STATISTICS FOR QSAR MODELS VALIDATION

SORANA D. BOLBOACĂ, CARMEN E. STOENOIU,
AND LORENTZ JÄNTSCHI

General validation principles for quantitative structure-activity relationship (qSAR) models in the context of chemical regulation were developed [1] due to the importance and implication in of these methods in drug design. A brief analysis of different techniques used in validation of multiple linear regression models [2] is reviewed. The hierarchical steps in models validation are highlighted and a validation technique is proposed. The following statistical approaches are considered: correlation analysis (Pearson, Spearman, Kendall and Gamma coefficients as parameters and associated significance levels [3]), regression analysis (leave-one-out cross-validation and determination coefficients), and other inferential statistics (cross correlation coefficients, training vs. test experiment, correlated correlations analysis). The proposed statistical validation technique is exemplified on a qSAR model obtained by applying the molecular descriptors family on the structure-activity relationship approach [4]. Acknowledgments: UEFISCSU Romania supports the research through projects ID 458/2007 and ID 1051/2007.

References

[1] Gramatica P. QSAR Comb Sci 2007;26(5):694-701.
[2] Bolboacă SD, Jäntschi L. Env Chem Lett DOI: 10.1007/s10311-007-0119-9.
[3] Bolboacă SD, Jäntschi L. Leonardo J Sci 2006;9:179-200.
[4] Jäntschi L. Leonardo El J Pract Techol 2005;6:76-98.

Univ. of Medicine and Pharmacy Cluj-Napoca, 13 E Isac, 400023, Ro
E-mail address: sbolboaca@umfcluj.ro

Technical Univ. of Cluj-Napoca, 15 C-tin Daicoviciu, 400020, Ro
E-mail address: carmen@j.academicdirect.org

Technical Univ. of Cluj-Napoca, 15 C-tin Daicoviciu, 400020, Ro
E-mail address: lori@j.academicdirect.org

1991 *Mathematics Subject Classification.* 62J05 (Linear regression); 62H20 (Measures of association); 62J99 (cross correlation analysis).

Key words and phrases. MDF-SAR (Molecular Descriptors Family on the Structure-Activity Relationship); Models Assessment and Validation; Statistical Methods.

