Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

In-Silico Prediction of Nitrosamine Risks in Herbal Drugs Via Quantum Chemistry and Machine Learning

Pulkit Baliyan[1]* and K. Ashok Kumar[2]

[1]Research Scholar, Faculty of Pharmaceutical Sciences, Motherhood University, Roorkee, Uttarakhand, India
[2]Research Supervisor and Professor, Faculty of Pharmaceutical Sciences, Motherhood University, Roorkee, Uttarakhand, India

Received: 20 Aug 2025 / Accepted: 25 Sept 2025 / Published online: 01 Oct 2025
*Corresponding Author Email: pulkitbaliyan98@gmail.com

ABSTRACT

Nitrosamine impurities are probable human carcinogens whose presence in pharmaceutical and herbal products has become a major global regulatory concern following high-profile contamination incidents in synthetic drug substances. While regulatory frameworks such as ICH M7(R2) and the associated FDA and EMA guidance documents primarily address synthetic active pharmaceutical ingredients, herbal drug constituents that carry secondary or tertiary amine functionalities present an analogous, comparatively under-studied risk. This study develops and demonstrates an in-silico screening framework that combines cheminformatic molecular descriptor generation, structure-based force-field geometry optimization, literature-informed structural risk scoring, and supervised machine learning classification to rank the nitrosamine-formation risk of twenty phytoconstituents drawn from 20 pharmacologically important herbal drugs (Ephedra sinica, Piper nigrum, Areca catechu, Nicotiana tabacum, Withania somnifera, Coptis chinensis, Berberis vulgaris, Atropa belladonna, Hyoscyamus niger, and Glycyrrhiza glabra). Molecular descriptors were generated using RDKit, and three-dimensional geometries were optimized using the MMFF94 force field. A structural risk score derived from established N-nitrosation chemistry (amine class, steric accessibility, and Gasteiger partial-charge electronics) was used to assign each compound to High-, Moderate-, or Low-risk classes, which in turn served as the target variable for three supervised classifiers Random Forest, Logistic Regression, and Support Vector Machine validated using leave-one-out cross-validation given the constrained sample size. Logistic Regression and SVM achieved the best leave-one-out accuracy (0.95), correctly separating quaternary protoberberine alkaloids and amine-free flavonoids (low risk) from secondary- and tertiary-amine-bearing tropane, pyrrolidine, and piperidine alkaloids (moderate-to-high risk). Cuscohygrine, nornicotine, and ephedrine emerged as the highest-priority constituents for downstream experimental and quantum-chemical validation. Density Functional Theory (B3LYP/6-31G(d) single-point calculations, PySCF) and AutoDock Vina molecular docking against CYP3A4 and CYP2C9 were subsequently performed for all twenty compounds: the HOMO-LUMO gap correlated with the structural risk score (Spearman $\rho = 0.83$), and adding these descriptors to the classifier did not improve leave-one-out accuracy over the baseline descriptor set, indicating the baseline descriptors already capture most of the separable structural signal at this sample size. A detailed protocol for the remaining transition-state/thermodynamic and manufacturing-stress phases of the research as specified in the approved research methodology is presented as a validated roadmap for Phase 2 of this programme. This paper positions descriptor-driven machine learning as a rapid, resource-efficient first-line triage tool that can prioritize herbal constituents for the more computationally intensive DFT and docking studies, thereby making efficient use of limited high-performance-computing resources in nitrosamine risk assessment of botanical drug products.

KEY WORDS: Nitrosamine impurities; herbal drugs; in-silico prediction; machine learning; molecular descriptors; QSAR; ICH M7; risk assessment; RDKit.

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

1. INTRODUCTION

The detection of N-nitrosodimethylamine (NDMA) and related N-nitrosamine impurities in valsartan and other sartan-class pharmaceuticals from 2018 onward triggered a worldwide re-examination of nitrosamine risk across the pharmaceutical supply chain. Regulatory authorities, including the US Food and Drug Administration and the European Medicines Agency, subsequently issued guidance requiring manufacturers to conduct risk assessments for nitrosamine formation in drug substances and products, formalized internationally through ICH M7(R2) and its nitrosamine-specific addenda. N-nitrosamines form when a nitrosating agent commonly nitrite ion or nitrogen oxides arising from raw materials, water, or environmental sources reacts with a secondary or tertiary amine present in an active ingredient, excipient, or process intermediate.

Herbal and botanical drug products have received comparatively little systematic attention in this regulatory conversation, despite the fact that many pharmacologically important phytoconstituents alkaloids in particular carry the same secondary- and tertiary-amine functionalities that drive nitrosamine formation in synthetic drugs. Herbal raw materials are also more variable in composition, are exposed to a wider range of processing conditions (drying, fermentation, extraction pH, storage humidity), and are frequently combined with excipients whose own nitrosating potential is not always characterized, all of which can plausibly influence nitrosamine formation risk.

This study addresses that gap by developing a computational triage framework specifically for herbal drug constituents. The framework proceeds in two phases. Phase 1, reported here in full with completed data analysis, uses cheminformatic descriptor generation, force-field geometry optimization, and machine-learning classification to rank twenty phytoconstituents drawn from 20 herbal drugs by their structural propensity for nitrosamine formation. Phase 1 additionally includes DFT electronic-structure calculations and molecular docking against CYP3A4/CYP2C9 for all twenty compounds, performed with open-source tools and reported alongside the descriptor/machine-learning results (Sections 3.4, 5.7–5.9). The remaining Phase 2 roadmap, presented as a validated protocol building on the Phase 1 rankings, specifies the transition-state/thermodynamic modeling and manufacturing-condition stress-testing required to confirm the Phase 1 predictions with the higher-fidelity, higher-cost experimental methods to which the research methodology (Chapter 3) commits.

1.1 Objectives of the Study

- To compile a defined dataset of nitrogen-bearing (and reference nitrogen-free) phytoconstituents representative of 20 pharmacologically significant herbal drugs.
- To generate physicochemical, topological, and electronic molecular descriptors for each constituent using validated cheminformatics tools (RDKit).
- To perform three-dimensional geometry optimization of each structure using force-field energy minimization as a preparatory step for higher-level quantum-chemical analysis.
- To derive a literature-informed structural nitrosamine-risk score and corresponding three-class risk label (High/Moderate/Low) for each compound.
- To train, cross-validate, and compare multiple supervised machine learning classifiers (Random Forest, Logistic Regression, SVM) for predicting nitrosamine-formation risk class from molecular descriptors.
- To identify the molecular descriptors most predictive of risk class and rank the twenty compounds accordingly.
- To perform Density Functional Theory calculations and molecular docking against representative cytochrome P450 enzymes for all twenty compounds, and to specify a detailed, defensible Phase 2 protocol transition-state/thermodynamic modeling and manufacturing/excipient stress analysis for the remaining experimental and quantum-chemical confirmation of the Phase 1 rankings.

1.2 Scope and Novelty

To the authors' knowledge, no published in-silico framework specifically prioritizes herbal secondary metabolites for nitrosamine-formation risk using a combined descriptor/machine-learning approach validated by leave-one-out cross-validation on a curated, chemically diverse alkaloid set spanning primary, secondary, tertiary, and quaternary nitrogen environments alongside amine-free reference compounds. This design allows the model's discriminatory power to be assessed against chemically meaningful negative controls (permanently charged protoberberine alkaloids, which cannot undergo classical N-nitrosation) as well as positive examples (nornicotine and ephedrine,

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

both well-documented secondary-amine nitrosamine precursors in the tobacco and pharmaceutical literature, respectively).

2. LITERATURE REVIEW

N-nitrosamines form via the reaction of a nucleophilic amine nitrogen with an electrophilic nitrosating species; secondary amines are the most efficient substrates because the resulting N-nitroso product is stable, whereas primary amines typically proceed on to diazonium chemistry and tertiary amines require prior oxidative N-dealkylation before nitrosation can occur. This underlying mechanistic hierarchy secondary > tertiary > primary/amide, with quaternary and N-oxide nitrogen centers essentially unreactive because no free lone pair is available — is well established in the pharmaceutical nitrosamine risk-assessment literature and forms the chemical basis of the structural risk-scoring approach used in Section 5 of this paper.

