

ThermInfo: Collecting, Retrieving, and Estimating Reliable Thermochemical Data

Ana L. Teixeira,^{†‡,*} Rui C. Santos,[‡] João P. Leal,^{‡,§} José A. Martinho Simões,[‡] and Andre O. Falcao,[†]

[†] LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal; [‡] Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal; [§] Unidade de Ciências Químicas e Radiofarmacêuticas, Instituto Tecnológico e Nuclear, Instituto Superior Técnico, Universidade Técnica de Lisboa, 2686-953 Sacavém, Portugal.

ABSTRACT.

Standard enthalpies of formation are used for assessing the efficiency and safety of chemical processes in the chemical industry. However, the number of compounds for which the enthalpies of formation are available is many orders of magnitude smaller than the number of known compounds. Thermochemical data prediction methods are therefore clearly needed. Several commercial and free chemical databases are currently available, the NIST WebBook being the most used free source.

To overcome this problem a cheminformatics system was designed and built with two main objectives in mind: collecting and retrieving critically evaluated thermochemical values, and estimating new data. In its present version, by using cheminformatics techniques, ThermInfo allows the retrieval of the value of a thermochemical property, such as a gas-phase standard enthalpy of formation, by inputting, for example, the molecular structure or the name of a compound. The same inputs can also be used to estimate data (presently restricted to non-polycyclic hydrocarbons) by using the Extended Laidler Bond Additivity (ELBA) method. The information system is publicly available at <http://www.therminfo.com> or <http://therminfo.lasige.di.fc.ul.pt>. ThermInfo's strength lies in the data quality, availability (free access), search capabilities, and, in particular, prediction ability, based on a user-friendly interface that accepts inputs in several formats.

Keywords: Information System, Thermochemical Properties, Web Database, Standard Reference Data, Organic Compounds.

* Corresponding author.

E-mail address: ateixeira@lasige.di.fc.ul.pt

INTRODUCTION

Chemical information keeps growing fast [1,2]. According to the Chemical Abstracts Service [3], there are currently more than 66 million known organic and inorganic substances and approximately 12,000 new entries are added daily. Cheminformatics plays, therefore, an increasingly important role, not only in archiving and retrieving information but also in implementing property estimation methods through suitable software.

There is a number of databases covering a variety of chemical and physical properties [4]. Some are hard copies (predominantly the oldest ones), other are electronic, and several have both formats [5-8]. The present trend, however, is creating either public (free access) or commercial web-based information systems [9-23]. The advantages and disadvantages of both options have been discussed by some authors [24-29].

Thermodynamic values are of major importance in chemistry and chemical engineering [30]. However, the main repositories for those values are printed handbooks and literature reviews [31-47] and only a minority of the web-based information systems dedicated to thermodynamic properties provide full[11,16,48-50] or limited[12,16,18,20] free access to their database.

Standard enthalpy of formation values are essential for assessing the efficiency and safety of any chemical process [51]. Enthalpies of formation can be obtained experimentally but this often requires complex and expensive methodologies [52]. *Ab initio* methods can also be reliable tools to determine enthalpies of formation [53,54]. However, the most accurate quantum chemistry methods are still computationally too expensive for large molecules [55]. Moreover, these methods are often validated with small molecules and their validity for large molecules is assumed. To avoid this problem, the theoretical methods are often used together with homodesmotic reactions [56], which are designed to maximize error cancellation.

Despite the work of experimental and theoretical chemists, it is thus not surprising that the number of compounds for which the enthalpies of formation are available is many orders of magnitude smaller than the number of known compounds. Thermochemical data prediction methods are therefore badly needed [24].

One of the most accurate ways to estimate enthalpies of formation relies on additivity methods [31,37,40,57-63], such as the so-called Benson group method (initially proposed in 1958), whose parameters (group enthalpies) have been refined and extended by several authors [57,58]. Another additivity method that is simple to use and very attractive to chemists (since it deals with bond enthalpies), is the Laidler method [63]. In this scheme the main parameters are assigned to the chemical bonds, and therefore should reflect the strengths of those bonds. The Laidler method has been recently refined for a variety of hydrocarbon families [59,60] and this new parameterization, called the Extended Laidler Bond Additivity (ELBA) method, improves the reliability of estimates (even for very large and bulky compounds) and extends the method to new families of hydrocarbon compounds.

As explained elsewhere [60], the prediction of accurate data by the ELBA method demands a large number of empirical parameters. Moreover, all the important parameters, including non-bonded interactions, must be recognized *a priori*, in order to obtain accurate predictions. The identification of all the parameters is a time-consuming task and often requires some familiarity with the prediction method. These disadvantages can only be removed by implementing an application suitable for non-expert users.

There are several computer and web-based applications based on empirical additivity schemes for estimating thermochemical data of organic compounds. Examples of the former include THERM/EST[64], CHETAH[51], DIPPR 801/DIADEM[65], DETHERM Software Suite[66], and NIST ThermoData Engine[67]. The latter applications include S&P[68], DDBSP-ARTIST[69], and DDBST-UNIFAC[70].

THERM/EST (NIST Estimation of Thermodynamic Properties for Organic Compounds) is a computer program developed to estimate thermodynamic data based on the principles of group additivity developed by Benson et al.[57] and later extended by Domalski et al.[58]. The commercial version of this software was discontinued [64]. Another application is the NIST Structures and Properties program [68], which also features an implementation of Benson's group additivity method [57], using a graphic interface that allows the user to draw the molecule and estimate several thermochemical properties. The commercial version of this software was also discontinued but it is still available at the NIST WebBook site [68]. Another NIST product, the ThermoData Engine software, provides critically evaluated thermodynamic and transport property data [67]. Finally, the ASTM Computer Program for Chemical Thermodynamics and Energy Release Evaluation (CHETAH)[51] software is a commercial tool for predicting thermochemical properties and certain hazards associated with a chemical compound or a reaction. This is accomplished through the knowledge of the molecular structure(s) of the components involved, and once again by an implementation of Benson's group additivity method [57].

Having in mind all the strengths and limitations of both the offline and online applications we set off to develop a free access information system, which we called ThermInfo, with the following features: (1) it should be user friendly; (2) all the data should be critically evaluated, either by ourselves or by a trusted source; (3) the data should include values for a wide range of compounds, viz. long-lived and transient organic, inorganic, and organometallic molecules in the gas- and condensed-phases; (4) a variety of empirical methods, selected on the basis of their reliability to predict data, would be included; (5) the parameterization of those empirical methods should be easy to recalculate on the basis of new data. In addition, we would pursue the search of new estimation procedures, based on structure-energy relationships and machine learning methods.

The initial ThermInfo dataset was based on Pedley's compilations of organic compounds properties [40,41]. The ELBA method (presently restricted to non-polycyclic hydrocarbons) [60], was selected for thermochemical prediction. This method yields reliable estimates but, as stressed above, it is only suited to computational use due to the large number of parameters involved.

The following section of this manuscript describes the database, the data assessment, the method used to predict properties, and the system basic architecture and requirements. Then the database implementation details as well as the technology options to implement the system are explained. After describing ThermInfo's system architecture and its dissemination facilities, the main conclusions and directions for future work are presented.

CONSTRUCTION AND CONTENT

Available data and their assessment

As mentioned before, the ThermInfo starting point was the experimental thermochemical database taken from two critically evaluated compilations by Pedley and coworkers,[40,41] together with the associated structural data. Although that thermochemical database is not error-free, it has the important advantage of being thermodynamically consistent, while some others are mere compilations. Detected faults will be addressed in future revisions, although they have a negligible impact on the ELBA parameters (see below).