STATISTICAL METHODS FOR qSAR MODELS VALIDATION

Sorana D. BOLBOACA

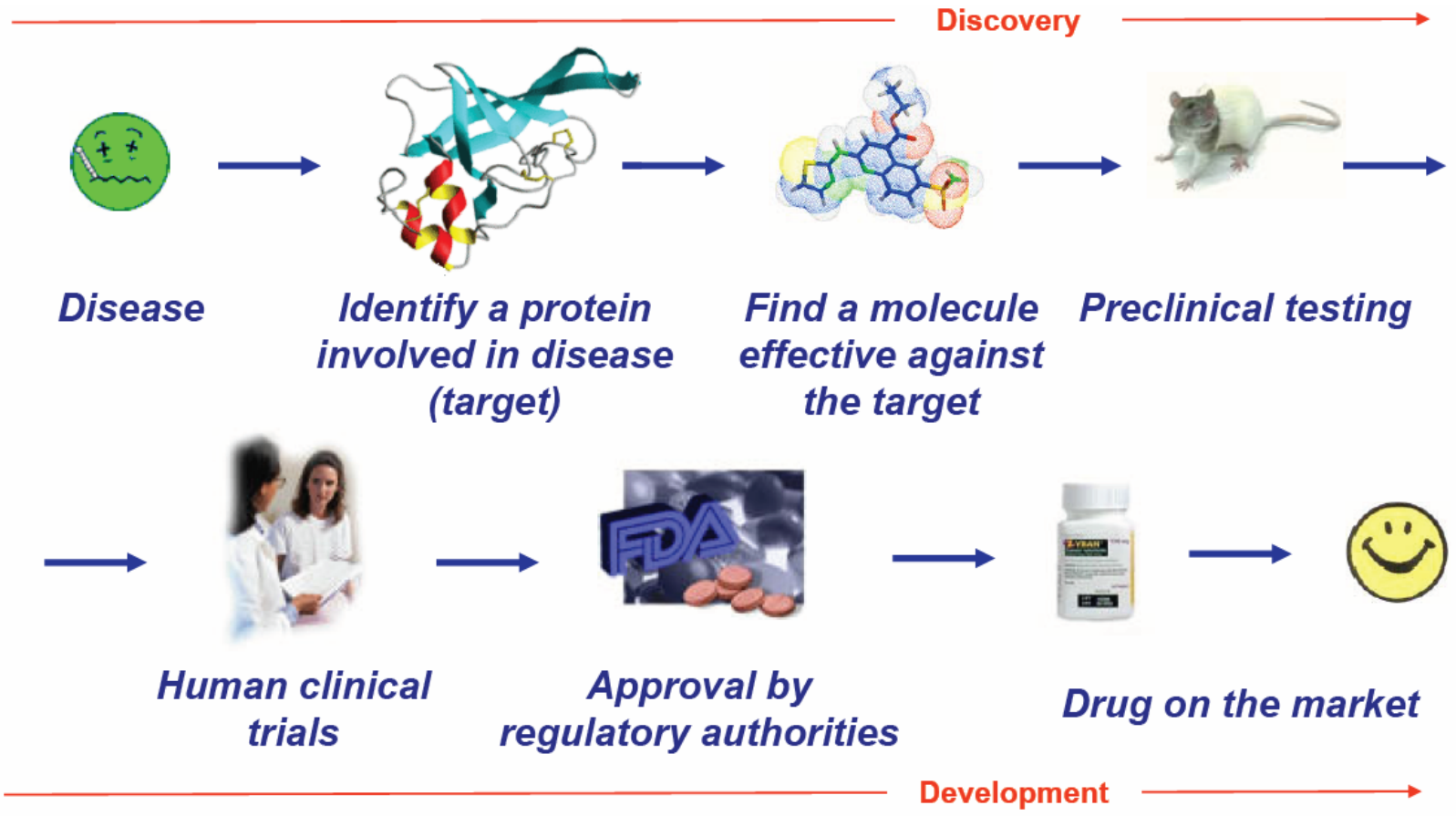
Carmen E. STOENOIU

Lorentz JÄNTSCHI

OUTLINE

- INTRODUCTION & BASIC ASSUMPTIONS
- DEFINITIONS
- STATISTICAL VALIDATION TECHNIQUES: MLR
  - CORRELATION ANALYSIS
  - LEAVE-ONE-OUT CROSS-VALIDATION
  - LEAVE-n-OUT CROSS-VALIDATION
  - TRAINING VERSUS TEST EXPERIMENT
  - CORRELATED CORRELATIONS
- CONCLUDING REMARKS

THE PATHWAY TO NEW DRUGS



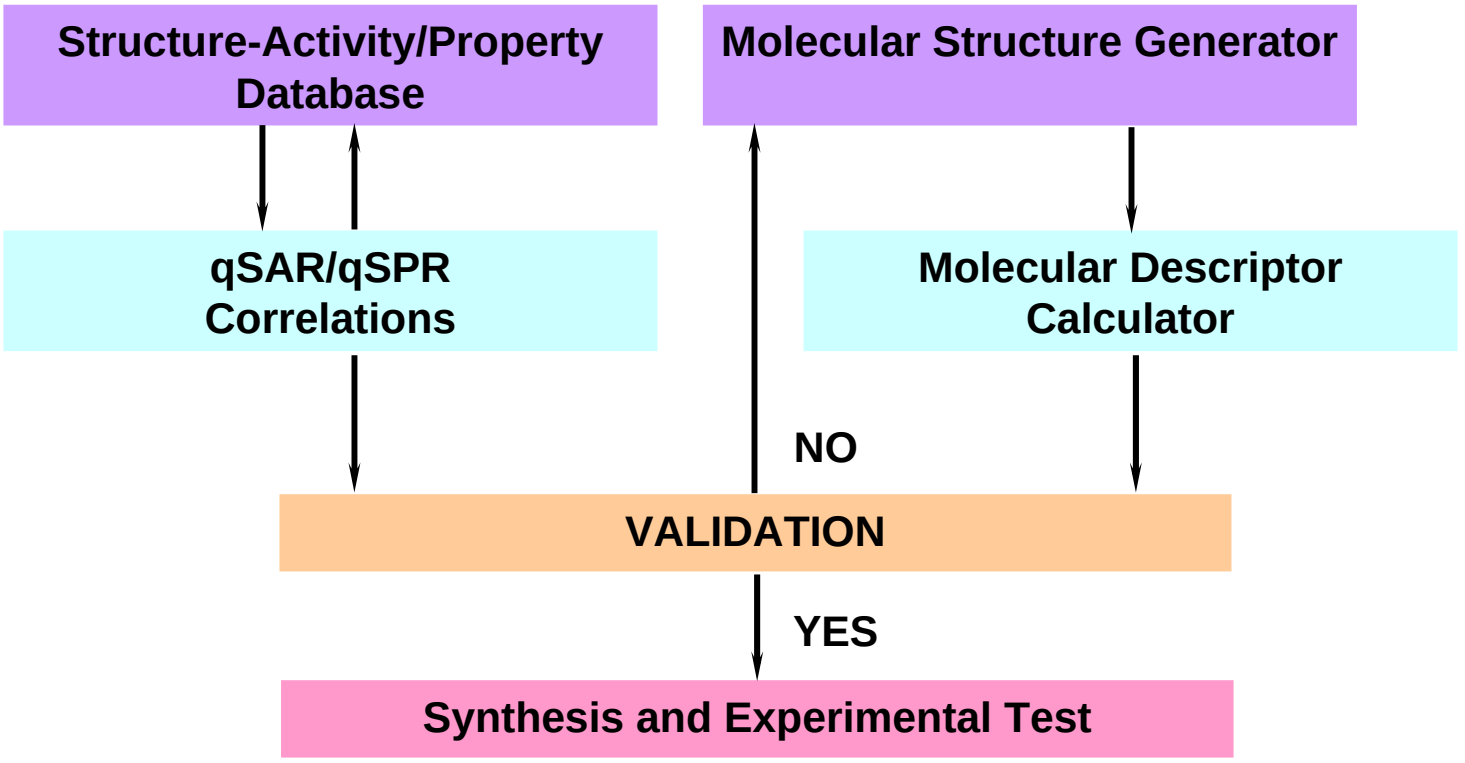
BASIC ASSUMPTIONS

- Molecular structures have:
  - different chemical molecular structures
  - different chemical properties/activities
- Similar molecular structures have similar molecular properties/activities

