Monatshefte für Chemie 105, 1—10 (1974)

Anwendung einiger Methoden der automatischen Zeichenerkennung auf die Interpretation von Massenspektren

Von

K. Varmuza*

Aus dem Institut für Allgemeine Chemie der Technischen Hochschule Wien, Österreich

Mit 4 Abbildungen

(Eingegangen am 18. Oktober 1973)

Application of Some Pattern Recognition Methods in Interpreting Mass Spectra

Some pattern recognition methods are briefly discussed from the geometric point of view: classification by distance measurements to prototypes, classification by decision-planes and classification by the nearest-neighbor-method. The last two methods give good results with regard to the automatic determination of molecular structures from low resolution mass spectra. It is emphasized, that pattern recognition methods may be useful for interpreting various types of data pools in physical and analytical chemistry.

Einleitung

Viele automatisierte und computerisierte Großgeräte, die in der instrumentellen Analytik verwendet werden, produzieren während einer Analyse eine riesige Datenmenge. Diese Datenflut enthält sicher oft eine Vielzahl wertvoller Informationen, die jedoch nur schwer herauszulesen sind. Die wichtigste Aufgabe des Computers besteht daher nicht in der Produktion von Daten, sondern in der Datenreduktion und in einer zweckmäßigen Darstellung der Ergebnisse. Das gewaltige Datenmaterial (z. B. Spektren), das sich im Laufe der Zeit ansammelt, kann selbst vom spezialisierten Fachmann kaum überblickt und bei jeder Analyse berücksichtigt werden. Es ist daher naheliegend, auch die chemisch-physikalische Interpretation der Daten

* Auszugsweise vorgetragen während der Vortragstagung des Vereins Österreichischer Chemiker und der Gesellschaft für Chemiewirtschaft, 26. bis 29. September 1973, Linz.

— zumindest eine Vor-Interpretation — von einem Computerprogramm ausführen zu lassen. Dabei ergeben sich jedoch beträchtliche Schwierigkeiten: Die gewünschte Information (z. B. die chemische Struktur einer unbekannten Substanz) ist im allgemeinen nur in verschlüsselter Form in den zur Verfügung stehenden Daten (z. B. Spektren) enthalten; der theoretische Zusammenhang zwischen gewünschter Information und der vorliegenden Datenstruktur ist meist nicht so genau bekannt, wie es für eine mathematische Formulierung notwendig wäre. Durch Anwendung von Strategien, die in Schachspiel-Computerprogrammen erprobt wurden und auch der Arbeitsweise des Chemikers bei der Spektrenauswertung nahekommen, konnten Programme entwickelt werden, die aus dem Massenspektrum und dem NMR-Spektrum eine chemische Struktur ableiten[1, 2]. Diese *heuristischen Verfahren* sind recht leistungsfähig, erfordern jedoch sehr umfangreiche und zeitaufwendige Computerprogramme, die in absehbarer Zeit kaum einem größeren Benützerkreis zugänglich sein werden.

Ein anderer Weg besteht darin, rein empirisch, ohne Berücksichtigung der physikalisch-chemischen Zusammenhänge, die Interpretation der Daten zu versuchen. Dazu sind Methoden geeignet, die auf dem Gebiet der automatischen Zeichenerkennung, z. B. zur Schriftzeichenerkennung oder Bildauswertung, entwickelt wurden. Die grundsätzliche Vorgangsweise ist stets folgende: Aus einer Datensammlung wird mit Hilfe mathematischer Methoden ein *Klassifikator* (das ist im wesentlichen eine Rechenvorschrift) abgeleitet, der die vorhandenen Daten mit möglichst kleiner Fehlerrate richtig klassifiziert[3, 4]. (Man denke etwa an einen Klassifikator, der auf Grund der Massenspektren einer größeren Spektrenbibliothek entscheiden kann, ob die betreffende Substanz einen Benzolring enthält oder nicht.) Falls die gesuchte Information in den Daten enthalten ist und ein zweckmäßiges Klassifizierungsverfahren gewählt wurde, ist anzunehmen, daß der erhaltene Klassifikator auch neue Daten in einem hohen Ausmaß richtig klassifizieren wird. Viele Methoden der automatischen Zeichenerkennung zeichnen sich durch Einfachheit und Übersichtlichkeit aus und erfordern nur grundlegende Mathematikkenntnisse. Die Ermittlung eines Klassifikators ist zwar meist mit umfangreicher — aber nicht komplizierter — Rechenarbeit verbunden, doch ist die Anwendung des Klassifikators einfach und schnell zu bewerkstelligen.

