FEATURE LEARNING USING STACKED AUTOENCODERS TO PREDICT THE ACTIVITY OF ANTIMICROBIAL PEPTIDES

Francy L. Camacho[1], Rodrigo Torres[2] and Raúl Ramos[3]

[1] School of Computer Science
[2] Grupo de Investigación en Bioquímica y Microbiología (GIBIM), School of Chemistry
[3] Center for High Performance and Scientific Computing, School of Computer Science
[1´2´3] Universidad Industrial de Santander, Carrera 27 calle 9, Bucaramanga, Colombia

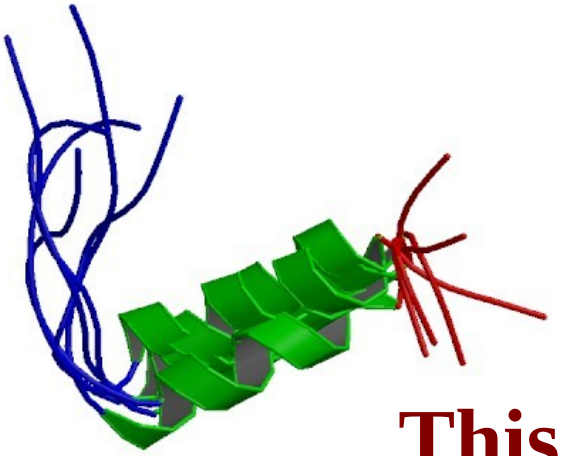
francy.camacho1@correo.uis.edu.co

Universidad Industrial de Santander

Super Computación y Cálculo Científico UIS

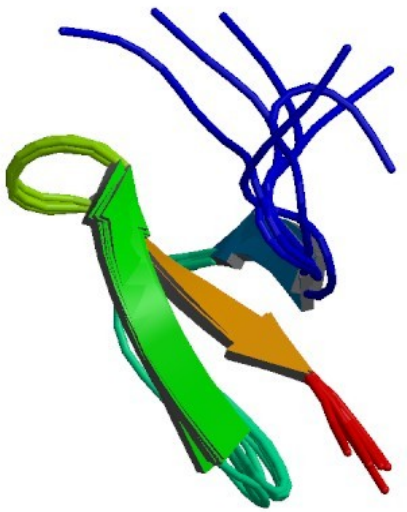
Content

- Introduction
- Prediction of antimicrobial peptides
  - Dataset and descriptors
  - Experimental configurations
- Results
- Conclusions



PDB ID: 1YTR

This work explores the machine learning methods to learn new descriptors from the existing ones, to predict the activity of antimicrobial peptides.



PDB ID: 1BNB

Figure taken from PDB. 1YTR. Activity: anti-Gram+ & Gram-, Cancer cells (http://aps.unmc.edu/AP/database/query_output.php?ID=00035).
1BNB. Activity: anti-Gram+ & Gram- (http://aps.unmc.edu/AP/database/query_output.php?ID=00047)

General Scheme

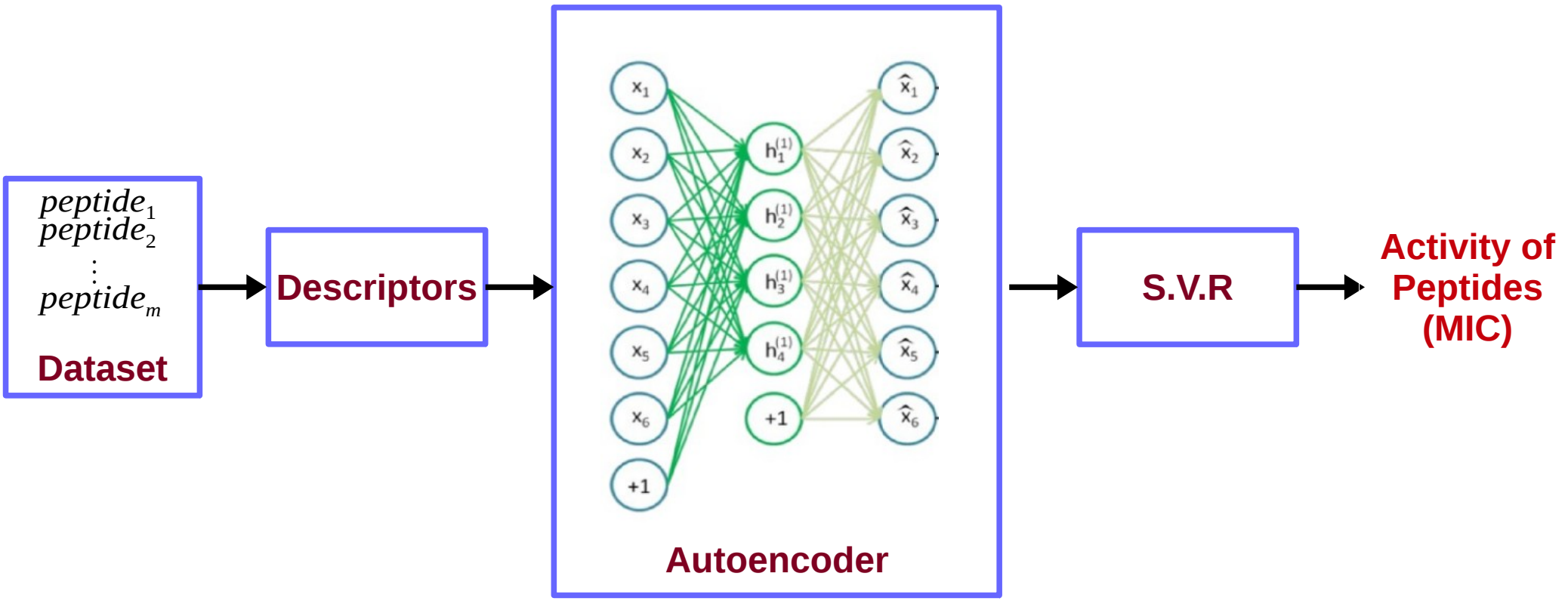


Fig. 1. Graphical representation of methodology used in this work

[1] Ng, A., Ngiam, J. ,Foo, C.Y., et al..: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

General Scheme

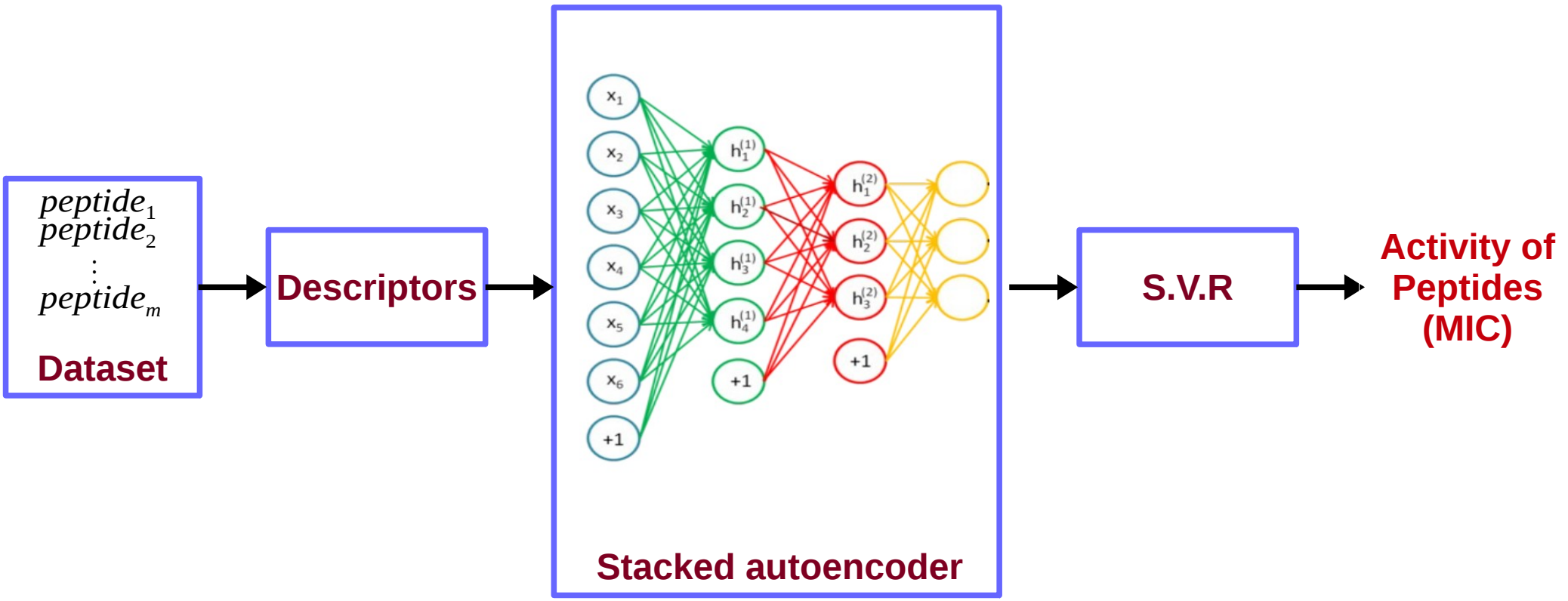


Fig. 1. Graphical representation of methodology used in this work

[1] Ng, A., Ngiam, J. ,Foo, C.Y., et al..: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

