Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HerbGNN: A Directed Message Passing Neural Network Benchmark for Early Enrichment in Herbicide Virtual Screening under Extreme Class Imbalance

Tonatiu Campos-García

tocampos_pos@ciatej.edu.mx

SECIHTI-CIATEJ

Juan Vazquez-Martínez

TecNM/ITS

Huitzimengari Campos-García

Colegio de Postgraduados

Research Article

Keywords: Graph neural network, Herbicide virtual screening, Class imbalance, Benchmarking, Early enrichment, Cheminformatics, Scaffold-stratified cross-validation, D-MPNN, Precision-recall, Applicability domain

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9132922/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Virtual screening in agrochemical research presents distinct cheminformatics challenges that remain underrepresented in benchmarking studies, most notably the extreme class imbalance inherent to curated herbicide datasets and the lack of systematic comparisons between modern deep learning architectures and classical machine learning baselines under realistic screening conditions. Here, we introduce HerbGNN, a directed message passing neural network (D-MPNN) framework for herbicide activity prediction, and present a comprehensive benchmark against Random Forest and XGBoost. The models were evaluated on a dataset comprising 363 annotated herbicidal actives from the HRAC 2026 classification and 50,000 structurally diverse background molecules from ChEMBL, resulting in an approximate class imbalance ratio of 1:140. We show that conventional discrimination metrics such as AUROC can provide misleading estimates of model performance under such conditions and argue that the area under the precision-recall curve (AUPRC) and the enrichment factor at 1% (EF1%) offer more informative evaluation criteria for early enrichment tasks. To mitigate scaffold bias and data leakage, all models were assessed using scaffold-stratified cross-validation based on Bemis-Murcko decomposition with bootstrap confidence intervals. HerbGNN achieved a mean AUROC of 0.956 ± 0.05, a mean AUPRC of 0.573 ± 0.19, and a mean EF1% of 66.6, corresponding to a 41% improvement in early enrichment over Random Forest and an approximately 80-fold enrichment relative to random selection. Applicability domain analysis and atomic importance mapping further indicate that the model captures chemically meaningful patterns associated with herbicidal activity.

Scientific Contribution

This work establishes a methodological benchmarking framework for machine learning in herbicide virtual screening under extreme class imbalance. By systematically comparing graph neural networks with descriptor-based machine learning baselines under scaffold-stratified validation, the study identifies key evaluation design decisions, particularly the use of precision-recall metrics and early enrichment criteria, that are critical for reliable performance assessment in imbalanced chemical datasets. These findings provide methodological guidance for cheminformatics benchmarking and are directly transferable to other data-limited chemical discovery problems.

Introduction

The global challenge of ensuring food security for a growing population is increasingly compromised by the rapid evolution of weed resistance to existing herbicides. Since the introduction of synthetic agrochemicals, chemical weed control has relied on a limited number of biochemical modes of action (MoA), such as the inhibition of acetolactate synthase (ALS) and protoporphyrinogen oxidase (PPO) [1, 2]. However, the overreliance on these targets has led to the emergence of resistant biotypes in more than 260 species worldwide, creating an urgent need for herbicides with novel mechanisms and improved selectivity [3, 4]. Addressing this challenge requires not only the discovery of new chemical entities but also a deeper understanding of molecular interactions underlying herbicide activity and

resistance [5, 6]. Traditional herbicide discovery pipelines, which rely heavily on high-throughput screening (HTS) and empirical field trials, are associated with high costs, long development timelines, and declining success rates [7, 8, 12]. Current estimates indicate that bringing a new active ingredient to market requires investments exceeding $300 million and over a decade of development, while only a small fraction of screened compounds ultimately reach commercialization [9]. These limitations have driven the adoption of computational approaches, including ligand-based and structure-based virtual screening, to reduce experimental burden and focus resources on promising candidates [10, 11].

In this context, machine learning (ML) methods have become widely used in cheminformatics for modeling structure-activity relationships. Classical algorithms such as Random Forest (RF) and gradient boosting methods, including XGBoost (XGB), have demonstrated strong performance in QSAR modeling and compound classification tasks due to their robustness, scalability, and ability to capture non-linear relationships in molecular descriptor space [97, 101, 102]. These approaches, typically based on fixed molecular representations such as fingerprints, are particularly effective at achieving high global discriminative performance. However, their reliance on predefined descriptors can limit their ability to generalize across structurally diverse scaffolds and to capture subtle, higher-order chemical patterns embedded in molecular graphs [17, 22]. In recent years, artificial intelligence (AI) and deep learning (DL) approaches have emerged as powerful alternatives, enabling direct learning from molecular structures and facilitating exploration of vast chemical spaces [10, 11]. Graph-based neural networks have shown strong performance in molecular property prediction by representing molecules as graphs of atoms and bonds and learning task-specific representations through message passing [16, 17]. Among these, Directed Message Passing Neural Networks (D-MPNN) have demonstrated improved capability to capture chemically relevant substructures and long-range dependencies compared to traditional descriptor-based methods [17]. Despite these advances, several challenges remain. First, many ML and DL models are evaluated primarily using ROC-based metrics, which can overestimate performance in highly imbalanced datasets, a common characteristic of herbicidal screening campaigns [32, 103]. Second, herbicidal datasets typically exhibit extreme class imbalance, where active compounds represent only a small fraction of the chemical space, leading to models that perform well globally but fail to prioritize true positives at early ranks [19–21, 48–50]. Third, both classical ML and DL models are often criticized for limited interpretability, which hinders their adoption in chemical research where mechanistic understanding is essential [17, 18].

Despite the widespread adoption of GNN-based frameworks in pharmaceutical drug discovery [16, 18], their application to agrochemical research, and specifically to herbicide virtual screening, remains largely unexplored. To date, the most comprehensive cheminformatics study of herbicide chemical space relied exclusively on classical machine learning approaches such as Random Forest [16], while recent reviews on AI-driven agrochemical discovery have identified graph-based deep learning as an emerging but underexplored strategy in this domain [11]. No prior work has reported the application of a D-MPNN framework to herbicide activity prediction under realistic imbalanced screening conditions, representing a critical gap in the current literature. Bridging this gap is particularly timely given the convergence of

expanding computational resources, curated herbicide databases such as the HRAC classification system, and the growing urgency for novel phytotoxic scaffolds to address herbicide resistance.

In this study, we present HerbGNN, a graph-based deep learning framework designed to address these limitations by integrating robust predictive performance, early enrichment capability, and mechanistic interpretability. Built upon a Directed Message Passing Neural Network (D-MPNN) architecture and optimized using a class-balanced focal loss, HerbGNN is specifically tailored to operate under severe class imbalance and to prioritize active compounds within large chemical libraries. Importantly, the model is evaluated against established baselines, including Random Forest and XGBoost, using identical scaffold-based splits to ensure a fair and rigorous comparison under realistic virtual screening conditions. Beyond predictive performance, the framework incorporates multiple layers of validation, including error distribution analysis, calibration assessment, and applicability domain characterization, to evaluate model reliability across chemical space. Furthermore, the predictive logic of HerbGNN is interrogated through targeted quantum chemical calculations on representative candidates, linking model-derived importance patterns to electronic descriptors such as frontier molecular orbitals and electrostatic potential distributions. This integration of data-driven learning with physically grounded validation aligns with emerging strategies that combine machine learning with quantum chemical insights to enhance interpretability and extrapolative power [10, 11, 14, 18]. In doing so, we establish a methodological reference framework for machine learning benchmarking in herbicide virtual screening and identify key evaluation design decisions with direct implications for the broader cheminformatics community working in data-limited chemical domains.

Results and Discussion

Chemical Dataset Architecture and MoA Composition

The curated dataset used for HerbGNN was designed to reflect the structural landscape of commercial herbicides while reproducing the extreme class imbalance (~ 1:140) typical of early-stage discovery pipelines [19, 22]. As summarized in Table 1, the active population represents a small fraction of the total chemical universe, comprising 363 compounds against a background of 50,000 molecules. As shown in Fig. 1A, the active subset spans multiple Modes of Action (MoA) according to the HRAC classification [23]. Among these, acetolactate synthase (ALS) inhibitors represent the most populated class, consistent with their historical dominance and commercial success since the late twentieth century [2, 4, 5, 9].

Table 1
Summary of training dataset composition.

Component	Category	Count	Proportion (%)
Active Set	HRAC Herbicides (Target)	363	0.72%
	Primary Class: ALS Inhibitors	124	34.1% of actives
Background Set	ChEMBL v.33 (Non-target)	50,000	99.28%
Total Universe	Combined Data	50,363	100%
Imbalance Ratio	Active: Inactive	~ 1 : 140	,

A closer inspection of the ALS-inhibitor subset (Fig. 1B) reveals a high structural density of sulfonylureas, which form the predominant chemical core of this group [24]. Nevertheless, additional families, including imidazolinones and triazolopyrimidines, were incorporated to preserve chemical diversity. This diversity is important because it exposes the model to multiple pharmacophoric arrangements within the same biochemical target class, enabling the identification of shared topological and electronic patterns across distinct scaffolds, a property associated with scaffold hopping behavior [25, 26].

The background subset, composed of 50,000 molecules extracted from ChEMBL, provides a chemically diverse environment that mimics the “noise” encountered in large-scale virtual screening campaigns [27, 28]. This large and structurally heterogeneous background forces the D-MPNN to discriminate herbicidal activity from general drug-like features, rather than relying on simple compositional or size-related biases [22]. Overall, the dataset architecture reflects both the chemical distribution of commercial herbicides and the statistical constraints of highly imbalanced screening scenarios, allowing model performance to be evaluated under conditions that resemble practical discovery workflows.

Chemical Space Characterization and Predictive Performance

To contextualize the predictive behavior of HerbGNN, the physicochemical boundaries of the explored chemical space were first examined. The distribution of lipophilicity (LogP) and polar surface area (TPSA) reveals a distinct “herbicidal niche” (Fig. 2A). Active compounds are predominantly clustered within a specific polarity region characterized by moderate LogP and elevated TPSA values, consistent with physicochemical requirements associated with systemic translocation in plants [29, 30]. This distribution suggests that herbicidal activity is concentrated within a constrained portion of the broader chemical space rather than being randomly dispersed. This observation is further supported by Principal Component Analysis (PCA) of the structural descriptors (Fig. 2B). The PCA scores indicate that active herbicides, particularly ALS inhibitors, form a statistically distinguishable cluster relative to the background ChEMBL molecules. Such separation suggests that active compounds occupy a specific structural and electronic manifold, which is well suited for representation through graph-based neural architectures capable of capturing local topological patterns and functional group arrangements [30, 31].

