Genome Publications
https://doi.org/10.61096/978-81-990998-7-6_3

Chapter 3

Physicochemical Descriptors and Multidimensional QSAR: 1D–6D Representations

Dr. S.P.R. Poonkodi
Associate Professor / HOD, Department of Chemistry, Government Arts College for Women, Sivagangai, Tamil Nadu, India.

Dr. Leslie V
Professor and Head, Department of Pharmacognosy, St. John's College of Pharmaceutical Sciences & Research, Kattappana, Idukki, Kerala, India.

Dr. Kannan Raman
Professor and Head, Department of Pharmacology, St. John's College of Pharmaceutical Sciences & Research, Kattapana, Idukki, Kerala, India.

Abstract: Quantitative Structure Activity Relationship (QSAR) modeling stands at the core of rational drug design, providing an essential framework for linking molecular structure to biological activity through quantifiable physicochemical descriptors. These descriptors translate chemical intuition into mathematical form, capturing steric, electronic, topological, and thermodynamic properties that define molecular behavior. As computational chemistry evolved, descriptor-based modeling expanded beyond one-dimensional (1D) linear relationships into higher-dimensional (2D–6D) representations that account for molecular geometry, conformational dynamics, and receptor flexibility. This chapter explores the theoretical and computational principles governing descriptor generation, categorization, and selection, as well as their integration into multidimensional QSAR models. It reviews the methodological progression from classical 1D-QSAR models, such as Hansch analysis and Free–Wilson approaches, to advanced 3D and 4D frameworks like CoMFA, CoMSIA, and grid-based conformational averaging. Furthermore, it examines how emerging 5D and 6D QSAR paradigms incorporate receptor adaptation and environmental fluctuations to enhance predictive accuracy. The chapter concludes by discussing descriptor computation tools (Dragon, PaDEL, MOE, RDKit, KNIME), validation strategies, and the convergence of descriptor science with machine learning and quantum mechanics, emphasizing reproducibility and interpretability in predictive modeling.

Keywords: Physicochemical descriptors, QSAR, molecular representations, multidimensional modeling, computational drug design.

Citation: S.P.R. Poonkodi, Leslie V, Kannan Raman. Physicochemical Descriptors and Multidimensional QSAR: 1D–6D Representations. *Comprehensive Approaches in Computer-Aided Drug Design: QSAR, Docking, Screening, Homology, Pharmacophore and AI-Driven Insights.* Genome Publication. 2025; Pp23-38. https://doi.org/10.61096/978-81-990998-7-6_3

1.0 INTRODUCTION

Physicochemical Descriptors in CADD

Physicochemical descriptors form the quantitative foundation of computer-aided drug design (CADD), representing a bridge between chemical structure and biological response. In essence, descriptors are numerical values derived from molecular structures that encapsulate information about properties such as size, shape, charge distribution, polarity, and lipophilicity. By correlating these computed values with experimentally observed biological activities, QSAR models provide an empirical yet mechanistically interpretable framework for predicting the activity of untested compounds. Historically, the field traces its origin to the pioneering works of Hansch and Fujita in the 1960s, who demonstrated that biological activity could be expressed as a mathematical function of physicochemical properties like partition coefficient (logP) and electronic constants (σ) [1]. These early models, though limited to linear relationships, laid the conceptual foundation for modern descriptor-driven computational modeling. In the contemporary era of CADD, descriptors are not merely computational conveniences they are molecular abstractions that define the informational granularity of a QSAR model. Whether derived from simple atomic counts or complex quantum mechanical calculations, descriptors encode how molecular features govern pharmacokinetic and pharmacodynamic outcomes. The evolution from 1D to 6D QSAR mirrors the growing recognition that drug activity depends on not just static molecular properties but dynamic interactions with biological macromolecules. This multidimensional expansion has been driven by advances in molecular modeling, cheminformatics, and data science, enabling the systematic exploration of vast chemical spaces while maintaining biological relevance.

The interplay between descriptors and biological activity has become increasingly sophisticated as molecular databases (such as ChEMBL, PubChem, and ZINC) provide millions of annotated compounds. These data-rich environments necessitate robust descriptor calculation pipelines and feature selection algorithms capable of identifying the most relevant molecular attributes. Consequently, physicochemical descriptors serve not only as the input to predictive models but also as interpretable markers of molecular mechanism, aiding medicinal chemists in hypothesis-driven design and lead optimization [2]. As discussed in this chapter, understanding how these descriptors are generated, classified, and applied in multidimensional contexts is essential for achieving predictive, reproducible, and mechanistically interpretable QSAR models.

1.1 Molecular Representation: From Chemical Structure to Numerical Features

A molecule's biological activity arises from its atomic composition, three-dimensional conformation, and electronic environment. Translating these complex features into numerical form is a central task of chemoinformatics. Molecular representations serve as the interface between structural data and computational modeling. They begin with the simplest one-dimensional (1D) notations, such as SMILES (Simplified Molecular Input Line Entry System) and InChI identifiers, and extend to multidimensional arrays capturing atomic coordinates, charge densities, or energy maps. These representations are the starting point for descriptor computation. where S*S* represents the molecular structure, and f*f* is the transformation function encoding structural features into numerical values [3]. Depending on the descriptor's nature, f*f* may compute atom-based properties (e.g., number of hydrogen bond donors), molecular graph invariants (e.g., Wiener index, Balaban index), or quantum chemical parameters (e.g., HOMO–LUMO gap, dipole moment). These descriptors collectively create a "molecular fingerprint" that uniquely characterizes a compound in a multidimensional feature space.