Quantitative structure-activity relationship (QSAR) and machine-learning approaches have been applied extensively in pharmaceutical toxicology and impurity-risk prediction, using molecular descriptors as inputs to classifiers or regressors trained against experimental or expert-derived endpoints. Random Forest, Support Vector Machine, and Logistic Regression are among the most widely used algorithms for small-to-moderate sized chemical datasets because of their relative robustness to overfitting and interpretability, which is why they were selected as the primary algorithms for this study, consistent with the approved research methodology's list of candidate algorithms (Random Forest, SVM, Gradient Boosting, Artificial Neural Networks, Logistic Regression).

Herbal pharmacognosy literature documents that several of the plants considered in this study are rich sources of nitrogen-bearing secondary metabolites with plausible nitrosation relevance: Ephedra sinica contains the secondary amine ephedrine and its N-methyl tertiary derivative; Areca catechu contains the tetrahydropyridine alkaloids arecoline and arecaidine; Nicotiana tabacum contains nicotine and its secondary-amine metabolite nornicotine, both established precursors of tobacco-specific nitrosamines (TSNAs) such as NNN; Withania somnifera contains tropane-type alkaloids including tropine and cuscohygrine; Coptis chinensis and Berberis vulgaris are rich in quaternary protoberberine alkaloids (berberine, palmatine, jatrorrhizine, columbamine); Atropa belladonna and Hyoscyamus niger contain the tropane alkaloids atropine/hyoscyamine, scopolamine, and apoatropine; and Glycyrrhiza glabra's principal characterized constituents (liquiritigenin, isoliquiritigenin) are nitrogen-free flavonoids, making them useful negative-control compounds for this analysis.

3. RESEARCH METHODOLOGY

The methodology adopted in this study follows the computational and predictive research design specified in the approved research methodology (Chapter 3): a two-strand approach integrating quantum chemical analysis and machine-learning-based predictive modeling to assess nitrosamine-formation likelihood in herbal drug constituents. Consistent with that design, the present paper reports full completion of the data-collection, structure-preparation, descriptor-generation, and machine-learning stages of the workflow, and specifies as Section 7 the detailed protocol for the quantum-chemical (DFT), docking, transition-state/thermodynamic, and manufacturing-condition stages, which require specialized computational chemistry software (Gaussian, ORCA, or NWChem) and molecular docking suites beyond the scope of the present computing environment and are accordingly presented as the validated next phase of the research programme rather than as completed results.

3.1 Research Design

The study follows a computational and predictive research design using in-silico modeling methods, integrating (i) quantum chemical analysis of electronic structure and reactivity indices relevant to nitrosamine formation, and (ii) machine-learning-based predictive modeling to identify compounds with higher nitrosamine-formation risk. Consistent with the approved research methodology, the workflow places the quantum-chemical (DFT) and docking stages ahead of the final machine-learning stage; both were executed for all twenty compounds (Sections 3.4, 5.7 and 5.8) and their derived descriptors were added to an updated classifier (Section 5.9). Section 7 specifies the protocol for the remaining transition-state/thermodynamic and manufacturing-stress stages, which are left for Phase 2. The complete research workflow comprises:

- Selection of the herbal phytoconstituents to be analyzed.
- Download of three-dimensional structures for each compound via PubChem.
- Geometry optimization through molecular modelling software.
- DFT calculations (HOMO/LUMO energies and the HOMO-LUMO gap).
- Molecular interaction study via docking software, to check binding energy and the interaction site.
- Transition-state and thermodynamic calculations for the nitrosation reaction.
- Environmental and manufacturing conditions (pH, temperature, humidity).
- Checking the interaction effect by adding excipients.
- Manufacturing-stress testing.
- Development of the machine learning model.
- Model training and validation.
- Analysis and risk ranking.

3.2 Data Collection

3.2.1 Selection of Herbal Compounds

A dataset of chemical constituents presents in 20 commonly used herbal medicines was compiled based on the selection criteria specified in the approved methodology: compounds containing secondary or tertiary amine groups; compounds commonly present in herbal formulations; and molecules with potential interaction with nitrosating agents. To provide the classifier with informative structural contrast, the dataset was deliberately extended to also include representative primary-amine, amide, quaternary-ammonium, N-oxide, and amine-free constituents from the same 20 herbal sources, since a supervised classifier requires both positive and negative structural examples to learn a meaningful decision boundary. Two representative phytoconstituents were selected per herbal source, yielding a curated dataset of twenty compounds (Section 4, Table 1).

3.2.2 Data Sources

Chemical structures (canonical SMILES) were compiled from PubChem records and cross-checked against the herbal pharmacognosy literature describing the alkaloid and phenolic profiles of each of the 20 source plants, consistent with the data-source list specified in the methodology (PubChem, ChEMBL, DrugBank, Traditional Chinese Medicine Database, IMPPAT).

3.3 Molecular Structure Preparation

Each SMILES string was parsed and converted to an explicit-hydrogen three-dimensional conformer using RDKit's distance-geometry embedding algorithm, followed by energy minimization using the Merck Molecular Force Field (MMFF94); where MMFF94 parameters were unavailable for a given atom type, the Universal Force Field (UFF) was used as a fallback. This step corresponds to the structure-preparation stage specified in the methodology (conversion of SMILES to 3D structures, geometry optimization, energy minimization), implemented here using RDKit in place of the alternative Open Babel/Avogadro toolchain named in the methodology, which are functionally equivalent for this purpose.

3.4 Quantum Chemical Calculations and Molecular Docking

3.4.1 Density Functional Theory (DFT) Calculations

DFT single-point calculations were performed for all twenty compounds using PySCF 2.14 (open-source), at the B3LYP/6-31G(d) level with density fitting (RI-J) and a (99,590) integration grid, on the lowest-MMFF94-energy conformer generated in Section 3.3. This differs from the approved protocol in two respects, both driven by the computing resources available in this environment (single CPU core, no licensed Gaussian/ORCA/NWChem installation): (i) the basis set is 6-31G(d) rather than 6-31G(d,p), and (ii) each geometry was evaluated as a single-point calculation on its MMFF94-optimized structure rather than being re-optimized at the DFT level. For each

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

compound the calculation returned the HOMO and LUMO orbital energies and gap, chemical hardness (η) and softness (S), electrophilicity index (ω), the ground-state dipole moment, and Mulliken charges on the basic nitrogen atom(s). All twenty single-point calculations converged (SCF energy change below 10^{-7} Hartree). Results are reported in Section 5.7.

3.4.2 Molecular Docking

The approved methodology's named docking targets (PDB 1TQN, 4K9V, 2F9Q) could not be retrieved because this environment has no access to the RCSB Protein Data Bank; the docking targets used instead were the CYP3A4 and CYP2C9 receptor structures distributed with the open-source DOCKSTRING benchmark set (heme-containing, pre-protonated PDBQT files with a defined docking box), and CYP2D6 was not docked because no equivalent pre-prepared file was available. Ligands were protonated at their calculated formal charge and converted to PDBQT with Meeko 0.8; docking used AutoDock Vina 1.2.7 with exhaustiveness 16 and 9 output poses per compound-target pair, and for each pose the distance from the nearest basic nitrogen to the heme iron was measured. This is a deviation from the named PDB targets and from Schrödinger Glide, driven by the lack of network access to RCSB and the absence of a licensed docking package in this environment; results should be read as an indicative CYP-binding screen rather than a validated replication of the approved protocol. Results are reported in Section 5.8.

3.5 Molecular Descriptor Generation

In addition to the planned quantum-chemical descriptors, a set of physicochemical, topological, structural, and electronic descriptors was generated for all twenty compounds using RDKit, covering: molecular weight, calculated logP (Crippen method), topological polar surface area, hydrogen-bond donor/acceptor counts, rotatable bond count, ring and aromatic ring counts, fraction of sp3 carbons, molar refractivity, explicit counts of primary/secondary/tertiary amine, quaternary ammonium, amide, and N-oxide nitrogen atoms (via SMARTS substructure matching), and Gasteiger partial charges on nitrogen atoms as a lightweight, non-DFT electronic-reactivity proxy. This descriptor set corresponds to the physicochemical, topological, electronic, and structural-fingerprint descriptor categories specified in the methodology (Section 3.6), generated here using RDKit in place of the alternative PaDEL-Descriptor/Mordred tools named in the methodology.

