BIOMATH 2013
International Conference on Mathematical
Methods and Models in Biosciences
Sofia, Bulgaria, 16-21 June 2013

# Quantitative Structure-Activity Relationships: Linear Regression Modelling and Validation Strategies by Example

Sorana D. Bolboacă & Lorentz Jäntschi

# Outline

- Objectives
- **Q**uantitative **S**tructure-**A**ctivity **R**elationships
- **L**inear **R**egression **M**odeling
  - Assumptions
  - Selection and Diagnostic
  - Validation
  - Predictive Power

# Objectives

- Exploratory data analysis on modeling (structure-activity) relationships: Linear Regression Models
- Beyond a LRM model
  - why … assumptions … selection … validation …
- Model by examples

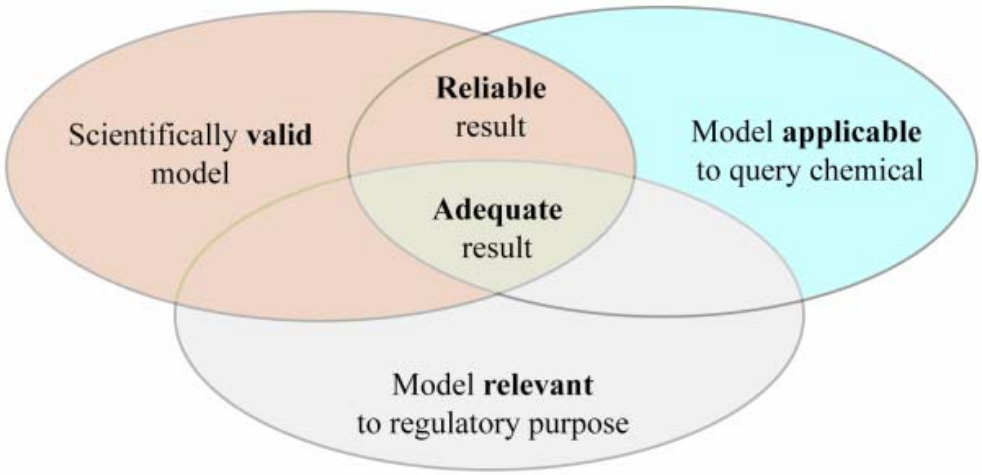
# QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

- = mathematical models linking chemical structure and pharmacological activity/property in a quantitative manner for a series of compounds [1]
- **Why** QSARs?
  - Identification and development of a new active compound is an extremely expensive and difficult process without a guaranteed result [2] (reflected in time and costs; often requires years before testing the new compound in human subjects)
  - ~ 90% of the initial candidates fail to be produces due to their toxicological properties [3].
  - The time needed to develop a drug varies from 10 to 15 years (Congressional Budget Office, Research and Development in the Pharmaceutical Industry (Washington, DC: CBO, October 2006).

1. Hammett LP. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. J Am Chem Soc 1937;59(1):96-103
2. Chen X-P, Du G-H. Target validation: A door to drug discovery. Drug Discov Ther 2007;1(1):23-29.
3. H. van de Waterbeemd and E. Gifford. Admet in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2003;2(3):192-204.

# QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

- Input data:
  - Chemical structure → structural information (descriptors – many approaches)
  - Activity/property (outcome variable)
- Output: model (regression vs. classification)
- Used to:
  - Supplement experimental data
  - Replace testing

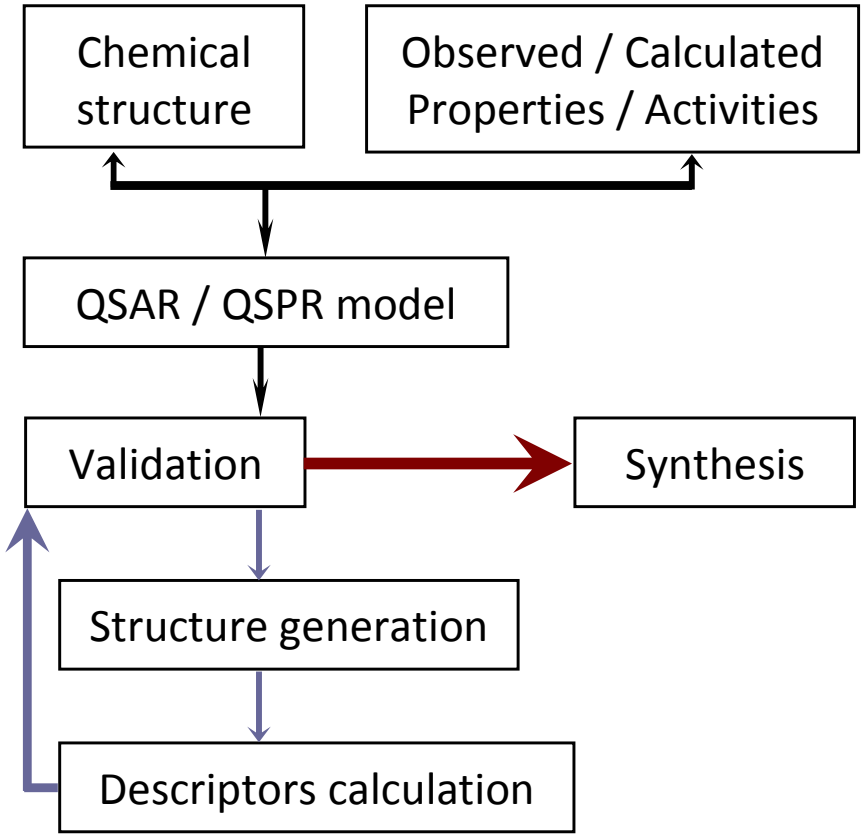


The interrelated concepts of (Q)SAR validity, reliability, applicability and adequacy

- Support priority setting of chemicals
- Guide experimental design (which?)
- Provide mechanistic information

- Group active compounds into chemical categories
- Data gaps for classification, labeling, risk assessment

# Quantitative Structure-Activity Relationships




- Assumptions:
  - structure of chemical compounds contains features responsible for its physical, chemical and/or biological properties
  - *similar compounds have similar properties* [4]



4. Johnson AM, Maggiora GM. Concepts and Applications of Molecular Similarity. New York: John Willey & Sons, 1990.

# QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

- OECD - Quantitative Structure-Activity Relationships Project - Guidance Document on the Validation of (Q)SAR Models [5]
- Principles:
  - a defined endpoint
  - an unambiguous algorithm
  - a defined domain of applicability
  - *appropriate measures of goodness-of-fit, robustness and predictivity*
  - a mechanistic interpretation, if possible


5. Organisation for Economic Co-operation and Development [online] [Accessed February 2, 2012] URL: http://www.oecd.org/officialdocuments/displaydocumentpdf?cote=env/jm/mono(2007)2&doclanguage=en

