

The CAPEC Database[†]

Thomas L. Nielsen, Jens Abildskov, Peter M. Harper, Irene Papaeconomou, and Rafiqul Gani*

CAPEC, Department of Chemical Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark

The Computer-Aided Process Engineering Center (CAPEC) database of measured data was established with the aim to promote greater data exchange in the chemical engineering community. The target properties are pure component properties, mixture properties, and special drug solubility data. The database divides pure component properties into primary, secondary, and functional properties. Mixture properties are categorized in terms of the number of components in the mixture and the number of phases present. The compounds in the database have been classified on the basis of the functional groups in the compound. This classification makes the CAPEC database a very useful tool, for example, in the development of new property models, since properties of chemically similar compounds are easily obtained. A program with efficient search and retrieval functions of properties has been developed.

Introduction

Measured data are essential to process modeling and to the selection of strategies for solving process design problems. They also form the basis of property models. Indeed, data are of key importance, when performing almost any task in computer-aided process engineering. Major advances made in property estimation in the past decade stem from our ability to comprehensively assess the accuracy and reliability of various model formulations, on the basis of access to even greater numbers of measured data, obtained over wider ranges of conditions. Recent advances in theory, experiment, and molecular simulation have also expanded the options for property determinations for use in process design. Diverse options for making computations from molecular-level principles¹ are currently being pursued with high-speed computers. Such efforts will probably provide increased guidance about systems at extreme conditions, reaction kinetics details, complex molecules, and so forth. Thus, future process and product engineers will be able to utilize a greater richness of property model options. Yet, paying attention to measurements will probably continue to be important. In the past, reconciliation of new ideas with measurements appears to have formed the basis of most successful innovations. A trend in recent years has been that corporations sell subscriptions to large collections of data that are updated as new data appear. Such subscriptions are expensive and inflexible for the needs of educational institutions or small industrial engineering R & D groups. The CAPEC database project was initiated in 1998 with the objective to establish a flexible, easy to use and easy to maintain, database on measured properties on pure components and mixtures. CAPEC is committed to research work in close collaboration with industry and to participate in educational activities. The research objectives of CAPEC are development of computer-aided systems for process simulation, design, analysis, and control/operation for chemical, petrochemical, pharmaceutical, and biochemical industries. In recent years, the "WebBook" of NIST² and Camsoft³ "Chemfinder"

on pure component data have initiated data exchange facilities by providing web-access to their databases. Other well-established data collections are those of DIPPR,⁴ Dortmund Data Bank,⁵ and the TRC.⁶

Classification of Properties

Pure Component Properties. The database includes approximately 13 000 compounds. When available, the following pure component properties are given: acentric factor, critical temperature, critical pressure, critical volume, critical compressibility, melting point, boiling point, triple-point temperature and pressure, boiling point at specified pressure, liquid volume at 298.15 K, ideal gas enthalpy at 298.15 K, ideal gas Gibbs energy at 298.15 K, ideal gas entropy at 298.15 K, density, solubility parameters, van der Waals surface area and volume, radius of gyration, dipole moment, octanol/water partition coefficient, refractive index, molecular refraction, enthalpy of fusion, enthalpy of combustion and flash point temperature, relative permittivity. Only experimental values are stored in the database. Properties are divided into primary, secondary, and functional properties. Primary properties are defined as the properties that can be estimated on the basis of molecular structure only. Examples are critical temperature, critical pressure, and normal boiling point, as estimated with the method of Constantinou and Gani.⁷ Secondary properties are properties that can be estimated on the basis of molecular structure and values of one or more primary properties. Properties such as the ideal gas enthalpy of formation at 298.15 K, liquid density at 298.15 K, or boiling point at a specified pressure (reduced pressure boiling point) are also classified as primary or secondary properties, since the intensive variables are fixed. The influence of pressure on liquid densities is neglected here. The characteristic feature of primary and secondary properties is that they are accurately specified by a single-value constant. Functional properties, on the other hand, depend on temperature, pressure, or both. Functional properties are usually represented by a parametrized mathematical correlation along with numerical parameter values that mimic the property behavior over a range of conditions. For example, vapor pressure is commonly represented by a set of Antoine parameters instead of raw (T , P^{sat}) data

[†] This contribution will be part of a special print edition containing papers presented at the Fourteenth Symposium on Thermophysical Properties, Boulder, CO, June 25–30, 2000.

* E-mail: rag@kt.dtu.dk.