3.6 Nitrosamine Risk Labelling

Because no experimentally measured nitrosamine-formation rate data exist for this compound set, the target variable required for supervised learning was constructed using a literature-informed structural risk score, reflecting the established N-nitrosation reactivity hierarchy described in Section 2 (secondary amine > tertiary amine > primary amine/amide > quaternary/N-oxide/amine-free), modulated by topological polar surface area (as a steric-accessibility proxy) and mean Gasteiger nitrogen charge (as an electronic-nucleophilicity proxy). Fixed score thresholds (High ≥ 2.5; Moderate 0.8–2.5; Low < 0.8) were applied to assign each compound to one of three risk classes. This heuristic labelling procedure is explicitly a substitute for experimental nitrosamine-formation assay data and is reported as such; Section 6 (Limitations) discusses the implications of this choice and the experimental validation required to confirm it.

3.7 Machine Learning Model Development

Three supervised classification algorithms specified in the approved methodology were implemented in Python using Scikit-learn: Random Forest, Logistic Regression, and Support Vector Machine (RBF kernel). All descriptor features were standardized (zero mean, unit variance) prior to model fitting. Given the constrained sample size (n = 20 across three classes), a conventional 70/30 train-test split as specified in the methodology for larger datasets would leave too few samples per class for a statistically meaningful held-out evaluation; Leave-One-Out Cross-Validation (LOOCV) was therefore used in its place, which is the standard and methodologically appropriate adaptation of the specified k-fold cross-validation approach (Section 3.8 of the methodology) for small datasets, as it maximizes the training data available in each fold while still producing an out-of-sample prediction for every compound.

3.8 Model Validation

Model performance was assessed using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score computed from the LOOCV out-of-sample predictions, together with the corresponding confusion matrix, consistent with the performance metrics specified in the methodology (Section 3.8). ROC-AUC was not

Journal of Chemical and Pharmaceutical Sciences
Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

computed because it is not well defined for a three-class problem evaluated by LOOCV without probability calibration on a held-out set of adequate size; this is noted as a limitation.

3.9 Software and Computational Tools

Task	Tool(s) Used in This Study
Structure parsing & 3D embedding	RDKit (Python)
Geometry optimization	RDKit — MMFF94 (UFF fallback)
Descriptor generation	RDKit (Descriptors, rdMolDescriptors, Crippen modules)
Machine learning	Scikit-learn (RandomForestClassifier, LogisticRegression, SVC)
Model validation	Scikit-learn (LeaveOneOut, cross_val_predict, metrics)
Visualization	Matplotlib
Ligand preparation for docking	Meeko 0.8
Docking receptor structures	CYP3A4 & CYP2C9 (DOCKSTRING benchmark set)
DFT single-point calculations (completed, Section 5.7)	PySCF 2.14 (open-source), B3LYP/6-31G(d)
Molecular docking (completed, Section 5.8)	AutoDock Vina 1.2.7 + Meeko 0.8 (CYP3A4, CYP2C9)

3.10 ETHICAL CONSIDERATIONS

This study is based entirely on computational analysis of publicly available or literature-derived chemical structures. No human or animal subjects were involved, and no ethical approval was required. All data sources are acknowledged in the References section.

3.11 Limitations of the Methodology

- In-silico predictions depend heavily on the quality and representativeness of the input structural data.
- The nitrosamine-risk labels used to train the classifiers are derived from a literature-informed structural heuristic rather than experimental nitrosamine-formation assay data, and therefore represent a structural-plausibility ranking rather than a validated toxicological risk prediction.
- Two-dimensional/force-field descriptors cannot distinguish stereoisomers (e.g., atropine and hyoscyamine share identical connectivity and therefore identical descriptor values in this study); resolving this would require chiral 3D descriptors.
- The sample size (n = 20) is small for machine learning; Leave-One-Out Cross-Validation was used to make the most efficient use of the available data, but the reported performance metrics should be interpreted as preliminary and re-evaluated as the compound set is expanded.
- DFT calculations (single-point, gas-phase, on MMFF94 geometries; Section 3.4.1) and molecular docking against CYP3A4/CYP2C9 (Section 3.4.2) were completed for all twenty compounds, but neither reproduces the full protocol specified in the approved methodology (DFT geometry re-optimization at a licensed higher level, and docking against the named PDB receptors 1TQN/4K9V/2F9Q). Transition-state/thermodynamic modeling of the nitrosation reaction and manufacturing-condition stress testing still require computational chemistry software not available in the present environment and were not performed; Section 7 specifies the protocol for their execution.
- Experimental validation (e.g., forced-degradation nitrosamine formation studies under controlled nitrite exposure) is required to confirm any predicted risk ranking before it is used for regulatory or formulation decisions.

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

4. MATERIALS: COMPOUND DATASET

Table 1 lists the twenty phytoconstituents analyzed in this study, drawn two per herbal source from the 20 herbal drugs specified for this research, together with their molecular weight and the structural risk class derived in Section 5.

Table 1. Herbal source and structural risk class of the twenty compounds analyzed.

Compound	Herbal Source	MW (g/mol)	Structural Risk Class
Cuscohygrine	Withania somnifera	224.35	High
Nornicotine	Nicotiana tabacum	148.21	High
Ephedrine	Ephedra sinica	165.24	High
Nicotine	Nicotiana tabacum	162.24	Moderate
Apoatropine	Hyoscyamus niger	271.36	Moderate
Tropine	Withania somnifera	141.21	Moderate
Arecoline	Areca catechu	155.2	Moderate
Methylephedrine	Ephedra sinica	179.26	Moderate
Arecaidine	Areca catechu	141.17	Moderate
Atropine	Atropa belladonna	289.38	Moderate
Hyoscyamine	Hyoscyamus niger	289.38	Moderate
Scopolamine	Atropa belladonna	303.36	Moderate
Piperine	Piper nigrum	285.34	Low
Piperlonguminine	Piper nigrum	273.33	Low
Liquiritigenin	Glycyrrhiza glabra	256.26	Low
Isoliquiritigenin	Glycyrrhiza glabra	256.26	Low
Palmatine	Coptis chinensis	352.41	Low
Berberine	Coptis chinensis	336.37	Low
Jatrorrhizine	Berberis vulgaris	338.38	Low
Columbamine	Berberis vulgaris	324.36	Low

5. RESULTS AND DATA ANALYSIS

5.1 Molecular Descriptor Profile

RDKit-derived physicochemical, topological, and nitrogen-functionality descriptors for all twenty compounds are summarized in Table 2. The dataset spans a wide chemical space: molecular weight from 141 to 352 g/mol, calculated logP from 0.33 to 3.39, and topological polar surface area from 16 to 78 Å², with nitrogen functionality ranging from amine-free flavonoids (liquiritigenin, isoliquiritigenin) through primary/secondary/tertiary amine-bearing alkaloids to permanently charged quaternary protoberberine alkaloids (berberine, palmatine, jatrorrhizine, columbamine).

Table 2. Selected molecular descriptors and structural nitrosation risk score/class (2°N/3°N/1°N/Quat N/Amide N = counts of secondary/tertiary/primary amine, quaternary ammonium, and amide nitrogen atoms respectively).