Dataset and descriptors

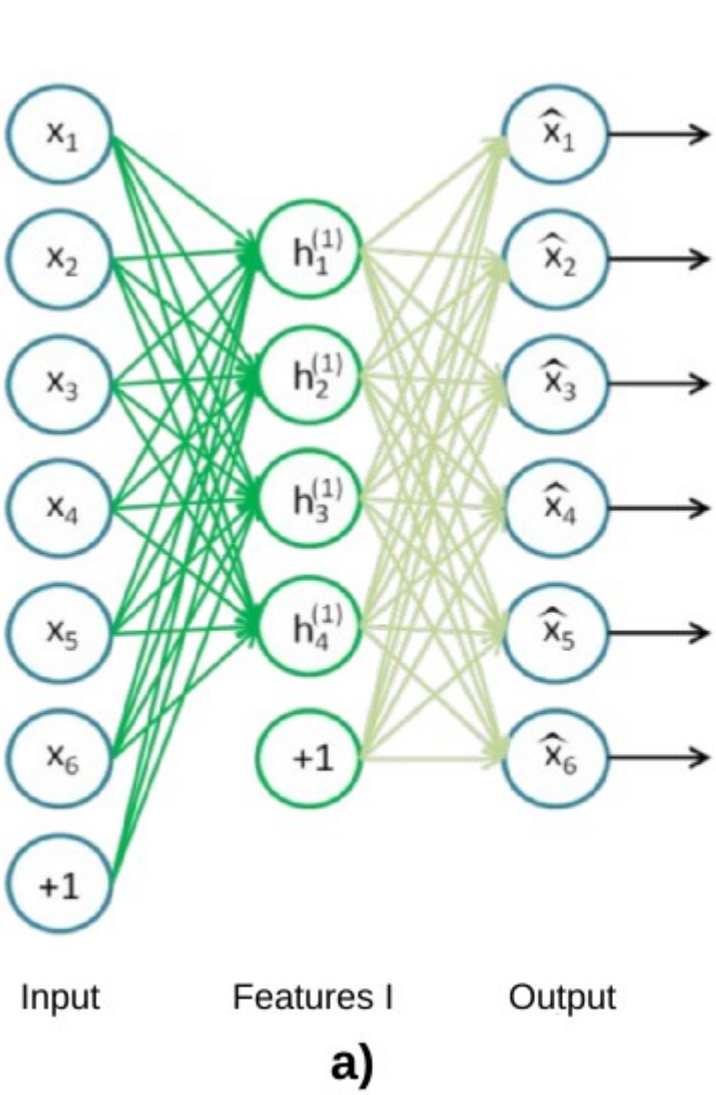
Dataset : CAMELs dataset (101 sequences of peptides)

Table 1. Nine groups of descriptors compute for the dataset. The second and third columns represent the number of descriptors before and after of preprocessing, respectively.

Descriptors	Initial	Final
Dipeptide Composition (Ddcd)	400	106
Normalized MoreauBroto autocorrelation (Dnmba)	240	112
Moran Autocorrelation (Dmad)	240	112
Geary Autocorrelation (Dgad)	240	112
Compositon, Transition and Distribution (Dctd)	147	147
Sequence order coupling number (Dsoc)	20	20
Quasi Sequence Order (Dqso)	50	46
Pseudoaminoacid compositon type I (Dpaac)	30	23
Pseudoaminoacid compositon type II (Dapaac)	30	23
All Descriptors (AllDesc)	1517	730

[2] Cao, D.S., Xu, Q.S., Liang, Y.Z. *Bioinformatics Applications Note* 29(7), 960–2 (Apr 2013)

Autoencoders and Stacked autoencoders



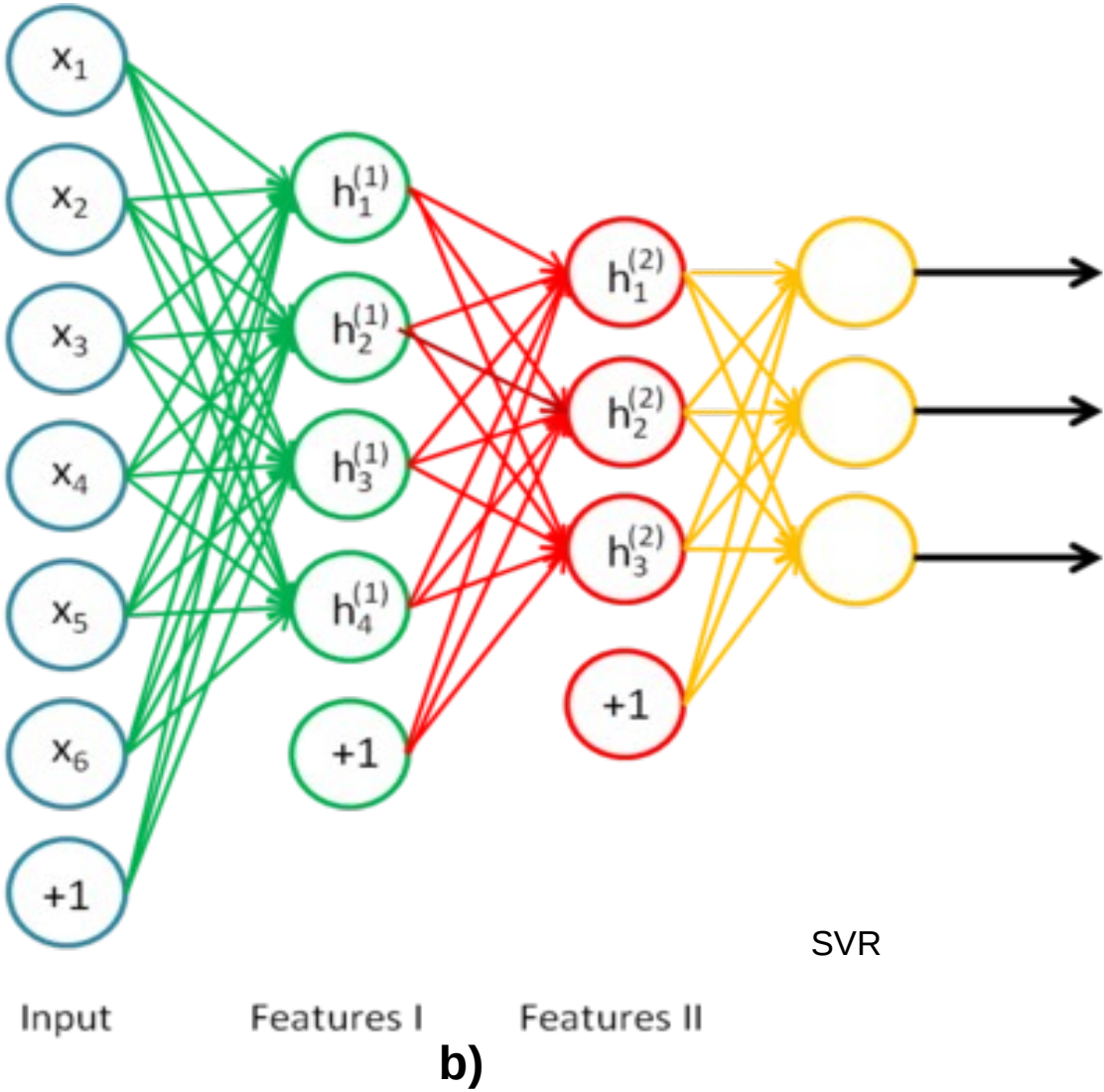
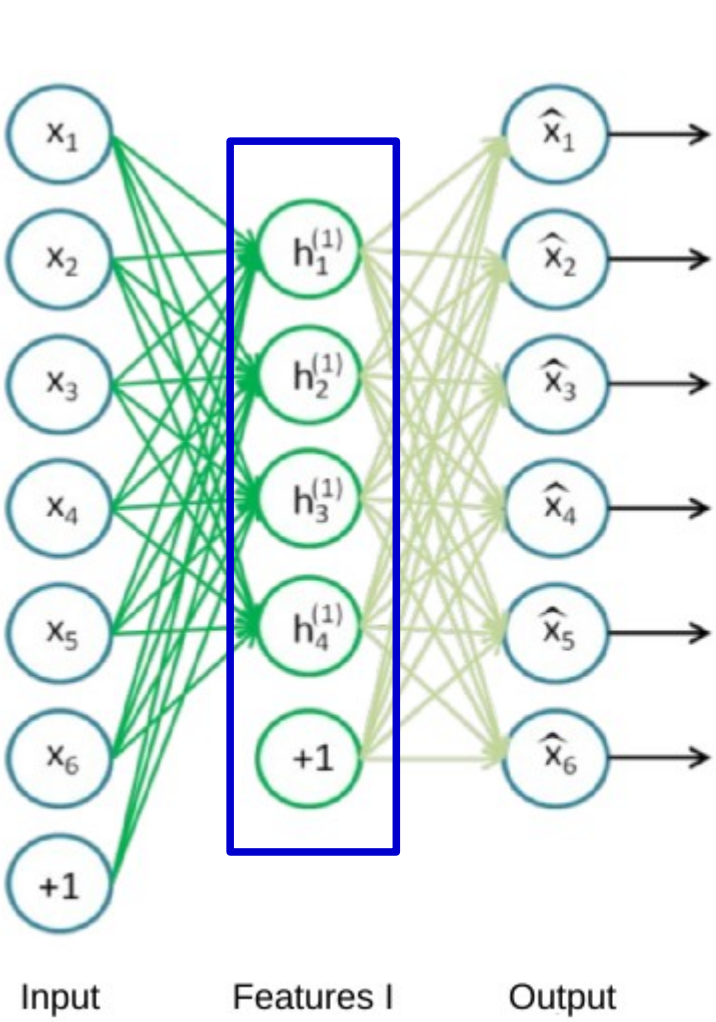


Fig. 2. Figure 1,a) is a Sparse Autoencoder and it contains 6 input neurons and four neurons in the hidden layer. Figure 1,b) is a stacked autoencoder with 2 hidden layer. Note this is a compressing autoencoder. It contains (6+1)*4+(4+1)*6 = 58 connections. Figure taken from [1]

