Migrating from PLS to Artificial Neural Networks – Adapting Interpretation Strategies

Kevin M Mendez, David I Broadhurst, Stacey N Reinke

The Centre for Integrative Metabolomics and Computational Biology

AUSTRALIA ECU EDITH COWAN UNIVERSITY

BACKGROUND

- Artificial Neural Networks (ANNs) are making a resurgence in systems biology and metabolomics. This is due to increased compute power, availability of code libraries, larger datasets, and societal acceptance.[1]
- We recently showed that ANNs have similar predictive ability to other contemporary machine learning algorithms (including PLS, Random Forest, and Support Vector Machines) for clinical metabolomics data with a binary outcome.[2]
- Interpretability of ANNs remains a key challenge for their wide spread use; however, single hidden layer ANNs have structural equivalence to PLS, in the form of projection to latent structures (Figure 1).
- **AIM:** To migrate standardised optimisation, visualisation, evaluation, and statistical inference techniques from PLS-DA to a fully connected non-linear (logistic), single hidden layer, ANN. This will provide a foundation for the implementation of more complex interpretable ANNs.

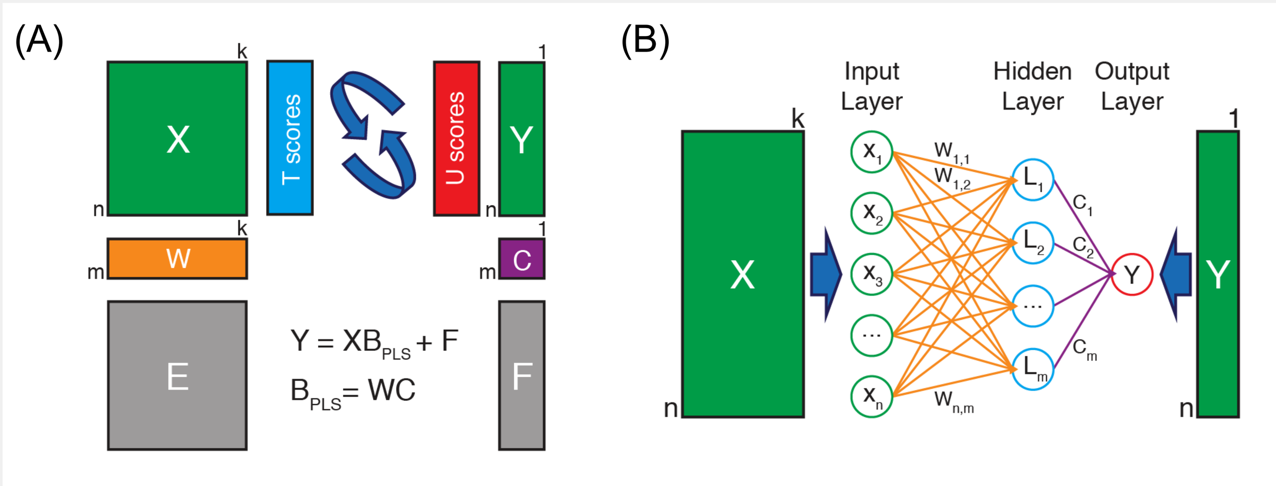


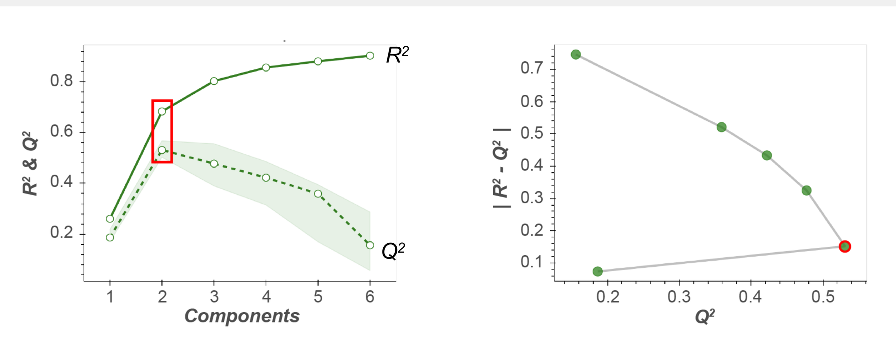
Figure 1. Structural Equivalence of ANNs to PLS. (A) Matrix representation of PLS. (B) Network representation of ANN. Adapted from [1].

