ISSN: 2584-0231(Online)
International Journal of Multidisciplinary Research in Arts, Science and Technology
Volume 4 | Issue 4 | April 2026
Available online at: https://ijmrast.com
DOI: https://doi.org/10.59828/ijmrast.v4i4.257

Computational Screening of Plant-Derived Biopesticide Candidates Using Molecular Descriptors, Insect-Target Scoring, and Machine Learning

Shishir Tripathi[1] and Naincy Srivastava[2]

[1]Assistant Professor, Department of Zoology, Shri Lal Bahadur Shastri Degree College, Gonda, U.P., India
[2]Department of Zoology, Shri Lal Bahadur Shastri Degree College, Gonda, U.P., India
[1]Corresponding Author Email Id: *shishir8686@gmail.com*

Abstract

Synthetic insecticides remain effective, but rising resistance and harm to non-target species have renewed interest in plant-derived alternatives. Compounds such as azadirachtin, the pyrethrins, rotenone, and nicotine have shaped the history of pest control, and many more plant metabolites are likely to have insecticidal potential. To explore this chemical space systematically, we assembled 97 phytochemicals, 51 with reported insecticidal activity and 46 common plant metabolites without such activity, and processed them through an open-source cheminformatics pipeline. For each molecule we computed 18 physicochemical descriptors, scored similarity to known actives across eight insect molecular targets (acetylcholinesterase, juvenile hormone esterase, ecdysone receptor, chitin synthase, voltage-gated sodium channel, GABA-gated chloride channel, nicotinic acetylcholine receptor, and midgut aminopeptidase N), and trained five classifiers. Random Forest performed best (test accuracy = 0.92, ROC-AUC = 0.987). The most discriminating features were lipophilicity (LogP), hydrogen-bond donor count, and molar refractivity, exactly the properties that govern insect cuticle penetration. A consensus score combining target similarity, similarity to actives, and insecticide-likeness ranked Tephrosin, Limonin, Quassin, Nimbin, Rotenone, Pyrethrin II, Gedunin, and Deguelin at the top, all compounds with established insecticidal activity. The pipeline recovers known biology from a small, heterogeneous dataset and offers a transparent triage tool for prioritising plant compounds in early biopesticide discovery.

Keywords: *biopesticides; phytochemicals; virtual screening; molecular descriptors; machine learning; insect targets; azadirachtin; rotenoids; pyrethrins*

1. Introduction

Insects destroy roughly a fifth to two-fifths of global crop production each year (Savary et al., 2019). Synthetic insecticides have controlled this loss for decades, but heavy use has produced widespread resistance, declines in pollinators, and contamination of soil and water (Sparks and Nauen, 2015; Goulson, 2014). Biopesticides are pest-control agents derived from microorganisms or plants and have therefore moved from a niche option to a central part of integrated pest management (Senthil-Nathan, 2015; Isman, 2006).

Plant secondary metabolites are an especially rich source of insecticidal chemistry because they evolved as defences against herbivores. Their biological appeal rests on three pillars. First, several act on conserved nervous-system targets that exist in insects and mammals but differ enough between the two for selective toxicity: pyrethrins prolong the open state of the insect voltage-gated sodium channel; nicotine and the neonicotinoids derived from it act at nicotinic acetylcholine receptors; many botanicals inhibit acetylcholinesterase (Casida and Durkin, 2013). Second, limonoids such as azadirachtin disrupt insect-specific developmental pathways such as juvenile hormone titres, ecdysteroid signalling, and chitin synthesis, and so cause

moulting failure and growth arrest without parallel effects in vertebrates (Mordue and Blackwell, 1993; Schmutterer, 1990). Third, the midgut aminopeptidase N that anchors Bacillus thuringiensis Cry toxins is also affected by certain plant alkaloids and saponins (Pardo-López et al., 2013).

Identifying new candidates by bioassay alone is slow and costly. Cheminformatic screening can rank molecules first and bioassay only the most promising ones (Lo et al., 2018). In this study we present a short, open-source pipeline that combines molecular descriptors, ligand-similarity scoring against eight insect targets, and machine learning to prioritise plant-derived biopesticide leads. Our aim is not to claim a new active, but to test whether a transparent, biology-grounded workflow can recover the known canonical botanicals from a small, heterogeneous training set.

2. Materials and Methods

2.1 Compound dataset

Ninety-seven small molecules were curated from the literature, using Senthil-Nathan (2015) as the principal source for insecticidal class assignments. Compounds were labelled 1 if they had documented insecticidal, antifeedant, or growth-regulatory activity (n = 51) and 0 if they were widely distributed phytochemicals or primary metabolites without such activity (n = 46). The positive set included neem and citrus limonoids, rotenoids, natural pyrethrins, nicotinic alkaloids, capsaicin, piperine, sanguinarine, ricinine, several monoterpenes (thymol, carvacrol, eugenol, citronellal, linalool, geraniol), sesquiterpenes (β-caryophyllene, α-humulene), phenylpropanoids (apiole, myristicin), furanocoumarins (psoralen, xanthotoxin), and quercetin. The negative set was drawn from amino acids, sugars and sugar alcohols, common phenolic acids, organic acids, and vitamins (Figure 1). All structures were entered as SMILES, parsed and canonicalised with RDKit (Landrum, 2024), and duplicates were removed.