DEFINITIONS

- quantitative Structure-Property/Activity Relationships (qSAR - qSPR)
  - mathematical approaches of linking chemical structure and activity/property of chemical compounds in a quantitative manner (Hansch, 1969)
  - 1868 (Crum-Brown and Fraser): the activity of a compound is a function of its chemical composition and structure
  - 1893 (Richet and Seancs): shown for a set of organic molecules that the cytotoxicity was inverse related with water solubility
  - Hammett (1935) and Taft (1952) put together the mechanistic basis of QSAR/QSPR development

DEFINITIONS: qSAR/qSPR



DEFINITIONS

OECD Principles (2004):

1. A defined endpoit
2. Un unambiguous algorithm
3. A defined domain of applicability
4. Appropriate measures of goodness-of-fit, robustness and predictivity
5. A mechanistic interpretation

STATISTICAL VALIDATION TECHNIQUES

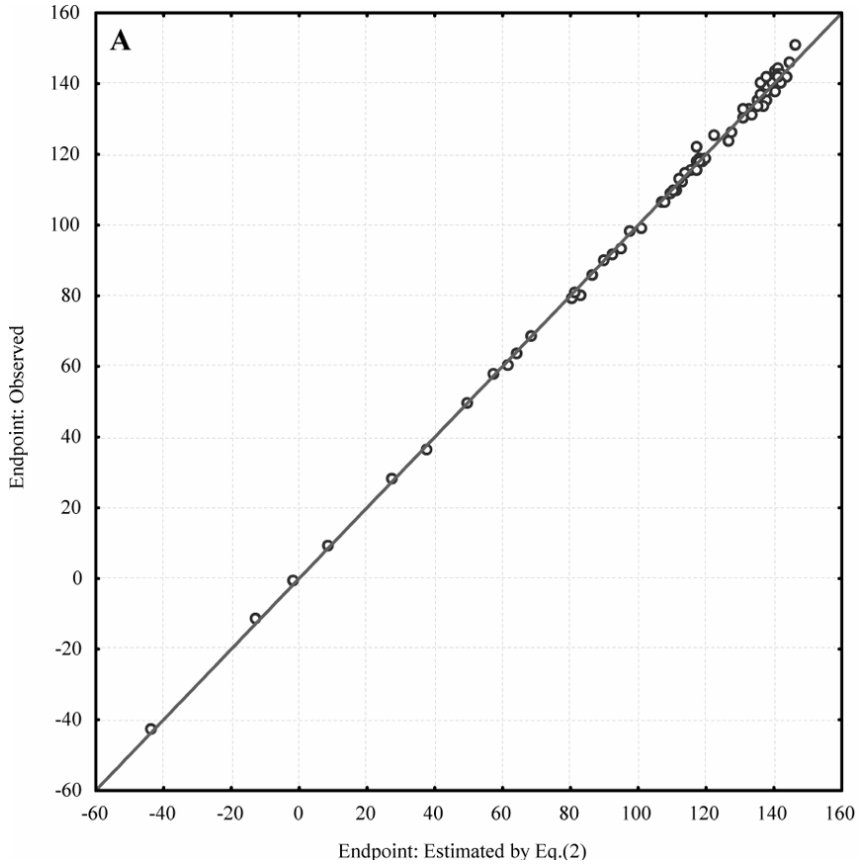
- Why?
  - CORRELATION ANALYSIS
  - LEAVE-ONE-OUT CROSS-VALIDATION
  - LEAVE-n-OUT CROSS-VALIDATION
  - TRAINING VERSUS TEST EXPERIMENT
  - CORRELATED CORRELATIONS

qSAR MODEL: REGRESSION

- Quantifies relationship between two variables (Y and X) with regression line:
  - Estimates average value of Y for any value of X
  - Assumes average of Y changes linearly with X
- Activity/Property = $a + c_1X_1 + c_2X_2 + \ldots$
  - a = intercept (Y_{mean} when X = 0)
  - c_i = regression coefficients
  - X_j = descriptor values

