Comparing the Performance of Quantum Descriptors and Classical Descriptors based ML Models for Drug-like Property Prediction

by

Riya
MT23235
Department of Computational Biology

Under the Supervision of Dr. N. Arul Murugan

Indraprastha Institute of Information Technology Delhi

August, 2025

Comparing the Performance of Quantum Descriptors and Classical Descriptors based ML Models for Drug-like Property Prediction

by

Riya
MT23235

Submitted in partial fulfillment of the requirements for the degree of Master of Technology in Computational Biology to

Indraprastha Institute of Information Technology Delhi

August, 2025

Certificate

This is to certify that the thesis titled **"Comparing the Performance of Quantum Descriptors and Classical Descriptors based ML Models for Drug-like Property Prediction "** being submitted by **Riya** to the Indraprastha Institute of Information Technology Delhi, for the award of the **Master of Technology**, is an original research work carried out by **her** under my supervision.

In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree.

August, 2025

Supervisor Name : Dr. N. Arul Murugan
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

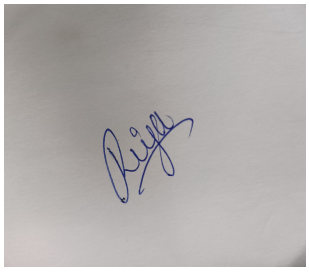
Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, **Dr. N. Arul Murugan**, for his constant support, insightful guidance, and unwavering encouragement throughout this research journey. His expertise and mentorship were instrumental in shaping the direction and depth of this thesis.

I would also like to thank my seniors, **Seraj** and **Abhishek Tripathi**, for their invaluable suggestions, motivation, and technical discussions that helped me overcome several challenges during the project. Their experience and patience played a significant role in my learning and confidence-building process.

My heartfelt thanks to all the faculty members and peers in the Department of Computational Biology at IIIT Delhi for fostering a collaborative and intellectually stimulating environment.

Finally, I am deeply grateful to my family and friends for their unconditional love, patience, and moral support that kept me going during this demanding phase of my academic life.



August 2025
Riya
MT23235
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Abstract

In drug discovery, the accurate prediction of pharmacokinetic properties plays a crucial role in identifying compounds with favorable absorption, distribution, metabolism, and excretion profiles. Among these, the partition coefficient (logP), blood–brain barrier permeability (logBB), and aqueous solubility (logS) are central to assessing drug–likeness and bioavailability. Traditional descriptor–based models often rely solely on classical cheminformatics features, which may fail to capture the underlying quantum–mechanical phenomena influencing molecular behavior.

This work integrates classical, quantum mechanical, and shape/surface descriptors (Quantum Mechanical Polar surface area(QMPSA)) to develop predictive models for logP, logBB, and logS. Quantum descriptors, including frontier orbital energies and Gibbs free energy difference (ΔG), were derived from Density Functional Theory (DFT) calculations, while surface and volumetric properties were computed using the Multiwfn program. These were complemented by classical Mordred descriptors, enabling the evaluation of six descriptor set combinations for LogP and LogS, while three descriptor set combinations were used for LogBB.

Models were systematically compared using curated datasets comprising 28,930 molecules for $\log P$, 6,460 for $\log S$, and 7,213 (classification) / 976 (regression) for $\log BB$. For $\log P$, the highest performance was achieved with an ExtraTrees regressor on Classical descriptors, attaining $R^2 = 0.9481$ and RMSE = 0.3847. For $\log S$, Gradient Boosting with Classical descriptors produced $R^2 = 0.8644$ and RMSE = 0.8488. The $\log BB$ models demonstrated complementary strengths: a classification model with Random Forest on a hybrid model with Classical descriptors and quantum descriptors achieved an accuracy of 0.9411 and an AUC of 0.9974 on hybrid classical and quantum descriptors, while a regression with ExtraTrees on a hybrid Classical+Quantum descriptor reached $R^2 = 0.665$ and RMSE = 0.403.

Together, these results highlight the dominance of Classical cheminformatics descriptors for lipophilicity and aqueous solubility, with hybridization yielding improvements in blood-brain barrier permeability tasks.

The results demonstrate that incorporating quantum–mechanical information alongside structural and surface descriptors enhances predictive accuracy and offers a generalizable, theory–driven framework for early–stage drug property estimation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Drug discovery is a multi–stage process in which early filtering based on pharmacokinetic properties can significantly reduce development time, cost, and late–stage attrition. Among these, three molecular properties are especially critical:

- **Lipophilicity** ($\log P$): the partition coefficient between water and n-octanol, which influences membrane permeability, absorption, and distribution.
- **Blood–brain barrier permeability** ($\log BB$): the ability of a molecule to cross the blood–brain barrier, essential for central nervous system (CNS)–targeted drugs.
- **Aqueous solubility** ($\log S$): the solubility of a molecule in water, affecting bioavailability, formulation stability, and delivery.

Traditional prediction methods for these properties often employ empirical, two–dimensional descriptors from cheminformatics tools such as Mordred or PaDEL. While computationally efficient, these descriptors may not fully capture solvation effects, electronic reactivity, and subtle structural characteristics that govern permeability and solubility.

Quantum mechanical descriptors, obtained from Density Functional Theory (DFT), provide a physics–informed alternative by encoding molecular electronic structure. Features such as HOMO, LUMO, SCF energy, dipole moment, ionization potential, and electrophilicity index offer insights into electronic behavior, while engineered surface descriptors from tools like Multiwfn capture solvent–accessible and volumetric properties. This integration of electronic and spatial descriptors can bridge the gap between empirical feature sets and the underlying molecular physics influencing ADMET–related endpoints.

In this study, unified prediction pipelines are developed for $\log P$, $\log BB$, and $\log S$ using combinations of classical, quantum, and engineered surface descriptors, along with Gibbs free energy differences (ΔG) where applicable.

1.2 Motivation

Accurate computational prediction of $\log P$, $\log BB$, and $\log S$ supports rational candidate prioritization in drug discovery. This work aims to improve such predictions by:

- Combining quantum mechanical descriptors with engineered surface features and classical descriptors.
- Incorporating solvent–specific calculations to better model partitioning and solubility behavior.
- Systematically benchmarking descriptor combinations to identify the most informative feature sets for each property.

For $\log P$, DFT calculations were performed in both water and n-octanol, enabling the computation of $\Delta G_{\text{octanol-water}}$ to represent partitioning energetics. For $\log S$, calculations were carried out in a single (water) environment. In both cases, Multiwfn was used to extract polar surface area, non–polar surface area, molecular volume, and density. For $\log BB$, descriptors were computed without Multiwfn features or ΔG, instead relying on classical and quantum descriptors for both classification and regression.

1.3 Scope of the Study

The scope of this work includes:

- Development of three unified pipelines for $\log P$, $\log BB$, and $\log S$ prediction.
- Use of six descriptor set combinations: Mordred + Quantum + Multiwfn + ΔG; Quantum + Multiwfn + ΔG; Quantum + Multiwfn; Quantum only; Multiwfn only; and Classical only.
- Model evaluation across multiple regressors/classifiers with selection based on R^2, RMSE, accuracy, and AUC–ROC.
- Comparative analysis of integrated descriptor sets versus single–source descriptors.

1.3.1 Quantum Descriptors as Predictive Tools

Quantum mechanical calculations capture molecular properties at the electronic level, reflecting polarity, charge distribution, and reactivity. The descriptors considered here

include HOMO, LUMO, dipole moment, electron affinity, ionization potential, electronegativity, chemical hardness/softness, electrophilicity index, molecular orbital energies, polarizability, and ΔG. These provide complementary information to surface and volumetric descriptors, enabling richer predictive models.

In detail, each quantum descriptor contributes unique insights:

- **HOMO (Highest Occupied Molecular Orbital):** Indicates the electron-donating capacity of a molecule. Molecules with higher HOMO energy are more prone to nucleophilic interactions.
- **LUMO (Lowest Unoccupied Molecular Orbital):** Represents the electron-accepting ability of a molecule. Lower LUMO energy corresponds to stronger electrophilic character. The HOMO–LUMO gap is widely used as a measure of overall stability and reactivity.
- **Dipole Moment:** Captures molecular polarity based on charge separation. It strongly influences solubility, intermolecular interactions, and transport across membranes.
- **Electron Affinity (EA):** The energy gained when a neutral molecule acquires an electron. A higher EA reflects a stronger tendency to act as an electron acceptor.
- **Ionization Potential (IP):** The energy required to remove an electron from a molecule. Molecules with higher IP are more resistant to oxidation and electron loss.
- **Electronegativity (χ):** Defined as $\chi = (IP + EA)/2$, this descriptor measures the overall tendency of a molecule to attract electrons.
- **Chemical Hardness (η) and Softness (S):** Hardness is defined as $\eta = (IP - EA)/2$, quantifying resistance to charge transfer. Softness, $S = 1/\eta$, is its reciprocal and indicates reactivity towards soft reagents.
- **Electrophilicity Index (ω):** Given by $\omega = \chi^2/2\eta$, it quantifies the stabilization energy when the system acquires charge, effectively capturing electrophilic strength.
- **Molecular Orbital Energies:** Beyond HOMO and LUMO, the full orbital spectrum reveals the electronic structure, excitation potential, and reactivity patterns.
- **Polarizability (α):** Reflects the extent to which a molecule's electron cloud can be distorted by an external electric field. Larger polarizability correlates with stronger dispersion interactions and solvation effects.