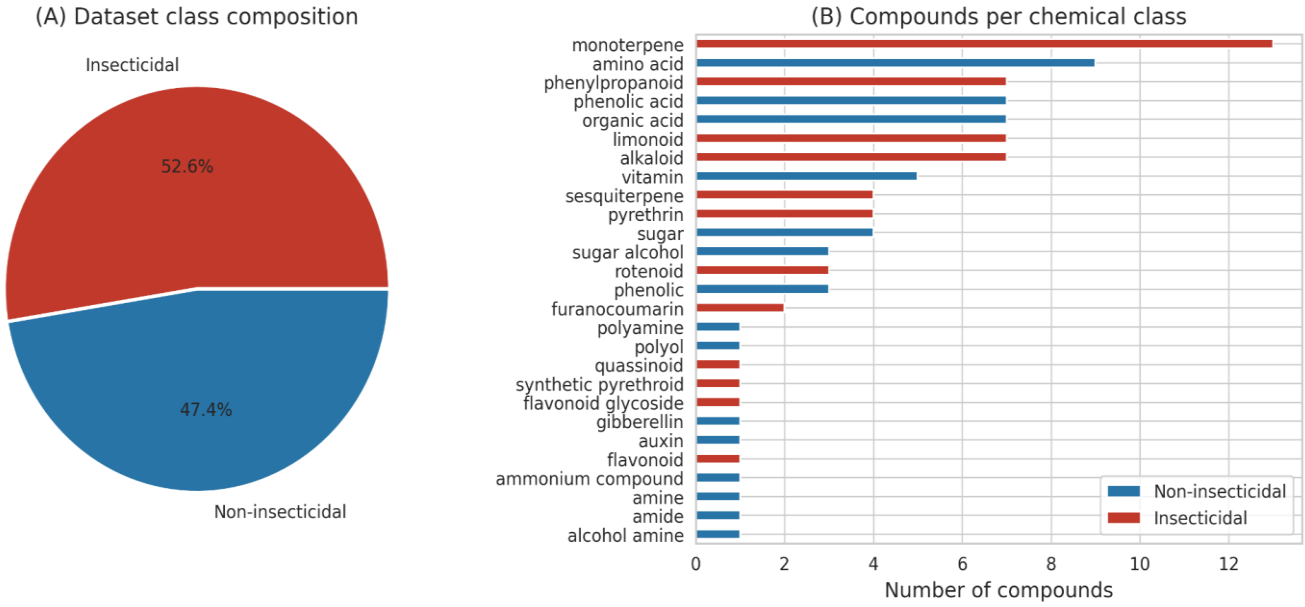

Figure 1. Composition of the curated dataset. (A) 51 insecticidal vs 46 non-insecticidal compounds. (B) Counts per chemical class.

2.2 Descriptors, targets, and scoring

Eighteen RDKit descriptors were computed per molecule: molecular weight, LogP (Crippen), hydrogen-bond donors (HBD) and acceptors (HBA), polar surface area (TPSA), rotatable bonds, aromatic and total ring counts, heavy atoms, molar refractivity, FractionCsp3, QED (Bickerton et al., 2012), heteroatom and stereocentre counts, saturated and aliphatic ring counts, BertzCT, and BalabanJ. Lipinski and Veber rule violations were also recorded, along with a five-criterion insecticide-likeness score (150 ≤ MolWt ≤ 500; 0 ≤ LogP ≤ 5; HBD ≤ 3; RotBonds ≤ 10; TPSA ≤ 120 Å^2) based on Tice (2001) and Hao et al. (2011). For machine learning, descriptors were complemented with 2048-bit Morgan circular fingerprints of radius 2 (Rogers and Hahn, 2010).

Eight insect targets were chosen to cover the major modes of action of botanical insecticides: acetylcholinesterase (AChE), juvenile hormone esterase (JHE), ecdysone receptor (EcR), chitin synthase (CHS), voltage-gated sodium channel (Na_v), GABA-gated chloride channel, nicotinic acetylcholine receptor (nAChR), and midgut aminopeptidase N. UniProt accessions and key residues are provided in supplementary file 02_insect_targets.csv. For each target, a small reference panel of confirmed

active ligands was assembled from the literature (for example, paraoxon and eserine for AChE; permethrin and the natural pyrethrins for Na_v; nicotine, imidacloprid, and acetamiprid for nAChR). Every compound was scored against every target as the maximum Tanimoto similarity of its Morgan fingerprint to any reference ligand, with a descriptor-based penalty for compounds far outside the property range of the reference set. The largest Tanimoto similarity to any insecticidal compound in the dataset (MaxTanimotoToActives) and the insecticide-likeness score were then combined with the per-target scores into a single ConsensusScore (equal weights, min–max normalised to [0, 1]).

