Computational and Theoretical Chemistry 1085 (2016) 46–55

Contents lists available at ScienceDirect

Computational and Theoretical Chemistry

journal homepage: www.elsevier.com/locate/comptc

Choosing an appropriate model chemistry in a big data context: Application to dative bonding

Qammar L. Almas, Benjamin L. Keefe, Trevor Profitt, Jason K. Pearson *

Department of Chemistry, University of Prince Edward Island, Charlottetown, PE C1A 4P3, Canada

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form 1 April 2016
Accepted 4 April 2016
Available online 8 April 2016

Keywords:
Computational chemistry
Dative bonding
Density functional theory
Cheminformatics
Coupled cluster theory
Molecular modeling
Molecular informatics
Chemical data repositories

ABSTRACT

We present an efficient, automated workflow for validating model chemistries in computational quantum chemistry by integrating several open-source web and semantic technologies within a discipline-specific context. We combine a range of open-source functionalities to (i) canonicalize the outputs of standard, popular computational chemistry software; (ii) store and index data within a central networked repository; (iii) query the data against a range of relevant properties; and (iv) compute robust statistical measures of model accuracy. Our workflow is tested by committing data from 10,304 *ab initio* potential energy surface calculations to a central repository and subsequently applying nested queries and analytics. Specifically, we investigate the performance of 44 different model chemistries (coupled with polarized, double-zeta basis sets) at reproducing CCSD(T)/CBS(D,T) potential energy surfaces of eight different Lewis acid/base pairs, whose dative bonds are known to be challenging to model for many electronic structure theories.

The performance of our workflow is measured by its computational speed and its ability to distill large datasets into robust, concise, and informative metrics. We report run times for each critical step in the workflow including data ingestion, canonicalization, querying, and post-query analytics, noting that querying is completed on a millisecond time scale. Employing the average absolute deviation (AAD) is shown to be a robust metric of the quality of underlying computational models, particularly when averaged over each unique chemical system in the dataset.

Our testing reveals the relatively poor performance of common density functional theories on reproducing the potential energy surfaces of some dative bonds. In particular, we note that boron-containing Lewis acids are less accurately modeled than aluminum-containing analogues, and phosphorus-containing Lewis bases are less accurately modeled than nitrogen-containing analogues. In particular, we find that the PBE0-D3(0)/6-31G* model chemistry is a robust model for the chosen dataset, though its accuracy varied widely depending on the chemical system.

1. Introduction

In a recent article, Clark et al. commented on the growing importance of data management in chemistry [1]. In particular, making research data immediately accessible and machine readable would be a major step forward in enabling algorithms to realize chemical insight on a scale not possible by traditional publication methods. It is imperative that we develop best practices for data stewardship if we are to take advantage of the vast potential and promise of "Big Data" in the chemical sciences. Specifically, because *computational* chemistry is inherently digital in nature, it is somewhat surprising that relatively slow gains have been made in terms of the development of trusted open repositories and access to the myriad of data produced on an ongoing basis. While there have been important advances both in terms of digital infrastructure [2–7] and applications [8–21] for "big data" in computational quantum chemistry, more is needed to broaden access and increase the value of such data to those who can benefit most from it.

Importantly, computational quantum chemistry can be used in principle to predict any observable chemical property; the problem is of course that the solution of the Born-Oppenhiemer Schrödinger wave equation (SWE) [22] (a usual target in computational quantum chemistry) is too computationally demanding to be tractable for the vast majority of chemical space. Accurate computational prediction of chemical properties on larger and larger swaths of chemical space are accessible though by *approximating* the solution

* Corresponding author.
E-mail address: jpearson@upei.ca (J.K. Pearson).

http://dx.doi.org/10.1016/j.comptc.2016.04.005
2210-271X/

to the SWE, or by resorting to density functional theory (DFT) [23], which has become a standard technique with broad applicability. However, the number of approximate models and density functionals for this purpose is large and continues to grow [24]. This growth is driven both by continuous algorithmic developments and the desire to probe novel chemical problems on larger scales. Unfortunately, the quality of computed properties varies significantly with the chosen model and also depends on the nature of the chemical system and the chemical property one is interested in. Consequently, the lack of a reasonably "universal" electronic structure model necessitates a validation step prior to any serious computational investigation into the behavior of chemical species. This is achieved most commonly by assessing the performance of a variety of selected models in predicting properties of interest by comparing against a standard reference, which are usually experimentally derived data or data from high-level *ab initio* calculations when possible. Once a model is found that yields a favorable comparison between its computed results and known references, one can then proceed confidently in predicting similar properties for other similar chemical systems.