In der vorliegenden Arbeit sollen einige Methoden der automatischen Zeichenerkennung mit Hilfe der anschaulichen geometrischen Betrachtungsweise einem präsumptiven Interessentenkreis vorgestellt werden. Diese Methoden wurden dahingehend untersucht, wieweit sie für die automatische Interpretation von niedrig aufgelösten Massenspektren geeignet sind. Das Ziel war Klassifikatoren zu entwickeln,

die aus dem niedrig aufgelösten Massenspektrum einer niedrigmolekularen Substanz direkte Hinweise auf ihre chemische Struktur liefern.

Darstellung von Merkmalsmengen als Vektor

Wir gehen davon aus, daß von einer größeren Anzahl von bekannten Objekten (Substanzen) jeweils eine Reihe von Merkmalen (Meßergebnissen) vorliegen. Die Merkmale $x_1, x_2, \ldots x_n$ einer Substanz x sind beispielsweise physikalisch-chemische Daten oder im Falle des Massenspektrums etwa die Peakhöhen im Spektrum bei den Massenzahlen 1 bis n. Die Merkmale einer Substanz werden zweckmäßigerweise in einem *Merkmalsvektor* $\vec{X}$ zusammengefaßt. Zur Darstellung dieses Vektors benötigt man einen n-dimensionalen Raum: Der Vektor geht vom Usprung aus und die Vektorspitze ist durch die Koordinaten $x_1, x_2, \ldots x_n$ gegeben. Im folgenden wird stets nur die Vektorspitze betrachtet. Zum Verständnis der Klassifizierungsmethoden reicht es meist aus, sich ein zwei- oder dreidimensionales Analogon zu vergegenwärtigen.

Jedes niedrig aufgelöste Massenspektrum kann daher ohne Informationsverlust als Punkt in einem n-dimensionalen Raum (Spektrenraum) dargestellt werden; jede Koordinatenachse entspricht einer Massenzahl; die Koordinaten des Punktes entsprechen den Peakhöhen im Spektrum.

Die Arbeitshypothese für die Entwicklung eines Klassifikators ist folgende: Man hofft, daß die Massenspektren chemisch ähnlicher Substanzen in diesem n-dimensionalen Spektrenraum Cluster bilden und daß der euklidische Abstand zwischen den Massenspektren ein Maß für die chemische Ähnlichkeit der Substanzen ist[5]. Der euklidische Abstand d zwischen den Punkten $X\ (x_1, x_2, \ldots x_n)$ und $Y\ (y_1, y_2, \ldots y_n)$ ist in Analogie zum anschaulichen zwei- bzw. dreidimensionalen Fall mit

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_n - y_n)^2}$$

gegeben. Im folgenden wird gezeigt, daß in Abhängigkeit von der Art der Clusterbildung verschiedene Methoden der automatischen Zeichenerkennung zweckmäßig sind.

Klassifizierung durch Abstandsmessung zu Prototypen

Abb. 1 zeigt eine theoretisch mögliche Clusterbildung für einen zweidimensionalen Fall. Die Ringe entsprechen Spektren der Substanzklasse A — beispielsweise Substanzen mit Benzolring; die Dreiecke entsprechen Spektren der Substanzklasse B — Substanzen ohne Benzol-

ring. S_A ist als Schwerpunkt aller Spektren der Klasse A der Prototyp dieser Klasse. In diesem einfachen Fall kann man durch eine Abstandsbestimmung entscheiden, welcher Klasse ein Spektrum angehört. Ist der Abstand eines Spektrums zum Schwerpunkt S_A größer als der kritische Abstand d_k, dann wird es der Klasse B zugeordnet, sonst der Klasse A.