The discriminative capacity of HerbGNN is quantitatively supported by the metrics summarized in Table 2. The model achieved a mean AUROC of 0.9556 (± 0.05), indicating strong global ranking performance between herbicidal actives and background molecules. However, given the severe ~ 1:140 class imbalance, the Area Under the Precision-Recall Curve (AUPRC) provides a more stringent and informative metric. The observed mean AUPRC of 0.5726 (± 0.19) represents an approximately 80-fold improvement over a random classifier ($AUPRC_{random} \approx 0.007$), demonstrating that the model effectively concentrates true actives at the top of the prioritized list despite the overwhelming presence of decoys. The practical utility of the framework is further highlighted by the Enrichment Factor at 1% (EF 1%), which reached a mean value of 66.64.

In a realistic virtual screening scenario, this level of early enrichment implies that an investigator would identify roughly 66 times more active herbicides within the top 1% of ranked candidates than would be obtained by random selection. Notably, Folds 0–2 exhibited near-perfect discrimination (AUROC > 0.99), whereas a moderate performance decrease was observed in Folds 3 and 4. This variability likely reflects the influence of scaffold diversity and the structural limits of the model's applicability domain under the scaffold-split validation scheme. Nevertheless, the consistently high EF 5% (Mean = 17.45) indicates that, even when global precision fluctuates, the model maintains a reliable hit-enriched region. Such early recognition is critical for reducing the experimental burden, logistical complexity, and costs associated with lead identification in modern agrochemical discovery pipelines [8, 11, 29, 30].

Table 2
Summary of Predictive Performance and Early Enrichment Metrics Across 5-Fold Cross-Validation.

Fold	AUROC	AUPRC	EF (1%)	EF (5%)
Fold 0	0.9906	0.6304	73.49	18.66
Fold 1	0.9950	0.8052	79.78	19.47
Fold 2	0.9971	0.6956	86.36	20.02
Fold 3	0.9063	0.3450	38.88	14.76
Fold 4	0.8888	0.3867	54.71	14.34
Mean (± SD)	**0.9556 ± 0.05**	**0.5726 ± 0.19**	**66.64 ± 19.3**	**17.45 ± 2.6**

AUROC: Area Under the Receiver Operating Characteristic curve; AUPRC: Area Under the Precision-Recall Curve; EF (1%) and EF (5%): Enrichment Factor at the top 1% and 5% of the ranked predictions, respectively. SD: Standard Deviation.

Predictive Performance and Precision-Recall Dynamics

The predictive performance of HerbGNN was evaluated under scaffold-split cross-validation and directly compared against classical machine learning baselines, namely Random Forest (RF) and XGBoost (XGB)

[97, 98], using identical data partitions to ensure a fair and rigorous comparison (Fig. 3). This validation strategy approximates realistic virtual screening scenarios and is widely recommended to assess model generalization across chemically distinct scaffolds rather than memorization of structural patterns [17, 33].

The Precision-Recall (PR) curves in Fig. 3A provide the most informative evaluation under severe class imbalance (~ 1:140), where PR-based metrics better capture model performance in rare-event detection tasks compared to ROC-based metrics [20, 32]. HerbGNN consistently maintains higher precision across a broad range of recall values compared to RF and XGB, particularly in the low-to-intermediate recall regime. This behavior indicates improved prioritization of active compounds during early ranking stages, a critical requirement in virtual screening workflows [35]. While performance variability across folds is observed, reflecting differences in scaffold complexity, HerbGNN exhibits a more gradual decline in precision, whereas baseline models show sharper drops, suggesting reduced discriminative capacity beyond top-ranked candidates [99, 100].

In contrast, the ROC curves (Fig. 3B) indicate that RF and XGB achieve comparable or slightly higher global discriminative performance, as reflected by marginally higher AUROC values. This observation is consistent with previous studies showing that tree-based ensemble methods, including Random Forest and gradient boosting models, are highly effective at capturing global decision boundaries in structured feature spaces such as molecular fingerprints [101, 102]. However, ROC-based metrics are less sensitive to class imbalance and may overestimate performance in highly skewed datasets [32, 103]. Thus, while all models effectively separate actives from inactives at a global level, these results do not capture differences in early enrichment behavior, where HerbGNN demonstrates clear advantages [104]. This distinction is further supported by early enrichment analysis (Fig. 3C), where HerbGNN achieves consistently higher EF1% values compared to RF and XGB. Early enrichment metrics are widely recognized as more relevant indicators of model utility in virtual screening, as they quantify the ability to identify active compounds within the top-ranked fraction of large libraries. The superior EF1% values obtained by HerbGNN indicate that it retrieves a greater proportion of active compounds within the top-ranked subset, directly enhancing screening efficiency and reducing experimental burden [35, 106]. Similarly, the distribution of AUPRC values (Fig. 3D) shows that HerbGNN achieves higher median performance and a favorable distribution across bootstrap samples, indicating improved robustness under scaffold-diverse conditions [32]. This result aligns with recent findings demonstrating that graph-based deep learning models [17], particularly message passing neural networks [84], outperform traditional machine learning approaches in tasks requiring generalization across diverse chemical scaffolds. Taken together, these results demonstrate that although classical machine learning models such as RF and XGB [97, 98] achieve competitive global discrimination, HerbGNN provides superior early enrichment and precision-recall performance [32]. These properties are particularly relevant for real-world virtual screening applications, where the ability to prioritize a small subset of high-confidence candidates is more critical than overall classification accuracy [35].

Error Analysis and Misclassification Drivers

To further investigate model behavior beyond aggregate performance metrics, an analysis of misclassified compounds was conducted by examining the distribution of prediction scores associated with false negatives (FN) and false positives (FP) across HerbGNN, RF, and XGB (Fig. 4). This analysis provides insight into how each model assigns confidence to incorrect predictions, offering an additional layer of interpretability that complements conventional evaluation metrics [14, 18]. The distribution of prediction scores (Fig. 4A) reveals marked differences in how models handle misclassifications. HerbGNN exhibits a highly polarized behavior, with false negatives strongly concentrated near zero and false positives predominantly located near high prediction scores (~ 0.9-1.0). This clear separation indicates that misclassifications occur at the extremes of the prediction spectrum, suggesting that the model establishes well-defined decision boundaries. In contrast, RF and XGB display broader and more overlapping distributions. False negatives for RF are spread across low to intermediate scores (~ 0.05–0.3), while XGB shows even wider dispersion, indicating greater uncertainty in distinguishing borderline compounds. Similarly, false positives for RF and XGB are distributed across mid-to-high score ranges, reflecting less consistent confidence assignment. These trends are further supported by the kernel density estimation (KDE) analysis (Fig. 4B), which highlights the asymmetry of prediction score distributions. HerbGNN shows minimal overlap between FN and FP densities, with FN sharply peaked near zero and FP concentrated at high scores. This bimodal separation suggests that the model makes errors in a structured and predictable manner. In contrast, RF and XGB exhibit substantial overlap between FN and FP distributions, indicating less stable decision boundaries and reduced confidence calibration. Such behavior is commonly observed in classical machine learning models trained on imbalanced datasets, where uncertainty in minority class predictions can lead to diffuse probability distributions [20, 48, 49].

From a mechanistic perspective, the observed behavior of HerbGNN is consistent with the representational advantages of graph-based neural networks. By encoding molecular structures as graphs and propagating information along chemical bonds, message passing neural networks capture higher-order structural dependencies and chemically meaningful features that may not be fully represented by fixed fingerprints [17, 22, 85]. This richer representation likely contributes to the sharper separation between FN and FP distributions, as the model can more effectively distinguish subtle structural patterns associated with herbicidal activity. Importantly, the concentration of false negatives near zero suggests that HerbGNN assigns low confidence to missed active compounds, indicating conservative behavior when uncertain. Conversely, false positives are associated with high prediction scores, suggesting that these compounds share strong structural features with known actives and may represent borderline or potentially novel candidates. This interpretation aligns with previous observations in virtual screening, where high-scoring false positives can still contain chemically relevant motifs and may warrant further investigation [104, 105]. In contrast, the broader and overlapping distributions observed for RF and XGB indicate less reliable confidence assignment. Tree-based ensemble models, while effective at capturing global decision boundaries, rely on predefined molecular descriptors such as fingerprints, which may limit their ability to resolve subtle structural differences

across diverse chemical scaffolds [97, 101, 102]. As a result, these models tend to assign intermediate probability values to misclassified compounds, reflecting increased uncertainty and reduced interpretability. Overall, this analysis demonstrates that HerbGNN not only improves predictive performance but also exhibits more structured and interpretable misclassification behavior. The clear separation between FN and FP prediction scores indicates enhanced confidence calibration and more stable decision boundaries, which are critical for prioritizing candidates in virtual screening workflows. These findings reinforce the advantage of graph-based deep learning approaches in capturing chemically meaningful patterns and provide additional evidence supporting the robustness and practical applicability of the HerbGNN framework.

Chemoinformatic Validation and Lead Prioritization

The final stage of the prioritization pipeline involved a multi-level validation of the selected candidates, integrating global physicochemical profiling, local structural interpretability, and applicability domain (AD) assessment. This combined strategy ensures that prioritized compounds are not only statistically relevant but also chemically plausible and predicted within a reliable region of the learned chemical space. The identification and roles of these molecules are summarized in Table 3.

Table 3
Chemical Identification of Prioritized Candidates and Control.

Candidate ID	Common Name	IUPAC	PubChem CID	Role in Study
Model 0	Chlorsulfuron	1-(2-chlorophenyl)sulfonyl-3-(4-methoxy-6-methyl-1,3,5-triazin-2-yl)urea	47491	Prioritized Lead (N-Sulfonylurea)
Model 1	Cinosulfuron	1-(4,6-dimethoxy-1,3,5-triazin-2-yl)-3-[2-(2-methoxyethoxy)phenyl]sulfonylurea	92420	Methoxy-triazine and a *N*-sulfonylurea
Model 2	Glyphosate	2-(phosphonomethylamino)acetic acid	3496	Contrastive Control

As shown in Fig. 5A, the radar plot indicates that Model 0 and Model 1 occupy the optimal “herbicidal niche,” exhibiting balanced lipophilicity (LogP), molecular weight, and polar surface area (TPSA), which are associated with adequate bioavailability and systemic transport in plants [29, 40]. These physicochemical profiles are consistent with established pesticide-likeness criteria and herbicide design principles, where a balance between polarity and hydrophobicity is required for efficient uptake and translocation [29, 30, 40]. In contrast, Model 2 (glyphosate) displays a markedly different profile characterized by high polarity and low structural complexity, reflecting its distinct mode of action and transport properties [69].

