Evolving Drug Design Methodology: from QSAR to AIDD

Jun Xu*[1,2]

[1] Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou 510006, China
[2] School of Biotechnology and Health Sciences, Wuyi University, Jiangmen 529020

ABSTRACT: When medicinal chemistry was born a hundred years ago, a drug design methodology was expected to be based on the knowledge of the relations among chemistry, biology and medicine. Originally, chemists believed that a drug molecule consists of a scaffold with several substituents. While the substituents were replaced by alternate functional groups (*aka* substructures), the activity value of the molecule would be changed accordingly. This is termed as structure–activity relationship (SAR), which can be used to guide chemists to chemically modify the molecule to improve its druggability. Along with the progress of computing technology, SAR evolved into QSAR (Quantitative SAR). QSAR method prevailed in the era when determinism dominated the scientific community. Therefore, the paradigm of QSAR studies was based on the thought of "discovering the analytical rules (analytical formula of functional) between independent variables and functions hidden in experimental data by curve fitting or regression". Earlier QSAR was based upon so called the similarity and additivity postulates. With the advent of the era of high-throughput experiments and big data, the two postulates are facing serious challenges. Coupled with the puzzling problems (such as substructure partitioning, "activity cliff", unbalanced data sampling, and the paradox of prediction accuracy and generalization), conventional QSAR was declining from mature. In the beginning of this century, artificial intelligence (AI), specifically deep learning (DL), significantly succussed in image pattern recognition and natural language processing (NPL). AI was soon adopted for QSAR studies as a disruptive approach. It is now believed that drug design can be data-driven instead of rule-based (curve-fitting or regression). QSAR can also be directly revealed by AI without knowing the mechanisms of actions. Thus, the two postulates of conventional QSAR are no longer required, and the associated puzzling problems or paradoxes could be resolved. By examining the historical pathway of QSAR evolving into AI assisted drug design (AIDD), this review summarizes the process how the drug design paradigm is transformed from determinism to causalitism + probabilitism. The principles and challenges of drug design methodology are explored, the pros and cons for QSAR and AIDD are discussed with perspectives. It is worth noting that although AIDD is powerful, it is not omnipotent and should be treated rationally. The essence of machine learning is to reveal the major trends of a data set; while the minor trends (*aka* outliers, which are often ignored or discarded) cannot be captured by AI algorithms. However, the outliers are likely to be the entrances to disruptive discoveries. Therefore, philosophically, it is unrealistic to develop innovative drugs only relying on AIDD. AIDD's achievements rely on inheriting the legacy of QSAR's theories, methods, technologies, and data.

Keywords: AI-driven Drug Discovery, QSAR, Drug Design, Machine Learning, ANN
* Corresponding author: xujun9@mail.sysu.edu.cn; junxu@biochemomes.com

CONTENTS

1. Introduction
   - 1.1 Basic QSAR Process
   - 1.2 Conditions for Linear QSAR Modeling
   - 1.3 QSAR Modelling Steps
2. QSAR-based Drug Design Methodology
   - 2.1 Chemical Substructure Partitioning
     - 2.1.1 Search Keys
     - 2.1.2 Bitmaps and Molecular Fingerprints
     - 2.1.3 Molecular Structure Linear Notation
     - 2.1.4 Bottle-necks of QSAR
   - 2.2 Molecular Descriptors
     - 2.2.1 Molecular Descriptor Types
     - 2.2.2 Molecular Structural Data Curation, Selection and Normalization
     - 2.2.3 Transforming and Combining Molecular Descriptors
   - 2.3 Additivity of Substituent Contribution
     - 2.3.1 Similarity and Additivity Postulates
     - 2.3.2 Intrinsic Non-additivity
     - 2.3.3 Non-additivity Caused by Experimental Errors
     - 2.3.4 Identifying Non-additive Data
   - 2.4 Activity Cliffs
     - 2.4.1 Exceptions to Similarity Postulate
     - 2.4.2 Identifying Activity Cliffs
     - 2.4.3 Pains in Activity Cliff Study
3. From QSAR to AIDD
   - 3.1 Patterns Related to Activities
   - 3.2 Numeric Pattern Recognitions
   - 3.3 Pattern Recognition Algorithms
     - 3.3.1 Naïve Bayes Classifiers (NBC)
     - 3.3.2 Decision Tree Classifiers
     - 3.3.3 Hierarchical Clustering
       - 3.3.3.1 Normalization of Data
       - 3.3.3.2 Euclidean Distance between Data Points
       - 3.3.3.3 Non-Euclidean Distance between Data Points
       - 3.3.3.4 Cluster Merging Strategies in Clustering Algorithms
     - 3.3.4 Non-hierarchical Clustering
       - 3.3.4.1 K-means Clustering and *k*-nearest Neighbor Clustering
       - 3.3.4.2 Dimension Reduction by Principal Component Analysis
       - 3.3.4.3 Dimension Reduction by Kohonen Neural Networks
   - 3.4 Pattern Recognition Based on Molecular Topology
   - 3.5 Artificial Neural Networks
     - 3.5.1 Neurons and Neural Networks
     - 3.5.2 Principles of Artificial Neural Networks
     - 3.5.3 ANN and von Neumann Architecture
     - 3.5.4 Drug Design, QSAR and ANN

3.5.4.1 Questions in Drug Design
3.5.4.2 Important Parameters for ADMET
3.5.4.3 Complexity of Multi-parameters Decision Making
3.5.5 Memory Footprint: Protein Structure Prediction
3.5.5.1 Brief History of Protein Structure Prediction
3.5.5.2 Physics that Drive Protein Folding
3.5.5.3 Protein Homology and Imperical Protein Structure Prediction
3.5.5.4 Protein Homology Modelling
3.5.5.5 AlphaFold 2: Success of DL
3.5.5.6 Unresolved Protein Structure Prediction Problems
3.5.6 Medicinal Chemistry Diversity Space Explorations
3.5.7 Target Identification and Validation
4. Summary and Prospect: Drug Design Methodology with AIDD
4.1 Paradigm of QSAR Studies
4.2 Challenges and Paradox to QSAR
4.2.1 Substructure Partitioning Issue
4.2.2 Activity Cliff Issue
4.2.3 Imbalanced Training Data
4.2.4 Paradox of Prediction Accuracy and Generality
4.3 Moving toward AI
4.3.1 Disruptive Thinking
4.3.2 Relations among ANNs and Natural Rule Types
4.3.2.1 Essence of ANN
4.3.2.2 ANNs and Natural Rule Types
4.3.2.3 AI and Metarecurion

1. INTRODUCTION

1.1 Basic QSAR Process

Drug discovery originated from the world traditional medical practician's natural medicine. Modern therapeutics developed from the screening of natural products based on animal models. This is generally called "phenotypic drug discovery" or "forward pharmacology". A hundred years ago, while drug design methodology was sprouted, Paul Ehrlich already proposed a systematic method to find pharmaceutical agents, and establish the relationship between chemistry, biology and medicine [1]. Natural products were not good enough for medical uses, and required to be chemically modified based upon the study of structure and activity relationship (SAR). With the progress of computing technology, SAR was expected to be quantified, therefore, SAR became QSAR (Quantitative SAR).

According to the survey of Hugo Kubinyi, a BASF medicinal chemist in Germany, QSAR can be traced back to 1863 a study on the relationship between the structure and activity of alkaloids done by A. Crum Brown and T. Fraser. [2]。 The beginning of QSAR history has no clear consensus, it is generally agreed that Corwin Hansch's work in 1960s pioneered modern QSAR. [3] while artificial intelligence (AI) was started in 1950s.

In medicinal chemistry, QSAR is broadly used compounds classifications, lead identifications and optimizations, predicting bioactivities or drug metabolism and

pharmacokinetics (DMPK) properties and toxicities for chemical compounds.

In earlier time of QSAR, Hansch analysis was based on following hypotheses: (Figure 1):

(1) the bioactivity of a molecule relies on its molecular structure (topology);

(2) the molecular structure consists of a scaffold and substituents at the scaffold;

(3) the scaffold (or privileged structure) is the main fragment that determines the bioactivity, and the substituents regulate the potency; [4]

(4) substituents' static and steric properties contribute to the potency;

(5) the contribution of each substituent to the potency is additive.

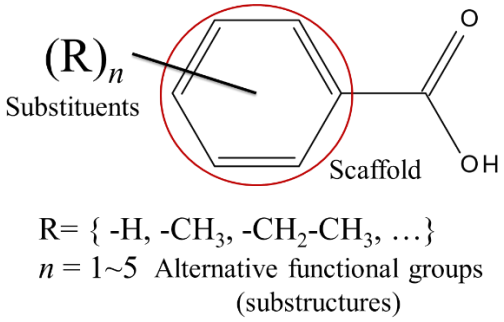


Figure 1. Concepts of scaffold and substituents in Hansch analyses

The electrostatic effect (electron pulling or pushing ability) of a substituent in an aromatic ring is measured by Hammett constant (denoted by σ), [5] A substituent has a specific σ value. Thus, a QSAR model of activity (*y*) and substituents can be established by linear regression method in equation (1).

$$y = f(\sigma(substituents(molecular\_substructure))) \quad (1)$$

Where, activity *y* is the function of σ , and σ is the function of a substituent, again a substituent is the function of a substructure. Equation (1) is a functional.

1.2 Conditions for Linear QSAR Modeling

There are two postulates hidden in the QSAR process:

(1) Similar molecular structures should have similar activities.

(2) The contribution of *n* (*n* > 1) substituents to the activity are of additivity.

Later, we will see the challenges or paradoxes against QSAR will associated with above-mentioned postulates.

To compute *y*, the *molecular_substructure* is replaced by an array of molecular descriptors that can be derived from a connection table (CT) of the molecular structure. Thus, CT is the lowest argument in equation (1). Furthermore, the precise representation of CT should be described by quantum chemistry eventually (first principles).

The simplest analytic formula of *y* usually is single-variable or multi-variable linear function. Correct fitting this function relies on the following conditions:

(1) training data used for modeling is correct and the positive/negative data points are evenly sampled;

(2) a scaffold is properly derived from the training set;

(3) substituents are properly derived, grouped, and represented in descriptors;
(4) when the scaffold has two or more substituents, they should be proved for being independent to each other. That is, the additivity is guaranteed;
(5) every selected descriptor ought to have strong linear relationship with *y*;
(6) there should have means to mutually transform continuous and discrete variables when descriptors come from continuous and discrete domains;
(7) the best algorithm is identified for modeling;
(8) efficient tools are selected to validate and select the final QSAR model;
(9) the QSAR model can be mapped to explainable chemistry.

1.3 QSAR Modelling Steps

Conventional QSAR modeling steps are listed as follows:
(1) Molecular structural data and bioactivity data curation, washing, and validation;
(2) Preprocessing molecular structure data or descriptor data;
(3) Selecting descriptors and algorithms;
(4) Generating and optimizing predictive models;
(5) Evaluating models for robustness, predictivity, and generality;
(6) Internally and externally testing models, testing the models with wet experiments if it is possible.

2、 QSAR-based Drug Design Methodology

2.1 Chemical Substructure Partitioning

To build a QSAR model, a scaffold has to be derived from a set of compounds with bioactivities (usually, this can be done by superimposing the molecular structures)[6,7]. The scaffold is a privileged substructure for a given biological target or activity, sets of alternative functional groups (R-groups) are then summarized at the specific positions of the scaffold (such as, ortho, para, or meta positions at an aromatic system). These R-groups represent how the substructures regulate the biological activity. Both substituents and the scaffold are substructures, the substructure partitioning is empirical without consensus rules.

Molecular structure data are prepared with chemical software, such as hemDraw or JChemPaint, and exported in SDF/MOL and many other formats that record atomic connection tables. [8]

There are many substructure partitioning approaches, such as, privileged structures, [9] graph pruning algorithm, and maximum common substructure. [10-12] Because a molecular scaffold definition relies on a specific drug target in medicinal chemistry, universal rules to define a drug scaffold or partition substructures for substituents are not existent.

2.1.1 Search Keys

In the earlier chemical database development, people were seeking a universal rule to produce the database search keys (*aka* substructural screens) to improve chemical structure search performance. MACCS (Molecular ACCess System) search keys were born for this purpose. The search keys are actually pre-defined substructural dictionary, which was the

consensus from a group of chemists organized by MDL(an early molecular design company). Therefore, MACCS search keys were purely biased on chemical experience without any medicinal chemistry rationale. The early version of MACCS had only 166 substructures, the later version was extended to 960 substructures. [13] MACCS search keys accelerated chemical database retrieval performance indeed. However, MACCS search keys should not be abused in QSAR modeling in medicinal chemistry. Because MACCS search keys are not necessarily associated with a drug target, not representing all important substructures in a compound library either. Furthermore, MACCS search keys are not guaranteed for being independent to each other (one search key can be a substructure of another search key). Again, MACCS search keys were not derived from drug-like compound libraries. There is no scientific foundation to use MACCS search keys as descriptors of drug leads.

2.1.2 Bitmaps and Molecular Fingerprints

To avoid the deficits of MACCS search keys, mathematic rule based systematic approaches to objectively derive substructures were proposed. Atom-center-fragment (ACF), [14] and extended-connectivity fingerprints (ECFP) [15] , and Daylight structure fingerprint (www.daylight.com/dayhtml/doc/theory/theory.finger.html) are representative approaches.

These rule-based approaches overcome empiric biases, result in objective and consistent substructure partitions. To improve the performance of informatic representation and computation, the data of substructures are store in bit-maps that are termed as molecular fingerprints (a new form of substructure dictionary). A disadvantage of the molecular fingerprints is that many substructure fragments produced by these methods lose their chemical meanings, and many fragments are rare in chemical databases. Therefore, molecular fingerprints can have many zero-bits resulting sparse bit-maps. However, molecular fingerprints used in chemical database search engines resulted in great performance, and became extremely popular.

Both ACF and ECFP systematically start at an atom (center atom) to derive a substructure by circular pruning (Figure 2). This type of substructure reflects the center atom's sense to specific chemical environment in an extent. It has chemical meaning, which can be used to elucidate chemical phenomenon (such as chemical shifts in NMR spectra). [14] Actually, molecular force field comes from the similar thought.

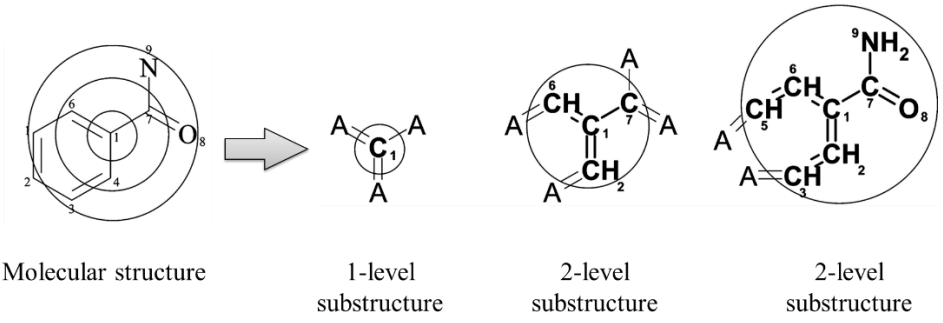


Figure 2. Substructures defined by ACF or ECFP approaches

Both search key and molecular fingerprint methods derive substructures from molecular structures. The former is based on the enumeration of chemists' experience, which retains

chemical intuition, but also inherits chemical biases; The latter automatically extracts substructures from molecular structures with algorithms based on pre-defined rules, which is rigorous, objective and reproducible, but they lose chemical meanings.

Search keys or molecular fingerprints can be used as feature vectors to characterize organic molecules. Taking MACCS-166 search key as an example, for any molecular structure ***S***, we can detect whether 166 substructures appear in ***S***. The binary bit-map ***B*** with a 166 bits is used to store the detection results. If a substructure appears in the molecule, the corresponding bit is set to "1", otherwise set to zero. In this way, each molecular structure ***S*** in a database is characterized as ***B*** (Figure 3). ***B*** is called substructure bitmap, which is a feature vector representing a molecule.

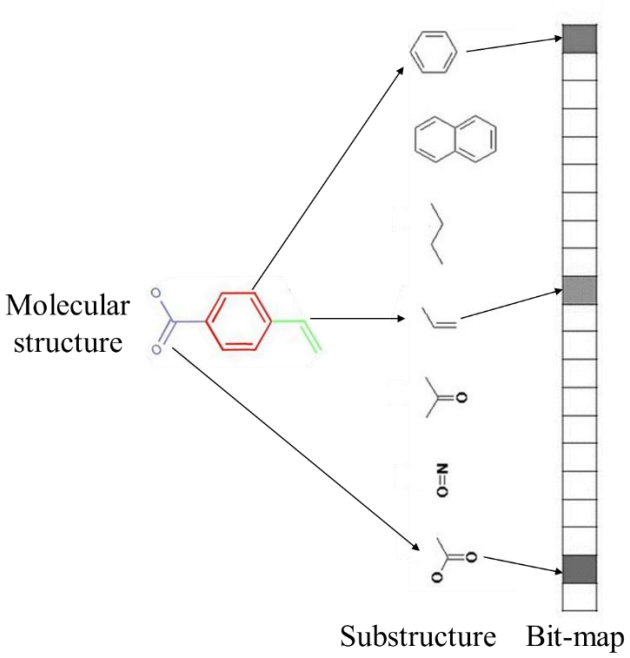


Figure 3. Molecular structure, substructure, and molecular bitmap

If you want to check whether molecule ***Q*** is in a chemical database, just convert ***Q*** into its molecular fingerprint $\boldsymbol{B_Q}$, and then compare $\boldsymbol{B_Q}$ with the fingerprint $\boldsymbol{S_i}$ of the *i*th molecule in the database. If $\boldsymbol{B_Q} \cap \boldsymbol{S_i} = 1$, it means that the query molecule ***Q*** is found in the database. The logical "and" operation of bitmap by computer is very fast, so the retrieval efficiency is very high.

For a 64-bit computer, the MACCS-166 requires three integers to store a molecule; MACCS-960 requires 15 integers. Daylight's molecular fingerprint calculation rules can produce many substructures, a longer bitmap (such as 256-bits or 512-bits) is required. When we use a longer bitmap to represent a molecule, the bitmap is informational sparse, resulting in waste of storage and increased computing costs. Daylight software folds the bitmap to shorten the length of the molecular fingerprint bitmap (for example, 2048-bits were folded to 1024-bits), which is called a Hashed fingerprint. This reduces computational costs, but the substructure information represented by the bitmap is lost.

When we feature a structure ***S*** with a bitmap ***B***, each bit in ***B*** is linked to a substructure, the bit is also called a descriptor. The number of bits in ***B*** is also called dimension. When a molecule is represented by a MACCS-166 fingerprint, the position of the molecule is actually determined

in 166-dimensional space.

Any molecule can be represented by a molecular fingerprint bitmap with the same length to calculate the similarity between molecules, save storage, and accelerate computation.

2.1.3 Molecular Structure Linear Notation

Early computers were rarely equipped with graphical terminals, so it was difficult to input chemical structure data into computers. Therefore, the technology of encoding the chemical structure diagram into a alphabetic string was invented to input the chemical structure diagram into a computer. The encoding resulted in chemical structure linear notation (LN).

LN uses a set of rules to convert the chemical structure CT to a string. The earliest LN was Wiswesser line notation (WLN). In 1968, ISI ® WLN was used to search the chemical database. WLN was also used internally by many pharmaceutical companies in the mid-1960s. [16] WLN uses letters to represent chemical structure fragments, which can effectively compress data and save storage, and is suitable for chemical structure database retrieval. However, WLN was difficult to be manually interpreted and eventually disappeared.

The nomenclature of organic chemistry [17] developed by the International Union of pure and applied chemistry (IUPAC) is also a LN, but it is not a graphic coding method in the sense of mathematical logic and cannot be directly used in the practice of chemical databases. The ideal LN should meet the following criteria:

(1) Non-ambiguity: a LN can only produce a unique molecular topology (structure);

(2) Uniqueness: a molecular topology can only produce a unique LN.

David Weininger combines the advantages of IUPAC system nomenclature and proposed an improved chemical structure linear notation (SMILES, simplified molecular input line entry system) [18]. SMILES is of non-ambiguity. However, the uniqueness is almost impossible. Therefore, The canonicalized SMILES is used to approach the goal of LN's uniqueness. Many chemical informatics software can convert SMILES strings into two-dimensional molecular structures and other data formats.

2.1.4 Bottle-necks of QSAR

Determining a scaffold and substituents is the prerequisite of traditional QSAR modeling, and it is also the bottleneck of QSAR studies. For more than half a century, the substructure classification issues never be completely resolved. The empirical method is not of generality. The rule-based methods are objective and systematic, but it produces too many mediocre substructures, sparse bitmaps, and fragments without chemical meaning.

These limits with other challenges make conventional QSAR decline after more than 60 years of development. [19]

As we will see in the following chapters, together with deep learning technology, SMILES as a rigorous chemical natural language brings hopes to resolve substructure partitioning issues.

2.2 Molecular Descriptors

Molecular fingerprint is a molecular descriptor, and its information carrier is Boolean vector. Each bit of the molecular fingerprint represents a substructure. Therefore, it is easy to think that the information carriers can also be integers, real numbers, complex numbers, and strings.

Molecular descriptors can be substructures, physical properties (such as molecular weight, molecular volume, refractive index, melting point), chemical properties (such as pH, enthalpy of formation, logP, logD), biological properties (such as IC_{50}, EC_{50}, bioavailability etc), or calculated molecular structure properties or indexes derived from theories (such as molecular topological index). [20]

Thus, molecules can be described in many different ways. A QSAR functional *y* can be represented in equation (2)

$$y = f(f'(\dots(descriptor(molecular\_structure)\dots)) \quad (2)$$

where *molecular_structure* is the lowest argument. If the scaffold and substituents were well defined, the QSAR's task is to seek the analytic form of (2).

molecular_structure is a topological graph *per se*, and can be represented in a SMILES string, CT or matrix. [21] Note that the functional *f (descriptor ())* may have many levels of nesting. Mapping molecular structure data to a vector or tensor space with molecular descriptors is not only a necessary for QSAR, but also makes it possible for us to use machine learning algorithms to build predictive models for the relationship of molecular structures and biological activities.

2.2.1 Molecular Descriptor Types

Molecular descriptors can be categorized into experimental and computational descriptors. The experimental descriptors are acquired from physical, chemical, and biological experiments with some measurement errors, which may cause the model inaccuracy.

The computational descriptors are calculated from fundamental data regarding molecular structures without experimental errors.

Molecular descriptor values can also be continuous or discrete. The continuous and discrete variables may co-existing in the same model, but the calculations are not allowed in a hybrid way.

Molecular descriptors with continuous values mainly take real values continuously in their definition domain (such as molecular weight, refractive index, melting point, solubility, pH value, logP, bioavailability, etc.);

Discrete valued molecular descriptors mainly take Boolean values (0 or 1, such as the existence or absence of the corresponding substructure) in their definition domain, integer value (such as: number of hydrogen bond acceptors / donors, number of rotatable bonds, molecular structure unsaturation, number of conjugated double bonds).

Over the past few decades, thousands of molecular descriptors for QSAR have been proposed [22] (github.com/mordred-descriptor). There are commercial or open-source packages that provide the descriptor calculations, such as Dragon[23], OpenBabel[24], RDkit[25], CDK2[26], PyDescriptor[27], Schrödinger, MOE, and DiscoveryStudio.

Because molecular descriptor calculations are within reach, people often ignore the scientific contents of various molecular descriptors, and the abuse of molecular descriptors in QSAR and deep learning practice often occurs.

2.2.2 Molecular Structural Data Curation, Selection and Normalization

Correct molecular descriptor data rely on the correctness of the raw molecular structure data.

Hence, washing the raw data is an important step before QSAR modeling begins. The steps of washing the raw data are depicted in Figure 4. The details can be found in references. [28,29]

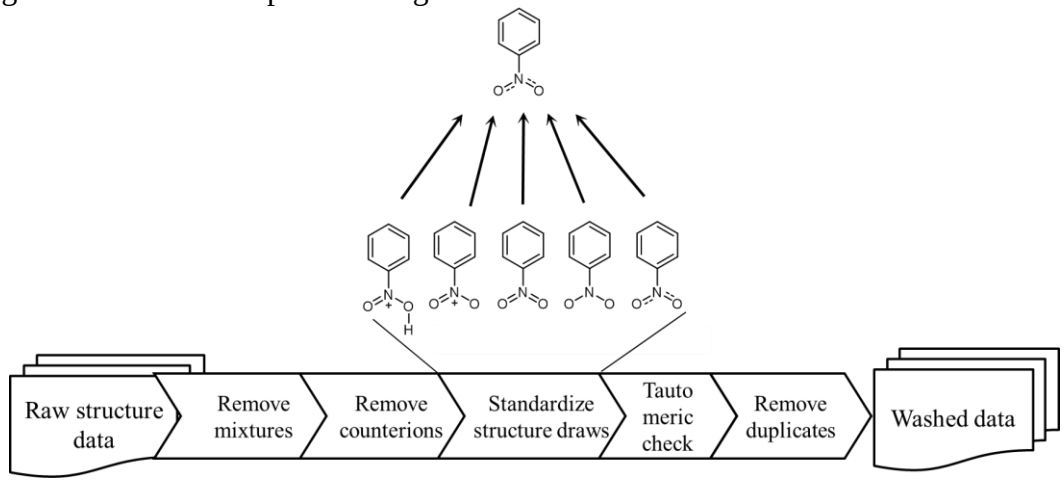


Figure 4. Steps of washing the raw chemical structure data

The washed raw data are also needed to be checked for consistency of compound naming, identifier and other text information, molecular structure normalization [30], valid atomic valence, correct abbreviations of elements or substituents. Finally, manual inspection is necessary, and there may be errors in the process of chemical structure data conversion. Figure 5 lists some examples of molecular structures that are easy to draw wrong. A free service for chemical and biological data collation processes is provided at: sites.google.com/site/dtclabdc/.

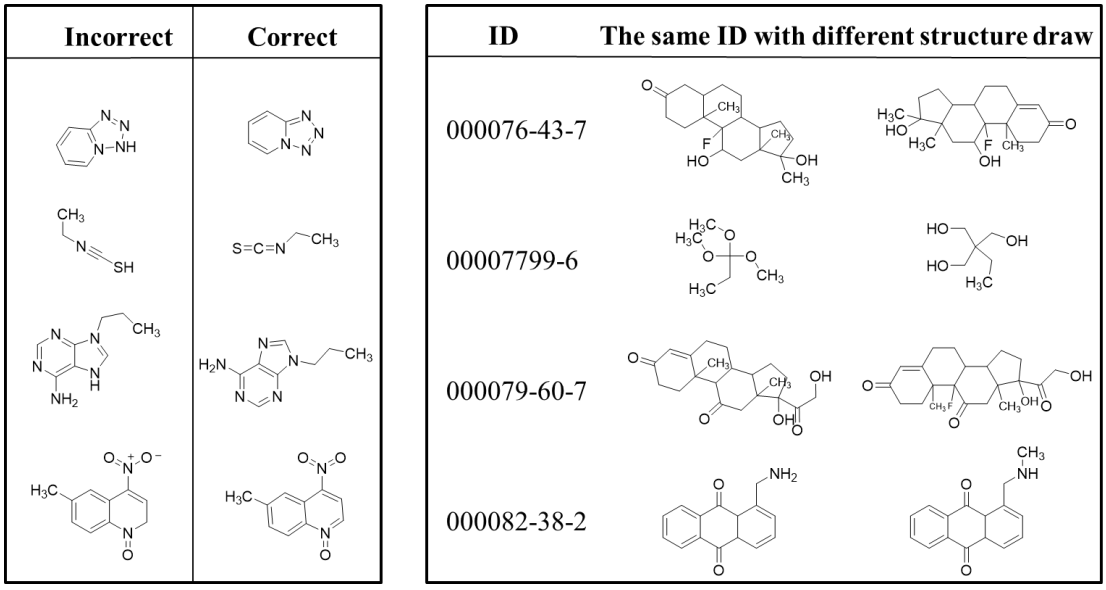


Figure 5. Possible structure drawing mistakes.
(Re-drawn based on the data in
cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=535257)

The following criteria are for selecting molecular descriptors in QSAR studies:
(1) A descriptor should have a clear correlation with the activity to be predicted;

(2) A descriptor and the activity to be predicted should have a wide range value distribution;

(3) If two descriptors are used in a model, they should be orthogonal to each other (that is, there is no correlation between them); For example, molecular weight and molecular volume are highly correlated with each other; they cannot be used in the same model;

(4) A selected descriptor has an explainable relationship with the biological activity.

