Comprehensive Medicinal Chemistry III

30010. Fingerprints and other molecular descriptions for database analysis and searching

Dávid Bajusz[1], Anita Rácz[2,3], and Károly Héberger[2]

[1] Medicinal Chemistry Research Group, Research Centre for Natural Sciences, Institute of Organic Chemistry Hungarian Academy of Sciences, Magyar tudósok krt. 2, H-1117 Budapest, Hungary
E-mail: bajusz.david@ttk.mta.hu Phone: + 36 1 382 69 74

[2] Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences,
Magyar tudósok krt. 2, H-1117 Budapest, Hungary
E-mail: racz.anita@ttk.mta.hu and heberger.karoly@ttk.mta.hu Phone: + 36 1 382 65 09

[3] Department of Applied Chemistry, Szent István University, Villányi út 29-43, H-1118 Budapest, Hungary

Cite as follows:

Dávid Bajusz, Anita Rácz, and Károly Héberger*, Chapter 3.14 – Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching.
In: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering
Comprehensive Medicinal Chemistry III - **Volume 3** Editor-in-Chiefs Samuel Chackalamannil, David P. Rotella and Simon E. Ward; *In silico* methods Eds: A. Davies, C. Edge,
Available online 13 June 2017, Pages 329–378
Oxford: Elsevier. Elsevier (2017)
http://dx.doi.org/10.1016/B978-0-12-409547-2.12345-5
ISBN: 9780128032008

Abstract

In this chapter we strive to provide a comprehensive but reasonably compact overview of the various possibilities for the computational representation of molecules. This includes a detailed introduction to the most commonly used chemical file formats (complemented with a few novel or more specific representations), a thorough overview of the theoretical backgrounds of various molecular fingerprints and descriptors, and a complete subchapter devoted to similarity measures and data fusion approaches. Finally, we provide a list of the most important online chemical databases and conclude the chapter with a short outlook on present trends and future expectations.

Keywords

molecular fingerprint, similarity, distance metric, molecular descriptor, database analysis, data fusion, cheminformatics, drug design, chemical file format, virtual screening

1. Introduction

Molecules possess an abundance of properties. 3D structure, atom connectivity, shape and physicochemical descriptors such as molecular weight or log*P* (logarithm of the *n*-octanol/water partition coefficient) are just a few of such properties that are usually of interest in the overlapping domains of cheminformatics, molecular modeling and drug design. Many of these properties have a quite intricate nature: for example 3D structure, which can rarely be characterized with a single conformation (but instead an ensemble of conformations), or even simple properties, such as molecular weight that depends on the isotope distribution of the composing atoms. The affinity of the molecule towards various types of environment (such as a physiological solution, a mixture of solvents or a particular protein binding pocket) – which is of particular interest in the mentioned fields – expands the property space of even a single compound beyond comprehension.

As such, it is currently impossible to give a perfectly accurate and complete computational representation of a molecule that accounts for its every possible aspect; however in most cases this is not necessary. Nonetheless, one should always be aware of the implied simplifications in the computational representation of molecules, especially when making predictions based on them. In other words, the main responsibility of the computational chemist (or drug designer or cheminformatician) is to know, which aspects of a molecule are important to be considered for the given application. If it is so, then one can utilize the predictive power of computational chemistry techniques to provide invaluable guidance for medicinal chemistry endeavors.

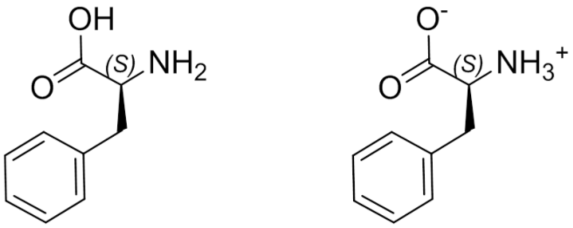
In the recent decades, medicinal chemistry has been armed with a growing number of large, highly featured and annotated databases. These databases contain the ever growing public (and proprietary) knowledge that is being aggregated by medicinal chemistry research and they provide an invaluable foundation for retrospective analyses and prospective/predictive work. Databases containing chemical information can be queried in a number of ways, each of which requires specific types of molecular representations. Substructure searches require an efficient chemical query language (*e.g.* SMARTS), while similarity searches need means to quantify the similarity of two molecules and appropriate structural representations that allow for such operations (*e.g.* molecular fingerprints). It is also common to query databases for specific values or ranges of quantifiable molecular properties such as molecular weight or log*P*.

In this chapter, we provide a comprehensive overview of molecular representations, including file formats, fingerprints and molecular descriptors – with an emphasis on those that are widely applied in works involving large databases. We also dedicate a subchapter to molecular similarity (and the diverse methodologies that have been developed to elaborate this sub-field), as a ubiquitously applied concept in cheminformatics and drug design. Last but not least, we provide a short overview of the most prominent online resources of molecular information that ultimately fuel most of the computational chemical research that involves a consideration of large compound sets.

2. Chemical file formats and data structures

The means to store chemical information started to develop in parallel with the first computers (in fact, earlier than computers became standard work tools). Since that time, generations of file formats have been developed (and have occasionally become obsolete). Since each file format was developed with a specific purpose, each encodes different aspects of the structure and properties of a molecule. Some of them are optimized for compactness, while others allow the storage of various properties or other information besides the 3D structure of the molecule.

A complete overview of the available chemical file formats would be a futile effort – and probably pointless, as there is little relevance (from a cheminformatics point of view) of covering file formats that are developed for *e.g.* specific molecular dynamics or quantum chemistry programs. Instead, we will dedicate this subchapter to introduce the reader to today's most widely used, cross-platform, human-readable chemical file formats, with an emphasis on their use in database searching and other subfields of cheminformatics. Throughout these subsections, we will occasionally refer to the structural representation of L-phenylalanine (see Figure 1) in the various file formats as a common example.



1 **2**

Figure 1. Neural (1) and zwitterionic (2) forms of L-phenylalanine. The α-carbon is annotated with the absolute configuration of the stereogenic center.

Before proceeding with the description of specific file formats, we make note of a convenient means for the interconversion of various chemical file formats. While many modeling software suites are capable of handling multiple file formats (and consequently, to convert them), computer programs dedicated specifically for this tasks also exist. Probably the best known example is Open Babel, an open source software for the interconversion of chemical structures between an abundance of file formats (over 110 for the current version).[1]

2.1 Line notation formats

The principle of line notation formats is to store chemical structures unambiguously, but as compactly as possible (usually on a single line). In fact, the IUPAC Recommendations on Organic and Biochemical Nomenclature can be thought of as a line notation system as well.[2] (Though ironically it is not necessarily more human-readable than *e.g.* the SMILES format, especially for large molecules.) In that "format", the representation of L-phenylalanine would be *(2S)-2-amino-3-phenylpropanoic acid* (or somewhat less intuitively, *(2S)-2-azaniumyl-3-*

phenylpropanoate for the zwitterionic form). While IUPAC names can unambiguously describe arbitrarily complicated molecules, the names themselves are often very complicated as well. Also, a substructure can usually be specified in multiple ways (depending on other groups, the order of priority, *etc.*), which essentially removes text-based IUPAC name queries from the toolbox of substructure searching.

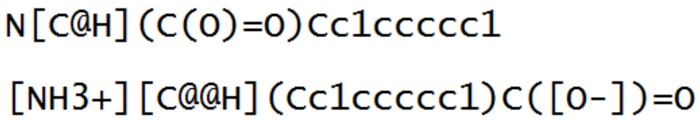
From the available tools for chemical identification and documentation, today's standard is the CAS (Chemical Abstracts Service) Registry number.[3,4] CAS maintains a constantly updated database (the CAS Registry) to store every reported chemical structure (*cca.* 111 million at the time these lines are written) with a uniquely assigned identifier, the CAS Registry number (*e.g.* for L-phenylalanine: *63-91-2*). In addition to the database itself, the CAS Registry powers the two major chemical information services of CAS: Scifinder (for querying chemical structures and reactions, primarily in the literature)[5] and STN (a search engine that provides access to patent content).[6] While CAS Registry numbers are quite compact linear notations, they do not provide direct, human-readable information on molecular structure, nor relation/proximity to another substance (*e.g.* the CAS number for racemic phenylalanine is *150-30-1*).

Efforts to provide compact means of chemical structure representation using line notations date back to the late 1940's when the Wiswesser Line Notation (WLN) was introduced.[7] Nowadays, multiple line notation formats are available. While we will only briefly cover those that are mostly of historic interest (*e.g.* WLN and ROSDAL), we introduce today's standard tools such as SMILES and InChI in more detail.

2.1.1 Simplified molecular-input line-entry system (SMILES)

SMILES was introduced by Weininger in 1987.[8] By that time, several line notation systems have already existed and their flaws were also well-known, which provided a need – but at the same time also a knowledge base – for the development of a mature, but easily understandable line notation language. SMILES provides a structure specification tool that is easily handled by both computers and humans, which is probably the main reason for its success in the past three decades. During this time, SMILES have become the *de facto* standard of molecular line notations, and a basis of inspiration for other line notation systems, such as SLN (see subchapter 2.1.3.3). A concise overview of the format follows (a more complete documentation is available in the original SMILES paper[8] and on the website of Daylight[9]).

A SMILES string is a series of characters specifying primarily the heavy atoms (and optionally, hydrogens and bonds) of the molecule. Atoms are represented by their atomic symbols (first letter in upper case, second letter in lower case), and by default must be enclosed in square brackets: however, this is not required for elements in the "organic subset": *B, C, N, O, P, S, F, Cl, Br* and *I*, if the number of hydrogens conforms to the lowest normal valence minus those bonds that are explicitly given. Neighboring atoms are connected by bonds (single by default), branches are enclosed in parentheses. Atoms in aromatic rings are denoted with lower case letters. Hydrogens must be explicitly specified by default, but they are implied in the absence of square brackets (which is most often the case). Single and aromatic bonds are implied, but they can be specified explicitly as well with "–" and ":", respectively (double and triple bonds are denoted by "=" and "#"). Attached hydrogens, charge, stereochemistry and isotope numbers can be specified inside the square brackets of the given atom, *e.g. [15NH4+]* denotes an ammonium ion with a ^{15}N atom. As an example, SMILES strings of L-phenylalanine are provided in Figure 2.



```
N[C@H](C(O)=O)Cc1ccccc1

[NH3+][C@@H](Cc1ccccc1)C([O-])=O
```

Figure 2. SMILES strings of the neutral (above) and zwitterionic (below) form of L-phenylalanine. Breaking points of the phenyl ring are labeled as number 1. Attached hydrogens, charges and stereochemical information are provided inside the square brackets. Note that the order of the substituents of the α-carbon is interchangeable, but affects the specified stereochemistry of the α-carbon: looking from the nitrogen towards the α-carbon, the three following substituents (in sequential order) can appear counter-clockwise (@) or clockwise (@@).

Probably the greatest drawback of SMILES is the lack of an internal canonicalization approach. (As seen on the above example, atoms can be mapped to the SMILES string in an arbitrary order.) While Weininger and colleagues have published a canonicalization algorithm shortly after the release of the SMILES format itself, their approach could not handle stereochemistry.[10] Moreover, this method was included in the commercial software of Daylight[11], resulting in many different variations of canonicalization being implemented in several cheminformatics toolkits. Recently, a universal SMILES representation based on the InChI canonicalization procedure was introduced by O'Boyle.[12] Schneider *et al.* have developed and published a robust, open-source, freely available SMILES canonicalization algorithm.[13] They conclude that their canonical labeling algorithm (which is also available in the open-source cheminformatics package RDKit[14]) "could be combined with the universal SMILES representation proposed by O'Boyle".

SMILES is also able to specify reactions. Notably, two available "grammars" for specifying reactions in SMILES are *reactant>>product* and *reactant>agent>product*. Disconnected molecules (*e.g.* two reactants) are separated by a dot (this is also available for single compounds that have *e.g.* counterions). The functionality of SMILES has also been extended to support substructure querying and matching (SMARTS) and the description of generic reactions (SMIRKS). We well briefly cover these extensions in the following subchapters.

2.1.1.1 Smiles arbitrary target specification (SMARTS)

SMARTS is an extension of the SMILES language that was developed with the explicit intention to provide a query language for substructure searching.[15] As a result, any SMILES string is an acceptable SMARTS query by definition. To facilitate substructure searching, a number of additional features have been implemented for SMARTS, to allow for more general queries. These include atomic primitives (for example * denotes any atom, while *a* denotes any aromatic atom), bond primitives (*e.g.* ~ for any bond, @ for any ring bond) and logical operators (*e.g. !* for *not*, *&* for *and*). Another interesting feature is recursive SMARTS, which enables to define an atomic environment starting with the atom of interest in the following form: *$(SMARTS)*. With these features, quite complex queries can be formulated, *e.g. C[$(aaO)&$(aaaN)]* would match any molecule that has a methyl carbon as a substituent of an aromatic ring with an oxygen in an ortho position and a nitrogen in a meta position.

Like SMILES, SMARTS is able to handle – and in particular, to query – reactions. A reaction query may consist of reactant, agent and product parts (all optional), as introduced briefly above. The queries can be run against a set of "target" reactions (also specified in the SMILES format). Atom mapping is also supported in the query and the target, thus changes on a specific atom can be specified. For example, while *C=O>>C-O* would match any reaction that has a molecule with a C=O group on the reactant side and a molecule with a C-O moiety on the product side, *[C:1]=O>>[C:1]-O* is specific to the same carbon atom with a doubly bonded oxygen on the reactant side and a singly bonded oxygen on the product side. However, this would still match *e.g.* a carboxyl group that is unchanged during the reaction as the oxygen was not specified to be the same atom on the reactant and product sides. Transformation of the double bond to a single bond is thus specified with the *[C:1]=[O:2]>>[C:1]-[O:2]* query. (However, to produce any matches, the atoms must be mapped in the target reactions as well!) A complete documentation of SMARTS is available on the website of Daylight Inc.[15]

2.1.1.2 SMIRKS

SMIRKS is a hybrid of SMILES and SMARTS, developed to provide a means to specify generic reactions.[16] It is a restricted version of reaction SMARTS with a set a rules that act as constraints. For example, the transformation of the oxo group in the SMARTS example above (see previous subchapter) can be specified with the following SMIRKS transform:

[:1][C:2](=[O:3])[*:4]>> [*:1][C:2]([O:3][*])[*:4]*

In some sense, SMIRKS can be considered the opposite of reaction SMARTS. While the latter can be used for querying already defined reactions, the former can be applied to enumerate reactions based on a generic reaction. (A recent example is a paper of Guasch *et al.*, describing the enumeration of ring-chain tautomers.[17]) A complete description of SMIRKS is available on the website of Daylight Inc.[16]

2.1.1.3 MQL

The Molecular Query Language (MQL) was developed in 2007 with the intention to provide a way to define more complex, feature-rich substructure queries.[18] MQL is based on the SMILES language and a context-free grammar, implementing concepts from Unix-style regular expressions such as brackets for grouping, "*(...)?*" for optional singular occurrence, "*(...)**" for optional occurrence zero or more times, *etc.* MQL is open-source and was developed to be compatible with external cheminformatics toolkits, such as the Chemistry Development Kit (CDK).[19,20]

2.1.2 International Chemical Identifier (InChI)

While it is slightly less human-readable than SMILES (especially for complicated structures), InChI is a fully featured, flexible and standardized line notation.[21,22] It was developed with the support of IUPAC (International Union of Pure and Applied Chemistry)[23] with principal contributions from NIST (U.S. National Institute of Standards and Technology)[24] and is maintained by InChI Trust, a non-profit organization established for this purpose.[25] This signifies some of the main advantages of InChI, *i.e.* that it is free, open-source and is maintained by a single organization (meaning that there are no concurrent, parallel implementations).

```
InChI=1S/C9H11NO2/c10-8(9(11)12)6-7-4-2-1-3-5-7/h1-5,8H,6,10H2,
(H,11,12)/t8-/m0/s1
```

Figure 3. InChI string of L-phenylalanine, composed of the following layers: prefix ("InChI=1S/"), empirical formula, skeletal connections ("/c"), hydrogens ("/h") and stereochemistry, composed of three sublayers: tetrahedral centers ("/t"), and two indicator layers ("/m" and "/s"). The InChI string for the zwitterionic form is identical to the neutral form, as they are derived from the same core parent structure and have the same number of protons.

Compared to SMILES, InChI employs a greatly different logic (see Figure 3 for the InChI of L-phenylalanine as an example). An InChI code is made up of several layers separated with forward slashes ("/"), each of which presents specific information on the molecular structure. While some layers are mandatory, other layers (providing more subtle structural features) are optional. Hence, InChI is capable of providing molecular information at different levels of detail. We summarize the most important InChI layers and their syntax in Table 1.

Table 1. InChI layers and sublayers (with L-phenylalanine as an example, where applicable).[22]

Layer	Syntax example	Comment

Prefix	InChI=1/
	InChI=1S/
		Denotes the major version number (currently 1, or 1S for standard InChI)

Formula	C9H11NO2	Represented according to the Hill convention (carbons first, then hydrogens, then other elements in alphabetical order). Canonical numbers are assigned to heavy atoms in the order they appear in the formula.
Skeletal connections	/c10-8(9(11)12)6-7-4-2-1-3-5-7	Connections between skeletal atoms specified with their canonical numbers. Branches are specified in parentheses.
Hydrogens	/h1-5,8H,6,10H2,(H,11,12)	Comma-separated list of the positions of hydrogen atoms in three sublayers: (i) bridging hydrogens (if applicable), (ii) immobile hydrogens, (iii) mobile hydrogens. For L-phenylalanine, no bridging hydrogens are present, heavy atoms 1 through 5 (phenyl ring) and 8 (α carbon) are bonded to one hydrogen atom, while 6 (methylene) and 10 (nitrogen) to two. One hydrogen is shared between atoms 11 and 12 (carboxyl oxygens).
Charge	/q+1	Charge sublayer: net charge of the core parent structure.
	/p+1	Protonation sublayer: net number of protons removed from or added to the core parent structure.
Stereo chemistry	/b4-3+	Z/E configuration of double bonds.

	/t8-	Configuration of stereogenic centers.
	/m0	Indicator sublayer of “/t”: specifies the stereo arrangement relative to the canonicalized core parent structure. (“/m0” for identical, “/m1” for inverse. Only for “/s1”.)
	/s1	Indicator sublayer of “/t”: specifies if absolute (“/s1”) or relative (“/s2”) stereochemistry is provided.
Isotopic	/i12+1	Specifies non-natural isotopes. In the example: atom 12 consists of the isotope with mass increased by unity with respect to the natural value. May have its own stereochemistry sublayer.
FixedH	/f10H3	Lists the exact positions of tautomeric (mobile) hydrogens. May have its own formula, charge and stereochemistry sublayers.

A great advantage of InChI codes is that they are readily canonicalized at the time of generation, thus the relationship between substance and InChI code is mutually unambiguous. (Canonical atom numbers are assigned to elements in the order they appear in the empirical formula, *i.e.* first to carbons, then nitrogens, *etc.*) Further standardization was achieved with the introduction of the Standard InChI, which was designed to enable interoperability between large databases and web resources and “distinguishes between chemical substances at the level of ‘connectivity’, ‘stereochemistry’, and ‘isotopic composition’.[22]

2.1.2.1 InChIKey

A drawback of InChI strings is that they can be really long for big molecules, which makes them unfit for web search queries. To remedy this deficiency, InChIKey, a compact hashed code derived from InChI was developed.[22] InChIKey compresses the information content of an InChI string to a fixed length of 27 characters in the following format:

AAAAAAAAAAAAAA-BBBBBBBBFV-P

The first 14 characters encode the core molecular constitution (formula, connectivity, hydrogens and charge), while the first 8 characters of the second block encode advanced structural features (stereochemistry, isotopic substitution, exact position of mobile hydrogens, metal ligation data). F is a flag for Standard (*S*) or non-standard (*N*) InChI, V denotes the version (currently *A*, meaning version 1), and P is a protonation flag (*N* for neutral and other letters for various protonation states). By definition, the first block is the same for substances sharing the same molecular skeleton. As an example, the InChIKey of L-phenylalanine is: *COLNVLDHVKWLRT-QMMMGPOBSA-N*.

A key consideration in the design of the InChIKey was the compatibility with search engines. Thus, InChIKeys only contain uppercase letters and are restricted to a fixed length. While InChIKey collisions (*i.e.* two structures having the same InChIKey) are theoretically possible, their likeliness has been shown to be practically nonexistent in theory and in practice as well.[26] A disadvantage of InChIKey is that the chemical structure cannot be restored algorithmically. However, publicly available InChI Resolvers provide a lookup service for this task.[27]

2.1.3 Other line notation formats

2.1.3.1 Wiswesser Line Notation (WLN)

Once a widespread tool in chemical database searching[28], the Wiswesser Line Notation is now mostly of historic interest. It was the first computer-processable line notation system, introduced in 1949.[7] It uses alphanumerical characters and a set of special characters (to indicate rings and substitution positions).[29] It makes use of letters not only to encode individual atoms, but also – quite intuitively – to denote common structural features (for example, the letters *X* and *Y* denote branched chains of X- and Y-shapes). This approach has enabled a great extent of compactification (for example the WLN code for phenylalanine would be "*VQYZ1R*" – only six characters!) and WLN was applied for indexing large chemical databases, such as the Chemical Structure Index (CSI) or the Index Chemicus Registry System (ICRS).[30] However, the notation was error-prone and difficult to code and with the advent of connection table formats, it has gradually lost its importance.

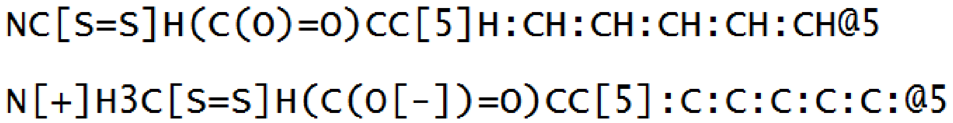
2.1.3.2 Representation of Organic Structures Description Arranged Linearly (ROSDAL)

The ROSDAL notation was developed in 1985 in the Beilstein Institute and has powered the Beilstein DIALOG system (or Beilstein-Online).[30] It is quite intuitive, but not easily readable: each atom (other than hydrogens) is assigned a unique number and the connections between the atoms are listed sequentially, *e.g.* for phenylalanine: "*1O-2=3O,2-4-5N,4-6-7=-12-7*". As seen, heteroatoms (but not carbons) are specified with their chemical symbols, while bonds are denoted with the special characters - (single), = (double), # (triple) and ? (for any connection), and alternating bonds can be simplified, *e.g.* the substring denoting the phenyl ring is "*7=-12*". A great drawback of this notation is that atoms are arbitrarily numbered and there is no unique representation of a molecule. Despite being mostly obsolete, some chemical drawing packages (such as ChemDoodle) still support the ROSDAL notation.[31] A common deficiency of WLN and ROSDAL is the lack of support for important structural features such as ions and stereochemistry.

2.1.3.3 Sybyl Line Notation (SLN)

The Sybyl Line Notation was developed at Tripos (now a subsidiary of Certara) and it was published in 1997.[32] In many senses, SLN is very similar to the SMILES notation, although they employ different concepts regarding the representation of certain features. For example in SLN, aromaticity is a property of the bonds, not the atoms (in contrast to SMILES): aromatic bonds are always explicitly represented with a colon. Each atom with a ring closure is assigned an ID number in square brackets (*e.g.* "*[1]*") and subsequent connections to a previously defined ring closure atom are specified with a commercial at symbol, *e.g.* "*@1*". In addition, square brackets are also used to specify properties of atoms and bonds in a [*property=value*] format, *e.g.* a double dative bond is encoded by the following substring: „*=[type=dative]*". As an example, SLN strings for L-phenylalanine are provided in Figure 4.

In 2008, the functionality of the SLN language was substantially expanded: support for reactions, queries and virtual combinatorial libraries was implemented.[33] While SLN is a fully featured and versatile line notation language, its use is less common nowadays, possibly due to the decreased popularity of the Sybyl modeling suite.



```
NC[S=S]H(C(O)=O)CC[5]H:CH:CH:CH:CH:CH@5

N[+]H3C[S=S]H(C(O[-])=O)CC[5]:C:C:C:C:C:@5
```

Figure 4. SLN strings for the neutral (above) and zwitterionic (below) forms of L-phenylalanine. Hydrogens may or may not be specified explicitly. Formal charges are implemented as atomic properties, as well as absolute configurations, here specified with the "[S=S]" substring.

2.2 Chemical table files

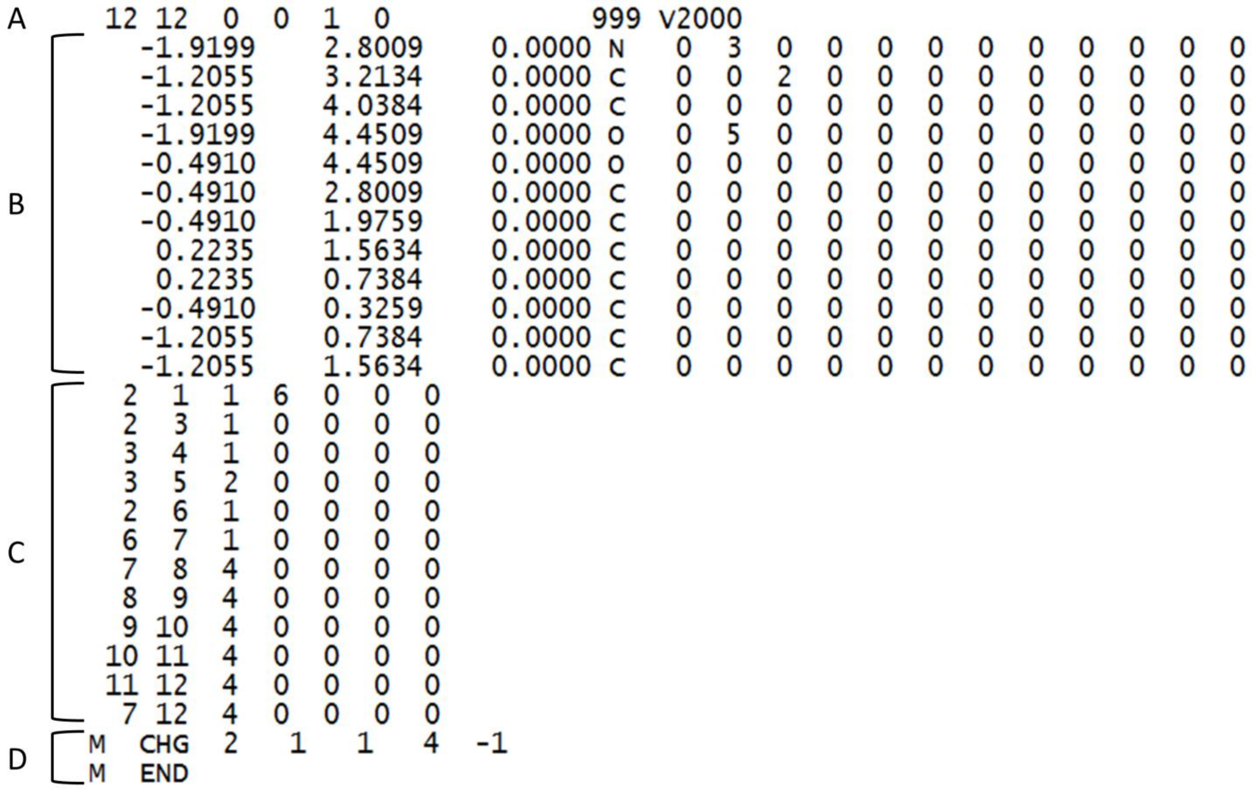
Chemical table files are the standard formats for the 3D representation of small molecules, but they are widely used to store 2D structures as well. Though not as compact as line notations, they are capable of storing specific 3D conformations, partial charges, database fields (such as molecular properties), *etc.* We briefly review the two most prominent tabular formats in this subchapter: the Ctab-based format family ("MDL molfiles") and the Tripos *mol2* format.

2.2.1 Connection table (Ctab) based file types

Several file formats have been developed at Molecular Design Limited that are built around the common concept of connection table (Ctab) blocks.[34] These include most notably *mol* files for storing a single molecule, *sdf* files (or SDfiles) for storing multiple molecules and associated data, as well as *rxn* and *rdf* files (or RDfiles) for storing single and multiple reactions, respectively. While the Ctab format is quite rich in features and optional fields, some of its functionalities are now rarely used or obsolete. Here we will only review its most important features, while we refer to the work of Dalby *et al.* for a more detailed description.[34]

The connection table (Ctab) consists of various blocks in a fixed order: the counts line, the atom block, the bond block, two optional blocks (atom list and Stext) and the properties block. The counts line specifies the total number of atoms and bonds in the molecule (among others), the atom and bond blocks list the coordinates (and other properties) of the atoms and the specification of the bond, while the properties block contains additional information on charge, radicals, isotopes, *etc.* and is always terminated by "*M END*". (See Figure 5 for the Ctab of L-phenylalanine.)

The Ctab itself is preceded by three lines – which can contain the molecule title and information about the program used for the production of the file – and is followed by data fields in *sdf* files, specified in a two-line format, where the first line is the *data header* – starting with ">" and specifying either the field name or the field number in the associated database – and the second line contains the data value. In multiple molecule formats (such as *sdf*), entries are separated with a line containing four dollar signs ("$$$$").