The interpretability analysis (Fig. 5C) provides a localized view of the structural features driving model predictions. For Model 0 and Model 1, the D-MPNN assigns high importance to the sulfonylurea bridge

and adjacent heterocyclic nitrogen-containing rings. These regions correspond to well-established pharmacophoric elements involved in binding to the acetolactate synthase (ALS) active site, where hydrogen-bond acceptors and heterocyclic scaffolds play a central role in enzyme inhibition [41–44, 73]. The alignment between model-derived importance and experimentally validated binding motifs suggests that HerbGNN captures chemically meaningful structure-activity relationships rather than spurious correlations. While Model 0 represents a classical sulfonylurea scaffold, the highlighted substructures in Model 1 illustrate the model's ability to prioritize structurally more complex and decorated heterocycles that retain key pharmacophoric features. This behavior is consistent with the concept of scaffold hopping, where novel chemical frameworks preserve essential functional motifs while introducing structural diversity [25, 26, 39]. Such capability is particularly relevant in herbicide discovery, where overcoming resistance and expanding chemical diversity are critical challenges [2, 4, 8].

Conversely, Model 2 lacks a coherent structural pattern consistent with the herbicidal niche defined by the training actives. Its small, highly polar structure does not contain extended aromatic systems or sulfonylurea-like motifs commonly associated with ALS inhibition [41–44, 47]. This explains its low prioritization score and supports the model's ability to discriminate between mechanistically relevant and unrelated compounds. To further assess prediction reliability, an applicability domain (AD) analysis was incorporated as an additional validation layer (Fig. 5B), relating prediction confidence to structural similarity with the training set and enabling estimation of whether compounds fall within the learned chemical space [36–38, 51]. The prioritized candidates (Model 0 and Model 1) are located within or near regions of high similarity and confidence, indicating that their predicted activity is supported by sufficient representation in the training data and driven by learned structure-activity relationships rather than uncontrolled extrapolation. As expected, compounds approaching the applicability boundary (~ 0.4) exhibit increased dispersion in prediction scores, reflecting higher uncertainty in regions with limited chemical representation [36, 38]. While high-confidence predictions are largely concentrated within or near this threshold, a subset of compounds close to the boundary still receive elevated scores, suggesting controlled generalization to structurally related scaffolds rather than unrestricted extrapolation. Additionally, a substantial proportion of compounds are assigned near-zero scores across the similarity range, indicating conservative behavior toward the majority inactive class. This behavior is consistent with recent advances in graph-based deep learning, which demonstrate improved capacity to capture transferable chemical patterns across diverse molecular spaces [17, 22, 85].

Overall, the integration of physicochemical profiling, structural interpretability, and applicability domain analysis provides a mechanistically informed and statistically robust framework for lead prioritization. This multi-level validation ensures that selected candidates are not only enriched in predicted activity but also chemically plausible and located within a region of the chemical space where model predictions can be considered reliable. Such an approach is consistent with current best practices in explainable artificial intelligence and cheminformatics-driven molecular design [10–14, 18], reinforcing the potential of HerbGNN as a practical tool for herbicide discovery.

Mechanistic Validation via Quantum Chemical Analysis: FMO and ESP Insights

The predictive logic of HerbGNN was further examined through quantum mechanical characterization at the B3LYP/6-31G(d) level of theory [58, 59]. Rather than applying these calculations to the entire screening library, the analysis was intentionally restricted to a small, representative subset of molecules: two high-confidence prioritized candidates (Model 0 and Model 1) and one contrastive control (Model 2). This targeted selection enables a focused mechanistic comparison between scaffolds strongly favored by the model and a structurally and electronically distinct herbicide that received a low prediction score. Such an approach is consistent with common practices in computational chemistry, where Density Functional Theory calculations are typically applied to representative leads due to their computational cost, rather than to large virtual screening datasets [60, 61]. The combined analysis of Frontier Molecular Orbitals (FMO) and Molecular Electrostatic Potential (ESP) on these representative structures provides a mechanistic interpretation of the model's prioritization behavior [60–63].

The energy gap between the Highest Occupied Molecular Orbital (E_{HOMO}) and the Lowest Unoccupied Molecular Orbital (E_{LUMO}) is a fundamental descriptor of molecular stability and chemical hardness [63, 64]. As summarized in Table 4, Model 0 and Model 1 exhibit relatively narrow energy gaps ($\Delta E \approx 3.03$–3.46 eV). According to the theory of chemical hardness and softness, molecules with smaller gaps are generally more polarizable and chemically reactive, a characteristic often associated with heterocyclic herbicides and other bioactive scaffolds [63, 65]. Conversely, the highest dipole moment found in Model 0 (9.11 D) highlights a superior electrostatic orientation, which facilitates long-range recognition and high-affinity binding within the enzymatic pocket. This contrast confirms that the neural network is not merely recognizing atom types but is effectively responding to the underlying quantum chemical stability and polarity of the candidates.

Visual inspection of the FMOs in Fig. 6A indicates that, for the prioritized leads, the electron density is predominantly localized over the sulfonylurea bridge and the adjacent heterocyclic rings. In Model 0, the E_{HOMO} is extensively delocalized across the sulfonylurea framework, while the E_{LUMO} is more localized toward the nitrogen-containing pyrimidine ring. Similarly, Model 1 exhibits a spatial redistribution of electron density, with the E_{HOMO} centered around the sulfur-containing moiety and the E_{LUMO} shifted toward the heterocyclic scaffold. This spatial differentiation between frontier orbitals is consistent with increased polarizability and electronic flexibility, features commonly associated with heterocyclic herbicidal scaffolds [66–68]. In contrast, Model 2 presents a significantly wider energy gap of 6.7419 eV, consistent with higher kinetic stability and a distinct electronic profile relative to the prioritized candidates. Moreover, the spatial distributions of its E_{HOMO} and E_{LUMO} are highly similar, reflecting a more localized electronic structure [65, 69]. This orbital overlap, together with its wide energy gap, reflects a more localized electronic structure compared to the prioritized candidates. Such differences provide a physically interpretable rationale for the HerbGNN prioritization logic: the model preferentially identifies scaffolds with delocalized and electronically adaptable frameworks, while assigning lower scores to

molecules whose electronic profiles differ substantially from those characteristics of the ALS-inhibitor chemical space.

Table 4
Computed FMO Energies, Energy Gaps, and Total Dipoles.

Molecule	E_{HOMO} (eV)	E_{LUMO} (eV)	Δ*E* (Gap, eV)	Dipole Moment (D)
Model 0	-5.9202	-2.4551	3.4651	9.1137
Model 1	-5.3411	-2.3097	3.0314	4.6108
Model 2	-7.1474	-0.4055	6.7419	3.3822

FMO: Frontier Molecular Orbital; EHOMO: Energy of the Highest Occupied Molecular Orbital; ELUMO: Energy of the Lowest Unoccupied Molecular Orbital; ΔE: Energy gap calculated as the difference between EHOMO and ELUMO.

The ESP maps, visualized in Chemcraft [70] from ORCA-generated files [71], provide a surface-level perspective on the model's selectivity (Fig. 6B). Molecular electrostatic potentials are widely used to interpret intermolecular interactions, hydrogen bonding, and reactive sites in organic and bioactive molecules [62, 73]. In Model 0 and Model 1, the electrostatic distribution reveals a well-organized dipolar profile. Most of the molecular scaffold exhibits a positive-to-neutral potential (visualized in red), while critical functional groups, specifically the sulfonyl oxygens (SO_2) and the carbonyl oxygen, display pronounced negative potential regions (transitioning to yellow-green with cyan tips). These regions correspond to zones of high electron density and act as primary nucleophilic centers, consistent with hydrogen-bond acceptor sites involved in interactions with polar residues within the acetolactate synthase (ALS) active site [73–77]. Structural and mechanistic studies of ALS show that the catalytic pocket is organized around the thiamine diphosphate (TPP) cofactor and a divalent metal ion, forming a highly polar environment that stabilizes negatively polarized functional groups of substrates and inhibitors through hydrogen bonding and electrostatic interactions [73–77]. High-resolution crystal structures and mechanistic analyses reveal that residues involved in TPP coordination and substrate positioning play key roles in stabilizing reaction intermediates [73–77]. The electrostatic features observed in the prioritized sulfonylurea-like candidates are therefore consistent with the known binding environment of ALS inhibitors, which rely on appropriately positioned electron-rich oxygen atoms and heterocyclic nitrogen centers to form stabilizing interactions within the catalytic pocket [73–77]. The ESP map for Model 2 shows a markedly different electrostatic profile dominated by highly localized negative regions on the phosphonate and carboxylate groups. Additionally, the visualization indicates a short contact between the protonated amine group and a neighboring oxygen atom, consistent with an intramolecular N-H···O hydrogen-bonding interaction. Spectroscopic and computational studies have shown that glyphosate commonly adopts zwitterionic conformations stabilized by intramolecular hydrogen bonds between the protonated amine and either the carboxylate or phosphonate oxygen atoms [78–80].

Overall, the observed correspondence between the neural network importance patterns and the quantum-chemical descriptors indicates that HerbGNN is not merely recognizing superficial structural motifs, but is capturing underlying electronic features associated with known herbicidal reactivity. The alignment between model-highlighted regions and the FMO and ESP reactive centers suggests that the network has internalized chemically meaningful patterns within the ALS-inhibitor chemical space. This convergence between data-driven predictions and quantum-chemical descriptors provides a mechanistically interpretable basis for candidate prioritization and supports the use of HerbGNN as a physically informed virtual screening tool.

Conclusions

In this study, HerbGNN is presented as a graph-based deep learning framework for herbicide discovery, specifically designed to operate under the constraints of highly imbalanced chemical spaces. Through a systematic comparison against established descriptor-based machine learning models, including Random Forest and XGBoost, HerbGNN demonstrated comparable global discrimination but consistently superior performance in Precision–Recall space and early enrichment metrics. This distinction highlights the importance of evaluating models using metrics aligned with real-world virtual screening objectives, where the prioritization of true positives at early ranks is critical [32, 103]. Importantly, HerbGNN exhibited improved reliability beyond predictive performance. Error distribution analysis revealed structured separation between false positives and false negatives, while calibration assessment indicated better agreement between predicted probabilities and observed outcomes. Applicability domain analysis further showed that high-confidence predictions are predominantly supported by the training chemical space, with evidence of controlled generalization near the similarity boundary rather than uncontrolled extrapolation [36–38, 51]. The integration of quantum chemical analysis provided an additional layer of mechanistic interpretability. The observed correspondence between model-prioritized regions and electronic descriptors, including frontier molecular orbitals and electrostatic potential distributions, supports the hypothesis that HerbGNN captures chemically meaningful patterns associated with herbicidal activity, rather than relying solely on statistical correlations. Despite these advances, some limitations should be acknowledged. The dataset size and class imbalance may constrain generalization to entirely novel chemical spaces, and the quantum chemical analysis was restricted to a limited set of representative compounds. Moreover, while the results demonstrate improved prioritization performance, experimental validation will be required to confirm biological activity of the proposed candidates. Overall, HerbGNN provides a robust, interpretable, and practically relevant framework for herbicide discovery. By combining improved early enrichment, probabilistic reliability, and chemically grounded interpretability, this approach offers a scalable strategy for prioritizing candidates in large chemical libraries and establishes a foundation for future integration with experimental validation and structure-based design.