The distributions of descriptors can vary greatly. For example, logP value ranges from -1 to 10, while drug-like molecular weight ranges from 100 to 1000 Dalton. If the two molecular descriptors are directly brought into the modeling without preprocessing, the contribution of logP to the model may be ignored because its range is too narrow, and the modeling results are dominated by the molecular descriptors with wide range. The purpose of data canonicalization (normalization and standardization) is to get rid of such problems. By checking the distribution of molecular descriptors, we can eliminate unreasonable data and select correct descriptors. In addition, there are many data standardization schemes. Researchers should use them on the basis of understanding their respective principles and scope of application in order to get a good effect of QSAR or deep learning modeling.

Min-max normalization and Z-score normalization are both sensitive to outliers (outliers) of a data set. Tangent function normalization are of robustness. Pre-calculating the statistical parameters of molecular descriptors (such as minimum, maximum, mean, mode, median, standard deviation) can help in choosing a simple normalization method such as minimum-maximum or Z-score normalization. If a molecular descriptor obeys Gaussian distribution (ideal random distribution, uniform sampling), Z-score normalization is the best. For noisy molecular descriptors (generally experimental data, especially biological experimental data), robust normalization techniques (such as tanh normalization) should be selected. If readers are interested in further understanding data normalization technology, they can read reference. [31]

2.2.3 Transforming and Combining Molecular Descriptors

A molecular descriptor has a definition domain and allowed value types as follows:

Boolean: such as molecular fingerprint bitmap, each byte bit can only take 0 or 1 (discrete value);

Positive integer: such as number of hydrogen bond receptors, number of hydrogen bond donors, number of rotatable bonds (discrete value);

Positive real number: such as IC_{50} (half inhibitory concentration) (continuous value).

Because the absolute value of IC_{50} with high activity is very small, it is generally taken as a negative logarithm, that is, pIC_{50}. In this way, when IC_{50} is far less than 1 M, pIC_{50} takes a positive real number. The larger the value is, the higher the activity is. pIC_{50} is in line with normal thinking pattern.

The free energy change (ΔG) of pIC_{50} and drug target binding affinity relationship is linear:

$$pIC_{50} = \frac{1}{RT}\Delta G + B \quad (3)$$

where, R and B are constant, T is temperature (experiments are usually conducted in a given temperature), also constant. Unsteady temperature could introduce noises to experiments.

The free energy change (ΔG) of ligand binding affinity to a target can be calculated by first principles or approximate models, so as to predict the IC_{50} value of small molecules inhibiting protein activity. [32]

Molecular Descriptor Combinations

Usually, the relationship of molecular structure and activity is nonlinear and complex with many layers. To simplify the modeling process, one often combines molecular descriptors. Therefore, we have to discuss a basic rule of combining different types of molecular descriptors.

(1) Discrete descriptors should not be hybridized with continuous descriptors;

(2) Boolean descriptors should not be hybridized with continuous descriptors or non-Boolean descriptors variables is often used, that is, the combination of multiple groups of molecular discrete descriptors;

(3) In principle, it is not suggested to combine descriptors that have significantly different meanings. For example, MACCS search keys should not be hybridized with Daylight fingerprints;

(4) Combining different types of independent descriptors can produce dependent descriptors. For example, combining MACCS search keys and Daylight fingerprints can be produce a new fingerprint vector, in which the substructure represented by Daylight fingerprints can intersect with the substructure represented by MACCS. This would violate the principle that molecular descriptors must be orthogonal to each other and will not produce correct results.

Molecular Descriptor Transformations

Either QSAR or AI is seeking a proper mathematical transformer to generate correct mapping from one set of parameters to others *per se*. A transformer is actually a generalized functional. For example, in a conventional QSAR model, the ionization constant *y* for a molecule is the function of the electrostatic constant σ in a benzene ring and the stereoscopic effect parameter *s*:

$$y = f(\sigma, s) \quad (4)$$

where *y*, σ, and *s* are real numbers， *f()* is a transformer that maps a real number vector to another real number vector.

In a virtual screening model, for example, an independent variable is a molecule that is represented in a SMILES string, and the function is the determination of whether the molecule is active or not (1 or 0) as shown in equation (5)：

$$y = f(\vec{S}) \quad (5)$$

where *y* is a Boolean variable, $\vec{S}$ is a string (a vector holds discrete components), *f()* transforms $\vec{S}$ to a Boolean variable.

In a conventional QSAR study, the most common and simplest transformer conducts linear "continuous variable vector→ continuous variable vector" transformation.

In essence, machine learning (ML) generalizes the transformer, which can conduct following transformations:

(I) Continuous variable vector → continuous variable vector;

(II) Discrete variable vector → discrete variable vector;

(III) Continuous variable vector → discrete variable vector;

(IV) Discrete variable vector → discrete variable vector.

A typical example of the (I) transformation is the motion law of a macro object described by classical mechanics. Force (F) acting on an object is proportional to the mass (*m*) and the acceleration (*a*) of the object, that is,

$$F = ma \quad (6)$$

where, *m*, *a*, and F take continuous values in real number domain.

A typical example of the (II) transformation is the use of quantum mechanics to describe the motion of electrons in atoms. For example, according to the law of atomic absorption or emission spectrum, when light irradiates on an atom, if the energy difference ΔE of electrons outside the atomic nucleus when transiting from low-energy orbit to high-energy orbit is equal to the energy carried by the photon ($h\nu$), the atom will absorb the energy of the photon, and the absorption line will be absorbed in the absorption spectrum:

$$\Delta E = R(\frac{1}{n_1^2} - \frac{1}{n_2^2}) \quad (7)$$

where ν is the frequency of a light wave (the reciprocal of the wave length λ), R is the constant related to the number of nuclear charges, n_1 and n_2 are the quantum numbers of electrons in low-energy and high-energy orbits, respectively. n_1 and n_2 cannot take continuous values, but only positive integer values. Therefore, equation (7) is a function that transforms integer discrete variables to a real discrete variable

A typical example of the (III) transformation is to describe the phenomenon that occurs in a neuron. A neuron consists of cell body and branches like roots that diverge outward from the cell body, called dendrites. The slender branches of dendrites are called axons, and there is a small gap between the terminal of axons and the dendrites of another neuron called synapse. A neuron can be regarded as a signal processing unit. The axon releases action potentials (electrical signals, such as high / low levels, whose values are discrete) and is the output of the signal processing unit, while the synapse receives neurotransmitters from the previous neuron, so the synapse is the input of the neuron cell. Because neurotransmitters are not electrical signals but chemical substances (such as neuropeptides and endogenous small molecule alkaloids). Therefore, the signal received by synapse is compound concentration data, and its value is continuous. In this way, the function of a neuron can be described as following:

$$V = f(\vec{C}) \quad (8)$$

where V denotes the released action potential signal (discrete variable), $\vec{C}$ denotes the concentration of neurotransmitters received from the last neuron (continuous variable), equation (8) is a function of transforming continuous variables into integer / Boolean discrete variables.

S functions (sigmoid functions) can be employed to transform "continuous variables" onto "discrete variables" in equation (8). S functions have many formulations, the commonly used are logical function (9) and hypertangent function (10):

$$y = f(\frac{1}{1+e^{-x}}) \quad (9)$$

$$y = f(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}) \quad (10)$$

Where the value range of equation (9) is [0, 1], while equation (10) is [-1, 1]. They have the same shape. Equation (9) is commonly employed to fit experimental data to obtain activity values such as half inhibition rate (IC_{50}). As shown in Figure 6, X-axis represents the logarithm of the molar concentration (x= log (C)), Y-axis represents the percentage of protein activity inhibited. y ranges between 0 (no inhibition) and 1(complete inhibition).

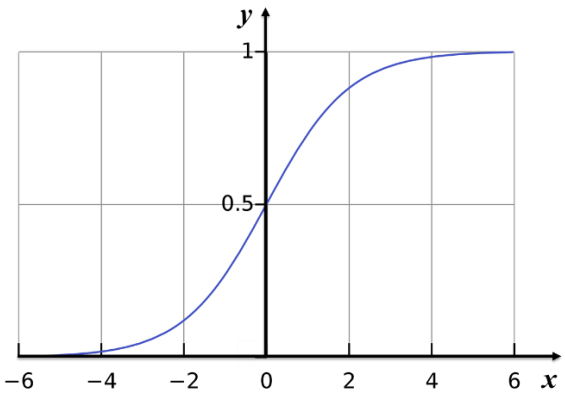


Figure 6. The shape of a logistic function

The *S* functions are monotonic. The first-order derivative of a *S* function is bell-shaped, which has only one local extreme value.

In biological cells, a protein (receptor, *R*) recognizes a drug (ligand, *L*) by through molecular docking at the active site of *R*. Under a given thermodynamic condition, the binding of *R* and *L* reaches an equilibrium:

$$R + L \rightleftharpoons R \cdot L \tag{11}$$

Given a receptor concentration [*R*], the protein activity inhibition rate *y* increases with the increase of ligand concentration [*L*]. When [*L*] is so large that every receptor molecule's active site is occupied by a ligand, the protein activity will be completely inhibited (*y*=100%). At this time, the protein activity inhibition rate will not grow by increasing [*L*].

S-function transformation represents a universal natural phenomenon. For examples, the biological responses to stimulus, the luminescence phenomenon caused by atomic absorbing energy (ground state electrons absorb energy, and after the energy accumulates to a certain extent and reaches the excited state, the frequency of the emitted light quanta can only take discrete values, that is, the multiple of Planck constant). These phenomena follow the same rule: quantitative changes eventually lead to qualitative changes. Mathematically, continuous variables are quantized. The quantized value represents an object's intrinsic feature, which is termed as eigenvalue. These features are commonly used to classify objects in science. The mission of QSAR and AI methods is to find the analytic function of the classifications. They both generate following mathematical transformers:

Numerical regression: continuous variables → continuous variables;

Logistic regression: discrete variables → discrete variables;

Deep learning: continuous variables ↔ Discrete variables.

Therefore, DL has broader applications.

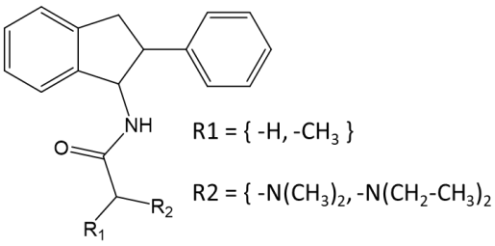
2.3 Additivity of Substituent Contribution

2.3.1 Similarity and Additivity Postulates

The similarity and additivity postulates of QSAR are related to each other. The similarity postulate (molecules with a similar scaffold should have similar activity/property) addresses the

commonness of a group of molecules.[33] This postulate is also the foundation of ligand-based virtual drug screening. [34,35]。

Additivity postulate was first proposed by S. M. Free and J. W. Wilson in 1964.[36] In essence, the postulate assumes that the similar molecules have a common scaffold specifically for a biological target. These molecules can be represented in a common scaffold, in which several substituents are attached. Figure 7 presents an example, a scaffold has two substituents R_1 and R_2. The additivity postulate assumes that the activity contributions of R_1 and R_2 are independent, the activity value of the molecule can be calculated by simply adding the contributions from R_1 and R_2.



R1 = { -H, $-CH_3$ }

R2 = { $-N(CH_3)_2$, $-N(CH_2-CH_3)_2$

R2	R1			LD_{50} Average
	-H	$-CH_3$		
$-N(CH_3)_2$	2.13	1.64	⇨	1.885
$-N(CH_2-CH_3)_2$	1.28	0.85	⇨	1.065
	⇩	⇩		
LD_{50} Average	1.705	1.245		1.475 (Average)

Linear Regression: $LD_{50} = a + x_{R1} + x_{R2}$

R_1 contributes LD_{50}: -H = +0.23; $-CH_3$ = -0.23

R_2 contributes LD_{50}: $-N(CH_3)_2$ = +0.41; $-N(CH_2-CH_3)_2$ = -0.41

Total LD_{50} average: a = 1.475

LD_{50}值 Predicted by additivity postulate:

M*i* (R1 R2):

M1 (-H, $-N(CH_2-CH_3)_2$) = 1.475 + 0.23 + (-0.41) = 1.295 (1.28 experiment)

M2 (-H, $-N(CH_3)_2$) = 1.475 + 0.23 + (+0.41) = 2.115 (2.13 experiment)

Figure 7. Example of Free-Wilson analysis.

With a given scaffold, the activity of a molecule is function of the substituents. If a substituents ***R*** in the scaffold has *n* alternative groups ($n > 1$), the additivity postulate assumes that the contributions of all substituent to the molecular activity can be simply summed up, that is, the molecular activity *y* can be expressed as a linear combination of the descriptor $\vec{\boldsymbol{x}}$ of the substituents ***R***

$$y = \sum_{i=1}^{n} c_i x_i + a \qquad (12)$$

c_i is the *i*th regression constant, x_i is the *i*th descriptor, regression constant *a* is related to the average activity of the all-training molecules.

In many cases, the postulates work well, and are consistent to medicinal chemists' SAR map. An SAR map is depicted in Figure 8.

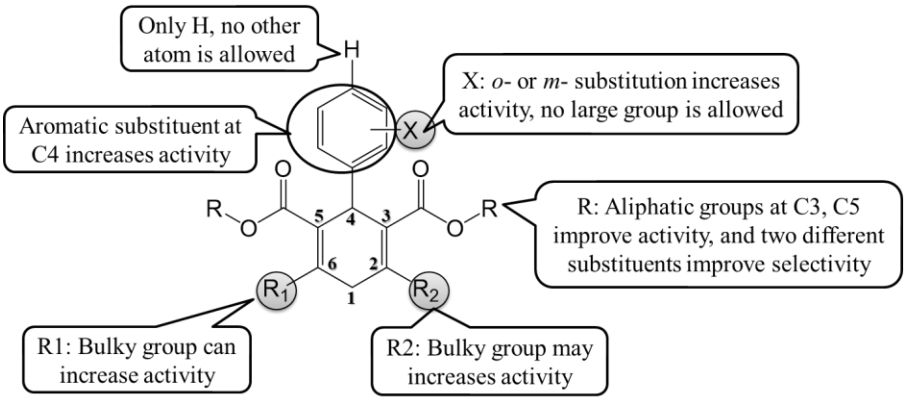


Figure 8. SAR map for dihydropyridine calcium channel blockers (redraw based on ref.). [37]

2.3.2 Intrinsic Non-additivity

If there is a synergistic effect among substituents, it will produce the non-additivity of substituents' contribution to molecular properties. For example, if the -X group in Figure 8 has a conjugation effect with the aromatic ring, then X, R, and R1 will synergistically contribute to the molecular property due to redistributing the electronic densities on the aromatic ring. This will result unexpected properties.

Following factors can lead to non-additive effects:

(1) Hydration effect of receptor-ligand, [38,39]

(2) Hydrogen bond formation or hydrophobicity, [40,41]

(3) Formation of intramolecular hydrogen bonds, [42]

(4) The ligand binding mode changes (for example, ligand turnover or side chain competition in the binding site can lead to a non additive change of more than 2 log units), [43]

(5) Ligands trigger protein side chain movement. [44]

Among them, (4) and (5) will cause stronger non-additive effects.

2.3.3 Non-additivity Caused by Experimental Errors

(1) Experimental errors: there are many interfering factors in biological experiments. Even if there is no essential non-additive effect, the experimental data may still produce large fluctuations. If the experimental process is complex, it will accumulate into a non additive effect caused by significant errors.

(2) Unbalanced experimental sampling: ideally, the collection of experimental data should be continuous and representative. For example, the alternative functional groups for substituent X (Figure 8) should be based on diversified physical properties (for examples, the groups with strong electron pulling to strong electron pushing properties, such as, -F, -Cl, -H, $-CH_3$, $-CH(CH_3)_2$). In this way, the X contributions to the biological activity may gradually increase or decrease. But in practice, the experimental data sampling is commonly unplannable (due to reasons such as no synthetic feasibility) resulting in deficit data.

Non-additive effects in QSAR data is common. In 2008, scientists from Peter Willett group of University of Sheffield used the Free-Wilson method to study a batch of nearly complete SAR

data sets, and found that only half of the data are significantly of additivity. [45]

Therefore, recognizing non-additivity data training data is critical for QSAR studies. Although non-additive data should not be used in QSAR studies. Non-additive data can be important for drug innovations, and indicate that either the receptor-ligand binding is significantly changed, or new SAR features are identified and, require elucidations. An in-depth discussions can be found in reference. [46]

In addition, communications between QSAR modelers and experimenters are critical to maintain the systematic design, acquisition, and quality of data. Medicinal chemists should have basic understanding of QSAR principles, and modelers should the basic knowledge regarding the pharmaceutical data and how the data are obtained *in vivo*, *in vitro*, and *in silico*.

Under the conditions of minimizing experimental errors and balanced samplings, a QSAR modeler should only select the data without synergistic effects among substituent groups to warrant the additivity postulate to get regression constants $\vec{\boldsymbol{c}}$ and a for equation (12).

2.3.4 Identifying Non-additive Data

To identify non-additivity data, the distribution of experimental errors for a data set needs to be examined. If the distribution is normal, significant non-additivity data including abnormal SAR data can be figured out by estimating the upper limit of experimental uncertainty in the data set.

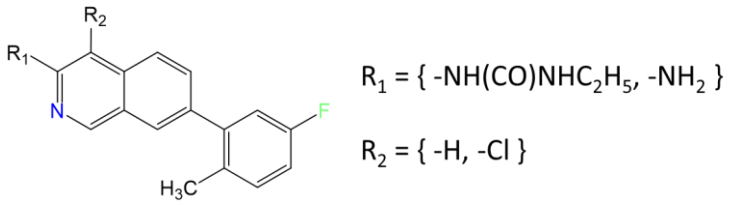


Figure 9. Tyrosine kinase ABL inhibitor's scaffold and two substituents (reproduced from ref.[46])

Chemical double mutant cycle (CDMC) was originally established by Diederich and Hunter of University of Sheffield to measure the interaction ability of two substituents. CDMC was performed to analyze whether there is non-additivity between substituents R1 and R2 in an example of Figure 9. [47,48] An example flow-chart of CDMC is depicted in Figure 10.

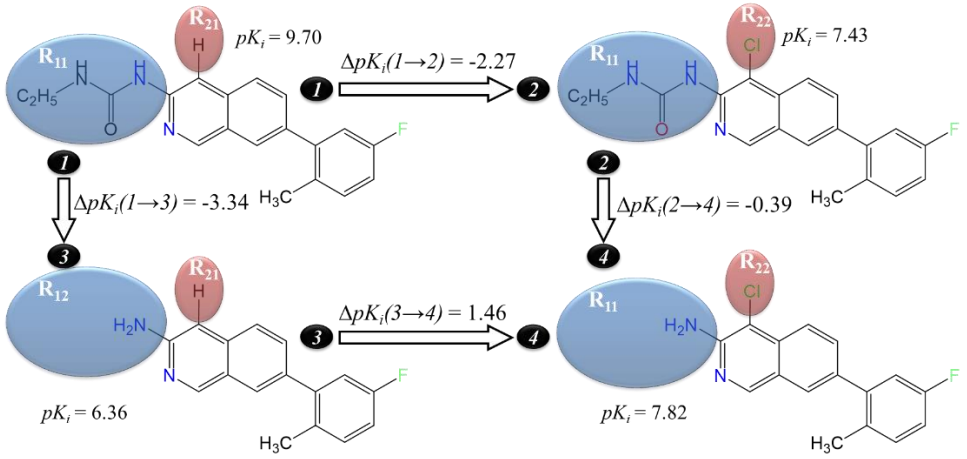


Figure 10. CDMC analysis for four ABL inhibitory compounds

To quantitatively examine the non-additivity, Kramer proposed a method to calculate non-additivity.[43] The non-additivity for substituents R_1 and R_2 can be computed through $\Delta\Delta pK_i$ as follows:

$$\Delta\Delta pK_i(R_1, R_2) = \Delta\Delta pK_i(3 \rightarrow 4) - \Delta\Delta pK_i(1 \rightarrow 2) \quad (12)$$

Thus, the non-additivity for R_1 and R_2 can be measured in $\Delta\Delta pK_i(R_1, R_2)$:

$\Delta\Delta pK_i(R_1, R_2)$ =1.46-(-2.27)=3.73 (see Figure 10).

Let's estimate the range of experimental error ε. The relationship of $pK_{i(exp)}$ and $pK_{i(true)}$ can be expressed as following:

$$pK_{i(\text{exp})} = pK_{i(\text{true})} + \varepsilon \quad (13)$$

Thus, $\Delta\Delta pK_{i(exp)}$ is:

$$\Delta\Delta pK_{i(\text{exp})} = \Delta\Delta pK_{i(\text{true})} + \varepsilon_2 + \varepsilon_3 - \varepsilon_1 - \varepsilon_4 \quad (14)$$

If R_1 and R_2 are completely additive, then $\Delta\Delta pK_{i(true)} = 0$

Hence, $\Delta\Delta pK_{i(exp)} = \varepsilon_2 + \varepsilon_3 - \varepsilon_1 - \varepsilon_4$, the errors are caused be random experimental errors.

For the general structure in Figure 9, if R_1 and R_2 have many alternative functional group members, the general structure is a library holding thousands of compounds. CDMC analysis can be applied on all the compounds in the library, and result in many $\Delta\Delta pK_{i(exp)}$ data (that is, the experimental errors ΔpK_i when we introduce R_1 and R_2).

Equation (14) tells us that the experimental errors consist of two portions: true non-additivity errors ($\Delta pK_{i(true)}$) and random experimental errors.

If $\Delta\Delta pK_{i(exp)} \neq 0$, Kramer and colleagues proposed to use the quantile-quantile plot (QQPlot) slope as a criterion to conclude if it is a non-additivity error or random error.

First, let's compute all errors ($\Delta\Delta pK_{i(exp)}$) from the compound in the library, denoted in $\vec{\boldsymbol{\varepsilon}}$:

$$\vec{\boldsymbol{\varepsilon}} = \sum_{j=1}^{n} \Delta\Delta pK_{i(exp)}^{j} \quad (15)$$

where， n is the total number of the library.

Array $\vec{\boldsymbol{\varepsilon}}$ are sorted ascendingly ($\varepsilon_{j-1} < \varepsilon_j$), measure means and standard deviation of $\vec{\boldsymbol{\varepsilon}}$, μ and σ:

$$\mu = \frac{\sum_{j=1}^{n} \varepsilon_j}{n}, \sigma = \sqrt{\frac{\sum_{j=1}^{n} (\varepsilon_j - \mu)^2}{n-1}} \quad (16)$$

The quantile(Q) of $\vec{\boldsymbol{\varepsilon}}$ is computed as following:

$$Q_j = \frac{\varepsilon_j - \mu}{\sigma}, t_j = \frac{j-0.5}{n} \quad (17)$$

Searching normal distribution table to obtain the quantile Q_j , t_j and corresponding theoretical Q'_j。

Plot Q_j against Q'_j to result in QQ-Plot (Figure 11).

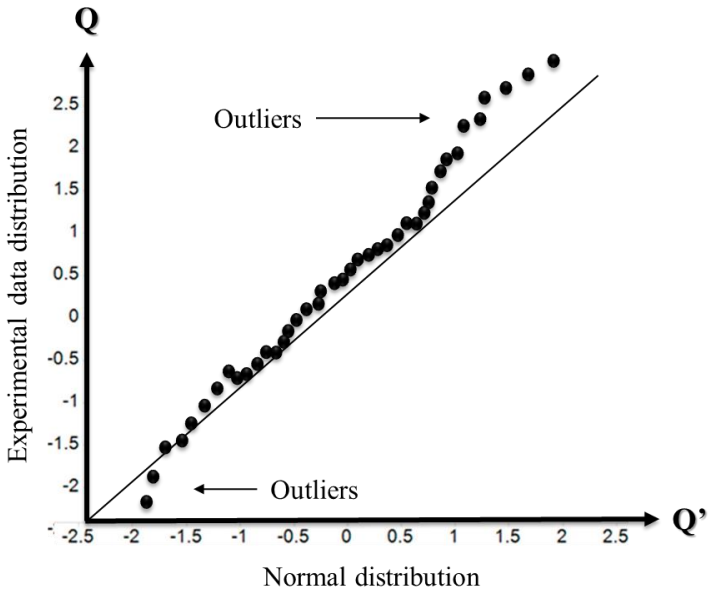


Figure 11. QQ-Plot for non-additivity analysis.

Quantile refers to dividing the probability distribution range of a random variable into several continuous intervals with the same probability. In Figure 11, if the quantile observed in the experiment is consistent with the quantile of the standard normal distribution (on the 45-degree slope line), it shows that the experimental error is a random error caused by the experimental uncertainty, that is, the data points belong to the additive data points. Those data points that obviously deviate from the normal distribution (points that deviate from the linear relationship, that is, outlier data points) may belong to non-additive data, at least the data whose experimental error is too large to be measured repeatedly.

Therefore, QQ-plot is an effective tool to identify potential non-additive data points. Relevant specific application examples can be read from Kramer et al. [43] [46]

Weak non-additivity can be an illusion formed by the accumulation of many experimental errors. Therefore, when analyzing non-additivity, the accumulation of experimental errors should be considered to avoid analyzing the non-additive effect that does not exist.

The main purpose of non-additive analysis is to exclude non-additive data from training data, so as to improve the prediction accuracy of QSAR models. However, non-additive data should not be discarded, but should be studied separately to capture the potential new mechanism of ligand-receptor interaction.

2.4 Activity Cliffs

2.4.1 Exceptions to Similarity Postulate

The challenges to the similarity postulate have been reported in many fields of chemistry. This phenomenon was originally termed as "property cliff", that is, a similar molecule has a very different property. This was regarded as a outlier [49]. Later, odor cliffs and activity cliffs were reported. Activity cliffs were mentioned by Michael Lajinessin a QSAR monograph edited by Silipo and Vittoria. [50]

Nowadays, there are often a pair of compounds with similar molecular structure, one of which is effective, but the other is ineffective; one is agonist, the other is antagonist; one is selective for a given target, but the other is non-selective.

2.4.2 Identifying Activity Cliffs

To study activity cliffs, many SAR visualization approaches were developed to single heterogeneous data out in biological assay data. Such as scaffold trees, SAR maps, clustering/decision trees, structure activity similarity maps (SAS maps), self-organization maps (SOM), dimension reduction maps, various network graphs (such as Sali graph, bipartite matching molecular series graph), etc. the commonly used visualization approaches are summarized in reference. [51]

Among these approaches, activity landscape (AL), also known as SAS map is interesting (Figure 12).

