# Reactions in the Beilstein Information System: Nonaporic Organic Synthesis

MARTIN G. HICKS
Beilstein Institute, Varrentrappstrasse 40-42, 6000 Frankfurt 90, West Germany

Received July 15, 1990

This paper describes how chemical reactions are presented in the Beilstein Database. The Database will be briefly described, and examples of reactions given. A short summary of some of the methods of reaction indexing is given with a view to the future of the handling of reactions in a reaction database. To illustrate the present use of the database to search for reactions, the concept of nonaporic organic synthesis has been introduced.

## INTRODUCTION

The aim of every synthetic chemist is to design successful syntheses. The definition of success is often relative to the novelty. It is acceptable, at the first attempt, to produce a novel compound in low yields; at least one has made it. Optimization of the steps in a synthesis leads to a route to the compound in yields which make it usable. The synthetic chemist's nightmare is to have spent a lot of time and money getting to the last synthon only to find that the final reaction does not work and to have used up all of the compound in the attempt. These routes can be described as "aporic". In the literal translation aporia is a "pathless path".

A good description of an aporia is as follows:[1] one is walking along a mountain path, the path narrows, it suddely gives out, and one is stranded on a ledge, unable to go backward or forward. This provides a good analogy to a poor synthesis: starting off with 50 g of compound "A", this was converted to 5 g of compound "B", then after some effort 500 mg of compound "C" was produced, on the last step of the synthesis path to compound "D" the reaction does not work. The chemist is left standing on the cliff ledge with nothing but an empty sample tube. He cannot go backward, this is an aporic path; he has to start again.

Thus a successful chemist is one who follows nonaporic paths. How does he achieve this? Either he has experience in the area and knows how to get from "A" to "D" or he requires the equivalent of a route map and if none is available, he needs a guide.

A route map tells him where people have been before and which ways are successful. Using a route map, he is unlikely to find a brand new route, but perhaps a short cut. The *Beilstein Handbook* is our route map, in fact, it is the route map of nonaporic synthesis paths in the world of organic chemistry. Each route is carefully defined and set out for chemists to follow.

The Beilstein Database, with its substructure searching facility, allows the chemist easier access to analogous reactions than with the Handbook. By careful choice of substructures, routes taken from similar areas can be transposed to other areas.

What Beilstein does not have is a guide. Expert systems can provide this function. Suggestions of direction can be made. The programs associated with reaction databases can provide a stepwise guide toward the goal. The success of the present-day reaction databases underlines the need that the chemist has in receiving help in finding his way through the world of organic synthesis. Therefore, Beilstein is considering methods that will enhance our present methods of access to the chemical data within our database and maximize the use of the information to the chemist.

## THE DATABASE

Beilstein differs from bibliographic data collections in the way the compounds are handled. In bibliographic systems, the basic unit is the cited paper (or abstract thereof), and all compounds are referenced to that. Physical or chemical information is usually sparsely included, if at all, and in a manner difficult for search and retrieval. Beilstein on the other hand is a structurally oriented database. The basic unit is a compound, and all the chemical and physical data and bibliographic data are included in this unit.

The Database implementation began in 1985, and the first on-line file was available in December 1988 on STN-International. The database now contains nearly 3.5 million organic compounds from the time period 1830–1979.

The Database is built from three sources: the Handbook data H-EIV, which was abstracted directly from the Handbook; the internal file cards 1960–1979 (these are internal Beilstein abstracts used to assist the production of the Handbook); and, shortly, the abstraction of the primary literature 1980 onward will be becoming available. There is a logical division of the file into the full-file and short-file. The full-file contains the Handbook data, which has been evaluated and made nonredundant, and the short-file where this processing has not yet taken place. Thus, during the Handbook production, there will be a constant flow of data from the short-file to the full-file, with a consequent increase in quality.

## CHEMICAL STRUCTURES

The structures are the cornerstone of the Beilstein Database, so it was of vital importance to design a registry program which would be able to recognize like structures as like and discriminate between different structures. Such a system has been developed at the Beilstein Institute, the structures being stored and registered in Beilstein Registry Connection Tables (BRCT), which form subunits, one per component, of the Structure Distribution Format (SDF). The lists present in the BRCT are shown in Table I. This system has several special features. All multiple bonds are described in terms of the atom-centered descriptor of individual valence $\pi$-electrons; this removes the problems associated with coding and searching for aromatic bonds. All tautomers are individually registered without normalization—thus preserving full information integrity. The specific information dealing with the coding and normalization of the tautomers is stored in nonregisterable lists, only used by the search system. The full version of S4, the substructure search system developed by Beilstein/Softron,[2] allows the user to search for individual or generic tautomers. Tetrahedral stereocenters, double bonds, and allene axes, which correspond to most of the stereochemistry, are stored in terms of a parity code, which is an atom-index-based descriptor, able to unambiguously define the stereochemistry, not only for registration, but also for substructure searching. Other types of stereochemistry are stored in terms of the standard stereochemical descriptors and are, thus, not searchable by using a structure search.