Methods

Curation of the Chemical Library and Dataset Architecture

The foundational dataset was constructed by integrating 363 active herbicides from the 2026 HRAC Global Herbicide MoA Classification Master List [26] with a background set of 50,000 molecules randomly sampled from ChEMBL v.33 [27]. Background compounds were selected under the assumption of no reported herbicidal activity. Active-compound retrieval and initial HRAC parsing were performed using a specialized Python (v3.11) framework [81], while background canonical SMILES were extracted directly from the ChEMBL repository. The extraction process involved parsing the latest *chemreps* data from the EMBL-EBI FTP, filtering for non-empty entries, and performing a randomized selection with a fixed seed to ensure dataset stability and reproducibility. Standardization and cleaning were executed using RDKit [82] following a multi-stage pipeline: (i) removal of inorganic and organometallic species; (ii) salt stripping, neutralization of formal charges, and selection of the largest molecular fragment; (iii) tautomer canonicalization and duplicate removal via canonical SMILES; and (iv) size-filtering to a molecular weight window of 50–800 Da. Elements were restricted to a standard organic subset (H, C, N, O, F, P, S, Cl, Br, I). The final processed dataset comprised 50,363 unique molecules.

Descriptor Calculation and Dataset Export

Molecular descriptors used for chemoinformatic characterization were computed with RDKit [82]: MolWt, TPSA, MolLogP, number of heteroatoms, number of rings, number of rotatable bonds, fraction sp3 carbons, and additional structural counts. Calculations were parallelized where applicable; invalid molecules were flagged and removed. A cleaned CSV/XLSX with selected columns (Mode of Action, Chemical Class, PubChem_CID, Canonical_SMILES) was exported as part of the pipeline for downstream modeling.

Deep Learning Architecture and Implementation

In this approach, molecules were represented as directed graphs, where atoms and bonds were encoded into initial feature vectors of 147 and 14 dimensions, respectively, following the representation implemented in the Chemprop (v2) framework [17, 83, 84]. Unlike conventional message-passing neural networks that propagate information at the atom level, the Directed Message Passing Neural Network (D-MPNN) associates messages with directed bonds, thereby preventing redundant backtracking of information between neighboring atoms, commonly referred to as “tottering” [17, 85, 86]. This edge-centric message passing scheme enables a more efficient and chemically meaningful encoding of molecular structure, improving the quality of learned representations for property prediction tasks [85, 86].

Graph Neural Network (GNN) Development, Training, and Evaluation

Molecular property prediction was performed using a Directed Message Passing Neural Network (D-MPNN) implemented in the Chemprop v2 framework and integrated with PyTorch Lightning (v2.6.1) for efficient and scalable training [83, 87]. In this representation, molecules are encoded as directed graphs

$G = (V, E)$, where atoms and bonds are treated as nodes and edges, respectively. Unlike conventional node-based graph neural networks, the D-MPNN propagates messages along directed bonds, reducing redundant information exchange (“tottering”) and enabling improved capture of chemically relevant substructures and long-range dependencies [83]. Atom-level features included atomic number, degree, formal charge, chirality, and hybridization state, while bond features encoded bond type, conjugation, and stereochemistry using one-hot representations. After message passing, atom embeddings were aggregated through a permutation-invariant pooling operation to generate fixed-length molecular representations suitable for classification tasks [83]. To ensure a realistic evaluation of model generalization, data were partitioned using a scaffold-balanced splitting strategy based on Bemis-Murcko frameworks, implemented via the Astartes library (v1.3.3) [45, 46, 88]. This approach enforces separation of structurally distinct scaffolds between training, validation, and test sets, providing a stringent benchmark that reflects prospective virtual screening conditions and prevents overestimation of model performance due to scaffold memorization [17, 33]. The same scaffold splits were consistently applied across all baseline models (Random Forest and XGBoost) to ensure fair and directly comparable evaluation.

Given the pronounced class imbalance (~ 1:140), model optimization incorporated class-balanced weighting combined with focal loss (γ = 2.0), which emphasizes hard to classify minority-class samples while reducing the contribution of well classified majority-class instances [19, 20]. This strategy is particularly suited for rare event detection problems such as herbicidal activity prediction [48–50]. Hyperparameter optimization was conducted using Bayesian optimization with a Tree-structured Parzen Estimator (TPE) over 50 iterations, using validation AUPRC as the objective function [89]. The final model employed a hidden dimension of 300, three message-passing steps, and a dropout rate of 0.2. Training was performed for up to 100 epochs with early stopping (patience = 10) based on validation performance [90]. Model performance was evaluated using five-fold scaffold-split cross-validation. In line with the severe class imbalance, primary evaluation metrics included the area under the precision-recall curve (AUPRC) and early enrichment (EF1%), which provide a more informative assessment than ROC-based metrics in rare-event settings [32, 35, 103].

To ensure robust statistical comparison with baseline models, performance distributions were further analyzed using bootstrap resampling to estimate 95% confidence intervals for PR curves, ROC curves, and enrichment metrics. Additional evaluation layers included calibration analysis (reliability diagrams and Brier scores) to assess probabilistic reliability, as well as error distribution analysis (false positives and false negatives) to characterize model behavior under misclassification scenarios. Finally, an applicability domain (AD) analysis based on Tanimoto similarity (ECFP4) was performed to relate prediction confidence with structural proximity to the training set, enabling assessment of prediction reliability across chemical space [36–38, 51]. All models were trained using GPU acceleration on an NVIDIA RTX 4060 Ti, while data preprocessing and feature generation were parallelized on a multi-threaded CPU environment.

Chemical Space Analysis

Global chemical space analyses were conducted using a curated dataset of herbicides annotated by Mode of Action (MoA). Each compound was represented using physicochemical descriptors capturing molecular size, polarity, lipophilicity, and structural complexity, including Molecular Weight (MolWt), Topological Polar Surface Area (TPSA), MolLogP, number of heteroatoms, number of rings, number of rotatable bonds, and fraction of sp^3 carbons. Missing values were imputed using descriptor-wise mean substitution, and all features were standardized via z-score normalization. Principal Component Analysis (PCA) was applied to the standardized descriptor matrix, and compounds were projected onto the first two principal components to visualize global chemical space organization. Molecules were stratified into Acetolactate Synthase (ALS) inhibitors and non-ALS herbicides, enabling identification of target-specific clustering patterns. Additionally, a two-dimensional projection based on TPSA and MolLogP was used to capture polarity-lipophilicity balance and identify physicochemical niches associated with herbicidal activity.

Applicability Domain

The Applicability Domain (AD) was defined using a similarity-based approach based on circular fingerprints (ECFP4, radius = 2, 2048 bits). For each test compound, the maximum Tanimoto similarity to the training set was calculated, and prediction confidence was analyzed as a function of structural proximity to the training distribution. A similarity threshold of Tanimoto ≥ 0.4 was used to distinguish in-domain and near-boundary regions. This framework enables the evaluation of prediction reliability by linking model confidence to structural similarity, allowing identification of regions where predictions are expected to be more or less reliable [36–38, 51].

Predictive performance was evaluated using threshold-independent metrics derived from Precision-Recall (PR) and Receiver Operating Characteristic (ROC) curves across five scaffold-split validation folds. PR curves were interpolated over a common recall grid to compute mean performance and variability across folds. To ensure robust statistical comparisons, bootstrap resampling was applied to estimate confidence intervals for PR curves, ROC curves, AUPRC, and early enrichment metrics (EF1%). Error analysis was performed by classifying predictions into false positives (FP) and false negatives (FN) using a fixed probability threshold. The distribution of prediction scores for each error type was analyzed using violin plots and kernel density estimation (KDE), enabling characterization of confidence patterns and identification of regions of uncertainty within chemical space.

Representative compounds were selected based on model prediction scores to enable targeted chemoinformatic validation. Top-ranked candidates were identified according to predicted probabilities, and two high-confidence molecules (Model 0 and Model 1) were selected as prioritized leads. Glyphosate (Model 2) was included as a contrastive reference due to its distinct mechanism of action and structural properties. Descriptor values were normalized using min-max scaling, and radar plots were constructed to compare individual compounds against class-level averages, enabling visualization of their alignment with the physicochemical space of ALS inhibitors.

Model Interpretation

Model interpretability was explored using a fragment-based attribution approach. Monte Carlo Tree Search (MCTS) was applied to identify substructures contributing most significantly to model predictions. These fragments were mapped onto molecular structures to visualize chemically relevant regions driving model decisions, enabling interpretation of predictions in terms of known pharmacophoric features.

Quantum Chemical Validation (DFT) and Mechanistic Interpretation.

Mechanistic validation of the prioritized leads was conducted using the ORCA v.6.1 program system [71] to establish a physical correlation between the structural representations learned by the D-MPNN architecture and molecular reactivity. Geometry optimizations were performed at the B3LYP/6-31G(d) level of theory, utilizing the RIJCOSX approximation to accelerate the calculation of exchange and Coulomb integrals [60, 92, 93]. To ensure chemical accuracy in biological environments, single-point energy calculations included Grimme's D3 dispersion corrections and the Polarizable Continuum Model (PCM) for implicit water solvation. Convergence criteria were set to "TightOpt" to ensure stable stationary points [71, 94–96].

Frontier molecular orbital energies (E_{HOMO} and E_{LUMO}) were extracted to compute the fundamental energy gap ($\Delta E = E_{HOMO} - E_{LUMO}$). Additionally, Electrostatic Potential (ESP) maps were generated on a high-density 0.2 Bohr grid to identify nucleophilic and electrophilic centers [60, 80]. This electronic characterization was performed on a small set of representative molecules selected from the model outputs: two high-confidence candidates (Model 0 and Model 1) and one structurally distinct negative control (Model 2, glyphosate). Model 0 was chosen as the top-ranked prediction to represent a prioritized lead, whereas Model 1 was selected to evaluate the model's ability to identify structurally diverse scaffolds. Glyphosate was included as a contrastive control because, despite being a potent commercial herbicide, it was assigned a low probability score by HerbGNN. This case was selected to investigate whether the model's predictive behavior was driven by specific structural and electronic signatures rather than a generic 'herbicide-likeness' bias. By analyzing glyphosate via DFT, we sought to verify if its distinct electronic profile provides a physical justification for its exclusion from the prioritized leads. To evaluate the physical significance of the model's predictions, atomic importance scores derived from the D-MPNN interpretability module were correlated with DFT-derived ESP values. This relationship was assessed using Spearman rank correlation and validated through permutation testing (n = 1,000) to ensure statistical significance. This integration aims to determine whether the chemical substructures identified as "highly important" by the deep learning architecture coincide with the regions of high electron density or specific reactivity centers identified via quantum mechanics.