In modern CADD workflows, descriptor generation typically begins with structure preprocessing using software such as RDKit, MOE, or Open Babel, which standardize valence states, remove counterions, and optimize geometry. Subsequent descriptor computation tools like Dragon or PaDEL generate thousands of features that encompass physicochemical, topological, geometrical, and electronic dimensions. While these features increase model richness, they also introduce redundancy and collinearity, requiring feature selection techniques such as principal component analysis (PCA), recursive feature elimination (RFE), or genetic algorithms (GA) [4]. The transition from molecular structure to mathematical representation thus transforms chemistry into data science, enabling machine learning algorithms to capture the quantitative essence of drug–receptor interactions.

1.2 Classification of Molecular Descriptors: Constitutional, Topological, Geometrical, Electronic, and Thermodynamic

The classification of molecular descriptors provides a systematic framework for understanding how structural information translates into quantifiable attributes relevant to drug activity. Although thousands of descriptors exist, they can be broadly categorized into five principal groups: constitutional, topological, geometrical, electronic, and thermodynamic [5]. Each class captures a different aspect of molecular behavior and contributes distinctively to QSAR model interpretability. Constitutional descriptors represent the simplest form, quantifying fundamental molecular features such as atom counts, molecular weight, or number of hydrogen bond donors/acceptors. They provide baseline information about chemical composition and are commonly used in early-stage filtering, such as in Lipinski's rule of five or Veber's rules [6]. Topological descriptors abstract molecular structure as a graph, where atoms are vertices and bonds are edges. Indices like the Wiener number, Kier–Hall connectivity indices, and Balaban's J index capture molecular branching and connectivity patterns, which often correlate with molecular transport and binding properties [7]. Geometrical descriptors incorporate three-dimensional (3D) information, including molecular volume, surface area, and shape indices derived from spatial coordinates. These descriptors are vital in modeling steric effects influencing receptor fit and selectivity [8].

Electronic descriptors quantify charge distribution and orbital properties, often obtained from quantum chemical calculations using methods such as density functional theory (DFT). Examples include dipole moment, polarizability, ionization potential, and HOMO–LUMO energy gap all critical for understanding electrostatic and charge transfer interactions during ligand binding [9]. Thermodynamic descriptors reflect molecular reactivity and stability through properties such as Gibbs free energy, enthalpy, or solvation energy. These parameters provide insight into binding energetics and conformational equilibria, bridging the gap between structural features and bioactivity [10]. Collectively, these descriptor categories form the multidimensional basis of QSAR modeling. By integrating multiple descriptor types, models achieve a more holistic representation of molecular behavior, enabling accurate prediction of diverse biological endpoints such as enzyme inhibition, receptor affinity, and membrane permeability.

1.3 Theoretical Basis of Descriptor Calculation and Its Role in QSAR Modeling

Descriptor calculation is rooted in theoretical chemistry, graph theory, and statistical mechanics. Each descriptor encodes a measurable or computable aspect of molecular structure based on physical laws or mathematical abstraction. For example, topological indices arise from graph invariants, while electronic descriptors stem from solutions to the Schrödinger equation. The underlying principle of descriptor theory is the structure–property relationship (SPR), which asserts

that molecules with similar structures will exhibit similar properties. where A*A* denotes the biological activity and D1…Dn*D*1…*Dn* are the computed descriptors. The function f*f* may be linear (e.g., multiple linear regression) or nonlinear (e.g., support vector machines, neural networks) depending on data complexity [11]. Physicochemical descriptors thus serve as the explanatory variables linking chemical structure to observed activity.

Descriptor computation involves either empirical correlations or ab initio calculations. Empirical descriptors such as logP or polar surface area (PSA) are often derived from experimental data or additive fragment contributions. In contrast, quantum mechanical descriptors require electronic structure calculations using semi-empirical (AM1, PM7) or DFT methods to estimate molecular orbitals, electron density, and electrostatic potentials [12]. Geometrical descriptors typically arise from optimized 3D structures generated by molecular mechanics force fields (MMFF94, OPLS4) or conformational sampling via Monte Carlo or molecular dynamics simulations. Importantly, descriptor reliability depends on structural accuracy, standardization of molecular orientation, and reproducibility of computational conditions. For instance, descriptors derived from multiple conformations must be averaged or weighted based on Boltzmann populations to capture realistic behavior in biological environments. As the dimensionality of QSAR increases (from 3D to 6D), descriptor computation incorporates dynamic and environmental effects, reflecting molecular flexibility and receptor adaptation key elements for modern predictive pharmacology [13].

1.4 Descriptor Selection and Redundancy Elimination: Statistical and Machine Learning Approaches

A critical step in descriptor-based QSAR modeling is the identification of relevant variables from a potentially massive feature space. Modern descriptor-generation tools can calculate over 5,000 features per molecule, leading to redundancy, multicollinearity, and overfitting if all are used indiscriminately. Therefore, feature selection is vital to enhance model interpretability, predictive performance, and computational efficiency [14]. Classical statistical approaches, such as stepwise regression, variance inflation factor (VIF) analysis, and principal component analysis (PCA), help detect and eliminate correlated descriptors. PCA, for instance, transforms correlated variables into orthogonal principal components, preserving variance while reducing dimensionality. Partial least squares (PLS) regression further identifies latent variables that maximize covariance between descriptor space and biological activity [15].