It is possible to convert from the BRCT format to other formats as required. For example, to the CAS format required to build the file searchable on STN. Test files have also been

0095-2338/90/1630-0352$02.50/0 

Table I: Lists in the Beilstein Registry Connection Table (BRCT)—Version 2.01

| | | |
|---|---|---|
| HD | Header Vector | *Contains BRN and list lengths and codes.* |
| **Registerable Lists** | | |
| PI | Pi-bonding Vector | *Specifies for each atom the number of valence electrons which contribute to one or more pi-bonds.* |
| FR | From List. | *Specifies for each atom the index of the lowest indexed atom to which the current atom is attached.* |
| RC | Ring Closure List. | *The attachments constitute the ring-closure bonds.* |
| AT | Atom List. | *Specifies the location of non-carbon atoms.* |
| LH | Localized Hydrogen List. | *Specifies the number of non-tautomeric hydrogens attached to an atom.* |
| AS | Stereo Atom List. | *Specifies by means of a parity code based on the atom index the configuration of tetrahedral stereocenters.* |
| BS | Stereo bond List. | *Specifies by means of a parity code based on the atom index the configuration of double bond stereochemistry.* |
| XS | Stereo Axis List. | *Specifies by means of a parity code based on the atom index the configuration of allene chirality axes.* |
| LC | Localized Charge List.. | *Specifies the location and value of localized charges.* |
| DC | Delocalized Charge List. | *Specifies the value and the atoms over which the charge is delocalized.* |
| LR | Localized Radical List. | *Specfies the location of an unpaired valence electron.* |
| DR | Delocalized Radical List. | *Specifies the atoms over which the radical is delocalized.* |
| MA | Abnormal Mass List. | *Specifies the mass and atom index of those atoms with abnormal mass.* |
| MU | Abnormal Mass (Unknown Location) List. | *Specifies the mass and atomic number of elements with abnormal mass.* |
| HI | Hydrogen Isotope List. | *Specifies the location of D and T isotopes.* |
| HU | Hydrogen Isotope (Unknown Location) List. | *Specifies the mass of D and T isotopes at unknown location.* |
| PA | Stereo Poly Atom List. | *Specifies by means of a parity code the configuration of non-tetrahedral stereocenters.* |
| PB | Stereo Poly Bond List. | *Specifies by means of a parity code the configuration of enantiomeric double bonds.* |
| PX | Stereo Poly Axis List. | *Specifies by means of a parity code the configuration of each non-tetrahedral stereo axis.* |
| NS | Non-interpreted Stereo Descriptors List. | *Contains stereo descriptors which have not been interpreted in terms of a parity code.* |
| **Non-registerable Lists** | | |
| VA | Non-default Valence List. | *Specifies the atoms with a non-default valence* |
| TM | Tautomer Group List. | *Specifies the mobile tautomer species in a molecule.* |
| TL | Tautomer Group (localized) list. | *Specifies the origin atoms of the mobile tautomer species.* |
| AS | Graph Atom Coordinates List. | *Specifies the coordinates of each atom.* |
| BG | Graph Bond List. | *Specifies the orienatation of each out of plane bond.* |
| NG | Non-graph Atom List. | *Specifies the coordinates or bond orientation of non-graph atoms.* |
| FI | Fischer Atom List. | *Specifies the location of Fischer atoms.* |
| BO | Bond Type List. | *Specifies the bond type of each bond.* |
| NC | Normalized Charges List. | *Specifies the orignal location of charges eliminated or relocated during normalization.* |
| CA | CIP Stereo Atom List. | *Specifies the CIP descriptor for each stereo atom of known configuration.* |
| CB | CIP Stereo Bond List. | *Specifies the CIP descriptor for each stereogenic double bond.* |
| CX | CIP Stereo Axis List. | *Specifies the CIP descriptor for each allene type axis.* |
| ZG | Graph Atom Z-coordinate List. | *Specifies the Z-coordinate for each graph atom.* |
| ZN | Non-graph Atom Z-coordinate List. | *Specifies the Z-coordinate for each non-graph atom.* |
| LV | Stereocenter Ligand Parity List. | *Specifies the priority vector for each parity vector.* |
| HS | Hash Code List. | *Contains one or more hash codes for the molecule.* |
| ON | Original Numbering List. | *Specifies for each atom index the original numbering.* |
| OG | Original Stereo Descriptors List. | *Contains the original stereo descriptor information.* |
| SD | Supplementary Descriptors List. | *Contains descriptors, codes and text which cannot be described elsewhere.* |

converted into DARC, MACCS, and HTSS formats.

The BRCT describes structures fully, but is not able itself to describe reactions. The vehicle for describing reactions would be a modified version of the SDF format. In the present SDF format the interrelationships between individual compounds in a multicomponent system (e.g., a salt) can be described. It is planned to extend this methodology to describe reactions, with the inclusion of the necessary lists needed to hold the atom–atom mappings, reaction site, and transformation information.

## FACTUAL DATA

The factural data in the Beilstein Database are stored in three types of fields:

**Numeric Fields.** There are over 70 different factual fields. Each factual field can be divided into subfields to contain parameter data, temperature, pressure, etc.