# LINEAR REGRESSION

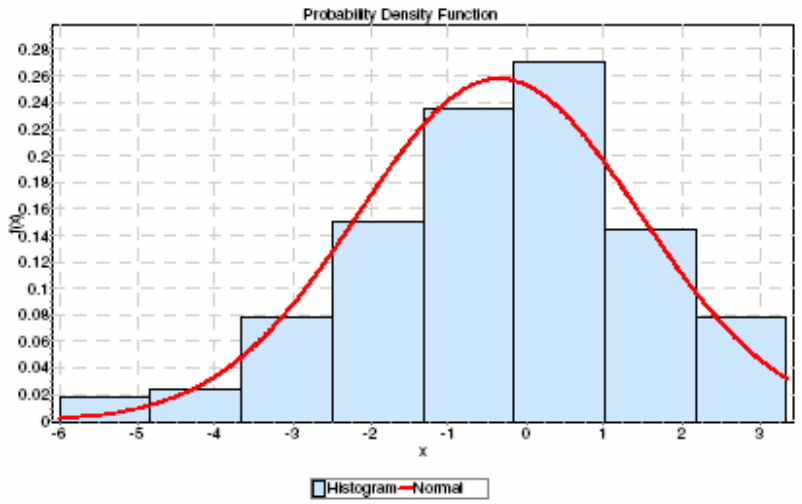
Objectives of linear regression analysis [6]:

- *to describe* - strength of the association between outcome and factors of interest
- *to adjust* - data for covariates or cofounders
- *to identify predictors* - factors that affect the outcome
- *to predict the outcome*

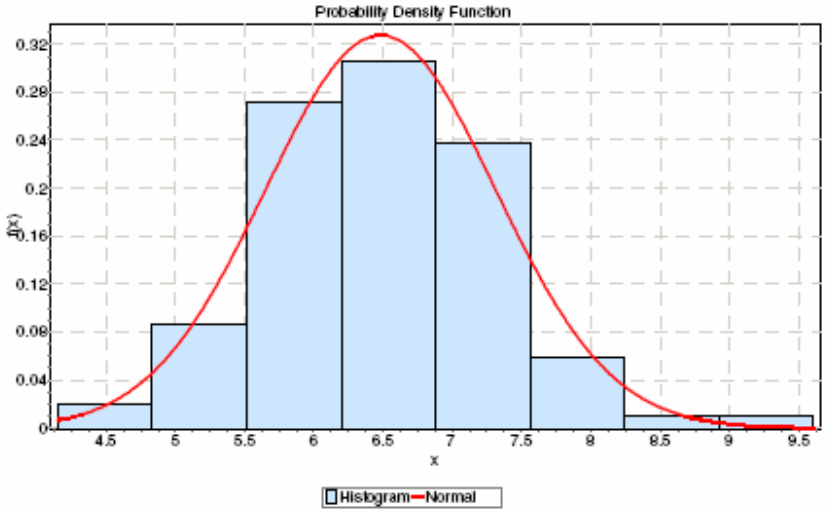
| Glaton (1886) [7] | to understand heredity |
|---|---|
| Pearson (1896) [8] | optimum values of slope and correlation coefficient could be calculated from the product-moment |
| Yule (1897) [9] | minimizing the sum of squares error |


6. Chan YM. Biostatistics 201: Linear Regression Analysis. Singapore Medical Journal 2004;45(2):55.
7. Galton F. Regression towards mediocrity in hereditary stature. J Anthropol Inst Great Brit Ireland 1886;15: 246-263.
8. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia, Proc R Soc Lond 1896 ;187:253-318.
9. Yule GU. On the significance of Bravais' formulae for regression, &c, in the case of skew correlation. Proc R Soc Lond 1897;60:477-489.

# LINEAR REGRESSION MODELING: ASSUMPTIONS




(Duchowicz et al., 2008) - N = 166




(Jäntschi et al, 2009) - N = 206

| Statistic | Value | Probability of observation | Reject the hypothesis of normality |
|---|---|---|---|
| Kolmogorov-Smirnov | 0.05508 | 67.43% | No |
| Anderson-Darling | 0.56539 | 14.1%; 12.5%; 14.3% | No |
| Chi Squared | 3(df=7) | 88.6% | No |
| Wilks-Shapiro | 0.98173 | 2.8% | Yes |
| $Z_{Skewness}$ | -2.58 | 1% | Yes |
| $Z_{Kurtosis}$ | 0.53 | 59.5% | No |
| Jarque-Bera | 6.61 | 3.7% | Yes |

| Statistic | Value | Probability of observation | Reject the hypothesis of normality |
|---|---|---|---|
| Kolmogorov-Smirnov | 0.03348 | 96.91% | No |
| Anderson-Darling | 0.44432 | 27.2%; 25.2%; 19.2% | No |
| Chi Squared | 11(df=7) | 13.8% | No |
| Wilks-Shapiro | 0.98709 | 5.8% | No |
| $Z_{Skewness}$ | 1.48 | 14% | No |
| $Z_{Kurtosis}$ | 2.51 | 1.2% | Yes |
| Jarque-Bera | 7.577 | 2.3% | Yes |

- But ... "normal law ... is not valid in a great many cases which are both common and important" [9]



10. Jäntschi L, Bolboacă SD. Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. UASVM Horticulture, 2009;66(2):691-697.

# DATA SET BY EXAMPLE

- Endocrine disrupting chemicals with experimental values of relative binding affinity expressed in logarithmic scale (logRBA) [11]
- → binders: weak (logRBA < -2.0), moderate (-2.0 ≤ logRBA ≤ 0), strong (logRBA > 0) [12]

| Set | n | Type of binder: n (% [95%CI]) | | |
|---|---|---|---|---|
| | | weak | moderate | strong |
| Training | 132 | 60 (45 [34; 55]) | 41 (31 [24; 39]) | 30 (23 [16; 31]) |
| Test | 23 | 3 (13 [5; 36]) | 16 (70 [48; 87]) | 4 (17 [5; 39]) |
| External | 9 | 4 (44 [12; 77]) | 5 (56 [23; 88]) | 0 (0 [0; 43]) |


11. Li J, Gramatica P. Mol Divers. 2010 Nov;14(4):687-96.
12. Blair RM , Fang H, Branham WS, Hass BS, Dial SL, Moland CL, et al. Toxicol Sci 2000;54:138-153

# LINEAR REGRESSION MODELING: ASSUMPTIONS

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i + \varepsilon$$

| Assumption | What is the effect? | How to detect it? | How to fix it? |
|---|---|---|---|
| Normality | Unreliable coefficients and confidence intervals | Plot: normal probability plot<br>Statistics: skewness & kurtosis [22]<br>Test[c]: Kolmogorov-Smirnov [23], [24], Anderson-Darling [25], Chi-Squared [26]; Shapiro-Wilks test [27] (n < 50) | Identify and withdrawn the outliers (if any) - Grubs test [28] |
| Linearity | Estimations and predictions are in error | Plot<br>▪ observed vs estimated values<br>▪ residuals versus estimated values | Transformation<br>(see Table 2) |
| Independence | Important in models where time is important | Plot: autocorelation plot of residuals<br>Test: Durbin-Watson [a] [29], [30]. If no autocorrelation exists in the sample DW ~ 2 | D-W < 1.00 › structural problem › reconsider the transformation (if any).<br>Add more independent variables. |
| Homoscedasticity | Too wide or too narrow confidence intervals | Plot (pattern of errors): residuals vs predicted value<br>Test: Breusch-Pagan[b] [31], Bartlett [32], Levene [33] | Use different variables.<br>Use Generalized Least Square |
| Collinearity (independent variables) | Predictors are related to each other | ▪ Correlation matrix: r ? 0.80 or 0.90 indicates collinearity [34]<br>▪ VIF ? 10 and/or T(tolerance) < 0.01 indicates the existence of collinearity [34] | Remove the variable that is correlated with others<br>Be aware that collinearity is not bad all time |

[a] the errors are serially uncorrelated; WD ∈ [0, 4], DW = 2 › no autocorrelation; [b] the variance of the residuals is the same for all values of *Y*; [c] EasyFit program uses it to test the normality of *Y*;

