Published in final edited form as:

Curr Top Med Chem. 2012 ; 12(18): 1946–1956.

A Survey of Quantitative Descriptions of Molecular Structure

Rajarshi Guha[†] and Egon Willighagen[‡]

[†]NIH Center for Advancing Translational Science, 9800 Medical Center Drive Rockville, MD 20850, U.S.A [‡]Department of Bioinformatics - BiGCaT, Maastricht University, P.O. Box 616, NL-6200 MD Maastricht, The Netherlands

Abstract

Numerical characterization of molecular structure is a first step in many computational analysis of chemical structure data. These numerical representations, termed descriptors, come in many forms, ranging from simple atom counts and invariants of the molecular graph to distribution of properties, such as charge, across a molecular surface. In this article we first present a broad categorization of descriptors and then describe applications and toolkits that can be employed to evaluate them. We highlight a number of issues surrounding molecular descriptor calculations such as versioning and reproducibility and describe how some toolkits have attempted to address these problems.

1 Introduction

Computational methods play an important role in many chemical disciplines ranging from drug discovery to materials science. There are a plethora of techniques that differ in terms of computational complexity, time requirements and so on. However the common requirement underlying all these methods is a formal description of a the molecular structure. There are many ways to “describe” a molecule. A common approach is to describe the connectivity, taking into account the types of atoms and bonds. In other words, explicit representations of chemical structure such as SMILES, MDL/Symyx SD files and so on. While these descriptions are vital to modern chemical information systems, they do not necessarily allow computational techniques to be directly applied to them.

Methods that aim to predict chemical and biological properties generally require a numerical description of chemical structures. Such numerical forms range from a set of 3D coordinates which coupled with appropriate atom types, is sufficient for methods such as quantum mechanical (QM) approaches and docking to more abstract numerical descriptions derived from 2D or 3D representations which can be useful in statistical approaches. It is now possible to evaluate thousands of numerical descriptors of chemical structure. As will be discussed later, many of these descriptors are closely related or capture the same information, allowing one to be substituted for another. The selection of relevant descriptors is a well-known problem and given a large collection of them, approaches to identify a suitable subset have been discussed extensively in the literature [1, 2]. Figure 1 is a summary depiction of the major types of descriptors and the form of molecular structure information that is required to compute them. The depiction is very general and focuses on small molecule descriptors. As will be described in the following sections, molecular descriptors can be calculated for many chemical entities, not just small organic molecules.

In addition to there being many possible descriptors defined in the literature, there are also multiple implementations of a give descriptor. These implementations are available in the form of libraries (which require one to write a program to use them) or complete

applications (graphical user interface (GUI) or command line). As a result, not only must one choose one or more descriptors that are relevant to the problem at hand, but one must be concerned about how they are calculated and whether such a calculation can be reproduced across different implementations of those descriptors. It is easy to understand *why* two implementations of the same descriptor can lead to different results. The primary reasons being differences in the chemistry model of the framework or toolkit used to implement the descriptor. For example, a descriptor that calculates the number of aromatic atoms may be implemented using two toolkits with differing aromaticity models, and hence it is possible that the values generated by the two implementations will differ. Other sources of differences include parameters that may be involved in the descriptor calculation and reference data values (such as atomic radii, electronegativity values) that are employed during descriptor calculation. While most implementations will employ the same data sources for standard concepts (e.g., atomic weights), slight differences in these types of input data can lead to differences in the final descriptor value [3]. As a result, in most cases, two implementations of a descriptor do not usually give the exact same value, though they are usually quite similar. Explicitly explaining the differences may or may not be possible (it is usually more difficult in the cases of commercial implementations for which source code is not available) [4].

2 A Categorization of Descriptors

2.1 Numerical descriptors

We have noted that a molecular structure can be characterized using many numerical descriptors. In this section we describe a categorization of descriptors. Admittedly, the grouping is somewhat arbitrary but serves as a useful guide. In addition, the categorization also takes into account the nature of the chemical structure being considered - some descriptors are only useful when applied to small molecules, whereas other descriptors are defined specifically for polymers or protein structures. Broadly, one can classify descriptors based on the nature of structural information they require.

Constitutional descriptors only require atom and bond labels and usually represent counts of different types of atoms or bonds. While very simplistic they can play a useful role in a variety of applications ranging from summaries of physicochemical properties to predictive modeling [5].

Topological descriptors take into account connectivity along with atom and bond labels. They consider the molecule as a labeled graph and characterize it using graph invariants. In addition to graph invariants, topological descriptors based on information theory such as entropy have also been defined[6]. An advantage of these descriptors is that do not require any intensive preprocessing steps such as 3D coordinate generation and conformational analysis. In addition, many graph invariants can be rapidly computed for graphs corresponding to small molecule structures (even though they may be computationally intractable on much larger graphs). An example of a topological descriptor is the Wiener index [7, 8]. It is simply the sum of the edge counts in the shortest paths between all pairs of non-hydrogen atoms. From a physical standpoint, the descriptor characterizes the branching in a molecular structure. While this class of descriptors have been used extensively in predictive modeling applications [9, 10, 11] (for example, the Wiener index described above has been useful in models of boiling point data), a key drawback is their general lack of interpretability. The abstract nature of many graph invariants and information theoretic descriptors, their use in predictive models makes it difficult to explain the model in simple physicochemical terms, though various attempts have been made at interpretations [12, 13].

