Review: Using Referenced Data Responsibly

Joseph C. Bloxham*, Mark E. Redd*, Neil F. Giles, Thomas A. Knotts IV and W. V. Wilding

April 9, 2020

1 Introduction

The American Institute of Chemical Engineers (AIChE) established the Design Institute for Physical Properties (DIPPR) in 1978 to respond to various industrial needs for reliable chemical data. Currently DIPPR maintains and expands the 801 Database as a part of the 801 Project. The 801 Database contains the most complete collection of critically-evaluated physical property values for pure compounds in the world. DIPPR uses a proprietary method to recommend the most trustworthy values *<!!–Data vs values vs correlations–>* and correlations to industrial professionals through evaluating all available information on each compound. Each database entry for a given pure compound includes a complete chemical profile with a recommended value or correlation. This expert evaluation process makes the DIPPR 801 Database the "Gold Standard" for pure-component property values. The DIPPR 801 project is sponsored by corporations and academic institutions worldwide. These sponsors use DIPPR recommended values for process design, simulation, and research purposes.

Some words should be defined in this context before our discussion:

- Data: information based on an experimental finding
- Value: a database entry that can include predicted or experimental information
- Correlation: an temperature-dependent equation for a property that has been fitted with with experimental or predicted values

DIPPR operates on 4 guiding principles. These are enumerated as follows:

1. Industrial sponsor control

   DIPPR is directed by a steering committee made up of made up of thermophysical *<!!–physical or thermophysical?–>* property experts from sponsoring companies. Sponsors select compounds to be reviewed and added to the database. This level of control affords sponsors of the project the opportunity to focus DIPPR's efforts on compounds and data that they need for their specific applications.

Additionally, sponsors are included in the review process. Sponsors critically review property values based on available data and can use their own expertise to ensure the quality of the database.

2. Critical evaluation

   DIPPR strives to comprehensively collect and evaluate all relevant data related to the properties included in the database. Before building a chemical profile for any compound, an exhaustive search of available data is undertaken. Peer-reviewed journals, databases, industrial handbooks, and other available sources are read and recorded. The collected data are then subjected to a thorough vetting process. Only the most reliable and consistent values are recommended.

   This careful evaluation relies on the expertise and training of both project staff and sponsors. Any value that is recommended in the database must first be examined by several staff and experts, each having carefully weighed the merits of each source, including experimental methods, unique issues to each property, institution and researcher. Other considerations influence the evaluation process as well. This process culminates in a final recommendation that must be approved both by sponsors and DIPPR's expert staff. This evaluation process makes the recommended values in the 801 database the "Gold Standard" for reliable thermophysical property values.

3. Consistency

   The 801 database is built on the idea that thermophysical properties are not independent of one another and often are strongly correlated. Therefore as chemical profiles are built and maintained, recommended values must be chosen to be consistent with one another. This is a primary consideration when choosing which values to recommend.

   Furthermore, the 801 project checks consistency by considering chemical family trends, comparison compounds, and other general rules. These trends inform the general reliability of a value under evaluation.

   These consistency checks ensure a more reliable database that considers these properties in every relevant context.

4. Completeness

   Finally, database chemical profiles require a value for every relevant property to be useful in process design. Therefore, a complete set of values for every property corresponding to each compound is a central principle of the database.

As mentioned before, experimental data are considered generally more reliable than smoothed or predicted values. However where reliable experimental data cannot be found for a given compound and property, DIPPR critically evaluates and employs the most accurate estimation methods to ensure each compound has a complete set of values. *<!!–The only exceptions to this are cases in which thermophysical properties are not applicable to a particular compound (e.g. water's chemical profile does not include flammability limits). –>*

This article is a response to instances where users misuse DIPPR 801 database content. This commonly arises from database users misinterpreting the content. All thermophysical property values in the database must be understood in the context of DIPPR's evaluation process to be correctly understood. Therefore, this work aims to promote a better understanding of the evaluation process and definitions used in the DIPPR 801 database, leading to less confusion and higher utility of the available data.