# LRM: SELECTION

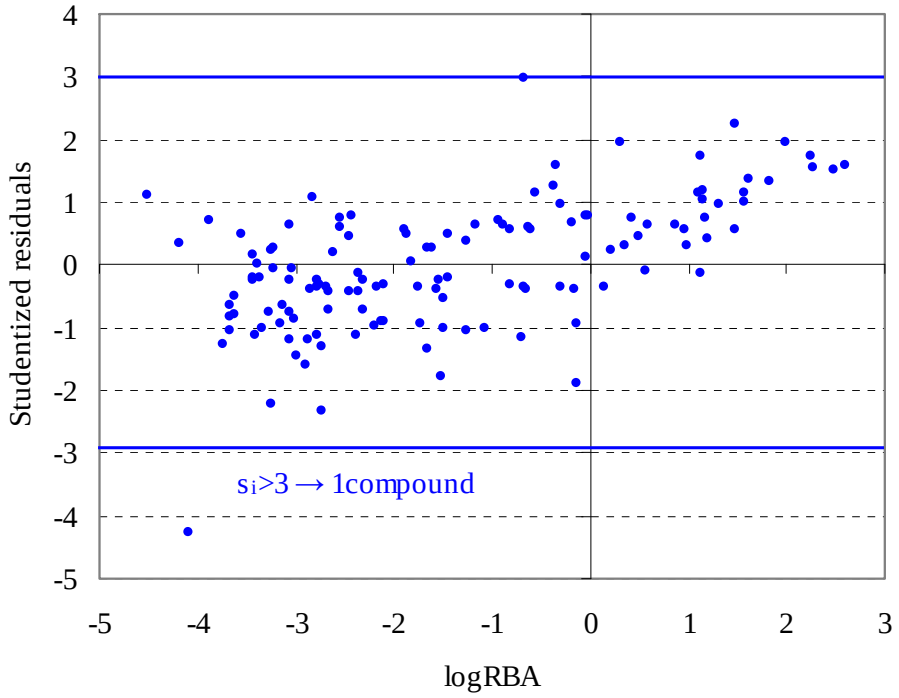
**Unusual data: not identify by usual parameter (r, F)**

- Outlier:
  - X's or Y
  - Regression outlier: ↑ |residuals|
- Leverage point: unusual combination of variables ($h_i$ model – threshold = ***2·(k+1)/n***)
- Influential point: influence on the regression coefficients ($D_i$ model – threshold = ***4/n*** )

- Neither ignore, nor throw them without thinking
- Think of reason why observation may be different
- Change the model
- Fit the model with and without the unusual data and see the effect