Machine learning approaches have revolutionized descriptor selection by introducing nonlinear, data-driven techniques. Recursive feature elimination (RFE) with support vector machines (SVMs), genetic algorithms (GAs), random forest (RF) importance ranking, and mutual information (MI)-based filtering are widely applied to identify the most informative descriptors [16]. Hybrid methods combining statistical and AI techniques have shown superior robustness, particularly when dealing with noisy or high-dimensional datasets. Descriptor interpretability remains a critical consideration. While AI-driven selection may yield high predictive accuracy, excessive automation can obscure mechanistic insight. Therefore, the best practice involves a balanced approach: using automated selection to reduce dimensionality while validating selected descriptors against known physicochemical principles. Software like QSARINS, KNIME, and Orange3 facilitate this process by integrating feature selection, visualization, and statistical validation modules within a unified workflow [17].

1.5 One-Dimensional QSAR (1D-QSAR): Linear Relationships between Physicochemical Properties and Activity

One-dimensional QSAR represents the foundational stage of quantitative modeling, describing biological activity as a direct function of simple physicochemical parameters. These models rely primarily on scalar descriptors such as partition coefficients, pKa values, molar refractivity, and substituent constants. The classic Hansch equation, introduced in the 1960s, mathematically captured these relationships by correlating biological activity with lipophilicity (logP), electronic effects (σ), and steric parameters (Es) [18]. The general form of the Hansch equation is expressed as:

$$\log(1/C)=a\log P+b(\log P)2+c\sigma+dEs+k\log(1/C)=a\log P+b(\log P)2+c\sigma+dEs+k$$

where C*C* represents the concentration required to elicit a biological response, and the coefficients a,b,c,*a*,*b*,*c*, and d*d* denote the relative contributions of each physicochemical property. The quadratic term accounts for parabolic relationships between lipophilicity and biological activity, acknowledging that both excessive hydrophobicity and hydrophilicity can reduce drug efficacy. 1D-QSAR models are often linear, enabling straightforward interpretation and mechanistic insight. They are particularly useful for series of congeneric compounds sharing a common scaffold but differing in substituent patterns. The Free–Wilson approach complements Hansch analysis by directly associating structural fragments with activity changes, thereby quantifying the additive effects of substituent modifications [19]. Despite their simplicity, 1D-QSAR models remain relevant for rapid screening and for guiding early lead optimization where limited structural diversity exists. Their main advantage lies in interpretability and low computational cost. However, limitations arise from their inability to account for conformational flexibility, receptor interactions, or three-dimensional shape effects. As the complexity of drug–target systems increased, 1D-QSAR gradually evolved into higher-dimensional frameworks incorporating spatial, electronic, and dynamic parameters to capture more realistic molecular behavior [20].

1.6 Two-Dimensional QSAR (2D-QSAR): Topological and Fragment-Based Representations

Two-dimensional QSAR extends the classical 1D approach by incorporating information derived from molecular connectivity and graph theory. Molecules are treated as mathematical graphs, allowing the calculation of topological indices that describe connectivity, branching, and cyclicity patterns. Common 2D descriptors include the Wiener index (path length), Balaban's J index, Kier–Hall connectivity indices, and molecular fingerprints such as ECFP and MACCS keys [21]. In 2D-QSAR, structural variations are encoded without explicit three-dimensional coordinates, relying instead on relational information among atoms. These topological descriptors are robust to conformational changes and provide a compact way to represent large chemical libraries. Moreover, fragment-based methods identify substructures functional groups or pharmacophores associated with particular biological responses, thereby linking chemical motifs to activity trends [22].

A typical 2D-QSAR workflow involves generating a molecular graph, calculating topological descriptors using tools like PaDEL, Dragon, or RDKit, and correlating them with biological activity through statistical or machine learning algorithms such as multiple linear regression (MLR), partial least squares (PLS), support vector machines (SVM), or random forests (RF). Feature selection methods help to identify the most relevant topological patterns contributing to biological potency [23]. While 2D-QSAR captures structure–activity relationships across chemically diverse datasets, it does not explicitly consider molecular geometry or conformational preferences. Nevertheless, its computational efficiency and interpretability make it indispensable in chemoinformatics pipelines for virtual screening, scaffold hopping, and preliminary hit prioritization. Notably, several successful drug

discovery campaigns, including β-blockers and ACE inhibitors, were informed by 2D-QSAR models that accurately predicted activity trends within related compound series [24].

1.7 Three-Dimensional QSAR (3D-QSAR): Molecular Fields and Spatial Descriptors (CoMFA, CoMSIA)

The introduction of three-dimensional QSAR represented a paradigm shift by incorporating molecular shape and electrostatic fields into predictive modeling. In 3D-QSAR, molecular alignment and spatial field mapping form the basis of correlating three-dimensional descriptors with biological activity. The two most influential methodologies in this domain are Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) [25]. In CoMFA, molecules are superimposed based on a common structural framework, and interaction energies (steric and electrostatic) are computed at grid points surrounding each molecule using a probe atom. These energy values form a field matrix that captures how molecular features interact with hypothetical receptor environments. The resulting dataset is analyzed using partial least squares (PLS) regression to identify spatial regions where variations in molecular fields correlate with biological activity [26].

