





---

**Untersuchung von chemometrischen Methoden zur Erstellung und  
Validierung von QSAR-Modellen**



Von der Fakultät für Lebenswissenschaften  
der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Desiree Ingrid Baumann  
aus Mainz

1. Referent:

2. Referent:

eingereicht am: 15.07.2015

mündliche Prüfung (Disputation) am: 18.12.2015

Prof. Dr. Knut Baumann

Prof. Dr. Hermann Wätzig

Druckjahr 2016

Dissertation an der Technischen Universität Braunschweig,  
Fakultät für Lebenswissenschaften

Berichte aus der Pharmazie

**Desiree Ingrid Baumann**

**Untersuchung von chemometrischen Methoden  
zur Erstellung und Validierung von QSAR-Modellen**

Shaker Verlag  
Aachen 2016

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Zugl.: Braunschweig, Techn. Univ., Diss., 2015

Copyright Shaker Verlag 2016

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8440-4551-2

ISSN 0945-0939

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen  
Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9  
Internet: [www.shaker.de](http://www.shaker.de) • E-Mail: [info@shaker.de](mailto:info@shaker.de)

## **Vorveröffentlichungen der Dissertation**

Teilergebnisse aus der vorliegenden Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor dieser Arbeit, in folgenden Beiträgen vorab veröffentlicht:

### **Publikationen:**

Baumann D, Baumann K: **Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation.** *J Cheminform* 2014, 6.

### **Tagungsbeiträge:**

Baumann D, Baumann K: **Reliable Estimation of externally validated prediction errors for QSAR models,** *Journal of Cheminformatics* 2013, 5 (Suppl 1).

Baumann D, Baumann K, **Beyond Search-Based Variable Selection: Predictivity, Model Selection and Stability,** Conferentia Chemometrica, Ungarn, Sopron, 2013.

Baumann D, Kreiß J, Baumann K, **Median  $R_{test}^2$  for Characterizing the Success of a Regression Model,** Scandinavian Conference, SSC14 Symposium, Sardinien (2015).

# Inhaltsverzeichnis

## I Einleitung und Zielsetzung der Arbeit

### 1. Quantitative Struktur-Aktivitäts-Beziehungen und das lineare Modell

1.1. Quantitative Struktur-Aktivitäts-Beziehung (QSAR) .....	1
1.2. Zielsetzung der Arbeit .....	3
1.3. Moleküldescriptoren .....	7
1.4. Das lineare Modell .....	8
1.4.1. Die Multiple lineare Regression zur Lösung des linearen Modells.....	9
1.4.2. Limitierungen der MLR .....	10

## II Grundlagen und Methoden

### 2. Modellerstellung und Modelloptimierung

2.1. Modellselektion .....	12
2.2. Variablenselektion.....	13
2.2.1. Die Tabu-Suche .....	14
2.2.2. „Simulated Annealing“ .....	16
2.3. Das einfache Kreuzvalidierungsschema .....	18
2.3.1. Lass'-ein Objekt-heraus Kreuzvalidierung.....	20
2.3.2. Die $k$ -fache Kreuzvalidierung .....	21
2.3.3. Lass'-mehrere Objekte-heraus Kreuzvalidierung .....	21
2.4. Der Einfluss der Kreuzvalidierung auf die Modellselektion .....	22
2.4.1. Das Phänomen der Modellüberanpassung („Overfitting“) .....	22
2.4.2. Kreuzvalidierte Gütekriterien und der „Model Selection Bias“ .....	23
2.4.3. Einfluss der Kreuzvalidierung auf die Konsistenz der Variablenselektion .....	24
2.4.4. Die Modellunteranpassung und die Validierdatensatzgröße .....	24

### 3. Überblick über Güteparameter

3.1. Absolute Gütekriterien .....	26
3.2. Relative Gütekriterien .....	27

### 4. Fehlerschätzungen in Kontext der Datenstruktur

4.1. Der Resubstitutionsfehler .....	30
4.2. Der „In-Sample Error“ und das „X-Fixed“-Design.....	30
4.3. Der Generalisierungsfehler und das „X-Random“- Design .....	32
4.4. Der Generalisierungsfehler in Abhängigkeit von der Modellkomplexität .....	34



## 5. Validierungskonzepte und die Schätzung des Vorhersagefehlers

5.1. Die doppelte Kreuzvalidierung (DCV) .....	37
5.1.1. Die wiederholte doppelte Kreuzvalidierung .....	39
5.1.2. Die doppelte Kreuzvalidierung als Ensemble-Methode und „Bagging“ .....	39
5.2. Die Testdatensatzmethode.....	41
5.3. Die doppelte Kreuzvalidierung in Kontext der internen und externen Validierung.....	42
5.4. Vergleich der Testdatensatzmethode mit der doppelten Kreuzvalidierung.....	45