Abbreviations

AD
Applicability Domain
AI

Artificial Intelligence
ALS
Acetolactate Synthase
AUPRC
Area Under the Precision-Recall Curve
AUROC
Area Under the Receiver Operating Characteristic
D-MPNN
Directed Message Passing Neural Network
DFT
Density Functional Theory
DL
Deep Learning
ECFP4
Extended-Connectivity Fingerprint 4
EF1%
Enrichment Factor at 1%
EHOMO
Highest Occupied Molecular Orbital Energy
ELUMO
Lowest Unoccupied Molecular Orbital Energy
ESP
Electrostatic Potential
FMO
Frontier Molecular Orbital
FN
False Negative
FP
False Positive
HRAC
Herbicide Resistance Action Committee
HTS
High-Throughput Screening
KDE
Kernel Density Estimation
LogP / MolLogP
Lipophilicity (Partition Coefficient)
MCTS
Monte Carlo Tree Search
ML

Machine Learning
MoA
Mode of Action
MolWt
Molecular Weight
PCA
Principal Component Analysis
PCM
Polarizable Continuum Model
PPO
Protoporphyrinogen Oxidase
PR
Precision-Recall
QSAR
Quantitative Structure-Activity Relationship
RF
Random Forest
ROC
Receiver Operating Characteristic
TPE
Tree-structured Parzen Estimator
TPP
Thiamine Diphosphate
TPSA
Topological Polar Surface Area
XGB
XGBoost

Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no external funding.

Author Contribution

C.G.T. formed the data sets and calculated descriptors, designed the experiments, performed all modelling, analyzed and interpreted the data. V.M.J. and C.G.H reviewed and edited the manuscript. C.G.T. conceived the presented idea and wrote the manuscript.

Acknowledgement

The author Tonatiu Campos-García thank SECIHTI-México for the postdoctoral fellowships awarded.

Data Availability

The datasets, scripts, and analysis pipeline used in this study will be made publicly available through a GitHub repository upon publication of the article. The repository will include the curated dataset, fingerprint generation scripts, baseline model training scripts (Random Forest and XGBoost), statistical analysis, bootstrap evaluation, and figure generation pipeline described in the manuscript. To ensure reproducibility of the benchmarking analyses without requiring retraining of the deep learning model, precomputed prediction outputs from the HerbGNN model will also be provided. During the peer-review process, the materials are available from the corresponding author upon reasonable request.

References

1. Heap I (2026) The International Herbicide-Resistant Weed Database. http://www.weedscience.org
2. Sparks TC, Lorsbach BA, Sparks JM, Duke SO (2025) Origins of new modes of action for fungicides, herbicides and insecticides: a review and analysis. Pest Manag Sci 81(10):6019–6028. 10.1002/ps.8962
3. Hsiao AS (2025) Agrochemical innovation for crop health: moving forward through dynamic disorder. Front Plant Sci 16:1719658. 10.3389/fpls.2025.1719658
4. Torra J, Alcantara-De la Cruz R (2022) Molecular mechanisms of herbicide resistance in weeds. Genes 13(11):2025. 10.3390/genes13112025
5. Alwarnaidu Vijayarajan VB, Torra J, Runge F, Nolan G, Hennessy M, Dermot Forristal P (2025) Target-site resistance to ALS-inhibiting herbicides in Stellaria media, Papaver rhoeas, Glebionis segetum and Veronica persica from Ireland. Weed Res 65(1). 10.1111/wre.70000
6. Ofosu R, Agyemang ED, Márton A, Pásztor G, Taller J, Kazinczi G (2023) Herbicide resistance: managing weeds in a changing world. Agronomy 13(6):1595. 10.3390/agronomy13061595
7. Berestetskiy A (2023) Modern approaches for the development of new herbicides based on natural compounds. Plants 12(2):234. 10.3390/plants12020234
8. Duke SO, Twitty A, Baker C, Sands D, Boddy L, Travaini ML et al (2024) New approaches to herbicide and bioherbicide discovery. Weed Sci 72(5):444–464. 10.1017/wsc.2024.54

9. AgbioInvestor (2024) Time and Cost of New Agrochemical Product Discovery, Development and Registration: A Study on Behalf of CropLife International. AgbioInvestor
10. Riter LS, Bienstock RJ, O'Sullivan W (2025) Building Trust: Validating Artificial Intelligence (AI) and Machine Learning in Agricultural Applications. J Agric Food Chem. 10.1021/acs.jafc.5c16871
11. Djoumbou-Feunang Y, Wilmot J, Kinney J, Chanda P, Yu P, Sader A et al (2023) Cheminformatics and artificial intelligence for accelerating agrochemical discovery. Front Chem 11:1292027. 10.3389/fchem.2023.1292027
12. Yuan H, Zhou H, Huo J, Zhang J, Cheng M (2025) A pesticide molecular generation model based on deep learning and scaffold structure MHA-RNN. Trans Chin Soc Agric Eng 41(1):200–211. 10.11975/j.issn.1002-6819.202409017
13. Zhang P, Cao H, Li T, Fu Y (2025) Virtual Screening and Bioassay of Novel Protoporphyrinogen Oxidase and p-Hydroxyphenylpyruvate Dioxygenase Dual-Target Inhibitors. Molecules 30(7):1491. 10.3390/molecules30071491
14. Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. IEEE Access 8:42200–42216. 10.1109/ACCESS.2020.2976199
15. Ciarlantini C, Francolini I, Silvestro I, Mariano A, d'Abusco AS, Piozzi A (2024) Design of bioactive and biomimetic scaffolds based on chitosan-alginate polyelectrolyte complexes for tissue engineering. Carbohydr Polym 327:121684. 10.1016/j.carbpol.2023.121684
16. Oršolić D, Pehar V, Šmuc T, Stepanić V (2021) Comprehensive machine learning based study of the chemical space of herbicides. Sci Rep 11(1):11479. 10.1038/s41598-021-90690-w
17. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H et al (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59(8):3370–3388. 10.1021/acs.jcim.9b00237
18. Lamens A, Bajorath J (2026) Explainable artificial intelligence for molecular design in pharmaceutical research. Chem Sci. 10.1039/D5SC08461J
19. Ross TY, Dollár GKHP (2017) Focal loss for dense object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2980–2988
20. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5(4):221–232. 10.1007/s13748-016-0094-0
21. Dablain D, Jacobson KN, Bellinger C, Roberts M, Chawla NV (2024) Understanding CNN fragility when learning with imbalanced data. Mach Learn 113(7):4785–4810. 10.1007/s10994-023-06326-9
22. Li Z, Jiang M, Wang S, Zhang S (2022) Deep learning methods for molecular representation and property prediction. Drug Discov Today 27(12):103373. 10.1016/j.drudis.2022.103373
23. HRAC (2026) HRAC Global Herbicide MoA Classification Master List. Herbicide Resistance Action Committee. https://hracglobal.com/tools/2026-hrac-global-herbicide-moa-classification. Accessed 07 Mar 2026

24. Andrea TA, Artz SP, Ray TB, Pasteris RJ (2024) Structure-activity relationships of sulfonylurea herbicides. Rational Approaches to Structure, Activity, and Ecotoxicology of Agrochemicals. CRC, pp 373–395
25. Wang S, Zhang R, Li X, Cai F, Ma X, Tang Y et al (2025) Recent advances in molecular representation methods and their applications in scaffold hopping. Npj Drug Discov 2(1):14. 10.1038/s44386-025-00017-2
26. Rossen L, Sirockin F, Schneider N, Grisoni F (2025) Scaffold hopping with generative reinforcement learning. J Chem Inf Model 65(13):6513–6525. 10.1021/acs.jcim.5c00029
27. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940. 10.1093/nar/gky1075
28. Smit I, Adasme MF, Manners E, Corbett S, Bosc N, Do HMA et al (2026) Integrating artificial intelligence and manual curation to enhance bioassay annotations in ChEMBL. J Cheminform 18(1):24. 10.1186/s13321-026-01165-x
29. Avram S, Funar-Timofei S, Borota A, Chennamaneni SR, Manchala AK, Muresan S (2014) Quantitative estimation of pesticide-likeness for agrochemical discovery. J Cheminform 6(1):42. 10.1186/s13321-014-0042-6
30. Huang JJ, Wang F, Ouyang Y, Huang YQ, Jia CY, Zhong H, Hao GF (2021) HerbiPAD: a free web platform to comprehensively analyze constitutive property and herbicide-likeness to estimate chemical bioavailability. Pest Manag Sci 77(3):1273–1281. 10.1002/ps.6140
31. Babar M, Hassan F, Ijaz M, Mohyuddin MT (2024) Chemoinformatics. Trends in Plant Biotechnology. Springer, pp 335–357. doi:10.1007/978-981-97-0814-7_11
32. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432. 10.1371/journal.pone.0118432
33. Sheridan RP (2013) Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 53(4):783–790. 10.1021/ci400084k
34. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874. 10.1016/j.patrec.2005.10.010
35. Truchon JF, Bayly CI (2007) Evaluating virtual screening methods: good and bad metrics for the early recognition problem. J Chem Inf Model 47(2):488–508. 10.1021/ci600426e
36. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R (2012) Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17(5):4791–4810. 10.3390/molecules17054791
37. Hanser T, Barber C, Marchaland JF, Werner S (2016) Applicability domain: towards a more formal definition. SAR QSAR Environ Res 27(11):865–881. 10.1080/1062936X.2016.1250229
38. Schultz LE, Wang Y, Jacobs R, Morgan D (2025) A general approach for determining applicability domain of machine learning models. npj Comput Mater 11(1):95. 10.1038/s41524-025-01573-x

