Pharmacophore

ISSN-2229-5402

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

AGENTIC AI WORKFLOW FOR VIRTUAL SCREENING WITH DOCKING, ADMET FILTERING, AND HUMAN HIT TRIAGE

Luis Herrera[1]*, Daniela Rojas[1], Andres Castro[2]

1. ***Department of Computational Pharmaceutical Systems, Faculty of Medicine, Pontifical Catholic University of Chile, Santiago, Chile.***
2. ***Department of AI Drug Engineering, Faculty of Pharmacy, University of Concepcion, Concepcion, Chile.***

ARTICLE INFO

Received:
18 November 2025
Received in revised form:
02 February 2026
Accepted:
09 February 2026
Available online:
28 February 2026

Keywords: Agentic AI, Virtual screening, Molecular docking, ADMET prediction, Human-in-the-loop, Drug discovery automation

ABSTRACT

Virtual screening can identify novel chemical starting points, but the workflow connecting docking, diverse filtering, and expert hit selection is largely manual. This fragmentation becomes especially limiting when screening campaigns expand from focused libraries to ultra-large chemical spaces. Current pipelines often require separate tools, manual file transfers, and subjective expert triage. These discontinuities can slow turnaround and make decision-making inconsistent from one project to another. This article proposes an agentic AI workflow that could autonomously manage the virtual screening cascade. The system docks a compound library, re-scores poses with learned models, filters candidates using multi-parameter ADMET profiles, and presents a prioritized, explainable hit list to a human expert for final triage. The proposed framework includes a docking job manager, a machine-learning re-scoring model, an ADMET prediction suite, a compound selection engine, a human-review dashboard, and an orchestrator agent. These modules would coordinate through standardized interfaces while preserving audit trails for each decision. Such a system would be expected to reduce repetitive data processing and improve consistency in screening documentation. It would allow medicinal chemists to focus attention on high-value judgment calls rather than routine docking, filtering, and ranking operations. An agentic virtual screening workflow could broaden access to advanced computational screening. By combining automation with expert oversight, it could support continuous and adaptive screening across multiple targets.

To Cite This Article: Herrera L, Rojas D, Castro A. Agentic AI Workflow for Virtual Screening with Docking, ADMET Filtering, and Human Hit Triage. Pharmacophore. 2026;17(1):81-90. https://doi.org/10.51847/MKbUv8irD5

Introduction

Virtual screening has become a central strategy for early hit identification because it can prioritize compounds before experimental testing, yet the practical workflow often remains fragmented across library preparation, docking, scoring, filtering, and expert inspection. Deep docking and ultra-large virtual screening frameworks have shown how computational pipelines can extend structure-based discovery into much larger chemical spaces [1], while open-source platforms have demonstrated that scalable docking infrastructure can be organized for broad community use [2]. However, even when docking engines are efficient, the surrounding operations still require substantial manual coordination. An agentic workflow is therefore motivated not by replacing established screening methods, but by connecting them into a coherent, traceable, and expert-supervised system.

Autonomous AI systems in chemistry have begun to show how tool use, planning, and iterative execution can reshape scientific workflows. Large language model systems augmented with chemistry tools can coordinate discrete computational operations in ways that resemble a scientific assistant rather than a single predictive model [3]. More broadly, autonomous chemical research systems illustrate how AI agents can plan tasks, invoke tools, interpret intermediate outputs, and generate structured reports under defined constraints. These developments suggest that the same orchestration paradigm could be adapted to virtual screening, where many steps are procedural, repetitive, and dependent on intermediate quality checks.

Virtual screening poses several challenges that are well suited to workflow-level intelligence. Docking scores can be heterogeneous across targets and protocols, so learned binding-affinity or pose-scoring methods such as KDEEP [4], OnionNet [5], and GNINA [6] provide complementary signals that could be used for re-ranking. At the same time, property prediction

Corresponding Author: Luis Herrera; Department of Computational Pharmaceutical Systems, Faculty of Medicine, Pontifical Catholic University of Chile, Santiago, Chile. E-mail: luis.herrera@gmail.com.

resources such as SwissADME [7], ADMETlab 3.0 [8], and ADMET-AI [9] show that pharmacokinetic and toxicity considerations can be represented computationally during prioritization. A workflow agent could coordinate these signals while maintaining traceability for why each compound was promoted, deprioritized, or sent for human review.