- ΔG **(Gibbs Free Energy):** Derived from quantum chemical calculations, this value indicates the thermodynamic favorability of processes such as solvation and conformational changes.

By combining these descriptors with shape, surface, and volumetric features, models achieve a more holistic representation of molecular behavior, ultimately improving predictions of drug-likeness and pharmacokinetic properties.

1.3.2 Surface and Volume-Based Descriptors as Predictive Tools

In addition to electronic descriptors, structural descriptors derived from surface and volumetric calculations provide valuable information about the geometric and physicochemical properties of molecules. In this study, these parameters were extracted using the **Multiwfn** program, which enables detailed analysis of molecular wavefunction outputs. The descriptors considered include non-polar surface area, polar surface area, molecular density, and molecular volume. Together, these features complement quantum mechanical descriptors by reflecting solubility, permeability, and packing behavior.

- **Non-Polar Surface Area (NPSA):** Extracted using Multiwfn, this quantifies the hydrophobic portion of the molecular surface. Molecules with larger NPSA are generally more lipophilic, influencing membrane permeability and affinity for hydrophobic protein pockets.

- **Polar Surface Area (PSA):** Computed in Multiwfn, PSA measures the hydrophilic portion of the surface arising from heteroatoms and polar bonds. It is strongly associated with hydrogen-bonding capacity and is widely used as a predictor of oral bioavailability and blood–brain barrier penetration.

- **Molecular Density:** Obtained from Multiwfn, density is defined as mass per unit volume, reflecting packing efficiency and compactness. Higher density values often indicate greater molecular stability but can be inversely related to solubility.

- **Molecular Volume:** Calculated by Multiwfn, molecular volume represents the three-dimensional space occupied by a molecule. It affects steric interactions, diffusivity across membranes, and the ability to fit into protein cavities.

These surface- and volume-based descriptors, extracted systematically using Multiwfn, provide complementary insights to quantum descriptors. While quantum parameters capture the electronic behavior of molecules, surface and volumetric descriptors describe how molecules occupy space and interact physically with their environment. The integration of both perspectives enhances the predictive capacity of computational models for solubility, permeability, and drug-likeness.

1.3.3 Property Definitions

Permeability ($\log P$) measures the relative affinity of a molecule for lipophilic versus hydrophilic environments:

$$\log P = \log_{10}\left(\frac{[\text{solute}]_{\text{octanol}}}{[\text{solute}]_{\text{water}}}\right)$$

A high $\log P$ value indicates strong lipophilicity, which generally enhances passive membrane diffusion, whereas excessively high values may reduce aqueous solubility and increase toxicity risk.

Aqueous solubility ($\log S$) is defined as the base–10 logarithm of a compound's molar solubility in water:

$$\log S = \log_{10}\left(C_{\text{water}}\right)$$

where C_{water} is the molar concentration of the solute in water at equilibrium. Solubility is influenced by intermolecular interactions, hydrogen bonding capacity, molecular polarity, and crystal lattice energy. Low $\log S$ values can lead to poor bioavailability, formulation challenges, and reduced therapeutic efficacy.

Blood–brain barrier permeability ($\log BB$) quantifies a compound's ability to cross the BBB:

$$\log BB = \log_{10}\left(\frac{[\text{drug}]_{\text{brain}}}{[\text{drug}]_{\text{blood}}}\right)$$

A $\log BB$ value greater than 0.3 generally indicates high brain penetration, while values below -1.0 suggest poor permeability across the barrier.

1.3.4 Machine Learning Approach

For each property, datasets were curated to remove duplicates and descriptor errors, resulting in 28,930 molecules for $\log P$, 6,460 for $\log S$, and 7,213 (classification) / 976 (regression) for $\log BB$. Classification and Regression models were trained and validated using cross–validation, with performance metrics including R^2, RMSE, accuracy, and AUC–ROC. Descriptor set combinations were compared to determine optimal feature integration.

1.4 Comparative Analysis with Traditional Methods

This work benchmarks integrated quantum–mechanical and engineered descriptors(extracted using Multiwfn) against traditional physicochemical models. The results highlight that

combining quantum–level information with surface and classical features consistently improves predictive accuracy and interpretability, offering a scalable framework for diverse ADMET property prediction.

Chapter 2

Background and Related Work

2.1 Overview of Related Work

The accurate prediction of pharmacokinetic properties such as lipophilicity ($\log P$), blood–brain barrier permeability ($\log BB$), and aqueous solubility ($\log S$) remains a key challenge in drug discovery. These properties strongly influence the absorption, distribution, metabolism, and excretion (ADME) profiles of drug candidates. While experimental determination is the gold standard, it is often laborious, costly, and unsuitable for large-scale screening. Computational models therefore play an increasingly important role in early-stage filtering of candidate molecules.

Initial predictive efforts were largely rooted in the Quantitative Structure–Activity Relationship (QSAR) framework, which uses statistical correlations between molecular descriptors and experimental endpoints. Early QSAR models relied on two-dimensional, handcrafted descriptors such as molecular weight, hydrogen bond donors/acceptors, and topological indices, computed from cheminformatics toolkits like Mordred, PaDEL, or RDKit. These descriptors have powered a variety of regression and classification models, including linear regression, support vector machines (SVM), and ensemble tree methods. While efficient and interpretable, purely classical descriptors can be insufficient for capturing solvent effects, three-dimensional conformation, and electronic properties that critically influence $\log P$, $\log BB$, and $\log S$.

2.2 Existing Methods and Theories

2.2.1 Mordred and Classical Descriptors

Mordred is an open-source descriptor calculator capable of generating over 1,800 descriptors covering topological, geometrical, constitutional, and electronic categories. Such descriptors have been widely applied to property prediction tasks including permeabil-

ity and solubility. A standard classical ML pipeline involves descriptor preprocessing, dimensionality reduction via feature selection, and model training using methods such as Random Forests, Gradient Boosted Trees, or Support Vector Regression. Although computationally lightweight, classical descriptors generally lack explicit representation of quantum mechanical effects or environment-specific behavior.

2.2.2 Quantum Chemical Descriptors

Quantum chemical methods, notably Density Functional Theory (DFT), provide a physics-based description of molecules by explicitly modeling electron density and molecular orbitals. Software packages like Gaussian and ORCA enable the calculation of descriptors including:

- HOMO and LUMO energies
- HOMO–LUMO gap
- Dipole moment
- Ionization potential (IP) and electron affinity (EA)
- Electronegativity (χ), chemical hardness (η), and electrophilicity index (ω)
- Gibbs free energy differences (ΔG) between solvent environments

These descriptors can be computed under different solvent models (e.g., water, octanol) using approaches such as the Polarizable Continuum Model (PCM), allowing the capture of environment-specific effects relevant to $\log P$ and $\log S$. Although more computationally demanding, quantum descriptors often yield richer insights into molecular behavior than purely empirical descriptors.

2.2.3 Hybrid Approaches

Hybrid descriptor frameworks integrate classical, quantum mechanical, and engineered surface/shape features. For example, augmenting Mordred descriptors with Multiwfn-derived polar surface area, non-polar surface area, molecular volume, and density—along with ΔG values from solvent-specific DFT calculations—can better capture both physical and chemical determinants of ADMET endpoints. Similarly, combining quantum descriptors with GNN-derived latent features provides a multi-perspective representation of molecular properties.

These approaches aim to leverage the interpretability and efficiency of classical descriptors while incorporating the physical realism of quantum calculations and the relational learning capabilities of graph-based models. Key challenges include aligning

descriptor scales, avoiding redundancy, and managing computational cost in large-scale applications.

2.3 Conclusion of Literature Review

Molecular property prediction has evolved from empirical, descriptor-based QSAR models to data-driven graph neural networks and quantum-informed hybrid frameworks. Classical descriptors offer speed and interpretability but lack the depth to model subtle physicochemical phenomena; quantum descriptors are accurate but costly to generate.

This thesis benchmarks integrated descriptor approaches for three critical pharmacokinetic endpoints—$\log P$, $\log BB$, and $\log S$—using curated datasets and multiple machine learning strategies. The objective is to determine whether hybridizing quantum mechanical and classical descriptors offers a measurable performance advantage and practical feasibility for large-scale drug discovery pipelines.

Chapter 3

Methodology

This chapter outlines the methodological framework for predicting three pharmacokinetic properties—$\log P$, $\log BB$, and $\log S$—using a combination of classical (Mordred), quantum mechanical, and engineered surface/shape descriptors. The study integrates cheminformatics, quantum chemistry, and machine learning techniques to produce a unified yet property-specific prediction pipeline.

