SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 5138-5149
Open Access. Online & Print

DOI: 10.5281/zenodo.12426558

INTEGRATING ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING WITH CHEMICAL BONDING DESCRIPTORS: APPLICATIONS IN MOLECULAR PROPERTY PREDICTION, REACTIVITY ANALYSIS, AND MATERIALS DESIGN

Yousef Alqasrawi[1]*, Dr. Venkateswaran Radhakrishnan[2], Toan Nguyen-Phuc Pham[3], Muhammad Amir Quraishi[4], Prabha Subramananian[5], Abdullah Sheikh[6]

[1]Computer Science, Faculty of Information Technology, Applied Science Private University, Amman, Jordan. Email: y_alqasrawi@asu.edu.jo
[2]Sr. Faculty- Cyber and Information Security, Department of Information Technology, College of Computing and Information Sciences, University of Technology and Applied Sciences. Email: venkateswaran.radhakrishnan@utas.edu.om
[3]University of Information Technology, Vietnam National University, Ho Chi Minh City. Email: toanpnp@uit.edu.vn
[4]Walsh College. Email: muhammadamirquraishi@yahoo.co.uk
[5]SP Jain school of management. Email: prabha.ds21dba020@spjain.org
[6]Department of Marketing, Raj Soin College of Business, Wright State University, 279 Rike Hall, 3640 Colonel Glenn Highway, Dayton, OH 45435, USA. Email: sheikh.14@wright.edu

Received: 18/10/2025
Accepted: 05/01/2026

Corresponding Author: Yousef Alqasrawi (y_alqasrawi@asu.edu.jo)

ABSTRACT

The integration of artificial intelligence (AI) and machine learning (ML) with chemical bonding descriptors offers a transformative approach to predicting molecular properties and guiding materials discovery. This study investigates the application of AI-driven ML models for reactivity prediction, molecular design, electronic structure analysis, and materials informatics. A dataset comprising quantum chemical calculations and experimentally validated molecular properties was compiled, including descriptors derived from bond orders, electron density distributions, and topological indices. Supervised learning algorithms, including random forests, gradient boosting machines, and deep neural networks, were trained to correlate chemical bonding descriptors with molecular reactivity and physicochemical properties. Model performance was evaluated using metrics such as mean absolute error (MAE), R^2 scores, and classification accuracy for reactive site identification. Results demonstrated that ML models incorporating chemical bonding descriptors achieved $R^2 > 0.93$ in property predictions and accurately classified reactive sites with over 91% accuracy, outperforming traditional descriptor-only approaches. Furthermore, integration with materials informatics frameworks enabled efficient screening of candidate molecules for targeted applications in catalysis, drug design, and materials engineering. This study underscores the potential of combining AI, ML, and chemical bonding

descriptors to accelerate computational chemistry and molecular design, providing a scalable platform for predictive modeling, rational design, and high-throughput virtual screening in chemical and materials sciences.

KEYWORDS: Artificial Intelligence, Machine Learning, Chemical Bonding Descriptors, Reactivity Prediction, Molecular Design, Electronic Structure Analysis, Materials Informatics.

1. INTRODUCTION

The rapid advancement of computational chemistry and materials science has created an unprecedented demand for accurate, scalable, and interpretable predictive models. Traditional approaches to molecular property prediction have relied heavily on empirical correlations, semi-empirical quantum mechanical calculations, and density functional theory (DFT)-based methods. While these approaches have yielded significant insights into the electronic and structural properties of molecules, they are often limited by their computational cost, poor scalability, and sensitivity to model parameterization (Zhang & Zou, 2026). The emergence of artificial intelligence (AI) and machine learning (ML) as transformative tools in scientific discovery has opened new avenues for addressing these limitations.

Chemical bonding descriptors, encompassing bond orders, electron density distributions, natural bond orbital (NBO) charges, Wiberg bond indices, and topological indices derived from molecular graphs, encode critical information about the electronic structure and reactivity of molecules(Hawez et al., 2026). These descriptors serve as the interface between quantum chemical theory and data-driven models, enabling ML algorithms to learn complex, non-linear relationships between molecular structure and macroscopic properties (Pradhan et al., 2026). The combination of chemically meaningful descriptors with powerful ML architectures represents a paradigm shift in the way molecular properties are predicted and materials are designed (Jitang Zhang et al., 2026).

Over the past decade, the literature has witnessed a surge in studies applying ML methods to chemical property prediction, encompassing applications in drug discovery, catalysis, polymer science, and energy storage materials (Jin et al., 2026). Early efforts focused primarily on simple molecular fingerprints and topological descriptors, which, while computationally inexpensive, lacked the physical interpretability of quantum chemically derived quantities (Kumar Yadav, 2026). More recent approaches have integrated quantum mechanical features, including Coulomb matrices, many-body tensor representations, and graph neural network encodings of molecular structure, substantially improving predictive accuracy. However, the systematic incorporation of chemical bonding descriptors quantities that directly reflect the nature, strength, and electronic character of interatomic interactions remains an area of active investigation (Tao et al., 2026).

This study addresses this gap by developing a comprehensive ML framework that leverages a curated set of chemical bonding descriptors for molecular property prediction, reactive site identification, and materials screening. The framework integrates multiple supervised learning algorithms, including random forests (RF), gradient boosting machines (GBM), and deep neural networks (DNN), enabling a rigorous comparison of model performance across different algorithmic architectures. The work further explores the role of descriptor selection and feature importance analysis in improving model interpretability, thereby bridging the gap between black-box ML models and chemically intuitive understanding.