CoMSIA extends this concept by employing Gaussian distance-dependent functions and incorporating additional similarity indices hydrophobic, hydrogen bond donor, and acceptor fields. This method overcomes CoMFA's sensitivity to grid placement and alignment by providing smoother potential surfaces and improved interpretability [27]. 3D-QSAR models are powerful in elucidating structure activity maps, offering medicinal chemists visual feedback on which molecular regions enhance or diminish activity. The resulting contour maps serve as design blueprints for modifying substituents to optimize potency or selectivity. However, the reliability of 3D-QSAR depends critically on accurate molecular alignment and conformational representation. Misalignment or incorrect conformer selection can lead to spurious correlations. To mitigate these issues, ensemble-based or alignment-independent variants have been developed, bridging toward the higher-dimensional QSAR paradigms discussed in subsequent sections [28].

1.8 Four-Dimensional QSAR (4D-QSAR): Conformational Ensembles and Grid-Based Averaging

Four-dimensional QSAR (4D-QSAR) introduces an additional layer of realism by incorporating molecular dynamics and conformational diversity into predictive modeling. While 3D-QSAR assumes a single static conformation for each ligand, 4D-QSAR acknowledges that molecules exist as ensembles of conformations in solution or at the receptor site. This dynamic representation allows the model to average interaction fields across conformational space, reflecting thermally accessible geometries [29]. The core principle of 4D-QSAR lies in the concept of grid cell occupancy descriptors (GCODs). A set of conformations for each ligand is generated typically via molecular dynamics (MD) or Monte Carlo simulations and superimposed in a 3D grid representing the interaction field. The frequency with which atoms occupy specific grid cells forms the descriptor matrix, effectively encoding conformational flexibility [30]. Machine learning models, often based on PLS or neural networks, are then trained to correlate these occupancy patterns with biological activity.

One of the pioneering implementations of this methodology was developed by Hopfinger and colleagues, who demonstrated that averaging molecular interaction energies across conformations improved predictive accuracy for flexible ligands. Modern 4D-QSAR workflows automate this process using software like Quasar, GQSAR, or integrated MD–QSAR pipelines in MOE or Schrödinger Maestro [31]. The inclusion of dynamic conformational behavior makes 4D-QSAR particularly useful for ligands with multiple bioactive states or flexible linkers. However, this added realism comes with

increased computational cost and the need for careful sampling of relevant conformations. Despite these challenges, 4D-QSAR represents a critical advance toward modeling ligand–receptor interactions under biologically realistic conditions, serving as a conceptual bridge to 5D and 6D QSAR methodologies [32].

1.9 Five-Dimensional and Six-Dimensional QSAR: Receptor Flexibility and Dynamic Environmental Effects

The evolution of QSAR into five and six dimensions reflects ongoing efforts to incorporate receptor adaptability, solvent effects, and environmental dynamics into predictive modeling. Whereas 4D-QSAR considers ligand conformations, 5D-QSAR further accounts for multiple receptor conformations (e.g., induced fit effects), while 6D-QSAR integrates environmental parameters such as solvation, temperature, or ionic strength, providing the most comprehensive representation currently feasible in computational frameworks [33]. In 5D-QSAR, the receptor is no longer treated as rigid. Multiple receptor conformations, obtained from experimental structures, homology models, or molecular dynamics simulations, are included to represent flexible binding environments. For each ligand–receptor combination, descriptors such as interaction energies, hydrogen bonding patterns, or contact surface areas are computed and statistically analyzed. The ensemble averaging over both ligand and receptor conformations enables the model to capture induced-fit phenomena, which are crucial for accurately predicting binding affinities in dynamic systems [34].

6D-QSAR further extends this by introducing environmental descriptors parameters that quantify the effects of solvent interactions, dielectric constant, and temperature on molecular behavior. For example, free energy of solvation or hydration shell density can be incorporated into the model as an additional descriptor layer. Advanced computational techniques like implicit solvent models (Poisson–Boltzmann, Generalized Born) or explicit solvent molecular dynamics are commonly employed to generate these data [35]. These multidimensional QSAR approaches bridge the gap between static molecular modeling and full-scale molecular simulations. They enable a holistic view of drug–receptor interactions by integrating structural, dynamic, and environmental variables. The computational frameworks supporting such models include advanced platforms like Schrödinger's FEP+, BioVia Discovery Studio's 6D-QSAR modules, and customized machine learning pipelines that combine QSAR descriptors with MD-derived energetics. The predictive performance of these models, while computationally intensive, has been shown to outperform traditional QSAR approaches for flexible targets such as GPCRs, kinases, and proteases [36].

However, 5D and 6D-QSAR face inherent challenges high dimensionality, limited interpretability, and the need for large training datasets. As computational resources and AI-driven dimensionality reduction improve, these barriers are gradually being overcome. The integration of deep learning and hybrid simulation–QSAR frameworks marks a new frontier where multidimensional models achieve both predictive accuracy and mechanistic transparency.

The transition from 1D to 6D QSAR represents an evolution from simple linear relationships to highly complex, dynamic, and multidimensional models. Each QSAR dimension introduces additional layers of structural and environmental information, improving predictive power but also increasing computational demands and interpretability challenges. A comparative understanding of these models highlights how advancements in descriptor theory and computational capabilities have expanded the scope of structure activity correlation in drug design. 1D-QSAR models remain favored for their simplicity and interpretability. They are highly effective in cases where molecular variation is limited

and biological activity correlates with scalar physicochemical parameters such as lipophilicity or electronic effects. These models are computationally inexpensive and suitable for early drug discovery screening but lack the capacity to account for three-dimensional or dynamic effects [37].