1. SELECT DATASET AND CREATE NOTEBOOK

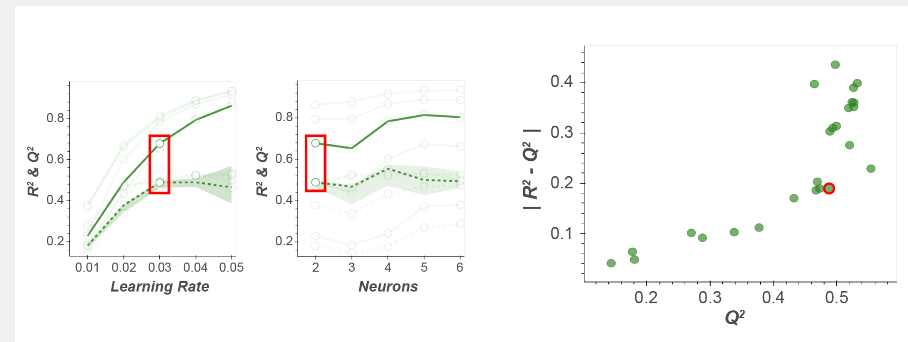
- Dataset retrieved from Metabolomics Workbench (ST0001047).
- Modelling performed using Python programming language in the Jupyter Notebook framework.

PLS-DA | Scan the QR codes to launch the interactive workflows (jupyter, binder) | **ANN**

2. OPTIMISE HYPERPARAMETERS

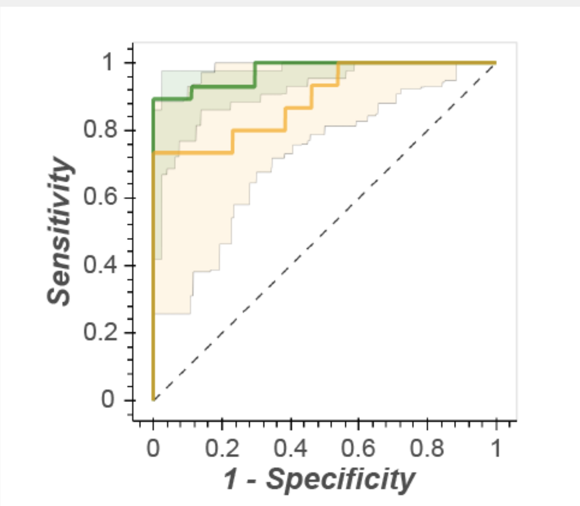


Both the standard R^2 & Q^2 plot (left) and | R^2 - Q^2 | vs Q^2 plot (right) are readily interpretable. 2 latent variables were chosen for the PLS model.



The standard R^2 & Q^2 plot approach was difficult to interpret for optimising two hyperparameters (learning rate, left; number of neurons, centre). The | R^2 - Q^2 | vs Q^2 plot (right) was readily interpretable. A learning rate of 0.03 and 2 neurons were chosen for the ANN model.