Geometric descriptors require a 3D conformation as input and therefore the computations are not as fast as topological descriptors which do not require the 3D coordinate generation step. Nearly all geometric descriptors work on a single conformation and for cases where multiple conformation must be considered, an averaging procedure is usually employed. Examples of these descriptors include the gravitational index [14] and moment of inertia descriptors. Geometric descriptors encompass many aspects of a molecular structure. For example, a number of them characterize the molecular surface. The simplest such example is the surface area descriptor (and the corresponding volume descriptor). A simple extension of such a descriptor is to characterize distributions of a physicochemical property over the molecular surface such as the Charged Partial Surface Area descriptors (CPSA) [15] that characterize the partial charge distribution over the molecular surface. Note that surface descriptor calculations based on analytical [16] or tessellation algorithms [17] can be slow and that parameterized methods such as the Topological Polar Surface Area (TPSA) [18] and MOE van der Waals Surface Area (VSA) descriptors [19] that are based on precalculated surface area values derived from a list of functional groups.

It is important to note that there is a difference between the information a descriptor captures and how it is calculated. The above categories are based on input requirements. This gives rise to situations where the molecular volume property can be described by geometrical descriptors, e.g. using an analytical approaches [17], and with approximate, constitutional descriptors, such as the very fast algorithm use by the van der Waals Volume as a Sum of Atomic and Bond Contributions (VABC) approach [20] which is based on group contributions. The latter descriptor has been implemented in the Chemistry Development Kit (CDK) and correlates well with other approximations to the molecular volume (Figure 2).

Another important class of geometrical descriptors are those that describe molecular shape. This property is a key factor in ligand-receptor interactions and is an important approach to virtual screening. A variety of descriptors have been devised to characterize molecular shape. The use of overlapping atom centered Gaussians to represent molecular shape [21] represents a significant computational advantage over the traditional analytical approach of intersecting hard spheres, supporting extremely rapid shape comparisons [22]. An alternative approach described by Ritchie et al [23] employed a spherical harmonic based representation of surfaces, that in contrast to the Gaussian approach allows one to generate characterizations of the surface and hence the shape at varying resolutions. Another approach to multi-resolution shape representations is the use of shapes [24], which represent a parameterization of the convex hull derived from the molecular coordinates. This approach has been employed to characterize shape and volume [25] though it is more computationally intensive than the methods mentioned previously. In addition to functional representations of molecular shape, a number of simpler formulations have been defined. For example, the “shape signature” approach described by Zauhar et al [26] is a novel application of a ray tracing method that generates a description of the molecular shape. Ballester et al [27] described the Ultrafast Shape Recognition (USR) approach that characterized molecular shape in terms of the distribution of distances between the atoms and a set of pre-defined points. This procedure characterizes the shape in the form of a 12-D real valued fingerprint. As a result, the calculation procedure is rapid and once calculated shape similarities can be evaluated quickly via a simple distance calculation. It should also be noted that shape indices based on the molecular topology have also been described [28], which are very rapid to calculate (somewhat analogous to the VSA descriptors) but do not directly represent the molecular shape and instead infer it from topological connectivity.

The last category consists of QM descriptors. These descriptors can also be applied to problems such as Quantitative Structure-Activity Relationship (QSAR) models (though of course the time complexity of QM derived descriptors do not make them amenable for large

scale virtual screening applications). These descriptors can be evaluated at various levels of theory. For example, partial charges, Highest Occupied Molecular Orbital (HOMO) and Lowest Occupied MOlecular Orbital (LUMO) energies are commonly evaluated using semi-empirical methods, while Eroglu and Turkmen [29] described density functional theory (DFT) derived descriptors such as electronegativity, hardness and ionization potential (though Puzyn et al question the need for such a high level of theory [30]). Sullivan et al [31] employed frontier orbital theory and DFT to generate quantum mechanical analogs of Hammett substituent constants [32] and showed that they agreed well with empirically derived constants.

The descriptor categories listed here are not necessarily distinct, many hybrid descriptors have been described that combine one or more categories. For example, the charge indices described by Galvez et al [33] combine partial charge information and topological connectivity. Similarly the CPSA descriptors combine surface area and partial charge information. When considering hybrid descriptors, it is useful to note the autocorrelation descriptors that characterize the distribution of molecular properties across a molecule in terms of an autocorrelation function. These descriptors may be based on topological or geometric distances and can be combined with arbitrary molecular properties [34, 35].

2.2 Fingerprints

Besides the numerical descriptors outlined so far, another useful approach are those of fingerprints. Traditionally, fingerprint descriptors are represented in the form of bit strings. These binary fingerprints can be divided into hashed fingerprints, where substructures (such as paths of length n) are converted to a string representation and then hashed into a randomly selected bit positions, or keyed fingerprints, where each bit position corresponds to a unique substructural feature. The nature of the features can range from simple functional groups (hydroxyl, carbonyl) and topological substructures (paths, chains, cycles) to atom environments [36, 37] and pharmacophoric elements [38]. Depending on the definition, fingerprints can work with topology only or may require 3D conformations. The latter class of fingerprints are usually related to 3D pharmacophores and most fingerprint definitions only require connectivity information.

Fingerprints are a compact representation of a chemical structure and are used in a variety of ways ranging from database searches looking for similar structures or substructures to virtual screenings studies [39]. In many search applications, the binary nature of fingerprints allow for efficient algorithms to compare molecules and evaluate similarity and a number of heuristics have been developed that allow for rapid similarity searches [40, 41]. In virtual screening scenarios, fingerprints can be used as features. In many cases, the entire fingerprint is employed as a set of independent variables. Since this can lead to poor predictive models, some sort of feature selection is usually employed and there have been studies that address the problem of identifying subsets of bits from a fingerprint that lead to optimal performance in predictive models (albeit, usually in a target-class specific manner) [42].