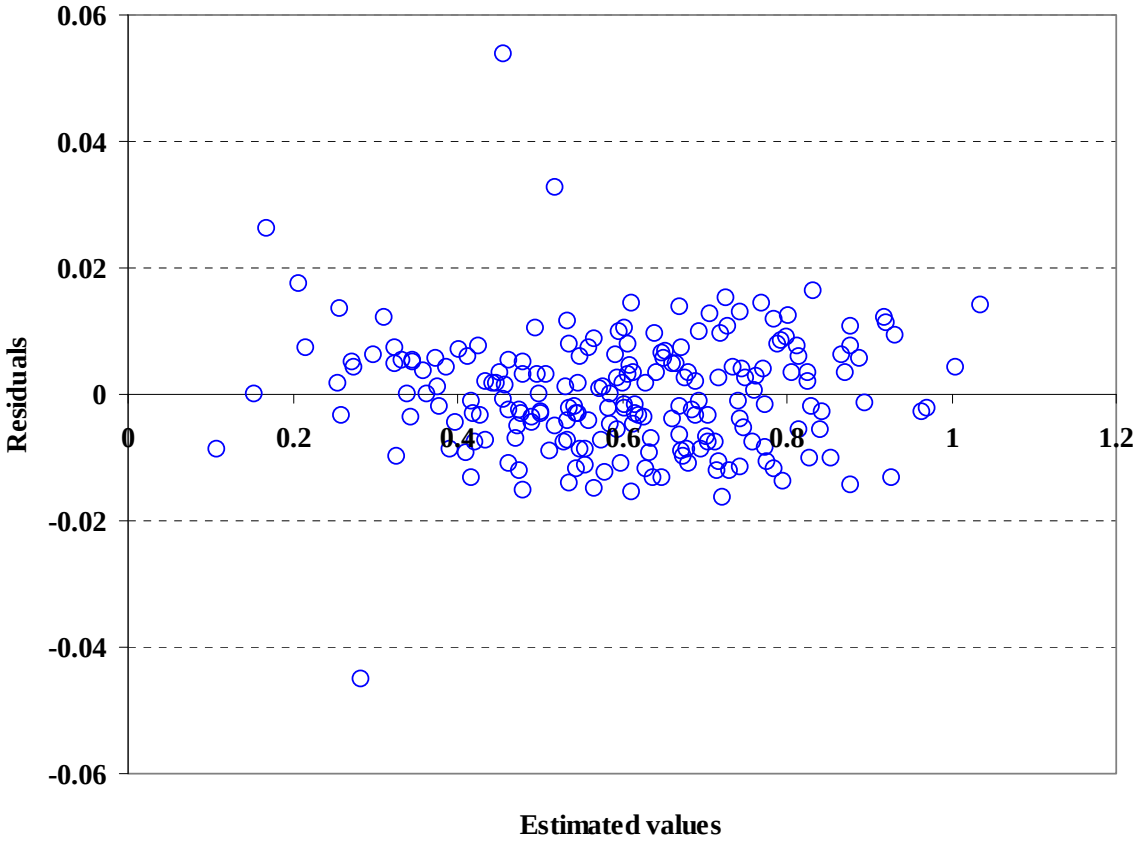
qSAR MODEL: REGRESSION

Bolboaca and Jäntschi, 2008 – *Environ. Chem. Letter*



qSAR MODEL: REGRESSION

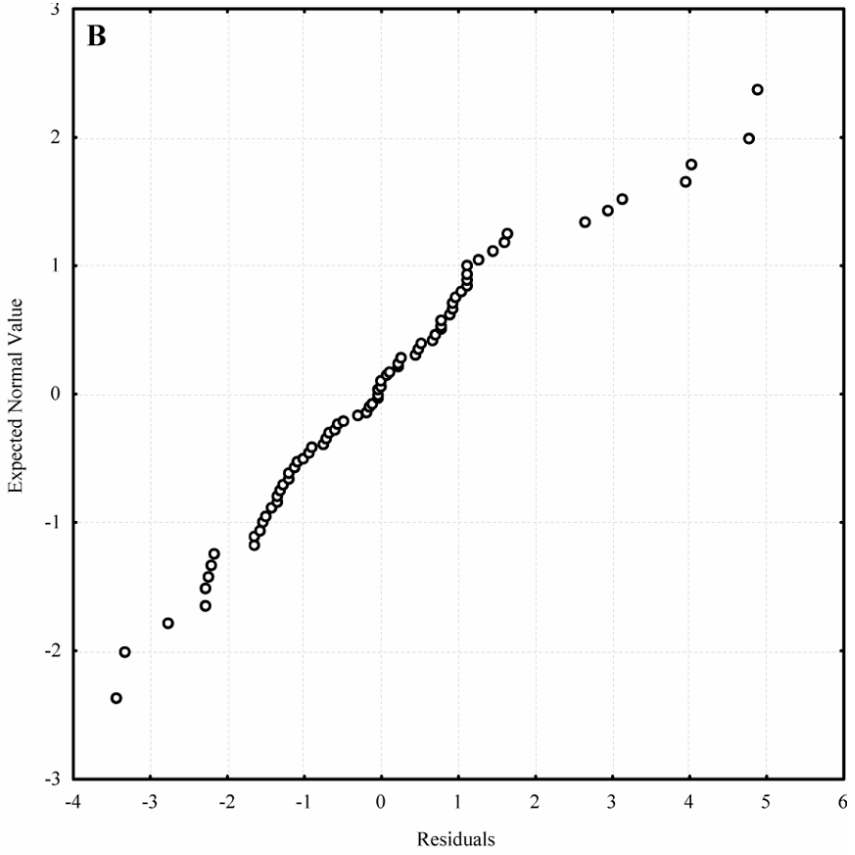
Residuals versus estimated values (Jäntschi et al., 2007)



qSAR MODEL: REGRESSION

Normal plot of residuals

(Bolboaca and Jäntschi, 2008 – *Environ. Chem. Letter)*



CORRELATION ANALYSIS

- Definition: correlation coefficient vs determination coefficient
- Types (Bolboaca and Jäntschi, 2006):
  - Pearson
  - Spearman
  - Semi-Quantitative
  - Kendall's tau (a, b, or c)
  - Gamma
- Implementation

  http://l.academicdirect.org/Statistics/linear_dependence/

CORRELATION ANALYSIS BY EXAMPLE

- Toxicity of polychlorinated organic compounds (Wei et al., 2001): $r^2 = 0.3660$
- Antimalarial activity of sulfonamide derivates (Agrawal et al., 2001): $r^2 = 0.7414$
- Inhibitory activity on carbonic anhydrase II of substituted 1,3,4-thiadiazole- and 1,3,4-thiadiazoline-disulfonamides (Supuran and Clare, 1999): $r^2 = 0.7190$
- Cytotoxicity of quinolines (Smith et al., 1997): $r^2 = 0.8700$
- Insecticidal activity of neonicotinoids (Hasegawa et al., 1999): $r^2 = 0.9850$