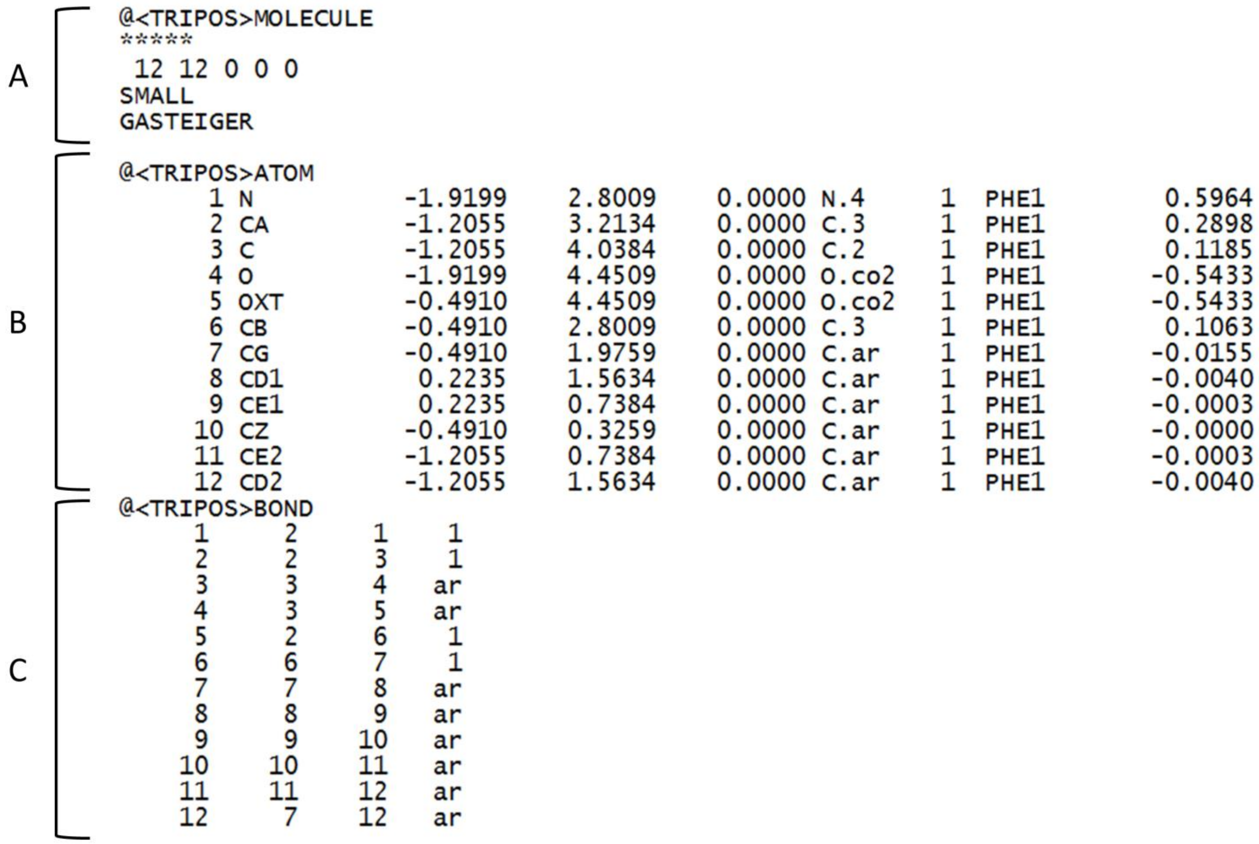
```
 12 12  0  0  1  0            999 V2000
   -1.9199    2.8009    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
   -1.2055    3.2134    0.0000 C   0  0  2  0  0  0  0  0  0  0  0  0
   -1.2055    4.0384    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.9199    4.4509    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
   -0.4910    4.4509    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4910    2.8009    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4910    1.9759    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.2235    1.5634    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.2235    0.7384    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.4910    0.3259    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2055    0.7384    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2055    1.5634    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  6  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  3  5  2  0  0  0  0
  2  6  1  0  0  0  0
  6  7  1  0  0  0  0
  7  8  4  0  0  0  0
  8  9  4  0  0  0  0
  9 10  4  0  0  0  0
 10 11  4  0  0  0  0
 11 12  4  0  0  0  0
  7 12  4  0  0  0  0
M  CHG  2   1   1   4  -1
M  END
```

Figure 5. Ctab block of the zwitterionic form of L-phenylalanine. (Written with Chemaxon Marvin.[35]) A) The counts line contains the numbers of atoms and bonds (both *12*), the chirality flag (*1*, as the molecule is chiral), the Ctab version (*V2000*) and some additional properties. B) The atom block specifies the 3D coordinates, atom symbol and various properties (such as mass difference, charge, hydrogen count) of each atom. In particular, the second column after the atom symbols contains formal charges (where *3* and *5* encode a charge of +1 and -1, respectively) and the third column specifies the atom stereo parity (see the work of Dalby *et al.* for a detailed specification).[34] C) The bond block lists the bonds, specifying the numbers of the first and second atoms, the bond type (*1* for single, *2* for double, *etc.*), the bond stereo specification and a few additional properties. D) The properties block can have a number of "*M XXX*" entries that specify additional properties, and is always terminated by "*M END*". Here, the "*M CHG*" line specifies (from left to right) that there are two atoms with formal charges: atom number *1* with a formal charge of *1* and atom number *4* with a formal charge of *-1*.

While Ctab formats are "the closest thing cheminformatics has to a universally-adopted standard", some argue that they are out-of-date.[36] It is worth to note that the format has been updated several times to resolve known issues. In 1995, the next-generation V3000 format was introduced by MDL (Molecular Design Limited) to overcome known limitations of the V2000 format, such as the hard limit on the number of atoms and bonds (999 in V2000) or the handling of partially defined stereochemistry. (However, the position of V2000 as the *de facto* default format has been more or less unchanged ever since.) In 2011, Clark proposed further changes to enable the handling of zero-order bonds and explicitly specified count of attached hydrogens in a backwards compatible manner, thus extending the scope of Ctab files to store non-organic substances, such as metal complexes.[37] Nonetheless, the superfluous atom specification and the lack of extensibility (among other known issues) constitute a valid base to criticism of the MDL Ctab format and call for an alternative, 21st century *lingua franca* for chemical information management.[38]

2.2.2 Tripos mol2 format

Compared to Ctab files, the *mol2* format (developed at Tripos) offers a more flexible (and customizable) structure for chemical data storage.[39] One of the most important changes is that *mol2* files are "free format" files, meaning that the widths of the various data fields are not fixed. (Also, empty lines are ignored and comments can be specified on lines starting with #.) Additionally, a *mol2* file consists of so-called "records" (similar to the sub-blocks of the Ctab block) that may specify diverse types of information. Most of the records can be omitted, but the *MOLECULE*, *ATOM* and *BOND* (and for some programs, the *SUBSTRUCTURE*) records are universally used. Other important features are the inclusion of partial charges by default and the concept of substructures, which effectively extends the usage of the *mol2* format to the domain of macromolecules (*e.g.* proteins). As a reference, the *mol2* entry of the zwitterionic L-phenylalanine is presented in Figure 6.



```
@<TRIPOS>MOLECULE
*****
 12 12 0 0 0
SMALL
GASTEIGER
@<TRIPOS>ATOM
      1 N          -1.9199    2.8009    0.0000 N.4     1  PHE1        0.5964
      2 CA         -1.2055    3.2134    0.0000 C.3     1  PHE1        0.2898
      3 C          -1.2055    4.0384    0.0000 C.2     1  PHE1        0.1185
      4 O          -1.9199    4.4509    0.0000 O.co2   1  PHE1       -0.5433
      5 OXT        -0.4910    4.4509    0.0000 O.co2   1  PHE1       -0.5433
      6 CB         -0.4910    2.8009    0.0000 C.3     1  PHE1        0.1063
      7 CG         -0.4910    1.9759    0.0000 C.ar    1  PHE1       -0.0155
      8 CD1         0.2235    1.5634    0.0000 C.ar    1  PHE1       -0.0040
      9 CE1         0.2235    0.7384    0.0000 C.ar    1  PHE1       -0.0003
     10 CZ         -0.4910    0.3259    0.0000 C.ar    1  PHE1       -0.0000
     11 CE2        -1.2055    0.7384    0.0000 C.ar    1  PHE1       -0.0003
     12 CD2        -1.2055    1.5634    0.0000 C.ar    1  PHE1       -0.0040
@<TRIPOS>BOND
     1     2     1    1
     2     2     3    1
     3     3     4   ar
     4     3     5   ar
     5     2     6    1
     6     6     7    1
     7     7     8   ar
     8     8     9   ar
     9     9    10   ar
    10    10    11   ar
    11    11    12   ar
    12     7    12   ar
```

Figure 6. Zwitterionic phenylalanine, *mol2* entry. (Written with Open Babel.[1]) A) The *MOLECULE* record by default lists the molecule title (here, ***** denotes that it was not supplied), a line containing the numbers of atoms (12), bonds (12), substructures (0), features (0) and sets (0), the molecule type (*SMALL*) and the charge type (*GASTEIGER*). B) The *ATOM*

record lists the atoms in the following format: atom number, name, 3D coordinates, atom type, substructure ID, substructure name and charge. C) The *BOND* record lists the bonds in the following format: bond number, 1st and 2nd atoms and bond type (with *ar* for aromatic).

Drawbacks of this file format include the lack of stereochemistry support in the absence of 3D coordinates, and the lack of a unified format specification: while *mol2* was originally developed for the Sybyl software suite of Tripos, now each program implements the *mol2* format with slight differences.

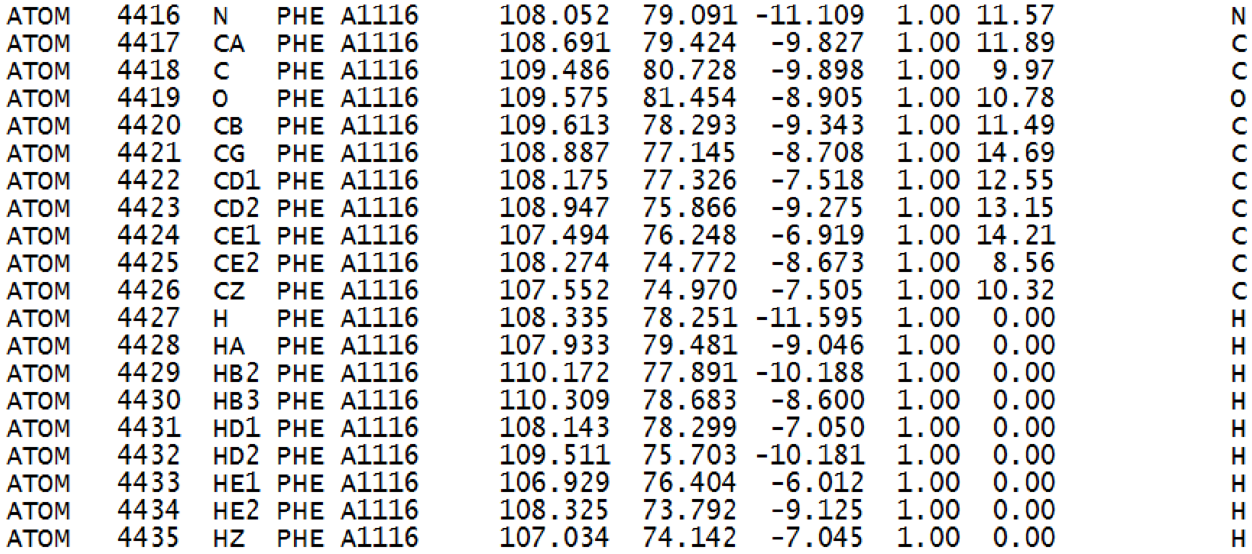
2.3 Other formats

The file formats covered so far are optimized for their usage in chemistry-related applications, mostly covering the domain of small molecules. The treatment of macromolecules (such as proteins and nucleotide chains) requires specific features, which led to the evolution of the specialized file formats of this domain. Of these, we will concisely cover the two that are probably most often encountered by computational medicinal chemists: pdb files and FASTA sequences. In addition, we briefly introduce the reader to novel chemical data management formats and tools, such as the Chemical Markup Language and structure recognition approaches.

2.3.1 Protein Data Bank file format (pdb)

The *pdb* format was introduced in 1992 as the official format specification of the Protein Data Bank (the standard global repository of experimentally solved macromolecular structures).[40] Since a great portion of medicinal chemistry-related modeling work deals with macromolecules, the use of *pdb* files is ubiquitous in this field.

Pdb files are fixed format and they consist of single-line records (starting always with the record type). The diversity of record types allows for the specification of highly detailed information. Atom information is stored in *ATOM* and *HETATM* records (the latter is used for non-polymer entries, such as waters or small molecules), while bond information is stored in *CONECT* records. Atom blocks (corresponding to polymer chains) are terminated with *TER* records. A more thorough overview of the different types of records is provided in the official documentation of the PDB file format.[41] As an example, we provide the *ATOM* section of a phenylalanine residue from an actual PDB entry in Figure 7.



```
ATOM   4416  N   PHE A1116     108.052  79.091 -11.109  1.00 11.57           N
ATOM   4417  CA  PHE A1116     108.691  79.424  -9.827  1.00 11.89           C
ATOM   4418  C   PHE A1116     109.486  80.728  -9.898  1.00  9.97           C
ATOM   4419  O   PHE A1116     109.575  81.454  -8.905  1.00 10.78           O
ATOM   4420  CB  PHE A1116     109.613  78.293  -9.343  1.00 11.49           C
ATOM   4421  CG  PHE A1116     108.887  77.145  -8.708  1.00 14.69           C
ATOM   4422  CD1 PHE A1116     108.175  77.326  -7.518  1.00 12.55           C
ATOM   4423  CD2 PHE A1116     108.947  75.866  -9.275  1.00 13.15           C
ATOM   4424  CE1 PHE A1116     107.494  76.248  -6.919  1.00 14.21           C
ATOM   4425  CE2 PHE A1116     108.274  74.772  -8.673  1.00  8.56           C
ATOM   4426  CZ  PHE A1116     107.552  74.970  -7.505  1.00 10.32           C
ATOM   4427  H   PHE A1116     108.335  78.251 -11.595  1.00  0.00           H
ATOM   4428  HA  PHE A1116     107.933  79.481  -9.046  1.00  0.00           H
ATOM   4429  HB2 PHE A1116     110.172  77.891 -10.188  1.00  0.00           H
ATOM   4430  HB3 PHE A1116     110.309  78.683  -8.600  1.00  0.00           H
ATOM   4431  HD1 PHE A1116     108.143  78.299  -7.050  1.00  0.00           H
ATOM   4432  HD2 PHE A1116     109.511  75.703 -10.181  1.00  0.00           H
ATOM   4433  HE1 PHE A1116     106.929  76.404  -6.012  1.00  0.00           H
ATOM   4434  HE2 PHE A1116     108.325  73.792  -9.125  1.00  0.00           H
ATOM   4435  HZ  PHE A1116     107.034  74.142  -7.045  1.00  0.00           H
```

Figure 7. *ATOM* section of a phenylalanine residue from a PDB entry. The record type (*ATOM*) is followed by the atom number, atom type, residue type, chain identifier, residue number, the 3D coordinates, the occupancy and temperature factor values and element symbol (followed by the formal charge, where applicable). This example highlights a drawback of fixed format files: structures with more than 9999 residues would be problematic to represent, as only four character positions are specified for residue numbers (characters 23-26) and the format does not allow any overflow (character 22 is reserved for the chain identifier).

A great advantage of the *pdb* format is the standard atom typing that was introduced for the common amino acids, nucleotides, cofactors, *etc.* These atom names are based on IUPAC rules[42] and they are also listed in an appendix of the original *pdb* format documentation.[43] Their use eliminates the need to explicitly specify the bonds in the mentioned residue classes, saving a great amount of space in *pdb* files (however, they do have to be specified if non-standard atom names are given).

Another widely used format for crystallographic data is the Crystallographic Information File (*cif*).[44,45] A merit of *cif* (in comparison with *pdb*) is that it is a free format – allowing for more flexibility (and removing any limitations on the number of atoms, residues or chains). On the other hand, it is slightly less human-readable than *pdb*. By default, structures can be downloaded in both formats from the Protein Data Bank (in fact, the standard archive format of PDB is *PDBx/mmCIF* – a customized version of *cif* – since 2014).[46] In addition, an XML-based representation of the *pdb* format, PDBML (Protein Data Bank Markup Language) has been developed and published, and is available in the Protein Data Bank.[47,48] While the XML implementation can provide better interoperability (see subsection 2.3.3), this format currently lacks the compactness of *pdb* files.

2.3.2 FASTA

FASTA is a format that was introduced in a DNA and protein sequence alignment software of the same name and it has been the *de facto* standard format for DNA and protein sequences ever since, owing to its simplicity.[49] It is essentially a sequence of one-letter amino acid (or nucleic acid) codes and special characters (*X* for any residue, - for gap and * for sequence termination) with a single description line (the first line of the file, starting with “>”).

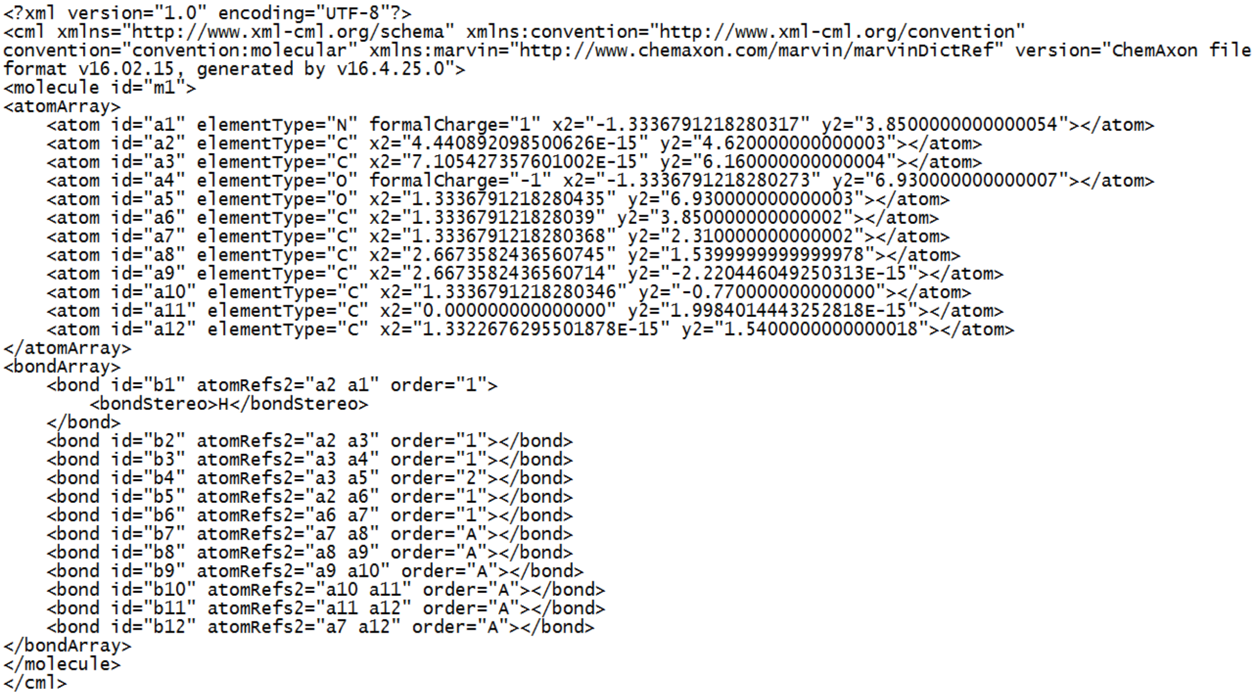
Due to their compactness, *fasta* files (or more generally, one-letter amino acid sequences) are the ideal format for querying macromolecular databases on the basis of sequence identity/similarity. (This type of database querying is often needed in homology modeling and other domains of computational medicinal chemistry.) Such tasks can be decomposed to two (consecutive) questions: (i) how can I find the optimal alignment of two protein sequences, and (ii) how can I quantify the homology/similarity between the two aligned sequences?

While the basics of sequence alignments have been established as early as in the 1970 work of Needleman and Wunsch[50], the BLAST (Basic Local Alignment Search Tool) algorithm (which is considered the standard tool for sequence-based similarity searches) was introduced in 1990 by Altschul *et al.*[51] In contrast to the approach of Needleman and Wunsch, which optimizes the overall alignment of the two sequences, BLAST is a local similarity algorithm, seeking only relatively conserved subsequences. BLAST (along with its numerous specialized versions) powers a web service of the same name, maintained by the National Center for Biotechnology Information (NCBI).[52]

For quantifying the similarity between two aligned sequences, scoring functions based on substitution matrices are usually applied. Substitution matrices contain additive score contributions for each possible exchange of amino acid *A* to amino acid *B*. After summing these score contributions, the alignment with a higher score is better. (The two most commonly used substitution matrix types are Point Accepted Mutation (PAM) matrices[53] and Blocks Substitution (BLOSUM) Matrices.[54]) The introduction of gaps is usually allowed, but penalized in the overall score. The BLAST web server also returns an Expect value (or E-value) for each alignment, which is the number of BLAST hits that are expected to result by chance with the observed score (or higher). While a low E-value alone does not prove that two sequences are homologous, it is a useful guideline to infer some sort of biological relationship.

2.3.3 Chemical Markup Language

The Chemical Markup Language (CML) is an application of XML (eXtensible Markup Language) for the management and integration of chemical data.[55] It was first introduced by Murray-Rust and Rzepa and has been extended numerous times in the following years.[56–63] The main goal of CML is to provide a portable data type that allows for the production of interoperable and reusable documents. In addition to chemical structure, it is capable of storing crystallographic, spectral[62] and reaction[61] data and has been integrated with *e.g.* RSS aggregators[60] and Microsoft Word.[64] It operates using various conventions, allowing for applicability by various subdomains of chemistry.[65] While CML provides interoperability with an explicit specification of properties via markup text, the *cml* format is still fairly compact in comparison with *e.g.* the tabular file formats presented above. (The *cml* entry for L-phenylalanine is provided as an example in Figure 8.) The Chemical Markup Language is entirely open-source and is developed on a voluntary basis. (In fact, it is considered as a project of the Blue Obelisk Movement, an Internet community dedicated to the development of open-source, interoperable cheminformatics software.[66]) Its flexibility, interoperability and open-source implementation make CML (in the authors' opinion) a strong candidate for being the next-generation standard for chemical data storage and exchange.



```
<?xml version="1.0" encoding="UTF-8"?>
<cml xmlns="http://www.xml-cml.org/schema" xmlns:convention="http://www.xml-cml.org/convention"
convention="convention:molecular" xmlns:marvin="http://www.chemaxon.com/marvin/marvinDictRef" version="ChemAxon file
format v16.02.15, generated by v16.4.25.0">
<molecule id="m1">
<atomArray>
    <atom id="a1" elementType="N" formalCharge="1" x2="-1.3336791218280317" y2="3.8500000000000054"></atom>
    <atom id="a2" elementType="C" x2="4.440892098500626E-15" y2="4.620000000000003"></atom>
    <atom id="a3" elementType="C" x2="7.105427357601002E-15" y2="6.160000000000004"></atom>
    <atom id="a4" elementType="O" formalCharge="-1" x2="-1.3336791218280273" y2="6.930000000000007"></atom>
    <atom id="a5" elementType="O" x2="1.3336791218280435" y2="6.930000000000003"></atom>
    <atom id="a6" elementType="C" x2="1.333679121828039" y2="3.850000000000002"></atom>
    <atom id="a7" elementType="C" x2="1.3336791218280368" y2="2.310000000000002"></atom>
    <atom id="a8" elementType="C" x2="2.6673582436560745" y2="1.5399999999999978"></atom>
    <atom id="a9" elementType="C" x2="2.6673582436560714" y2="-2.220446049250313E-15"></atom>
    <atom id="a10" elementType="C" x2="1.3336791218280346" y2="-0.770000000000000"></atom>
    <atom id="a11" elementType="C" x2="0.000000000000000" y2="1.9984014443252818E-15"></atom>
    <atom id="a12" elementType="C" x2="1.3322676295501878E-15" y2="1.5400000000000018"></atom>
</atomArray>
<bondArray>
    <bond id="b1" atomRefs2="a2 a1" order="1">
        <bondStereo>H</bondStereo>
    </bond>
    <bond id="b2" atomRefs2="a2 a3" order="1"></bond>
    <bond id="b3" atomRefs2="a3 a4" order="1"></bond>
    <bond id="b4" atomRefs2="a3 a5" order="2"></bond>
    <bond id="b5" atomRefs2="a2 a6" order="1"></bond>
    <bond id="b6" atomRefs2="a6 a7" order="1"></bond>
    <bond id="b7" atomRefs2="a7 a8" order="A"></bond>
    <bond id="b8" atomRefs2="a8 a9" order="A"></bond>
    <bond id="b9" atomRefs2="a9 a10" order="A"></bond>
    <bond id="b10" atomRefs2="a10 a11" order="A"></bond>
    <bond id="b11" atomRefs2="a11 a12" order="A"></bond>
    <bond id="b12" atomRefs2="a7 a12" order="A"></bond>
</bondArray>
</molecule>
</cml>
```

Figure 8. CML entry for L-phenylalanine. (Written with Chemaxon Marvin.[35]) The <*cml*> element contains the CML specifications (such as the applied schema and convention), the <*molecule*> element summarizes all the sub-elements that belong to the given molecule (titled *m1* in this example). The <*atomArray*> and <*bondArray*> elements consist of <*atom*> and <*bond*> elements, respectively, with the relevant properties listed in a *name="value"* general format.

Although it is independent of CML, we make note of a novel XML-based query language, the Chemical Subgraphs and Reactions Markup Language (CSMRL).[67] CSRML was recently developed (together with the ChemoTyper software and the ToxPrint library) to support chemotypes as a novel approach in substructure querying. (Here, chemotypes are defined as a way of representing chemical entities with three objectives: they should be (i) publicly accessible, (ii) coded in a unique and reproducible manner and (iii) capable of combining both connected and nonconnected chemical patterns as well as atom, bond and molecule-based properties into a single query.) CSRML provides serious advancements over existing query languages in multiple respects, for example it provides a way to produce canonical representations of queries (*i.e.* a query can be specified exactly one way). Also, complex queries can be formulated by defining chemotypes based on molecular properties (in addition to structural patterns). CSRML was developed based on the CML language, extending it with additional features.

2.3.4 Structure recognition tools

Besides the enormous amount of data that is publicly available in standard chemical information formats (such as those detailed above), chemical knowledge of similar (or even greater) multitude is gathered in “traditional” formats, such as text and images in scientific publications and patents. While the conversion of images and text to chemical data is manually cumbersome and time-demanding, recent advances in text mining and optical structure recognition (OSR) are already paving the way towards the seamless, automated processing of textual and pictorial chemical information. These applications provide researchers access to invaluable chemical information that can be utilized for many purposes, *e.g.* for supporting drug discovery and navigating in the intellectual property (IP) space. In this subchapter, we present a brief overview of current sources and services dealing with chemical text mining and image recognition. (Since a detailed review of the related techniques and methodologies would be out of the scope of this chapter, we only refer the reader to recent, well-written reviews in this respect.[68–70])

Currently the most complete set of chemical text mining tools is provided by ChemAxon in their commercial Chemistry Text Mining Suite.[71] The software package includes naming applications (such as Name to Structure), supports text mining from Asian languages (with Chinese Name to Structure and Japanese Name to Structure) and processes whole documents for extracting chemical data (Document to Structure) with the integration of current optical structure recognition software (see below). Most of these functionalities are also provided free of charge on the web server *chemicalize.org*.[72,73]

Current, freely available chemical text mining and named entity recognition (NER) tools include OSCAR (Open-Source Chemistry Analysis Routines)[74], CheNER (Chemical Named Entity Recognizer)[75] and OPSIN (Open Parser for Systematic IUPAC Nomenclature)[76]. Although there are many challenges still unsolved in optical structure recognition – such as extracting chemical data from complex arrangements, such as SAR tables –, this field is also quickly progressing. Several commercial and open source solutions are available as well, including CLiDE[77], Imago[78], OSRA[79] and MLOCSR[80] (the latter two also operate as web services).

Since traditionally the analysis of the patent literature is a cumbersome task, possible applications in this sub-field provide great motivation for the development of text mining and structure recognition tools. In particular, Markush structures (*i.e.* generic representations of a compound class consisting of a core structure and multiple R-groups) are quite extensively used in patent documents.[81] Recent developments describe novel methods for the visualization of Markush structures[82], as well as deconvoluting complex (nested) Markush structures[83], mapping specific structures to Markush structures[84] or encoding and searching Markush structures.[85] We anticipate that coupling these techniques with the text mining and OSR approaches mentioned above will effectively multiply the speed of patent analysis tasks, enabling the extraction of even more information from even more complex data structures than what state-of-the-art programs can currently handle.

3. Molecular fingerprints

Fingerprints are an important and ubiquitous concept in the domain of cheminformatics. Their primary purpose is to provide numerical representations of the structure or certain features of molecules, thus enabling the quantification of the similarity of two molecules. While fingerprints are often represented as bit strings (streams of zeros and ones), in a general sense, any vector of continuous, discrete or categorical (such as 0 and 1) numerical values can be considered a fingerprint. Depending on the fingerprint, various similarity metrics can be used for similarity calculations between molecules. We will cover these options in detail in subchapter 5.

In general, molecular fingerprints are not applicable for chemical data storage, as they are generated with algorithms that *e.g.* check for the presence of a predefined set of substructures or use hashing functions to set the values of certain bits in a bit string – thus converting fingerprints back to structures is not possible for most fingerprint types. This is not necessary however, as there are various options for the compact storage of 2D structures, as well as for more detailed representations including 3D structures and an arbitrary number of molecular properties, including even molecular fingerprints (see subchapter 2).

On the other hand, various molecular fingerprints are used in ligand-based virtual screening approaches. In typical setups, novel bioactive molecules are sought based on their similarities to one or more reference compounds with known activity profiles. The greatest advantages of fingerprint-based virtual screening approaches are computational feasibility and minimal setup and configuration requirements.[86] On the other hand, fingerprint-based methods often fail to identify activity cliffs[87] and their performance depends greatly on the particular fingerprint type.[88] (Although the latter issue can be addressed with data fusion techniques[89], which we cover in more detail in subchapter 5.) Fingerprint similarity searching was compared to other virtual screening methods (such as shape similarity searching and ligand docking) in several works, with varying conclusions, but such comparisons are out of the focus of the present work.[90–92]

This subchapter is dedicated to a thorough overview of current fingerprinting methods. Before moving on to an itemized description however, we highlight some recent, well-written reviews of this field[86,93], as well as some detailed, in-depth analyses of molecular fingerprints, dealing primarily with the similarities and differences among the fingerprinting methods themselves[88], and the effects of various parameters (such as the addressable space, atom typing schemes and bit scaling rules) on fingerprints and their virtual screening performances (in particular on hashed fingerprints implemented in Schrödinger's Canvas).[94,95] In addition, a review by Heikamp and Bajorath summarizes fingerprint engineering strategies, *i.e.* methods for designing fingerprints with an optimized search performance.[96]

Addressing the need for standardized methods for the evaluation and comparison of fingerprints and their screening performances, Riniker and Landrum have assembled and published a standard benchmarking platform for fingerprints[97], based on the open-source RDKit cheminformatics toolkit.[14] Earlier, the same authors have published *similarity maps,* a useful, open-source tool for the visualization of atomic contributions to the similarity between two molecules, which – depending on the fingerprint method – is not always straightforward to see.[98]

While a comprehensive collection of cheminformatics and molecular modeling applications have recently been published by Cereto-Massagué *et al.*[86] (and we also refer to them throughout this chapter), we make note of Cinfony, a Python-based common application programming interface (API) for integrating several cheminformatics toolkits, including Open Babel, RDKit, CDK and others ("the toolkit of cheminformatics toolkits").[99,100] Another, recent addition to the set of publicly available fingerprint-related tools is ChemDes, an online platform capable of generating a rich selection of molecular fingerprints and descriptors, integrating the functionality of multiple popular cheminformatics packages.[101,102] While the former encompasses a larger scope of functionality, the latter offers a user-friendly graphical interface.

3.1 Substructure key-based

In key-based fingerprints, the bits are set according to the presence or absence of predefined substructures (*structural keys*), as shown in Figure 9. The fingerprint length is determined by the number of structural keys and each bit corresponds of a single, specific key. While key-based fingerprints are useful for molecules that are likely to be covered by the structural keys, the treatment of novel or less common substructural features is problematic. Nonetheless, several programs allow for customized keysets, thus the substructural keys can be easily updated, if needed.

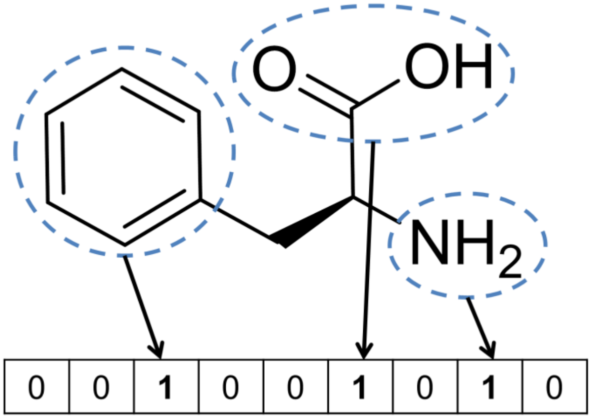


Figure 9. In substructure key-based fingerprints, bits are set according to the substructures that are present in the molecule. (**1** or "on" if the given substructure is present and **0** or "off" if absent.) Thus, each bit position corresponds to a specific substructure.