[1] Ng, A., Ngiam, J. ,Foo, C.Y., et al..: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

Autoencoders and Stacked autoencoders



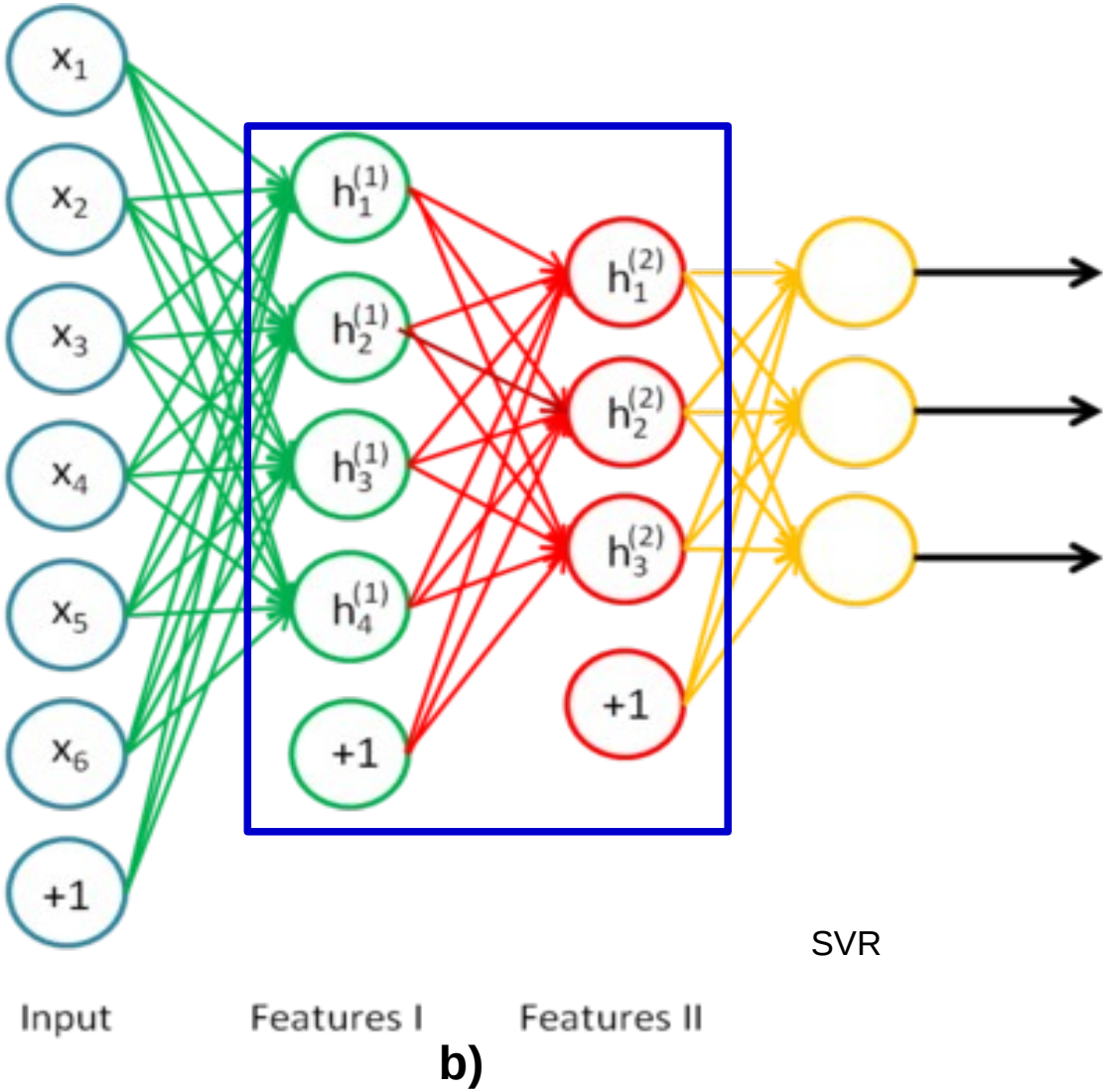


Fig. 2. Figure 1,a) is a Sparse Autoencoder and it contains 6 input neurons and four neurons in the hidden layer. Figure 1,b) is a stacked autoencoder with 2 hidden layer. Note this is a compressing autoencoder. It contains (6+1)*4+(4+1)*6 = 58 connections. Figure taken from [1]

[1] Ng, A., Ngiam, J. ,Foo, C.Y., et al..: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

Processing workflow

1. **Preprocessing:** all descriptors are preprocessed by standardizing and removing constant descriptors.

2. **Unsupervised feature learning:** differents configurations of autoencoders (AE) and stacked autoencoders (SAE).

3. **Supervised prediction:** Support Vector Regression task are run on the new representation.

Experimental configurations

Original: We performed SVR on original descriptors.

AE: We trained different configurations of AE (20-1000 neurons in the hidden layer).

SAE2: Second layer was the half of first one.

SAE4: Second layer was a quarter of first one.

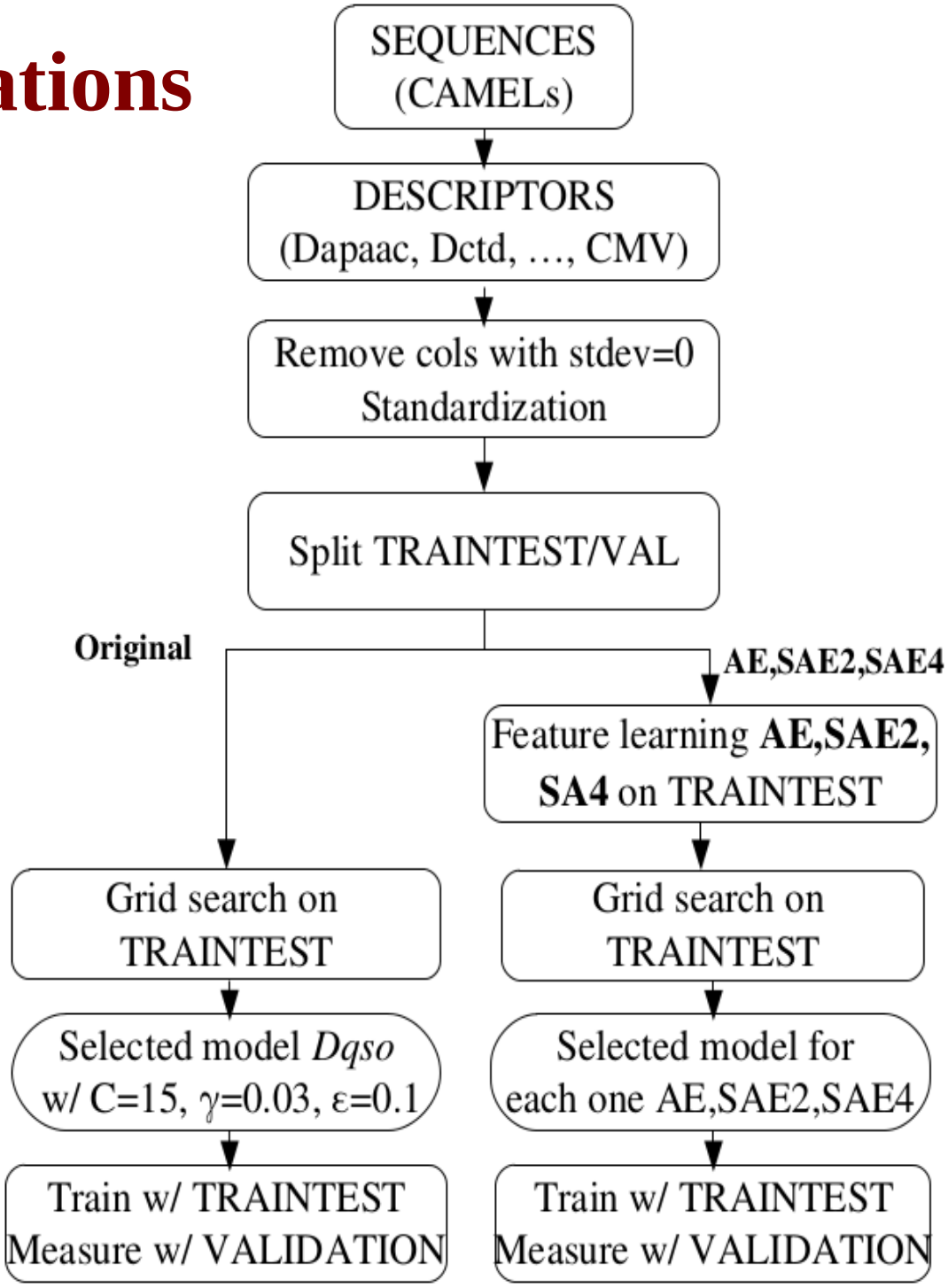


Fig 3. Graphical representation of methodology used in this work

Validation and supervised training

- Pearson correlation coefficient of prediction R_{ext}
- Correlation coefficient of multiple determination R^2_{ext}
- Root Mean Square Error of Prediction $RMSE_{ext}$
- R^2 predictive

[3] Kiralj, R., Ferreira, M.M.C. Theory and Application. *Journal of the Brazilian Chemical Society* 20(4), 770–787 (2009)
[4] Pratim, P., Paul S., Mitra, I. and Roy, K.. *Molecules* (Basel, Switzerland), 14:1660–1701, (2009).