CORRELATION ANALYSIS BY EXAMPLE

Coefficient	Eq.(1)		Eq.(2)		Eq.(3)	
	r^2	t	r^2	t	r^2	t
Pearson	0.9694	29.77†	0.9712	30.71†	0.9850	42.86†
Spearman	0.6648	7.45†	0.7485	9.13†	0.8163	11.15†
Semi-Quantitative	0.8028	10.68†	0.8526	12.73†	0.8967	15.59†
	r^2	Z	r^2	Z	r^2	Z
Kendall τ_a	0.4084	4.96†	0.5079	5.53†	0.5931	5.98†
Kendall τ_b	0.4299	4.98†	0.5174	5.54†	0.6028	5.98†
Kendall τ_c	0.3816	4.81†	0.4746	5.35†	0.5542	5.78†
Gamma	0.4423	3.43†	0.5246	4.07†	0.6098	4.73†

Eq.(1) (Diudea et al., 2002);
Eq.(2) (Bolboaca and Jäntschi, 2006)
Eq.(3) (Bolboaca and Jäntschi, 2006)
r^2 = squared correlation coefficient; t = Student t parameter; Z = Z test parameter

CORRELATION ANALYSIS BY EXAMPLE

- Bolboaca at al., 2008 – Molecules

RFD			
Pearson	$r = 0.5520$	$3.35 \cdot 10^{-1}$	$t_{Prs,1} = 1.15$
Spearman	$\rho = 0.9000$	$3.74 \cdot 10^{-2}$	$t_{Spm,1} = 3.58$
Semi-Q	$r_{sQ} = 0.7049$	$1.84 \cdot 10^{-1}$	$t_{sQ} = 1.72$
Kendall tau-a	$\tau_{Ken,a} = 0.8000$	$5.00 \cdot 10^{-2}$	$Z_{Ken,\tau a} = 1.96$
Kendall tau-b	$\tau_{Ken,b} = 0.8000$	$5.00 \cdot 10^{-2}$	$Z_{Ken,\tau b} = 1.96$
Kendall tau-c	$\tau_{Ken,c} = 0.6400$	$1.17 \cdot 10^{-1}$	$Z_{Ken,\tau c} = 1.57$
Gamma	$\Gamma = 0.8000$	$1.17 \cdot 10^{-1}$	$Z_{\Gamma} = 1.57$

LEAVE-ONE-OUT CROSS-VALIDATION

- Employ *n* training sets and from each of these one compound is excluded
- Training set → model → predict$_{\text{excluded compound}}$
- $r^2_{\text{cv-loo}} - Q^2_{\text{LOO}}$
- Incomplete measure of a model's predictive power (Shao, 1993; Baumann and Stiefl, 2004)
- Inadequate for model predictivity assessment for completely new chemicals (Gramatica, 2007)
- Implementation:

  http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/

LEAVE-ONE-OUT CROSS-VALIDATION

- $r^2_{cv\text{-}loo} < 0.7$ (Gramatica, 2007)
- $r^2 - r^2_{cv\text{-}loo} < 0.3$ (Bolboaca and Jäntschi, 2008)

Example:

- $\hat{Y}_{4v} = 5.75 + 199 \cdot iSMmEQt - 9010 \cdot iSMMWHg - 0.071 \cdot LADmkQt + 2.86 \cdot INPRJQg$
- $n = 30$; $r = 0.994$, $r^2 = 0.988$; $F_{est} = 537$ ($p < 0.001$)
- $r_{cv\text{-}loo}{}^2 = 0.985$

LEAVE-ONE-OUT CROSS-VALIDATION BY EXAMPLE

Eq.	r^2	$r^2_{cv\text{-}loo}$	$F_{cv\text{-}loo}$	$r^2 - r^2_{cv\text{-}loo}$
2	0.9913	0.9908	7654†	0.0004
3	0.9982	0.9980	17562†	0.0002
† $p < 0.0001$				

LEAVE-n%-OUT CROSS-VALIDATION

External Validation:

- n% = 2/3·n
- 2/3·n → model
- model → 1/3·n

LEAVE-n%-OUT CROSS-VALIDATION BY EXAMPLE

- $\hat{Y}_{ChP} = -63.06 - 218.52 \cdot P(1/100) + 4.72 \cdot P(2/5) + 6.18 \cdot 10^{-3} \cdot P(77/24)$

