Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

AUTONOMOUS AI AGENT FOR QSAR MODELING WITH DATASET CURATION, DESCRIPTOR SELECTION, AND DOMAIN ASSESSMENT

Chen Hao[1]*, Liu Fang[1], Zhao Lin[2]

1. ***Department of Drug Discovery Informatics, Faculty of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China.***
2. ***Department of AI Pharmaceutical Systems, Faculty of Engineering, Nanjing University, Nanjing, China.***

ARTICLE INFO

Received:
03 May 2025
Received in revised form:
20 July 2025
Accepted:
25 July 2025
Available online:
28 August 2025

Keywords: Autonomous AI agent, QSAR, Cheminformatics, AutoML, Dataset curation, Descriptor selection

ABSTRACT

QSAR modeling is central to computational drug discovery because it links molecular structure to biological activity before synthesis or testing. However, the practical construction of a reliable QSAR model still depends on expert judgment across data preparation, feature design, validation, and interpretation. The QSAR workflow is difficult to reproduce because each stage can involve subjective choices about chemical standardization, activity normalization, descriptor filtering, model selection, and applicability domain definition. These choices can limit the routine use of QSAR by medicinal chemists who need rapid, transparent, and fit-for-purpose predictive models. An autonomous QSAR agent would accept raw chemical–biological data together with a target endpoint specification and then execute the full modeling workflow with minimal human intervention. The agent would produce a documented predictive model, a data-quality summary, and an applicability-domain assessment suitable for decision support. The proposed system would include a data-cleaning engine, a descriptor-calculation and feature-selection module, an AutoML-based model trainer, an applicability-domain assessor, and a natural-language reporting interface. These components would operate as a coordinated workflow rather than as disconnected scripts. Such an agent could reduce routine modeling burden, enforce best-practice checks automatically, and make QSAR workflows more accessible to non-specialists. Its most important contribution would not be replacing expert judgment, but making each modeling decision traceable, reviewable, and reproducible. An autonomous QSAR agent could democratize predictive modeling in drug discovery by shifting expert effort from repetitive implementation toward strategic interpretation. The concept represents an emerging AI direction in which cheminformatics tools become active workflow participants rather than passive software components.

To Cite This Article: Hao C, Fang L, Lin Z. Autonomous AI Agent for QSAR Modeling with Dataset Curation, Descriptor Selection, and Domain Assessment. Pharmacophore. 2025;16(4):11-21. https://doi.org/10.51847/Bhyi6RjARa

Introduction

Quantitative structure–activity relationship modeling has become a core method for connecting chemical structure with biological or physicochemical endpoints in drug discovery, yet the growth of available molecular data has not eliminated the bottleneck of constructing trustworthy models. Automated QSAR frameworks have shown that model building can be systematized, but they still depend on careful choices about input data, representation, validation, and reporting [1]. More recent QSAR platforms have moved toward integrated automation, suggesting that a single system could coordinate data preparation, descriptor handling, model search, and deployment-oriented documentation [2]. Benchmark resources such as MoleculeNet and Therapeutics Data Commons have also encouraged standardized evaluation, but they do not by themselves solve the challenge of transforming raw project data into a reliable model [3, 4].

Current practice often requires chemists and data scientists to manually inspect molecular structures, resolve inconsistent activity labels, remove problematic records, and select descriptors before any model is trained. These expert-intensive steps are vulnerable to implicit bias, especially when data cleaning rules, feature filters, and validation splits are applied inconsistently across projects [5]. Tools such as DeepMol and QSPRpred demonstrate that automated cheminformatics

Corresponding Author: Chen Hao; Department of Drug Discovery Informatics, Faculty of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China. E-mail: chen.hao@outlook.com

pipelines can reduce some of this manual burden by combining preprocessing, representation, and model-building utilities in reproducible software environments [6, 7]. However, a fully autonomous agent would need to go further by deciding when a workflow step should be repeated, when a model should be rejected, and when uncertainty should be communicated to the user.

The emergence of autonomous systems in chemical discovery provides a conceptual foundation for end-to-end QSAR automation. Closed-loop molecular discovery platforms have shown how prediction, decision-making, measurement planning, and feedback can be organized into iterative workflows rather than isolated computational tasks [8]. In parallel, large language models and chemistry-aware tool use have opened a route toward agents that can coordinate software modules, inspect intermediate outputs, and produce human-readable explanations of technical decisions [9]. Reinforcement-learning systems for molecular design further illustrate how autonomous decision policies can guide chemical search, although QSAR automation would emphasize trustworthy prediction and documentation rather than molecule generation alone [10, 11].

This article proposes an Emerging AI concept: an autonomous AI agent that ingests raw chemical–biological data and produces a fully documented QSAR model with automated curation, descriptor selection, model building, validation, applicability-domain assessment, and interpretability reporting. The agent would combine the systematic workflow design of automated QSAR platforms [1, 2], the model-development flexibility of open-source cheminformatics toolkits [6, 7], and the self-evaluating behavior associated with AI agents and chemistry tool augmentation [9, 12]. Its purpose would be to make QSAR modeling reproducible, inspectable, and accessible while preserving human oversight at critical decision points. Rather than claiming experimental superiority, the concept defines how such a system could be designed, evaluated, and integrated into medicinal chemistry decision support.