This work will cover the following topics:

- Common mistakes made while using the database and how to avoid them
- The DIPPR 801 Project's role in providing high quality chemical profiles
- Correctly interpreting reported uncertainties in the database
- The best practices for using the DIPPR database in scientific literature and research

2 Common Mistakes in Publications

As the "Gold Standard" in chemical property data, DIPPR is frequently referenced in many publications. While many researchers use the database correctly, often mistakes are made in the use of DIPPR values. These errors often involve a fundamental misunderstanding of DIPPR's purpose and definitions and can lead to significant issues with published work. By pointing out common mistakes in using DIPPR data, we hope that authors and reviewers will be able to improve the clarity and utility of the literature. The most common mistakes include interpreting DIPPR "accepted" and correlation values as experimental, and neglecting to properly credit original data sources. While we give specific examples in order to illustrate some of these errors, we do not intend to discredit any of the work in the papers we will cite as examples.

2.1 Insufficient Citations

In a 1996 article, Laugier and Richon report a phase envelope for 2 binary systems [Laugier and Richon, 1996]. In order to regress binary interaction parameters, Laugier et al. use 4-methyl-1-pentene's critical parameters and an acentric factor attributed to DIPPR. DIPPR did not have experimental data for these values, and the accepted DIPPR values are from Nokay's method for estimating critical temperature, Lyderson's method for predicting critical pressure, and the acentric factor is defined from those predicted properties. Though the phase envelope presented in the paper fit the data well, the values used come from predicted sources. Thus, the fit of the binary interaction parameters will have uncertainty associated with the prediction method.

Neglecting to cite the prediction methods and treating the DIPPR recommended values as a primary source is a frequent mistake in literature.

2.2 Using Correlation Values as Experimental

In a 2009 article, experimental data for the vapor pressure of several compounds are reported [Mokbel et al., 2009]. However, when comparing their data to literature values, Mokbel et al. compared their findings to a DIPPR correlation rather than the data that informed the DIPPR correlation. Crucially, they referred to the correlation results as "data from the DIPPR data bank" rather than identifying the data sources directly or stating that the comparison was to the DIPPR accepted correlation. While their results match well with theory and literature, more accurately describing the DIPPR references would make their paper more precise.

Temperature-dependent properties in the database are often fit to correlation equation to make simulation and process design easier. However, often researchers will treat this correlation as experimental data. DIPPR recommended correlations can be sourced from a variety of methods. Treating the correlations as experimental data are incorrect, and can be avoided by referencing DIPPR's cited reference .

2.3 Interpreting Recommended Values as Experimental

In a 2018 article, Keshavarz et al. published a quantitative-state-property-relationship (QSPR) for the prediction of autoignition temperatures (AIT) [Keshavarz et al., 2018]. In it, they claim to use experimental data for 54 compounds to relate molecular descriptors to AIT. However, upon closer inspection,19 AIT values they reference to the DIPPR database are not experimental at all, but are values recommended by DIPPR prediction methods. This is a flagrant error when building a prediction method, and by better understanding the designations of DIPPR values this error could have been avoided.

Often, researchers will assume that "accepted", the designation of DIPPR recommended values in the database, means that the value is experimental. While this is often the case, this is not universally true. As shared above, an accepted designation means that the value is the best estimate of the property in the expert eyes of the DIPPR evaluators. When possible, this is based on experimental data, but it could also be based on group contribution methods, family property charts, or from a relationship with a related property. This can cause issues, especially when researchers use accepted, non-experimental values to inform prediction methods or when evaluating equations of state.

3 Explanation of DIPPR Evaluation Processes

Many of the mistakes made in using DIPPR data could be avoided by a better understanding of the DIPPR evaluation process. The DIPPR database is composed of pure-component property data, correlations or prediction methods from various sources. This information is compiled to maintain a comprehensive collection of relevant information on each compound in the database. Relevant properties include 32 property constants and 15 temperature-dependent properties shown in Table 1. Data, correlations or prediction methods may be collected from a variety of sources including peer-reviewed literature, technical reports, handbooks, data compilations, and safety data sheets (SDS).

