Introduction

Since the publication of the first Glossary of Terms Used in Computational Drug Design over 15 years ago the practice of both medicinal chemistry and computational drug design have undergone a rapid and continuous change that has resulted in a considerable expansion of terminology. In addition, medicinal chemists are increasingly required to understand and interpret language that was formerly the predominant domain of computational chemists. To reflect these changes the authors have compiled this supplementary Glossary.
of over 200 terms that were not previously defined or whose meaning has changed somewhat since the first version.

To avoid a repetition of terms included in the original Glossary we have chosen to keep this supplement as a separate document and to identify it by the designation Part II. By inference, therefore, Part I is the earlier Glossary (H. van de Waterbeemd, R. E. Carter, G. Grassy, H. Kubinyi, Y. C. Martin, M. S. Tute and P. Willett. Glossary of terms used in computational drug design (IUPAC Recommendations 1997). Pure Appl. Chem., 1997, Vol. 69, No. 5, pp. 1137–1152. http://dx.doi.org/10.1351/pac199769051137). Those searching for specific terminology are advised to refer to both Glossaries.

Alphabetical entries

1D property descriptor
An observed or calculated property of the whole molecule.
Note: Examples: molar mass or octan-1-ol-water log P.

1D structure
The structure of a molecule encoded into a string such as SMILES [1, 2] or InChI [3].

2D property
A molecular property that is calculated from the structure diagram of the molecule.
Note: Examples include counts of hydrogen bond donors and acceptors, topological polar surface area, topological indices, or a molecular fingerprint.

2D structure
The structure of a molecule presented as a drawing or a file that contains a description of the topology, stereochemistry, and atomic symbol of its atoms and the bonds connecting them, but no explicit information about its three-dimensional structure.

2D substructure searching
See substructure searching.

2D-QSAR (two-dimensional quantitative structure-activity relationships)
A computational model of the quantitative relationship between the observed independent 1D property of a set of compounds and their dependent 1D properties or 2D properties [4].

3D property
A property that is observed or calculated from a 3D representation of a molecule, which may depend on one or more conformations of the molecule.
Note: Examples are: dipole moment, polar surface area, distances between key atoms, or the vector of electrostatic interaction energy calculated at a number of points surrounding the molecule.

3D searching
A virtual screening method that processes a chemical database to discover those compounds that match a query that either contains a 3D pharmacophore, a molecular shape or field, the 3D structure, or a combination of these [5].

3D structure
The structure of the conformation of a molecule presented as a drawing or a file that describes atomic symbols of its atoms and the bonds connecting them as well as their coordinates in three-dimensional space.
3D structure generation
A method to generate one or more 3D structures, conformations, of a molecule from its topological molecular graph.

3D-QSAR (three-dimensional quantitative structure-activity relationships)
A computational model of the quantitative relationship between the target observed independent 1D properties of a set of compounds and their 3D properties calculated from a single conformation [6–8]. Modified from [9].

4D-QSAR (four-dimensional quantitative structure-activity relationships)
A computational model of the quantitative relationship between the target observed independent 1D properties for a set of compounds and the 3D properties of several of their conformations [10].

algorithm
A step-by-step procedure for solving a problem or performing a function, usually by a computer.

applicability domain
The property or structure space for which the predictions of a computational model are considered to be reliable [11].

area under the curve (AUC)
The area under a graph curve, such as the number of actives identified by an algorithm as a function of the number of compounds tested, commonly used as a measure of the discriminating ability of an algorithm to correctly classify a test molecule.
Note 1: See also ROC Curve.
Note 2: In pharmacokinetics AUC is the area under the plot of plasma concentration versus time, which is used to evaluate drug exposure.

autocorrelation vector
A vector that describes a molecular structure in which each element corresponds to a distance (number of bonds in a 2D structure or a binned interatomic distance in a 3D structure) between atoms of a particular type and the count of the number of times that distance is found in the structure [12].
Note: For example, a simple autocorrelation vector of a 2D structure might consist of elements corresponding to seven distances (1–7 bond distances) and seven types of atom pairs (C–C, C–O, C–N, C–other, O–N, O–O, N–N).

basis function
A one-electron function used in the expansion of the molecular orbital function. Basis functions are commonly represented by atomic orbitals centered on each atom of the molecule [13].

basis set
In quantum chemistry, a set of basis functions employed for the representation molecular orbitals [13].

basis set superposition error (BSSE)
An artfactual increase in calculated stability of the supersystem (the system formed by noncovalent interaction between two or more molecular entities, e.g. hydrogen bond system) resulting from the basis set of the supersystem being larger than for the component subsystems. The BSSE arises from a lowering of the quantum mechanical energy when the electron density of each subsystem spreads into the basis functions provided by the other subsystems [13].
**Bayesian classifier**
A largely *supervised learning* classification algorithm that classifies an object such as a chemical structure using the relative frequency of each of the object’s properties in the various classes. If it assumes the features are independent it is called a naïve Bayesian classifier. The classifier aims to minimize the probability of misclassification [14].

**Bayesian regularized neural network**
A feed-forward neural network that uses *Bayesian statistics* to optimize the complexity and predictive power of the model [15].

**Bayesian statistics**
A branch of statistics in which the evidence of the state of a system is expressed in terms of degrees of belief (probabilities) [16].

**Bayes’s theorem**
A method for calculating *prior probability* estimates of an event that can be revised in accordance with new observations.

**belief theory**
A method to combine probability estimate states that given two or more probabilities $P_j$ that a particular event is true, the combined probability of the event is given by [17]:

$$P = 1 - \Pi (1 - P_j)$$

**bilinear equation**
A QSAR equation that describes the non-linear dependence of the relative biological potency (log 1/$C_i$) of a molecule on log $P$ by the following form [18]:

$$\log(1/C_i) = a \log P_i - b \log(\beta P_i + 1) + c$$

Note: See also *Hansch equation*, which is also a nonlinear equation in log $P$.

**bit string, bitmap**
A description of a molecule in a fixed length vector, each element of which is set from corresponding *structural keys* or calculated by hashing a molecular fingerprint.