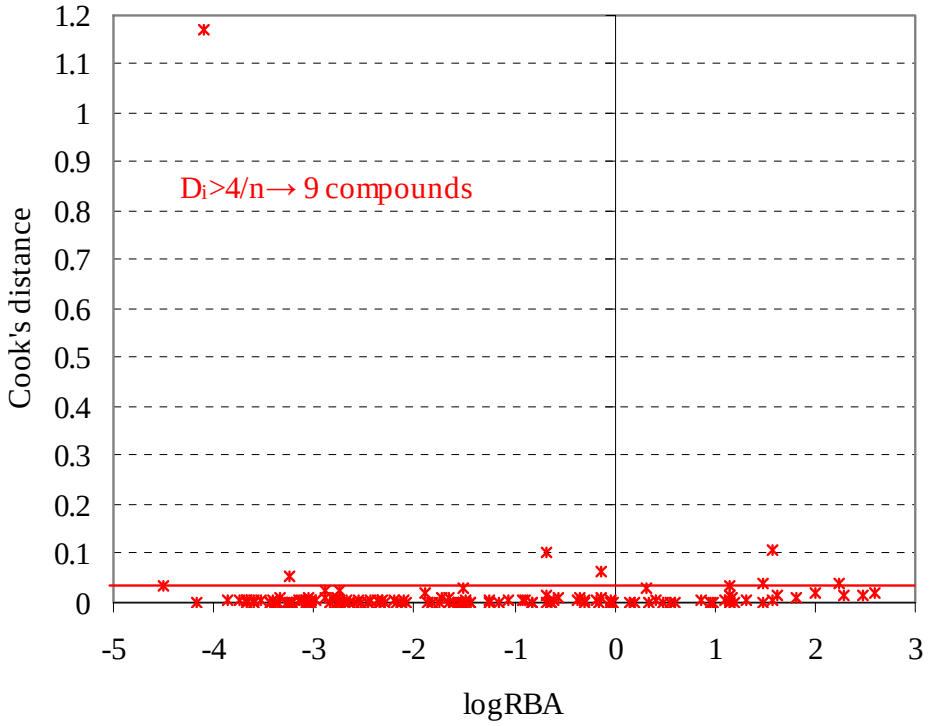
# LRM: SELECTION

## Studentized residuals (N=132)



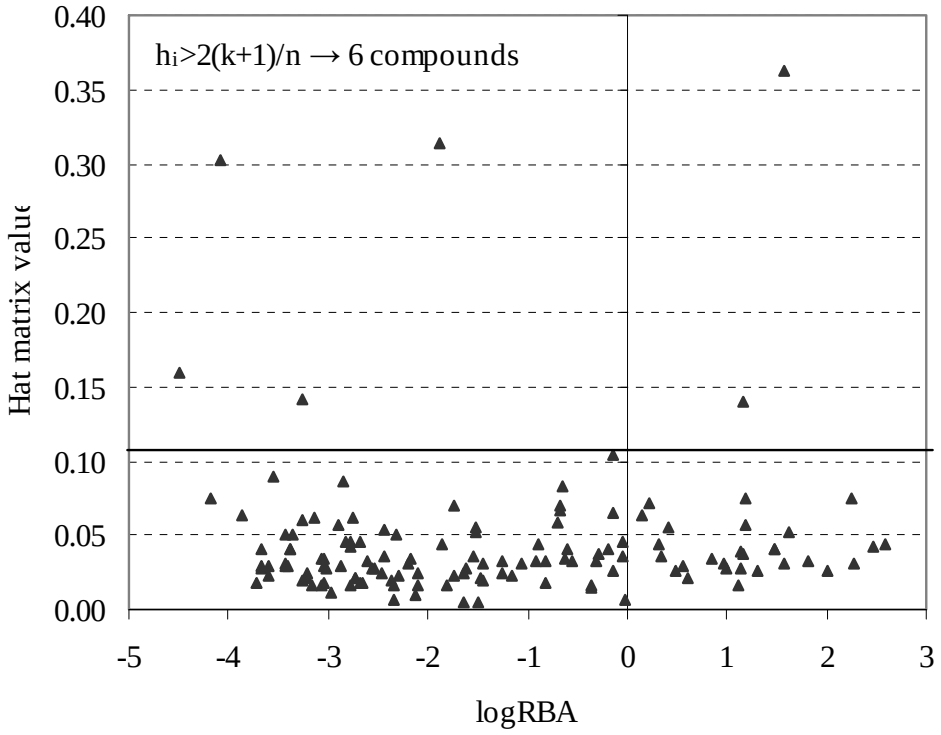
# LRM: SELECTION

## Cook's distance (N=132)



# LRM: SELECTION

## Hat matrix (N=132)



# LRM: DIAGNOSTIC

| Parameter (Abbreviation) | Formula [ref] | Remarks |
|---|---|---|
| Residual Mean Square (RMS)<br>- Error variance | $RMS = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$ | RMS: the smaller the better<br>$0 < RMS < \infty$ |
| Average Prediction Variance (APV) | $APV = \frac{RMS}{n} \cdot (n+k)$ [51] | The smaller the better |
| Total Squared Error (TSE) | $TSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} + 2 \cdot k - n$ [52]<br>$TSE = \frac{SSE}{MSE} - (n - 2 \cdot k) + 2$ [39] | The smaller the better<br>$TSE > (k+1) \rightarrow$ bias due to incompletely specified model<br>$TSE < (k+1) \rightarrow$ the model is over specified (contains too many variables) |
| Average Prediction Mean Squared Error (APMSE) | $APMSE = \frac{RMS}{n-k-1}$ [53] | The smaller the better |
| Mean Absolute Error (MAE)<br>- Measures the average magnitude of the errors; could be also used to compare two models | $MAE = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert}{n}$ | $MAE = 0 \rightarrow$ perfect accuracy<br>$0 < MAE < \infty$ |
| Root Mean Square Error (RMSE):<br>- Measures the average magnitude of the error | $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ | $RMSE > MAE \rightarrow$ variation in the errors exists<br>$0 < RMSE < \infty$ |
| Mean Absolute Percentage Error (MAPE)<br>- Measure of accuracy expressed as percentage | $MAPE = \frac{\sum_{i=1}^{n}\lvert (y_i - \hat{y}_i)/y_i \rvert}{n}$ [54], [55] | $MAPE \sim 0 \rightarrow$ perfect fit |
| Standard Error of Prediction (SEP) | $SEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-1}}$ | The smaller the better |
| Relative Error of Prediction (REP%) | $REP(\%) = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$ | The smaller the better |

n = sample size; k = number of independent variables in the model; $\bar{y}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the $i^{th}$ compound in the sample; $y_i$ = observed/measured activity/property of $i^{th}$ compound; SSE = sum of squared errors; MSE = mean of squared errors

# LRM: Predictive Power

| Parameter (Abbreviation) | Formula [ref] | Remarks |
|---|---|---|
| Predictive Squared Correlation Coefficient in Training Set ($Q_{F1}^2$) | $Q_{F_1}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}} (y_i - \bar{y}_{TR})^2}$ [56] | Prediction is considered accurate if the predictive power of the model is > 0.6 [57] |
| Predictive Squared Correlation Coefficient in Test Set ($Q_{F2}^2$) | $Q_{F_2}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TS}} (y_i - \bar{y}_{TS})^2}$ [58] | |
| External Predictive Ability ($Q_{F3}^2$) | $Q_{F_3}^2 = 1 - \frac{\sum_{i=1}^{n_{TS}} (\hat{y}_i - y_i)^2 / n_{TS}}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2 / n_{TR}}$ [59] | |
| Predictive Power (PP): Fisher's approach | $t = \frac{\overline{res}_{TS} - 0}{stdev(res_{TS}) / \sqrt{n_{TS}}}$ [60]<br>$p = TDIST(abs(t), n_{TS}-1, 1)$ | Evaluate if the mean of residual is statistically different by the expected value (0) |

n = sample size; v = number of independent variables in the model; $\bar{y}$ = the mean of estimated/predicted activity/property; $\hat{y}_i$ = predicted value of the $i^{th}$ compound in the sample; $y_i$ = observed/measured activity/property of $i^{th}$ compound; $\overline{res}$ = mean of residuals; stdev = standard deviation; TR = training set; TS = test set; EXT = external set; abs = absolute value

# LINEAR REGRESSION MODELING: log(RBA)

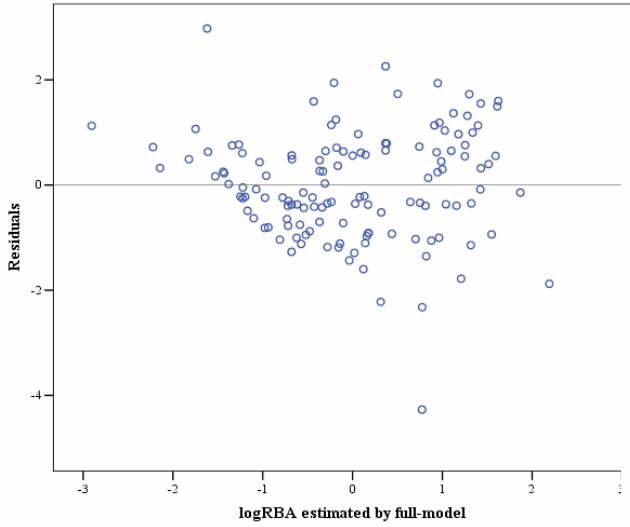
| **Statistical parameter** | **Full-model (n=132)** | **$D_i$-model (n=115)[a]** | **$h_i$-model (n=123)[b]** |
|---|---|---|---|
| Normality tests: KS-AD-CS | 0.116* - 2.409* - 14.862** | 0.124* - 2.432* - 12.613* | 0.120* - 2.428* - 12.083* |
| Durbin-Watson | 1.275 | 1.292 | 1.263 |
| Collinearity: highest R<br>higher VIF & lower T | 0.7700<br>TIE: 3.367& 0.297 | 0.7889<br>ATS4m: 4.082&0.245 | 0.7752<br>ATS4m: 4.516&0.221 |
| $R^2$ | 0.6559 | 0.7797 | 0.6928 |
| $R^2_{adj}$ | 0.6394 | 0.7675 | 0.6769 |
| $s_{est}$ | 1.0701 | 0.8293 | 0.9977 |
| F-value (p-value) | 39.711 ($9.89 \cdot 10^{-27}$) | 63.721 ($3.12 \cdot 10^{-33}$) | 43.59 ($1.62 \cdot 10^{-27}$) |
| $Q^2$ | 0.5832 | 0.7543 | 0.6497 |
| $s_{loo}$ | 1.1827 | 0.8764 | 1.0668 |
| $F_{loo}$-value (p-value) | 28.74 ($9.49 \cdot 10^{-22}$) | 55.17 ($1.85 \cdot 10^{-31}$) | ($1.62 \cdot 10^{-27}$) |
| $\|R^2-Q^2\|$ | 0.0727 | 0.0254 | 0.0431 |
| $C_p$-statistic | 7.00 | 7.00 | 7.00 |
| AIC ($w_i$-AIC) | 18.9639 (0.2856) | 18.3078 (0.3965) | 18.7490 (0.3180) |
| $AIC_{R2}$ ($w_i$- $AIC_{R2}$) | 8.0504 (0.3137) | 7.7421 (0.3659) | 8.0077 (0.3204) |
| $AIC_c$ ($w_i$- $AIC_c$) | 1.2657 (0.2990) | 0.7766 (0.3819) | 1.1358 (0.3191) |
| BIC | 52.0750 | 9.8317† | 33.1255 |
| HQC | 26.2887 | 34.7113† | 7.8043 |
| FIT | 1.3058 | 2.3097 | 1.5076 |

