Review paper

doi:10.5633/amm.2026.0115

Molecular Descriptors – A Cornerstone in Translational Drug Discovery

Aleksandar M. Veselinović

University of Niš, Faculty of Medicine, Departmant of Chemistry, Niš, Serbia

Contact: Aleksandar M. Veselinović,
81 Dr Zorana Djindjića Blvd., 18000 Niš, Serbia
E-mail: aveselinovic@medfak.ni.ac.rs

Abstract

Molecular descriptors are a cornerstone of cheminformatics, enabling the transformation of molecular structures into quantitative representations used in QSAR and QSPR modeling. This review offers a comprehensive and structured overview of descriptor classes, spanning from zero-dimensional (0D) constitutional and one-dimensional (1D) fragment-based to two-dimensional (2D) topological, three-dimensional (3D) geometrical, and four-dimensional (4D) conformational descriptors. The survey extends to pharmacophore-based, quantum-chemical, empirical, SMILES-based, and SHAP-based descriptors, presenting their theoretical foundations, computational methods, and relevance across drug discovery, toxicology, and materials science. Each descriptor type is examined with respect to the information it captures—from elemental composition and functional group patterns to spatial configurations, electronic properties, conformational ensembles, and symbolic encodings. Emphasis is placed on how these descriptors are computed, what structural or dynamic features they represent, and how they contribute to predictive modeling. Special categories, such as 4D descriptors that incorporate molecular flexibility and SHAP-based descriptors that enable model interpretability, illustrate the field's ongoing evolution toward more informative and explainable representations. The review also considers hybrid approaches and the incorporation of descriptors into modern machine learning frameworks, including Monte Carlo optimization, ensemble modeling, and AI-based

prediction systems. Finally, future directions are discussed, including the development of transferable, dynamic, and biologically contextual descriptors capable of capturing multi-scale chemical behavior. These innovations aim to bridge theoretical precision with data-driven modeling, enhancing both predictive performance and mechanistic insight across cheminformatics applications.

Keywords: Molecular descriptors; QSAR/QSPR modeling; Drug discovery; ADMET; Cheminformatics

Pregledni rad

doi:10.5633/amm.2026.0115

Molekulski deskriptori – temelj translacionog otkrivanja lekova

Aleksandar M. Veselinović

Univerzitet u Nišu, Medicinski fakultet, Katedra Hemija, Niš, Srbija

Kontakt: Aleksandar M. Veselinović,
Bulevar dr Zorana Đinđića 81, 18000 Niš, Srbija
E-mail: aveselinovic@medfak.ni.ac.rs

Molekulski deskriptori predstavljaju kamen temeljac hemoinformatike, omogućavajući transformaciju molekulskih struktura u kvantitativne reprezentacije koje se koriste u QSAR i QSPR modelovanju. Ovaj pregled nudi sveobuhvatan i jasno strukturiran prikaz klasa deskriptora, od nula-dimenzionalnih (0D) konstitucionih i jednodimenzionalnih (1D) fragmentnih, preko dvodimenzionalnih (2D) topoloških i trodimenzionalnih (3D) geometrijskih, do četvorodimenzionalnih (4D) konformacionih deskriptora. Pregled se dalje proteže na farmakoforne, kvantno-hemijske, empirijske, SMILES-zasnovane i SHAP-zasnovane deskriptore, uz predstavljanje njihovih teorijskih osnova, računarskih metoda i relevantnosti u otkrivanju lekova, toksikologiji i nauci o materijalima. Svaka vrsta deskriptora razmatra se prema informacijama koje obuhvata — od elementarnog sastava i obrazaca funkcionalnih grupa do prostorne organizacije, elektronskih svojstava, konformacionih ansambala i simboličkih kodiranja. Poseban akcenat stavljen je na to kako se ovi deskriptori računaju, koje strukturne ili dinamičke karakteristike predstavljaju i kako doprinose prediktivnom modelovanju. Posebne kategorije, poput 4D deskriptora koji uključuju molekulsku fleksibilnost i SHAP-zasnovanih deskriptora koji omogućavaju interpretabilnost modela, ilustruju kontinuiranu evoluciju ka informativnijim i objašnjivijim reprezentacijama. U radu se razmatraju i hibridni pristupi, kao i uključivanje deskriptora u savremene okvire mašinskog učenja, uključujući Monte Karlo optimizaciju, ansambl-modelovanje i sisteme za predikciju zasnovane na veštačkoj inteligenciji. Na kraju se diskutuje o budućim pravcima, uključujući razvoj prenosivih, dinamičkih i biološki

kontekstualizovanih deskriptora sposobnih da obuhvate višeskalno hemijsko ponašanje. Ove inovacije imaju za cilj da premoste teorijsku preciznost i modele vođene podacima, unapređujući i prediktivne performanse i mehanistički uvid u primenama hemoinformatike.

Ključne reči: Molekulski deskriptori, QSAR/QSPR modelovanje, otkrivanje lekova, ADMET, hemoinformatika

Introduction

Molecular descriptors are fundamental to building Quantitative Structure–Activity Relationship (QSAR) and Quantitative Structure–Property Relationship (QSPR) models, which are critical tools in drug discovery, toxicology, and materials science. Originating from the seminal work of Hansch and Fujita in the 1960s, these descriptors bridge chemical structures with numerical data, enabling statistical and machine learning analyses (1–3). By predicting the properties of novel compounds, descriptors streamline experimental research, reducing costs and timelines, which highlights their increasing importance in both academic and industrial contexts (4–6).

Descriptors are numerical values derived from molecular representations, such as Simplified Molecular Input Line Entry System (SMILES) notations, graph-based encodings, connectivity matrices, or quantum-chemical calculations. They capture diverse molecular characteristics, including topological, geometric, electronic, steric, and pharmacophoric properties (7-14). Todeschini and Consonni note that descriptors are not arbitrary but result from well-defined mathematical transformations that convert chemical information into quantifiable metrics for modeling (12,13).

The development of thousands of descriptors, documented in detailed monographs and integrated into widely used software, has significantly enhanced molecular characterization (12–14). However, this expansion introduces challenges, such as descriptor redundancy, interpretability issues, and the risk of overfitting in predictive models. As a result, recent studies focus not only on the variety of descriptors but also on strategies for their careful selection, validation, and integration into advanced machine learning frameworks to ensure robust and reliable predictions (7–14).

The Role of Molecular Descriptors in QSAR/QSPR Modeling

Molecular descriptors are essential to QSAR and QSPR modeling, providing a quantitative, objective way to represent molecular structures. These descriptors enable researchers to analyze and model diverse chemical compounds without relying on specific experimental conditions, serving as the core input for a range of modeling techniques—from traditional linear regression and stochastic methods to advanced approaches like artificial neural networks, random forests, and XGBoost (15). By facilitating predictions of critical properties — such as biological activity, toxicity, ADME/T profiles, and physicochemical stability —descriptors not only support predictive modeling but also enhance mechanistic insights by identifying structural features, such as toxicophores or pharmacophores, that

drive molecular behavior. This dual role aids rational drug design and ecotoxicological risk assessment (16).