**Boolean Fields.** These fields store the presence of a keyword or a parameter.

**String Fields.** Chemical names, literature citations, and comments are stored as strings.

The logical data hierachy of the Beilstein Database is shown in Table II.

Chemical reactions are stored in four fields in the Beilstein Database: preparation, chemical behavior, chemical derivative, and isolation from natural products.

The preparation field is by far the most important field for reactions in the data base. This field, whose subfields are shown in Table III, contains the description of a compound formation from its educts, with reagents and solvents, etc. These fields are the basic requirement to be able to build a reaction database.

The chemical behavior field, sometimes known as the reaction field, is only present when a reaction has been studied for a particular purpose, such as from the mechanism, rate, etc.

The chemical derivative field contains information concerning the formation of standard chemical derivatives (hydrazones, etc.) of the compound in question.

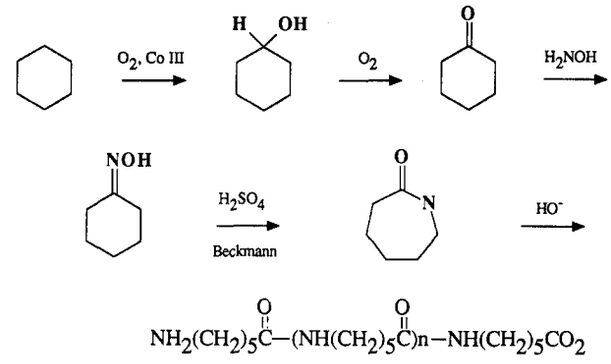
The isolation from natural products field contains a description of the method of isolation.

Table IV shows the number of occurrences of these fields for the short-, full-, and whole-file. Clearly the most important field is the preparation. The vast number of preparations that we have in the database 4824761 make it essential for us to develop a system which will enable the user to get to the information that he requires, and avoid being swamped by thousands of hits. Whereas for those reaction databases already in existence, which contain selected general examples, and are thus suitable for general queries, the queries put to a Beilsten Reaction Database would be sensibly more specialized in nature. The fact that Beilstein is oriented toward preparative organic chemistry is illustrated by the statistical analysis of the preparation data in Table V. It can be seen that 86% of all compounds have at least one preparation associated with them; giving on average the whole database, 1.4 preparations per compound.

Examination of the educt field carried out on 1 461 501 compounds gives the list of most frequent educts as shown in Table VI. There are no real surprises for the organic chemist who well knows the utility of bromine and acetic anhydride. This list does of course not give any information concerning the type of reaction or of its importance.

## EXAMPLES OF REACTIONS IN THE BEILSTEIN DATABASE.

Examples will be given in the form of some searches that a typical organic chemist would be interested in.




Nylon 6

**Figure 1.** Synthesis of Nylon 6.

BRN 106934 Beilstein
MF C6 H11 N O
CN hexahydro-azepin-2-one
Hexahydro-azepin-2-on
FW 113.16
SO 4-21-00-03196; 2-21-00-00216; 0-21-00-00240; 5-21
LN 25312
RN 105-60-2; 45604-12-4

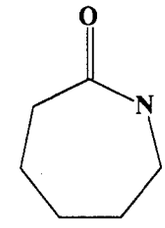



**Figure 2.** Display of caprolactam identification data.

The first example is that of caprolactam. Caprolactam is a very important chemical made in vast tonnage by the chemical industry. Its chief use is in the production of Nylon 6. When Nylon 6,6 was discovered and the route from adipic acid and hexamethylenediamine patented, other routes to similar polymers were investigated. The base-catalyzed polymerization of caprolactam provided a suitable alternative.

The synthesis, shown in Figure 1, starts from cyclohexane, which undergoes oxidation to cyclohexanol, which is subsequently oxidized to cyclohexanone. This reacts with hydroxylamine to give the cyclohexanone oxime, which undergoes a Beckmann rearrangement, to caprolactam.

A full structure search yields, as expected, one hit; the display is shown in Figure 2. To show how much information we have present in the Beilstein Database, a display of the field availabilities is shown in Table VII. This demonstrates that the Database is not just applicable for use by synthetic chemists but also by physical chemists and theoreticians.

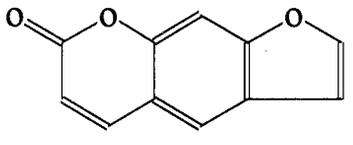
Searching for cyclohexanone oxime as the educt is shown in Figure 3 with the display of the first two hits showing some of the early patent and papers on the reaction.

The second example is that of furocoumarin.[3] A substructure search for the compound yields 492 hits with 399 having a preparation (see Figure 4). This list (L3) forms the starting point for further investigations.