3.1 Dataset Preparation

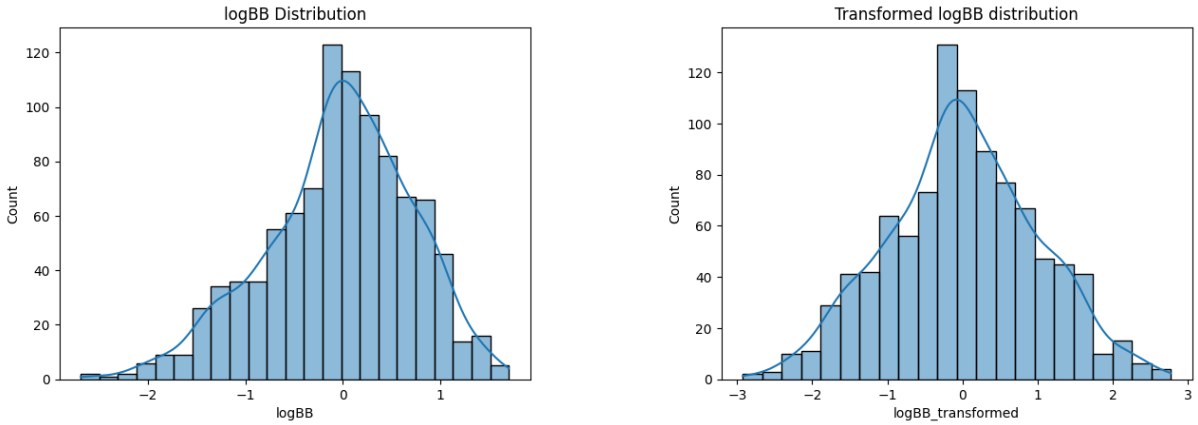
Three curated datasets were used:

- **logP:** 28,930 molecules with experimentally determined partition coefficients. Dataset compiled from multiple public sources and processed for two solvent environments: n-octanol and water.
- **logS:** 6,460 molecules with experimentally measured aqueous solubility values. Dataset processed for a single solvent environment: water.
- **logBB:** 7,213 molecules (classification: BBB+/BBB−) and 976 molecules (regression: continuous $\log BB$ values). Dataset compiled from publicly available sources and processed for water, and octanol environments.

logBB Data Transformation

For the $\log BB$ dataset, exploratory data analysis revealed that the raw distribution was slightly skewed, with heavy tails in both positive and negative ranges. Such skewness can reduce regression model stability and predictive accuracy. To address this, a transformation was applied to normalize the distribution and center the data closer to zero, enhancing numerical stability for downstream machine learning models.

Figure 3.1 shows a side-by-side comparison of the original and transformed log BB distributions. The transformation reduces skewness, producing a more symmetric, Gaussian-like distribution suitable for regression modeling.



(a) Raw log BB distribution

(b) Transformed log BB distribution

Figure 3.1: Comparison of log BB distributions before (left) and after (right) transformation. The transformation reduces skewness and produces a more symmetric distribution suitable for regression analysis.

All datasets underwent preprocessing steps to ensure consistency:

1. Canonicalization of SMILES strings to remove duplicates.
2. Conversion to 3D molecular structures (`mol2`) using OpenBabel.
3. Geometry optimization with UCSF Chimera to remove steric clashes.
4. Conversion to `xyz` format for quantum chemistry calculations.

Chimera Minimisation and Energy Stabilisation

To obtain physically realistic conformations, all molecules were subjected to geometry minimisation in UCSF Chimera. This step uses molecular mechanics force fields (such as AMBER ff14SB) to iteratively reduce steric clashes, unrealistic torsion angles, and high-energy contacts. By applying steepest descent followed by conjugate gradient methods, structures are driven into local minima on the potential energy surface. This stabilisation ensures that downstream quantum chemistry calculations (DFT-based) begin from physically reasonable starting geometries, improving both accuracy and convergence.

Below molecular structure illustrates an example of a molecule before and after minimisation. High-energy strained bonds and steric clashes (visible as distorted geometries) are relaxed into stable, low-energy conformations.

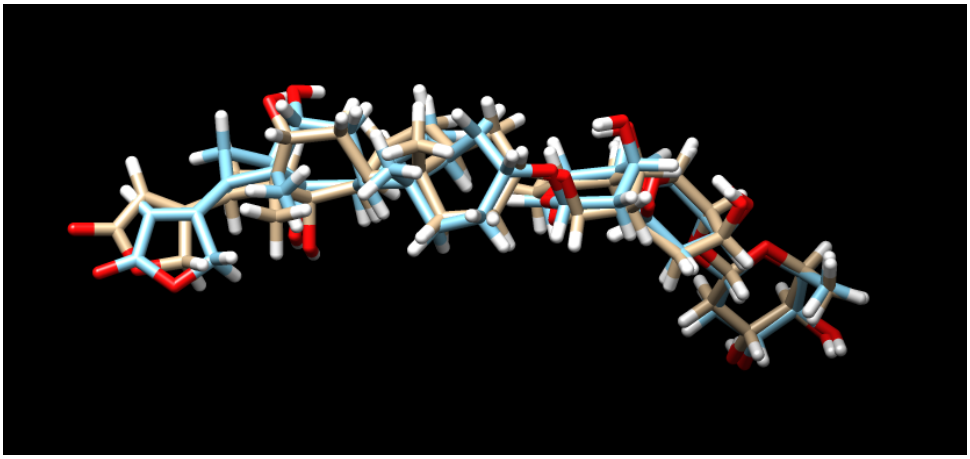


Figure 3.2: Example of UCSF Chimera minimisation showing energy stabilisation of a molecular structure. Minimisation reduces steric clashes and unrealistic bond angles, generating physically stable conformations for subsequent analysis.

3.2 Descriptor Generation

3.2.1 Classical Descriptors (Mordred)

The Mordred Python package was used to compute over 1,600 two-dimensional descriptors per molecule, covering constitutional, topological, geometrical, and electronic categories. Descriptors with missing values or zero variance were removed. All values were standardized prior to model training.

3.2.2 Quantum Descriptors

Quantum mechanical descriptors were calculated using Density Functional Theory (DFT):

- **logP:** Gaussian 16, B3LYP/6-31+G(d) with PCM solvent model for octanol and water. Gibbs free energy difference ($\Delta G_{\text{octanol-water}}$) was calculated to model partitioning energetics.
- **logS:** Gaussian 16, B3LYP/6-31+G(d) with PCM solvent model for water only.
- **logBB:** ORCA 6.0, B3LYP/def2-TZVP in octanol, water for single point based calculations. No Multiwfn descriptors were used.

Extracted quantum descriptors included:

- HOMO, LUMO, and HOMO–LUMO gap
- Self-consistent field (SCF) energy
- Dipole moment

- Potential Energy to Kinetic Energy ratio
- Ionization potential (IP) and electron affinity (EA)
- Electronegativity (χ), hardness (η), softness (S), electrophilicity index (ω)
- Gibbs free energy differences (logP only)

3.2.3 Engineered Surface Descriptors (Multiwfn)

For logP and logS, Multiwfn was applied to Gaussian `.fchk` files to extract:

- Polar surface area (PSA)
- Non-polar surface area (NPSA)
- Molecular volume
- Molecular density

These descriptors provide solvent-accessible shape information complementary to electronic properties.

3.3 Descriptor Set Combinations

In case of $\log P$ and $\log S$, Six descriptor combinations were evaluated for each endpoint:

1. Mordred + Quantum + Multiwfn + ΔG
2. Quantum + Multiwfn + ΔG
3. Quantum + Multiwfn
4. Quantum only
5. Multiwfn only
6. Mordred only

For $\log BB$, only Mordred and quantum descriptors individually and in combination were used.

3.4 Feature Selection

To reduce dimensionality and improve generalization, a three-step selection process was applied:

- Random Forest feature importance
- XGBoost feature importance
- Univariate selection via `SelectKBest` (f-regression or f-classification)

The intersection of top features from each method was used as the final feature set for model training.

3.5 Model Building

Classical machine learning models were implemented for regression (logP, logS, logBB regression) and classification (logBB):

- Extra Trees Regressor
- Gradient Boosting Regressor
- Random Forest Regressor / Classifier
- XGBoost Regressor / Classifier

Initial benchmarking used LazyPredict to identify strong candidates, followed by hyperparameter tuning with 5-fold cross-validation.

3.6 Evaluation Metrics

- **Regression:** R^2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE)
- **Classification:** Accuracy, Precision, Recall, AUC–ROC

Metrics were chosen to reflect both overall predictive power and error magnitude.