**Boltzmann enhanced discrimination of receiver operating characteristic**
A generalization of the area under the ROC curve to weight more heavily the early recognition of active compounds [19].

**bootstrap resampling**
A procedure to evaluate the accuracy of the statistics of a model, such as the overall $R^2$ or the contribution of particular properties, by systematically recomputing the statistics generated from many models developed using sample sets that contain the same number of observations as the original set, but for which certain observations are randomly omitted and other observations are included more than once.

See also: *cross-validation* and *jackknifing*.

**calculated molar refractivity (CMR)**
The calculated molar refractivity of a molecule or substituent used as a measure of size and polarizability in 2D-QSAR.
canonical structure representation
A unique representation of the chemical structure of a molecule that is independent of the sequence in which the atoms were ordered in the original input file but is dependent on the algorithm used for the canonicalization.
Note 1: This representation is used to ensure the uniqueness of molecules in a database and to assist identification of molecules in internet searches.
Note 2: At least one canonicalization algorithm exists for any type of structure file that is the basis of a structure search system.

chance correlation
An artificial correlation that can arise when too many properties are screened relative to the number of available observations [20].
Note: For example, if one tests 20 possible random descriptors for statistical significance in a multiple regression equation of the properties of 15 compounds, the average fitted $R^2$ is 0.81 even though an average of only four descriptors were included in the equation.

chemical fingerprints
See molecular fingerprints.

cheminformatics (or chemoinformatics)
The science of handling, indexing, archiving, searching, and evaluating information that is specific to chemical structures and is used in data mining, information retrieval, information extraction, and machine learning.

circular fingerprints
Hashed molecular fingerprints that describe, either by atom type or properties, each atom in the molecule; the atoms connected to it (for path length 2); the atoms connected to them (for path length 4); and the atoms connected to them (for path length 6); etc. [21].
Note: See also path fingerprints and structural keys.

classification
The discovery or application of a rule set that uses descriptors to assign objects such as chemical structures to one of several classes, such as mutagenic/non-mutagenic or active/inactive [22].
See recursive partitioning and Bayesian classifier.

clog $P$
Calculated log of the octan-1-ol/water partition coefficient.
Modified from [23].

cluster analysis
A procedure that partitions large data sets into distinct groups each of which contains objects, e.g. chemical structures, with similar properties but that are different from the properties of members of the other groups [24].
Modified from [9].

cluster centroid
The geometric center of a cluster, often exemplified as the object that is closest to the center.

collinearity
A linear relationship between two or more of the descriptors in a model, which can lead to difficulties in interpreting the relative importance of these descriptors.
comparative molecular field analysis (CoMFA)
A 3D-QSAR method that models the quantitative relationship between the biological activity of a set of compounds and 3D properties calculated by sampling interaction energies with probe atoms located on a lattice around their aligned three-dimensional structures [25]. Modified from [9].

comparative molecular similarity analysis (CoMSIA)
A 3D-QSAR method similar to CoMFA that instead uses Gaussian-type functions to calculate the 3D properties of the molecules by the similarity of each molecule to each probe at each lattice position surrounding the molecules [26].

component loading
See principal component loading.

component score
See principal component score.

concordance
The ability of a two-group classification model to correctly classify all molecules.

confusion matrix
A table layout that shows the results of a two-class/binary classifier in which columns correspond to the predicted classifications from the model and rows correspond to the observed classifications.
Note: A confusion matrix highlights the numbers of true positives TP (positives classified as positive), true negatives TN (negatives classified as negative), false positives FP (negatives classified as positive) and false negatives FN (positives classified as negatives).

continuum
The computational representation of a solvent as a continuous medium instead of individual explicit solvent molecules.
Note: See also explicit solvent, implicit solvent, Poisson–Boltzmann equation.

correlation coefficient, \( r \)
A measure of the degree of the interrelationship which exists between two measured quantities, \( x \) and \( y \) [27].
Note 1: In computational drug design this is frequently used to describe the relationship between a dependent biological variable \( y \) (e.g. a log \( K_i \) of binding) and one or more independent variable(s) \( x \). It ranges from 0 for no relationship to -1.0 for a perfect negative correlation or +1.0 for a perfect positive correlation.
Note 2: The Pearson correlation coefficient describes the relationship between two continuous variables whereas the Spearman correlation coefficient describes the relationship between the continuous dependent variable and the rank order of the predictor variable.
Note 3: See also \( R^2 \).

counterpoise correction
A method to correct for basis set superposition error.

cross-validation
A measure of the robustness of a computational model that is obtained by successively omitting a subset of the molecules of the original training set and forecasting their target property from a revised model, repeat-
ing the process a number of times, and using the difference between the observed and predicted values as a measure of the model predictivity.

Note: See also leave-one-out, leave-some-out, bootstrap resampling and jackknifing.

data mining
The extraction of useful information from large sets of observations.

Daylight fingerprints
Path fingerprints that are generated from a structure where atoms are described by atomic number and aromaticity; bonds are described as single, double, triple, or aromatic; and the results are hashed into a bit string of fixed length [28].

Note: Although these fingerprints were originally described by Daylight CIS, many other investigators have also implemented them.

de novo design
The process whereby a computer algorithm designs new molecules to meet certain single or multiple criteria, defined by an objective function, for example predicted affinity for a binding site or predicted affinity plus appropriate lipophilicity.

Modified from [9].

decision tree
The result of a recursive partitioning classification method that shows at each branching point the value of the property responsible for the split and the resulting classifications after the split.

density functional theory
A quantum mechanical modeling method used to investigate the electronic structure of molecules using functionals that describe the spatially dependent electron density, rather than the wavefunction as in ab initio and semi-empirical quantum chemical methods.

Note: See also [13].

descriptor (in computational drug design)
A qualitative or quantitative property of a molecule or a part of a molecule.

docking
Computational methods that optimize the placement of a ligand in a macromolecular binding site of known or proposed 3D structure and provide a score as to the quality of the fit [29].