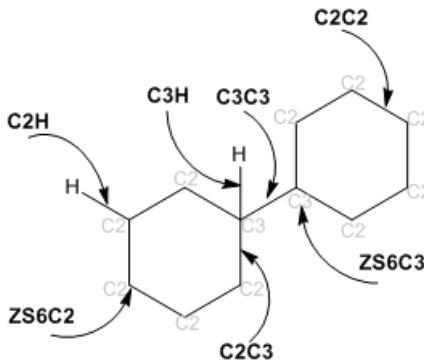
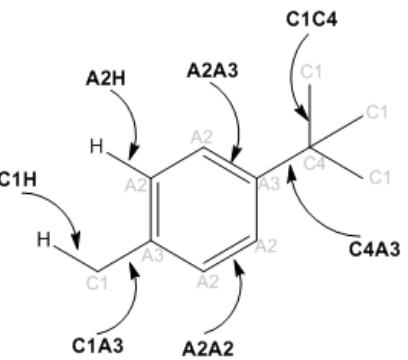
New databases and individual values will be added in the near future. The procedure will be as follows:

New data may come from other compilations or from the primary literature. In both cases, a critical evaluation of the values will be performed. Checks of the structural information will also be performed, based on different sources. When several values are available for a given property of the same compound, a single value will be chosen to be the "recommended value", although the rest of the values will be stored in the database. This choice will rely on several criteria, such as the experimental method(s) used, and also on consistency tests performed by applying available estimation methods and/or high-level quantum chemistry calculations.

The ELBA method and its implementation

The ELBA version used in ThermInfo includes a set of 165 parameters [60]. Each one of these new parameters has an assigned physical meaning (i.e. they are not fudge parameters). An example of the method, illustrating the calculation of gas- and liquid-phase standard enthalpies of formation, is presented in Table 1.

Table 1. ELBA parameters required to estimate the gas- and liquid-phase standard enthalpies of formation at T=298.15 K (kJ mol⁻¹) of bicyclohexyl and 4-methyl-1-*tert*-butylbenzene.

Compound name	bicyclohexyl	4-methyl-1- <i>tert</i> -butylbenzene
SMILES	C1(C2CCCCC2)CCCCC1	CC(C1=CC=C(C)=C1)(C)C
Compound structure		
Sum of ELBA parameters	8 C2C2 + 4 C2C3 + 1 C3C3 + 20 C2H + 2 C3H + 10 ZS6C2 + 2 ZS6C3	3 C1C4 + 12 C1H + 1 C1A3 + 1 C4A3 + 2 A2A2 + 4 A2A3 + 4 A2H
$\Delta_f H_m^\circ$ (g) (kJ mol ⁻¹)	Experimental 215.7±1.5 ^a Estimated 214.6	Experimental 57.6±1.0 ^b Estimated 56.0
$\Delta_f H_m^\circ$ (l) (kJ mol ⁻¹)	Experimental 273.7±1.4 ^a Estimated 274.0	Experimental 109.7±0.9 ^b Estimated 108.4

^a Reference [40]; ^b Verevkin, S. P. Thermochemical properties of branched alkylsubstituted benzenes. *J. Chem. Thermodyn.* **1998**, *30*, 1029–1040.

The automatic generation of structural descriptors from a computer-readable representation of a chemical structure diagram was not a straightforward task. These representations are analyzed automatically to derive the frequencies of occurrence of the ELBA parameters using a procedure that iterates the molecular structure (bonds and atoms) and extract the 165 ELBA parameters based on a set of structural characteristics (viz. the number of atoms, number of bonds, number of single, double, triple and aromatic bonds, number of rings, atom multiplicity, bond order, atom environment, maximum and minimum bond order, ring size, and cis/trans configurations) and a set of interrelated rules (Figure 1). The structure of the compounds is encoded in the widely used notations in chemical information systems, the SMILES and 3D MDL MOL File [71-73]. To validate ELBA implementation, along with other extensive tests, enthalpies of formation were manually calculated by the ELBA method for a set of more than 450 compounds and compared with the data computed by ThermInfo.

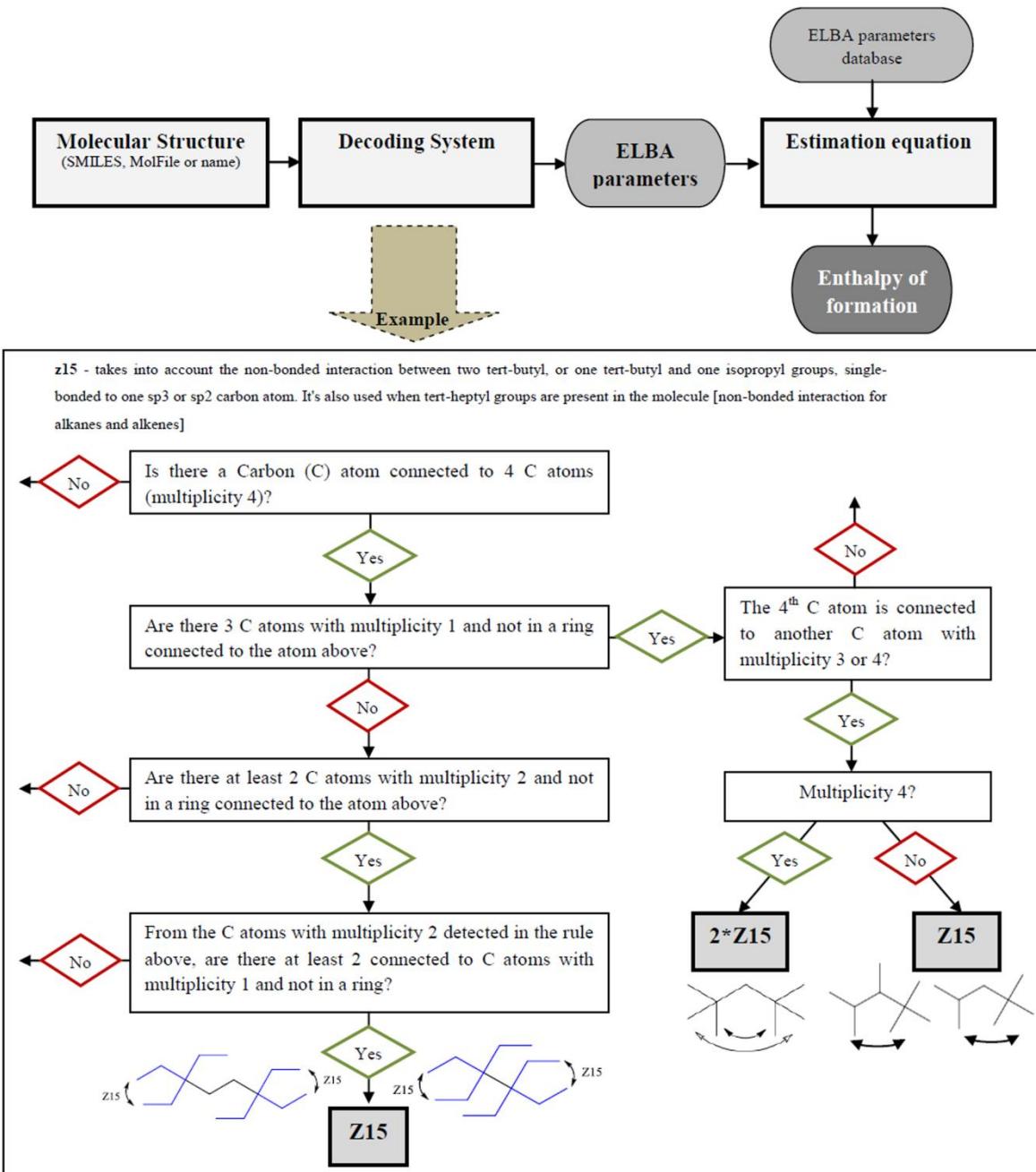


Figure 1 Diagram representing the implementation of the ELBA method. The steps needed to calculate the enthalpy of a compound with a specific example of the decoding system for the parameter z15.

System architecture and requirements

We have designed a relational database to accommodate and organize the heterogeneous data in a way that is compatible with our goals of efficient loading, updating, querying, and eliminating redundancy. The relational model was chosen because it has a good

performance and powerful querying of data, using a high-level language, Structured Query Language (SQL) [74]. In addition, it is easy to administer, has wide acceptance, and is well documented, which facilitates the use and development of applications that work on the database [75,76]. To allow a quick, easy, and intuitive access to the data, we have designed a user-friendly web tool based on three main requirements: the visual design, the interaction design, and the functionality. In the development process, we are performing multiple interactions with the following steps: problem analysis and project planning; requirements analysis; system modelling; system implementation; system evaluation; and system maintenance [77-79]. All steps were monitored by users in order to assess the system usefulness, design, and workability [80]. The client/server architecture was adopted to design the system [77]. On the client side, the browser is used for controlling user's input/output and communicating with the web server. The server side is responsible for manipulating the data, communicating with the database and retrieving the data requested from the client side.