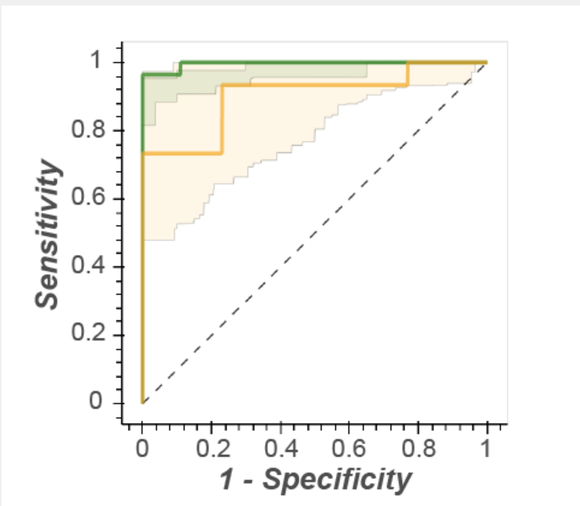
3. MODEL EVALUATION



AUC_{Train} = 0.97
AUC_{IB} = 0.92-0.99
AUC_{Test} = 0.89
AUC_{OOB} = 0.72-0.98

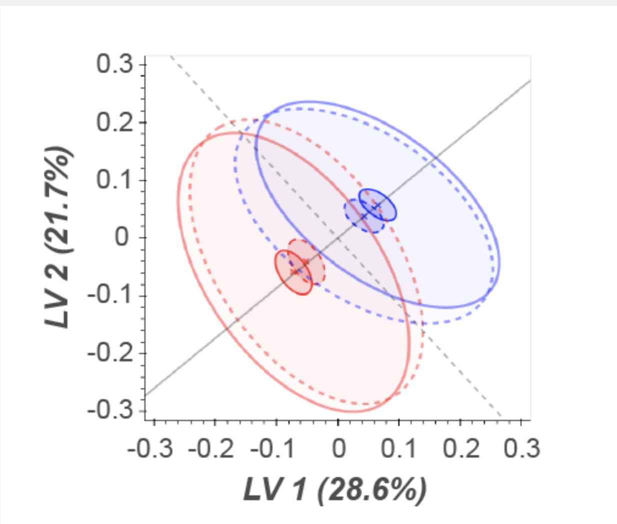
Predicted scores (train and test sets) split into the respective binary classification. ROC curves are presented with 95% confidence intervals (CIs) derived from 100 iterations of bootstrap remodelling.

Green line, predicted scores for training set; green 95% CIs, in-bag (IB) predictions; yellow line, predicted scores for test set; yellow 95% CIs, out-of-bag (OOB) predictions.



AUC_{Train} = 1.00
AUC_{IB} = 0.95-0.99
AUC_{Test} = 0.90
AUC_{OOB} = 0.77-1.00

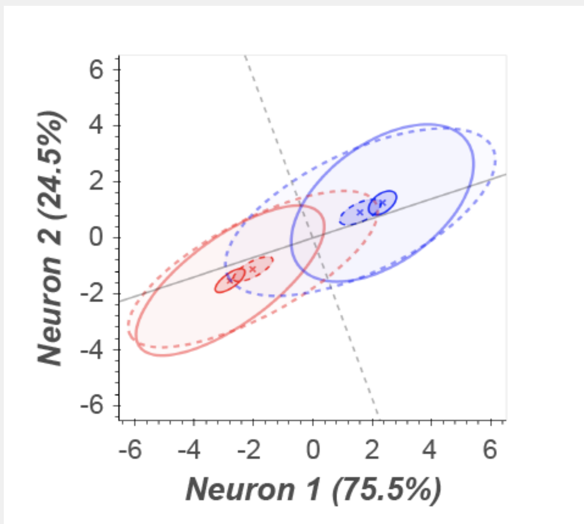
4. STATISTICAL INFERENCE



Latent Variable 2 vs Latent Variable 1

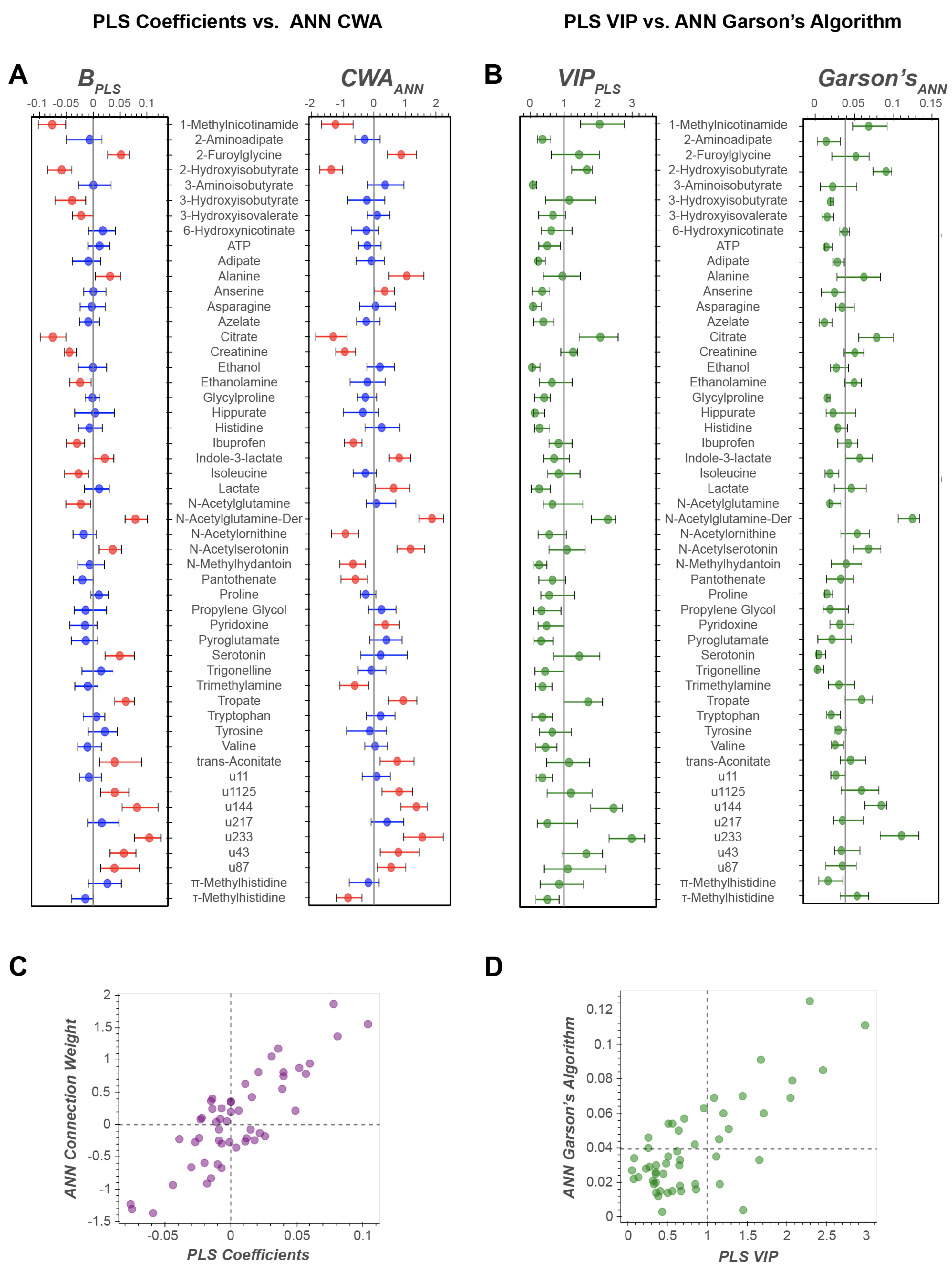
Bootstrap projection (scores) plots.