2.3 Describing other chemistries

We next consider descriptors that are designed for specific classes of molecular systems.

Descriptors for polymeric systems—These systems represent a challenge for descriptor development. There have been reports that employed traditional small-molecule descriptors [43, 44] that consider the polymer in terms of the individual monomers. A number of approaches have been described that includes both deterministic, geometry derived descriptors [45] as well as stochastic methods [46]. For the case of polypeptides the

descriptors generally tend to be based on the individual amino acids. Examples include the Vector of Hydrophobic, Steric, and Electronic (VHSE) family of descriptors described by Mei et al [47], a set of side chain descriptors from Collantes and Dunn [48], field based descriptors [49] and the AAindex database [50] which allows rapid lookup of amino acid properties.

Descriptors for inorganic compounds—Inorganic materials are sufficiently different from small molecules that traditional descriptors are not very useful. Recently, a number of descriptors for inorganic materials are very similar to constitutional descriptors - essentially counts of various crystal parameters such as accessible surface areas and volumes [51, 52, 53]. Other descriptors characterize the periodic voids of porous materials via tessellation methods [54, 55].

Other methods developed for crystal structures have been developed too, for example, using a radial distribution functions [56], which capture the packing of organic crystal structures. Such approaches have been used to cluster crystal structures [57] and compare properties [58].

Bioactivity as a descriptor—While the descriptors mentioned so far are derived from some form of chemical structure, an alternative approach is to employ observed biological activities of a molecule as descriptors of that molecule. This approach was taken by Sedykh et al [59] who hypothesized that dose response data points from high throughput dose-response assays could be employed as biological descriptors. They observed that when such descriptors were combined with traditional chemical structure descriptors, the predictive performance of models developed using the combined set was better than those developed using conventional approaches. A similar approach can be taken using the Prediction of Biological Activity Spectra for Chemical Substances (PASS) methodology [60], though this is an indirect method - one must predict the PASS profile from the chemical structure, which can then be used in subsequent predictive models.

Atom and bond descriptors—It should also be noted that various toolkits also calculate descriptors for chemical entities like atoms and bonds. These descriptors include simple element property descriptors, but also geometrical descriptor, such as radial distribution functions that describe the chemical environment of those atoms. These, for example, have been used to predict proton shift in nuclear magnetic resonance [61]. Similarly, bond descriptors can be used to predict bond properties.

3 What is a Useful Descriptor?

Given that we can generate thousands of descriptors, it is critical to ask, what makes a descriptor useful? Fundamentally, a descriptor must correlate structural features with some physicochemical property and show minimal correlation with other descriptors. In addition, a generally useful descriptor will be applicable to a broad class of molecules. Furthermore, a descriptor should show minimal degeneracy. That is, the descriptor should generate different values for structurally different molecules, even if the structural differences are small. An example is the case of stereochemistry. Most constitutional and topological descriptors that are not designed to take this into account will generate the same numerical values for all the stereoisomers of a given chiral molecule. We would consider such a descriptor to be degenerate. In addition to degeneracy, a descriptor should be continuous, in the sense, that small structural changes should lead to small changes in the value of the descriptor.

The requirements listed here ensures that subsequent analyses involving descriptors are robust. But there are other aspects of a descriptor that can make it more or less useful. For

example, a descriptor that can be rapidly calculated and is not dependent on experimental properties can be considered more useful than one that is computationally intensive and / or dependent on experimental results. In addition, it is important for many scenarios that the descriptor have some form of physical interpretability. For example, a model that employs descriptors to predict cytotoxicity will be most useful if it is numerically accurate as well as providing insight into which structural features confer toxicity. In such a scenario, abstract descriptors do not allow us to provide a physical explanation of the models predictions. Another scenario is the ability to map descriptor values back to the structure for visualization purposes such as the “Glowing Molecule” depictions of Segall et al [62]. Such visualizations are practical only when descriptor values can be related to structural features. Examples include surface based descriptors such as CPSA’s as well as explicit structural descriptors such as paths, chains etc.

However, there are certain scenarios where some of the above characteristics are *undesirable*. One such scenario is in the safe exchange of structures, such that two parties can perform analyses on the structure-derived data, without having to know the structures explicitly. In such cases, the ability to go from a set of descriptor values back to a chemical structure should be minimized. It has not been conclusively proven that such a “one-way” descriptor can be defined. For example Masek et al [63] have shown that one can regenerate a target structure (or substructural features) or a set of closely related analogs given a combination of descriptors such as BCUTs. Tetko *et al.* [64] suggest an approach to prevent reverse engineering but avoid the use of descriptors and instead sharing structures of related molecules, rather than the molecules of interest. One could also employ degenerate descriptors, which would represent a large family of molecules, such that the target molecule would be hidden. This is likely unreliable as it would be difficult to perform analyses (e.g., predictive models) with such descriptors.

This discussion has focused on aspects of descriptors that make them useful in computational models. It is also relevant to note that the utility of a descriptor can be dependent on the problem being considered. More specifically, certain end points may require specific molecular features to be taken into account. Descriptors that do not, will not be useful in developing models for that endpoint. For example, an accurate boiling point model will require descriptors that capture information regarding molecular branching, hydrogen bonding potential and so on. Similarly, a model to predict activity against a cytoplasmic or nuclear target, will likely require that molecular characteristics enabling passage across the cell membrane be taken into account. In such a case, a descriptor characterizing the lipophilicity of the molecule will be useful. These points also apply to other types of computational approaches, such as clustering. In the end, the utility of a descriptor is determined by whether it encapsulates sufficient and relevant information. In general a single descriptor will not satisfy these requirements and thus one usually works with sets of descriptors.