UTILITY AND DISCUSSION

Data preprocessing and database

The database has been loaded with data collected by the Molecular Energetics Group (Centro de Química e Bioquímica) [76,81]. The database fields can be divided into three categories: structural data, thermochemical data, and the corresponding references (see *Supporting Information*, Table S1).

The database consists of twelve tables segmented into five logical categories of data: structural, thermochemical, references, registered users, and database statistics/evolution. The statistics of the database in June, 2012 are displayed in Table 2 and evidence the completeness and representativeness of the data set. Currently, it contains a completely non-redundant set of 2,956 organic compounds, corresponding to more than 15,500 different compound names and synonyms. The analysis of Table 2 shows also that the 2,956 compounds are well characterized in terms of structural properties and are divided into 6 classes, 12 subclasses, and 387 families, according to their constitution and structural arrangement. All the compounds in the current version of the database are characterized with at least one thermochemical property. The database contains 4,687 values for the standard molar enthalpy of formation (crystalline, liquid, and gas phases) and 1,790 values for the standard molar enthalpy of phase change (fusion, vaporization, and sublimation). The property data include the experimental uncertainties [82,83]. The size of ThermInfo dataset is comparable to other widely used databases. The NIST WebBook (2011 version) [49], for instance, contains more than 48,000 organic and inorganic compounds but only about 6,000 have at least a value of enthalpy of formation and/or phase change. The CRC Handbook of Chemistry and Physics (2010 version)[5] contains approximately 10,000

organic compounds, but enthalpies of formation and/or phase change are reported for 1,500.

Table 2. ThermInfo database statistics: number of records in the several categories/fields of the dataset (June, 2012).

Data category	Data field	Number of records in the database
Structural Data	SMILES	2,956
	Chemical Structure	2,956
	CASRN	2,952
	Compound Name	2,956
	Synonyms	12,598
	Classes	6
	Sub-Classes	12
Thermochemical Data	Families	387
	Crystalline Phase	1,461
	Standard Molar Enthalpy of Formation	Liquid Phase 1,486
		Gas Phase 1,740
	Fusion	83
	Standard Molar Enthalpy of Phase Change	Vaporization 1,093
		Sublimation 614
Total Number of Compounds		2,956

Technologies options for the system implementation

The database is implemented using MySQL [84]. Many of the basic application tools, scripts, and web interfaces, were developed using Hypertext Preprocessor (PHP), a server-side programming language designed especially for the web, with the possibility to be embedded in Hypertext Markup Language (HTML) code. PHP is very well documented, supports intensive transactions, runs fast, and works well with other programming

languages chosen for the development of this project, MySQL and JavaScript. Both data presentation and property prediction features allow the user to draw the chemical structure in a JAVA applet (JChemPaint) [85] and export it as a SMILES or 3D MDL MOL file. The use of JavaScript allows a dynamic interaction with the structure. A drawn chemical structure can be converted to a downloadable file format using a Python library (OASA) [86]. Conversions between a given structure identifier and another structure identifier or representation are made by (a not yet completely validated structure-name lookup) the Chemical Identifier Resolver provided by the NCI/CADD group [87], using a simple Uniform Resource Locator (URL) Application Programming Interface (API) scheme or Open Babel [88,89]. Pybel [90] (a Python library that provides access to the Open Babel toolkit) was also used to convert file formats, calculate molecular fingerprint to compare molecules, and to access data and information about structural attributes of the molecule in order to extract the ELBA [60] parameters and predict thermochemical properties.

The web interface is delivered using the open-source Apache web server [91]. Control over access to administrative functions is performed by using Apache Hypertext Access.

Overview of the ThermInfo web-based system

A schematic summary of the ThermInfo system architecture is presented in Figure 2, evidencing the relationships between the system main components.

Data sources. Data collected by the Molecular Energetics Group[76,81] and data submitted to a temporary database by users, via a web browser (these data are subject to a validation process by an administrator).

Administrative features. Enables administrators to interact with the database via a web browser. This interaction involves data validation (manual curation of the data submitted by users, stored in temporary database) and data removal/update.

Data presentation. Queries the database and displays information about the compounds in a structured way, via a web browser, based on four types of data-search: quick, advanced, structural, and substructure.

Properties prediction. Predicts thermochemical data and displays information in a structured way, via a web browser, based on two types of data input: quick and structural.

ThermInfo is officially available at <http://www.therminfo.com> or <http://therminfo.lasige.di.fc.ul.pt>.

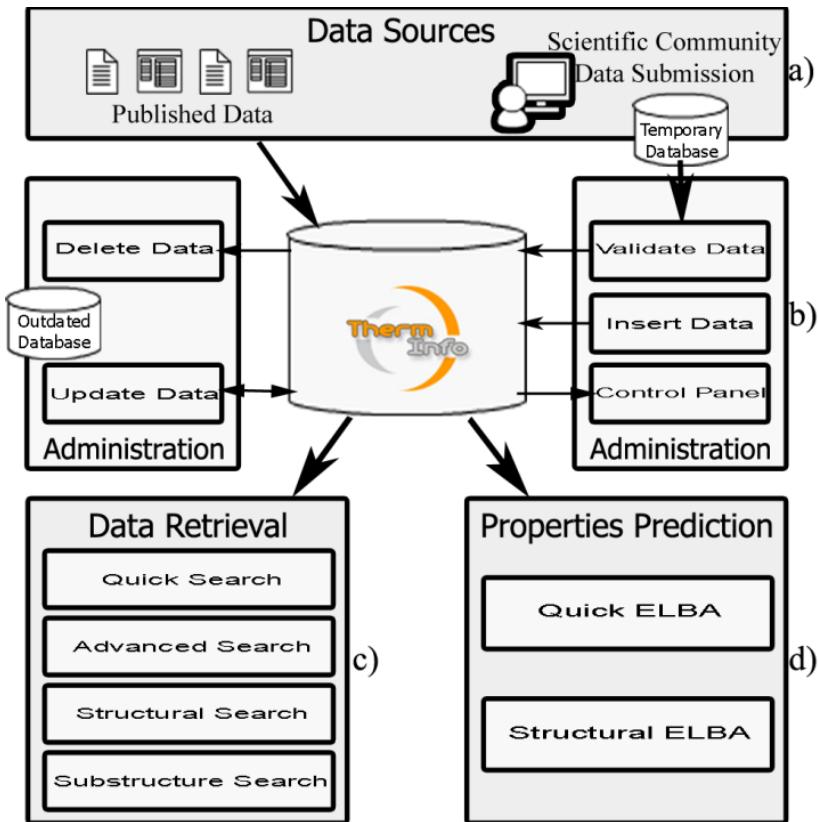


Figure 2. Simplified scheme of ThermInfo system architecture. Relationships between the database and the data sources, administrative features, data presentation, and properties prediction. **(a)** data sources; **(b)** administrative features; **(c)** data retrieval; **(d)** properties prediction.

Dissemination facilities

ThermInfo has been conceived to provide fast and easy access to information related to structural and thermochemical properties. For this purpose we have implemented a user-friendly web interface to query, insert new data to the database, predict thermochemical properties, and administrate the system. All the features are described in detail in the ‘Help’ section of the system. In addition, many forms contain guiding messages.

The major features of ThermInfo can be divided into four groups, according to its functions.

Finding and presenting data. ThermInfo has multiple search interfaces available, enabling both simple and complex queries (Figure 3). To search a compound in ThermInfo, the user selects the desirable search option and then specifies the query mode or draws a chemical structure. The data are retrieved by the server, which performs a data preprocessing, queries the database, and outputs a results list.

Quick Search [Help] **1a)**

dihydroxycyclobutenedione Name Please do not use wildcards. [\[Help\]](#)

Security code: **91U3**

[Type only **numerical characters**. Ignore letters and special characters.]