Key-based fingerprints account for a set of specific structural features, which limits their ability to identify molecules that are similar to the query (on the level of *e.g.* atom environments), but contain diverse fragments, rings, ring systems, *etc.* (In contrast, this is usually considered advantageous for several applications in drug discovery, such as hit expansion or scaffold hopping.) On the other hand, they are the tools of choice if the aim is to retrieve molecules that contain identical substructures with those found in the query molecule. As such, the use of key-based fingerprints with similarity metrics can be thought of as a transition between substructure searching and similarity searching. (In fact, structural keysets were historically used for substructure searching: in this case, molecules that did not have the same bits set as the query molecule were filtered out.[103])

MACCS (for Molecular Access System, a program developed by MDL) can be considered the prototype of key-based fingerprints – or at least it is the best-known. It was developed by MDL (Molecular Design Limited, now a subsidiary of BIOVIA[104]) and has two variants: one contains 166 keys, while the other contains 960 keys. While the former is almost universally implemented in cheminformatics applications (*e.g.* in RDKit[14] or Open Babel[1]), the latter is quite rare (it is available in BIOVIA's Discovery Studio however[105]). The reason for the popularity of the 166-bit version is that while short, it covers most of the interesting features for drug discovery and virtual screening.[86] However, its application in virtual screening is somewhat ambiguous: while MACCS fingerprints show reasonably good retrieval rates in the work of Bender *et al.*[88], they clearly perform weakly in the comparative study of Sastry and colleagues[94]. Nevertheless, Durant and colleagues have published useful guidelines for the reoptimization of MACCS keys, allowing for its optimization for specific sceniarios.[106]

The PubChem fingerprint consists of 881 structural keys that encode quite diverse features.[107] It can be divided into seven sections, each corresponding to a given feature type (*e.g.* atom counts, bonded atom pairs, rings, *etc.*). It is implemented in the PubChem database for similarity searching, as well as some cheminformatics toolkits, such as CDK.[19,20] Further examples of key-based fingerprints include the modifiable BCI (Barnard Chemical Information Ltd., now Digital Chemistry) fingerprints,[108] two fingerprints of Open Babel (termed FP3 and FP4)[1] and two fingerprints implemented in CDK, based on the work of Klekota and Roth (a set of "privileged" substructures for biological activity),[109] and the electrotopological state formalism of Hall and Kier.[110]

A somewhat special substructure-based fingerprint is the set of MQNs (molecular quantum numbers) introduced by Nguyen *et al.*[111] MQNs are a set of 42 integer value descriptors of molecular structure, which count atoms, bonds, polar groups (such as H-bond donors and acceptors) and topological features (such as 5-membered rings or acyclic tetravalent nodes). Consequently, they are not "traditional" key-based fingerprints in the sense that instead of encoding the presence or absence of substructures (either 0 or 1), they encode substructure counts. (On the other hand, feature counts are much more commonly applied in topological fingerprints, as detailed in the next subchapter.) MQNs were primarily introduced as a tool to define, analyze and visualize large chemical spaces (*i.e.* large databases), for which principal component analysis is proposed as a convenient dimension reduction method.[112]

3.2 Topological

Many molecular fingerprinting methods are inspired by more abstract concepts than substructure matching. In general, these methods perceive unique, non-predefined (sub)structural features of molecules. Since there are no predefined substructures, no bit positions are assigned to specific features. Instead, mapping the features to a bit position (or more bit positions) is usually carried out with an appropriate *hashing function* (thus, topological fingerprints are often referred to as *hashed* fingerprints). Since the number of bit positions (the length of the fingerprint or *addressable space*) is finite, it can occur that two features are mapped to the same bit position: this phenomenon is called a *bit collision*. Bit collisions cause a loss of information and have been shown to deteriorate virtual screening performance in the work of Sastry *et al.*[94] (They can be avoided by increasing the addressable space, although that results in a loss of speed as the individual fingerprints get larger.) Due to the possibility of bit collisions, hashed fingerprints would be – by default – inappropriate for substructure searching. However, with a large enough addressable space, the frequency of bit collisions becomes negligible: in this case, substructure searching can be carried out in the same way as mentioned in the previous subchapter.

3.2.1 Path-based

Path-based topological fingerprints operate by enumerating all fragments of a molecule based on linear or branched paths up to a certain number of bonds and hashing them to the addressable space (see Figure 10 as an example). This removes the limitation of predefined substructures, *i.e.* any imaginable molecule produces a meaningful fingerprint.

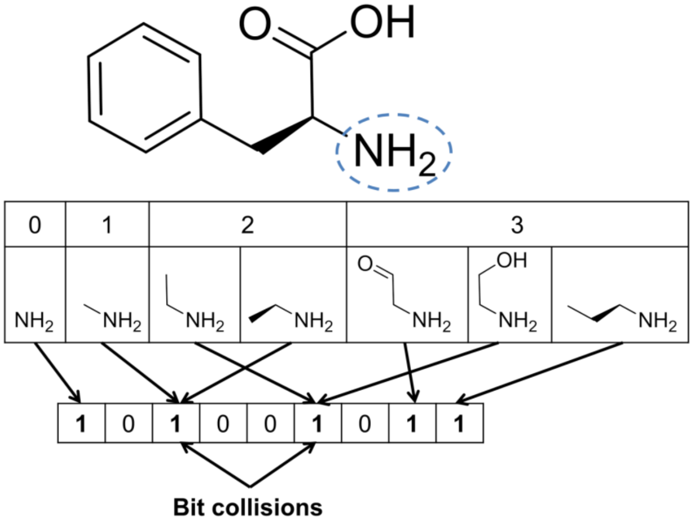


Figure 10. In path-based fingerprints, linear (or in some cases, branched) paths up to a certain length (here three bonds) are enumerated and encoded. (In this example, the paths start from the nitrogen atom of L-phenylalanine, but in practice, the procedure is repeated for each heavy atom.) Note that depending on the applied fingerprint, the two paths (fragments) with 2 bonds of length can be treated as identical or different: this depends on the atom typing scheme used (see main text).

An early example of path-based fingerprints is the Daylight fingerprint, consisting of 2048 bits and encoding all possible connectivity pathways of a molecule up to a predefined length.[113] (Interestingly, the original implementation employed a pseudo-random number generator as the hashing function that produces typically 4 or 5 bit positions per pattern.) The Daylight fingerprint and its variants are still popular: for example, Open Babel has implemented a similar fingerprint (FP2), which consists of 1024 bits and enumerates fragments with 1-7 atoms.[1]

For generating path-based fingerprints, a variety of atom-typing schemes are available. In Pipeline Pilot, functional classes (similar to pharmacophore features), AlogP codes and *mol2* atom types can be used for this purpose (besides element types).[114] For example in Figure 10, the two paths (fragments) with 2 bonds of length would be treated as identical (N-C-C) if element symbols are used for atom typing, but they are differentiated by mol2 atom types (N-CA-C *vs*. N-

CA-CB, see Figure 6 for reference). Schrödinger’s Canvas implements an even wider set of atom typing schemes, ranging from “generic” (where all atoms and bonds are equivalent) to complex ones such as RTXHB (which takes into account ring size, aromaticity, H-bond acceptor/donor type, *etc.*).[115] Both programs implement the possibility to account for feature counts (rather than the mere presence or absence of the features), while Canvas also includes a variety of bit scaling rules to account for the fact that larger fragments usually occur with higher frequencies.[94]

Some path-based fingerprints employ an extended representation of molecular paths (or fragments) by involving branched fragments in the fingerprint generation. Notable examples are OpenEye’s Tree fingerprints[116] and Schrödinger’s Dendritic fingerprints.[115]

3.2.2 Circular

Instead of paths, circular fingerprints encode circular atom environments starting from the central atom and expanding to a given diameter. Circular fingerprints cannot be used for substructure searching (since a particular substructure in a different environment sets a different bit), but are ideal (and popular) choices for similarity searching. Two typical examples for circular fingerprints are Molprint2D and extended connectivity fingerprints.

Molprint2D was introduced in 2004 by Bender and colleagues.[117] This fingerprint encodes atom environments (for each heavy atom) within a distance of two bonds in the following way: the atom types found at distances of one and two bonds are listed and converted to the form *type-freq-distance*, where *freq* is the number of *type* atoms at the given *distance* from the central atom. These entries are sorted by distance and then by type, and hashed to a bit position. A 3D version called Molprint3D was also developed and published in the same year.[118] Molprint2D is available in many software packages, including Open Babel[1] and Canvas[115].

Extended connectivity fingerprints (ECFP) can be considered as the *de facto* standard of circular fingerprints. They were first introduced in 2000 in Pipeline Pilot[114,119], but have since been implemented in various cheminformatics packages, including ChemAxon’s JChem[120] and Schrödinger’s Canvas.[115] A thorough description of extended connectivity fingerprints was published in 2010 by Rogers and Hahn.[121]

ECFPs are based on a modified version of the Morgan algorithm, which was originally introduced as a solution for the molecular isomorphism problem, *i.e.* to determine whether two molecules – with different atom numberings – are the same.[122] The ECFP generation process consists of three consecutive steps: (i) initial assignment, where each atom has an integer identifier assigned to it (based on atomic properties, such as atomic number, formal charge, *etc.*), (ii) iterative updating, where each atom identifier is updated to an array containing the atom identifiers of the central atom and the neighboring atoms (up to *n* bonds of distance, where *n* is the iteration number), which is hashed back into a new, single-integer identifier, and (iii) duplicate removal, where multiple occurrences of the same feature are removed (or counted, if feature counts are requested). The process is illustrated in Figure 11.

ECFPs are highly customizable: in particular the diameter, the addressable space, the consideration of feature presence/absence *vs*. feature counts, and the atomic properties used for generating the atom identifiers are all influential parameters that affect the virtual screening performance either slightly or substantially. In Pipeline Pilot, there are a number of options for atom typing (similarly to path-based fingerprints), such as *mol2* atom types, functional classes or AlogP codes and the diameter is variable as well. To distinguish between the various alternatives, a naming convention was introduced: *xyFz_n*, where *x* is the atom typing (*E* for atom type, *F* for functional class, *A* for AlogP code and *S* for Sybyl *mol2* atom type), *y* is the fingerprint type (*C* for extended connectivity, *P* for path fingerprints), *z* indicates the presence/absence of a feature (*P*) or feature counts (*C*) and *n* is the diameter. (For example, ACFC_4 would be an extended connectivity fingerprint with AlogP-based atom types, a diameter of 4 and feature counts.) Canvas includes some other options for configuration, such as bit scaling and bit filtering rules, while in JChem, the atomic properties used for identifier generation are highly customizable and the whole configuration can be supplied in an *xml* file.

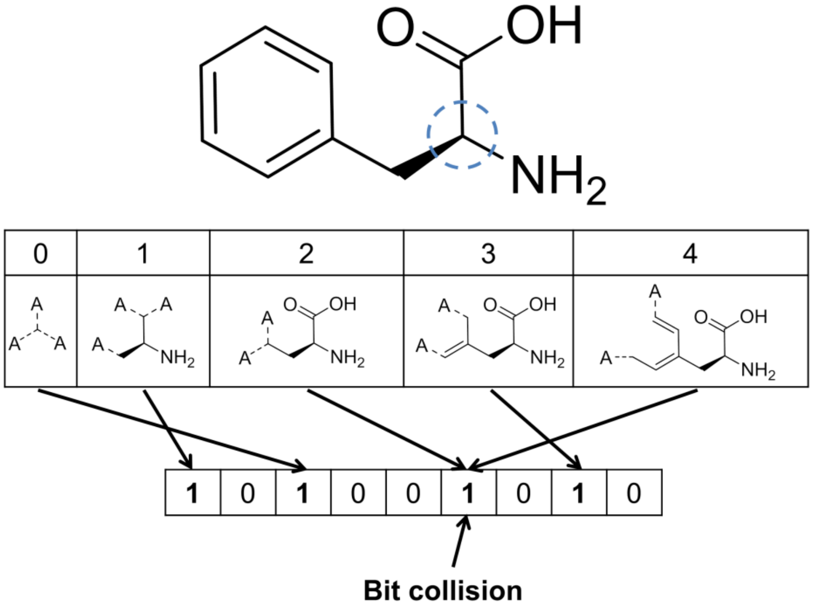


Figure 11. Extended connectivity fingerprints encode full atom environments up to a predefined diameter (here, 4 bonds). For the terminal atoms, any other bonds (that are not part of the fragment) are also accounted for (here, the letter “A” denotes “any atom”). In this example, the central atom is the α-carbon, but in practice, the procedure is repeated for each heavy atom.

3.2.3 Topological subsets

Besides paths and circular atom environments, other topological subsets can also be used for fingerprint generation. Here, we briefly mention two examples whose theoretical basics have been laid down about three decades ago, but are still implemented in current software packages.

Atom pairs have been introduced in 1985 by Carhart *et al.* as a topological descriptor for structure-activity studies.[123] Two years later the same laboratory has published topological torsions for the same purpose.[124] The authors have also developed an atom typing scheme for these descriptors, where an atom is characterized by its element type, the number of its heavy (*i.e.* non-hydrogen) neighbor atoms and the number of its π electrons. (For example, the α-carbon of

L-phenylalanine has three heavy neighbors and zero π electrons, so it could be represented as *C_3_0*.) While alternative atom types (based on physicochemical properties) have been proposed by Kearsley *et al.*, no general superiority is observed for any of these atom typing schemes.[125]

Atom pairs are defined as two atom types and the topological distance (*i.e.* number of bonds in the shortest path) that separates them, and are calculated for each pair of atoms in the molecule to produce the (hashed) atom pair fingerprint. For a molecule with *n* heavy atoms, the total number of atom pairs will be *n*(n-1)/2*. On the other hand, a topological torsion is defined as four, consecutively bonded atom types and represent the topological analogue of the basic conformational element, the torsion angle. Similarly to atom pairs, topological torsions are also enumerated for each quartet of consecutively bonded atoms of the molecule and hashed to a fingerprint.

The two descriptors capture different aspects of molecular topology. For example the original authors note that while atom pairs are generally sensitive to small changes (*e.g.* changing a single atom changes the *n-1* atom pairs involving that atom), topological torsions provide local information and are hence less sensitive to such small local changes. In addition, atom pairs and topological torsions have provided a conceptual basis for the development of further, similar methods, such as geometric atom pairs (the 3D analogue of atom pairs)[126] or fluorine environment fingerprints (a specialized implementation of topological torsions, focusing on fluorine atoms).[127] Meanwhile, the two core methods are implemented in a number of current software packages, including Canvas[115] and RDKit.[14]

APfp, a 2D atom pair fingerprint (not to be confused with the APFP_n fingerprints of Pipeline Pilot) published recently by Awale and Reymond is a simplified version of the atom pair fingerprints detailed above.[128] APfp is a set of 20 integer values, counting the total number of atom pairs in a molecule at distances of 1 to 20 bonds, while Xfp is a related 55-dimensional extended fingerprint that also encodes atomic properties: both have been shown to perceive molecular shape quite effectively. 3D analogues of these methods (based on through-space distances and atom pair Gaussian functions) have been published in 2015 by the same research group.[129]

3.2.4 Pharmacophore

While the category of pharmacophore fingerprints substantially overlaps with the fingerprints that were already covered (*e.g.* topological fingerprints with "functional class" or similar atom typing schemes), we dedicate a separate subchapter for pharmacophore fingerprints as a group of methods that are based on different considerations or design concepts. Pharmacophore fingerprints employ a similar concept to that of atom pair fingerprints (see previous subchapter) in the sense that they provide a simplified representation of key interacting atoms/groups (*i.e.* pharmacophores), which are not necessarily bonded to each other – they can be separated by any number of bonds. On the other hand, in pharmacophore fingerprints, subsets consisting of more than two (typically three or four) features are often encoded: these subsets are called three[130,131] and four-point pharmacophores respectively.[132] In addition to comparing and screening small molecules, pharmacophore fingerprints can also be used for protein binding site similarity calculations.[133]

Among popular software packages, the Molecular Operating Environment (MOE) is particularly rich in pharmacophore fingerprints, with a total of seven choices, encompassing two, three and four-point pharmacophores.[134] The two-point fingerprints TGD (Typed Graph Distances) and TAD (Typed Atom Distances) can be considered an implementation of atom pair and geometric atom pair fingerprints respectively, with a different atom typing scheme (H-bond donor, H-bond acceptor, cation, anion, hydrophobic, polar). Thus, TGD uses a 2D (graph-based) representation of the molecule, while TGT uses a 3D conformation. The three-point fingerprints TGT (Typed Graph Triangles) and TAT (Typed Atom Triangles) are extensions of TGD and TAD, using triplets of atoms instead of pairs. It was shown that the implementation of overlapping atom types, as well as feature counts, could improve the virtual screening performance of TGT fingerprints.[135] GpiDAPH3 and piDAPH3 are also three-point pharmacophore fingerprints, but they employ a different atom typing scheme, in which atoms are differentiated based on three properties (in π system, is H-bond donor, is H-bond acceptor; a total of eight possible combinations). GpiDAPH3 is graph-based, while piDAPH3 is 3D conformation-based. The four-point fingerprint piDAPH4 is an extension of piDAPH3, which considers quartets of features instead of triangles. 2D (configurable) pharmacophore fingerprints are also available in JChem[120] and RDKit[14], while 3D (three and four-point) fingerprints are available in Canvas.[115]

3.3 Hybrid

In a general sense, the term “hybrid fingerprint” encompasses any method that utilizes more than one fingerprinting concepts, but in a wider sense, we can also include in this definition single fingerprints that are optimized *post hoc* based on some known rationale or pre-defined goal.

An early example is the Unity 2D fingerprint of Tripos (now Certara), which is a 988-bit fingerprint that includes both structural keys and path fragments.[136] More recently, Xue and colleagues have introduced the concept of MFPs or “minifingerprints” (short binary fingerprints)[137] and designed various improved fingerprinting methods[138], including MP-MFP, a hybrid fingerprint that contains 61 property-based and 110 structural key-based bits.[139] The more recent PDR-FP (property descriptor range fingerprint) employs solely property-based descriptors, selected systematically to correlate with activity-relevant molecular features.[140]

A common design concept behind these fingerprints is the selection of the most relevant pieces of information that will constitute the final fingerprint, which is often shorter than “traditional” fingerprints, but retains their screening performance (or even improves upon it). In general, there are two (somewhat alternative) principles that underlie the design of such hybrid fingerprints: the recombination of bits (or features) from two or more different fingerprints[141,142] and the weighting (or scaling) of the bits (or features) of a single fingerprint, based on relative importance.[143,144] (The latter is not to be confused with the bit scaling rules implemented in Canvas, which is related to the feature counts of certain fingerprints.[94,115]) Due to their highly customized implementation, hybrid fingerprints can provide serious improvements over individual fingerprints even in tasks that are traditionally considered to be particularly difficult for 2D fingerprints – such as scaffold hopping.[142,145]

3.4 Other fingerprint types

All of the fingerprint methods presented so far are based (one way or another) on the topology or structural features of the molecules. In this subchapter, we present several additional fingerprint types that are based on slightly or radically different ideas. (Nonetheless, these fingerprints are influenced by the molecular topologies as well, but this influence is more or less indirect.)

3.4.1 Text-based

SMILES strings (see subchapter 2.1.1) are very compact representations of molecular structure. Thus, it has been proposed that introducing a means to compare SMILES strings (and calculate their similarities) would speed up virtual screening, since the generation of molecular graphs or connection tables could be omitted. Here, we briefly include two, conceptually diverse SMILES-based fingerprints, LINGO and SMIfp. For a recent, detailed comparison of SMILES-based similarity methods, we refer to the work of Öztürk *et al.*[146]

LINGO is based on the fragmentation of the SMILES string into substrings.[147] In particular, "a *q*-LINGO is a *q*-character string, including letters, numbers and symbols, such as "(", ")", "[", "]", "#", *etc.* obtained by stepwise fragmentation of a canonical SMILES molecular representation". For producing these substrings, LINGO uses canonical SMILES strings. (Although multiple SMILES canonicalization approaches exist – see subchapter 2.1.1 –, the use of alternative canonicalization schemes provides very similar or identical results for property or similarity calculations, as long as the same canonicalization scheme is used consistently for the whole process.) For a SMILES string of length *n*, (*n-q+1*) substrings of length *q* are extracted and their occurrences are counted to provide the final LINGO profile. Such profiles can be thought of as fingerprints and their similarities can be calculated with *e.g.* the Tanimoto coefficient.

In contrast, SMIfp applies a fixed-size chemical space by counting the occurrences of 34 different symbols in the SMILES strings.[148] For SMIfp, the *city block* (or Manhattan) distance is proposed for similarity calculations. This combination has been shown to perform well in recovering several series of active molecules from multiple databases. SMIfp powers a series of SMIfp-browsers for searching popular chemical databases, available at the website of the authors.[149]

3.4.2 EDPrints

The recently developed EDPrints (Electron Density Fingerprints) approach of Kooistra *et al.* is based on a unique idea: chemical shifts (*i.e.* shifts in frequency of NMR spectroscopy signals), as well as partial atomic charges can represent the particular molecular environment of an atom (due to the asymmetric distribution of electrons in chemical bonds).[150] Therefore, both values are useful representations of the chemical and structural properties of molecules, thus they can be utilized in the construction of a fingerprint (as alternatives to the topology of the molecule). EDPrints calculates ^{13}C and ^{1}H chemical shifts with the BatchNMRPredictor program[151], and Merck molecular force field (MMFF94)-type nonpolarized partial atomic charges with Balloon.[152] The calculated values are transformed into non-negative integers and mapped to a bitstring, where each bit position reflects a particular descriptor value. The authors have found that EDPrints is able to achieve similar (and sometimes better) screening accuracies compared to other 2D ligand-based screening methods and exceeds their speed in terms of the time needed to compare a pair of molecules.

3.4.3 Affinity fingerprints

Instead of focusing on the molecular structure, affinity fingerprints encode the experimentally determined binding potencies of a compound against several target proteins. (Hence, an activity fingerprint is by default a vector of real numbers.) The concept was introduced in 1995 in the work of Kauvar *et al*., reporting the design of a small reference panel of eight proteins, which was used to define the affinity fingerprints of 122 structurally diverse compounds.[153] With the application of statistical methods (in particular, multivariate regression techniques), the authors were able to predict binding potencies quite effectively and concluded that affinity fingerprints can be useful tools in a number of applications, including chemical classification, compiling diversified chemical libraries and guiding combinatorial chemistry efforts. Interestingly, a linear combination of the measured binding affinities can provide quite precise predictions for the binding affinities to proteins that are not included in the panel.[153,154] The methodology was further elaborated and compared with a structural fingerprint (MOLSKEYS) by Dixon and Villar.[155] Possibly the best known and most extensively used public reference panel is the cancer cell line panel operated by the National Cancer Institute (NCI).[156] The panel consists of 60 human cancer cell lines, on which 50-percent growth-inhibitory concentrations (GI50) are determined for compounds submitted by research groups from around the world and the results are published in NCI-operated databases.

The concept of affinity fingerprints was later expanded to *in silico* data with the introduction of “virtual affinity fingerprints”, encoding interaction energies derived from docking (instead of experimentally determined affinities).[157] Since docking was not (and is still not) a reliable method for the prediction of ligand affinities, Bender and colleagues have introduced “Bayes Affinity Fingerprints” as an alternative of virtual affinity fingerprints.[158] In this approach, class-specific Bayes models are employed to predict a wide panel of bioactivities and the resulting Bayes scores constitute the Bayes Affinity Fingerprint. Besides reporting consistently higher retrieval rates compared to ECFP_4 fingerprints, another valuable conclusion of the study is that information-optimal “bioactivity spaces” of low dimensionality can be utilized to answer complex questions about bioactivity profiles (*e.g.* “can I modulate off-target activity B independently of on-target activity A?”). The further development of biological fingerprints – and chemogenomics (the study of systematic relationships between targets, based on their ligands) in general – was reviewed by Bender *et al*.[159]

3.4.4 Interaction fingerprints

As the name suggests, instead of chemical features, interaction fingerprints (IFP) encode information about the interactions between the molecule and its environment, usually a protein target. In this manner, protein-ligand complexes resolved with X-ray crystallography (or other experimental methods) or predicted with docking (or other computational methods) can be processed and compared. Consequently, interaction fingerprints can mostly be utilized in docking-based virtual screens or database analysis studies.

Interaction fingerprints consist of a fixed number of bits per residue, corresponding to a predefined set of general or specific interactions between the ligand and the residue. These interactions are checked based on standard cutoffs (in terms of *e.g.* H-bond distance or angle) and the corresponding bits are set to either 1 (there is an interaction) or 0 (no interaction). The process is repeated for every residue of the protein and the resulting substrings are concatenated in the sequential order of the residues to produce the final interaction fingerprint.

There are multiple reported methods for generating interaction fingerprints, the earliest of which to our knowledge is SIFt (Structural Interaction Fingerprints),[160,161] which implements seven different interaction types (including "any contact", "any backbone contact" and "any sidechain contact"). A somewhat modified version of SIFt is implemented in the Schrödinger Small-Molecule Drug Discovery Suite.[162]

Several similar (and similarly named) methodologies were introduced in the following years. The IFP implementation of Marcou and Rognan omits the most general interaction types of SIFt, and enables the configuration of the fingerprinting algorithm, including the addition of weaker and scarcer interaction types such as cation-π interactions or metal complexation.[163] Mpamhanga *et al.* have designed interaction fingerprint based on an alternative concept: their IFPs are fixed-length bit strings that are as long as the number of heavy atoms in the binding site.[164] Pérez-Nueno and colleagues have introduced an atom-pair based interaction fingerprint (APIF) that considers the relative positions of interacting atom pairs.[165] An implementation of interaction fingerprints (termed PLIF, or Protein-Ligand Interaction Fingerprints) is also included in the Molecular Operating Environment (MOE).[134] A more recent approach is SPLIF (Structural Protein-Ligand Interaction Fingerprints), developed by Da and Kireev, in which "three-dimensional structures of interacting ligand and protein fragments are explicitly encoded" (as a result, the interactions themselves are encoded implicitly, in contrast to *e.g.* SIFt).[166]

Trivially, interaction fingerprints can only be compared between the complexes of the same protein structure (as the length of the fingerprint varies according to the total number of residues). Alternatively, with a prior residue selection, such comparisons can be extended to multiple structures, or even multiple, similar proteins (in this case, a prior 3D alignment is needed to match the residues in the different proteins). These methodologies are utilized in recent works of the research group of de Graaf at the VU University of Amsterdam, where interaction fingerprints are employed to power class-specific protein-ligand structural databases such as KLIFS (for kinases)[167–169] or PDEstrian (for phosphodiesterases).[170,171]

3.4.5 FLAP

FLAP (Fingerprints for Ligands And Proteins) was developed in 2007 and is currently distributed by Molecular Discovery.[172,173] FLAP uses a unique approach for fingerprint generation, and employes a wide range of important concepts from cheminformatics/molecular modeling, such as pharmacophores, molecular interaction fields and molecular shape. As such, it is quite distinct from the interaction fingerprints detailed in the previous subchapter, even though it also encodes information about the interactions between the protein and the ligand.

FLAP uses molecular interaction fields (MIFs) based on the GRID force field to identify the active site of the protein. MIFs are then condensed into a few target-based pharmacophoric points (*i.e.* points where specific interactions with the protein would be favorable). To define FLAP fingerprints, one has to select (manually or automatically) a set of such pharmacophoric points. Then, all possible arrangements of four pharmacophoric points are generated and stored in a so-called protein fingerprint. Ligands are fitted to these pharmacophoric points to identify favorable interactions and the resulting poses are filtered according to the complementarity of the shapes of the ligand and the binding site (to account for steric clashes).

3.4.6 Reaction fingerprints

Similarly to line notations, the concept of fingerprints has also been extended to encode reactions (in addition to compounds). Reaction fingerprint generation has been a quite intuitive process since the earliest reported examples of such methods: molecular fingerprints or vectors (consisting of various descriptors) are generated for both the reactants and the products, from which a difference fingerprint (or vector) is calculated. The difference is calculated with a *bitwise OR* operation in the case of Daylight Structural Reaction Fingerprints[113], and with a vector subtraction in the case of descriptor-based reaction vectors.[174] These methods have been successfully applied for a number of tasks, including metabolite prediction[175], *de novo* molecular design[176] and local QSAR analysis.[177] Most recently, Schneider and colleagues have developed a novel reaction fingerprint method based on atom-pair fingerprints of the products and reactants and a physicochemical property based representation of the agents.[178] The method was successfully applied for building machine-learning models, assessing the similarities of reactions and classifying a large set of reactions based on reaction type.

4. Numerical descriptors and similar representations

In the past few decades numerous books with hundreds of pages were published in the field of molecular descriptors. Although the topic is indeed huge, more and more novel molecular descriptors appear from year to year. In this section we would like to provide a general view of the world of molecular descriptors.

A basic definition of molecular descriptors (molecular representations) is that they are measured or computationally calculated properties of molecules. Another definition is stated in the popular and largely useful encyclopedia of Todeschini and Consonni:[179] "*The molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.*" These molecular representations can be used mostly for similarity searching, virtual screening and QSAR (or related) models.

The molecular descriptors can be classified in several ways. There are binary, discreet, and continuous values and we can also calculate other new descriptors from previously defined ones. Usually they are classified based on their dimensionality as 0, 1, 2, 3 … *n* D descriptors. We should mention the obvious confusion that while fingerprints can be calculated from the 2D or

3D representations or properties of the molecules, they are often termed 1D descriptors.[180] Another, truly 1D descriptor is reported in the work of Dixon and Merz, based on the projection of a 2D graph or a 3D conformation to a single dimension.[181] As for higher dimensional descriptors, the term "2D descriptor" means that it is calculated from the 2D (graph) representation of the molecule and in the same manner 3D descriptors are based on the 3D representation of the molecule. In the group of zero dimension descriptors are the molecular weight, the number of different atoms, *etc*. Thus, 0D descriptors are independent of the molecular structure and can be calculated from the molecular formula.

On the other hand, we can differentiate between physicochemical, structural, topological, electronic, geometric and simple indicator parameters (based on their theoretical backgrounds). The term "whole-molecular descriptors" encompasses descriptors that can be calculated from the 2D representation of the molecule. Here we can distinguish between simple descriptors (number of H-bond donors and acceptors, number of ring systems and rotatable bonds *etc*.) and physicochemical descriptors. When using molecular descriptors, one always has to decide between the computational demanding but precise, and the time-saving but less precise methods. Usually 2D descriptors are less time-demanding than 3D descriptors.

4.1 2D descriptors

As it was mentioned above, 2D descriptors are always calculated from the 2D representation of molecules. 2D descriptors are also called graph invariants,[179] as their values do not depend on the numbering of the atoms in the molecule. Here the two main categories: whole-molecular and topological descriptors, as well as other smaller groups will be discussed.

4.1.1 Whole-molecular descriptors

Whole-molecular descriptors can be categorized into simple and the physicochemical descriptors. Simple 2D descriptors are the number of H-bonds acceptors, H-bond donors, the number of rings or other countable parts of a molecule. They can be easily calculated.

