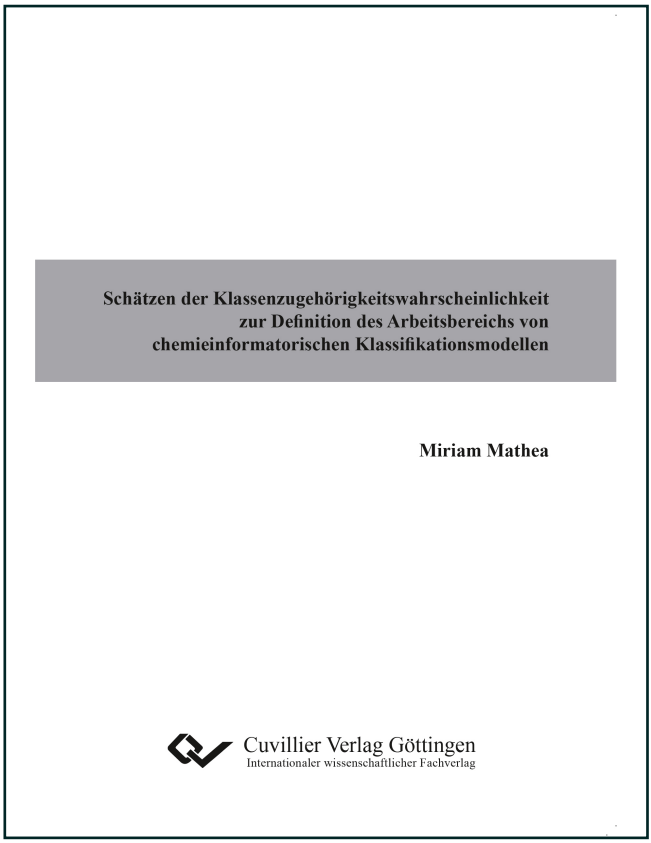
Miriam Mathea (Autor)

Schätzen der Klassenzugehörigkeitswahrscheinlichkeit zur Definition des Arbeitsbereichs von chemieinformatorischen Klassifikationsmodellen



https://cuvillier.de/de/shop/publications/7732

1 Theoretische Grundlagen

1.1 Quantitative-Struktur-Wirkungs-Beziehungen

In der modernen Arzneistoffentwicklung werden u.a. computergestützte Verfahren für das Wirkstoffdesign verwendet. Diese Verfahren werden auch als *In-Silico*-Techniken bezeichnet. Hierzu gehört auch die Analyse von quantitativen Struktur-Aktivitäts-Beziehungen (engl.: Quantitative-Structure-Activity Relationships (QSAR)). QSAR Analysen haben das Ziel entweder die biologische Aktivität einer Substanz selbst oder bestimmte Faktoren, die die Aktivität bestimmen, vorherzusagen [1]. Hierbei wird die Struktur-Aktivitäts-Beziehung mit Hilfe einer mathematischen Funktion ausgedrückt. Wenn eine solche Analyse durchgeführt werden soll, werden zunächst die biologischen Aktivitäten von verwandten Molekülen sowie deren chemische Struktur benötigt. Damit die chemische Struktur der Analyse zugänglich gemacht werden kann, werden Deskriptoren berechnet (siehe 2.2). Danach wird der funktionelle Zusammenhang zwischen den Moleküldeskriptoren und der biologischen Aktivität modelliert.

Die QSAR Analyse nahm vermutlich ihre Anfänge mit der Publikation zweier schottischer Pharmakologen (Crum-Brown und Fraser), welche 1868 zu der Erkenntnis kamen, dass die physiologische Aktivität ϕ einer Substanz eine Funktion ihrer chemischen Konstitution C sei.

$$\phi = f(C)$$

Im Jahre 1964 knüpften Hansch und Fujita [2] sowie Free und Wilson [3] an diese Idee an, indem sie die biologische Aktivität und die physikalisch-chemischen, sowie strukturellen Eigenschaften von Molekülen korrelierten. Heutzutage ist die QSAR Analyse eine gut etablierte Methode, welche vielfältig angewendet wird, nicht nur zur Vorhersage der biologischen Aktivität, sondern auch zur Vorhersage anderer Moleküleigenschaften im Rahmen quantitativer Struktur-Eigenschafts-Beziehungen (engl.: Quantitative Structure-Property Relationships (QSPR)). Durch die Möglichkeit Eigenschaften von noch nicht vorhandenen Molekülen vorherzusagen, lässt sich gegebenenfalls der zeit- und kostenintensive Syntheseaufwand reduzieren bzw. effizienter in eine bestimmte Richtung lenken (Leitstrukturoptimierung) [4, 5,6].