The evolution of molecular descriptors has closely paralleled advancements in computational chemistry. Early descriptors focused on basic structural counts, such as the numbers of atoms, bonds, and rings, followed by topological indices introduced by pioneers such as Wiener, Balaban, and Hosoya. Over time, descriptors grew more complex, incorporating two- and three-dimensional geometric properties, quantum-chemical metrics like HOMO–LUMO energies and dipole moments, and, more recently, four-dimensional descriptors that account for conformational dynamics and interactions with biological targets (17,18). This progression reflects a shift from simple structural metrics to descriptors capturing intricate aspects of molecular function.

Today, the field of molecular descriptors is expansive, with over 5,000 distinct descriptors documented across various research domains, including constitutional, information-theoretic, topological, geometric, electrostatic, quantum-chemical, pharmacophoric, and experimentally derived types. Software tools such as Dragon, PaDEL, RDKit, ChemDes, and alvaDesc have become indispensable, automating the calculation of thousands of molecular descriptors from molecular structures. While this diversity empowers predictive modeling, it also poses challenges related to redundancy, high dimensionality, and interpretability (19,20).

Despite the rise of deep learning and "black-box" modeling approaches, evidence underscores that the success of QSAR/QSPR models hinges on the thoughtful selection, engineering, and interpretation of descriptors. Well-chosen descriptors improve model robustness and predictive accuracy while providing meaningful insights into structure–function relationships. Thus, a deep understanding of the theoretical basis, classification, and practical applications of molecular descriptors remains fundamental to advancing QSAR and QSPR research (21-23).

Extended Categories of Molecular Descriptors

Beyond traditional categories like constitutional, topological, geometrical, and quantum-chemical descriptors, several specialized subcategories have enriched the QSAR/QSPR modeling landscape. These include descriptors rooted in chemical graph theory, which convert atom and bond connectivity into precise mathematical indices; information-theoretic descriptors that quantify structural complexity using concepts like entropy; experimentally derived descriptors from physicochemical measurements; and pharmacophore-based descriptors that emphasize 3D configurations critical for

molecular recognition. Additionally, fragment-based descriptors, which indicate the presence or absence of specific functional groups or substructures, provide clear insights into chemical reactivity and biological activity, enhancing interpretability (24,25).

The vast array of descriptors available in modern cheminformatics tools, such as Dragon, PaDEL, RDKit, ChemDes, and similar platforms, presents both opportunities and challenges. These tools can generate thousands of descriptors from a single molecular structure, enabling robust modeling but also introducing issues such as redundancy, collinearity, and overfitting. To address these, feature selection and dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), have become essential in QSAR/QSPR workflows. These methods help identify the most informative descriptors, eliminating noise and redundancy to improve model robustness and interpretability (26).

Another critical aspect is the assessment of chemical similarity in descriptor space. For continuous descriptors, similarity is typically measured using distance metrics like Euclidean or Mahalanobis, often weighted by feature importance. For binary descriptors, which capture the presence or absence of structural fragments, metrics such as the Tanimoto coefficient, Jaccard index, or Manhattan distance are commonly used. These similarity measures are foundational to organizing chemical libraries, navigating structure–activity relationships, and conducting virtual screening. The choice of metric significantly impacts perceived molecular similarity, influencing which compounds are prioritized for experimental evaluation (27-29).

Zero-Dimensional (0D) Descriptors

Zero-dimensional (0D) descriptors represent the simplest form of molecular characterization, relying solely on the molecular formula and counts of constituent atoms, without considering their structural connectivity. Examples include molecular weight, counts of specific elements (e.g., carbon, hydrogen, nitrogen, oxygen, or halogens), total atom count, elemental mass fractions, empirical formulas, and atomic composition indices (30).

Despite their lack of information on bonding patterns or stereochemistry, 0D descriptors provide a basic yet valuable snapshot of molecular size and elemental makeup. In applications such as early-stage drug discovery or chemical database screening, they serve as efficient filters to eliminate molecules outside the desired property ranges. For instance, molecular weight is a key parameter in medicinal chemistry, integral to Lipinski's "rule of five," and influences pharmacokinetic properties

such as absorption, distribution, and metabolic stability. Similarly, atom counts and elemental composition can correlate with properties such as polarity or lipophilicity, thereby impacting membrane diffusion and overall pharmacokinetic behavior (31).

Far from being outdated, 0D descriptors remain relevant in modern cheminformatics. They are often used as baseline variables in large-scale QSAR/QSPR datasets, complementing more complex descriptors. Their straightforward interpretability makes them particularly useful for rapid chemical similarity assessments or applying constraints in virtual screening. Thus, 0D descriptors demonstrate that even the most fundamental numerical representations of chemical composition can retain significant predictive utility when integrated into comprehensive modeling frameworks (32).

Fragment-Based (1D) Descriptors

Fragment-based, or one-dimensional (1D) descriptors, are molecular encodings that capture the presence or absence of specific functional groups, substructures, or chemical motifs. Often called expert-based descriptors, they are grounded in established chemical knowledge, offering a direct, interpretable link to molecular functionality. Unlike continuous descriptors focused on size or topology, these provide clear insights into the structural components driving molecular behavior.

Common examples include counts of functional groups (e.g., carbonyls, hydroxyls, amines), aromatic or non-aromatic rings, heteroatoms (e.g., N, O, S, P, halogens), and rotatable bonds. The number of rotatable bonds is particularly significant as a measure of molecular flexibility. Higher counts often correlate with reduced oral bioavailability, lower membrane permeability, and decreased metabolic stability, making this descriptor a key component of Lipinski's "rule of five," which limits rotatable bonds to ten or fewer in drug-like molecules (33).

Other fragment-based descriptors offer additional insights into molecular complexity. The count of rings reflects structural rigidity and planarity, influencing binding affinity and specificity to biological targets. Heteroatoms enable directional interactions, such as hydrogen bonding or acid–base complementarity, which are critical for defining pharmacophoric patterns and ensuring selective molecular recognition. Similarly, the number of heavy atoms (all non-hydrogen atoms) serves as a simple yet effective indicator of molecular mass and complexity. Molecules with higher heavy atom counts may exhibit greater binding specificity and diverse interactions, but can face challenges like reduced permeability or increased toxicity risks due to bioaccumulation (33).

Fragment-based descriptors strike a balance between computational simplicity and chemical interpretability, making them highly valuable in cheminformatics. Their ease of calculation and ability to connect structural motifs to pharmacokinetic and toxicological properties ensure their continued relevance in QSAR/QSPR modeling. By linking molecular features to functional outcomes, these descriptors bridge numerical modeling with chemically intuitive insights, enhancing both predictive accuracy and mechanistic understanding (34).

