The Interpretation of Neural Network QSAR Models Using Weights and Biases

R. Guha, D.T. Stanton and P.C. Jurs

Department of Chemistry
Pennsylvania State University
and
Miami Valley Laboratories
Procter & Gamble

March 15th, 2005

Outline

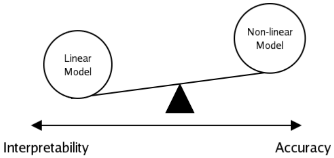
1. Background
   - Why Do We Need An Interpretation?
   - Some Aspects of Interpretability
2. Interpreting a Neural Network
3. Summary

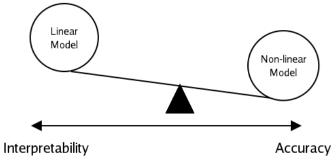
Isn't a Prediction Enough?

- Predictive models are good for screening purposes
- To understand *why* a compound is active we need an interpretation
- Interpretation is one way to approach the inverse QSAR problem
- Interpretability depends on modeling technique & descriptors involved

Interpretability & Accuracy

- Interpretability generally involves a trade off with accuracy
- Linear regression models are amenable to interpretation, but are often not very accurate
- Neural networks are black boxes, but are often more accurate
- Some techniques lie in between (random forests)





Types of Interpretation

Broad Interpretation

- Essentially describes which descriptors are important
- Good for understanding which descriptors to focus on
- Based on randomization

Detailed Interpretation

- Describes how the property (activity) relates to the descriptor
- Gives us conclusions like:
  high value of DESC leads to **low** values of activity
- Allows for a detailed understanding of the SAR in QSAR

CNN Interpretation in the Literature

- Relative importance of input neurons
- Uses the training set to develop measures of importance
- In many cases the methods depend on the nature of the network

Guha, R. et al., *J. Chem. Inf. Model.*, **2005**, *in press*

Tickle, A.B. et al., *Intl. Conf. on Neural Networks*, **1997**, *4*, 2530-2534

Yao, S. et al., *Proc. Fifth IEEE Intl. Conf. on Fuzzy Systems*, **1996**, *1*, 361-367

Outline

Goals

Analogy with PLS Interpretations

The method is analogous to the PLS approach for linear models which considers the linear combination coefficients for each latent variable as indicating the *effect* of a descriptor on the output

Utilizing CNN weights and biases . . .

- Correlate input descriptors to network output through each hidden neuron
- Order the hidden neurons
- Consider hidden neurons as *latent variables*

Goals

Analogy with PLS Interpretations

The method is analogous to the PLS approach for linear models which considers the linear combination coefficients for each latent variable as indicating the *effect* of a descriptor on the output

Utilizing CNN weights and biases ...

- Correlate input descriptors to network output through each hidden neuron
- Order the hidden neurons
- Consider hidden neurons as *latent variables*

Some Preliminaries

We know ...

- The transfer function is sigmoidal

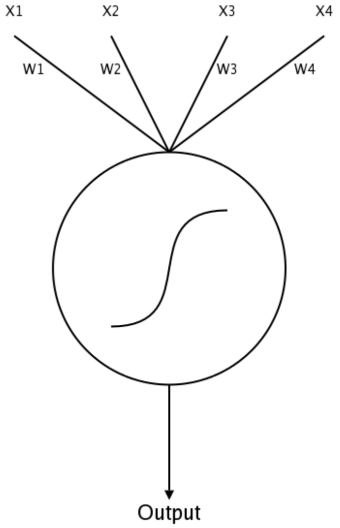
$$O = \frac{1}{1 + \exp(-\sum w_i x_i)}$$

- We can approximate this as

$$O \sim \exp\left(w_1 x_1 + \cdots + w_n x_n\right)$$

This indicates ...

- O is an increasing function of its inputs
- Output from a hidden neuron is always positive



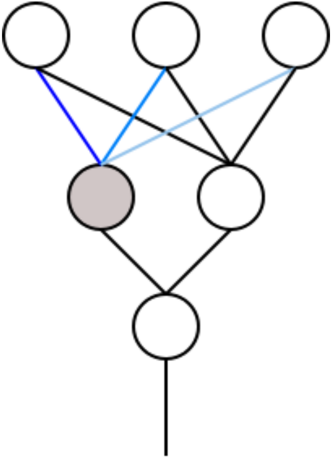
What Do The Weights Tell Us?