The thesis of this article is that an agentic AI workflow could orchestrate docking, ADMET filtering, and human hit triage into a single automated and auditable pipeline. Such a system would draw from machine-learning advances in molecular property prediction [10], binding prediction [11], and active screening strategies [12], but its defining contribution would be the coordination layer that manages decisions across tools. The proposed workflow keeps the medicinal chemist as the final authority while using the agent to perform repetitive execution, consistency checking, and rationale generation. In this sense, the framework is a system design for expert augmentation rather than a claim of autonomous discovery.

Background

The Virtual Screening Pipeline – Docking to Hit Selection

A conventional virtual screening pipeline usually begins with target preparation and library curation, continues through docking, and ends with post-processing, compound ranking, visual inspection, and hit selection. Ultra-large screening studies have emphasized the importance of infrastructure for library handling, docking execution, and result management [2], while synthon-based virtual library screening has shown how chemical-space design can be integrated into structure-guided discovery [13]. Benchmarks such as DOCKSTRING frame docking as an accessible task for ligand design evaluation, but they also reveal the need for standardized workflows that can compare methods under consistent assumptions [14]. In practice, the bottleneck is often not only the docking calculation, but the coordination of docking outputs with chemical judgment, property filters, and project-specific constraints.

Machine-Learning Scoring Functions for Pose and Compound Re-Ranking

Machine-learning scoring functions offer a route to re-evaluate docking poses and compound rankings using representations that capture protein-ligand interactions beyond classical scoring terms. GNINA combines molecular docking with deep learning so that pose evaluation and scoring can be informed by learned structural patterns [6], while related virtual screening use cases illustrate how such models can be integrated into practical docking workflows [15]. Other approaches, including 3D convolutional binding prediction in KDEEP [4] and contact-based convolutional modeling in OnionNet [5], support the broader idea that learned scoring can complement docking scores. In an agentic workflow, these models would not be treated as independent endpoints, but as re-scoring components invoked after docking to generate a more interpretable and multi-signal ranking. **Figure 1** illustrates how machine-learning scoring functions can operate as a post-docking re-ranking layer, integrating learned protein–ligand interaction signals with classical docking outputs to support more interpretable pose selection and compound prioritization.

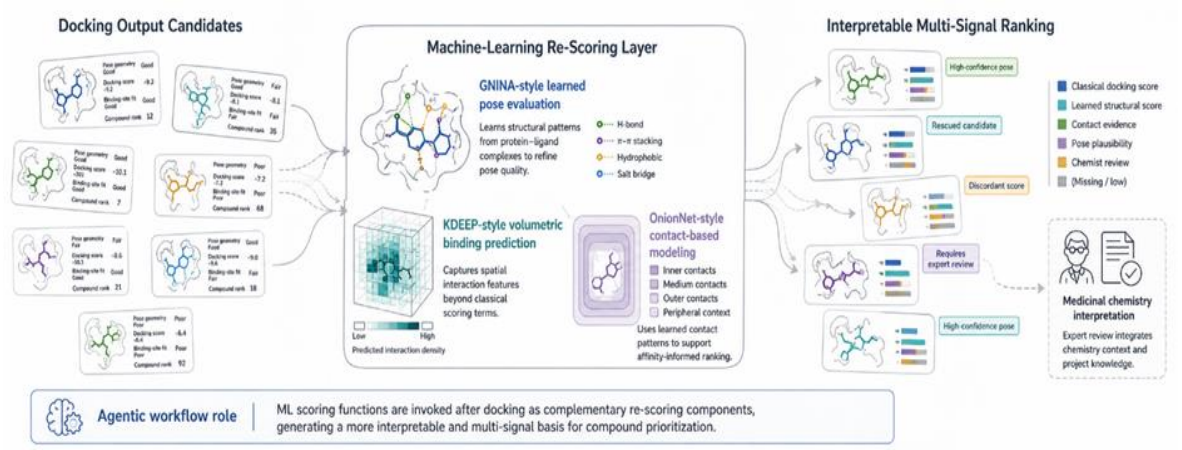

Figure 1. Machine-Learning Re-Scoring as a Multi-Signal Layer for Docking Pose and Compound Prioritization

Table 1 shows key examples of machine-learning-based scoring functions and their integration with docking workflows.

