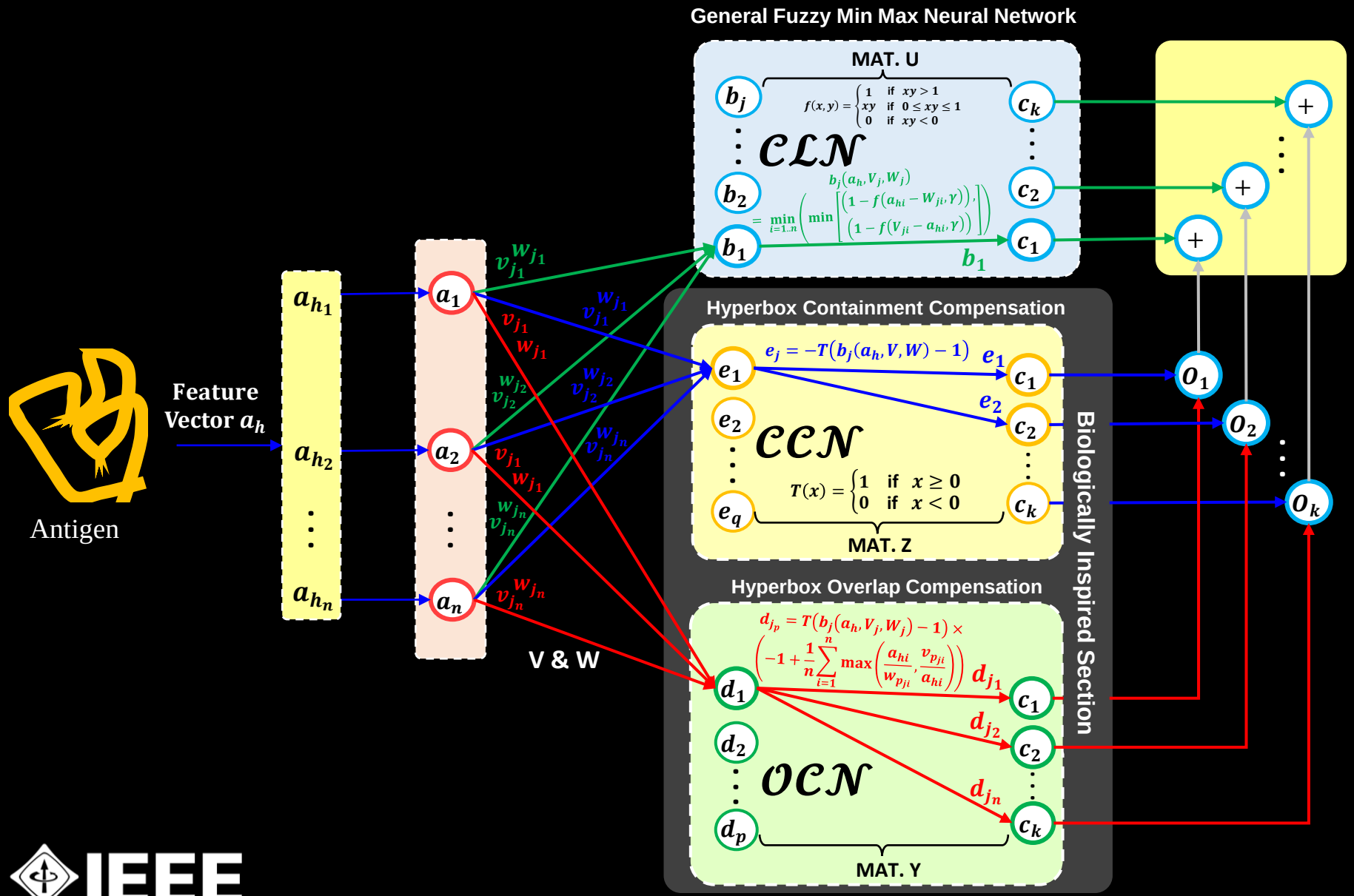
Mitacs Globalink

UBC THE UNIVERSITY OF BRITISH COLUMBIA

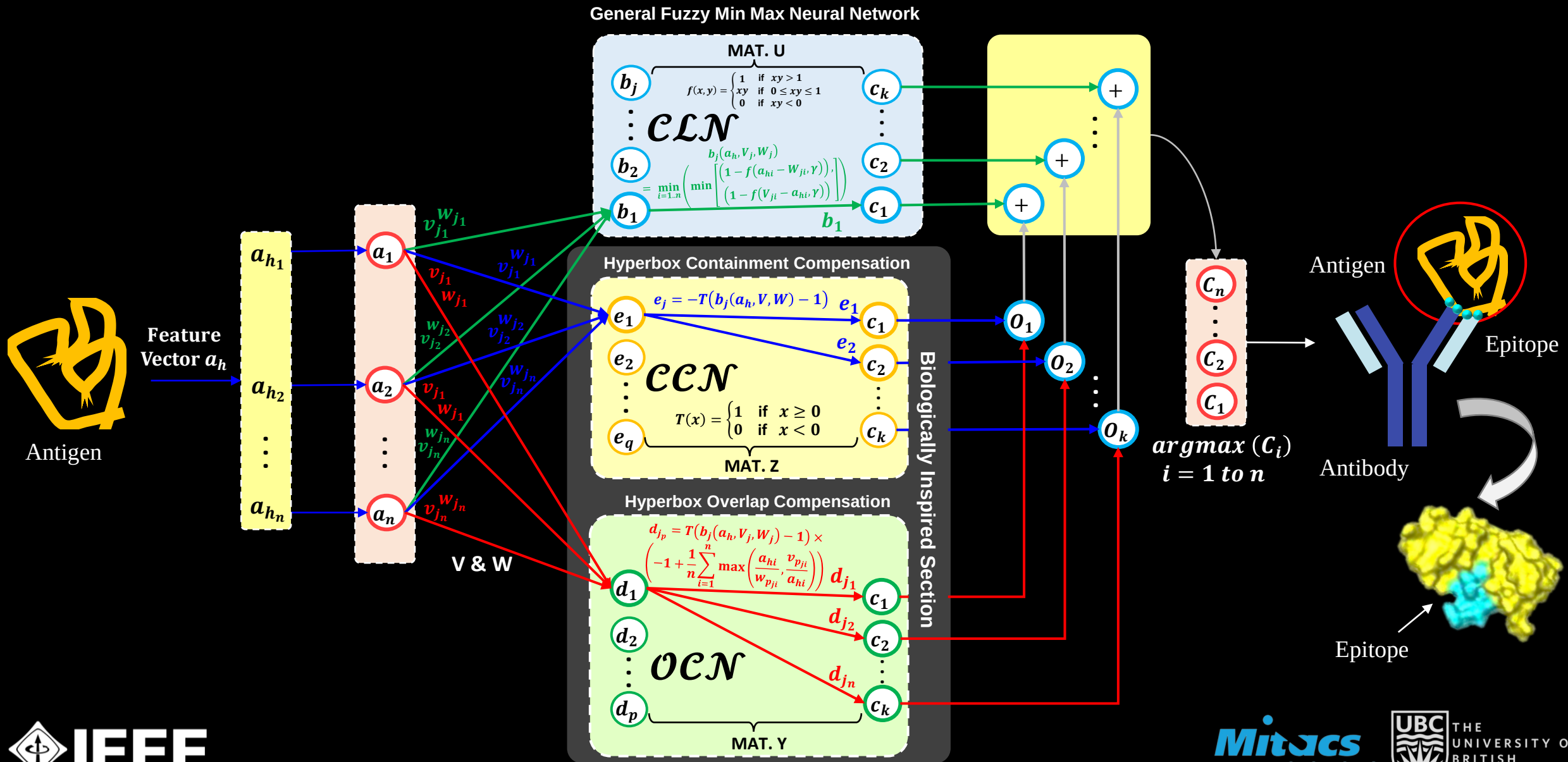
Model Architecture



Model Architecture



Model Architecture



Experiments

Dataset

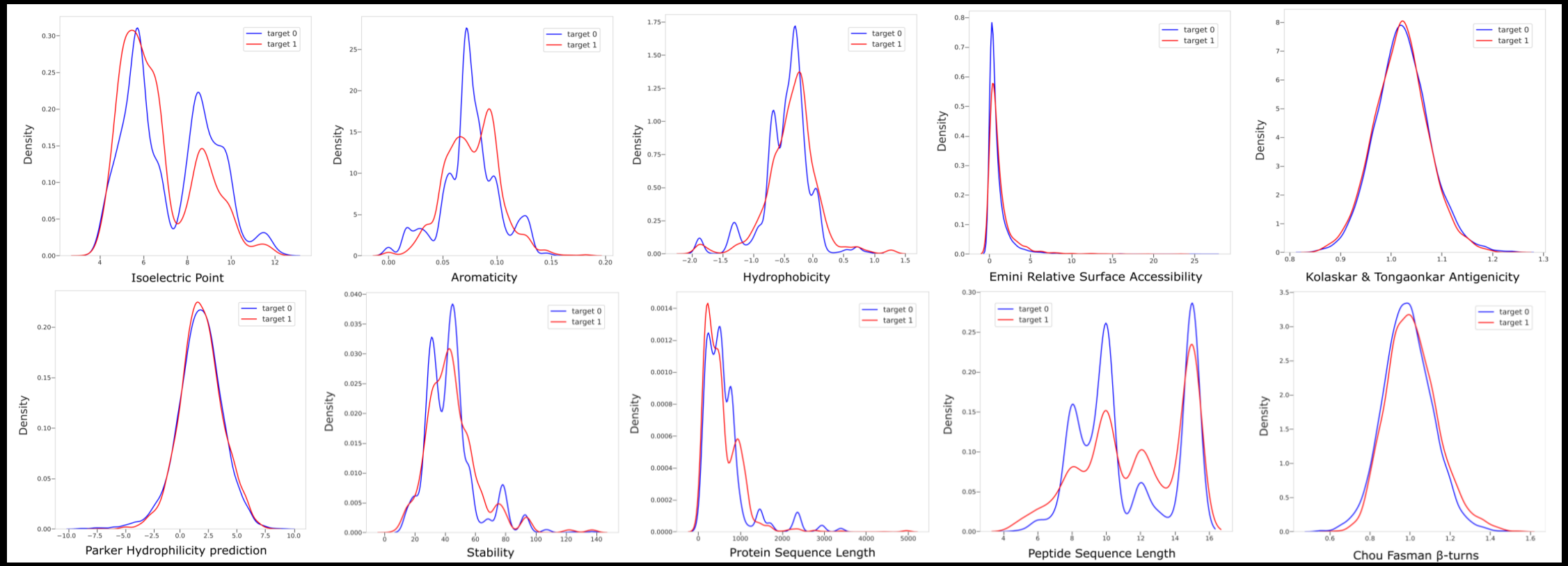
- Epitope candidate amino acid sequence (IgG) and the activity label data from IEDB and UniProt was used. To minimize high-class imbalance, the epitope data was converted to binary classification. For this study, ‘Positive-High,’ ‘Positive-Intermediate,’ and ‘Positive Low’ samples were all regarded as ‘Positive’ samples. Nonetheless, high-class imbalance makes it harder to get positive samples compared to negative ones, making DL models inefficient.

Performance Evaluation Metrics

- In datasets skewed towards a particular class, accuracy alone may be misleading while accessing model performance. Thus, Precision & MCC along with $Acc.$ is considered a better evaluation metric.
- Same dataset, feature selection and pre-processing techniques are used while implementing baselines.

R. Vita et al., “The Immune Epitope Database (IEDB): 2018 update.” (2019).
UniProt Consortium, “UniProt: the universal protein knowledgebase in2021” (2021).

Feature-wise KDE-plots



Feature-wise KDE-plots for epitope targets **(in Red)** and non-epitope **(in Blue)** antigen regions. A weak negative co-relation can be seen.

Results and Comparison

- Proposed model outperforms ML models by a significant margin & can predict B-cell epitopes with comparable accuracy & higher precision, while being less data intensive.
- As number of training samples varies, the second-best performing model loses consistency, i.e., for n = 200, the Random Forest is the second best-performed model, its performance decreases for n = 300, where Ridge classifier & GBC classifiers are seen performing better. This illustrates that no single model exhibits robust and consistent performance while training on limited data.