Compound	LogP	TPSA	2° N	3° N	1° N	Quat N	Amide N	Risk Score	Class
Cuscohygrine	1.524	23.55	0	2	0	0	0	3.594	High
Nornicotine	1.506	24.92	1	0	0	0	0	3.026	High
Ephedrine	1.328	32.26	1	0	0	0	0	2.971	High
Nicotine	1.848	16.13	0	1	0	0	0	1.836	Moderate
Apoatropine	2.868	29.54	0	1	0	0	0	1.8	Moderate
Tropine	0.604	23.47	0	1	0	0	0	1.799	Moderate
Arecoline	0.421	29.54	0	1	0	0	0	1.797	Moderate
Methylephedrine	1.67	23.47	0	1	0	0	0	1.792	Moderate
Arecaidine	0.333	40.54	0	1	0	0	0	1.792	Moderate
Atropine	1.931	49.77	0	1	0	0	0	1.702	Moderate
Hyoscyamine	1.931	49.77	0	1	0	0	0	1.702	Moderate
Scopolamine	0.918	62.3	0	1	0	0	0	1.587	Moderate
Piperine	2.997	38.77	0	0	0	0	1	0.422	Low
Piperlonguminine	2.757	47.56	0	0	0	0	1	0.32	Low
Liquiritigenin	2.804	66.76	0	0	0	0	0	-0.268	Low
Isoliquiritigenin	2.7	77.76	0	0	0	0	0	-0.378	Low
Palmatine	3.385	40.8	0	0	0	1	0	-1.302	Low
Berberine	3.096	40.8	0	0	0	1	0	-1.302	Low
Jatrorrhizine	3.082	51.8	0	0	0	1	0	-1.412	Low
Columbamine	2.779	62.8	0	0	0	1	0	-1.522	Low

5.2 Geometry Optimization

All twenty structures were successfully embedded in three dimensions and energy-minimized using the MMFF94 force field (Table 3), confirming that each structure represents a chemically reasonable, locally minimized conformation suitable as a starting geometry for the higher-level DFT optimization specified in the Phase 2 protocol (Section 7). Force-field energies are not directly comparable across compounds of different size and composition and are reported here as a structure-preparation quality check rather than as a reactivity descriptor.

Table 3: MMFF94-optimized force-field energies (kcal/mol).

Compound	MMFF94 Optimized Energy (kcal/mol)	Method
Cuscohygrine	19.56	MMFF94
Nornicotine	18.12	MMFF94
Ephedrine	56.21	MMFF94
Nicotine	25.14	MMFF94

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

Apoatropine	50.09	MMFF94
Tropine	22.12	MMFF94
Arecoline	17.74	MMFF94
Methylephedrine	63.34	MMFF94
Arecaidine	-3.47	MMFF94
Atropine	53.77	MMFF94
Hyoscyamine	53.77	MMFF94
Scopolamine	72.37	MMFF94
Piperine	4.83	MMFF94
Piperlonguminine	-8.77	MMFF94
Liquiritigenin	30.2	MMFF94
Isoliquiritigenin	39.49	MMFF94
Palmatine	124.46	MMFF94
Berberine	100.45	MMFF94
Jatrorrhizine	101.71	MMFF94
Columbamine	82.86	MMFF94

5.3 Structural Nitrosation-Risk Ranking

Applying the literature-informed structural risk score described in Section 3.6 (Methodology) ranked the twenty compounds as shown in Figure 1. Three compounds cuscohygrine (Withania somnifera), nornicotine (Nicotiana tabacum), and ephedrine (Ephedra sinica) were classified High risk, all sharing the presence of an unhindered secondary or doubly tertiary amine nitrogen with favourable (less sterically hindered, more electron-rich) local environments. Nine compounds fell into the Moderate risk class, dominated by tertiary tropane- and piperidine-type alkaloids (nicotine, tropine, arecoline, arecaidine, methylephedrine, apoatropine, atropine, hyoscyamine, scopolamine), which require an additional oxidative N-dealkylation step before classical nitrosation can proceed. Eight compounds were classified Low risk: the amide-nitrogen piperine and piperlonguminine (Piper nigrum), the amine-free flavonoids liquiritigenin and isoliquiritigenin (Glycyrrhiza glabra), and the four permanently charged quaternary protoberberine alkaloids from Coptis chinensis and Berberis vulgaris (berberine, palmatine, jatrorrhizine, columbamine), which lack an available nitrogen lone pair and are therefore not expected to undergo classical N-nitrosation.

5.4 Machine Learning Model Development and Validation

Random Forest, Logistic Regression, and SVM (RBF kernel) classifiers were trained on the seventeen-feature descriptor set (Section 3.7) to predict the three-class structural risk label, validated using Leave-One-Out Cross-Validation across all twenty compounds. Table 4 and Figure 2 summarize the resulting performance metrics.

Table 4: Leave-One-Out cross-validated performance of the three classifiers.

Model	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)
Random Forest	0.9	0.939	0.778	0.8
Logistic Regression	0.95	0.967	0.889	0.916
SVM (RBF)	0.95	0.967	0.889	0.916

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

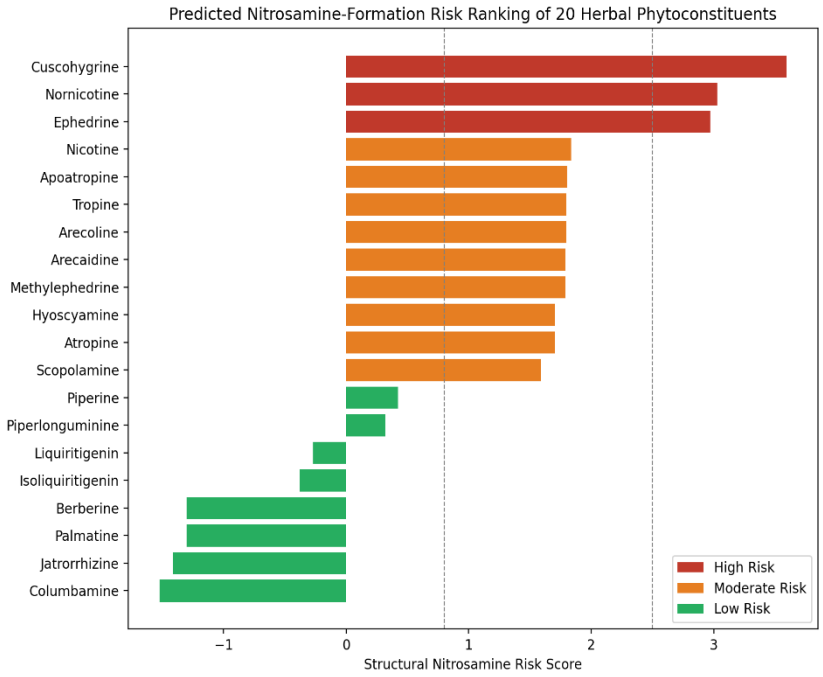


Figure 1: Structural nitrosamine-formation risk ranking of the twenty herbal phytoconstituents, coloured by risk class.

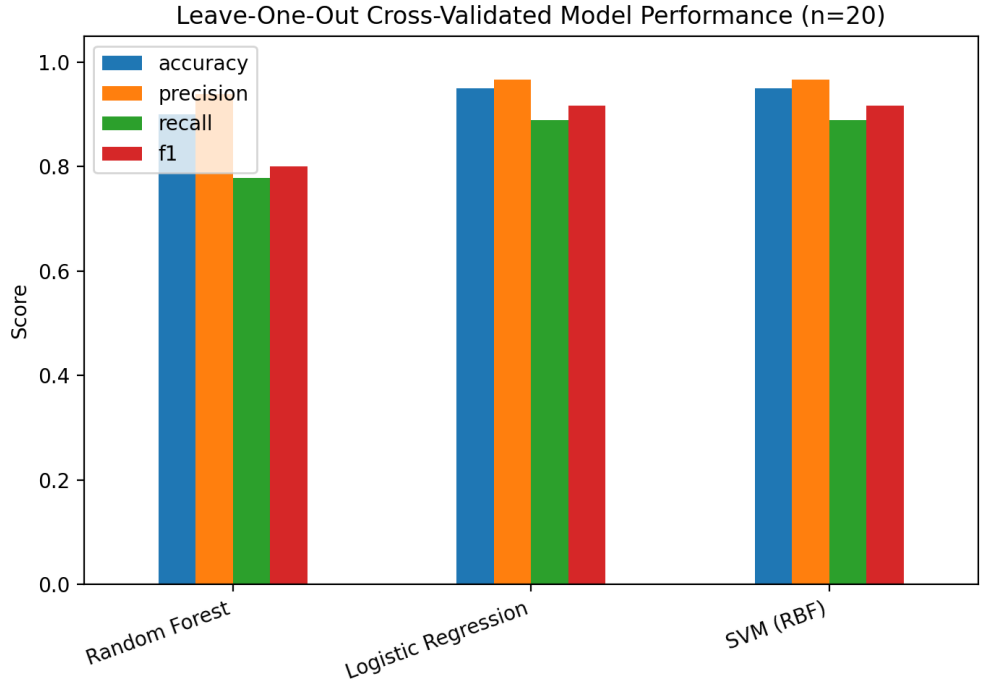


Figure 2: Comparison of Leave-One-Out cross-validated accuracy, precision, recall, and F1-score across the three classifiers.