Physicochemical descriptors are a larger and more diverse group. There are several parameters and several calculation types. The easiest ones (like molecular weight) are included in 0D descriptors, while we summarize the more complex ones here.

One of the most important physicochemical descriptors is the lipophilicity index, log*P*. Log*P* is part of Lipinski's rule of five to predict drug-likeness. Usually, log*P* means the logarithm of the *n*-octanol (organic)-water (aqueous) partition coefficient, but if there are more lipophilicity indices at hand, it should be denoted as $\log K_{\mathrm{o/w}}$. It plays a role in drug absorption and distribution, thus this descriptor is one of the oldest and most important members of the descriptors applied in drug discovery. The log*P* and hydrophobicity descriptors are frequently used in quantitative structure-activity (and structure-property) relationships (QSAR, QSPR). Lipophilicity can be decomposed into two parts: hydrophobicity and polarity. Hydrophobicity entails the contribution of nonpolar interactions to lipophilicity, but they are not synonyms of each other. The equations relating the partition coefficient (log*P*) to the hydrophobicity constant (π) are the following:

$$\log P \; = \; \log K_{o/w} = \log \frac{[C]_{n-octanol}}{[C]_{water}} \, , \tag{1}$$

where [C] is the concentration of the solute in different phases, and

$$\pi_X = \log P_X - \log P_H \; , \tag{2}$$

where P_{X} and P_H values are connected to the substituted and the unsubstituted molecules' partition coefficients.

Several opportunities are developed for not just measuring, but also for calculating log*P* descriptors. Nowadays the use of *in silico* log*P* descriptors is also acceptable in the field of drug design. The work of Hansch and Fujita has pioneered the calculation of *in silico* log*P* descriptors.[182] Log*P* can be measured by HPLC (high performance liquid chromatography) or TLC (thin layer chromatography) methods manually, or it can be calculated *in silico* in several ways based on for example quantum chemical or half-empirical calculations. Two larger groups of *in silico* calculations are substructure-based and property-based methods. Some calculations – based on the summation of log*P* contributions of molecule fragments – were reviewed by Leo.[183] Typical *in silico* log*P* calculation methods are ALOGP, MLOGP, ClogP, XLOGP2, XLOGP3, *etc*. Nowadays there are several software packages that are freely accessible online. Due to the large number of choices, a careful evaluation of this descriptor is very important in every field of pharmaceutical science. In the past years some detailed comparisons of log*P* calculation methods were published.[184–187] They can help to decide and to answer questions like "what should be used" or "what is the best technique", *etc*. In the latter article the authors show that *in silico* calculations usually give more reliable results than TLC measurements.

Another highly important descriptor is the acidic dissociation constant and its negative logarithm (pK_a). It is connected not just to the *pH* value (and other descriptors like lipophilicity), but to many other ADME related properties, such as membrane and blood-brain barrier permeability. The definition of the acidic dissociation constant is the following:

$$K_a = \frac{[A^-][H^+]}{[HA]} \tag{3}$$

, where *HA* is the acidic form and A^- is the conjugate base.

There are several ways to measure *pKa* values, for example direct potentiometric titration and measurements based on chromatographic retention or UV absorption, *etc*.[188] One can also use *in silico pKa* predictions, which have a serious advantage compared to experimental methods. Namely, the preparation of the molecules is not needed for the analysis. Most of the in silico *pKa* prediction methods and commercial software are based on linear free energy relations (LFER) and the Hammett equation:[189,190]

$$pK_a = pK_a^u + \rho\Sigma\sigma \tag{4}$$

, where pK_a^u is the dissociation constant of the unsubstitued molecule, ρ is the constant for an ionizable group and σ is the constant of an exact substituent.

Similarly to *in silico* lipophilicity indices, comparisons between the *pKa* predication methods and software have also been published.[191]

Last but not least the third crucially important physicochemical descriptor is the (aqueous) solubility (and its negative logarithm). In drug discovery the aqueous solubility of a drug candidate cannot be increased endlessly, because strongly hydrophilic molecules cannot cross lipid membranes. Thus solubility has to be in an optimum range. There are many options for the *in silico* calculation of solubility, some of which are based on other physicochemical descriptors, such as the melting point or lipophilicity. The well-known Yalkowsky equation is a good example for this:[192]

$$\log(S_0) = -0.01(MP - 25°C) - logP + 0.50 \tag{5}$$

, where S_0 is the water solubility, *MP* is the melting point and *logP* is the lipophilicity index.

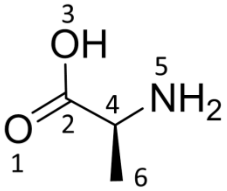
This equation has larger bias, due to the use of the melting point, which is hard to predict based on the molecular structure. The other opportunity for *in silico* solubility calculation is the use of different QSPR models. In this case we can use several chemometric tools for the prediction, such as partial least square regression (PLSR)[193], multiple linear regression (MLR)[194], random forest (RF)[195] and artificial neural network (ANN)[196]. A comparison of the general solubility equation (GSE) and the above mentioned chemometric tools has also been published.[197]

One can find further physicochemical (molecular) descriptors, which are not discussed here in Todeschini's encyclopedia of descriptors.[179]

4.1.2 Topological descriptors

Topological descriptors (or topological indices) are numerical descriptors that can be calculated from the molecular structure. Topological descriptors are one of the most populated (and popular) members of the molecular descriptor (and 2D descriptor) family. Their importance is emphasized, as they can provide simple and useful information about the structures of the molecules. However, they do not contain any information about the stereochemistry and the 3D conformation. (With the words of Hugo Kubinyi: "The topology and topography of a molecule may be compared to the skeleton of a human being. How boring would life be, if we recognize each other only by our skeletons and not by shape, behavior and spirit!")[198] Thus, topological descriptors can be calculated without the optimization of molecular structure. With the combination of other molecular descriptors (such as physicochemical descriptors), they can be quite useful in QSAR and QSPR studies.

These descriptors are based on the 2D graph of a molecule. The calculation of the topological descriptors is closely connected to graph theory, where the molecular structures are built up from vertices (which symbolize the atoms) and edges (which denote covalent bonds). There are two types of these molecular graphs (or maps), the H-filled (contains the hydrogen atoms) and H-depleted (without hydrogen atoms). From these molecular graphs we can also create a graph theoretical matrix, where the relations between edges or vertices can be examined (defined). From another point of view we can distinguish between the adjacency matrices and the distance matrices (based on the original molecular graph). The former matrix contains the non-hydrogen adjacent atoms (for the H-depleted version) and assigns each pair of them as one (if they are adjacent) or zero (if not), while the latter one counts the number of edges between two vertices in the shortest path (topological distance). See Figure 12 for an example of a simple molecule’s adjacency and distance matrices. Topological indices can be calculated from different molecular maps in an unequivocal way as they are based on the topological distances between the atoms of the molecule (through-bond indices). The topological indices can be topostructural (using only the atomic distances) or topochemical (using also atomic properties, like chemical identity or hybridization states).[199] The molecular graphs are also used for the creation of molecular signatures, which are also an appropriate way to calculate several topological indices.



A	*1*	*2*	*3*	*4*	*5*	*6*
1	-	**1**	0	0	0	0
2	**1**	-	**1**	**1**	0	0
3	0	**1**	-	0	0	0
4	0	**1**	0	-	**1**	**1**
5	0	0	0	**1**	-	0
6	0	0	0	**1**	0	-

D	*1*	*2*	*3*	*4*	*5*	*6*
1	0	1	2	2	3	3
2	1	0	1	1	2	2
3	2	1	0	2	3	3
4	2	1	2	0	1	1
5	3	2	3	1	0	2
6	3	2	3	1	2	0

Figure 12. Adjacency (A) and distance (D) matrices of L-alanine. In the adjacency matrix, an entry $a_{i,j}$ is **1** if atoms *i* and *j* are connected with a bond and 0 otherwise, while an entry $d_{i,j}$ in the distance matrix is the topological distance of atoms *i* and *j*. Both matrices are symmetric (*i.e.* their information content can be represented by either their upper or their lower triangular alone).

In the following part the most popular and “good to know” examples from the huge pool of topological indices will be discussed in detail. Nowadays, it is a very hard task to select those, which are highly recommended to know from the enormous amount of literature in this field. In the work of Randic and Basak the question has already been posed: do we need more molecular descriptors?[200] (Naturally we do, because we want to have as diverse of a toolbox as possible and to cover the various possible descriptions of molecules as fully as possible. It is worth noting that this is true for other fields of science as well, in fact it is the essential part of making progress in science.) Although there are some novel topological descriptors in the literature, the most of the widespread ones were developed in the second part of the last century.

Wiener index: The most well-known first generation topological index. First generation means that the integers are based on such graph properties like topological distances.[201] It is calculated by the half sum of all distance matrix entries. It was defined originally in 1947 by Wiener,[202] but there are several modified versions of it in the literature, for example Hyper-Wiener or All-path Wiener indices, *etc*.[179] Basically we can conclude that in the past century the Wiener index has not just become popular, but it was used for the development of several other descriptors. The basic equation is the following:

$$W = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij}\ , \tag{6}$$

where N is the number of atoms and δ_{ij} is the distance matrix entry (shortest possible way between atoms *i* and *j*).

Zagreb group index: Another first generation topological index, which is based on the vertex degrees of the atoms (denoted with δ and representing the number of σ bonds of the atom) in the molecule graph. The two types of this index were published in 1975.[203]

$$M_1 = \sum_a \delta_a^2 \text{ and } M_2 = \sum_b \left(\delta_i \cdot \delta_j\right)_b\ , \tag{7}$$

where *a* runs over the *A* atoms of the molecule and *b* over all of the *B* bonds of the molecule. δ_i, δ_j are the vertex degrees of the atoms incident to the considered bond.

The 1st Zagreb index (M_1) is nowadays also called *Gutman index* and the 2nd Zagreb index (M_2) is also a part of the *Schultz molecular topological index*.

Balaban J index: It was called "averaged distance sum connectivity" in the original article of Balaban.[199] The formula is connected to the Randic index (see below), but here the vertex degrees are changed to the averaged distance sums, thus the equations was modified.

$$J = \frac{1}{\mu+1}\sum_b\left(\bar{\sigma}_i\bar{\sigma}_j\right)_b^{-0.5} = \frac{B}{\mu+1}\sum_b\left(\sigma_i\sigma_j\right)_b^{-0.5}, \tag{8}$$

where *B* is the number of edges (bonds), μ is the cyclomatic number (the number of bonds that needs to be deleted to break all rings – which in the majority of cases is equal to the number of rings), σ_i and σ_j are the distance sums of the vertices *i* and *j* (*i.e.* the sum of the distances in the *i*th and *j*th row of the molecule's distance matrix) and $\bar{\sigma}_i = \sigma_i/B$ is the averaged distance sum of i.

Kappa shape index: these are second generation topological indices, based on the work of Kier.[204,205] Second generation indices are real numbers based on integer graph properties. This group of indices has current applications in drug discovery.[201]

The first, second and third ordered shape attributes can be calculated in the following ways:

$$^1\kappa = n(n-1)^2/(^1P_i)^2 \tag{9}$$

$$^2\kappa = (n{-}1)(n-2)^2/(^2P_i)^2 \tag{10}$$

$$^3\kappa = (n{-}1)(n-3)^2/(^3P_i)^2 \text{ , when } n \text{ is odd} \tag{11}$$

$$^{3}\kappa = (n{-}3)(n-2)^2/(^{3}P_i)^2 \text{ , when } n \text{ is even} \quad (12)$$

Where $^{l}P_i$ represents the number of paths of length l in molecule i, and n is the number of atoms. The maximum and minimum values of $^{l}P_i$ are:

$$^{1}P_{max} = \frac{n(n-1)}{2} \quad (13)$$

$$^{1}P_{min} = (n-1) \quad (14)$$

The kappa shape index was modified with a specific correction (α value), which can be calculated as:

$$\alpha_X = \left(\frac{r_X}{r_{Csp^3}}\right) - 1 \text{ ,} \quad (15)$$

where α is the decrement or increment of n for a non-carbon element (X), r represents the covalent radius of atom X relative to sp^3 carbon atom. The modified kappa shape indices can be found in the work of Kier and Hall[206] or the article of Estrada *et al*.[201]

Randic branching index: also called connectivity index, it was the first connectivity index, developed by Randic[207] more than forty years ago. He introduced it at first for alkanes, with the intention to produce a graph invariant topological index. The equation is the following:

$$\chi = \sum\nolimits_{all\ edges} (\delta_i \delta_j)^{-0.5}, \quad (16)$$

where δ_i and δ_j are the number of non-hydrogen atoms bonded to atoms i and j and forming an edge (bond) ij.

In the past decades several modifications of connectivity indices were created, such as *mean Randic index*, *modified Randic index* and *edge connectivity index (ε)*.[208] Regarding these, the reader is referred to the encyclopedia of Todeschini.[179]

Other important topological indices are the E-state index[209,210] or the valence molecular connectivity index.[211,212]

There is a third generation of topological indices as well, which are real numbers based on real number local vertex invariants (LOVIs). Some third generation types are based on information theory, which is applied to non-symmetrical matrices or terms of distance sums. Balaban was the pioneer of this topic in the nineties.[213–215]

4.1.3 Other 2D descriptors

4.1.3.1 Indicator variables

Indicator variables are one of the simplest or easiest way to describe constituents of a molecule, thus they have been applied in QSAR and related analyses for many decades. They can take positive, negative or zero values based on the state of a property (characteristic), for example it can indicate the *cis*, *trans* or no *cis*/*trans* isomer forms in the following way: –1 represents the *cis* isomers, +1 represents the trans isomers and 0 means the characteristic is not applicable, *i.e.* no *cis*/*trans* isomers. Some well-known examples for indicator variables are the following: Free-Wilson descriptors (*de novo approach*),[216] Fujita-Ban analysis (modified version of Free-Wilson analysis),[217] Hansch or Hansch-Free-Wilson models.[218] In the latter cases the models or descriptors contain not just indicator variables but also their combinations with other parameters such as physicochemical descriptors.

We discuss indicator variables here, including the large group of substructure based descriptors. Indicator variables are also called dummy or Boolean descriptors, if they define the presence or lack of structural elements (1 or 0). They can indicate the presence of hydroxyl, carbonyl *etc*. groups or even the different positions of substituents like *ortho*, *meta* and *para*.[219] (Note that they are closely related to substructure key-based fingerprints, see subchapter 3.1.) A typical example for Boolean descriptors can be found in the work of Devillers in which the descriptors indicate biodegradability.[220]

In a related example, the recent work of Borysov *et al.* introduces the concept of extreme descriptors as "those variables that have the same value for almost all compounds and only a few values that are different from the common median". Although it is a common practice in QSAR modeling to exclude such descriptors from model building, the authors have shown that they can be useful for activity prediction in a standard binary classification setting, and for the identification of mislabeled compounds.

4.1.3.2 Thermodynamic descriptors

Thermodynamic descriptors are a smaller segment of 2D descriptors, with some of them using 3D parameters of the molecules as well (at least optionally). A typical example is the *heat of formation* (ΔH_f) which can be calculated based on the method of Singh *et al*.[221]

4.1.3.3 Molecular identification numbers

In his 1989 work, Burden has introduced the concept of "molecular identification numbers", calculated from the matrix representation of the connection table of a molecule (without hydrogens).[222] It is useful for substructure searching and diversity/similarity searching tasks. This approach was refined and further developed, first by Rusinko and Lipkus, and later by Pearlman and Smith to produce what is now known as the BCUT approach.[223] BCUT descriptors are calculated as the first few highest and lowest eigenvalues of such matrices, with several options for the property that defines the diagonal elements (including *e.g.* partial charges, polarizabilities and hydrogen-bonding properties).[224]

4.2 3D descriptors

3D descriptors are calculated from the 3D representation of the molecule. Thus, 3D chemical structures (conformations) are needed, which in best case can be bioactive conformations of the molecules in the field of drug design. 3D descriptors are an as useful, large and well-known group as their 2D siblings, but they were typically developed later than 2D descriptors and their calculation is more time-consuming and computationally demanding. 3D descriptors (like 2D ones) are frequently applied for QSAR and similarity searching. Some of the descriptors are sensitive to the position and orientation of the molecular structure, but some of them are insensitive for these properties (in other words, they are *alignment-independent* or *alignment-free*). The alignment of molecules is often a time-consuming step of the procedure and not necessarily unambiguous. 3D descriptors can be classified in several ways for example based on the work of Kunal *et al.*[180] or Todeschini and Consonni.[179] Some of the typical and frequently used descriptor classes in the literature are the electronic, size, volume, shape, molecular shape analysis (MSA), molecular field analysis (MFA) and receptors surface analysis (RSA) descriptors. The family of 3D descriptors is too large to cover every aspect of it, thus in the following section the most popular ones will be discussed.

4.2.1 Electronic descriptors

Electronic descriptors can be defined in the group of 2D descriptors as well (see the Hammett equation and related descriptors). They can be local or global descriptors, *i.e.* they can define the electronic properties either of specific regions of the molecule or the whole molecule.[225] Electronic charges and electron densities play an important role in different chemical reactions and physicochemical parameters. Charged partial surface area descriptors are also built up from electronic descriptors with the combination of shape and steric descriptors. One important class of electronic descriptors is *charge descriptors*, which contains for example the *total absolute atomic charge (electronic charge index) descriptor*. Some typical examples for electronic descriptors are detailed below.

(Electric) Dipole moment: This is a quantum-chemical electronic descriptor (or also classified as electric polarization descriptor), based on the strength and orientation of the interaction of the examined molecule with an external electrostatic field. The magnitude of the dipole moment and its components μ_x, μ_y and μ_z are usually calculated. The SI unit of dipole moment is *coulomb meter (Cm),* but for practical reasons, *Debye* ($\approx 3.34\times10^{-30}$ Cm) is used more frequently. The components of μ are calculated in the following way:

$$\mu_x = \sum_i q_i \cdot x_i, \qquad (17)$$

$$\mu_y = \sum_i q_i \cdot y_i, \qquad (18)$$

$$\mu_z = \sum_i q_i \cdot z_i, \qquad (19)$$

where *x*,*y* and *z* are the coordinates and *q* is the charge of atom *i*.

Highest occupied (HOMO) and lowest unoccupied (LUMO) molecular orbital energies: They are also quantum-chemical descriptors. ε_{HOMO} is the energy of the highest energy orbital that contains electrons. A molecule with high HOMO energy values can donate electrons easily. The ionization potential (IP) is closely connected to this: IP = -ε_{HOMO}. Sometimes the energies of the second or third highest occupied orbitals are also used as descriptors. These descriptors are closely related to the nucleophilicity of a molecule.

ε_{LUMO} is the energy of the lowest energy molecular orbital that does not contain any electrons. A molecule with low LUMO energy values can accept electrons easily, thus this is connected to the definition of electronic affinity (EA): EA = -ε_{LUMO}. In other words, it measures the electrophilicity of a molecule.

4.2.2 Size descriptors

In addition to 3D descriptors, size descriptors can also be simple ones like molecular weight or the number of bonds and atoms. We can also find size descriptors in the group of topological indices, but several 3D size descriptors exist as well. Size, volume and shape descriptors are often connected to each other, for example steric descriptors account for the size and the shape of the molecule. On the other hand there are some volume descriptors, such as van der Waals volume, which can also be considered a size descriptor (see later). A popular 3D example is detailed below.

Radius of gyration: This descriptor characterizes the distribution of atomic masses in a molecule,[226] and provides an absolute measure of molecular compactness.[227] It is defined as:

$$R_G = \sqrt{\frac{\sum_{i=1}^{A} m_i \cdot r_i^2}{MW}}, \qquad (20)$$

where r_i is the distance of i-th atom from the center of mass of the molecule, m_i means the atomic mass, A is the number of atoms and MW represents the molecular weight.

4.2.3 Volume descriptors

Volume descriptors are either steric or size descriptors (or both). There are both experimental and theoretical ways for determining the volume of the molecules. Some examples of volume descriptors are listed here (van der Waals volume is discussed later in detail).

McGowan's characteristic volume (V_x): As a steric descriptor it is calculated as the sum of the atomic volume parameters in the molecule. It can be used for example for cavity term measurements. The equation is the following:

$$V_x = \sum_{i=1}^{A} w_i - 6.56 \cdot B, \qquad (21)$$

where w_i means the McGowan volume parameter and B is the number of bonds. McGowan atomic parameters for the calculations can be found in his article from 1987.[228]

Molecular volume index (MVI): This volume descriptor is based on the summation of van der Waals volumes of each groups of the molecule.[229] With the double summation, each pair of non-hydrogen atoms is involved. The equation is the following:

$$MVI = \sum_{i=1}^{A} \sum_{j=i+1}^{A} \frac{V_i^{vdw} \cdot V_j^{vdw}}{D_{ij}^2}\,, \qquad (22)$$

where V^{vdw} represents the van der Waals volume of each group and D_{ij}^2 is the square of the topological distance of v_{i} and v_{j}.

Geometric volume: The atoms can be defined as point masses, which construct the solid geometric shape of the molecule. Thus the geometric volume is the volume of this solid geometric shape. In the concept of this volume, the atoms are interconnected and form some regular and irregular tetrahedrons. The volumes of these tetrahedrons are computable in analytical and numerical ways as well.[230]

4.2.4 Shape descriptors

Molecular shape descriptors are very useful in the modeling of physicochemical processes. This topic has a large scientific literature and in the past 30-40 years several novel shape descriptors were developed.[227,231] Some of the most popular ones are *shadow indices*, *WHIM shape descriptors*[232] or *molecular shape analysis* descriptors (see below in details).

Shadow indices (Jurs shape indices): geometrical descriptors related to the size and shape of the molecule. The basic principle is the projection of the molecular surface onto three mutually perpendicular planes XY, XZ and YZ.[225] The descriptors encode the conformation and also the orientation of the molecule. Rotational invariance is obtained by the previous alignment of the X,Y, and Z axes along the three axes of principal inertia.[233,234]

4.2.5 van de Waals parameters

Another interesting part is the *van der Waals molecular surface (or area)* (SA^{vdw}), which can be included in the aforementioned descriptor groups as well. It is also called *total molecular surface area (TSA)* and it is connected to the hard-sphere model. The total van der Waals surface is calculated as the sum of the atomic van der Waals surfaces. This is related to binding, solubility *etc*.

The *van der Waals volume (V^{vdw})* is the volume of the space inside the van der Waals molecular surface. The van der Waals volume is closely connected to the *van der Waals radius* (R^{vdw}). The van der Waals radius is the distance, where the attractive and repulsive forces between the two nonbonded atoms are equal. Its calculation was originally described by Bondi,[235] but less time-consuming calculations have been developed as well by Zhao *et al.*[236] There is a vast literature in this field between the 80's and 90's, for instance molecular mechanics based methods,[237] as well as other techniques.[238]

4.2.6 Molecular Shape Analysis descriptors

Molecular shape analysis (MSA) is a common technique in QSAR analysis, which combines the molecular shape similarity and commonality measures (with other descriptors) to determine the similarities between molecules and build appropriate QSAR models.[239] In this section we do not cover MSA QSAR in detail, only the commonly used molecular shape similarity descriptors (MSA descriptors). Molecular shape similarity is applied to the comparison of 3D molecular shapes, which are represented by atomic properties. On the other hand, molecular shape commonality is using conformational energy and molecular shape together to measure molecular similarity. The basic concept of MSA in QSAR analysis is that the shape of the molecule is related to the binding site cavity (or pocket), thus it is related to biological activity as well. Some popular molecular shape similarity descriptors are listed below:

Common Overlap Steric Volume (COSV): It represents the overlapping volume of two superimposed molecules.[240] Van der Waals volume of the molecules is used for the determination.

$$M_0(i,j) \equiv V_0(i,j) = V_i \cap V_j \tag{23}$$

Non (Common) Overlap Steric Volume (NCOSV or V_{non}): It is defined as the volume of molecule i, which does not overlap with the volume of the reference molecule j. (Thus, it can be related to steric misfit.) Basically V_{non} is the difference between the $V_{\mathrm{i,j}}$ composite steric volume of two aligned molecules and V_{j} (reference molecule).[240]

4.2.7 Molecular Field Analysis descriptors

Molecular field analysis (MFA) is also a frequently used technique in QSAR. It is a grid-based QSAR technique, mostly applied in Comparative Molecular Field Analysis (CoMFA) based on the work of Cramer *et al.*[241] (Here, we only briefly mention CoMFA, as a detailed introduction would be out of the scope of this chapter.) The molecular field can be represented by a 3D rectangular grid. MFA analysis is based on the calculation of interaction energies (steric and electrostatic interactions in the case of CoMFA) between some probes (H^+ or CH_3) and the molecule, represented by a rectangular grid. Thus the field of molecules can be described by MFA grids, and the energies associated with MFA grid points may serve as inputs (descriptors) for the calculation of QSAR models.

4.2.8 Receptor Surface Analysis descriptors

Receptor surface analysis (RSA) is another modeling method, primarily for such cases when the 3D structure of a receptor is unknown.[242] A hypothetical model can be created with this predictive tool for the receptor site. On the other hand, RSA descriptors are also useful in QSAR model building when the receptor surface is known. The RSA approach clearly differs from pharmacophore approaches, as it captures information about the receptor instead of the ligands. From receptor surface models, one can derive descriptors, which provide 3D information about the (steric or electrostatic) interaction energies between each point of the receptor surface and the ligand. RSA descriptors can be combined with other 3D or 2D descriptors for QSAR analysis. Some typical examples for RSA descriptors are *IntraEnergy, InterEnergy, StrainEnergy, MinIntraEnergy, etc.*[180]

4.2.9 3D descriptor families

4.2.9.1 WHIM descriptors

Weighted holistic invariant molecular (WHIM) descriptors are based on statistical indices calculated on the projections of atoms around the *x*, *y* and *z* axes.[232] The algorithm involves a principal component analysis (PCA) of the centered molecular matrix (coordinates of the molecule). Depending on the six different weighting schemes (such as atomic mass, van der Waals volume, *etc*.), different covariance matrices (and principal components) can be built. Directional WHIM descriptors can also be calculated. Typical groups of them are size descriptors (constructed from the eigenvalues of the covariance matrix), shape descriptors, symmetry descriptors, density descriptors, *etc*. In summary, there are 66 directional WHIM descriptors (11 for each weighting scheme). Global WHIM descriptors are calculated as combinations of directional WHIM descriptors.

4.2.9.2 GETAWAY descriptors

GETAWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors are based on the molecular influence matrix (H), which is calculated in the following way:[243]

$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^{\mathbf{T}} \times \mathbf{M})^{-1} \times \mathbf{M}^{\mathbf{T}} \qquad (24)$$

, where M represents the molecular matrix (containing the centered Cartesian coordinates of the atoms of the molecule).

The diagonal elements of the **H** matrix are called leverages. The combination of the molecular influence matrix with geometry matrix (**G**) is called the influence/distance matrix (**R**). It is very important, because a set of GETAWAY descriptors are calculated from both of them based on different concepts and matrix operators. The GETAWAY approach also employs different atom weighting schemes, similarly to WHIM. Some examples are: geometric mean of the leverage magnitude ($H_{GM)}$, total information content on the leverage equality (I_{TH}), standardized information content on the leverage equality (I_{SH}), mean information content on the leverage magnitude (HIC), average row sum of the influence/distance matrix (RARS), *etc*.

There is another group of GETAWAY descriptors (denoted as R-GETAWAY), which combines the information mentioned above with geometric interatomic distances in the molecule.

4.2.9.3 EVA descriptors

EVA or EigenVAlue descriptors are based on the extraction of chemical structures from mid- and near-infrared spectra. The basic concept was developed in 1997 by Ferguson *et al.*[244] The fundamental molecular properties can be characterized potential energy functions (vibrations). The normal coordinate eigenvalues (and eigenvectors) from vibrational frequencies or atomic displacements can be calculated with the use of quantum and molecular mechanical methods.

EVA descriptors are 3D descriptors providing information about molecular size, shape and electronic properties. A modification of the original conception is EEVA (Electronic EigenVAlues), which defines a set of vectorial descriptors.[245,246] Here the eigenvalues of the Schrödinger equation are used instead of the vibrational frequencies.

4.2.9.4 MEDV (MEDV-13) descriptors

The Molecular Electronegativity Distance Vector (MEDV) is a vectorial descriptor, which includes 91 terms that contain information about the relative electronegativities.[247] The electronegativities are represented by the modified E-state indices and topological distances between each possible pair of 13 atom types. In the first step of the MEDV descriptor calculation, one has to assign each atom in the molecule to one of the aforementioned atom types. The atom types are based on the most frequently occurring atoms in organic molecules and also the number of bonded non-hydrogen atoms (vertex degree).[225] The single molecular descriptor can be calculated with this equation:

$$h(u,v) = \sum_{i\in u}\sum_{j\in v}\frac{S_i^*\cdot S_j^*}{d_{ij}^2}, u,v = 1,2,\dots,13 \,, \qquad (25)$$

where u and v are the atom types, S^* represents the modified E-state index, d_{ij} is the topological distance between v_i and v_j vertices.

There is also an extension of MEDV-13, called Molecular Holographic Distance Vector (MHDV).[248] It is developed to describe more specific molecular structures like peptide sequences, which contain heteroatoms and multiple bonds.

4.2.9.5 GRIND descriptors

Grid-Independent Descriptors (GRIND) are based on molecular interaction fields (MIF), which can be calculated with the GRID method and software[249] (or other programs).[250] The calculation of the descriptors involves two main steps. The first step is the computation of molecular interaction field (MIF). The second step involves a particular type of autocorrelation transform to create alignment-independent variables. On the other hand, we can also construct the steps in the following way: a) computing the MIF, b) filtering the MIF and c) encoding the virtual receptor site (VRS) into GRIND (GRIND encodes geometrical relationships between the VRS regions). The molecular descriptors can be used for the creation of “correlograms” and also for different types of chemometric analysis. The original MIF descriptors can be generated with appropriate software from the autocorrelation transform.

4.2.9.6 VolSurf descriptors

VolSurf descriptors have a quite similar basis to G-WHIM and GRIND descriptors. These descriptors also encode information about molecular interaction fields (MIF) with a GRID force field parametrization.[251] The final descriptors are alignment independent and are related to molecular size or shape, distribution of hydrophobic or hydrophilic parts of the molecule, *etc*. Different probes such as H_2O, DRY (hydrophobic) or O (carbonyl oxygen, represents hydrogen-bonding donor groups) are used for the calculation of interaction fields. Some typical descriptors are: water-excluded volume (*V*) with H_2O probe, accessible surface (traced also with H_2O probe) of the water interaction field, rugosity (R), polarizability (POL, traced with DRY probe), elongation (E), volume of the interactions with the probe O (Wp1-Wp8), Integy moments (INTEraction enerGY moments), *etc*.