Background

The Qsar Modeling Pipeline

A conventional QSAR pipeline begins with molecular structures and associated activity values, then proceeds through structure normalization, endpoint definition, descriptor calculation, feature selection, model training, validation, and interpretation. Each stage introduces decisions that can influence downstream conclusions, including how duplicate structures are handled, how inconsistent bioactivity values are reconciled, and which validation split is used [5]. Automated frameworks show that these steps can be formalized into repeatable workflows, but the pipeline still requires transparent logging so users can understand why specific records, descriptors, or models were retained [1]. An autonomous agent would therefore treat QSAR modeling as a governed sequence of decisions rather than a single model-fitting operation.

Autonomous Agents and AutoML in Chemistry

Autonomous agents in chemistry can be understood as systems that perceive the state of a scientific workflow, choose the next computational action, execute that action through software tools, and update their plan based on intermediate evidence. AutoML provides part of this capability by automating algorithm selection, hyper-parameter search, and model comparison, while chemistry-specific platforms such as QSARtuna, DeepMol, QSPRpred, PoseidonQ, and QSPRmodeler adapt these ideas to molecular prediction workflows [2, 6, 7, 13, 14]. Large language models extend the agent concept by enabling natural-language planning, tool calling, and explanation, especially when paired with chemistry software rather than used as standalone predictors [9, 12]. In an autonomous QSAR system, these capabilities would be orchestrated so that the agent can reason over curation outputs, descriptor diagnostics, model validation, and applicability-domain warnings.

Dataset Curation – Challenges and Automation

Dataset curation is often the most consequential part of QSAR modeling because noisy chemical structures or inconsistent activity annotations can distort even carefully tuned models. Automated systems must identify salts, mixtures, duplicate compounds, incompatible units, censored activity records, and assay artifacts before training begins [5]. Scaffold-aware and temporally motivated splitting strategies are also needed because random splitting can overstate the usefulness of a model for prospective chemical series, and algorithms such as SIMPD provide a way to simulate time-aware validation scenarios when true temporal order is unavailable [15]. Privacy-preserving and federated data-partitioning work further shows that dataset organization itself can be treated as a modeling design problem rather than as a clerical preprocessing step [16].

Molecular Descriptors and Feature Selection

Molecular representation spans physicochemical descriptors, topological indices, structural fingerprints, learned graph embeddings, and other encodings that capture different aspects of chemical structure. Because descriptor spaces can become large and redundant, feature-selection methods are needed to reduce dimensionality while preserving useful chemical information [17]. Tools such as ECoFFeS and MoDeSuS illustrate how evolutionary computation and dedicated descriptor-selection workflows can support QSAR model development by identifying compact and informative feature subsets [17, 18]. Broader reviews of molecular representation and machine-learning best practices emphasize that descriptor choice should be linked to endpoint type, data quality, interpretability needs, and validation strategy rather than treated as a purely technical preference [19, 20].

Applicability Domain and Model Validation

Applicability domain assessment is essential because a QSAR prediction is meaningful only when the query compound is sufficiently related to the chemical and biological space represented in the training data. Conformal prediction provides a principled route for communicating prediction confidence, while cheminformatics implementations such as CPSign show how confidence-based reporting can be integrated into molecular modeling workflows [21, 22]. Large-scale comparisons of QSAR and conformal prediction methods indicate that domain-aware prediction should be evaluated alongside conventional validation rather than added as an afterthought [23]. For an autonomous agent, applicability domain assessment would be part of the required output, enabling users to distinguish between predictions that are likely within the learned domain and predictions that should prompt caution or additional data collection.

Agent System Overview

High-Level Architecture

The proposed agent would operate as a closed-loop QSAR workflow manager that receives a molecular dataset and a target endpoint, then activates modules for curation, descriptor computation, feature selection, model search, validation, applicability-domain assessment, and report generation. The architecture would draw on automated QSAR systems that already connect multiple modeling stages, while adding explicit state tracking and self-critique so that failed checks can trigger workflow revision [1, 2]. Each module would write structured logs describing its inputs, actions, rejected alternatives, and outputs, creating a traceable record suitable for later audit. This architecture would make the agent more than a script runner, because it would evaluate intermediate workflow states and decide whether the process should proceed, pause for review, or return to an earlier step.

Figure 1 presents the proposed autonomous QSAR agent as a traceable left-to-right workflow that converts raw chemical–biological data into a validated predictive model, applicability-domain assessment, and reviewer-ready modeling report.