Table 1: Available Pure-Component Thermophysical Properties Studied and Recommended in the DIPPR 801 Database

Constant Properties	Constant Properties (cont.)	Temperature Dependent Properties
Molecular Weight	Heat of Fusion at Melting Point	Solid Density
Critical Temperature	Standard Net Heat of Combustion	Liquid Density
Critical Pressure	Flash Point	Heat Capacity of Ideal Gas
Critical Volume	Lower Flammability Limit	Heat Capacity of Liquid
Critical Compressibility Factor	Lower Flammability Limit Temperature	Heat Capacity of Solid
Acentric Factor	Upper Flammability Limit	Heat of Vaporization
Normal Boiling Point	Upper Flammability Limit Temperature	Second Virial Coefficient
Melting Point	Autoignition Temperature	Surface Tension
Triple Point Temperature	Radius of Gyration	Thermal Conductivity of Liquid
Triple Point Pressure	Solubility Parameter	Thermal Conductivity of Solid
Liquid Molar Volume	Dipole Moment	Thermal Conductivity of Vapor
Ideal Gas Enthalpy of Formation	Van Der Waals Volume	Vapor Pressure of Liquid

Constant Properties	Constant Properties (cont.)	Temperature Dependent Properties
Ideal Gas Gibbs Energy of Formation	Van Der Waals Area	Vapor Pressure of Solid or Sublimation Pressure
Ideal Gas Absolute Entropy	Refractive Index	Viscosity of Liquid
Standard Heat of Formation	Heat of Sublimation	Viscosity of Vapor
Standard Gibbs Energy of Formation	Parachor	
Standard Absolute Entropy	Dielectric Constant	

3.1 Evaluation Overview

Each of the following sections describes a general step in the evaluation process. This process does not happen chronologically as listed below. The process is iterative as correlation models, methods, and trusted data sources are changed until a “Gold Standard” chemical profile is built that represents the best information related to the physical properties.

By design, the process involves many different individuals carefully examining all the relevant information and coming to a consensus about the property value to be recommended. The final say is given by a set of DIPPR expert staff and sponsors. This human element in the process allows careful alterations to achieve a result that is complete, self-consistent and consistent with literature.

The following topics are relevant to understanding DIPPR’s evaluation process.

- The role of industrial sponsors
- An explanation of “Accepted” values
- The dynamic nature of the database
- The process of ensuring property consistency
- The focus on complete profiles for each compound

3.2 Industrial Sponsorship

More than forty corporations sponsor the 801 project to aid the design and operation of their chemical processes. Sponsorship of the 801 project ensures access to the most up-to-date version of the 801 database with the best, most accurate property values recommended. These corporations understand that having accurate chemical data is the difference between success and failure.

In the case that data are lacking or no suitable estimation method exists, sponsorship also funds original research continually tailored to sponsor needs and areas where the database may lack. Depending on the specific need, this research may experimentally measure property data, develop and validate new or improved estimation methods, or use molecular modeling to better understand and predict property values. Generally, the results of this research are published in peer-reviewed literature and then are evaluated at DIPPR before being added to the database officially.

Furthermore, DIPPR develops and distributes software and database interface tools to allow easy access to relevant data. This software is exclusively for sponsors *<!!–I believe this is the case. Is this wrong?–>* and contains tools for comparing, plotting and quickly evaluating compound properties and estimating compound properties where possible. All aspects of this software is tailored to sponsor needs and therefore act as a set of tools that are applicable to a variety of use cases.

Sponsors also direct DIPPR efforts. As sponsors submit compounds to be added to the 801 database, they ensure the database contains industrially relevant chemicals. This likewise helps to keep DIPPR's focus on the most important chemicals. The database is composed entirely of compounds that sponsors have submitted and therefore contains compounds common to a large variety of chemical processes. The staff and experts that maintain DIPPR ensure each compound added to the database is done so at the behest of sponsors and that great care is taken to find the best information possible for each compound.