4 Descriptor Implementations

A variety of descriptor implementations are available, with most cheminformatics toolkits providing implementations of a common set of descriptors and fingerprints. Table 1 summarizes the descriptor classes implemented by several toolkits and applications. Note that due to space limitations this is not a comprehensive list of implemented descriptors across different applications and the reader is referred to the documentation accompanying the software packages.

While library calls for descriptor evaluation is the most flexible approach, it is not useful for non-programmers as well as users who simply need a descriptor matrix for modeling or

searching. As a result, GUI or command line wrappers around descriptor APIs enable easier usage. The CD-KDescUI application (<http://rguha.net/code/java/cdkdesc.html>) is a Java Swing application that exposes the CDK descriptor and fingerprint API's. It allows the users to specify a SMILES or SD file and evaluate a selected set of descriptors or fingerprints and export them to a text file. The application also has a command line mode that can be run in server environments. A similar application is PaDEL [65] that includes the CDK descriptors and also implements a number of additional descriptors such as the extended topochemical atom descriptors [66], McGowan Volume [67], and the Klekota & Roth substructure counts [68].

There are also web interfaces wrapping descriptor calculation tools and some of these interfaces have a programmable feature. For example, AMBIT2 implement the OpenTox API providing a REST-like interface to descriptor calculation [69, 70] (see Figure 3). Another online application programming interface for descriptor calculation is provided by the SADI platform, which, like AMBIT2, uses semantic web technologies for descriptor calculation [71].

In addition to libraries and their wrappers there are various GUI applications that allow easy descriptor calculations. Examples include Bioclipse [72] (see Figure 3), MOE (Chemical Computing Group) and Maestro (Schrodinger, Inc.) as standalone tools, but also as a plugin to Microsoft Excel, called LICSS [73]. Workflow tools are also a useful class of applications for descriptor generation and allow one to easily generate descriptors from multiple sources [74] (see Figure 3). Given the focus of this article on Open Source implementations, we have noted the availability in Table 1. While a number of implementations are not strictly Open Source according to OSI definitions, the fact that they provide free academic licenses does allow them to be used somewhat freely.

Due to our experience with the CDK project we provide a brief overview of the architecture of the descriptor calculations implemented by the CDK. First, the toolkit explicitly differentiates between atom, bond and whole molecule descriptors. There are a total of 30, 6 and 51 descriptor algorithms for atoms, bonds and molecules respectively. Note that each individual descriptor algorithm may calculate multiple descriptors. For example, the CPSA descriptor evaluates 30 individual descriptor values. Thus for a given molecule it is possible to generate approximately 300 descriptor values.

Every descriptor class implements an interface allowing them to return values in a uniform manner, viz., via a DescriptorValue object. This class can represent the result of a descriptor calculation and includes both the descriptor values (via a IDescriptorResult object) as well as meta-data on the descriptor itself via a DescriptorSpecification object (see below for more details). The descriptor classes are designed such that in the event of a calculation error (invalid, missing or nonsensical input), null values are returned, rather than throwing an exception. This is useful during automated descriptor calculation, where one desires a "rectangular" descriptor matrix, from which one could filter out the undefined descriptors. While the user can pick and choose specific descriptors to calculate, the API provides a class (DescriptorEngine) to specifically evaluate all or a subset of descriptors for a molecule at one go. Currently the descriptor implementations are not threadsafe, making them unsuitable for parallelization. This is a drawback since many of the descriptors can be slow. For example, the descriptors [75] are implemented using SMARTS matching to identify the various paths and chains described by Kier *et al.*, rather than explicitly walking the molecular graph.

4.1 Comparing implementations

Given the multiple implementations of many descriptors, it is natural to compare their performance. As noted previously, it is desirable, but unlikely, that implementations of the same descriptor using different cheminformatics toolkits will be identical. Gupta et al employed SMARTS based descriptors from the CDK and MOE to develop decision tree models to predict human liver microsomal metabolic stability [76]. Their results indicated very similar performance between the two implementations. In general, one would expect a high degree of correlation between different implementations of well-defined descriptors. Here “well defined” indicates that the underlying parametrization or chemistry model is completely specified, so that different implementations use them identically. For example, correct implementations of the TPSA descriptor are expected to correlate very well, and this is shown in Figure 4 for the ACD Labs and CDK implementations.

A number of physicochemical descriptors commonly employed in predictive modeling are themselves models of an experimental property. For example, log P can be experimentally measured and a number of algorithms have been developed to predict log P [77, 78]. Given the utility of log P in drug discovery, a number of implementations are available. Given that this descriptor is a surrogate for an actual experimental property, it is reasonable to compare calculated log P values from different implementations to the experimentally observed values, rather than between themselves. Figure 5 compares computed log P values from the CDK (specifically, an implementation of the XlogP method), ChemAxon and ACD Labs for a set of 10,000 molecules taken from the logPstar dataset.

4.2 Descriptor Naming & Versioning

As noted above, different implementations of the same descriptor can generate slightly different values depending on the underlying chemistry model, differences in reference values and so on. Even within the same software implementation, different versions of the same descriptor can generate different values. Thus to maintain reproducibility in studies that employ molecular descriptors, some form of versioning is crucial. The simplest form of version is to manually note the version of the implementation used for the calculation and report it along with the results. Though useful, it is not amenable to automation. In some cases, toolkits will provide an API call to retrieve the version information, which can then be associated with a descriptor calculation.