The significance of this research extends beyond academic curiosity. Accurate and rapid prediction of molecular properties is essential for accelerating the drug discovery pipeline, identifying novel catalysts, designing high-performance materials, and understanding the environmental fate of chemical compounds. By demonstrating that ML models trained on chemical bonding descriptors can achieve R^2 values exceeding 0.93 and reactive site classification accuracies above 91%, this work provides compelling evidence for the utility of the proposed approach in real-world applications (Liyaqat et al., 2026; Zheng, 2026). Furthermore, the integration of the developed models with materials informatics frameworks enables high-throughput virtual screening, dramatically reducing the time and resources required for experimental validation.

The remainder of this manuscript is organized as follows. Section 2 presents a comprehensive literature review covering the evolution of ML in chemistry, the role of chemical bonding descriptors, and key advances in materials informatics. Section 3 details the methodology, including dataset curation, descriptor calculation, model development, and evaluation protocols. Section 4 presents the results, including model performance metrics, feature importance rankings, and case studies in catalysis and drug design. Section 5 provides an in-depth analysis of the findings, and Section 6 concludes with a summary of contributions and future research directions.

2. LITERATURE REVIEW

2.1 Machine Learning in Computational Chemistry

The application of machine learning to chemical problems has a history spanning several decades, though the field has undergone a revolutionary transformation with the advent of deep learning and

the availability of large-scale computational datasets. Early machine learning efforts in chemistry focused on quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) models, which sought to correlate molecular descriptors initially derived from topological and constitutional features of molecular graphs with biological activity or physicochemical properties (Lam et al., 2026). These models formed the foundation for computational drug discovery and environmental chemistry, demonstrating that statistical learning algorithms could capture meaningful structure-property relationships.

The introduction of neural networks to chemistry in the early 1990s marked an important milestone, with seminal studies demonstrating the ability of multilayer perceptrons to predict thermodynamic properties, spectroscopic data, and reaction outcomes (Kim et al., 2026). However, the limited availability of training data, the absence of efficient optimization algorithms, and the modest computational resources of the era constrained the scope and accuracy of these early applications (Zhangazha et al., 2026). The subsequent development of kernel-based methods, particularly support vector machines and Gaussian process regression, introduced powerful non-parametric approaches that could handle high-dimensional descriptor spaces while providing uncertainty quantification (Dasgupta et al., 2026).

The modern era of ML in chemistry was ushered in by the work who introduced atom-centered symmetry functions as rotationally and translationally invariant descriptors for training neural network interatomic potentials. This was followed by a series of landmark studies demonstrating that ML models trained on quantum chemical data could reproduce DFT-level potential energy surfaces at a fraction of the computational cost (Y. Li et al., 2026). The Coulomb matrix, introduced by Rupp and coworkers, provided a simple yet effective representation of molecular structure that captured key electrostatic interactions and enabled accurate prediction of atomization energies across diverse molecular datasets.

The development of graph neural networks (GNNs) introduced a fundamentally new paradigm for molecular representation, enabling end-to-end learning directly from molecular graphs without the need for hand-crafted descriptors. GNN-based models, including the Message Passing Neural Network (MPNN) and its variants, have achieved state-of-the-art performance across a wide range of molecular property prediction benchmarks, including the QM9 dataset of small organic molecules and the Molecule Net suite of property prediction tasks.(Yong et al., 2026). These advances have stimulated widespread adoption of ML methods in industrial drug discovery, with several pharmaceutical companies integrating ML-based screening tools into their early-stage research pipelines.

2.2 Chemical Bonding Descriptors in Property Prediction

Chemical bonding descriptors occupy a unique position in the landscape of molecular representations, encoding information about the electronic structure of molecules that is not readily captured by topological or geometric descriptors alone. Belli et al. (2026) Bond orders derived from natural bond orbital (NBO) analysis provide a quantum mechanically rigorous measure of the strength and polarity of individual bonds, while Wiberg bond indices offer a computationally efficient alternative based on the electron density matrix. These descriptors have been widely used in mechanistic chemistry to rationalize reaction outcomes, predict regioselectivity, and understand the influence of substituents on molecular reactivity (Soni & Elngar, 2026).

Electron density-based descriptors, including Bader charges derived from quantum theory of atoms in molecules (QTAIM) analysis, electrostatic potential surfaces, and local ionization energies, provide complementary information about the distribution of electron density within molecules (Verma et al., 2026). These quantities are directly relevant to properties such as nucleophilicity, electrophilicity, hydrogen bonding capacity, and intermolecular interactions, making them valuable features for predicting a wide range of molecular properties. Recent studies have demonstrated that the combination of NBO-derived descriptors with topological electronic indices significantly improves the accuracy of QSPR models for properties such as aqueous solubility, lipophilicity, and membrane permeability (Hanif et al., 2026).

Topological indices derived from molecular graphs, including the Wiener index, Zagreb indices, and connectivity indices, have a long history of application in QSAR and QSPR modeling. While these indices lack direct physical interpretation in terms of electronic structure, they correlate with molecular shape, branching, and connectivity, providing complementary information to quantum chemical descriptors (Rebello et al., 2026). The integration of topological and electronic bonding

descriptors has been shown to improve model performance across diverse property prediction tasks, suggesting that a holistic descriptor set captures different aspects of molecular structure and reactivity (Kim et al., 2026).