39. Hu Y, Stumpfe D, Bajorath J (2017) Recent advances in scaffold hopping: miniperspective. J Med Chem 60(4):1238–1246. 10.1021/acs.jmedchem.6b01437
40. Ouyang Y, Huang JJ, Wang YL, Zhong H, Song BA, Hao GF (2021) In silico resources of drug-likeness as a mirror: what are we lacking in pesticide-likeness? J Agric Food Chem 69(37):10761–10773. 10.1021/acs.jafc.1c01460
41. Pang SS, Guddat LW, Duggleby RG (2003) Molecular basis of sulfonylurea herbicide inhibition of acetohydroxyacid synthase. J Biol Chem 278(9):7639–7644. 10.1074/jbc.M211648200
42. Wang C, Yang GY, Wei YL, Li SQ, Li HY, Xia ZW et al (2025) Design, synthesis, and bioevaluation of sulfonylamides as acetohydroxyacid synthase inhibitors. J Mol Struct 1326:141075. 10.1016/j.molstruc.2024.141075
43. Xu Q, Gao Y, Sun Z, Shi JR, Tang JY, Wang Y et al (2024) Chemical synthesis, herbicidal activity, crop safety, and molecular basis of ortho-fluoroalkoxy substituted sulfonylureas as novel acetohydroxyacid synthase inhibitors. J Agric Food Chem 72(41):22595–22605. 10.1021/acs.jafc.4c05201
44. Cheng Y, Lonhienne T, Garcia MD, Williams CM, Schenk G, Guddat LW (2023) Crystal structure of the commercial herbicide, amidosulfuron, in complex with Arabidopsis thaliana acetohydroxyacid synthase. J Agric Food Chem 71(13):5117–5126. 10.1021/acs.jafc.2c08528
45. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39(15):2887–2893. 10.1021/jm9602928
46. Bemis GW, Murcko MA (1999) Properties of known drugs. 2. Side chains. J Med Chem 42(25):5095–5099. 10.1021/jm9903996
47. Duke SO (2012) Why have no new herbicide modes of action appeared in recent years? Pest Manag Sci 68(4):505–512. 10.1002/ps.2333
48. Jiang J, Zhang C, Ke L, Hayes N, Zhu Y, Qiu H et al (2025) A review of machine learning methods for imbalanced data challenges in chemistry. Chem Sci 16(18):7637–7658. 10.1039/D5SC00270B
49. Altalhan M, Algarni A, Alouane MTH (2025) Imbalanced data problem in machine learning: A review. IEEE Access 13:13686–13699. 10.1109/ACCESS.2025.3531662
50. Zakharov AV, Peach ML, Sitzmann M, Nicklaus MC (2014) QSAR modeling of imbalanced high-throughput screening data in PubChem. J Chem Inf Model 54(3):705–712. 10.1021/ci400737s
51. Luque Ruiz I, Gómez-Nieto MÁ (2018) Study of the applicability domain of the QSAR classification models by means of the rivality and modelability indexes. Molecules 23(11):2756. 10.3390/molecules23112756
52. Aniceto N, Freitas AA, Bender A, Ghafourian T (2016) A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. J Cheminform 8(1):69. 10.1186/s13321-016-0182-y
53. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23(8):1538–1546. 10.1016/j.drudis.2018.05.010

54. Esposito C, Landrum GA, Schneider N, Stiefl N, Riniker S (2021) GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. J Chem Inf Model 61(6):2623–2640. 10.1021/acs.jcim.1c00160
55. Niazi SK, Mariam Z (2023) Recent advances in machine-learning-based chemoinformatics: a comprehensive review. Int J Mol Sci 24(14):11488. 10.3390/ijms241411488
56. Liu Y, Guo Y, Wu W, Xiong Y, Sun C, Yuan L, Li M (2019) A machine learning-based QSAR model for benzimidazole derivatives as corrosion inhibitors by incorporating comprehensive feature selection. Interdiscip Sci Comput Life Sci 11(4):738–747. 10.1007/s12539-019-00346-7
57. Ambure P, Cordeiro MNDS (2020) Importance of data curation in QSAR studies especially while modeling large-size datasets. Ecotoxicological QSARs. Springer, pp 97–109. doi:10.1007/978-1-0716-0150-1_5
58. Lee C, Yang W, Parr RG (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys Rev B 37(2):785. 10.1103/PhysRevB.37.785
59. Hariharan PC, Pople JA (1973) The influence of polarization functions on molecular orbital hydrogenation energies. Theor Chim Acta 28(3):213–222. 10.1007/BF00533485
60. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98(7):5648–5652. 10.1063/1.464913
61. Li SC, Wu H, Menon A, Spiekermann KA, Li YP, Green WH (2024) When do quantum mechanical descriptors help graph neural networks to predict chemical properties? J Am Chem Soc 146(33):23103–23120. 10.1021/jacs.4c04670
62. Politzer P, Murray JS (2002) The fundamental nature and role of the electrostatic potential in atoms and molecules. Theor Chem Acc 108(3):134–142. 10.1007/s00214-002-0363-9
63. Fukui K (1982) Role of frontier orbitals in chemical reactions. Science 218(4574):747–754. 10.1126/science.218.4574.747
64. Pearson RG (1963) Hard and soft acids and bases. J Am Chem Soc 85(22):3533–3539
65. Zhan CG, Nichols JA, Dixon DA (2003) Ionization potential, electron affinity, electronegativity, hardness, and electron excitation energy: molecular properties from density functional theory orbital energies. J Phys Chem A 107(20):4184–4195. 10.1021/jp0225774
66. Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH et al (2019) A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci 10(2):370–377. 10.1039/C8SC04228D
67. Jeschke P (2016) Progress of modern agricultural chemistry and future prospects. Pest Manag Sci 72(3):433–455. 10.1002/ps.4190
68. Vitaku E, Smith DT, Njardarson JT (2014) Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among US FDA approved pharmaceuticals: miniperspective. J Med Chem 57(24):10257–10274. 10.1021/jm501100b

69. Duke SO (2018) The history and current status of glyphosate. Pest Manag Sci 74(5):1027–1034. 10.1002/ps.4652
70. Chemcraft - graphical software for visualization of quantum chemistry computations. Version 1.8, build 682. https://www.chemcraftprog.com
71. Neese F (2025) Software update: the ORCA program system,version 6.0. Wiley Interdiscip Rev Comput Mol Sci 15(2):e70019. 10.1002/wcms.70019
72. Politzer P, Murray JS, Clark T (2013) Halogen bonding and other σ-hole interactions: a perspective. Phys Chem Chem Phys 15(27):11178–11189. 10.1039/C3CP00054K
73. McCourt JA, Pang SS, King-Scott J, Guddat LW, Duggleby RG (2006) Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase. Proc Natl Acad Sci 103(3):569–573. 10.1073/pnas.0508701103
74. Liu Y, Li Y, Wang X (2016) Acetohydroxyacid synthases: evolution, structure, and function. Appl Microbiol Biotechnol 100(20):8633–8649. 10.1007/s00253-016-7809-9
75. Lonhienne T, Garcia MD, Noble C, Harmer J, Fraser JA, Williams CM et al (2017) High resolution crystal structures of the acetohydroxyacid synthase-pyruvate complex provide new insights into its catalytic mechanism. ChemistrySelect 2(36):11981–11988. 10.1002/slct.201702128
76. European Bioinformatics Institute (EBI) (2026) Acetolactate Synthase: Catalytic Site Atlas M-CSA ID: 289. CSA Database. https://www.ebi.ac.uk/thornton-srv/m-csa/entry/289/
77. European Bioinformatics Institute (EBI) (2026) Acetolactate Synthase: Catalytic Site Atlas M-CSA ID: 722. CSA Database. https://www.ebi.ac.uk/thornton-srv/m-csa/entry/722/
78. Sheals J, Persson P, Hedman B (2001) IR and EXAFS Spectroscopic Studies of Glyphosate Protonation and Copper(II) Complexes of Glyphosate in Aqueous Solution. Inorg Chem 40(17):4302–4309. 10.1021/ic000849g
79. Obeid G et al (2023) Solvation Effects on Glyphosate Protonation and Conformation: A First-Principles Study. J Chem Phys 158:054306. 10.1063/5.0134003
80. Meddeb-Limem S et al (2022) Hydrogen-Bonding Patterns and Dimerization of Glyphosate: A Combined DFT and Molecular Dynamics Study. J Mol Liq 120154. 10.1016/j.molliq.2022.120154
81. Van Rossum G (2007) Python programming language. In: USENIX ATC
82. Landrum G (2026) RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org. Accessed Feb 2026
83. Graff DE, Morgan NK, Burns JW, Doner AC, Li B, Li SC et al (2025) Chemprop v2: An Efficient, Modular Machine Learning Package for Chemical Property Prediction. J Chem Inf Model. 10.1021/acs.jcim.5c02332
84. Heid E, Greenman KP, Chung Y, Li SC, Graff DE, Vermeire FH et al (2023) Chemprop: a machine learning package for chemical property prediction. J Chem Inf Model 64(1):9–17. 10.1021/acs.jcim.3c01250

85. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: International conference on machine learning. PMLR, pp 1263–1272
86. Xu S, Xie L, Dai R, Lyu Z (2025) Dumpling GNN: hybrid GNN enables better ADC payload activity prediction based on the chemical structure. Int J Mol Sci 26(10):4859. 10.3390/ijms26104859
87. Falcon W (2019) PyTorch Lightning: A Lightweight PyTorch Wrapper for High-Performance AI Research. In: Proceedings of the NeurIPS 2019 Workshop on ML Systems, Vancouver, WA, Canada. Vol. 13. https://github.com/Lightning-AI/pytorch-lightning
88. Burns JW, Spiekermann KA, Bhattacharjee H, Vlachos DG, Green WH (2023) Machine learning validation via rational dataset sampling with astartes. J Open Source Softw 8(91):5996. 10.21105/joss.05996
89. Luo G (2016) A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw Model Anal Health Inf Bioinform 5(1):18. 10.1007/s13721-016-0125-6
90. Ji Z, Li J, Telgarsky M (2021) Early-stopped neural networks are consistent. Adv Neural Inf Process Syst 34:1805–1817
91. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A (2019) Descriptive statistics and normality tests for statistical data. Ann Card Anaesth 22(1):67–72. 10.4103/aca.ACA_157_18
92. Hehre WJ, Ditchfield R, Pople JA (1972) Self-consistent molecular orbital methods. XII. Further extensions of Gaussian-type basis sets for use in molecular orbital studies of organic molecules. J Chem Phys 56(5):2257–2261. 10.1063/1.1677527
93. Guevara-Vela JM, Rocha-Rinza T, Pendás ÁM (2017) Performance of the RI and RIJCOSX approximations in the topological analysis of the electron density. Theor Chem Acc 136(5):57. 10.1007/s00214-017-2084-0
94. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132(15). 10.1063/1.3382344
95. Mennucci B, Tomasi J (1997) Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries. J Chem Phys 106(12):5151–5158. 10.1063/1.473558
96. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):A1133. 10.1103/PhysRev.140.A1133
97. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32. 10.1023/A:1010933404324
98. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 785–794. 10.1145/2939672.2939785
99. Unterthiner T, Mayr A, Klambauer G, Steijaert M, Wegner JK, Ceulemans H, Hochreiter S (2014) Deep learning as an opportunity in virtual screening. In: Proceedings of the deep learning workshop at NIPS, vol 27. MIT Press, Cambridge, MA, pp 1–9