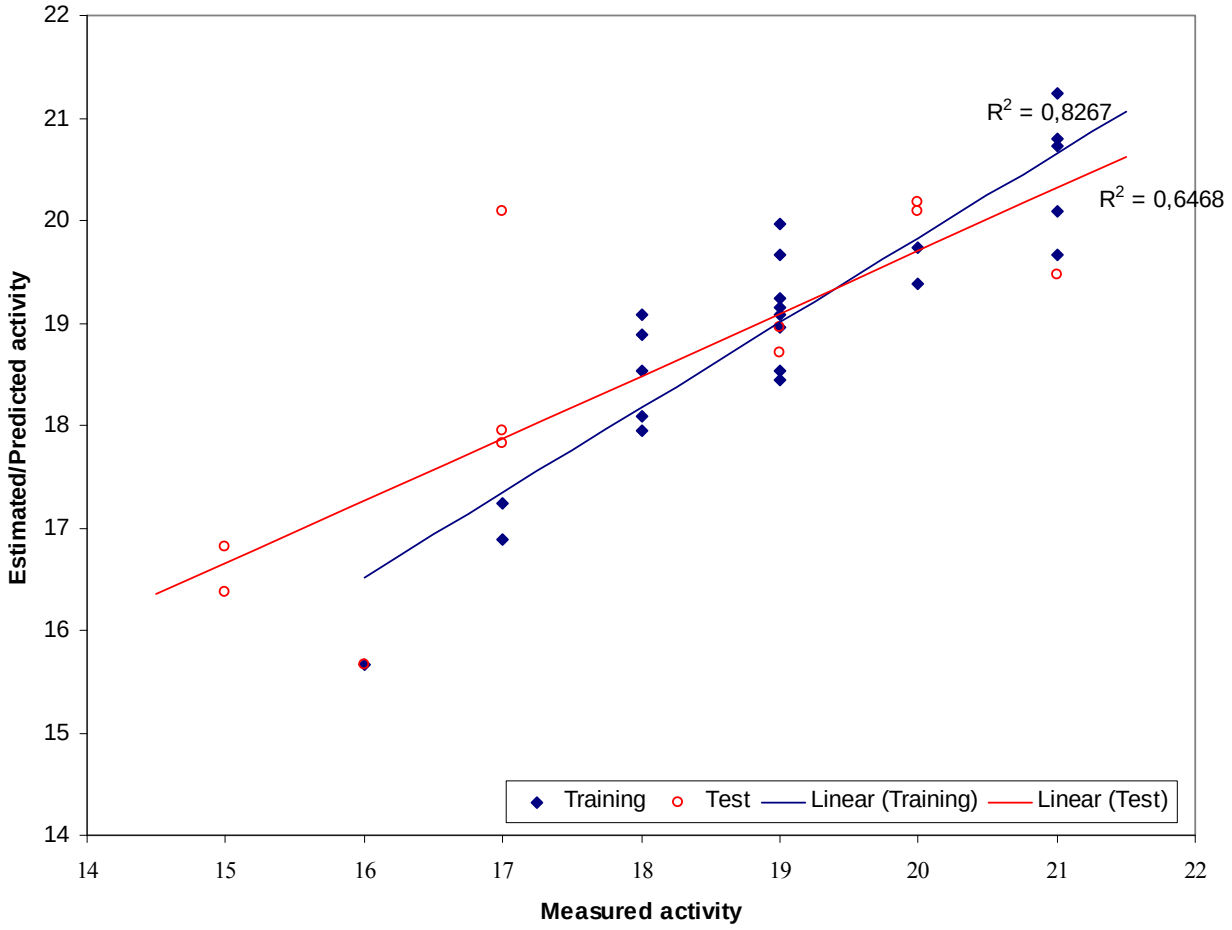
Training:

- $r_{tr} = 0.9092$ [0.7948-0.9611]
- $SErr_{tr} = 0.63$
- $F_{tr} = 30$ ($1.95 \cdot 10^{-7}$)

Test

- $r_{ts} = 0.8042$
- $F_{ts} = 16$ ($2.84 \cdot 10^{-3}$)