Comparison of B-Cell Epitope Prediction Results with ML Classifiers for $n = 200, Imb = 0.75$. Here, learning parameters $\theta = 0.25, \gamma = 2$ for proposed method

Classification Model	Data Subset Configuration - I					
	Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)	MCC (↑)	TT (*sec*) (↓)
Ada Boost Classifier	66.25	26.50	21.50	23.33	0.0376	0.364
Support Vector Machine	66.88	38.42	37.00	35.31	0.1728	0.013
Logistic Regression	68.12	29.33	16.00	19.13	0.0587	0.022
Decision Tree	68.12	41.90	35.00	34.23	0.1710	**0.012**
Ridge Classifier	68.12	15.00	14.00	14.44	0.0226	0.013
Gradient Boosting Classifier	69.38	46.67	29.00	35.44	0.1776	0.083
K-Neighbors Classifier	70.62	45.83	26.50	32.90	0.1842	0.063
Random Forest	72.50	46.67	21.00	27.19	0.1966	0.412
Proposed Method	**72.50**	**90.00**	**47.37**	**62.07**	**0.4914**	0.614

Comparison of B-Cell Epitope Prediction Results with ML Classifiers for $n = 300, Imb = 0.75$. Here, learning parameters $\theta = 0.20, \gamma = 2$ for proposed method

Classification Model	Data Subset Configuration - II					
	Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)	MCC (↑)	TT (*sec*) (↓)
Support Vector Machine	65.42	36.87	30.48	31.68	0.0993	0.015
K-Neighbors Classifier	69.58	34.50	12.62	17.58	0.0619	0.062
Decision Tree	71.67	48.09	**46.19**	**45.98**	0.2780	0.013
Ada Boost Classifier	73.75	53.20	41.19	45.56	0.2962	0.092
Random Forest	74.17	50.17	22.38	30.12	0.2031	0.411
Logistic Regression	74.17	49.83	22.38	29.78	0.2089	0.023
Gradient Boosting Classifier	74.58	55.17	33.57	40.79	0.2777	0.098
Ridge Classifier	**75.42**	57.50	28.57	36.87	0.2792	**0.013**
Proposed Method	75.00	**80.00**	22.22	34.78	**0.3289**	1.814

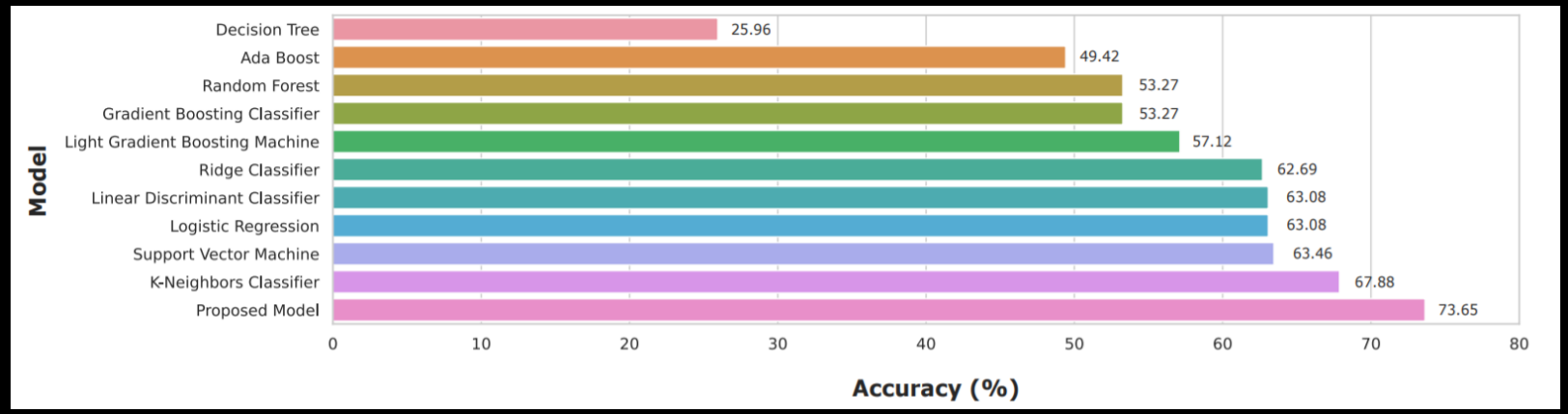
IEEE

Mitacs Globalink

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Results and Comparison

- In contrast to the previous experiment, where the model performance was assessed on a small & severely unbalanced data, here, we analyze the model on SARS-CoV-1 antigen sequence. It should be noted that while the model was trained using IEDB and Uniprot data, the antigen sequence was previously unknown to the model. This translates to real-time circumstances involving the prediction of B-cell epitopes of new antigen sequences. The Proposed model outperforms ML models by a significant margin (≈ 5.77% on *Acc.)*