Note. Docking programs differ as to whether conformational and chemical (tautomer, protoner) changes in the binding site are allowed and if only precalculated conformations of the ligand are used or if ligand flexibility is part of the docking.

Modified from [9].

drug-likeness
An estimate of the probability that a chemical structure is similar to known drugs based on chemical or physical properties [30].

electrotopological state descriptors
See topological indices.

energy of the highest occupied molecular orbital ($E_{HOMO}$)
The energy of the highest energy level in the ground state of a molecule that contains electrons, sometimes used as a measure of nucleophilicity in QSAR models.
energy of the lowest unoccupied molecular orbital (E\text{LUMO})
The energy of the lowest energy level of a molecule that contains no electrons in the ground state, sometimes used as a measure of electrophilicity in QSAR models.
Note: Also known as the energy of the lowest empty molecular orbital, E\text{LEMO}.

enrichment factor
The concentration of actives among the top-scoring virtual screening hits; for example, the fraction of true actives that are retrieved in the top 1% scoring molecules compared to the fraction of true actives in the whole database.
Note: See also precision and recall.

expert system
A system that predicts properties or activities of chemicals from rules devised by domain experts.
Note: Examples are the CLOGP program [31] to predict log \( P \) or the Derek Nexus system [32] to predict toxicity.

explicit solvent
Individual solvent molecules that are included in a computation, in contrast to those implied by a continuum approximation.
Note: See also continuum.

extended connectivity fingerprints (ECFP fingerprints)
Hashed circular fingerprints that encode each atom according to its atomic symbol and hybridization for path length 2, ECFP2; path length 4, ECFP4; and path length 6, ECFP6 [33].

false negative (FN)
A model prediction of a negative result, such as no biological activity, when the true result is positive, i.e. activity.
Note: See also confusion matrix.

false positive (FP)
A model prediction of a positive result, such as biological activity, when the true result is negative, i.e. inactivity.
Note: See also confusion matrix.

force field
A set of mathematical functions and their associated parameters used in a molecular mechanics or dynamics calculation of conformations, flexibility, and interactions of molecules [34].
Note: Within the molecular mechanics approach, a set of potential functions defining bond stretch, bond angle (both valence and dihedral) distortion energy of a molecule as compared with its nonstrained conformation (that characterized by standard values of bond lengths and angles). A set of transferable empirical force constants is preassigned and the harmonic approximation is usually employed. Some force fields may contain terms for interactions between non-bonded atoms, electrostatic, hydrogen bond and other structural effects as well as account for anharmonicity effects. In vibrational spectroscopy, the inverse problem is solved of determining a set of force constants and other parameters of a chosen potential energy functions which would match with experimentally observed vibrational frequencies of a given series of congeneric molecules [13].
Modified from [9].
fragment keys
Description of the structure of a molecule as the presence or absence of each of a pre-determined set of molecular substructures, usually associated with a particular location in a bit string.
Note: An example is the publically described ISIS keys [35].

free energy perturbation
A method that uses molecular dynamics or Metropolis Monte Carlo simulations to compute the relative free energy differences between two conditions, typically the solution and bound states of a ligand-protein system [36]. Modified from [9].

functional class fingerprints (FCFP)
Hashed circular fingerprints that encode each atom according to its property (hydrophobic, hydrogen bond donor, hydrogen bond acceptor, negatively charged, or positively charged) for path length 2, FCFP2; path length 4, FCFP4; and path length 6, FCFP6 [33].

global model
A computational model designed to cover all of chemistry space although it typically includes a measure of if the molecule for which the properties are to be predicted is within or outside the applicability domain of the model.

graph-based method
A computational method that converts the problem to be solved into a graph characterized by the character of the objects, such the atoms of a molecule or molecules in a database, and the distances between them [37]. Note: Examples are the Ullman algorithm used in substructure searching [38] or the detection of a pharmacophore or a maximum common substructure (MCS) within a set of molecules using the Bron–Kerbosch algorithm.

Hammett equation
The equation that describes the relationship between the logarithm of the relative equilibrium or rate constant for a chemical reaction and the electronic properties of the substituents on the molecules [27, 39]:

$$\log \left( \frac{K_i}{K_o} \right) = \rho \sigma_i + X$$

in which $\rho$ describes the sensitivity of a reaction to electronic effects and $\sigma_i$ describes the electronic effect of the substituent.
Note: For the reference system benzoic acid, the value of $K_o$ is the $K_a$ of unsubstituted benzoic acid and $\rho$ is 1.00.

Hansch equation
A QSAR equation that describes the relationship between the logarithm of the relative biological potency of a molecule (log 1/C) and its electronic and hydrophobic properties:

$$\log \left( \frac{1}{C_i} \right) = \rho \sigma_i + b \log P - c (\log P)^2$$

in which fitting to the square term provides a parabola with an associated optimum log $P$ [40].
Note: See also bilinear equation.

Hansch–Fujita π-constant
A constant that describes the contribution of a substituent to the octan-1-ol-water log $P$ of a compound [41]. Modified from [9].
hashing
An algorithm that maps data of variable length, such as the raw descriptors of path or circular fingerprints or InChIs, to a smaller fixed length vector.
Note: Hashed InChIs are named InChIKeys.

implicit solvent
See continuum.

InChI™
See international chemical identifier.

InChIKey
A fixed length (character) condensed digital representation of the InChI that is not human-understandable.
Note: this is sometimes referred to as a hashed InChI.

International chemical identifier (InChI™)
A non-proprietary canonical identifier of chemical substances that can be used in printed and electronic data sources to enable easier linking of diverse data compilations [3].
Note: Examples are chloroacetic acid, “InChI=1/C2H3ClO2/c3-1-2(4)5/h1H2,(H,4,5)/p-1”; and 2-methylpyridine, “InChI=1S/C6H7N/c1-6-4-2-3-5-7-6/h2-5H,1H3”.

intrinsic water (solvent)
See explicit solvent.