Two-Dimensional (2D) Descriptors

Two-dimensional (2D) molecular descriptors are a cornerstone of QSAR and QSPR modeling, capturing the topological structure of molecules based on atom connectivity, independent of their three-dimensional conformation (13,14). Rooted in graph theory, these descriptors represent molecules as graphs, G=(V,E), where vertices (V) denote atoms and edges (E) represent chemical bonds. This abstraction enables quantitative analysis of structural features like chain length, branching, cycles, aromatic systems, heteroatom placement, and multiple bond distributions, transforming chemical intuition into mathematical indices. A major strength of 2D descriptors is their utility when three-dimensional conformations are unavailable, making them ideal for high-throughput virtual screening and database mining. Their rapid computation, without the need for geometrical optimization, ensures their prominence in cheminformatics workflows (12).

The development of topological descriptors reflects a progressive formalization of chemical structure. The Wiener index (W), one of the earliest measures, calculates the sum of shortest path lengths between all atom pairs in a molecular graph. Linear molecules yield higher W values, while branching reduces them, making branching effective for predicting properties such as boiling points, melting points, lipophilicity (logP), toxicity, and biodistribution. Extensions like the Hyper-Wiener index (WW), which incorporates distances and their squares, and the Modified Wiener index (W*), which weights central bonds, enhance sensitivity to local structural variations and are particularly useful in pharmacophore modeling and reactivity studies (17,35).

The Randić index (χ), or connectivity index, quantifies branching by using reciprocal values of adjacent atom degrees, emphasizing terminal or less-connected atoms. Its extensions, incorporating atomic properties such as valence, electronegativity, and polarizability, enable predictions of pharmacokinetic and pharmacodynamic properties, including logP, membrane permeability, toxicity, and receptor selectivity. The Kier–Hall indices, higher-order extensions of the Randić concept, analyze

connectivity patterns across multiple levels, capturing shape, symmetry, and centrality, and are effective in distinguishing structural isomers and modeling ligand–receptor interactions (35).

The Balaban index (J) normalizes average topological distances by the number of bonds and cycles, enabling comparisons across molecules of varying sizes. Its robustness makes it valuable for predicting ADMET-related properties in diverse chemical datasets, with applications in medicinal chemistry and toxicology (36,37).

Other indices further enrich 2D descriptor applications. The Zagreb indices (M_1 and M_2) use atomic degrees to quantify electronic density distribution and molecular flexibility. The Augmented Zagreb Index (AZI) accounts for specific bond types and excels at modeling complex, multicentric structures. The Hosoya index (Z) counts independent edge pairs, showing sensitivity to cycles and branching, while its modified version (Z*) is tailored for carbon-rich frameworks with local symmetries. The Szeged index (Sz) refines topological distance by partitioning atoms around each bond, offering superior accuracy for systems with local asymmetry or uneven mass distribution compared to the Wiener index (38).

2D topological descriptors demonstrate the power of connectivity-based encodings to capture essential molecular architecture. Their computational efficiency, broad applicability, and proven predictive accuracy ensure their enduring role in QSAR/QSPR modeling, even as more complex three- and four-dimensional descriptors emerge. By providing a balance of simplicity and insight, these descriptors remain indispensable for understanding and predicting molecular behavior across diverse chemical and biological contexts.

SMILES-Based Descriptors in QSAR/QSPR Modeling

SMILES-based descriptors represent a unique and innovative approach in QSAR and QSPR modeling, leveraging the SMILES to transform linear string notations into predictive numerical features. Unlike traditional descriptors that rely on explicit geometric, topological, or physicochemical representations, SMILES-based descriptors directly utilize the text-based molecular graph encoded in SMILES, offering a seamless bridge between symbolic chemical representations and quantitative modeling (39).

The core of SMILES-based descriptors lies in decomposing the SMILES string into smaller fragments, such as sequences representing specific atoms, bond types, ring closures, or functional groups. Each fragment is assigned a numerical weight, which is iteratively optimized using Monte Carlo techniques

to maximize correlation with the target endpoint, such as biological activity or physicochemical property. The CORAL software has been instrumental in formalizing this methodology, employing stochastic optimization to adjust correlation weights and build robust predictive models by balancing training and validation datasets. This process ensures that the descriptors are both predictive and computationally efficient (40,41).

A key advantage of SMILES-based descriptors is their versatility and scalability. As SMILES is a standardized notation widely adopted in cheminformatics databases, these descriptors can be rapidly calculated for large chemical libraries without requiring detailed 2D or 3D structural data. This computational simplicity makes them particularly valuable for early-stage virtual screening, where speed and scalability are critical. Moreover, their interpretability enhances their utility individual SMILES fragments can be directly linked to structural motifs that drive or hinder the modeled property, facilitating mechanistic insights into structure–activity or structure–property relationships. SMILES-based descriptors are categorized into local and global types. Local descriptors focus on specific fragments within the SMILES string, such as "C=O" for carbonyl groups, "–NH–" for amines, or "c1ccccc1" for benzene rings, enabling precise identification of substructures that influence the target property. This granularity supports the recognition of pharmacophores or toxicophores, enhancing the interpretability of QSAR/QSPR models. Global descriptors, in contrast, consider the entire SMILES sequence, capturing the cumulative effect of all fragments and their interdependencies. This holistic approach contributes to model robustness and stability, particularly in Monte Carlo optimization and CORAL-based frameworks, where the combination of local and global descriptors balances predictive accuracy with mechanistic clarity (40,41).

An advanced development in this domain is the introduction of quasi-SMILES descriptors, which extend the standard SMILES notation by incorporating non-structural information, such as experimental conditions, solvents, or formulation parameters. By embedding contextual metadata, quasi-SMILES descriptors enable predictions of not only intrinsic molecular properties but also system-level behaviors influenced by environmental factors. This innovation broadens their applicability across fields such as medicinal chemistry, ecotoxicology, and nanomaterials research, where external conditions significantly influence outcomes (42).

SMILES-based descriptors have proven robust across diverse endpoints, including physicochemical properties like partition coefficients and solubility, toxicological outcomes such as mutagenicity and aquatic toxicity, and pharmacological activities. In many instances, models built with SMILES

descriptors match or surpass those using traditional 2D or 3D descriptors, particularly when paired with Monte Carlo optimization, which systematically explores descriptor space while minimizing overfitting. From a broader perspective, SMILES-based descriptors embody a key principle in cheminformatics: symbolic representations, when mathematically formalized, can yield powerful predictive models without the need for resource-intensive structural calculations. As cheminformatics increasingly integrates with big data, text mining, natural language processing, and machine learning, the role of SMILES and quasi-SMILES descriptors is poised to grow, offering a computationally efficient and interpretable approach to molecular modeling in an era of expanding data-driven research (40,41).

Physicochemical Descriptors in QSAR/QSPR Modeling

Physicochemical descriptors quantify essential physical and chemical properties of molecules, typically derived from two-dimensional (2D) structural data. These descriptors are pivotal in evaluating the pharmacokinetic profiles of drug candidates, as they encode attributes that dictate a compound's behavior in biological systems. Key examples include lipophilicity, aqueous solubility, permeability, molecular weight, and counts of hydrogen bond donors and acceptors (42).