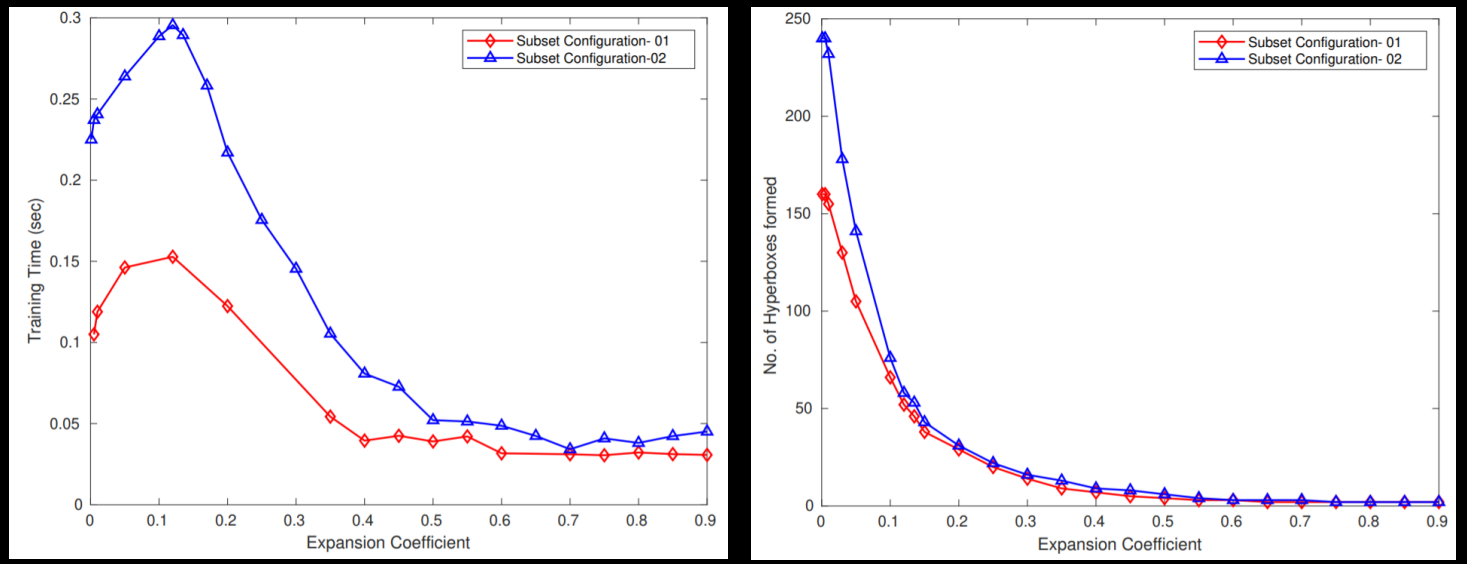


Comparison bar-plot between various models and the obtained classification results on the SARS-CoV-1 antigen sequence

R. Vita et al., “The Immune Epitope Database (IEDB): 2018 update.” (2019).
UniProt Consortium, “UniProt: the universal protein knowledgebase in2021” (2021).

Parametric Analysis

- The model's overall training time is extremely short (i.e.,1 to 5 sec), the sample test time is rather long, ranging from 5 to10 sec. Such a substantial gap is not found in low-dimensional data categorization tasks.
- As hyperbox expansion coefficient (θ) increases, number of hyperboxes created during training increases in an 'exponential' manner rather than a 'linear' one.
- The model training time climbs steeply until 0.2, after which it reduces 'exponentially' on both configurations.



Parametric study results (a) No. of hyperboxes formed vs. expansion coefficient (θ) (b) Training time (sec) vs. expansion coefficient (θ)

Conclusion

Summary

- Prediction of B-cell epitopes with high accuracy prior to lab-oratory tests can greatly reduce experimental costs while also accelerating the identification process.
- Proposed a *fuzzy* approach towards B-cell epitope identification.
- Addressed the problem of limited availability of datasets and high-class imbalance seen in them.

Key findings

- Fuzzy classifier-based models are more suited towards problems with limited/ highly imbalanced data.
- Prediction through ‘fuzzy models’ is a promising approach towards B-cell epitope identification.

Future work

- Future work may focus on reducing model’s lengthy sample testing time.
- Developing an algorithm for tuning hyperbox expansion coefficient (θ) & fuzziness coefficient (γ).

Thanks

Feel free to contact us at
achharia_be18@thapar.edu