ISIS keys
Predefined fragment keys that are used in 2D substructure and similarity searching with the ISIS software but are also generated by many other software programs and used for 2D-QSAR or cluster analysis of molecules [42].

jackknifing
A procedure to evaluate the accuracy of the statistics of a model, such as the overall $R^2$ or the contribution of particular properties, by systematically recomputing the statistics generated from models developed by leaving out one or more observations at a time from the sample set.
Note: See also cross-validation and bootstrap resampling.

k-nearest neighbor (kNN)
A classification or quantitative method that predicts the property, such as biological activity, of a molecule based on the property of $k$ (usually 3, but can be larger) most similar molecules, sometimes weighted so that the most similar molecules contribute more to the prediction [43].

kappa shape descriptor
A type of topological index.

kernel method
Algorithms for pattern analysis that use functions such as Gaussians to enable them to operate in a high-dimensional descriptor space without ever computing the coordinates of the data in that space.
Note: The support vector machine (SVM) is the most widely used kernel method [44, 45].

Kohonen map or Kohonen neural net
See self-organizing map, SOM.
leave-one-out cross-validation (LOO)
A special case of cross-validation in which each observation is left out once and only once. The difference between the observed property and that predicted when the observation is omitted from the fit is used to calculate $q^2$, the formal equivalent of the squared correlation coefficient, $R^2$.

leave-some-out cross-validation (LSO)
A form of cross-validation in which more than one observation is left out at each run. Typically many runs with random samplings are performed.

leverage
The influence of an observation on the coefficients of a fitted equation [46].
Note: An observation with high leverage may be an outlier or in poorly explored property space.

ligand efficiency (LE)
Measure of the free energy of binding per heavy atom count (i.e. non-hydrogen) of a molecule [23].
Note 1: It is used to rank the quality of molecules in drug discovery, particularly in fragment-based lead discovery.
Note 2: An LE value of 1.25 kJ mol per non-hydrogen atom is the minimum requirement of a good lead or fragment.

linear interaction energy (LIE)
A method that forecasts ligand binding free energies using force field estimations of the receptor-ligand interactions and thermal conformational sampling [47].

Lipinski rules
See rule of five.

lipophilic ligand efficiency (LLE)
A measure of the efficiency of ligand binding that is corrected for binding driven by lipophilicity.
Note: It is calculated as [48]:

$$LLE = pK_a - \log P$$

values of seven or greater are considered characteristic of a good clinical candidate.

loading of a property
The contribution of the property to a principal component or partial least squares latent variable.

local model
A computational model that applies to only a subset of molecules, typically a closely related series.
Note: See also global model.

log $D$
The logarithm of the ratio of the total concentration (neutral and charged species) of a compound in a nonpolar phase, traditionally water-saturated octan-1-ol, to that in water at a given pH.

log $P$
The logarithm of the ratio of the concentration of the uncharged form of a compound in a nonpolar phase, traditionally water-saturated octan-1-ol, to that in water.
Note: Common algorithms to calculate log $P$ are CLOGP and ALOGP.
**machine learning**
A computer algorithm that generates empirical models, such as a model of biological potency as a function of the properties of the molecules, that is derived from the analysis of a training set for which all the necessary data are available.

Note: Examples include artificial neural networks, recursive partitioning, support vector machines, and clustering.

**macromolecular Crystallographic Information File (mmCIF)**
A file that describes in detail the features of a macromolecular structure and the x-ray diffraction experiment that was used to derive that structure [49].

**Markush structure**
A structure diagram that describes a set of related molecules with the structure of the common core and the structures of the substituents at each position.

Note 1: Markush structures are commonly used in patent claims.
Note 2: See also [23].

**molecular connectivity indices**
See topological indices.

**molecular diversity**
A measure of the spread of various properties or chemotypes within a set of compounds.

**molecular fingerprints**
Descriptions of the structure of a molecule calculated from the properties of each of its atoms and bonds that are then usually condensed into a fixed-length string by a hashing algorithm [50].

Note: See also circular fingerprints, Daylight fingerprints, extended connectivity fingerprints, functional class fingerprints, and path fingerprints.

**molecular graph**
The graph with differently labeled (colored) vertices which represent different kinds of atoms and differently labeled (colored) edges related to different types of bonds [13].

Note: For 3D structures the edges are the distances between the atoms [51].

**molecular mechanics**
See force field.

**molecular similarity**
The degree to which two molecules resemble one another as calculated from their respective 2D or 3D properties, molecular fingerprints, fragment keys, or superimposed 3D structures that usually ranges from 1 (identical) to 0 (dissimilar) [52].

Note: Examples include Tanimoto or Tversky similarities for 2D structures and Carbo or Hodgkin for 3D structures.

Modified from [9].

**multidimensional scaling (MDS)**
A visualization method that converts a dataset of distances between molecules in property space into coordinates in lower dimension space with emphasis on preserving the relative distances between dissimilar molecules [53].
multiobjective optimization (MOO)
The search for those combinations of properties that produces the best compromise between the various objectives of the search, for example properties of compounds that are the best trade-off between potency, bioavailability, and selectivity.
Note: See also Pareto front and Pareto optimization.

normalization
A technique for putting various molecular descriptors on a common scale that involves subtracting the population mean of that descriptor from an individual raw score and then dividing the difference by the population standard deviation of the values for that descriptor.
Note: See also [13].

Orthogonal projections to latent structures (O-PLS)
A preprocessing method for partial least squares calculations that filters out variation that is not correlated to the property to be fit and so results in improved interpretability of a PLS model while maintaining its predictivity [54].

overfitting
Deriving a statistical model that is more complex than the underlying data allows with the result that model predictivity is degraded.

er overruning
Training a machine learning model for so long that the original input data is memorized with the result that model predictivity is degraded.

Pareto front
The various combinations of predictor properties that lead to optimum predicted compromises in multiple target properties.

