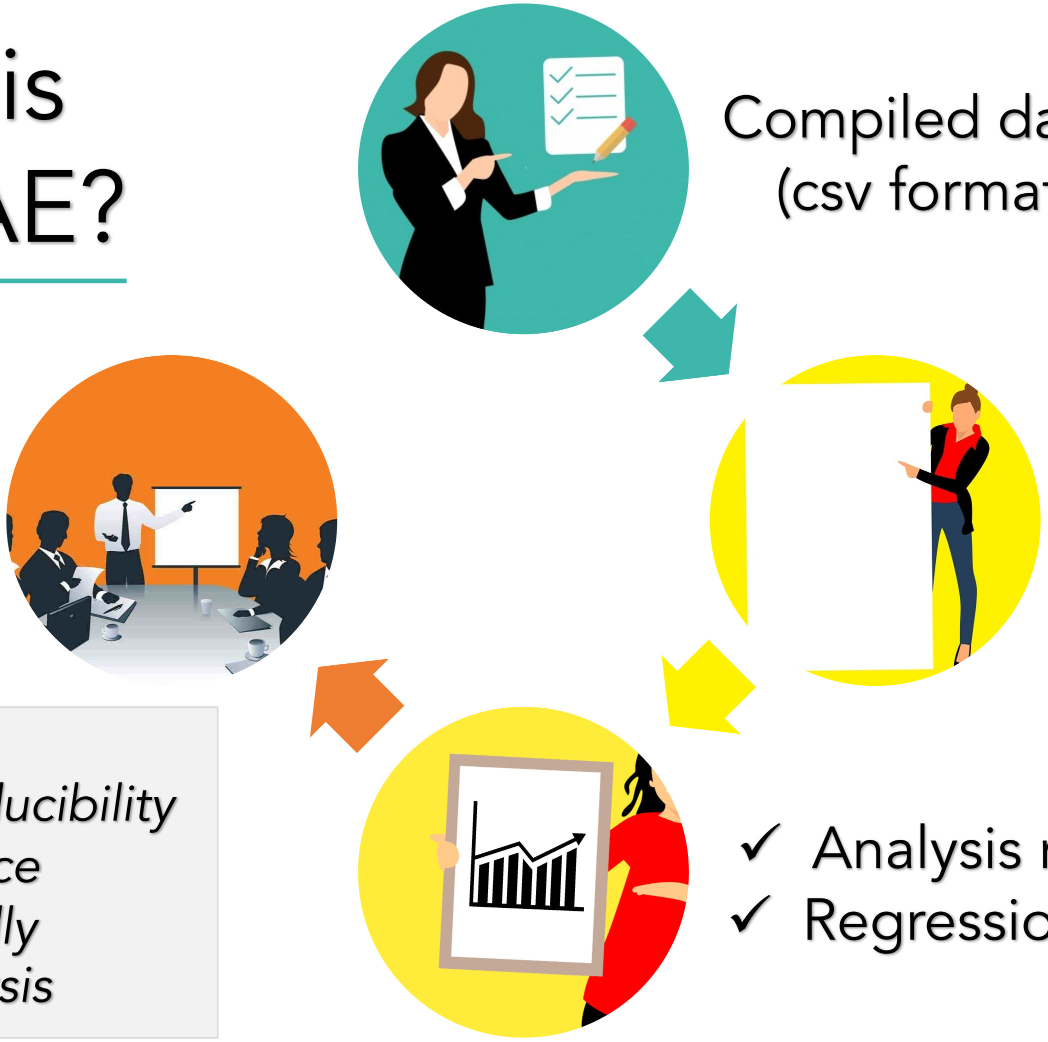
REGGAE

Regression Generator & Analyzer

How-to-reggae at Coding Camp

By Liliana Gallegos

What is REGGAE?



Compiled data
(csv format)

Select analysis
Run regression

✓ Analysis report
✓ Regression plot

<u>Aims:</u>

Statistical reproducibility
open-source
user-friendly
quick analysis

Setup

□ Create r environment and install r-essentials:

```
conda create –n r_env r-essentials r-base
```

□ Activate environment:

```
conda activate r_env
```

□ Install packages required:

```
conda install --yes --file R-requirements.txt
```