A topic closely related to versioning is descriptor naming. The preceding sections have discussed a categorization of descriptors and have used a number of commonly used names. When performing calculations in an automated manner, or employing remote services for descriptor calculations [79], it is useful to be able to refer to descriptors via a standardized naming scheme. Currently the QSAR modeling community lacks such a scheme, though the CDK, a Java library for cheminformatics [80] does implement such a scheme. The “descriptor specification” approach defines a set of classes that include information of the implementation title, identifier and vendor. The identifier may or may not include a version number. The actual details of the a descriptor are listed in a dictionary, along with an optional namespace. The dictionary entry is referenced by the descriptor specification and this design thus allows one to seamlessly work with multiple implementations of the same descriptor from different vendors and keep track of which version is employed throughout a calculation or workflow. In addition, the CDK provides a hierarchical annotation of descriptors, grouping them into different classes. Since this annotation is based on the Resource Description Framework (RDF) [81], it allows one to go beyond simple groupings and perform reasoning on descriptors. A simple example is to identify functionally equivalent descriptors (such as simple transformations of a base descriptor - square and cube roots of the gravitational index is an example [82]) in an automated fashion. While this

infrastructure is currently only supported by the CDK, adoption of it (or some similar framework) would enable easier interoperability between descriptor implementations and enhance reproducibility. The framework is, however, actively used by software using the CDK descriptors, such as Bioclipse [4] and Ambit [83]. Moreover, the ideas have been adopted by the CHEMINF ontology for cheminformatics [84], which is used by the CHESS framework which also uses the CDK for descriptor calculations [71].

5 Summary

Molecular descriptors play a fundamental role in cheminformatic and chemometric analyses and as we have described in this article it is possible to evaluate thousands of descriptors using a variety of software. Though the bulk of descriptors are well defined in the literature, multiple implementations of the same descriptor can yield different results. These differences arise from differences in the underlying chemistry models and reference data used by the implementations. As shown by Gupta et al [76], the performance of predictive models developed using commercial or Open Source descriptor implementations is very similar. While different tools differ in the specific number and types of descriptors that they calculate, the fact many descriptors are correlated with others suggests that it does not matter too much which set of descriptors are used in an application. Of course, this cannot be a general rule - specific applications may require a specific set of descriptors. For example, a predictive model for boiling points may not perform very well if branching and intermolecular forces are not taken into account by the descriptors in the model. This is not to say that all descriptor tools exhibit similar performance. For example, it is clear from Figure 5 that the ACD Labs implementation of log P performs significantly better than other implementations. Though this is not a completely rigorous comparison (the methods underlying the ACD Labs, CDK and ChemAxon implementations are different), it does highlight that certain implementations of a descriptor can fare better than others - especially those cases where the descriptor is based on a predictive model.

Given these challenges it is imperative that descriptor tools provide access to version information. This allows the user to provide an exact specification of how the descriptor was calculated. However, one aspect that has not been standardized across different implementations is the naming scheme. In other words, tools that evaluate the same descriptor may name them in a similar but not exactly identical manner. This makes automated comparisons and merging of descriptors from different sources problematic. To address this, the CDK has developed the concept of a “descriptor specification” which is associated with a descriptor value and includes information on the vendor, descriptor title and a reference to an entry in a descriptor dictionary that contains more details of the descriptor in RDF. While the extra descriptor metadata is currently used to implement a simple classification scheme, it is conceivable that in the future the specification approach could be adopted by multiple vendors, allowing for automated reasoning over descriptor implementations.

In summary, there is no dearth of tools to generate molecular descriptors. Many of them are available under liberal Open Source licenses, though these implementations do not necessarily cover all the descriptors described to date. The tools range from toolkits (which implies that one must write a program to generate descriptors) to self-contained GUI or command line tools. Given the caveats regarding implementation specific differences, it is important to keep track of provenance information when calculating descriptors to ensure reproducibility.