4.3 4D and other special descriptors

4.3.1 4D descriptors

4D descriptors are also grid-based descriptors (but they are not derived from molecular interaction fields.) The fourth dimension represents ensemble sampling or conformational flexibility, which is defined by the *conformational ensemble profile* (CEP). The CEP is calculated with molecular dynamics (MD) simulations. 4D descriptors help to identify the active conformations of the flexible molecules and they are also useful tools in alignment problems. The pioneering work of Hopfinger *et al*. in 1997[252] has introduced 4D descriptors to QSAR analysis. Hence, 4D descriptors are among the youngest members of the family of (now thousands of) molecular descriptors. There are two forms of 4D QSAR: receptor dependent and receptor independent.[253]

4D descriptors can be assigned to the occupancy frequencies of the different atom types in the cubic grid cells during the molecular dynamics simulation.[254,255] The generation of the descriptors has some basic steps, which are discussed here in detail. First the molecular structures are generated with a conformational search and the conformers with the minimum energies are kept as the initial structures. The cell grid (with a default grid spacing of 1.0 Å) is constructed based on the largest compound in the data set. In the next step the interaction pharmacophore elements (IPE) are introduced. Every atom of each molecule can be classified into one of the IPE classes: "any type" or generic, nonpolar (NP), polar-positive charge (P+), polar negative charge (P-), hydrogen bond acceptor (HA), hydrogen bond donor (HB), aromatic (Ar). These types are created based on the interactions between the active site and the pharmacophore groups. The third step of the process is the estimation of the conformational ensemble profiles (CEP) with MD simulations. After that, each conformation of the CEP for each molecule is implemented in a reference grid cell based on the trial alignment. Finally, the grid cell occupancy descriptors are calculated from CEPs. This is important because we use a conformational ensemble profile of each compound instead of using only one starting conformation (which is the essence of the fourth dimension).

4.3.2 Other special descriptors and approaches (outlook)

The list does not stop at 4D descriptors, as one can easily find interesting articles about the application of higher (such as five or six) dimensional descriptors.[253] These are connected more to QSAR and related topics, thus they are only briefly mentioned. Five dimensional QSAR descriptions are calculated with the Quasar technology (and software) and are an extension of the 4D siblings with the multiple representation of the topology of the quasi-atomistic receptor surrogate.[256,257] Another dimension is introduced in 6D descriptors for QSAR modeling, enabling the simultaneous consideration of different solvation models. It can be achieved for example by mapping parts of the surface area with solvent properties .[258]

Another interesting field is the discrimination of molecular shapes. Gaussian approximations can be used to optimize the alignment of two molecules, but Gaussian functions can also provide very good and realistic representations of molecular shapes.[259] The rapid overlay of chemical structures (ROCS) program applies the atom-centered Gaussians for the discrimination and optimal alignment of molecules.[260,261] ROCS alignments can be applied not just for similarity searching and virtual screening, but also for 3D QSAR and SAR analyses. Further examples of shape-based approaches include the normalized PMI (principal moments of inertia) ratios introduced by Sauer and Schwarz[262], and Ultrafast Shape Recognition (USR), which employs three statistical moments (average, standard deviation and kurtosis) of atomic distances in four reference locations, providing a vector of 12 shape descriptors for each molecule.[263]

5. Similarity - dissimilarity measures, distance metrics

5.1 Introduction

The old adage dates back to the ancient Roman-Greek period: “similis simili gaudet” [like rejoices in like] and corresponds to the observation that similar entities behave similarly.

In the scientific domain of (medicinal) chemistry, there are some common formulations of the similarity principle, for example:

i) Structurally similar molecules are presumed to exhibit similar properties and similar biological activities.[264]

ii) “Molecular similarity in descriptor space is often called the “neighborhood principle” or “neighborhood behavior axiom”[265]

iii) Other scientists formulate the same principle as the “concept of similarity”, where molecules may be grouped according to their biological effects or physicochemical properties.[266]

iv) “Structure-property similarity principle”: similar structures generally have similar properties[267]

The similarity principle plays a ubiquitous role in rational drug design, lead discovery, synthesis design, molecular diversity analysis and compound optimization (*e.g.* optimization of ADMET properties (*e.g.* similar absorption, distribution, metabolism, excretion and toxicity) Virtual screening is able to find similar molecules from large databases. Finding a “patentable” compound with a desired property value is an important aim to be pointed out.

Naturally there are many examples violating the similarity principle. A small change in the molecular structure (changing the configuration of a stereo center, introducing a small molecular weight substituent, *etc.*) can lead to a dramatic increase (or decrease) in biological activity (*e.g.* dramatic difference in the toxicity of dioxins, or the carcinogenicity of polycyclic aromatic hydrocarbons (PAHs), *etc*.). On the other hand, a lot of compound classes (steranes, amphetamines, sulphonamides, *etc*.) possess very similar biological activities. If there is insufficient information about a new compound, the scientific community should resort to similar compounds and make estimations from “similar” molecules.

Similarity analysis can be simplified according to the following: First a molecule should be represented with an appropriate computational construct, such as a fingerprint or a set of descriptors (see subchapters 3 and 4). Then, a similarity measure (metric) should be applied for quantifying the similarities between pairs of molecules, and finally an algorithm can be applied for *e.g.* clustering the molecules according to their biological activity, properties, *etc.*

As the available computing power has dramatically increased in recent years, application of similarity methods to large databases became feasible even with moderate computing resources. The goal of such calculations is usually either the identification of molecules that are similar to one or more reference molecules, or the compilation of a subset of molecules that are as diverse as possible. (The latter application is particularly important in early hit discovery where diverse molecular databases are desired.) The reader is referred to some straightforward reviews about similarity and diversity in the cheminformatics field.[265,268–270]

5.2 Common measures of similarity and dissimilarity (distance)

Similarity and dissimilarity (distance) are used more or less interchangeably, in spite of some important differences*, e.g*. their directions are reversed, their scales are different by definition, *etc*. It is also important to note that there is an inherent asymmetry in the distributions of these metrics.

Holiday *et al*. have used three different clustering approaches and distinguished three types of measures: distance metrics, association coefficients and similarity (correlation) coefficients.[271]

A distance metric $D_{\mathrm{A,B}}$ between objects (molecules) *A* and *B* should obey four rules:[272]

Distance values (*D*) must be positive for non-identical objects: $D_{\mathrm{A,B}}$, > 0 (1)

The distance from an object to itself must be zero: $D_{\mathrm{A,A}}$, = $D_{\mathrm{B,B}}$, = 0 (2)

A distance value must be symmetric: $D_{\mathrm{A,B}}$, = $D_{\mathrm{B,A}}$ (3)

A distance metric must obey the *triangular inequality*: $D_{\mathrm{A,B}} \leq D_{\mathrm{A,C}} + D_{\mathrm{C,B}}$ (4)

If a distance coefficient fails to obey either of these four rules, it cannot be called a metric.

Similarity measures (*S*) are reversely scaled (the higher the more similar) as opposed to dissimilarity or distance (the lower the more similar). A similarity measure has to obey three rules.[273] It should be noted that Eq. (6) is erroneous in the original publication]:

Similarity values for non-identical objects must be: $0 < S_{\mathrm{A,B}} < 1$ (5)

The similarity of an object to itself must be unity: $S_{\mathrm{A,A}}$, = $S_{\mathrm{B,B}}$, =1 (6)

A similarity value must also be symmetric: $S_{\mathrm{A,B}}$, = $S_{\mathrm{B,A}}$ (7)

Both similarity and distance values are always non-negative, but their scales greatly differ: $S \in [0; +1]$ (or $S \in [-1; +1]$ for some correlation-based similarity definitions) and $D \in [0; +\infty]$. In practice, an inherent asymmetry is present in similarity/dissimilarity calculations in terms of the resulting value ranges. It is relatively simple to convert to similarities to dissimilarities:

$D_{A,B} = 1 - S_{A,B}$, such special distances will bear the constraint of $D \in [0; +1]$. (8)

However, conversion in the opposite direction is somewhat less straightforward, as similarity values should be recalculated between zero and one, while distances can be arbitrarily large. Thus, a suitable conversion rule is:

$$S_{A,B} = \frac{1}{1+D_{A,B}} \tag{9}$$

It is easy to see that similarity values calculated with this equation will always have a value between 0 and 1 (with 1 corresponding to identical objects, where the distance is 0).

It should be noted that the typical value ranges of different similarity metrics can be different,[274] even though they cannot fall outside of the predefined range (*i.e.* $0 \leq S \leq 1$).

As distances are measured on different scales, their comparison should be preceded by appropriate scaling. Four basic scaling methods are to be used: (i) range scaling (between 0 and 1), (ii) standardization (autoscaling, *i.e.* centered and scaled to unit standard deviation), (iii) rank transformation and (iv) normalization (scaled to unit length).

Table 2 summarizes the most frequently used similarity and distance measures. The first ten items in the table are distance measures, the remaining ones are similarity coefficients. For more binary similarity measures, the reader is referred to the work of Todeschini *et al.*[275] Additionally, several graph-based similarity measures have been proposed in the literature for molecular similarity calculations.[276,277] Some more specific similarity measures are defined in part 5.5.

Table 2. Formulas for frequently used similarity and distance measures.

Similarity/Distance measure	**Definition/Formula for numerical variables**[a]	**Formula for dichotomous variables**[b]
Hamming (Manhattan) distance	$D_{A,B} = \sum_{j=1}^{n} \lvert x_{j,A} - x_{j,B} \rvert$	$D_{A,B} = a + b - 2c$
Euclidean distance	$D_{A,B} = \sqrt{\sum_{j=1}^{n} \left(x_{j,A} - x_{j,B}\right)^2}$	$D_{A,B} = \sqrt{a + b - 2c}$
Soergel distance	$D_{A,B} = \frac{\sum_{j=1}^{n} \lvert x_{jA} - x_{jB} \rvert}{\sum_{j=1}^{n} max\left(x_{jA}, x_{jB}\right)}$	$D_{A,B} = 1 - \frac{c}{a + b - c}$
Pearson correlation (*r*) distance (*PCD*)	$\mathrm{PCD}_{\mathrm{A,B}} = 1 - \mathrm{abs}(r)$ for centered *x* variables *r* is identical with cosine coefficient otherwise: $r = \frac{\sum_{j=1}^{n}(x_{j,A} - \bar{x}_j)(x_{j,B} - \bar{x}_j)}{\sqrt{\sum_{j=1}^{n}(x_{j,A} - \bar{x}_j)^2 \sum_{j=1}^{n}(x_{j,B} - \bar{x}_j)^2}}$	$\mathrm{PCD}_{\mathrm{A,B}} = 1 - \frac{c}{\sqrt{ab}}$
Spearman rank correlation (*ρ*) distance	$\mathrm{SRCD}_{\mathrm{A,B}} = 1 - \mathrm{abs}(\rho)$ $\rho = \frac{\sum_{j=1}^{n} R(x_{j,A})R(x_{j,B}) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\left(\sum_{j=1}^{n} R(x_{j,A})^2 - n\left(\frac{n+1}{2}\right)^2\right)\left(\sum_{j=1}^{n} R(x_{j,B})^2 - n\left(\frac{n+1}{2}\right)^2\right)}}$	*Not defined*

Pearson correlation squared distance	$\mathrm{PCSD}_{\mathrm{A,B}} = 1 - r^2$	$\mathrm{PCSD}_{\mathrm{A,B}} = 1 - \frac{\mathrm{c}^2}{ab}$
Spearman rank correlation squared distance	$\mathrm{SRCD}_{\mathrm{A,B}} = 1 - \rho^2$	*Not defined*
Minkowski distance	$D_{A,B} = \left(\sum_{j=1}^{n} \lvert x_{j,A} - x_{j,B} \rvert^{p}\right)^{1/q}$ Where *p* and *q* > 0 and generally *p* = *q*	*Not defined*
Chebishev distance	$D_{A,B} = \lim_{\mathrm{n}\to\infty} \left(\sum_{j=1}^{n} \lvert x_{j,A} - x_{j,B} \rvert^{p}\right)^{1/q}$	*Not defined*
Correlation coefficient	Pearson: $r = \frac{\sum_{j=1}^{n}(x_{j,A} - \bar{x}_j)(x_{j,B} - \bar{x}_j)}{\sqrt{\sum_{j=1}^{n}(x_{j,A} - \bar{x}_j)^2 \sum_{j=1}^{n}(x_{j,B} - \bar{x}_j)^2}}$	Matthews: $S_{A,B} = \frac{cd - (a - c)(b - c)}{\sqrt{ab(a - c + d)(b - c + d)}}$
Ochiai/Cosine coefficient	$S_{A,B} = \frac{\sum_{j=1}^{n} x_{j,A}x_{j,B}}{\sqrt{\sum_{j=1}^{n}\left(x_{j,A}\right)^2 \sum_{j=1}^{n}\left(x_{j,B}\right)^2}}$	$S_{A,B} = \frac{c}{\sqrt{ab}}$
Dice coefficient	$S_{A,B} = \frac{2\sum_{j=1}^{n} x_{j,A}x_{j,B}}{\sum_{j=1}^{n}\left(x_{j,A}\right)^2 + \sum_{j=1}^{n}\left(x_{j,B}\right)^2}$	$S_{A,B} = \frac{2c}{a + b}$
Tanimoto coefficient	$S_{A,B} = \frac{\left[\sum_{j=1}^{n} x_{j,A}x_{j,B}\right]}{\left[\sum_{j=1}^{n}\left(x_{j,A}\right)^2 + \sum_{j=1}^{n}\left(x_{j,B}\right)^2 - \sum_{j=1}^{n} x_{j,A}x_{j,B}\right]}$	$S_{A,B} = \frac{c}{a + b - c}$
Russell-Rao coefficient	*Not defined*	$S_{A,B} = \frac{c}{n}$

Forbes coefficient	*Not defined*	$S_{A,B} = \frac{cn}{(a+c)(b+c)}$
Simpson coefficient	*Not defined*	$S_{A,B} = \frac{c}{min(a,b)}$
Simple matching coefficient	*Not defined*	$S_{A,B} = \frac{c+d}{n}$
Tversky coefficient	*Not defined*	$S_{A,B} = \frac{c}{\alpha(a-c)+\beta(b-c)+c}$

[a] In the definitions for continuous variables (such as physicochemical properties or biological activities), $x_{j,A}$ and $x_{j,B}$ are the values of feature j for molecules A and B respectively, $\bar{x}_j$ is the average value of feature j, $R(x_{j,A})$ is the rank number of feature j of molecule A, and n is the number of features.

[b] In the definitions for binary variables (such as fingerprints), a is the number of *on* bits in molecule A, b is number of *on* bits in molecule B, c is the number of bits that are *on* in both molecules, d is the number of common *off* bits and n is the bit length (total number of bits) of the fingerprint: $n = a + b - c + d$.

The Minkowski distance is a generalization of distances, if $p = q = 1$, it equals the Manhattan distance whereas $p = q = 2$ equals the Euclidean distance. The Tversky coefficient can also be considered as a generalization of similarity coefficients. The α and β parameters are always non-negative, but in medicinal chemistry applications, they are usually set within the unit interval [0, 1]. If $\alpha \neq \beta$, then the Tversky coefficient is asymmetric. The case $\alpha = \beta = 1$ equals the Tanimoto similarity. Similarity indices are intrinsically symmetric in nature, but asymmetric indices have some advantages: asymmetric forms allow for measuring and modulating the similarity of one molecule in the context of another *i.e.* they have the potential of alleviating the size dependency often observed in chemical similarity searching. Inspired by Tversky's work, Mestres and Maggiora have defined an entire family of field-based molecular similarity indices.[278]

Distance and similarity metrics have been the subject of several comparative studies focusing on their applications in several fields. Bender *et al.* have compared a vast selection of molecular fingerprints and visualized their similarities (*i.e.* the similarities between the methods themselves) with principal component analysis.[88] Additionally, they have included some of the fingerprints in combination with the Cosine similarity metric (replacing the Tanimoto coefficient). Sum of ranking differences (SRD)[279] corroborates the authors' original observation that the exchange of the similarity metric from Tanimoto to Cosine does not affect the ranking behavior of these fingerprints significantly (see Figure 13). In addition, all of the examined fingerprint methods have ranked the datasets significantly differently than random ranking.

More recently, Bajusz *et al.* have established that the Tanimoto index, Dice index, Cosine coefficient and Soergel distance were identified to be the best and in some sense equivalent measures for similarity calculations (in combination with the path-based Chemaxon Chemical Fingerprint), whereas the similarity metrics derived from the Euclidean and Manhattan distances are not recommended on their own, although their variability and diversity from other similarity metrics might be advantageous in certain cases. The behavior of these similarity coefficients was very similar for fragment-sized, leadlike and druglike compounds. Analysis of variance (ANOVA) showed that the choice of the data pre-processing method (such as interval scaling, standardization and rank transformation) does not affect the results significantly.[274]

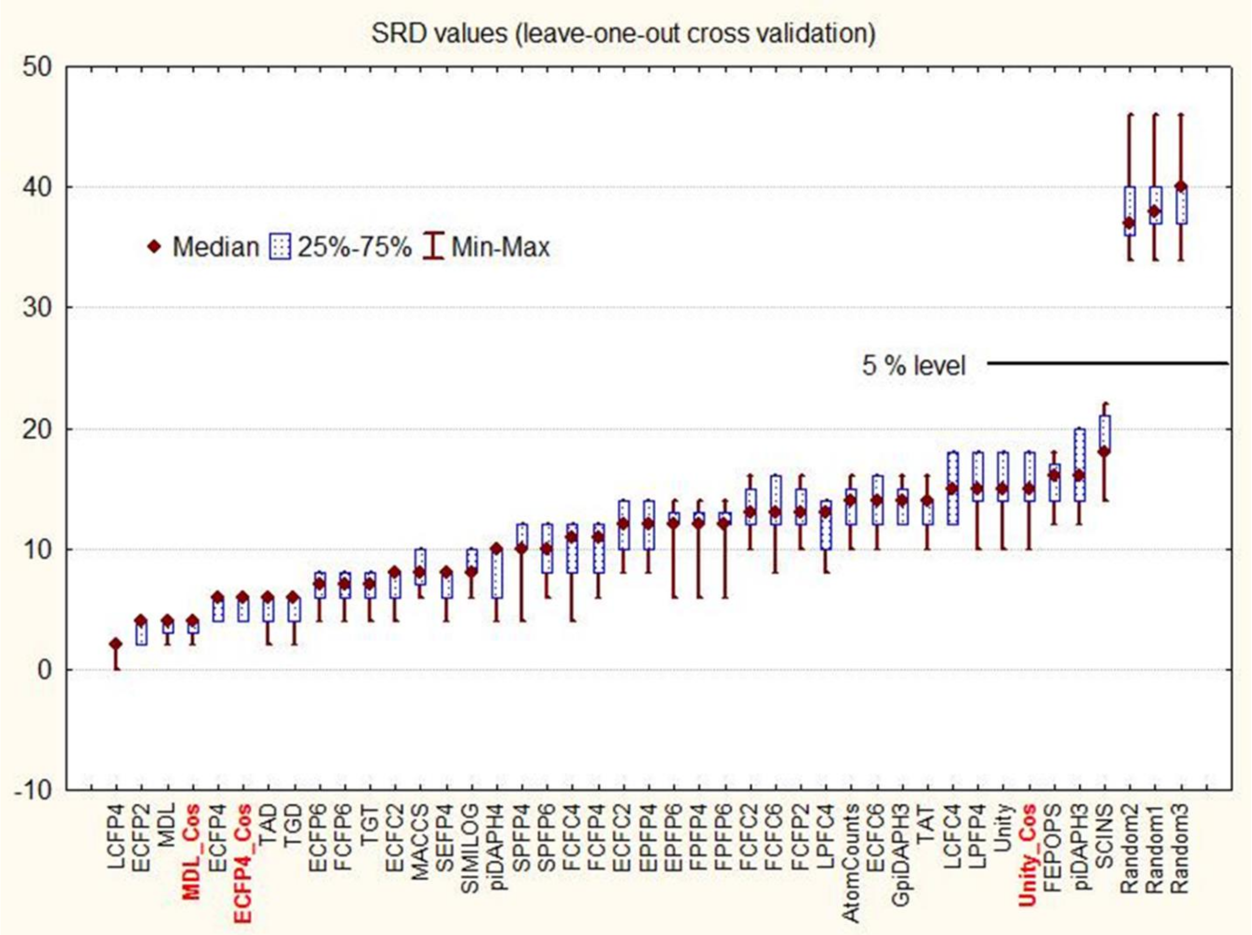
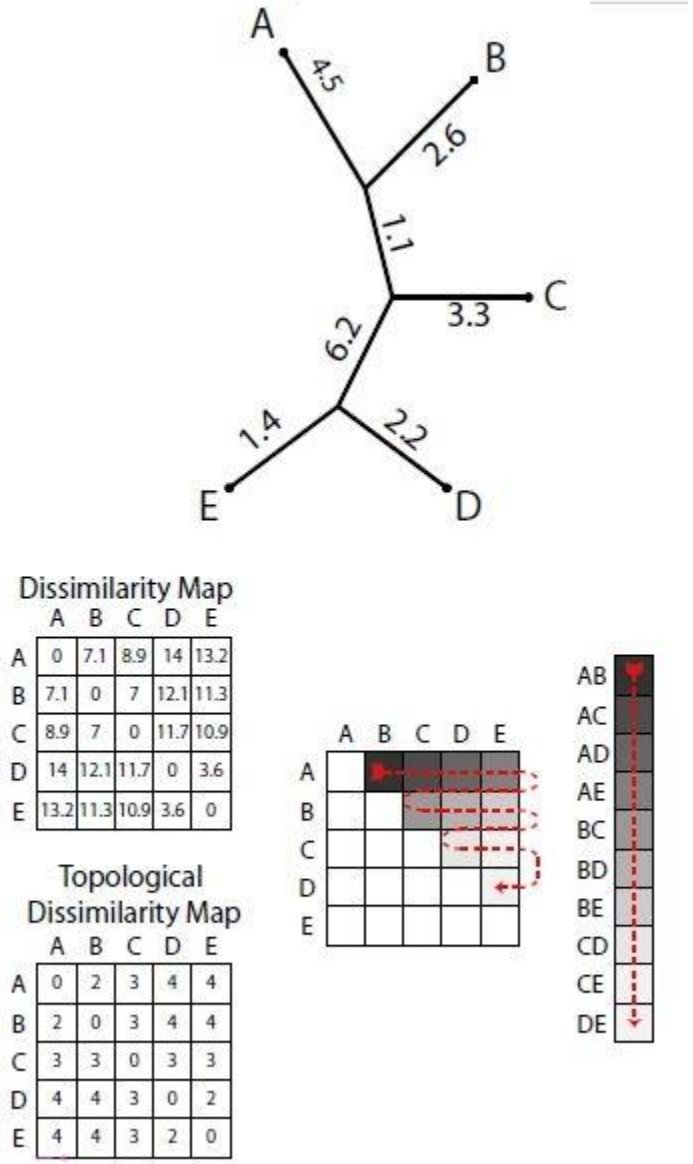


Figure 13. Leave-one-out cross-validated SRD (sum of ranking differences) values for the fingerprint methods compared by Bender *et al.*[88] As SRD values measure a method's distance from the consensus (in terms of ranking the observations – here, molecules), the smaller SRD values are better. Here, LCFP4 proved to be the best (most consistent) fingerprint for these eleven activity classes. The ranking has passed the randomization test, 5 % level of random ranking is shown in the figure. It can also be concluded that the similarity metrics (Tanimoto or Cosine, the latter is marked with red) do not affect the ranking behavior significantly.

Haws *et al*. have described an easy way to unfold symmetric diagonal matrices into vectors (so that they can be compared with *e.g*. the similarity metrics included in Table 2).[280] Their first figure uses the example of a dissimilarity map, but it can also be used for correlation matrices as well (see Figure 14).



Dissimilarity Map

	A	B	C	D	E
A	0	7.1	8.9	14	13.2
B	7.1	0	7	12.1	11.3
C	8.9	7	0	11.7	10.9
D	14	12.1	11.7	0	3.6
E	13.2	11.3	10.9	3.6	0

Topological Dissimilarity Map

	A	B	C	D	E
A	0	2	3	4	4
B	2	0	3	4	4
C	3	3	0	3	3
D	4	4	3	0	2
E	4	4	3	2	0

Figure 14. A process to "unfold" diagonal dissimilarity maps into vectors according to Haws *et al*.[280] The dissimilarity map contains distances whereas the topological dissimilarity map is calculated with path counts. While the original work revolves around phylogenetic trees, the concept presented can be easily translated to be applicable to distance, similarity or correlation matrices, as well as topological distance matrices (as presented in subchapter 4.1.2).

Schuffenhauer and Brown have grasped the problem reversely:[281] "diversity selection can be … only of value if used with dissimilarity cut-offs in ranges where the similar property principle is at least partially valid." Mean pairwise intermolecular dissimilarity has been calculated for a rapid selection of external datasets.[282]

"There is clearly a lot of 'art' involved in defining similarity, and different definitions are useful for different purposes." "Different methods select different subsets of actives for the same biological activity and the same method might work better on some activities than others in a way that is difficult to predict beforehand. In retrospect, this makes sense because receptors are diverse, and chemical groups that appear equivalent to one descriptor might not be equivalent to another."[283] Novel types of plots for the comparison of binary similarity coefficients have been elaborated by Salim *et al*. These plots clearly show the differences between coefficients in size, and the size bias of similarity coefficients is revealed.[284]

There is no doubt that the Tanimoto coefficient is the most frequently applied similarity measure for bit strings. It can be advantageously applied in many cases, but it is not the best (or even deficient) in other cases. Below we summarize some (contradictory) findings concerning the Tanimoto coefficient. If a compound has a Tanimoto similarity coefficient (based on 'Unity' fingerprints) larger than 0.85 to an active compound, then the compound has an 80 % chance of itself being active in the same assay.[285] Martin *et al*. have shown that similar biological activity might be expected from structurally similar compounds: "as the structural similarity is increased, so is the biological similarity." The enrichment in active compounds is higher than the same of docking to proteins of known 3D structures. If setting the limit for Tanimoto coefficients higher than 0.85 in the case of Daylight fingerprints only 30% of compounds proved to be active.[264] Cosine and Tanimoto coefficients were compared. If the Cosine coefficient is used for the calculation of the intermolecular (dis)similarities, a set of dissimilar molecules is selected faster than earlier algorithms. The algorithm is applicable to any type of representation that characterizes a molecule by a set of attribute values and to any procedure that involves calculating a sum of inter-molecular similarities.[286] The Tanimoto coefficient is not a perfect similarity measure, it also has some limitations. The distribution of Tanimoto coefficients for a comparison of 54-bit strings is peculiar: it is a multimodal skewed distribution.[287] "The Jaccard coefficient, also known as the Tanimoto coefficient is the most widely used in practice." "There cannot be one similarity measure and one descriptor that correlates with every molecular property at the same time. In different "similarities", different features emerge as being important (and in our case, different bioactivities invariably require different descriptors)."[266] Even the efficiency of the well-known Tanimoto coefficient can be increased using a single, bioactive reference compound (or using several reference structures) with data fusion (see later) and machine-learning techniques.[288]

"It was also found that different coefficients perform better in certain ranges of molecular size (or bit density). The Russell–Rao coefficient was found to perform better in the case of large queries, while the Forbes coefficient performed better on small queries. The Tanimoto coefficient was outperformed in many cases, but not consistently. The good performance of Russell–Rao and the weak performance of Forbes were also observed in an application using the dictionary of natural products database (DNP), where the Tanimoto coefficient was often outperformed by a factor of two."[289] "Certain coefficients whether single or in combination, appear repeatedly as best performers. One would definitely include the Russell/Rao, Forbes, and Simple Match in this pool and would probably add the Tanimoto, Cosine and others."[284] It was revealed that the size-bias and asymmetries that are inherent in most similarity coefficients lead to a bias in the selection of active compounds (altogether 14 coefficients were examined, including the Tanimoto coefficient).[290] Considering the large number of studies about the inconsistent and not satisfactory behavior of the Tanimoto coefficient, it is somewhat surprising that it is still the "default" similarity metric.

Association coefficients may be more suitable for 2D fragment-based similarity searching than distance values. Atom sequences seem to perform best among the studied descriptors. Combination of atom sequences and the set theoretic Tanimoto coefficient is strongly recommended in similarity searching.[291] Fingerprint-based iterative similarity search with multiple active compounds as references performs better than 2D fingerprint similarity searching in a large-scale comparative analysis carried out on 208 well-defined compound activity classes.[292]

In the work of Reisen *et al.*, six types of similarity measure were tested for their use in high content screening (HCS):[293] i) distance measures (Euclidean, Manhattan, and Mahalanobis distances), ii) linear correlation measures (Pearson correlation coefficient, Cosine similarity), iii) nonlinear correlation measures on a ratio scale (Maximum information coefficient, Distance correlation), iv) nonlinear correlation measures on an ordinal scale (Kendall’s τ, Spearman’s ρ), v) comparison of up- and downregulated features using a threshold (Tanimoto index, Dice coefficient), vi) CMAP-like similarity measures (see Figure 15).

Receiver operating characteristic (ROC) curves are used to evaluate the performance of the similarity functions using the area under the curve (AUC). Data preprocessing greatly influences the results. The right choice of a similarity measure is a prerequisite for high-quality, high-content screening fingerprints. The nonlinear rank-based correlation methods (Kendall’s τ and Spearman’s ρ) seem to be suitable for identifying similarities and dissimilarities among the high-dimensional high-content screening readouts.

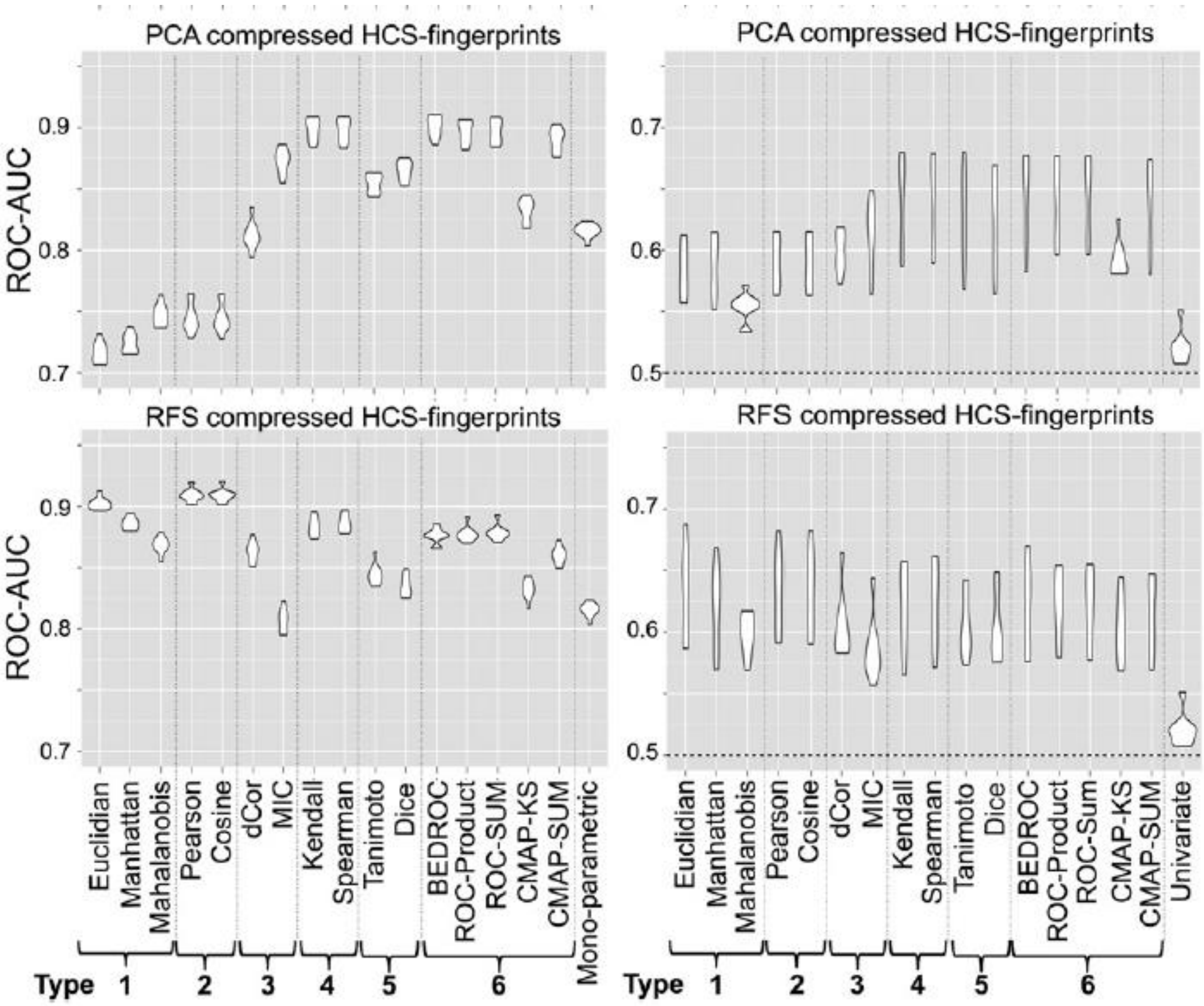


Figure 15. Comparison of similarity metrics for high content screening (HCS) fingerprints in the work of Reisen *et al.*[293] Left column: similarity analysis among replicates, right column: correlation between chemical and HCS fingerprints. Violin plots show the performance distribution over the 10 separate evaluation sets. © 2013 Society for Laboratory Automation and Screening. Copyright permission needed!