3.7 Workflow Overview

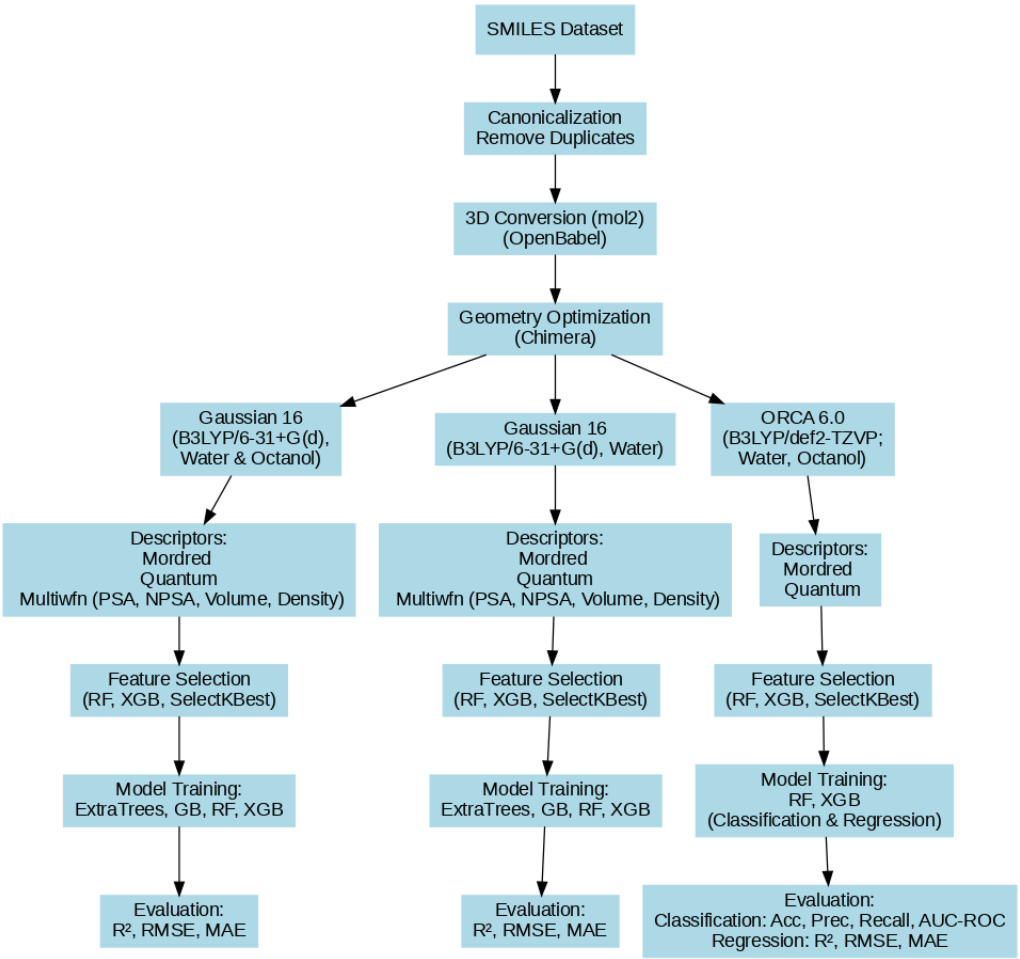


Figure 3.3: Overview of complete pipeline from dataset curation to descriptor extraction, feature selection, model training, and evaluation for $\log P$, $\log BB$, and $\log S$.

The workflow begins with canonicalization of SMILES strings and removal of duplicates, followed by 3D conversion of molecules using `OpenBabel`. Molecular geometries are optimized with `Chimera`, and quantum chemical calculations are performed using either `Gaussian 16` (B3LYP/6-31+G(d)) in water and octanol environments or `ORCA 6.0` (B3LYP/def2-TZVP). From these calculations, different sets of descriptors are extracted: classical Mordred descriptors, quantum chemical descriptors, and Multiwfn-derived properties such as polar surface area (PSA), nonpolar surface area (NPSA), molecular volume, and density. Feature selection is carried out using methods such as Random Forest (RF), XGBoost (XGB), and SelectKBest to identify the most informative features. Models including ExtraTrees, Gradient Boosting (GB), RF, and XGB are then trained for both regression and classification tasks. Finally, performance is evaluated using appropriate metrics: R^2, RMSE, and MAE for regression tasks, and accuracy, precision, recall, and AUC–ROC for classification tasks.

Chapter 4

Implementation and Experimental Work

This chapter details the computational setup, descriptor extraction workflows, and training procedures used in the three pipelines for predicting $\log P$, $\log S$, and $\log BB$.

4.1 Software and Tools

- **Operating system / Hardware:** Linux (x86_64); computations run on CPU nodes for descriptor generation; model training on CPU; random seeds fixed for reproducibility.
- **OpenBabel** (for SMILES → 3D conversion to `mol2`).
- **UCSF Chimera** (for geometry minimization prior to DFT).
- **Gaussian 16** (DFT descriptors for **logP** and **logS**).
- **ORCA 6.0** (DFT descriptors for **logBB**).
- **Multiwfn** (surface/shape descriptors: PSA, NPSA, volume, density; used for **logP** and **logS** only).
- **Python stack:** `rdkit`, `mordred`, `numpy`, `pandas`, `scikit-learn`, `xgboost`, `matplotlib`, `lazypredict`.

4.2 Descriptor Extraction Workflow

For each molecule, the following sequence was executed:

1. **Canonicalization:** SMILES were canonicalized and deduplicated.

2. **3D construction:** SMILES → `mol2` (OpenBabel).

3. **Geometry optimization:** Energy minimization in UCSF Chimera to remove steric clashes and obtain a low-energy conformation.

4. **Quantum inputs:** Conversion to `xyz`; Gaussian/ORCA input files prepared for the required solvent environment(s).

5. **DFT calculations:**

   - **logP:** Gaussian 16, B3LYP/6-31+G(d) with PCM (water and *n*-octanol); Gibbs free energy difference $\Delta G_{\text{octanol-water}}$ computed.
   - **logS:** Gaussian 16, B3LYP/6-31+G(d) with PCM (water).
   - **logBB:** ORCA 6.0, B3LYP/def2-TZVP (water, octanol).

6. **Descriptor parsing:** Quantum descriptors extracted from `.out` files (HOMO/LUMO, gap, SCF energy, dipole, IP, EA, χ, η, ω, and ΔG where applicable).

7. **Surface based descriptors (Multiwfn):** From Gaussian `.fchk` for **logP** and **logS**: PSA, NPSA, molecular volume, and density.

8. **Classical descriptors (Mordred):** >1600 2D descriptors per molecule; NaN/zero-variance features removed.

9. **Aggregation:** All descriptors merged into a single matrix per endpoint; consistent molecule IDs maintained to avoid leakage.

4.3 Geometry Optimization

All structures were minimized before DFT to reduce SCF failures and ensure descriptor stability. A standard protocol (steepest descent followed by conjugate gradient to convergence) was used. Problematic molecules (rare) were re-minimized or re-seeded.

4.4 Model Training

4.4.1 Data Splitting and Preprocessing

- **Splits:** 80 % train / 20 % test. For $\log BB$ classification, splits were stratified (BBB+/BBB−).

- **Scaling/Imputation:** Numeric features standardized with `StandardScaler` (fit on train only; transform applied to test). Occasional missing values imputed with median (train statistics only).

- **Feature selection:** Intersection of top-ranked features from Random Forest importance, XGBoost importance, and `SelectKBest` (f-statistic). Selection was performed within cross-validation to prevent information leakage.

4.4.2 Model Families and Tuning

- **Initial triage:** `lazypredict` used only for quick baseline ranking.
- **Final training:** ExtraTrees, Gradient Boosting, Random Forest, and XGBoost (regressors for $\log P$ and $\log S$; classifier + regressor for $\log BB$).
- **Validation:** 5-fold cross-validation on the training set; hyperparameters tuned via `GridSearchCV`/`RandomizedSearchCV`. Final models refit on the full training set with best settings and then evaluated once on the held-out test set.

4.5 Evaluation Metrics

- **Regression (logP, logS, logBB-reg):** R^2 and RMSE.
- **Classification (logBB-cls):** Accuracy, Precision, Recall, AUC–ROC.

4.6 Limitations and Challenges

- **SCF convergence:** A small subset failed at default thresholds; resolved via re-minimization, tighter SCF, or adjusted basis set for robustness.
- **Dimensionality:** Large descriptor spaces required careful filtering and consensus feature selection to avoid overfitting.
- **Compute cost:** DFT over thousands of molecules is time-intensive; batching and checkpointing were used to recover from interruptions.
- **Class balance (logBB):** Managed with stratified splits and threshold-agnostic AUC–ROC reporting.

Chapter 5

Results and Discussion

This chapter reports results for the three endpoints—partition coefficient ($\log P$), aqueous solubility ($\log S$), and blood–brain barrier permeability ($\log BB$)—across all trained models and descriptor bundles. Unless stated otherwise, regression models are evaluated with R^2, MAE, and RMSE; classification models with accuracy and AUC–ROC. We denote descriptor sets as:

- **CSS** = Mordred descriptors (classical/2D)
- **QM** = quantum descriptors from DFT
- **MWF** = Multiwfn surface descriptors (PSA, NPSA, volume, density)
- **G** = ΔG (free energy difference; used only for $\log P$ and $\log S$)

5.1 Log P: Lipophilicity

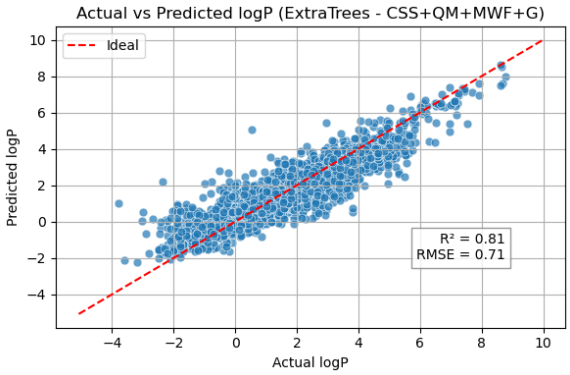
5.1.1 Model Comparison

Model ID	**Descriptor Set**	**Regressor**	**Features**	R^2
Model 1	CSS + QM + MWF + ΔG	ExtraTrees	1858	0.814
Model 2 (*Best Physics-Based*)	QM + MWF + ΔG	ExtraTrees	32	**0.844**
Model 3	QM + MWF	ExtraTrees	31	0.698
Model 4	QM only	Random Forest	23	0.482
Model 5	MWF only	ExtraTrees	8	0.593
Model 6 (*Best Overall*)	CSS only	ExtraTrees	1826	**0.9481**

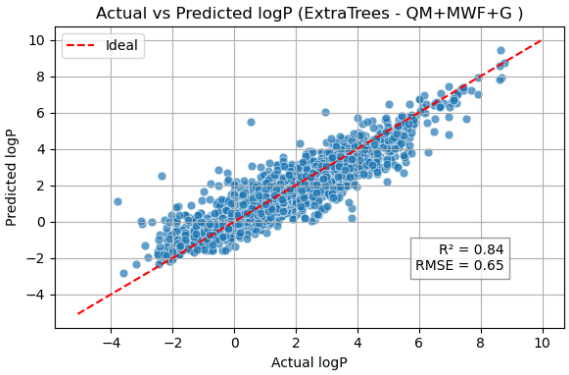
Table 5.1: Performance summary for $\log P$ across descriptor bundles. Model 2 is the strongest physics-based configuration; Model 6 gives the best overall accuracy.