2D-QSAR models provide a richer structural context by encoding atomic connectivity, branching, and fragment distributions. They excel in virtual screening and scaffold-based drug discovery where topology is a key determinant of activity. The use of topological indices and molecular fingerprints allows for rapid high-throughput analysis across diverse datasets. However, they still treat molecules as static entities and cannot capture conformational preferences or spatial interactions [38]. 3D-QSAR methodologies like CoMFA and CoMSIA revolutionized the field by linking spatial fields to biological activity. They provide visual contour maps illustrating steric and electrostatic contributions to potency. These models are highly informative for structure-guided optimization, enabling chemists to pinpoint regions favorable for substituent modification. Yet, their dependence on molecular alignment and single-conformation representation introduces uncertainty, especially for flexible ligands [39].

4D-QSAR and higher models address these issues by integrating molecular flexibility and conformational ensembles. 4D-QSAR accounts for the thermally accessible conformations of ligands, whereas 5D-QSAR introduces receptor flexibility, and 6D-QSAR integrates environmental and solvent parameters. Collectively, these models simulate a dynamic biochemical reality, closely approximating the complexity of ligand–receptor interactions. Their predictive accuracy is generally higher than lower-dimensional models, though this comes at the cost of interpretability and computational intensity [40]. From a practical standpoint, model selection depends on the balance between accuracy, interpretability, and computational feasibility. While higher-dimensional QSARs are theoretically superior, their advantages manifest only when high-quality structural and activity data are available. Conversely, 1D and 2D models remain invaluable in early discovery phases or when data scarcity limits model generalization. Modern CADD workflows often integrate multiple QSAR dimensions using 2D-QSAR for large-scale screening and higher-dimensional models for refined lead optimization [41].

1.11 Software Platforms and Computational Workflows for Descriptor Generation and Multidimensional QSAR

The practical implementation of QSAR modeling relies heavily on computational platforms that facilitate descriptor calculation, data preprocessing, model building, and validation. These tools range from specialized descriptor generators to integrated modeling environments and machine learning frameworks. Each plays a critical role in transforming molecular data into predictive models.

Descriptor Generation Tools

- **Dragon** (by Kode Chemoinformatics) computes over 5,000 molecular descriptors across 0D–3D categories, including constitutional, topological, geometrical, and electronic types. It is widely used in both academic and industrial QSAR pipelines [42].
- **PaDEL-Descriptor** provides an open-source alternative capable of calculating over 1,400 descriptors and fingerprints. It integrates seamlessly with KNIME and Python-based workflows for high-throughput analysis [43].
- **RDKit**, a Python chemoinformatics library, allows customized descriptor computation, molecular fingerprints (e.g., ECFP, MACCS), and 3D structure handling within machine learning pipelines.

- **MOE** (Chemical Computing Group) and **Discovery Studio** (Dassault Systèmes) combine descriptor generation with molecular modeling, docking, and pharmacophore tools, enabling holistic CADD workflows.

Modeling and Workflow Integration Tools:

- **QSARINS**, developed by the University of Insubria, is tailored for linear regression-based QSAR modeling and adheres to OECD principles for model validation and interpretability.
- **KNIME** offers a modular workflow platform for integrating descriptor computation, feature selection, and machine learning. Nodes for RDKit, Weka, and Python enable flexible QSAR pipeline design.
- **DeepChem** and **TensorFlow** provide deep learning frameworks for constructing nonlinear QSAR and molecular representation models using graph neural networks (GNNs) or convolutional neural networks (CNNs).
- **Schrödinger Maestro**, **BIOVIA Pipeline Pilot**, and **Simca** provide end-to-end suites for advanced multidimensional QSAR, 3D contour visualization, and statistical validation.

Workflow Overview

The typical QSAR workflow involves the following steps:

1. **Data Collection and Curation:** Extraction of chemical structures and biological activities from databases such as ChEMBL or PubChem.
2. **Structure Optimization:** Energy minimization and standardization using tools like Open Babel or MOE.
3. **Descriptor Calculation:** Generation of molecular descriptors via PaDEL, RDKit, or Dragon.
4. **Feature Selection:** Redundancy elimination using PCA, GA, or random forest-based ranking.
5. **Model Building:** Application of statistical or ML algorithms (PLS, SVM, RF, ANN).
6. **Validation:** Internal (cross-validation) and external (test-set) performance evaluation.
7. **Interpretation and Visualization:** Mapping of important descriptors or 3D contour fields.

The convergence of these software ecosystems ensures reproducibility, regulatory compliance, and interoperability between QSAR models and other CADD components such as docking and pharmacophore analysis [44].

1.12 Challenges, Validation, and Reproducibility in Descriptor-Based QSAR Modeling

Despite their widespread application, QSAR models face enduring challenges related to data quality, model overfitting, reproducibility, and interpretability. Descriptor-based modeling is only as reliable as the data underpinning it. Poorly curated datasets, inconsistent biological assay conditions, or ambiguous endpoint definitions can produce misleading correlations [45]. Hence, rigorous data preprocessing and standardization remain critical prerequisites. Another major issue is descriptor redundancy and collinearity, where multiple descriptors encode similar information. This inflates model complexity without enhancing predictive accuracy. Robust feature selection and dimensionality reduction are thus essential to mitigate multicollinearity effects. However, aggressive feature pruning risks discarding mechanistically relevant information, highlighting the trade-off between simplicity and completeness.