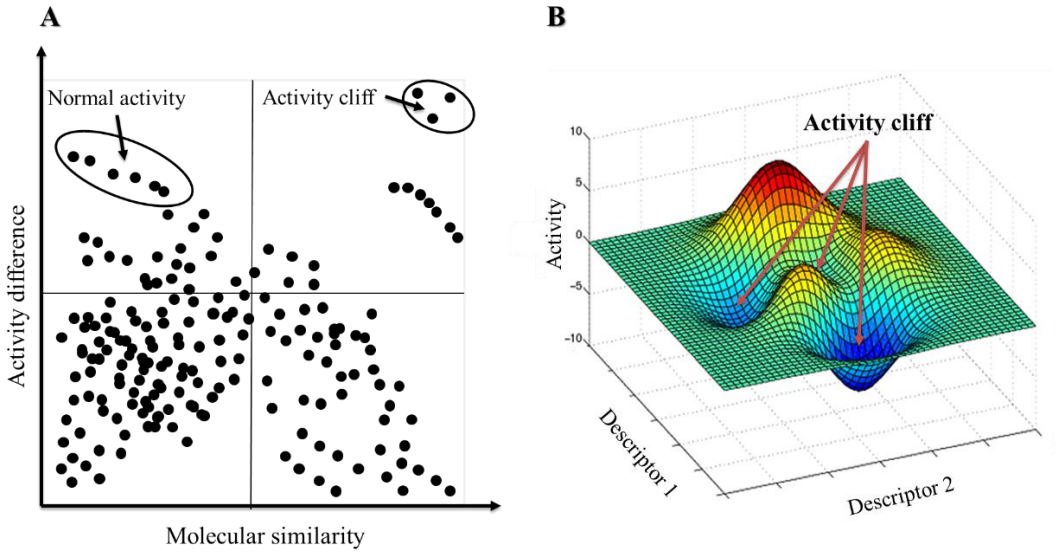


Figure 12. 2D and 3D activity landscapes

SAS maps can be two-dimensional and three-dimensional, represent the distribution of active compounds in chemical space, and demonstrate the determinants or key compounds of SAR. The vertical axis of Figure 12A represents the activity similarity difference of a pair of compounds, the horizontal axis represents the molecular structure similarity of a pair of compounds, and the dots represent a pair of compounds. [52-54] Under normal circumstances, the higher the molecular structure similarity of compound pairs, the smaller the activity difference. If the higher the similarity of the molecular structure of a pair of compounds, the greater the difference of their activities, it is the activity cliff. Figure 12D is a three-dimensional landscape map, x-axis and y-axis represent different molecular descriptors (they should be continuous

variables), z-axis represents the biological activity of molecules [55] and the Activity cliff is more intuitive on this kind of map.

When there are many ligand data for a given target, the ligand shows great structural diversity. At this time, the SAS map should be discontinuous. Bajorath team proposed the activity cliff network graphic method. The edge of the network represents the Activity cliff pair (the nodes at both ends of the edge represent the Activity cliff pair with large activity difference, the strong activity is represented by the red dot, the weak activity is represented by the green dot, and the activity difference of the Activity cliff pair is $\geqslant 2$; the compounds related to the Activity cliff pair but not strong and weak by themselves are represented by the yellow dot). [56]

The visualization approaches can manifest "isolated cliffs" and "coordinated cliffs". These exceptional Activity cliffs could guide chemists suddenly enter the realm of "hidden willows and bright flowers" while they were confused and frustrated in lead optimization process. Figure 13 shows an example, a heteroatom effect [57] and methyl effect [58] to and aromatic ring.

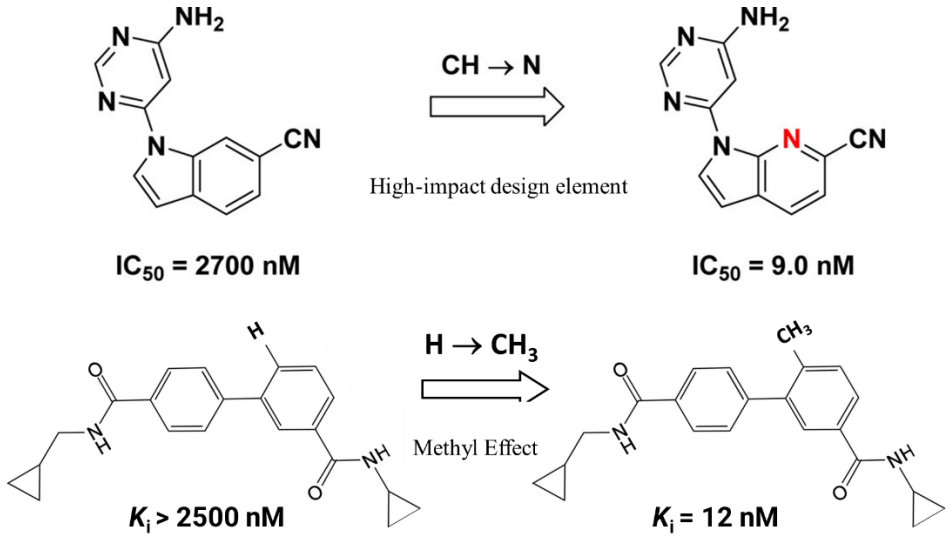


Figure 13. Examples of activity cliffs in medicinal chemistry.

It can be seen that identifying Activity cliffs is an important strategy for lead optimization processes. The development of various visualization methods of SAR is helpful to identify Activity cliffs. Bajorath team has done a lot of work in these aspects [59] and systematically study on the evolution of Activity cliffs. [56,60] From about 287000 compounds targeting about 1900 targets, they found about 3500 pairs of similar compounds with diatomic substitution. There are 852 pairs of compounds with significant differences in biological activities (they generally regard the two compounds with pK_i differences greater than 2 as Activity cliff pairs). [56]

However, it is still questioned: "Do activity cliffs exist?" [61,62]

2.4.2 Pains in Activity Cliff Study

The difficulties of Activity cliff studies lie in:

(1) The concepts of "molecular similarity" and "activity similarity" are not clearly unified defined. There are many methods to calculate molecular similarity. [63] The traditional concept of similarity of pharmaceutical chemists is qualitative, that is, two molecules have a common "scaffold", and their differences from substituents. In order to quantify the concept of

similarity, similarity calculation methods based on molecular fingerprints were proposed; In order to consider the stereoscopic effects, similarity calculation methods based on the three-dimensional superposition of molecules were proposed. But, chemists have no consensus on unified similarity calculation.

If the characteristics of molecules A and B are characterized by feature vectors $\vec{a}$ and $\vec{b}$, then the similarity $S_T(A,B)$ of molecules A and B can be calculated by Tanimoto coefficient: [64]

$$S_T(A,B) = \frac{\vec{c}}{\vec{a}+\vec{b}-\vec{c}} \quad (18)$$

where, $\vec{c}$ is the common feature vector of A and B.

Let ($\vec{a}$-$\vec{c}$) be only features for A, ($\vec{b}$-$\vec{c}$) be only features for B. α and β are weights for molecules A and B (α and β are real number within 0 and 1), The similarity of A and B, $S_{Tv}(A,B)$ called Tversky coefficient [65]:

$$S_{Tv}(A,B) = \frac{\vec{c}}{\alpha(\vec{a}-\vec{c})+\beta(\vec{b}-\vec{c})+\vec{c}} \quad (19)$$

Let α=β=0.5, it is called Sørensen–Dice coefficient [66,67]:

$$S_D(A,B) = \frac{2\vec{c}}{\vec{a}+\vec{b}} \quad (20)$$

Tanimoto coefficient is well employed in chemoinformatics. Chemical structure similarity approaches have been reviewed by Peter Willett and colleagues. [63]

(2) The types of Activity cliff phenomena are diverse and the situations are complicated. For example, from the perspectives of scaffold and substituents, there are topological cliffs, scaffold cliffs, chiral cliffs, R-group cliffs, scaffold / topological cliffs, and combinations of these cliff types. From the perspectives of calculating the structural similarity of three-dimensional superposition of molecules, many factors can cause cliffs, such as hydrogen binding or ion interactions, lipophilic or aromatic group interactions, water molecules, stereoisomerisms, and the combinations of the factors. It is difficult to classify the cliffs and figure out mechanisms.

(3) Random errors (data produced by the same laboratory) and systematic errors (data produced by different laboratories) are not avoidable in experiments. Therefore, the true or false Activity cliffs are seldomly distinguished. A long biological experimental protocol can accumulate errors, many experiments are hard to be repeated, and many biological experiments are costly and time-consuming.

(4) The essence of the Activity cliff is whether the substituents in a molecular scaffold contribute to the specific properties of the molecule. If the contributions of substituents to a specific property are not additive, there will be Activity cliffs. If the Activity cliff is not very steep, it is difficult to distinguish from the experimental error accumulations. If there are multiple substituents on a molecular scaffold, it is necessary to investigate whether there is a synergistic effect between the substituents. If there is a synergistic effect (such as conjugation effect and stereoscopic effect), there should be an activity cliff phenomenon, but it may not be observed experimentally, because the synergistic effect can lead to the increase, decrease, or no-effect to an activity.

(5) An activity cliff can also be treated as an outlier due to inadequate data. However, due to the difficulty of chemical synthesis, inadequate data can be common cases.

(6) Some activity cliff problems are not the problem of ligands themselves, but the leap from quantitative change to qualitative change when receptors are regulated. For example, the activity pocket of the receptor is large enough, and a substituent on the ligand smoothly increases its activity with the increase of atomic weight, but once it reaches the limit, it leads to a sudden transition of the receptor conformation. For this kind of phenomenon, molecular dynamics

simulations are required for the activity cliff investigation.

Activity cliffs have been studied for many years in QSAR field. [56] There are many achievements in activity cliff studies from data visualization, to the improvement of similarity measurement, mechanism classifications, and monitoring activity cliffs on a time scale. The resurgence of this wave of AI may bring new opportunities for QSAR and activity cliff studies. [68]

3. From QSAR to AIDD

3.1 Patterns Related to Activities

In the earlier of last century, QSAR was simplified as linear relationship due to the limit of computing performance. With the rapid development of computer hardware, software, and data accumulation, it is imperative to study QSAR with nonlinear approaches. Thus, the AI era of QSAR is coming.

A QSAR method is pattern recognition *per se*. The patterns can be graphic (substructural or subgraphic) [16] or numerical [69] , and derived from molecular structure data. The common patterns related to activities or properties are as follows:

(1) Patterns represented in molecular structure/substructures or general structure (*aka* Markush structure [70]). These patterns are mainly empirical, and manifest to medicinal chemists that some portions of a molecular graph are important to activities; some portions can be replaced by alternative functional groups; some portions have positive or negative impacts on activities. These patterns directly point out the directions for chemists to modify or optimize a drug lead. The disadvantages of these patterns are empirical and lack of consensus.

(2) Patterns represented in molecular fingerprints. These patterns are substructures tailored from molecular structures using a specific algorithm, and derived from a compound library for profiling a chemical library, or computing similarities of compounds [71]. These patterns are objective but, they don't have chemical meaning, therefore, they cannot manifest intuitive guidance for a chemist to modify chemical structure for drug lead optimizations.

(3) Patterns represented in regression. These patterns address the mathematical relationship between data (independent variable x) and activity (dependent variable y) expressed in mathematical function forms (or curves). The process of generating the patterns is called numeric pattern recognition assuming that variables x (*e.g.* descriptor) and y (*e.g.* activity) can take continuous values in their own definition domain. If x is an vector, the components of the vector should be orthogonal to each other and, are a highly correlated to y. The advantage of these patterns is to quantitatively predict activities. The disadvantage of these patterns is that the patterns ought to fit to an analytical formula (straight line or curve), and over-fittings are possible.[72]

(4) Patterns represented in pharmacophores. Pharmacophore is a necessary substructure feature for a drug target to recognize drug molecules. Pharmacophore consists of the set of relative positions of several special atoms or atomic groups in three-dimensional space in modern QSAR. The concept of pharmacophore has experienced the evolution from qualitative description to quantitative studies. [73] The concept evolution led to the development of 2D-QSAR into multidimensional QSAR. Its advantage is that the mechanisms of actions between drug molecules and targets are related, and the virtual screening based on this concept can produce

lead compounds with novel scaffolds.

3.2 Numeric Pattern Recognitions

In QSAR, a numeric pattern recognition is to build a classifier based on a molecular feature vector $f(\vec{\boldsymbol{x}})$, which will map the feature vector $\vec{\boldsymbol{x}}$ on to a member c_i with a tag set $\boldsymbol{C}$:

$$\vec{\boldsymbol{x}} = \{x_1, x_2, \dots, x_{M-1}, x_M\} \quad (21)$$
$$\boldsymbol{C} = \{c_1, c_2, \dots, c_{L-1}, c_L\} \quad (22)$$

where $\boldsymbol{C}$ a tag set for a classifier, L is the number of tags. For example, a set of compounds are classified into actives and inactives classis, then L=2, $\boldsymbol{C}$={0,1}. If, however, the compounds are classified into inactives, middle-actives, and high-actives, then L=3, $\boldsymbol{C}$={0,1,2}. Usually, members (tags) in $\boldsymbol{C}$ are not ordered, however, they are mutually exclusive. That is that a compound can have only one tag. A compound cannot be both active and inactive. In other situations, a compound might be assigned with multiple tags. This involves in multiple tagging issue, which will not be discussed here.

The independent variable vector $\vec{\boldsymbol{x}}$ of $f()$ has M components, and the ith component x_i represents a property (a feature value, or descriptor) of a molecule. Therefore, $\vec{\boldsymbol{x}}$ is the feature vector of a molecule.

The properties of a molecule are mathematically different, each component of $\vec{\boldsymbol{x}}$ has its own datum type, such as Boolean, integer, real, or enumeric, and its value can be continuous or discrete. Also, the property value range varies significantly. Such as, melting point value ranges from -10 ℃ to 300 ℃, while pH value ranges from 0 to 14.

Additionally, the property value distributions can be very different. The ideal distribution is Gaussian distribution; however, many property value distributions are not Gaussian (non-random distribution) in chemoinformatics.

A training data set $\boldsymbol{D}$ consists of pairs of N molecular structures and their activities:

$$\boldsymbol{D} = \{(\vec{x}_1 \to c_1), (\vec{x}_2 \to c_2), \dots, (\vec{x}_k \to c_k), \dots, (\vec{x}_N \to c_N)\}, \ c_k \in \boldsymbol{C} \quad (23)$$

where, $\vec{x}_k$ is the feature vector of the kth molecule, c_k is the kth molecule's classification tag belongs to set $\boldsymbol{C}$.

Supervised learning is a pattern recognition process, which assigns feature vectors (patterns) to all member tags in $\boldsymbol{C}$ by learning from $\boldsymbol{D}$ *per se*.

Unsupervised learning is pattern recognition process, which figures out number of intrinsic patterns by studying the input data set $\boldsymbol{D}$, in which $\boldsymbol{C}$ is not a tag set. Instead, c_k is the kth molecule's activity value (a real number).

Both learning processes are inductive, but different from reasoning analysis, which deduces the conclusion of the evolution of things based on known theorems or knowledge rules.

3.3 Pattern Recognition Algorithms

3.3.1 Naïve Bayes Classifiers (NBC)

NBC is based on Bayes' theorem [74], that is, when the number of samples is large enough to be close to the total number of objects, the probability that a property (or activity) of a sample is true will be close to the probability that the property in the total samples is true:

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} \quad (24)$$

where, *p(y)* or *p(x)* represents the probability while *y* or feature *x* is true, *p(x|y)* or *p(y|x)* is termed conditional probability.

p(x|y) is the probability of the feature *x* presents when *y* (result) presents. It is termed the priori probability.

p(y|x) is the probability of *y* presents when the feature *x* (premise) presents. It is termed the posterior probability. It is the priori probability modified with (*p(y)/p(x)*).

If the components from a feature vector ($\vec{\boldsymbol{x}}$) in conditional probability are independent to each other, then NBC is a pattern recognition algorithm based on Bayes' theorem (equation (24)).[75] The algorithm is simple with fewer parameters and theoretical errors, and insensitive to inadequate data. However, if the components from a feature vector ($\vec{\boldsymbol{x}}$) are highly correlated to each other, the algorithm can produce more errors.

NBC has been applied in the predictions for drug chemical stabilities,[76] metabolic stabilities,[77] and cellular toxicities. [78]

3.3.2 Decision Tree Classifiers

Decision tree is a supervised machine learning algorithm (knowing in advance that the object can be divided into a given number of classes, such as active and inactive; or patient and non-patient).

In a training set ***D*** (see equation (23)), each molecule is represented with a feature vector ($\vec{x}$), which has *M* components. A decision tree can be generated in the following ways (if the objects can be classified into two classes, the tree is a binary tree):

(1) Select the *i*th component $\vec{x}_I$ from $\vec{x}$, find a threshold (within its defined value domain) t_i , which divides the unclassified molecules in to groups A and B (the condition is to maximize the feature differences between A and B in terms of the feature components $\vec{x}_i$ (the *i*th *property*), thus, A and B are the new nodes (*aka* branches) in the binary tree. A and B divide up the molecules owned by their parent node in the tree.

(2) Go to step (1) until no more partitionable node in the binary tree.

A typic binary tree is depicted in Figure 14. This is a binary tree with a maximal depth of 7. The nodes that can form bifurcation are called burst nodes, and the nodes that cannot be divided again are called leaf nodes (also known as convergence nodes). Among them, age, family history, working pressure, BMI (body mass index), salty food, gender, and activity are features (descriptors) of diabetes. Among the 18 leaf nodes, only 2 nodes are pure convergence nodes. Pure convergence nodes refer to nodes without false positives or false negatives. In the case of Figure 15, these two nodes are pure convergence nodes without false positives. It can be seen that in practice, false positives or false negatives are difficult to avoid.

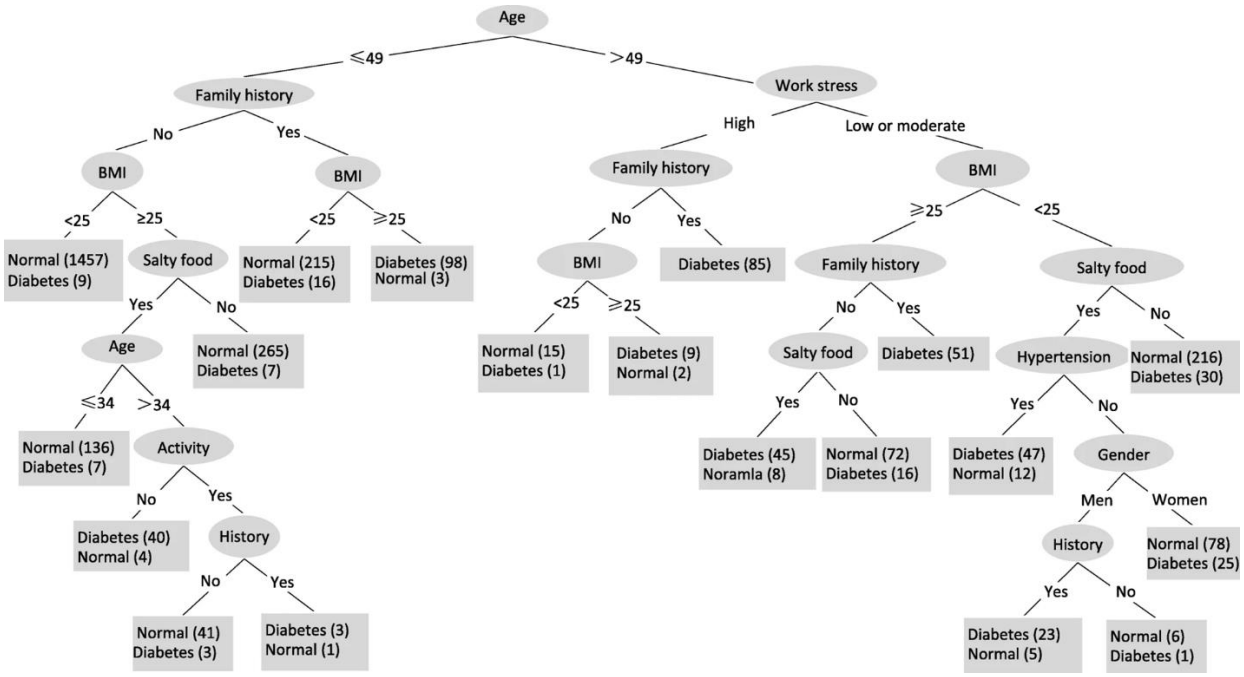


Figure 14. Binary decision tree for diabetes diagnosis (from Pei and co-workers [79])

The decision tree algorithm repeatedly divides the "non pure convergence" nodes into sub-nodes until no better convergence nodes can be generated. It is worth noting that some components of features can be applied repeatedly (such as BMI, family history, age, etc.). In order to terminate the loop, the definition of "non pure convergence" nodes needs to be flexible. If absolutely no false positive or false negative nodes must be sought as the termination condition, the algorithm may not converge.

It can be seen that the simple decision tree algorithm can divide the data into subsets and minimize the false positive and false negative rates.

For a given data set, changes in parameters such as the order of decision sequences or thresholds may produce many decision trees, namely random forest. [80] The classification performance of each decision tree is different, and the idea of the random forest is to select the best decision tree. [81] Since the decision tree algorithm transforms the problem of data division into the problem of finding the best "decision sequence", each decision tree can be locally optimized, and the global optimization needs to generate a forest of decision trees. Because there is no exploration backtracking mechanism, the best decision search of a single decision tree may fall into a "trap" and cannot pull itself out from a local optimization. [82] This problem might be solved by combining decision tree with genetic algorithm (GA). [83]

3.3.3 Hierarchical Clustering

Hierarchical clustering is common in nature.[84] Phylogenetic tree of life (Figure 15) is a typic example.[85]

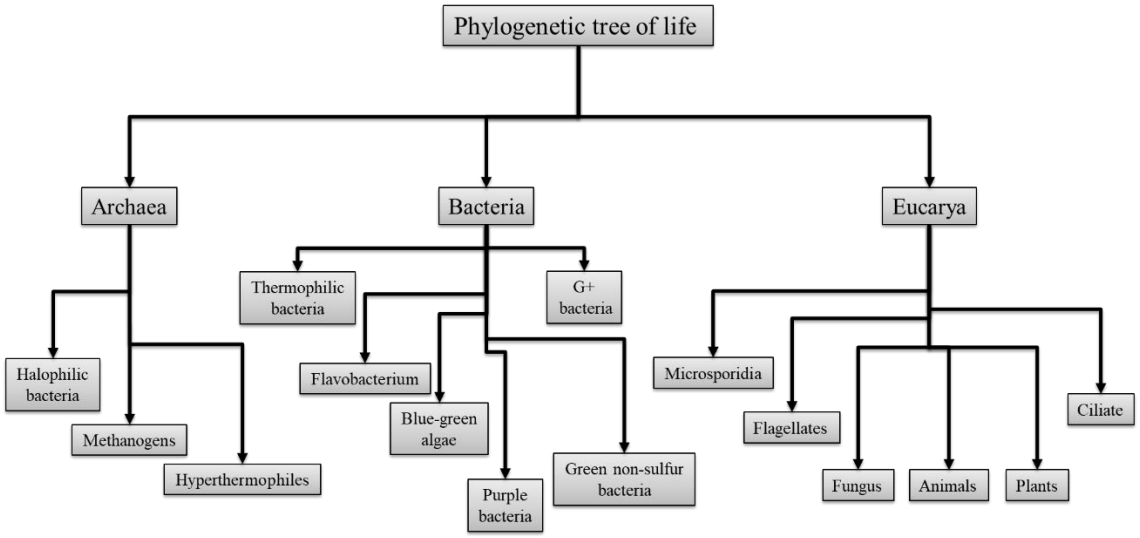


Figure 15. Phylogenetic tree of life (data from ref [85])

The typical algorithms for generating hierarchical classifications are clustering algorithms. The clustering results depends on many parameters such as the selection of molecular descriptors, distance calculations (essentially molecular similarity measures), cluster linkage strategies, and cluster scaling parameter (zoom).

Let $\vec{a}$ and $\vec{b}$ be the feature vectors of molecules A and B, and all the components in $\vec{a}$ and $\vec{b}$ have the same value distribution (Gaussian distribution), the distance (Euclidean distance) of A and B can be calculated in equation (25): [86]

$$D_E(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{25}$$

where, a_i and b_i are the ith components of $\vec{a}$ and $\vec{b}$, respectively. Fewer D_E value means A an 的 B are more similar. The similarity can also be computed as following:

$$cos\alpha_E(A,B) = \frac{\vec{a}\times\vec{b}}{|\vec{a}|\times|\vec{b}|} \tag{26}$$

Thus, the similarity is normalized to 100%.

The problem is that molecular descriptors can have very different definition domains and distributions. Lipinski descriptors (molecular weight, logP, hydrogen bonds donors and acceptors, and rotatable bonds) are typical examples. Data normalizations are required to get rid of this problem.

3.3.3.1 Normalization of Data

(1) Linear normalization:

$$x_j' = x_{min}' + \frac{(x_{max}'-x_{min}')\times(x_j-x_{min})}{(x_{max}-x_{min})} \tag{27}$$

where *x* is the original value, and *x'* is normalized value.

(1) Ratio normalization:

$$x_j' = \frac{x_j}{\sum_{j=1}^{m} |x_j|} \tag{28}$$

Here, *m* is the total number of samples.

(2) Z−score normalization:

$$x_j' = \frac{x_j - \bar{x}}{\sigma} \tag{29}$$

Here, $\bar{x}$ is average, σ is standard deviation.

If a descriptor value obeys Gaussian distribution, it can be normalized with Z−score. If the Gaussian distribution is not obeyed, Z−score normalization will distort the data patterns (that is, the variance will be far away from the standard deviation), resulting in incorrect pattern recognition.

3.3.3.2 Euclidean Distance between Data Points

Let $\vec{x}$ (with *m* components) be the feature vector of molecule X, then X is viewed as a point in *m*-dimensional space (the feature vector $\vec{x}$ is the coordinate of X). Let a compound library *L* has *n* molecules, thus, *L* has *n* points in the *m*-dimensional space (*aka* molecular diversity space).

Assume $\vec{x}$ is highly correlated with a specific bioactivity B, then the QSAR similarity postulate can be expressed as: compounds with similar activities gather together in the *m*-dimensional space at scene B.

In this way, the distance between two data points in *m*-dimensional space is a measure of the similarity between the two molecules. The shorter the distance, the higher the similarity.

A hierarchical clustering algorithm is to divide *n* data points into *C* clusters (subsets, *C* >1), so that the sum of distances among data points in each cluster is minimized (or the sum of similarities among homologous molecules is maximized), and the sum of inter-cluster distances is maximized.

However, the Euclidean distance between data points calculated by equation (25) is defective. In order to minimize the sum of distances within a cluster, it is better to calculate the distance of a point and the cluster center of the same cluster. Therefore, the Euclidean distance is modified to equation (30), which is called Mahalanobis distance (also known as generalized squared interpoint distance): [87]

$$D_M(A) = \sqrt{(\vec{a} - \vec{\mu})^T \boldsymbol{S}^{-1} (\vec{a} - \vec{\mu})} \tag{30}$$

$D_M(A)$ is the distance of molecule A and its cluster center, $\vec{\mu}$ is the compound library L's mean of the feature vectors in the same cluster, and $\boldsymbol{S}$ the covariance matrix.

If the feature vectors $\vec{a}$ and $\vec{b}$ of molecules A and B have the same value distributions, their Mahalanobis distance $D_M(A,B)$ is calculated as following:

$$D_M(A,B) = \sqrt{\left(\vec{a}-\vec{b}\right)^T \boldsymbol{S}^{-1}(\vec{a}-\vec{b})} \quad (31)$$

Therefore, an Euclidean distance is a special case of a Mahalanobis distance when the covariance matrix $\boldsymbol{S}$ is the identity matrix. If the covariance matrix is diagonal, the resulting special Mahalanobis distance is called the normalized Euclidean distance D_{SE}:

$$D_{SE}(A,B) = \sqrt{\sum_{j=1}^{m} \frac{(a_j-b_j)^2}{\sigma_j^2}} \quad (32)$$

Here, m is the number of descriptors， σ is standard deviation.

3.3.3.3 Non-Euclidean Distance between Data Points

When a molecular feature vector takes discrete values (such as eigenvalues, or bit maps), the distances of molecules in the generalized space have to be calculated with non-Euclidean geometries including:

Minkowski distance (*aka* Manhattan distance, or L^1 distance),

Chebyshev distance, and

Hamming distance.