Abb. 2 zeigt die tatsächlichen Verhältnisse bei 500 Massenspektren. Es wurde der Schwerpunkt der Massenspektren von 113 Substanzen

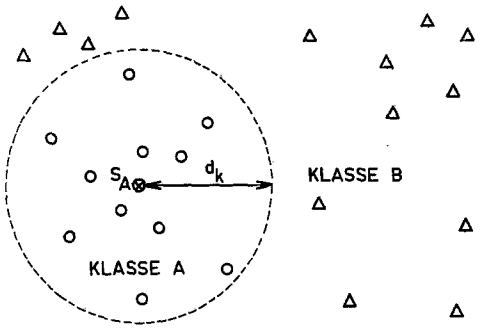

Abb. 1. Klassifizierung durch Abstandsmessung zum Schwerpunkt S_A der Klasse A. Innerhalb des kritischen euklidischen Abstandes d_k liegen nur Mitglieder der Klasse A

mit Benzolring, der Schwerpunkt der Massenspektren von 387 Substanzen ohne Benzolring, sowie die euklidischen Abstände d aller 500 Spektren zu den beiden Schwerpunkten berechnet. Die Häufigkeitsverteilungen der Abstände zeigen, daß es keinen kritischen Abstand gibt, der die beiden Klassen ausreichend trennt. Eine Normierung der Peakhöhen in den Massenspektren auf gleiche Summe der Peakhöhen in jedem Spektrum oder eine Normierung auf gleiche Vektorlänge (die Vektorspitzen liegen dann auf einer Hyperkugel[6]) ermöglichten keine bessere Separierung der beiden Klassen. Die Verwendung logarithmierter Peakhöhen anstatt der üblichen linearen Werte hatte dagegen einen günstigen Einfluß. Da auch die Klassifikationen anderer chemischer Strukturen (z. B. aliphatischer Alkohol, Sauerstoff im Molekül) ähnliche Ergebnisse brachte, scheint diese einfache Methode der Abstandsbestimmung für Massenspektren wenig wertvoll. Bei anderen Datenstrukturen mit ausgeprägterer Clusterbildung kann sie aber durchaus erfolgreich sein. Auch eine Gewichtung[3] der einzelnen Merkmale auf Grund des Beitrages zur Unterscheidung der Klassen wäre zu überlegen.

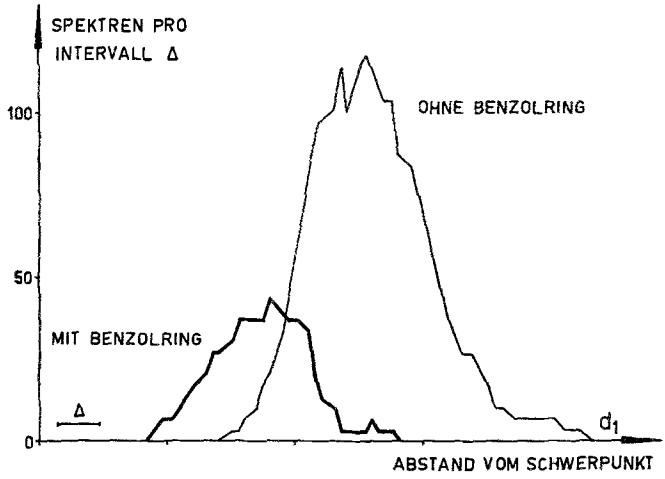

Abb. 2a

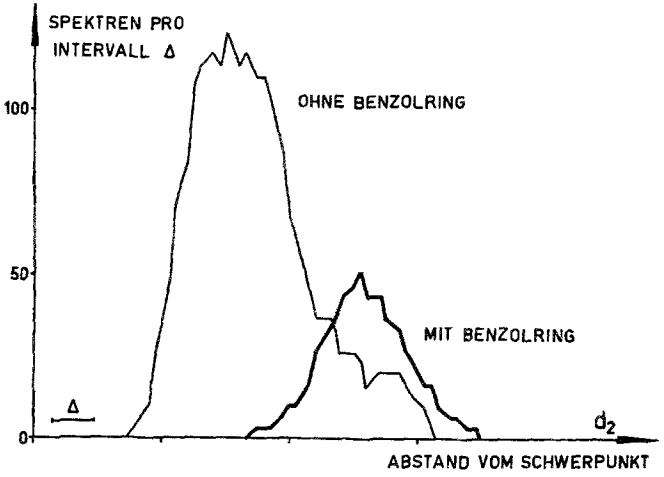

Abb. 2b

Abb. 2. Häufigkeitsverteilungen der euklidischen Abstände d_1 und d_2 von 500 Massenspektren zu den Schwerpunkten der Substanzen mit und ohne Benzolring. d_1 ist der Abstand vom Schwerpunkt der Substanzen mit Benzolring, d_2 ist der Abstand vom Schwerpunkt der Substanzen ohne Benzolring. (Der Maßstab ist linear und in willkürlichen Einheiten.) Die Peakhöhen der Massenspektren wurden logarithmiert und auf den Basispeak normiert