Results

Table 2. Comparative results for different algorithms used for prediction of antimicrobial peptides

METHOD *	R_{ext}	$RMSE_{ext}$	R^2_{ext}	R^2_{pred}	Ref
GA-SVM	0.78	1.39	-	-	[5]
PSO-GA-SVM	0.90	0.96	-	-	[5]
STR-MLR	-	-	0.33	-	[6]
G-PLS	0.80	-	0.67	0.64	[7]
ANN	-	-	0.72	-	[8]
Setup Original (Dqso + SVR)	0.87	1.10	0.73	0.74	This work
Setup AE (Dctd (900) + SVR)	0.90	1.10	0.74	0.74	This work
Setup SAE2 (Dqso(140,70)+SVR)	0.96	0.86	0.84	0.84	This work
Setup SAE4 (Dqso(800,200)+SVR)	0.97	0.84	0.85	0.85	This work

* GA = Genetic Algorithms, SVR = Support Vector Regression, STR = , MLR = Multiple Linear Regression, G/PLS = Genetic Function Approximation/Partial Least Square, ANN = Artificial Neural Network

[5] Zhou, X., Li, Z., Dai, Z., Zou, X. *Journal of molecular graphics & modelling* 29(2), 188–196 (Jun 2010)
[6] Wang, Y., Ding, Y., Wen, H., et al. *Combinatorial chemistry & high throughput screening* 15(4), 347–353 (May 2012)
[7] Borkar, M.R., Pissurlenkar, R.R.S., Coutinho, E.C. *Journal of computational chemistry* 34(30), 2635–46 (2013)
[8] Torrent, M., Andreu, D., Nogués, V.M., Boix, E. *PLoS one* 6(2), e16968 (Jan 2011),

Results

Table 2. Comparative results for different algorithms used for prediction of antimicrobial peptides

METHOD *	R_{ext}	$RMSE_{ext}$	R^2_{ext}	R^2_{pred}	Ref
GA-SVM	0.78	1.39	-	-	[5]
PSO-GA-SVM	0.90	0.96	-	-	[5]
STR-MLR	-	-	0.33	-	[6]
G-PLS	0.80	-	0.67	0.64	[7]
ANN	-	-	0.72	-	[8]
Setup Original (Dqso + SVR)	0.87	1.10	0.73	0.74	This work
Setup AE (Dctd (900) + SVR)	0.90	1.10	0.74	0.74	This work
Setup SAE2 (Dqso(140,70)+SVR)	0.96	0.86	0.84	0.84	This work
Setup SAE4 (Dqso(800,200)+SVR)	0.97	0.84	0.85	0.85	This work

* GA = Genetic Algorithms, SVR = Support Vector Regression, STR = , MLR = Multiple Linear Regression, G/PLS = Genetic Function Approximation/Partial Least Square, ANN = Artificial Neural Network

[5] Zhou, X., Li, Z., Dai, Z., Zou, X. *Journal of molecular graphics & modelling* 29(2), 188–196 (Jun 2010)
[6] Wang, Y., Ding, Y., Wen, H., et al. *Combinatorial chemistry & high throughput screening* 15(4), 347–353 (May 2012)
[7] Borkar, M.R., Pissurlenkar, R.R.S., Coutinho, E.C. *Journal of computational chemistry* 34(30), 2635–46 (2013)
[8] Torrent, M., Andreu, D., Nogués, V.M., Boix, E. *PLoS one* 6(2), e16968 (Jan 2011),

Results

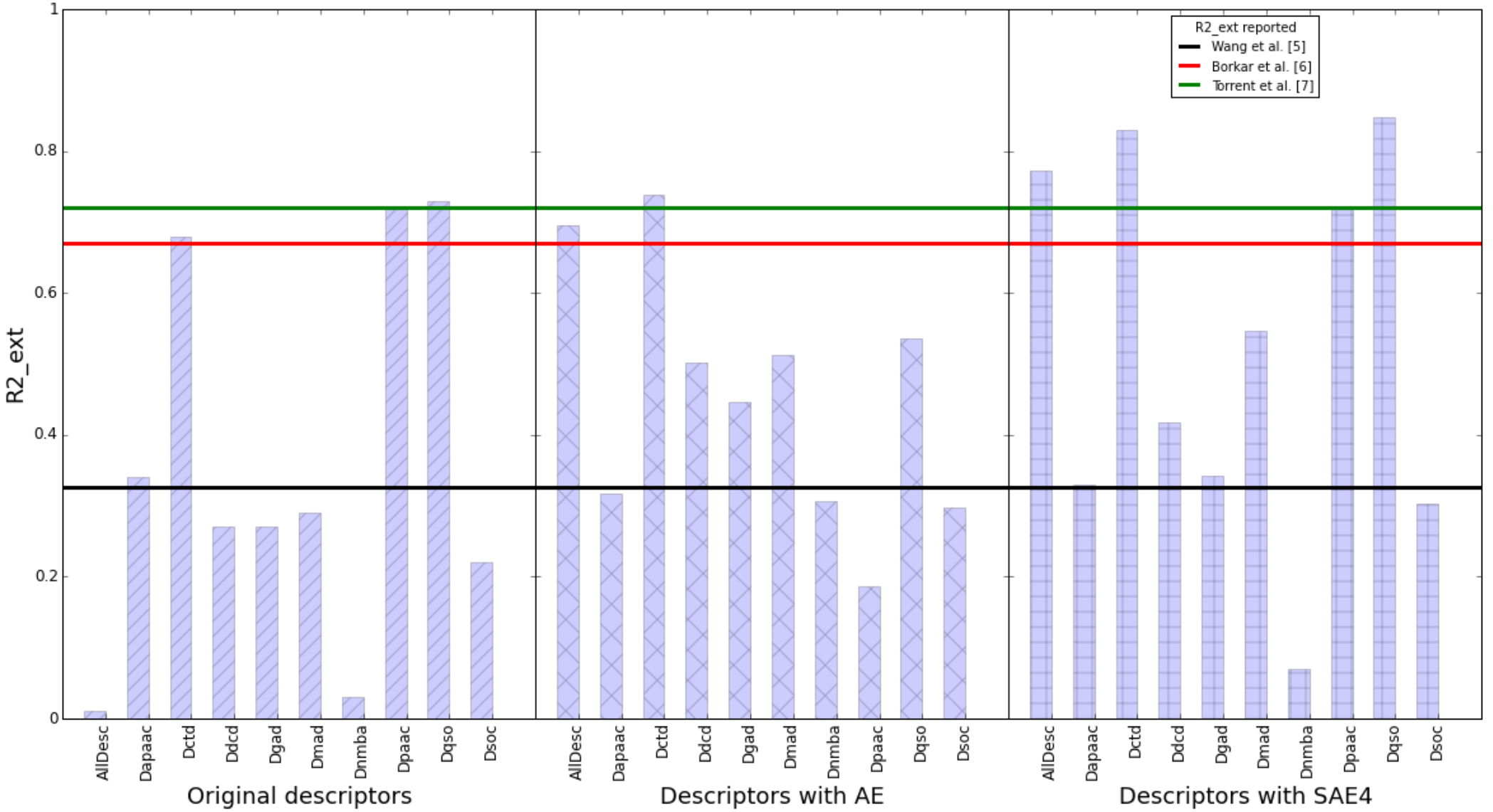


Fig. 4. Best performance for R^2_{ext} for each group of descriptors in tree experimental setup. The R^2_{ext} is the best when is 1.

[6] Wang, Y., Ding, Y., Wen, H., et al. *Combinatorial chemistry & high throughput screening* 15(4), 347–353 (May 2012)

[7] Borkar, M.R., Pissurlenkar, R.R.S., Coutinho, E.C. *Journal of computational chemistry* 34(30), 2635–46 (2013)

[8] Torrent, M., Andreu, D., Nogués, V.M., Boix, E. *PLoS one* 6(2), e16968 (Jan 2011),