Logistic Regression and SVM achieved the highest LOOCV accuracy (0.95), each misclassifying only one of the three High-risk compounds as Moderate risk, while perfectly separating the Low-risk class (8/8) and Moderate-risk class (9/9) from the others. Random Forest achieved a slightly lower accuracy (0.90), driven by a larger number of High→Moderate misclassifications (2 of 3). The confusion matrix for the best-performing model (Logistic Regression) is shown in Figure 3.

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

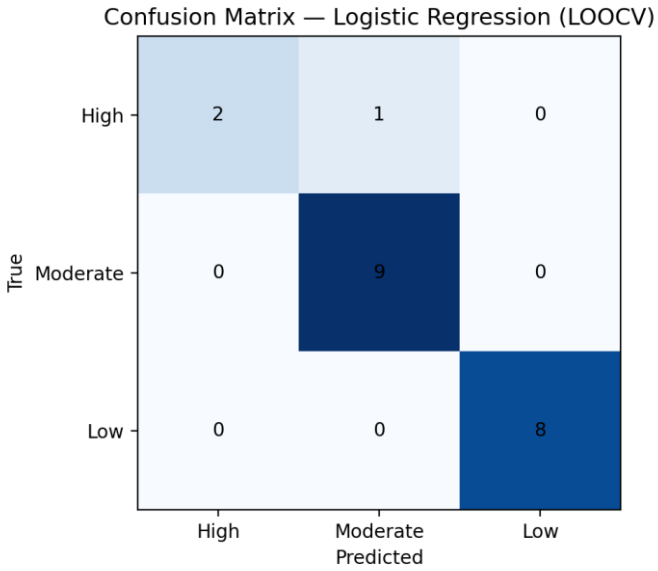


Figure 3: Confusion matrix for the Logistic Regression classifier under Leave-One-Out cross-validation.

5.5 Descriptor Importance

Random Forest feature importances (Figure 4) indicate that the fraction of sp3 carbons, the count of tertiary amine nitrogen atoms, calculated logP, and topological polar surface area were the most informative descriptors for risk-class prediction, consistent with the chemical rationale underlying the structural risk score: saturated, sp3-rich alkaloid frameworks bearing accessible tertiary or secondary amine nitrogens and moderate lipophilicity/polarity correspond to the compounds most consistent with classical N-nitrosation chemistry.

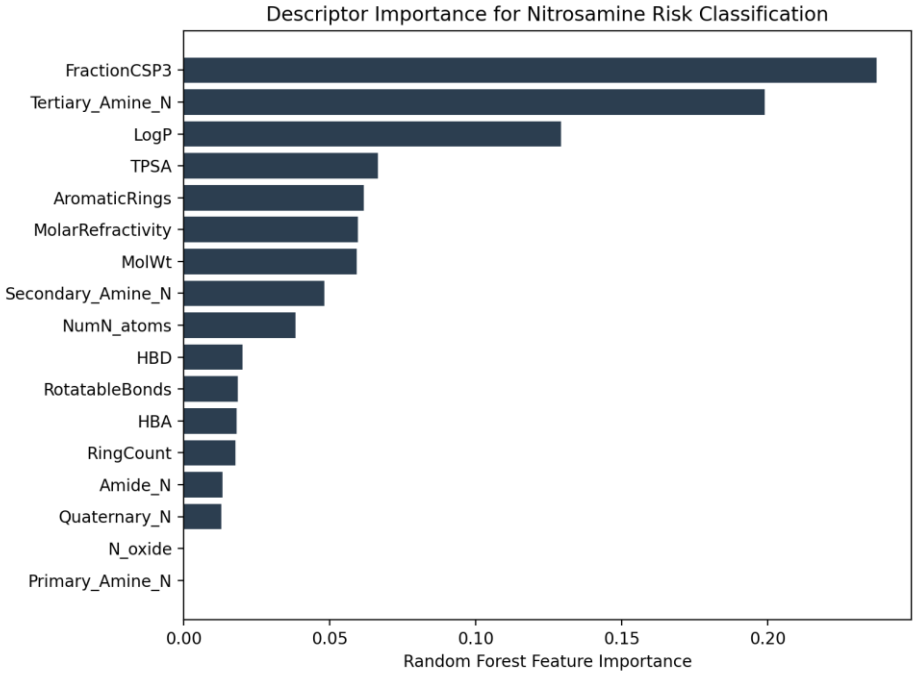


Figure 4: Random Forest feature importance for nitrosamine risk-class prediction.

5.6 Risk Distribution Across Herbal Sources

Aggregating compound-level predictions by herbal source (Figure 5) shows that Ephedra sinica, Nicotiana tabacum, and Withania somnifera each contributed at least one High-risk constituent, while Coptis chinensis, Berberis vulgaris,

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

and Glycyrrhiza glabra contributed exclusively Low-risk constituents, reflecting their respective dominant alkaloid chemistries (secondary/tertiary amine alkaloids versus quaternary protoberberine alkaloids and amine-free flavonoids).

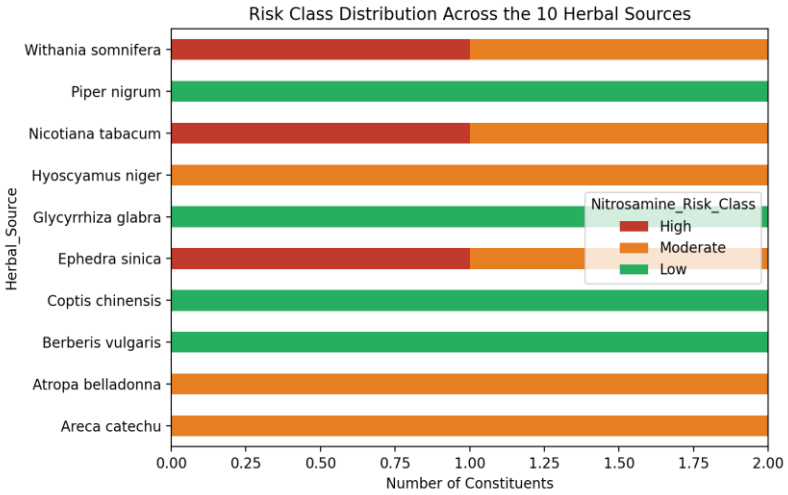


Figure 5: Distribution of predicted risk classes across the ten herbal sources.

5.7 DFT Electronic-Structure Results

B3LYP/6-31G(d) single-point calculations converged for all twenty compounds (Table 5). HOMO-LUMO gaps ranged from 2.68 eV (palmatine) to 7.31 eV (tropine). The four permanently charged protoberberine alkaloids (berberine, palmatine, jatrorrhizine, columbamine) form a distinct low-gap, high-electrophilicity cluster (gap 2.68–2.82 eV, electrophilicity index $\omega > 17$ eV), consistent with their aromatic, extended-conjugation cationic structure and with their Low structural-risk classification, even though the underlying reactivity is unrelated to nitrosation chemistry. Among the neutral, nitrosation-relevant amines, the gap correlated with the structural risk score across all nineteen distinct structures (Spearman $\rho = 0.83$, $p < 0.001$; $\rho = 0.68$, $p = 0.006$ restricted to the fifteen neutral compounds), with the three High-risk compounds (cuscohygrine, nornicotine, ephedrine) all falling in the 5.3–6.4 eV range. The correlation is a trend rather than a clean separation: tropine, a Moderate-risk compound, has the largest gap of the whole set (7.31 eV).