Validation is the cornerstone of trustworthy QSAR models. The OECD (Organisation for Economic Co-operation and Development) has established principles outlining the criteria for a valid QSAR model: (1) a defined endpoint, (2) an unambiguous algorithm, (3) a defined applicability domain, (4) appropriate measures of goodness-of-fit, robustness, and predictivity, and (5) mechanistic

interpretation if possible [46]. Statistical metrics such as R2*R*2, Q2*Q*2, root mean square error (RMSE), and external predictive Rpred2*Rpred*2 are routinely employed to assess model performance. Additionally, Y-randomization tests and bootstrapping help confirm that observed correlations are not due to chance. Reproducibility challenges arise when descriptor calculation parameters, molecular alignments, or preprocessing steps are not standardized. Even subtle variations in force field selection or geometry optimization methods can yield different descriptor values. Consequently, documentation of computational protocols, versioning of software tools, and adherence to FAIR (Findable, Accessible, Interoperable, Reproducible) principles are increasingly emphasized in QSAR research [47].

Interpretability also remains a concern in complex, nonlinear QSAR models. While deep learning algorithms may achieve exceptional predictive performance, their "black-box" nature complicates mechanistic understanding. To address this, explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are now being integrated into descriptor-based workflows to identify which molecular features drive predictions [48]. Through these advances, the community strives toward QSAR models that are not only accurate but also transparent, reproducible, and regulatory-compliant.

1.13 Applications of Multidimensional QSAR in Modern Drug Discovery

Multidimensional QSAR approaches have found applications across nearly every stage of the drug discovery pipeline, from hit identification to lead optimization and toxicity prediction. By combining descriptor-based modeling with experimental feedback, researchers can efficiently explore chemical space, prioritize compounds, and elucidate mechanisms of action.

Hit Identification and Virtual Screening

QSAR models particularly 2D and 3D variants serve as virtual screening filters for large compound libraries. By ranking molecules based on predicted potency or ADMET properties, these models drastically reduce the number of candidates requiring experimental validation. For example, 3D-QSAR contour maps have been used to guide the design of new HIV protease inhibitors and tyrosine kinase blockers, highlighting steric and electrostatic regions crucial for potency [49].

Lead Optimization

Multidimensional QSAR supports iterative structure refinement. 4D-QSAR models incorporating conformational sampling have successfully predicted binding affinities for flexible ligands, such as GPCR agonists and enzyme inhibitors, enabling targeted structural modifications. 5D-QSAR approaches incorporating receptor flexibility have improved selectivity modeling in kinase inhibitor design, where induced-fit effects are prevalent [50].

Toxicity and ADMET Prediction

Descriptor-based QSAR models remain indispensable in predicting absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles. Machine learning-enhanced 2D and 3D-QSAR frameworks accurately forecast hepatotoxicity, cardiotoxicity, and blood–brain barrier permeability, reducing late-stage attrition. Regulatory agencies such as the U.S. Environmental Protection Agency (EPA) and European Chemicals Agency (ECHA) increasingly accept validated QSAR models as in silico alternatives to animal testing under the REACH initiative [51].

Polypharmacology and Off-Target Prediction

Multidimensional QSAR enables the exploration of multi-target interactions by integrating descriptors reflective of ligand flexibility and receptor conformational diversity. 5D- and 6D-QSAR frameworks have been particularly effective in mapping cross-reactivity patterns across kinase families and GPCR subtypes, supporting the rational design of safer and more selective drugs [52].

Drug Repurposing

Large-scale QSAR models trained on multi-target bioactivity data have facilitated drug repurposing initiatives, identifying unexpected therapeutic potentials for existing drugs. Integration with deep learning and network pharmacology further enhances these predictive capabilities, highlighting QSAR's growing role in translational bioinformatics [53]. Through these diverse applications, multidimensional QSAR continues to evolve as a central analytical pillar of modern computer-aided drug design, complementing experimental and AI-driven methodologies.

1.14 Future Perspectives: Integrating AI, Quantum Mechanics, and Multi-Omics into Descriptor Science

The future of descriptor-based QSAR lies at the intersection of artificial intelligence, quantum chemistry, and systems-level biology. As computational power and data availability expand, descriptor science is transitioning from handcrafted numerical features to learned representations derived from neural networks and quantum mechanical simulations.

AI and Deep Learning Integration

Graph neural networks (GNNs) and message-passing neural architectures now learn molecular features directly from atomic graphs, bypassing manual descriptor calculation. These data-driven representations capture intricate structure activity relationships and generalize across diverse chemical classes. Hybrid QSAR models that combine classical descriptors with AI-derived embeddings exhibit improved predictive performance and interpretability [54].

Quantum Mechanically Derived Descriptors

Advances in quantum computation and density functional theory are enabling high-precision electronic descriptors, such as frontier orbital distributions, polarizability tensors, and reaction field energies. These descriptors enrich QSAR models with fundamental physical information, linking molecular reactivity to biological function. Quantum machine learning (QML) approaches are emerging as a bridge between ab initio calculations and statistical modeling, offering sub-chemical-accuracy predictions for complex systems [55].

