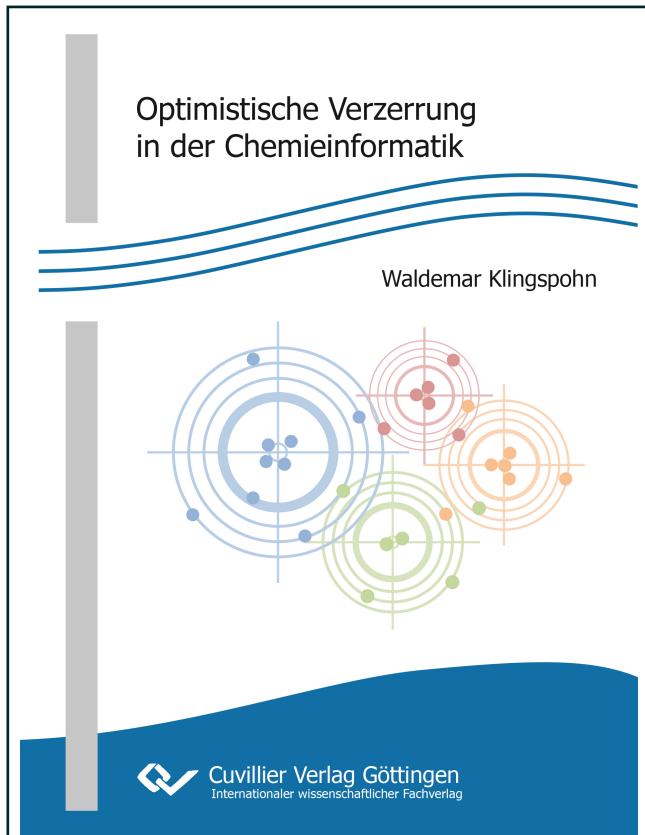




Waldemar Klingspohn (Autor)  
**Optimistische Verzerrung in der Chemieinformatik**



<https://cuvillier.de/de/shop/publications/8031>

Copyright:  
Cuvillier Verlag, Inhaberin Annette Jentsch-Cuvillier, Nonnenstieg 8, 37075 Göttingen,  
Germany  
Telefon: +49 (0)551 54724-0, E-Mail: [info@cuvillier.de](mailto:info@cuvillier.de), Website: <https://cuvillier.de>



# Inhaltsverzeichnis

---

Abbildungsverzeichnis [xv](#)

Tabellenverzeichnis [xix](#)

Abkürzungen und Symbole [xxi](#)

Mathematische Notation [xxv](#)

## I Einleitung [1](#)

### 1 Einleitung [3](#)

- 1.1 Der Beginn moderner Wirkstoffentwicklung [3](#)
- 1.2 Analyse der QSAR [7](#)
- 1.3 Validität einer QSAR-Analyse nach OECD [11](#)

### 2 Optimistische Verzerrung im Maschinellen Lernen [13](#)

- 2.1 Aspekte der optimistischen Verzerrung [14](#)

### 3 Zielsetzung der Arbeit [17](#)

## II Grundlagen und Methoden [19](#)

### 4 Moleküldeskriptoren [21](#)

- 4.1 Fingerprints [21](#)
- 4.2 Datenstruktur [22](#)

### 5 Datenvorbehandlung [23](#)

- 5.1 Normalisierung [23](#)
- 5.2 Zentrierung [24](#)
- 5.3 Standardisierung [24](#)

### 6 Die Klassifikation [27](#)

- 6.1 Das Klassifikationsproblem [28](#)



6.2	Modellanpassung und das „Bias-Varianz-Dilemma“	30
6.3	Güteparameter der Klassifikation	33
<b>7</b>	<b>Klassifikationsmodelle</b>	<b>37</b>
7.1	$k$ -Nächste-Nachbarn ( $k$ -NN)	37
7.2	Support Vector Machines (SVM)	39
7.2.1	Hyperparameter der SVM	43
7.3	Random Forest (RF)	44
7.3.1	Der Entscheidungsbaum	44
7.3.2	Vom Baum zum Wald	48
7.4	Rotation Forest (RotF)	51
7.5	Local <sub>SVM</sub> -Methode (Local <sub>SVM</sub> )	55
<b>8</b>	<b>Evaluierung der Leistungsfähigkeit</b>	<b>59</b>
8.1	„Hold-out“-Methode	60
8.2	$k$ -fache Kreuzvalidierung	61
8.3	Modellselektion im Rahmen eines Optimierungsverfahrens	63
8.4	Doppelte Kreuzvalidierung zur Modellselektion und Evaluierung	65
<b>9</b>	<b>Datensätze</b>	<b>69</b>
9.1	Das Klassenungleichgewicht	69
9.1.1	Regulierung des Klassenungleichgewichts	70
9.2	Software	74
<b>III</b>	<b>Ergebnisse und Diskussion</b>	<b>75</b>
<b>10</b>	<b>Variabilität im Zuge der Kreuzvalidierung</b>	<b>77</b>
10.1	Variabilität der Kreuzvalidierung	78
10.1.1	Methodik	78
10.1.2	Ergebnisse	79
10.2	Vergleichsoptimierung unter dem Aspekt der Variabilität der Kreuzvalidierung	83
10.2.1	Methodik	84
10.2.2	Ergebnisse	86
10.3	Evaluierung einer früheren Publikation	94
10.4	Diskussion	95
<b>11</b>	<b>Optimierung der Hyperparameter</b>	<b>101</b>
11.1	Methodik	102
11.2	Ergebnisse Modelloptimierung SVM	105



11.3 Ergebnisse *Model Selection Bias* 112

11.4 Diskussion 118

## 12 Local<sub>SVM</sub> 125

12.1 Das Multi-Klassen-Problem 126

12.1.1 Methoden zur Zerlegung des Multi-Klassen-Problems 127

12.2 Methodik 128

12.3 Ergebnisse 130

12.4 Diskussion 140

## 13 Rotation Forest 145

13.1 Methodik 146

13.2 Ergebnisse 148

13.3 Diskussion 157

# IV Zusammenfassung – Summary 161

## Anhang 173

### A Ergänzendes Material 175

A.1 Variabilität im Zuge der Kreuzvalidierung 175

A.2 Optimierung der Hyperparameter 197

A.3 Local<sub>SVM</sub> 215

A.4 Rotation Forest 235

### B MATLAB Quellcode 237

B.1 Standardisierung 237

B.2 *k*-fache Kreuzvalidierung 238

B.3 *Random Under- und Oversampling* 239

B.4 Random Forest 242

B.5 Rotation Forest 245

B.6 Support Vector Machines 250

B.7 *k*-Nächste-Nachbarn 252

B.8 Local<sub>SVM</sub> 255

B.9 Wiederholte Doppelte Kreuzvalidierung für die SVM 263

### C Publikation 269

### Literaturverzeichnis 287