Table 1. Examples of Machine-Learning Scoring Functions in Docking and Virtual Screening

Model/Method	Approach	Role in Docking Workflow	Key Feature
GNINA [6]	Deep learning combined with docking	Re-scores docking poses using learned structural patterns	Integrates molecular docking with CNN-based pose evaluation
KDEEP [4]	3D convolutional neural networks	Predicts binding affinity from 3D structures	Uses volumetric representation of protein-ligand complexes
OnionNet [5]	Contact-based convolutional modeling	Enhances docking score interpretation	Learns interaction patterns from atom-pair contacts

Virtual Screening Integration [15]	Workflow-level application	Applies ML scoring after docking to refine compound ranking	Demonstrates practical incorporation of ML models into docking pipelines

In-Silico ADMET Profiling and Multi-Parameter Optimization

ADMET profiling is essential because a compound with a favorable docking pose may still be unsuitable if it carries liabilities in solubility, permeability, metabolism, transporter interaction, cardiotoxicity, mutagenicity, or other developability dimensions. SwissADME provides a practical example of web-accessible pharmacokinetic and drug-likeness estimation [7], while ADMETlab 3.0 extends this concept into a broader decision-support environment for ADMET prediction [8]. ADMET-AI further illustrates how machine-learning platforms can evaluate large chemical libraries for multiple ADMET-related properties [9]. Within the proposed framework, ADMET predictions would function as both hard alerts and soft prioritization signals, allowing the agent to support multi-parameter decision-making rather than rank compounds by docking alone.

Agentic AI and Autonomous Scientific Workflows

Agentic AI differs from a single predictive model because it plans, invokes tools, monitors intermediate states, and adapts its next action according to the evolving workflow context. Tool-augmented large language models in chemistry show how an AI system can call specialized software and assemble outputs into a broader reasoning process [3]. Autonomous chemical research systems provide an additional precedent for agents that can coordinate experimental or computational steps while documenting task execution. For virtual screening, this means the agent would function as a workflow controller that launches docking jobs, detects failures, invokes re-scoring and ADMET modules, and prepares human-readable evidence for triage.

Human-in-the-Loop as a Design Principle

Human-in-the-loop design is essential because virtual screening decisions depend on context that may not be fully represented in docking scores or predictive models. Reviews of artificial intelligence in drug discovery emphasize that AI can assist medicinal chemistry, but model outputs still require expert interpretation and domain constraints [16]. Perspectives on AI-era drug design similarly argue for rethinking workflows around collaboration between algorithmic suggestions and human judgment [17]. In the proposed system, the agent would advise, summarize, and document, while the human expert would retain authority to accept, reject, override, or redirect screening decisions.

Agentic Workflow Architecture Overview

High-Level Agent Design

The high-level architecture centers on a workflow orchestrator that receives a target structure, a compound library, and project-specific constraints, then decomposes the screen into executable stages. Deep docking demonstrates how AI-guided selection can support structure-based screening at large scale [1], and VirtualFlow Ants illustrates how algorithmic strategies can coordinate docking across very large libraries [18]. An LLM-based planner could add a natural-language interface and tool-calling layer, drawing on chemistry-tool augmentation concepts [3], while still constraining execution to validated docking, scoring, and ADMET modules. The agent would plan docking runs, monitor completion, apply re-scoring and ADMET profiling, generate a prioritized list, and pause for expert review before any downstream commitment.

Figure 2 illustrates the proposed agentic virtual-screening architecture, showing how docking execution, machine-learning re-scoring, ADMET profiling, exception handling, and human hit triage are coordinated through an auditable workflow orchestrator.