Analyzing the parent in terms of the synthon approach provides, among others, two possible synthons; benzofuran and coumarin (see Figure 5). Searching for benzofuran as an educt and combining this with the list L3 gives 6 hits. One example is shown in Figure 6. Similarly, searching for coumarin as an educt and combining this with the original list gives 58 hits, so this seems to be the chemist's preferred route. An example is shown in Figure 7.

```
=> s cyclohexanone oxime/pre.edt

L3          13 CYCLOHEXANONE OXIME/PRE.EDT

=> s l2 and l3

L4          1 L2 AND L3

=> d hit


Preparation:
PRE
        Start:      cyclohexanone oxime
        Reag:       sulfuric acid
        Detail:     Isolierung durch Zersetzen des Reaktionsgemisches mit Alkalilauge in der Kaelte
        Reference(s):
        1.          Helferich,Malkomes, Chem.Ber. 55 <1922>,703, CODEN: CHBEAM
        2.          Marvel,Eck, Org.Synth. 17<1937>60, CODEN: ORSYAT Coll.Vol., II<1943>371
        3.          Ruzicka, Helv.Chim.Acta 4 <1921>,447, CODEN: HCACAV
        Note(s):
        4.          Handbook Data


=> d hit


Preparation:
PRE
        Start:      cyclohexanone oxime
        Reag:       sulfuric acid
        Temp:       150.0 Cel
        Detail:     im Luftstrom
        Reference(s):
        1.          Pat. No.: 755944, D.R.P., I.G. Farbenind., 1939 DRP/DRBP Org. Chem.: 1950-1951 6
                    1527
        2.          Pat. No.: 736735, D.R.P., I.G. Farbenind., 1939 DRP/DRBP Org. Chem.: 6 1321
        Note(s):
        3.          Handbook Data
```

**Figure 3.** Search for preparation of caprolactam from cyclohexanone oxime.



Furocoumarin

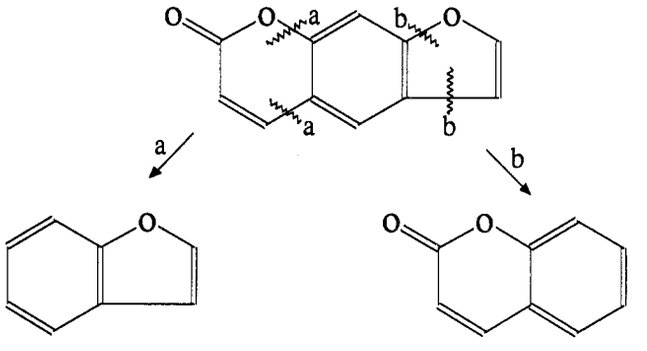
```
=>S L1 Full

L2   492  SEA SSS FUL L1

=>S L2 AND PRE/FA

L3   399  L2 AND PRE/FA
```

**Figure 4.** Search for furocoumarin.

The complexity of the Database means that these very simple searches demonstrated here are just at the very tip of the iceberg of possibilities. Experienced searchers will be able, with suitable combination of other searchable fields, to carry out sophisticated searches giving them just the results that they require. Good use of the Database will limit the unnecessary following of aporic synthesis paths.

## A BEILSTEIN REACTION DATABASE

The first part of this paper described the Beilstein Database as it now is and shows that the basic requirements for a reaction database are already met. The institute is presently carrying out a systems analysis with a view to the development of the software needed to index, register, store, and search reactions.




Search for synthon "a" - benzofuran

```
=>S L3 AND BENZOFURAN?/PRE.EDT

   L4   6    L3 AND BENZOFURAN?/PRE.EDT
```

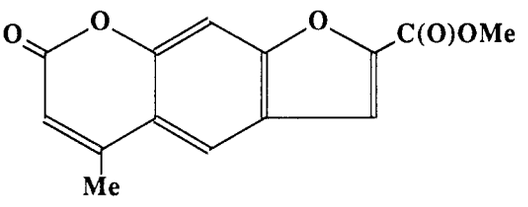
Search for synthon "b" - coumarin or chromene

```
=>S L3 AND (COUMARIN OR CHROMEN?)/PRE.EDT

   L5   58   L3 AND (COUMARIN OR CHROMEN?)/PRE.EDT
```

**Figure 5.** Search for furocoumarin synthons.

The structure search system will be based on S4, which is the system used at the Institute and for on-line searching on Dialog. The system has also demonstrated its capabilities for CD-ROM applications.[4]

BRN 273720 Beilstein
MF C14 H10 O5
CN 5-methyl-7-oxo-7H-furo<3,2-g>chromene-2-carboxylic acid methyl ester
5-Methyl-7-oxo-7H-furo<3,2-g>chromen-2-carbonsaure-methyl ester
FW 258.23
SO 4-19-00-03832
LN 23087; 289
RN 106738-73-2

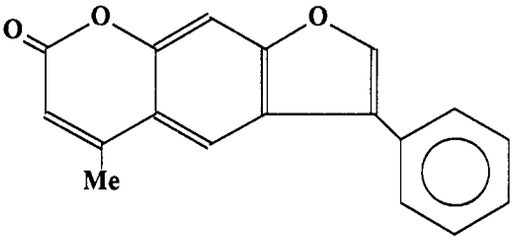



Preparation:
PRE

Educt: 6-hydroxy-benzofuran-2-carboxylic acid methyl ester,
3-chloro-cis-crotonic acid
Reag: polyphosphoric acid
Reference(s):
1. Dann, Illing, Leibigs Ann. Chem. 605 <1957> 146, 152 CODEN: LACHDL
Note(s):
1. Handbook Data