2.3 Statistical analysis and machine learning

Descriptor distributions in the two groups were compared with the Mann–Whitney U test, with effect sizes given by Cohen's d. Compounds were projected into three principal components on standardised descriptors (scikit-learn; Pedregosa et al., 2011), and k-means clustering with silhouette scoring was used to probe the chemical space. Five classifiers i.e. Random Forest, Gradient Boosting, Logistic Regression, SVM (RBF), and XGBoost, were trained on the combined descriptor and fingerprint matrix. Performance was estimated by stratified five-fold cross-validation; the best model was further evaluated on a 75/25 stratified held-out split.

3. Results

3.1 Insecticidal compounds occupy a distinct region of chemical space

The two groups differed strongly across most descriptors (Table 1). Insecticidal molecules were larger (mean MolWt ≈ 250 vs 165 Da), more lipophilic (mean LogP ≈ 2.74 vs −0.80), less polar (TPSA ≈ 45 vs 88 Å^2), and had fewer hydrogen-bond donors (≈ 0.6 vs 3.2). Twelve of the eighteen descriptors were significantly different at $\alpha = 0.05$, with ten showing large Cohen's d ($|d| \geq 0.8$). The largest single effect was for LogP ($|d| = 2.25$), followed by HBD ($|d| = 1.63$) and molar refractivity ($|d| = 1.35$). These shifts have a clear biological reading: insecticides must cross the lipid-rich cuticle and reach internal targets, so moderately lipophilic molecules with few polar groups have a built-in advantage, while small polar primary metabolites are largely excluded from insect tissues at biologically meaningful rates.

Table 1. Most discriminating descriptors (Mann–Whitney U test; Cohen's d).

Descriptor	Mean (Insect.)	Mean (Non)	p-value	Cohen's d	Effect
LogP	2.74	-0.80	1.3e-14	2.25	Large
HBD	0.59	3.24	2.1e-14	-1.63	Large
MolarRefract.	69.7	38.6	1.4e-10	1.35	Large
RingCount	2.31	0.72	5.8e-08	1.14	Large
TPSA	45.0	88.4	2.1e-08	-0.99	Large
BertzCT	499.8	222.4	2.6e-07	0.95	Large
MolWt	250.1	164.9	1.8e-05	0.84	Large
QED	0.554	0.442	3.5e-06	0.88	Large

Principal-component analysis on the standardised descriptors gave a clear visual separation between the two groups along PC1, which loads on size, lipophilicity, and ring count (Figure 2). K-means clustering returned the highest silhouette score at k = 2 (0.415), confirming that the dataset is driven by a single biological dichotomy rather than by finer chemical-class structure.

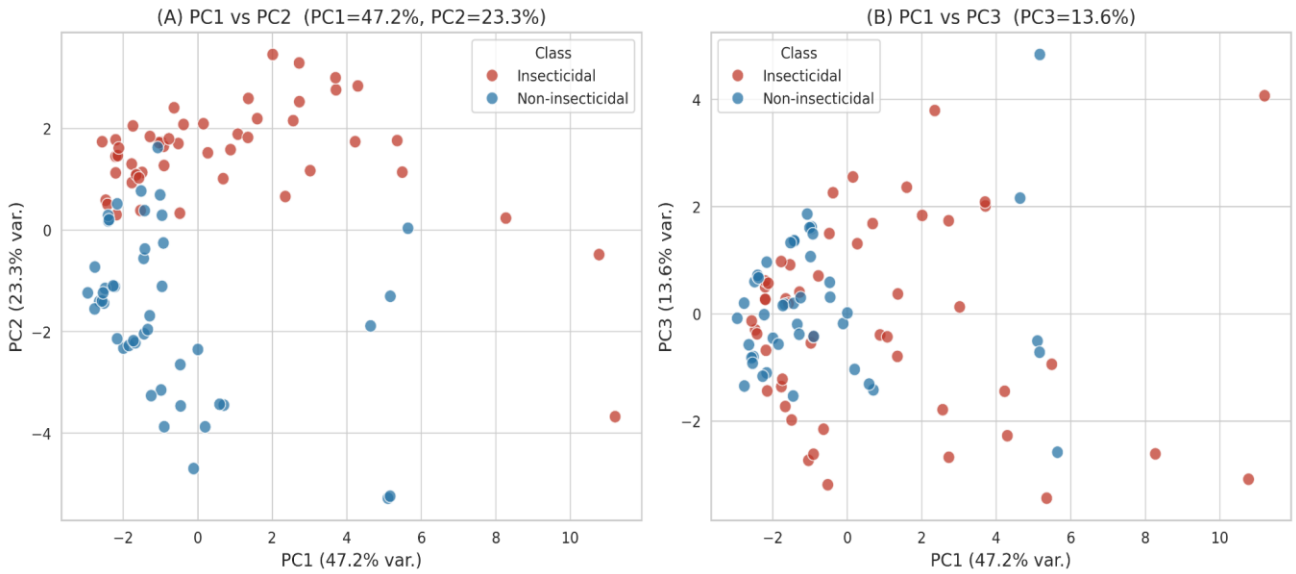

Figure 2. PCA projection of the 97 compounds on standardised molecular descriptors. PC1 separates insecticidal from non-insecticidal compounds.