The absolute values tell us . . .

The weights, w_i, determine which input neuron dominates the contribution to a hidden neuron

The signs tell us . . .

The nature of the correlation between an input to a neuron and the output from the neuron



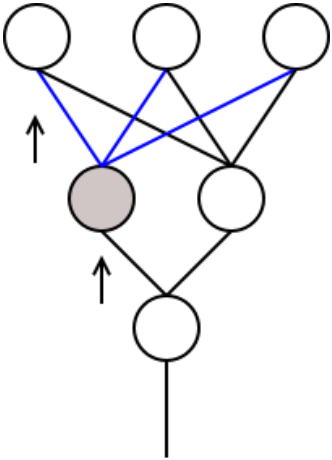
What Do The Weights Tell Us?

The absolute value tells us . . .

The weights, w_i, determine which input neuron dominates the contribution to a hidden neuron

The signs tell us . . .

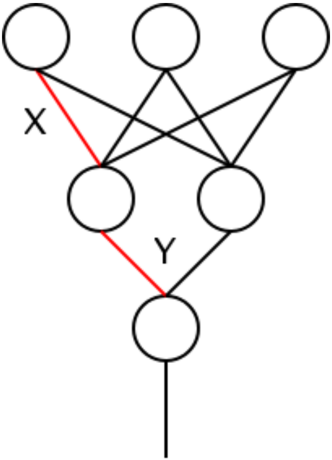
The nature of the correlation between an input to a neuron and the output from the neuron



Effective Weights

What are they?

- As input flows from an input neuron to the output neuron it is acted on by two weights
- The effective weight for an input neuron is thus XY



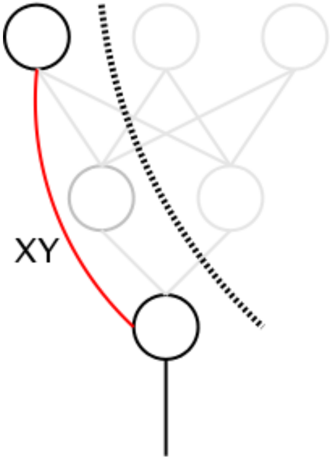
Effective Weights

What are they?

The result is that the network looks like a single connection between the input neuron and the output neuron with a weight XY

Effective Weight Matrix

	Hidden Neuron	
Descriptor	1	2
Desc 1	52.41	29.30
Desc 2	37.65	22.14
Desc 3	-10.50	-16.85



Why Do We Ignore The Bias Term?

Equipartitioning View

- When considering effective weights via a given hidden neuron, the bias term must be partitioned.
- The simplest approach is to equipartition the bias term
- The net result is that the same value is added to each effective weight.

Constant Bias View

- CNN's exhibit the universal function approximation property
- A sufficient condition for this is that the transfer function has a non-zero derivative at the origin
- This implies that the bias can be taken as a constant rather than trainable weight

Hornik, K. *Neural Networks*, **1993**, *6*, 1069-1072

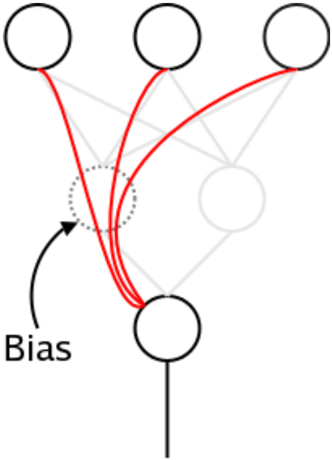
Ordering Hidden Neurons

Contribution of a hidden neuron . . .

- Depends on the output of the neuron
- Depends on the inputs to the neuron

Quantifying Contributions

- Take the column means of the effective weight matrix
- Also include bias terms for each hidden neuron
- Convert to a proportional scale for ease of use (SCV)



Validation of the Method

- Build a linear model with N descriptors and interpret it
- Build a CNN model with the same descriptors and interpret it

The two interpretations should match since both models should encode similar SPR trends

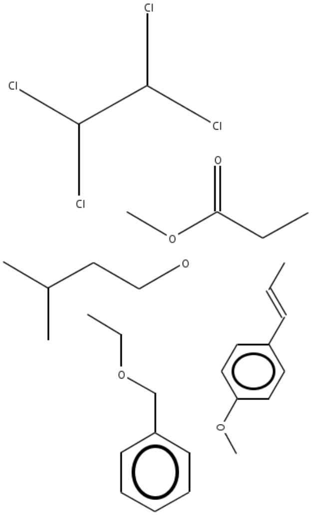
Boiling Points

Dataset

- 277 compounds
- Original work reported CNN models
- No interpretations
- $145K < BP < 653K$