**Figure 6.** Display of a furocoumarin preparation from a benzofuran.

BRN 256589 Beilstein
MF C18 H12 O3
CN 5-methyl-3-phenyl-furo<3,2-g>chromene-7-one
5-Methyl-3-phenyl-furo<3,2-g>chromen-7-on
FW 276.29
SO 4-19-00-01890
LN 21922
RN 68454-22-8




Preparation:
PRE

Educt: 4-methyl-7-phenacyloxy-coumarin,
Reag: sodium ethylate, ethanol
Detail: anschliessendes Behandeln mit wss. Salzsaure
Reference(s):
1. Caporale, Antonello, Farmaco Ed. Sci. 13 <1958> 363, 366 CODEN: FRPSAX
Note(s):
1. Handbook Data

**Figure 7.** Display of a furocoumarin preparation from a coumarin.

Most of the present day reaction databases have two main software units required for the indexing of reactions: (1) the determination of the atom–atom correspondences between the educts and products with subsequent determination of the reaction site and hence transformation; (2) the coding of the reaction.

There have been many methods used to determine the atom–atom correspondences and hence reaction sites. One of the first methods was that of Lynch and Willett[5] who used extended connectivities to determine the maximal common subgraph (MCS). This method has been refined by McGregor and Willett.[6] Methods using superposition of structures have been implemented by Dubois[7] and Fujita.[8] The method of Weise[9] which uses the superposition of fragments not only gives the correspondences, but because of its synthon-type approach, has further applications for synthesis design. Wilcox and Levinson[10] developed a new approach to the finding of the MCS. Wipke and Rogers[11] have implemented analysis methods based on tree-structured graphs to find the MCS. A different approach is that of Dugundji and Ugi,[12] whose use

Table II: Logical Data Structure of the Beilstein Database

**Identification**
- Beilstein Registry Number
- CAS Reg.No.
- Lawson No.
- Chemical Name
- Synonym
- Beilstein Citation
- Molecular Formula
- Linear Structure Formula
- Multi-component System Mol. Form.
- Number of Charges
- Atom Counts
- Number of Atoms
- Element Counts
- Number of Components
- Molecular Weight
- Structure

**Chemical Data**
- Preparation
- Chemical Behaviour
- Isolation From Natural Products
- Derivative
- Elementary Analysis
- Purity
- Related Structure
- Purification

**Physical Data of Single Component Systems**

**Structure and Energy Parameters**
- Conformation
- Interatomic Distances and Angles
- Dipole Moment
- Quadrupole Moment
- Bond Moment
- Molar Polarization
- Electrical Polarizability
- Optical Anisotropy
- Coupling Constants
- Nuclear Quad. Coupling Consts.
- Molecular Deformation
- Rotational Constants
- Moment of Inertia
- Energy Barriers
- Molecular Energy
- Dissociation Energy
- Ionization Potential
- Affinity

**Physical State**
- Crystal Property Description
- Melting Point
- Crystal Phase
- Decomposition
- Sublimation
- Triple Point
- Transition Point Crystalline Mod.
- Crystal System
- Space Group
- Dimensions of the Unit Cell
- Density of the Crystal
- Boiling Point
- Liquid Phase
- Transition Points of Liquid Mod.
- Critical Temperature
- Critical Pressure
- Critical Density
- Critical Volume
- Association In the Gas Phase
- Vapour Pressure

**Other Physical & Mechanical Properties**
- Density
- Molar Volume
- Mechanical Properties
- Coefficient of Expansion
- Compressibility
- Ultrasonic Properties
- Surface Tension

**Transport Phenomena**
- Dynamic Viscosity
- Kinematic Viscosity
- Bulk Viscosity
- Self-diffussion
- Thermal Conductivity

**Calorific Data**
- Enthalpy of Combustion
- Enthalpy of Formation
- Enthalpy of Hydrogenation
- Enthalpy of Melting
- Enthalpy of Vaporization
- Enthalpy of Sublimation
- Enthalpies of Other Phase Trans.
- Entropy Delta S
- Entropy of Formation
- Heat Capacity cp
- Heat Capacity cpO
- Heat Capacity cv
- Zero-point Energy
- Gibbs Energy of Formation
- Calorific Data

**Optical Properties**
- Refractive Index
- Optics
- Optical Rotatory Power
- Mutarotation
- Circular Dichroism
- ORD

**Spectra**
- NMR Spectrum
- NMR Abs.
- NMR Data
- ESR Data
- Nuclear Quadrupole Resonance
- Rotational Spectrum
- IR Spectrum
- IR Bands
- Vibrational Spectrum
- Raman Spectrum
- Raman Bands
- Electronic Spectrum
- UV/VIS Spectrum
- Absorption Maxima
- Emission Spectrum
- Fluorescence Spectrum
- Fluorescence Maxima
- Phosphorescence Spectrum
- Phosphorescence Maxima
- Other Spectroscopic Methods
- Mass Spectrum