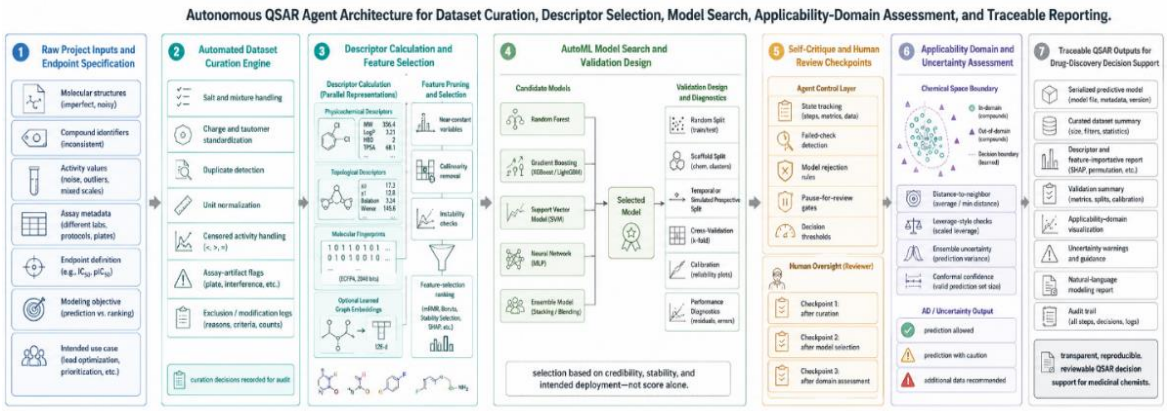

Figure 1. Autonomous QSAR Agent Architecture for Dataset Curation, Descriptor Selection, Model Search, Applicability-Domain Assessment, and Traceable Reporting

Core Inputs and Outputs

The core inputs would be a structured compound file or tabular dataset containing molecular identifiers, chemical structures, activity values, and metadata, together with an endpoint specification that defines the modeling objective. The outputs would include a serialized predictive model, a validation and curation report, descriptor and feature-importance summaries, and an applicability-domain visualization for future compounds. Existing tools such as PoseidonQ and QSPRpred show the value of packaging QSAR model development in reusable software environments, while an autonomous agent would extend this by producing a decision-oriented narrative for non-expert users [7, 13]. The system would therefore treat the report as an essential scientific artifact rather than a cosmetic supplement to the model file.

Design Principles

The agent should be automated by default but interruptible by design, allowing human review after data curation, model selection, and domain assessment. Its design should follow machine-learning best practices in chemistry by prioritizing reproducibility, transparent validation, interpretable reporting, and careful handling of domain shift [20]. Large language model components could assist with workflow explanation and user-facing summaries, but they should remain grounded in outputs from deterministic cheminformatics tools and validated modeling modules [9]. This separation would reduce the risk of unsupported narrative generation while still using language interfaces to make technical QSAR decisions easier to inspect.

Table 1 consolidates the proposed QSAR agent into a decision architecture that clarifies what the agent evaluates, what evidence it uses, when it should pause, and which scientific artifact each module must produce.

Table 1. Agent Decision Architecture across the Autonomous QSAR Workflow

Agent workflow stage	Primary scientific decision	Evidence inspected by the agent	Automated action	Required pause or escalation condition	Traceable output artifact
Endpoint intake and task framing	Whether the dataset and endpoint specification define a coherent QSAR task	Molecular identifiers, structures, assay labels, activity units, endpoint description, intended prediction use	Classify task as regression, classification, ranking, or uncertainty-aware prediction	Endpoint is ambiguous, assay context is inconsistent, or intended use is not compatible with available data	Endpoint specification sheet with modeling objective and decision-use statement
Chemical structure curation	Whether each compound record is chemically usable for modeling	Salt forms, mixtures, stereochemistry, charge state, tautomer handling, duplicate structures, invalid structures	Standardize structures, flag ambiguous records, remove or merge duplicates according to predefined rules	High duplicate burden, unresolved structure conflicts, or extensive excluded records	Structure-curation log with modified, excluded, and retained compounds
Activity normalization	Whether activity values can be compared within a single modeling target	Units, censored values, qualitative labels, assay conditions, replicate measurements, conflicting annotations	Normalize units, transform activity values, reconcile replicates, flag inconsistent records	Incompatible assay formats, unresolved replicate disagreement, or severe class imbalance	Activity-processing report with transformation and exclusion rationale
Validation-split design	Whether the validation strategy reflects intended prospective use	Dataset size, chemical scaffold diversity, temporal metadata, series structure, deployment scenario	Select random, scaffold-aware, temporal, simulated prospective, or cross-validation design	Random split would overstate performance or scaffold diversity is too limited for robust assessment	Validation-design statement with split rationale and limitations
Descriptor and representation selection	Which molecular representation is appropriate for the endpoint and dataset	Physicochemical descriptors, topological descriptors, fingerprints, learned embeddings, descriptor stability	Calculate descriptors, remove redundant or unstable features, rank representation families	Descriptor space is unstable, sparse, excessively redundant, or poorly aligned with interpretability needs	Descriptor inventory and feature-selection rationale
AutoML model search	Which model is credible rather than merely high-scoring	Candidate algorithms, hyperparameters, validation metrics, calibration, variance across splits, interpretability constraints	Train and compare candidate models under predefined validation rules	Performance is unstable, overfitting is detected, or selected model conflicts with domain assumptions	Model-comparison report with selected and rejected alternatives
Applicability-domain assessment	Whether future predictions should be considered in-domain, uncertain, or unreliable	Chemical-space distances, leverage diagnostics, ensemble uncertainty, conformal confidence, query-compound similarity	Assign prediction-status categories and uncertainty warnings	Large fraction of relevant compounds fall outside modeled domain	Applicability-domain map and prediction-use warning rules
Model explanation and reporting	Whether the final model can be communicated responsibly to users	Feature importance, descriptor families, validation results, uncertainty indicators, curation decisions	Generate natural-language report grounded in logged workflow outputs	Explanation implies unsupported mechanism or hides major data limitations	Reviewer-ready QSAR report with audit-linked explanations
Human oversight checkpoint	Whether the workflow is ready for project-level use	Curation report, validation design, model diagnostics, domain assessment, limitations	Present concise approval or revision options to expert user	Expert rejects endpoint framing, curation assumptions, validation design, or deployment readiness	Human-review record with accepted, revised, or rejected decisions

