Thermophysical Property Reliability Issues in the Context of Automated Consumption

Kenneth Kroenlein [C, S], Rob Chirico, Vladimir Diky, Ala Bazyleva and Joe Magee
Applied Chemicals and Materials Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.
kenneth.kroenlein@nist.gov

Development of empirical and semi-empirical correlations and models, including force fields, is a data-intensive task. Such activities are frequently performed presuming an error-free set of targets, ignoring ordinary experimental uncertainty, possibility of misreported information, and from where a particular recommended value was ultimately derived. This point is emphasized by the large number of correlations in the open literature which report single-value deviation metrics from target collections which are substantially smaller than achievable experimental accuracy. The required effort to effectively curate an experimentally derived optimization targets is not insubstantial: exponential growth in publication rates and data generation in thermophysical properties has yielded tremendous challenges as well as potential rewards for data analysis groups. Data volumes have grown to such a degree that many traditional data collection and interpretation approaches cannot scale to remain comprehensive and current, to track shifting interests within research and industrial communities, or to effectively filter erroneous data from input streams. It is thus necessary to rely on a substantially increased role for digital archives, automated analysis and machine learning approaches. The approach adopted at the Thermodynamics Research Center (TRC) at the National Institute of Standards and Technology (NIST) is dynamic data evaluation, whereby a reliable and comprehensive data archive is used in conjunction with an algorithmically-encoded expert analysis in order to generate up-to-date property recommendations. These efforts have facilitated a decade's long collaboration with five major journals which report thermophysical and thermochemical property information, where reported data are vetted for consistency by TRC before being made available in a free and open context. This collaborative effort has yielded the unexpected statistic [1] that roughly 20% of published manuscripts have major errors in data or metadata independent of uncertainty characterization. Statistical analysis of error rates and types in analyzed manuscripts will be discussed, with a particular emphasis on potential impacts in the context of machine learning and other data-hungry technologies applied without significant data discrimination.