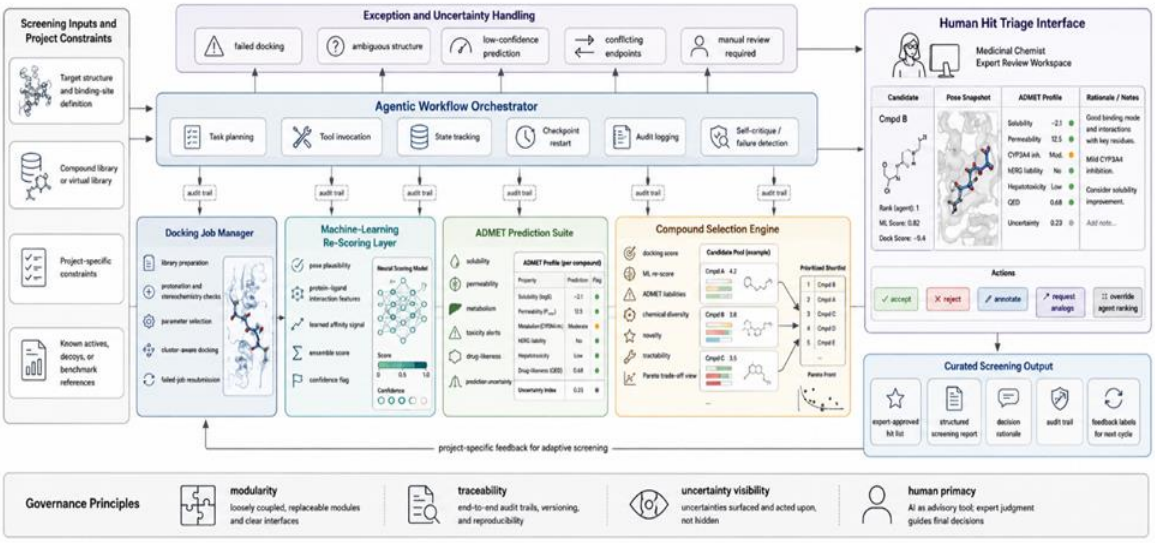

Figure 2. Agentic AI Workflow for Virtual Screening with Docking, ADMET Filtering, and Human Hit Triage

Core Modules and Their Interaction

The core modules include a containerized docking engine, a re-scoring or ensemble model, an ADMET prediction suite, a compound selection engine, and a triage dashboard connected through standardized APIs. The docking layer could draw design lessons from open screening platforms that organize computational resources for reproducible virtual screening [2], while the re-scoring layer could integrate learned binding models such as PotentialNet [11] or protein-ligand affinity predictors [19]. The ADMET layer would attach property predictions and confidence information using tools conceptually aligned with SwissADME [7], ADMETlab 3.0 [8], and ADMET-AI [9]. The compound selection engine would then combine these outputs into a ranked and explainable candidate set for expert triage.

Table 2 defines the functional architecture of the proposed agentic workflow by linking each screening layer to its inputs, agentic actions, human oversight points, and traceability requirements.

Table 2. Functional Architecture of the Proposed Agentic Virtual-Screening Workflow