Results

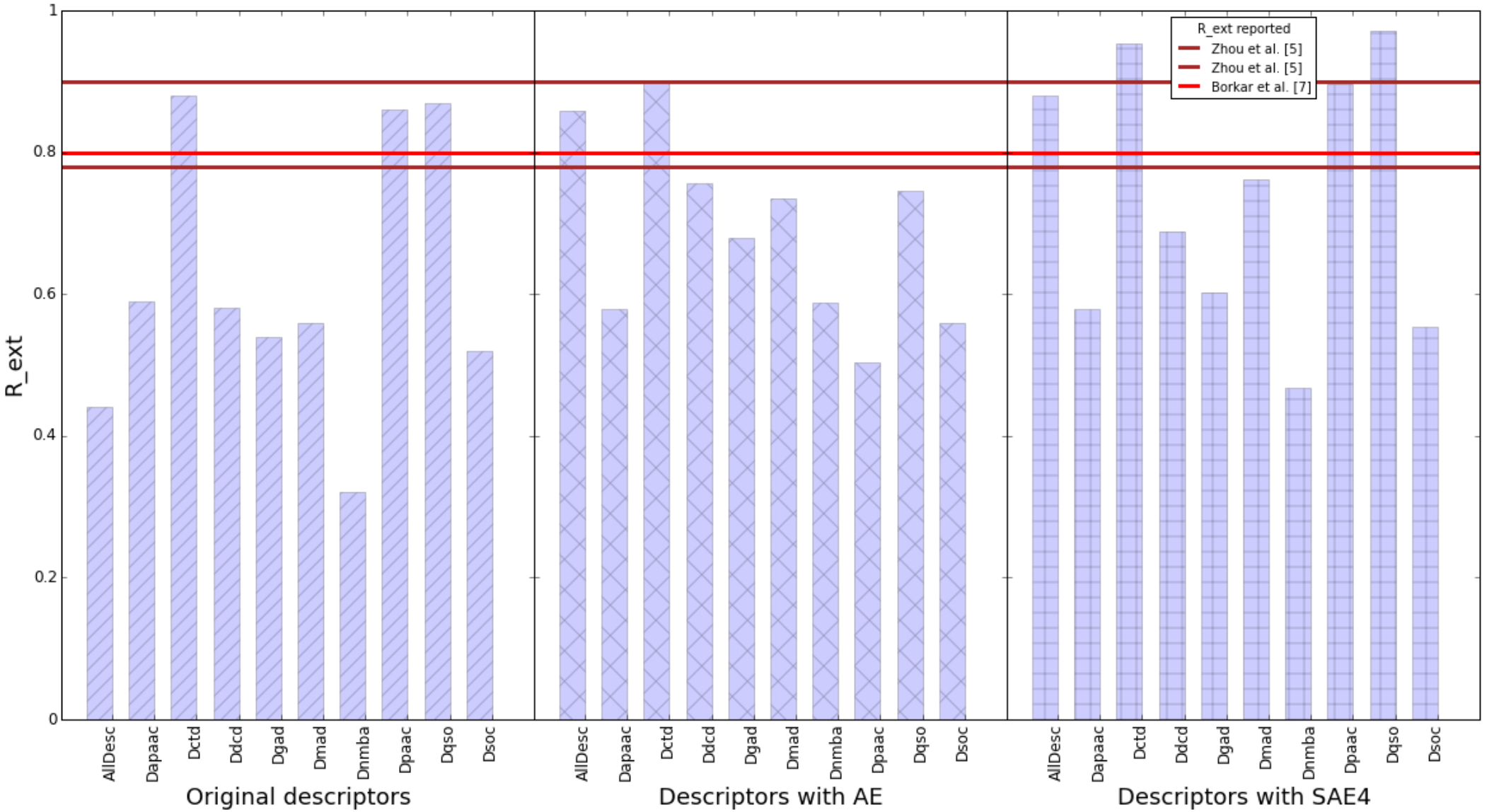


Fig. 5. Best performance for R_{ext} for each group of descriptors in three experimental setup. The R_{ext} is the best when is 1.

[5] Zhou, X., Li, Z., Dai, Z., Zou, X. *Journal of molecular graphics & modelling* 29(2), 188–196 (Jun 2010)

[7] Borkar, M.R., Pissurlenkar, R.R.S., Coutinho, E.C. *Journal of computational chemistry* 34(30), 2635–46 (2013)

Results

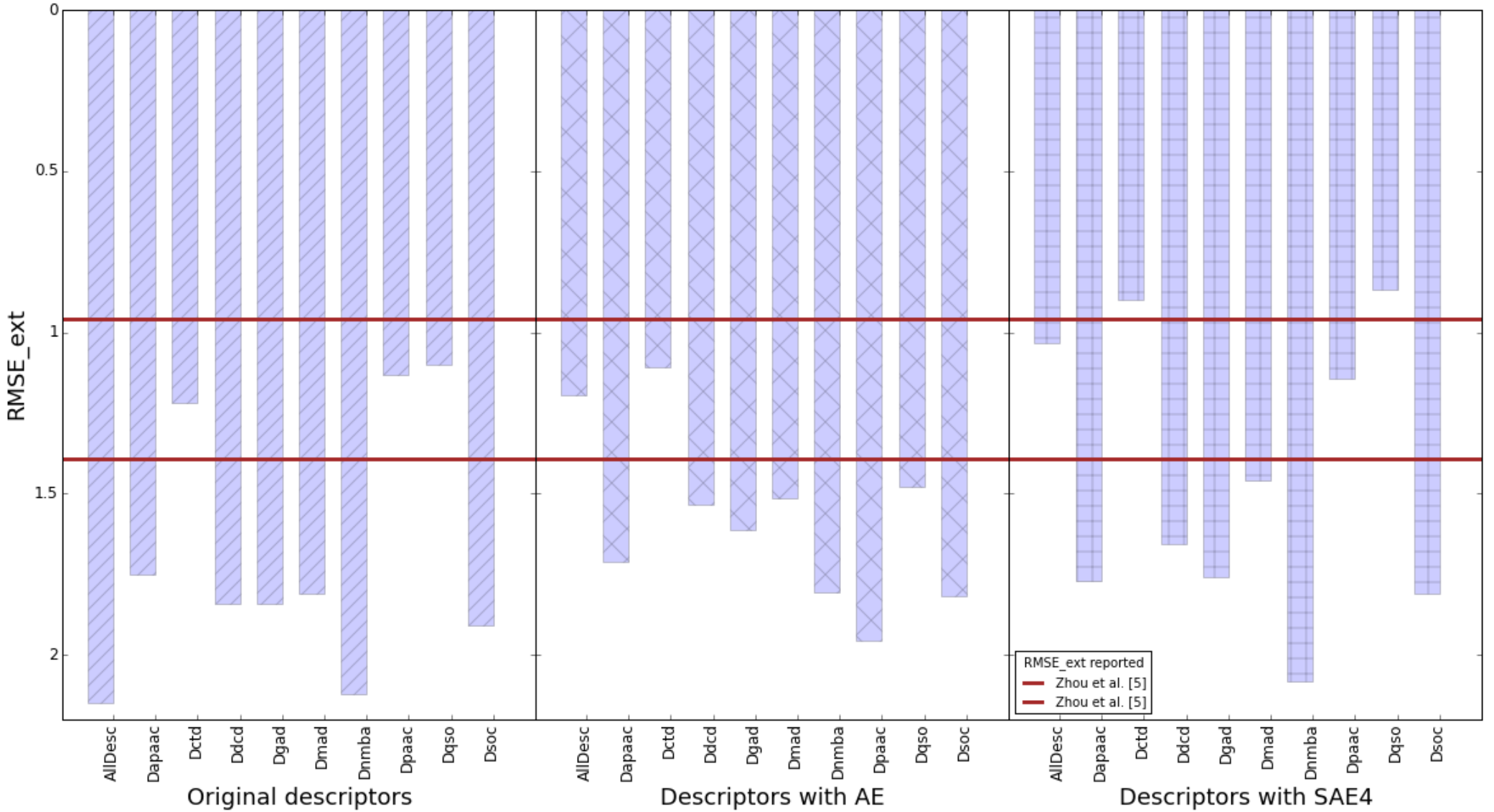


Fig. 6. Best performance for $RMSE_{ext}$ for each group of descriptors in three experimental setup. $RMSE_{ext}$ closer to zero is better performance.

[5] Zhou, X., Li, Z., Dai, Z., Zou, X. *Journal of molecular graphics & modelling* 29(2), 188–196 (Jun 2010)

Results

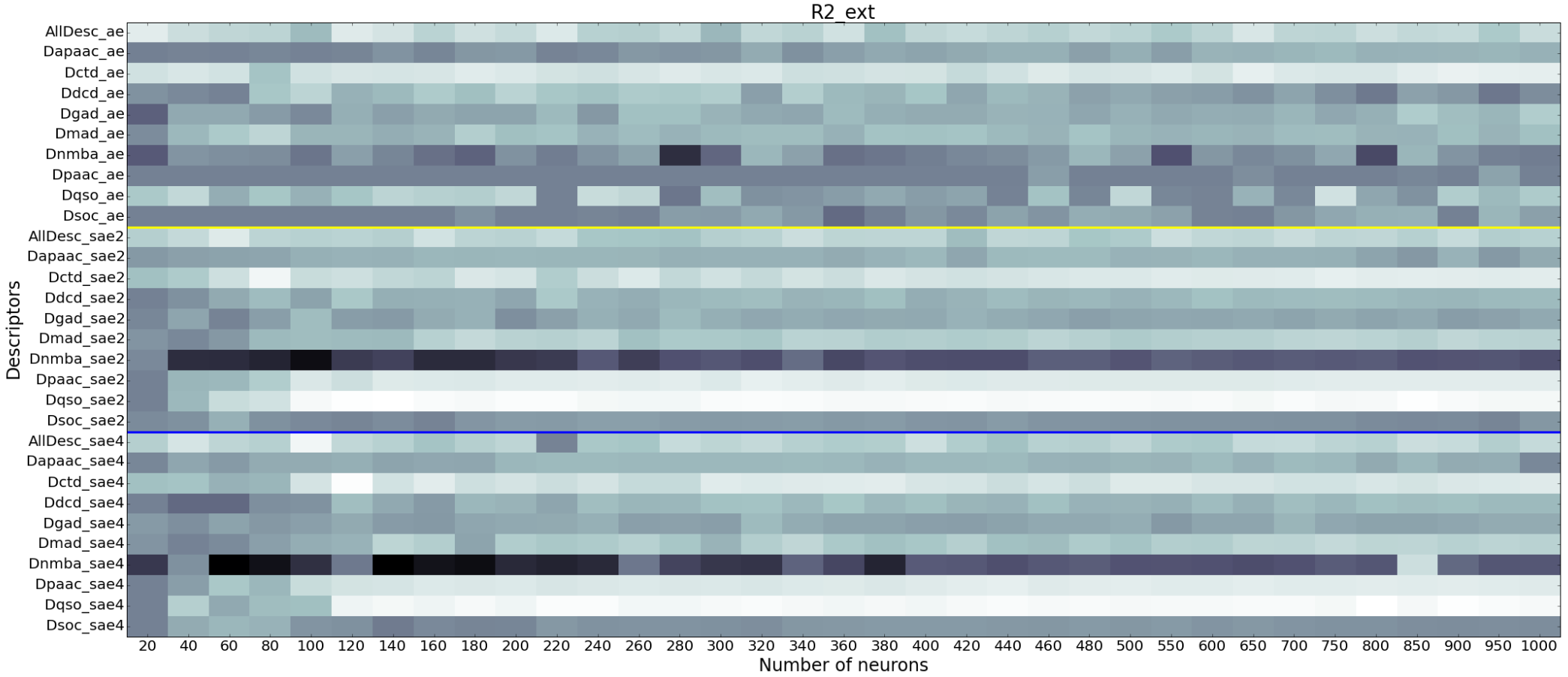


Fig. 7. Graphical representation of results for autoencoder and stacked autoencoder for each variation of number of neurons in the first hidden layer. The best result for AE was Dctd, SAE2 was Dqso and SAE4 was Dqso. The best score is represented with white and the worst is shown with dark blue. Configurations with more neurons in the hidden layers seem to work better.