1.2 Moleküldeskriptoren

1.2.1 Einleitung

Die Analyse der strukturellen Information von Molekülen ist mit der Hilfe von Moleküldeskriptoren möglich. Hierbei handelt es sich um eine numerische Repräsentation des Moleküls. Deskriptoren können einerseits das Ergebnis standardisierter Experimente umfassen, zum Beispiel physikochemische Eigenschaften repräsentieren oder sie sind das Ergebnis eines standardisierten Algorithmus. Es gibt folglich sehr viele unterschiedliche Moleküldeskriptoren, die für verschiedene Anwendungsgebiete mehr oder weniger gut geeignet sind. Eine sehr ausführliche und umfassende Übersicht wurde von Todeschini verfasst [7].

Moleküldeskriptoren lassen sich nach ihrer Dimensionalität in unterschiedliche Klassen einteilen. 1D-Deskriptoren können beispielsweise einfache Eigenschaften wie das Molekulargewicht oder aber auch die Anzahl bestimmter Atome oder Bindungen kodieren. Die populärsten Deskriptoren sind die 2D-Deskriptoren und die 3D-Deskriptoren, welche zusätzlich die Topologie bzw. die Konformation des Moleküls mit einbeziehen. Darüber hinaus gibt es auch 4D- und höher dimensionale Deskriptoren [5].

1.2.2 Fingerabdruck-Deskriptoren (engl.: Fingerprints)

Die 2D-Fingerabdruck-Deskriptoren gehören zu den am weitesten verbreiteten Deskriptoren. Es sind Vektoren [8, 3, 9–11], welche ein Molekül bezüglich der An- oder Abwesenheit und/oder der Frequenz bestimmter Substrukturen charakterisieren. So kann beispielsweise die Anwesenheit einer Hydroxylgruppe mit 1 für anwesend oder 0 für abwesend gekennzeichnet werden (Abbildung 1).

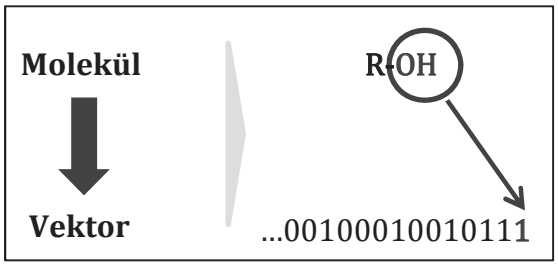


Abbildung 1: Ausschnitt der Erzeugung eines Vektors. Dieser Vektor wird auch als molekularer Fingerabdruck bezeichnet. Die chemischen Eigenschaften des Moleküls werden numerisch repräsentiert, in diesem Beispiel wird die Anwesenheit der Hydroxylgruppe durch eine 1 im Vektor gekennzeichnet.

1.2.3 Topologische Deskriptoren

Die topologischen Deskriptoren gehören ebenfalls zu den 2D-Deskriptoren. Sie sind weit verbreitet und werden vielfältig angewendet. Topologische Deskriptoren kodieren chemische Verknüpfungsinformationen von Molekülen. Diese Verknüpfungsinformationen, wie zum Beispiel die Verknüpfungsart (auch Konnektivität genannt) oder die Größe eines Rings, lassen sich aus der Strukturformel ableiten [12, 7, 13, 14].

Da 2D-Deskriptoren keine Informationen über die genaue räumliche Anordnung der Moleküle (die sog. Konformation) benötigen, sind sie oftmals beliebter als 3D-Deskriptoren. In vielen Fällen ist die Konformation der aktiven Verbindung unbekannt und somit ist eine Vielzahl an vorbereitenden und z.T. rechenintensiven Schritten notwendig, bevor mit diesen 3D-Deskriptoren gearbeitet werden kann [4].