Pareto optimization
The search for solutions to functions that provide a compromise between several desirable outcomes, such as potency, lack of toxicity, and good ADME properties [55].

partial atomic charges
Charges assigned to atoms, arising from the uneven distribution of electrons in bonds, that are used to calculate intermolecular interaction energy or to characterize the charge distribution of a molecule.
Note 1: The charges can be derived from quantum chemical population analyses, fitted dipole moments or electrostatic potentials, spectroscopic data, or from partitioning electron density.
Note 2: There is no firm theoretical basis for the assignment of partial atomic charges, hence there is no “correct” one.

partial least squares, projection to latent structures (PLS)
A multivariate analysis method that is especially suitable for developing a regression model when there are more predictors than observations and/or the predictors are correlated with each other [56].
Note: PLS operates by projecting the target and calculated properties into a new space and successively extracting orthogonal (latent) variables that relate the target to the calculated properties, using cross-validation to select the number of latent variables to include in the model.
Modified from [9].
path fingerprints
Vectors of Boolean arrays generated from the properties of the atoms and bonds in linear paths, typically two to seven bonds long, hashed into an integer of fixed determined length [50].
Note: See also circular fingerprints, Daylight fingerprints, and structure keys.

PDB file
A text file that stores the atomic coordinates of a macromolecule, usually a protein or nucleic acid with associated ligands, solvent molecules and ions [49].
Note: The documentation for the file format is at http://www.wwpdb.org/documentation/format33/v3.3.html. Modified from [9].

pharmacophore
A proposal for the ensemble of steric and electronic features that define the optimal supermolecular intermolecular interaction of a ligand with a specific biological target structure with the result that it triggers or blocks its biological response [57].

pipelining programs
Visual programming of the computational execution of sequential operations on a dataset, with each node specifying an operation with options, and the nodes processed sequentially [58].
Note: Examples are KNIME and pipeline pilot.

PLS
See Partial Least Squares.

PLS latent variable
One of the vectors of a linear combination of the input descriptors that a partial least squares calculation extracts from a dataset.

PLS loading
The contribution of a particular property to a PLS latent variable.

Poisson–Boltzmann equation
A differential equation that uses a mean-field, continuum model to describe the electrostatic interactions in ionic solutions where both ions and solvent are treated implicitly, resulting in less computational effort to describe biomolecules in ionic aqueous solutions.
Note: See also continuum.

polar surface area (PSA)
Surface area over all polar atoms (usually oxygen and nitrogen), including any attached hydrogen atoms, of a molecule [23].
Note 1: Polar surface area is a commonly used metric (c.f. molecular descriptor) for the optimisation of cell permeability. Molecules with a PSA of greater than 1.4 nm² are usually poor at permeating cell membranes. For molecules to penetrate the blood–brain barrier, the polar surface area should normally be smaller than 0.6 nm², although values up to 0.9 nm² can be tolerated [23].
Note 2: Sometimes sulfur atoms are included in the definition.
Note 3: The value of PSA depends on the conformation of the molecule and also the algorithm and atomic radii used in the calculation.
Note 4: See also topological polar surface area, TPSA, the 2D approximation of PSA.
pose diagram
A 2D diagram that shows the interactions between a ligand and a protein in the 3D structure of the complex [59].
Note: LIGPLOT is frequently used for this purpose.

posterior probability
The conditional probability that is assigned to a quantity after relevant evidence is taken into account. The prior and posterior probabilities are linked by the (normalized) likelihood function (Bayes Theorem).
Note: See also prior probability and Bayes theorem.

potential energy surface (PES)
A function that gives the potential energy of a chemical system, usually a molecule but could be a reacting system, as a function of the coordinates of the nuclei due to bond stretches, angle bends, and bond rotation.

precision
In virtual screening, the fraction of actives in the top-scoring hits, such as the fraction of actives in the top 1% scoring molecules.
Note: See also recall, sensitivity, and specificity.

prediction set
Molecules that have been set aside from model development to test the reliability of the derived model.
Note: See also test set.

principal component loading
The contribution, scaled between –1.0 and 1.0, of an original variable to a particular principal component.

principal component score
A property of a molecule calculated from a principal components analysis of the original properties and the value of these properties for the molecule.

principal components analysis (PCA)
A variable reduction method that operates on the correlation matrix of the variables to construct a small set of new orthogonal, i.e. non-correlated, variables (principal components) derived from linear combinations of the original variables [60].
Modified from [9].

prior probability
The conditional probability that is assigned to a quantity before relevant evidence is taken into account, for example, the prior probability of an active compound would be the fraction of the compounds in the dataset that are active.
Note: The prior and posterior probabilities are linked by the (normalized) likelihood function (Bayes theorem).

Protein Data Bank (PDB)
An archive of freely available macromolecular structural data of proteins, nucleic acids, and complex assemblies.
Note 1: The PDB is maintained by the Worldwide Protein Data Bank (wwPDB) with members RSCB PDB (USA), PDBe (Europe), PDBj (Japan), and BMRB (USA) that act as centers for deposition, data processing and distribution of PDB data.
Note 2: The url for the organization is http://www.wwpdb.org/.
Modified from [9] and [23].
The square of the leave-one-out cross-validation regression coefficient calculated from the variance of the observed versus predicted values of the target property, usually potency, of each compound when eliminated during cross-validation.

Note: This is related to $r^2$, which is calculated as the fit of the observed to calculated for the whole dataset.

Note: See also leave-one-out cross-validation.

**QM/MM**
Methods that apply quantum mechanics to a (central) part of the system such as an enzyme active site or the bonds that change during a reaction and simultaneously apply molecular mechanics to another part (environment) [61].

**Quantitative structure–activity relationships (QSAR)**
An equation or other function that describes the relationship between a biological property of compounds, usually a measure of relative potency, and one or more properties of the compounds.

See Hansch equation and bilinear equation.

**$R^2$ or $r^2$**
The ratio of the sum of squares explained by a regression model (SSR) to the “total” sum of squares around the mean (SST), or, because SST equals SSR plus the sum of squares of the error or residuals from the fit (SSE):

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Note 1: $R^2$ values can range between 0.0 and 1.0.

Note 2: Except for the comparison of different models for the same dataset, $R^2$ values are not a good indication of fit of a model because the calculation depends on SST, which is larger if there is more spread in the observed values.

**random forest method**
A classification method that produces many recursive partitioning models, each with a random selection of predictor properties, and then combines the models for predictions [62].