Model details

- 7 descriptor OLS model
- $R^2 = 0.98$, $RMSE = 9.98$ K
- CNN model was 7-4-1
- $R^2 = 0.91$, $RMSE = 15.21$ K



Goll, E.S. et al., *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 974-983

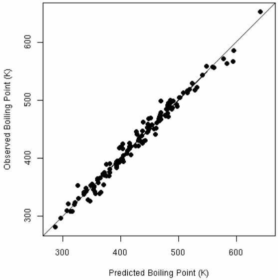
Linear Interpretation

Component 1 focuses on ...

- Size effects
- Higher molecular weight
- Longer paths

Component 2 focuses on ...

- Hydrogen bonding ability
- Charge weighted negative surface area
- Lower hydrophobic surface area



	Component	
Descriptor	1	2
PNSA-3	**-0.30**	**-0.42**
RSHM-1	0.19	**0.77**
V4P-5	**0.48**	-0.15
S4PC-12	0.28	-0.07
MW	**0.49**	-0.085
WTPT-2	**0.48**	-0.05
DPHS-1	0.26	**-0.41**

CNN Interpretation - Effective Weight Matrix

	Hidden Neuron			
Descriptor	1	3	2	4
PNSA-3	-1.80	**-6.57**	0.39	-1.43
RSHM-1	**4.03**	6.15	1.50	1.01
V4P-5	**9.45**	2.15	**3.24**	0.60
S4PC-12	3.36	2.73	**1.99**	0.56
MW	3.94	**8.42**	1.94	0.76
WTPT-2	1.71	2.61	1.17	-0.13
DPHS-1	0.66	0.44	0.33	1.65
SCV	0.52	0.33	0.13	0.01

- The most weighted descriptors are very similar to those in the OLS model
- The signs of the effective weights match those from the OLS model as well as chemical reasoning

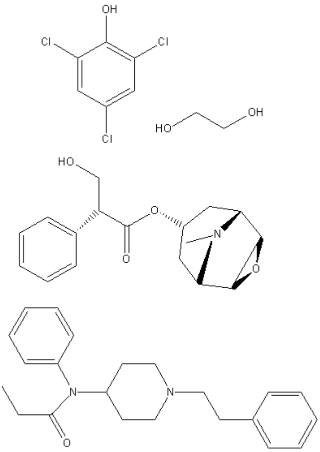
Skin Permeability

Dataset

- Original work reported linear models
- Measured activity was the permeability coefficient (K_p)
- $-5.03 < log(K_p) < -0.85$

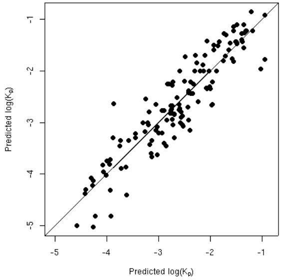
Model details

- 7 descriptor OLS model
- $R^2 = 0.84$, $RMSE = 0.37$ log units
- CNN model was 7-5-1
- $R^2 = 0.94$, $RMSE = 0.23$ log units



Patel, H. et al., *Chemosphere*, **2002**, *48*, 603-613

Linear Interpretation



Component 1 focuses on ...

- Smaller size
- Lower polar surface area
- Larger hydrophobic surface area

Component 2 focuses on ...

- Larger hydrophobic surface area
- Larger surface area
- Corrections for the overestimation or underestimation of some molecules in component 1

	Component	
Descriptor	1	2
SA	−0.08	**0.52**
FPSA-2	**−0.52**	0.14
NN	−0.36	−0.03
MOLC-9	**0.61**	0.11
PPHS-1	0.03	**0.69**
WPHS-3	0.09	**0.48**
RNHS	**0.46**	−0.04