## 6. Methoden der multivariaten Statistik

6.1. Die Singulärwertzerlegung (SVD) .....	48
6.1.1. Die Hauptkomponentenanalyse.....	49
6.1.2. Die SVD zur Schätzung der Regressionskoeffizienten .....	49
6.1.3. Die SVD zur Erkennung von Multikollinearitäten.....	50
6.1.4. Die SVD zur Approximation der X-Matrix .....	51
6.2. Alternative Regressionstechniken zur MLR .....	52
6.2.1. Die Hauptkomponentenregression (PCR) .....	52
6.2.1.1. Der Begriff der Rangreduktion in Kontext der PCR-Schätzung .....	53
6.2.2. Die Ridge-Regression.....	54
6.2.3. Das Lasso .....	56
6.2.4. Das Elastische Netz .....	58
6.2.5. „L <sub>2</sub> -Boosting“ .....	60
6.2.6. „Twin-Boosting“ .....	62
6.2.7. CAR-scores .....	63
6.3. Klassifikationsmethoden .....	64
6.3.1. Lineare Diskriminanzfunktion nach Fisher .....	64
6.3.2. „Support Vector Machines“ (SVMs) .....	66
6.4. Hypothesentests.....	69
6.4.1. Der t-Test auf verbundene Stichproben.....	69
6.4.2. Der Wilcoxon-Paardifferenztest.....	71

## III Herleitungen und Versuchsaufbau

### 7. Theoretischer Hintergrund zur Simulationsstudie

7.1. Systematischer Fehler und Varianz der Regressionskoeffizientenschätzung (volles Modell) .....	72
7.1.1. Die MLR-Schätzung unter dem Gauss-Markov Theorem .....	72
7.1.2. Die Varianz der MLR-Schätzung für das volle Modell .....	73
7.1.3. Varianz und systematischer Fehler der PCR-Schätzung für das volle Modell.....	74
7.1.4. Varianz und systematischer Fehler der Ridge-Schätzung für das volle Modell.....	76
7.2. Die Zusammensetzung des Vorhersagefehlers.....	78

7.2.1. Der Vorhersagefehler der MLR unter Gültigkeit des Gauss-Markov Theorems .....	80
7.2.2. Der Vorhersagefehler der $k$ -nächsten-Nachbarn-Methode .....	82
7.3. Auswirkung der Variablenselektion im Fall der MLR-Schätzung .....	83
7.3.1. Der „Omitted-Variable Bias“ im Fall der MLR .....	83
7.3.2. Der „Omitted-Variable-Bias“ und die Modellunterspezifikation .....	84

## **8. Simulationsstudie zur Untersuchung der DCV für den Regressionsfall**

8.1. Die Simulationsmodelle A.1 und A.2 .....	85
8.2. Durchführung der Simulationsstudie .....	86
8.3. Kriterien zur Auswertung der Simulationsstudie .....	88
8.3.1. Zerlegung des Modellfehlers für die MLR .....	93
8.3.2. Zerlegung des Modellfehlers für die PCR .....	98
8.3.3. Zerlegung des Modellfehlers für die Ridge-Regression .....	102
8.3.4. Zerlegung des Modellfehlers für das Lasso .....	105
8.3.5. Zusammenfassung und kurzes Fazit der Zerlegung .....	107

# **IV Ergebnisse**

## **9. Untersuchung der Simulationsmodelle A.1 und A.2**

9.1. Ergebnisse für das Simulationsmodell A.1 .....	109
9.2. Ergebnisse für TS-MLR und TS-Ridge (Simulationsmodell A.2) .....	115
9.3. Ergebnisse für TS-MLR, TS-PCR, TS-Ridge, Lasso (Simulationsmodell A.2) .....	124
9.4. Untersuchung der Validität der doppelten Kreuzvalidierung .....	134
9.4.1. Vergleich der Vorhersagefehler mit den theoretischen Vorhersagefehlern .....	134
9.4.2. Hypothesentests zur Untersuchung der Validität der DCV .....	140
9.4.3. Einfluss der Testdatensatzgröße auf die Variabilität des Vorhersagefehlers .....	142

## **10. Experimentelle Daten zur Untersuchung der DCV für den Regressionsfall**

10.1. Der Löslichkeitsdatensatz .....	146
10.2. Der Artemisinin Datensatz .....	155
10.3. Der LogP-Datensatz .....	161
10.4. Der Adenosin Datensatz .....	162

## **11. Untersuchung der doppelten Kreuzvalidierung (DCV) als Ensemble-Methode**

11.1. Simulationsstudie zur Untersuchung der DCV als Ensemble-Methode .....	164
11.1.1. Kriterien zur Analyse der DCV unter Ensemble-Bildung .....	165
11.1.2. Simulationsergebnisse zur DCV unter Ensemble-Bildung .....	168

11.2. Reales Datenbeispiel zur Untersuchung der DCV als Ensemble-Methode .....	174
--	-----

## **12. Untersuchung der doppelten Kreuzvalidierung für den Klassifikationsfall**

12.1. Simulationsstudie zur Untersuchung der DCV für die Klassifikation .....	177
12.1.1. Simulationsergebnisse für den Klassifikationsfall.....	179
12.1.2. Hypothesentests zur Untersuchung der Validität der DCV .....	182
12.2. Datenbeispiel zur Untersuchung der Validität der DCV für den Klassifikationsfall.....	184