**recall**
The fraction of the total number of actives, in the top-scoring docking hits, such as the fraction of true actives that are retrieved in the top 1% scoring molecules.

Note: See also precision, sensitivity, and specificity.

**receiver-operator characteristic curve (ROC curve)**
A plot that shows the result of a test of the performance of a binary classifier by plotting the fraction of true positives identified versus the fraction of false positives as the discrimination threshold is varied [19].

Note 1: ROC curves are frequently used to compare different docking/scoring combinations or virtual screening protocols by the retrieval results from seeding a database of inactive molecules with known actives.

Note 2: The area under the ROC curve (AUC) provides a measure of the superiority of the model over random predictions: If the value for the AUC for a ROC curve has the value of 0.9–1.0, the fit is considered excellent, whereas an AUC of 0.5 suggests that the algorithm has no discriminatory power.

Note 3: See also Boltzmann enhanced discrimination of receiver operating characteristic (BEDROC).

Note 4: This is an example of using area under the curve as a measure of model performance.
recursive partitioning
A classification method that uses predictor properties to successively divide the data set into subsets such that each resulting subset is enriched in molecules that are in one class [63].

regularization
Introducing additional information in order to solve an ill-posed problem or to prevent overfitting, usually taking the form of a penalty for complexity of the model.

response scrambling
Evaluation of the robustness of a QSAR or classification model by repeatedly randomizing the target property of compounds, developing models based on this randomized property, and comparing the statistics of fit of the scrambled models with those of the true model, which should be superior.

R-group decomposition
The process whereby the molecules in a dataset are partitioned into the core ("scaffold"), usually user-specified, and the specific substituents ("R-groups") found at each specific position.

ridge regression
A regularized statistical method to fit a model that takes into account the correlations between some of the predictors.

RMSE
See root-mean-square error.

ROC curve
See receiver-operator characteristic curve.

root-mean-square deviation (RMSD)
See standard errors of estimates.

root-mean-square error (RMSE)
See standard errors of estimates.

rule of five
The rule of five states that molecules that violate two or more of the following rules are likely to have permeability problems: (1) CLOGP calculated octan-1-ol-water log $P$ greater than 5.0; (2) molecular weight greater than 500; (3) more than five hydrogen bond donors; and (4) the sum of oxygen and nitrogen atoms is greater than 10 [64].

Note 1: Natural products, peptides and other substrates for biological transporters are exceptions.
Note 2: The original authors also noted that one can calculate the octan-1-ol-water log $P$ with the program MLOGP for which the cut-off is 4.15. Users often use log $P$'s calculated with other programs or measured values.
Note 3: Modified from [23].

scoring
A mathematical formula that estimates the binding affinity of a ligand to a macromolecular target based on the structure of the complexes.

Note 1: Scoring is used to select poses and to rank compounds in docking.
Note 2: See also docking.
SDF file (SDfile)
A computer text file that follows a specified format in which each molecule is described by 2D or 3D coordinates, atom type, and connectivity to other atoms and also specific user-definable fields that contain other information such as name and biological or chemical properties [65].
Note: The format of the files is described in the following document: http://download.accelrys.com/free-ware/ctfile-formats/ctfile-formats.zip.

self-organizing map (SOM)
A type of artificial neural network that uses unsupervised machine learning to project high-dimensional data into two dimensions that are usually presented as a contour plots [66].
Note: This is sometimes called a Kohonen map.

semi-empirical quantum chemical methods
Methods that use a variety of parameterizations that allow the user to perform quantum chemical calculations on large molecules or many molecules in a reasonable length of time.
Note: See also [13].
Modified from [9].

sensitivity analysis
Analysis, often by finding approximate derivatives of the output with respect to each input, that determines which of the descriptors in a model has the greatest influence on the output.

sensitivity of a two-class model
The ability of a two-class model to detect active or toxic molecules.
Note: See also specificity, precision, and recall.

Similarity ensemble approach (SEA)
A method that predicts the probability that a molecule or set of molecules will be active against a particular target by considering its pairwise similarities to all known ligands for that target [67].

similarity searching
A virtual screening method that calculates the molecular similarity of an input 2D or 3D structure to each of the molecules in a database to identify a requested number of most similar molecules or those above some similarity threshold [68].
Note: See also molecular similarity, Tanimoto similarity, and Tversky similarity.

SMARTS
An expansion of the SMILES language that describes a particular specific or generalized substructure [69]. For example: the SMARTS that describes any aliphatic ester of a carboxylic acid is “C(=O)OC”; the SMARTS that describes any aromatic ester is “C(=O)OC”; and two SMARTS that describes both are “C(=O)O[c,C]” and “C(=O)[#6]”.

SMILES (simplified molecular input line entry system)
A chemical language that describes molecules in a string of ASCII characters that completely specifies the structure of a molecule as a hydrogen-suppressed graph with nodes as atoms and edges as bonds, parentheses to indicate branching points, lower case to describe aromatic atoms, and numbers to designate ring connection points [1, 2].
Note 1: Examples are chloroacetic acid, “ClCC(=O)O”; and 2-methylpyridine, “Cc1ncccc1.”
Note 2: The SMILES of a molecule is an example of a 1D structure.
Modified from [9] and [23].
SMIRKS
A derivative of the SMILES language that describes the transformation in a chemical reaction [69].
Note: For example the hydrolysis of methyl formate is written “C(=O)OC > HOH > C(=O)OOC.”

specificity of a two-class model
The ability of a two-class model to detect known inactive or non-toxic molecules.
Note: See also sensitivity, precision, and recall.

standard error of estimates (SEE)
The standard error of the errors or residuals, observed minus calculated, from a computational model:

$$\text{SEE} = \sqrt{\frac{\sum (Y_o - Y_c)^2}{n}}$$

in which $Y_o$ is the observed value of $Y$, $Y_c$ is the calculated value of $Y$, and $n$ is the number of observations.
Note 1: Sometimes also labeled RMSD.
Note 2: See also standard errors of predictions.