For example, in a molecular graph, the minimal number of chemical bonds from atom A to atom B is an integer, so it is meaningless to take a non-integer value. Another example, when pedestrians walk from one intersection to the next in Manhattan, the distance they have to walk is the sum of street-block edge lines, not street-block diagonal lines. Therefore, the distance of $\vec{a}$ and $\vec{b}$ (discrete variable vectors) is termed Minkowski distance:

$$D_{L1}(A,B) = ||\vec{a}-\vec{b}|| = \sum_{j=1}^{m} |a_j - b_j| \quad (33)$$

When the components of $\vec{a}$ and $\vec{b}$ take discrete values and, abide certain “traffic rules” (for examples, some streets are one-way streets, and some intersections are not allowed to turn left or right), the distance between two points cannot be calculated with European distance or Manhattan distance algorithms. Playing chess on a chessboard is a typical example. At this time, the minimum number of moves required from one point (square) on the chessboard to another point (square) becomes the distance between the two points, which is called Chebyshev distance. Extended to generalized space, Chebyshev distance is the value of the maximal difference between two sets of coordinates $\vec{a}$ and $\vec{b}$:

$$D_C(A,B) = MAX_{j=1}^{m} |a_j - b_j| \quad (34)$$

When a molecular feature vector is a string (for examples, chemical structure linear notation SMILES, or a gene sequence), the distance between molecules A and B is actually the minimal numbers of operations required to convert one string into another string. The calculated distance is called Hamming distance, which was originally proposed by Richard Hamming, an American information scientist, [88] in 1950. Hamming distance is mainly used to process bit string data. The calculation of Hamming distance needs to be treated in a specific way. There is no unified

algorithm and unified mathematical expression.

3.3.3.4 Cluster Merging Strategies in Clustering Algorithms

In hierarchical clustering algorithm, *m* clusters in the next level need to be merged into *n* clusters ($n<m$). Merging is based on the distances between different clusters. There are two merging strategies:

(1) Agglomerative method. Starting from the original data points, the nearest data points are merged into a cluster from bottom to top. As shown in Figure 15, Halophilic bacteria, Methanogens, and Hyperthermophiles are merged to form a up-level group Archaea. The groups are further merged until to form one class.

(2) Divisive method. Starting from the top, the whole data set is divided into sub-clusters from top to bottom, and then divided into sub-clusters repeatedly until each data point forms a cluster by itself.

In the aggregation method, according to the cluster linkage strategy, hierarchical clustering includes single-linkage, complete-linkage, and average-linkage, and centroid-linkage, and other clustering strategies. [89] Agglomerative hierarchical clustering methods, such as Ward clustering algorithm [90] are more commonly used in chemoinformatics because they can produce more balanced and reliable clusters (especially compared with the numerical based non-hierarchical cluster analysis method introduced in the next section).

3.3.4 Non-hierarchical Clustering

3.3.4.1 K-means Clustering and *k*-nearest Neighbor Clustering

Natural laws are both hierarchical and non-hierarchical. For example, shape-recognizing the handwritings of Arabic numeral symbols from 0 to 9 is not hierarchical.

Jarvis-Patrick clustering [91] was broadly employed in early chemical diversity analyses. [92] This is a *k*-nearest neighbors algorithm (*k*-NN): if two molecules *i* and *j* share more than *k* (predefined integer) nearest neighbors, then *i* and *j* belong to the same cluster. Hence, *k*-NN is a supervised learning method. Because the algorithm is based on distance, normalizing feature vectors is critical. Because the algorithm is sensitive to local data structures, *k*-NN can produce imbalanced clusters.

In order to overcome the deficit of imbalanced clusters, K-means clustering (another unsupervised non-hierarchical clustering method) was proposed. K-means clustering can be confused with *k*-NN. K-means clustering assumes that the data can be divided into K (also predefined integer) clusters, randomly selects K cluster centers (or seeds), and then calculates average distances for the K clusters. If the position of a molecule is closer to an average of one cluster rather than the seed of the cluster, then the molecule can be partitioned into other clusters. This clustering algorithm is fast, but the result varies on the initial random seeds and the number assigned to K.

In a word, *k*-NN and K-means (or K-median) clustering are non-decisive, which need to be tried and optimized, and it is difficult to avoid the interference of human factors.

Non-hierarchical clustering process for a compound library classification consists of following steps:

(1) Derive molecular descriptors from the molecular connection tables;

(2) Select principal components from the descriptors by means of principal component analysis (PCA) or factor analysis (FA); [93]

(3) Normalize the main components to make them comparable in values;

(4) Select a similarity or distance metric to calculate the similarity or distance between a pair of molecules;

(5) Select a clustering algorithm to classify molecules in the library.

These clustering algorithms are based on numerical pattern recognition, which cannot directly figure out how many clusters in a compound library. The clustering result depends on many parameters, such as descriptor selection, data normalization, similarity metric, initial clustering centers (random selection), selecting a number of shared nearest neighbors, selecting stratification threshold (in hierarchical clustering algorithms).

In addition, because the computational complexity of K-means algorithms is the factorial of the number of data points, the number of iterations must be limited to make the calculation feasible. However, the condition of potential infinite iteration is that the user must specify the number of clusters before a clustering process starts.

It can be seen that the key to a clustering algorithm based on feature vectors is to guess (if it is unknown) the number of clusters in the data set in advance.

If a high-dimensional feature vector can be reduced to the 2-dimensional or 3-dimensional space that can be seen by the naked eye, we could have a better judgment on the problem divided into several clusters. This kind of dimensionality reduction technology mainly includes principal component analysis (PCA) [94] and artificial neural network dimensionality reduction.

3.3.4.2 Dimension Reduction by Principal Component Analysis

PCA is often used to reduce the dimension [95,96]. It uses orthogonal transformation to reduce $m \times n$ matrix to a new matrix in which the components in the new n-dimensional vector are independent to each other. In order to transform the covariance matrix of the original data matrix into a diagonal matrix, the original coordinate system is transformed into a new orthogonal coordinate system, so that it points to the p orthogonal directions where the sample points are scattered most widely ($p<n$). The essence of PCA is to replace the original feature vector $\boldsymbol{Y}(y_1, y_2, \ldots, y_n)$ with a new feature vector $\boldsymbol{X}(x_1, x_2, \ldots, x_n)$. Every new component x_{ij} is a linear combination of the original components $y_1, y_2, \ldots, y_n$:

$$x_{ij} = \sum_{j=1}^{n} c_{ji} y_{ji} \tag{35}$$

Here, $\boldsymbol{c}$ is the parameter matrix derived by PCA. If a component $\boldsymbol{c_i}$ of $\boldsymbol{c}$is close to zero, meaning the information contribution from this component can be ignored, and the corresponding new variable can be excluded. Thus, all components in $\boldsymbol{c}$ are sorted based on their information contributions to the new matrix, if the top-2 or top-3 components in which their information contribution to the new matrix exceeds significantly (say greater than 85%), the PCA dimension reduction is successful, the first two or three principal components can be used to graph the matrix. The data matrix is reduced to 3 or 2 dimensions, and the topology of the data points is approximately preserved, and the dimension reduction is successful.

It is worth noting that when using PCA, the number of data points (m) should be greater than the number of descriptors (n). In chemoinformatics, thousands of molecular descriptors can be available, the number of data points can be much less. In this case, the number of descriptors

should be compressed before dimensionality reduction with PCA.

3.3.4.2 Dimension Reduction by Kohonen Neural Networks

Kohonen neural network, also known as self-organization map (SOM), [97] belongs to unsupervised machine learning technology. SOM maps high-dimensional data to two-dimensional space, but clusters the data. SOM adopts competitive learning mechanism, with only input layer and output layer (also known as feature map). At first, the neuron w_{ij} of the characteristic graph is initialized by random numbers. Then, each data point x_{ij} finds the neuron that is most similar to itself (the shortest distance between x_{ij} and w_{ij}) in the feature map and "resides" until all the data in the training set find their own attributions. The resulting two-dimensional SOM map automatically aggregates molecules with similar feature vectors.

Gasteiger and his team composed six parameters related to molecular static property (atomic partial charge q_σ, total molecular charge q_{tot}, σ-electronegativity χ_σ, π-electronegativity χ_π, lone electron pair electronegativity χ_{LP}, and atomic polarization α) together with topological autocorrelation vectors, and form a new molecular feature vector. SOM was applied to project the molecular feature vector into a two-dimensional map, and dopamine agonists and benzodiazepine receptor agonists are automatically divided into different areas of the SOM map as shown in figure 16. [98]

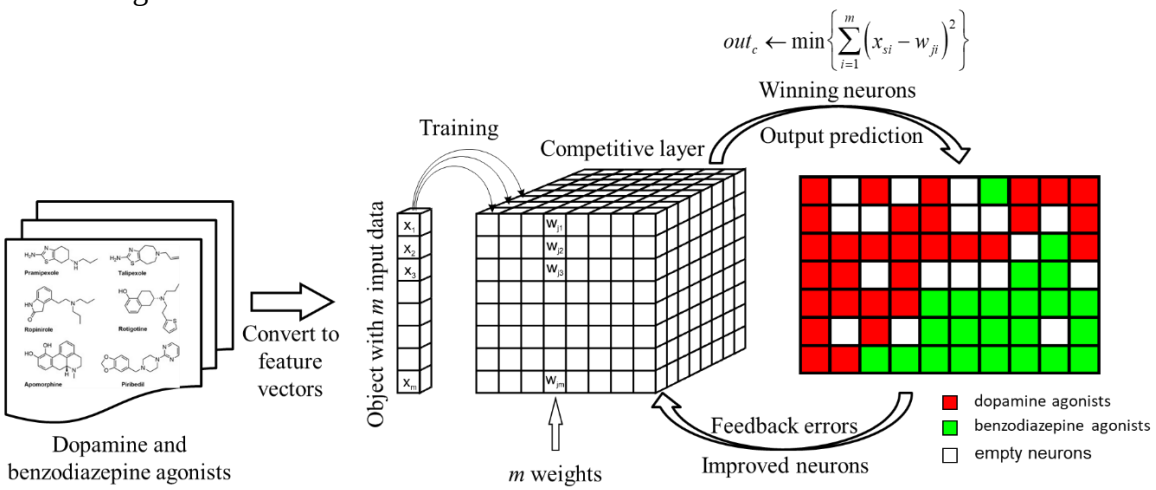


Figure 16. SOM maps two different types of molecules represented by feature vectors to 2D space, and manifests two different types of molecules into two areas. Redrawn based [98].

SOM and its applications in chemoinformatics was reviewed by Gasteiger and co-workers.[99]

3.4 Pattern Recognition Based on Molecular Topology

As mentioned above, hierarchical clustering algorithms need a pre-defined stratification threshold. Non-hierarchical clustering algorithms need pre-defined parameters manifesting the guesses the number of clusters in the data set for a compound library. However, a chemist cannot define these parameters before obtaining the results from the clustering algorithms.

In order to resolve this paradox, a scaffold based classification approach (SCA) [100] was proposed. According to SCA, a molecular structure can have three topological bond types as shown in Figure 18:

(1) chain bond: one end of the bond is directly or indirectly connected to a ring atom, while the other end is directly or indirectly connected to a monatomic substituent;

(2) ring bond: atoms at both ends of the bond are in the same ring;

(3) linker bond: one end of the bond is directly or indirectly connected to a ring atom, while the other end is directly or indirectly connected to atoms in a different ring.

Thus, a molecular scaffold can be defined as a structure without chain bonds. The definition usually conforms to chemical experience.

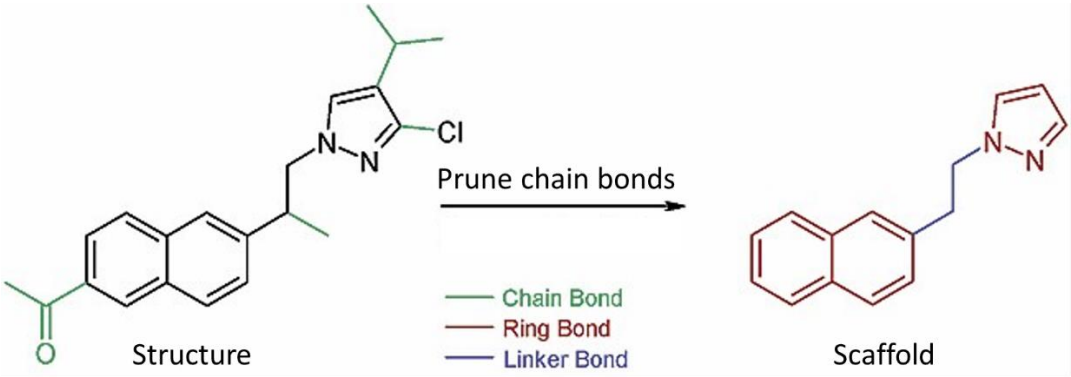


Figure 18. Scaffold defined by SCA

In this way, with graph theory algorithm, SCA derives *n* scaffolds from a compound library containing *m* structures, then assigns scaffold IDs to the *m* structures based on their scaffolds, therefore the library is divided into *n* clusters. Each scaffold is a cluster center for a cluster of compounds. Acyclic compounds have no ring containing scaffold, which can be divided into:

(1) saturated acyclic compound group, which has no scaffold (i.e. a cluster of compounds without a cluster center),

(2) unsaturated acyclic compound group (fragments containing unsaturated bonds are regarded as scaffolds).

The steps of SCA clustering are as follows:

(1) Scan all molecular structures in a compound library to generate non-redundant scaffolds used the rules described in Figure 18. Let *n* be the total number of scaffolds;

(2) Sort the scaffolds in ascending order of molecular complexities, assigned an ID (*aka CID*, and $CID \in (0..n)$) to each scaffold based on the complexity order.

In this way, a compound library with *m* molecules is clustered into *n* +1 clusters, and the compounds in cluster 0 holds saturated acyclic compounds (a cluster without scaffold). Any compound S_j has a complexity (S_j), which can be computed as follows:

A virtual complexity feature vector V_V for the *n* scaffolds (cluster centers) is calculated. V_V consists of four descriptors, namely, the smallest set of small rings (*SSSRS*) [101], the number of heavy atoms (*Hatoms*), the number of non-hydrogen bonds in a molecule(*Bonds*), the sum of non-hydrogen atomic numbers (*San*). Let V_J be the complexity feature vector of the *j*th molecule,

$$V_j = \{SSSRs_j, Hatoms_j, Bonds_j, San_j\} \quad (36)$$

Thus, the virtual cluster center feature vector V_v is computed as the following,

$$V_v = \left\{ \max_{i=1..n}(SSSRs_i), \max_{i=1..n}(Hatoms_i), \max_{i=1..n}(Bonds_i), \max_{i=1..n}(San_i) \right\} \quad (37)$$

Then, the complexity of *j*th compound S_j is computed in (38),

$$Complexity(S_j) = \frac{\|V_v + V_j\| - \|V_v - V_j\|}{\|V_v + V_j\|} \quad (38)$$

(3) For compounds in a same cluster, their structural differences on the substituents. Its cluster center (scaffold) has no substituents, and its *cyclicity* is the highest. The *cyclicity* for other compounds in the same cluster is based on the comparison against the cluster center. The feature vector of *cyclicity* consists of six descriptors: the number of heavy atoms (*Hatoms*), the number of rotatable bonds (*RBs*), the number of single substitute atoms (*SAs*), the number of double substitute atoms (*DAs*), the number of double bonds (*DBs*), and the number of triple bonds (*TBs*). Thus, the feature vector for cluster center V_{cs} is defined as following:

$$V_{cs} = \{Hatoms_s, RBs_s, SAs_s, DAs_s, SBs_s, TBs_s\} \quad (39)$$

where, subscript s denotes scaffold, the *cyclicity* feature vector (V_{cj}) of a compound *j* in the same cluster (S_j) is defined as the following,

$$V_{cj} = \{Hatoms_j, RBs_j, SAs_j, DAs_j, SBs_j, TBs_j\} \quad (40)$$

Thus, *Cyclicity*(S_j) is computed as the following:

$$Cyclicity(S_j) = \frac{\|V_{cj} + V_{cs}\| - \|V_{cj} - V_{cs}\|}{\|V_{cj} + V_{cj}\|} \quad (38)$$

At this point, each compound S_j in a compound library has both *complexity* and *cyclicity* values (*complexity* (S_j), *cyclicity* (S_j)). That is, SCA maps the compound library into two-dimensional space. The map is termed as SCA-plot.

SCA-plot has interesting characteristics, such as the computation is fast, it is suitable for the chemical diversity comparison of large compound libraries, compounds with the same scaffold are distributed on the same curve according to different *cyclicities*, scaffolds with similar substitution patterns are distributed on the same curve, compounds with higher *cyclicity* and simpler scaffold aggregate in the upper left corner of SCA-plot, compounds with higher *complexity* and fewer substituents aggregate in the upper right corner of SCA-plot (such as caged or fused ring compounds are located here), compounds with longer chains and simpler scaffolds (such as soap-like agents or fatty acid compounds) aggregate in the lower left corner of SCA-plot, compounds with long chain and complicated scaffolds aggregate in the lower right corner of SAC-plot, however, this corner is left empty due to the infeasibility for synthesizing these types of compounds, which are not drug-like.

Since *CID*(S_j) (integer) calculated by SCA is monotonically consistent with the cycle *Cyclicity*(S_j) (real), in practice, SCA-plot can be drawn with *Cyclicity*(S_j) and *Complexity*(S_j). The advantage of this type of SCA plot is that the complexity coordinates do not cause data point overlapped. An example of SCA-plot is depicted in Figure 19. [100]

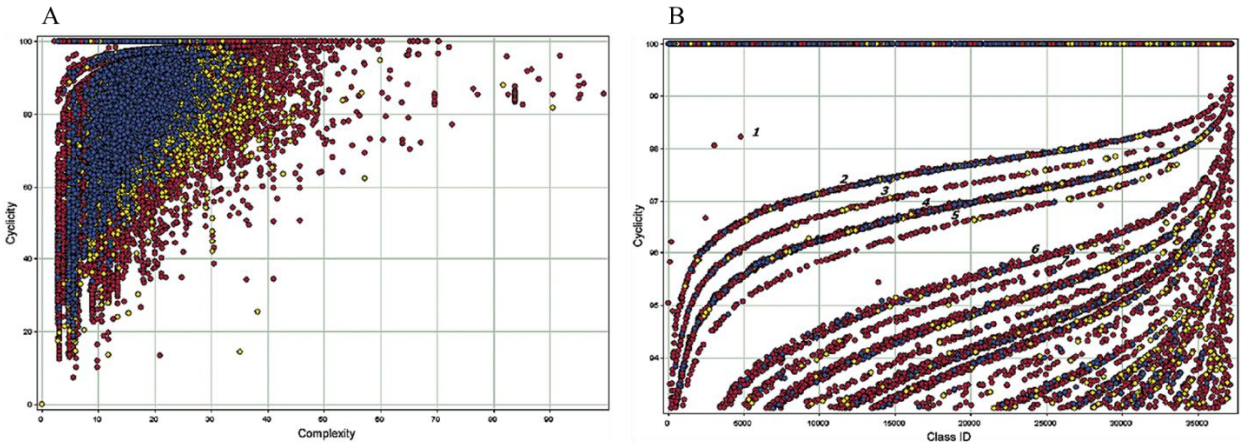


Figure 19. SCA-plot。A: Comparison of chemical diversity for four libraries ACD (red), NCI (green), MDDR (yellow), and CMC (blue); B: The X-axis is CID (locally enlarged view) of chemical complexity with CID. There are naturally formed curves in the figure, on which the numbers 1, 2, 3... manifest scaffolds with single-, double- and triple- substituents.

3.5 Artificial Neural Networks

3.5.1 Neurons and Neural Networks

Neurons are the main functional units of human nerves. Although they differ in shape and function, they have four main components:

(1) Cell body: the main body of a cell, which is equivalent to the central processing unit (CPU) of a computer;

(2) Dendrite: multiple protrusion branches sent from the cell body. Protein receptors receive external signals or signals sent by other neuronal synapses, which can extend hundreds of microns;

(3) Axon: a branch from the cell body, which can extend more than 1 meter. Its surface is wrapped (electrically insulated) by a cholesterol rich membrane with mitochondria in it Axoplasm that delivers neurotransmitters. Each neuron has only one axon responsible for transmitting output signals;

(4) Synapse: the bifurcation point of axon terminals, which contacts the cell bodies or dendrites of multiple other neurons to output signals, resulting in excitation or inhibition of other neurons. The neurotransmitters (chemical messengers) that output signals are called chemical synapses; those that output signals with electrical signals are called electrical synapses (common in fish).

Neurons have three functional types:

(1) Sensory neurons: collect touch, sound, light, chemical and other signals, and transmit the signals to the central nervous system;

(2) Motoneurons: receive command signals from the central nervous system and control physiological functions, such as muscle contraction;

(3) Intermediate neuron: connect other neurons to form a neural network to form memory and computing ability.

The dendrites of neurons sense mechanical movement (tactile) from the environment sound, light, chemical signals, whose intensity is measured in continuous value. Taking chemical signal

as an example, the signal is expressed by the concentration (from nM to μM) of signal molecules (such as hormones and messenger molecules). Signals enter the cell body to change the concentration or activity of molecules. After the change exceeds the threshold, electrical signals are generated, and signals are output to the next neuron through axons. Therefore, neurons have electrical excitability. Neurons are actually gated switches that convert analog signals into digital signals (Figure 20). When the input signal strength is lower than a threshold, the axon has no signal output, and the gating switch is in a silent state. When the input signal strength exceeds a threshold, the axon has a signal output, the gating switch is turned on, and the neuron will produce an all or nothing electrochemical pulse, called action potential. This potential propagates rapidly along the axon and activates synaptic connections when it reaches the axon. Synaptic signals may be excitatory or inhibitory, increasing or decreasing the net voltage reaching the cell body. It is transmitted to other neurons to regulate the functions of other organs of the body.

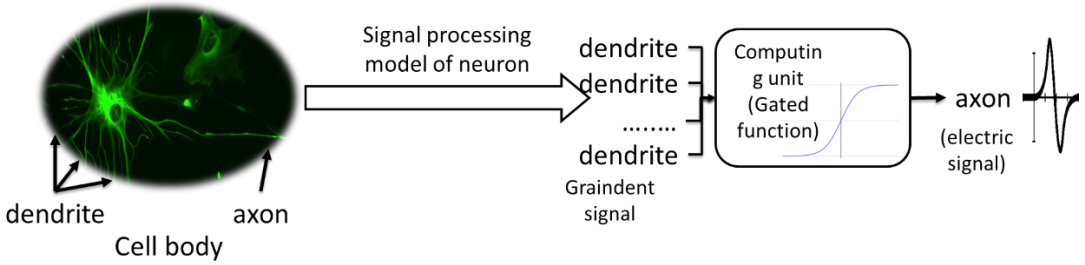


Figure 20. A neuron and its signal processing process

In most synapses, signals are transmitted from the axons of one neuron to the dendrites of another. However, axons or dendrites can also be linked to each other to form neural networks. Neurons are differentiated from neural stem cells during brain development in childhood. They are basically fixed after adulthood without increase.

3.5.2 Principles of Artificial Neural Networks

In 1943, Warren McCulloch and Walter Pitts published a paper on logical calculus of internal thoughts in neural activities, and proposed the mathematical model of neural network. [102] In 1949, Donald Hebb proposed the Hebb learning mechanism in order to explain the behavior of associative learning (also known as Hebbian learning). He believed that when the axons of *a* cell were close enough to *b* cell and, stimulated and activated B cell repeatedly, lasting metabolic changes would occur in the cells to improve the efficiency of cell activation. Simultaneous activation of cells leads to a significant increase in the strength of synaptic connections between cells. This is the basis for neurons to achieve unsupervised learning. [103] In 1954, B. Farley and W. Clark of MIT first proposed a computer program to simulate the learning mechanism of Hebb neural network. [104]

Artificial neural network (ANN) is a nonlinear pattern recognition technology inspired by the functional model of neurons established by biologists. ANN consists of a set of simplified mathematical models of neurons. An artificial neuron is a function, which consists of a set of inputs (equivalent to dendrites), a computing unit (equivalent to a cell body) containing a gating function and a set of outputs (equivalent to a synapse). Each neuron (function) is a node in an ANN, and the output of one neuron is used as the input of another neuron, thus establishing the

connection between neuron nodes and forming an ANN. The connection of neurons has a weight, which determines the influence intensity of one neuron node on another.

Therefore, ANN is a function network formed by nonlinear functions in parallel or series. To build an ANN, one has to decide which type of function to use, and what kind of connection between functions (called ANN architecture). Training an ANN with the data of marked premise-conclusion pair (for example, the premise can be molecular structure data, and the conclusion can be a certain property / activity data) to determine the weight required for the connection between functions (neurons), establish a neural network model to predict a certain property / activity based on molecular structure, and save the learning results with the determined ANN network architecture and the connection weight between network nodes.

3.5.3 ANN and von Neumann Architecture

In 1945, J. Presser Eckert and John W. Mauchly of the University of Pennsylvania invented the first computer, ENIAC. Based on their work, John von Neumann summarized the basic architecture of modern electronic digital computers (Figure 21, redrawing based on en.wikipedia.org/wiki/Von_Neumann_architecture).

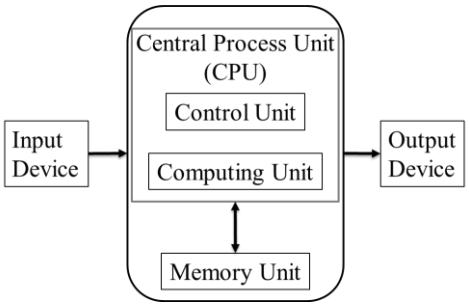


Figure 21. Von Neumann architecture of computer.

The von Neumann architecture features with that data and programs are stored in the same address space of computer memory, CPU operation unit performs arithmetic / logic operations, control unit includes instruction register and program counter, memory unit (memory) stores data and instructions, and external mass storage devices and input and output devices.

Although mankind has already entered the era of billion times of ultra-high-speed computing, any modern computer is still based on von Neumann architecture, and quantum computing is no exception. [105] Due to sharing a common bus, instruction acquisition and data operation cannot occur at the same time. This is called the von Neumann bottleneck, which limits the performance of modern computer systems.

Comparing Figures 20 and 21, the calculation of a single neuron is consistent with von Neumann's architecture. To our best knowledge, the human brain has 86 billion neurons and about the same number of non-neuron cells. [106] They form complicated networks and perform parallel computing, which is the future goal of human made computers.

3.5.4 Drug Design, QSAR and ANN

3.5.4.1 Questions in Drug Design

Modern drug design relies on the understanding of mechanisms of actions at molecular and atomic levels. The essence of drug molecular interaction mechanism is the interaction between receptors (generally biological macromolecules) and ligands (which can be biological macromolecules or small molecules). Therefore, the basic problems of drug design can be summarized as follows:

(1) Molecular recognitions and related algorithms: Molecules are graphs *per se*. Studying the molecules involving graphical (topological) pattern recognizing isomorphism graphs (structure retrieval. structure encoding), homomorphic graphs (substructure search, chemical structure database search engine), equivalence graphs (bioelectronic isosteric body, scaffold-hopping), graphical features (stereochemistry, chemical reaction expression, chemical synthesis design), hypergraphs (Markush / general structure search, patent search engine), fuzzy graphs (protein NMR spectrum elucidation), similar graphs (QSAR modeling), clustered graphs (molecular diversity analysis), generated graphs (two-dimensional structures, three-dimensional structures and conformations), superimposed graphs (graph coincidence and graph complementarity);

(2) Macromolecules related algorithms: sequence alignments, folding predictions (α - helix. β – sheet, loop, co-evolutionary residue pairing), tertiary structure predictions (*ab initio* 3D structure prediction, homology modeling), quaternary structure prediction (domain, thermal stability, and protein-protein binding), structural superposition (geometric superposition, active region superposition), structural complementarity (ligand receptor flexible docking), polymerization (dimer, polymer and function, prediction of active sites);

(3) Intermolecular interaction: inhibition (ligand inhibition. Ligand coordination), activation (ligand directly binding active sites), allosteric action (ligand indirectly binding active sites), flexible docking, shape complementarity, pharmacophore complementarity (steric, static and lipophilic complementarities), membrane protein bindings (GPCR, ion channel and other proteins), environmental effects (metal ions, water molecules, and multiple ligands)

(4) System problems: gene annotations, function predictions, proteome cluster analysis, biological function predictions, constructions and analyses of signal transduction networks, metabolic pathway network analyses, cell simulations, transitional state analyses of catalytic reactions, virtual screening, and target identification and validation;

(5) Application problems: personalized medicine, precision medicine, drug-drug interactions, chemical synthetic planning, ADMET parameters predictions, and physical chemistry, biochemical stability predictions etc.