1.3 Einführung in die Multivariate Datenanalyse

1.3.1 Einleitung

Im letzten Kapitel (1.2) wurde die Funktion von Deskriptoren beschrieben. Hierauf wird nun aufgebaut. Es wird beispielhaft angenommen, dass die Bioaktivität von verschiedenen Molekülen an einer bestimmten Zielstruktur gemessen wurde und zu jedem Molekül jeweils Deskriptoren berechnet wurden. Nun wird die Frage gestellt, wodurch die Bioaktivität beeinflusst wird und ob diese gegebenenfalls modelliert werden kann, um die Bioaktivität für ein unbekanntes Molekül vorhersagen zu können. Es gibt verschiedene Parameter von denen die Bioaktivität abhängen kann, beispielsweise die Molekülgröße oder die An- oder Abwesenheit bestimmter funktioneller Gruppen (alle Spalteneinträge des Deskriptors). In diesem Beispiel wird die Bioaktivität als abhängige oder beobachtete Variable $\boldsymbol{y}$ bezeichnet und die Spaltennamen des Deskriptors als unabhängige Variable $\boldsymbol{x}$ oder unabhängige Variablen X. Für p verschiedene unabhängige Variablen gilt: $X = (\boldsymbol{x_1}, \boldsymbol{x_2} \dots \boldsymbol{x_p})$. Wenn nun angenommen wird, dass $\boldsymbol{y}$ und X voneinander abhängen, kann dies folgendermaßen ausgedrückt werden:

$$\boldsymbol{y} = f(X) + \boldsymbol{e}.$$

Bei f handelt es sich um eine unbekannte Funktion von *X,* bei $\boldsymbol{e}$ um einen zufälligen Fehlerterm. Dieser Fehlerterm ist unabhängig von *X* und hat einen Mittelwert von Null. Die Funktion f kann auch mehr als nur einen Parameter miteinbeziehen. Das Ziel ist es *f* zu schätzen. In dieser Arbeit werden Vektoren mit einem kleinen, **fetten**, *kursiven* Buchstaben, Matrizen mit einem großen, *kursiven* Buchstaben und Skalare mit einem kleinen, *kursiven* Buchstaben gekennzeichnet.

Wenn beispielsweise nur *X* bekannt ist und $\boldsymbol{y}$ unbekannt ist und die Annahmen für den Fehlerterm (zufällig, unabhängig von *X*, Mittelwert ist Null) zutreffen, kann $\boldsymbol{y}$ durch:

$$\widehat{\boldsymbol{y}} = \hat{f}(X)$$

vorhergesagt werden. $\hat{f}$ repräsentiert die Schätzfunktion für f und $\widehat{\boldsymbol{y}}$ repräsentiert die Vorhersage für $\boldsymbol{y}$. Die Richtigkeit der Vorhersage von $\boldsymbol{y}$ hängt wesentlich von zwei Größen ab. Diese werden als reduzierbarer Fehler und nicht-reduzierbarer Fehler bezeichnet. Im Allgemeinen wird $\hat{f}$ keine perfekte Schätzfunktion für f sein. Dieser Fehler ist reduzierbar, weil die Richtigkeit von $\hat{f}$ potentiell durch unterschiedliche Techniken verbessert werden kann. Aber auch wenn f perfekt geschätzt werden würde, wären noch nicht alle Fehler beseitigt. Deshalb ist $\boldsymbol{y}$ auch eine Funktion von $\boldsymbol{e}$, welche per Definition nicht durch *X* vorhergesagt werden kann. Variabilität assoziiert mit $\boldsymbol{e}$ beeinträchtigt die Präzision der Vorhersage. Dieser Fehler wird auch nicht-reduzierbarer Fehler genannt. Die Größe $\boldsymbol{e}$ kann auch unbestimmte oder ungemessene, abhängige Variablen enthalten, welche nützlich wären um $\boldsymbol{y}$ zu bestimmen. Würden diese bekannt oder messbar sein, könnte der Fehler reduziert werden [15, 16].

$$E(\boldsymbol{y} - \widehat{\boldsymbol{y}})^2 = E[f(X) + \boldsymbol{e} - \hat{f}(X)]^2$$

$$= [f(X) - \hat{f}(X)]^2 \quad + \quad Var(\boldsymbol{e})$$

reduzierbar nicht-reduzierbar

Der nicht-reduzierbare Fehler wird immer eine obere Grenze für die Präzision von Vorhersagen vorgeben, welche in der Praxis fast immer unbekannt ist.