## **13. Vergleich von verschiedenen Variablenselektionsmethoden**

13.1. Simulationsmodelle zur Untersuchung verschiedener Variablenselektionsmethoden .....	185
13.1.1. Durchführung der Simulationsstudie.....	186
13.1.2. Simulationsergebnisse .....	188
13.2. Steroid-Datenbeispiel zur Untersuchung der Variablenselektionsverfahren .....	195

## **14. Untersuchung von relativen Güteparametern**

14.1. Herleitung der Erwartungswerte für den Mittelwert und Median des $R_{test}^2$ .....	198
14.1.1. Herleitung des Erwartungswertes des $R_{test}^2$ .....	198
14.1.2. Der Median des $R_{test}^2$ .....	202
14.1.3. Berechnung der Erwartungswerte für den Mittelwert und Median des $R_{test}^2$ .....	205
14.2. Untersuchung des $R_{test}^2$ an Simulationsmodellen.....	207
14.2.1. Simulationsmodell C.1 .....	208
14.2.2. Simulationsmodell C.2 .....	208
14.2.3. Simulationsmodell C.3 .....	208
14.2.4. Simulationsmodell C.4 .....	209
14.2.5. Die Simulationsmodelle C.1-C.4 und ihre Annahmen .....	210
14.2.6. Durchführung der Simulationsstudie.....	210
14.2.7. Kriterien zur Analyse der Simulationsmodelle.....	211
14.2.8. Ergebnisse der Simulationsmodelle C.1 und C.2 .....	212
14.2.9. Ergebnisse der Simulationsmodelle C.3 und C.4 .....	217
14.3. Reale Datenbeispiele zur Untersuchung des $R_{test}^2$ .....	218
14.3.1. Der Löslichkeitsdatensatz.....	219
14.3.2. Der LogP-Datensatz .....	220
14.4. Einfluss der wiederholten Stichprobenziehung auf die Variabilität des $R_{test}^2$ .....	222
14.5. Fazit des Kapitels 14 .....	223

<b>15. Zusammenfassung und Schlussfolgerung.....</b>	<b>224</b>
--	------------

## V Anhang

### 16. Ergänzende Ergebnisse zu den Simulationsmodellen für die DCV

16.1. Variablenselektionsergebnisse (Simulationsmodell A.1, <b>Abbildungen A1-A2</b> ) .....	229
16.2. Vorhersagefehler der DCV ( <b>Tabellen A1a-d</b> ) .....	230
16.3. Bias- und Varianzterm für das Lasso (Simulationsmodell A.1, <b>Tabelle A2</b> ) .....	232
16.4. Variablenselektion ( <b>Abbildungen A3-A5</b> : TS-MLR, TS-PCR, Simulationsmodell A.2) .....	233
16.5. Variablenselektion ( <b>Tabellen A3a-j</b> : TS-MLR, TS-PCR, Simulationsmodell A.2) .....	235
16.6. Variablenselektion ( <b>Abbildung A6-A8</b> : TS-MLR, TS-PCR, TS-Ridge, Lasso) .....	245
16.7. Verteilung der Differenzen der Vorhersagefehler ( <b>Abbildung A9</b> ).....	246
16.8. Bias- und Varianzterme (Simulationsmodell A2, <b>Tabellen A4a-d</b> ) .....	247
16.9. Konfidenzintervalle der Vorhersagefehler ( <b>Tabelle A5a-i</b> : Simulationsmodell A.1).....	249
16.10. Konfidenzintervalle der Vorhersagefehler ( <b>Tabelle A6a-i</b> : Simulationsmodell A.2).....	254
16.11. t-Test auf verbundene Stichproben ( <b>Tabelle A7a-d</b> : Simulationsmodelle A1-A2) .....	259
16.12. Wilcoxon-Paardifferenztest ( <b>Tabelle A8a-d</b> : Simulationsmodelle A1-A2) .....	261

### 17. Ergänzende Ergebnisse zu den realen Datenbeispielen

17.1. Indices der Objekte für die Variablenvorselektion, Löslichkeitsdaten ( <b>Tabelle A9</b> ) .....	263
17.2. Vorhersagefehler (,Orakel'- und Testdaten, Löslichkeitsdaten) ( <b>Abbildung A10</b> ) .....	264
17.3. Vorhersagefehler für den Artemisinin-Datensatz, LMO: $\alpha=30\%$ ( <b>Abbildung A11</b> ).....	265
17.4. Vorhersagefehler (,Orakel'-, Testdaten, Artemisinin-Daten), <b>Abbildung A12</b> ) .....	266
17.5. Variabilität der ,Orakel'-Vorhersagefehler (Artemisinin-Daten, <b>Abbildung A13</b> ) .....	267
18. DCV mit Ensemble-Bildung ( <b>Tabelle A10</b> : TS-PCR, Simulationsmodell A.1) .....	267
19. Ergänzende Herleitung zum $R^2_{test}$ .....	268
20. Programmiercode.....	269
Literaturverzeichnis .....	290