Dataset Curation and Preprocessing Module

Chemical Structure Cleaning and Standardization

The curation module would standardize structures by removing salts, normalizing charges where appropriate, handling tautomers consistently, detecting duplicates, and flagging compounds that may interfere with assay interpretation. Such operations are necessary because automated model building cannot compensate for structurally inconsistent or chemically ambiguous inputs, a concern highlighted by reproducibility studies of QSAR modeling by non-experts [5]. The agent would document every rejected or modified record and would distinguish reversible standardization from exclusion decisions. By connecting these cleaning rules to downstream model diagnostics, the system could make curation an explicit scientific decision rather than a hidden preprocessing routine.

Activity Data Processing

The activity-processing component would normalize endpoint units, handle censored measurements conceptually, convert qualitative labels when classification is requested, and detect values that appear inconsistent with the rest of the assay context. Automated frameworks for imbalanced screening data show that endpoint-specific preprocessing and validation design are especially important when activity classes are uneven or assay data are noisy [24]. Scaffold-aware, privacy-aware, or temporally motivated splitting would be available so that the selected validation design reflects how the model is intended to be used [15, 16]. The agent would record the splitting logic because validation design is one of the main points at which overly optimistic QSAR conclusions can enter a workflow.

Automated Data Quality Report

Before model training, the agent would generate a data-quality report describing structure standardization actions, duplicated records, activity conflicts, excluded compounds, endpoint transformations, and the final validation design. This report would support reproducibility by making visible the same decisions that are often buried in notebooks or informal project notes [5]. Automated QSAR platforms demonstrate that systematic reporting can be embedded in modeling software, but the proposed agent would expand this principle by linking each curation decision to a later modeling consequence [1, 14]. A user could then review whether the curated dataset is scientifically plausible before the agent invests effort in descriptor selection and model search. **Figure 2** illustrates how the proposed agent converts heterogeneous assay activity data into a reproducible, validation-ready QSAR dataset through endpoint normalization, conflict detection, split-design selection, and transparent data-quality reporting.

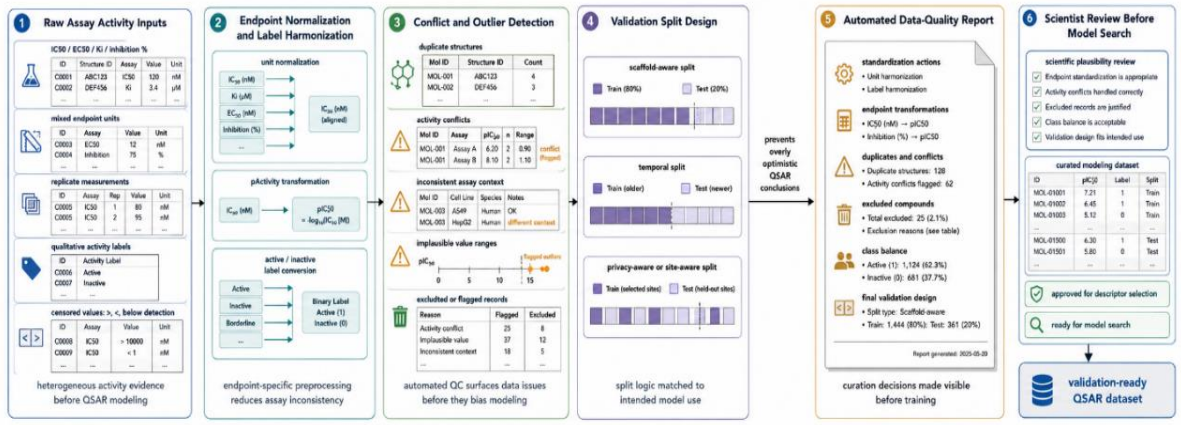

Figure 2. Activity Data Processing and Data-Quality Reporting Workflow for Reproducible QSAR Modeling

Automated Descriptor Selection and Model Building

Descriptor Calculation and Pruning

The descriptor module would compute a broad representation set, including physicochemical descriptors, structural fingerprints, and potentially learned graph-based features, then remove near-constant, redundant, or unstable variables before deeper feature selection. Reviews of molecular representation emphasize that no single representation is universally optimal, so an autonomous agent should evaluate descriptor families in relation to the endpoint, chemical diversity, and interpretability needs [19]. Feature-selection tools such as ECoFFeS and MoDeSuS illustrate how evolutionary search and structured descriptor reduction can be incorporated into QSAR workflows [17, 18]. The agent could also use recursive elimination or Boruta-style strategies guided by baseline models, while preserving a transparent record of why descriptors were discarded or retained.