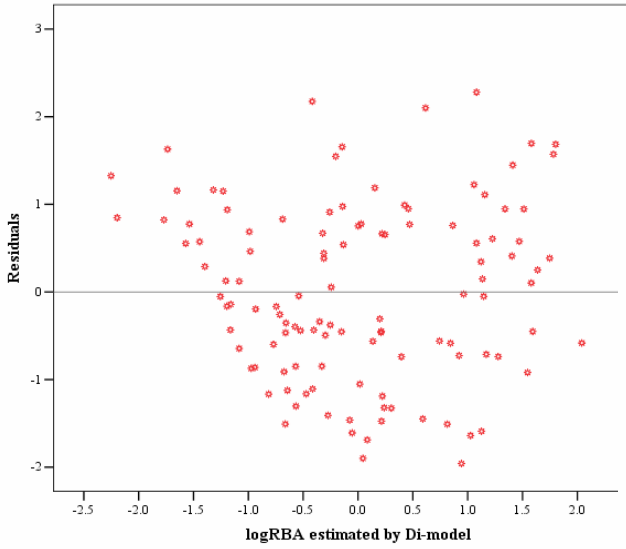
* $p \geq 0.05$; ** $p = 0.0378$; † = absolute values; KS = Kolmogorow-Smirnov; AD = Anderson Darling; CS = Chi-Squared; R = correlation coefficient; VIF = Variance Inflation Factor; T = tolerance; $R^2$ = determination coefficient; $R^2_{adj}$ = adjusted determination coefficient; $s_{est}$ = standard error of the estimate; F-value = Fisher's statistics; $Q^2$ = determination coefficient in leave-one-out analysis; $s_{loo}$ = standard error of the predict; Cp-statistic = Mallows' statistic; AIC = Akaike's information criterion; $AIC_{R2}$ = AIC based on the determination coefficient; $AIC_c$ = AIC corrected by McQuarrie and Tsai; BIC = Bayesian Information Criterion; HQC = Hannan-Quinn Criterion; FIT = Kubinyi's function;

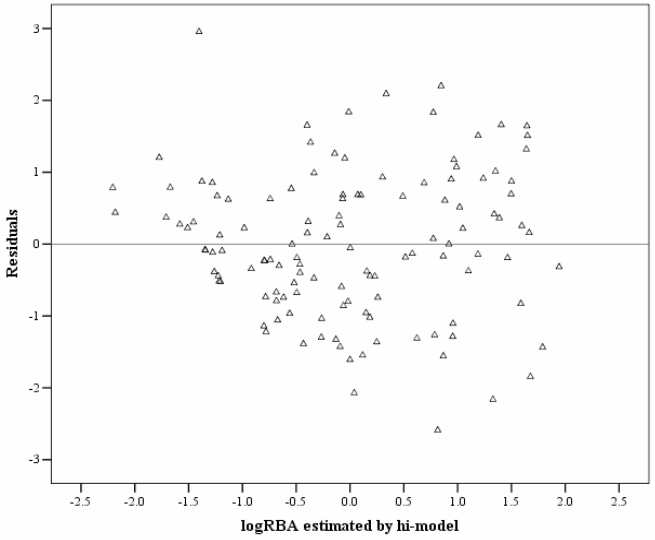
[a] 56 weak binders, 35 moderate binders, and 24 strong binders; withdrawn (16 compounds): 4 weak binders, 6 moderate binders and 6 strong binders;

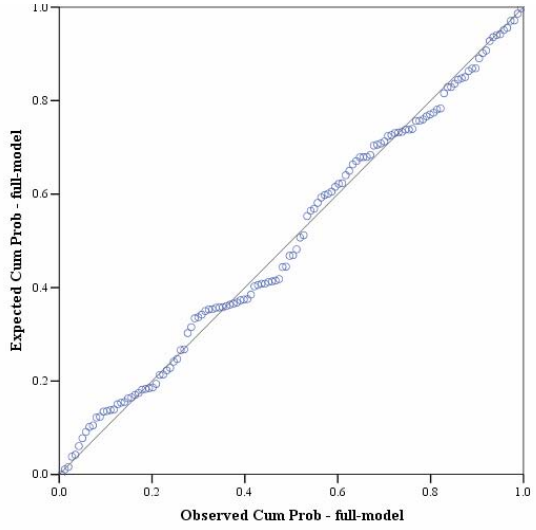
[b] 57 weak binders, 38 moderate binders, and 28 strong binders; withdrawn (8 compounds): 3 weak binders, 3 moderate binders and 2 strong binders;

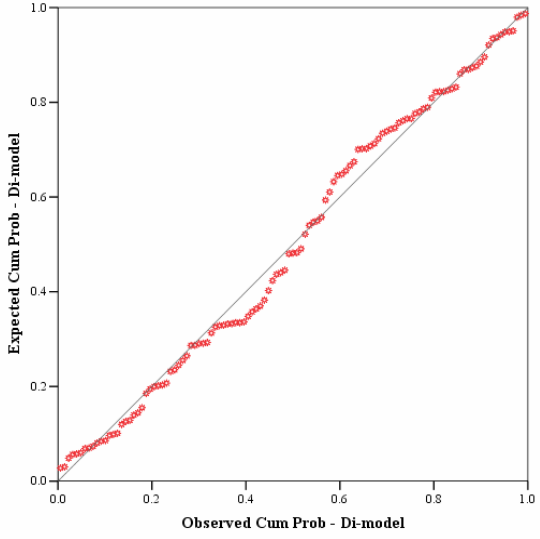
# LINEAR REGRESSION MODELING: log(RBA)