References

1. Miller, A. Subset Selection in Regression. 2. Chapman & Hall/CRC; Boca Raton, FL: 2002.
2. Kohavi R, John G. Wrappers for Feature Subset Selection. *Artif Intell.* 1997; 97:273–324.
3. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL. The Blue Obelisk–Interoperability in Chemical Informatics. *J Chem Inf Model.* 2006; 46:991–998. [PubMed: 16711717]
4. Spjuth O, Willighagen EL, Guha R, Eklund M, Wikberg JEE. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *Journal of Cheminformatics.* 2010;2. [PubMed: 20222949]
5. Bender A, Glen R. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J Chem Inf Model.* 2005; 45:1369–1375. [PubMed: 16180913]
6. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F. On Entropy-Based Molecular Descriptors: Statistical Analysis of Real and Synthetic Chemical Structures. *J Chem Inf Model.* 2009; 49:1655–1663. [PubMed: 19530677]
7. Nikolic S, Trinajstić N, Randić M. Wiener Index Revisited. *Chem Phys Lett.* 2001; 333:319–321.
8. Wiener H. Structural Determination of Paraffin Boiling Points. *J Am Chem Soc.* 1947; 69:2636–2638.
9. Garcia-Domenech R, Galvez J, Julian-Ortiz Jd, Pogliani L. Some New Trends in Chemical Graph Theory. *Chem Rev.* 2008; 108:1127–1169. [PubMed: 18302420]
10. Randić M, Zupan J. On Interpretation of Well-Known Topological Indices. *J Chem Inf Comput Sci.* 2001; 41:550–560. [PubMed: 11410029]
11. Besalu E, Gallegos A, Carbo-Dorca R. Topological Quantum Similarity Indices and Their Use in QSAR: Application to Several Families of Antimalarial Compounds. *Commun Math Comp Chem.* 2001; 44:41–64.
12. Todeschini R, Cazar R, Collina E. The Chemical Meaning of Topological Indices. *Chemom Intell Lab Sys.* 1975; 97:6609–6615.
13. Stanton D. On The Physical Interpretation of QSAR Models. *J Chem Inf Comput Sci.* 2003; 43:1423–1433. [PubMed: 14502475]
14. Katritzky AR, Mu L, Lobanov VS, Karelson M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J Phys Chem A.* 1996; 100:10400–10407.
15. Stanton D, Jurs P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assisted Quantitative Structure Property Relationship Studies. *Anal Chem.* 1990; 62:2323–2329.
16. Connolly M. Analytical Molecular Surface Calculation. *J Appl Crystallography.* 1983; 16:548–558.
17. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies. *J Comp Chem.* 1995; 16:273–284.
18. Ertl P, Rohde B, Selzer P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J Med Chem.* 2000; 43:3714–3717. [PubMed: 11020286]
19. Labute P. A Widely Applicable Set of Descriptors. *J Mol Graph Model.* 2000; 18:464–477. [PubMed: 11143563]
20. Zhao YH, Abraham MH, Zissimos AM. Fast calculation of van der Waals volume as a sum of atomic and bond contributions and its application to drug compounds. *J Org Chem.* 2003; 68:7368–7373. [PubMed: 12968888]
21. Grant J, Pickup B. A Gaussian Description of Molecular Shape. *J Phys Chem.* 1995; 99:3503–3510.
22. Grant J, Gallardo M, Pickup B. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J Comp Chem.* 1999; 17:1653–1666.

23. Ritchie D, Kemp G. Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *J Comp Chem*. 1999; 20:385–395.
24. Edelsbrunner H, Mücke E. Three Dimensional Alpha Shapes. *ACM Trans Graph*. 1994; 13:43–72.
25. Wilson JA, Bender A, Kaya T, Clemons PA. Alpha shapes applied to molecular shape characterization exhibit novel properties compared to established shape descriptors. *J Chem Inf Model*. 2009; 49:2231–2241. [PubMed: 19775113]
26. Zauhar RJ, Moyna G, Tian L, Li Z, Welsh WJ. Shape signatures: A new approach to computer-aided ligand- and receptor-based drug design. *J Med Chem*. 2003; 46:5674–5690. [PubMed: 14667221]
27. Ballester P, Graham Richards W. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J Comp Chem*. 2007; 28:1711–1723. [PubMed: 17342716]
28. Kier L. A Shape Index from Molecular Graphs. *Quant Struct-Act Relat Pharmacol, Chem Biol*. 1985; 4:109–116.
29. Eroglu E, Turkmen H. A DFT-based quantum theoretic QSAR study of aromatic and heterocyclic sulfonamides as carbonic anhydrase inhibitors against isozyme, CA-II. *J Mol Graph Model*. 2007; 26:701–708. [PubMed: 17493855]
30. Puzyn T, Suzuki N, Haranczyk M, Rak J. Calculation of quantum-mechanical descriptors for QSPR at the DFT level: Is it necessary? *J Chem Inf Model*. 2008; 48:1174–1180. [PubMed: 18510372]
31. Sullivan J, Jones A, Tanji K. QSAR treatment of electronic substituent effects using frontier orbital theory and topological parameters. *J Chem Inf Comput Sci*. 2000; 40:1113–1127. [PubMed: 11045804]
32. Hammett L. The Effect of Structure Upon the Reactions of Organic Compounds. Benzene Derivatives. *J Am Chem Soc*. 1937; 59:96–103.
33. Galvez J, Garcia R, Salabert M, Soler R. Charge Indexes. New Topological Descriptors. *J Chem Inf Comput Sci*. 1994; 34:520–525.
34. Moreau G, Broto P. Autocorrelation of Molecular Structures: Application to SAR Studies. *Nouv J Chim*. 1980; 4:757–764.
35. Broto P, Moreau G, Vandycke C. Molecular Structures – Perception, Auto-correlation Descriptor and SAR Studies - Auto-correlation Descriptor. *Eur J Med Chem*. 1984; 19:66–70.
36. Bremser W. HOSE - A Novel Substructure Code. *Anal Chim Act*. 1978; 103:355–365.
37. Bender A, Mussa H, Glen R, Reiling S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J Chem Inf Comput Sci*. 2004; 44:1708–1718. [PubMed: 15446830]
38. Renner, S.; Fechner, U.; Schneider, G. Pharmacophores and Pharmacophore Searches. In: Langer, T.; Hoffmann, R., editors. Chapter Alignment-free Pharmacophore Patterns - A Correlation Vector Approach. Vol. 32. Wiley-VCH; Weinheim, Germany: 2006. p. 49-79.
39. Sun H. An accurate and interpretable bayesian classification model for prediction of hERG liability. *ChemMedChem*. 2006; 1:315–322. [PubMed: 16892366]
40. Baldi P, Hirschberg D, Nasr R. Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive OR. *J Chem Inf Model*. 2008; 48:1367–1378. [PubMed: 18593143]
41. Swamidass SJ, Baldi P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J Chem Inf Comput Sci*. 2007; 47:302–317.
42. Xue L, Godden JW, Stahura FL, Bajorath J. Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. *J Chem Inf Comput Sci*. 2004; 44:1275–1281. [PubMed: 15272835]
43. Mattioni BE, Jurs PC. Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. *J Chem Inf Comput Sci*. 2002; 42:232–240. [PubMed: 11911692]
44. Katritzky AR, Sild S, Lobanov V, Karelson M. Quantitative Structure Property Relationship (QSPR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J Chem Inf Comput Sci*. 1998; 38:300–304.