CNN Interpretation - Effective Weight Matrix

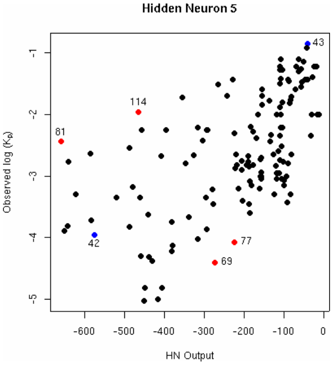
	Hidden Neuron				
Descriptor	5	2	4	3	1
SA	-44.17	**67.34**	8.33	8.18	5.96
FPSA-2	**-156.82**	-10.72	**20.85**	-13.07	-92.47
NN	**-97.81**	2.22	-6.65	1.71	-12.70
MOLC-9	-28.85	**17.79**	15.40	-11.36	-1.20
PPHS-1	**106.55**	**31.30**	-16.76	-13.99	34.55
WPHS-3	-11.36	**-14.31**	-2.31	-10.01	54.16
RNHS	20.16	-5.89	**-49.57**	23.88	27.09
SCV	0.85	0.13	0.02	0.01	0.00

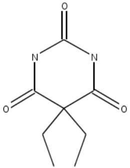
- The most important neuron focuses on hydrophobic & polar effects
- The next most important neuron focuses on size effects

CNN Interpretation - *Score* Plot for Hidden Neuron 5

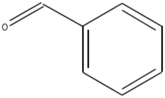
Observations . . .

- SCV = 0.85
- Active molecules area characterized by low polar surface area and larger hydrophobic surface area
- Does not perform too well on inactive molecules
- 69,77 and 81,114 are mispredicted





42 (-3.95)

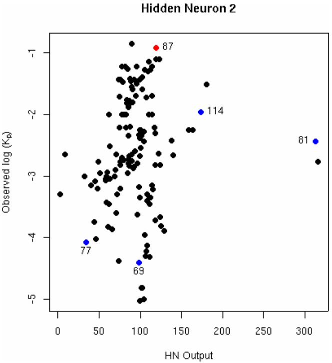


43 (-0.85)

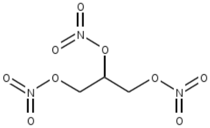
CNN Interpretation - *Score* Plot for Hidden Neuron 2

Observations . . .

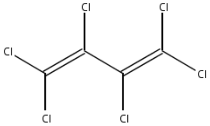
- SCV = 0.13
- Corrects for 69,77 and 81,114
- Describes larger molecules with higher hydrophobic surface area
- MOLC-9 balances the effect of MW
- Molecule 87 is underestimated



77 (-4.07)



114 (-1.96)

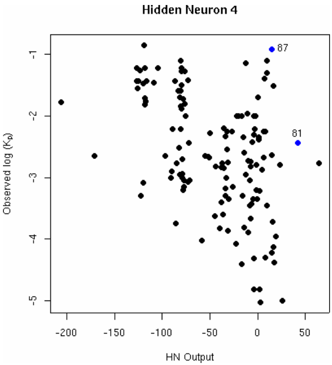


87 (-0.92)

CNN Interpretation - *Score* Plot for Hidden Neuron 4

Observations . . .

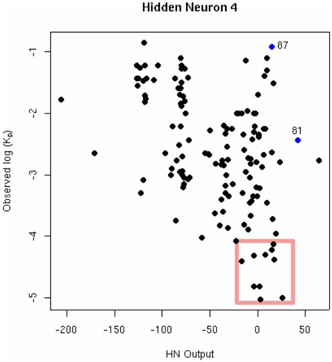
- Corrects underestimation of molecule 87 by HN 2
- Further corrects for molecule 81
- Does not perform well for inactive molecules



CNN Interpretation - *Score* Plot for Hidden Neuron 4

Observations . . .

- Corrects underestimation of molecule 87 by HN 2
- Further corrects for molecule 81
- Does not perform well for inactive molecules
- Overestimation corrected by HN 1



Outline

Caveats

- The method *linearizes* the network
- Clearly, the interpretations will loose some of the details of the encoded SPR's

Conclusions

- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools

Caveats

- The method *linearizes* the network
- Clearly, the interpretations will loose some of the details of the encoded SPR's

Conclusions

- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools

Caveats

- The method *linearizes* the network
- Clearly, the interpretations will loose some of the details of the encoded SPR's

Conclusions

- CNN interpretations appear to be valid
- Discrepancies may be present if we do not select optimal descriptor subsets for the CNN model
- The method avoids complexity and uses only the weights and biases and hence does not use the training set explicitly

The method should help CNN models to be used as design tools as well as predictive tools