Abbreviations: CSS—Mordred; QM—quantum; MWF—Multiwfn; ΔG—free energy.

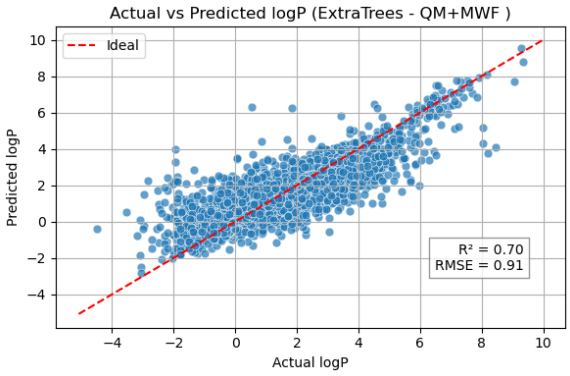
5.1.2 Visual Diagnostics



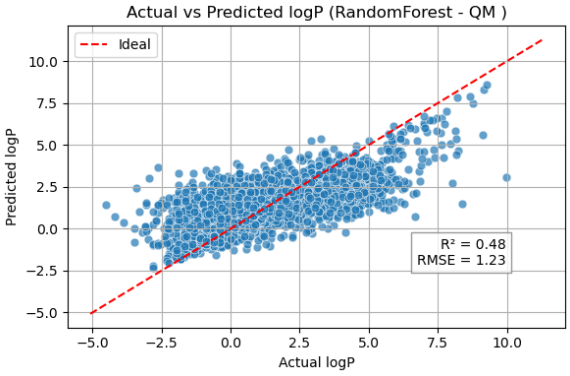
(a) CSS + QM + MWF + ΔG (ExtraTrees)



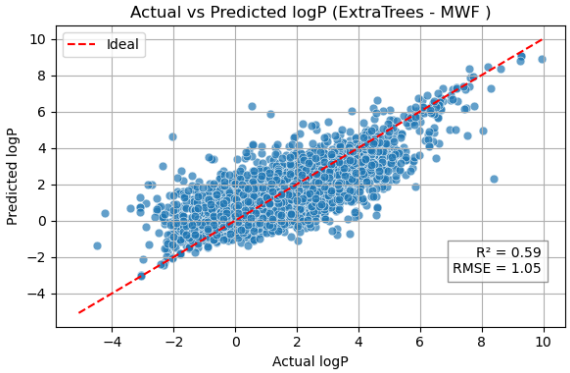
(b) QM + MWF + ΔG (ExtraTrees)



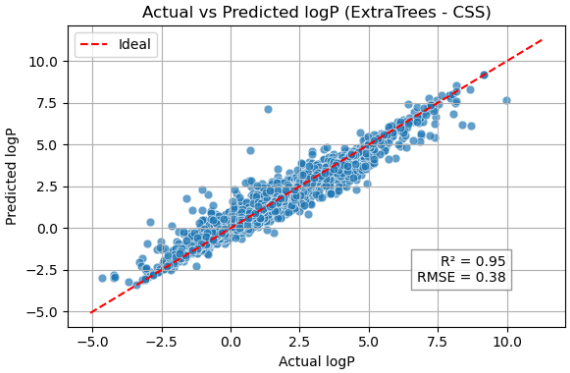
(c) QM + MWF (ExtraTrees)



(d) QM only (Random Forest)



(e) MWF only (ExtraTrees)



(f) CSS only (ExtraTrees)

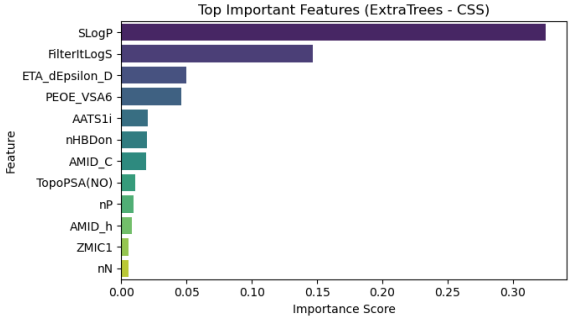
Figure 5.1: Actual vs. predicted log P for all six models (larger panels, two per row).

Performance interpretation.

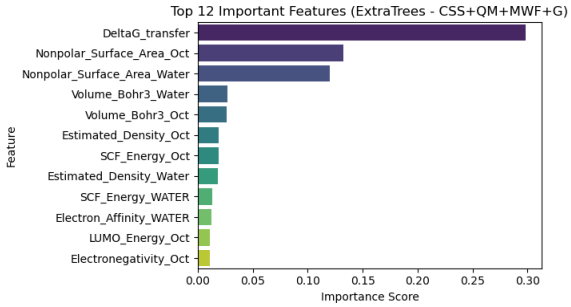
- **CSS only (Model 6)** shows the tightest clustering along the identity line with minimal spread at the extremes, matching its top metrics ($R^2 \approx 0.95$, RMSE ≈ 0.38).
- **QM + MWF + ΔG (Model 2)** is the best physics-based model: points follow the 45° trend with moderate dispersion at high log P, consistent with $R^2 \approx 0.84$.

- **CSS + QM + MWF + ΔG (Model 1)** underperforms CSS-only despite richer inputs—plots show slightly wider residual band, indicating added features didn't translate to extra signal (possible redundancy/high dimensionality).
- **QM + MWF (Model 3)** loses some linearity vs. Model 2, especially at the tails—matching its lower R^2—highlighting the benefit of ΔG.
- **Single-source baselines** (*QM only*, *MWF only*) show the broadest scatter and bias at extremes, confirming that electronic properties or shape/area alone are insufficient for lipophilicity.

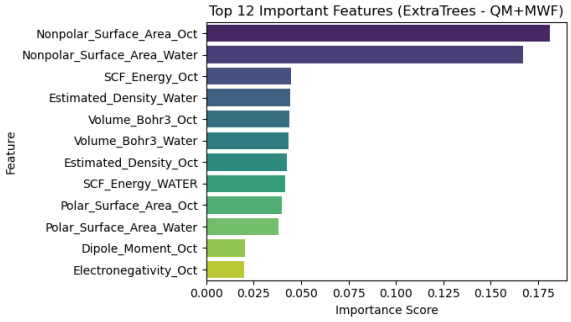
5.1.3 Feature Importance



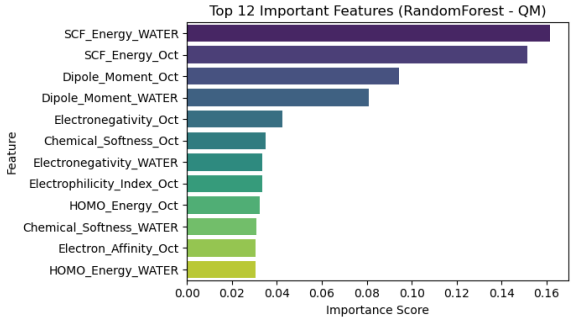
(a) CSS only (ExtraTrees)



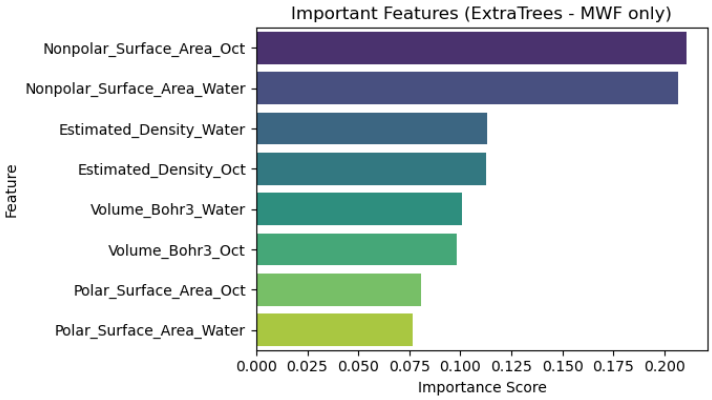
(b) CSS + QM + MWF + ΔG (ExtraTrees)



(c) QM + MWF (ExtraTrees)



(d) QM only (Random Forest)



(e) MWF only (ExtraTrees)

Figure 5.2: Top/important features across descriptor bundles (two per row; last row centered as only five FI plots are available).

Performance interpretation.

- **CSS only:** *SLogP* and *FilterItLogS* dominate—these encode lipophilicity directly, explaining the strong predictive power of the cheminformatics-only model.
- **Physics-enriched (CSS + QM + MWF + ΔG):** nonpolar surface areas in octanol/water, density/volume surrogates, and $\Delta G_{transfer}$ rank highly—capturing size/shape and solvent preference energetics.