Klassifizierung mit Entscheidungsebenen

Abb. 3 zeigt eine Clusterbildung, wo man mit der eben beschriebenen Schwerpunkts-Methode auch nicht zum Ziel kommt. Es ist jedoch möglich, eine Gerade zu finden (im mehrdimensionalen Fall eine Hyperebene), die die beiden Klassen separiert. Diese Methode der automatischen Zeichenerkennung wurde als erste bei Massenspektren erprobt[7, 8, 9]. Die erforderlichen Entscheidungsebenen können — falls sie überhaupt existieren — mit einem Iterationsverfahren

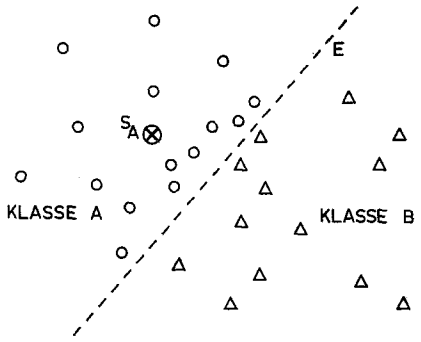

Abb. 3. Klassifizierung mit einer Entscheidungsebene E, welche die Klassen A und B vollständig separiert. Der Schwerpunkt S_A der Klasse A ist für eine Klassifizierung durch Abstandsmessung nicht geeignet

(in amerikanischen Veröffentlichungen als „learning machine" bezeichnet) gefunden werden, das mehrfach ausführlich beschrieben wurde[7, 10, 11]. Der Klassifikator enthält die Lage der Ebene in Form eines *Entscheidungsvektors;* ein Spektrum wird zur Klasse A oder B zugeordnet, je nachdem auf welcher Seite der Entscheidungsebene es liegt.

Tab. 1 enthält einige eigene Ergebnisse, die mit dieser Klassifikationsmethode erhalten wurden. Mit Hilfe von 250 Massenspektren (dem sogenannten *Trainingsset*) wurden Entscheidungsebenen berechnet, die jeweils 2 Klassen vollständig separieren (z. B. die Alkylbenzole von allen übrigen Verbindungen). Mit 250 anderen Massenspektren (dem sogenannten *Prediction Set*) wurde die Brauchbarkeit der Entscheidungsebenen geprüft. Jeder Klassifikator liefert für eine bestimmte chemische Struktur die Antworten „ja" oder „nein". Die in Tab. 1 angegebene Verläßlichkeit der Antwort gibt an, zu wieviel Prozent die entsprechende Antwort bei der Klassifikation des Predictionsets richtig war. Ausführliche Untersuchungen[12] zeigten, daß im Gegensatz zu früheren optimistischen Angaben[9, 10] nur wenige Struktur-

merkmale für dieses Klassifizierungsverfahren geeignet sind. Von insgesamt 31 untersuchten chemischen Strukturen hatten die in Tab. 1 angegebenen 6 Klassen eine Verläßlichkeit der Antworten von über 70%.

Klassifizierung mit Hilfe von Nachbarspektren

Bei einer wenig ausgeprägten Clusterbildung nach Abb. 4 ist keine einfache Entscheidungsfläche zur Trennung der beiden Klassen möglich. Man könnte zwar versuchen, durch Anwendung mehrerer Gera-

Tabelle 1. *Klassifizierung mit Entscheidungsebenen*

Chemische Struktur (Klasse)	Anteil der Klasse am Predictionset (%)	Verläßlichkeit der Antworten bei Klassifizierung des Predictionsets (%)	
		ja	nein
Alkylbenzol	6	75	99
Benzolring	22	91	97
>CO am Benzolring	5	71	99
—OH am Benzolring	4	80	99
Sauerstoff im Molekül	55	84	83
Stickstoff im Molekül	24	78	95

den oder durch eine Funktion höherer Ordnung eine Separierung durchzuführen, aber das ist mathematisch sehr aufwendig. Trotzdem kann man in dieser Verteilung jeden beliebigen Punkt als unbekannt annehmen und seine Klasse nach einem einfachen Verfahren bestimmen: Der jeweils am nächsten liegende Nachbar (der sogenannte 1. Nachbar) zeigt die richtige Klasse an. Diese gedanklich sehr einfache *Nachbarmethode* ist in der Theorie der automatischen Zeichenerkennung seit langem bekannt und eingehend untersucht[4, 13, 14].