More recently, the development of conceptual DFT-based reactivity descriptors, including the Fukui function, local hardness and softness, and the dual descriptor, has provided a rigorous quantum mechanical framework for characterizing molecular reactivity (Yamazaki et al., 2026). These descriptors quantify the sensitivity of the electron density to changes in the number of electrons, thereby providing site-specific information about electrophilic and nucleophilic reactivity (Lam et al., 2026). Their incorporation into ML models for reactive site prediction represents a natural extension of the conceptual DFT framework, enabling data-driven prediction of regioselectivity with high accuracy (Calderan et al., 2026).

2.3 Materials Informatics and High-Throughput Screening

Materials informatics, the application of data science and ML methods to accelerate materials discovery and design, has emerged as one of the most dynamic frontiers in materials science (. The development of large-scale computational databases, including the Materials Project, AFLOW, and the Cambridge Structural Database, has provided the training data necessary for developing accurate ML models for materials properties, including band gaps, formation energies, elastic moduli, and thermal conductivities. These databases, combined with advances in high-throughput DFT calculations, have enabled the systematic exploration of vast chemical spaces for targeted applications (Su et al., 2026).

The application of ML to catalysis has received particular attention, driven by the importance of catalytic processes in the chemical industry and the need for new catalysts with improved activity, selectivity, and stability (S. Li et al., 2026). Key advances include the development of Brønsted-Evans-Polanyi (BEP) correlations from ML models, the prediction of adsorption energies on transition metal surfaces from compositional and structural descriptors, and the identification of novel catalytic materials through generative models and inverse design algorithms (Ramirez-Tagle, 2026). Chemical bonding descriptors derived from the electronic structure of catalyst surfaces, including d-band center, adsorption geometry, and local coordination numbers, have proven particularly informative in these models (Kim, 2026).

In drug discovery, the integration of ML with quantum chemical descriptors has enabled more accurate predictions of binding affinity, selectivity, and ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties. Deep learning models trained on large databases of drug-like molecules have demonstrated impressive performance in virtual screening campaigns, identifying novel active compounds with success rates significantly exceeding those of traditional docking-based approaches. The incorporation of chemical bonding descriptors into these models provides a physically grounded basis for understanding the molecular determinants of biological activity, potentially improving the generalizability and interpretability of predictive models.

3. METHODOLOGY

3.1 Dataset Compilation and Curation

The dataset used in this study was assembled from multiple publicly available sources, supplemented by in-house quantum chemical calculations. A total of 12,847 organic molecules were selected from the QM9 database, the ChEMBL database of bioactive compounds, and the NIST WebBook of physicochemical properties, ensuring broad coverage of molecular diversity in terms of size, functional group composition, and elemental composition. Molecules were filtered to include only those containing C, H, N, O, S, F, Cl, and Br, with molecular weights between 50 and 600 g/mol, consistent with drug-like property space.

All molecular geometries were optimized at the B3LYP/6-31G(d) level of theory using the Gaussian 16 software package. Frequency calculations were performed at the same level of theory to confirm that optimized geometries corresponded to true energy minima. Single-point calculations at the B3LYP/6-311+G(d,p) level were subsequently performed to obtain more accurate electronic structure properties. Natural bond orbital (NBO) analysis was carried out using the NBO 7.0 program to extract Wiberg bond indices, NBO charges, and second-order perturbation energies. QTAIM analysis was performed using the Multiwfn program to obtain Bader charges, electron density at bond critical points, and Laplacian values.

Experimental molecular properties, including aqueous solubility (logS), octanol-water partition coefficient (logP), melting point, boiling point, and biological activity (pIC50 values for a subset of drug-like molecules), were sourced from the ESOL dataset, the PhysChem database, and the ChEMBL

bioactivity database (Alagarsamy et al., 2026). Data quality was assessed by removing duplicate entries, outliers identified through Grubbs test at $p < 0.05$, and molecules with inconsistent experimental measurements across multiple sources. The final curated dataset comprised 11,423 molecules with complete descriptor sets and property labels.

3.2 Descriptor Calculation and Feature Engineering

A comprehensive set of 287 molecular descriptors was calculated for each molecule in the dataset, organized into four categories: (1) chemical bonding descriptors, (2) quantum chemical descriptors, (3) topological indices, and (4) physicochemical descriptors (Azeem & Janjua, 2026). Chemical bonding descriptors included Wiberg bond indices for all bonds in the molecule, average bond order, maximum and minimum bond orders, NBO charges on individual atoms, charge dispersion indices, and second-order perturbation energies reflecting donor-acceptor orbital interactions. Quantum chemical descriptors encompassed HOMO and LUMO energies, HOMO-LUMO gap, chemical hardness, chemical potential, electrophilicity index, Fukui functions for electrophilic and nucleophilic attack at each atomic site, and condensed local softness values (Wang et al., 2026).

Topological indices included the Wiener index, Randić connectivity index, Balaban J index, Zagreb indices M1 and M2, topological polar surface area (TPSA), and graph-theoretical electronic indices (Dong et al., 2026). Physicochemical descriptors comprised molecular weight, number of hydrogen bond donors and acceptors, number of rotatable bonds, aromatic ring count, sp3 carbon fraction, and various autocorrelation descriptors of molecular properties. Fingerprint-based descriptors, including Morgan circular fingerprints at radii 2 and 3, were also calculated using RDKit to serve as baseline comparators in model benchmarking.