3.3 Accepted Values

Once a compound has been selected for inclusion in the database, all relevant data are analyzed and evaluated by project staff. Data is sourced by searching available reference materials, online databases, and relevant journals. The evaluation focuses on proper vetting of experimental data, consistency across properties and chemical families, and building a complete chemical profile. Project staff then make recommendations about the "best" values and correlations for each property and their recommendations are reviewed by three additional experts. Unanimous approval is required before the chemical is added to the database. Relevant details of this process are discussed below, though the methodology cannot be discussed completely in this space.

The "best" values as determined by this vetting process appear in the database with an "Acceptance" field marked "accepted". Other available data can be found in the database with different acceptance values and express something about the results of DIPPR's expert review process. All "Acceptance" values are listed below with their attendant meanings in **Table 2**.

Table 2: Possible values and corresponding meanings of "Acceptance" values in the DIPPR 801 database

Acceptance Value	Meaning
Accepted	Indicates the value was vetted by DIPPR and found to be the most consistent and reliable value for the given property.
Rejected	The value was found to be questionable or incorrect based on DIPPR criteria for reliability and consistency and should not be recommended.
Not Accepted	The value was vetted by DIPPR and found to be reliable but was not deemed to be the “best” value to be recommended. Often, these values may be equivalent to the accepted value but are not from the primary source.
Unevaluated	This value has been entered into the database for consideration but has not yet been evaluated by DIPPR.

These Acceptance values are used in the database for quickly identifying which sources and values should be recommended to users of the database. The values which do not fall under the Accepted category are kept in the database both to avoid duplicate evaluations of the same data and to allow reevaluation of data that can happen as new research or information becomes available.

An Accepted designation does not mean that the value is experimental or that the uncertainty in the point is low. This simply means that the value reflects the most consistent and reliable information available. Accepted values can be experimental, predicted, or in the case of correlations, both. This information is clearly available by viewing the Data Type of a given property. DIPPR chooses the best properties available by constantly reviewing new data, checking for consistency between properties, and building a comprehensive view of each chemical species.

3.4 Dynamic Nature of the Database

Often, new data are found and entered into the database after a compound has been added and been given a complete chemical profile. When this occurs, the new data are submitted as “Unevaluated” until they can be evaluated. In light of new data, another evaluation may reassign accepted values to reflect the newer and better values. In this way, the 801 database is a dynamic and perpetually improving database. A particular snapshot of the database will reflect the best recommendations available at that time. However being a dynamic database, any property value may later be supplanted by better values as they are found and evaluated.

Changes in the DIPPR recommended values can be triggered by new experimental measurements, the development of better prediction methods, or the addition of similar compounds for comparison. New data are assigned an impact factor based on the amount, type, and the time since the last review of the chemical. The impact factor helps DIPPR investigators know when a review of a compound is merited. This process ensures that the database stays up to date and prioritizes the most significant improvements.

3.5 Inter-Property Consistency

The analysis method DIPPR uses allows for a more holistic picture of a chemical's properties than can be found in other data sources. Many properties are dependent on other properties, as illustrated in Figure 1. Analyzing these properties independently can lead to inaccuracies<*!!–COME BACK HERE–*>, so DIPPR evaluators ensure that properties are consistent with known relationships.