Die Nachbarmethode kann auf dem Gebiet der Spektreninterpretation beim Vergleich eines unbekannten Spektrums mit einer Spektrenbibliothek angewendet werden[15]. Als Ähnlichkeitskriterium zweier Spektren dient der Abstand der Vektorspitzen im mehrdimensionalen Spektrenraum. Erste Versuche mit Massenspektren zeigten[16], daß der euklidische Abstand zwischen den Spektren tatsächlich in engem Zusammenhang mit der chemischen Ähnlichkeit der betreffenden Substanzen steht (d. h., je kleiner der Abstand, desto mehr Strukturmerkmale stimmen im allgemeinen überein). Weiters zeigte sich, daß der jeweils 1. Nachbar die beste Klassifizierungsfähigkeit besitzt.

Eine wesentliche Verbesserung der Ergebnisse wird erzielt, wenn *mehrere* Nachbarspektren zur Klassifizierung herangezogen werden, wobei die Majorität der Nachbarn über die Klassenzugehörigkeit des unbekannten Spektrums entscheidet. Von der 500 Massenspektren umfassenden Bibliothek wurde jedes einzelne Spektrum als unbekannt angenommen. Aus den jeweils restlichen 499 Spektren wurden die 6 am nächsten liegenden Nachbarspektren zur Klassifizierung

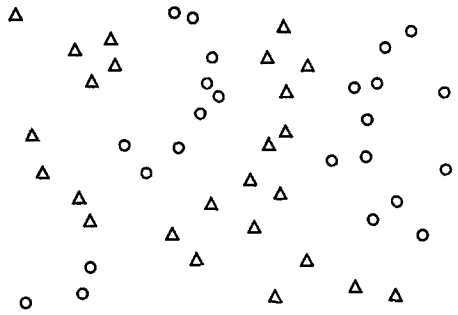

Abb. 4. In dieser Clusterbildung kann die Klasse jedes Punktes aus dem am nächsten liegenden Nachbarpunkt abgelesen werden

verwendet. Die in Tab. 2 zusammengefaßten Ergebnisse zeigen die hohe Erfolgsrate dieses Klassifizierungsverfahrens. Beispielsweise war beim Strukturmerkmal „Benzolring" die ja-Antwort des Klassifikators zu 96%, die nein-Antwort zu 99% richtig. In 2% (von insgesamt

Tabelle 2. *Klassifizierung mit 6 Nachbarspektren*

Chemische Struktur (Klasse)	Anteil der Klasse (%)	Verläßlichkeit der Antworten (%) ja	nein	Rückweisungen (%)
Alkan	8	75	99	3
Aliphatischer Alkohol	12	88	93	4
Benzolring	23	96	99	2
Pyridinring	7	77	97	3
>CO	28	86	79	9
—NR_2	14	97	92	2
—OH am Benzolring	4	82	98	1
—CH_3	78	85	73	9
—C_2H_5	40	67	82	19
Sauerstoff im Molekül	53	99	77	11
Stickstoff im Molekül	25	94	90	8

500 Spektren) endete die Abstimmung unter den 6 Nachbarn unentschieden und der Klassifikator reagierte mit einer Rückweisung. Zu bemerken ist, daß jedes Massenspektrum durch die Massenspektren von 6 anderen Substanzen klassifiziert wurde, da in der verwendeten Spektrenbibliothek jede Substanz nur einmal vorkommt. In der Praxis wird in vielen Fällen ein Massenspektrum der untersuchten Substanz in der Bibliothek enthalten sein.