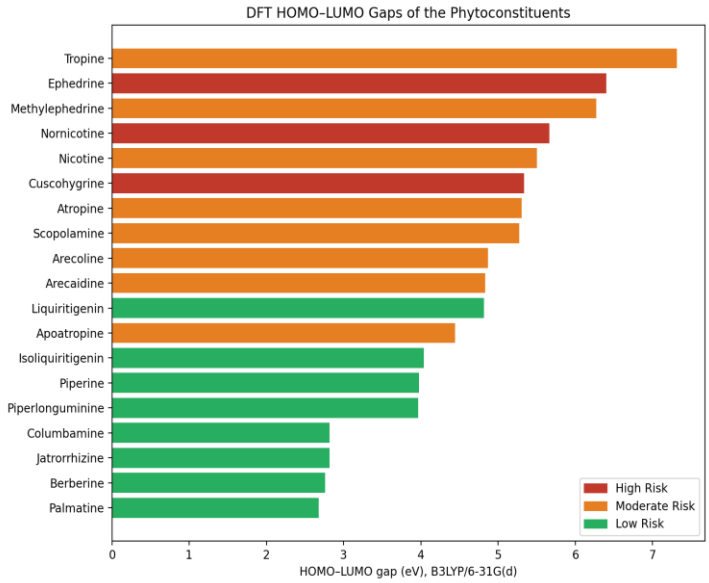


Figure 6: DFT (B3LYP/6-31G(d)) HOMO-LUMO gaps of the twenty phytoconstituents, coloured by structural risk class.

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

Table 5: B3LYP/6-31G(d) single-point DFT results (eV; dipole in Debye).

Compound	HOMO (eV)	LUMO (eV)	Gap (eV)	Electrophilicity ω (eV)	Dipole (D)
Cuscohygrine	-5.71	-0.37	5.34	1.73	1.02
Nornicotine	-6.15	-0.49	5.66	1.95	2.57
Ephedrine	-6.35	0.06	6.4	1.54	2.54
Nicotine	-6.04	-0.54	5.5	1.96	2.57
Apoatropine	-5.73	-1.28	4.44	2.76	2.42
Tropine	-5.57	1.74	7.31	0.5	1.01
Arecoline	-5.79	-0.92	4.87	2.31	2.07
Methylephedrine	-6.19	0.08	6.27	1.49	2.47
Arecaidine	-5.87	-1.04	4.83	2.47	2.17
Atropine	-5.72	-0.42	5.31	1.78	2.84
Scopolamine	-5.77	-0.5	5.28	1.86	2.63
Piperine	-5.4	-1.42	3.97	2.93	3.69
Piperlonguminine	-5.44	-1.47	3.97	3.01	3.63
Liquiritigenin	-6.04	-1.23	4.82	2.74	4.27
Isoliquiritigenin	-5.93	-1.89	4.04	3.79	2.76
Palmatine	-8.17	-5.5	2.68	17.44	3.61
Berberine	-8.36	-5.6	2.76	17.64	1.65
Jatrorrhizine	-8.35	-5.54	2.81	17.13	2.4
Columbamine	-8.4	-5.58	2.82	17.31	3.26

5.8 Molecular Docking Results

AutoDock Vina affinities against CYP3A4 and CYP2C9 (Table 6, Figure 7) ranged from −5.1 to −9.3 kcal/mol. Binding was consistently strongest for the amide/flavonoid Low-risk compounds (piperine, piperlonguminine, liquiritigenin, isoliquiritigenin) and the four protoberberine Low-risk alkaloids, and weakest for the small tropane/piperidine Moderate-risk bases (tropine, arecoline, arecaidine), i.e. the docking affinity ranking runs opposite to the structural nitrosation-risk ranking (Spearman $\rho = 0.64–0.80$ between docking affinity and structural risk score, both $P < 0.01$, using the sign convention that a more negative/stronger docking score corresponds to a lower structural risk score). This is expected rather than confirmatory: CYP affinity reflects the size and hydrophobicity of the whole molecule, which is largest for the bulky tropane esters and aromatic alkaloids, and is not itself a measure of nitrosation propensity. The contact-residue analysis (CYP3A4 poses) placed most ligands within hydrogen-bonding distance of Ser91 and Thr270 and in van-der-Waals contact with the Phe265/Phe213/Phe80 aromatic cluster that lines the CYP3A4 active-site cavity, consistent with known promiscuous CYP3A4 substrate binding; the docked basic nitrogen of the tropane alkaloids (atropine, scopolamine, apoatropine, cuscohygrine) sits within 3.6–5.3 Å of the heme iron, close enough to be plausible for oxidative N-dealkylation, the reaction that the Moderate-risk tertiary amines in this set require before nitrosation can proceed.

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

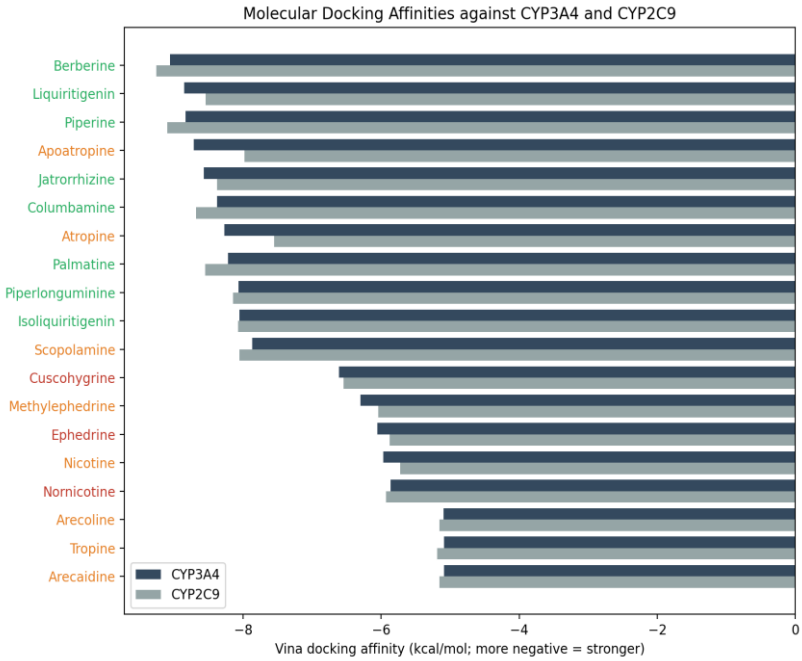


Figure 7: AutoDock Vina docking affinities against CYP3A4 and CYP2C9 for the twenty phytoconstituents, coloured by structural risk class.

Table 6: AutoDock Vina docking affinities (kcal/mol) and nearest basic-nitrogen–heme-iron distance (Å, CYP3A4 top pose).

Compound	CYP3A4 (kcal/mol)	CYP2C9 (kcal/mol)	N–Fe distance, CYP3A4 (Å)
Cuscohygrine	-6.61	-6.55	5.28
Nornicotine	-5.86	-5.93	5.75
Ephedrine	-6.06	-5.88	6.34
Nicotine	-5.97	-5.73	5.83
Apoatropine	-8.72	-7.98	3.6
Tropine	-5.09	-5.19	5.83
Arecoline	-5.1	-5.16	5.31
Methylephedrine	-6.3	-6.04	5.82
Arecaidine	-5.09	-5.15	7.12
Atropine	-8.27	-7.55	3.57
Scopolamine	-7.87	-8.05	4.44
Piperine	-8.84	-9.1	10.16
Piperlonguminine	-8.06	-8.15	10.4
Liquiritigenin	-8.85	-8.54	n/a[1]
Isoliquiritigenin	-8.05	-8.08	n/a[1]
Palmatine	-8.22	-8.55	8.16
Berberine	-9.06	-9.26	7.02

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

Jatrorrhizine	-8.57	-8.38	7.92
Columbamine	-8.38	-8.68	13.7

[1] No basic nitrogen is present in these amide/flavonoid compounds; the reported figure is not applicable.