- **QM + MWF:** solvent-accessible areas (octanol/water) and SCF-related terms are top contributors—shape plus coarse electronic stability.
- **QM only:** SCF energies and dipole-related terms lead—useful but incomplete without size/area terms, mirroring lower performance.
- **MWF only:** nonpolar/polar surface areas and densities drive signal—reasonable mid-tier accuracy but limited without orbital/electronic descriptors.

5.2 LogS : Aqueous Solubility

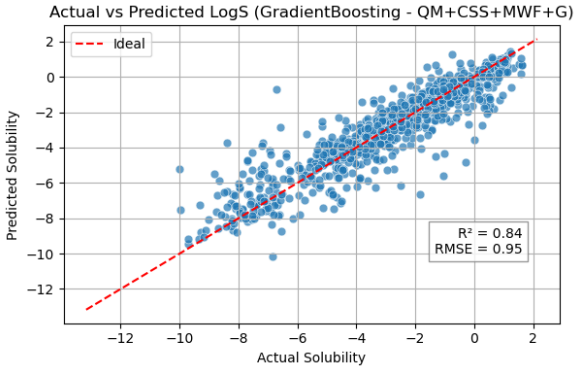
5.2.1 Model Comparison

Model ID	**Descriptor Set**	**Regressor**	R^2	**RMSE**
Model 1 (*Best Physics-Based*)	CSS + QM + MWF + ΔG	Gradient Boosting	0.8432	0.9458
Model 2	QM + MWF + ΔG	Random Forest	0.5214	1.7379
Model 3	QM + MWF	Random Forest	0.4982	1.7861
Model 4	QM only	Random Forest	0.5467	1.6653
Model 5	MWF only	Random Forest	0.5094	1.7661
Model 6 (*Best Overall*)	CSS only	Gradient Boosting	**0.8644**	**0.8488**

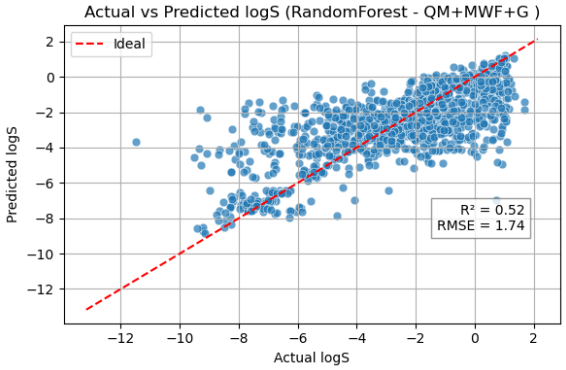
Table 5.2: Performance summary for $\log S$ across descriptor bundles. Although Model 6 (CSS only) provides the best statistical fit, Model 1 is highlighted as the best physics-based integrative model.

Abbreviations: CSS—Mordred; QM—quantum; MWF—Multiwfn; ΔG—free energy.

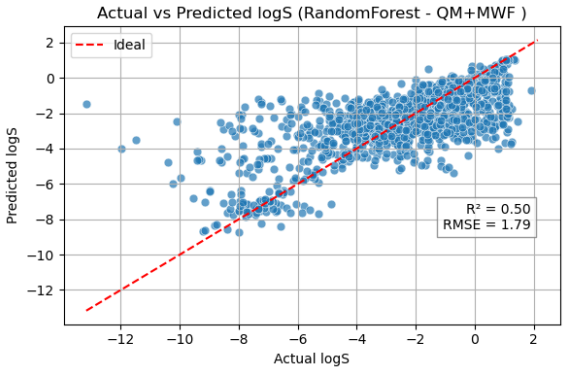
5.2.2 Visual Diagnostics (All Models)



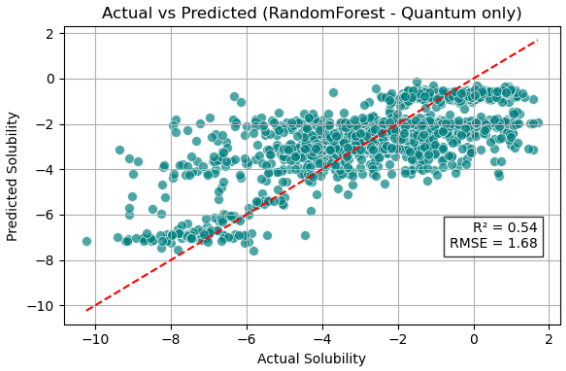
(a) CSS + QM + MWF + ΔG (Gradient Boosting)
Best Physics-Based



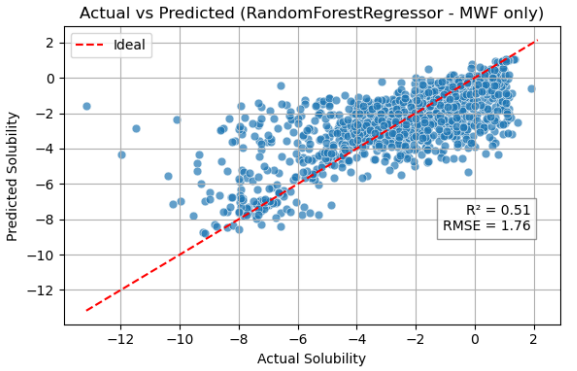
(b) QM + MWF + ΔG (Random Forest)



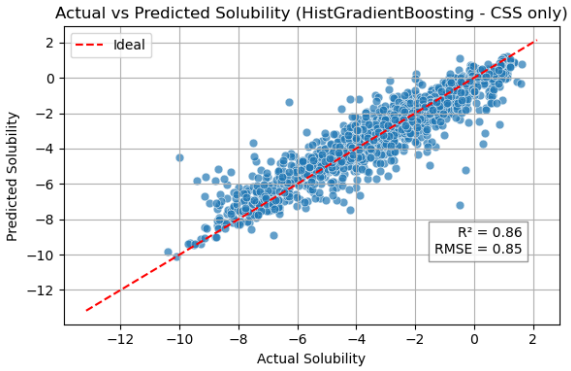
(c) QM + MWF (Random Forest)



(d) QM only (Random Forest)



(e) MWF only (Random Forest)



(f) CSS only (Gradient Boosting)
Best Overall

Figure 5.3: Visual diagnostics for $\log S$ prediction across descriptor sets. Two models stand out: CSS-only with Gradient Boosting as the most accurate, and CSS+QM+MWF+ΔG as the best physics-based rationale.

For solubility, two trends emerge. The *CSS only* model with Gradient Boosting achieves the highest statistical accuracy ($R^2 = 0.8644$, RMSE=0.8488). However, the integrative *CSS + QM + MWF* + ΔG model offers a balanced physics-based interpretation, highlighting how polarity, surface accessibility, and free energy cues jointly govern aqueous behaviour. Models relying on QM or MWF alone show higher dispersion, indicating insufficient explanatory power when electronic or geometric descriptors are isolated.

5.3 $\log BB$: Blood–Brain Barrier Permeability

5.3.1 Classification Model (BBB^+/BBB^-)

Model	Descriptor Set	Classifier	Accuracy	AUC–ROC
Model 1 (*Best Model*)	Classical + Quantum	Random Forest	**0.9411**	**0.9974**
Model 2	Quantum only	Random Forest	0.8739	0.6828
Model 3	Classical only (Mordred)	Random Forest	0.9287	0.8721

Table 5.3: Comparison of Random Forest classification models for $\log BB$ prediction using different descriptor sets.

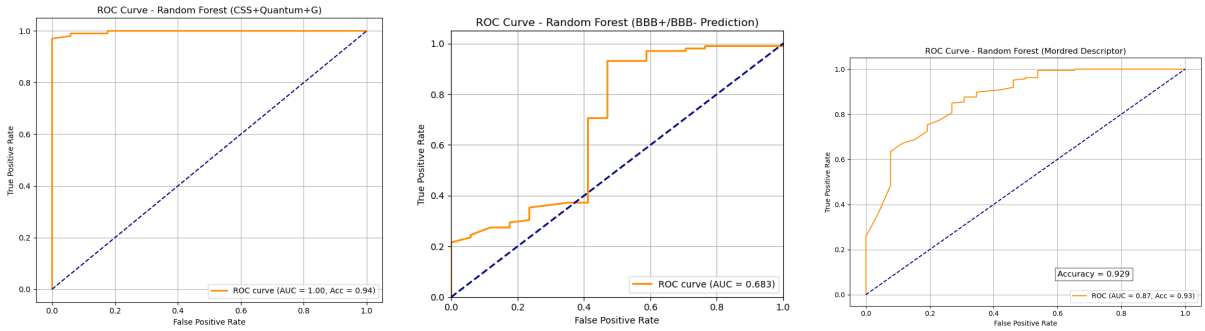


Figure 5.4: AUC–ROC curves for BBB^+/BBB^- classification using (left) Classical + Quantum , (middle) Quantum descriptors, and (right) Classical (Mordred) descriptors.

The results in Table 5.3 and Fig. 5.4 show a clear distinction between descriptor sets. The combined descriptor model (Model 1) achieved the highest accuracy (94.1%) and near-perfect AUC (0.997), reflecting strong predictive power and generalization as confirmed by 5–fold cross-validation (Accuracy = 0.927 ± 0.016, ROC–AUC = 1.000 ± 0.000). In contrast, the Quantum-only model (Model 2) showed limited discriminative ability (AUC = 0.683), while the Mordred-only model (Model 3) performed strongly with balanced accuracy and ranking (AUC = 0.872). These findings highlight the complementary role of combining classical and quantum descriptors.