Conventional QSAR and AI technologies have been employed in the above-mentioned problems, especially in predicting parameters related to ADMET.

3.5.4.2 Important Parameters for ADMET

Absorption

Passive diffusion (driven by the concentration difference between the two sides of the cell membrane) membrane penetration and active transport (pulled in by transporters on cell membrane) are two classes of drug absorptions in cells.

Passive diffusion absorption can be estimated by molecular properties such as logD or hydrogen bonding ability. If the molecule is a substrate of a transporter (e.g., cytochrome CYP3A4. P-gp glycoprotein), it belongs to active transported absorption, and there is no suitable

prediction model.

Caco-2 or Madin Darby canine kidney (MDCK) monolayer membrane test is used to evaluate oral drug absorption *in vitro*. The bovine microvessel endothelial cell (BMEC) model is an *in vitro* model of the ability of molecules to cross the blood-brain barrier and be absorbed. Calculation methods to predict oral absorption generally use logP or logD, PSA, hydrogen bonding ability as feature descriptors, and multiple linear regression (MLR), partial least square (PLS) method and, ANN is used to build a absorption predictive models.

Hydrogen bonding ability has an essential impact on oral absorption. Gastro Plus of Simulation Plus is a commercial program for predicting drug absorption [108].

Bioavailability

The dominant factors of bioavailability are the abilits of drugs to be absorbed and the first pass metabolism of drugs in the liver. Drug absorption is closely related to molecular solubility and cell permeability, intestinal wall transporters and metabolic enzymes. The cell permeability of drugs is the function of molecular size, hydrogen bonding ability, lipophilicity, molecular shape, and molecular flexibility (can be measured by number of rotatable bonds).

Blood brain barrier penetration

Drugs for the central nervous system (CNS) diseases need to cross the blood brain barrier (BBB), while other types of drugs should avoid penetrating the BBB to prevent side effects on the CNS.

The prediction of BBB permeability requires experimental data of drug uptake by the brain, such as drug concentration in the whole brain, drug concentration in extra cellular fluid (ECF) or cerebrospinal fluid (CSF). These data were obtained through micro-dialysis experiments. [109] The permeability of blood-brain barrier and the time interval after administration are also crucial. These problems lead to the limited experimental data that can be used for predictive modeling. The BBB permeability of drugs is related to molecular weight and polar surface area (PSA). It is generally believed that molecules with molecular weight <450 and PSA<100 $Å^2$ are more likely to pass through the BBB. The predictive models are usually based on the physical and chemical parameters such as logP, the number of hydrogen bond receptors and PSA using multiple linear regression approaches. [110,111]

Protein transporters regulations

Protein transporters are on cell surfaces and actively absorb or eliminate endogenous or exogenous molecules to maintain the balance of cell survival. [112] If a drug molecule is a substrate of a protein transporter, the ADME properties of the drug cannot be predicted by the passive diffusion model. P-gp of the ATP binding cassette (ABC) family is the main protein transporter family, which actively transports antitumor drugs from cells, resulting in multiple drug resistance (MDR) of tumors. Other protein transporters that may affect drug absorption include the MDR associated proteins (MRP1 and MRP2), breast cancer resistance protein (BCRP), the peptide transporter (PepT1), and the apical sodium dependent transporter (ASBT). The data of receptor ligand binding mode should be used to predict transporter mediated drug molecular absorption, and accurate simulation results cannot be obtained only by the descriptors from ligand molecules (such as hydrogen bond receptor, molecular weight, lipophilicity, amino group, molecular surface area, and polarizability).

Dermal and ocular penetration

Some drugs are delivered through skin or eyes. The feature descriptors for their absorption prediction model are similar to the one for oral absorption or BBB penetration predictions. Namely, logP, and parameters related to water solubility (hydrogen bond, molecular weight and,

molecular flexibility). The QikProp of Schrödinger, the DermWin of EPA are the commonly used software packages for this purpose.

Plasma protein binding (PPB)

Free state drug molecules pass through cell membrane and bind to their targets to exert efficacy. [113] It is important to know the drug molecules that bind with plasma proteins. In blood, except for various particles such as red blood cells, leukocytes, and platelets bind to drug molecules, albumin (binds to acidic drugs) α1-acid glycoprotein (binds to alkaline drugs), lipoprotein (binds to neutral and alkaline drugs) α, β, γ - globulins will bind to drug molecules at pH 7.4, the plasma protein binding rate (the ratio of bound drugs to unbound drugs) shows a sigmoidal curve against logD. [114]

Volume of distribution (V_d)

The half-life of drugs *in vivo* is determined by the distribution volume and clearance rate. Although $\log V_d$ is not correlated with logD, $\log V_{du}$ (plasma free drug molecular volume) is linearly correlated with logD. [114] V_d should be the feature parameter, which is the function of $\log D_{7.4}$, ionization constant (pK_a) and PPB. [115]

Plasma clearance (CL)

CL is the total volume of plasma cleared by the drug per unit body-weight and time, and measured in L/(kg • h). It is related to V_d , blood drug half-life $t_{½}$, together with drug administration frequency. It is also believed that human liver drug clearance and rat liver clearance data can be used to predict CL. [116]

Plasma half-life, $t_{½}$

After a drug is administrated, the time for the blood-drug concentration reduced to half is called drug plasma half-life ($t_{½}$), which determines the frequency of drug administration. Usually, the frequency is determined by both CL and V_d. It is known that CL and V_d are the functions of logP, pK_a, molecular weight or molar refraction.

Lipophilicity

Lower water solubility and higher lipophilicity of molecules lead to lower oral bioavailability. Hydrophilic molecules have poor cellular permeability and absorption. Ionization constants directly affect solubility and lipophilicity. The physicochemical parameters related to lipophilicity, membrane permeability, drug absorption, distribution and clearance pathways are automatically measurable. The gold standard of lipophilicity is the partition coefficient in octanol / water system, that is logP. Other experimental measurements are immobilized artificial membranes (IAM), immobilized liposome chromatography (ILC), liposome/water partitioning. In addition to logP, $\log D_{7.4}$ or $\log D_{6.5}$ (pH 7.4 [blood pH] or pH 6.5 [intestinal pH]) are better parameters for lipophilicity, especially for ionizable drug molecules. Although, there are fewer data or predictive models for logD.

Solubility

Oral drugs become active after disintegrated in digestive tract. Low solubility is not conducive to oral absorption. Solubility can be measured by turbidimetry and nephelometry. At present, there is no accurate solubility prediction program, and the main problem is lack of experimental data measured under unified conditions.

pKa

Solubility, lipophilicity, cellular permeability and absorption are all effected by pKa. Many experimental pKa data and models are available. Popular chemoinformatics packages such as, ACD/pKa, Pallas/pKa, SPARC can predict pKa.[117]

Hydrogen binding

The ability of drug molecules forming hydrogen bonds is the determinant of their ability to penetrate cells. Drug molecules need to overcome the hydrogen bond formed with water to penetrate cells. At first, people used ΔlogP (the difference between $logP_{octanol/water}$ and $logP_{alkane/water}$) as a measure of hydrogen bonding ability, but many molecules have poor solubility in alkane phase, which makes the experimental measurements difficult. The hydrogen bonding capacity of molecules can be estimated, through the number of nitrogen atoms and oxygen atoms, and PSA. These descriptors can be calculated in most chemoinformatics software.

Cellular permeability

The ability of molecules penetrating cell membranes can be predicted by experimental models, such as Caco-2 cells (human intestinal cell absorption model), and the parallel artificial membrane permeability assay (PAMPA), Surface plasmon resonance (SPR) measures the bindings of molecules to liposomes.

3.5.4.3 Complexity of Multi-parameters Decision Making

Many chemoinformatics software packages, such as Cerius[2], Dragon, and Molconn-Z [118], have modules to calculate ADMET parameters. The ultimate goal of these calculations is to support decisions in drug development processes. There many ADMET parameters, in which exist complicated relations among them as shown in Figure 22. The parameters at the upper layer of the decision tree/network can be viewed as the functions of the parameters at the lower layer of the tree/network.[119] Sometimes, the parameters in the same layer are not independent to each other.

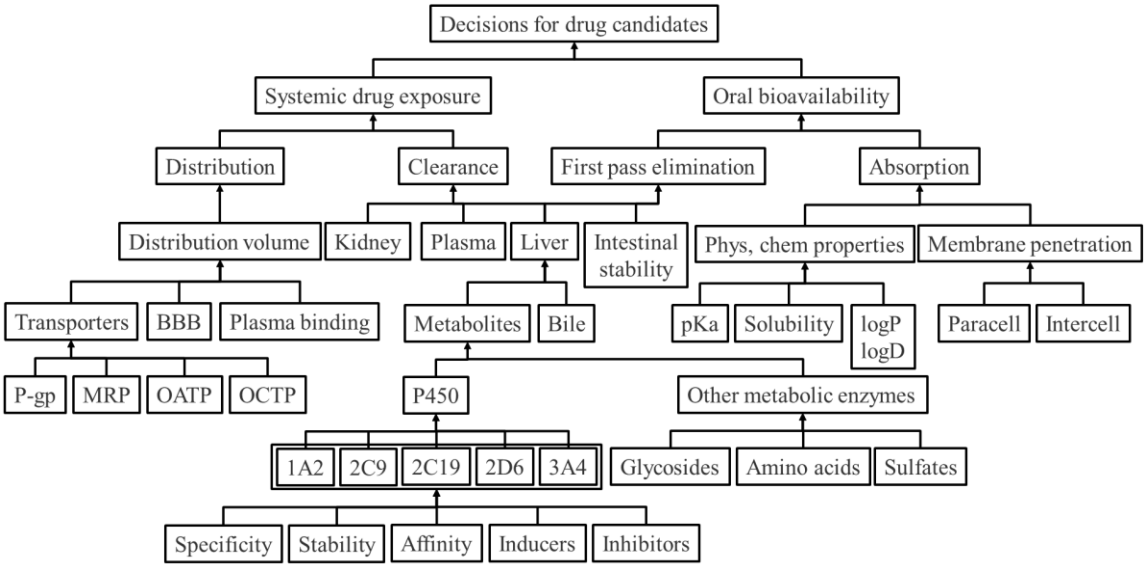


Figure 22. Drug candidate decision network based on ADME data (Based the data from ref.[120])

Such multi-parameters decision-making is a one of grand challenges to conventional QSAR. Hence, QSAR trended to explore the applications of various AI approaches after 1980s, such as recursive partitioning (RP), supporting vector machines (SVM), SOM and ANN.

After 2010s, the era of QSAR-based drug design methodology has been moving towards history. Artificial intelligence aided drug design (AIDD) is coming.

3.5.5 Memory Footprint: Protein Structure Prediction

3.5.5.1 Brief History of Protein Structure Prediction

The three-dimensional structure data of proteins are the fundation of structure-based drug design (SBDD). Using experimental methods (such as NMR, X-ray crystallography, low temperature electron microscopy) to determine the three-dimensional structure of proteins are costly, but require months to years of efforts. [121] As of The 2nd of August 2022, although 193,455 proteins have been tested experimentally (www.rcsb.org), they still account for less than one thousandth of the total proteins to be tested. Therefore, prediction of protein structure by computational methods is often the only feasible solution.

In 1961, American biochemist Christian Anfinsen proved through experiments that ribonuclease A was immersed in urea solution to destroy the four pairs of disulfide bonds that maintain the protein structure, resulting in the denaturation of ribonuclease A. While urea was removed from the solution, as a result, ribonuclease A automatically restored its physiological function, that is, the three-dimensional structure was recovered. Therefore, Anfenson believed that although proteins also need ribosomes mRNA, chaperones to help them to fold correctly, and the tertiary structure information of proteins should be encoded in only one-dimensional amino acid sequences, known as Anfinsen's dogma, which has become the theoretical basis for *ab initio* protein structure prediction. [122]

In 1969, Cyrus Levinthal, an American molecular biologist, pointed out that even for small peptides with 100 residues, it would take astronomical time to find the natural three-dimensional folding conformation through random folding, while biochemical experiments showed that protein folding only took from microseconds to hours. [123] The reasonable explanation for this paradox (Levinthal's paradox) can only be that protein folding cannot be a completely random process, and the three-dimensional structure of a protein should be encoded in a gene sequence.

There are two strategies to predict the three-dimensional structure of proteins from sequences: predictions based on the physical laws and predictions based on statical fitting to experimental data.

3.5.5.2 Physics that Drive Protein Folding

According to traditional methods, in order to predict the three-dimensional structure of proteins, it is necessary to fully understand the physical laws driving protein folding [121] and the statistical laws of protein conformation evolution. [124] The physical factors that cause proteins to change from a sequence to three-dimensional folding are summarized as follows: [125]

(1) Hydrogen bond. *a*-helixs and β-sheets of proteins are maintained by hydrogen bonds. [126]

(2) van der Waals force. Short distance dispersion forces between residues.

(3) Stereochemistry. The dominant stereoscopic configuration of adjacent residues on the backbone, such as the amino group of proline on the ring, glycine has no side chain, Cα is a chiral carbon atom. These factors limit the conformational selection of residues.

(4) Electrostatic interaction. Coulombic force attraction or repulsion of amino acids due to charge. Polar groups are also mutually exclusive or attractive because of the opposite static electricity.

(5) Hydrophobic interaction. Due to the action of surrounding water molecules, hydrophobic amino acids tend to lie in the core of proteins, and polar amino acids tend to be exposed in

aqueous solvents. [127,128]

(6) Side chain entropy. Protein folding is a process of side chain entropy reduction, which requires energy to help a protein change from a high entropy state (random free conformation) to a compact and orderly natural state. [127]

Because a protein can have a great number of atoms, the computational complexity is high. There are too many factors affecting protein stability, and the prediction of protein three-dimensional structure based on physical laws is greatly challenged by computational complexity.

3.5.5.3 Protein Homology and Empirical Protein Structure Prediction

Because physical law based *ab initio* protein structure prediction methods were greatly challenged by computational complexity, the structure prediction methods based on experimental data have been extensively explored.

Based on the homology analyses of proteins, it is assumed that similar sequences have similar 3D structures. Homologous proteins from different species need to maintain their 3D structures in order to conserve their biological functions. Although the 1D sequence continues to mutate and evolve, the 3D structure remains relatively conserved. [129] Therefore, the 3D structure of proteins is more conservative than their sequence structure. [130] Experience has shown that proteins with homologous similarity higher than 20% can have very similar three-dimensional structures. [131]

The relationship between protein sequence homology and homology modeling prediction accuracy (RMSD) is depicted in Figure 23.

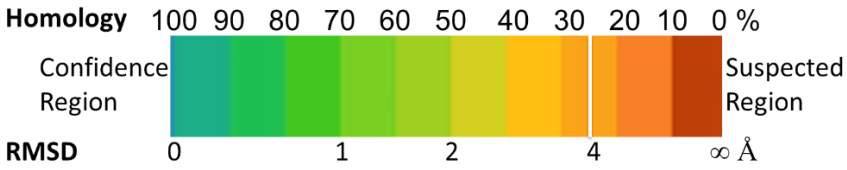


Figure 23. Relationship between protein homology and 3D structure prediction reliability

Although, proteins with a long evolutionary relationship may still have highly similar 3D structures as shown in Figure 24 (Carlos Outeiral Rubiera's blog: www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/ accessed on June 20, 2022).

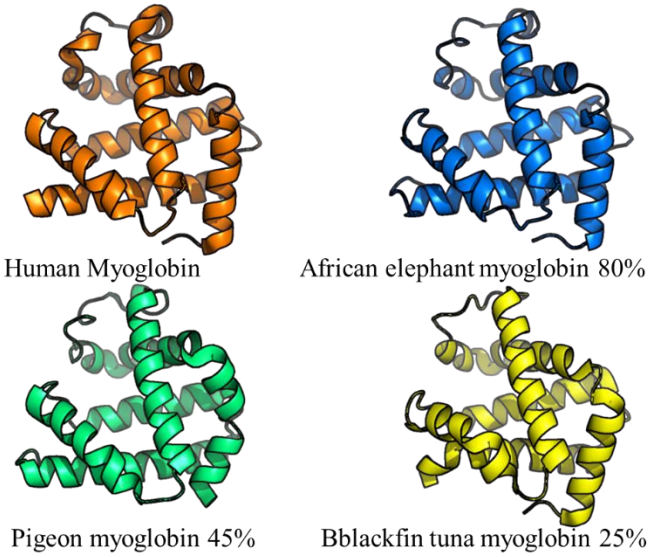


Figure 24. Proteins with lower homologous similarity may have similar 3D structures.

Homology modeling depends on the quantity and quality of protein sequence alignments and experimental template data. Traditional sequence alignment methods inevitably have alignment gaps (or indels), that is, some parts of the query sequence cannot find the corresponding templates. These gaps are often the main sources of errors.

Generally, when a sequence similarity is greater than 70%, the root mean square deviation (RMSD) of Cα atom between the predicted structure and the experimental measured structure is about 1Å. If the similarity is greater than 50%, the predicted protein structure will be more reliable, and the error of side chain filling and rotation state can be reduced (the RMSD can be in the range of 1~2 Å). The typical resolution of the structure solved by NMR is 1~2 Å. If the similarity is in the range of 30-50%, the RMSD can increase, the major errors would be from the loop regions. If the similarity is less than 30%, the RMSD could be too large to accept due to the misjudgment of basic foldings. If the similarity were less than 25% (usually referred to as "twilight zone"), protein structure prediction would be difficult, and the RMSD could increase to 2~4 Å or even greater (Figure 23). [132,133]

Loop regions are mainly flexible fragments connecting folding domains, and its secondary structure random coils are the most difficult parts to predict in homologous modeling. In addition, the conformation of the side chains are also difficult to predict, and have higher degrees of flexibility affected by the co-evolution of remote residues. This will increase the RMSD. [134]

The 3D folding of a protein is driven by two factors: the minimization of free energy and the co-evolution of protein residues. With the steady growth of experimental data in the worldwide protein database (wwpdb, website: wwpdb.org) [135], the surge of protein sequence data With the rapid development of deep learning technology, many protein structure prediction methods improve the prediction accuracy by combining the data of protein residue co-evolution. [136-139]

Some multinational pharmaceutical companies and the Wellcome Trust jointly funded the establishment of a non-profit Structural Genomics Consortium (SGC) to establish a high-quality public database of protein structure and accumulate experimental data of various representative folding templates of proteins. [140]

In order to promote the development of protein structure prediction technology, a critical assessment of protein structure prediction (CASP) competition [141] was established. Since 1994,

the CASP protein prediction competition has been held every two years, with the participation of more than 100 teams from all over the world. The competition is conducted in a double-blind way: neither contestants, organizers nor evaluators knew the 3D structures of target proteins in advance. The contest molecules were proteins that the structures have recently or would be determined experimentally (provided by the designated structural genomics team). The evaluating method is to superimpose the predicted results with the Cα coordinates of the measured structure, and calculate the RMSD of the two structures. The organizers also proposed a GDT-TS (global distance test total score) method to score each team. GDT-TS represents the percentage of the correctly predicted residues in a model in the total protein residues. [142] As of 2020, CASP has been held for 14 sessions, and the evaluating criteria were changed from time to time.

3.5.5.4 Protein Homology Modelling

The homology modeling process consists of following steps: [143]
(1) Select templates,
(2) Multiple sequence alignments,
(3) Assemb structure fragments (domains)
(4) Model loops,
(5) Optimize side chain conformations,
(6) Optimize global structures,
(7) Evaluate structure quality.

Template selection depends on the results of multiple sequence alignments (MSA). [144] Gene sequences encounter multiple gene insertions and deletions during evolution. This can result in a target protein failed in hitting a template from protein sequence databases by MSA. In this case, it is necessary to assemble multiple templates, or use *ab initio* calculation method to model the gaps. [125]

Predicting side chain conformations is another difficulty in protein structure prediction. Many side chains in crystal structures are not in their "lowest energy" states due to the energy factor in the accumulation of hydrophobic nuclei and single molecules in protein crystals. Generally, the side chain conformations with the lowest energies are predicted by searching the rotameric library.

There are three domain-assembly methods:

(1) Fragment assembly: first, construct a conservative core fragment, and then replace the variable region of the protein with the fragment that has been analyzed. The variable regions are usually from a fragment library. [145,146] The fragment assembly method varies due to different ways of dealing with areas that are not conservative or lack templates.

(2) Fragment matching: a protein sequence can be divided into several fragments, and a template matching each fragment is retrieved from a sequence database. In this way, sequence alignment is conducted on fragments rather than on the whole sequence. Each fragment template is selected based on sequence similarity Cα-coordinates comparisons, and spatial conflicts cases within van der Waals radius. [147]

(3) Satisfaction of spatial constraints: align templates to build a set of geometric criteria; then transform the geometric criteria to probability density functions as constraints. Apply the functions to internal the protein coordinates (protein backbone distances and dihedral angles) as the bases for the global optimization of protein structure, and the position of non-hydrogen atoms

in proteins is optimized by minimizing conjugate gradient energy. [148] This method is often used in loop area modeling. [149]

Poor template selection can introduce significant errors, [150,151] multiple template recognition (MTR) and MSA are used to reduce such errors. Of course, the experimental data of the template structure itself may also be wrong. Therefore, the construction of the PDB-REPORT database has reported millions of errors in experimental template structures from PDB database (most of the errors are very minor). The last step of this protein structure prediction method is to further optimize the structure with molecular dynamics, but the force field parameters may also introduce errors. [152]

In the absence of experimental data comparison, Ramachandran plot [153] (also known as Rama map. Ramachandran graph or [ϕ, ψ] map) can be used to evaluate the rationality of the predicted structure (Figure 25).

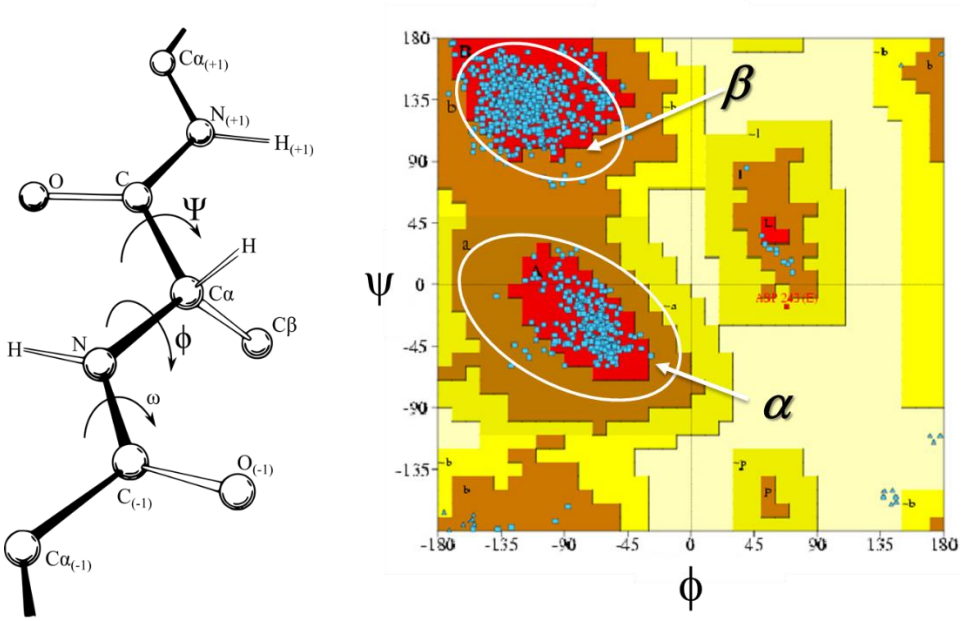


Figure 25. Rama map for evaluating protein structure quality. Left: Definition of two dihedral angles of protein backbone ϕ and ψ ; Right: Map of ϕ to ψ , red, brown, and dark yellow areas represent the energy reasonable, allowable, and alert areas, respectively. Light yellow areas are unacceptable areas. α and β areas represent the regions where α and β folds.

Allowable dihedral angle values of a protein backbone ψ And φ are restricted. The definition of dihedral angle of φ and ψ is depicted in Figure 25 in the left. At peptide bond angle ω is usually 180° for the partial double bond keeps peptide bonds flat.

In Figure 25 (right graph), blue dots represent the (ψ , Φ) coordinates for the residues of a protein (PDB: 1AXC). The relationship of dihedral angles ψ and Φ should remain inside red areas for an ideal protein structure, inside brown or dark yellow areas for an in an reasonable protein structure, and outside these areas are very rare indicating significant errors, which should be checked and corrected. Usually, the RMSD of a predicted protein structure is close to 3Å.

A predicted protein structure can also be evaluated by semi-quantitative scoring functions:

(1) Statistical potentials: Based on residue-residue contact frequency in PDB database, each residue-residue pair is assigned a probability or energy score for a predicted protein structure,

and the final score is calculated from the score of every residue-residue pair. Some scoring methods can find out residue-residue pairs lower scores, although the overall score for the protein is good. These methods may be good for globular proteins, which often have hydrophobic cores and solvent exposed polar amino acids. [154]

(2) Physics based energy calculations: The methods are based on the energy landscape hypothesis of protein folding, that is, a natural protein state is also in the lowest potential energy state. Implicit solvation is usually used to provide a continuous approximate solvent bath for a single protein molecule without explicit representation of a single solvent molecule. The interaction between atoms related to the stability of proteins in solution (mainly van der Waals and electrostatic interaction) are calculated. The protein molecules are too large, and it is not feasible to use semi-empirical quantum mechanics to compute the score. Generally, CHARMM effective force field (EFF) is used. [155]

These scoring methods are often inconsistent with experimental data. This may be related to the lack of representation of proteins in the PDB database. In 2006, a SVM based comprehensive scoring function was reported. The scoring function linearly combined six scoring functions (DOPE total weight atomic statistical electrostatic potential, MODPIPE surface contact and combined statistical potential, and two PSIPRED/DSSP secondary structure consistency scores). It is said that the result is improved. [156]

The proteins for CASP contest do have experimental data, hence, there is no need to use these scoring functions. RMSD based evaluations are good enough. However, the overall RMSD, including the loop domain, tends to underestimate the quality of contest models, because correct modeling of non-loop structures is important. CASP's GDT-TS scoring method minimizes the impact of the side chain errors. [157]

3.5.5.5 AlphaFold 2: Success of DL

In the previous CASP 13 contests, protein structure prediction algorithms keep improving without disruptive progress until the 14th CASP contest in 2020.

Google DeepMind team's AlphaFold 2 of won the first place of the contest (Figure 26), surpassing the second place (Rosetta@home team led by David Baker, University of Washington, Seattle, USA) scores more than twice.

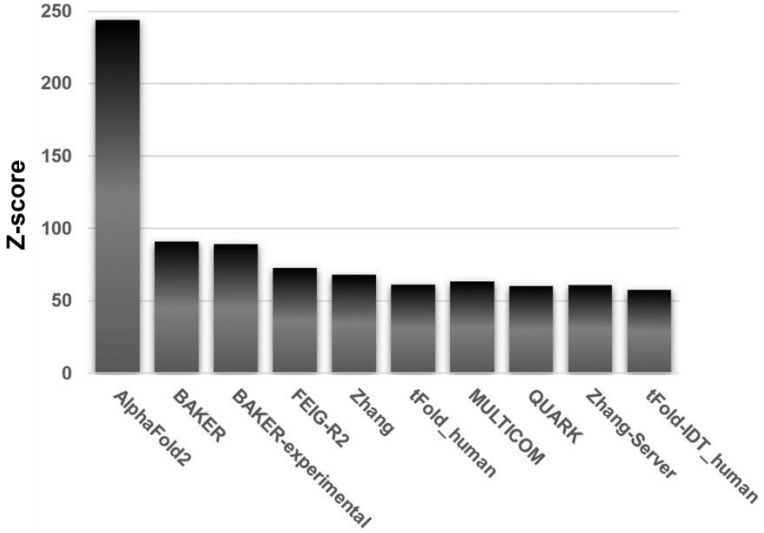


Figure 26. The top ten scores in CASP14 (Total 146 teams).