Workflow layer	Primary function	Key inputs	Agentic actions	Decision-relevant outputs	Human oversight point	Traceability requirement
Screening intake and constraint definition	Converts the scientific screening question into executable workflow conditions	Target structure, binding-site definition, compound library, project goals, known actives or reference compounds	Parses inputs, checks completeness, identifies missing parameters, proposes initial workflow plan	Structured screening configuration; target and library readiness status	Chemist confirms binding-site assumptions, library scope, and project-specific constraints	Record target version, ligand library source, preparation rules, and user-approved constraints
Library preparation and molecular standardization	Ensures compounds are chemically consistent before docking	SMILES, SDF files, stereochemical information, protonation assumptions, salts, tautomers	Standardizes compounds, flags ambiguous structures, prepares docking-ready formats	Cleaned compound set; excluded or flagged molecules; preparation warnings	Chemist reviews ambiguous molecules or compounds with uncertain states	Preserve original and processed structures, transformation rules, and exclusion reasons
Docking job management	Executes structure-based screening at scale	Prepared target, prepared library, docking parameters, computational resources	Submits docking jobs, monitors completion, detects failure, resubmits when appropriate	Docking poses, docking scores, job status, failed-compound list	Chemist reviews repeated docking failures or questionable pose assumptions	Store docking engine, version, parameter files, grid definition, job logs, and failure records
Machine-learning re-scoring	Adds learned protein–ligand interaction signals beyond docking scores	Docking poses, ligand structures, protein–ligand complexes, docking scores	Invokes learned scoring models, generates ensemble or confidence-weighted re-scores	Re-ranked compounds; pose plausibility indicators; confidence flags	Chemist evaluates whether re-scoring agrees with visual pose inspection and project chemistry	Record model identity, version, input representation, confidence indicators, and score transformations
ADMET prediction suite	Profiles developability liabilities before final prioritization	Candidate structures, predicted physicochemical descriptors, ADMET model outputs	Runs multi-endpoint ADMET prediction, attaches uncertainty indicators, flags severe liabilities	Solubility, permeability, metabolism, toxicity, drug-likeness, and uncertainty profile	Chemist decides whether liabilities are exclusionary, acceptable, or target-context dependent	Preserve endpoint-level predictions, model sources, thresholds, and uncertainty notes
Compound selection engine	Integrates docking, re-scoring, ADMET, novelty, diversity, and tractability	Docking scores, ML re-scores, ADMET profiles, similarity metrics, diversity measures	Applies hard filters, soft weighting, Pareto-style trade-off logic, and diversity-aware selection	Prioritized candidate set; trade-off rationale; scaffold-grouped shortlist	Chemist reviews whether ranking logic matches medicinal chemistry priorities	Store ranking formula, filter thresholds, promoted/deprioritized reasons, and alternative candidates
Exception and uncertainty handling	Prevents silent loss or overconfident promotion of uncertain compounds	Failed jobs, conflicting model outputs, low-confidence predictions, inconsistent molecular states	Routes uncertain cases to review, marks unresolved issues, pauses unsafe automation steps	Exception queue; uncertainty report; manual-review requests	Chemist adjudicates uncertain compounds or revises workflow assumptions	Record exception type, agent action, reviewer decision, and final disposition
Human hit triage dashboard	Converts computational outputs into	Candidate cards, poses, scores, ADMET profiles,	Summarizes evidence, explains ranking, supports	Expert-approved hit list; rejected compounds;	Medicinal chemist retains	Preserve review actions, rationale text,

	expert-reviewable evidence	rationales, scaffold groupings	accept/reject/annotate/request-analog actions	annotations; override decisions	final authority over hit selection	user overrides, and reviewer annotations
Adaptive screening feedback	Uses expert decisions and experimental follow-up to refine subsequent cycles	Accepted hits, rejected compounds, uncertain cases, assay outcomes, expert comments	Updates project-specific heuristics, proposes revised compound subsets, reprioritizes exploration	Next-cycle screening plan; refined selection criteria; active-learning labels	Chemist determines whether feedback is sufficient to narrow or broaden the search	Record feedback source, label meaning, update rule, and cycle-to-cycle changes

Design Principles

The proposed system follows four design principles: modularity, reliability, transparency, and human primacy. Modularity is supported by the broader molecular machine-learning ecosystem, where benchmarks such as MoleculeNet encourage interchangeable evaluation of predictive models across tasks [10]. Reliability requires the agent to detect failed docking jobs, inconsistent molecular states, or low-confidence ADMET predictions rather than silently continue. Transparency and human primacy are equally important, because the agent's ranking should be accompanied by a rationale that experts can audit, challenge, or override before compounds proceed further.

Docking Module – Automated Execution and Re-Scoring

Library Preparation and Automated Parameter Selection

The docking module begins with automated inspection of the input library and target structure, including compound standardization, protonation-state handling, stereochemical checks, and selection of docking parameters consistent with the project context. Ultra-large screening systems highlight the importance of robust preparation and workflow control before docking can be meaningfully scaled [2]. Synthon-based screening further shows that chemical library representation can influence how virtual screening explores accessible chemical space [13]. In the proposed agentic system, automated parameter selection would remain auditable, so the chemist could review assumptions such as binding-site definition, treatment of flexible residues, and handling of ambiguous chemical structures.