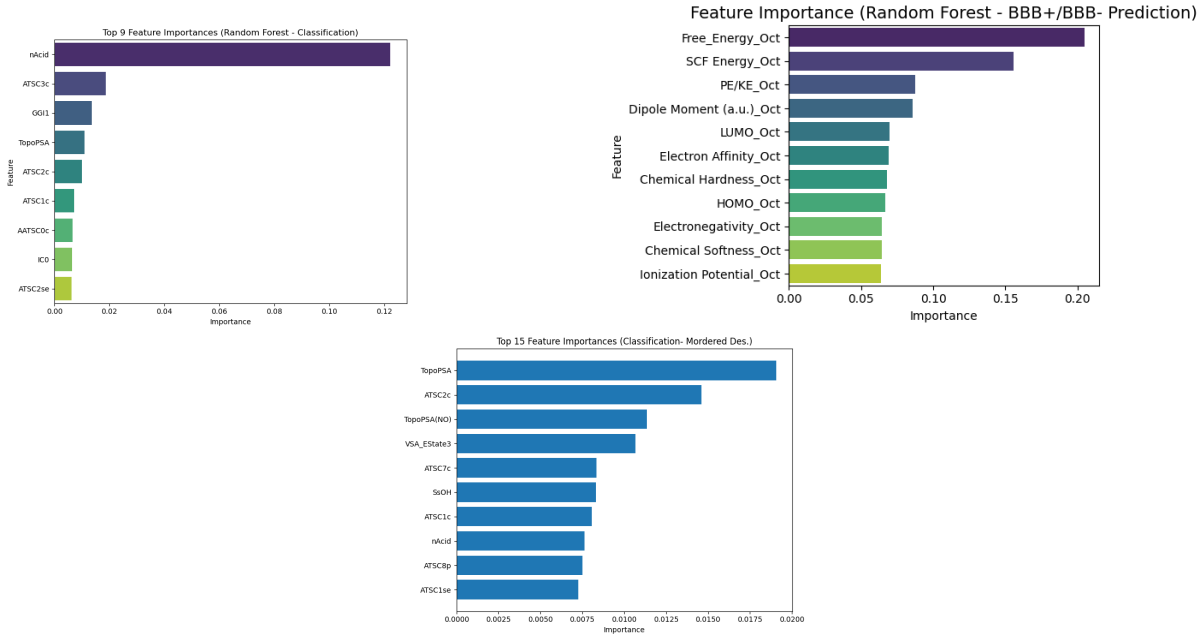


Figure 5.5: Feature importance for BBB^+/BBB^- classification using (left) Classical + Quantum + ΔG, (middle) Quantum descriptors, and (right) Mordred descriptors.

Feature importance analysis (Fig. 5.5) provides chemical interpretability. The combined model emphasized descriptors such as **TopoPSA**, **nAcid**, and autocorrelation indices (ATSC series), consistent with the known influence of polar surface area and ionization on BBB penetration. The Quantum-only model prioritized **dipole moment**, **SCF energy**, and orbital energies (HOMO/LUMO), highlighting the role of molecular energetics. In the Mordred-only model, topological indices such as **TopoPSA**, **ATSC2c**, and **VSA_Estate** dominated, reaffirming that polar surface area and structural topology are decisive for passive diffusion across the BBB.

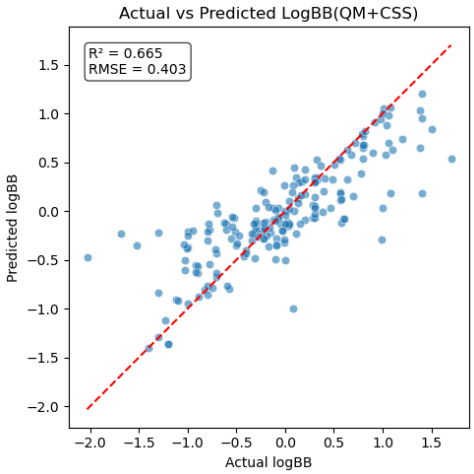
Overall, the classification results show that combining descriptor families (Classical + Quantum + ΔG) yields superior predictive performance, while still retaining meaningful chemical insights from both electronic and structural properties.

5.3.2 Regression (Continuous $\log BB$)

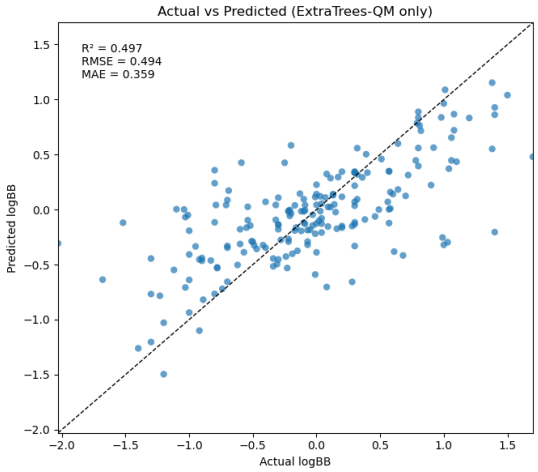
Model ID	**Descriptor Set**	**Regressor**	**Features**	R^2	**RMSE**
Model 1 (*Best Model*)	Classical + Quantum	ExtraTrees	758	**0.665**	**0.403**
Model 2	Quantum only	ExtraTrees	29	0.497	0.494
Model 3	Classical only	ExtraTrees	729	0.569	0.498

Table 5.4: Performance summary for $\log BB$ regression models. The hybrid Classical+Quantum feature set (Model 1) achieves the best accuracy and lowest error.

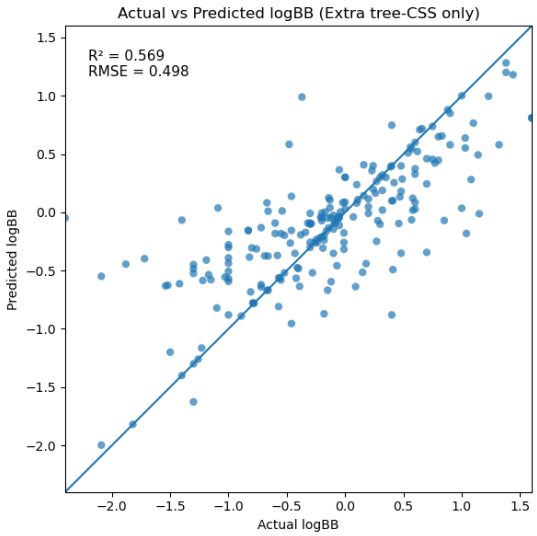
Visual Diagnostics



(a) Classical + Quantum (ExtraTrees)



(b) Quantum only (ExtraTrees)



(c) Classical only (ExtraTrees)

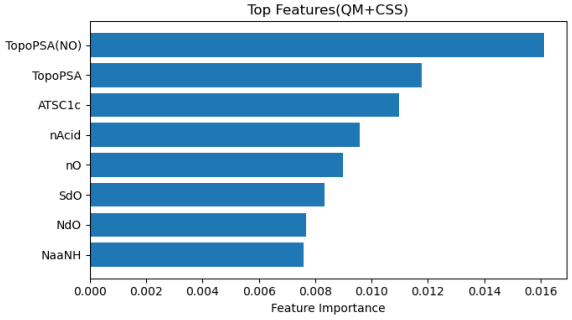
Figure 5.6: Actual vs. predicted log BB values for regression models.

Performance interpretation.

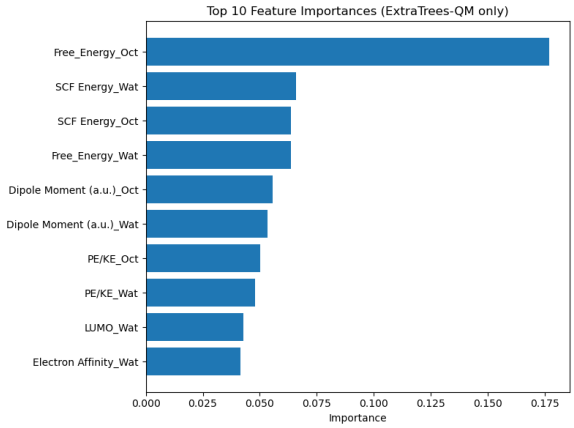
- **Model 1 (Classical + Quantum)** shows the strongest predictive alignment ($R^2 = 0.665$, RMSE=0.403). Points follow the $y = x$ line with moderate dispersion, especially at higher logBB values.
- **Model 2 (Quantum only)** has the weakest fit ($R^2 = 0.497$), showing broader scatter and bias at extremes. This indicates that quantum descriptors alone cannot fully capture permeability.
- **Model 3 (Classical only)** explains more variance than QM-only ($R^2 = 0.569$) but

yields higher RMSE (0.498), suggesting that cheminformatics descriptors capture global patterns but miss finer energetic effects.