100. Martin EJ, Zhu XW, Riley P, Kearnes S, Sosnina EA, Tian L et al (2026) Comparing massively-multitask regression algorithms for drug discovery. J Comput Aided Mol Des 40(1):58
101. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43(6):1947–1958
102. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM (2016) Extreme gradient boosting as a method for quantitative structure-activity relationships. J Chem Inf Model 56(12):2353–2360
103. Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. pp 233–240
104. Scior T, Bender A, Tresadern G, Medina-Franco JL, Martínez-Mayorga K, Langer T et al (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52(4):867–881
105. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection,what can we learn from earlier mistakes? J Comput Aided Mol Des 22(3):213–228
106. Li H, Sze KH, Lu G, Ballester PJ (2021) Machine-learning scoring functions for structure-based virtual screening. Wiley Interdiscip Rev Comput Mol Sci 11(1):e1478

Figures

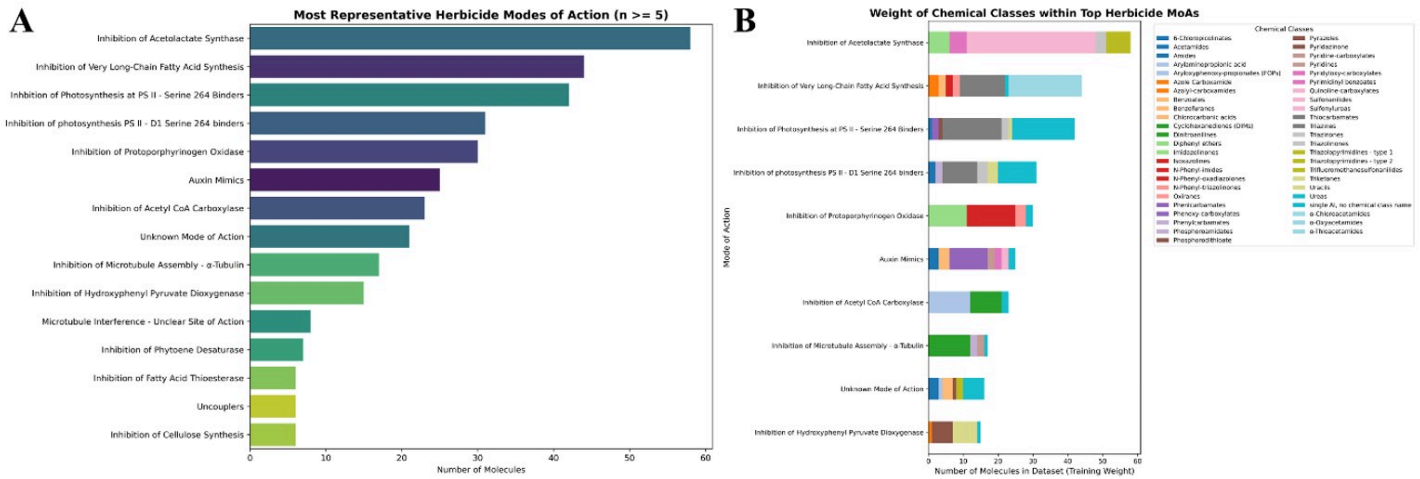


Figure 1

Dataset architecture and chemical diversity**.** (A) Frequency of herbicide modes of action (MoA) in the curated dataset. (B) Chemical class composition within the top MoAs, highlighting the prevalence of sulfonylureas in the Acetolactate Synthase (ALS) inhibition category.

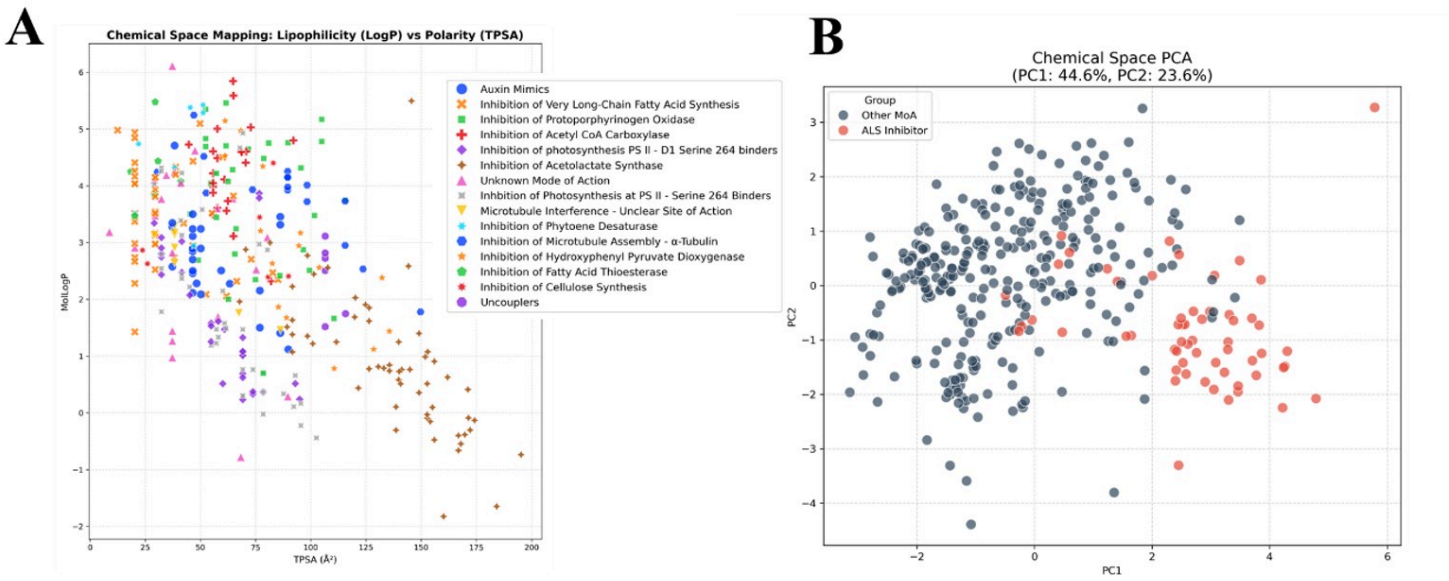


Figure 2

Chemical space characterization of herbicides. (A) Scatter plot of Lipophilicity (MolLogP) versus Polarity (TPSA) demonstrating the unique polar niche of ALS inhibitors. (B) Principal Component Analysis (PCA) showing the structural clustering of active candidates against the global herbicide space.

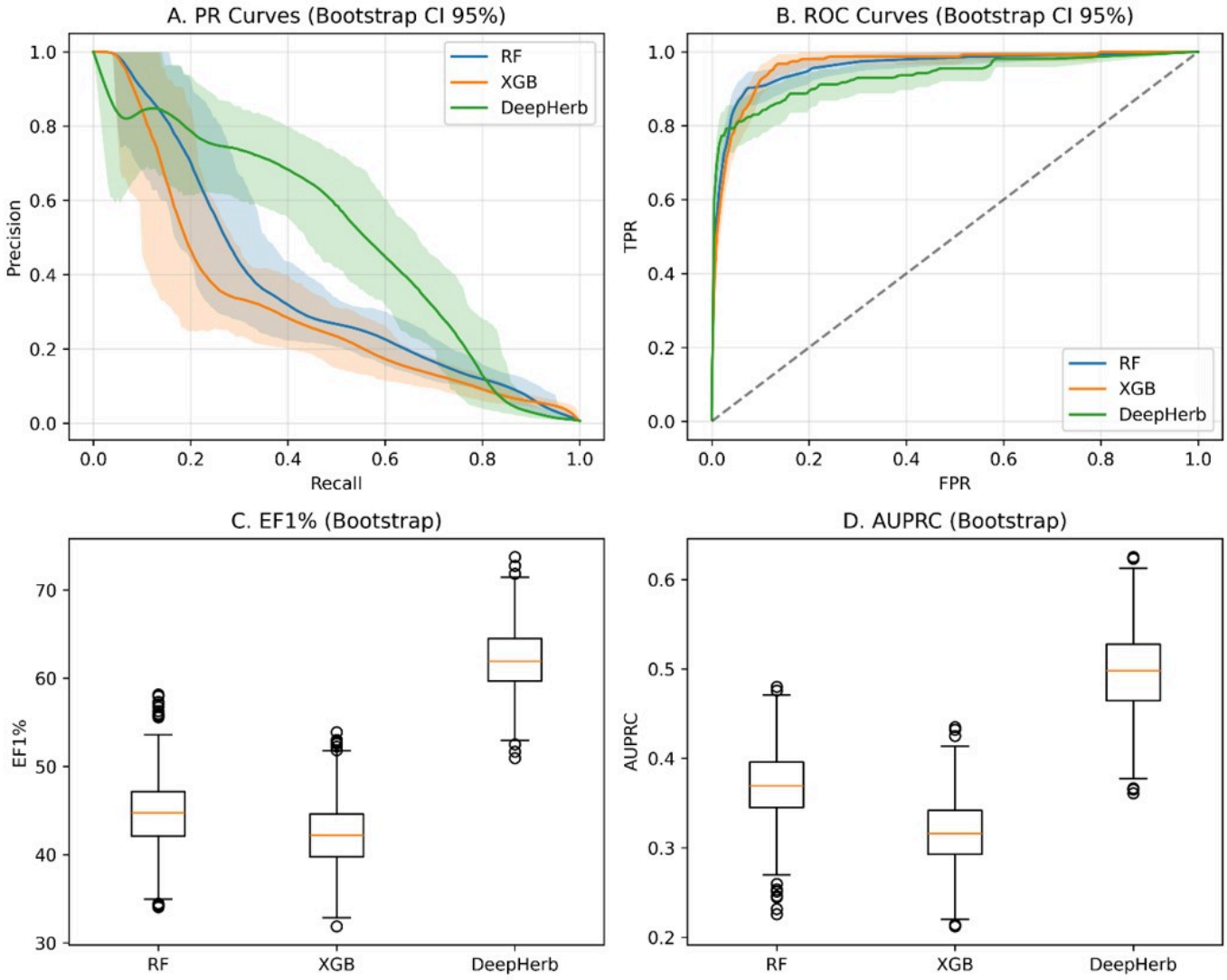


Figure 3

Comparative performance of HerbGNN and baseline models under scaffold-based cross-validation. (A) Precision-Recall (PR) curves showing mean performance with 95% bootstrap confidence intervals, highlighting early retrieval capability under severe class imbalance. (B) Receiver Operating Characteristic (ROC) curves with 95% bootstrap confidence intervals, illustrating global discriminative performance across models. (C) Distribution of early enrichment factor at 1% (EF1%) obtained by bootstrap resampling, reflecting the ability to prioritize active compounds in the top-ranked fraction. (D) Distribution of area under the precision-recall curve (AUPRC) obtained by bootstrap resampling, summarizing model performance in imbalanced classification scenarios.