45. Edvinsson T, Arteca GA, Elvingson C. Path-space ratio as a molecular shape descriptor of polymer conformation. *J Chem Inf Comput Sci*. 2003; 43:126–133. [PubMed: 12546545]
46. González-Díaz H, Gia O, Uriarte E, Hernández I, Ramos R, Chaviano M, Seijo S, Castillo JA, Morales L, Santana L, Akpaloo D, Molina E, Cruz M, Torres LA, Cabrera MA. Markovian chemicals “in silico” design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J Mol Model*. 2003; 9:395–407. [PubMed: 13680309]
47. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers*. 2005; 80:775–786. [PubMed: 15895431]
48. Collantes ER, Dunn WJ. Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogs. *J Med Chem*. 1995; 38:2705–2713. [PubMed: 7629809]
49. Norinder U. Theoretical amino acid descriptors. Application to bradykinin potentiating peptides. *Peptides*. 1991; 12:1223–1227. [PubMed: 1815210]
50. Kawashima S, Ogata H, Kanehisa M. AAindex: Amino Acid Index Database. *Nucl Acids Res*. 1999; 27:368–369. [PubMed: 9847231]
51. Willems T, Rycroft C, Kazi M, Meza J, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous and Mesoporous Materials*. 2012; 149:134–141.
52. Haranczyk M, Sethian J. Automatic structure analysis in high-throughput characterization of porous materials. *J Chem Theory Comput*. 2010; 6:3472–3480.
53. Mackay A. Descriptors for Complex Inorganic Structures. *Croatica Chemica Acta*. 1984; 57:725–736.
54. Anurova N, Blatov V. Analysis of ion-migration paths in inorganic frameworks by means of tilings and Voronoi-Dirichlet partition: A comparison. *Struct Sci*. 2009; B65:426–434.
55. Carr DA, Lach-hab M, Yang S, Vaisman II, Blaisten-Barojas E. Machine Learning Approach for Structure-Based Zeolite Classification. *Microporous and Mesoporous Materials*. 2009; 117:339–349.
56. Willighagen EL, Wehrens R, Verwer P, de Gelder R, Buydens LMC. Method for the computational comparison of crystal structures. *Acta crystallographica Section B, Structural science*. 2005; 61:29–36.
57. Wehrens R, Willighagen E. Mapping databases of X-ray powder patterns. *RNews*. 2006; 6:24–28.
58. Willighagen EL, Wehrens R, Melssen W, de Gelder R, Buydens LMC. Supervised Self-Organizing Maps in Crystal Property and Structure Prediction. *Crystal Growth & Design*. 2007; 7:1738–1745.
59. Sedykh A, Zhu H, Tang H, Zhang L, Richard A, Rusyn I, Tropsha A. Use of in vitro HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of in vivo toxicity. *Environ Health Perspect*. 2011; 119:364–370. [PubMed: 20980217]
60. Poroikov V, Filimonov D, Lagunin A, Glorizova T, Zakharov A. PASS: identification of probable targets and mechanisms of toxicity. *SAR and QSAR Env Res*. 2007; 18:101–110. [PubMed: 17365962]
61. Aires-De-Sousa J, Hemmer MC, Gasteiger J. Prediction of ¹H NMR Chemical Shifts Using Neural Networks. *Analytical Chemistry*. 2002; 74:80–90. [PubMed: 11795822]
62. Segall M, Champness E, Obrezanova O, Leeding C. Beyond profiling: using ADMET models to guide decisions. *Chem Biodivers*. 2009; 6:2144–2151. [PubMed: 19937845]
63. Masek B, Shen L, Smith K, Pearlman R. Sharing Chemical Information without Sharing Chemical Structure. *J Chem Inf Model*. 2008; 48:256–261. [PubMed: 18254609]
64. Tetko IV, Abagyan R, Oprea TI. Surrogate data—a secure way to share corporate data. *J Comput Aided Mol Des*. 2005; 19:749–764. [PubMed: 16267691]
65. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comp Chem*. 2011; 32:1466–1474. [PubMed: 21425294]
66. Roy K, Ghosh G. QSTR with extended topochemical atom indices. 2. Fish toxicity of substituted benzenes. *J Chem Inf Comput Sci*. 2004; 44:559–567. [PubMed: 15032536]