Lipophilicity, a critical descriptor, measures a compound's affinity for lipid environments and influences pharmacokinetic processes such as passive membrane transport, intestinal absorption, plasma protein binding, tissue distribution, and intracellular bioavailability. Suboptimal lipophilicity can lead to poor absorption, reduced oral bioavailability, or limited penetration across barriers like the blood–brain barrier. Conversely, excessive lipophilicity may lead to poor aqueous solubility, non-specific binding, and increased metabolic clearance, posing a key challenge for achieving an optimal balance in drug design.

Studies, both in vitro and in silico, highlight the interplay between lipophilicity and other molecular properties, including size, polarity, hydrogen-bonding capacity, and ionization state. For instance, a high number of hydrogen bond donors increases polarity, potentially hindering absorption, while elevated molecular weight often correlates with reduced membrane permeability and impaired transport. Aqueous solubility, another vital descriptor, determines a compound's availability in free form for pharmacological action and elimination. Insufficient solubility can limit efficacy and complicate formulation development (43).

Physicochemical descriptors offer a dual benefit: they serve as robust predictors of drug efficacy and safety, guiding lead optimization toward compounds with favorable ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiles, and they can be calculated directly from molecular structures without experimental data. This computational efficiency makes them invaluable for large-scale virtual screening and early-stage predictions, enabling rapid evaluation of thousands of compounds before resource-intensive experimental testing.

By bridging fundamental chemical principles with practical drug design, physicochemical descriptors combine interpretability with predictive power for ADMET outcomes. Their continued prominence in QSAR and QSPR modeling underscores their essential role in modern cheminformatics (44).

Quantum-Chemical Descriptors in QSAR/QSPR Modeling

Quantum-chemical descriptors stand at the forefront of modern cheminformatics, offering a sophisticated means of capturing the electronic properties of molecules, which significantly enhances the interpretability and physiological relevance of QSAR and QSPR models. Unlike traditional two-dimensional or three-dimensional descriptors that focus on geometry or topology, these descriptors provide direct insight into electronic characteristics, enabling predictions of binding affinity, selectivity, and reactivity while offering a mechanistic understanding of molecular behavior in biological systems (45).

Historically, the application of quantum-chemical descriptors was constrained by high computational costs, limiting early studies to semi-empirical methods like AM1 and PM3. Advances in computational power and software have shifted the preference toward density functional theory (DFT), which strikes an optimal balance between accuracy and efficiency. Hybrid functionals, such as mPW1PW91 and B3LYP, have become particularly effective, delivering reliable estimates of electronic density, charge distribution, and total system energy without excessive computational demands (46).

Among the most significant quantum-chemical descriptors are the energies of frontier molecular orbitals, specifically the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). Rooted in frontier molecular orbital theory, these descriptors are critical for estimating ionization potential, electron affinity, and overall chemical reactivity. The HOMO energy indicates a molecule’s tendency to donate electrons, while the LUMO energy reflects its capacity to accept them. The HOMO–LUMO gap, the energy difference between these orbitals, serves as a key indicator of kinetic stability and reactivity. Derived properties, such as hardness, softness,

electronegativity, and electrophilicity, further extend these insights within the Hard and Soft Acids and Bases (HSAB) framework, providing a rational basis for predicting molecular selectivity and reaction pathways (47).

Another vital category includes partial atomic charges, calculated using methods such as Mulliken, Natural Population Analysis (NPA), or Merz–Kollman. These charges facilitate modeling of electrostatic interactions, prediction of hydrogen-bonding patterns, and assessment of electrostatic complementarity in protein–ligand interactions. Related properties, such as dipole moment and polarizability tensors, describe a molecule's response to external fields and contribute to understanding van der Waals forces, dispersion interactions, and solvation effects (48).

Quantum-chemical calculations also yield thermodynamic parameters, including enthalpy, free energy, and heat capacity, as well as specialized descriptors such as protonation energies and quadrupole moments. These parameters not only enhance predictive modeling but also provide mechanistic explanations for processes such as proton transfer, tautomerism, and noncovalent binding, enriching the understanding of molecular interactions. The true strength of quantum-chemical descriptors lies in their ability to integrate structural, electronic, and reactivity-related information into a unified framework. This dual role, serving as both predictive tools and conceptual bridges between molecular structure and mechanism of action, makes them indispensable in drug design and molecular modeling. By combining predictive accuracy with explanatory depth, quantum-chemical descriptors solidify their position as critical tools at the intersection of theoretical chemistry and practical applications in cheminformatics (49,50).

Three-Dimensional (3D) Molecular Descriptors in QSAR/QSPR Modeling

Three-dimensional (3D) molecular descriptors go beyond the limitations of two-dimensional topological descriptors by incorporating the spatial arrangement of atoms within a molecule, offering a richer perspective on a molecule's geometric and steric properties. These descriptors provide critical insights into biological activity by directly connecting a molecule's spatial structure to pharmacological behaviors such as receptor binding, membrane permeability, and drug–target interactions. Unlike their 2D counterparts, which rely solely on atomic connectivity, 3D descriptors capture intricate details like molecular volume, surface area, shape, orientation, and spatial interactions, making them essential for understanding complex molecular recognition processes that underpin biological responses (51).

The computation of 3D descriptors begins with geometric optimization of molecular structures, typically performed using quantum-chemical methods like PM6 or DFT, or molecular mechanics approaches such as MMFF94. These optimized atomic coordinates serve as the foundation for calculating a wide array of parameters, including molecular volume, solvent-accessible surface area, bond lengths, bond angles, torsional angles, and measures of molecular rigidity or flexibility. This ability to account for conformational variability is particularly valuable for molecules that can adopt multiple low-energy conformations, as the bioactive conformation often determines the pharmacological outcome (52).

Among the most prominent surface-related descriptors is the polar surface area, defined as the cumulative surface contributions of polar atoms, such as oxygen and nitrogen, along with their attached hydrogens. This descriptor is a powerful predictor of oral bioavailability and blood–brain barrier penetration, with well-established empirical thresholds that guide medicinal chemists in refining molecular scaffolds to optimize absorption and distribution properties (53).

Significant advancements in 3D descriptor methodologies have come through grid-based approaches, notably Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA). These methods project molecules onto a three-dimensional grid, where steric, electrostatic, hydrophobic, and hydrogen-bonding fields are calculated at each grid point. The resulting spatial maps, when correlated with biological activity, offer vivid visualizations that highlight molecular features driving enhanced potency or selectivity, making these techniques invaluable for ligand-based drug design (54,55).