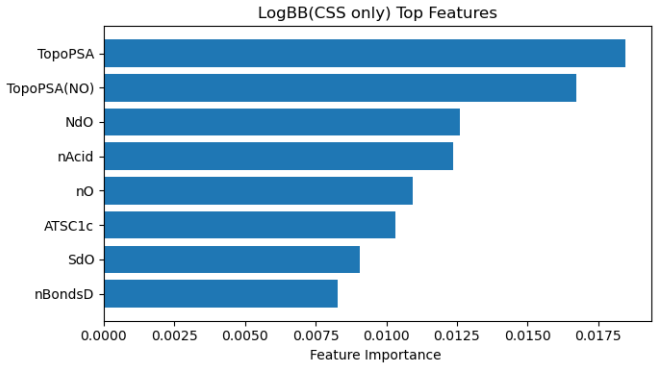
Feature Importance



(a) Classical + Quantum (ExtraTrees)



(b) Quantum only (ExtraTrees)



(c) Classical only (ExtraTrees)

Figure 5.7: Feature importance across regression models. Top descriptors vary between cheminformatics (surface area, counts), quantum (energetics, dipole), and hybrids (integration of both).

Feature interpretation.

- **Hybrid model (Classical + Quantum):** TopoPSA(NO), TopoPSA, and ATSC1c dominate, with quantum terms (dipole, SCF energy) contributing—showing that both structural accessibility and molecular energetics are important.
- **Quantum only:** Free Energy and SCF Energy descriptors dominate, followed by dipole moment and orbital-level terms (LUMO, electron affinity). This reflects a purely electronic/energetic view of permeability.
- **Classical only:** Topological polar surface area (TopoPSA, TopoPSA(NO)) and

atom count features (nO, nAcid, nBondsD) dominate, capturing size and polarity effects but missing energetic balance.

5.4 Cross-Property Comparison

Property	Best Model	Performance	Winning Descriptor Bundle
$\log P$	ExtraTrees	$R^2 = 0.948$, RMSE=0.385	CSS (Classical descriptors)
$\log S$	Gradient Boosting	$R^2 = 0.8644$, RMSE=0.8488	CSS (Classical descriptors)
$\log BB$ (Class.)	Random Forest	Acc=0.9411, AUC=0.9974	CSS + QM (Hybrid)
$\log BB$ (Regr.)	ExtraTrees	$R^2 = 0.665$, RMSE=0.403	CSS + QM (Hybrid)

Table 5.5: Across-property summary of best configurations. Each row lists the highest-performing model per endpoint, with corresponding R^2/RMSE or classification metrics.

5.5 Key Observations and Implications

- **Cheminformatics still dominates:** CSS-only descriptors (Mordred) consistently give the best overall performance for $\log P$ and $\log BB$ classification, thanks to engineered features like SLogP and polar surface area.
- **Hybridization benefits specific tasks:** For $\log S$ and $\log BB$ regression, combining CSS with QM and MWF descriptors improves stability by capturing both molecular energetics and solvent-accessible geometry.
- **Energetics drive $\log P$:** Models including ΔG (octanol–water free energy transfer) show significant gains, reinforcing partitioning as a free-energy-driven phenomenon.
- $\log BB$ **classification > regression:** The classification task achieves high accuracy and AUC, while continuous regression is noisier but improved by hybrid descriptor bundles.
- **Descriptor complementarity:**
  - CSS encodes coarse physicochemical rules (lipophilicity, polarity, atom counts).
  - QM adds electronic reactivity (SCF energy, frontier orbitals, dipole).
  - MWF captures solvent-accessible shape/size and density.

  Their complementarity enhances interpretability and robustness across endpoints.

- **General insight:** While ML models can extract strong predictive signals from classical cheminformatics alone, integrating QM and MWF features enhances mechanistic interpretability and extends applicability to endpoints with higher noise (e.g., $\log S$, $\log BB$ regression).

Chapter 6

Future Work

Based on the observations and performance results in this study, several avenues can be explored to further improve the prediction of pharmacokinetic properties using machine learning integrated with quantum chemistry descriptors.

6.1 Expansion of Descriptor Coverage

Future work could incorporate a broader range of quantum descriptors, including higher-order multipole moments, molecular electrostatic potential (MEP)-derived features, polarizability tensors, entropy terms, and solvation parameters from models such as COSMO-RS or SMD. These additions may capture subtle intermolecular interactions and solvent effects, further refining property predictions.

6.2 Advanced Model Architectures

While this study primarily focused on classical machine learning models, future research can explore more expressive algorithms such as:

- Transformer-based molecular representations (e.g., ChemBERTa, MolFormer)
- Message Passing Neural Networks (MPNNs)
- Ensemble regressors/classifiers with stacked generalization

These methods may capture non-linear descriptor interactions more effectively, particularly for complex endpoints such as $\log BB$.

6.3 Transfer Learning and Multi-task Learning

Transfer learning from large-scale, pre-trained molecular models (e.g., trained on PubChem or ChEMBL) could provide improved model initialization for smaller datasets.

Multi-task learning—training a shared model to predict multiple properties such as $\log P$, $\log S$, and $\log BB$ simultaneously—may improve generalization by leveraging correlations between endpoints.

6.4 Integration with Broader ADMET Profiling

Extending the pipeline to include additional ADMET endpoints—such as toxicity, hERG inhibition, metabolic stability, and general drug-likeness—would make the models more practical for real-world drug discovery workflows.

6.5 Pipeline Automation and Scalability

For large-scale applications, the descriptor generation workflow (Gaussian/ORCA execution, error handling, multiprocessing) should be fully automated. Integration with cloud computing or HPC clusters would enable scaling to datasets with tens of thousands of molecules without prohibitive runtimes.

6.6 Hybrid and Ensemble Descriptor Models

Future studies can investigate hybrid architectures that fuse classical Mordred descriptors, quantum chemical features, and surface/shape descriptors (Multiwfn) into unified feature spaces. Ensemble methods—combining different base learners—may also help achieve a stronger bias–variance trade-off.

6.7 Uncertainty Quantification and Model Interpretability

Incorporating methods for uncertainty estimation (e.g., prediction intervals, Bayesian inference) and interpretability tools such as SHAP values or attention weight analysis would increase transparency and trust in predictive outcomes, particularly for clinical or regulatory use.

In summary, future advancements in descriptor diversity, scalable computation, and advanced learning architectures can significantly enhance predictive accuracy and robustness, making such pipelines central to modern in silico drug discovery.

Chapter 7

Conclusion

This thesis investigated predictive modeling of three key pharmacokinetic properties—lipophilicity ($\log P$), aqueous solubility ($\log S$), and blood–brain barrier permeability ($\log BB$)—by integrating machine learning with complementary molecular descriptor families. We systematically compared:

- **Classical (CSS):** Mordred-generated 2D descriptors,
- **Quantum (QM):** DFT-derived features (Gaussian/ORCA), including frontier orbital energies and thermodynamic terms,
- **MWF:** Surface/shape descriptors (*Multiwfn*) such as polar/non-polar surface areas, density, and volume,
- $\boldsymbol{\Delta G}$: solvent-phase free-energy differences (applied in the $\boldsymbol{\log P}$ pipeline).

Key Findings

- **Permeability / $\log P$:** The *classical-only* model achieved the strongest performance with $R^2 = 0.948$ and RMSE = 0.3847. Physics-based hybrids (CSS + QM + MWF + ΔG) improved interpretability but did not exceed the predictive accuracy of CSS alone.
- **Aqueous Solubility / $\log S$:** The *classical-only* model again performed best ($R^2 = 0.8644$, RMSE = 0.8488). Hybrid models (CSS + QM + MWF + ΔG) were competitive ($R^2 = 0.8432$) and highlighted the value of solvation-relevant features (e.g., polar surface area, molecular volume) over QM-only or MWF-only sets.
- **Blood Brain Barrier Permeability / $\log BB$:** *Classification* ($\text{BBB}^+/\text{BBB}^-$) clearly outperformed regression. The Classical and Quantum hybrid descriptor set delivered the top classification scores (Accuracy = 0.9411, AUC–ROC = 0.9974).

For regression, hybrid model with Classical and Quantum Descriptor also surpassed single-source models ($R^2 = 0.665$ vs. ~ 0.569 for CSS-only and ~ 0.497 for QM-only).

- **Role of Quantum Mecahnical descriptors:** Although computationally heavier, Quantum features enriched mechanistic interpretability (frontier orbital alignment, ionization/electron affinity trends, and ΔG effects) and improved ranking ability (AUC), especially when combined with MWF surface/volume descriptors.

Methodological Contribution

Beyond model benchmarking, this work delivers a reusable pipeline—*descriptor computation* (CSS, QM, MWF, ΔG) $\rightarrow$ *feature selection* $\rightarrow$ *multi-model evaluation*—that is scalable across endpoints and adaptable to new descriptor sources. The framework balances accuracy and computational cost, enabling practical deployment for large-scale screening.

Practical Implications

For early-stage discovery, *classical* descriptors provide fast, high-accuracy baselines for $\log P$ and $\log S$, while *hybrids* (CSS + QM, optionally + MWF/ΔG) are preferred when interpretability and robust ranking (AUC) are critical—particularly for $\log BB$ classification. Multiwfn-derived surface/volume features add solvation- and transport-relevant signal that complements QM detail.

Overall Conclusion

A balanced integration of *classical efficiency*, *quantum-level detail*, and *modern machine learning* yields a powerful, scalable toolset for drug-likeness assessment. Classical descriptors provide robust predictive baselines; quantum and Multiwfn-derived surface/volume features add mechanistic clarity and improve ranking robustness when hybridized. The resulting pipeline advances accuracy, interpretability, and practical usability for early-stage drug discovery.