□ Confirm by running reggae help options:

```
Rscript reggae.r -h
```

Available: *https://github.com/Liliana-Gallegos/REGGAE*

Statistical analysis options

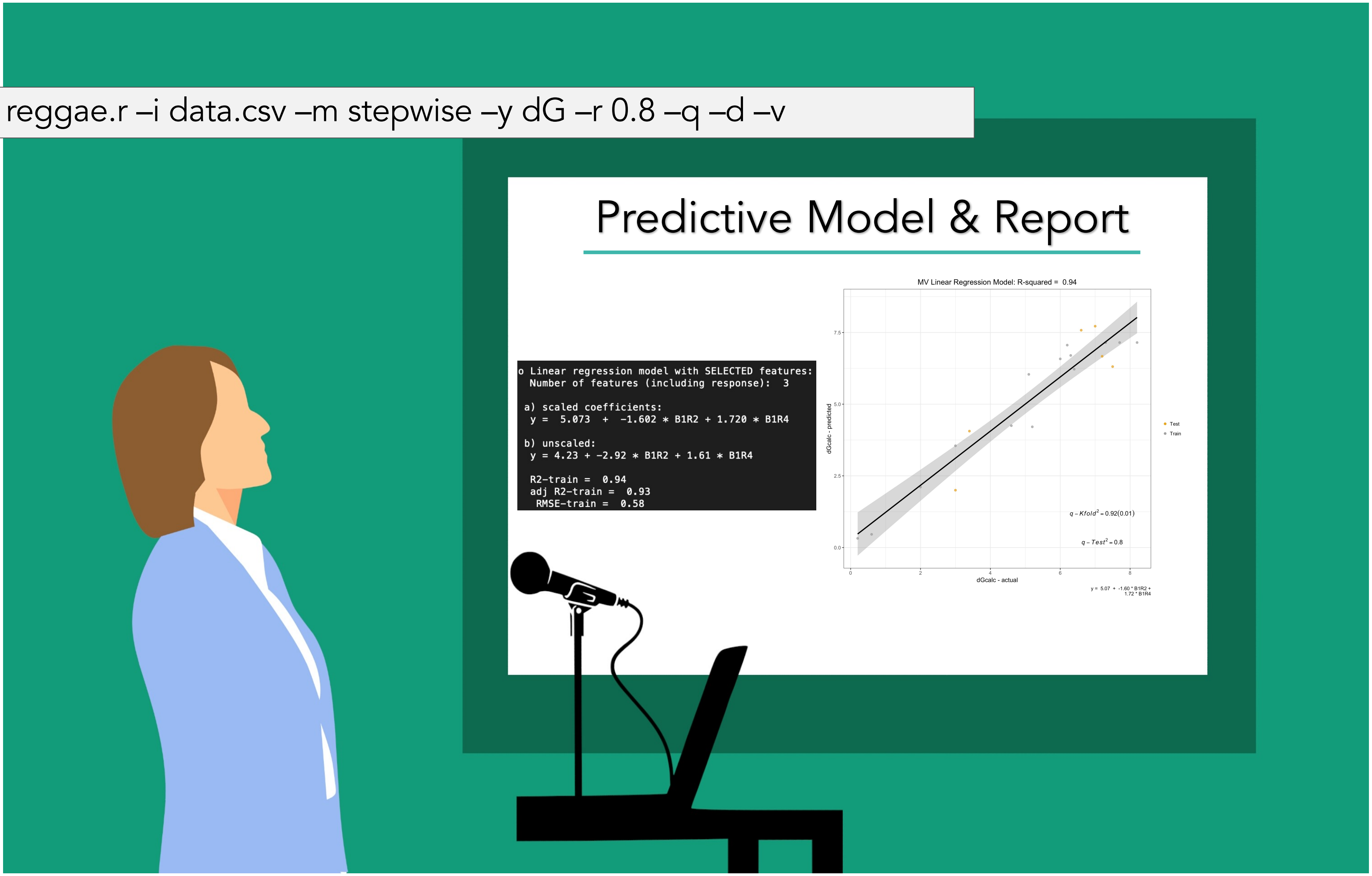
Regression analysis: Multivariate Linear or Random Forest (*quick not opt)*

–m , --model	Feature selection: *full, stepwise, dredge, mincorr*
–b , --buildmodel	Build a model: x_1, x_2, x_n
–y , --yresponse	Define y-response
	Split into Train/ Test datasets: [Default = 1 full dataset]
–r , --randsample	(a) ratio between 0 and 1 or (b) 0 = predefined test/train
–p , --pca	Principal Component analysis with kmeans clustering.
	Build scaled/unscaled train and test sets random or universal training set.
–c , --corrplot	Pairwise Correlation
–q , --crossvalidation	Cross Validation q^2 values from leave-one-out, K-fold, and external.
–d , --diagnostics	Diagnostics QSAR criteria for an acceptable model, F-value comparisons, collinearity, *outlier testing*.
	ANOVA analysis.
–v , --verbose	Plots for selected analysis.

Other options

–i , --inputfile	Requires csv file name.
–o , --outputfile	Optional: output file name (default = REGGAE-analysis-output.txt)
–x , --extdata	To validate model using external csv data set. Requires the name of csv data file.
–e , --exportdata	Export data sets. Select from option: scaled, unscaled, predicted.
–s , --seed	Optional: specify the seed for random sample split. (default seed = 42)
–k , --kfoldvalue	Optional: To adjust k-value for K-fold cross validation. (default = 5)
–v , --verbose	Optional: To print extra output and plots. (defaul = FALSE)
–h , --help	

```
Rscript reggae.r –i data.csv –m stepwise –y dG –r 0.8 –q –d –v
```