Scaffold hopping plays an increasing role in the last decade. Sufficient degree of diversity should be preserved not to overlook promising drug candidates or classes. A set of methods is presented that are designed to find compounds that are structurally different to a certain query compound while retaining its bioactivity properties (scaffold hops). These methods utilize various indirect ways of measuring the similarity between the query and a compound that take into account additional information beyond their structure-based similarities.[294] Stiefl *et al.* have defined scaffold hops based on whether the extended reduced graph approach (ErG) is able to switch among different chemotypes (and if yes, to what extent).[295] Group fusion is suitable for scaffold-hopping applications as well. Scaffold similarity searches were introduced by Ertl.[296] Classical similarity search and scaffold keys similarity search was compared and the superiority of the latter has been proven. He also stated that the successful Tanimoto similarity measure is unfortunately not working well for scaffolds.

Structural unit analysis identifies the molecular substructures or fragments that distinguish compounds with high activity from those with average activity. The method is suitable to scaffold hopping, as well. A set of techniques were elaborated using the nearest-neighbor graph-based similarity.[297] Maggiora, in his editorial, discussed the reasons why QSAR modeling often fails, and mentioned the idea of activity cliffs: "identifying and removing outliers may not necessarily always be a statistical problem as some outliers may only be *apparent* and may, in fact, arise from activity cliffs in the data."[298] Maggiora *et al*. outlined the various similarity concepts in an easily perceivable way.[299] Chemical, molecular, biological, global and local similarity all have different meanings. They reformulated the definition of the Tanimoto coefficient in a more intuitive way to interpret it "as the ratio of the number of features shared by *A* and *B* molecules to the number of their unique features." The different results of similarity searching has been shown if using different fingerprints and similarity coefficients.

Interestingly Muchmore *et al.* utilize a different terminology in their work, where they call fingerprints as similarity measures. A novel probabilistic framework was used for their interpretation. Tanimoto coefficients were calculated for ten different similarity methods (MACCS keys, Daylight fingerprints, maximum common subgraphs, rapid overlay of chemical structures (ROCS) shape similarity, and six connectivity-based fingerprints) combined with a database of more than 150 000 compounds and activity data against 23 protein targets. Different similarity measures were compared with receiver operating characteristic curves (ROC) and the Boltzmann-enhanced discrimination of receiver operating characteristics (BEDROC). Decision theory helped in the data fusion and the probability that any two molecules will exhibit similar biological activity is calculated.[300]

Density in chemical space is one of the limitations when using similarity methods. Not sufficient density causes heavy computational time consumption and it can lead to sometimes inaccurate results.[93] Todeschini *et al.* compared the largest pool of binary similarity coefficients.[275] Five pairs and a triplet of coefficients were found to yield identical similarity values, *i.e.* the same coefficients were rediscovered and bear different names. The similarity coefficients were grouped differently: symmetric, asymmetric intermediate and correlation-based binary coefficients (their shapes were plotted for the simulated data sets). Other partitioning options include metric-nonmetric, and – based on the shape behavior – increasing exponential, quasi-linear, logarithmic and sigmoidal. The best ranked coefficient was CT4 (Consonni-Todeschini coefficient, fourth version),[301] which is basically the logarithmic transformation of the (shifted) Tanimoto index used for similarity analysis. Nonetheless, the Harris−Lahey, Tanimoto, Gower−Legendre, Sokal−Sneath and Jaccard (2012) coefficients are ranked similarly well on two real data sets (MDDR and WOMBAT).

5.3 Similarity fusion, fusion rules, consensus scoring

In analytical chemistry data fusion usually means uniting data of vastly different origins and scales, for example the fusion of NMR and mass spectrometry data[302] or the combination of "fluorescence with the biomarkers… and traditional metabolomics measurements in the form of ^{1}H NMR spectroscopy", as reported in the work of Bro *et al.*[303]

In chemoinformatics and medicinal chemistry, data fusion usually means "merging" different, eventually seriously conflicting rankings. The basic idea is far-reaching: multicriteria optimization, method comparison, feature ranking can all be considered as data fusion techniques.

Selection of structurally diverse subsets of chemical structures is a valuable aim. Several different algorithms have been compared and all of them have been suggested for dissimilarity-based compound selection, provided that they are sufficiently rapid in execution for use with large files of compounds. MaxMin is the best algorithm currently available for non-focused, dissimilarity-based compound selection.[304]

Virtual screening based on fingerprints is rarely used as a standalone method. Methods such as group fusion, data fusion or voting have been introduced to combine similarities calculated from multiple fingerprint methods.[93] Binary similarity coefficients (22) were compared using Unity fingerprints (2D fragment bit-strings), the coefficients were clustered and a consensus scoring was calculated.[271] Two, three and four binary similarity coefficients were merged in the work of Salim *et al*., who have established that the best fusions were better than the best single coefficients for 12 data sets out of 15. Consensus scoring has been shown to improve the hit rates in virtual library screening.[284] There are different possibilities to combine information from several scoring algorithms to provide a single prediction. Using several data sets, Ginn *et al*. used MIN, MAX and SUM rules, defining the combined prediction by the lowest, highest and average prediction of the individual methods. Both SUM and MAX are, overall, to be preferred to the individual results. The most effective for similarity fusion is generally the usage of the SUM rule with rank data. It was also found that consensus scoring performed better (and significant at $p < 0.05$) in 28 out of 30 runs, if the SUM rule is used. In practice, one can use this information to determine how to deal with multiple known active structures: in particular it was shown that adding up individual scores of each pair of query and library compound improves the overall results.[305]

As so many similarity methods have been defined, it became a valid research goal to provide guidance on which similarity measure is the most suitable for solving a special task. Sheridan and Kearsley provided some justification to use a large number of molecular similarity methods.[283]

As some of the available similarity coefficients quantify different types of structural resemblance, it seems to be advantageous to use them in combination. Willett distinguishes between *similarity fusion* (combining the results of database searches that use a common reference structure but use different similarity coefficients) and *group fusion* (when several, structurally diverse reference structures are available, the reference molecule is allowed to vary and the similarity measure is kept constant).[89,306,307] *Consensus scoring* (combination of the results of different search algorithms and/or scoring functions) is a similar data fusion technique for structure-based approaches, such as docking.

Later, Willett extended the three accepted fusion rules (MIN, MAX and SUM) to include the average, median, geometric and harmonic means, Euclidian norm, and some other rules concentrating on the top ranks only. Fusion rules are to be used for similarity scores or the ranks alike. Supervised fusion rules are also interpreted in the Bayesian framework, when biological activity is also considered in the search.[307] Willett devoted a chapter to answer the question: why does data fusion work? He has established that "the SUM rule is likely to out-perform the MAX rule in similarity fusion; that the converse applies in group fusion; and that group fusion is generally far superior to similarity fusion". A further fusion rule is emphasized in Willett's work: the reciprocal rank fusion (RRF) rule, which is applicable only to rank data and which derives from the fact that virtual screening often involves applying a cut-off on the similarity scores (such as the top-1%) so that only a small fraction of the database is considered further in a project.[89] Let p ($p \leq n$) be the number of times that an individual database structure dy, occurs above the chosen cut-off. Then the RRF rule involves summing the reciprocal ranks for those p occurrences to give a fused score:

$$\sum_{x=1}^{p} \frac{1}{RANK_x(dy)} \qquad (10)$$

The reciprocal rank fusion rule outperformed all of the other rules that Chen *et al.* considered in their comparative study of 15 different fusion rules.[308] Group fusion is most effective when: i) as many reference structures as possible are used; ii) only a small proportion (1-5 %) of each ranked similarity list is submitted to the final fusion rule; and iii) when the reciprocal rank rule is used to combine the individual search outputs. The Pareto data fusion approach has been elaborated by Cross *et al*.[309] The Pareto approach counts for each molecule the number of times other molecules achieve a better rank in all of the lists, thus the best molecules will receive a Pareto score of 0. (The ties are managed by successive interactive ranking).

Independently from virtual screening and similarity coefficients, data fusion is extensively used for method and model comparison.[279,310,311] Sum of ranking differences (SRD) are supported with theoretical considerations:[312,313] in particular, the average of ranks (or scores) is a better option than any of the individual ranks (or scores), as derived from the maximum likelihood principle. The SRD procedure involves two kinds of validation: a randomization test and a leave-one-out or leave-many-out cross-validation. Maximum for correct classification rates or minimum for error rates is the natural data fusion choice.[312]

Performance evaluations of ranking methods in the context of virtual screening cannot be evaluated without accepted, consensual metrics. Area under the receiver operating characteristic curve is not suitable to the "early recognition" problem. Performance indicators such as enrichment factors, robust initial enhancement, Boltzmann enhanced discrimination of receiver operating characteristic, area under the accumulation curve corresponding to an empirical cumulative density function, and their weighted variants were analyzed theoretically in detail.[314]

5.4 Clustering algorithms

Clustering is the collective name for a group of methods, where the molecules (samples, objects, *etc.*) are arranged in groups (or "clusters") based on their distances from (in other words their similarities to) each other. Since clustering is not the main focus of this chapter, we refer the reader to recent, well-written reviews on clustering methods and applications from the fields of chemometrics[273] and cheminformatics,[315] and collect only a small set of diverse developments and applications in this subchapter. (Also, it would be virtually impossible to enumerate all clustering algorithms that were provided by the machine learning community.)

“Using various stopping rules the grouping of similarity coefficients resulted in three, 11 or 13 clusters.”[284]

“The average clustering coefficient similarity threshold function can be characterized by the presence of a peak that covers a range of similarity threshold values. This peak is preceded by a steep decline in the number of edges of the similarity network. The maximum of this peak is well aligned with the best clustering outcome. If no reference set is available, choosing the similarity threshold associated with this peak would be a near-ideal setting for the subsequent network cluster analysis.”[316]

“Despite the long tradition of pattern recognition research, there is no technique that yields the best classification in all scenarios. Therefore, as many techniques as possible should be considered in high accuracy applications. Typical related works either focus on the performance of a given algorithm or compare various classification methods.” Amancio *et al*. compared the performance of nine well-known classifiers and found that the *k*-nearest neighbor method frequently allowed the best accuracy.[317]

Large margin nearest neighbors approach and its multi-metric extensions have recently been elaborated by Kireeva *et al*.[318] Their algorithms cluster the compounds in the training set with the same property label together while the compounds from different classes are separated by a large margin. In most of the cases the metric learning algorithm leads to better classification; the performance dependence from the data density has been discussed. The *k*-medoids clustering method was favored by Jaskowiak *et al*.[319] Saeh *et al*. generated robust models while combining 3D pharmacophore fingerprints and the support vector machine classification algorithm. Lead-hopping was also simulated: an entire class of compounds was excluded from the training set. Still, the model trained on the remaining compounds was able to recall 75% of the actives from the “new” lead series and correctly classifying >99% of the 5000 inactive compounds included in the validation set.[320]

The generalized metric swarm learning (GMSL) algorithm has been developed by Zhang and Zhang, where a sample pair is represented as a similarity vector *via* the well-learned metric swarm. The sample pairs are transformed into a vectorized similarity space (metric swarm space) *via* an established joint similarity function, whereas SVM-like classification can be easily implemented.[321] They have presented the efficiency of their approach on the example of face recognition but it should have great potential in medicinal chemistry, as well. Gaussian Ensemble Screening (GES) was developed by Perez-Nueno *et al*.[322] “GES is a new and fast way to predict polypharmacological relationships between drug classes; it quantitatively provides an efficient way to measure the similarity between clusters of arbitrary numbers of members.”

5.5 Similarity measures for special tasks

5.5.1 Symmetric field-based similarity indices

The first quantum chemical similarity index has been defined by the analogy of correlation coefficients in 1980:

$$r_{AB} = \frac{\int \rho_A \rho_B dV}{\sqrt{\int \rho_A^2 dV} \sqrt{\int \rho_B^2 dV}} \tag{11}$$

where ρ_A and ρ_B are the first-order density functions for molecules *A* and *B*.[323] This is the most frequently used field-based similarity index. Similar ones are defined by Hodgkin and Richards[324] and by Petke.[325] A comparative analysis of quantum chemical similarity and dissimilarity indices has been carried out not long ago.[326]

Standard Quantum-Based (SQB) similarity methods were compared with Tanimoto based similarity and it was established that "the use of a complex number format for molecular representation proved to be superior compared to real representation and benchmark Tanimoto method (TAN), where the complex SQB method outperformed TAN in nine cases".[327]

5.5.2. Similarities of 1D structures

Gene expression data analysis requires special similarity measures. The Jackknife correlation coefficient (successively removed the outlying features) provided the best results, showing better enrichments than other distance measures whereas the second ranked measure is the Kendall tau. Good results were also shown by Manhattan distance, Popular distance measures in the gene expression clustering literature, namely, Euclidean distance, Spearman, and, Pearson coefficients displayed inferior results to at least five other distances under evaluation.[319]

In another example, Dixon and Merz, Jr. have introduced a truly 1D representation of chemical structure. A 3D molecular model or a 2D chemical graph is projected onto a single coordinate of atomic positions. A novel measure of overall structural similarity has been defined after the alignment of 1D representations to match identical atom types:

$$Sim_{A,B} = \frac{S_{A,B}^{max}}{\sqrt{S_{A,A}^{max} S_{B,B}^{max}}} \tag{12}$$

1D similarity has been defined such a way: molecule *B* is aligned at various offsets with respect to molecule *A*, and the total overlap area S_{AB} between rectangular regions of the same type is computed. The largest overlap area that can be achieved, $S_{A,B}^{max}$, is combined with normalization factors from aligning each molecule with itself. The constraint is valid: $0 < Sim_{A,B} < +1$. These 1D similarities have been reported to consistently outperform both Daylight 2D fingerprints and Cerius2 pharmacophore fingerprints.[181]

Similarities of drug chemical structures, of drug protein targets and of drug side-effect profiles (Tanimoto similarities and protein sequence similarities) were calculated and fused into an overall prediction score using the *k*-nearest neighbor algorithm.[328]

A web-interfaced target identification program (called TargetHunter) with a built-in powerful data-mining algorithm (TAMOSIC) has been elaborated for predicting potential biological targets of query compounds. Its performance has been compared with the multiple-category models (MCM) for predicting protein targets and therapeutic activities.[329]

A revised algorithm based on Shannon's information theory and the Neumann entropy characterizes each individual protein residue position with the number of significantly correlated pairs by computing the corresponding mutual information matrix.[330] Chemogenomics aspects were summarized by Bender *et al*. focusing on target prediction of small molecules with molecular descriptor based models.[159] Semantic links between compounds and proteins, including *similarity neighboring links* and interaction links were utilized to predict drug target interactions.[331] Features are structured according to prior knowledge into groups based on similarity in gene expression.[332] Pearson correlation similarity was calculated and heatmaps were plotted.

5.5.3 Graph (tree) based similarities

Direct similarity between the two feature trees have been created by Rarey and Dixon.[333] The similarity value of the feature trees is computed by a weighted average over the similarity values of the matches. The weight factor for each match is the sum of the subtree sizes *size(m)* of the match. The total weight is the sum of the sizes of the matched subtrees plus a scaling factor *u* times the total size of the unmatched subtrees:

$$S_M(A,B) = \frac{\sum_{x\in M} size(m) sim(m)}{u(size(A)+size(B))+(1-u)\sum_{x\in M} size(m)} \tag{13}$$

There is a notable research gap on comparing different approaches retrieving those models in the repository that most closely resemble a given process model or fragment thereof.[334] Business process graphs use various notations and customs. Three types of parameterized similarity metrics between business process models have been elaborated: i) node matching similarity metrics based on properties of business process model elements (such as their labels and their other attributes); ii) structural similarity metrics based on the relations between these elements; and iii) behavior similarity metrics based on the intended behavior of process models.

A detailed, comprehensive survey on techniques to define and calculate business similarity measures has been carried out. Nine properties of distance measures are enumerated and twenty-one business similarity measures have been compared in their work.[335] Sum of ranking differences using the average as reference provides the ordering presented in Figure 16.

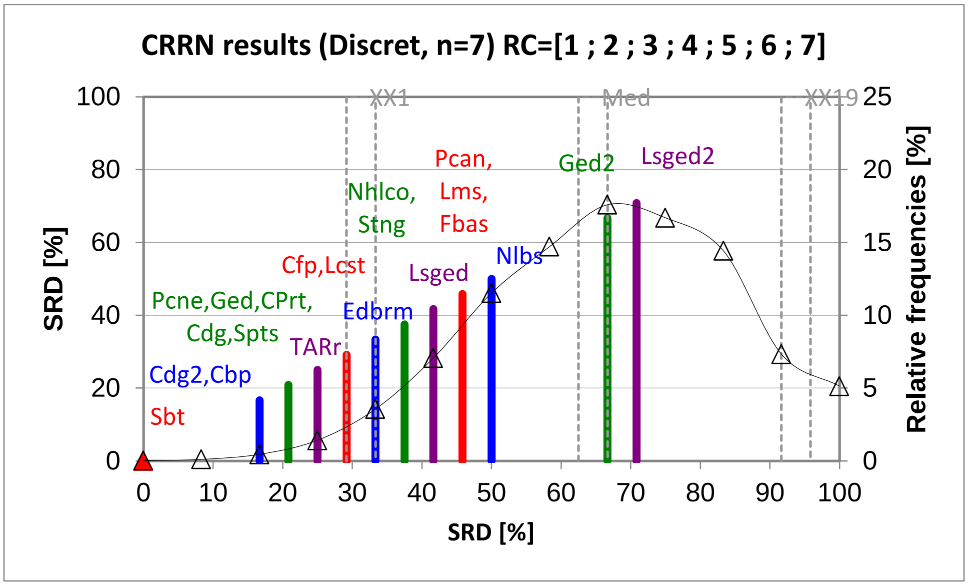


Figure 16. Ordering of business similarity measures by sum of ranking differences. Notations: Percentage of common activity names (Pcan), Label matching similarity (Lms), Similarity of activity labels (Sal), Feature-based activity similarity (Fbas), Percentage of common nodes and edges (Pcne), Node-and link-based similarity (Nlbs), Graph edit distance (Ged), Graph edit distance (Ged2), Label similarity and graph edit distance (Lsged) Label similarity and graph edit distance (Lsged2), Number of high-level change operations (Nhlco) Comparing PMs represented as trees (CPrt) Edit distance between reduced models (Edbrm), Comparing dependency graphs (Cdg), Comparing dependency graphs (Cdg2), TAR-relationship (TARr) Causal behavioural profiles (Cbp) Causal foot prints (Cfp), Sets of traces as n-grams (Stng) Longest common subsequence of traces, Lcst, Similarity based on principal transition sequences (Spts) Similarity based on traces (Sbt). XX1 band corresponds to 5 % first kind error, *i.e.* many of the similarity measures are indistinguishable from random ranking (black distorted Gaussian like curve and right y axis).

In a ligand-based virtual screening experiment, the similarity between every library molecule and a query molecule is measured by some similarity function. An extension of the optimal assignment method for chemical graphs was produced that uses evolutionary algorithms to optimize edge weights. A variety of similarity functions can be improved by optimizing the edge weights.[336] The similarity between two chemical graphs can be evaluated by means of the maximum common subgraph approach.[337] Ralaivola *et al*. reviewed graph kernels and developed new graph kernels (Tanimoto, MinMax and Hybrid kernel) for chemical molecules. These kernels measure the similarity between feature vectors, or molecular fingerprints, consisting of binary vectors or vectors of counts.[338]

5.5.4 Semantic similarity measures

Semantic similarity was defined by Resnik based on information content:[339]

$$sim(c_1c_2) = \max_{c\in S(c_1c_2)}[-logp(c)] \qquad (14)$$

where $sim(c_1c_2) \in [0; +\infty]$ and $-logp(c)$ is the negative log likelihood, or the *information content* of a concept *c*.

A similar definition is given by Lin.[340] Lin defines the similarity between two terms as the ratio of the commonality of the terms and the information needed to fully describe the two terms:

$$sim(c_1c_2) = \max_{c \in S(c_1c_2)} \left[\frac{2logp(c)}{logp(c_1)+logp(c_2)}\right] \tag{15}$$

where $sim(c_1c_2) \in [0; +1]$.

Schlicker *et al*. have combined Lin's and Resnik's similarities into relevance similarity:[341]

$$sim(c_1c_2) = \max_{c \in S(c_1c_2)} \left[\frac{2logp(c)}{logp(c_1)+logp(c_2)}\big(1-p(c)\big)\right] \tag{16}$$

Their approach enables the comparison of the underlying molecular biology of different taxonomic groups and provides a new comparative genomics tool identifying functionally related gene products.

A novel similarity measure for ligand-based virtual screening has been created from a text processing similarity measure called Adapted Similarity Measure of Text Processing (ASMTP).[342] ASMTP was reported to outperform the Tanimoto coefficient-based virtual screening.

Kendall's τ has been the most widely used measure of similarity between two orderings, R* and R^F; a novel similarity function is defined as:

$$\tau(R^*, R^F) = \frac{P}{P+Q} = \frac{Q}{\binom{|D|}{2}} \tag{17}$$

where *P* is the number of concordant pairs and *Q* is the number of discordant pairs on a data set *D*.[343] An optimal method has been derived based on rank support vector machine that selects the most ambiguous objects for ranking so that the ordering on the set maximizes the degree of learning.

5.5.5 Supervised similarity measures

Ligand similarity measures, such as shape similarity are defined such a way: each atom of a molecule is mapped to a five dimensional space, where the first three coordinates are determined by the 3D conformation of the molecule. The two remaining dimensions encode the atomic partial charges and the atomic lipophilicity (AlogP).[344,345] Multiple logistic regression was applied to combine different similarity values.

One popular measure of the roughness of a structure–activity landscape is the structure–activity landscape index, SALI:[346]

$$SALI_{A,B} = \frac{|A_A - A_B|}{1 - Sim_{A,B}}, \tag{18}$$

where A_A and A_B are the affinities/activities of compounds A and B, while $Sim_{A,B}$ is their similarity. Basic characteristics of molecular similarity and dissimilarity *networks* are reviewed by Sukumar *et al*.[347] "The choice of descriptors affects the computed similarity—for instance, two molecules might be constructed from the same molecular scaffold and thus be very similar in size and shape, but have very different properties because of the different chemical natures of the functional groups or substituent atoms. Conversely, molecules with very different molecular scaffolds might look similar for binding to a protein."

Cuissart *et al*. have constituted two more similarity indices in the biodegradability context:[348]

$$Sim_1(QS, IS) = \frac{|MCS|}{|QS|} \tag{19}$$

and

$$Sim_2(QS, IS) = \frac{|MCS|}{|QS|} \times \frac{|MCS|}{|IS|} \tag{20}$$

where $|S|$ equals the number of non-hydrogen atoms within molecule S and MCS is the maximum common substructure between two molecules. "IS" means "Instance Structure", whose biodegradability value is used to compute the prediction model, whereas QS means "Query Structure", whose biodegradability is predicted by this model. Both similarity indices are discriminant to a group of compounds of similar activity. The index Sim_2 operates more in conformity with the SAR principles than Sim_1.

Dropout training of artificial neural networks (ANNs) outperforms conventional ANNs, when using two types of fingerprints and two similarity metrics (Tanimoto and Buser).[349] Dropout means randomly omitting a large portion of the nodes from an artificial neural network.

Naderi *et al*. have predicted drug-likeness for molecular synthesis. Each active compound in the DUD-E library was decomposed into fragments and a molecular synthesis was simulated. A search space was constructed from small molecules, followed by the stochastic exploration of the chemical space by constructing multi-fragment molecules. Tanimoto coefficient was used to reveal the chemical similarity. Parent compounds are compared to those constructed with the eSynth computational package using molecular fingerprint matching.[350]

Sixteen different feature ranking methods were compared by Jankowski and Usowicz in combination of six weighting schemes and the counts of winnings, defeats and draws were presented.[351]

5.5.6 Chemometric/Spectral similarity

An adaptive similarity measure (sample similarity) was proposed to construct highly accurate locally weighted partial least squares (LW-PLS) calibration models. The method is based on the weighted Euclidean distance and was used for weighting in regression.[352]

Common mathematical methods to express similarity in NIR (near-infrared) spectroscopy are correlation coefficients and distances.[353] The authors of a recent work have successfully transformed differences in spectra into Bayesian hypothesis testing.[354] Logarithm values of the posterior odds ratios for the spectra are able to detect subtle differences, *e.g.* changes in the analytical process.

A novel spectral similarity measure was introduced by Bodis *et al*.[355] De Gelder and colleagues have shown that "various similarity and dissimilarity criteria previously described in literature can be written as special cases of a general expression." The have introduced a new similarity criterion, based on this generalized expression.[356]

The spectral-contrast-angle method introduced by Wan *et al.* is based on the vector representation of mass spectra and has been shown to perform better than the similarity index method for spectral comparison.[357] The spectral contrast angle is defined as the angle between two mass spectrum vectors.

The correlation and congruence coefficients (*i.e.* similarity values) of all total ion chromatograms relative to the reference chromatogram were calculated with median or averaged data. Principal component analysis and cluster analysis can successfully discriminate between mountain origins of green teas, and therefore can be further applied to identify and authenticate Pu-Erh green teas.[358]

Three mass spectral similarity measures (NIST composite measure, the real part of Discrete Fourier Transform and the detail of Discrete Wavelet Transform) were compared and integrated with retention indices. As a consequence compound identification was enhanced by 1.7-3.5 %.[359]

Similarity of infrared spectra were characterized by: correlation coefficient of mean centered absorbance (*i.e.* cosine measure), mean of the absolute absorbance differences, mean of the squared absorbance differences, dot product of the absorbance vectors, each normalized to unit length, whereas Tanimoto index has provided the chemical structure similarity. The definitions were given in vector notations and scaled properly. The first hits corresponding to the most similar spectra yield the highest structural similarity with the query compound. Among the four investigated spectral similarity measures the correlation coefficient of mean-centered absorbances performed best.[360]

5.5.7 Fuzzy similarity measures

Enormous progress took place in the development of fuzzy similarity measures. Fuzzy sets are sets whose elements have degrees of membership. Thus, instead of 0 (not a member) and 1 (is a member), the membership of an element in a fuzzy set can be any real number between 0 and 1.

Fuzzy similarity measures are defined for the comparison of such fuzzy sets. A small but characteristic selection can be found below.

Three new equations were suggested to calculate the distance between intuitionistic fuzzy sets (IFSs) on the basis of the Hausdorff distance. The proposed similarity measure is much simpler than the existing methods and is well suited to use with linguistic variables.[361]

Existing similarity measures can provide unreasonable results in some special cases. Therefore, several new similarity measures were proposed to differentiate different IFSs. The proposed similarity measures can deal with problems more effectively and reasonably than some of the existing methods.[362]

The degree of similarity between intuitionistic fuzzy sets *A* and *B* (B^c is the complement of *B*) is defined as:

$$\dot{s}(A,B) = \sum_{j=1}^{n} \dot{s}\left(\alpha_j,\beta_j\right) = \frac{1}{n}\sum_{j=1}^{n} \frac{d\left(\alpha_j,\beta_j^c\right)}{d\left(\alpha_j,\beta_j\right)+d\left(\alpha_j,\beta_j^c\right)} \tag{21}$$

where α_j and β_j are the *j*th intuitionistic fuzzy values of *A* and *B*, respectively. The analogous formula has been used for interval-valued intuitionistic fuzzy sets, as well. Then, the developed similarity measure was applied for consensus analysis in group decision making based on intuitionistic fuzzy preference relations, and it was extended to the interval-valued intuitionistic fuzzy sets.[363]

"Reasonable" measures to calculate the degree of similarity between intuitionistic fuzzy sets were proposed based on the L_p metric.[364] A new similarity measure for IFSs was suggested and its usefulness in medical diagnostic reasoning was proven.[365]

An axiomatic definition of a similarity measure between dual hesitant fuzzy sets was proposed and the shortcomings in existing similarity measures were enumerated.[366]

Some new distance measures between intuitionistic fuzzy sets (IFSs) were suggested. Maximum degree of similarity between IFSs was applied for pattern recognition.[367]

Some similarity measures were introduced between two triangular fuzzy numbers based on the vector similarity measures in vector space (which can be used to aggregate the decision information). A methodology for multiple criteria group decision-making (MCGDM) problems with triangular fuzzy information is proposed; the criteria values take the form of linguistic values. The weighted similarity measures between each alternative and ideal alternative, can be used to rank alternatives and select the most desirable alternative.[368]

Dual hesitant fuzzy sets include fuzzy sets, intuitionistic fuzzy sets and hesitant fuzzy sets as its special cases. Some distance and similarity measures based on the Hamming distance, Euclidean distance and Hausdorff distance were derived for usage in decision making, pattern recognition, *etc*. Two examples illustrate these distance and similarity measures and their applications in pattern recognition.[369]

Novel discrete and continuous hesitant fuzzy distance measures and hesitant fuzzy ordered distance measures between hesitant fuzzy sets were elaborated. A hesitant fuzzy clustering algorithm based on novel similarity measures was also created.[370]

5.5.8 Other similarity measures

An object is described by sets of features instead of geometric points in a metric space. A new measure of remoteness between sets of nominal values is proposed instead of considering the distance between two sets: a new measure of perturbation type 1 of one set by another was introduced. The consideration is based on set-theoretic operations and the proposed measure describes changes of the second set after adding the first set to it, or *vice versa*. The measure of the sets' perturbation returns a value in the range [0, 1], and this measure is not symmetric in general.[371]

The cosine formula often yields a similarity measure in citation studies that is twice the number of that obtained by the Jaccard index.[372]

Most online shopping sites and many other applications now use collaborative recommender systems. The measurement of similarity plays a fundamental role in such systems. The accuracy of the most well-known similarity measures (Pearson's correlation coefficient, cosine similarity and mean squared differences) decreases due to data sparsity. Therefore, a user-user potential matrix is calculated; potential similarities between users from this matrix are computed; the potential similarities are modified based on the users' preliminary neighborhoods, and *k* users with the highest modified similarity values are selected.[373]

6. Online web resources

Since the pool of cheminformatics and molecular modeling software is quickly and constantly renewed, we refrain from providing a detailed overview of them: a current and prominent selection was presented recently by Cereto-Massagué *et al.*,[86] and we have already referred to several of them throughout this chapter. However, online resources such as molecular databases are more permanent over time, thus we provide a short overview of top-level databases to provide some guidelines to anyone who wishes to apply them. These databases implement diverse ways of submitting search queries: besides recognizing chemical names and SMILES strings, many of them have an integrated sketching interface and support similarity searches (based on the entered queries). Most of these databases offer some possibilities for batch downloading and query automation as well, through a graphical user interface (GUI), an application programming interface (API) or other mechanisms.