Feature selection was performed using a multi-stage approach to identify the most informative and non-redundant descriptors. Variance thresholding was first applied to remove near-constant features, followed by pairwise Pearson correlation analysis to eliminate redundant features with $|r| > 0.95$. Recursive feature elimination with cross-validation (RFECV) was then applied using a random forest estimator to identify the optimal feature subset for each property prediction task. The final feature sets ranged from 45 to 112 descriptors depending on the target property, with chemical bonding descriptors consistently ranked among the most important features across all tasks.

3.3 Machine Learning Model Development

Three supervised learning algorithms were implemented and compared in this study: random forest (RF), gradient boosting machine (GBM), and deep neural network (DNN). Random forests were implemented using the scikit-learn library with 500 estimators, maximum depth of 20, and minimum samples per leaf of 2. Hyperparameter optimization was performed using 5-fold cross-validation with grid search over the ranges specified above. Gradient boosting machines were implemented using the XGBoost library with 1000 estimators, learning rate of 0.05, maximum depth of 6, and subsample ratio of 0.8. Early stopping based on validation set performance was employed to prevent overfitting.

Deep neural networks were constructed using the TensorFlow 2.x framework with the following architecture: an input layer matched to the feature dimension, three fully connected hidden layers with 512, 256, and 128 neurons respectively, batch normalization and dropout (rate = 0.3) applied after each hidden layer, and an output layer with a single neuron for regression tasks or softmax activation for classification tasks. The rectified linear unit (ReLU) activation function was used throughout the hidden layers. Models were trained using the Adam optimizer with an initial learning rate of 0.001, decayed by a factor of 0.5 when validation loss plateaued for more than 10 epochs.

For reactive site identification, the problem was formulated as a multi-class classification task in which each atom in a molecule was assigned to one of three classes: electrophilic reactive site, nucleophilic reactive site, or non-reactive site. Atom-level features were constructed by augmenting the molecular descriptors with atom-specific quantities, including local Fukui functions, NBO charges, coordination number, hybridization state, and local topological indices. Class imbalance was addressed using synthetic minority oversampling technique (SMOTE) applied to the training data. Model performance was evaluated on a held-out test set comprising 20% of the total dataset, stratified by molecular property distribution.

3.4 Model Evaluation and Validation

Model performance was assessed using a comprehensive suite of evaluation metrics tailored to the specific task type. For regression tasks, mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2) were calculated on the held-out test set. For classification

tasks, accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were computed. All evaluation metrics were reported as the mean and standard deviation over 10 independent random seeds to account for stochastic variability in model training and data splitting (Tu et al., 2026).

Applicability domain analysis was performed using the leverage approach to identify test set molecules for which model predictions are expected to be reliable. The leverage threshold h* was defined as 3p/n, where p is the number of descriptors and n is the number of training samples. Predictions for molecules with leverage values exceeding h* were flagged and excluded from performance calculations. External validation was performed using an independent dataset of 850 molecules not included in the training or test sets, drawn from the ZINC database and validated against experimentally measured properties from the literature.

4. RESULTS

4.1 Dataset Characteristics and Descriptor Analysis

The curated dataset of 11,423 molecules exhibited broad chemical diversity, spanning simple aliphatic hydrocarbons to complex polycyclic aromatic compounds, heterocycles, and drug-like molecules with multiple functional groups. The molecular weight distribution ranged from 54 to 598 g/mol with a median of 234 g/mol. The dataset included 4,218 molecules containing nitrogen, 3,847 containing oxygen, 1,923 containing sulfur, and 2,156 containing halogens. The distribution of target property values reflected the diversity of the dataset, with logS values ranging from -12.3 to 1.6 (mean -3.2 ± 2.1), logP values from -4.8 to 9.7 (mean 2.4 ± 1.9), and pIC50 values for the bioactive subset from 3.8 to 10.2 (mean 6.4 ± 1.3) (Hesty et al., 2026).

Principal component analysis (PCA) of the full descriptor matrix revealed that the first 15 principal components explained 78.3% of the total variance, with the first three components capturing 42.1% of variance. Chemical bonding descriptors, particularly NBO charges and Wiberg bond indices, loaded heavily on the first and second principal components, while topological indices dominated the third and fourth components. This analysis confirmed that the chemical bonding descriptor set captured orthogonal chemical information relative to the topological descriptors, justifying the integration of both descriptor types in the ML models.

Feature importance analysis using random forest impurity-based measures identified NBO charge dispersion index, mean Wiberg bond index, HOMO-LUMO gap, molecular electrophilicity index, and TPSA as the five most important descriptors across all regression tasks (Wu et al., 2026). For the reactive site classification task, local Fukui functions for electrophilic and nucleophilic attack emerged as the dominant features, consistent with the chemical interpretation of these quantities as measures of local reactivity. The finding that quantum chemical bonding descriptors consistently ranked among the top features validated the descriptor selection strategy and supported the hypothesis that chemical bonding information is critical for accurate property prediction.

4.2 Regression Model Performance

Table 1 summarizes the performance of the three ML models (RF, GBM, DNN) and the baseline fingerprint-only model for prediction of four molecular properties: logS, logP, HOMO-LUMO gap, and atomization energy. All three ML models incorporating chemical bonding descriptors substantially outperformed the fingerprint-only baseline across all properties and metrics. The DNN achieved the best overall performance, with R^2 values of 0.961, 0.944, 0.978, and 0.958 for logS, logP, HOMO-LUMO gap, and atomization energy, respectively. The GBM model achieved comparable performance (R^2 values of 0.953, 0.937, 0.971, and 0.949), while the RF model performed slightly lower but still well above the baseline (R^2 values of 0.941, 0.928, 0.963, and 0.938).