standard error of prediction (SEP)
The standard deviation of the errors or residuals, observed minus calculated, from a computational model:

$$\text{SEP} = \sqrt{\frac{\sum (Y_o - Y_p)^2}{n-k}}$$

in which $Y_o$ is the observed value of $Y$, $Y_p$ is the predicted value of $Y$, $n$ is the number of observations, and $k$ is the number of terms in the model.
Note 1: Sometimes also labeled RMSEP.
Note 2: See also standard error of estimates.

structural keys
A pre-established set of substructures the presence or absence of which are used to describe a molecule and are used during substructure searching as a filter to eliminate molecules that cannot match a query, for clustering or similarity searching, or for developing classification or regression models [50].
Note 1: ISIS keys is an example.
Note 2: See also molecular fingerprints.

structure diagram
A 2D drawing that shows the atoms of a molecular structure and the bonds that connect them.

substructure searching
A virtual screening method that uses a graph-based method to discover which molecules in a chemical structure database contain the substructure specified in the query [70].

supervised learning
A machine learning method that aims to discover a relationship between the predictor and the response values, such as exists in a QSAR or recursive partitioning model.

support vector machine (SVM)
A supervised learning technique, applicable to both classification and regression, that non-linearly maps the input property space into a very high dimensional feature space in which it either constructs an optimal separating hyperplane for classification or performs linear regression without penalizing small errors [66].
Tanimoto similarity
A scale that ranges from 0.0 to 1.0 when calculated as the ratio of the fingerprint bits or structural keys that both molecules have set divided by the number of bits or keys set in both molecules plus those set either molecule.
Note 1: In set theory terms the Tanimoto similarity is equal to the ratio of the intersection set to the union set.
Note 2: See also Tversky similarity.
Note 3: The value of a Tanimoto similarity of a pair of molecules depends on the molecular fingerprint or structural keys used in the calculation [17].
Note 4: The Tanimoto similarity can also be calculated for molecules described by continuous variables. The similarities in this case range from –1/3 to 1.0.

test set
The set of molecules not used in any way to devise a computational model but instead are used to test its predictivity.
Note: A test set differs from a prediction set in that the molecules and their associated activities may be derived from a source other than that from which the model was derived, for example, compounds tested after the model was developed.

topological index
A numerical value associated with chemical constitution for correlation of chemical structure with various physical properties, chemical reactivity or biological activity. The numerical basis for topological indices is provided (depending on how a molecular graph is converted into a numerical value) by either the adjacency matrix or the topological distance matrix. In the latter the topological distance between two vertices is the number of edges in the shortest path between these [13].
Note: Examples include: molecular connectivity chi (\(\mu_X\)); Kappa Shape (\(1K\), \(2K\), \(3K\), etc.); electrotopological state (\(S\)); and Dragon descriptors.
Modified from [9].

topological polar surface area (TPSA)
An approximation to polar surface area that is calculated from the 2D structure [71].

training set
The molecules and their associated properties that are used to generate a computational model.
Note: Modified from [23].

true negative (TN)
A model prediction of a negative result, such as no biological activity, when the true result is also negative.
Note: See also confusion matrix.

true positive (TP)
A model prediction of a positive result, such as biological activity, when the true result is also positive.
Note: See also confusion matrix.

Tversky similarity
An asymmetric measure in which a target structure is compared to a reference structure with user-selectable weighting of the relative importance of the fingerprint bits or structural keys that are present only in the reference versus the fingerprint bits or structural keys that are only in the target, weightings for which the Tanimoto coefficient uses 1.0.
Note: See also Tanimoto similarity.
underdetermined system
A system in which the number of descriptors is much larger than the number of observations.

validation
The process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose.

validation set
The molecules not used to devise a computational model but instead are used during model development to test for possible over-fitting as seen when additional properties are added to a model but the predicted potencies of the validation set become less accurate.

virtual reaction
The computer encoding of a possible chemical reaction between two explicit or Markush structures or sub-structures, often with a mapping between the atoms in the starting material and those in the product.

Note: see also SMIRKS.

virtual screening
Computational methods that rank the molecules in a database by their forecast continuous or categorical biological or chemical properties [72, 73].

Note: Virtual screening is often used to predict the ability of molecules to bind to a macromolecular target of known 3D structure, to fit a ligand-based hypothesis of bioactivity, for their similarity to known actives, or to be mutagenic.

Modified from [23].

wavefunction, \( \psi \)
“A mathematical expression whose form resembles the wave equations of physics, supposed to contain all the information associated with a particular atomic or molecular system. When a wavefunction is operated on by certain quantum mechanical operators, a theoretical evaluation of physical and chemical observables for that system (the most important one being energy) can be carried out” [13].

workflow
The sequence of steps used to perform a certain task.

Note: Pipeline Pilot and KNIME are two popular computer programs that support custom and changeable workflows.

Membership of sponsoring bodies
Membership of the Division Committee of the Chemistry and Human Health Division during the preparation of this report (2010–2012) was as follows:

Active Membership of the Subcommittee on Medicinal Chemistry and Drug Development (2010–2012) was as follows:

C.R. Ganellin (UK, Chairperson), J. Proudfoot (USA, Secretary, 2010) and D. Buckle (UK, Secretary, 2011–2012), S.O. Bachurin (Russia), E. Breuer (Israel), D.R. Buckle (UK), M.S. Chorghade (USA), P.W. Erhardt (USA), J. Fischer (Hungary), A. Ganesan (UK), G. Gaviraghi (Italy), T. Kobayashi (Japan), P. Lindberg (Sweden), Y. Martin (USA), P. Matyus (Hungary), A. Monge (Spain), T.J. Perun (USA), F. Sanz (Spain), J. Senn-Bilfinger (Germany), N.J. de Souza (India), G. Tarzia (Italy), H. Timmerman (Netherlands), J. Ulander (Sweden), M. Varasi (Italy), Yao, Zhu-Jun (PR China).