- Chemical Abstracts Service (CAS) Registry: The CAS Registry (operated by the American Chemical Society) is the largest molecular database to date.[3,4] It is constantly updated and stores every reported chemical structure (*cca.* 111 million at the time these lines are written) with a uniquely assigned identifier, the CAS Registry number. In addition to the database itself, the CAS Registry powers the two major chemical information services of CAS: Scifinder (for querying chemical structures and reactions, primarily in the literature)[5] and STN (a search engine that provides access to patent content).[6] While these are paid services, a freely available website covering a subset of high interest of the CAS Registry entries – titled Common Chemistry – is also operated by CAS.[374]

- Pubchem: An open chemistry database operated by the National Center for Biotechnology Information.[107,375] The database covers a diverse set of information about the stored compounds, from chemical names, 2D and 3D structures and line notations (InChI, InChIKey, SMILES) to calculated properties and even vendor and patent information. Pubchem operates separate (overlapping) databases for substances (*i.e.* depositor-supplied molecules) and compounds (*i.e.* unique chemical structures, including stereoisomers and even tautomers). In addition, the Pubchem Bioassay database is a rich source of deposited biochemical screening data.[376]
- ChemSpider: An open chemistry database operated by the Royal Society of Chemistry.[377] In addition to 2D and 3D representations and chemical names, ChemSpider also stores experimentally determined properties and NMR spectra, as well as links to literature articles and reference works.
- ChEMBL: Operated by the European Bioinformatics Institute (EMBL-EBI), ChEMBL is today's probably largest database of experimental bioactivity data.[378–380] In addition to bioactivity data, line notations, links to external databases and drug-related information (availability type, route of administration, *etc.*) are also stored. ChEMBL integrates bioactivity data from various sources, including full-matrix datasets such as DrugMatrix or the GSK Published Kinase Inhibitor Set, as well as screening results contributed by pharma companies. Lately, ChEMBL has integrated the full bioactivity content of PubChem BioAssays, which currently accounts for more than half of the data points in the ChEMBL database. EMBL-EBI also operates an open, searchable patent database called SureChEMBL.[381]
- Protein Data Bank (PDB): The Protein Data Bank is without question the largest and most commonly used repository for experimentally determined protein structures, hence it is one of the most important data sources for medicinal chemical and modeling work.[40] The database is thoroughly linked and annotated and offers a detailed and highly configurable "Advanced search" mechanism for exploring and analyzing not only the protein-ligand complexes, but also the ligands themselves.
- ZINC (ZINC Is Not Commercial): Probably the largest molecular database focusing on commercially available compounds (storing approx. 35 million compounds the time these lines are written).[382,383] In addition to storing the molecules with calculated properties and links to vendor databases, ZINC also offers a plethora of subsets for downloading, categorized based on various attributes, such as physicochemical properties, targets, vendors or availability. Thus, it a flexible and customizable source of information for virtual screening. Other comprehensive vendor databases include eMolecules[384] and Mcule[385], with the latter offering a wide set of online virtual screening tools as well.
- GDB databases: Besides commercially available compounds and experimentally determined data, databases of virtual compounds are also available, the most prominent of which are probably the GDB databases based on the work of the research group of Reymond.[149] GDB databases are generated with a systematic enumeration of all possible combinations of a predefined number of common heavy atoms (C, N, O, S and halogens).

For example, the latest database, GDB-17 enumerates molecules of up to 17 heavy atoms: a total of 166 billion molecules![386] Such approaches serve not only as a basis for analytical work, but also define the future directions in synthetic medicinal chemistry by providing an abundance of ideas for new scaffolds. In addition to downloadable subsets, the same group has implemented online browsers based on various fingerprints, as well as molecular quantum numbers (MQN)[111] for the efficient exploration of the chemical universe.

7. Outlook

Cheminformatics – and by extension, computational medicinal chemistry – is a field that has the potential to change very quickly. Novel file formats and fingerprints, based on original ideas, continue to emerge, providing an always renewing toolbox for the researchers of these fields. However, applications seem to lag behind these technological advances. If we want to look for the reasons behind that, we might consider factors such as publication delays, customs and personal preferences (*i.e.* if I have learned to apply one method that seems to work, I might not be open to learning a new one), or a sort of "canonization" (the process of a scholarly community accepting some works as the most important – not to be confused with canonicalization!) similar to that in literature. For example, if ECFP4 fingerprints are widely accepted for quantifying the 2D similarities of molecules, one might feel a motivation to prefer them over *e.g.* topological torsion fingerprints, when the latter might be in fact more suited for the given application. Similarly, the MDL *mol* (and *sdf*) formats continue to be the most prevalent chemical file formats despite their flaws and limitations. Since we maintain that true progress requires the thorough and independent testing, validation and application of the new computational tools, we expect a steady increase in the number of application-related works in this field. (Nonetheless, we hope to have provided useful guidance for that with this chapter.)

In terms of descriptors, the question is always open: „Do we need more and more molecular descriptors?" This question has been raised already in 2001 by Randic and Basak.[200] On the other hand, a distinguished publication of Cherkasov *et al.* paraphrases Mark Twain about QSAR (and thus molecular descriptors): „reports of [QSAR] death are an exaggeration".[387]

After carefully re-reading many publications in the field of molecular descriptors, we can infer the following: although there are some novel descriptors in the recent literature, most of the widespread ones were developed in the second part of the last century. In the near future we expect more application-related articles and somewhat fewer works about the development of new descriptors and fingerprints. However, it is also clear based on the aforementioned works and our previous experience that we want to have as diverse of a toolbox as possible, to cover the various possible descriptions of molecules as fully as possible. (It is worth noting that this is true for other fields of science as well, in fact it is an essential part of making progress in science.)

Like molecular descriptors, similarity measures are abundantly developed recently. The machine learning community provides a plethora of novel algorithms, whose limitations, usefulness and applicability have been only scarcely studied as of yet. More and more specific applications are taken into account, *e.g.* the early recognition problem or the need for diverse data sets. Unfortunately, a lot of novel measures and algorithms require an elaborate computational and mathematical skill set and expertise from the user to understand and implement them. In addition, some factors (weighting, desirability) are too subjective to let the measures and algorithms spread widely and unambiguously.

Consensus modeling (the usage of the average for more similarity measures) becomes more and more popular, and multicriteria optimization seems to be unavoidable nowadays. Multicriteria decision making (MCDM) and multiresponse optimization (MRO) intend to find compromising (optimal, consensus) solutions to unsolved problems. Recently the equivalency of MCDM with sum of ranking differences (SRD) has been proven.[388] SRD provides optimal solutions without the above mentioned subjectivity, and its propagation can be predicted without much risk. Like molecular representations, various similarity measures and algorithms can be used for solving different types of problems; a complex interplay of the mentioned methods should be considered and are expected as a future trend.

Novel similarity measures are to be expected from diverse disciplines (graph theoretical- business similarity-, fuzzy measures, *etc.* are transferrable to pharmaceutical applications) with great probability. Such measures will be tested and filtered by the drug design community – possibly slowly but definitely.

While there are well-established methods and algorithms in cheminformatics, new problems and new approaches arise abundantly and the pharma industry needs economic and effective techniques. As the amount of the generated data increases, cheminformatics and rational drug design approach the domain of big data. Nevertheless, extensions of the current data analysis methodologies are already being developed and successfully applied in related fields. Machine learning algorithms can operate with extremely complex models (*e.g.* neural networks with many thousands or even millions of nodes, ant colony, support vector machines, random forest, *etc.*) and may solve problems that scientists only dreamed of – even just several years ago. However, the validation practices are developing slowly and more emphasis should be placed on developing suitable validation techniques.

Finally, we refer to some recent reviews showing the present state of arts and some future trends in the cheminformatics field.[268–270]

References

(1) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (1), 33.

(2) IUPAC Nomenclature Home Page http://www.chem.qmul.ac.uk/iupac/ (accessed Apr 27, 2016).

(3) CAS Registry System. *J. Chem. Inf. Comput. Sci.* **1978**, *18* (1), 58–58.

(4) Chemical Substances - CAS REGISTRY http://www.cas.org/content/chemical-substances (accessed Apr 28, 2016).

(5) Scifinder https://scifinder.cas.org (accessed Apr 28, 2016).

(6) STN - The Choice of Patent Experts http://www.cas.org/products/stn (accessed Apr 28, 2016).

(7) Wiswesser, W. J. How the WLN Began in 1949 and How It Might Be in 1999. *J. Chem. Inf. Comput. Sci.* **1982**, *22* (2), 88–93.

(8) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36.

(9) SMILES - A Simplified Chemical Language http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed May 2, 2016).

(10) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29* (2), 97–101.

(11) Daylight Chemical Information Systems http://www.daylight.com/ (accessed May 2, 2016).

(12) O’Boyle, N. M. Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4* (1), 22.

(13) Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order—An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55* (10), 2111–2120.

(14) RDKit: Open-Source Cheminformatics Software http://rdkit.org/ (accessed May 2, 2016).

(15) SMARTS - A Language for Describing Molecular Patterns http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed May 2, 2016).

(16) SMIRKS - A Reaction Transform Language http://daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed May 2, 2016).

(17) Guasch, L.; Sitzmann, M.; Nicklaus, M. C. Enumeration of Ring–Chain Tautomers Based on SMIRKS Rules. *J. Chem. Inf. Model.* **2014**, *54* (9), 2423–2432.

(18) Proschak, E.; Wegner, J. K.; Schüller, A.; Schneider, G.; Fechner, U. Molecular Query Language (MQL) - A Context-Free Grammar for Substructure Matching. *J. Chem. Inf. Model.* **2007**, *47* (2), 295–301.

(19) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493–500.

(20) 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* (17), 2111–2120.

(21) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminform.* **2013**, *5* (1), 7.

(22) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, 23.

(23) IUPAC - International Union of Pure and Applied Chemistry: Home http://www.iupac.org/ (accessed Apr 28, 2016).

(24) National Institute of Standards and Technology http://www.nist.gov/ (accessed Apr 28, 2016).

(25) InChI Trust: Home http://www.inchi-trust.org/ (accessed Apr 28, 2016).

(26) Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. InChIKey Collision Resistance: An Experimental Testing. *J. Cheminform.* **2012**, *4* (1), 39.

(27) Unichem https://www.ebi.ac.uk/unichem/.

(28) Fritts, L. E.; Schwind, M. M. Using the Wiswesser Line Notation (WLN) for Online, Interactive Searching of Chemical Structurest. *J. Chem. Inf. Comput. Sci* **1982**, *22*, 106–109.

(29) Vollmer, J. J. Wiswesser Line Notation: An Introduction. *J. Chem. Educ.* **1983**, *60* (3), 192.

(30) Engel, T. Representation of Chemical Compounds. In *Chemoinformatics: A Textbook*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2003; pp 15–169.

(31) ChemDoodle | Chemical Drawing Software https://www.chemdoodle.com/ (accessed Apr 27, 2016).

(32) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 71–79.

(33) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model.* **2008**, *48* (12), 2294–2307.

(34) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32* (3), 244–255.

(35) Marvin 16.4.25. ChemAxon LLC (http://www.chemaxon.com) 2016.

(36) Apodaca, R. On the (F)utility of Extending the Molfile Format http://depth-first.com/articles/2012/01/11/on-the-futility-of-extending-the-molfile-format/ (accessed May 3, 2016).

(37) Clark, A. M. Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting. *J. Chem. Inf. Model.* **2011**, *51* (12), 3149–3157.

(38) Clark, A. M. Why Not to Use MDL MOL/SDF http://molmatinf.com/whynotmolsdf.html (accessed May 3, 2016).

(39) Tripos Mol2 File Format http://www.tripos.com/data/support/mol2.pdf (accessed May 17, 2016).

(40) RCSB Protein Data Bank http://www.rcsb.org/pdb/home/home.do (accessed May 18, 2016).

(41) wwPDB: File Format http://www.wwpdb.org/documentation/file-format (accessed May 18, 2016).

(42) IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. Tentative Rules (1969). *J. Biol. Chem.* **1970**, *246* (24), 6489–6497.

(43) PDB format 1992 http://www.wwpdb.org/docs/documentation/file-format/PDB_format_1992.pdf (accessed May 18, 2016).

(44) Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1991**, *47* (6), 655–685.

(45) Brown, I. D.; McMahon, B. CIF: The Computer Language of Crystallography. *Acta Crystallogr. Sect. B Struct. Sci.* **2002**, *58* (3), 317–324.

(46) PDBx/mmCIF Dictionary Resources http://mmcif.wwpdb.org/ (accessed May 19, 2016).

(47) Westbrook, J.; Ito, N.; Nakamura, H.; Henrick, K.; Berman, H. M. PDBML: The Representation of Archival Macromolecular Structure Data in XML. *Bioinforma. Orig. Pap.* **2005**, *21* (7), 988–99210.

(48) PDBML Schema Resources http://pdbml.pdb.org/ (accessed May 19, 2016).

(49) Lipman, D. J.; Pearson, W. R. Rapid and Sensitive Protein Similarity Searches. *Science* **1985**, *227* (4693), 1435–1441.

(50) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443–453.

(51) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410.

(52) BLAST: Basic Local Alignment Search Tool http://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed May 18, 2016).

(53) Dayhoff, M. O.; Schwartz, R.; Orcutt, B. C. A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*; Nat. Biomed. Res. Found., 1978; pp 345–358.

(54) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Biochemistry* **1992**, *89*, 10915–10919.

(55) Chemical Markup Language | CML http://www.xml-cml.org/ (accessed May 18, 2016).

(56) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 928–942.

(57) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 2. Information Objects and the CMLDOM. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1113–1123.

(58) Gkoutos, G. V.; Murray-Rust, P.; Rzepa, H. S.; Wright, M. Chemical Markup, XML, and the World-Wide Web. 3. Toward a Signed Semantic Chemical Web of Trust. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1124–1130.

(59) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 4. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 757–772.

(60) Murray-Rust, P.; Rzepa, H. S.; Williamson, M. J.; Willighagen, E. L. Chemical Markup,

XML, and the World Wide Web. 5. Applications of Chemical Metadata in RSS Aggregators. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 462–469.

(61) Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *J. Chem. Inf. Model.* **2006**, *46* (1), 145–157.

(62) Kuhn, S.; Helmus, T.; Lancashire, R. J.; Murray-Rust, P.; Rzepa, H. S.; Steinbeck, C.; Willighagen, E. L. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. *J. Chem. Inf. Model.* **2007**, *47* (6), 2015–2034.

(63) Adams, N.; Winter, J.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *J. Chem. Inf. Model.* **2008**, *48* (11), 2118–2128.

(64) Chemistry Add-in for Word - Microsoft Research http://research.microsoft.com/en-us/projects/chem4word/ (accessed May 18, 2016).

(65) Chemical Markup Language | Conventions http://www.xml-cml.org/convention/ (accessed May 18, 2016).

(66) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk—Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46* (3), 991–998.

(67) Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; *et al*. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* **2015**, *55* (3), 510–528.

(68) Vazquez, M.; Krallinger, M.; Leitner, F.; Valencia, A. Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. *Mol. Inform.* **2011**, *30* (6-7), 506–519.

(69) Gurulingappa, H.; Mudi, A.; Toldo, L.; Hofmann-Apitius, M.; Bhate, J. Challenges in Mining the Literature for Chemical Information. *RSC Adv.* **2013**, *3* (37), 16194.

(70) Eltyeb, S.; Salim, N. Chemical Named Entities Recognition: A Review on Approaches and Applications. *J. Cheminform.* **2014**, *6* (1), 17.

(71) Chemistry Text Mining Suite ChemAxon https://www.chemaxon.com/products/chemistry-text-mining-suite/ (accessed May 19, 2016).

(72) chemicalize.org http://www.chemicalize.org/ (accessed May 19, 2016).

(73) Southan, C.; Stracz, A. Extracting and Connecting Chemical Structures from Text Sources Using Chemicalize.org. *J. Cheminform.* **2013**, *5* (1), 20.

(74) Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. OSCAR4: A Flexible Architecture for Chemical Text-Mining. *J. Cheminform.* **2011**, *3* (1), 41.

(75) Usié, A.; Alves, R.; Solsona, F.; Vázquez, M.; Valencia, A. CheNER: Chemical Named Entity Recognizer. *Bioinformatics* **2014**, *30* (7), 1039–1040.

(76) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739–753.

(77) Valko, A. T.; Johnson, A. P. CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.* **2009**, *49* (4), 780–787.

(78) Imago OCR - Life Sciences 0.3.0 documentation http://lifescience.opensource.epam.com/imago/ (accessed May 20, 2016).

(79) Filippov, I. V.; Nicklaus, M. C. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740–743.

(80) Frasconi, P.; Gabbrielli, F.; Lippi, M.; Marinai, S. Markov Logic Networks for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.* **2014**, *54* (8), 2380–2390.

(81) Cosgrove, D. A. Markush Structures and Chemical Patents. In *Scaffold Hopping in Medicinal Chemistry*; Brown, N., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2013.

(82) Deng, W.; Berthel, S. J.; So, W. V. Intuitive Patent Markush Structure Visualization Tool for Medicinal Chemists. *J. Chem. Inf. Model.* **2011**, *51* (3), 511–520.

(83) Deng, W.; Scott, E.; Berthel, S. J.; Venus So, W. Deconvoluting Complex Patent Markush Structures: A Novel R-Group Numbering System. *World Pat. Inf.* **2012**, *34* (2), 128–133.

(84) Deng, W.; Schneider, G.; So, W. V. Mapping Chemical Structures to Markush Structures Using SMIRKS. *Mol. Inform.* **2011**, *30*, 665–671.

(85) Cosgrove, D. A.; Green, K. M.; Leach, A. G.; Poirrette, A.; Winter, J. A System for Encoding and Searching Markush Structures. *J. Chem. Inf. Model.* **2012**, *52* (8), 1936–1947.

(86) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

(87) Cruz-Monteagudo, M.; Medina-Franco, J. L.; Pérez-Castillo, Y.; Nicolotti, O.; Cordeiro, M. N. D. S.; Borges, F. Activity Cliffs in Drug Discovery: Dr Jekyll or Mr Hyde? *Drug Discov. Today* **2014**, *19* (8), 1069–1080.

(88) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods?: A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49* (1), 108–119.

(89) Willett, P. Fusing Similarity Rankings in Ligand-Based Virtual Screening. *Comput. Struct. Biotechnol. J.* **2013**, *5*, e201302002.

(90) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519.

(91) Tresadern, G.; Bemporad, D.; Howe, T. A Comparison of Ligand Based Virtual Screening Methods and Application to Corticotropin Releasing Factor 1 Receptor. *J. Mol. Graph. Model.* **2009**, *27* (8), 860–870.

(92) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data Set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50* (12), 2079–2093.

(93) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opin. Drug Discov.* **2016**, *11* (2), 137–148.

(94) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010**, *50* (5), 771–784.

(95) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29* (2), 157–170.

(96) Heikamp, K.; Bajorath, J. Fingerprint Design and Engineering Strategies: Rationalizing and Improving Similarity Search Performance. *Future Med. Chem.* **2012**, *4* (15), 1945–1959.

(97) Riniker, S.; Landrum, G. A. Open-Source Platform to Benchmark Fingerprints for Ligand-Based Virtual Screening. *J. Cheminform.* **2013**, *5* (1), 26.

(98) Riniker, S.; Landrum, G. A. Similarity Maps - a Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminform.* **2013**, *5* (1), 43.

(99) O'Boyle, N. M.; Hutchison, G. R. Cinfony – Combining Open Source Cheminformatics Toolkits behind a Common Interface. *Chem. Cent. J.* **2008**, *2* (1), 24.

(100) Cinfony - Cinfony 1.2 documentation http://cinfony.github.io/ (accessed Jun 3, 2016).

(101) Dong, J.; Cao, D.-S.; Miao, H.-Y.; Liu, S.; Deng, B.-C.; Yun, Y.-H.; Wang, N.-N.; Lu, A.-P.; Zeng, W.-B.; Chen, A. F. ChemDes: An Integrated Web-Based Platform for Molecular Descriptor and Fingerprint Computation. *J. Cheminform.* **2015**, *7* (1), 60.

(102) ChemDes - An integrated web-based platform for molecular descriptor and fingerprint computation http://www.scbdd.com/chemdes/ (accessed May 30, 2016).

(103) Christie, B. D.; Leland, B. A.; Nourse, J. G. Structure Searching in Chemical Databases by Direct Lookup Methods. *J. Chem. Inf. Model.* **1993**, *33* (4), 545–547.

(104) BIOVIA - Scientific Enterprise Software for Chemical Research, Material Science R&D http://accelrys.com/ (accessed Jun 3, 2016).

(105) Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017, San Diego: Dassault Systèmes, 2016.

(106) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.

(107) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. Chapter 12 – PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; 2008; Vol. 4, pp 217–241.

(108) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software [§]. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 141–142.

(109) Klekota, J.; Roth, F. P. Chemical Substructures That Enrich for Biological Activity. *Bioinformatics* **2008**, *24* (21), 2518–2525.

(110) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf.*

Model. **1995**, *35* (6), 1039–1045.

(111) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4* (11), 1803–1805.

(112) Deursen, R. van; Blum, L. C.; Reymond, J.-L. A Searchable Map of PubChem. *J. Chem. Inf. Model.* **2010**, *50* (11), 1924–1934.

(113) Fingerprints - Screening and Similarity http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed May 23, 2016).

(114) Dassault Systèmes BIOVIA, Pipeline Pilot, Version 9.5, San Diego: Dassault Systèmes, 2016.

(115) Schrödinger Release 2016-2: Canvas, Version 2.8. Schrödinger, LLC: New York, NY, USA 2016.

(116) Cheminformatics and Molecular Modeling | OpenEye Scientific Software http://eyesopen.com/ (accessed Jun 9, 2016).

(117) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.

(118) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47* (26), 6569–6583.

(119) Hassan, M.; Brown, R. D.; Varma-O’Brien, S.; Rogers, D. Cheminformatics Analysis and Learning in a Data Pipelining Environment. *Mol. Divers.* **2006**, *10* (3), 283–299.

(120) JChem 15.7.27 (Http://www.chemaxon.com). ChemAxon LLC 2016.

(121) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(122) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113.

(123) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Model.* **1985**, *25* (2), 64–73.

(124) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Model.* **1987**, *27* (2), 82–85.

(125) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 118–127.

(126) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (1), 128–136.

(127) Vulpetti, A.; Hommel, U.; Landrum, G.; Lewis, R.; Dalvit, C. Design and NMR-Based

Screening of LEF, a Library of Chemical Fragments with Different Local Environment of Fluorine. *J. Am. Chem. Soc.* **2009**, *131* (36), 12949–12959.

(128) Awale, M.; Reymond, J.-L. Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of ZINC and GDB-17. *J. Chem. Inf. Model.* **2014**, *54* (7), 1892–1907.

(129) Awale, M.; Jin, X.; Reymond, J.-L. Stereoselective Virtual Screening of the ZINC Database Using Atom Pair 3D-Fingerprints. *J. Cheminform.* **2015**, *7* (1), 3.

(130) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (3), 569–574.

(131) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 117–125.

(132) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42* (17), 3251–3264.

(133) Wood, D. J.; Vlieg, J. de; Wagener, M.; Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **2012**, *52* (8), 2031–2043.

(134) Molecular Operating Environment (MOE), 2013.08. Chemical Computing Group Inc.: Montreal, QC, Canada 2016.

(135) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D Fingerprints for Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2423–2431.

(136) Certara http://www.certara.com/ (accessed Jun 14, 2016).

(137) Xue, L.; Godden, J. W.; Bajorath, J. Database Searching for Compounds with Similar Biological Activity Using Short Binary Bit String Representations of Molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 881–886.

(138) Xue, L.; Godden, J. W.; Bajorath, J. Mini-Fingerprints for Virtual Screening: Design Principles and Generation of Novel Prototypes Based on Information Theory. *SAR QSAR Environ. Res.* **2003**, *14* (1), 27–40.

(139) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Model.* **2003**, *43* (4), 1151–1157.

(140) Eckert, H.; Bajorath, J. Design and Evaluation of a Novel Class-Directed 2D Fingerprint to Search for Structurally Diverse Active Compounds. *J. Chem. Inf. Model.* **2006**, *46* (6), 2515–2526.

(141) Nisius, B.; Bajorath, J. Molecular Fingerprint Recombination: Generating Hybrid Fingerprints for Similarity Searching from Different Fingerprint Types. *ChemMedChem* **2009**, *4* (11), 1859–1863.

(142) Nisius, B.; Bajorath, J. Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of Hits. *Chem. Biol. Drug Des.*

2010, *75* (2), 152–160.

(143) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 746–753.

(144) Arif, S. M.; Holliday, J. D.; Willett, P. The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening. In *Advances in Mathematical Chemistry and Applications*; Basak, S. C., Restrepo, G., Villaveces, J. L., Eds.; Bentham Science Publishers, 2014; pp 92–112.

(145) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53* (15), 5707–5715.

(146) Öztürk, H.; Ozkirimli, E.; Özgür, A. A Comparative Study of SMILES-Based Compound Similarity Functions for Drug-Target Interaction Prediction. *BMC Bioinformatics* **2016**, *17* (1), 128.

(147) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45* (2), 386–393.

(148) Schwartz, J.; Awale, M.; Reymond, J.-L. SMIfp (SMILES Fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules. *J. Chem. Inf. Model.* **2013**, *53* (8), 1979–1989.

(149) Research Group of Prof. Reymond http://www.gdb.unibe.ch/ (accessed May 31, 2016).

(150) Kooistra, A. J.; Binsl, T. W.; van Beek, J. H. G. M.; de Graaf, C.; Heringa, J. Electron Density Fingerprints (EDprints): Virtual Screening Using Assembled Information of Electron Density. *J. Chem. Inf. Model.* **2010**, *50* (10), 1772–1780.

(151) BatchNMRPredictor, Version 1.1a. Porta Nova Software GmbH: Zürich, Switzerland 2010.

(152) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47* (6), 2462–2474.

(153) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2* (2), 107–118.

(154) Beroza, P.; Villar, H. O.; Wick, M. M.; Martin, G. R. Chemoproteomics as a Basis for Post-Genomic Drug Discovery. *Drug Discov. Today* **2002**, *7* (15), 807–814.

(155) Dixon, S. L.; Villar, H. O. Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting. *J. Chem. Inf. Model.* **1998**, *38* (6), 1192–1203.

(156) Weinstein, J. N.; Myers, T. G.; O'Connor, P. M.; Friend, S. H.; Fornace, A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V; Anderson, N. L.; *et al*. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **1997**, *275* (5298), 343–349.

(157) Briem, H.; Lessel, U. F. In Vitro and in Silico Affinity Fingerprints: Finding Similarities beyond Structural Classes. *Perspect. Drug Discov. Des.* **2000**, *20* (1), 231–244.

(158) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. “Bayes Affinity Fingerprints” Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46* (6), 2445–2456.

(159) Bender, A.; Young, D. W.; Jenkins, J. L.; Serrano, M.; Mikhailov, D.; Clemons, P. A.; Davies, J. W. Chemogenomic Data Analysis: Prediction of Small-Molecule Targets and the Advent of Biological Fingerprints. *Comb. Chem. High Throughput Screen.* **2007**, *10* (8), 719–731.

(160) Deng, Z.; Chuaqui, C.; Singh, J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47* (2), 337–344.

(161) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural Interaction Fingerprints: A New Approach to Organizing, Mining, Analyzing, and Designing Protein-Small Molecule Complexes. *Chem. Biol. Drug Des.* **2006**, *67* (1), 5–12.

(162) Small-Molecule Drug Discovery Suite 2016-1. Schrödinger, LLC: New York, NY, USA 2016.

(163) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47* (1), 195–207.

(164) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-Based Interaction Fingerprint Scoring: A Simple Method for Improving the Effectiveness of Fast Scoring Functions. *J. Chem. Inf. Model.* **2006**, *46* (2), 686–698.

(165) Pérez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixidó, J. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2009**, *49* (5), 1245–1260.

(166) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54*, 2555–2561.

(167) Kinase-Ligand Interaction Fingerprints and Structures database (KLIFS) http://klifs.vu-compmedchem.nl/ (accessed May 26, 2016).

(168) van Linden, O. P. J.; Kooistra, A. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Knowledge-Based Structural Database to Navigate Kinase-Ligand Interaction Space. *J. Med. Chem.* **2014**, *57* (2), 249–277.

(169) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2015**, *44* (D1), D365–D371.

(170) PhosphoDiEsterase Structure and ligand Interaction Annotated database (PDEStrIAn) http://pdestrian.vu-compmedchem.nl/ (accessed May 26, 2016).

(171) Jansen, C.; Kooistra, A. J.; Kanev, G. K.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. PDEStrIAn: A Phosphodiesterase Structure and Ligand Interaction Annotated Database As a Tool for Structure-Based Drug Design. *J. Med. Chem.* **2016**, acs.jmedchem.5b01813.

(172) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And

Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47* (2), 279–294.

(173) FLAP - (Fingerprints for Ligands and Proteins) http://www.moldiscovery.com/software/flap/ (accessed May 31, 2016).

(174) Broughton, H.; Hunt, P.; Mackey, M. Methods for Classifying and Searching Chemical Reactions. US20030182094 A1, 2003.

(175) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, *3* (5), 821–832.

(176) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-Based Approach to *de Novo* Design Using Reaction Vectors. *J. Chem. Inf. Model.* **2009**, *49* (5), 1163–1184.

(177) Sheridan, R. P.; Hunt, P.; Culberson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46* (1), 180–192.

(178) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39–53.

(179) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley–VCH, Weinheim, Germany, 2000.

(180) Roy, K.; Kar, S.; Das, R. N. *A Primer on QSAR/QSPR Modeling - Fundamental Concepts*; Springer, 2015.

(181) Dixon, S. L.; Merz, K. M. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44* (23), 3795–3809.

(182) Hansch, C.; Fujita, T. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

(183) Leo, A. J. Calculating logPoct from Structures. *Chem. Rev.* **1993**, *93* (4), 1282–1306.

(184) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More than 96,000 Compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861–893.

(185) Andrić, F.; Bajusz, D.; Rácz, A.; Šegan, S.; Héberger, K. Multivariate Assessment of Lipophilicity Scales—computational and Reversed Phase Thin-Layer Chromatographic Indices. *J. Pharm. Biomed. Anal.* **2016**, *127*, 81–93.

(186) Andrić, F.; Héberger, K. Chromatographic and Computational Assessment of Lipophilicity Using Sum of Ranking Differences and Generalized Pair-Correlation. *J. Chromatogr. A* **2015**, *1380*, 130–138.