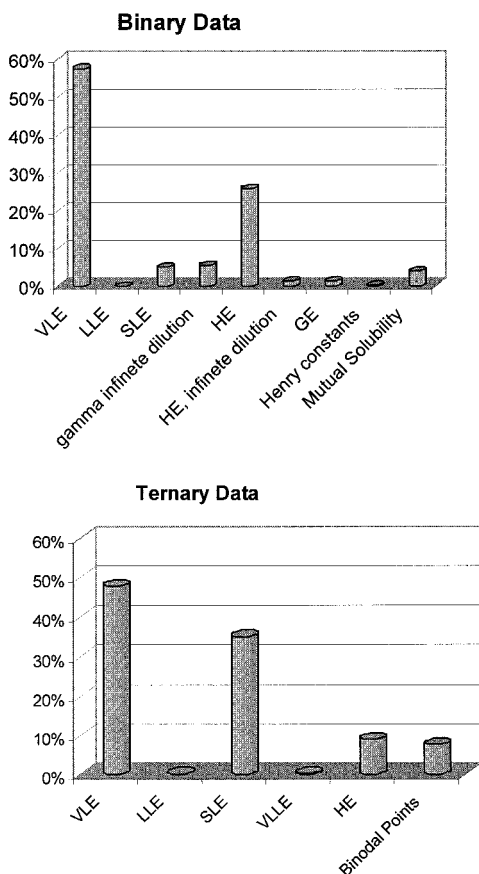


Figure 1. Distribution of collected mixture data among the different categories of data.

points. Therefore, functional properties are described as numerical values of correlation parameters that have been fitted to the measured data. Examples of functional properties in the database are ideal gas heat capacities, solid heat capacities, and second virial coefficients.

Mixture Properties. The mixture properties have been divided into two main categories, namely binary and ternary data. Experimental data for quaternary or higher properties are rare and are therefore not considered in the first version of the database. Binary data have been collected on vapor–liquid equilibrium (VLE), liquid–liquid equilibrium (LLE), solid–liquid equilibrium (SLE), infinite dilution activity coefficients, enthalpies of mixing, partial molar enthalpies of mixing at infinite dilution, excess Gibbs energies, Henry’s law constants, and mutual solubilities. Collected ternary mixture data comprise VLE, LLE, SLE, VLE, enthalpies of mixing, and binodal data. The data with the original references were stored. Approximately 41 000 binary and 10 000 ternary data points have been collected, and Figure 1 shows how the data are distributed in the different categories. It is noted that VLE data is the largest class of data in the database for both binary and ternary data.

The mixture data were carefully evaluated for errors such as obvious outliers. If any data seemed suspicious, cross-checks were made with the original reference. Perhaps unexpectedly, this procedure often revealed that the errors could be traced back to the original data. This is not a test for thermodynamic consistency but merely a check for obvious errors such as errors in typing when transcribing the data from paper to the electronic version.

Special Drug Solubility Data. This class of data covers solubilities of drugs such as steroids, penicillins, and

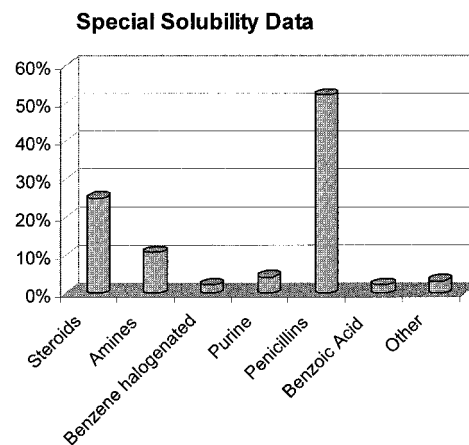


Figure 2. Distribution of special solubility data in the seven main categories.

Solubility Indicators with Water as Solvent

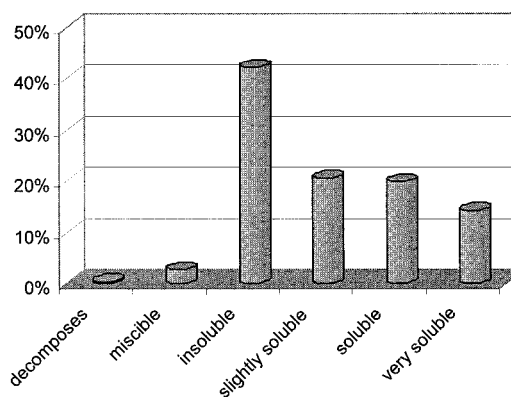


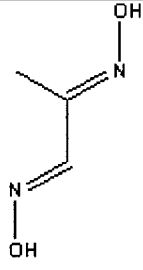
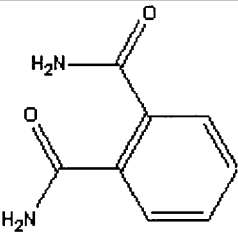
Figure 3. Distribution of solutes with respect to water as solvent.