3.2 Random Forest separates the two groups with high accuracy

Across the five classifiers, Random Forest gave the strongest cross-validation performance (accuracy 0.99 ± 0.02; ROC-AUC 0.998 ± 0.004) and held up on the 75/25 held-out test split (accuracy = 0.92; F1 = 0.92; ROC-AUC = 0.987; confusion matrix with one false positive and one false negative out of 25 test compounds; Figure 3). Logistic Regression and XGBoost came close; Gradient Boosting and SVM (RBF) performed less well, mainly because of their greater sensitivity to the small training set.

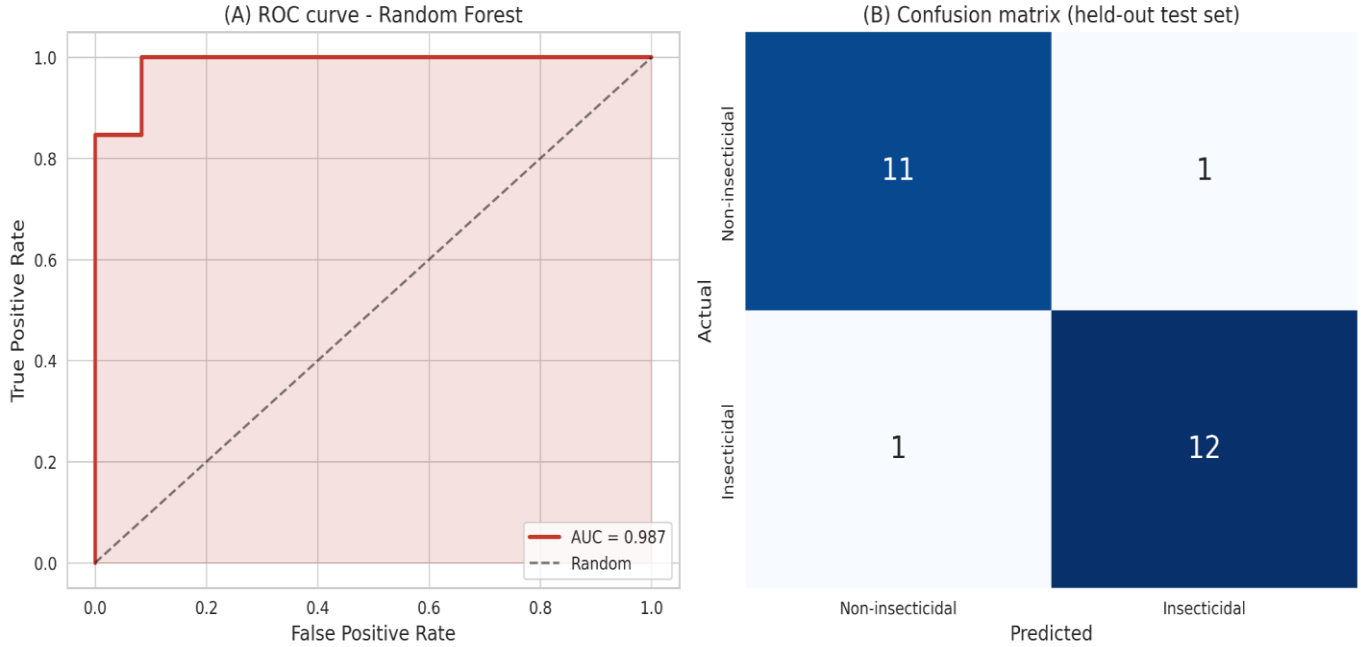

Figure 3. Random Forest performance on the held-out test set. Left: ROC curve (AUC = 0.987). Right: confusion matrix.

Feature importances from the trained Random Forest were dominated by classical physicochemical descriptors rather than by individual fingerprint bits: LogP (0.087), HBD (0.052), molar refractivity (0.034), BalabanJ (0.028), heavy atoms (0.027), BertzCT (0.025), TPSA (0.022), QED (0.022), and MolWt (0.020) made up the top ten. Six Morgan fingerprint bits also entered the top 25, but bulk physicochemistry led the ranking. The features that mattered to the model are the same ones that matter biologically: they govern whether a molecule can reach an insecticidal target at all.