Red, control; blue, case. Inner ellipses, 95% CI of the mean; outer ellipses, 95% CI of the population. Solid lines, IB predictions; dashed lines, OOB predictions.



Neuron 2 vs Neuron 1

5. VARIABLE CONTRIBUTION



Several variable contribution metrics have been proposed for ANNs. The most comparable to PLS coefficients and Variable Influence on Projection (VIP) are Connection Weight Approach[3] and Garson's Algorithm[4], respectively.

(A) Median (and 95% CI) B_{PLS} (left) and CWA_{ANN} (right). Blue, contribution not significant based on 95% CIs; red, contribution significant based on 95% CIs. **(B)** Median (and 95% CI) VIP_{PLS} (left) and $Garson_{ANN}$ (right). **(C)** Scatter plot of CWA_{ANN} vs B_{PLS}, Pearson's r = 0.85 (p-value = 2.79 x 10^{-15}). **(D)** Scatterplot of $Garson_{ANN}$ vs. VIP_{PLS}, Pearson's r = 0.75 (p-value = 1.33 x 10^{-10}). Dashed lines at respective "importance" cut-off: $Garson_{ANN}$ = 0.038, VIP_{PLS} = 1.00.

CONCLUSIONS & FUTURE DIRECTIONS

- Migration of visualisation strategies was successful.
- | Q^2 – R^2 | vs Q^2 plot aids interpretability for choosing ANN hyperparameters.
- Using bootstrapping strategies enables clear visual interpretation and statistical inference.
- CWA and Garson metrics suitable alternatives to B_{PLS} and VIP, respectively.
- VIP and Garson cut-offs not statistically justified – recommend reporting B_{PLS} and CWA with 95% CIs.
- This work provides a foundation for ANN use, including more complicated architectures.

REFERENCES

1. Mendez *et al.* (2019) *Metabolomics*. 15(11): 142
2. Mendez *et al.* (2019) *Metabolomics*. 15(12): 150.
3. Olden & Jackson (2002) *Ecological Modelling*. 154: 135-150.
4. Garson (1991) *AI Expert*. 6: 47-51.

FUNDING

This work was partly funded by the Australian Research Council (LE170100021).

Stacey N Reinke | Senior Research Fellow | Centre for Integrative Metabolomics & Computational Biology (www.cimcb.info)
School of Science | Edith Cowan University | Perth, Australia | email: stacey.n.reinke@ecu.edu.au | twitter: @StaceyReinke