amino acids. The data have been divided into seven categories: steroids, amines, benzene halogenated compounds, purine, penicillins, and other. This type of data is found in pharmaceutical journals, as not many traditional engineering journals publish such data to a large extent. Presently, the drug solubility database contains approximately 1500 data points covering 233 solutes and 90 solvents. As shown by Figure 2 most of the solubility data are found for the penicillins and steroids categories. In addition to the solubility data, a solvent list with solubility indicators (decomposes, miscible, insoluble, slightly soluble, soluble, and very soluble) has been constructed for most of the 13 000 compounds in the database. For instance, Figure 3 shows the distribution of solutes in the six categories when water is the selected solvent.

Classification of Compounds

Besides the classification of properties, the compounds in the database are classified in nine main categories such as polar compounds, nonassociating compounds, electrolytes, and steroids. Each main category is divided into several subcategories to gain further information about the functional groups in the compound. For instance, “Esters” and “Nitro” are two subcategories in “Polar Non-Associating Compounds” and “Alcohols” and “Amides” are two subcategories in “Polar Associating Compounds”. An overview of the nine different main categories is given in Figure 4, where for illustration purposes the subcategories of “Polar Associating Compounds” are shown. This classification of compounds facilitates easy and efficient search and retrieval of the properties of special families of compounds.

Table 1. Classification of 2-Hydroxyimino Propanol Oxime and 1,2-Benzenedicarboxamide

	
Name: 2-(Hydroxyimino) propanol oxime	Name: 1,2-Benzenedicarboxamide
Classified: "Polar Associating Comp."	Classified: "Polar Associating Comp."
Subcategory: "Oximes"	Subcategory: "Amides"

Compound Classification

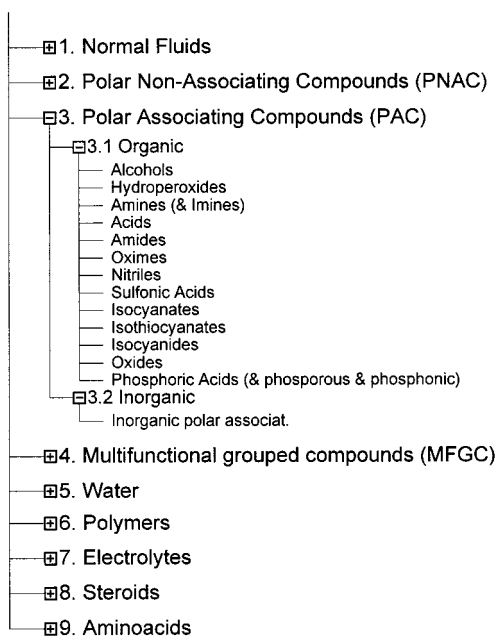
**Figure 4.** Nine main categories for the classification of compounds.

Table 1 shows an example of the classification of 2-(hydroxyimino) propanol oxime and 1,2-benzenedicarboxamide. On the basis of the structure and functional oxime groups ($=N-OH$), 2-(hydroxyimino) propanol oxime is classified as a "Polar Associating Compound" and subcategorized as an "Oxime" compound. 1,2-Benzenedicarboxamide is also classified as a "Polar Associating Compound" but is subcategorized as "Amides" due to the functional amide groups. Figure 5 shows how the compounds in the database distribute in the nine main categories. It is seen that more than one-third of the 13 000 compounds are multifunctional-grouped compounds. This underlines the necessity of research on property models that are able to handle multifunctional systems.

Database Structure

A database engine has been made to browse/search and retrieve data that are stored in a Microsoft Access database. An overview of the program structure is shown in