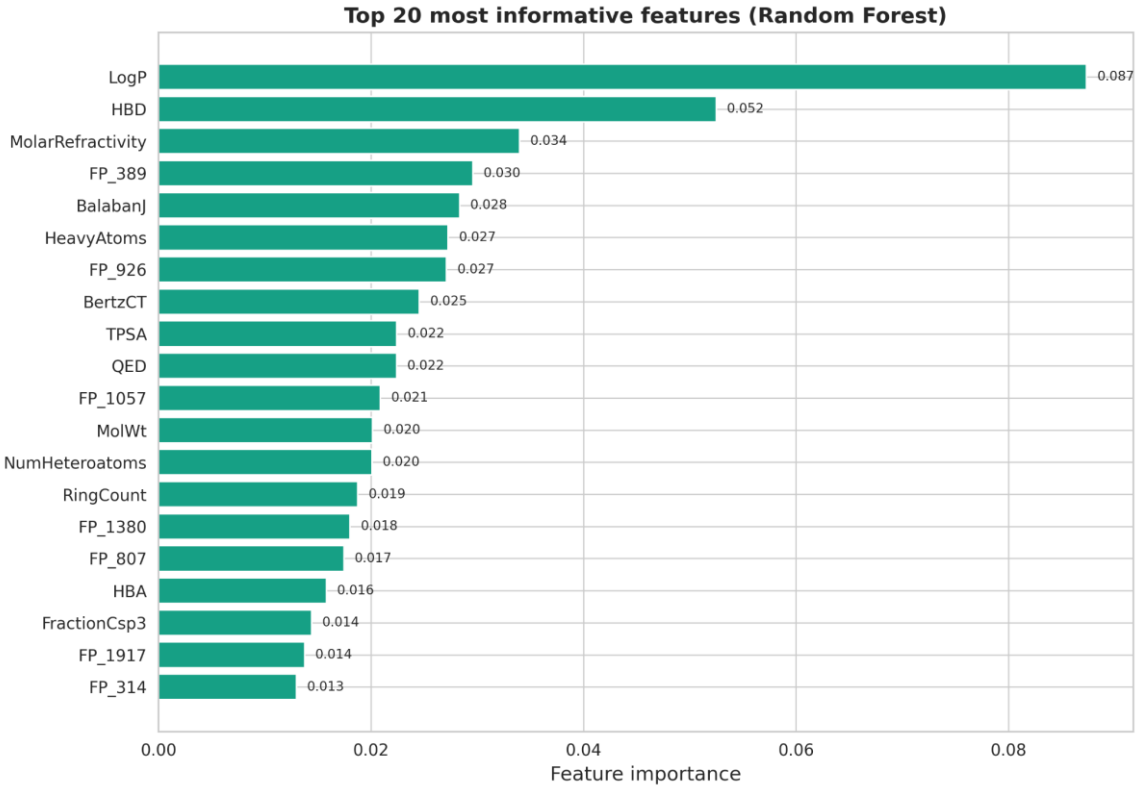

Figure 4. Top features ranked by mean decrease in Gini impurity in the Random Forest classifier.

3.3 Consensus scoring recovers the canonical botanical insecticides

Per-target scores showed clear class-level patterns (Figure 5). Limonoids and rotenoids scored most strongly against AChE and JHE; pyrethrins scored highest at the voltage-gated sodium channel; nicotinic and other alkaloids scored at nAChR; monoterpenes spread their predicted activity across several targets, in line with their known multi-target mode of action (Regnault-Roger et al., 2012). Non-insecticidal primary metabolites scored low across all eight targets.