(187) Andrić, F.; Héberger, K. Towards Better Understanding of Lipophilicity: Assessment of in Silico and Chromatographic logP Measures for Pharmaceutically Important Compounds by Nonparametric Rankings. *J. Pharm. Biomed. Anal.* **2015**, *115*, 183–191.

(188) Doğan Daldal, Y.; Çakır, C.; Yılmaz, H.; Demiralay, E. Ç.; Özkan, S. A.; Alsancak, G. Liquid Chromatographic, Spectrophotometric and Potentiometric Pka Determination of Ranitidine and Famotidine. *Curr. Drug ther.* **2015**, *9* (4), 277–284.

(189) Hammett, L. P. Some Relations between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17*, 125–136.

(190) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

(191) Liao, C.; Nicklaus, M. C. Comparison of Nine Programs Predicting pKa Values of Pharmaceutical Substances. *J. Chem. Inf. Model.* **2009**, *49* (9), 2801–2812.

(192) Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H. Estimation of Aqueous Solubility by the General Solubility Equation (GSE) the Easy Way. *QSAR Comb. Sci.* **2003**, *22* (2), 258–262.

(193) Puzyn, T.; Mostrag, A.; Falandysz, J.; Kholod, Y.; Leszczynski, J. Predicting Water Solubility of Congeners: Chloronaphthalenes-A Case Study. *J. Hazard. Mater.* **2009**, *170*, 1014–1022.

(194) Ghasemi, J.; Saaidpour, S. QSPR Prediction of Aqueous Solubility of Drug-like Organic Compounds. *Chem. Pharm. Bull. (Tokyo).* **2007**, *55* (4), 669–674.

(195) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mutchell, J. B. O. Random Forest Models to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(196) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 450–456.

(197) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Model.* **2001**, *41* (5), 1208–1217.

(198) Kubinyi, H. A General View on Similarity and QSAR Studies. In *Computer-Assisted Lead Finding and Optimization*; Waterbeemd, H. van de, Testa, B., Folkers, G., Eds.; Wiley-VHC, 1997; pp 9–28.

(199) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

(200) Randic, M. A New Descriptor for Structure - Property and Structure - Activity Correlations †. *J. Chem. Inf. Comput. Sci.* **2001**, 650–656.

(201) Estradal, E.; Patlewicz, G.; Uriarte, E. From Molecular Graphs to Drugs . A Review on the Use of Topological Indices in Drug Design and Discovery. *Indian J. Chem.* **2003**, *42*, 1315–1329.

(202) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69* (1), 17–20.

(203) Gutman, I.; Ruščić, B.; Trinajstić, N.; C. F. Wilcox, J. Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62* (9), 3399–3405.

(204) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct. Relationships* **1985**, *4* (3), 109–116.

(205) Kier, L. B. Distinguishing Atom Differences in A Molecular Graph Shape Index. *Quant. Struct. Relationships* **1986**, *5* (1), 7–12.

(206) Kier, L. B.; Hall, L. H. The Kappa Indices for Modeling Molecular Shape and Felxibility. In *Topological indices and related descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Amsterdam, 1999; pp 455–489.

(207) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(208) Estrada, E. Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (1), 31–33.

(209) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801–807.

(210) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Model.* **1991**, *31* (1), 76–82.

(211) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(212) Kier, L.; Hall, L. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 792–795.

(213) Balaban, A. T.; Balaban, T.-S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8* (1), 383–397.

(214) Balaban, A. T. Using Real Numbers as Vertex Invariants for Third-Generation Topological Indexes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23–28.

(215) Balaban, A. T. Local versus Global (Le. Atomic versus Molecular) Numerical Modeling of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398–402.

(216) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Chem. Inf. Model.* **1964**, *7* (4), 395–399.

(217) Fujita, T.; Ban, T. Structure-Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters. *J. Med. Chem.* **1971**, *14* (4), 148–152.

(218) Kubinyi, H. Quantitative Structure-Activity Relationships. 2. A Mixed Approach, Based on Hansch and Free-Wilson Analysis. *J. Med. Chem.* **1976**, *19* (5), 587–600.

(219) Devillers, J. No-Free-Lunch Molecular Descriptors in QSAR and QSPR. In *Topological indices and related descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Amsterdam, 2000; pp 1–21.

(220) Cambon, B.; Devillers, J. New Trends in Structure-Biodegradability Relationships. *Quant. Struct. Relationships* **1993**, *12*, 49–56.

(221) Singh, V. K.; Tewari, V. P.; Gupta, D. K.; Srivastava, A. K. Calculation of Heat of Formation :- Molecular Connectivity and IOC-ω Technique, a Comparative Study. *Tetrahedron* **1984**, *40* (15), 2859–2863.

(222) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Model.* **1989**, *29* (3), 225–227.

(223) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discov. Des.* **1998**, *9*, 339–353.

(224) ChemAxon. Fingerprint and descriptor generation - GenerateMD.

(225) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics, Volumes I & II*, 2nd ed.; Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VHC, 2009.

(226) Tanford, C. *Physical Chemistry of Macromolecule*; Wiley: New York, 1961.

(227) Arteca, G. A. Molecular Shape Descriptors. In *Reviews in Computational Chemistry*; VHC Publishers: New York, 1996; Vol. 9, pp 191–253.

(228) Abraham, M. H.; McGowan, J. C. The Use of Characteristic Volumes to Mesaure Cavity Terms in Reversed Phase Liquid Chromatography. **1987**, *23* (4), 243–246.

(229) Cheng, Y. Y.; Yuan, H. Quantitative Study of Electrostatic and Steric Effects on Physicochemical Property and Biological Activity. *J. Mol. Graph. Model.* **2006**, *24* (4), 219–226.

(230) Bhattacharjee, S.; Basak, A. C.; Dasgupta, P. Molecular Property Correlation in Haloethanes with Geometric Volume. *Comput. Chem.* **1992**, *16* (3), 223–228.

(231) Zyrianov, Y. Distribution-Based Descriptors of the Molecular Shape. *J. Chem. Inf. Model.* **2005**, *45* (3), 657–672.

(232) Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *J. Chemom.* **1994**, *8* (4), 263–272.

(233) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/activity and Structure/property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99–109.

(234) Rohrbaugh, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987**, *59*, 1046–1054.

(235) Bondi, A. Van Der Waals Volumes and Radii. *J. Phys. Chem.* **1964**, *68*, 441–451.

(236) Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. 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* (19), 7368–7373.

(237) Meyer, A. Y. Molecular Mechanics and Molecular Shape. Part 1. van Der Waals Descriptors of Simple Molecules. *J. Chem. Soc. Perkin Trans. 2* **1985**, *8*, 1161–1169.

(238) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation , Quantification , and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4–12.

(239) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based upon Molecular Shape Analysis. *J. Am. Chem. Soc.* **1980**, *102* (24), 7196–7206.

(240) Tokarski, J. S.; Hopfinger, A. J. Three-Dimensional Molecular Shape Analysis-Quantitative Structure-Activity Relationship of a Series of Cholecystokinin-A Receptor Antagonists. *J. Med. Chem.* **1994**, *37* (21), 3639–3654.

(241) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110* (12), 5959–5967.

(242) Hahn, M. Receptor Surface Models. 1. Definition and Construction. *J. Med. Chem.* **1995**, *38*, 2080–2090.

(243) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.

(244) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 143–152.

(245) Tuppurainen, K. EEVA (Electronic Eigenvalue): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. *SAR QSAR Environ. Res.* **1999**, *10*, 39–46.

(246) Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Peräkylä, M. Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (3), 607–613.

(247) Liu, S.; Cai, S.; Cao, C.; Li, Z. Molecular Electronegative Distance Vector (MEDV) Related to 15 Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1337–1348.

(248) Liu, S. S.; Yin, C. S.; Cai, S. X.; Li, Z. L. A Novel MHDV Descriptor for Dipeptide QSAR Studies. *J. Chinese Chem. Soc.* **2001**, *48* (2), 253–260.

(249) Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPCA: A New Computational Tool to Design Selective Ligands. *J Med Chem* **2000**, *43* (16), 3033–3044.

(250) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43* (17), 3233–3243.

(251) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: The VolSurf Approach. *J. Mol. Struct. THEOCHEM* **2000**, *503* (1-2), 17–30.

(252) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119* (43), 10509–10524.

(253) Damale, M.; Harke, S.; Kalam Khan, F.; Shinde, D.; Sangshetti, J. Recent Advances in Multidimensional QSAR (4D-6D): A Critical Review. *Mini-Reviews Med. Chem.* **2014**, *14* (1), 35–55.

(254) Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15* (5), 3281–3294.

(255) De Melo, E. B.; Ferreira, M. M. C. Four-Dimensional Structure-Activity Relationship Model to Predict HIV-1 Integrase Strand Transfer Inhibition Using LQTA-QSAR Methodology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1722–1732.

(256) Vedani, A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45* (11), 2139–2149.

(257) Vedani, A.; Dobler, M. MultiDimensional QSAR: Moving from From 3 to 5 Dimensional Concepts. *QSAR Quant. Struct. Relationships Drug Des.* **2002**, *21*, 382–390.

(258) Vedani, A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR.

Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor. *J. Med. Chem.* **2005**, *48* (11), 3700–3703.

(259) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.

(260) Hawkins, P. C. D.; Skillman, A G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools Comparison of Shape-Matching and Docking as Virtual Screening Tools. **2007**, *50*, 74–82.

(261) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48* (5), 1489–1495.

(262) Sauer, W. H. B.; Schwarz, M. K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 987–1003.

(263) Ballester, P. J.; Richards, W. G. Ultrafast Shape Recognition to Search Compound Databases for Similar Molecular Shapes. *J. Comput. Chem.* **2007**, *28* (10), 1711–1723.

(264) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45* (19), 4350–4358.

(265) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity– a Review. *QSAR Comb. Sci.* **2003**, *22* (910), 1006–1026.

(266) Bender, A.; Glen, R. C. Molecular Similarity: A Key Technique in Molecular Informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.

(267) Wilkins, C. L.; Randić, M. A Graph Theoretical Approach to Structure-Property and Structure-Activity Correlations. *Theor. Chim. Acta* **1980**, *58* (1), 45–68.

(268) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Mol. Divers.* **2006**, *10* (1), 39–79.

(269) Mathea, M.; Klingspohn, W.; Baumann, K. Chemoinformatic Classification Methods and Their Applicability Domain. *Mol. Inform.* **2016**, *35* (5), 160–180.

(270) Vogt, M.; Bajorath, J. Chemoinformatics: A View of the Field and Current Trends in Method Development. *Bioorg. Med. Chem.* **2012**, *20* (18), 5317–5323.

(271) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings. *Comb. Chem. High Throughput Screen.* **2002**, *5* (2), 155–166.

(272) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(273) Drab, K.; Daszykowski, M. Clustering in Analytical Chemistry. *J. AOAC Int.* **2014**, *97* (1), 29–38.

(274) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7*, 20.

(275) Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity

Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901.

(276) Yan, X.; Yu, P.; Han, J. Substructure Similarity Search in Graph Databases. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*; Illinois, U. O., Ed.; ACM Press, 2005; pp 766–777.

(277) Klinger, S.; Austin, J. Weighted Superstructures for Chemical Similarity Searching. In *Proceedings of the 9th Joint Conference on Information Sciences*; 2006.

(278) Mestres, J.; Maggiora, G. M. Putting Molecular Similarity into Context: Asymmetric Indices for Field-Based Similarity Measures. *J. Math. Chem.* **2006**, *39* (1), 107–118.

(279) Héberger, K. Sum of Ranking Differences Compares Methods or Models Fairly. *TrAC Trends Anal. Chem.* **2010**, *29* (1), 101–109.

(280) Haws, D. C.; Huggins, P.; O'Neill, E. M.; Weisrock, D. W.; Yoshida, R. A Support Vector Machine Based Test for Incongruence between Sets of Trees in Tree Space. *BMC Bioinformatics* **2012**, *13* (1), 210.

(281) Schuffenhauer, A.; Brown, N. Chemical Diversity and Biological Activity. *Drug Discov. Today Technol.* **2006**, *3* (4), 387–395.

(282) Turner, D. B.; Tyrrell, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 18–22.

(283) Sheridan, R. P.; Kearsley, S. K. Why Do We Need so Many Chemical Similarity Search Methods? *Drug Discov. Today* **2002**, *7* (17), 903–911.

(284) Salim, N.; Holliday, J.; Willett, P. Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 435–442.

(285) Brown, R. D.; Martin, Y. C. An Evaluation of Structural Descriptors and Clustering Methods for Use in Diversity Selection. *SAR QSAR Environ. Res.* **1998**, *8* (1-2), 23–39.

(286) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm For Selecting Sets Of Dissimilar Molecules From Large Chemical Databases. *Quant. Struct. Relationships* **1995**, *14* (6), 501–506.

(287) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (1), 163–166.

(288) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.* **2006**, *46* (2), 462–470.

(289) Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of Similarity Measures for Searching the *Dictionary of Natural Products* Database. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 449–457.

(290) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 819–828.

(291) Chen, X.; Reynolds, C. H. Performance of Similarity Measures in 2D Fragment-Based

Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1407–1414.

(292) Yu, X.; Geer, L. Y.; Han, L.; Bryant, S. H. Target Enhanced 2D Similarity Search by Using Explicit Biological Activity Annotations and Profiles. *J. Cheminform.* **2015**, *7*, 55.

(293) Reisen, F.; Zhang, X.; Gabriel, D.; Selzer, P. Benchmarking of Multivariate Similarity Measures for High-Content Screening Fingerprints in Phenotypic Drug Discovery. *J. Biomol. Screen.* **2013**, *18* (10), 1284–1297.

(294) Wale, N.; Watson, I. A.; Karypis, G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model.* **2008**, *48* (4), 730–741.

(295) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46* (1), 208–220.

(296) Ertl, P. Intuitive Ordering of Scaffolds and Scaffold Similarity Searching Using Scaffold Keys. *J. Chem. Inf. Model.* **2014**, *54* (6), 1617–1622.

(297) Wolohan, P. R. N.; Akella, L. B.; Dorfman, R. J.; Nell, P. G.; Mundt, S. M.; Clark, R. D. Structural Unit Analysis Identifies Lead Series and Facilitates Scaffold Hopping in Combinatorial Chemistry. *J. Chem. Inf. Model.* **2006**, *46* (3), 1188–1193.

(298) Maggiora, G. M. On Outliers and Activity Cliffs--Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46* (4), 1535.

(299) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (8), 3186–3204.

(300) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48* (5), 941–948.

(301) Consonni, V.; Todeschini, R. New Similarity Coefficients for Binary Data. *MATCH Commun. Math. Comput. Chem.* **2012**, *68*, 581−592.

(302) Spiteri, M.; Dubin, E.; Cotton, J.; Poirel, M.; Corman, B.; Jamin, E.; Lees, M.; Rutledge, D. Data Fusion between High Resolution (1)H-NMR and Mass Spectrometry: A Synergetic Approach to Honey Botanical Origin Characterization. *Anal. Bioanal. Chem.* **2016**, *408* (16), 4389–4401.

(303) Bro, R.; Nielsen, H. J.; Savorani, F.; Kjeldahl, K.; Christensen, I. J.; Brünner, N.; Lawaetz, A. J. Data Fusion in Metabolomic Cancer Diagnostics. *Metabolomics* **2013**, *9* (1), 3–8.

(304) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graph. Model.* **1997**, *15* (6), 372–385.

(305) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?*; Kluwer Academic Publishers: Dordrecht, 2000; pp 1–16.

(306) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today* **2006**, *11* (23-24), 1046–1053.

(307) Willett, P. Combination of Similarity Rankings Using Data Fusion. *J. Chem. Inf. Model.* **2013**, *53* (1), 1–10.

(308) Chen, B.; Mueller , C.; Willett, P. Combination Rules for Group Fusion in Similarity-Based Virtual Screening. *Mol. Inform.* **2010**, *29* (6-7), 533–541.

(309) Cross, S.; Baroni, M.; Carosati, E.; Benedetti, P.; Clementi, S. FLAP: GRID Molecular Interaction Fields in Virtual Screening. Validation Using the DUD Data Set. *J. Chem. Inf. Model.* **2010**, *50* (8), 1442–1450.

(310) Kalivas, J. H.; Héberger, K.; Andries, E. Sum of Ranking Differences (SRD) to Ensemble Multivariate Calibration Model Merits for Tuning Parameter Selection and Comparing Calibration Methods. *Anal. Chim. Acta* **2015**, *869*, 21–33.

(311) Tencate, A. J.; Kalivas, J. H.; White, A. J. Fusion Strategies for Selecting Multiple Tuning Parameters for Multivariate Calibration and Other Penalty Based Processes: A Model Updating Application for Pharmaceutical Analysis. *Anal. Chim. Acta* **2016**, *921*, 28–37.

(312) Héberger, K.; Kollár-Hunek, K. Sum of Ranking Differences for Method Discrimination and Its Validation: Comparison of Ranks with Random Numbers. *J. Chemom.* **2011**, *25* (4), 151–158.

(313) Kollár-Hunek, K.; Héberger, K. Method and Model Comparison by Sum of Ranking Differences in Cases of Repeated Observations (Ties). *Chemom. Intell. Lab. Syst.* **2013**, *127*, 139–146.

(314) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the “Early Recognition” Problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488–508.

(315) MacCuish, J. D.; MacCuish, N. E. Chemoinformatics Applications of Cluster Analysis. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4* (1), 34–48.

(316) Zahoránszky-Kőhalmi, G.; Bologa, C. G.; Oprea, T. I. Impact of Similarity Threshold on the Topology of Molecular Similarity Networks and Clustering Outcomes. *J. Cheminform.* **2016**, *8* (1), 16.

(317) Amancio, D. R.; Comin, C. H.; Casanova, D.; Travieso, G.; Bruno, O. M.; Rodrigues, F. A.; da Fontoura Costa, L. A Systematic Comparison of Supervised Classifiers. *PLoS One* **2014**, *9* (4), e94137.

(318) Kireeva, N. V; Ovchinnikova, S. I.; Kuznetsov, S. L.; Kazennov, A. M.; Tsivadze, A. Y. Impact of Distance-Based Metric Learning on Classification and Visualization Model Performance and Structure-Activity Landscapes. *J. Comput. Aided. Mol. Des.* **2014**, *28* (2), 61–73.

(319) Jaskowiak, P. A.; Campello, R. J. G. B.; Costa, I. G. On the Selection of Appropriate Distances for Gene Expression Data Clustering. *BMC Bioinformatics* **2014**, *15 Suppl 2*, S2.

(320) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, *45* (4), 1122–1133.

(321) Zhang, L.; Zhang, D. MetricFusion: Generalized Metric Swarm Learning for Similarity Measure. *Inf. Fusion* **2016**, *30*, 80–90.

(322) Pérez-Nueno, V. I.; Venkatraman, V.; Mavridis, L.; Ritchie, D. W. Detecting Drug Promiscuity Using Gaussian Ensemble Screening. *J. Chem. Inf. Model.* **2012**, *52* (8), 1948–1961.

(323) Carbó, R.; Leyda, L.; Arnau, M. How Similar Is a Molecule to Another? An Electron Density Measure of Similarity between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17* (6), 1185–1189.

(324) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem.* **1987**, *32* (S14), 105–110.

(325) Petke, J. D. Cumulative and Discrete Similarity Analysis of Electrostatic Potentials and Fields. *J. Comput. Chem.* **1993**, *14* (8), 928–933.

(326) Miranda-Quintana, R.-A.; Cruz-Rodes, R.; Codorniu-Hernandez, E.; Batista-Leyva, A. J. Formal Theory of the Comparative Relations: Its Application to the Study of Quantum Similarity and Dissimilarity Measures and Indices. *J. Math. Chem.* **2010**, *47* (4), 1344–1365.

(327) Al-Dabbagh, M.; Salim, N.; Himmat, M.; Ahmed, A.; Saeed, F. A Quantum-Based Similarity Method in Virtual Screening. *Molecules* **2015**, *20* (10), 18107–18127.

(328) Zhang, P.; Agarwal, P.; Obradovic, Z. Computational Drug Repositioning by Ranking and Integrating Multiple Data Sources. In *Machine Learning and Knowledge Discovery in Databases*; Blockeel, H., Kersting, K., Nijssen, S., Železný, F., Eds.; Springer Berlin Heidelberg, 2013; pp 579–594.

(329) Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. TargetHunter: An in Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013**, *15* (2), 395–406.

(330) Janda, J.-O.; Popal, A.; Bauer, J.; Busch, M.; Klocke, M.; Spitzer, W.; Keller, J.; Merkl, R. H2rs: Deducing Evolutionary and Functionally Important Residue Positions by Means of an Entropy and Similarity Based Analysis of Multiple Sequence Alignments. *BMC Bioinformatics* **2014**, *15*, 118.

(331) Fu, G.; Ding, Y.; Seal, A.; Chen, B.; Sun, Y.; Bolton, E. Predicting Drug Target Interactions Using Meta-Path-Based Semantic Network Analysis. *BMC Bioinformatics* **2016**, *17*, 160.

(332) Ghalwash, M. F.; Cao, X. H.; Stojkovic, I.; Obradovic, Z. Structured Feature Selection Using Coordinate Descent Optimization. *BMC Bioinformatics* **2016**, *17*, 158.

(333) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput. Aided. Mol. Des.* **1998**, *12* (5), 471–490.

(334) Dijkman, R.; Dumas, M.; van Dongen, B.; Käärik, R.; Mendling, J. Similarity of Business Process Models: Metrics and Evaluation. *Inf. Syst.* **2011**, *36* (2), 498–516.

(335) Becker, M.; Laue, R. A Comparative Survey of Business Process Similarity Measures. *Comput. Ind.* **2012**, *63* (2), 148–167.

(336) Rosenbaum, L.; Jahn, A.; Dörr, A.; Zell, A. Optimization and Visualization of the Edge Weights in Optimal Assignment Methods for Virtual Screening. *BioData Min.* **2013**, *6* (1), 7.

(337) Mohr, J.; Jain, B.; Sutter, A.; Laak, A. Ter; Steger-Hartmann, T.; Heinrich, N.; Obermayer, K. A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test. *J. Chem. Inf. Model.* **2010**, *50* (10), 1821–1838.

(338) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Netw.* **2005**, *18* (8), 1093–1110.

(339) Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1*; Morgan Kaufmann Publishers Inc., 1995; pp 448–453.

(340) Lin, D. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers, 1998; pp 296–304.

(341) Schlicker, A.; Domingues, F. S.; Rahnenführer, J.; Lengauer, T. A New Measure for Functional Similarity of Gene Products Based on Gene Ontology. *BMC Bioinformatics* **2006**, *7*, 302.

(342) Himmat, M.; Salim, N.; Al-Dabbagh, M. M.; Saeed, F.; Ahmed, A. Adapting Document Similarity Measures for Ligand-Based Virtual Screening. *Molecules* **2016**, *21* (4).

(343) Yu, H. Selective Sampling Techniques for Feedback-Based Data Retrieval. *Data Min. Knowl. Discov.* **2011**, *22* (1-2), 1–30.

(344) Armstrong, M. S.; Finn, P. W.; Morris, G. M.; Richards, W. G. Improving the Accuracy of Ultrafast Ligand-Based Screening: Incorporating Lipophilicity into ElectroShape as an Extra Dimension. *J. Comput. Aided. Mol. Des.* **2011**, *25* (8), 785–790.

(345) Gfeller, D.; Michielin, O.; Zoete, V. Shaping the Interaction Landscape of Bioactive Molecules. *Bioinformatics* **2013**, *29* (23), 3073–3079.

(346) Guha, R.; Van Drie, J. H. Structure--Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48* (3), 646–658.

(347) Sukumar, N.; Krein, M. P.; Prabhu, G.; Bhattacharya, S.; Sen, S. Network Measures for Chemical Library Design. *Drug Dev. Res.* **2014**, *75* (6), 402–411.

(348) Cuissart, B.; Touffet, F.; Crémilleux, B.; Bureau, R.; Rault, S. The Maximum Common Substructure as a Molecular Depiction in a Supervised Classification Context: Experiments in Quantitative Structure/biodegradability Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (5), 1043–1052.

(349) Mendenhall, J.; Meiler, J. Improving Quantitative Structure–activity Relationship Models Using Artificial Neural Networks Trained with Dropout. *J. Comput. Aided. Mol. Des.* **2016**, *30* (2), 177–189.

(350) Naderi, M.; Alvin, C.; Ding, Y.; Mukhopadhyay, S.; Brylinski, M. A Graph-Based Approach to Construct Target-Focused Libraries for Virtual Screening. *J. Cheminform.* **2016**, *8*, 14.

(351) Jankowski, N.; Usowicz, K. Analysis of Feature Weighting Methods Based on Feature Ranking Methods for Classification. In *Neural Information Processing*; Lu, B.-L., Zhang, L., Kwok, J., Eds.; Springer Berlin Heidelberg, 2011; pp 238–247.

(352) Kim, S.; Okajima, R.; Kano, M.; Hasebe, S. Development of Soft-Sensor Using Locally Weighted PLS with Adaptive Similarity Measure. *Chemom. Intell. Lab. Syst.* **2013**, *124*, 43–49.

(353) Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N. A Review of

near Infrared Spectroscopy and Chemometrics in Pharmaceutical Technologies. *J. Pharm. Biomed. Anal.* **2007**, *44* (3), 683–700.

(354) Gan, F.; Hopke, P. K.; Wang, J. A Spectral Similarity Measure Using Bayesian Statistics. *Anal. Chim. Acta* **2009**, *635* (2), 157–161.

(355) Bodis, L.; Ross, A.; Pretsch, E. A Novel Spectra Similarity Measure. *Chemom. Intell. Lab. Syst.* **2007**, *85* (1), 1–8.

(356) de Gelder, R.; Wehrens, R.; Hageman, J. A. A Generalized Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification. *J. Comput. Chem.* **2001**, *22* (3), 273–289.

(357) Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (1), 85–88.

(358) Wu, Y.; Lv, S.; Wang, C.; Gao, X.; Li, J.; Meng, Q. Comparative Analysis of Volatiles Difference of Yunnan Sun-Dried Pu-Erh Green Tea from Different Tea Mountains: Jingmai and Wuliang Mountain by Chemical Fingerprint Similarity Combined with Principal Component Analysis and Cluster Analysis. *Chem. Cent. J.* **2016**, *10*, 11.

(359) Zhang, J.; Koo, I.; Wang, B.; Gao, Q.-W.; Zheng, C.-H.; Zhang, X. A Large Scale Test Dataset to Determine Optimal Retention Index Threshold Based on Three Mass Spectral Similarity Measures. *J. Chromatogr. A* **2012**, *1251*, 188–193.

(360) Varmuza, K.; Karlovits, M.; Demuth, W. Spectral Similarity versus Structural Similarity: Infrared Spectroscopy. *Anal. Chim. Acta* **2003**, *490* (1), 313–324.

(361) Hung, W.-L.; Yang, M.-S. Similarity Measures of Intuitionistic Fuzzy Sets Based on Hausdorff Distance. *Pattern Recognit. Lett.* **2004**, *25* (14), 1603–1611.

(362) Liang, Z.; Shi, P. Similarity Measures on Intuitionistic Fuzzy Sets. *Pattern Recognit. Lett.* **2003**, *24* (15), 2687–2693.

(363) Xu, Z.; Yager, R. R. Intuitionistic and Interval-Valued Intutionistic Fuzzy Preference Relations and Their Measures of Similarity for the Evaluation of Agreement within a Group. *Fuzzy Optim. Decis. Mak.* **2009**, *8* (2), 123–139.

(364) Hung, W.-L.; Yang, M.-S. Similarity Measures of Intuitionistic Fuzzy Sets Based on Lp Metric. *Int. J. Approx. Reason.* **2007**, *46* (1), 120–136.

(365) Szmidt, E.; Kacprzyk, J. A Similarity Measure for Intuitionistic Fuzzy Sets and Its Application in Supporting Medical Diagnostic Reasoning. In *Artificial Intelligence and Soft Computing - ICAISC 2004*; Rutkowski, L., Siekmann, J. H., Tadeusiewicz, R., Zadeh, L. A., Eds.; Springer Berlin Heidelberg, 2004; pp 388–393.

(366) Singh, P. A New Method for Solving Dual Hesitant Fuzzy Assignment Problems with Restrictions Based on Similarity Measure. *Appl. Soft Comput.* **2014**, *24*, 559–571.

(367) Wang, W.; Xin, X. Distance Measure between Intuitionistic Fuzzy Sets. *Pattern Recognit. Lett.* **2005**, *26* (13), 2063–2069.

(368) Zhang, L.; Xu, X.; Tao, L.; Zhang, L.; Xu, X.; Tao, L. Some Similarity Measures for Triangular Fuzzy Number and Their Applications in Multiple Criteria Group Decision-Making. *J. Appl. Math.* **2013**, *2013*, 538261.

(369) Su, Z.; Xu, Z.; Liu, H.; Liu, S. Distance and Similarity Measures for Dual Hesitant Fuzzy

Sets and Their Applications in Pattern Recognition. *J. Intell. Fuzzy Syst.* **2015**, *29* (2), 731–745.

(370) Zhang, X.; Xu, Z. Novel Distance and Similarity Measures on Hesitant Fuzzy Sets with Applications to Clustering Analysis. *J. Intell. Fuzzy Syst.* **2015**, *28* (5), 2279–2296.

(371) Krawczak, M.; Szkatuła, G. On Asymmetric Matching between Sets. *Inf. Sci. (Ny).* **2015**, *312*, 89–103.

(372) Hamers, L.; Hemeryck, Y.; Herweyers, G.; Janssen, M.; Keters, H.; Rousseau, R.; Vanhoutte, A. Similarity Measures in Scientometric Research: The Jaccard Index versus Salton's Cosine Formula. *Inf. Process. Manag. an Int. J.* **1989**, *25* (3), 315–318.

(373) Leng, Y.; Lu, Q.; Liang, C. A Collaborative Filtering Similarity Measure Based on Potential Field. *http://dx.doi.org/10.1108/K-10-2014-0212* **2016**, *45* (3), 434–445.

(374) Common Chemistry http://www.commonchemistry.org/ (accessed Jul 5, 2016).

(375) PubChem https://pubchem.ncbi.nlm.nih.gov/ (accessed Jul 5, 2016).

(376) PubChem BioAssay - NCBI http://www.ncbi.nlm.nih.gov/pcassay (accessed Jan 27, 2016).

(377) ChemSpider | Search and share chemistry http://www.chemspider.com/ (accessed Jul 5, 2016).

(378) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; *et al*. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100–D1107.

(379) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; *et al*. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42* (Database issue), D1083–D1090.

(380) ChEMBL database (release 20) https://www.ebi.ac.uk/chembl/ (accessed Jan 21, 2016).

(381) SureChEMBL https://www.surechembl.org/search/ (accessed Jul 6, 2016).

(382) Irwin, J. J.; Shoichet, B. K. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.

(383) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757–1768.

(384) eMolecules https://www.emolecules.com/ (accessed Jul 6, 2016).

(385) Kiss, R.; Sándor, M.; Szalai, F. A. http://Mcule.com: A Public Web Service for Drug Discovery. *J. Cheminform.* **2012**, *4 (Suppl 1*, 17.

(386) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875.

(387) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; *et al*. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57* (12), 4977–5010.

(388) Rácz, A.; Bajusz, D.; Héberger, K. Consistency of QSAR Models: Correct Split of

Training and Test Sets, Ranking of Models and Performance Parameters. *SAR QSAR Environ. Res.* **2015**, *26* (7-9), 683–700.