**Funding:** This manuscript (PAC-REC-12-12-04) was prepared in the framework of IUPAC project 2010-057-3-700.

### Annex 1: abbreviations and acronyms used in computational drug design literature

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almond</td>
<td>A 3D-QSAR program [74].</td>
</tr>
<tr>
<td>ALOGP</td>
<td>A program to calculate log P [75].</td>
</tr>
<tr>
<td>AM1</td>
<td>A semi-empirical quantum chemistry program [76].</td>
</tr>
<tr>
<td>AM1-BCC</td>
<td>A method that post-processes AM1 results to generate partial atomic charges for use in molecular modeling [77, 78].</td>
</tr>
<tr>
<td>Amber</td>
<td>A force field [79].</td>
</tr>
<tr>
<td>Amoeba</td>
<td>A force field [80].</td>
</tr>
<tr>
<td>AMSOL</td>
<td>A software package that combines a number of semi-empirical quantum chemical methods with several solvation models [81].</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve.</td>
</tr>
<tr>
<td>AutoDock</td>
<td>A program that docks molecules into protein binding sites [82].</td>
</tr>
<tr>
<td>B3LYP</td>
<td>A density functional theory approximation that is widely used in quantum chemical calculations [83].</td>
</tr>
<tr>
<td>BCUT descriptors</td>
<td>Encoding the intermolecular interaction properties of a molecule [84].</td>
</tr>
<tr>
<td>BEDROC</td>
<td>Boltzmann enhanced discrimination of receiver operating characteristic.</td>
</tr>
<tr>
<td>BSSE</td>
<td>Basis set superposition error.</td>
</tr>
<tr>
<td>Cactvs</td>
<td>A chemistry information toolkit, one function of which generates 3D structures [85].</td>
</tr>
<tr>
<td>CAESAR</td>
<td>A program that generates 3D structures [86].</td>
</tr>
<tr>
<td>Catalyst</td>
<td>A 3D-QSAR program [87].</td>
</tr>
<tr>
<td>CLOGP</td>
<td>A program that calculates log P [88].</td>
</tr>
<tr>
<td>CMR</td>
<td>Calculated molar refractivity.</td>
</tr>
<tr>
<td>CoMFA</td>
<td>Comparative molecular field analysis, a 3D-QSAR program.</td>
</tr>
<tr>
<td>CoMSIA</td>
<td>Comparative molecular similarity analysis, a 3D-QSAR program.</td>
</tr>
<tr>
<td>Concord</td>
<td>A program that generates 3D structures [89].</td>
</tr>
<tr>
<td>CORINA</td>
<td>A program that generates 3D structures [90].</td>
</tr>
<tr>
<td>Derek Nexus</td>
<td>A program that predicts various toxicity endpoints [91].</td>
</tr>
<tr>
<td>DOCK</td>
<td>A program that docks molecules into protein binding sites [92].</td>
</tr>
<tr>
<td>DRAGON descriptors</td>
<td>Molecular descriptors for 2D- and 3D-QSAR [93].</td>
</tr>
<tr>
<td>E_HOMO</td>
<td>Energy of the highest occupied molecular orbital.</td>
</tr>
<tr>
<td>E_LUMO</td>
<td>Energy of the lowest unoccupied molecular orbital.</td>
</tr>
<tr>
<td>FCFP</td>
<td>Functional class fingerprint.</td>
</tr>
<tr>
<td>FEP</td>
<td>Free energy perturbation.</td>
</tr>
<tr>
<td>FRED</td>
<td>A program that docks molecules into protein binding sites [94].</td>
</tr>
</tbody>
</table>
Glide: A program that docks molecules into protein binding sites [95].
GOLD: A program that docks molecules into protein binding sites [96].
ICM: Internal coordinate mechanics program for protein structure prediction, modeling, cheminformatics and ligand docking [97].
InChI: international chemical identifier.
KNIME: An open-source pipelining program [58].
kNN: k nearest neighbor.
LIE: Linear interaction energy [98].
LIGPLOT: A program that generates a pose diagram from a PDB file [59].
LLE: Lipophilic ligand efficiency.
LOO: Leave-one-out cross-validation.
LUMO: Lowest unoccupied molecular orbital.
LSO: Leave-some-out cross-validation.
MDS: Multidimensional scaling.
MLOGP: A program that calculates log P [99].
MM4: A molecular mechanics program [100].
mmCIF: macromolecular crystallographic information file.
MMFF: Merck molecular force field [34].
MNDO: A semi-empirical quantum chemistry program.
MOE: A general molecular modeling program that can be used to generate 3D structures [101].
Omega: A program that generates 3D structures [102].
Molconn-Z: A program that generates descriptors for 2D-QSAR [103].
MOO: Multiobjective optimization.
OPLS: A force field parameterized from the properties of various liquids [104].
O-PLS: Orthogonal projections to latent structures, a method that makes PLS results more interpretable [54].
PCA: Principal components analysis.
PDB: Protein data bank.
Pentacle: A 3D-QSAR program [105].
PES: Potential energy surface.
Phase: A 3D-QSAR program [106].
Pipeline Pilot: A pipelining program [107].
PLS: Partial least squares or projections to latent structures.
PM3: A semi-empirical quantum chemistry program.
PSA: Polar surface area.
QM/MM: Quantum mechanics/molecular mechanics calculations.
RMSD: Root mean square deviation.
RMSE: Root mean square error.
ROC: Receiver operator characteristic curve.
ROF: Rule of five.
SEA: Similarity Ensemble Approach.
SEE: Standard error of estimates.
SEP: Standard error of prediction.
SIMCA: A program for the statistical analysis of data [108].
SOM: Self-organizing map.
SVM: Support vector machine.
TPSA: Topological polar surface area.
VolSurf: A program that predicts pharmacokinetic properties [109].
References


**Note:** Republication or reproduction of this report or its storage and/or dissemination by electronic means is permitted without the need for formal IUPAC or De Gruyter permission on condition that an acknowledgment, with full reference to the source, along with use of the copyright symbol ©, the name IUPAC, the name De Gruyter, and the year of publication, are prominently visible. Publication of a translation into another language is subject to the additional condition of prior approval from the relevant IUPAC National Adhering Organization and De Gruyter.