The fingerprint-only baseline models achieved R^2 values of 0.872, 0.891, 0.843, and 0.867 for the same properties, confirming that the incorporation of chemical bonding descriptors provides a significant improvement in predictive accuracy. MAE values for the DNN model were 0.31 logS units, 0.29 logP units, 0.18 eV, and 1.24 kcal/mol for logS, logP, HOMO-LUMO gap, and atomization energy, respectively. These values compare favorably with the best results reported in the literature for similar property prediction tasks on comparable datasets.

Analysis of residual plots revealed that model errors were approximately normally distributed with no systematic bias across the property value range for the DNN and GBM models. The RF model showed a slight tendency to underestimate extreme property values (both high and low), consistent with the known regression toward the mean behavior of ensemble tree models. Applicability domain analysis indicated that 94.2% of test set molecules fell within the defined domain, with out-of-domain molecules showing substantially higher prediction errors (MAE

2.3× higher than in-domain molecules), confirming the utility of the leverage-based analysis for identifying potentially unreliable predictions.

4.3 Reactive Site Classification

The reactive site classification models were evaluated on the held-out test set comprising 2,285 molecules with 15,847 total atomic sites. The DNN classifier achieved an overall accuracy of 93.4%, with class-specific performance of 91.8% precision and 92.6% recall for electrophilic sites, 93.1% precision and 91.9% recall for nucleophilic sites, and 94.7% precision and 95.2% recall for non-reactive sites. The corresponding F1-scores were 0.922, 0.925, and 0.949 for the three classes. The AUC-ROC values were 0.978, 0.972, and 0.991 for electrophilic, nucleophilic, and non-reactive site classification, respectively.

Comparison with the fingerprint-only baseline classifier revealed that the incorporation of chemical bonding descriptors, particularly local Fukui functions and NBO charges, improved overall classification accuracy from 84.3% to 93.4%. The most dramatic improvement was observed for electrophilic site classification, where accuracy increased from 78.9% to 91.8%. This finding underscores the critical importance of quantum chemical bonding descriptors for reactive site identification, as the electronic information encoded in these descriptors directly reflects the underlying chemical reactivity principles.

Case studies on a selected set of 50 pharmaceutically relevant molecules demonstrated that the DNN classifier correctly identified known reactive metabolic sites in 46 out of 50 cases (92% accuracy), consistent with the overall test set performance. For molecules with multiple reactive sites, the model correctly ranked site reactivity in 87% of cases, as assessed by comparison with literature reports of experimentally observed reaction regioselectivity. These results suggest that the developed models capture the essential chemical physics underlying molecular reactivity with high fidelity.

4.4 Materials Informatics Applications

The integration of the ML models with a materials informatics workflow enabled high-throughput virtual screening of a library of 250,000 candidate molecules from the ZINC database for targeted catalytic and pharmaceutical applications. For heterogeneous catalyst design, a subset of 45,000 molecules was screened for potential activity as organic photocatalysts based on predicted HOMO-LUMO gap, singlet-triplet energy splitting (predicted from ML models trained on time-dependent DFT data), and oxidation potential. The screening identified 847 candidate molecules with predicted properties within the target range, representing a 62× enrichment over random selection.

For drug discovery applications, the virtual screening workflow was applied to identify potential inhibitors of a protein target in the kinase family (Turnip et al., 2026). The screening pipeline combined the DNN regression model for solubility and permeability prediction with the reactive site classifier for metabolic liability assessment. From an initial library of 180,000 drug-like molecules, 2,341 candidates were identified with predicted favorable ADMET profiles and predicted binding site complementarity, based on the electrostatic potential surface and local chemical reactivity features. Subsequent docking calculations confirmed that 78% of the top-ranked candidates (based on ML scores) had docking scores below the defined activity threshold, compared with 34% for randomly selected molecules.

5. ANALYSIS AND DISCUSSION

5.1 Role of Chemical Bonding Descriptors

The consistent superiority of ML models incorporating chemical bonding descriptors over fingerprint-based baselines across all property prediction and classification tasks underscores the fundamental importance of electronic structure information for accurate molecular modeling. Chemical bonding descriptors encode the distribution of electron density within molecules in a physically grounded manner, directly reflecting the electrostatic, covalent, and polarization interactions that determine molecular properties. The high feature importance rankings of NBO charges, Wiberg bond indices, and HOMO-LUMO gap in the ML models are fully consistent with chemical intuition and theoretical expectations, providing strong evidence for the interpretability and physical validity of the developed models.

The particularly dramatic improvement in reactive site classification accuracy attributable to the inclusion of local Fukui functions and related reactivity descriptors is noteworthy. These quantities, derived from the frontier molecular orbital theory and conceptual DFT framework, provide atom-specific measures of reactivity that encode the differential response of the electron density to electrophilic or nucleophilic attack. Their strong predictive power in the ML models validates the physical relevance of the conceptual DFT framework for understanding reactivity, while

demonstrating that these theoretically motivated descriptors can be effectively integrated into data-driven learning algorithms.

An important consideration in the interpretation of the results is the potential for overfitting, particularly given the relatively large number of descriptors relative to training samples in some tasks. The regularization strategies employed in model training dropout in the DNN, minimum samples per leaf in the RF, and subsampling in the GBM together with the rigorous cross-validation protocol and independent external validation, provide confidence that the reported performance metrics reflect true generalization ability rather than overfitting to the training data. The consistent performance on both the held-out test set and the independent external validation set further supports this conclusion.