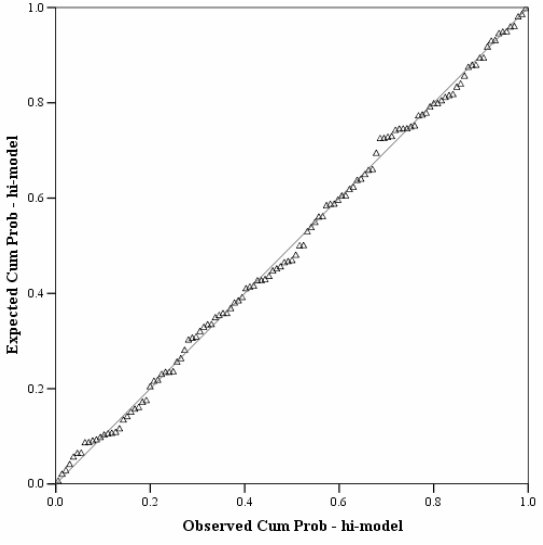


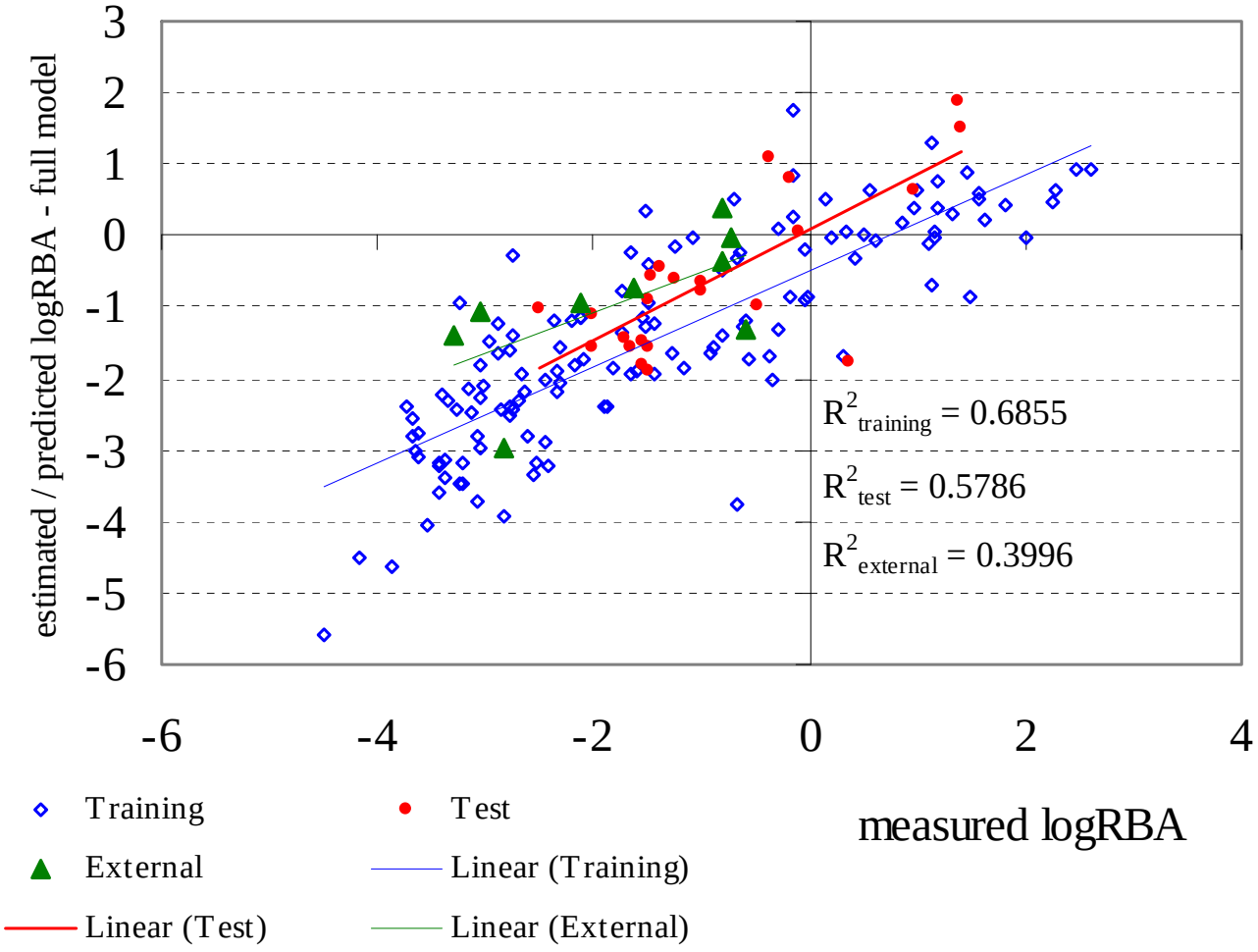



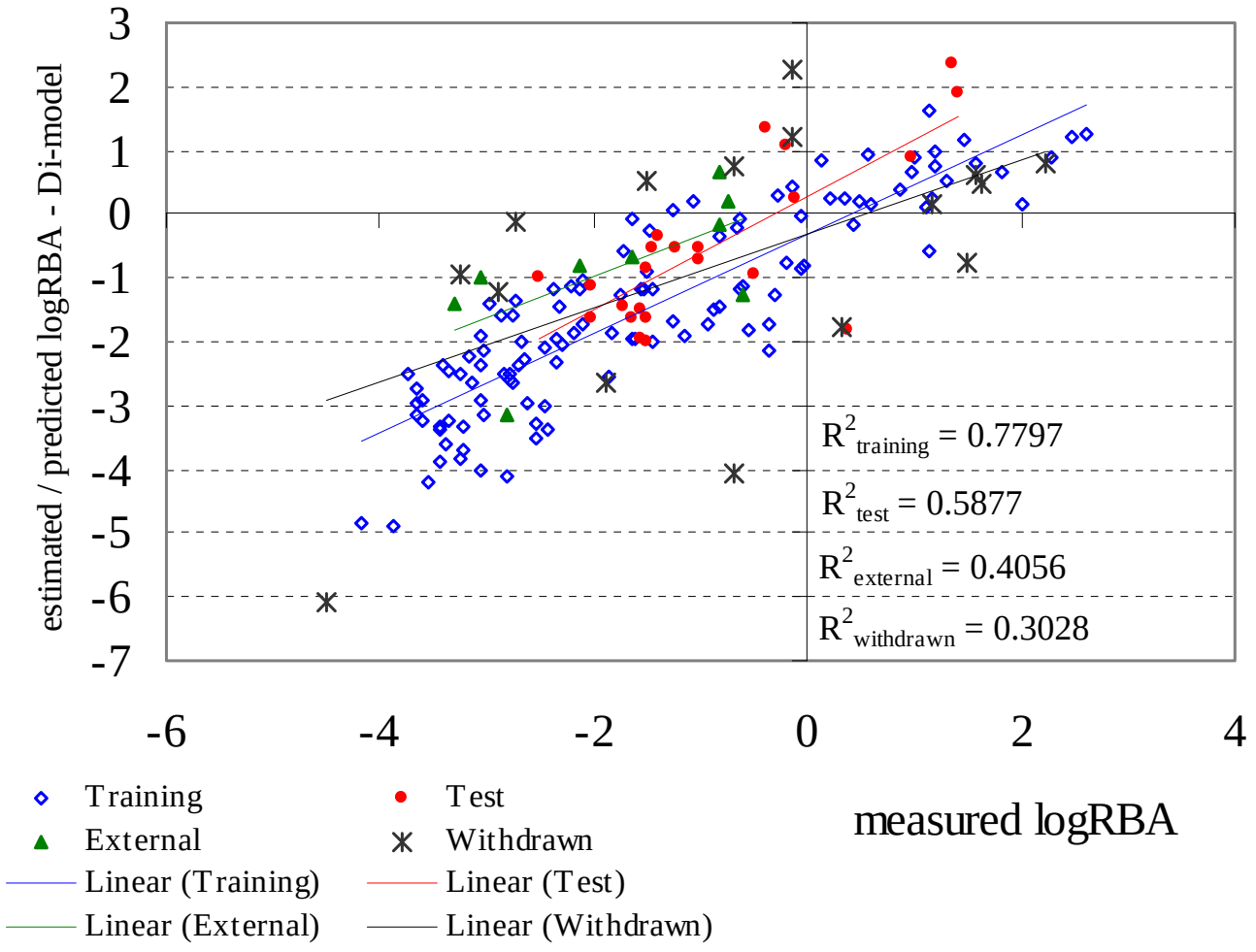
# LINEAR REGRESSION MODELING: log(RBA)

| Parameter (Abbreviation) | Full-model (n=132) | $D_i$-model (n=115) | $h_i$-model (n=123) |
|---|---|---|---|
| Residual Mean Square (RMS) | 1.1361 | **0.6815** | 0.9870 |
| Average Prediction Variance (APV) | 1.1877 | **0.7170** | 1.0351 |
| Total Squared Error (TSE) | **7.0000** | **7.0000** | **7.0000** |
| Average Prediction Mean Squared Error (APMSE) | 0.0091 | **0.0063** | 0.0085 |
| Mean Absolute Error (MAE) | 0.8356 | **0.6812** | 0.7827 |
| Root Mean Square Error (RMSE): | 1.0414 | **0.8037** | 0.9689 |
| Mean Absolute Percentage Error (MAPE) | 1.3033 | **1.0797** | 1.1649 |
| Standard Error of Prediction (SEP) | 1.0453 | **0.8072** | 0.9729 |
| Relative Error of Prediction (REP%) | 73.9756 | **58.0395** | 70.9144 |