Cluster-Aware Job Management

Cluster-aware job management is a natural responsibility for the agent because docking campaigns involve many independent tasks that can fail, stall, or complete with inconsistent outputs. VirtualFlow Ants provides a conceptual example of AI-driven docking coordination across very large computational screens [18], and deep docking protocols demonstrate the need for structured execution when screening pipelines scale beyond manual supervision [20]. The agent would submit jobs, monitor status, detect abnormal termination, and resubmit with adjusted settings when appropriate. When repeated failures occur, the system should flag the compound or target configuration for human review instead of hiding the exception.

Machine-Learning Re-Scoring

After docking, the agent would invoke a machine-learning re-scoring module to evaluate poses and compounds using learned representations of protein-ligand interaction patterns. GNINA's integration of docking with deep learning provides one model for combining conventional search with learned scoring [6], while KDEEP [4] and OnionNet [5] show alternative neural approaches for estimating protein-ligand binding affinity from structural information. PotentialNet also illustrates how graph-based molecular learning can support property prediction relevant to compound prioritization [11]. In this workflow, the re-scoring output would be presented as an additional decision layer rather than as an unquestioned replacement for docking scores or expert visual inspection.

Admet Filtering Module – In-Silico Profiling and Prioritization

ADMET Prediction Suite Integration

The ADMET filtering module would attach a structured developability profile to each docked compound before final ranking. SwissADME demonstrates how pharmacokinetic, drug-likeness, and medicinal chemistry friendliness estimates can be made accessible for compound evaluation [7], while ADMETlab 3.0 provides a broader platform for ADMET prediction and decision support [8]. ADMET-AI further illustrates how machine-learning systems can be used to evaluate large chemical libraries across multiple ADMET-relevant endpoints [9]. The agent would record predictions, confidence indicators, and warning flags so that a human reviewer can distinguish promising compounds from candidates with unresolved liabilities.

Rule-Based and Machine-Learning Filters

Filtering should combine rule-based alerts with machine-learning predictions because neither approach alone is sufficient for medicinal chemistry decision-making. Reviews of AI in drug discovery emphasize that predictive models can support prioritization, but their outputs must be interpreted in relation to known chemical liabilities and project goals [21]. Molecular benchmarks such as MoleculeNet provide a foundation for evaluating predictive models across chemically relevant tasks [10],

while ADMET-focused platforms demonstrate how property predictions can be operationalized in screening contexts [8]. In the proposed workflow, hard filters would flag severe liabilities for review, while softer filters would adjust ranking according to the target product profile.

Table 3 shows key strategies for combining rule-based alerts and machine-learning predictions in medicinal chemistry workflows.

Table 3. Integration of Rule-Based and Machine-Learning Filtering in Drug Discovery

Approach	Description	Role in Workflow	Example/Platform
Rule-Based Alerts	Hard-coded chemical rules to flag liabilities	Identify severe chemical issues requiring review	Toxicity alerts, PAINS filters
Machine-Learning Predictions	Models trained on molecular data to predict properties	Support compound prioritization and ranking	MoleculeNet [10], ADMET prediction platforms [8]
Combined Filtering	Integration of rule-based and ML outputs	Balances hard safety flags with softer ranking adjustments	Adjust ranking according to target product profile while flagging critical liabilities
Review & Interpretation	Human assessment in context of project goals	Ensures decisions align with chemical priorities	Medicinal chemist review based on flagged alerts and predicted properties [21]

Multi-Parameter Scoring and Pareto-Optimal Selection

Final prioritization should reflect a multi-parameter balance among docking quality, learned re-scoring, ADMET profile, novelty, and chemical diversity. Pareto-based virtual screening concepts show how multi-objective ranking can expose trade-offs rather than collapsing them prematurely into a single opaque score [22]. Active learning approaches to virtual screening also suggest that compound selection can be guided by uncertainty, diversity, and expected value of subsequent testing [12, 23]. The agent would therefore present a ranked list and an interpretable trade-off view, allowing the human expert to decide whether to favor potency-like signals, cleaner ADMET profiles, novelty, or chemical tractability.