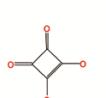
Advanced Search [Help] **1b)**

Compound Name: Physical State: All
Molecular Formula: Molecular Weight:
Wildcard: ? represents one character
SMILES: Identical Structures
Class: 02 - Ring Systems Containing Only Isolated Non-Benzenoid Rings
Sub-Class: All
Family: All
Characteristic:

CH Groups	<input type="checkbox"/> Alkane <input type="checkbox"/> Alkene <input type="checkbox"/> Alkyne <input type="checkbox"/> Arene [R-H] <input type="checkbox"/> [R ₂ C=CR] [R-C≡C-R] <input type="checkbox"/> [Ar-H]
CHO Groups	<input checked="" type="checkbox"/> Alcohol <input type="checkbox"/> Ether <input type="checkbox"/> Peroxide <input type="checkbox"/> Aldehyde [R-OH] <input type="checkbox"/> [R-O-R] <input type="checkbox"/> [R-O-R, R-O-H] <input type="checkbox"/> [R(C=O)H] <input type="checkbox"/> Ketone <input checked="" type="checkbox"/> Carboxilic Acid <input type="checkbox"/> Ester [R(C=O)R] <input type="checkbox"/> [R(C=O)OH] <input type="checkbox"/> [R(C=O)O-R]
CHN Groups	<input type="checkbox"/> Amine <input type="checkbox"/> Hydrazine <input type="checkbox"/> Imine <input type="checkbox"/> Nitrile/Isonitrile [R ₂ N] <input type="checkbox"/> [R-NH-NH-R] <input type="checkbox"/> [R=N=R, R-N=N-R] <input type="checkbox"/> [R-C≡N, R-N ⁺ ≡ C] [R-NH ⁻ -C, R-N=O, R-O-N=O, R-N ⁺ (=O)O, R-O-N ⁺ (=O)O]
CHON Groups	<input type="checkbox"/> NO _x <input type="checkbox"/> Amide [R ₂ N ⁻ -C, R-N=O, R-O-N=O, R-N ⁺ (=O)O, R-O-N ⁺ (=O)O] <input type="checkbox"/> [R(C=O)NR ₂]
CHS Groups	<input type="checkbox"/> Thiol <input type="checkbox"/> Thioether <input type="checkbox"/> Polysulphide [R-S-S-] <input type="checkbox"/> Thiocarbonyl [R=C=S] [R-SH] <input type="checkbox"/> [R-S-R] <input type="checkbox"/> R, R-SH] <input type="checkbox"/> R, R(C=S)H]
CHOS Groups	<input type="checkbox"/> SOX [R ₂ S=O, R ₂ S(=O) ₂ , R-O-S(=O)-O-R, R-O-S(=O) ₂ -O-R]
CHX Groups (X=F, Cl, Br, I)	<input type="checkbox"/> Halogen [R-X]
Physical	<input type="checkbox"/> Radical <input type="checkbox"/> Charges <input type="checkbox"/> Ionic <input type="checkbox"/> Solvation <input type="checkbox"/> Polymer

Structural Search [Help] **1c)**

Arquivo Editar Ver Atom Bond Tools R-groups Templates Ajuda



You are searching for:
- Characteristic: Alcohol
- Class: 02 - Ring Systems Containing Only Isolated Non-Benzenoid Rings
Number of compounds found: 79

2)

1. Molecular ID: C001290
Compound Name: (Hydroxymethyl)oxirane
Molecular Formula: C₃H₆O₂
CAS Registry Number: 556-52-5
SMILES: OCC1OC1
More Info:

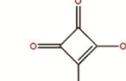
2. Molecular ID: C001332
Compound Name: 3,4-Dihydroxy-3-cyclobutene-1,2-dione
Molecular Formula: C₄H₆O₄
CAS Registry Number: 2892-51-5
SMILES: O=C1C(CO)=C1O=O
More Info:

3. Molecular ID: C001358
Compound Name: Cyclopentanol
Molecular Formula: C₅H₁₀O
CAS Registry Number: 96-41-3
SMILES: OC1CCCC1
More Info:

4. Molecular ID: C001359
Compound Name: cis-2-Methylcyclopentanol
Molecular Formula: C₆H₁₂O
CAS Registry Number: 25144-05-2
SMILES: O[C@H]1[C@H](O)CCCC1
More Info:

3)

3,4-Dihydroxy-3-cyclobutene-1,2-dione



Structural Data

Molecular ID: C001332 Compound Name: 3,4-Dihydroxy-3-cyclobutene-1,2-dione
Other Names: Dihydroxycyclobutenedione; Quadratic acid; Squaric acid; 1,2-Dihydroxycyclobutene-3,4-dione; 3,4-Diketo-1,2-dihydroxycyclobutene; 1,2-Dihydroxycyclobutenedione; 1,2-Dihydroxy-3,4-cyclobutenedione; 3,4-Dihydroxycyclobutene-1,2-dione; 4-Hydroxycyclobutene-1,2-dione

CASRN: 2892-51-5 Molecular Formula: C₄H₆O₄
Molecular Weight: 114.06 Physical State: Crystal

SMILES: O=C1C(CO)=C1O=O
Unique SMILES: [C]1([C]([C](=[C]1[OH])=O)=O)=O

Thermochemical Data

Standard Molar Enthalpy of Formation at 298.15 K [kJ/mol]:

- Crystalline Phase -598.2 Uncertainty 0.4
- Liquid Phase n.a. Uncertainty n.a.
- Gas Phase -514.5 Uncertainty 16.6

Standard Molar Enthalpy of Phase Change at 298.15 K [kJ/mol]:

- Solid - Liquid n.a. Uncertainty n.a.
- Liquid - Gas n.a. Uncertainty n.a.
- Solid - Gas 83.7 Uncertainty 16.7

Observations: Possible keto-enol tautomeric compound.

References

1. J. B. Pedley
Thermochemical Data and Structures of Organic Compounds, Vol. 1, 1994, 1-571
2. J. B. Pedley, R. D. Naylor, S. P. Kirby
Thermochemical Data of Organic Compounds, 2nd ed., 1986, 1-792

Figure 3. Composite screenshot example of data retrieval features. **(1)** The three types of data input: **(a)** Quick Search, term-based search; **(b)** ‘Advanced Search’, multiple search fields based on specific structural characteristics; **(c)** ‘Structural or Substructure Search’, based on the molecular structure drawn in a Java applet. **(2)** Search result list: the query description and the list of compounds found in the database. **(3)** Detailed information available for the selected compound.

ThermInfo has four search features available (Table 3): ‘Quick Search’, which provides a single text box that allows users to search for a chemical compound based on the compound name, the molecular formula, the molecular ID, CASRN, or SMILES; ‘Advanced Search’, which provides multiple search fields that allow users to limit the search results, based on specific characteristics, namely the compound name, the molecular formula, the physical state, the molecular weight, class, sub-class, family, functional groups, and other; ‘Structural Search’ and ‘Substructure Search’, which provide a Java applet that allows users to draw chemical structures or substructures that will be internally converted into a SMILES string (alternatively a SMILES string can be directly typed).

A successful search displays the query description and a summary of the information for the 100 most relevant compounds. By clicking ‘View’ for a specific compound, the user will be able to obtain all the information available for that compound. The results are ordered according to the selected search type (Table 3).

User Contributions. The feature ‘Insert Data’ is restricted to registered users, to prevent the inclusion of inappropriate data, and allows them to submit data for new compounds. Before disclosing these data, an automatic pre-processing stage and a curation process performed by an administrator are implemented, to verify and validate possibly incorrect data. The main purpose of this feature is to support the expansion of the database by the scientific community. The features ‘Suggestion’, ‘Erroneous Data’, ‘Question’, etc. are within ‘Contact Us’ and allow an easy interaction between users and the ThermInfo team. This is very important not only to understand the user’s questions and needs but also because it allows the participation of the scientific community in the process of assessing data quality.

System Management. ThermInfo administrators manage the system via a simple administrative web interface. In ‘Administration’ new records can be added, updated or deleted. To update or delete records the administrator needs to search for the compound using its molecular ID. If it exists, it will display the compound information that can be deleted or changed. The outdated data are moved to outdated tables. The ‘Validate Data’ feature allows the administrator to check for new compounds added by users and approve or reject their insertion into the ThermInfo database. The approval or rejection of the new data is automatically reported to the depositor. The ‘Control Panel’ feature allows the administrator to monitor the current data suppliers, as well as the usage and the growth of the database over time.

Table 3. Description of the search types available on each search feature.