Conceptually advanced 3D descriptor families have further enriched the field. Weighted Holistic Invariant Molecular (WHIM) descriptors, derived from principal component analysis of atomic properties like mass, electronegativity, and volume, are alignment-independent, making them particularly effective for large datasets with non-standardized conformations. Molecular Representation of Structures based on Electron diffraction (MoRSE) descriptors simulate electron diffraction patterns using distance-weighted atomic property functions, demonstrating predictive power for properties such as lipophilicity, infrared spectra, toxicity, and binding affinity. GETAWAY descriptors, which integrate geometry, topology, and atomic weighting information, effectively bridge 2D and 3D representations by capturing the combined influence of molecular shape and electronic distribution on behavior. Additionally, 3D autocorrelation descriptors, derived from atomic

distributions across molecular surfaces, provide rotational and translational invariance, ensuring robustness for highly flexible molecules with complex conformational dynamics (56,57).

Despite their higher computational demands than 2D descriptors, 3D descriptors are indispensable in modern QSAR and QSPR modeling due to their ability to capture nuanced details of molecular recognition. They are particularly critical in scenarios where biological activity hinges on spatial complementarity with macromolecular targets, thereby supporting rational drug design strategies aimed at enhancing selectivity, efficacy, and safety profiles. By linking fine-grained structural details to functional outcomes, 3D descriptors form a cornerstone of cheminformatics, driving progress in predictive modeling and drug discovery.

Pharmacophore-Based Descriptors in QSAR/QSPR Modeling

Pharmacophore-based descriptors hold a distinctive role in QSAR and QSPR modeling by focusing on the structural and physicochemical features essential for molecular recognition by biological targets. Unlike geometric or quantum-mechanical descriptors that capture broad molecular properties, pharmacophore descriptors emphasize the spatial arrangement of functional groups critical for intermolecular interactions, such as hydrogen bonding, electrostatic complementarity, hydrophobic contacts, and aromatic stacking. This targeted approach makes them uniquely suited for linking molecular structure to biological activity (58).

The concept of a pharmacophore, first articulated by Paul Ehrlich, refers to the minimal structural features required for a molecule to elicit a biological response. In contemporary cheminformatics, this idea is translated into quantitative descriptors that encode the three-dimensional distribution of key molecular features, including hydrogen bond donors and acceptors, charged groups, hydrophobic regions, aromatic rings, and steric volumes. These features are represented in formats such as vectors, grids, or distance matrices, which capture both the identity of functional groups and their spatial relationships, providing a robust framework for modeling molecular interactions (59).

A primary advantage of pharmacophore descriptors is their interpretability. Unlike many descriptors that yield abstract numerical values, pharmacophore features can be directly mapped to specific molecular structures and correlated with binding sites on biological targets, such as proteins or nucleic acids. For instance, a pharmacophoric motif defined by two hydrogen bond donors at a specific distance and orientation relative to an aromatic ring may be critical for enzyme inhibition. This direct linkage between descriptor and molecular function positions pharmacophore descriptors as a vital

bridge between ligand-based modeling and structure-based drug design, enabling researchers to translate computational predictions into actionable chemical insights (60).

Several computational tools have streamlined the application of pharmacophore descriptors. Platforms like VolSurf+, LigandScout, and MOE generate pharmacophore fingerprints or three-dimensional pharmacophore models from molecular datasets, which can be correlated with biological activity or used in virtual screening. Additionally, alignment-free pharmacophore fingerprints have been developed to efficiently handle large chemical libraries, facilitating rapid similarity searches and clustering analyses, which are critical for high-throughput drug discovery (61).

Pharmacophore descriptors are particularly valuable for flexible ligands, where multiple conformations may fulfill a pharmacophoric hypothesis. When integrated with conformational sampling or molecular dynamics simulations, these descriptors evolve into dynamic pharmacophores, or dynophores, which account for the temporal stability and adaptability of pharmacophoric features in dynamic environments. This approach significantly enhances predictive reliability in systems where conformational plasticity is a key determinant of activity, such as in peptide ligands, macrocycles, and natural products (62,63).

In essence, pharmacophore-based descriptors combine biological relevance, interpretability, and predictive power. Their ability to pinpoint the critical interaction motifs driving molecular recognition makes them indispensable in modern cheminformatics, particularly in drug discovery. They guide scaffold optimization, lead expansion, and virtual screening of extensive compound libraries, offering a powerful tool for advancing the design of effective and selective therapeutic agents.

Four-Dimensional (4D) Molecular Descriptors in QSAR/QSPR Modeling

While three-dimensional (3D) descriptors mark a significant improvement over two-dimensional and topological approaches by incorporating spatial molecular features, they are limited by their reliance on a single, static conformation, typically optimized in vacuum or with implicit solvation models. In biological systems, molecules are not rigid but exist as dynamic ensembles of conformers that fluctuate among various low-energy states. This conformational flexibility means that a single geometry often fails to capture a compound's full structural and functional behavior, particularly for molecules with significant flexibility (64).

Four-dimensional (4D) descriptors address this limitation by introducing conformational dynamics as the fourth dimension, capturing the time-dependent behavior of molecules across their

conformational landscape. Unlike 3D descriptors, which depend on a single optimized structure, 4D descriptors incorporate ensembles of conformations generated by computational methods such as Monte Carlo sampling, molecular dynamics simulations, or docking. These descriptors use averaged values, weighted factors, or distribution maps to reflect the probabilities and contributions of individual conformers, thereby enhancing the realism and predictive accuracy of QSAR and QSPR models. This approach is particularly valuable for flexible molecules, such as peptides, macrocycles, and compounds with multiple rotatable bonds, where conformational variability significantly influences biological activity (65).

Several computational frameworks have been developed to leverage 4D descriptors. The Molecular Comparative Fields (MCF) approach, for instance, generates statistical maps from multiple conformers to quantify interaction potentials across conformational ensembles. In molecular docking, tools like AutoDock4, combined with interfaces such as RACCOON, enable flexible modeling of both ligands and biomolecular targets, producing 4D docking profiles that account for binding variability. Similarly, VolSurf+ employs 4D pharmacophore fields to compute spatiotemporal distributions of molecular interactions, linking ligand flexibility to pharmacophoric hotspots critical for target recognition (66).

The strength of 4D descriptors lies in their ability to model the dynamic nature of ligands in environments where conformational adaptability governs binding affinity, selectivity, and overall biological activity. Compared to traditional 3D QSAR models, 4D ensemble-based approaches deliver more stable and physiologically relevant predictions, particularly in pharmacodynamic studies where the interplay between molecular flexibility and target recognition is paramount (67). As computational power and molecular simulation techniques continue to advance, 4D descriptors are poised to play an increasingly central role in cheminformatics, bridging the gap between static structural models and the dynamic complexity of biological systems.

Empirical and Experimental Descriptors in QSAR/QSPR Modeling

Empirical and experimental descriptors remain a vital component of QSAR and QSPR modeling, complementing theoretical descriptors derived from quantum-chemical or molecular modeling approaches. These descriptors, grounded in experimental measurements or established chemical observations, capture the nuanced effects of substituents, particularly in aromatic systems, by quantifying electronic, steric, and hydrophobic properties. Their ability to describe molecular

characteristics that are challenging to model through geometry or electronic structure alone makes them indispensable for predictive and mechanistic insights in cheminformatics (68).