AutoML Model Search

The model-building module would perform automated search across candidate algorithms and hyper-parameter configurations, using internal validation logic that matches the endpoint and the intended deployment setting. Chemistry-oriented AutoML systems such as QSARtuna, DeepMol, QSPRpred, PoseidonQ, and QSPRmodeler show that molecular prediction workflows

can be made configurable, repeatable, and accessible through integrated software platforms [2, 6, 7, 13, 14]. The proposed agent would add self-monitoring around this search, checking whether a selected model is consistent with curation assumptions, descriptor stability, and applicability-domain requirements. Model selection would therefore be framed not as choosing the most attractive validation score, but as identifying a model that is credible, interpretable, and appropriate for the chemical space under study.

Ensemble and Model Explanation

The final predictive system could be an ensemble when the agent determines that combining complementary models would improve stability or uncertainty characterization without sacrificing interpretability. Ensemble-based uncertainty estimation has been studied for molecular prediction, and such methods could inform how the agent reports uncertainty and flags fragile predictions [25]. Feature-attribution summaries could be generated for model explanation, but they should be interpreted as aids to chemical reasoning rather than definitive mechanistic claims, consistent with broader best-practice guidance for machine learning in chemistry [20]. The agent's report would translate these explanations into a concise narrative describing influential molecular features, known limitations, and the relationship between model interpretation and applicability-domain status.

Applicability Domain Assessment and Model Reporting

Domain Definition and Calculation

The applicability-domain module would estimate whether a new compound lies within the chemical and response space represented by the curated training set, using approaches such as distance-to-neighbor analysis, leverage-style diagnostics, ensemble uncertainty, and conformal prediction. Large-scale comparisons of QSAR and conformal prediction methods support the idea that confidence-aware prediction should be treated as a core modeling output rather than a secondary annotation [23]. Conformal prediction frameworks for drug discovery provide a useful foundation because they can attach interpretable confidence information to predictions without requiring the user to inspect raw model internals [21]. CPSign further illustrates how conformal prediction can be implemented in cheminformatics workflows, making it a natural candidate for automated domain reporting in the proposed agent [22].

Automated Model Documentation

The agent would generate a structured report describing how the dataset was curated, which descriptors were selected, how the model was chosen, and how the applicability domain should be interpreted for future predictions. Automated QSAR tools already demonstrate the value of packaging model construction and validation into reproducible workflows, but an autonomous agent would extend this by producing documentation that is readable by both computational specialists and medicinal chemists [1, 13]. The report would include conceptual descriptions of validation design, calibration behavior, domain boundaries, and interpretation warnings, without presenting unsupported claims of prospective performance. By linking each section of the report to the corresponding workflow log, the system would make model documentation part of the scientific record rather than an after-the-fact summary.

Model Quality Self-Assessment

The agent would perform self-assessment after model selection by checking for signs of overfitting, unstable descriptors, inconsistent validation behavior, or excessive out-of-domain predictions for relevant query compounds. Ensemble uncertainty studies suggest that repeated modeling procedures can be useful for identifying fragile predictions, so the agent could use such signals to decide whether additional diagnostic steps are needed [25]. If weaknesses are detected, the agent would conceptually re-enter earlier stages, such as feature pruning, split redesign, or curation review, rather than presenting a questionable model as complete. This diagnostic loop would reflect the broader principle that machine-learning models in chemistry should be judged by reliability, traceability, and domain appropriateness, not only by apparent fit to historical data [20].

Agent Orchestration, Self-Critique, And Human-In-The-Loop

Workflow Orchestration and State Management

The agent would be organized as a directed workflow in which curation, descriptor generation, feature selection, model search, applicability-domain estimation, and report writing are treated as dependent tasks with recorded inputs and outputs. Chemistry-tool-augmented large language models show how language interfaces can coordinate specialized computational tools while keeping the scientific operations grounded in external software [9]. Reviews of autonomous agents in chemistry suggest that such systems should not merely call tools, but should maintain state, evaluate intermediate outcomes, and revise their plans when a workflow fails to satisfy predefined criteria [12]. For reproducibility, every action would be written to an audit file that allows a user to reconstruct which data, parameters, descriptors, and decision rules produced the final model. Recent LLM-based autonomous chemistry systems further demonstrate that agentic workflows can dynamically plan, execute, and revise domain-specific computational tasks, supporting the feasibility of a QSAR agent that manages curation, modeling, diagnostics, and reporting as coordinated scientific actions [26]. **Table 2** summarizes the main orchestration checkpoints that allow the proposed QSAR agent to connect automated task execution with reproducibility, self-critique, and human review.