An example of this principle can be seen when evaluating the vapor pressure (VP), heat of vaporization (HVAP), and liquid heat capacity (C_p^l) of a given chemical. As shown by previous researchers (Cite Hogge and BLOX), these properties are related through the following three relationships:

$$VP = exp\left(A + \frac{B}{T} + Cln(T) + DT^E\right)$$

$$\Delta H_{vap} = T\Delta V \frac{dVP}{dT}$$

$$C_p^l = C_p^{IG} - T\int_0^{VP}\left(\frac{d^2V_v}{dT^2}\right)dP - \frac{d\Delta H_{vap}}{dT} + \left(\Delta V - T\left(\frac{dV_v}{dT}\right)\right)\left(\frac{d^2VP}{dT}\right)$$

These equations are exact mathematical relationships, and thus the fit of the vapor pressure curve should allow prediction of the other two properties. If there are disagreements between the values, DIPPR evaluators can carefully choose the correct sources or prediction methods on which to base the DIPPR accepted values. This also allows property information to inform other properties. Experimental data in one property can then be used to predict or verify the other two properties. These inter-relationships occur in many properties, and ensuring consistency across a chemical profile improves property prediction and makes DIPPR recommendations the "Gold Standard."

A useful example of using inter-property consistency to improve property prediction is illustrated by dimethyl oxalate, an industrially important chemical that was added to the database in 2019. When investigators looked at the available data, there were 14 different available sources for vapor pressure values. However, these values were relatively varied, and depending on the trusted data sources the prediction