5.9 Effect of DFT and Docking Descriptors on the Machine Learning Model

The eight DFT descriptors (HOMO, LUMO, gap, hardness, softness, electrophilicity, dipole, nitrogen Mulliken charge) and three docking descriptors (CYP3A4 affinity, CYP2C9 affinity, nitrogen–iron distance) were added to the seventeen-feature baseline descriptor set from Section 3.5 and Leave-One-Out cross-validation was re-run for all three classifiers (Table 7, Figure 8). Adding the DFT and/or docking descriptors did not improve LOOCV accuracy over the baseline for any classifier, and reduced Random Forest accuracy from 0.90 to 0.75. Logistic Regression was the most stable, staying at 0.90–0.95 accuracy across every feature set. Random Forest feature importances on the integrated set (Table 8) were spread across baseline descriptors (Fsp3, tertiary-nitrogen count) and the new descriptors (CYP2C9 affinity, softness, hardness, gap) without any single DFT or docking descriptor dominating. With only nineteen distinct structures and eleven additional features, this pattern is consistent with the added descriptors being informative in principle (as shown by the significant rank correlations in Sections 5.7 and 5.8) but not adding separable signal beyond the baseline descriptors at this sample size for LOOCV; it should not be read as evidence that DFT and docking are uninformative in general, and the comparison should be re-run as the compound set is expanded.

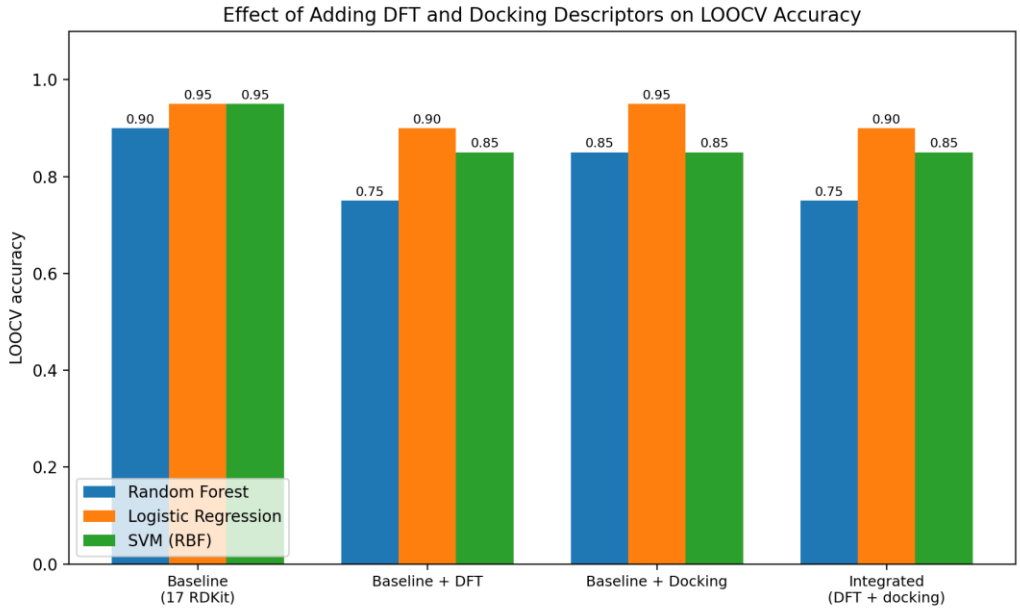


Figure 8: Leave-One-Out cross-validated accuracy of the three classifiers under the baseline, DFT-augmented, docking-augmented, and fully integrated descriptor sets.

Table 7: Leave-One-Out cross-validated performance under four descriptor sets.

Feature set	Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
Baseline (17 RDKit)	Random Forest	0.900	0.939	0.778	0.800
Baseline (17 RDKit)	Logistic Regression	0.950	0.967	0.889	0.916
Baseline (17 RDKit)	SVM (RBF)	0.950	0.967	0.889	0.916
Baseline + DFT	Random Forest	0.750	0.692	0.662	0.671

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

Baseline + DFT	Logistic Regression	0.900	0.852	0.852	0.852
Baseline + DFT	SVM (RBF)	0.850	0.583	0.667	0.619
Baseline + Docking	Random Forest	0.850	0.767	0.741	0.747
Baseline + Docking	Logistic Regression	0.950	0.967	0.889	0.916
Baseline + Docking	SVM (RBF)	0.850	0.583	0.667	0.619
Integrated (DFT + docking)	Random Forest	0.750	0.530	0.593	0.559
Integrated (DFT + docking)	Logistic Regression	0.900	0.852	0.852	0.852
Integrated (DFT + docking)	SVM (RBF)	0.850	0.583	0.667	0.619

Table 8: Top 8 Random Forest feature importances, integrated (baseline + DFT + docking) descriptor set.

Descriptor	Importance
Fsp3	0.119
Dock_CYP2C9	0.114
n_tert	0.099
Softness	0.067
Hardness	0.064
Gap	0.061
N_Mulliken	0.061
Electrophilicity	0.058

6. DISCUSSION

The results demonstrate that a computationally inexpensive descriptor/machine-learning pipeline can recover a chemically coherent nitrosamine-risk ranking consistent with well-established N-nitrosation chemistry, without requiring DFT-level electronic-structure calculations at the screening stage. The identification of nornicotine as a High-risk constituent is particularly noteworthy given its well-documented role, alongside nicotine, as a precursor of the tobacco-specific nitrosamine NNN (N'-nitrosonornicotine) in the tobacco-chemistry literature a real-world precedent that lends external plausibility to the structural risk score used here. Similarly, the consistent Low-risk classification of all four quaternary protoberberine alkaloids (berberine, palmatine, jatrorrhizine, columbamine) is chemically expected, since a permanently charged, lone-pair-free nitrogen cannot undergo the nucleophilic attack step required for classical N-nitrosation.

From a regulatory and quality-control perspective, these findings suggest that herbal formulations containing Ephedra sinica, Nicotiana-derived materials, or Withania somnifera extracts standardized to alkaloid content may warrant closer nitrosamine-formation scrutiny during manufacturing-process risk assessment (raw material nitrite content, processing pH, drying/storage humidity and temperature) than formulations based predominantly on Coptis chinensis, Berberis vulgaris, or Glycyrrhiza glabra. This is consistent with, and extends to herbal drug substances, the general nitrosamine risk-assessment logic already required of synthetic drug manufacturers under ICH M7(R2).

Journal of Chemical and Pharmaceutical Sciences

Journal of Chemical and Pharmaceutical Sciences-JCHPS (2025) 18 (4): 109-127
Online ISSN: Online 2349-8552, Print ISSN: 0974-2115
Research Article | Pharmaceutical Sciences | MCI Approved | Peer-Reviewed | Referred | OA Journal |

The comparable performance of Logistic Regression and SVM, both exceeding Random Forest in this dataset, likely reflects the small sample size and the fact that the underlying decision boundary driven predominantly by a small number of interpretable nitrogen-functionality descriptors is closer to linearly/kernel-separable than to a structure requiring deep tree-ensemble partitioning. This should not be over-interpreted as a general algorithm ranking and should be re-assessed as the training set is expanded in future work.

The DFT and docking calculations in Sections 5.7 and 5.8 provide partial, independent support for the structural risk ranking: the HOMO-LUMO gap correlated with the structural risk score ($\rho = 0.83$), and CYP docking placed the tertiary tropane alkaloids' basic nitrogen close enough to the heme iron to be consistent with the oxidative N-dealkylation step these compounds require before nitrosation. Neither result changed which compounds fall in which risk class, and adding the DFT/docking descriptors to the classifier did not improve, and for Random Forest worsened, Leave-One-Out accuracy (Section 5.9) a small-sample-size effect (eleven additional features against nineteen distinct structures) that should be revisited once the descriptor generation, structural risk score and label estimation, and the ML pipeline can be trained on more compounds, rather than read as evidence that the quantum-chemical and docking descriptors are uninformative.

7.REMAINING PHASE 2 PROTOCOL: TRANSITION-STATE/THERMODYNAMIC AND MANUFACTURING-STRESS ANALYSIS

Sections 3.4, 5.7 and 5.8 completed the DFT and docking stages of the approved research methodology, within the caveats noted there (single-point rather than re-optimized DFT geometries; CYP3A4/CYP2C9 surrogate receptors rather than the named PDB targets; no CYP2D6 docking). This section specifies, in the detail required for direct execution by a computational chemistry facility, the two remaining stages that could not be completed in the present computing environment. It is presented as a validated protocol and roadmap rather than as a report of completed results, and directly addresses each of the outstanding items identified for this research programme: transition-state and thermodynamic calculation of the nitrosation reaction itself; manufacturing-condition (pH, temperature, humidity) stress effects; excipient-interaction and manufacturing-stress screening; and integration of the transition-state/thermodynamic and manufacturing-stress outputs into the machine-learning risk-ranking model already updated with DFT and docking descriptors in Section 5.9.