Search Type	Description	Search Feature		
		Quick	Advanced	Structural Substructure
Compound Name	The search term is the IUPAC name of a compound. However, searching by using alternative names is also possible for many compounds. Wildcards are not required since the search method retrieves all the compounds whose names contain the search string. The results list will be ordered according to the index of the first occurrence of the query term in the matching string and the difference between the molecular weight of the query structure and the molecular weight of each compound containing the query term, found in the database. When the textual search has no results and knowing the diversity of homophonic names, as well as the propensity for humans to misspell names or misplaced letters and numbers, the system retrieves those compounds which the name phonetically matches the search term. The phonetic matching is based on the MySQL's function - soundex().	✓	✓	
Molecular Formula	Searches for compounds with the requested chemical formula. In the database, the atoms are in CHXNOS (X = halogen) order, but the search term may be written in any order. The symbol ? can replace the number of atoms of an element. For example, using C?H11 molecular formulas with ? atoms of carbon [? = 2-9] and 11 atoms of hydrogen will be retrieved.	✓	✓	
Molecular ID	Performs an exact string matching for the unique ID associated with the chemical compounds in the database. Molecular ID is a unique and stable identifier assigned to each compound of the database by ThermInfo. It has the format CONNNNN (N = digit).	✓		
CASRN	Performs an exact string matching for the CAS Registry Number (CASRN) associated with the chemical compounds in the database. It has the format NNNNNNNN-NN-N (1-7 digits, hyphen, 2 digits, hyphen, 1 digit). The last digit is a check digit used to verify the validity and uniqueness of the entire number.	✓		
SMILES	This search feature allows to retrieve chemical structures similar to a given query structure, according to a selected similarity threshold. A threshold of 100% (Identical Structures option) retrieves identical chemical structures with different notations (ignoring stereo or isotopic information). Various predefined thresholds between 70-95% are allowed. The calculation of the similarity between the chemical structures is based on Open Babel[88,89] Fingerprints and it is measured using the Tanimoto coefficient [92-96]. The results list will be ordered according to the similarity coefficient (all compounds shown have a similarity score greater than the selected threshold) [71-73] and the difference between the molecular weight of the query structure and the molecular weight of each similar compound found in the database.	✓	✓	✓
Physical state	Searches for compounds in a certain physical state selected from a drop down menu with the options: gas, liquid or crystal.	✓		
Molecular weight	Searches for compounds matching a specific molecular weight or within a specified interval.	✓		
Class		✓		
Sub-class	Searches for compounds of a specific class, subclass and/or family, selected from a drop down menu.	✓		
Family		✓		
Characteristics	Searches for compounds according to the functional groups and other characteristics of the compound. Multiple selections can be made using the check-boxes.	✓		

Properties Prediction. As mentioned above, it is possible to predict new thermochemical data with ThermInfo. For that purpose, we have developed an ergonomic and functional interface (Figure 4). In the present version of ThermInfo only one estimation method (ELBA) has been implemented. Two different kinds of input can be used. The simplest is ‘Quick ELBA’, which allows predictions based on a query in a text box (the compound name or its SMILES). If the compound name is used, it will be converted to SMILES using the Chemical Identifier Resolver [87].

Extended Laidler Bond Additivity (ELBA) Method [Help]

Ethylicyclohexane Name Predict Properties
0 double bonds in trans configuration (E) [cyclic compounds: 8- or 12-atoms ring]

1 a)

1 b)

You are predicting properties for:

SMILES: CCC1CCCCC1

Compound Name: Ethylicyclohexane Molecular Weight: 112.21264
* Compound Name(s) and Structure provided by the Chemical Identifier Resolver.

Just Predict Properties

:: Number of compounds found on the database: 8 ::

- Experimental and Estimated Properties for similar compounds found on the database:

Molecular ID: C001598
Compound Name: Ethylicyclohexane
Molecular Formula: C₈H₁₆
Molecular Weight: 112.21
SMILES: CCC1CCCCC1
Properties Prediction: View

Molecular ID: C001608
Compound Name: 1,1-Dimethylcyclohexane
Molecular Formula: C₈H₁₆
Molecular Weight: 112.21
SMILES: CC1(C)CCCCC1
Properties Prediction: View

2)

3)

You are predicting properties for:

Compound Name: Ethylicyclohexane
SMILES: CCC1CCCCC1
More Info: View

Experimental Standard Molar Enthalpy of Formation at 298.15 K for CCC1CCCCC1 [kJ/mol]:		
Gas-phase	Liquid-phase	
-171.5	-212.1	

Estimated Standard Molar Enthalpy of Formation at 298.15 K using ELBA for CCC1CCCCC1 [kJ/mol]:		
Gas-phase	Liquid-phase	
-172.3	-212.8	

Set of ELBA parameters used to predict properties of CCC1CCCCC1

ELBA parameter	Used Frequency	Short Description <small>(More Info)</small>
C2C2	4	single bond between two sp ³ carbon atoms single-bonded to one sp ³ carbon atom [C-C bond for alkanes]
C2C3	3	single bond between a sp ³ carbon atom single-bonded to one sp ³ carbon atom, and a sp ³ carbon atom single-bonded to two sp ³ carbon atoms [C-C bond for alkanes]
C1H	3	single bond between a sp ³ carbon atom single-bonded to one sp ³ carbon atom, and a hydrogen atom [C-H bond for alkanes]
C3H	1	single bond between a sp ³ carbon atom single-bonded to three sp ³ carbon atoms, and a hydrogen atom [C-H bond for alkanes]
ZS6C2	5	takes into account the strain of a secondary sp ³ carbon atom in a six-carbon ring [strain parameter for cycloalkanes]
ZS6C3	1	takes into account the strain of a tertiary sp ³ carbon atom in a six-carbon ring [strain parameter for cycloalkanes]
C1C2	1	single bond between a sp ³ carbon atom single-bonded to three hydrogen atoms, and a sp ³ carbon atom single-bonded to one sp ³ carbon atom [C-C bond for alkanes]
C2H	12	single bond between a sp ³ carbon atom single-bonded to two sp ³ carbon atoms, and a hydrogen atom [C-H bond for alkanes]

Figure 4. Composite screenshot example of properties prediction feature. **(1)** The two types of data input: **(a)** ‘Quick ELBA’, based on the compound name in a text-box; **(b)** ‘Structural ELBA’, based on the molecular structure drawn in a Java applet. **(2)** Search result list: the query description and the list of similar compounds found in the database. **(3)** The experimental and predicted values for the properties of the selected compound and the set of ELBA parameters used for the predictions.

The most sophisticated input is under the ‘Structural ELBA’ feature, which provides a Java applet that allows users to draw the chemical structure of the target compound. This chemical structure will be internally converted by the Java applet into a 3D MDL MOL file to include geometrical information. Once a structure has been drawn or converted from a chemical name into a SMILES string, a structure search is performed in our database. The output includes: the query description (SMILES, compound name, molecular structure, and molecular weight); the option to ‘Just Predict Properties’ for the target molecule; a list with the target molecule and similar compounds (isomers), if found on the database. By selecting ‘Just Predict Properties’ for a given compound in the results list, the predicted values and the set of ELBA parameters used for the prediction are presented in tabular form. When ‘View’ is selected the predicted data are compared to the experimental values. When the ‘View’ button is not visible, it means that no experimental data are available.

System performance with respect to response time

System response time is obviously a very important factor for users but it is also highly dependent on the test environment (for example, internet connection speed, internet traffic level, device characteristics and web browser). This introduces a large amount of error variation into the measured times. ThermInfo performance with respect to response time to retrieve and predict data was assessed in a simple way within a closed Local Area Network environment, to suppress external causes of variance. For that purpose, each compound of a sample of 500 randomly selected compounds was submitted to the system and the time it takes to retrieve and predict properties (excluding the non-hydrocarbon compounds) was measured. This procedure was repeated three times for each search parameter. The average response time and result set size with corresponding standard deviation of the three repetitions of each experiment were recorded (Table 4).

Table 4. Recorded average number of hits and response time to retrieve and predict properties with different inputs.

Input	Average number of hits ^a	std dev ^b	Average response time (seconds) ^c	std dev ^b
<i>Search Performance</i>				
Compound Name	4.18	27.29	0.687	0.173
Molecular Formula	9.71	13.62	0.055	0.046
CASRN	1.00	-	0.076	0.038
SMILES with similarity threshold = 100%	3.84	8.50	1.284	0.216
SMILES with similarity threshold = 90%	4.89	9.08	1.269	0.212
SMILES with similarity threshold = 80%	11.20	23.33	1.295	0.216
<i>Prediction Performance</i>				
SMILES	-	-	3.927	4.396

^a Number of compounds found by the searching engine during the search; ^b std dev = standard deviation;

^c Average response time of three repetitions for each experiment.