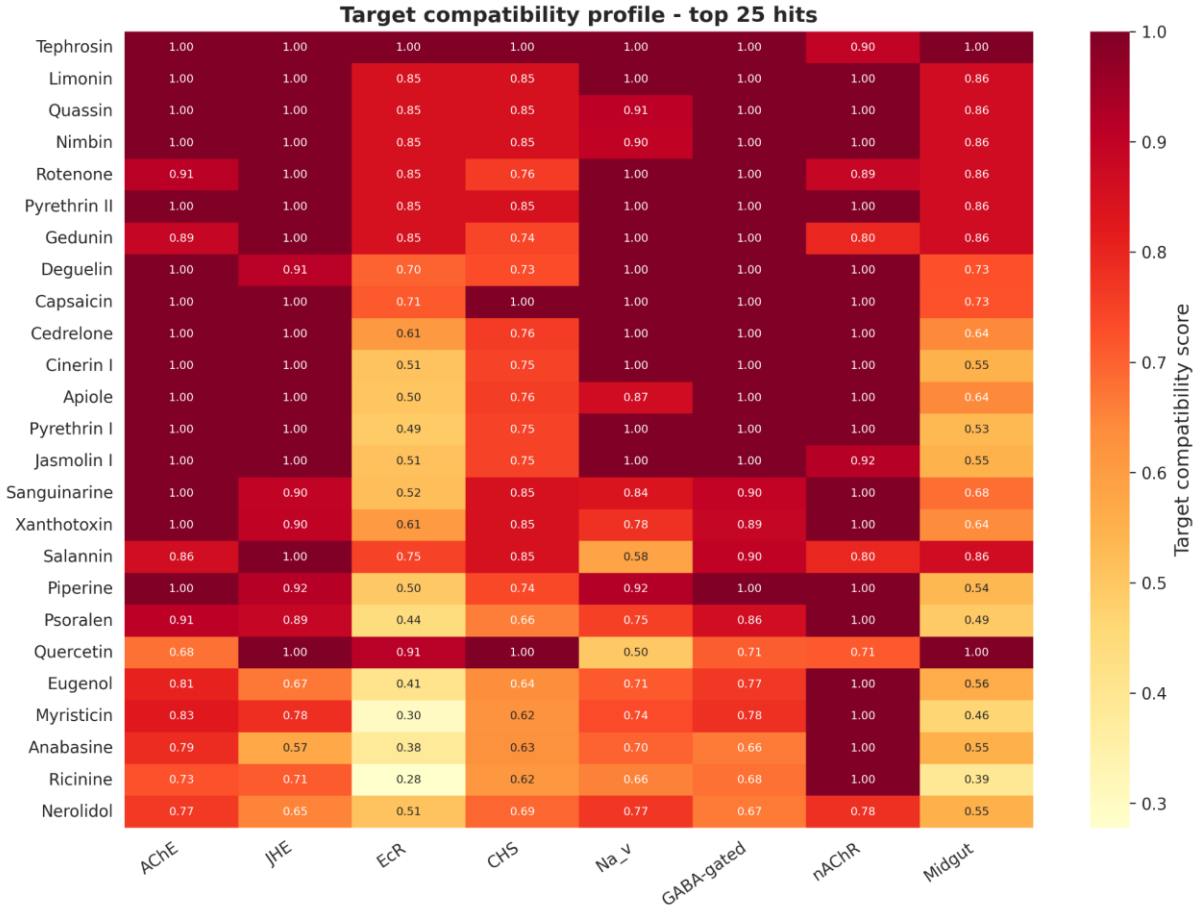

Figure 5. Per-compound, per-target score heatmap. Darker red indicates a stronger predicted interaction.

The ConsensusScore distribution was clearly bimodal (Supplementary Figure S1), and the top 20 ranked compounds i.e. Tephrosin, Limonin, Quassin, Nimbin, Rotenone, Pyrethrin II, Gedunin, Deguelin, Capsaicin, Cedrelone, Cinerin I, Apiole, Pyrethrin I, Jasmolin I, Sanguinarine, Xanthotoxin, Salannin, Piperine, Psoralen, and Quercetin, were all drawn from the positive label set. No primary metabolite or known non-insecticide reached the cut-off. The top five (Table 2) include the two iconic neem-family limonoids Limonin and Nimbin, the bitter quassinoid Quassin, the rotenoid Tephrosin, and rotenone itself. Their structures are shown in Figure 6.

Table 2. Top five compounds by ConsensusScore.

#	Compound	Class	Plant source	Best target	Score
1	Tephrosin	rotenoid	Tephrosia	AChE	1.000
2	Limonin	limonoid	Citrus/Meliaceae	AChE	0.981
3	Quassin	quassinoid	Quassia amara	AChE	0.976
4	Nimbin	limonoid	Azadirachta indica	AChE	0.976
5	Rotenone	rotenoid	Derris/Lonchocarpus	JHE	0.965