Prominent among these descriptors are Hammett constants, which measure the electronic effects of substituents at the *meta*- and *para*- positions of aromatic rings, providing insight into how electronic modifications influence molecular behavior. These are often paired with Swain–Lupton parameters, which distinguish between inductive and resonance contributions, and Taft parameters, which extend the framework to account for steric effects in aliphatic systems. Another key descriptor, logP, quantifies hydrophobicity and is critical for evaluating lipophilicity and pharmacokinetic properties, such as membrane permeability and bioavailability. Verloop parameters further enrich this category by describing the steric dimensions of substituents; length, width, thickness, and angular eccentricity; offering a detailed perspective on spatial effects. Additional descriptors, such as molecular mass, molar volume, refractivity, and polarizability, provide complementary insights into molecular size and interaction potential (69).

Rather than being used in isolation, empirical descriptors are typically integrated with theoretical ones to build more robust and versatile QSAR/QSPR models. Principal component analysis (PCA) reveals that these descriptors span multiple dimensions of molecular space; the first dimension is often dominated by electronic effects, driven by Hammett constants, resonance contributions, and chemical shifts; the second reflects steric influences, captured by Verloop parameters, polarizability, and volume; and the third encompasses lipophilic and surface-related properties, including logP, solvent-accessible surface area, and the HOMO–LUMO gap. Notably, descriptors like Taft parameters and logP often contribute to multiple dimensions, underscoring their complex, multifaceted roles in molecular characterization (70).

Despite challenges in obtaining consistent, high-quality experimental data, empirical and experimental descriptors remain highly relevant in modern cheminformatics. Their integration with theoretical descriptors enhances both the predictive accuracy and interpretability of QSAR/QSPR models, serving as a crucial link between experimental evidence and computational abstraction. By grounding models in measurable chemical realities and leveraging advanced computational methods, these descriptors ensure that cheminformatics remains both scientifically robust and practically applicable (71,72).

SHAP-Based Descriptors in QSAR/QSPR Modeling

The integration of advanced machine learning into QSAR and QSPR modeling has significantly enhanced predictive accuracy, but it has also underscored a critical challenge: the lack of interpretability in high-performing "black-box" models. These models often deliver accurate predictions without revealing the molecular features driving those outcomes. SHAP (SHapley Additive exPlanations) values, rooted in cooperative game theory, address this limitation by providing a rigorous and practical framework for attributing the contribution of each descriptor; whether an atom-based feature, molecular fragment, topological index, or physicochemical parameter; to a model's predictions, offering both local and global interpretability (73).

In QSAR/QSPR studies, SHAP assigns a numerical value to each descriptor, quantifying its impact on the prediction for a specific molecule. Unlike traditional regression coefficients or static feature importance rankings, SHAP provides molecule-specific explanations, enabling researchers not only to predict activity but also to pinpoint which structural elements enhance or diminish it. Locally, SHAP highlights how individual molecular features influence a compound's outcome, identifying pharmacophoric or toxicophoric motifs and reactive regions. Globally, by aggregating contributions across datasets, SHAP uncovers broader trends, revealing how structural motifs correlate with biological or physicochemical properties. Visualizations such as heatmaps or feature attribution plots make these insights intuitive, guiding rational molecular modifications by identifying regions to optimize or eliminate in drug design (74). SHAP-based descriptors transcend conventional feature importance metrics by enabling QSAR models to serve as both predictive and explanatory tools. This dual role bridges statistical learning with chemical reasoning, transforming models into tools for ranking compounds and elucidating the chemical logic behind their activity. Such interpretability is invaluable in personalized molecular design, where transparent decision-making supports targeted exploration of chemical space with mechanistic clarity (75).

By embedding interpretability into machine-learning–based QSAR/QSPR frameworks, SHAP-based descriptors represent a technical and philosophical advancement. They redefine QSAR not only as a predictive tool but also as an explanatory paradigm, accelerating drug discovery, toxicological assessment, and rational design across the molecular sciences by providing clarity and actionable insights into the complex interplay between molecular features and their functional outcomes (76,77).

Future Directions in Molecular Descriptor Development

The evolution of molecular descriptors has been instrumental in advancing QSAR and QSPR modeling, yet significant challenges persist, shaping the path for future innovation. Traditional descriptors (topological, physicochemical, or quantum-chemical) have proven their predictive strength, but they often lack generalizability across diverse chemical spaces, suffer from redundancy, and struggle to adapt to emerging molecular classes. The future of descriptor development lies in addressing these limitations through innovative, integrative, and dynamic approaches that combine theoretical rigor with data-driven flexibility, ultimately enhancing both predictive accuracy and mechanistic insight (78).

A key direction is the development of hybrid descriptors that seamlessly integrate structural, electronic, dynamic, and experimental information into cohesive representations. By synthesizing data from multiple molecular domains, these descriptors can capture a more comprehensive picture of molecular behavior while minimizing redundancy through advanced dimensionality-reduction techniques, such as principal component analysis or autoencoders. This holistic approach promises to improve model robustness and applicability, enabling QSAR/QSPR models to tackle complex chemical systems with greater precision (79).

The rise of data-driven descriptors, derived directly from molecular graphs, SMILES notation, or three-dimensional conformations, using deep learning architectures such as graph neural networks (GNNs), marks another transformative trend. These representations have demonstrated exceptional predictive performance, particularly in large-scale virtual screening. However, their often-opaque nature poses interpretability challenges. A promising avenue is integrating these data-driven descriptors with classical ones, combining the predictive power of GNN embeddings with the mechanistic clarity of traditional descriptors. By incorporating explainable AI techniques, such as SHAP or attention mechanisms, this synergy can preserve the chemical intuition critical for rational molecular design while leveraging the scalability of modern machine learning (80).

Another critical frontier is the development of descriptors that account for the dynamic and context-dependent nature of molecules in biological systems. Unlike static 3D models, which rely on single, optimized conformations, real-world molecules exist as dynamic ensembles influenced by solvents, membranes, and macromolecular interactions. Advancing to four-dimensional (4D) and even five-dimensional (5D) descriptors; capturing conformational ensembles, temporal evolution, and multi-scale interactions; offers a path toward more physiologically relevant modeling. Coupling these descriptors with molecular dynamics simulations, enhanced sampling techniques, or quantum-

mechanical/molecular-mechanical (QM/MM) hybrid methods will enable a deeper understanding of how molecular flexibility and environmental factors shape biological activity, paving the way for more accurate predictions in complex systems (81).

From an application perspective, future descriptors must prioritize transferability and adaptability to meet the demands of modern drug discovery, materials science, and toxicology. As these fields increasingly rely on high-throughput screening of diverse chemical libraries, descriptors need to remain robust across varied molecular classes, including peptides, macrocycles, and nanomaterials. This requires the development of standardized protocols for descriptor calculation and the establishment of interoperable, open-access databases to ensure reproducibility and cross-study comparability. Additionally, descriptors tailored to emerging domains, such as biologics or nanostructured materials, will be essential to address the unique challenges posed by these systems (82).