Table 3.1. Comparative Characteristics of 1D–6D QSAR Models

QSAR Dimension	Key Descriptor Type	Structural Representation	Flexibility Considered	Computational Demand	Major Applications	Principal Limitations
1D-QSAR	Scalar physicochemical properties (logP, σ, MR, pKa)	Molecular constants and substituent parameters	None	**Low**	Early SAR analysis, preliminary lead identification	Oversimplified; neglects 3D structure and receptor interaction

2D-QSAR	Topological, connectivity, and fragment-based indices	Atom–bond graph representation	Implicit (fixed topology)	**Low–Moderate**	Virtual screening, scaffold hopping, molecular similarity analysis	Ignores 3D orientation and conformational dynamics
3D-QSAR	Spatial field-based descriptors (steric, electrostatic, hydrophobic fields)	Superimposed molecular conformations in 3D grid	Limited (single conformer)	**Moderate**	Structure-guided ligand optimization, visualization of SAR maps	Alignment dependence; sensitive to conformer choice
4D-QSAR	Grid cell occupancy descriptors (GCODs), averaged field maps	Ensemble of conformations representing molecular dynamics	Explicit ligand flexibility	**High**	Flexible ligand modeling, dynamic SAR prediction	High computational cost; requires extensive sampling
5D-QSAR	Receptor–ligand ensemble interaction descriptors	Multiple receptor conformations (induced-fit representation)	Explicit receptor and ligand flexibility	**Very High**	Induced-fit modeling, allosteric site exploration	Requires accurate receptor data and dynamic binding models
6D-QSAR	Environmental and solvent-related descriptors (dielectric constant, solvation energy)					

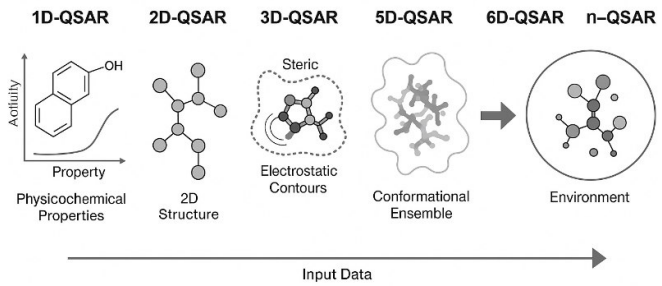

Figure 3.1. Evolution of Multidimensional QSAR Models from 1D to 6D

Multi-Omics and Systems-Level Integration

Next-generation QSAR frameworks are expanding beyond molecular-level descriptors to integrate genomic, proteomic, metabolomic, and transcriptomic data. This convergence, termed systems-QSAR, captures the biological context of drug action, enabling predictions of not only potency but also tissue specificity and patient response. Such integrative modeling aligns with the paradigm of precision medicine, where chemical and biological descriptors co-evolve within shared computational ecosystems [56].

Cloud and High-Performance Computing (HPC)

The increasing dimensionality of QSAR models demands scalable computational infrastructure. Cloud-based CADD platforms and GPU-accelerated workflows now support real-time descriptor generation and model retraining, fostering reproducibility and collaborative development. Ultimately, the integration of AI-driven feature learning, quantum mechanical precision, and systems-level biological context will redefine QSAR as a multidimensional science capable of bridging chemical theory and translational pharmacology. This evolution will not replace classical descriptor methodologies but rather amplify them, creating a continuum from interpretable empirical models to self-learning predictive systems a vision that embodies the next frontier of computer-aided drug design.

CONCLUSION

The systematic evolution of physicochemical descriptors and multidimensional QSAR modeling represents a cornerstone of modern computer-aided drug design. Beginning with the early one-dimensional models of Hansch and Fujita, where biological activity was correlated with simple parameters such as lipophilicity and electronic constants, QSAR has transformed into a multidimensional, data-rich discipline capable of simulating complex biochemical interactions. Through the progressive incorporation of topological, spatial, conformational, receptor, and environmental dimensions, the predictive scope of QSAR has expanded from static relationships to dynamic, mechanistically interpretable frameworks.

The conceptual journey from 1D to 6D QSAR demonstrates how chemical and biological realism can be systematically embedded into mathematical models. 1D and 2D approaches remain invaluable for their interpretability and computational simplicity, forming the backbone of early-stage screening and regulatory risk assessment. Conversely, 3D to 6D QSAR models leverage advanced descriptors, conformational ensembles, and receptor flexibility to approximate the dynamic nature of molecular recognition processes, thus improving predictive accuracy in lead optimization and selectivity profiling.

The integration of descriptor science with artificial intelligence, quantum chemistry, and multi-omics data heralds a new era of intelligent QSAR, where models evolve from empirical correlations to knowledge-driven systems capable of autonomous learning and mechanistic reasoning. Despite these advances, challenges persist in data quality, model interpretability, and reproducibility. Adherence to OECD validation principles, the adoption of FAIR data standards, and the incorporation of explainable AI will be critical in ensuring the reliability and ethical deployment of QSAR technologies in drug discovery.

Ultimately, physicochemical descriptors remain the quantitative language through which chemical structures communicate their biological intent. Their continued evolution driven by theoretical innovation, computational power, and interdisciplinary collaboration will sustain QSAR's

role as both a predictive science and a translational bridge between molecular design and pharmacological reality. In the broader context of computational drug design, multidimensional QSAR stands not merely as a modeling tool but as a conceptual framework that unifies chemistry, biology, and artificial intelligence in the pursuit of precision therapeutics.

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] Free SM Jr, Wilson JW. A mathematical contribution to structure–activity studies. J Med Chem. 1964;7(4):395–9.

[3] 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.

[4] 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.

[5] Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. GRid-INdependent Descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. J Med Chem. 2000;43(17):3233–43.