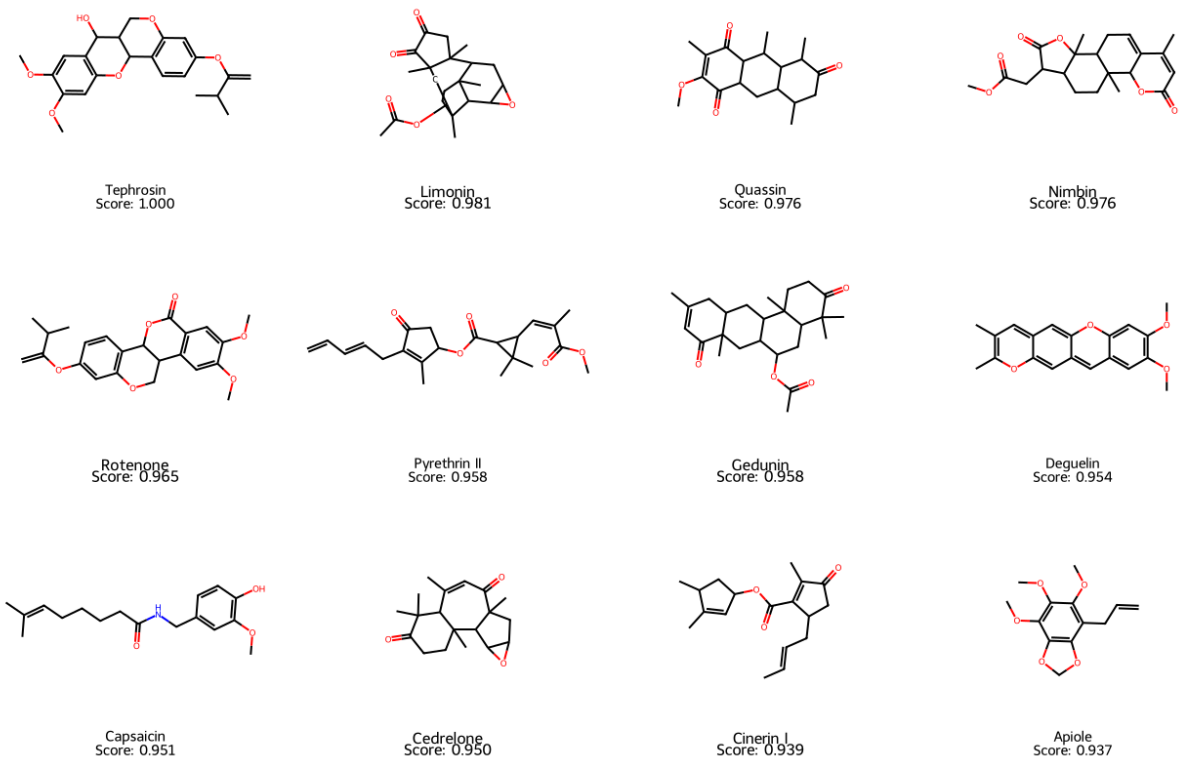

Figure 6. Two-dimensional structures of the top consensus hits.

4. Discussion

The top 20 consensus hits read like a roll-call of well-known botanical insecticides, which is what a successful retrospective virtual screen should produce. The neem-family limonoids — Nimbin, Salannin, Gedunin, Limonin, Cedrelone, and the related quassinoid Quassin — are structural relatives of azadirachtin and act through three overlapping mechanisms: antifeedant deterrence at chemosensory neurons, disruption of juvenile-hormone and ecdysteroid signalling leading to abnormal moulting, and direct cytotoxicity to midgut epithelial cells (Mordue and Blackwell, 1993; Schmutterer, 1990; Senthil-Nathan, 2013). The pipeline assigned most of them AChE or JHE as their best target, consistent with the documented inhibition of head-region acetylcholinesterase activity in Lepidoptera treated with neem (Senthil-Nathan et al., 2006). The three rotenoids (Tephrosin, Rotenone, Deguelin) actually inhibit mitochondrial complex I (Hollingworth and Tomalski, 1990), which was not in our target panel; their high rank therefore comes from structural similarity to other large polycyclic

actives, illustrating both the value and a limitation of similarity-based scoring. The four pyrethrins (I, II, Cinerin I, Jasmolin I) are well established as voltage-gated sodium channel modulators (Soderlund, 2012), and their physicochemical profile sits cleanly in insecticide-like property space.

Among the alkaloids and phenylpropanoids, Capsaicin activates the TRPV1 channel and shows insect repellency and toxicity (Regnault-Roger et al., 2012); Sanguinarine is a documented acetylcholinesterase inhibitor (Schmeller et al., 1997); and Piperine, the pungent principle of black pepper, both acts directly on insects and synergises with pyrethrins (Scott et al., 2008). Apiole, myristicin, and the furanocoumarins psoralen and xanthotoxin are classical photoactive defensive metabolites that increase in plant tissues under herbivore attack (Berenbaum, 1991). Quercetin's appearance at rank 20 reflects its established antifeedant and growth-modulatory role across many insect species (Simmonds, 2003) without putting it in the same league as the cuticle-penetrating polycyclic actives.

The descriptor analysis ties these biological observations to physicochemical first principles. The strongest single discriminator was LogP, and the next most important features were HBD, molar refractivity, and ring complexity. These are exactly the properties that govern whether a molecule can cross the wax-rich insect cuticle or the peritrophic membrane and reach internal targets. A polar primary metabolite such as glutamic acid simply cannot enter insect tissues fast enough to be toxic, no matter what its in-vitro affinity might be. The ranges that worked well are MolWt 200–500 Da, LogP 1–5, HBD ≤ 3, at least one ring system, closely match the empirical insecticide-likeness filters of Tice (2001) and Hao et al. (2011), reinforcing the idea that botanical insecticides occupy a definable region of property space that can be flagged before any wet-lab work.

5. Limitations

The dataset is small (n = 97) and biased toward heavily studied classes. The negative set is composed of primary metabolites, which makes the classification task structurally easy and probably inflates apparent performance; a harder negative set drawn from non-insecticidal plant secondary metabolites would be a stiffer test. Target scoring is ligand-based, so a high score against AChE indicates structural similarity to AChE inhibitors, not experimentally confirmed binding. Several biologically important targets such as mitochondrial complex I, ryanodine receptors, glutamate-gated chloride channels, and TRPV channels, were not in our panel. All predictions require wet-lab validation, and the pipeline does not address non-target effects on pollinators, beneficial insects, or mammals.