Table 2. Orchestration checkpoints for a reproducible QSAR agent workflow

Agent workflow checkpoint	What the agent records	Why it adds scientific value
Task-state tracking	Current workflow step, completed actions, failed actions, and dependencies between tasks	Prevents the agent from treating QSAR modeling as disconnected tool calls and supports reconstruction of the full modeling path
Intermediate self-checks	Data-quality flags, descriptor-generation errors, feature-selection changes, and model-search outcomes	Allows the system to detect weak or unstable steps before final model reporting
Decision-rule logging	Splitting logic, applicability-domain thresholds, model-selection criteria, and rejection rules	Makes methodological choices explicit rather than hidden inside automated execution
Human review trigger	Cases requiring user approval, such as conflicting activity labels, unstable validation results, or poor applicability-domain coverage	Keeps expert judgment involved when automated decisions could affect scientific interpretation
Final audit trail	Dataset version, parameters, selected descriptors, validation design, model outputs, and report-generation history	Supports reproducibility, peer review, and later comparison with alternative QSAR workflows

Human-Review Interface

Although the system would be autonomous, it should pause at scientifically important checkpoints and present concise summaries for human review, especially after curation and after model selection. Studies of QSAR reproducibility by non-experts show that users need support in understanding how modeling choices affect reliability, so the interface should make assumptions visible rather than hiding them behind automation [5]. Large language models designed for chemistry assistance could help convert technical logs into readable explanations, while the underlying decisions would remain tied to deterministic curation, modeling, and validation outputs [27]. This human-in-the-loop design would allow the agent to automate routine work while preserving expert authority over endpoint definition, data acceptance, and deployment readiness.

Integration Into Drug Discovery Workflows

Deployment as a Command-Line or Rest API Tool

The agent could be deployed as a containerized command-line application or REST API so that project teams can invoke QSAR modeling from electronic lab notebooks, compound registration systems, or virtual screening pipelines. Existing open-source QSAR and molecular prediction tools show that reproducible packaging is essential for adoption, because users must be able to rerun workflows and compare outputs across projects [6, 7]. Integration with broader therapeutic data resources would also help the agent align project-specific models with community benchmarks and reusable endpoint definitions [4]. In practice, the agent would function as a service that converts project data into a documented predictive artifact while preserving the provenance needed for later review.

Enabling Non-Experts and High-Throughput QSAR

By automating curation, descriptor selection, model construction, and domain reporting, the agent would allow medicinal chemists to request fit-for-purpose QSAR models without manually assembling a full cheminformatics workflow. Platforms such as PoseidonQ and QSPRmodeler already indicate that accessible interfaces can lower barriers to QSAR model development [13, 14]. Autonomous molecular discovery systems further suggest that prediction tools become more valuable when they are embedded into iterative design cycles rather than treated as isolated analyses [8]. The proposed agent would therefore support high-throughput decision-making while still communicating uncertainty, data limitations, and applicability-domain status in a form that non-specialists can use responsibly.

Evaluation Strategy

Quality and Reproducibility Of Generated Models

The agent should be evaluated by comparing its generated QSAR workflows with expert-built workflows across benchmark resources and project-like datasets, while emphasizing reproducibility, validation appropriateness, interpretability, and robustness to domain shift. MoleculeNet provides a widely used molecular machine-learning benchmark, and Therapeutics Data Commons extends benchmarking toward therapeutic tasks and standardized evaluation settings [3, 4]. Molecular generation benchmarks such as MOSES are not direct QSAR benchmarks, but they illustrate the broader need for clear, reusable evaluation suites in AI-driven chemistry [28]. The evaluation should avoid claiming universal superiority and should instead ask whether the agent produces consistent, inspectable, and scientifically defensible models under clearly defined conditions.

Table 3 provides an evaluation and governance framework for judging the autonomous QSAR agent as a complete scientific modeling system rather than only as a predictive algorithm.