AlphaFold 2 team's success highlights the great power of DL algorithms. About two-thirds of the proteins have a score of more than 90 (the full score is 100). [158]

On July 15, 2021, AlphaFold 2 team published a paper entitled "highly accurate protein structure prediction with AlphaFold" in *Nature*, and the software is also openly accessible to the world. [159]

AlphaFold 2 is a sophisticated software project. Readers interested in its implementation details can read the original paper [159] and its supplementary materials. Here, we highlight some significantly bright points:

(1) New architecture. Alphafold1 (the 1st version of AlphaFold) predicted protein structures by modifying an ANN architecture that was for image process. Realizing the architecture was not proper for protein structure prediction, the team designed a *de novo* architecture, which changed the independent modules in the previous architecture into mutually coupled neural network systems to form a single differentiable end-to-end model. The model was trained as an integrated architecture. The protein sequence was input into the model, and then the model directly outputted 3D structure of the protein (end-to-end system). After the prediction result of neural network converged, the predicted protein structure was optimized based on energy by Amber program, and the error of chirality was corrected.

(2) New Database. A high-quality protein sequence database (BFD, big fantasy database) was created specifically AlphaFold 2. BFD was composed of more than 65.98 million protein families represented by MSA and hidden Markov models (HMMs), with a total of more than 2.2 billion protein sequences. These data are constantly optimized, and the amount of data continues to increase.

(3) Integrating advanced technologies. The MSA search engine of AlphaFold 2 combined HHBlits, HHSearch, and HMMER3; its 3D structure optimization technology adoped OpenMM v.7.3.1 and Amber99sb force field, and the ANN was constructed using TensorFlow, Sonnet, NumPy, Python, and Colab.

(4) New hardware. To resolve the computing complexity problem, Google team developed new hardware including a dedicated tensor processing unit (TPU), about 128 TPUv3 cores (equivalent to 100-200 GPU), the computing power exceeds the sum of the computing resources available to other CASP teams.

(5) New algorithm. In order to reduce the computational complexity, affine transformation was used to realize 3D structure assembly. The residues were placed at the coordinate origin, and the affine matrix was used to replace and rotate the residues in the space in each iteration, so that the structure complied with physical and geometric constraints, which greatly reduced the computational complexity.

(6) New transformer. A machine learning architecture, Evoformer, was constructed to reflect the evolution of protein folding. This architecture consisted of two transformers, between which there was a direct communication channel, information feedback and iterative optimization.

(7) Strengthened teamwork. Scientists and engineers work closely together, and scientific problems were effectively transformed into engineering problems. Software engineers repeatedly invented, debugged, and screened various network architectures, until the optimized program system was implemented.

AlphaFold 2 is mainly used to predict the molecular structure of a single protein, and the team has also made significant progress in predicting the structure of protein complexes. [160] It is believed that AlphaFold 2 is probably one of the most important scientific achievements of this

century (https://www.blopig.com/blog/)

3.5.5.6 Unresolved Protein Structure Prediction Problems

Remarkable progress has been achieved in the field of AI-assisted protein structure prediction, and AlphaFold 2 has proved that the protein 3D structure can be predicted quite accurately from an amino acid sequence. However, AlphaFold 2 can only predict the structures that are most likely to be found in PDB database, and many problems remain unsolved:

(1) Multiple privileged conformations. Under a given condition, a protein can have multiple privileged conformations, [161] and each conformer can have a specific function. For example, Abl tyrosine kinase has inactive and active conformers, which are switched by a loop domain as shown in Figure 27. Human potassium channel protein (hERG) has conformers corresponding to the states of open, close, and silent. If one of the conformer was inhibited due to mutation or by a small molecular binding, it would lead to arrhythmia. [162] However, neither cryo-electron microscopy nor AlphaFold 2 can determine or predict these conformers [163,164]. Technologies are been developing to deal with this challenge. [165]

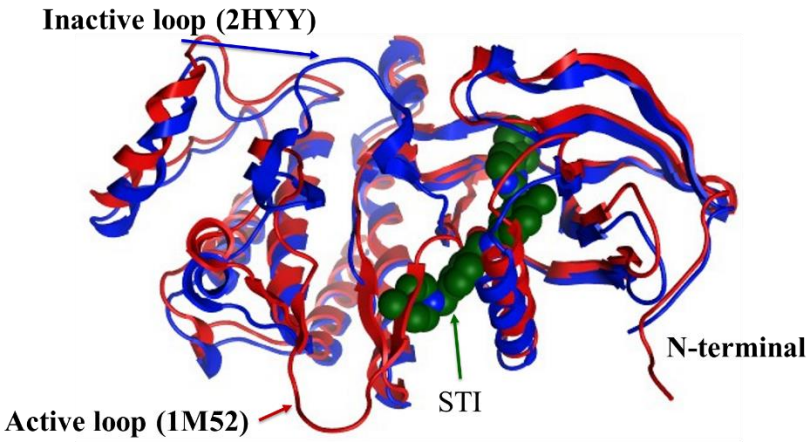


Figure 27. Abl tyrosine kinase active and inactive conformers (STI is an inhibitor)

(2) Conformations due to protein modifications. Protein properties regulated by both conformational turnover, and post-translational modifications (PTM), [166] such as ubiquitination, phosphorylation, sulfation, acetylation, and methylation. Phosphorylation and dephosphorylation of the key residues of kinases and phosphatases are the major mechanisms of cell regulations. PTMs move protein domains, flip loops and, change protein polymerizations. These structural changes cannot only be encoded in protein sequences. For example, human mitogen activated protein kinase 1 (MAPK1) has PTM induced active and inactive conformations, which are reflected in 113 structures recorded in the PDB database. After Thr-185 and Tyr-187 of MAPK1 are phosphorylated, the loop region is activated and the conformation changes. [167] However, AlphaFold 2 cannot distinguish the active and inactive conformers of MAPK1. [168] Although these structural changes caused by PTMs are very important for drug design.

(3) Protein complex prediction. The complex structures from with other proteins, cofactors, small molecules, DNA or RNA are the main concerns for drug design. The binding of endogenous or exogenous small molecules at orthosteric or allosteric sites will regulate protein conformation and change protein properties. These informers cannot only be encoded in a primary sequence of protein; [169] The structural features of proteins in PDB database cannot fully

reflect the structural features of human proteome. More than 40% of UniProt protein families do not have crystal structure data. Some protein families (such as kinases) have a lot of data, while many protein families have no data, which will inevitably lead to unbalanced predictions.

(4) Binding mode prediction. Currently predicted protein structures are at apo states (no ligand binding). A protein structure has many pockets, identifying a druggable binding pocket for a ligand is crucial to drug design. It is important for a protein to encode these binding modes into a primary sequence. In a simplest case, protein-ligand binding is based on shape (steric) and electrostatic (static) complementarity. However, because proteins can have multiple conformers, and a ligand can also have many conformations, the binding of a protein and a ligand is a dynamic and mutual molecular recognition. It can bake things more complicated that a ligand can also induce the conformational changes of a protein (induced fit) (Fig. 28). [170]

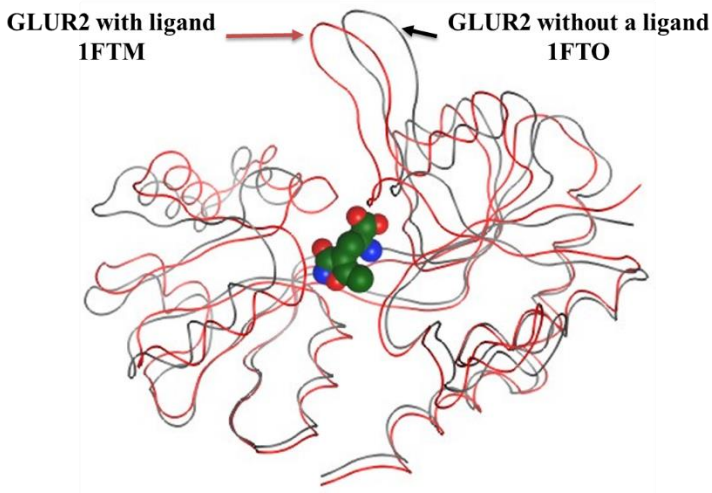


Figure 28. GLUR2 apo structure (1FTM) and GLUR2-ligand co-structure (ligand induced fit)[168]

(5) Protein folding dynamics. Protein structural instability caused by single or multiple point mutations can lead to genetic diseases, such as cystic fibrosis, Tay Sachs, Creutzfeldt Jakob disease, sickle cell anemia, or cancer. [171] Another example is that p53 mutations reduce thermodynamic stability, interfere with p53 DNA interaction or cause misfolding. [172] Antibodies usually need long loops and cavities, which also makes proteins unstable. Due to the lack of non-protein molecule conformation data in PDB database, it is very difficult to predict the structure of unstable proteins or misfolded proteins.

3.5.6 Medicinal Chemistry Diversity Space Explorations

Since the inventions of high-throughput synthesis and screening technology (HTS), ten million of molecules can be synthesized and screened. [173] The number of drug-like small molecules [174,175] is estimated to be from 10^{18} to 10^{200}. The ability of high-performance syntheses is limited by the limited knowledge of synthetic reactions, but by the available reagents (*aka* molecular fragments) of synthetic reactions. In order to estimate the organic chemical space accessible to humans, Ertl studied more than 3 million known molecules and derived 3.1 million chemical substituents. Using known synthetic methods, it is estimated that the number of accessible organic chemicals ranges from 10^{20} to 10^{24}. [176] Most of these molecules have nothing

to do with pharmaceutical chemistry. In order to explore the chemical diversity space that pharmaceutical chemists are concerned about, Reymond and co-workers enumerated a pure virtual small molecule database called GDB-17 based on chemical rules, 17 carbon, nitrogen, oxygen, sulfur and halogen atoms, with the constraints of valence bond rules, functional group instability and synthesis feasibility. GDB-17 hosted 166.4 billion ($1664x10^{11}$) molecules in SMILES format. [18] The authors claimed that 99.9% of the molecules are molecules that have never been found by human beings, and these molecules meet the criteria for drug-likeness [177], lead-likeness, and fragment-likeness. [179,180]

Even if this chemical space were compressed to hundreds of billions, it would still far beyond human synthetic ability. Up to now, about 125 million small molecules can be purchased from markets. [181]. After the emergence of DNA encoded libraries (DELs) synthesis technology, the era of synthesizing and screening 10 billion small molecules has come. [182]

However, the overlapped portion of "human accessible organic chemistry space" and "the organic chemistry space that can bind to drug targets" is the medicinal chemistry space.

In order to explore the medicinal chemistry space, many biological experiments (such as protein specific antibodies, engineered recombinant proteins, gene knockouts, gene knockins, RNA interference, intrabodies, and proteomics) have been developed to study the relationship between proteins and diseases. [183] To identify and validate drug targets from these proteins, many structural biological techniques (such as X-ray crystallography, multi-dimensional NMR, freeze electron microscopy) have been developed to determine the 3D structures and their active sites of proteins. Biophysical experimental techniques (such as surface plasma resonance (SPR), isothermal titration calorimetry (ITC), saturation transfer difference NMR (STD-NMR)) and computational techniques (such as molecular docking technologies and molecular dynamics simulations (MD)) have been developed to verify the binding affinities and binding modes of small molecules and proteins.

Experimental methods are expensive and limited by experimental conditions. Therefore, *in silico* approaches are highly expected. The success of AlphaFold 2 has set off an upsurge of using AI techniques to explore the medicinal chemistry space. One attractive idea is to generate virtual drug-like molecules based on the known of the interaction between drug targets and small molecules with physical, chemical and biological activity data of small molecules from scratch.

Designing new drug like molecules based on known drug-like molecules can be regarded as a natural language process (Figure 27).

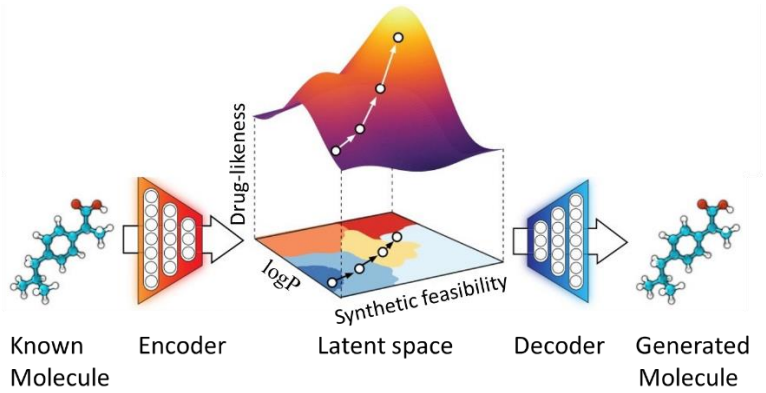


Figure 27. ANN for generating drug-like molecules (Modified from reference[186])

Known molecules are regarded as chemical sentence written in canonicalized SMILES (chemical natural language) strings, which are translated into new chemical sentences with "the same semantics" by a translator composed of ANN. In this way, discrete pharmaceutical chemistry space (such as medicinal chemistry space composed of small molecules in ZINC database) can be mapped by "translation machine" to three-dimensional (logP, drug-likeness [184], and synthesis feasibility [185]) or higher dimensional continuous space.

The translator is composed of encoder and decoder. The encoder maps the known small molecule SMILES strings to a latent space, and the decoder reads the latent space information and "translates" the new used SMILES strings (predicted new molecules). Encoders and decoders can be constructed using recurrent neural networks (RNNs) to perform sequence-to-sequence learning. The encoder uses three one-dimensional convolution layers to transmit the data to a full connection layer and then outputs it. The decoder is a three-layer gated recurrent unit (GRU) network. The last layer of the decoder defines the probability distribution of all possible characters at each position in the SMILES strings. This means that the "translation" operation is random. According to the random seed used to sample characters, the same point in the latent space may be decoded into different SMILES strings (molecules). [186]

The above-mentioned "translator" can be termed as a molecular generative model based on chemical language or syntax. [187] This model can also be used to search the seed space of medicinal molecules, such as "natural-product-like" molecules [188], chemical stability molecules [189], target-focused molecules [187,190], specific protein-bound molecules [191,192], reaction-produced molecules [193], high-yielding combinatorial chemistry molecules [196], Fragment-assembled molecules [195], and phenotypic active molecules [196].

In a word, as long as this model is trained with relevant molecular subsets, new molecules similar to the training sets can be generated. What the model learns is the range distribution of various molecular properties and the distribution of molecular structure features with certain properties. The SMILES-based molecular generators were evaluated with a conclusion that this model is powerful and should be widely applied. The disadvantage of the model is that it cannot capture molecular steric features, which are very important to medicinal chemists. [197]

In 2021, a team from New York University proposed a masked graph modeling for molecular generation, which learned the structural features of molecules by capturing the conditional probability of a given atom adjacent to other atoms and bonds, trained the model [198] by iteratively masking and replacing different parts of the initial graph, and evaluated the model with the CHMBL [199] and QM9 [200] data sets. According to the authors' report, this model seems to be superior to the graph based and SMILES based methods.

To synthesize computer designed virtual molecules, synthetic plans have to be drawn either by men or by a program. The early theoretical work of AI assisted synthetic planning program can be traced back to E J. Corey's retro-synthon model. [201] Computer-aided synthetic planning was viewed as a landmark achievement of AI application. The 1970s and 1980s were the first heyday. Since the resurgence of a new wave of AI represented by DL at the beginning of this century, AI assisted synthetic planning has once again attracted people's interest. [202]

There are two layers of evaluating the feasibility for a designed molecule:

(I) Synthetic plan layer: identifying chemical reactions and reagents to assemble the molecule,

(II) Yielding layer: identifying proper reaction conditions (such as temperature, pressure, catalyst, pH, etc.) for producing the molecule.

The solution of problem (I) does not automatically solve problem (II), and both problems

must be solved. Currently, most of efforts are on solving problem (I), because there is a large amount of data available for (I). It is more challenging to solve problem (II) because there are no much reliable data (including reaction condition data and quantitative data of yieldings). In order to solve problem (II), it is necessary to screen the reaction conditions and sample a large number of reaction yields based on the chemical diversity of substituents. [194]

Synthetic planning requires the establishment of atom and chemical bond mapping from reactant molecules to product molecules. Nugmanov team recently reported GraphormerMapper, which directly processes molecular graphs into atoms and bond sets based on transformer neural network, and derive molecular features related to atoms and bonds in combination with the bidirectional encoder of transformers (BERT) network, which solves problem (I). [203]

3.5.7 Target Identification and Validation

The modern drug discovery process still follows a protocol of "one disease–one target–one drug", which requires figuring out a drug target of each drug or active compound. Although the regulatory authorities do not force drug applications to be accompanied by target data, they only require drugs to be safe and effective. If reliable target data is attached, it is obviously more convincing, and the design and optimization of lead compounds also have a theoretical basis. Therefore, target-based drug discovery (*aka* reverse pharmacology) has become the mainstream paradigm of drug discovery, and it is successful indeed. From 1999 to 2018, the US FDA approved 326 drugs, and more than 80% of the drugs were based on this paradigm.

However, in recent years, people gradually realize that many drugs (such as small molecule kinase inhibitors) are multi-targeted. [204] Therefore, this paradigm is being challenged and to be improved.

Incorrect target identifications and validations are the major causes of clinical trial failures. However, experimental studies on drug mechanisms of actions are time-consuming and very expensive. Therefore, *in silico* studies on target identifications and validations are demanding. [205]

Drug targets interact with drugs (small molecules or biological macromolecules), which regulate physiological processes or pathological states. The main drug types are G protein coupled receptor (GPCR), various enzymes (such as proteases, kinases, esterases, HDACs, etc.), ion channels, and nuclear receptors. Among the small molecule drugs on the market, the main drug target types are GPCR 33%, ion channels 18%, nuclear receptors 16%, kinases 3-6%. Enzymes are the second largest proteome in the human genome and the second largest group of drug targets on the market. However, the drugs that targeting enzymes are not the mainstream.

Drugs are divided into three categories: small molecule drugs, biologics, and nucleic acid drugs. The ways of drugs regulate targets are as follows:

Small molecule drugs

- receptors: activation / antagonism / reverse activation / regulation / allosteric regulation / sensitization;
- enzymes: inhibition / excitation;
- transcription factors: inhibition / excitation;
- ion channels: blocking / opening;
- transporters: inhibition;
- protein protein complex: inhibition;
- nucleic acid (DNA): binding / alkylation / complexation / intercalation

Biologics

- cellular proteins: antibodies;
- transmembrane receptors: recombinant proteins;
- cell surface receptors: antibody drug conjugates (ADC);
- substrates and metabolites: enzymolyses;
- other proteins: inhibition / activation.

Nucleic acid drugs

- RNA: interferes (RNAi/siRNA/miRNA/shRNA).

Good drug targets have the following (ideal) features:

(1) The targets have definite experimental proofs (such as knockin / knockout phenotypic data) that been regulated for being beneficial to cure;

(2) Regulating targets will not bring safety problems with safe profiling and without adverse reaction;

(3) Targets are expressed in specific cells / tissues / organs or pathogens;

(4) Targets have 3D structures measured in the experiments, and the structural changes have definite relationships with the cures;

(5) Targets are testable, have HTS protocols detecting the activities and specific biomarkers for precision medicine;

(6) Targets have animal models or predictable phenotypic data (such as gene mutation data);

(7) There are no intellectual property barriers.

Target identification relies on the understanding of pathology, molecular mechanisms (studied through molecular biology and chemical biology), and *in silico* models. Chemical proteomics (such as affinity chromatography, expression cloning, protein microarray, reverse transfection cell microarray, and biochemical inhibition), functional genomics (such as gene knockout, random chemical induction of gene mutation, DNA + zinc finger protein binding, antisense RNA binding to mRNA, gene silencing with siRNA, shRNA, RNAi, miRNA, CRISPR-Cas9 endonuclease gene editing, cell overexpression or up-regulation by cDNA or plasmid or viral vector), and *in silico* methods are used for target identification and validation.

Conventional *in silico* target discovery approaches are summarized as follows:

(1) Database search

To find targets for oligo compounds (OC, the compounds with unknown targets), ligand substructure[14] or similarity[6] based searches can be applied against medicinal chemistry databases with target annotations (such as, ChEMBL [199], PubChem [206], BindingDB [207], cBinderDB [208]). SEA [209], SwissTarget (www.swisstargetprediction.ch) [210], and SPiDER (modlabcadd.ethz.ch/software/spider) databases are searchable with fuzzy pharmacophores[211], SuperPred is searchable with ECFP fingerprint-based similarity [212]、PPB (gdbtools.unibe.ch:8080/PPB) uses multiple descriptors based similarity search to find targets [213].Random forest method was also used for target fishing from a database [214]

The databases commonly used for fishing targets are listed in Table 1.

Table 1. Commonly used medicinal chemistry databases and URLs [215]

Database	URL
BindingDB	http://www.bindingdb.org/bind/
ChEMBL	http://www.ebi.ac.uk/chembl
DrugBank	http://www.drugbank.ca/
DrugPort	http://www.ebi.ac.uk/thornton-srv/databases/drugport/
HumanCyc	http://humancyc.org/
Human Metabolome Database	http://www.hmdb.ca/
KEGG	http://www.genome.jp/kegg/
MDL Drug Data Report	http://accelrys.com/products/databases/bioactivity/mddr.html
PubChem	http://pubchem.ncbi.nlm.nih.gov/
SuperTarget	http://bioinf-apache.charite.de/supertarget_v2/
WOMBAT	http://www.sunsetmolecular.com/
ZINC	https://zinc.docking.org

(2) Bioactivity spectrum

A bioactivity spectrum is graphed with bioactivities and compounds, where the bioactivities with biological targets. [216] Search the bioactivity spectrum with OC as the query structure, and results hits being similar to OC. The targets associated with the hits are proposed targets for an OC. [217]

(3) Association based target identification

It is better to identify drug targets based on multiple sources of data including omics data, experimental results of animal models, data of gene expression changes, text data from scientific and technological literature, the data of relationship between genes and diseases [218] or protein-protein interaction network. [219] The target identification based on a single-sourced data is more susceptible to systematic errors,

(4) Reverse pharmacophore search and reverse molecular docking

Docking a ligand structure (a 3D structure or pharmacophore) to the active site of a receptor is a common practice for drug target identification. INVDock [220], IdTarget [221], DRAR-CPI [222] , and TarFisDock [223] are examples. They simulate the interactions of ligands and protein targets, and require that the candidate proteins' 3D structure is known. Thus, these methods are limited by the 3D structure availability.

The premise of fishing for protein targets with the 3D structures of small molecules is that the structures of small molecules should be close to its lowest energy conformations. In turn, crystal conformations are postulated to be close to the lowest energy conformations. When a small molecule binds to the active site of a specific protein target, its active conformation should change from its lowest energy conformation, and the energy difference (activation energy) from the lowest energy conformation to the active conformation should be in the range of translational energy fluctuation. Otherwise, even if the receptor activity pocket has proper steric cavity, it is still difficult to bind. It can be seen that the method of reverse molecular docking fishing target requires an algorithm of *ab initio* generated 3D conformation of small molecules based on experimental crystal structure data.

(5) Receptor binding site similarity

A common postulate is that similar ligands should have similar targets. [224] The following packages, such as CavBase(relibase.ccdc.cam.ac.uk) [225], SuMo(sumo-pbil.ibcp.fr) [226], PocketMatch (proline.physics.iisc.ernet.in/pocketmatch) [227], PARIS (cbio.ensmp.fr/paris) [228], and

IsoMIF (bcb.med.usherbrooke.ca/imf) [229] are based on this postulate, which involves collecting binding sites, simplifying binding-pocket representation, and scoring functions. A complete review regarding these methods has published. [230]

(6) Machine learning (DL)

The relationship between small molecules and drug targets may be latent in a data set represented in descriptors. Elucidating the relationship can be beyond human reasoning capacity with conventional approaches, such as QSAR.

To simplify the problem, the first step is to reduce the data dimension so the data patterns can be inspected in 2D or 3D space. DL (Yann Lecun, Yoshua Bengio and, Geoffrey Hinton have published a review in 2015 [231]) can do better work than principal component analysis (PCA) in dimensionality reduction. [232]

In order to improve the accuracy of prediction, BANDIT method integrates data from six different sources (drug efficacy, post-treatment transcription reaction, drug molecular structure, adverse reactions, bioassay results, known targets), and uses Bayesian machine learning technology to predict drug targets. [233] This method proved that integrating data from multiple sources could significantly improve the performance of target prediction.

Mamoshina used the Keras Python library of Tensorflow to integrate five ML technologies (ElasticNet, SVM, *k*-NN, RF and, depth feature selection feedforward neural network) to establish a supervised learning model to predict the biomarkers and tissue-specific targets of aging on human skeletal muscle from human muscle transcriptome data. [234]

deepDTnet is a target prediction program based on DL. It predicts the new uses of drug targets or small molecule drugs based on heterologous drug gene disease network (embedded with 15 types of chemical, genomic, phenotypic and cellular characteristic data). The training data are 732 small molecule drugs approved by FDA. Experiments proved that topotecan's (approved topoisomerase inhibitor) new predicted target is the human retinoic acid receptor related orphan receptor γt(ROR-γt) (inhibitor, IC_{50}=0.43 μM) for the treatment of multiple sclerosis. [235]

Lu's team used the restricted Boltzmann machine (RBM) to construct a deep belief network (DBN) DL architecture to predict drug targets. [236] The training data comes from the DrugBank database, covering 1412 drugs, 1520 targets and 146240 drug target pairs.

Target prediction algorithms based on the data of biological network can be divided into two categories: network-based and DL based. The network-based algorithms use a variety of alternative network methods to identify targets, study network data from different angles, and explain the mechanism of drug action. [237] ML based algorithms integrate heterogeneous data (such as experimental data from biology, chemistry, physics, and theoretical calculation data), and mine the relationship between latent features and various biological entities (such as proteins, nucleic acids, small molecules, metabolites) [238] to identify targets. [239]

4. Summary and Prospect: Drug Design Methodology with AIDD

4.1 Paradigm of QSAR Studies

QSAR originated from chemists' studies on molecular structure and activity relationship (SAR). Along with the progress of computing technology, SAR has evolved into quantitative SAR research, namely QSAR. Its original intention is to mine the mathematical relationship between molecular structure data and properties or activities.

The paradigm of QSAR studies is consistent with traditional scientific paradigm and outlook: to predict a molecular property *y*, we must first find the independent variable x which has a functional relationship with *y*, and then determine the mathematical formula of $y=f(x)$ that can explain the experimental data of *Y* and *X*, thus the molecular property can be predicted.

A molecular structure is a graph composed with atoms (nodes) and chemical bonds (edges), which is described in discrete mathematics and can be represented in a chemical formula or linear notation (1D array), topological diagram (2D diagram), stereochemical diagram (2.5D diagram), or conformational graph (3D structure).

These discrete data can be transformed to continuous data with fitting methods or word-embedding methods (originated from NLP) as shown in Table 2.

Table 2. Data transformations in QSAR

Original data type	**Data type after transformation**
Chemical formula or Linear notation	One-hot encoding Bit-map, fingerprint
Topology	2D graph Bitmap fingerprint Attribute graph Descriptors Linear notation, such as SMILES
Stereochemical graph	2D graph (wedge bonds, stereo tags) Bitmap fingerprint Attribute graph Descriptors Linear notation, such as SMILES
Conformational graph	Voxel Spatial coordinates 3D descriptors Coulomb matrix

The transformed molecular structure data can be regressed with linear regressions, kernel regressions, random forest, SVM, ANN, message passing neural networks (MPNN), sequence modeling, and CNN.

QSAR did explore many AI algorithms, and develop many molecular descriptors to solve more complicated problems. Still, conventional QSAR faces great challenges that call for the new era of drug design methodology and new paradigm to come.

4.2 Challenges and Paradox to QSAR

Two basic postulates of QSAR are 1) similar structures have similar properties/activities, and 2) the contributions to properties/activities of different functional groups in the same molecule are additive.

With the advent of the era of big data and high-performance computing, these postulates are facing more and more challenges, which also bring puzzling problems and paradoxes to modern QSAR studies.