Results

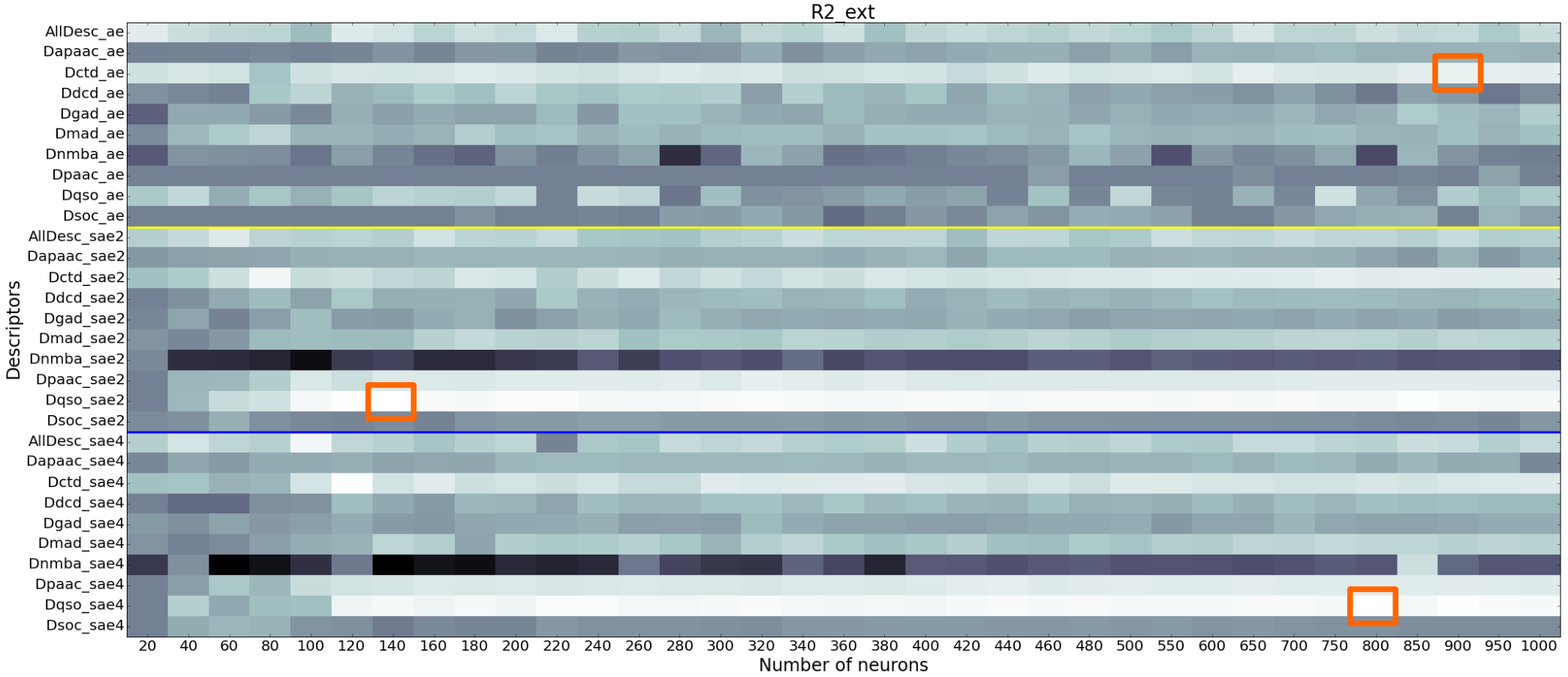
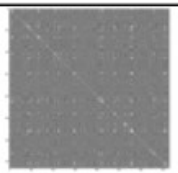
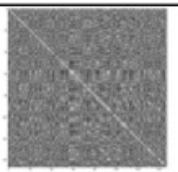
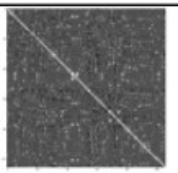
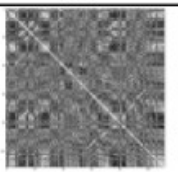
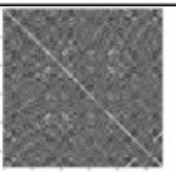
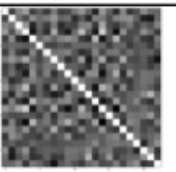
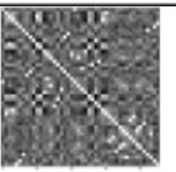
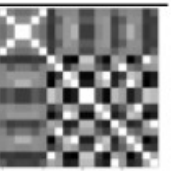
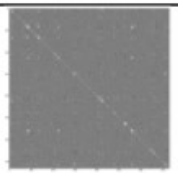
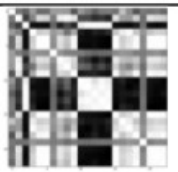
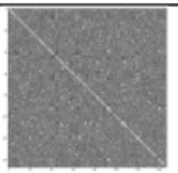
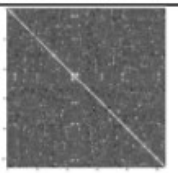
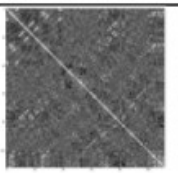
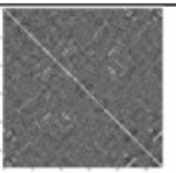
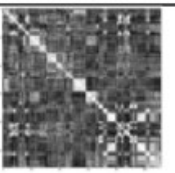
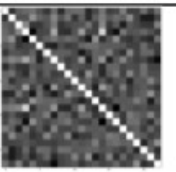
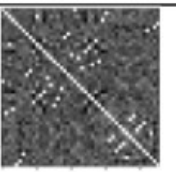
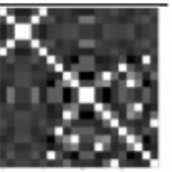


Fig. 7. Graphical representation of results for autoencoder and stacked autoencoder for each variation of number of neurons in the first hidden layer. The best result for AE was Dctd, SAE2 was Dqso and SAE4 was Dqso. The best score is represented with white and the worst is shown with dark blue. Configurations with more neurons in the hidden layers seem to work better.

Results

Table 3. Correlation among descriptors within each group for SAE configurations compared with the original representation. The lowest correlation is represented with dark pixel and the highest correlation is shown with white pixel.

-	AllDesc	Dapaac	Dctd	Ddcd	Dgad	Dmad	Dnmba	Dpaac	Dqso	Dsoc
Original										
SAE										
Neurons	100,25	460,230	120,30	220,110	100,50	180,90	40,10	320,160	800,200	60,15

Conclusions

When feeding the new features to a supervised machine learning method, we show how learnt representations consistently provide satisfactory results compared with recent works.

Besides, we also show how, sparse representations seem to be preferable to more compact ones, giving a better chance for data separability for the supervised prediction task later on.

Moreover, we also show how the learnt representations also enhance the independence of the initial descriptors reducing the correlation among them.

Conclusions

We believe this approach to be worthwhile exploring in other areas in protein prediction sharing data characteristics and problem complexity.

However, we also observed the importance of the selection of the original set of descriptors from which the learning process starts.

This suggests probably hybrid approaches where specialists hand-craft a base collection of descriptors and the unsupervised learning process complements them.

Acknowledgments

The authors thank the support of the High Performance and Scientific Computing Centre and Grupo de Investigación en Bioquímica y Microbiología at Universidad Industrial de Santander. This project was funded by COLCIENCIAS (Project number: 1102-5453-1671) and Vicerrectoría de Investigación y Extensión (VIE) from UIS.

Cañón del Chicamocha, Colombia

https://commons.wikimedia.org/wiki/File:Ca%C3%B1%C3%B3n_del_Chicamocha_2014-09-21_%282%29.jpg

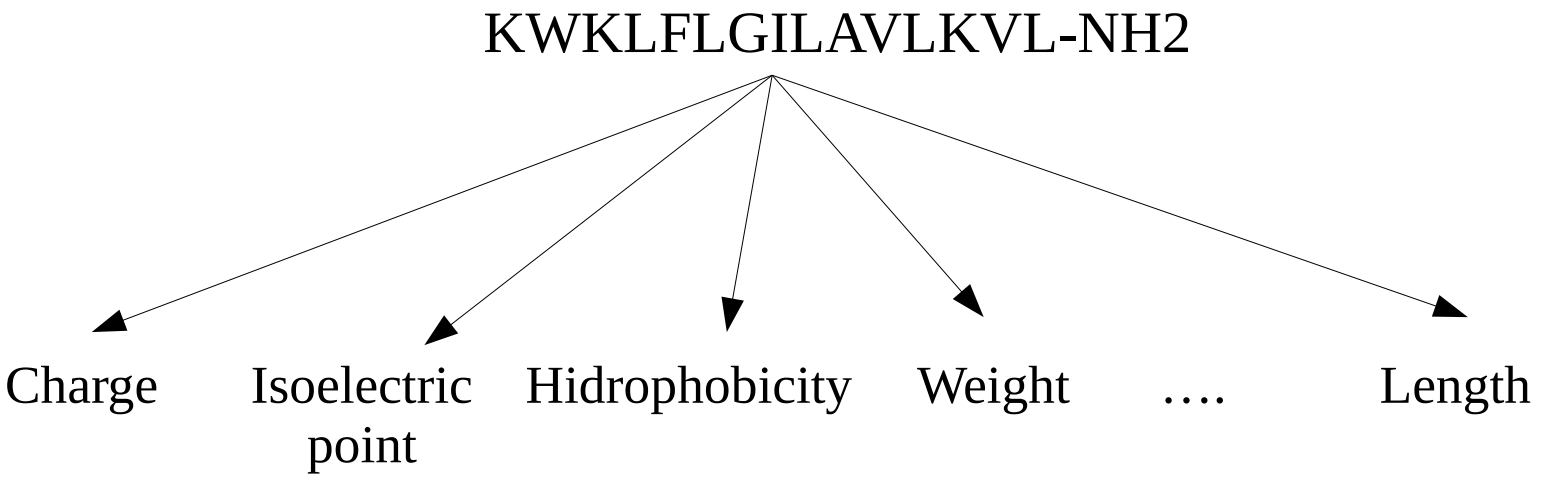
CAMELs

- High identity
- Same size
- MIC between 0.5 – 7
- 101 sequences
- Amino terminal
- Alpha helix