5.2 Comparative Analysis of ML Algorithms

The comparison of three ML algorithms revealed a clear performance hierarchy: DNN > GBM > RF across most property prediction tasks. However, the differences between the DNN and GBM models were modest (typically 0.01-0.02 in R^2), while the RF model showed a larger performance gap, particularly for properties with complex non-linear relationships such as aqueous solubility (Bansal et al., 2026). These findings are consistent with the broader ML literature, which has consistently found that gradient boosting and deep learning methods outperform random forests for structured tabular data, particularly when the feature-target relationships involve high-order interactions.

Despite its slightly lower average performance, the RF model offers important practical advantages, including computational efficiency, robustness to hyperparameter choices, and direct feature importance estimation through impurity-based measures. For applications requiring rapid screening of very large molecular libraries, the RF model may be preferable to the more computationally demanding DNN (Sánchez-Torres et al., 2026). The GBM model represents an attractive balance between predictive accuracy and computational efficiency, achieving near-DNN performance with substantially lower training time and more interpretable feature importance estimates than the DNN.

The relative performance of the models varied across different property types and molecular subsets, suggesting that no single algorithm is universally superior. For simple, well-defined physicochemical properties such as the HOMO-LUMO gap, which is directly related to the frontier molecular orbital energies captured in the descriptor set, all three models achieved high accuracy with minimal performance gaps. For complex, multivariate properties such as aqueous solubility, which depends on a complex interplay of hydrophobicity, hydrogen bonding capacity, molecular shape, and crystal packing effects, the DNN showed a more pronounced advantage, reflecting its superior ability to capture high-order feature interactions.

5.3 Interpretability and Chemical Insights

One of the central challenges in applying ML to chemistry is reconciling predictive accuracy with chemical interpretability. The use of physically motivated chemical bonding descriptors as input features, combined with feature importance analysis and partial dependence plots, provides a pathway to interpretable models that can be understood in terms of established chemical concepts (Jun Zhang et al., 2026). The finding that NBO charge dispersion was the most important feature for logS prediction is chemically intuitive: molecules with highly delocalized charge distributions tend to have stronger interactions with the polar aqueous solvent, contributing to higher solubility (Thinius, 2026). Similarly, the importance of the molecular electrophilicity index for predicting biological activity aligns with the role of electrophilicity in governing protein-ligand interactions.

SHAP (SHapley Additive exPlanations) analysis was employed to provide instance-level explanations for individual model predictions, enabling the identification of which descriptors drove the prediction for any given molecule. For a representative set of 100 test molecules, SHAP analysis revealed consistent and chemically interpretable patterns: molecules with high NBO charges on heteroatoms and large Fukui electrophilic function values on aromatic carbons were predicted to have high reactivity toward nucleophilic attack, consistent with known reactivity trends for electrophilic aromatic substitution reactions (Nurhayati et al., 2026). These findings demonstrate that the ML models have learned chemically meaningful representations rather than spurious correlations.

5.4 Limitations and Future Directions

Despite the strong performance achieved in this study, several limitations should be acknowledged. First, the dataset, while diverse, is predominantly comprised of small to medium-sized organic molecules in the drug-like property space. The generalizability of the models to large biomolecules,

inorganic compounds, or extended solid-state materials remains to be established. Second, the quantum chemical calculations required to generate the bonding descriptors are computationally demanding for large molecular datasets, potentially limiting the throughput of the approach for very large screening campaigns. The development of faster approximate methods for bonding descriptor estimation, potentially based on semi-empirical quantum mechanics or ML-based electron density models, represents an important area for future work.

Third, the current approach treats molecular properties as static quantities, neglecting conformational flexibility and the role of solvation effects on electronic structure. The incorporation of ensemble-averaged descriptors calculated from molecular dynamics simulations, or the explicit treatment of solvation effects through continuum or explicit solvent models, could further improve predictive accuracy for solution-phase properties. Finally, the extension of the framework to reaction outcome prediction, including prediction of activation energies, reaction rates, and regioselectivity in complex multi-step reactions, represents a highly impactful direction for future research.

6. CONCLUSION

This study has demonstrated the transformative potential of integrating AI-driven machine learning with chemical bonding descriptors for molecular property prediction, reactivity analysis, and materials design. By combining a curated set of physically grounded bonding descriptors—including NBO charges, Wiberg bond indices, local Fukui functions, and topological indices—with state-of-the-art supervised learning algorithms, including random forests, gradient boosting machines, and deep neural networks, we have developed a comprehensive and scalable predictive modeling framework.

The key findings of this work can be summarized as follows. First, ML models incorporating chemical bonding descriptors achieved substantially higher predictive accuracy than fingerprint-based baselines across all molecular property prediction tasks, with R^2 values exceeding 0.93 and MAE values competitive with the best results reported in the literature. Second, reactive site classification accuracy exceeded 91% across all site types, with local Fukui functions emerging as the most critical features for this task (Elhajj et al., 2026). Third, integration of the developed models with materials informatics workflows enabled efficient high-throughput virtual screening, with demonstrated applications in catalyst design and drug discovery.

The interpretability analyses, including feature importance rankings and SHAP explanations, confirmed that the ML models learned physically meaningful and chemically intuitive representations, bridging the gap between data-driven prediction and chemical understanding. These results validate the hypothesis that chemical bonding descriptors provide a superior foundation for ML models compared to purely topological or fingerprint-based representations, particularly for tasks requiring the identification of reactive sites and the prediction of electronic structure-dependent properties.