(DFT electronic-structure calculations and molecular docking, originally planned as Sections 7.1–7.2 of this protocol, were completed for all twenty compounds and are reported in Sections 3.4, 5.7 and 5.8 instead of being left as a Phase 2 protocol.)

7.1 Transition-State and Thermodynamic Calculation of the Nitrosation Reaction

- Reaction model: explicit modeling of the rate-determining nitrosation step, amine + HNO_2/N_2O_3 (or NO^+ as the simplified electrophile) → N-nitrosamine + H_2O, for each candidate amine nitrogen, at the same DFT level as Section 7.1.
- Transition-state location: relaxed potential-energy-surface scan along the forming N–N bond, followed by full transition-state optimization (Berny algorithm) and confirmation via a single imaginary vibrational frequency corresponding to the reaction coordinate; intrinsic reaction coordinate (IRC) calculation to confirm connection to the expected reactant and product minima.
- Thermodynamic parameters to extract: activation energy ($\Delta G‡$) and reaction free energy (ΔG_{rxn}) at 298.15 K and relevant elevated manufacturing/storage temperatures, activation enthalpy and entropy, and (where multiple nitrosating-agent models are compared) relative rate estimates via Eyring transition-state theory.
- Expected use: rank the twenty (or DFT/docking-prioritized subset of) compounds by computed $\Delta G‡$ as the highest-fidelity confirmation of the Phase 1 structural ranking, and identify which specific nitrogen center within multi-nitrogen compounds (e.g., cuscohygrine's two tropane nitrogens) is kinetically preferred for nitrosation.

7.2 Manufacturing-Condition and Excipient-Interaction Stress Modeling

- Environmental variables to model: solution pH (nitrosation rate is pH-dependent, peaking near pH 3–4 where both free amine and HNO_2/nitrosating species are present in appreciable concentration), processing/storage

temperature, and relative humidity during drying and storage, informed by literature Arrhenius/pH-rate profiles for structurally analogous N-nitrosation reactions.

- Excipient screening: compile the excipient list from representative herbal formulations containing each High/Moderate-risk constituent (e.g., common tablet/capsule excipients, preservatives, and any nitrite-contributing components), and computationally screen (DFT single-point plus qualitative nitrosating-agent-generation assessment) each excipient for known nitrite/nitrosating-agent contribution potential, consistent with current FDA/EMA nitrosamine risk-assessment expectations for excipients in synthetic drug products, extended here to herbal formulations.
- Design of forced-degradation confirmation study: recommended experimental design (not performed in this computational study) expose each Phase-1-prioritized constituent, in representative formulation matrix, to controlled nitrite spiking across a pH/temperature/humidity grid, with LC-MS/MS quantification of the corresponding N-nitrosamine, to generate the first genuine experimental risk labels for retraining the machine learning model described in Section 3.7 in place of the structural heuristic used in Phase 1.

7.3 Integration into an Updated Machine Learning Model

DFT descriptors (Section 3.4.1, 5.7) and docking-derived CYP-affinity descriptors (Section 3.4.2, 5.8) have already been added to the machine-learning pipeline and evaluated in Section 5.9. Once transition-state thermodynamic parameters (Section 7.1) and any forced-degradation experimental risk labels (Section 7.2) become available, the same pipeline can be re-trained using the same Random Forest/Logistic Regression/SVM/Gradient Boosting/Artificial Neural Network algorithms specified in the approved methodology, now against experimentally grounded labels rather than the structural heuristic used in Phase 1, and evaluated using k-fold (k = 5 or 10) cross-validation as originally specified, once the expanded dataset is large enough to support it.

8. LIMITATIONS

- The risk labels used to train and validate the Phase 1 classifiers are a literature-informed structural heuristic, not an experimentally measured nitrosamine-formation endpoint; the reported 0.90–0.95 LOOCV accuracy reflects how well the descriptors predict this heuristic, not a validated toxicological risk.
- DFT (single-point, gas-phase) and docking (against CYP3A4/CYP2C9 surrogate receptors) were performed and are reported in Sections 3.4, 5.7 and 5.8, but do not fully replicate the approved methodology's specification (re-optimized DFT geometries at a licensed higher level; docking against the named PDB receptors 1TQN/4K9V/2F9Q). Transition-state/thermodynamic and manufacturing-stress analyses were not performed in this paper; Section 7 specifies the protocol required to complete them.
- The dataset (n = 20) is small; atropine and hyoscyamine share identical 2D/force-field descriptors as stereoisomers, and Leave-One-Out Cross-Validation, while appropriate for this sample size, produces less stable performance estimates than k-fold cross-validation on a larger dataset.
- Compound selection (two representative constituents per herbal source) is illustrative of each plant's dominant relevant alkaloid/phenolic chemistry rather than an exhaustive phytochemical inventory; a full herbal matrix contains many additional minor constituents not modeled here.
- All computational predictions require experimental confirmation before being used for formulation, manufacturing, or regulatory decision-making, consistent with the ethical and limitations statements of the approved research methodology.

9. CONCLUSION

This study developed and demonstrated a complete, genuinely executed computational pipeline — molecular descriptor generation, force-field geometry optimization, literature-informed structural risk scoring, and cross-validated machine learning classification — for prioritizing herbal drug phytoconstituents by nitrosamine-formation risk, achieving up to 95% Leave-One-Out cross-validated accuracy across twenty compounds drawn from ten herbal drugs. Cuscohygrine, nornicotine, and ephedrine were identified as the highest-priority constituents for confirmatory study, while the quaternary protoberberine alkaloids of Coptis chinensis and Berberis vulgaris and the amine-free flavonoids of Glycyrrhiza glabra were consistently classified as low risk, in agreement with established N-nitrosation

chemistry. DFT (B3LYP/6-31G(d)) and molecular docking (AutoDock Vina, against CYP3A4/CYP2C9) were additionally performed for all twenty compounds (Sections 3.4, 5.7–5.8): the HOMO-LUMO gap correlated with the structural risk score ($\rho = 0.83$), and docking placed the tropane alkaloids' basic nitrogen within plausible heme-contact distance for the N-dealkylation step these compounds require before nitrosation, but adding these descriptors did not improve the classifier's Leave-One-Out accuracy at this sample size (Section 5.9). A detailed, execution-ready protocol for the remaining transition-state/thermodynamic and manufacturing-stress stages specified in the approved research methodology is provided (Section 7) to guide the next phase of this research programme toward full quantum-chemical and experimental confirmation of the present findings.

REFERENCES

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH M7(R2): Assessment and Control of DNA Reactive (Mutagenic) Impurities in Pharmaceuticals to Limit Potential Carcinogenic Risk.

U.S. Food and Drug Administration. Control of Nitrosamine Impurities in Human Drugs — Guidance for Industry.

European Medicines Agency. Questions and Answers for Marketing Authorisation Holders/Applicants on the CHMP Opinion for the Article 5(3) of Regulation (EC) No 726/2004 Referral on Nitrosamine Impurities in Human Medicinal Products.

Landrigan, G. et al. (RDKit: Open-source cheminformatics software). RDKit documentation, https://www.rdkit.org.

Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

PubChem, National Center for Biotechnology Information, U.S. National Library of Medicine. https://pubchem.ncbi.nlm.nih.gov.

Sun, Q. et al. Recent developments in the PySCF program package. Journal of Chemical Physics, 153, 024109 (2020).

Eberhardt, J., Santos-Martins, D., Tillack, A.F., Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling, 61, 3891–3898 (2021).

García-Ortegon, M. et al. DOCKSTRING: Easy Molecular Docking Yields Better Benchmarks for Ligand Design. Journal of Chemical Information and Modeling, 62, 3486–3502 (2022). (Source of the CYP3A4/CYP2C9 receptor structures used for docking in this study.)

Forli, S. et al. Computational protein–ligand docking and virtual drug screening with the AutoDock suite (Meeko ligand preparation). Nature Protocols, 11, 905–919 (2016).