Sequence	Activiy
KWKLFLGILAVLKVL-NH2	0.159
KWNLNGNINAVLKVL-NH2	0.209
KWKGELEIEAELKVL-NH2	0.376
GWKLGLKILNVLKVL-NH2	0.496
KWKLFKKNNNNNKHN-NH2	0.498

Artem Cherkasov and Bojana Jankovic. Application of 'inductive' QSAR Descriptors for Quantification of Antibacterial Activity of Cationic Polypeptides. Molecules (Basel, Switzerland), 9(12):1034–52, January 2004.

Descriptors



Descriptors

Charge	Isoelectric point	Hidrophobicity	Weight	….	Length
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	…	$X_{1,n}$
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	…	$X_{1,n}$
$X_{m,1}$	$X_{m,2}$	$X_{m,3}$	$X_{m,4}$	…	$X_{m,n}$

Descriptors

Charge	Isoelectric point	Hidrophobicity	Weight	….	Length
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	…	$X_{1,n}$
$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	…	$X_{1,n}$
$X_{m,1}$	$X_{m,2}$	$X_{m,3}$	$X_{m,4}$	…	$X_{m,n}$

Descriptors	
Dipeptide Composition (Ddcd)	The fraction of dipeptide type in a sequence.
Normalized MoreauBroto autocorrelation (Dnmba)	Based on the distribution of amino acid properties along the sequence (Hydrophobicity scales,Average flexibility indices, Polarizability parameter, Free energy of solution in water, Residue accessible surface area in trepeptide, Residue volume, Steric Parameter, Relative mutability)
Moran Autocorrelation (Dmad)	
Geary Autocorrelation (Dgad)	
Compositon, Transition and Distribution (Dctd)	It is computed for a given attribute to describe the global percent composition of each of the three groups in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein, and the distribution pattern of the attribute along the sequence, respectively
Sequence order coupling number (Dsoc)	They are derived from the physicochemical distance matrix between each pair of the 20 amino acids
Quasi Sequence Order (Dqso)	
Pseudoaminoacid compositon type I (Dpaac)	It is made up of a 50-dimensional vector in which the first 20 components reflect the effect of the amino acid composition and the remaining 30 components reflect the effect of sequence order
Pseudoaminoacid compositon type II (Dapaac)	

[2] Cao, D.S., Xu, Q.S., Liang, Y.Z. *Bioinformatics Applications Note* 29(7), 960–2 (Apr 2013)

Autoencoders and Stacked autoencoders

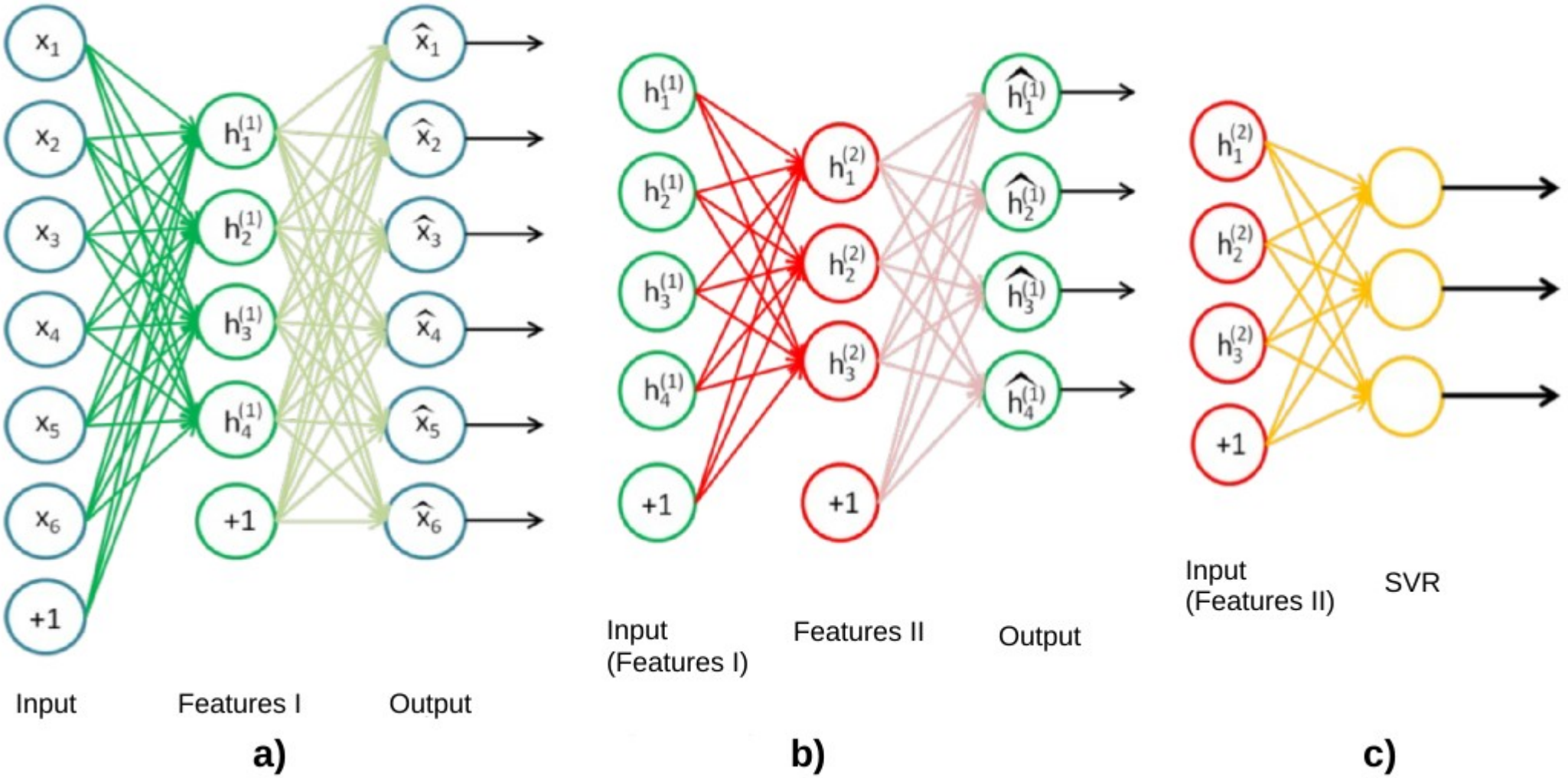


Figure taken from Ng, A., Ngiam, J., Foo, C.Y., Mai, Y., Suen, C.: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