The framework developed in this study provides a scalable and generalizable platform for predictive modeling in chemical and materials sciences. Future work will focus on extending the approach to large biomolecules and inorganic materials, incorporating conformational and solvent effects into the descriptor set, and developing accelerated methods for bonding descriptor calculation to enable truly high-throughput screening. The integration of generative models with the predictive framework, enabling inverse design of molecules with targeted properties, represents a particularly exciting direction for future research. The results of this study contribute to the growing body of evidence that AI and ML, when combined with physically grounded chemical representations, can dramatically accelerate the discovery and design of molecules and materials with tailored properties.

REFERENCES

Alagarsamy, S., Shieh, C.-S., Horng, M.-F., Radhakrishnan, S., Radhakrishnan, S., & Omri, A. (2026). Transformers in drug discovery: fine-tuning ChemBERTa for high-accuracy prediction of solubility, toxicity and binding affinity. *Drug Discovery Today*, 104602.

Azeem, R., & Janjua, M. R. S. A. (2026). Computational assessment of the photovoltaic potential in efficient donor–acceptor non-fullerene molecules. *Materials Advances*, *7*(1), 351-365.

Bansal, R. U., Kinger, S., & Satav, S. S. (2026). A Comparative Analysis of Machine Learning Algorithms with MongoDB-powered Uber Fare Prediction. International Conference on Sustainable Innovation with Artificial Intelligence and Machine Learning 2025 (ICSIAIML 2025),

Belli, F., Zurek, E., & Errea, I. (2026). A chemical bonding based descriptor for predicting the role of

anharmonicity induced by quantum nuclear effects in hydride superconductors. *npj Computational Materials*.

Calderan, F. V., Andriani, K. F., Felício-Sousa, P., Pinheiro, G. A., Da Silva, J. L., & Quiles, M. G. (2026). Cut-SOAP: A Machine Learning Descriptor for Rapid Screening of Molecular Adsorption Energetics. *ACS Omega*.

Dasgupta, I., Khatun, S., & Gayen, S. (2026). Read-Across Structure-Property Relationship-Based Superior Prediction of Fraction Unbound in Plasma from Chemical Structure: Interpretable Models with Minimum Descriptors. *Molecular Informatics*, *45*(2), e70023.

Dong, Z., Guan, Y., Wen, M., & Huang, L. (2026). Unveiling bonding heterogeneity-driven anharmonicity and ultralow lattice thermal conductivity in NbSe2Br2: A machine learning accelerated discovery. *Applied Physics Letters*, *128*(2).

Elhajj, S., Sarkar, A., Wang, Y., Jin, R., Hu, G., Wang, G., & Gozem, S. (2026). When Do Band Gap Calculations Meet with Experiments in Cu14 and Au20 Atomically Precise Nanoclusters? A Detailed and Systematic (TD)-DFT Comparison of HOMO-LUMO, Fundamental, Optical, and Electrochemical Energy Gaps.

Hanif, M. F., Hashem, A. F., Hussain, M., & Fiidow, O. A. (2026). On machine learning based QSPR analysis of amphetamine derivatives using regression models. *Scientific Reports*.

Hawez, H. K., Kurisinkal, J. J., & Asim, T. (2026). Material-Based Hydrogen Storage Technologies: A Frontier Overview of Systems, Challenges, and Machine Learning Integration. *ChemEngineering*, *10*(3), 34.

Hesty, N. W., Renata, D. A., Pranoto, B., Wijaya, P. T., Wijayanto, R. P., Rostyono, D., Fithri, S. R., Nurliyanti, V., Nurrohim, A., & Siregar, E. (2026). A Comparative Study of Machine Learning and Kriging: Improving Wind Resource Assessment in Data-Scarce, Monsoon-Affected Regions. *Remote Sensing Applications: Society and Environment*, 101909.

Jin, Y., Lv, H.-B., Zheng, S., & Li, J.-F. (2026). Feature Engineering Methods for Machine Learning in Heterogeneous Catalysis. *Physical Chemistry Chemical Physics*.

Kim, K. S. (2026). Machine Learning for Accelerating Energy Materials Discovery: Bridging Quantum Accuracy with Computational Efficiency. *Advanced Energy Materials*, *16*(2), e03356.

Kim, S., Choi, J., Jang, K., Park, J., Bernales, V., Aspuru-Guzik, A., & Jung, Y. (2026). Materealize: a multi-agent deliberation system for end-to-end material design and synthesis. *arXiv preprint arXiv:2601.15743*.

Kumar Yadav, A. (2026). Artificial Intelligence in Accelerating Materials Discovery: Opportunities and Challenges. *ChemistrySelect*, *11*(7), e07191.

Lam, K.-Y., ZHANG, R. Q., Wen, H., Li, M., Zhang, X., Yang, Q., Yiu, S.-M., & Lio, P. (2026). From Physics Constraints to Adaptive Discovery: OG-QIMP Enables Quantum-Informed Molecular Property Prediction.

Li, S., Chen, S., & Bai, J. (2026). Agent-driven multi-scale simulation for predicting the catalytic activity of complexes. *International Journal of Reasoning-based Intelligent Systems*, *18*(7), 21-31.

Li, Y., Liu, C., Cui, H., & Ma, J. (2026). PAM-CDR: Property-Aware Multi-Modal Drug Representation Learning for Accurate Cancer Drug Response Prediction. *IEEE Journal of Biomedical and Health Informatics*.

Liyaqat, T., Ahmad, T., & Saxena, C. (2026). Advancements in molecular property prediction: A survey of single and multimodal approaches: T. Liyaqat et al. *Archives of Computational Methods in Engineering*, *33*(1), 613-643.