CONCLUSIONS AND FUTURE WORK

ThermInfo is a publicly available, web-accessible, information system that brings together critically evaluated values of thermochemical properties of pure substances (for now restricted to organic compounds) and structural data. It also implements a reliable method to predict enthalpies of formation of hydrocarbons. ThermInfo database has already a significant size but it will grow considerably in the near future. This growth will include an extension of the organic compounds database as well as the addition of critically evaluated data for long-lived and transient organic, inorganic, and organometallic molecules in the gas- and condensed-phases. It will also extend the ELBA scheme to other families of compounds, implement a variety of other empirical methods, selected on the basis of their reliability to predict data, and search for new estimation procedures, based on structure-energy relationships and machine learning methods. Finally, ThermInfo scope will also increase in the near future. The database will include values for other molecular properties such as melting point, boiling point, density, refractive index, solubility, standard molar entropy and standard molar Gibbs energy of formation. All this will be possible by implementing a flexible and modular architecture of the database.

ThermInfo strength lies in the data quality, availability (free access), searchability (using different criteria and format input), and, in particular, prediction ability, based on a user-friendly interface that accepts inputs in several formats. It also allows the collaboration of the scientific community, which is encouraged to participate in the project by submitting suggestions and requests for new features and by submitting new data. ThermInfo is available free of charge at <http://www.therminfo.com> or <http://www.therminfo.lasige.di.fc.ul.pt>.

Availability and Requirements

Project name: ThermInfo

Project home page: <http://www.therminfo.com>

Operating system: Platform independent

System requirements: Web browser HTML 4.0 compatible with Java (JRE) plug-in and cookies enabled

Programming language: PHP, JavaScript, Python

Database system: MySQL

Acknowledgment

ALT and RCS gratefully acknowledge Fundação para a Ciência e a Tecnologia for a doctoral grant (SFRH/BD/64487/2009) and a post-doctoral grant (SFRH/BPD/26610/2006), respectively.

Supporting Information

A table with *ThermInfo database fields description* was included in Supporting Information.

Author Contributions

ALT carried out the implementation of the ThermInfo Information System and drafted the manuscript, supervised by AOF and JPL. All authors were involved in the design and testing of the ThermInfo. RCS was responsible for data collection, supervised by JAMS. JPL, RCS, and JAMS developed the ELBA method. All authors read, reviewed, and approved the final version of the manuscript.

REFERENCES

- (1) Bachrach, S. M. Chemistry publication – making the revolution. *J. Cheminf.* **2009**, *1*.
- (2) Chen, W. L. Chemoinformatics: Past, Present, and Future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255.
- (3) *Chemical Abstracts Service (CAS)*; American Chemical Society: Columbus, OH: 2012. <http://www.cas.org/> (accessed May, 2012).
- (4) Lide, D. R., Sources of Physical and Chemical Data. In *CRC Handbook of Chemistry and Physics*, 90th (CD-ROM Version) ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2010.
- (5) Lide, D. R., *CRC Handbook of Chemistry and Physics*, 90th (CD-ROM Version) ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2010.
- (6) Marsh, K. N., *Recommended Reference Materials for the Realization of Physicochemical Properties*, Blackwell Scientific Publications: Oxford, 1987.
- (7) O'Neil, M. J., *The Merck Index: An Encyclopedia of Chemical, Drugs and Biologicals*, 14th ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 2006.
- (8) Speight, J. G.; Lange, N. A., *Lange's Handbook of Chemistry*, 16th ed.; McGraw-Hill Professional: New York, 2005.
- (9) DETHERM - Thermophysical Properties & Phase Equilibrium Data of Pure Substances & mixtures; DECHEMA e. V. <http://i-systems.dechema.de/detherm/mixture.php> (accessed March, 2012).
- (10) CHEMnetBASE; Taylor and Francis group, LLC. <http://www.chemnetbase.com/> (accessed March, 2012).
- (11) Ionic Liquids Database (ILThermo) – NIST Standard Reference Database 147; National Institute of Standards and Technology. <http://ilthermo.boulder.nist.gov> (accessed May, 2012).
- (12) DIPPR® Project 801; American Institute of Chemical Engineers. www.aiche.org/dippr (accessed March, 2012).
- (13) Reaxys; Elsevier Properties SA: Switzerland. <http://www.reaxys.com> (accessed March, 2012).
- (14) Springer Materials - The Landolt-Börnstein Database; Taylor and Francis group, LLC. <http://www.springermaterials.com> (accessed March, 2012).
- (15) Chemical Entities of Biological Interest (ChEBI); European Bioinformatics Institute, European Molecular Biology Laboratory. <http://www.ebi.ac.uk/chebi/> (accessed March, 2012).
- (16) CODATA, The Committee on Data for Science and Technology; International Council for Science. <http://www.codata.org/resources/databases/index.html> (accessed March, 2012).
- (17) Knovel; Knovel Corporation. <http://www.knovel.com/web/portal/main> (accessed March, 2012).

- (18) Protein Data Bank (PDB); Research Collaboratory for Structural Bioinformatics <http://www.rcsb.org> (accessed March, 2012).
- (19) The PubChem Project; National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health. <http://pubchem.ncbi.nlm.nih.gov> (accessed March, 2012).
- (20) Dortmund Data Bank (DDB); DDBST – Dortmund Data Bank Software and Separation Technology GmbH. <http://www.ddbst.de> (accessed March, 2012).
- (21) Infotherm - Database of Thermophysical Properties; FIZ CHEMIE. <http://www.fiz-chemie.de/infotherm/> (accessed March, 2012).
- (22) NIST Standard Reference Data; National Institute of Standards and Technology. <http://www.nist.gov/srd/> (accessed March, 2012).
- (23) SciFinder; American Chemical Society, Chemical Abstracts Service: Columbus, OH. <https://scifinder.cas.org> (accessed February, 2012).
- (24) *Computational Thermochemistry: Prediction and Estimation of Molecular Thermodynamics*; Irikura, K. K.; Frurip, D. J., Eds.; ACS Symposium Series 677; American Chemical Society: Washington, DC, 1998.
- (25) Roth, D. L. Web-Accessible Chemical Compound Information. *J. Electron. Resour. Med. Libr.* **2008**, 5, 228–242.
- (26) Southan, C.; Várkonyi, P.; Muresan, S. Complementarity Between Public and Commercial Databases: New Opportunities in Medicinal Chemistry Informatics. *Curr. Top. Med. Chem.* **2007**, 7, 1502–1508.
- (27) Southan, C.; Várkonyi, P.; Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminf.* **2009**, 1.
- (28) Williams, A. J. A perspective of publicly accessible/open-access chemistry databases. *Drug Discovery Today* **2008**, 13, 495–501.
- (29) Williams, A. J. Public chemical compound databases. *Curr. Opin. Drug Discovery Dev.* **2008**, 11, 393–404.
- (30) Dohrn, R.; Pfohl, O. Thermophysical properties - Industrial directions. *Fluid Phase Equilibria* **2002**, 194, 15-29.
- (31) Cox, J. D.; Pilcher, G., *Thermochemistry of Organic and Organometallic Compounds*, Academic Press: London and New York, 1970.
- (32) Donald, D. W.; William, H. E.; Vivian, B. P.; Richard, H. S.; Iva, H.; Sylvia, M. B.; Kenneth, L. C.; Ralph, L. N. Erratum: The NBS tables of chemical thermodynamic properties. Selected values for inorganic and C₁ and C₂ organic substances in SI units [J. Phys. Chem. Ref. Data **11**], Suppl. 2 (1982)]. *Journal of Physical and Chemical Reference Data* **1989**, 18, 1807-1812.
- (33) Stull, D. R. W. J., E. F.; Sinke, G. C., *The Chemical Thermodynamics of Organic Compounds*, John Wiley & Sons, Inc.: New York, 1969.