As a result of the necessity of such a validation, there is a large body of literature highlighting *specific* use cases and applications of *specific* models to *specific* problems (for example see Ref. [25]). Though it is often possible for one to make use of previous model validations more generally, there are two important problems with this paradigm. The first is that the literature is an inherently *static* record. New models are developed often and the scope of chemical space and property space of interest to the community is a moving target. As such, a static reference is not always appropriate for assessing the ever-changing landscape of model chemistries and applicable problems. Second, there is often a wealth of data "hidden" from the literature that could have otherwise been quite useful, the so-called "dark data". For example, it is routine to predict harmonic vibrational frequencies to assess the nature of a stationary point along a potential energy hypersurface when optimizing molecular structures. However, it is far less common for the numerical values of these computed frequencies (and associated atomic displacement vectors and absorption intensities, etc.) to ever see the light of day even within supporting information files. Consequently, the reader learns about the performance of model chemistries on properties of interest to the author but nothing about their performance on other response properties, which may otherwise be readily available if one had access to the original data.

A much more attractive solution to the problem of validating model chemistries would be to refer to a "living" central repository of benchmark data "on the fly", where research results are continuously contributed and accessible (in their entirety) by the scientific community. This is not an entirely new idea [2,5,7,26], though much is left to be done to realize it. If one imagines a suitable repository housing the detailed computational results published over the past several decades of development, one immediate and obvious utility of such a resource would be to automate practices that are common, effectively standardized, and of significant importance within the computational chemistry community, *like model validation*. With large-scale datasets becoming an emerging currency in chemistry [7,5,21,26,27] (particularly in computational chemistry) and given the large and growing number of publications in computational quantum chemistry (for example, the interested reader is directed to the "Quantum Chemistry Literature Database", http://qcldb2.ims.ac.jp/pub.html) one may reasonably expect that a database sufficient for this purpose is possible and probably even likely within the near future. In fact, we have ongoing work in our laboratory to this end. Therefore, in the current work we focus on the development of an efficient computational workflow for harnessing the power of such a repository for the purpose of validating model chemistries, while simultaneously presenting new data of interest to the broader chemical community.

In particular, as a test case we investigate the performance of 24 computational electronic structure models in reproducing the CCSD(T)/CBS(D,T) potential energy surface of dative bond stretching between 8 unique Lewis acid–base pairs. Such systems (at stretched bond lengths) can be useful approximations to modeling the electronic structure of "frustrated Lewis pairs" (FLPs) [28,29], which are of emerging interest in many areas and are novel species to which we may apply our recently developed *Localized Pair Model* [30,31] analysis techniques. The electrostatic attraction between a Lewis pair, countered by the steric hindrance due to bulky ligands, elongates the equilibrium dative bond and creates a so-called "frustrated" environment. In order to overcome such steric frustration, these species give rise to novel reactivity and are the first examples of metal-free hydrogenation [28,32–41] in addition to being efficient catalysts for a wide range of synthetic transformations [42–44]. FLP chemistry has been investigated intensively by experimental techniques in recent years but there have been relatively few theoretical investigations of the electronic structure and reactivity of FLPs [45] owing to the theoretical challenge in modeling them accurately [46]. The properties of our FLP models, such as long-range dispersion forces, non-equilibrium structures, and dative bonding have all been shown to be particularly problematic for modern DFT [47] and so these species will afford a particularly relevant and interesting test case where it is not otherwise clear how one should proceed with modeling [46]. Moreover, FLPs (by definition) consist of bulky ligands, which generally equates to them being intractable to model with highly accurate *ab initio* electronic structure techniques and so we chose to investigate the performance of a wide variety of DFT models in reproducing the potential energy surface of dative bond stretching in small molecule model systems.

By examining the fidelity of each of our chosen approximate electronic structure models against a coupled cluster complete basis set CCSD(T)/CBS(D,T) reference dataset, we can offer insight into appropriate models for such systems but also demonstrate a generalized and automated workflow of model validation. Such a generalized workflow requires open repositories that simultaneously (i) fully index computational chemistry data, (ii) offer customized, nested queries against that data, and (iii) cover a suitable breadth of model chemistries, properties and chemical space for our current purposes. To the best of our knowledge we are not aware of any suitable repositories that presently fit these criteria, though we expect that this won't always be the case. For this reason we resorted to generating our own data and consequently chose systems that were both relevant to our interests and could offer reasonable complexity for the purpose of demonstrating our general approach.

2. Methods

2.1. Description of workflow

To facilitate the storage and retrieval of computational chemical information we have chosen to host our data within an *Islandora* instance [48]. Islandora is an open-source software framework designed to facilitate the management and discovery of digital assets, and was originally developed at the authors' institution. It integrates Drupal [49], Fedora [50], and Solr [51] to provide a robust platform capable of facilitating a wide range of applications and queries. What is particularly attractive about such a platform are the advanced full-text searching capabilities as well as the built-in support for open standards such as extensible markup