Es ist allerdings nicht immer das Ziel Vorhersagen für $\boldsymbol{y}$ zu tätigen, in manchen Fällen soll auch die Beziehung zwischen *X* und $\boldsymbol{y}$ untersucht werden, um zu verstehen, wie sich

$\boldsymbol{y}$ als Funktion von $\boldsymbol{x_1}, \boldsymbol{x_2} \ldots \boldsymbol{x_p}$ verändert. Es kann beispielsweise auch von Interesse sein lediglich einige wichtige Variablen zu identifizieren.

An dieser Stelle stellt sich natürlich die Frage, wie die unbekannte Funktion f geschätzt werden kann. Hierfür gibt es unterschiedliche lineare und nicht-lineare Ansätze. Generell haben diese Methoden bestimmte Charakteristiken, nach welchen sie unterschieden werden. Die meisten Methoden lassen sich entweder in die Gruppe der parametrischen oder in die Gruppe der nicht-parametrischen Methoden einteilen [15, 16].

1.3.2 Datenvorbehandlung

Häufig sind unterschiedliche Variablen nicht miteinander vergleichbar, da sie auf verschiedenen Skalen gemessen wurden. Mathematische Funktionen können aber sensibel für solche Unterschiede sein und diese mit modellieren.

Die Datenmatrix der unabhängigen Variablen *X* wird als **zentriert** bezeichnet, wenn von jedem Variablenvektor $\boldsymbol{x}$ der Mittelwert berechnet wird und der Mittelwertvektor schließlich von der Rohmatrix subtrahiert wird. Folglich wird von jedem Element von *X* sein entsprechender Spaltenmittelwert abgezogen.

Die Datenmatrix *X* wird als **autoskaliert** bezeichnet, wenn von jedem Variablenvektor $\boldsymbol{x}$ die Standardabweichung berechnet wird und jede Variable der zentrierten Matrix durch die zugehörige Standardabweichung geteilt wird. Wenn anstelle der zentrierten Matrix die Rohmatrix verwendet wird, so wird die Datenmatrix als **skaliert** bezeichnet.

Wenn die Datenmatrix sowohl zentriert als auch skaliert wurde, wird sie als autoskaliert oder **z-transformiert** bezeichnet.

1.3.3 Parametrische Methoden

Bei den parametrischen Methoden wird als erstes eine Annahme über die Gestalt von f gemacht. Beispielsweise wäre eine einfache Annahme, dass f linear in *X* ist.

$$f(X) = b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \cdots + b_p\boldsymbol{x_p}$$

Das Schätzproblem wurde durch die Annahme vereinfacht und somit müssen nur die Koeffizienten (auch Parameter genannt) $p + 1$: $b_0, b_1, b_2 \ldots b_p$ geschätzt werden. Nun

werden Trainingsdaten/Moleküle benötigt um das Modell zu trainieren. Es müssen $b_0, b_1 \dots b_p$ geschätzt werden, wobei Werte gefunden werden sollen, sodass:

$$\boldsymbol{y} \approx b_0 + b_1\boldsymbol{x_1} + b_2\boldsymbol{x_2} + \cdots + b_p\boldsymbol{x_p}.$$

Die geläufigste Methode um das Modell anzupassen wird als "Methode der kleinsten Quadrate" bezeichnet. Dieser Ansatz reduziert das Problem f zu schätzen darauf eine Reihe von Parametern zu schätzen. Die möglichen Nachteile sind, dass das Modell welches gewählt wird, normalerweise nicht das wahre f sein wird und je weiter es davon entfernt liegt, desto dürftiger wird die Schätzung. Eine Lösung hierfür wäre es flexiblere Modelle zu wählen, welche sich verschiedenen, möglichen Formen von f anpassen können, allerdings müssten hierfür mehr Parameter geschätzt werden. Ein komplexeres Modell neigt leichter zu einer „Überanpassung". Dieses Phänomen wird später noch einmal detaillierter betrachtet [15, 16].