Table 4 operationalizes the proposed triage logic by showing how docking quality, learned re-scoring, ADMET risk, novelty, diversity, tractability, uncertainty, and expert judgment can be converted into transparent hit-prioritization decisions.

Table 4. Decision Logic for Multi-Parameter Hit Prioritization and Human Triage

Decision dimension	What the agent evaluates	Why it matters for hit triage	Example decision rule	When the compound should be promoted	When the compound should be flagged or deprioritized	Human-review question
Docking pose plausibility	Binding orientation, pocket fit, key contacts, steric clashes, pose consistency	A high docking score is not useful if the pose is chemically implausible or structurally unstable	Promote only if the pose is compatible with the binding site and does not depend on obvious artifacts	Compound shows a coherent binding mode and retains meaningful interactions across plausible poses	Pose is strained, clashes with the protein, depends on unrealistic geometry, or fails repeated docking	Does the pose make medicinal chemistry sense for this target?
Docking score strength	Classical docking score relative to screened library and reference compounds	Provides a fast first-pass estimate of structure-based fit	Treat score as a screening signal, not a final potency claim	Compound ranks favorably relative to internal reference compounds or known actives	Score is weak, unstable across runs, or inconsistent with pose plausibility	Is the docking score credible enough to justify further inspection?
Machine-learning re-score	Learned binding-affinity or pose-quality signal from re-scoring models	Can capture interaction patterns missed by classical scoring functions	Use ML re-score as a complementary ranking layer rather than a replacement for docking	ML re-score supports docking rank and improves confidence in the candidate	ML re-score strongly conflicts with docking without a clear explanation	Does learned re-scoring strengthen or weaken confidence in the compound?
ADMET developability profile	Solubility, permeability, metabolism, toxicity alerts, drug-likeness, and uncertainty	Prevents compounds with attractive docking from dominating despite major developability liabilities	Apply severe-liability flags as review triggers and use moderate liabilities as soft ranking penalties	Compound has no severe predicted liabilities or has manageable liabilities for the project context	Compound shows severe predicted toxicity, poor permeability, poor solubility, or multiple unresolved warnings	Are predicted liabilities acceptable for an early hit, or do they outweigh binding promise?
Chemical novelty	Similarity to known actives, scaffold uniqueness, unexplored chemotype status	Encourages discovery of useful starting points rather than redundant analogs	Preserve a subset of chemically novel candidates even when scores are not top ranked	Compound offers a new scaffold or interaction hypothesis with acceptable risk	Compound is redundant with higher-quality candidates or lacks	Does this molecule add new chemical information to the campaign?

					novelty and developability	
Chemical diversity	Scaffold distribution, cluster representation, analog redundancy	Reduces over-selection of one chemical family and supports broader experimental learning	Select representative compounds across high-value scaffold clusters	Compound represents a distinct cluster with acceptable docking and ADMET evidence	Compound duplicates a better-ranked analog without adding interpretive value	Should this scaffold be represented in the experimental shortlist?
Synthetic and procurement tractability	Commercial availability, synthetic accessibility, analog availability, structural complexity	A strong virtual hit is less useful if it cannot be obtained or modified efficiently	Penalize compounds with poor feasibility unless they offer exceptional biological rationale	Compound is purchasable, synthetically plausible, or has accessible analogs	Compound is difficult to source, unstable, or synthetically unattractive	Can this compound realistically enter hit-confirmation or hit-to-lead work?
Prediction uncertainty	Model confidence, applicability-domain fit, conflicting outputs, endpoint instability	Prevents overconfident decisions when models operate outside reliable chemical space	Route low-confidence or conflicting cases to manual review instead of automatic exclusion	Compound has consistent evidence across models or uncertainty is scientifically manageable	Compound has conflicting docking, ML, or ADMET signals with no clear rationale	Is this uncertainty tolerable, or should the compound be held for clarification?
Expert override and annotation	Human accept/reject decisions, medicinal chemistry comments, requested analog directions	Ensures the agent remains advisory and captures project-specific judgment	Expert decision supersedes automated ranking and becomes structured feedback	Chemist accepts compound despite moderate liabilities because it fits project strategy	Chemist rejects compound despite high score because of unmodeled chemistry or target knowledge	What project-specific knowledge changes the automated recommendation?
Adaptive next-cycle value	Ability of the compound to inform the next screening or analog-selection cycle	Supports continuous and learning-oriented screening rather than one-time ranking	Prioritize compounds that improve knowledge of structure–activity, liabilities, or scaffold direction	Compound can test a new interaction mode, scaffold, or ADMET trade-off hypothesis	Compound is unlikely to teach anything new even if it ranks reasonably	Will testing this compound improve the next screening decision?