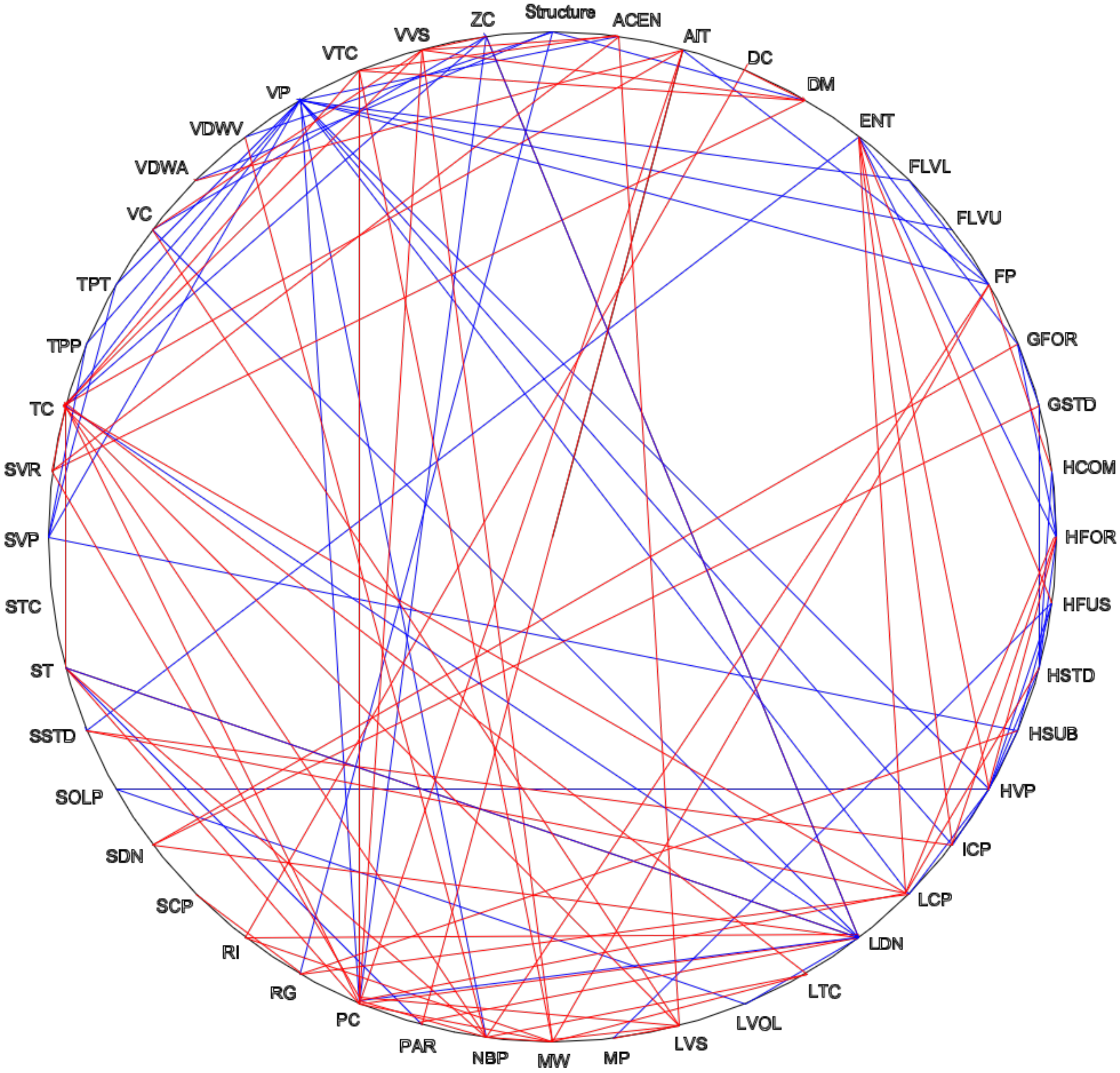

Figure 1: Interconnected properties used by the DIPPR database

could vary significantly at higher temperatures. Liquid heat capacity data were obtained, and by using the above thermodynamic relationships, the most likely correlation for the vapor pressure line was obtained. By ensuring inter-property consistency, the property correlations for all three of these properties were improved.

In addition to experimental data and consistency, comparisons to similar compounds and family relationships are considered in selecting the best values. As many have noted, properties often follow a predictable trend in chemical families [Costa et al., 2018]. With additional information about family members, the most correct experimental value can be selected among several sources by looking to see what family trends would predict.

*<!!–Need to rework this paragraph–>*Likewise, compounds with similar structures generally have similar properties. This is the foundational principle of group contribution prediction methods and QSPR, and is also often true of temperature dependent properties as well. While this principle is not always as certain as thermodynamic consistency, consistency in family trends and comparable compounds can be a useful tool. *<!!–(Perhaps show an example of data with comparison compounds, would make a nice figure)–>*

3.6 Completeness

DIPPR chemical profiles require a value for every relevant property to be useful in process design. In this way, DIPPR can give engineers and scientists a complete look at any chemical. Where experimental data are not available, DIPPR utilizes the highest accuracy prediction methods to fill any property values left without data. These methods are constantly updated and reviewed, and new methods vetted are as they become available. These range from group contribution methods, QSPRs, *ab initio* calculations, and others. By approaching each chemical holistically, individual inaccuracies can become apparent and be avoided. This self-consistency ensures DIPPR values are of the highest quality possible for a given compound.

4 DIPPR Uncertainty

DIPPR assigns uncertainty levels to both constants and temperature-dependent correlations. These error designations are assigned by DIPPR based on data type, acquisition method, and original author reported values. For predicted values, uncertainty is assigned based on general knowledge about the prediction method given the chemical family and property. For the sake of simplicity and to be conservative with uncertainty, DIPPR uses 9 quantized uncertainty levels to assign to any property value. These levels are given in Table 3. Due to the quantized nature of DIPPR uncertainty levels, the reported DIPPR uncertainty will rarely be exactly the same as author estimates.

Table 3: DIPPR's 9 Quantized Error Levels

Uncertainty Levels	
$< 0.2\%$	$< 25\%$
$< 1\%$	$< 50\%$
$< 3\%$	$< 100\%$
$< 5\%$	$> 100\%$
$< 10\%$	

While DIPPR provides uncertainty levels for thermophysical data, these are guidelines and still require some consideration in their use, especially if correlations are being used to fit to new prediction methods. Error designations for temperature dependent correlations are only based on the reliability of the data fitting the correlation and not the error associated with regression. Therefore, regression statistics are also available and should be examined in lieu of assuming the quantized error. For example, it is possible that, at the edges of the given temperature range, the error of the correlation may be higher than the assigned error designation. For this reason, we recommend examining the data sources when using a temperature based correlation near the edges of the available range.