1.3.4 Nichtparametrische Methoden

Nichtparametrische Methoden machen keine ausdrücklichen Annahmen über die Form von f. Stattdessen streben sie eine Schätzung von f an, welche die Abweichung der Schätzwerte und der Trainingsdaten minimiert. Dieser Ansatz hat einen großen Vorteil gegenüber den parametrischen Methoden. Dadurch, dass Annahmen über die Form von f vermieden werden, können nichtparametrische Methoden, durch eine größere Spanne an möglichen Formen, potentiell eine genauere Anpassung an f ermöglichen. Ein großer Nachteil jedoch ist, dass, solange das Problem f zu schätzen nicht auf eine kleine Anzahl Parameter reduziert werden kann, eine größere Anzahl an Daten/Molekülen benötigt wird um f genau zu schätzen [15, 16].

1.3.5 Das Dilemma zwischen Vorhersagegenauigkeit und Interpretierbarkeit

Allgemein lässt sich sagen, dass bei Erhöhung der Flexibilität eines Modells, sich die Interpretierbarkeit erniedrigt. In einigen Fällen würde ein unflexibles Modell bevorzugt werden. Wenn beispielsweise Interesse am Zusammenhang zwischen $\boldsymbol{y}$ und X besteht, ist es vorteilhafter ein leicht zu interpretierendes Modell (z.B. lineares Modell mit wenig Parametern) vorliegen zu haben. Im Gegensatz dazu wären flexible Ansätze weniger geeignet, da in diesem Fall Zusammenhänge mit einzelnen Variablen und $\boldsymbol{y}$ nur schwer zu erkennen sind. Auch wenn ausschließlich Interesse an der Vorhersage besteht, wäre

es trotzdem nicht immer sinnvoll die flexibelste Methode zu wählen aufgrund der Problematik der Überanpassung [15, 16].

1.4 Regression

1.4.1 Einfache Lineare Regression

Die Einfache Lineare Regression modelliert die abhängige Variable $\boldsymbol{y}$ mit nur einer einzigen unabhängigen Variablen $\boldsymbol{x}$. Hierbei wird angenommen, dass ein linearer Zusammenhang zwischen $\boldsymbol{x}$ und $\boldsymbol{y}$ besteht.

$$\boldsymbol{y} = b_0 + b_1\boldsymbol{x} + \boldsymbol{e}$$

Bei $b_0$und $b_1$handelt es sich um die Koeffizienten und bei $\boldsymbol{e}$ handelt es sich um den Gesamtfehler des Modells. Nun wird ein Teil der Moleküle des Datensatzes, die sogenannte Trainingsdatenpartition, benutzt um die Koeffizienten $\widehat{\boldsymbol{b}}$ zu schätzen. Danach kann eine Vorhersage für ein zukünftiges ungesehenes Molekül x_0 gemacht werden.

$$\widehat{y_0} = \widehat{b_0} + \widehat{b_1}x_0$$

Das Ziel ist es, die Koeffizienten so zu schätzen, dass die resultierende Gerade möglichst nah an den vorhandenen Punkten verläuft. Es gibt unterschiedliche Verfahren um zu messen, was denn eigentlich nah ist. Die wohl geläufigste Methode ist die bereits erwähnte „Methode der kleinsten Quadrate“. Bei dieser Methode werden zunächst die Residuen berechnet ($\boldsymbol{y} - \widehat{\boldsymbol{y}}$) und danach werden diese quadriert und aufsummiert. Schließlich werden die Koeffizienten $\widehat{\boldsymbol{b}}$ so ausgewählt, dass die Summe der quadrierten Residuen (engl. Residual Sum of Squares (RSS)) minimiert wird.

1.4.2 Modellvalidierung

Im Rahmen der Regression wird meist der **Mittlere Quadratische Fehler (engl.: Mean Squarred Error (MSE))** verwendet um die Leistungsfähigkeit einer Methode zu beurteilen.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x_i}))^2$$