Ultimately, the future of molecular descriptors lies in their transformation from static numerical encodings to dynamic, context-aware representations that reflect the complexity of molecular systems. By integrating classical cheminformatics with cutting-edge AI, advancing dynamic and multi-scale modeling, and prioritizing transferability, the next generation of descriptors will not only enhance predictive performance but also deepen our understanding of the intricate interplay between molecular structure and function. This evolution will empower cheminformatics to drive breakthroughs in drug discovery, materials design, and toxicological assessment, bridging the gap between computational prediction and real-world molecular behavior.

Conclusion

Molecular descriptors remain the foundation of QSAR and QSPR modeling, uniting a diverse array of approaches, from zero-dimensional (0D) to four-dimensional (4D), pharmacophore-based, quantum-chemical, physicochemical, SMILES/quasi-SMILES, and empirical descriptors. These descriptors collectively encode molecular composition, connectivity, three-dimensional geometry, conformational dynamics, electronic properties, and experimental characteristics, enabling robust predictive and mechanistic modeling across drug discovery, toxicology, materials science, and beyond. Established methodologies, including topological indices, alignment-independent 3D descriptors, field-based techniques like CoMFA and CoMSIA, quantum-chemical parameters, and dynamic pharmacophore models, provide complementary perspectives that facilitate high-throughput virtual screening, precise molecular optimization, and mechanistic insights into molecular interactions.

Looking forward, the evolution of QSAR/QSPR hinges on addressing persistent challenges, such as descriptor redundancy, limited generalizability, and the need for physiological relevance. Hybrid descriptors that seamlessly integrate structural, electronic, dynamic, and experimental data, refined using advanced dimensionality-reduction techniques such as principal component analysis or autoencoders, promise to enhance model robustness and applicability across diverse chemical spaces. Data-driven descriptors, derived from molecular graphs or SMILES notations via graph neural networks, offer unparalleled predictive power but require integration with classical descriptors to ensure chemical interpretability. By leveraging explainable AI techniques, such as SHAP and attention mechanisms, these hybrid approaches can balance scalability with mechanistic clarity, guiding rational molecular design. Furthermore, the shift toward four-dimensional (4D) and five-dimensional (5D) descriptors; capturing conformational ensembles, temporal dynamics, and multi-scale interactions via molecular dynamics, enhanced sampling, or quantum-mechanical/molecular-mechanical methods; will enable more physiologically relevant predictions, particularly for flexible molecules like peptides, macrocycles, and biologics in complex biological environments.

To meet the demands of modern applications, future descriptors must prioritize transferability, adaptability, and standardization. As drug discovery, toxicology, and materials science increasingly rely on screening diverse chemical libraries, descriptors must remain robust across molecular classes, including nanomaterials and biologics. Standardized protocols for descriptor calculation, coupled with interoperable, open-access databases, will ensure reproducibility and cross-study comparability, fostering collaborative innovation. By harmonizing theoretical rigor with data-driven flexibility, these advancements will redefine QSAR/QSPR as a translational framework, bridging predictive accuracy with chemically informed, mechanistically transparent design. This evolution will accelerate the development of novel therapeutics, safer materials, and sustainable chemical solutions, solidifying molecular descriptors as indispensable tools in the molecular sciences.

References

1. Hansch C, Fujita T. ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure. J Am Chem Soc. 1964;86(8):1616-26.
2. Hansch C, Leo A, Hoekman D. Exploring QSAR: Fundamentals and applications in chemistry and biology. Washington, DC: American Chemical Society; 1995.
3. Achary PGR. Applications of Quantitative Structure-Activity Relationships (QSAR) based Virtual Screening in Drug Design: A Review. Mini Rev Med Chem. 2020;20(14):1375-88.
4. Kubinyi H. QSAR and 3D QSAR in drug design. Part 1: Methodology. Drug Discov Today. 1997;2(11):457-67.
5. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: Where have you been? Where are you going to? J Med Chem. 2014;57(12):4977-5010.
6. Li J, Zhao T, Yang Q, Du S, Xu L. A review of quantitative structure-activity relationship: The development and current status of data sets, molecular descriptors and mathematical models. Chemom Intell Lab Syst. 2025;256:105278.
7. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Des. 2007;13(34):3494-504.
8. Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O. Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct Prop Relat. 2016;1(1):45–63
9. Polishchuk P. Interpretation of quantitative structure–activity relationship models: Past, present, and future. J Chem Inf Model. 2017;57(11):2618-39.
10. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018 Aug;23(8):1538–1546.
11. Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, et al. QSAR without borders. Chem Soc Rev. 2020;49(11):3525-64. DOI: 10.1039/D0CS00098A
12. Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim: Wiley-VCH; 2000.
13. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. 2nd ed. Weinheim: Wiley-VCH; 2009.
14. Karelson M. Molecular descriptors in QSAR/QSPR. New York: Wiley-Interscience; 2000.
15. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: A classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947-58.

16. Cronin MTD, Schultz TW. Pitfalls in QSAR. J Mol Struct THEOCHEM. 2003;622(1-2):39-51
17. Wiener H. Structural determination of paraffin boiling points. J Am Chem Soc. 1947;69(1):17-20.
18. Hosoya H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Soc Jpn. 1971;44(9):2332-9.
19. Mauri A, Consonni V, Pavan M, Todeschini R. Dragon software: An easy approach to molecular descriptor calculations. Match Commun Math Comput Chem. 2006;56(2):237-48.
20. Yap CW. PaDEL-Descriptor: An open-source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-74.
21. Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res. 2009;20(3-4):241-66.
22. Gramatica P. Principles of QSAR models validation: Internal and external. QSAR Comb Sci. 2007;26(5):694-701.
23. Lo YC, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. Drug Discov Today. 2018;23(8):1538-46.
24. Danishuddin, Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discov Today. 2016 Aug;21(8):1291–1302.
25. Leach AR, Gillet VJ. An introduction to chemoinformatics. Dordrecht: Springer; 2007.
26. Dong J, Cao DS, Miao HY, Liu S, Deng BC, Yun YH, et al. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform. 2015;7:60. DOI: 10.1186/s13321-015-0109-z
27. Nikolova N, Jaworska J. Approaches to measure chemical similarity – a review. QSAR Comb Sci. 2004;22(9-10):1006-26. DOI: 10.1002/qsar.200330831
28. Willett P, Barnard JM, Downs GM. Chemical similarity searching. J Chem Inf Comput Sci. 1998;38(6):983-96.
29. Maggiora GM, Shanmugasundaram V. Molecular similarity measures. Methods Mol Biol. 2011;672:39-100.
30. Todeschini R, Consonni V. Molecular descriptors for chemoinformatics. Volume I: Alphabetical listing. Weinheim: Wiley-VCH; 2009.
31. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46(1-3):3-25.