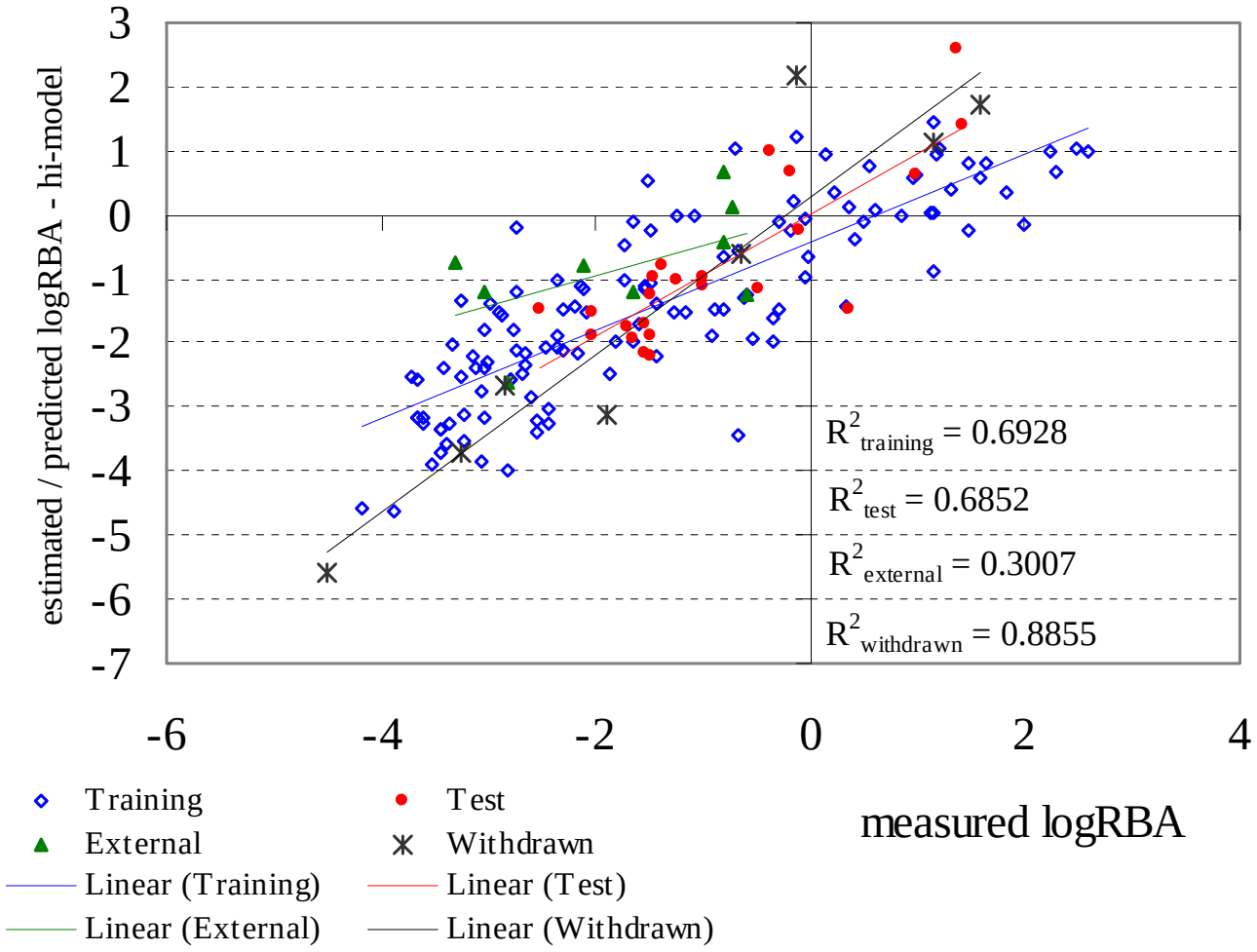
# LINEAR REGRESSION MODELING: log(RBA)