- (34) Zwolinski, B. J.; Wilhoit, R. C., *Handbook of Vapor Pressures and Heats of Vaporization of Hydrocarbons and Related Compounds*, Thermodynamics Research Center: College Station, TX, 1971.
- (35) Domalski, E. S.; Evans, W. H.; Hearing, E. D. Erratum: Heat Capacities and Entropies of Organic Compounds in the Condensed Phase [J. Phys. Chem. Ref. Data **13**, Suppl. 1 (1984)]. *Journal of Physical and Chemical Reference Data* **1990**, *19*, 1075.
- (36) Acree, J. W.; Chickos, J. S. Phase Transition Enthalpy Measurements of Organic and Organometallic Compounds. Sublimation, Vaporization and Fusion Enthalpies From 1880 to 2010. *Journal of Physical and Chemical Reference Data* **2010**, *39*, 043101-043942.
- (37) Domalski, E. S.; Hearing, E. D. Heat Capacities and Entropies of Organic Compounds in the Condensed Phase Volume II. *Journal of Physical and Chemical Reference Data* **1990**, *19*, 881-1047.
- (38) Domalski, E. S.; Hearing, E. D. Erratum: Heat Capacities and Entropies of Organic Compounds in the Condensed Phase, Volume III [J. Phys. Chem. Ref. Data **25**, 1--525 (1996)]. *Journal of Physical and Chemical Reference Data* **1997**, *26*, 1501-1505.
- (39) *Enthalpies of Vaporization of Organic Compounds: A Critical Review and Data Compilation*; Majer, V.; Svoboda, V., Eds.; Blackwell Scientific Publications: Oxford, 1985.
- (40) Pedley, J. B., *Thermochemical Data and Structures of Organic Compounds*, TRC Data Series; Thermodynamics Research Center (TRC): College Station, TX, 1994; Vol. 1.
- (41) Pedley, J. B.; Naylor, R. D.; Kirby, S. P., *Thermochemical Data of Organic Compounds*, 2nd ed.; Chapman and Hall: New York, 1986.
- (42) Roux, M. V.; Temprado, M.; Chickos, J. S.; Nagano, Y. Critically Evaluated Thermochemical Properties of Polycyclic Aromatic Hydrocarbons. *Journal of Physical and Chemical Reference Data* **2008**, *37*, 1855-1996.
- (43) Stephenson, R. M.; Malanowski, S., *Handbook of the Thermodynamics of Organic Compounds*, Elsevier: New York, 1987.
- (44) Yaws, C. L., *Thermophysical Properties of Chemicals and Hydrocarbons*, William Andrew: 2008.
- (45) Zabransky, M.; Kolska, Z.; Ruzicka, J. V.; Domalski, E. S. Heat Capacity of Liquids: Critical Review and Recommended Values. Supplement II. *Journal of Physical and Chemical Reference Data* **2010**, *39*, 013103-013404.
- (46) Zabransky, M.; Ruzicka, J. V.; Domalski, E. S. Heat Capacity of Liquids: Critical Review and Recommended Values. Supplement I. *Journal of Physical and Chemical Reference Data* **2001**, *30*, 1199-1689.
- (47) Zabransky, M.; Ruzicka, J. V.; Majer, V.; Domalski, E. S. Erratum: ``Heat Capacity of Liquids. Critical Review and Recommended Values. Volumes I and II.'' [J. Phys. Chem. Ref. Data, Monograph No. 6 (1996)]. *Journal of Physical and Chemical Reference Data* **2001**, *30*, 441-443.

- (48) Ideal Gas Thermochemical Database with Updates from Active Thermochemical Tables; Technion – Israel Inst of Technology; Goos, E.; Burcat, A.; Ruscic, B., Eds. <ftp://ftp.technion.ac.il/pub/supported/aetdd/thermodynamics> mirrored at <http://garfield.chem.elte.hu/Burcat/burcat.html> (accessed March, 2012).
- (49) NIST Chemistry WebBook, NIST Standard Reference Database Number 69; National Institute of Standards and Technology, Gaithersburg MD, 20899; Linstrom, P. J.; Mallard, W. G., Eds. <http://webbook.nist.gov/chemistry/> (accessed March, 2012).
- (50) Korean Thermophysical Properties Data Bank (KDB); Chemical Engineering Research Information Center (CHERIC). <http://www.cheric.org/research/kdb/> (accessed March, 2012).
- (51) *CHETAH – The Computer Program for Chemical Thermodynamics and Energy Release Evaluation*, Version 9.0; ASTM International: 2009. <http://www.southalabama.edu/engineering/chemical/chetah/> (accessed March, 2012).
- (52) Martinho Simões, J. A.; Minas Da Piedade, M. E., *Molecular Energetics. Condensed-Phase Thermochemical Techniques*, Oxford University Press: New York, 2008.
- (53) Cramer, C. J., *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; John Wiley & Sons Ltd.: Chichester (England), 2004.
- (54) Jensen, F., *Introduction to Computational Chemistry*, 2nd ed.; John Wiley & Sons Ltd: Chichester (England), 2007.
- (55) Curtiss, L. A.; Redfern, P. C.; Frurip, D. J., Theoretical Methods for Computing Enthalpies of Formation of Gaseous Compounds. In *Reviews in Computational Chemistry*, Lipkowitz, K. B.; Boyd, D. B., Eds. Wiley–VCH: New York, 2000; Vol. 15, pp 147–211.
- (56) Wheeler, S. E.; Houk, K. N.; Schleyer, P. v. R.; Allen, W. D. A Hierarchy of Homodesmotic Reactions for Thermochemistry. *Journal of the American Chemical Society* **2009**, *131*, 2547-2560.
- (57) Cohen, N.; Benson, S. W. Estimation of heats of formation of organic compounds by additivity methods. *Chemical Reviews* **1993**, *93*, 2419-2438.
- (58) Domalski, E. S.; Hearing, E. D. Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15 K. *Journal of Physical and Chemical Reference Data* **1993**, *22*, 805–1159.
- (59) Leal, J. P. Additive Methods for Prediction of Thermochemical Properties. The Laidler Method Revisited. 1. Hydrocarbons. *Journal of Physical and Chemical Reference Data* **2006**, *35*, 55–76.
- (60) Santos, R. C.; Leal, J. P.; Martinho Simões, J. A. Additivity methods for prediction of thermochemical properties. The Laidler method revisited. 2. Hydrocarbons including substituted cyclic compounds. *J. Chem. Thermodyn.* **2009**, *41*, 1356–1373.
- (61) Salmon, A. Prediction of Enthalpy of Formation in the Solid State (at 298.15K) using Second-Order Group Contributions. Part 1. Carbon-Hydrogen and Carbon-Hydrogen-Oxygen Compounds. *Journal of Physical and Chemical Reference Data* **2006**, *35*, 1443–1457.

- (62) Salmon, A. Prediction of Enthalpy of Formation in the Solid State (at 298.15 K) Using Second-Order Group Contributions Part 2: Carbon-Hydrogen, Carbon-Hydrogen-Oxygen, and Carbon-Hydrogen-Nitrogen-Oxygen Compounds. *Journal of Physical and Chemical Reference Data* **2007**, *36*, 19–58.
- (63) Laidler, K. J. A System of Molecular Thermochemistry for Organic Gases and Liquids. *Can. J. Chem.* **1956**, *34*, 626–648.
- (64) Domalski, E. S.; Hearing, E. D.; Hearing, V. J. *NIST Estimation of the Chemical Thermodynamic Properties for Organic Compounds at 298.15 K; NIST Standard Reference Database 18*, National Institute of Standards and Technology: Gaithersburg, 1994. (accessed May, 2012).
- (65) *DIADEM Software Interface*, American Institute of Chemical Engineers: 2011. <http://www.aiche.org/dippr/products/801.aspx> (accessed March, 2012).
- (66) *DETERM Software Suite*, DECHEMA e.V.: 2012. <http://www.dechema.de/en/Publications/Engineering+Software.html> (accessed March, 2012).
- (67) NIST ThermoData Engine – NIST Standard Reference Database 103b, version 6.0; Thermophysical Properties Division – National Institute of Standards and Technology. <http://www.trc.nist.gov/tde.html> (accessed March, 2012).
- (68) Stein, S. S. *NIST Structures and Properties; NIST Standard Reference Database 25*, Version 2.0; National Institute of Standards and Technology: Gaithersburg, 1994. <http://webbook.nist.gov/chemistry/grp-add/> (accessed May, 2012).
- (69) Dortmund Data Bank Software and Separation Technology GmbH, *ARTIST Property Estimation*; Oldenburg, 2010. <http://www.ddbst.com/en/products/DDBSP-Artist.php> (accessed May, 2012).
- (70) Dortmund Data Bank Software and Separation Technology GmbH, *DDBSP: UNIFAC Calculation Software*; Oldenburg, 2010. <http://www.ddbst.com/en/products/DDBSP-Artist.php> (accessed May, 2012).
- (71) James, C. A.; Weininger, D.; Delany, J., *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008. <http://www.daylight.com/dayhtml/doc/theory/index.html> (accessed May, 2012).
- (72) Weininger, D. SMILES, a Chemical Language and Information System .1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (73) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
- (74) Lemmon, E. W.; McLinden, M. O.; Wagner, W. Thermodynamic Properties of Propane. III. A Reference Equation of State for Temperatures from the Melting Line to 650 K and Pressures up to 1000 MPa. *Journal of Chemical & Engineering Data* **2009**, *54*, 3141-3180.
- (75) Ramakrishnan, R.; Gehrke, J., *Database Management Systems*, 3rd ed.; McGraw Hill Higher Education: Boston, 2003.