Autoencoders and Stacked autoencoders

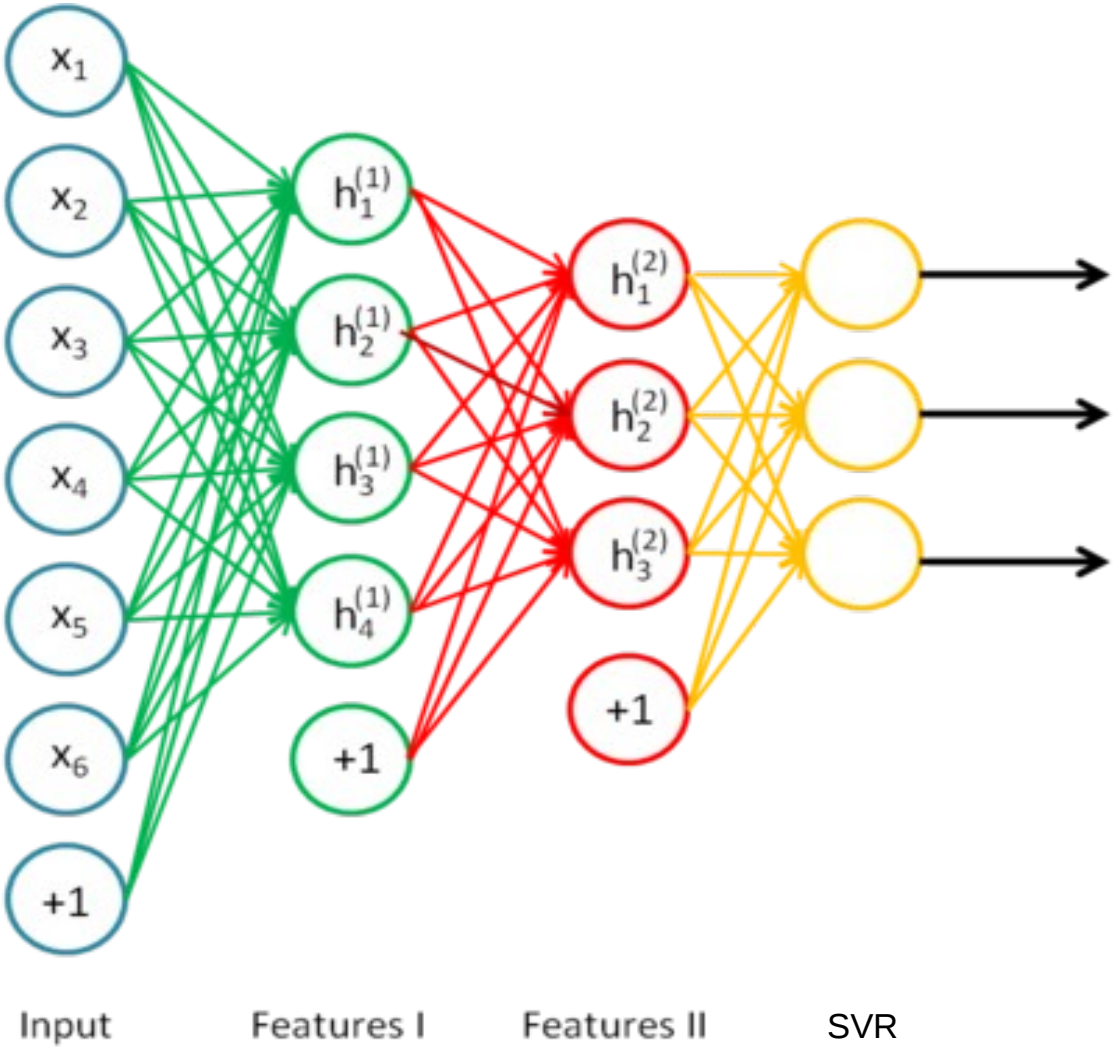
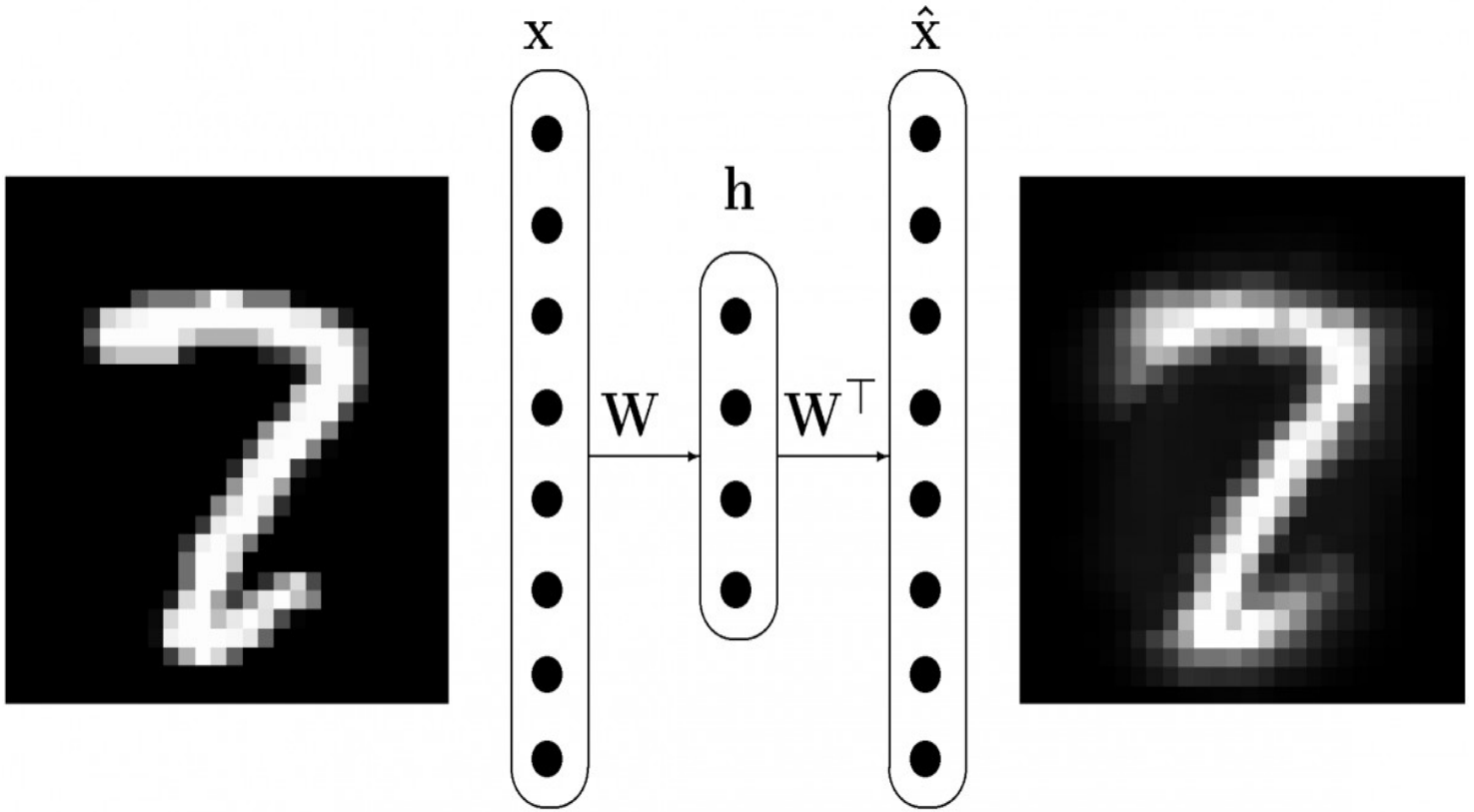


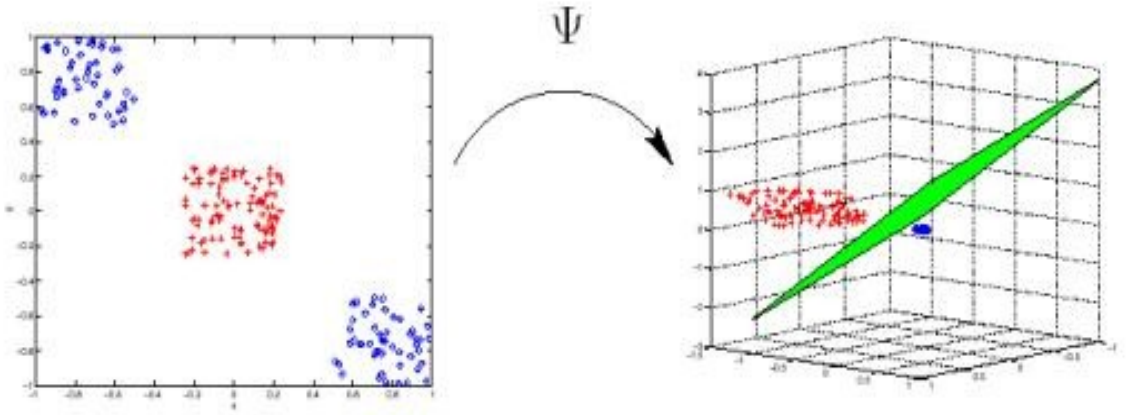
Figure taken from Ng, A., Ngiam, J., Foo, C.Y., Mai, Y., Suen, C.: Unsupervised Feature Learning and Deep Learning, http://ufldl.stanford.edu/wiki/index.php/UFLDL_Tutorial

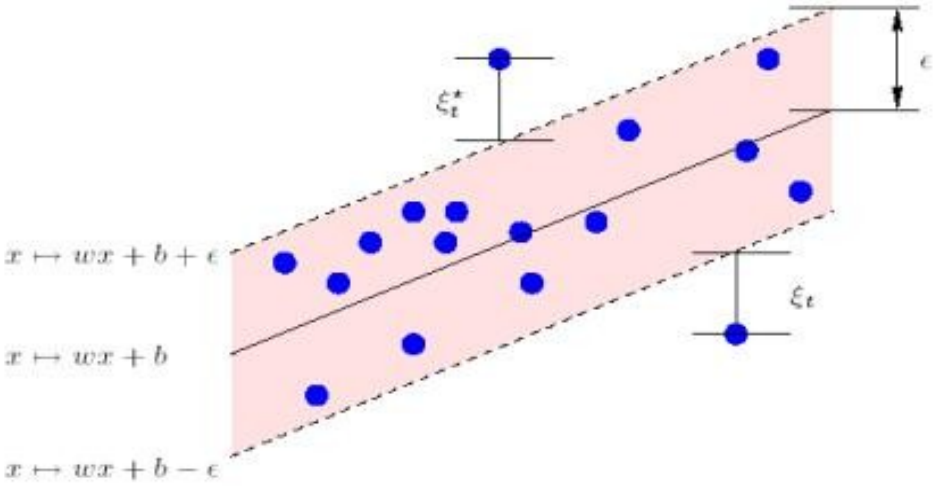
Sparse Autoencoders



(a) Input (b) Neural Encoding (c) Reconstruction

Support Vector Regression





- Source of http://kernelsvm.tripod.com/

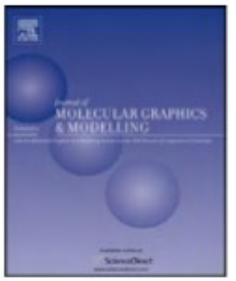
Baseline

Journal of Molecular Graphics and Modelling 29 (2010) 188–196

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



QSAR modeling of peptide biological activity by coupling support vector machine with particle swarm optimization algorithm and genetic algorithm

Xuan Zhou[a,b], Zhanchao Li[a], Zong Dai[a], Xiaoyong Zou[a,*]

[a] School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, PR China
[b] School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou 510006, PR China