4.2.1 Substructure Partitioning Issue

With chemists' intuition, a molecular structure consists of a scaffold and a set of substituents (alternative substructures for designated sites in the molecule). This substructure partition is done before further QSAR studies. However, neither there is a unified standard to derive a scaffold from a chemical structure, nor to derive substituents or distinguish a substructure from a scaffold. People in different fields (biology, chemistry, or mathematics) never agree to a general standard to partition substructures from a molecular structure.

In order to objectively partition molecular substructures, many methods including chemist experience consensus method (such as MACCS search keys), graph theoretic rule based method (rules to systematically derive substructural fragments, such as daylight's fingerprint, ECFP fingerprint). However, every method has its own deficits. Empirical consensus methods kept chemical meanings, but were empirical biased; while mathematical rule based methods avoided empirical biases, but produced chemical meaningless fragments. It seems to have no solution for substructure partitioning. Although QSAR is the quantitative **sub**structure-activity relationship *per se*.

When there are multiple substituents on a scaffold (for example, two substituents R_1 and R_2), when constructing a linear regression function $\boldsymbol{y=f(R_1, R_2)}$, we should prove whether R_1 and R_2 are independent (*aka* R_1 and R_2 are of additivity).

A protocol for rigorously prove the additivity was convincedly proposed, [36] but it did not become a feasible practice due to ineluctable experimental error acumination and impractically experimental costs.

4.2.2 Activity Cliff Issue

"Activity cliff" problem (that is, molecules with high similarity have significant activity differences) is a persistent challenge to QSAR. [55] The main reason is that the activity of molecules is often not determined by the overall structure of molecules, but by the local substructure of molecules.

For example, in aromatic ring systems, small perturbations (such as the substitution of highly electronegative groups) will lead to significant changes in molecular surface electrostatic charges; A planar molecule will cause a huge reversal of the molecular conformation because individual atoms change from achiral to chiral. If QSAR cannot characterize the above phenomena, the problem of these activity cliffs will remain problems.

4.2.3 Imbalanced Training Data

The training data set of typical QSAR model is small (from dozens to hundreds), the distribution of active data and inactive data can be extremely imbalanced (the most of cases, inactive data are missing), and the value range of active data is often narrow and uneven (ideally, it should be close to Gaussian distribution).

In the practice of medicinal chemistry, when an active molecule is found, many homologues of it will be synthesized to explore the chemical diversity space of the positive compound. However, few people are willing to explore the chemical diversity space of inactive compounds.

The biased data lead to the ill-prediction in QSAR models. "Garbage in and garbage out" is always true.

4.2.4 Paradox of Prediction Accuracy and Generality

The paradox of the prediction accuracy and generality for a QSAR model states: if molecular structures in a training set are not diverse, the QSAR model will have more accurate predictions, and less generality; however, if molecular structures in a training set are diverse, the QSAR model will have less accurate predictions, and more generality. You can't have your cake and eat it.[16]

4.3 Moving toward AI

After more than 80 years' development, QSAR, as a core of traditional drug design methodology, has developed many theories and methods to solve problems for drug lead discovery and optimization, aggregated big data, and laid the foundation for a new era of drug design methodology.

However, QSAR was born in an era when human computing power was limited and determinism dominated. In the post information time characterized by high-throughput experiments, high-performance computing and big data, based on the legacy of conventional QSAR, drug design methodology will take deep learning technology as a new starting point and move towards the AI aided drug design (AIDD) era.

4.3.1 Disruptive Thinking

AI was born in the 1940s, [240]- the time when determinism dominated. People believed that the world should be driven by Nature laws (logically, premise-conclusion pairs), in which the premises are based on a axiomatic system. The system is functional, logical and recursively evolving along time. Therefore, AI focused on developing reasoning and deductive systems based on the axioms and knowledge rules to simulate human logical thinking and natural processes in that time. For this reason, AI specific programming languages LISP (LISt Processor) and Prolog (*programmation en logique*, French for programming in logic) were also invented. The former makes recursive process easy, and the latter specially deals with logical reasoning calculus based on the "premise-conclusion" rules.

Early AI milestone achievements included automatic theorem proving (Chinese mathematician Wu Wenjun was famous in for his work of automatic geometric theorem proving), Feigenbaum's mass spectra elucidation expert system, and E J Corey's organic synthesis design reasoning system (based on his retro-synthon method).

QSAR, which evolved at the same time as that time, was also deeply branded with the era. Determinism was the mainstream thinking in QSAR. It was believed that prediction had to be based on clearly understanding the rules behind the relationship of structure and activity. The mission of QSAR is to figure the rules (represented in analytical formula of functional) between the independent variables and functions hidden in the data by curve-fitting or regression. The degree of fitting determines prediction performance. Therefore, AI was regarded as tools for nonlinear data fitting in QSAR. Those unsuccessful AI projects in the past, on the one hand, are because the software and hardware technology were not ready, and the deeper reason was that disruptive thinking, the idea of determinism thinking could be an alternative epistemology, was still fiercely controversial and, not widely accepted.

After the 2010's, DL has led to a disruptive change in people's thinking, that is, prediction is not necessarily to be based on clearly understanding the rules behind the relationship of structure and activity prediction models, instead, can be data-driven or statistic process. The prediction does not need to have hypothetic continuous mathematical models in advance.

Based on this epistemology, DL initially achieved great success in the field of image processing and natural language processing, so that DL was soon applied in QSAR. With new paradigm, the mission of QSAR is to find the statistically dominated correlation between independent variables and functions, without reading whether the independent variables are linear independent or whether the independent variables and functions are linear or nonlinear correlation in advance.

In this way, the two postulates of QSAR are no longer needed. Therefore, the related problems and paradoxes (such as the problems of substructure partitioning and substituent additivity, the paradox of prediction accuracy and generality) can no longer be existing.

However, the activity cliff problem and imbalanced data sampling problem still exist. The former is due to the "black box" nature of AI algorithm, while the latter makes DL algorithm unable to give full play to its advantages.

Big-data-based DL becomes mainstream practice in modern QSAR. AlphaFold 2 is a milestone achievement in the fields of drug discovery and development, and the era of drug design characterized by AIDD is arriving.

It is a common sense that the world works under the Nature laws, which we only know very little due to many limits. However, in the era of high-throughput, big scientific experiments, high-performance computing, and big data, unprecedented big data have been obtained. Determinism-driven QSAR is incapable of dealing with the big data. AIDD paradigm characterized by data-driven (actually statistics driven) came into play (Figure 28).

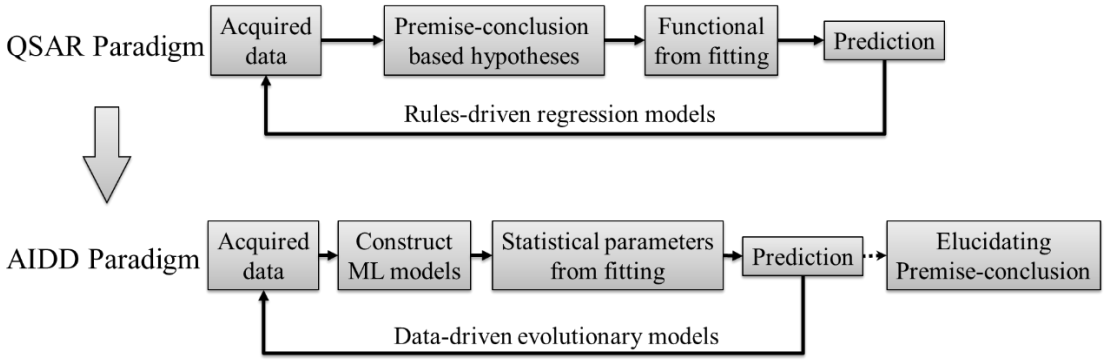


Figure 28. Transition from QSAR paradigm to AIDD paradigm

With AIDD paradigm, human's ability to discover knowledge from data has been dramatically expanded. The ability is supported by known causality rules, but by the probability laws (statistical laws) derived from big data. Although the mechanisms behind the AIDD predictions are to be elucidated, people can solve practical drug design problems sooner. The predictions can be interpreted later.

The achievements of AIDD are based on inheriting QSAR theories, methods, and data. SMILES, a typical example, has now become the natural language for AIDD to precisely describe molecular structure features. Natural language processing algorithms (such as word embedding technology) can be used to extract substructure features related to properties /

activities, which may solve the substructure partitioning and feature extraction problems that have plagued QSAR for many years; Because there is no need to define a molecular scaffold or the combination rules of substituents in advance, and SAR now completely depends on the results of statistical learning; the paradox of prediction accuracy and generality of QSAR can also be solved.

Historically, many molecular descriptors (multiple independent variables) were developed for QSAR studies. It is a great challenge for scientists to find a suitable model to drive QSAR from big biological data with thousands of descriptors. The essence of machine learning is to use statistical algorithms to find the correlation between variables from such a big data. AIDD technologies opens up new ways to solve the problems that were unable to be solved by QSAR before.

It is worth noting that ML technology is to conserve the features presented in the existing data to the greatest extent, only the minority of the data features can lead to innovations. The minority of the data points are often treated as outliers, and cannot be captured by AI methods. And these seemingly "exotic" outliers are likely to be the entrance to the world of new discoveries. It can be seen that the essence of machine learning results is conservative, and innovation is a subversion of conservatism.

Therefore, philosophically, it is unrealistic to discover innovative drugs using AIDD alone.

4.3.2 Relations among ANNs and Natural Rule Types

4.3.2.1 Essence of ANN

The human brain consists of about 86 billion neurons [241]. Each neuron receives a set of signal inputs (vector matrix as an independent variables, generally analog quantity, belonging to continuous mathematics), and outputs signals to the environment or another neuron through the cell body (functions include: numerical calculation, process control, information storage and update, information transmission) (vector matrix as a function, generally digital quantity or analog quantity, belonging to discrete or continuous mathematics).

Therefore, the essence of artificial neuron is a function, which transforms inputs (independent variables $\vec{x} = \{x_1, x_2, \dots, x_{n-1}, x_n\}$) into outputs ($\vec{y} = \{y_1, y_2, \dots, y_{m-1}, y_m\}$). There is at least one gating function (such as a *S*-function) inside the artificial neuron, which is responsible for transforming the value of $\vec{x}$ into the value of $\vec{y}$. Because both $\vec{x}$ and $\vec{y}$ can come from discrete or continuous mathematical space. Therefore, artificial neurons allow information to travel between discrete and continuous mathematical spaces.

ANN is derived from the mathematical model of neurons and can be regarded as a mini von Neumann structure computer. A group of neurons are connected in cascading, parallel, or recursion through input / output ports to form a powerful statistical computing ability. Its mathematical essence is a functional network. The dimensions of the network, the connection mode of function nodes, and the different combinations of various types of activation functions form a variety of ANN learning machines to execute different learning tasks.

4.3.2.2 ANNs and Natural Rule Types

There many types of ANN learning machines, they can only produce correct predictions if they match with nature rules. Nature rules have common features as summarized in Table 3.

Table 3. ANN and common features of nature rules

ANN Type	Common feature
Convolutional neural network (CNN)	Logicality
Deep neural network (DNN)	Hierarchy
Recurrent neural network (RNN)	Recursion
Parallel neural network (PNN)	Evolvability
Bidirectional long short memory network (BiLSTM)	Spatiotemporal
Graph neural network (GNN)	Discreteness

4.3.2.3 AI and Metarecursion

The current development of deep neural network (DNN) shows that even if we superficially simulate the working principle of neural network, it has greatly promoted science and technology.

AI, which sprouted in the 1940s, dormant for a long time due to the limits of early serial computing and chip speed, until the beginning of this century.

However, re-studying the basic theories and concepts laid by the sages of AI in the last century is still of great practical significance to the current AIDD development.

For example, the concept of "code as data and data as code" (CDDC) were also known as dual coding theory.

In 1971, Allan Paivio of the University of Western Ontario in Canada pointed out: if an object information could be represented by both image and language signal, then the image and language were processed and expressed in different ways in the human brain. Psychological codes corresponding to these representations are used to organize incoming information, which can be manipulated, stored and retrieved for subsequent use. When recalling the information, images and language codes can be used. For example, assuming that the concept of "dog" is stored in the human brain as the word "dog" and the image of the dog (shape, sound, smell and other sensory representations), when it is required to describe the concept of "dog", the human brain can retrieve the word or image about "dog" alone or at the same time. Although the word is retrieved, the dog's image will not be lost and can still be retrieved later. The ability to encode stimuli in two different ways increases the chance of remembering this concept, compared with stimuli encoded in only one way. [242]

This cognitive theory was applied in computer science; and the concept of treating computer codes as data and executing data as codes was formed.

Now, CDDC concept has evolved into a family of metarecursion concepts, that is, a new concept prefixed by meta-X, a new concept was recursively defined from the concept of X. For example, metascience is the use of scientific methodology to study science itself; Metaphysics is a the use of scientific methodology to study physics itself; Metamaterials are new materials with new functions created from existing materials. To keep this review concise, we only comment a few metarecusion concepts related to AIDD as follows.

(1) LISP language: as the second high-level programming language (the first is FORTRAN language) and the special language of early AI, CDDC is the unique attribute of LISP language, that is, the code written in LISP has the same data structure as its input and output [243]; This feature has been reserved until now and has been applied in contemporary bioinformatics. [244]

(2) Metadata: metadata refers to "data about data". For example, for a data file, the basic

information about it includes: source, purpose, creation time and date, authors, address on the network, usage standard, file size, data quality, process of creating the data, etc. This group of data is the metadata of a data file. Metadata can be divided into following classes:

- descriptive metadata
- structural metadata
- reference metadata
- administrative metadata
- legal metadata
- statistical metadata

With metadata, people can efficiently gain the data back from hundreds of millions of data files in the era of big data. These metadata can be interpreted by the metadata of the upper layer to drive the operation of other data parsing programs. This is a typical case of DBCD. In fact, metadata is a function of the original data. Theoretically, the series or parallel connection between different metadata can also form a new type of artificial neural network.

(3) Metalanguage: a metalanguage is a language about other languages. In computer science, a typical logic metalanguage (also known as formal language) is the Backus Naur form (BNF), which was invented by John Backus and Peter Naur in the 1960s. It was first used to strictly define the computer programming language [245] and specify the character set, syntax, grammer, operators, and control process of a programming language. All computer programming languages can be accurately defined by BNF. In the era of big data, in order to accelerate the mastery of data processing and shorten the learning cycle of programming languages, people have developed many scripting languages and mark-up languages. The execution of programs written in these languages is controlled by data elements of non-command sequences. Programming with these languages is also metaprogramming. Programs can be designed to read, generate, analyze or convert other programs, and even modify themselves at runtime, allowing programmers to minimize the number of lines of code expressing solutions, reduce development time, and allow programs to be modified and executed without recompiling. Artificial intelligence language LISP supports meta programming.

(4) Metalearning: metalearning was originally the theory of learning process proposed by Donald Maudsley in 1979. [246] It was used by John Biggsy in 1985 to describe the understanding and control of the learning state. [247] It originally belongs to the field of educational psychology. In recent years, metalearning has been introduced into the field of machine learning to study the self-evolution and adaptation of machine learning systems to the environment. Metalearning studies the adaptive learning process (Figure 29), studies how the learning system dynamically selects hypotheses, constantly updates / improves the prior parameters from basic learning, and improves the knowledge level of the learning machine and its adaptability to the environment. Obviously, the ability of metaprogramming or self-programming is very important for metalearning. In a typical inductive learning scenario, the learning system generates preliminary hypotheses from the data through conventional learning machines (such as decision trees, neural networks, or support vector machines), then evaluates the deviations generated by the hypotheses, and self corrects the errors. [248]

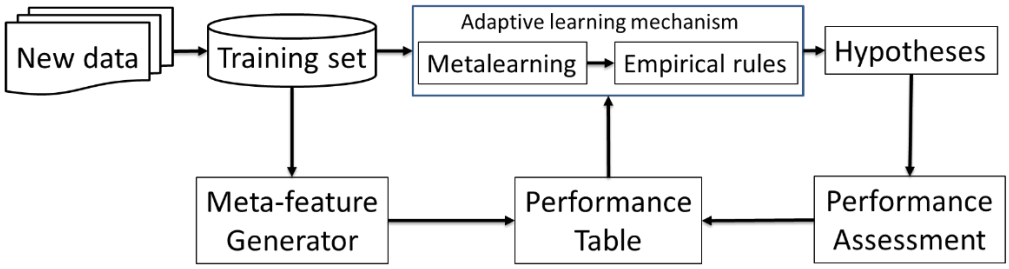


Figure 29. Metalearning process.

In 2022, Joel Lehman and co-workers proposed a large language model (LLM) used genetic algorithm to automatically generate and improve programming codes. LLM can be viewed as a newer metalearning case. [249] Jane X Wang of DeepMind team published a review on metalearning in natural learning and AI. [250]

CORRESPONDING AUTHOR INFORMATION

Jun Xu - *Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, 132 East Circle at University City, Guangzhou 510006, China*
School of Biotechnology and Health Sciences, Wuyi University, Jiangmen 529020
Emails: xujun9@mail.sysu.edu.cn; junxu@biochemomes.com
ORCID
0000-0002-1075-0337

NOTES

The authors declare no competing interests are declared.

BIOGRAPHY

Jun Xu is a full professor at Sun Yat-sen University (SYSU), holding the Director for Drug Design. He received his Ph.D. in chemistry from University of Science and Technology of China (USTC). His research focuses on drug design methodology, artificial intelligence, graph theory algorithms and the applications in discovering anti-aging, anti-metabolic-syndrome, anti-infective drugs from natural products and synthetic compounds. His career has led him from BIO-RAD Sadtler Division, Philadelphia, the Medicinal Chemistry Department at Boehringer Ingelheim, Ridgefield, to Discovery Partners International where he held the Director and Fellow for Drug Design, San Diego and then to his current position at Sun Yat-sen University in Guangzhou. He is a Fellow of Royal Chemical Society.

ACKNOWLEDGEMENTS

This review was written based on the author's teaching materials for the senior undergraduates in School of Pharmaceutical Sciences, Sun Yat-Sen University and School of Biotechnology and Health Sciences, Wuyi University.

ABBREVIATIONS

ABC	ATP binding cassette
ACF	atom center fragment
ADC	antibody drug conjugates
ADMET	absorption, distribution, metabolism, excretion, and toxicity
AI	artificial intelligence
AIDD	AI assisted drug design
AL	activity landscape
ANN	artificial neural network

BBB	blood brain barrier
BFD	big fantasy database, a high-quality protein sequence database
BiLSTM	bidirectional long short memory network
BMEC	bovine microvessel endothelial cell
BNF	the Backus Naur form
Caco-2	human colon adenocarcinoma
CASP	a critical assessment of protein structure prediction
CDDC	code as data and data as code
CDMC	chemical double mutant cycle
CL	plasma clearance
CNN	convolutional neural network
CSF	cerebrospinal fluid
CT	a connection table
DEL	DNA encoded library
DNN	deep neural network
DT	decision tree
ECF	extra cellular fluid
ECFP	extended-connectivity fingerprints
ENIAC	electronic numerical integrator and computer
FA	factor analysis
GA	genetic algorithm
GDT-TS	global distance test total score
GPCR	G protein-coupled receptors
GPU	graphics processing unit
GRU	gated recurrent unit
hERG	human potassium channel protein
HMM	hidden Markov model
HTS	high-throughput screening
IAM	immobilized artificial membranes
ILC	immobilized liposome chromatography
ITC	isothermal titration calorimetry
k-NN	k-nearest neighbor
log D	pH-dependent distribution coefficient
log P	octanol–water partition coefficient
log S	aqueous solubility
LSTM	long–short-term memory
MACCS	Molecular ACCess System
MAPK1	human mitogen activated protein kinase 1
MD	molecular dynamics
MDCK	Madin-Darby canine kidney
MDR	multiple drug resistance
MLR	multiple linear regression
MSA	multiple sequence alignments
MTR	multiple template recognition
NBC	naïve Bayes classifiers
OC	oligo compounds, the compounds with unknown targets

PAMPA	parallel artificial membrane permeability assay
PCA	principal component analysis
PLS	partial least square
PSA	polar surface area
PPB	plasma protein binding
PTM	post-translational modifications
QSAR	quantitative structure–activity relationship
RBF	radial basis function network
RBM	restricted Boltzmann machine
RF	random forest
RMSD	root mean square deviation
RNN	recurrent neural network
SAS	structure activity similarity
SBDD	structure-based drug design
SCA	scaffold based classification approach
SD	standard deviation
SDF	structure data format
SMILES	simplified molecular input line entry specification
SOM	self-organizing map
SPR	surface plasma resonance
STD-NMR	saturation transfer difference NMR
SVM	support vector machine
TPU	Google’s tensor processing unit
WLN	Wiswesser line notation

REFERENCES

1 Gibaud, S. & Jaouen, G. in *Medicinal Organometallic Chemistry* (eds Gérard Jaouen & Nils Metzler-Nolte) 1-20 (Springer Berlin Heidelberg, 2010).

2 Kubinyi, H. From narcosis to hyperspace: the history of QSAR. *Quantitative Structure‐Activity Relationships* **21**, 348-356 (2002).

3 Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **194**, 178-180 (1962).

4 DeSimone, R., Currie, K., Mitchell, S., Darrow, J. & Pippin, D. Privileged structures: applications in drug discovery. *Combinatorial chemistry & high throughput screening* **7**, 473-493 (2004).

5 Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *Journal of the American Chemical Society* **59**, 96-103, doi:10.1021/ja01280a022 (1937).

6 Yan, X., Li, J., Gu, Q. & Xu, J. gWEGA: GPU‐accelerated WEGA for molecular superposition and shape comparison. *Journal of computational chemistry* **35**, 1122-1130 (2014).

7 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. *Journal of Medicinal Chemistry* **48**, 1489-1495, doi:10.1021/jm040163o (2005).

8 Rzepa, H. S., Murray-Rust, P. & Whitaker, B. J. The Application of Chemical Multipurpose Internet Mail Extensions (Chemical MIME) Internet Standards to Electronic Mail and World Wide Web Information Exchange. *Journal of Chemical Information and Computer Sciences* **38**, 976-982, doi:10.1021/ci9803233 (1998).

9 Schneider, P. & Schneider, G. Privileged Structures Revisited. *Angewandte Chemie (International ed. in English)* **56**, 7971-7974, doi:10.1002/anie.201702816 (2017).

10 Xu, J. & Stevenson, J. Drug-like index: a new approach to measure drug-like compounds and their diversity. *Journal of Chemical Information and Computer Sciences* **40**, 1177-1187 (2000).

11 Garey, M. R. & Johnson, D. S. *Computers and intractability*. Vol. 174 (freeman San Francisco, 1979).

12 Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **39**, 2887-2893 (1996).

13 Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **42**, 1273-1280, doi:10.1021/ci010132r (2002).

14 Xu, J. 13C NMR Spectral Prediction by Means of Generalized Atom Center Fragment Method. *Molecules* **2**, 114-128 (1997).

15 Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742-754, doi:10.1021/ci100050t (2010).

16 Xu, J. & Hagler, A. Chemoinformatics and Drug Discovery. *Molecules* **7**, 566-600 (2002).

17 Favre, H. A. & Powell, W. H. *Nomenclature of organic chemistry: IUPAC recommendations and preferred names 2013*. (Royal Society of Chemistry, 2013).

18 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31-36, doi:10.1021/ci00057a005 (1988).

19 T Devinyak, O. & B Lesyk, R. 5-Year trends in QSAR and its machine learning methods. *Current Computer-Aided Drug Design* **12**, 265-271 (2016).

20 Stankevich, M. I., Stankevich, I. V. & Zefirov, N. S. Topological indices in organic chemistry. *Russian Chemical Reviews* **57**, 191 (1988).
21 Balaban, A. T. Applications of graph theory in chemistry. *Journal of chemical information and computer sciences* **25**, 334-343 (1985).
22 Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **10**, 4, doi:10.1186/s13321-018-0258-y (2018).
23 Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *Match* **56**, 237-248 (2006).
24 O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 1-14 (2011).
25 Landrum, G. (Academic Press Cambridge, 2013).
26 Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of cheminformatics* **9**, 1-19 (2017).
27 Masand, V. H. & Rastija, V. PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors. *Chemometrics and Intelligent Laboratory Systems* **169**, 12-18 (2017).
28 Fourches, D., Muratov, E. & Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **50**, 1189-1204, doi:10.1021/ci100176x (2010).
29 Fourches, D., Muratov, E. & Tropsha, A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *Journal of Chemical Information and Modeling* **56**, 1243-1252, doi:10.1021/acs.jcim.6b00129 (2016).
30 Gadaleta, D., Lombardo, A., Toma, C. & Benfenati, E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. *Journal of Cheminformatics* **10**, 60, doi:10.1186/s13321-018-0315-6 (2018).
31 Jain, A., Nandakumar, K. & Ross, A. Score normalization in multimodal biometric systems. *Pattern Recognition* **38**, 2270-2285, doi:https://doi.org/10.1016/j.patcog.2005.01.012 (2005).
32 Wan, S., Bhati, A. P., Zasada, S. J. & Coveney, P. V. Rapid, accurate, precise and reproducible ligand–protein binding free energy prediction. *Interface Focus* **10**, 20200007 (2020).
33 Johnson, M. A. & Maggiora, G. M. *Concepts and applications of molecular similarity*. (Wiley, 1990).
34 Cheeseright, T. J., Mackey, M. D., Melville, J. L. & Vinter, J. G. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *Journal of chemical information and modeling* **48**, 2108-2117 (2008).
35 Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics* **5**, 1-17 (2013).
36 Free, S. M. & Wilson, J. W. A mathematical contribution to structure-activity studies. *Journal of medicinal chemistry* **7**, 395-399 (1964).
37 Frey, K. M. Structure activity relationship (SAR) maps: A student-friendly tool to teach medicinal chemistry in integrated pharmacotherapy courses. *Currents in Pharmacy Teaching and Learning* **12**, 339-346, doi:https://doi.org/10.1016/j.cptl.2019.12.014 (2020).
38 Biela, A., Betz, M., Heine, A. & Klebe, G. Water Makes the Difference: Rearrangement of Water Solvation Layer Triggers Non‐additivity of Functional Group Contributions in Protein–Ligand Binding. *ChemMedChem* **7**, 1423-1434 (2012).
39 Nasief, N. N., Tan, H., Kong, J. & Hangauer, D. Water mediated ligand functional group cooperativity: the contribution of a methyl group to binding affinity is enhanced by a COO– group through changes in the structure and thermodynamics of the hydration waters of ligand–thermolysin complexes. *Journal of medicinal chemistry* **55**, 8283-8302 (2012).