67. Zhao YH, Abraham MH, Zissimos AM. Determination of McGowan volumes for ions and correlation with van der Waals volumes. *J Chem Inf Comput Sci*. 2003; 43:1848–1854. [PubMed: 14632432]
68. Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics*. 2008; 24:2518–2525. [PubMed: 18784118]
69. Jeliaskova N, Jeliaskov V. AMBIT RESTful web services: an implementation of the Open-Tox application programming interface. *Journal of Cheminformatics*. 2011; 3:18. [PubMed: 21575202]
70. Hardy B, et al. Collaborative development of predictive toxicology applications. *Journal of Cheminformatics*. 2010; 2:7. [PubMed: 20807436]
71. Chepelev L, Dumontier M. Chemical Entity Semantic Specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. *Journal of Cheminformatics*. 2011; 3:20. [PubMed: 21595881]
72. Spjuth O, Helmus T, Willighagen E, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg J. Bioclipse: An Open Source Workbench for Chemo- and Bioinformatics. *BMC Bioinformatics*. 2007; 8:59. [PubMed: 17316423]
73. Lawson K, Lawson J. LICSS - a chemical spreadsheet in microsoft excel. *Journal of Cheminformatics*. 2012; 4:3. [PubMed: 22301088]
74. Truszkowski A, Jayaseelan KVV, Neumann S, Willighagen EL, Zielesny A, Steinbeck C. New developments on the cheminformatics open workflow environment CDK-Taverna. *Journal of cheminformatics*. 2011; 3:54. [PubMed: 22166170]
75. Kier L, Hall L. Molecular Connectivity VII: Specific treatment to heteroatoms. *J Pharm Sci*. 1976; 65:1806–1809. [PubMed: 1032667]
76. Gupta RR, Gifford EM, Liston T, Waller CL, Hohman M, Bunin BA, Ekins S. Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties. *Drug Metab Dispos*. 2010; 38:2083–2090. [PubMed: 20693417]
77. Wang R, Lai L. A New Atom-Additive Method for Calculating Partition Coefficients. *J Chem Inf Comput Sci*. 1997; 37:615–621.
78. Wildman S, Crippen G. Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci*. 1999; 39:868–873.
79. Dong X, Gilbert K, Guha R, Heiland R, Kim J, Pierce M, Fox G, Wild D. A Web Service Infrastructure for Chemoinformatics. *J Chem Inf Model*. 2007; 47:1303–1307. [PubMed: 17602467]
80. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr Pharm Des*. 2006; 12:2110–2120.
81. Taylor K, Gledhill R, Essex J, Frey J, Harris S, De Roue D. Bringing Chemical Data onto the Semantic Web. *J Chem Inf Model*. 2006; 46:939–952. [PubMed: 16711712]
82. Wessel M, Jurs P, Tolan J, Muskal SM. Prediction of human intestinal absorption of drug compounds from molecular structure. *J Chem Inf Comput Sci*. 1998; 38:726–735. [PubMed: 9691477]
83. Jeliaskova N, Jeliaskov V. AMBIT RESTful web services: an implementation of the Open-Tox application programming interface. *Journal of Cheminformatics*. 2011; 3:18. [PubMed: 21575202]
84. Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. *PLoS ONE*. 2011; 6:e25513. [PubMed: 21991315]
85. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*. Wiley-VCH; Berlin: 2002.
86. Godden J, Xue L, Stahura F, Bajorath J. Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pac Symp Biocomput*. 2000:566–575. [PubMed: 10902204]
87. Liu K, Feng J, Young S, Power MV. A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. *J Chem Inf Model*. 2005; 45:515–522. [PubMed: 15807517]

Figure 1.

A graphical summary of descriptor types and the type of input information required. As one goes from top to bottom the calculations become more intensive, but the results capture aspects of molecular structure more realistically.

Figure 2.

A comparison of the VABC descriptor for molecular volume (as implemented by the CDK) and other approximate implementations of molecular volume (a 2D approach and a grid based approach), as implemented by MOE 2008.10. All values are in $\text{\AA}^3$ and both plots are overlaid with a loess fit.

Figure 3. Four tools to calculate descriptors: Bioclipse (top left), which allows selecting descriptor implementations from various independent tools, CDK-Taverna (top right), which is an extension to Taverna to calculate descriptors with the CDK, Ambit2 (bottom left), which implements the OpenTox API providing a REST-based API for descriptor calculation and wraps various descriptor calculation tools. The bottom right screenshot shows the Microsoft Excel plugin LICSS that uses the CDK for 2D diagrams and descriptor calculation. This screenshot is provided by Kevin Lawson, Syngenta, UK, and reproduced with permission.

Figure 4.

A comparison of Topological Polar Surface Area values generated using the CDK and ACD Labs software, for 57,857 molecules taken from PubChem AID 1996.

Figure 5.

A comparison of experimental versus calculated logP values generated using the CDK, ACD Labs and ChemAxon software, for 10,000 molecules taken from the proprietary logPstar dataset.

Table 1

A summary of descriptor implementations covering Open Source and commercial cases.

Name	License	Toolkit	Application	Descriptors	Fingerprints
CDK	LGPL	Yes	Yes	Constitutional, topological, geometric, electronic and hybrid	MACCS, PubChem, Kier & Hall Estate keys, Hashed linear and circular
CODESSA	Commercial	No	Yes	Constitutional, topological, geometric, electronic, QM, thermodynamic	
Dragon	Commercial	No	Yes	Extensive list of all classes of descriptor types and implements all those described in Ref. [85]	
Chem	Commercial	Yes	Yes	Constitutional, geometric, topological, electronic, hybrid	Hashed, circular, ECFP/FCFP
JOELib	GPL	Yes	Yes		Keys from Xue et al [86]
Molconn-Z	Commercial	No	Yes	topological	
OpenBabel	GPL	Yes		constitutional	MACCS, hashed
PowerMV [87]	GPL		Yes	constitutional, atom pairs, Burden numbers, pharmacophoric	
RDKit	New BSD	Yes		Constitutional, topological, EState, partial charges and VSA combining EState, partial charges and MR	MACCS, hashed, atom pairs, topological torsions