**Magnetic Properties**
- Magnetic Susceptibility
- Magnetic Data

**Electrical Properties**
- Static Dielectric Constant
- Dielectric Constant
- Electrical Data

**Electrochemical Behaviour**
- Electrochemical Behaviour
- Diss. Exponent
- Enthalpy of Dissociation (electrolytic)
- Isoelectric Point pH
- Redox Potential
- Polarographic Half-wave Potential

**Physical Data of Multi-component Systems**
- Solution Behaviour
- Solubility
- Solubility Product
- Solution Behaviour
- CMC (Critical Micelle Concentration)
- Liquid/Liquid Systems
- Liquid/Solid Systems
- Liquid/Vapour Systems
- Liquid/Vapour Data
- Azeotropes (components)
- Other Mechanical Properties
- Transport Phenomena
- Energy Data
- Boundary Surface Phenomena
- Adsorption
- Association

**Physiological Behaviour & Application**
- Use
- Toxicity
- Biological Function
- Ecological Data

**Bibliographic Information**
- Author
- Coden
- Journal

of reaction matrices was one of the first applications of mathematical methods in chemistry. This was later combined with Jochum et al.'s[13] principle of minimum chemical distance to find the optimal matrix.

After having defined the reaction centers, the reaction must be coded. While present-day reaction databases use similar methods, over the years other systems have been developed. One of the first systems was Gremas,[14] which uses substitution changes at an atom to describe a reaction. Hendrickson[15] in his SYNGEN system uses descriptors also based on substitution changes. Significant work has been carried out by Littler,[16] whose bond-change schemes have been adopted as an IUPAC standard. Zefirov[17] also has developed a scheme based on bond changes. Roberts[18] has defined reactions in

**Table III:** Preparation Data Fields

| subfields | field type |
|---|---|
| starting material | structure/string |
| reagent | string |
| solvent | string |
| catalyst | string |
| yield | numeric |
| time | string |
| temperature | numeric |
| pressure | numeric |
| reflux | y/n |
| room temperature | y/n |
| irradiation | y/n |
| byproducts | structure/string |
| other conditions | string |

**Table IV:** Reactions in the Beilstein Database

| field | | source | | |
|---|---|---|---|---|
| | | handbook | file cards | total |
| preparation | compounds | 844 829 | 2 276 211 | 2 957 198 |
| | occurrences | 1 290 018 | 3 534 743 | 4 824 761 |
| reaction | compounds | 143 599 | 328 966 | 441 288 |
| | occurrences | 458 913 | 964 813 | 1 423 726 |
| isolation | compounds | 6 888 | 30 991 | 35 587 |
| | occurrences | 20 553 | 53 356 | 73 909 |
| derivative | compounds | 26 189 | 67 178 | 89 501 |
| | occurrences | 39 883 | 93 618 | 133 501 |

**Table V:** Reaction Statistics

| | compounds | | occurrences | |
|---|---|---|---|---|
| total | 3 420 089 | | | |
| preparations | 2 957 198 | 86% | 4 824 761 | 1.4/cmpd |
| reactions | 441 288 | 13% | 1 423 726 | 0.4/cmpd |

**Table VI:** Most Frequent Educts[a]

| educt | frequency | educt | frequency |
|---|---|---|---|
| bromine | 21 294 | phenylmagnesium bromide | 4 131 |
| acetic anhydride | 16 639 | DNP | 4 110 |
| benzoyl chloride | 13 336 | benzene | 4 007 |
| ammonia | 12 170 | alcohol | 3 808 |
| aniline | 11 690 | thionyl chloride | 3 727 |
| benzaldehyde | 8 537 | phenyl isocyanate | 3 617 |
| chlorine | 8 135 | benzyl chloride | 3 320 |
| nitric acid | 7 418 | dimethylamine | 3 215 |
| methanol | 6 093 | *p*-toluidine | 2 952 |
| phenylhydrazine | 5 851 | methyl iodide | 2 828 |
| dimethyl sulfate | 5 251 | PCl5 | 2 718 |
| phenol | 4 466 | ethylmagnesium bromide | 2 588 |
| diazomethane | 4 141 | | |

[a] From 1 461 501 compounds.

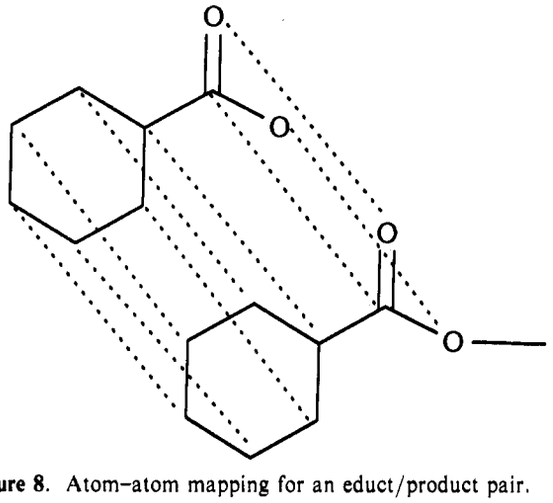
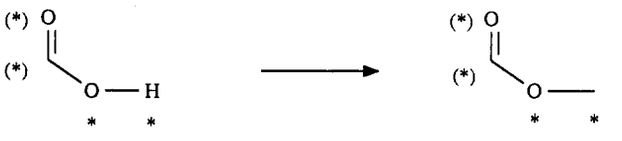