40 Baum, B. *et al.* Non-additivity of functional group contributions in protein–ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *Journal of molecular biology* **397**, 1042-1054 (2010).
41 Muley, L. *et al.* Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. *Journal of medicinal chemistry* **53**, 2126-2135 (2010).
42 Kuhn, B., Mohr, P. & Stahl, M. Intramolecular hydrogen bonding in medicinal chemistry. *Journal of medicinal chemistry* **53**, 2601-2611 (2010).
43 Kramer, C., Fuchs, J. E. & Liedl, K. R. Strong nonadditivity as a key structure–activity relationship feature: distinguishing structural changes from assay artifacts. *Journal of chemical information and modeling* **55**, 483-494 (2015).
44 Gomez, L. *et al.* Mathematical and Structural Characterization of Strong Nonadditive Structure–Activity Relationship Caused by Protein Conformational Changes. *Journal of Medicinal Chemistry* **61**, 7754-7766 (2018).
45 Patel, Y. *et al.* Assessment of additive/nonadditive effects in structure− activity relationships: implications for iterative drug design. *Journal of medicinal chemistry* **51**, 7552-7562 (2008).
46 Kramer, C. Nonadditivity Analysis. *Journal of Chemical Information and Modeling* **59**, 4034-4042, doi:10.1021/acs.jcim.9b00631 (2019).
47 Cockroft, S. L. & Hunter, C. A. Chemical double-mutant cycles: dissecting non-covalent interactions. *Chemical Society Reviews* **36**, 172-188 (2007).
48 Fischer, F. R., Schweizer, W. B. & Diederich, F. Molecular torsion balances: Evidence for favorable orthogonal dipolar interactions between organic fluorine and amide groups. *Angewandte Chemie International Edition* **46**, 8270-8273 (2007).
49 Maggiora, G. M. Vol. 46 1535-1535 (ACS Publications, 2006).
50 Silipo, C. & Vittoria, A. in *European Symposium on Quantitative Structure-Activity Relationships 1990: Sorrento, Italy).* (Distributors for the US and Canada, Elsevier Science).
51 Stumpfe, D. & Bajorath, J. Methods for SAR visualization. *RSC advances* **2**, 369-378 (2012).
52 Shanmugasundaram, V. & Maggiora, G. in *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY.* U271-U271 (AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA).
53 Pérez-Villanueva, J. *et al.* Structure–activity relationships of benzimidazole derivatives as antiparasitic agents: dual activity-difference (DAD) maps. *MedChemComm* **2**, 44-49 (2011).
54 Yongye, A. B. *et al.* Consensus models of activity landscapes with multiple chemical, conformer, and property representations. *Journal of chemical information and modeling* **51**, 1259-1270 (2011).
55 Peltason, L., Iyer, P. & Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *Journal of Chemical Information and Modeling* **50**, 1021-1033, doi:10.1021/ci100091e (2010).
56 Stumpfe, D., Hu, H. & Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **4**, 14360-14368, doi:10.1021/acsomega.9b02221 (2019).
57 Pennington, L. D. & Moustakas, D. T. The Necessary Nitrogen Atom: A Versatile High-Impact Design Element for Multiparameter Optimization. *Journal of Medicinal Chemistry* **60**, 3552-3579, doi:10.1021/acs.jmedchem.6b01807 (2017).
58 Leung, C. S., Leung, S. S. F., Tirado-Rives, J. & Jorgensen, W. L. Methyl Effects on Protein–Ligand Binding. *Journal of Medicinal Chemistry* **55**, 4489-4500, doi:10.1021/jm3003697 (2012).
59 Bajorath, J. Modeling of activity landscapes for drug discovery. *Expert Opinion on Drug Discovery* **7**, 463-473 (2012).

60 Hu, H. & Bajorath, J. Systematic identification of activity cliffs with dual-atom replacements and their rationalization on the basis of single-atom replacement analogs and X-ray structures. *Chemical Biology & Drug Design* **99**, 308-319 (2022).
61 Horvath, D. Quantitative structure–activity relantionships: In silico chemistry or high tech alchemy. *Rev Roum Chim* **55**, 783-801 (2010).
62 Medina-Franco, J. L. Activity Cliffs: Facts or Artifacts? *Chemical Biology & Drug Design* **81**, 553-556, doi:https://doi.org/10.1111/cbdd.12115 (2013).
63 Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **38**, 983-996, doi:10.1021/ci9800211 (1998).
64 Tanimoto, T. T. IBM internal report. *Nov* **17**, 1957 (1957).
65 Tversky, A. Features of similarity. *Psychological review* **84**, 327 (1977).
66 Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297-302 (1945).
67 Sorensen, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skar.* **5**, 1-34 (1948).
68 Schneider, N., Lewis, R. A., Fechner, N. & Ertl, P. Chiral cliffs: investigating the influence of chirality on binding affinity. *ChemMedChem* **13**, 1315-1324 (2018).
69 Pirhadi, S., Shiri, F. & Ghasemi, J. B. Multivariate statistical analysis methods in QSAR. *Rsc Advances* **5**, 104635-104665 (2015).
70 Law, J. & Rennie, R. *A Dictionary of Chemistry*. (Oxford University Press, 2020).
71 Willett, P. Similarity searching using 2D structural fingerprints. *Chemoinformatics and computational chemical biology*, 133-158 (2010).
72 Harrell, F. E. Regression modeling strategies. *Bios* **330**, 14 (2017).
73 Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug discovery today* **15**, 444-450 (2010).
74 Efron, B. Bayes' theorem in the 21st century. *Science* **340**, 1177-1178 (2013).
75 Murphy, K. P. Naive bayes classifiers. *University of British Columbia* **18**, 1-8 (2006).
76 Liu, Z. *et al.* ChemStable: a web server for rule-embedded naive Bayesian learning approach to predict compound stability. *Journal of computer-aided molecular design* **28**, 941-950 (2014).
77 Podlewska, S. & Kafel, R. MetStabOn—online platform for metabolic stability predictions. *International journal of molecular sciences* **19**, 1040 (2018).
78 Perryman, A. L. *et al.* Naïve bayesian models for Vero cell cytotoxicity. *Pharmaceutical research* **35**, 1-10 (2018).
79 Pei, D., Gong, Y., Kang, H., Zhang, C. & Guo, Q. Accurate and rapid screening model for potential diabetes mellitus. *BMC medical informatics and decision making* **19**, 1-8 (2019).
80 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32, doi:10.1023/A:1010933404324 (2001).
81 Tanaka, K. *et al.* in *Intelligent Computing Theories and Application.* (eds De-Shuang Huang *et al.*) 628-644 (Springer International Publishing).
82 Katoch, S., Chauhan, S. S. & Kumar, V. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications* **80**, 8091-8126 (2021).
83 Datta, S., Dev, V. A. & Eden, M. R. Hybrid genetic algorithm-decision tree approach for rate constant prediction using structures of reactants and solvent for Diels-Alder reaction. *Computers & Chemical Engineering* **106**, 690-698 (2017).
84 Murtagh, F. & Contreras, P. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 86-97 (2012).

85 Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**, 4576-4579 (1990).
86 Ward, J. Hierarchical grouping to optimize an objective function J Am Stat Assoc 58: 236–244. *Find this article online* (1963).
87 Chandra, M. P. in *Proceedings of the National Institute of Sciences of India.* 49-55.
88 Hamming, R. W. Error detecting and error correcting codes. *The Bell system technical journal* **29**, 147-160 (1950).
89 Szekely, G. J. & Rizzo, M. L. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification* **22**, 151-183, doi:10.1007/s00357-005-0012-9 (2005).
90 Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* **58**, 236-244 (1963).
91 Jarvis, R. A. & Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on computers* **100**, 1025-1034 (1973).
92 Malhat, M. G., Mousa, H. M. & El-Sisi, A. B. in *2014 9th International Conference on Informatics and Systems.* DEKM-61-DEKM-66 (IEEE).
93 Jöreskog, M. K. G. in *Principals of Modern Psychological Measurement* 217-228 (Routledge, 2012).
94 Cox, M. A. & Cox, T. F. in *Handbook of data visualization* 315-347 (Springer, 2008).
95 Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).
96 Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572, doi:10.1080/14786440109462720 (1901).
97 Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics* **43**, 59-69 (1982).
98 Bauknecht, H. *et al.* Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *Journal of Chemical Information and Computer Sciences* **36**, 1205-1213, doi:10.1021/ci960346m (1996).
99 Gasteiger, J., Teckentrup, A., Terfloth, L. & Spycher, S. Neural networks as data mining tools in drug design. *Journal of physical organic chemistry* **16**, 232-245 (2003).
100 Xu, J. A new approach to finding natural chemical structure classes. *Journal of medicinal chemistry* **45**, 5311-5320 (2002).
101 Zamora, A. An algorithm for finding the smallest set of smallest rings. *Journal of Chemical Information and Computer Sciences* **16**, 40-43 (1976).
102 McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5**, 115-133, doi:10.1007/BF02478259 (1943).
103 Hebb, D. O. *The organization of behavior: A neuropsychological theory*. (Psychology Press, 2005).
104 Farley, B. & Clark, W. Simulation of self-organizing systems by digital computer. *Transactions of the IRE Professional Group on Information Theory* **4**, 76-84, doi:10.1109/TIT.1954.1057468 (1954).
105 Mariantoni, M. *et al.* Implementing the quantum von Neumann architecture with superconducting circuits. *Science* **334**, 61-65 (2011).

106 Herculano-Houzel, S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences* **109**, 10661-10668, doi:doi:10.1073/pnas.1201895109 (2012).
107 Gumbleton, M. & Audus, K. L. Progress and limitations in the use of in vitro cell cultures to serve as a permeability screen for the blood-brain barrier. *Journal of pharmaceutical sciences* **90**, 1681-1698 (2001).
108 Agoram, B., Woltosz, W. S. & Bolger, M. B. Predicting the impact of physiological and biochemical processes on oral drug bioavailability. *Advanced drug delivery reviews* **50**, S41-S67 (2001).
109 de Lange, E. & Danhof, M. Considerations in the use of cerebrospinal fluid pharmacokinetics to predict brain target concentrations in the clinical setting. *Clinical pharmacokinetics* **41**, 691-703 (2002).
110 Raevsky, O. A. *et al.* Physicochemical property profile for brain permeability: comparative study by different approaches. *Journal of drug targeting* **24**, 655-662 (2016).
111 Prajapati, J., Patel, H. & Agrawal, Y. K. Targeted drug delivery for central nervous system: a review. *Int J Pharm Pharm Sci* **3**, 32-38 (2012).
112 Ayrton, A. & Morgan, P. Role of transport proteins in drug absorption, distribution and excretion. *Xenobiotica* **31**, 469-497 (2001).
113 Smith, D. A., van de Waterbeemd, H. & Walker, D. K. *Methods and principles in medicinal chemistry: Pharmacokinetics and metabolism in drug design*. (Wiley-VCH, 2006).
114 van de Waterbeemd, H., Smith, D. A. & Jones, B. C. Lipophilicity in PK design: methyl, ethyl, futile. *Journal of computer-aided molecular design* **15**, 273-286 (2001).
115 Lombardo, F., Obach, R. S., Shalaeva, M. Y. & Gao, F. Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data. *Journal of medicinal chemistry* **45**, 2867-2876 (2002).
116 Schneider, G., Coassolo, P. & Lavé, T. Combining in vitro and in vivo pharmacokinetic data for prediction of hepatic drug clearance in humans by artificial neural networks and multivariate statistical techniques. *Journal of medicinal chemistry* **42**, 5072-5076 (1999).
117 Hilal, S., Karickhoff, S. & Carreira, L. A rigorous test for SPARC's chemical reactivity models: Estimation of more than 4300 ionization pKas. *Quantitative Structure -Activity Relationships* **14**, 348-355 (1995).
118 Hong, H. *et al.* Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of chemical information and modeling* **48**, 1337-1344 (2008).
119 Todeschini, R. & Consonni, V. *Handbook of molecular descriptors*. (John Wiley & Sons, 2008).
120 Podlogar, B., Muegge, I. & Brice, L. Computational methods to estimate drug development parameters. *Current Opinion in Drug Discovery & Development* **4**, 102-109 (2001).
121 Thompson, M. C., Yeates, T. O. & Rodriguez, J. A. Advances in methods for atomic resolution macromolecular structure determination. *F1000Research* **9** (2020).
122 Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181**, 223-230, doi:doi:10.1126/science.181.4096.223 (1973).
123 Naganathan, A. N. & Muñoz, V. Scaling of Folding Times with Protein Size. *Journal of the American Chemical Society* **127**, 480-481, doi:10.1021/ja044449u (2005).
124 Sippl, M. J. Calculation of conformational ensembles from potentials of mena force: an approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology* **213**, 859-883 (1990).
125 Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338**, 1042-1046, doi:doi:10.1126/science.1219021 (2012).

126 Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences* **37**, 205-211 (1951).
127 Dill, K. A. Theory for the folding and stability of globular proteins. *Biochemistry* **24**, 1501-1509 (1985).
128 Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685 (1993).
129 Martí-Renom, M. A. *et al.* Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure* **29**, 291-325, doi:10.1146/annurev.biophys.29.1.291 (2000).
130 Kaczanowski, S. & Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures? *Theoretical Chemistry Accounts* **125**, 643-650, doi:10.1007/s00214-009-0656-3 (2010).
131 Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* **5**, 823-826, doi:https://doi.org/10.1002/j.1460-2075.1986.tb04288.x (1986).
132 Baker, D. & Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **294**, 93-96, doi:doi:10.1126/science.1065659 (2001).
133 Blake, J. D. & Cohen, F. E. Pairwise sequence alignment below the twilight zone11Edited by B. Honig. *Journal of Molecular Biology* **307**, 721-735, doi:https://doi.org/10.1006/jmbi.2001.4495 (2001).
134 Kabir, M. N. & Wong, L. EnsembleFam: towards more accurate protein family prediction in the twilight zone. *BMC bioinformatics* **23**, 1-20 (2022).
135 Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research* **47**, D520-D528 (2019).
136 Shindyalov, I., Kolchanov, N. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Engineering, Design and Selection* **7**, 349-358 (1994).
137 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67-72 (2009).
138 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
139 Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190 (2012).
140 Williamson, A. R. Creating a structural genomics consortium. *Nature Structural Biology* **7**, 953-953, doi:10.1038/80726 (2000).
141 Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics* **23**, ii-iv, doi:https://doi.org/10.1002/prot.340230303 (1995).
142 Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res* **31**, 3370-3374, doi:10.1093/nar/gkg571 (2003).
143 Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **29**, 291-325 (2000).
144 Kaur, K., Chakraborty, S. & Gupta, M. K. in *Journal of Physics: Conference Series.* 012028 (IOP Publishing).

145 Greer, J. Comparative model-building of the mammalian serine proteases. *Journal of Molecular Biology* **153**, 1027-1042, doi:https://doi.org/10.1016/0022-2836(81)90465-4 (1981).
146 Wallner, B. & Elofsson, A. All are not equal: A benchmark of different homology modeling programs. *Protein Science* **14**, 1315-1327, doi:https://doi.org/10.1110/ps.041253405 (2005).
147 Levitt, M. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology* **226**, 507-533, doi:https://doi.org/10.1016/0022-2836(92)90964-L (1992).
148 Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* **234**, 779-815, doi:https://doi.org/10.1006/jmbi.1993.1626 (1993).
149 Fiser, A. & Šali, A. in *Methods in enzymology* Vol. 374 461-491 (Elsevier, 2003).
150 Venclovas, Č. & Margelevičius, M. Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins: Structure, Function, and Bioinformatics* **61**, 99-105, doi:https://doi.org/10.1002/prot.20725 (2005).
151 Ginalski, K. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology* **16**, 172-177, doi:https://doi.org/10.1016/j.sbi.2006.02.003 (2006).
152 Flohil, J. A., Vriend, G. & Berendsen, H. J. C. Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins: Structure, Function, and Bioinformatics* **48**, 593-604, doi:https://doi.org/10.1002/prot.10105 (2002).
153 Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95-99, doi:https://doi.org/10.1016/S0022-2836(63)80023-6 (1963).
154 Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics* **17**, 355-362, doi:https://doi.org/10.1002/prot.340170404 (1993).
155 Lazaridis, T. & Karplus, M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation 11Edited by A. R. Fersht. *Journal of Molecular Biology* **288**, 477-487, doi:https://doi.org/10.1006/jmbi.1999.2685 (1999).
156 Eramian, D. *et al.* A composite score for predicting errors in protein structure models. *Protein Science* **15**, 1653-1666, doi:https://doi.org/10.1110/ps.062095806 (2006).
157 Gollery, M. Bioinformatics: sequence and genome analysis. *Clinical Chemistry* **51**, 2219-2220 (2005).
158 Service, R. F. (American Association for the Advancement of Science, 2020).
159 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
160 Callaway, E. What's next for AlphaFold and the AI protein-folding revolution. *Nature* **604**, 234-238 (2022).
161 Zweckstetter, M. NMR hawk‐eyed view of AlphaFold 2 structures. *Protein Science* **30**, 2333-2337 (2021).
162 Hedley, P. L. *et al.* The genetic basis of long QT and short QT syndromes: a mutation update. *Human mutation* **30**, 1486-1511 (2009).
163 Zhou, P.-z., Babcock, J., Liu, L.-q., Li, M. & Gao, Z.-b. Activation of human ether-a-go-go related gene (hERG) potassium channels by small molecules. *Acta pharmacologica Sinica* **32**, 781-788 (2011).
164 Wang, W. & MacKinnon, R. Cryo-EM structure of the open human ether-à-go-go-related K+ channel hERG. *Cell* **169**, 422-430. e410 (2017).
165 Stein, R. A. & Mchaourab, H. S. Modeling alternate conformations with AlphaFold 2 via modification of the multiple sequence alignment. *bioRxiv* (2021).

166 Nishi, H., Shaytan, A. & Panchenko, A. R. Physicochemical mechanisms of protein regulation by phosphorylation. *Frontiers in genetics* **5**, 270 (2014).
167 Huse, M. & Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **109**, 275-282 (2002).
168 Schauperl, M. & Denny, R. A. AI-Based Protein Structure Prediction in Drug Discovery: Impacts and Challenges. *Journal of Chemical Information and Modeling*, doi:10.1021/acs.jcim.2c00026 (2022).
169 Laskowski, R. A., Gerick, F. & Thornton, J. M. The structural basis of allosteric regulation in proteins. *FEBS letters* **583**, 1692-1698 (2009).
170 Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Nature Precedings*, 1-1 (2010).
171 Zhu, G. *et al.* Mutant p53 in cancer progression and targeted therapies. *Frontiers in oncology* **10**, 595187 (2020).
172 Chaudhuri, T. K. & Paul, S. Protein-misfolding diseases and chaperone-based therapeutic approaches. *The FEBS journal* **273**, 1331-1349 (2006).
173 Gong, Z. *et al.* Compound libraries: recent advances and their applications in drug discovery. *Current drug discovery technologies* **14**, 216-228 (2017).
174 Drew, K. L., Baiman, H., Khwaounjoo, P., Yu, B. & Reynisson, J. Size estimation of chemical space: how big is it? *Journal of Pharmacy and Pharmacology* **64**, 490-495 (2012).
175 Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* **16**, 3-50, doi:https://doi.org/10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6 (1996).
176 Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *Journal of Chemical Information and Computer Sciences* **43**, 374-380, doi:10.1021/ci0255782 (2003).
177 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **23**, 3-25, doi:https://doi.org/10.1016/S0169-409X(96)00423-1 (1997).
178 Teague, S. J., Davis, A. M., Leeson, P. D. & Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angewandte Chemie International Edition* **38**, 3743-3748, doi:https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U (1999).
179 Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today* **8**, 876-877, doi:10.1016/s1359-6446(03)02831-9 (2003).
180 Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. Fragment-based lead discovery. *Nature Reviews Drug Discovery* **3**, 660-672, doi:10.1038/nrd1467 (2004).
181 Lucas, X., Grüning, B. r. A., Bleher, S. & Günther, S. The purchasable chemical space: a detailed picture. *Journal of chemical information and modeling* **55**, 915-924 (2015).
182 Favalli, N., Bassi, G., Scheuermann, J. & Neri, D. DNA-encoded chemical libraries–achievements and remaining challenges. *FEBS letters* **592**, 2168-2180 (2018).
183 Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855-861, doi:10.1038/nature03193 (2004).
184 Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **4**, 90-98 (2012).
185 Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics* **1**, 1-11 (2009).

186 Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **4**, 268-276, doi:10.1021/acscentsci.7b00572 (2018).
187 Zheng, S., Yan, X., Yang, Y. & Xu, J. Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of chemical information and modeling* **59**, 914-923 (2019).
188 Zheng, S. *et al.* QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *Journal of cheminformatics* **11**, 1-12 (2019).
189 Li, X. *et al.* Deepchemstable: chemical stability prediction with an attention-based graph convolution network. *Journal of chemical information and modeling* **59**, 1044-1049 (2019).
190 Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science* **4**, 120-131, doi:10.1021/acscentsci.7b00512 (2018).
191 Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence* **2**, 134-140 (2020).
192 Zhang, J. & Chen, H. De Novo Molecule Design Using Molecular Generative Models Constrained by Ligand–Protein Interactions. *Journal of Chemical Information and Modeling*, doi:10.1021/acs.jcim.2c00177 (2022).
193 Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling* **60**, 47-55 (2019).
194 Su, S. *et al.* Predicting the feasibility of copper (i)-catalyzed alkyne–azide cycloaddition reactions using a recurrent neural network with a self-attention mechanism. *Journal of Chemical Information and Modeling* **60**, 1165-1174 (2020).
195 Yang, Y. *et al.* SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical science* **11**, 8312-8322 (2020).
196 Liu, Z. *et al.* Deep learning enables discovery of highly potent anti-osteoporosis natural products. *European Journal of Medicinal Chemistry* **210**, 112982 (2021).
197 Flam-Shepherd, D., Zhu, K. & Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nature Communications* **13**, 3293, doi:10.1038/s41467-022-30839-x (2022).
198 Mahmood, O., Mansimov, E., Bonneau, R. & Cho, K. Masked graph modeling for molecule generation. *Nature Communications* **12**, 3156, doi:10.1038/s41467-021-23415-2 (2021).
199 Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic acids research* **45**, D945-D954 (2017).
200 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. *Journal of chemical information and modeling* **52**, 2864-2875 (2012).
201 Corey, E. J. General methods for the construction of complex molecules. *Pure and Applied Chemistry* **14**, 19-38, doi:doi:10.1351/pac196714010019 (1967).
202 Wang, Z., Zhang, W. & Liu, B. Computational Analysis of Synthetic Planning: Past and Future. *Chinese Journal of Chemistry* **39**, 3127-3143 (2021).
203 Nugmanov, R., Dyubankova, N., Gedich, A. & Wegner, J. K. Bidirectional Graphormer for Reactivity Understanding: Neural Network Trained to Reaction Atom-to-Atom Mapping Task. *Journal of Chemical Information and Modeling*, doi:10.1021/acs.jcim.2c00344 (2022).
204 Li, Y. H. *et al.* The human kinome targeted by FDA approved multi-target drugs and combination products: A comparative study from the drug-target interaction network perspective. *PLoS One* **11**, e0165737 (2016).
205 Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery* **16**, 19-34, doi:10.1038/nrd.2016.230 (2017).

206 Ihlenfeldt, W. D. PubChem. *Applied Chemoinformatics: Achievements and Future Opportunities*, 245-258 (2018).
207 Gilson, M. K. *et al.* BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045-D1053 (2016).
208 Du, J. *et al.* cBinderDB: a covalent binding agent database. *Bioinformatics* **33**, 1258-1260 (2017).
209 Wang, Z., Liang, L., Yin, Z. & Lin, J. Improving chemical similarity ensemble approach in target prediction. *Journal of cheminformatics* **8**, 1-10 (2016).
210 Gfeller, D. *et al.* SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic acids research* **42**, W32-W38 (2014).
211 Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proceedings of the National Academy of Sciences* **111**, 4067-4072 (2014).
212 Nickel, J. *et al.* SuperPred: update on drug classification and target prediction. *Nucleic acids research* **42**, W26-W31 (2014).
213 Awale, M. & Reymond, J.-L. The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *Journal of cheminformatics* **9**, 1-10 (2017).
214 Lee, K., Lee, M. & Kim, D. Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server. *BMC Bioinformatics* **18**, 567, doi:10.1186/s12859-017-1960-x (2017).
215 Cereto-Massagué, A. *et al.* Tools for in silico target fishing. *Methods* **71**, 98-103, doi:https://doi.org/10.1016/j.ymeth.2014.09.006 (2015).
216 Cheng, T., Li, Q., Wang, Y. & Bryant, S. H. Identifying Compound-Target Associations by Combining Bioactivity Profile Similarity Search and Public Databases Mining. *Journal of Chemical Information and Modeling* **51**, 2440-2448, doi:10.1021/ci200192v (2011).
217 Fliri, A. F., Loging, W. T., Thadeio, P. F. & Volkmann, R. A. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proceedings of the National Academy of Sciences* **102**, 261-266, doi:doi:10.1073/pnas.0407790101 (2005).
218 Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic acids research* **47**, D1056-D1065 (2019).
219 Agamah, F. E. *et al.* Computational/in silico methods in drug target and lead prediction. *Briefings in bioinformatics* **21**, 1663-1675 (2020).
220 Chen, Y. & Zhi, D. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins: Structure, Function, and Bioinformatics* **43**, 217-226 (2001).
221 Wang, J.-C., Chu, P.-Y., Chen, C.-M. & Lin, J.-H. idTarget: a web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. *Nucleic acids research* **40**, W393-W399 (2012).
222 Luo, H. *et al.* DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical–protein interactome. *Nucleic acids research* **39**, W492-W498 (2011).
223 Li, H. *et al.* TarFisDock: a web server for identifying drug targets with docking approach. *Nucleic acids research* **34**, W219-W224 (2006).
224 Xie, L., Xie, L. & Bourne, P. E. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **25**, i305-i312 (2009).
225 Schmitt, S., Kuhn, D. & Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *Journal of molecular biology* **323**, 387-406 (2002).
226 Jambon, M. *et al.* The SuMo server: 3D search for protein functional sites. *Bioinformatics* **21**, 3929-3930 (2005).

227 Yeturu, K. & Chandra, N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC bioinformatics* **9**, 1-17 (2008).
228 Hoffmann, B., Zaslavskiy, M., Vert, J.-P. & Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC bioinformatics* **11**, 1-16 (2010).
229 Chartier, M., Adriansen, E. & Najmanovich, R. IsoMIF Finder: online detection of binding site molecular interaction field similarities. *Bioinformatics* **32**, 621-623 (2016).
230 Kellenberger, E., Schalon, C. & Rognan, D. How to measure the similarity between protein ligand-binding sites? *Current Computer-Aided Drug Design* **4**, 209 (2008).
231 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).
232 Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science* **313**, 504-507 (2006).
233 Madhukar, N. S. *et al.* A Bayesian machine learning approach for drug target identification using diverse data types. *Nature communications* **10**, 1-14 (2019).
234 Mamoshina, P. *et al.* Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Frontiers in genetics* **9**, 242 (2018).
235 Zeng, X. *et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science* **11**, 1775-1797 (2020).
236 Wen, M. *et al.* Deep-learning-based drug–target interaction prediction. *Journal of proteome research* **16**, 1401-1409 (2017).
237 Chen, L. & Wu, J. Vol. 7 185-186 (Oxford University Press, 2015).
238 Muzio, G., O'Bray, L. & Borgwardt, K. Biological network analysis with deep learning. *Briefings in bioinformatics* **22**, 1515-1530 (2021).
239 Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581-1592 (2018).
240 Kaplan, A. & Haenlein, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons* **62**, 15-25, doi:https://doi.org/10.1016/j.bushor.2018.08.004 (2019).
241 Azevedo, F. A. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology* **513**, 532-541 (2009).
242 Reed, S. K. *Cognition: Theories and applications*. (CENGAGE learning, 2012).
243 McCarthy, J. Recursive functions of symbolic expressions and their computation by machine, part I. *Communications of the ACM* **3**, 184-195 (1960).
244 Harrington, K. I., Rueden, C. T. & Eliceiri, K. W. FunImageJ: a Lisp framework for scientific image processing. *Bioinformatics* **34**, 899-900 (2018).
245 Knuth, D. E. Backus normal form vs. backus naur form. *Communications of the ACM* **7**, 735-736 (1964).
246 Maudsley, D. B. *A Theory of Meta-learning and Principles of Facilitation : an Organismic Perspective*. (Thesis (Ed.D.)--University of Toronto, 1979).
247 BIGGS, J. B. THE ROLE OF METALEARNING IN STUDY PROCESSES. *British Journal of Educational Psychology* **55**, 185-212, doi:https://doi.org/10.1111/j.2044-8279.1985.tb02625.x (1985).
248 Vilalta, R. & Drissi, Y. A Perspective View and Survey of Meta-Learning. *Artificial Intelligence Review* **18**, 77-95, doi:10.1023/A:1019956318069 (2002).
249 Lehman, J. *et al.* Evolution through Large Models. *arXiv preprint arXiv:2206.08896* (2022).
250 Wang, J. X. Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences* **38**, 90-95, doi:https://doi.org/10.1016/j.cobeha.2021.01.002 (2021).

251 Wooley, J. C., Godzik, A. & Friedberg, I. A primer on metagenomics. *PLoS Comput Biol* **6**, e1000667-e1000667, doi:10.1371/journal.pcbi.1000667 (2010).
252 Kshetrimayum, R. S. A brief intro to metamaterials. *IEEE Potentials* **23**, 44-46, doi:10.1109/MP.2005.1368916 (2005).