Falls der vorhergesagte Vektor $\boldsymbol{y} = (f(\boldsymbol{x}))$ nah an den experimentell ermittelten Vektor $\boldsymbol{y}$ ist, wird der MSE klein. Nachdem der MSE für das betrachtete Modell berechnet wurde, stellt sich die Frage, wie gut dieses Modell für zukünftige Daten/Molekülen geeignet ist, welche bisher nicht bei der Modellbildung zum Einsatz gekommen sind. In der Regel werden Modelle nicht nur erstellt um einen Zusammenhang zwischen den unabhängigen und abhängigen Variablen herzustellen, sondern auch um mit ihnen Eigenschaften (z.B. Bioaktivität) zukünftiger bisher nicht vorhandener Moleküle vorhersagen zu können. Um die Modellgüte testen zu können, wird der verwendete Datensatz in eine Trainingsdatenpartition und eine Testdatenpartition unterteilt. Mit den Trainingsdaten $\{(x_1, y_1), \dots, (x_n, y_n)\}$ wird die Schätzfunktion $\hat{f}$ erhalten, dieser Prozess wird auch als „Modelltraining“ bezeichnet. Anschließend können $\hat{f}(x_1), \hat{f}(x_2), \dots, \hat{f}(x_n)$ berechnet werden. Wenn diese Werte ungefähr gleich $y_1, y_2, \dots y_n$ sind, dann ist der MSE_{Train} (MSE der Trainingsdaten) klein. Wie bereits erwähnt, ist es zusätzlich interessant, nicht nur den MSE_{Train} zu berechnen, sondern zu wissen, ob $\hat{f}(x_0)$ ungefähr gleich y_0 ist. Bei x_0 handelt es sich um ein bisher ungesehenes Testdatum, welches bisher nicht als Trainingsdatum benutzt wurde. Am vielversprechendsten ist die Methode, welche den niedrigsten MSE_{Test} (MSE der Testdaten) aufweist. Denn wenn neue Moleküle hinzukommen, von denen beispielsweise der jeweilige experimentelle Wert für y_0 nicht bekannt ist, so kann davon ausgegangen werden, dass der MSE vergleichbar ist mit dem MSE_{Test}.

Wenn eine Methode einen niedrigen MSE_{Train} aber einen hohen MSE_{Test} aufweist, ist dies ein Anzeichen für eine Überanpassung. Dies passiert, weil die Methode nicht nur die unbekannte Funktion modelliert, sondern auch den Zufallsfehler. Unabhängig von der Überanpassung wird immer ein höherer MSE_{Test} als MSE_{Train} erwartet, da die meisten Methoden direkt oder indirekt versuchen den MSE_{Train} zu minimieren. Weniger flexible Methoden neigen weniger zur Überanpassung. Es ist oftmals deutlich schwieriger aufgrund geringer Datenlage den MSE_{Test} zu bestimmen und somit das Modell mit dem niedrigsten MSE_{Test} zu finden. Eine wichtige Methode, welche effizient die vorhandenen Daten ausschöpft um aus den Trainingsdaten den MSE_{Test} zu bestimmen, ist die Kreuzvalidierung. [15, 16].

Der MSE_{Test} ist ein Qualitätsmaß für die Vorhersagekraft von QSAR Modellen. Es gibt darüber hinaus noch weitere Qualitätsmaße. Auf eines von diesen wird im nächsten Abschnitt kurz eingegangen. Neben dieser Funktion dient der MSE_{Test} auch zur Auswahl

von Modellparametern *p.* Modelle werden beispielsweise für verschiedene Parameter *p* erstellt und untereinander verglichen. Anschließend wird dann das Modell mit dem entsprechenden Parameter *p* ausgewählt, welches die beste Vorhersagekraft mit sich bringt. Dieser Prozess wird auch als interne Validierung bezeichnet, da alle Moleküle (inklusive der Testmoleküle) die Modellauswahl beeinflussen und somit der MSE_{Test} möglicherweise verzerrt geschätzt wird [15].

Ein weiteres Qualitätsmaß ist der **quadrierte Korrelationskoeffizient R^2**, welcher auch als Bestimmtheitsmaß bezeichnet wird. Er beschreibt zu welchem Anteil das gebildete Modell die Varianz der abhängigen Variable erklären kann. Der R^2 nimmt Werte zwischen 0 und 1 an.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Das eigentliche Qualitätsmaß ist analog zum MSE_{Test} der R^2_{Test}, welcher mit bisher ungesehenen, von der Modellbildung unabhängigen, Molekülen berechnet wird. Bei $\bar{y}$ handelt es sich um den Mittelwert der Trainingsdaten. Eine wichtige Methode, welche effizient die Trainingsdaten nutzt um den R^2_{Test} zu berechnen, ist genau wie beim MSE_{Test}, die Kreuzvalidierung, welche im Folgenden noch näher erläutert wird [15].