LEAVE-n%-OUT CROSS-VALIDATION BY EXAMPLE

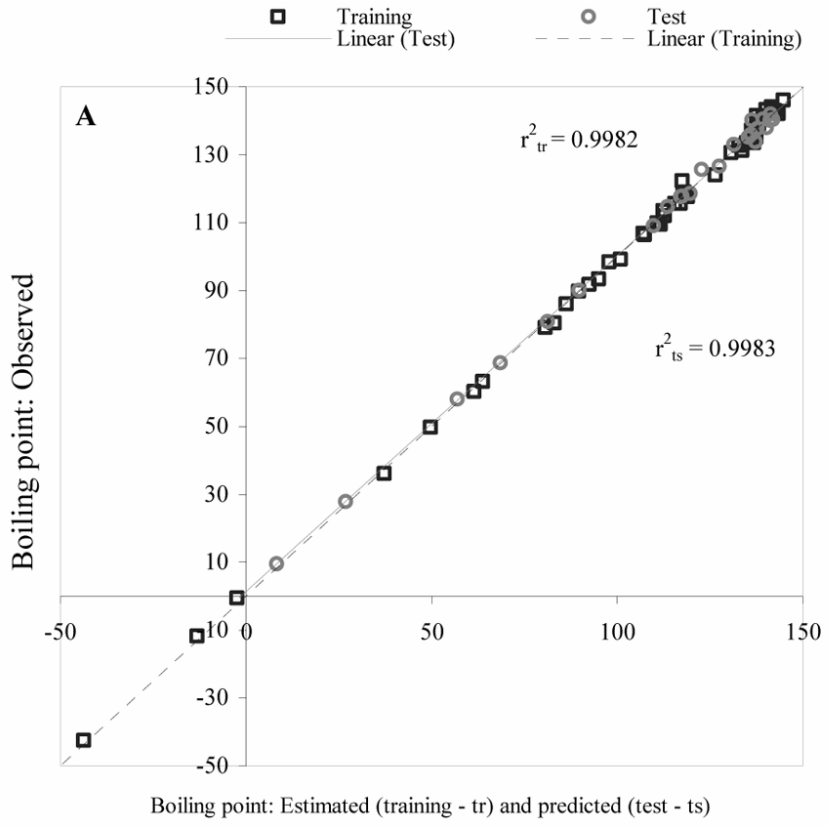


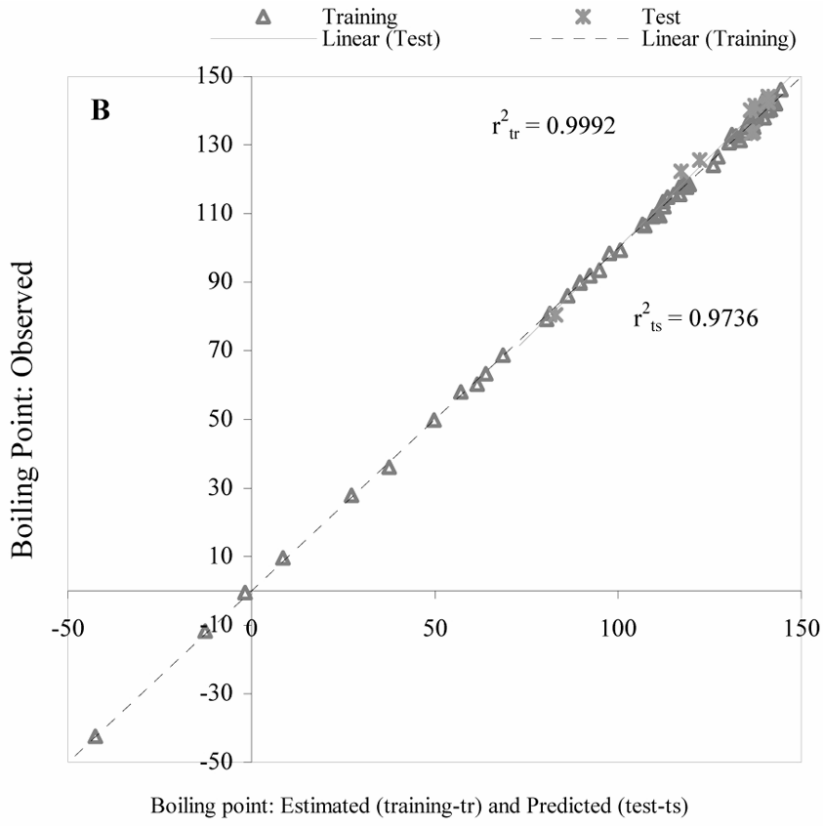
TRAINING VERSUS TEST EXPERIMENT

- A model is obtained for a sample of compounds
- Splitting the sample into training and test sets
- Training set → model → test set
- Model valid and stable:
  - r^2_{tr} is not statistically different by r^2_{ts}
- Implementation:

http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/qsar_qspr_s/

TRAINING VS TEST EXPERIMENT BY EXAMPLE





TRAINING VS TEST EXPERIMENT BY EXAMPLE

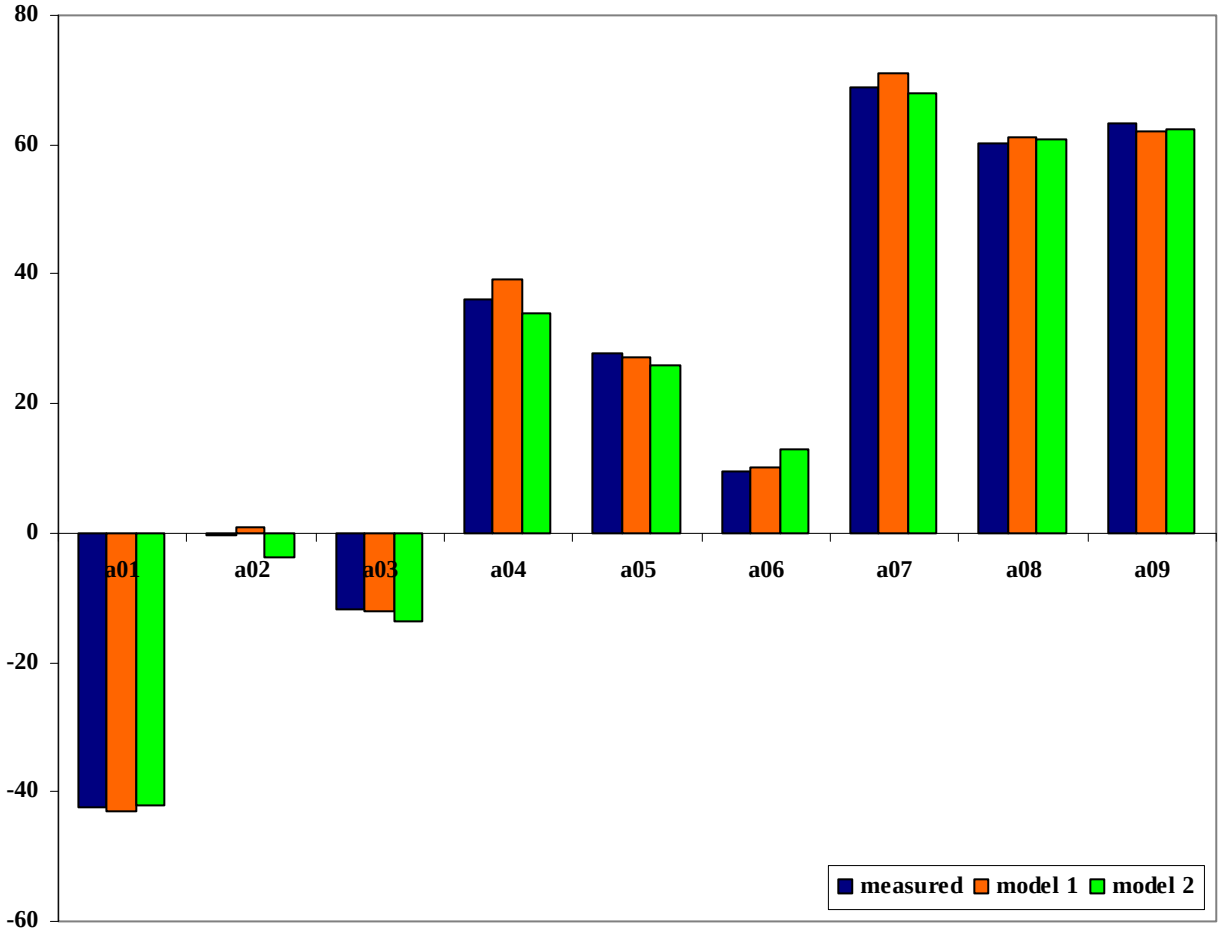
No_{tr}	r	F_{tr}	No_{ts}	r	F_{ts}
40	0.9992 [0.9987-0.9994]	11365*	33	0.9989	6987*
41	0.9993 [0.9988-0.9995]	1403*	32	0.9986	4693*
42	0.9989 [0.9982-0.9993]	9088*	31	0.9993	8648*
43	0.9988 [0.9980-0.9992]	8138*	30	0.9994	10549*
44	0.9987 [0.9979-0.9991]	7768*	29	0.9994	9881*
45	0.9993 [0.9988-0.9995]	14734*	28	0.9986	4191*
46	0.9991 [0.9985-0.9994]	11750*	27	0.9993	5926*
47	0.9990 [0.9984-0.9993]	11234*	26	0.9992	6706*
.....					
60	0.9990 [0.9984-0.9993]	14715	13	0.9996	3833*
61	0.9992 [0.9987-0.9994]	18102	12	0.9962	475*
62	0.9992 [0.9987-0.9994]	17566	11	0.9921	174*
63	0.9992 [0.9987-0.9994]	19681	10	0.9971	545*

CORRELATED CORRELATIONS

- A method of comparison two correlation coefficients taking into account the sample sizes on which those were obtained (Steiger 1980)
- Fisher's Z-test: test differences between correlation coefficients obtained by two different model (at a significance level of 5%)
- Implementation:

http://l.academicdirect.org/Statistics/tests/Steiger/

CORRELATED CORRELATIONS BY EXAMPLE



CORRELATED CORRELATIONS BY EXAMPLE

Eq(1)-Eq(2)			Eq(1)-Eq(3)			Eq(2)-Eq(3)	
$Y-\hat{Y}_{Eq(2)}$	**0.9956**		$Y-\hat{Y}_{Eq(3)}$	**0.9991**		$Y-\hat{Y}_{Eq(3)}$	**0.9991**
$Y-\hat{Y}_{Eq(1)}$	**0.9983**		$Y-\hat{Y}_{Eq(1)}$	**0.9983**		$Y-\hat{Y}_{Eq(2)}$	**0.9956**
$\hat{Y}_{Eq(1)}-\hat{Y}_{Eq(2)}$	**0.9957**		$\hat{Y}_{Eq(1)}-\hat{Y}_{Eq(3)}$	**0.9991**		$\hat{Y}_{Eq(2)}-\hat{Y}_{Eq(3)}$	**0.9976**
Steiger's Z	-4.24*		Steiger's Z	2.80‡		Steiger's Z	7.02‡
‡ $p < 0.05$; * $p \geq 0.05$							

CONCLUDING REMARKS

- The analytical methods are necessary for models interpretation and validation. The validity and reliability of a qSAR/qSPR model must be assessed carefully and consistently.
- The proposed analytical methods did not cover whole methods. The online software offers to researchers all over the world the opportunity to analyze models at the level of implemented statistical methods.
- The qSAR/qSPR models will have an important place in future chemical management as priority settings, risk assessment, classification and labelling.

ACKNOWLEDGEMENTS

- UEFISCSU Romania:
  - ID458/2007
    "Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca
  - ID1051/2007
    Technical University of Cluj-Napoca

- Thank you!