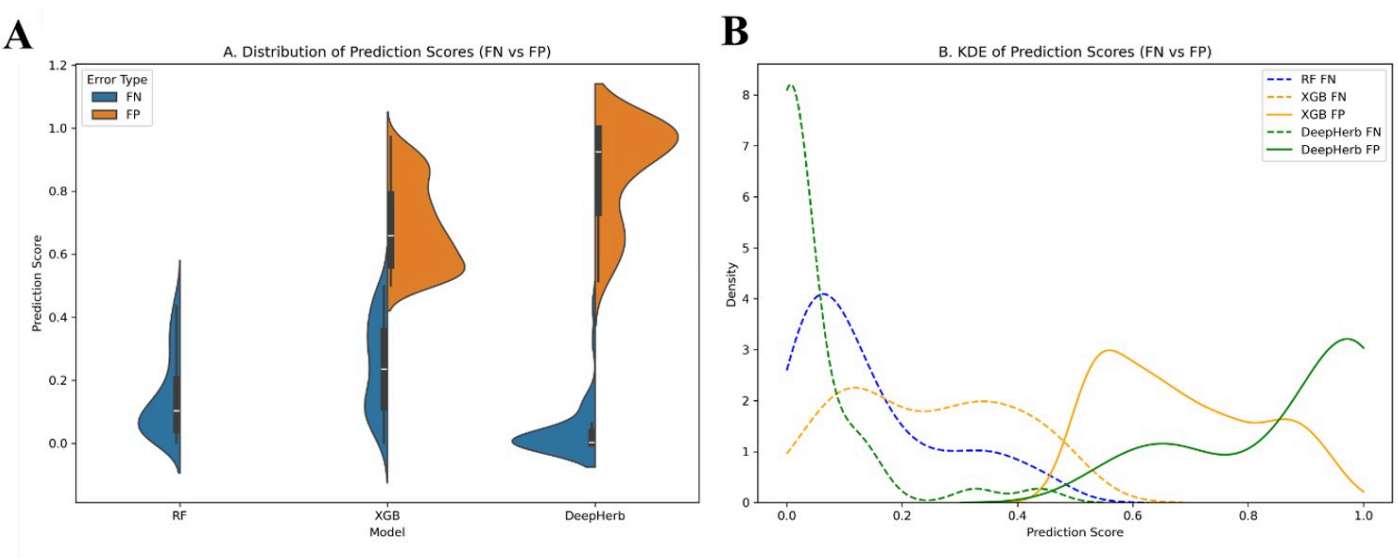


Figure 4

Error analysis of prediction score distributions across models. (A) Violin plots showing the distribution of prediction scores for false negatives (FN) and false positives (FP) in Random Forest (RF), XGBoost (XGB), and HerbGNN. HerbGNN exhibits a clear separation, with FN concentrated near zero and FP near one, whereas RF and XGB show broader and overlapping distributions. (B) Kernel density estimation (KDE) plots illustrating the asymmetric distribution of prediction scores.

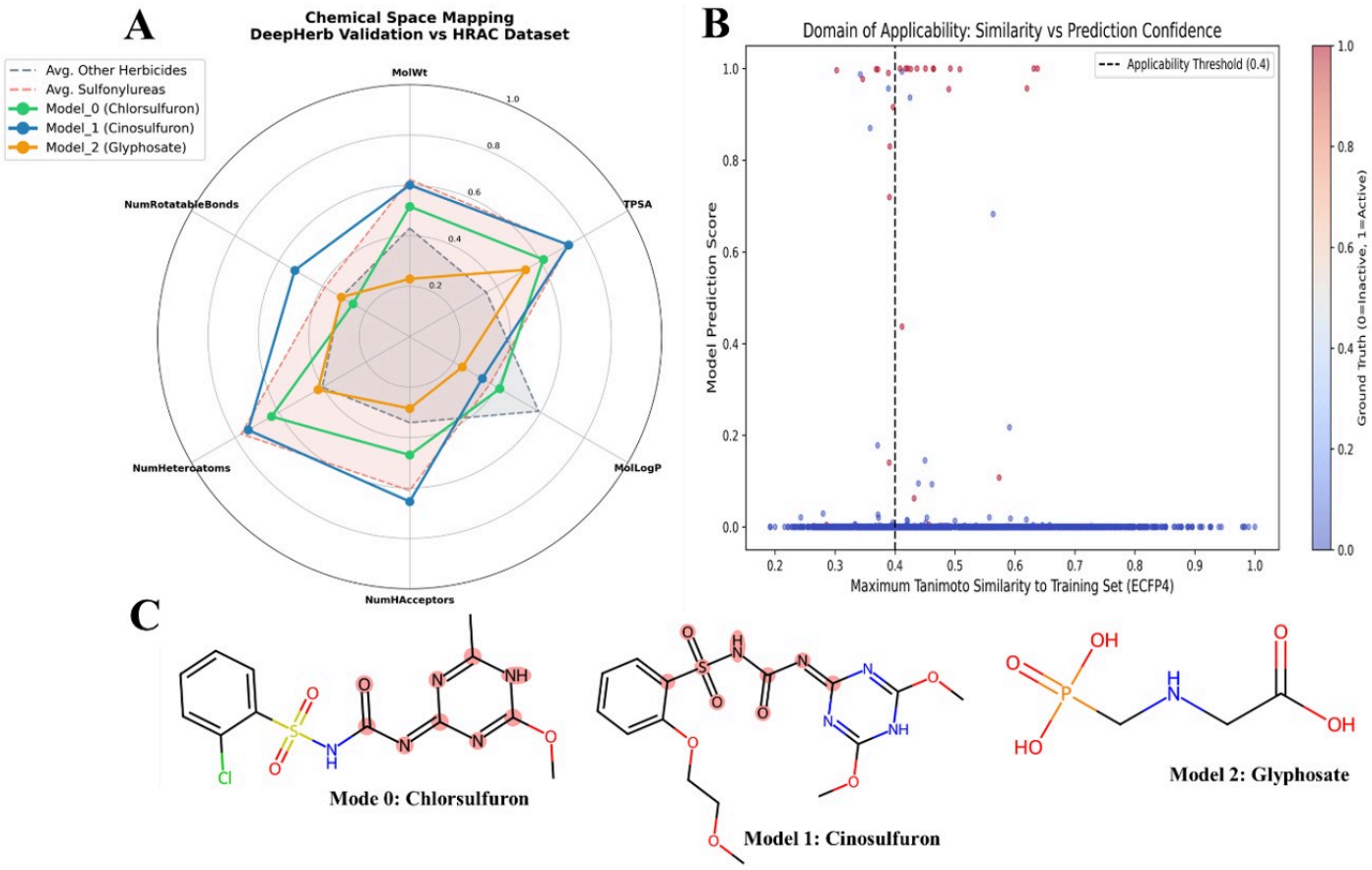


Figure 5

Chemoinformatic validation and interpretability of prioritized candidates. (A) Radar plot comparing normalized physicochemical descriptors (LogP, TPSA, MW, H-Bond Donors/Acceptors) of Model 0, Model 1, and Model 2 against the active set average. (B) Structural rationale identified by the D-MPNN; red highlights indicate atoms with the highest contribution to the herbicidal activity score, showing a focus on the sulfonylurea bridge and heterocyclic scaffolds.

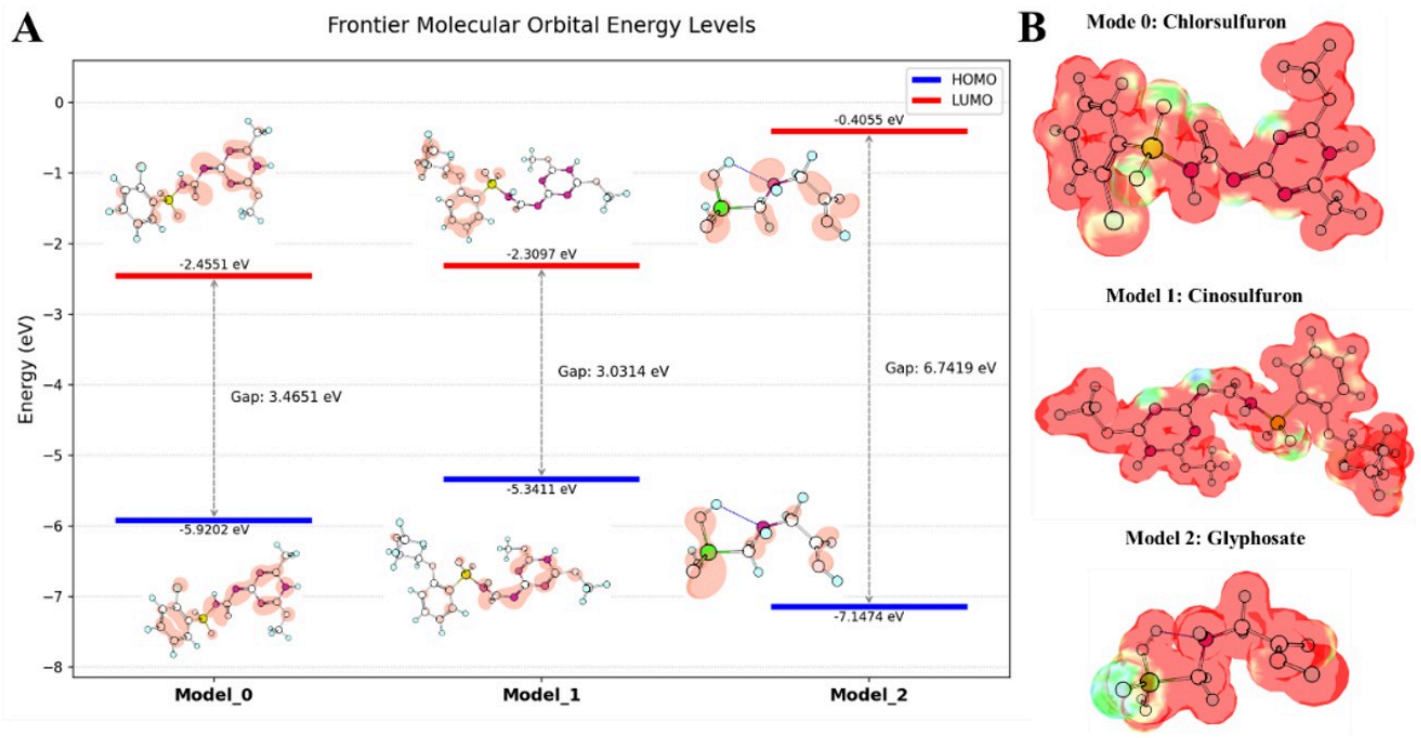


Figure 6

Mechanistic validation via Density Functional Theory (DFT). Comparative analysis of frontier molecular orbitals (E_{HOMO}-E_{LUMO}) (A) and Electrostatic Potential (ESP) maps (B) for active candidates and control, correlating electronic reactivity with GNN importance weights.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Graphicalabstract.jpg