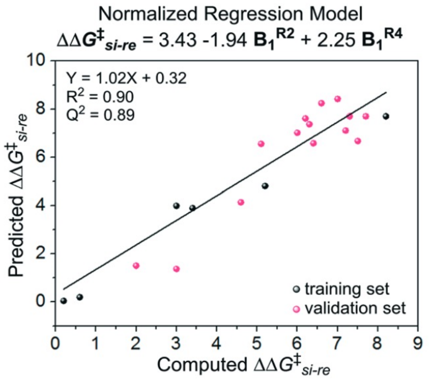
Predictive Model & Report

```
o Linear regression model with SELECTED features:
  Number of features (including response):  3

 a) scaled coefficients:
  y =  5.073  +  -1.602 * B1R2 + 1.720 * B1R4

 b) unscaled:
  y = 4.23 + -2.92 * B1R2 + 1.61 * B1R4

  R2-train =  0.94
  adj R2-train =  0.93
   RMSE-train =  0.58
```

MV Linear Regression Model: R-squared = 0.94
dGcalc - predicted
0.0
2.5
5.0
7.5
0
2
4
6
8
dGcalc - actual
Test
Train
$q - Kfold^2 = 0.92(0.01)$
$q - Test^2 = 0.8$
y = 5.07 + -1.60 * B1R2 +
1.72 * B1R4

Example: Quantitative Structure–Selectivity Relationships

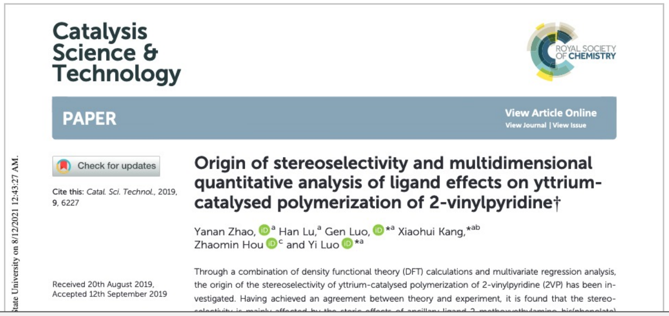
Volume 9 Number 22 21 November 2019 Pages 6201–6534
Catalysis Science & Technology
rsc.li/catalysis
R
R'
ISSN 2044-4761
ROYAL SOCIETY OF CHEMISTRY
Celebrating IYPT 2019
PAPER
Gen Luo, Xiaohui Kang, Yi Luo et al.
Origin of stereoselectivity and multidimensional quanti
ligand effects on yttrium-catalysed polymerization of 2



Total of 21 catalysts

Steric and electronic features:

- Sterimol descriptors: B_1, B_5, L and %Vbur
- Natural Population Analysis (NPA) charges



Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Catal. Sci. Technol.*, 2019, 9, 6227

Origin of stereoselectivity and multidimensional quantitative analysis of ligand effects on yttrium-catalysed polymerization of 2-vinylpyridine†

Yanan Zhao, [a] Han Lu,[a] Gen Luo, *[a] Xiaohui Kang,*[ab] Zhaomin Hou [c] and Yi Luo *[a]

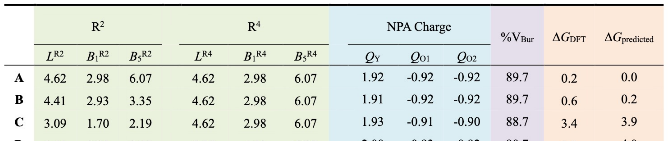
Received 20th August 2019, Accepted 12th September 2019

Through a combination of density functional theory (DFT) calculations and multivariate regression analysis, the origin of the stereoselectivity of yttrium-catalysed polymerization of 2-vinylpyridine (2VP) has been investigated. Having achieved an agreement between theory and experiment, it is found that the stereo-

State University on 8/12/2021 12:43:27 AM.

<u>**For statistical reproducibility:**</u>

- Available dataset (ESI: **.xlsx, .csv**, *table*)
- Available code
- Labeled split datasets or methods used



	R²			R⁴			NPA Charge			%V$_{Bur}$	ΔG_{DFT}	Δ$G_{predicted}$
	L^{R2}	B_1^{R2}	B_5^{R2}	L^{R4}	B_1^{R4}	B_5^{R4}	Q_Y	Q_{O1}	Q_{O2}			
A	4.62	2.98	6.07	4.62	2.98	6.07	1.92	-0.92	-0.92	89.7	0.2	0.0
B	4.41	2.93	3.35	4.62	2.98	6.07	1.91	-0.92	-0.92	89.7	0.6	0.2
C	3.09	1.70	2.19	4.62	2.98	6.07	1.93	-0.91	-0.90	88.7	3.4	3.9

```
Rscript reggae.r –i 2vp-labeled.csv –b B1R2,B1R4 –y dGcalc –r 0 –q –d –v
```

Zhao, Y.; Lu, H.; Luo, G.; Kang, X.; Hou, Z.; Luo, Y. *Catal. Sci. Technol.*, **2019**, *9*, 6227–6233.
Krafczyk, M. S.; Shi, A.; Bhaskar, A.; Marinov, D. Stodden, V. *Phil. Trans. R. Soc. A.* **2021,** *379.*