32. Oprea TI, Gottfries J. Chemography: The art of navigating in chemical space. J Comb Chem. 2001;3(2):157-66.
33. Hou TJ, Xia K, Zhang W, Xu XJ. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approaches. J Chem Inf Comput Sci. 2004;44(1):266-75.
34. Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Comb Chem High Throughput Screen. 2000;3(5):363-72.
35. Randić M, Basak SC. On use of the variable connectivity index 1chi(f) in QSAR: toxicity of aliphatic ethers. J Chem Inf Comput Sci. 2001;41(3):614-8.
36. Kier LB, Hall LH. Molecular connectivity in chemistry and drug research. New York: Academic Press; 1976.
37. Balaban AT. Chemical graphs: Looking back and ahead. J Chem Inf Comput Sci. 1995;35(3):339-50.
38. Gutman I, Das KC. The first Zagreb index 30 years after. Match Commun Math Comput Chem. 2004;50:83-92.
39. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–36.
40. Zivkovic M, Zlatanovic M, Zlatanovic N, Golubović M, Veselinović AM. The Application of the Combination of Monte Carlo Optimization Method based QSAR Modeling and Molecular Docking in Drug Design and Development. Mini Rev Med Chem. 2020;20(14):1389-1402.
41. Veselinović AM, Veselinović JB, Živković JV, Nikolić GM. Application of SMILES notation based optimal descriptors in drug discovery and design. Curr Top Med Chem. 2015;15(18):1768-1779.
42. Yang S, Kar S, Leszczynski J. SMILES and Quasi-SMILES in QSAR Modeling for Prediction of Physicochemical and Biochemical Properties. In: Toropova AP, Toropov AA, editors. QSPR/QSAR Analysis Using SMILES and Quasi-SMILES, Challenges and Advances in Computational Chemistry and Physics, vol. 33. Cham: Springer Science and Business Media B.V.; 2023. p. 327–348.
43. van de Waterbeemd H, Gifford E. ADMET in silico modelling: Towards prediction paradise? Nat Rev Drug Discov. 2003;2(3):192-204.
44. Waring MJ. Lipophilicity in drug discovery. Expert Opin Drug Discov. 2010;5(3):235-48.
45. Hou TJ, Xu XJ. Recent development and application of virtual screening in drug discovery: An overview. Curr Pharm Des. 2004;10(9):1011-33.

46. Karelson M, Lobanov VS, Katritzky AR. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev. 1996;96(3):1027-44.
47. Cramer CJ, Truhlar DG. Density functional theory for transition metals and transition metal chemistry. Phys Chem Chem Phys. 2009;11(46):10757-816.
48. Parr RG, Yang W. Density-functional theory of the electronic structure of molecules. J Phys Chem. 1996;100(31):12974-12980.
49. Besler BH, Merz KM Jr, Kollman PA. Atomic charges derived from semiempirical methods. J Comput Chem. 1990;11(4):431-9.
50. Wang L, Ding J, Pan L, Cao D, Jiang H, Ding X. Quantum chemical descriptors in quantitative structure–activity relationship models and their applications. Chemom Intell Lab Syst. 2021;217:104384.
51. Leach AR. Molecular modelling: Principles and applications. 2nd ed. Harlow: Prentice Hall; 2001.
52. Stewart JJP. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. J Mol Model. 2007;13(12):1173-213.
53. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem. 2000;43(20):3714-7.
54. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988;110(18):5959-67.
55. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem. 1994;37(24):4130-46.
56. Todeschini R, Gramatica P, Provenzani R, Marengo E. Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons. Chemom Intell Lab Syst. 1995;27(2):221-9.
57. Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci. 2002;42(3):682-92.
58. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl Chem. 1998;70(5):1129-43.
59. Yang SY. Pharmacophore modeling and applications in drug discovery: Challenges and recent advances. Drug Discov Today. 2010;15(11-12):444-50.

60. Leach AR, Gillet VJ, Lewis RA, Taylor R. Three-dimensional pharmacophore methods in drug discovery. J Med Chem. 2010;53(2):539-58.
61. Wolber G, Langer T. LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model. 2005;45(1):160-9.
62. Wieder M, Garon A, Perricone U, Boresch S, Seidel T, Almerico AM, Langer T. Common Hits Approach: Combining Pharmacophore Modeling and Molecular Dynamics Simulations. J Chem Inf Model. 2017;57(2):365-385.
63. Mortier J, Dhakal P, Volkamer A. Truly Target-Focused Pharmacophore Modeling: A Novel Tool for Mapping Intermolecular Surfaces. Molecules. 2018 Aug 6;23(8):1959.
64. Hopfinger AJ, Wang S, Tokarski JS, et al. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. J Am Chem Soc. 1997;119(43):10509-24.
65. Vedani A, Dobler M. 5D-QSAR: The key for simulating induced fit. J Med Chem. 2002;45(11):2139-49.
66. Cruciani G, Pastor M, Guba W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. Eur J Pharm Sci. 2000;11(Suppl 2):S29-39.
67. Albuquerque MG, Hopfinger AJ, Barreiro EJ, de Alencastro RB. Four-dimensional quantitative structure-activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A2 receptor antagonists. J Chem Inf Comput Sci. 1998;38(5):925-38.
68. Hansch C, Leo A, Taft RW. A survey of Hammett substituent constants and resonance and field parameters. Chem Rev. 1991;91(2):165-95.
69. Verloop A, Hoogenstraaten W, Tipker J. Development and application of new steric substituent parameters in drug design. In: Ariëns EJ, editor. Drug design. Vol. 7. New York: Academic Press; 1976. p. 165-207.
70. Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models in QSAR. J Med Chem. 2005;48(3):661-3.
71. Furuhama A, Hasunuma K, Aoki Y. Interspecies quantitative structure-activity relationships (QSARs) for eco-toxicity screening of chemicals: the role of physicochemical properties. SAR QSAR Environ Res. 2015;26(10):809-30.
72. Soares TA, Nunes-Alves A, Mazzolari A, Ruggiu F, Wei G-W, Merz KM Jr. The (Re)-Evolution of Quantitative Structure-Activity Relationship (QSAR) Studies Propelled by the Surge of Machine Learning Methods. J Chem Inf Model. 2022 Nov 28;62(22):5317–5320.
73. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. Artif Intell Rev. 2022;55(3):1947–99.

74. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573-84.
75. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647–665.
76. Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. J Chem Inf Model. 2020;60(11):4950–4963
77. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(7):3370-88. DOI: 10.1021/acs.jcim.9b00237
78. Niazi SK, Mariam Z. Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review. Int J Mol Sci. 2023 Jul 15;24(14):11488.
79. Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des. 2013;27(8):675-79.
80. Wieder O, Kohlbacher O, Kuenemann M, Garon A, Ducrot P, Seidel T, et al. A compact review of molecular property prediction with graph neural networks. Drug Discov Today Technol. 2020;37:1-12.
81. Lin X, Li X, Lin X. A Review on Applications of Computational Methods in Drug Screening and Design. Molecules. 2020 Mar 18;25(6):1375.
82. Tetko IV, Gasteiger J, Todeschini R, et al. Virtual computational chemistry laboratory – design and description. J Comput Aided Mol Des. 2005;19(6):453-63.