Zusammenfassung

In der vorliegenden Arbeit wurden aus der Fülle von Methoden, die Theoretiker der automatischen Zeichenerkennung anbieten, einige kurz beschrieben. Leider gibt es keine erkennbaren Zusammenhänge zwischen Datenstruktur und zweckmäßiger Klassifizierungsmethode, so daß erst die Praxis zeigt, welche Methode für ein bestimmtes Problem nützlich ist. Die Methoden der automatischen Zeichenerkennung (und die entsprechenden Computerprogramme) sind so allgemein, daß sie für eine Vielzahl von Problemen anwendbar sind[5]. Wenn die Programme für verschiedene Methoden der automatischen Zeichenerkennung vorhanden sind, dann können sie bei allen größeren Datensammlungen routinemäßig ohne großen Aufwand eingesetzt werden. Einige der beschriebenen Methoden sind geeignet, aus den niedrig aufgelösten Massenspektren niedrigmolekularer Verbindungen automatisch direkte Hinweise auf die chemische Struktur der betreffenden Verbindungen zu liefern. Die automatische Spektreninterpretation wird aber den Chemiker bei der Strukturaufklärung und Substanzidentifizierung nicht ersetzen können; sie liefert zusätzliche Ergebnisse, die — nach kritischer Beurteilung — die Lösung einiger Probleme erleichtern können.

Daten und Programme

In dieser Arbeit wurden aus Spektrensammlungen[17, 18] 500 niedrig aufgelöste Massenspektren verwendet, die von 500 verschiedenen Substanzen mit Molekulargewichten zwischen 16 und 150 und den chemischen Formeln C_{1-11}, H_{1-22}, O_{0-2} und N_{0-2} stammten. Alle Programme wurden in FORTRAN IV geschrieben. Die Rechenarbeiten erfolgten an einem IBM 7040-Computer (Kernspeicher 32 K/36 bit) im Rechenzentrum der Techn. Hochschule Wien.

Herrn Prof. Dr. *A. Maschka* danke ich für seine freundliche Unterstützung dieser Arbeit. Herrn Dr. *P. Krenmayr* danke ich für wertvolle Diskussionen und Herrn *H. Urban* für seine tatkräftige Mithilfe.

Literatur

[1] *B. G. Buchanan, A. M. Duffield* und *A. V. Robertson*, in: Mass spectrometry-Techniques and applications (Hrsg. *W. A. Milne*), S. 121. New York: Wiley. 1971.

[2] *L. R. Crawford* und *J. D. Morrison*, Anal. Chem. **43**, 1790 (1971).

[3] *G. Meyer-Brötz* und *J. Schürmann*, Methoden der automatischen Zeichenerkennung. München: Oldenburg. 1970.

[4] *N. J. Nilsson*, Learning machines. New York: McGraw-Hill. 1965.

[5] *B. R. Kowalski* und *C. F. Bender*, J. Amer. Chem. Soc. **94**, 5632 (1972).

[6] *L. R. Crawford* und *J. D. Morrison*, Anal. Chem. **40**, 1469 (1968).

[7] *P. C. Jurs, B. R. Kowalski* und *T. L. Isenhour*, Anal. Chem. **41**, 21 (1969).

[8] *P. C. Jurs, B. R. Kowalski, T. L. Isenhour* und *C. N. Reilley*, Anal. Chem. **41**, 690 (1969).

[9] *P. C. Jurs, B. R. Kowalski, T. L. Isenhour* und *C. N. Reilley*, Anal. Chem. **42**, 1387 (1970).

[10] *T. L. Isenhour* und *P. C. Jurs*, Anal. Chem. **43**, 20 A (1971).

[11] *P. Krenmayr* und *K. Varmuza*, Allgem. Prakt. Chem. **23**, 289 (1972).

[12] *K. Varmuza* und *P. Krenmayr*, Z. Anal. Chem. **266**, 274 (1973).

[13] *T. M. Cover* und *P. E. Hart*, IEEE Trans. **IT 13**, 21 (1967).

[14] *W. S. Meisel*, Computer-oriented approaches to pattern recognition. New York-London: Academic Press. 1972.

[15] *B. R. Kowalski* und *C. F. Bender*, Anal. Chem. **44**, 1405 (1972).

[16] *K. Varmuza*, Z. Anal. Chem. (1973), im Druck.

[17] *E. Stenhagen, S. Abrahamsson* und *F. W. McLafferty*, Atlas of mass spectral data. New York: Interscience. 1969.

[18] *W. Benz*, Massenspektrometrie organischer Verbindungen. Frankfurt a. M.: Akad. Verlagsges. 1969.

Dr. K. Varmuza
Institut für Allgemeine Chemie
Technische Hochschule Wien
Lehárgasse 4
A-1060 Wien
Österreich