# LINEAR REGRESSION MODELING: log(RBA)



# LINEAR REGRESSION MODELING: log(RBA)

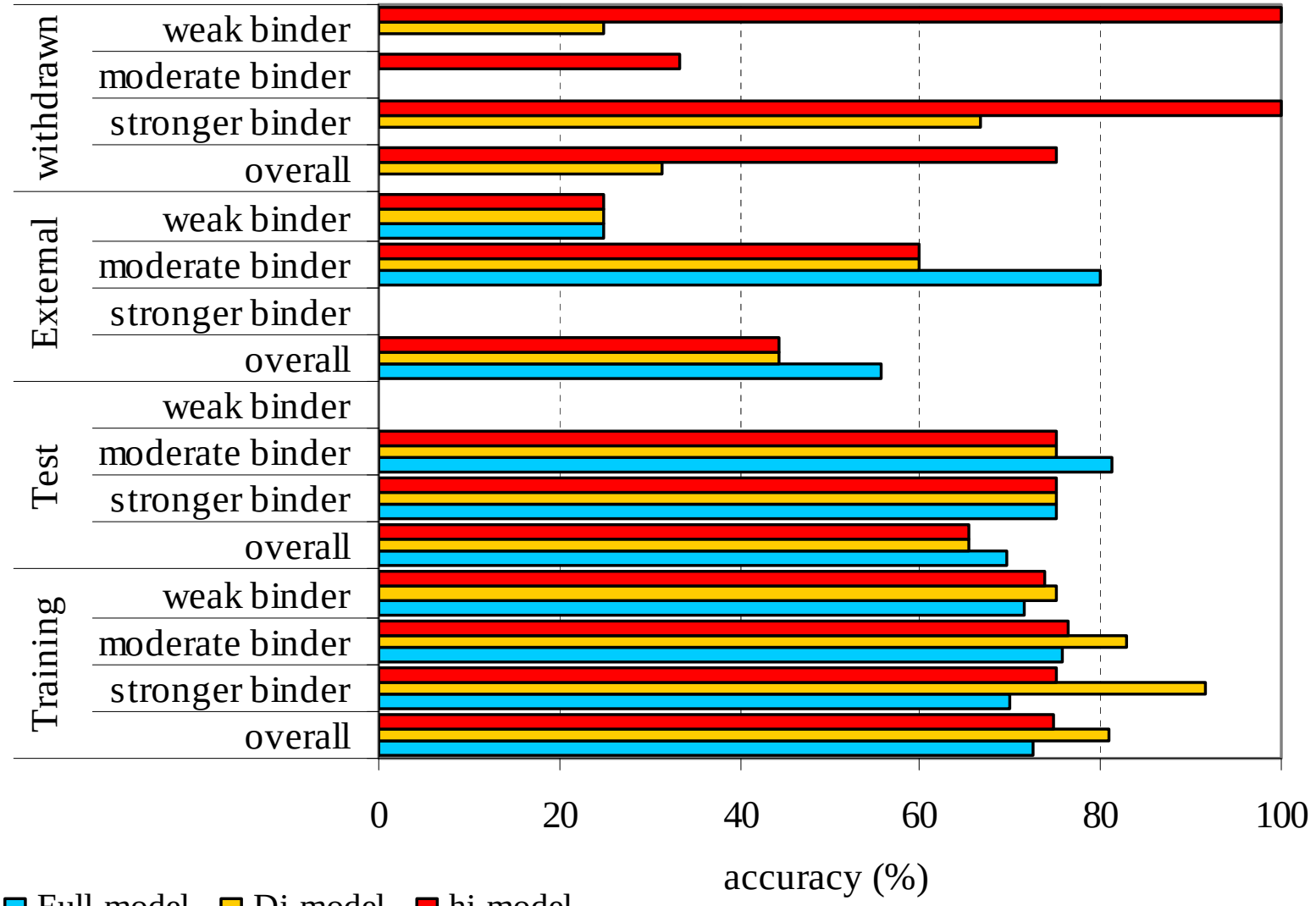


# LINEAR REGRESSION MODELING: log(RBA)

| Criterion | Full-model (n=132) | | $D_i$-model (n=115) | | | $h_i$-model (n=123) | | |
|---|---|---|---|---|---|---|---|---|
| | test[a] | external[b] | test[a] | external[b] | withdrawn[c] | test[a] | external[b] | withdrawn[d] |
| $Q_{F1}^2$ | 0.5498 | -0.1890 | 0.4796 | -0.4581 | 0.2009 | 0.6476 | -0.4444 | 0.7434 |
| $Q_{F2}^2$ | 0.4804 | 0.2010 | 0.3875 | 0.1450 | 0.0443 | 0.5738 | 0.1112 | 0.7431 |
| $Q_{F3}^2$ | 0.5527 | -16.3066 | 0.7809 | -17.6311 | -4.4056 | 0.7813 | -18.5792 | -2.9125 |
| PP (p) | -1.7852 (0.0440) | -2.8228 (0.0112) | -2.0961 (0.0239) | -3.0020 (0.0085) | 0.1039 (0.4593) | -0.4239 (0.3379) | -2.9139 (0.0097) | 0.0489 (0.4812) |

$Q^2_{F1}$ = predicted squared correlation coefficient in training set;
$Q^2_{F2}$ = predicted squared correlation coefficient in test set; $Q^2_{F3}$ = external predictivity ability; PP = predictive power;
PP = Predictive Power: Fisher's approach; [a] n=23; [b] n = 9; [c] n = 16; [d] n = 8

# LINEAR REGRESSION MODELING: log(RBA)



# SUMMARY

- Choosing a proper linear model is crucial in QSAR analysis: model able to predict accurately the activity of interest of new chemical compounds is desired under the hypothesis that changes in molecular structure directly reflect in the compound activity/property.
- Input data and data preparation for regression analysis are of great importance but these subjects were beyond the aim of the present paper.
- Regression analysis answer to the following questions: *Does the biological activity depend on structural information*? If so, *The nature of the relationship is linear*? If yes, *How good is the model in prediction of the biological activity of new compounds*?

# KEY MESSAGE

① test the assumption of linear regression (normality, linearity, independence, homoscedascity, and/or collinearity)

② construct the model(s) if assumptions are accomplished - analyze the data (choose the best performing model)

③ assess and diagnose the alternative models - analyze the LRM

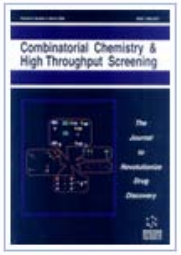
④ decide which model fit best to your objectives

# KEY MESSAGE

- Following these steps in linear regression analysis certainly led to a performing estimation model but
  - the model prediction power will always depend on the structure of compounds on which the model was obtained and on their biological activity

  and it will be proper to be applied on
  - similar compounds as structure ± activity/property

# Some Useful Articles



Combinatorial Chemistry & High Throughput Screening

**ISSN (Print):** 1386-2073
**ISSN (Online):** 1875-5402
**VOLUME:** 16
**ISSUE:** 4
**DOI:** 10.2174/138620731131604003 **Price:** $58

Back | Mark Item | Add to Cart

The Effect of Leverage and/or Influential on Structure-Activity Relationships

**Author(s):** Sorana D. Bolboaca and Lorentz Jantschi
Pages 288-297 (10)

Journal List > Int J Mol Sci > v.12(7); 2011 > PMC3155355

MDPI International Journal of MOLECULAR SCIENCES

Int J Mol Sci. 2011; 12(7): 4348–4364. PMCID: PMC3155355
Published online 2011 July 5. doi: 10.3390/ijms12074348

**Predictivity Approach for Quantitative Structure-Property Models. Application for Blood-Brain Barrier Permeation of Diverse Drug-Like Compounds**

Sorana D. Bolboacă[1] and Lorentz Jäntschi[2,*]

TheScientificWorldJOURNAL
Volume 9 (2009), Pages 1148-1166
doi:10.1100/tsw.2009.131

Abstract | Full-Text | How to Cite this Article

**Methods Paper**

**Comparison of Quantitative Structure-Activity Relationship Model Performances on Carboquinone Derivatives**

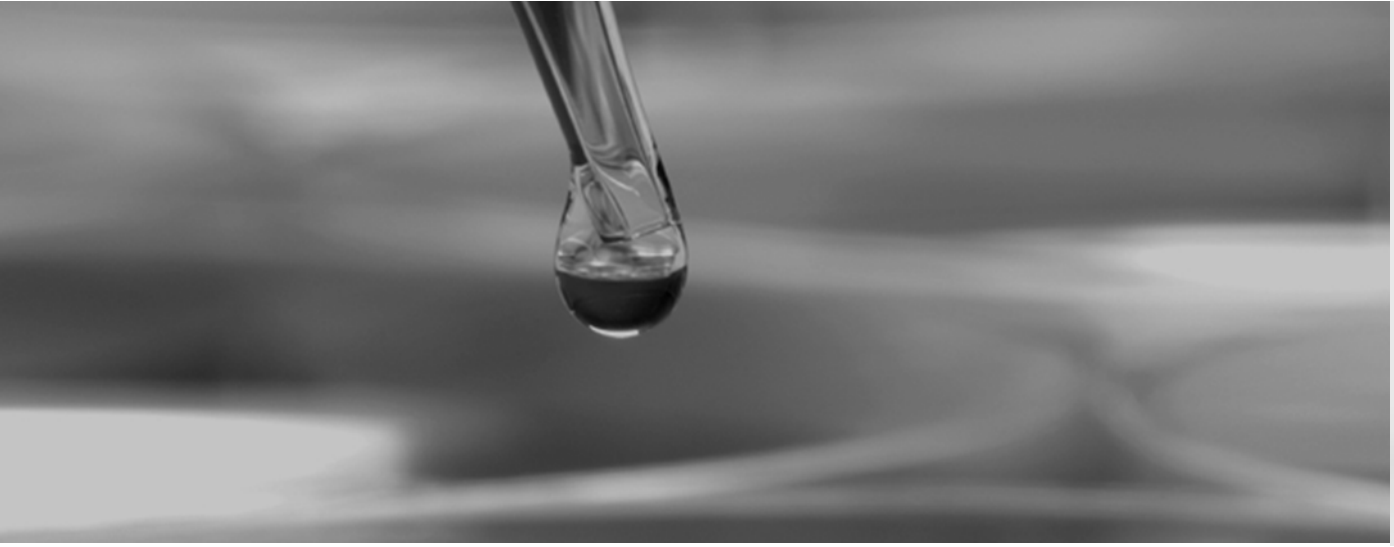
Sorana D. Bolboacă[1] and Lorentz Jäntschi[2]

Environmental Chemistry Letters
August 2008, Volume 6, Issue 3, pp 175-181

Modelling the property of compounds from structure: statistical methods for models validation

Sorana-Daniela Bolboacă, Lorentz Jäntschi

Download PDF (279 KB) | View Article



# THANK YOU FOR ATTENTION!