**Figure 8.** Atom–atom mapping for an educt/product pair.

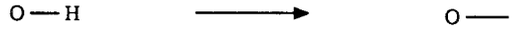
**Table VII:** Field Availability Display for Caprolactam

```
=> d fa

L2 ANSWER 1 OF 1
```

| Code | Field Name | Occur. |
|---|---|---|
| MF | Molecular Formula | 1 |
| CN | Chemical Name | 1 |
| FW | Formula Weight | 1 |
| SO | Beilstein Citation | 1 |
| LN | Lawson Number | 1 |
| PRE | Preparation | 77 |
| MP | Melting Point | 72 |
| BP | Boiling Point | 34 |
| SP | Sublimation Point | 1 |
| REA | Chemical Reaction | 179 |
| RSTR | Related Structure | 1 |
| INP | Isolation from Natural Product | 1 |
| PUR | Purification | 1 |
| CTCFM | Conformation | 2 |
| CTSKC | Skeletal Characteristics | 1 |
| DM | Dipole Moment | 3 |
| CTMEN | Molecular Energy | 1 |
| CTMEN | Molecular Energy | 1 |
| CTCRY | Crystal Phase | 2 |
| CTLIQ | Liquid Phase | 2 |
| VP | Vapour Pressure | 3 |
| DEN | Density (crystal) | 2 |
| CTMEC | Mechanical Properties | 1 |
| HFUS | Enthalpy of Fusion | 3 |
| CTCAL | Calorific Data | 3 |
| HSUB | Enthalpy of Sublimation | 2 |
| CP | Heat Capacity Cp | 2 |
| SREF | Entropy | 1 |
| HCOM | Enthalpy of Combustion | 1 |
| NMRS | NMR Spectrum | 1 |
| NMRA | NMR Absorption | 2 |
| CTNMR | NMR Data | 1 |
| CTESR | ESR Data | 1 |
| IRS | Infrared Spectrum | 15 |
| IRM | Infrared Maximum | 6 |
| CTVIB | Vibrational Spectrum | 2 |
| EAS | Electronic Absorption Spectrum | 3 |
| CTMS | Mass Spectrum | 2 |
| CTELE | Electrical Data | 2 |
| CTECB | Electrochemical Behaviour | 1 |
| DE | Dissociation Exponent | 7 |
| SLB | Solubility | 7 |
| CTLLSM | Liquid/Liquid Systems | 7 |
| CTLSSM | Liquid/Solid Systems | 5 |
| CTLVSM | Liquid/Vapour Systems | 6 |
| CTGASM | Gas Phase Behaviour | 1 |
| CTTRAM | Transport Phenomena | 2 |
| CTENEM | Energy of MCS | 1 |
| CTADSM | Adsorption | 2 |
| CTASSM | Association | 1 |
| CTUNCH | Unchecked Data | 12 |



Reaction Sites (in brackets influencing groups)



Reaction Transformation

**Figure 9.** Extracted reaction sites and transformation.

terms of concerted processes and thus electron changes. Electron changes at atoms are also the cornerstone of Dugundji–Ugi's[17] reaction matrix approach. Both Dubois[7] and Weise[9] code the transformations as connection tables. There are certain similarities in the approaches of Vladutz[19] and Fujita[8] in that they both describe a reaction in terms of a single connection table. Vladutz's superimposed reaction skeleton graph (SRSG) and Fujita's imaginary transition state (ITS) and reaction center graph (RCG) use special notation to stipulate the fate of bonds in a reaction. A different type of

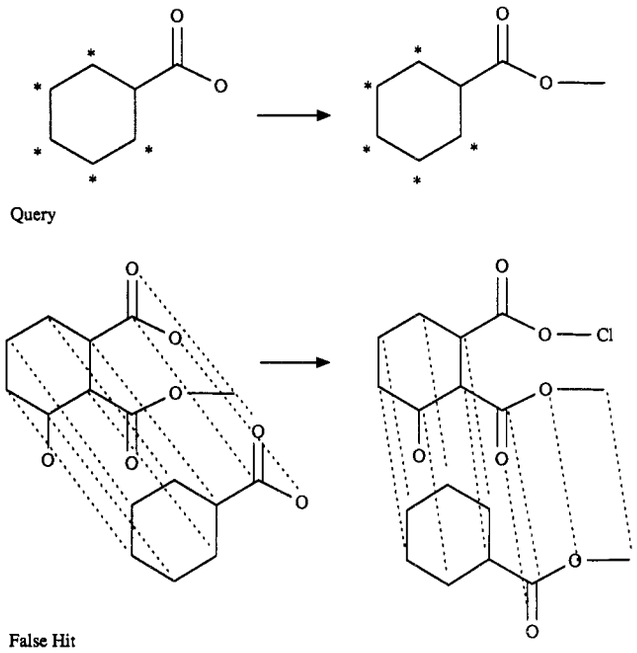



**Figure 10.** False drop for a reaction search carried out without atom–atom mapping.

representation based on a bond-centered graph is used by Wilcox and Levinson[10] to describe the whole reaction.