- (76) Bowman, J. S.; Emerson, S. L.; Darnovsky, M., *The Practical SQL Handbook: Using Structured Query Language*, 3rd ed.; Pearson Education – Addison-Wesley: Harlow, 1996.
- (77) Haag, S.; Cummings, M.; Phillips, A., *Management Information Systems for the Information Age*, 7th ed.; McGraw Hill Higher Education: Boston, 2007.
- (78) Nielsen, J., *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing: Thousand Oaks, CA, 1999.
- (79) Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, 3rd ed.; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, 1997.
- (80) Teixeira, A. L.; Santos, R. C.; Couto, F. M. In *ThermInfo: Collecting and Presenting Thermochemical Properties*, INForum – Simpósio de Informática 2009, Lisboa, Portugal, Sep 10–11, 2009; Rodrigues, L.; Lopes, R., Eds. Faculdade de Ciências da Universidade de Lisboa: Lisboa, Portugal, 2009; pp 273–284.
- (81) *Molecular Energetics Group of Centro de Química e Bioquímica from Faculdade de Ciências da Universidade de Lisboa*. <http://molenergetics.fc.ul.pt/> (accessed March, 2012).
- (82) *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, NCSL International: Boulder, CO, 1997.
- (83) Dong, Q.; Chirico, R. D.; Yan, X.; Hong, X.; Frenkel, M. Uncertainty Reporting for Experimental Thermodynamic Properties. *Journal of Chemical & Engineering Data* **2005**, *50*, 546–550.
- (84) *MySQL 5.1 Reference Manual*; Oracle: Redwood City, CA.
<http://dev.mysql.com/doc/refman/5.1/en/> (accessed May, 2012).
- (85) Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint – Using the Collaborative Forces of the Internet to Develop a Free Editor for 2D Chemical Structures. *Molecules* **2000**, *5*, 93–98.
- (86) Kosata, B. *OASA Library*, Version 0.13.1; 2009. http://bkchem.zirael.org/oasa_en.html (accessed May, 2012).
- (87) Sitzmann, M., *NCI/CADD Chemical Identifier Resolver, beta 3.1*; NCI/CADD Group: 2011. <http://cactus.nci.nih.gov/chemical/structure> (accessed May, 2012).
- (88) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.
- (89) *Open Babel: The Open Source Chemistry Toolbox*, Version 2.3.0; 2011.
<http://www.openbabel.org> (accessed May, 2012).
- (90) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*.
- (91) *Apache HTTP Server Project*, Version 2.2.17; The Apache Software Foundation: Los Angeles, CA, 2010. <http://httpd.apache.org> (accessed May, 2012).
- (92) *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*; Bajorath, J., Ed.; Methods in Molecular Biology 275; Humana Press: Totowa, NJ, 2004.

- (93) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, 2, 3204–3218.
- (94) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity – a Review. *QSAR Comb. Sci.* **2003**, 22, 1006–1026.
- (95) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, 11, 1046–1053.
- (96) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 983–996.

SUPPORTING INFORMATION

Table S1. ThermInfo database fields description.

1. Structural data, consist of a set of descriptors that specify the molecular structure of the compounds, showing how the atoms are connected, the molecular size, and other properties.

Molecular Identity Descriptor (ID)	is a unique and stable identifier for the compound, with the format CONNNNN (N = digit).
Compound Name	is the name provided for a compound, based on the current recommendations of the International Union of Pure and Applied Chemistry (IUPAC).
CAS Registry Number (CASRN)	is a unique numerical identifier created and assigned to a chemical substance by the Chemical Abstracts Service (CAS). It does not have any chemical significance and it is assigned in sequential order to assure uniqueness. It has the format NNNNNNNN-NN-N (1-7 digits, hyphen, 2 digits, hyphen, 1 digit). The right-most digit is a check digit used to verify the validity and uniqueness of the entire number and it is calculated by taking the last digit times 1, the next digit times 2, the next digit times 3 and so on, adding all these up, and computing the sum modulo 10. For example, the CAS number of methanol is 67-56-1: the checksum 1 is calculated as $(6 \cdot 1 + 5 \cdot 2 + 7 \cdot 3 + 6 \cdot 4) = 61$; $61 \bmod 10 = 1$. ^{S1}
Molecular Formula	identifies each constituent element of a compound by its chemical symbol and indicates the number of atoms of each element in subscript after the chemical symbol. The atoms are in CHXNOS (X = halogen) order.
Chemical Structure	is a bidimensional structural diagram of the compound in JPG format.
Molecular Weight	is the mass of one molecule of the compound, relative to the unified atomic mass unit.
Physical State	are the distinct forms of different phases of matter (gas, liquid or crystalline).
SMILES	is a specification for describing the structure of chemical molecules using short ASCII strings. This description is case-sensitive. For example, the SMILES for cyclohexane is C1CCCCC1 while for benzene is c1ccccc1. ^{S2-S4}
USMILES	is a special and unique SMILES amongst all valid possibilities for a given compound. ^{S2-S4}
Class, Subclass, Family	are hierarchical classifications according to the compound structure.
Characteristics	are tags according to the functional groups present in the molecule and other characteristics of the compound.

2. Thermochemical data are related to the energy released or absorbed in chemical reactions or in physical transformations.^{S5,S6}

Standard Molar Enthalpy of Formation	of a pure substance at 298.15 K is the enthalpy of the reaction where 1 mol of that substance in its standard state is formed from its elements in their standard reference states, all at 298.15 K. The so-called <i>reference states</i> of the elements at 298.15 K are their most stable physical states at that conventional temperature. ThermInfo contains values (in kJ mol^{-1}) of standard molar
---	---

Standard Molar Enthalpy of Phase Change

enthalpies of formation at 298.15 K and their associated uncertainties for crystalline, liquid, and gaseous compounds.

of a pure substance at 298.15 K is the enthalpy associated with the physical transformation of 1 mol of that substance from one phase to another, where the substance is in its standard state in both phases. Therefore, the standard molar enthalpy of a phase change is simply the difference between the standard molar enthalpies of formation of the substance in the two phases involved.

ThermInfo contains values (in kJ mol^{-1}) of standard molar enthalpies of fusion (transition from solid to liquid state), vaporization (transition from liquid to gaseous state) and sublimation (transition from solid to gaseous state), at 298.15 K, and their associated uncertainties.

Observations provide additional information, if applicable.

3. Bibliographic data provide complete references regarding the source of thermochemical data, including: author(s), journal/book title, year, volume and page(s).

SUPPORTING BIBLIOGRAPHY

- S1.** *CAS Registry and CAS Registry Numbers*; American Chemical Society: Columbus, OH. <http://www.cas.org/expertise/cascontent/registry/regsys.html> (accessed May, 2012).
- S2.** James, C. A.; Weininger, D.; Delany, J., *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008. <http://www.daylight.com/dayhtml/doc/theory/index.html> (accessed May, 2012).
- S3.** Weininger, D. SMILES, a Chemical Language and Information System .1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31–36.
- S4.** Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, 29, 97–101.
- S5.** Pedley, J. B., Thermochemical Data and Structures of Organic Compounds, TRC Data Series; Thermodynamics Research Center (TRC): College Station, TX, 1994; Vol. 1.
- S6.** Pedley, J. B.; Naylor, R. D.; Kirby, S. P., Thermochemical Data of Organic Compounds, 2nd ed.; Chapman and Hall: New York, 1986.