Distribution of Compounds in the 9 Classification Categories

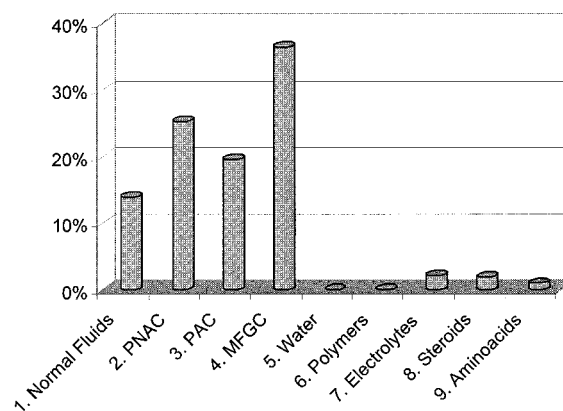
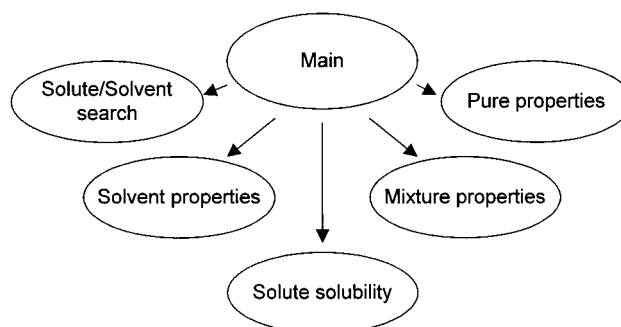
**Figure 5.** Distribution of the compounds in the CAPEC database.**Figure 6.** CAPEC database structure.

Figure 6. From the main interface one can access the different data categories; for example, pure component properties can be viewed by entering the "Pure Properties" section, and mixture properties and the references can be viewed by entering the "Mixture Properties" section. For the pure component properties and mixture properties the search engine includes analysis tools that provide some guidance to the accuracy of the data. The search for special solubility data is performed in the "Solute Solubility" section, where a list of solvents and the solubility data are shown when a solute is specified. If entering the "Solvent Properties", a list of solvents and the solubility indicators are shown for a specified solute. The last section is the "Solute/Solvents search", where dynamic queries can be set up to search for solvent or solute candidates that satisfy a

certain physicochemical property behavior. For example, the search for a solvent with a melting point greater than 350 K and a solubility parameter between 18.0 MPa and 20.0 MPa can easily be accomplished. In this example naphthalene is one of many possible solvent candidates with a melting point at 353.35 K and the solubility parameter 19.45 MPa. The CAPEC database will allow the user to add his/her own data to the original database. It is also possible to delete/change the data added by the user but not the original data belonging to the CAPEC database. In this way, all users of the CAPEC database will share a common database while, at the same time, each user will have his or her own in-house/confidential data for his or her own use.

Discussion and Conclusion

The CAPEC database project has been initiated, and a significant amount of data has been collected including pure component properties, mixture properties, and special drug solubility data. The database will continue to grow as new data appear in the literature, and especially the mixture properties are the target for future growth. The classification of compounds is very useful in the development of property models such as Constantinou and Gani's method for the prediction of pure component properties⁷ or UNIFAC⁸ and ASOG⁹ methods for the prediction of liquid activity coefficients. With the appropriate tools (currently under development at CAPEC) these methods could easily be extended to handle new families of compounds/mixtures. Also, special group contribution tables could be generated to handle special families of compounds/mixtures to increase the accuracy of the predictions compared to the predictions with the original group contribution tables. These are just some of many examples of the usefulness of the classification of the compounds. Also, the

CAPEC database and the database engine can easily be accessed and used by external programs such as process simulators and other Computer-Aided Process Engineering (CAPE) programs. The CAPEC database is available, free of charge, for educational use and for members of the CAPEC consortium.

Acknowledgment

The authors thank Jakob M. Harper and the master students at CAPEC for contributing to the development of the CAPEC database.

Literature Cited

- (1) Gubbins, K. E. The Future of Thermodynamics. *Chem. Eng. Prog.* **1989**, Feb, 38–49.
- (2) The NIST Chemistry Webbook can be found at <http://webbook.nist.gov/>.
- (3) Chemfinder can be found at <http://www.chemfinder.com/>.
- (4) Information about the DIPPR database project can be found at <http://www.aiche.org/dippr/>.
- (5) Information about the DECHEMA database can be found at <http://www.dechema.com/>.
- (6) *TRC Thermodynamic Tables—Hydrocarbons and Nonhydrocarbons*; Thermodynamics Research Center, Texas A&M University: College Station, TX, 1986.
- (7) Constantinou, L.; Gani, R. A New Group-Contribution Method for the Estimation of Properties of Pure Compounds. *AIChE J.* **1994**, *40*, 1697–1710.
- (8) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086–1099.
- (9) Palmer, D. A. Predicting Equilibrium Relationships for Maverick Mixtures. *Chem. Eng.* **1975**, June, 80–85.

Received for review August 4, 2000. Accepted December 20, 2000. The authors wish to thank Mitsubishi Chemical, Japan, and Union Carbide, USA, for financial support.

JE000244Z