5 How to Correctly Report DIPPR Values

Correctly using and citing DIPPR values can make data collection and processing easier, as well as increase the legitimacy of published findings. So far, we have discussed common mistakes in literature and DIPPR processes. With this information, we can now discuss best practices for authors and reviewers. Following these suggested best practices will ensure the database is interpreted correctly and is most useful to any who access it.

5.1 Understand Data Type

<!–Define data type more?–> Before using values collected from DIPPR, make sure the Data Type you have selected is appropriate for your application. For process design, the accepted DIPPR value is the best choice and is the central use case for the database. For creating prediction methods, parameterizing group contribution methods, or other scientific work, use only values that are based on experimental data. DIPPR software and tools make this easy by allowing database searches based on Data Type. If accepted values are used without reference to Data Type, there is a risk of only replicating the effectiveness of past methods rather than building new ones. This sort of error can introduce unforeseen uncertainty or even invalidate an

estimation method.

5.2 Reference Original Source

Where applicable, reference the primary source including the original author or method used. The 801 database includes the source of each value or correlation where possible. Using the primary source will ensure the property values are understood and reviewed in their original context. This often allows a lower uncertainty to be assigned to property values. Primary source use will also improve transparency in published papers, and allow for proper credit to be given to original researchers. Finally, this will also allow more meaningful comparisons between researchers and measurement methods.

In some cases, DIPPR has done original work when measuring or predicting. Where this is the case, DIPPR should be directly referenced. The database will clearly indicate this as measured or predicted by staff, or as an internal method. While these values are not experimental data, these estimates are the best property values available. Cite these values as a DIPPR recommended value based on internal methods.

5.3 Check Uncertainty

As discussed previously, to simplify the database and allow for staff insight into data reliability, DIPPR error designations are quantized. Though a point may be accepted, the prediction methods for a given property could have a large uncertainty. More exact predictions or data may be needed before a large capital investment.

5.4 Examine Correlation Limitations

When using DIPPR recommended correlations, ensure the temperature range of the correlation is representative of the application. As discussed previously, the error assigned to a correlation is based on the data informing the correlation, and regression error is still possible, and correlation values far from experimental data can have a much higher error than the assigned error level. For these reasons, checking the value sources and regression statistics are important. Finally, use caution when using a correlation near the edges of its temperature range.

6 Conclusion

We have discussed DIPPR terminology, DIPPR evaluations and processes, common mistakes in using DIPPR resources, and best practices in using DIPPR in scientific work. DIPPR is committed to maintaining the 801 database as the “Gold Standard” for pure-component physical property data. Using DIPPR’s proprietary methodology ensures the best results to be recommended to professionals. These recommendations will make correct use of DIPPR more accessible and meaningful for engineers and scientists everywhere.

References

[Costa et al., 2018] Costa, J. C. S., Mendes, A., and Santos, L. M. N. B. F. (2018). Chain length dependence of the thermodynamic properties of n-alkanes and their monosubstituted derivatives. *Journal of Chemical & Engineering Data*, 63(1):1–20.

[Keshavarz et al., 2018] Keshavarz, M. H., Jafari, M., Esmaeilpour, K., and Samiee, M. (2018). New and reliable model for prediction of autoignition temperature of organic compounds containing energetic groups. *Process Safety and Environmental Protection*, 113:491 – 497.

[Laugier and Richon, 1996] Laugier, S. and Richon, D. (1996). High-pressure vapor-liquid equilibria for ethylene + 4-methyl-1-pentene and 1-butene + 1-hexene. *Journal of Chemical & Engineering Data*, 41(2):282–284.

[Mokbel et al., 2009] Mokbel, I., Razzouk, A., Sawaya, T., and Jose, J. (2009). Experimental vapor pressures of 2-phenylethylamine, benzylamine, triethylamine, and cis-2,6-dimethylpiperidine in the range between 0.2 pa and 75 kpa. *Journal of Chemical & Engineering Data*, 54(3):819–822.