1.4.2.1 Das Dilemma zwischen Bias (systematischer Fehler) und Varianz

Der erwartete MSE_{Test} für ein ungesehenes Molekül x_0 kann zerlegt werden in die Summe aus der Varianz von $\hat{f}(x_0)$, dem quadrierten Bias von $\hat{f}(x_0)$ und der Varianz des Fehlerterms *e*.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(e)$$

$E(y_0 - \hat{f}(x_0))^2$ definieren den erwarteten MSE_{Test} und beziehen sich auf den gemittelten MSE_{Test} über alle Objekte x_0 aus dem Testdatensatz. Um den zu erwartenden Testfehler zu minimieren wird eine Methode benötigt, welche zugleich eine niedrige Varianz und einen niedrigen Bias erreicht. Der Bias eines Schätzers ist definiert als Differenz zwischen seinem Erwartungswert und der zu schätzenden Größe. Die Varianz und der quadrierte Bias sind positiv. Somit lässt sich erkennen, dass der zu erwartende MSE_{Test} niemals kleiner sein kann als die Varianz von *e*, dem nicht reduzierbaren Fehler. Die

$Var(\hat{f}(x_0))$ bezieht sich auf den Grad der Änderung von $\hat{f}$, wenn zur Schätzung verschiedene Trainingsdatensatzpartitionen benutzt werden. Aus unterschiedlichen Trainingsdatensatzpartitionen resultieren unterschiedliche $\hat{f}$s. Im Idealfall sollten die Unterschiede nicht zu groß sein. Falls eine Methode eine hohe Varianz aufweist, können kleine Unterschiede in den Trainingsdaten große Unterschiede in $\hat{f}$ hervorrufen. Allgemein weisen flexiblere Methoden eine höhere Varianz auf. Der Bias bezieht sich auf den Fehler, der gemacht wird, wenn ein komplexes Problem auf ein viel einfacheres Modell reduziert wird. Flexiblere Methoden weisen in der Regel einen geringeren Bias auf, aber eine höhere Varianz. Das Verhältnis in dem sich diese beiden Größen verändern entscheidet darüber, ob der MSE steigt oder sinkt. Wenn die Flexibilität von einer Methode erhöht wird, dann neigt der Bias dazu stärker zu sinken als die Varianz steigt und der MSE verringert sich. Ab einem bestimmten Punkt jedoch hat eine Steigerung der Flexibilität keinen großen Einfluss mehr auf den Bias aber die Varianz steigt signifikant, folglich vergrößert sich der MSE_{Test}. Das Ziel ist es, eine Methode zu finden, welche eine niedrige Varianz und einen niedrigen Bias aufweist. Angenommen das wahre f ist linear, dann würde die lineare Regression keinen Bias haben und flexiblere Methoden hätten Schwierigkeiten mitzuhalten. Wenn aber das wahre f hochgradig nicht-linear ist, funktionieren flexiblere Methoden vermutlich besser [15–18].

1.4.2.2 Kreuzvalidierung (engl.: Cross-Validation (CV))

Prinzipiell werden bei der Kreuzvalidierung der gesamte zur Verfügung stehende Datensatz in einen Konstruktionsdatensatz und einen Validierdatensatz aufgeteilt. Mit den Konstruktionsdaten wird ein Modell gebildet. Daraufhin werden die Eigenschaften der Validierdaten mit dem erstellten Modell vorhergesagt und ein Gütekriterium, wie beispielsweise der MSE, wird berechnet. Dieser Prozess wird mehrfach wiederholt, allerdings werden unterschiedliche Konstruktions- bzw. Validierdatenpartitionen gebildet. Je nach Vorgehensweise und Aufbau werden unterschiedliche Varianten der Kreuzvalidierung unterschieden.

Die **„Lass–ein–Objekt-heraus-Kreuzvalidierung“ (engl.: Leave-One-Out-Cross-Validation (LOO-CV))** ist eine Variante der Kreuzvalidierung, bei der einem Datensatz bestehend aus n Molekülen immer ein Molekül entzogen wird. Mit den restlichen $n - 1$ Molekülen wird das Modell gebildet und das separierte Molekül wird anschließend vor-