6. Conclusions

A short, open cheminformatics pipeline built from RDKit descriptors, ligand-similarity scoring against eight insect targets, and a Random Forest classifier was able to recover the canonical plant-derived insecticides (azadirachtin-family limonoids, rotenoids, pyrethrins, capsaicin, piperine, sanguinarine, quercetin, and others) from a small, heterogeneous training set. Lipophilicity, hydrogen-bond donor count, and molar refractivity were the most discriminating descriptors, in line with the long-established insecticide-likeness rules. The framework is transparent, reproducible, and easy to extend with new targets or compounds. It is best used as a fast triage step that reduces the size of phytochemical libraries before in-vivo bioassays, not as a replacement for them.

References

[1]. Berenbaum, M. R. (1991). Coumarins. In G. A. Rosenthal & M. R. Berenbaum (Eds.), Herbivores: Their Interactions with Secondary Plant Metabolites (2nd ed., Vol. 1, pp. 221–249). Academic Press.

[2]. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. Nature Chemistry, 4(2), 90–98. https://doi.org/10.1038/nchem.1243

[3]. Casida, J. E., & Durkin, K. A. (2013). Neuroactive insecticides: Targets, selectivity, resistance, and secondary effects. Annual Review of Entomology, 58, 99–117.

[4]. Goulson, D. (2014). An overview of the environmental risks posed by neonicotinoid insecticides. Journal of Applied Ecology, 51(4), 977–987.

[5]. Hao, G., Dong, Q., & Yang, G. (2011). A comparative study on the constitutive properties of marketed pesticides. Molecular Informatics, 30(6–7), 614–622.

[6]. Hollingworth, R. M., & Tomalski, M. D. (1990). Inhibitors of NADH:ubiquinone reductase as insecticides. ACS Symposium Series, 356, 65–82.
[7]. Isman, M. B. (2006). Botanical insecticides, deterrents, and repellents in modern agriculture and an increasingly regulated world. Annual Review of Entomology, 51, 45–66.
[8]. Landrum, G. (2024). RDKit: Open-source cheminformatics. https://www.rdkit.org
[9]. Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. Drug Discovery Today, 23(8), 1538–1546.
[10]. Mordue (Luntz), A. J., & Blackwell, A. (1993). Azadirachtin: An update. Journal of Insect Physiology, 39(11), 903–924.
[11]. Pardo-López, L., Soberón, M., & Bravo, A. (2013). Bacillus thuringiensis insecticidal three-domain Cry toxins. FEMS Microbiology Reviews, 37(1), 3–22.
[12]. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[13]. Regnault-Roger, C., Vincent, C., & Arnason, J. T. (2012). Essential oils in insect control: Low-risk products in a high-stakes world. Annual Review of Entomology, 57, 405–424.
[14]. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50(5), 742–754.
[15]. Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution, 3(3), 430–439.
[16]. Schmeller, T., Latz-Brüning, B., & Wink, M. (1997). Biochemical activities of berberine, palmatine and sanguinarine. Phytochemistry, 44(2), 257–266.
[17]. Schmutterer, H. (1990). Properties and potential of natural pesticides from the neem tree, Azadirachta indica. Annual Review of Entomology, 35, 271–297.
[18]. Scott, I. M., Jensen, H. R., Philogène, B. J. R., & Arnason, J. T. (2008). A review of Piper spp. (Piperaceae) phytochemistry, insecticidal activity and mode of action against insects. Phytochemistry Reviews, 7(1), 65–75.
[19]. Senthil-Nathan, S. (2013). Physiological and biochemical effect of neem and other Meliaceae plants secondary metabolites against Lepidopteran insects. Frontiers in Physiology, 4, 359.
[20]. Senthil-Nathan, S. (2015). A review of biopesticides and their mode of action against insect pests. In P. Thangavel & G. Sridevi (Eds.), Environmental Sustainability (pp. 49–63). Springer India.
[21]. Senthil-Nathan, S., Kalaivani, K., & Murugan, K. (2006). Effects of neem limonoids on lipid peroxidation and detoxification enzymes in Cnaphalocrocis medinalis. Chemosphere, 64(10), 1650–1658.
[22]. Simmonds, M. S. J. (2003). Flavonoid–insect interactions: Recent advances. Phytochemistry, 64(1), 21–30.
[23]. Soderlund, D. M. (2012). Molecular mechanisms of pyrethroid insecticide neurotoxicity. Archives of Toxicology, 86(2), 165–181.
[24]. Sparks, T. C., & Nauen, R. (2015). IRAC: Mode of action classification and insecticide resistance management. Pesticide Biochemistry and Physiology, 121, 122–128.
[25]. Tice, C. M. (2001). Selecting the right compounds for screening: Does Lipinski's Rule of 5 for pharmaceuticals apply to agrochemicals? Pest Management Science, 57(1), 3–16.

Cite this Article:

Tripathi, S., & Srivastava, N. (2026). Computational screening of plant-derived biopesticide candidates using molecular descriptors, insect-target scoring, and machine learning. International Journal of Multidisciplinary Research in Arts, Science and Technology (IJMRAST), 4(4), 106–113.

***Journal URL**: https://ijmrast.com/ **DOI:** https://doi.org/10.59828/ijmrast.v4i4.257*