Pharmacophore, 16(4) 2024, Pages 11-21

Table 3. Evaluation and Governance Framework for an Autonomous QSAR Agent

Evaluation dimension	Core question	Suggested assessment approach	Strong-performance indicator	Failure signal	Manuscript-level implication
Dataset-curation reliability	Does the agent improve the transparency and consistency of raw-data preparation?	Compare agent curation logs with expert-curated datasets across multiple endpoints	Most structure, duplicate, unit, and activity conflicts are detected and documented with reproducible rules	Silent exclusions, undocumented transformations, or inconsistent handling of similar records	Establishes whether automation strengthens or obscures the most consequential QSAR preprocessing stage
Validation appropriateness	Does the agent choose validation designs that match prospective use?	Compare random, scaffold-aware, temporal, and simulated prospective splits across benchmark and project-like datasets	Validation design is justified by chemical diversity, endpoint type, and deployment scenario	Overreliance on random splits or inflated performance under scaffold shift	Determines whether the agent supports credible decision-making rather than optimistic retrospective fitting
Descriptor-selection defensibility	Are selected descriptors chemically meaningful, stable, and not unnecessarily complex?	Audit retained and removed descriptor families across repeated runs and endpoint types	Feature set is compact, reproducible, interpretable, and aligned with endpoint biology or chemistry	Unstable descriptor selection, redundant features, or unexplained removal of chemically relevant variables	Clarifies whether the agent can support medicinal chemistry interpretation rather than only numerical prediction
Model-selection robustness	Does AutoML identify models that are reliable under realistic data constraints?	Compare candidate models across repeated splits, calibration checks, and uncertainty diagnostics	Selected model balances performance, calibration, interpretability, and stability	Highest-scoring model is selected despite overfitting, instability, or poor calibration	Frames model selection as scientific judgment embedded in automation
Applicability-domain usefulness	Does the agent help users know when predictions should not be trusted?	Test in-domain and out-of-domain query compounds using chemical-space, uncertainty, and conformal indicators	Domain warnings are specific, interpretable, and linked to prediction-use categories	Predictions are presented without domain status or uncertainty explanation	Makes domain assessment a required output rather than a post hoc annotation
Self-critique behavior	Can the agent detect when its own workflow is scientifically weak?	Trigger controlled failure cases, including noisy labels, small datasets, scaffold imbalance, and descriptor instability	Agent pauses, rejects, or revises workflows when quality thresholds fail	Agent completes workflow despite known data or validation defects	Supports the claim that the system is an autonomous modeling agent rather than a linear script
Human-review effectiveness	Do users understand and appropriately act on agent-generated warnings?	Conduct structured review with cheminformaticians and medicinal chemists	Users can identify key assumptions, limitations, and deployment cautions from the report	Users misinterpret confidence, domain warnings, or feature explanations	Evaluates whether the agent democratizes QSAR responsibly for non-specialists
Reproducibility and auditability	Can the full modeling workflow be reconstructed from saved outputs?	Rerun workflows using stored data, parameters, software versions, descriptor lists, and random seeds	Independent reruns reproduce curation, descriptors, splits, model selection, and reports	Missing provenance prevents reconstruction of results	Establishes the agent's value as a traceable scientific workflow system
Integration readiness	Can the agent be used within real drug-discovery infrastructure?	Test command-line, REST API, containerized, or notebook-based deployment scenarios	Outputs can be consumed by medicinal chemistry teams, ELNs, compound systems, or virtual-screening pipelines	Model artifact, report, or domain output cannot be reused downstream	Links conceptual architecture to practical adoption in drug-discovery workflows
Governance and limitation reporting	Does the system avoid overstating model certainty or biological meaning?	Review generated reports for unsupported claims, missing caveats, and inappropriate	Report clearly separates prediction, uncertainty, feature association, and	Agent-generated narrative implies causal or prospective validity without evidence	Protects scientific credibility and aligns the system with responsible AI use in cheminformatics

mechanistic interpretation	biological hypothesis

Workflow Efficiency and Time Savings

Workflow efficiency should be assessed conceptually by examining how much expert intervention is required to move from raw data to a documented model, how often the agent pauses for review, and whether its audit trail allows the workflow to be repeated. Automated QSAR and AutoML platforms provide a foundation for this comparison because they already reduce manual implementation effort in model construction and configuration [1, 2]. The proposed agent would be expected to improve efficiency mainly by coordinating steps that are often performed separately, such as curation, descriptor filtering, model search, and report generation. Evaluation should therefore focus on process quality and reproducibility rather than unsupported claims about absolute time reduction.

User Acceptance and Trust

User acceptance should be evaluated by asking cheminformaticians and medicinal chemists whether the generated reports are clear, whether the applicability-domain warnings are actionable, and whether the agent's decisions are sufficiently traceable for project use. Prior work on machine-learning best practices in chemistry emphasizes that trust depends on validation design, domain awareness, and transparent reporting, not simply on automated prediction [20]. Feature-selection studies in chemical modeling also show that users need to understand why particular molecular variables were retained or discarded, especially when model interpretation influences design decisions [29]. Feedback from users would be used to refine the human-review interface, the explanation style, and the thresholds that trigger self-critique.

Limitations

Garbage-In-Garbage-Out and the Limits of Automation

The agent cannot overcome fundamentally flawed input data, because misassigned structures, inconsistent endpoints, assay artifacts, or systematic measurement errors can compromise any downstream model. Automated curation can detect many routine problems, but it cannot guarantee that a historical assay is biologically appropriate for a new medicinal chemistry question [5, 24]. Applicability-domain methods can warn when predictions are being made outside the modeled chemical space, but they cannot transform irrelevant training data into a reliable basis for prospective decision-making [22, 23]. Human awareness remains essential when judging whether the available data truly represent the intended biological endpoint and chemical series.

Limited Creativity in Problem Formulation

The agent would automate the standard QSAR paradigm, but it would not automatically invent new mechanistic hypotheses, redefine the biological endpoint, or determine whether QSAR is the right modeling framework for a given project. Reinforcement-learning systems for molecular design show that autonomous algorithms can optimize within a specified objective, but their behavior remains shaped by the goals, representations, and constraints supplied by developers and users [10, 11, 30]. Similarly, large language models can assist with chemistry workflows and explanations, but they should not be treated as independent sources of scientific truth without grounding in curated data and validated tools [9, 12, 27]. The agent would therefore support expert work rather than replace the conceptual creativity required for problem formulation.

Conclusion

An autonomous AI agent for QSAR modeling would coordinate the full workflow from raw chemical–biological data to a documented predictive model. Its core functions would include data curation, descriptor selection, automated model search, validation, applicability-domain assessment, and natural-language reporting.

The main strength of such an agent would be reproducibility. By recording each decision and applying consistent best practices, the system could reduce hidden variation across projects while making QSAR more accessible to users who are not cheminformatics specialists.

The remaining challenges are substantial. The agent would remain dependent on input data quality, would need careful integration into real medicinal chemistry decision-making, and would require prospective validation before being trusted for high-impact project decisions.