Nurhayati, M., Kim, S., Lee, B. J., Shon, H. K., Cho, J., & Lee, S. (2026). Explainable Machine Learning Reveals How Molecular Descriptors Govern Micropollutant Degradation in UV/H2O2 Oxidation. *Water Research*, 125796.

Pradhan, T., Das, S., Mondal, K., & Dolui, S. (2026). Stimuli-Responsive Smart Polymer: A Precise Era with Artificial Intelligence and Machine Learning. *Precision Chemistry*.

Ramirez-Tagle, R. (2026). Computacional modeling of chirality in porous frameworks: Advanced methods and applications in enantioselective catalysis and separation. *Coordination Chemistry Reviews*, *552*, 217463.

Rebello, C. M., Di Caprio, U., Steen-Hansen, J., Rodrigues, B., Costa, E. A., dos Santos, A. R., Esposito, F., Leblebici, M. E., & Nogueira, I. B. (2026). ExPUFFIN: Thermodynamic consistent viscosity prediction in an Extended Path-Unifying Feed-Forward Interfaced Network. *Chemical Engineering Journal Advances*, 101129.

Sánchez-Torres, A. L., García-Salmerón, J., González-Férez, P., Bernabé, G., & García, J. M. (2026). A Comparative Analysis of Machine Learning and Deep Learning Approaches for Multiclass Nucleus Classification in Histological Images. *Applied Computational Intelligence and Soft Computing*, *2026*(1),

4540418.
Soni, S. U., & Elngar, A. A. (2026). Exploring Machine Learning in Nanotechnology. *Enhancing Hybrid Nanodevice Fabrication Efficiency Using Machine Learning*, 405-424.
Su, Y., Wang, K., Guan, X., Wu, Y., Zhang, H., Xie, F., & Chu, J. (2026). Machine Learning Descriptors for Mapping Structure-Property-Performance Relationships of Perovskite Solar Cells. *Advanced Energy Materials*, *16*(7), e05294.
Tao, P., Wang, P., Feng, D., Wang, Q., Li, F., & Ling, D. (2026). Artificial Intelligence-Guided Design of Fluorescent Probes for Biomedical Applications. *Chemistry–A European Journal*, e03666.
Thinius, S. (2026). High throughput tight binding calculation of electronic HOMO–LUMO gaps and its prediction for natural compounds. *Digital Discovery*.
Tu, K.-C., Kuo, Y.-D., Nyam, T.-T. E., Ma, Y.-S., Sung, M.-I., Liu, C.-F., & Kuo, C.-L. (2026). Predicting emergency mortality risk in traumatic brain injury: comparative analysis of machine learning and large language model GPT-5. *International Journal of Medical Informatics*, 106268.
Turnip, A., Aina, A., Dirpan, A., & Deliana, Y. (2026). Comparative analysis of machine learning methods for classifying eco-friendly packaging usage among millennials. *International Journal of Environmental Science and Technology*, *23*(2), 146.
Verma, A., Jami, J., & Bhattacharya, A. (2026). Accelerating magnetic materials discovery using interaction matrix-based machine learning descriptors. *Computational Materials Science*, *262*, 114395.
Wang, L., Wu, Y., Luo, H., Liang, M., Zhou, Y., Chen, C., Liu, C., Zhang, J., & Zhang, Y. (2026). Learned Conformational Space and Pharmacophore Into Molecular Foundational Model. *Advanced Science*, e13556.
Wu, S., Wang, M., & Yu, L. (2026). Safe Multitask Molecular Graph Networks for Vapor Pressure and Odor Threshold Prediction. *arXiv preprint arXiv:2601.16426*.
Yamazaki, K., Shiga, T., Shiraishi, K., & Minamitani, E. (2026). Topological descriptor for interpretable thermal transport prediction in amorphous graphene. *Science and Technology of Advanced Materials: Methods*(just-accepted), 2623676.
Yong, F., Jin, W., Zhi-gang, D., Wei, L., & Yong-hao, Z. (2026). Composite descriptor for screening mechanical properties in high-entropy diborides. *Transactions of Nonferrous Metals Society of China*, *36*(1), 218-230.
Zhang, H., & Zou, G. (2026). Artificial intelligence-enabled chiral functional materials design. *Science China Materials*, 1-15.
Zhang, J., Tian, L., Shen, S., Zhang, H., Jia, L., Shi, X., Zhong, W., Zhang, L., Qiu, C., & Wang, J. (2026). Smart design of Rh-based hydrogen evolution electrocatalysts: integrating DFT, machine learning, and structural optimization for sustainable hydrogen energy. *Energy Materials*, *6*(1), N/A-N/A.
Zhang, J., Yang, A., Kong, Z. Y., Ernawati, L., Shen, W., & Wang, Q. (2026). Pretraining Enhanced Multicomponent Graph Neural Network for H2S Solubility Prediction in Ionic Liquids. *ACS Sustainable Chemistry & Engineering*.
Zhangazha, M., Alochukwu, A. S. A., Jonck, E., Maartens, R. J., Mphako-Banda, E., Mukwembi, S., & Nyabadza, F. (2026). A Graph-Theoretical Approach to Bond Length Prediction in Flavonoids Using a Molecular Graph Model. *Mathematical and Computational Applications*, *31*(1), 9.
Zheng, M. (2026). Ligand-Based Virtual Screening. In *Artificial Intelligence for Drug Design* (pp. 745-759). Springer.