[6] Todeschini R, Lasagni M, Marengo E. New molecular descriptors for 2D and 3D structures. Theory. J Chemom. 1994;8:263–72.

[7] Cruciani G, Crivori P, Carrupt PA, Testa B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. J Mol Struct THEOCHEM. 2000;503:17–30.

[8] Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions. J Med Chem. 2000;43(20):3714–7.

[9] Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comput Sci. 1999;39(5):868–73.

[10] Lipinski CA. Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. Adv Drug Deliv Rev. 2016;101:34–41.

[11] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002;45(12):2615–23.

[12] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6–7):476–88.

[13] Chirico N, Gramatica P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model. 2011;51(9):2320–35.

[14] Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269–76.

[15] Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci. 2008;27(3):302–13.

[16] Roy PP, Chakraborty P, Mitra I, Ojha PK, Kar S, Das RN, et al. On two novel parameters for validation of predictive QSAR models. Molecules. 2009;14(5):1660–701.

[17] Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. OECD Series on Testing and Assessment No. 69. Paris: OECD; 2007.

[18] Organisation for Economic Co-operation and Development (OECD). (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure–Activity Relationship

models and predictions. 2nd ed. OECD Series on Testing and Assessment No. 405. Paris: OECD; 2024.
[19] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
[20] Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945–D954.
[21] Méndez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2019;47(D1):D930–D940.
[22] Zdrazil B, Williams AJ, Bento AP, et al. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024;52(D1):D464–D472.
[23] Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
[24] Mauri A. Dragon software: a software package for the calculation of molecular descriptors. WIREs Comput Mol Sci. 2020;10(1):e1458.
[25] Hawkins PCD, Nicholls A. Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Data Bank and the Cambridge Structural Database. J Chem Inf Model. 2010;50(4):572–84.
[26] Friedrich NO, Meyder A, de Bruyn Kops C, et al. Benchmarking commercial conformer ensemble generators. J Chem Inf Model. 2017;57(11):2719–28.
[27] Seidel T, Sander T, Becker T, et al. CONFORGE: a unified conformer ensemble generator for small molecules in cheminformatics. J Cheminform. 2023;15:68.
[28] Schaller D, Šribar D, Noonan T, et al. Next generation 3D pharmacophore modeling. WIREs Comput Mol Sci. 2020;10(6):e1468.
[29] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2:573–84.
[30] Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513–30.
[31] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning. PMLR. 2017;70:1263–72.
[32] Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702.e13.
[33] von Lilienfeld OA, Müller KR, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning. Nat Rev Chem. 2020;4(7):347–58.
[34] Cherkasov A, Muratov EN, Fourches D, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977–5010.
[35] Balaban AT. Highly discriminating distance-based topological index. Chem Phys Lett. 1982;89(5):399–404.
[36] Pires DEV, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015;58(9):4066–72.
[37] Hornberger KR, Savojardo C, Schreiber S, et al. Physicochemical property determinants of oral absorption for macrocycles and beyond-Rule-of-5 compounds. J Med Chem. 2023;66(20):13766–86.
[38] Khan AU, Asiri AM, Nasser IAM, et al. Descriptors and their selection methods in QSAR analysis: a review. J Mol Model. 2016;22:225.
[39] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound

classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947–58.
[40] Gianibbi G, Bartolini M, Andrisano V, et al. 3D-QSAR in drug discovery: recent advances and current challenges. Int J Mol Sci. 2024;25(2):913.
[41] Vedani A, Dobler M. 5D-QSAR: the key for simulating induced fit? J Med Chem. 2002;45(11):2139–49.
[42] Bak A, Skowronek P, Arczewska M, et al. Two decades of 4D-QSAR: a dying art or staging a comeback? Int J Mol Sci. 2021;22(10):5212.
[43] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform. 2011;3:33.
[44] Pastor M, Cruciani G, McLay I, Pickett S, Clementi S. Incorporating molecular shape into the alignment-free GRid-INdependent descriptors (GRIND). J Med Chem. 2004;47(26):6807–16.
[45] Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci. 2007;26(5):694–701.
[46] Organisation for Economic Co-operation and Development (OECD). Guidance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. OECD Series on Testing and Assessment No. 69. Paris: OECD Publishing; 2014.
[47] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
[48] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell. 2020;2(10):573–84.
[49] Bottegoni G, Kufareva I, Totrov M, Abagyan R. Four-dimensional docking: a fast and accurate account of discrete receptor flexibility in ligand docking. J Med Chem. 2016;59(17):7862–77.
[50] Sirimulla S, Bailey JB, Vegesna R, Narayan M. A comparative study of 3D and 4D-QSAR methods for predicting kinase inhibitor activity. J Chem Inf Model. 2018;58(12):2552–64.
[51] OECD QSAR Toolbox v4.6. The QSAR Application Toolbox. Helsinki: European Chemicals Agency (ECHA); 2023. Available from: https://qsartoolbox.org
[52] Ferreira LG, dos Santos RN, Oliva G, Andricopulo AD. Molecular docking and structure-based drug design strategies. Curr Top Med Chem. 2020;20(11):942–61.
[53] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688–702.e13.
[54] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2023;14(7):1811–26.
[55] von Lilienfeld OA, Müller KR, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning. Nat Rev Chem. 2020;4(7):347–58.
[56] Chen B, Ding Y, Liu C, Wu C, Chen H, Huang L. Integrating multi-omics data in drug discovery and development: progress and future perspectives. Trends Pharmacol Sci. 2022;43(9):721–39.