Open-source implementation would be important for transparency, peer review, and community improvement. Dedicated benchmarks for autonomous QSAR workflow quality would also be needed so that future systems can be evaluated not only as predictors, but as complete scientific modeling agents.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1):1.
2. Mervin L, Voronov A, Kabeshov M, Engkvist O. QSARtuna: an automated QSAR modeling platform for molecular property prediction in drug design. J Chem Inf Model. 2024;64(14):5365-74.
3. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.
4. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, et al. Artificial intelligence foundation for therapeutic science. Nat Chem Biol. 2022;18(10):1033-6.
5. Patel M, Chilton ML, Sartini A, Gibson L, Barber C, Covey-Crump L, et al. Assessment and reproducibility of quantitative structure-activity relationship models by the nonexpert. J Chem Inf Model. 2018;58(3):673-82.
6. Correia J, Capela J, Rocha M. DeepMol: an automated machine and deep learning framework for computational chemistry. J Cheminform. 2024;16(1):136.
7. van den Maagdenberg HW, Šícho M, Araripe DA, Luukkonen S, Schoenmaker L, Jespers M, et al. QSPRpred: a flexible open-source quantitative structure-property relationship modelling tool. J Cheminform. 2024;16(1):128.
8. Koscher BA, Canty RB, McDonald MA, Greenman KP, McGill CJ, Bilodeau CL, et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. Science. 2023;382(6677):eadi1407.
9. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35.
10. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. J Cheminform. 2017;9(1):48.
11. Blaschke T, Arús-Pous J, Chen H, Margreitter C, Tyrchan C, Engkvist O, et al. REINVENT 2.0: an AI tool for de novo drug design. J Chem Inf Model. 2020;60(12):5918-22.
12. Ramos MC, Collison CJ, White AD. A review of large language models and autonomous agents in chemistry. Chem Sci. 2025;16(6):2514-72.
13. Kabier M, Gambacorta N, Ciriaco F, Mastrolorito F, Kumar S, Mathew B, et al. PoseidonQ: A free machine learning platform for the development, analysis, and validation of efficient and portable QSAR models for drug discovery. J Chem Inf Model. 2025;65(8):3944-54.
14. Bachorz RA, Nowak D, Ratajewski M. QSPRmodeler-An open source application for molecular predictive analytics. Front Bioinform. 2024;4:1441024.
15. Landrum GA, Beckers M, Lanini J, Schneider N, Stiefl N, Riniker S. SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches. J Cheminform. 2023;15(1):119.
16. Simm J, Humbeck L, Zalewski A, Sturm N, Heyndrickx W, Moreau Y, et al. Splitting chemical structure data sets for federated privacy-preserving machine learning. J Cheminform. 2021;13(1):96.
17. Liu ZZ, Huang JW, Wang Y, Cao DS. ECoFFeS: a software using evolutionary computation for feature selection in drug discovery. IEEE Access. 2018;6:20950-63.
18. Martínez MJ, Razuc M, Ponzoni I. Modesus: A machine learning tool for selection of molecular descriptors in QSAR studies applied to molecular informatics. Biomed Res Int. 2019;2019(1):2905203.
19. Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. Wiley Interdiscip Rev Comput Mol Sci. 2022;12(5):e1603.
20. Artrith N, Butler KT, Coudert FX, Han S, Isayev O, Jain A, et al. Best practices in machine learning for chemistry. Nat Chem. 2021;13(6):505-8.
21. Alvarsson J, McShane SA, Norinder U, Spjuth O. Predicting with confidence: using conformal prediction in drug discovery. J Pharm Sci. 2021;110(1):42-9.
22. Arvidsson McShane S, Norinder U, Alvarsson J, Ahlberg E, Carlsson L, Spjuth O. CPSign: conformal prediction for cheminformatics modeling. J Cheminform. 2024;16(1):75.
23. Bosc N, Atkinson F, Felix E, Gaulton A, Hersey A, Leach AR. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. J Cheminform. 2019;11(1):4.
24. Casanova-Alvarez O, Morales-Helguera A, Cabrera-Pérez MÁ, Molina-Ruiz R, Molina C. A novel automated framework for QSAR modeling of highly imbalanced Leishmania high-throughput screening data. J Chem Inf Model. 2021;61(7):3213-31.
25. Dutschmann TM, Kinzel L, Ter Laak A, Baumann K. Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation. J Cheminform. 2023;15(1):49.
26. Zou Y, Zhang Y, Wang S, et al. El Agente: An autonomous agent for quantum chemistry. Matter. 2025.
27. Ishida S, Sato T, Honma T, Terayama K. Large language models open new way of AI-assisted molecule design for chemists. J Cheminform. 2025;17(1):36.
28. Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, et al. Molecular sets (MOSES): a benchmarking platform for molecular generation models. Front Pharmacol. 2020;11:565644.

29. Janet JP, Kulik HJ. Resolving transition metal chemical space: Feature selection for machine learning and structure-property relationships. J Phys Chem A. 2017;121(46):8939-54.
30. Ståhl N, Falkman G, Karlsson A, Mathiason G, Bostrom J. Deep reinforcement learning for multiparameter optimization in de novo drug design. J Chem Inf Model. 2019;59(7):3166-76.