An example of the atom–atom mapping for the reaction of cyclohexanoic acid to methyl cyclohexanoate is shown in Figure 8. The extracted reaction site and transformation are shown in Figure 9. The bracketed atoms define the influencing groups that a chemist would also want to be included in the reaction site information but are not straightforwardly detectable by the above algorithmic methods.

The advantage of having the aforementioned information is to be able to define a reaction very precisely for registration and search purposes and thus be able to eliminate the retrieval of false drops. Such a false drop is shown in Figure 10. If just the substructures of the product and educt are searched, the hit in Figure 10 is retrieved. This is because both substructures can be mapped onto the retrieved structure even though one of them does not play a part in the retrieved reaction.

The problems which a database producer has are many fold, not only are the above software units required, but there are several other problems which these systems must overcome.

The systems must be able to check whether the reaction as defined in the input has correct stoichiometry or not, and if not, then it must be able to be automatically corrected. Stereochemistry and tautomerism must be correctly handled. The reaction sites must in all cases be automatically correctly found, no mean task. Not only that but it would be desirable to help the chemist and extend the reaction site, to not only include that determined from graph theory, but also be close to that as described by the chemist himself, thus influencing groups need to be taken into account. The searching for multistep syntheses especially across document boundaries has also not yet been solved. One possible solution for this has been shown by Lawson and Kallies.[20]

## CONCLUSIONS

The present Beilstein Database allows access to the over 4 million preparations. While this access is enough to allow retrieval of individual steps of nonaporic synthesis paths, any analogy searching is limited to very similar molecules. Development of a Beilstein Reaction Database, with the inherent advantages in precision and analogy searching, will bring another valuable tool to the chemist's bench.

## REFERENCES AND NOTES

(1) Lodge, D. *Nice Work*; Penguin: London, 1989.
(2) Hicks, M. G.; Jochum, C. Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191–199.
(3) *Online Searching of Beilstein on STN: How to Find Preparations and Reactions*; Springer-Verlag: New York, 1989.
(4) Hicks, M. G.; Jochum, C.; Maier, H. Substructure Search System for Large Chemical Databases. Proceedings of the XIth ICCCRE. *Anal. Chim. Acta*, in press.
(5) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 154–159.
(6) McGregor, J. J.; Willett, P. Use of Maximal Common Subgraph Algorithm in the Automatic Identification of the Ostensible Bond Changes Occurring in Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 137–140.
(7) Dubois, J.-E. Computer Assisted Modelling of Reactions and Reactivity. *Pure Appl. Chem.* **1981**, 53, 1313–1327.
(8) Fujita, S. Description of Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205–212.
(9) Weise, A. Computeranalyse chemischer Reaktionsgleichungen mit Stöchiometriekorrektur. *J. Prakt. Chem.* **1980**, *322* (5), 761–768.
(10) Wilcox, C. S.; Levinson, R. A. A Self-Organized Knowledge Base for Recall, Design and Discovery in Organic Synthesis. In *Artificial Intelligence Applications in Chemistry*; Pierce, T. H.; Hohne, B. A., Eds.; American Chemical Society Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 209–230.
(11) Wipke, W. T.; Rogers, D. Tree-Structured Maximal Common Subgraph Searching. An Example of Parallel Computation with a Single Sequential Processor. *Tetrahedron Comput. Methodol.* **1989**, *2* (3), 177–202.
(12) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.
(13) Jochum, C.; Gasteiger, J.; Ugi, I. Dugundji, J. The Principle of Minimum Chemical Distance and the Principle of Minimum Structure Change. *Z. Naturforsch.* **1982**, *37B*, 1205–1215.
(14) Fricke, C.; Fugmann, R.; Kusemann, G.; Nickelsen, T.; Ploss, G.; Winter, J. H. Experience with Reaction Indexing and Searching in the IDC System. In *Modern Approaches to Chemical Reaction Searching*; Willett, P. Ed.; Gower: Aldershot, England, 1986; pp 68–77.
(15) Hendrickson, J. B. A. Systematic Organization of Synthetic Reactions. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–136.
(16) Littler, J. S. The Detailed Linear Representation of Reaction Mechanisms. *Pure Appl. Chem.* **1989**, *61*, 57–81.
(17) Zefirov, N. S. An Approach to Systematization and Design of Organic Reactions. *Acc. Chem. Res.* **1987**, *20*, 237–243.
(18) Roberts, D. C. A Systematic Approach to the Classification and Nomenclature of Reaction Mechanisms. *J. Org. Chem.* **1978**, *43*, 1473–1480.
(19) Vladutz, G. Do We Still Need a Classification of Reactions? In *Modern Approaches to Chemical Reaction Searching*; Willett, P., Ed.; Gower: Aldershot, England, 1986; pp 202–220.
(20) Lawson, A. J.; Kallies, H. Multistep Reactions: The RABBIT Approach. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 426–430.