Human-In-The-Loop Hit Triage Interface

Dashboard and Visualisation

The triage interface would present compounds through a dashboard that links docking poses, key interaction diagrams, re-scoring outputs, ADMET profiles, and similarity to known actives in a single review space. This design follows the broader view that AI in drug discovery should support expert interpretation rather than merely produce ranked lists [24]. Compounds could be grouped by scaffold, liability pattern, or predicted interaction mode so that the chemist can compare alternatives rather than inspect isolated records. The interface would allow the expert to accept, reject, annotate, or request analogs, preserving a transparent record of the decision path.

Explainability and Agent Rationale

For each recommended hit, the agent would generate a concise rationale that connects the compound's docking pose, learned re-scoring behavior, predicted ADMET profile, and chemical novelty. Explainability is particularly important because AI-assisted drug design requires confidence in why a compound is being prioritized, not only awareness that it has been scored highly [17]. The rationale could describe whether a compound was selected because it combines a plausible binding mode with a clean predicted liability profile, or because it represents a diverse scaffold worth expert review. Such explanations should remain evidence-linked and cautious, avoiding claims of biological activity before experimental confirmation.

Closing the Loop – Review Actions Feed Back into the Agent

Human review actions would become structured feedback for subsequent screening cycles. Active learning frameworks for virtual screening show how model-guided selection can iteratively focus the search space based on prior decisions and new observations [12], while self-focusing screening concepts illustrate how design spaces can be pruned adaptively rather than screened uniformly [23]. The agent could record accepted, rejected, and uncertain compounds, then use those labels to adjust re-ranking heuristics or propose a revised compound subset. Importantly, feedback would be treated as project-specific guidance rather than universal truth, because medicinal chemistry decisions depend on target biology, chemical tractability, and development context.

Agent Orchestration, Error Handling, and Self-Critique

Workflow Orchestration and State Management

The orchestrator would manage the screening campaign as a stateful workflow in which every compound has a documented status, such as prepared, docked, failed, re-scored, ADMET-profiled, filtered, or sent for human triage. Tool-augmented chemistry agents provide a useful conceptual precedent for systems that plan actions, invoke external software, and assemble intermediate outputs into a coherent workflow [3]. Autonomous chemical research systems further suggest that agents should maintain task context and produce auditable records of actions taken. In the proposed design, a directed task graph would allow restart from checkpoints and prevent repeated manual reconstruction of failed or interrupted screening runs.

Handling Exceptions and Uncertain Predictions

The agent should treat workflow exceptions as scientific signals requiring attention rather than as operational noise. If docking repeatedly fails, if molecular preparation yields inconsistent forms, or if an ADMET model returns low-confidence predictions, the compound should be flagged for manual review instead of being silently discarded. ADMET-AI and ADMETlab 3.0 show how predictive platforms can support large-scale compound evaluation, but their outputs still need careful interpretation when uncertainty or conflicting endpoints arise [8, 9]. This exception-handling strategy would protect scientific integrity by making uncertainty visible at the point where decisions are made.

Integration Into Hit-To-Lead And Iterative Screening Cycles

From Hit List to Hit-to-Lead

After triage, the agent would export the curated hit list with a structured screening report for the medicinal chemistry team. The report would summarize the target setup, docking assumptions, re-scoring logic, ADMET filters, human decisions, and unresolved uncertainties, aligning with the need for transparent AI-supported drug discovery workflows [16]. Deep-learning-enabled antibiotic discovery illustrates how computational prioritization can support experimental follow-up when candidate selection is carefully documented and interpreted [25]. In this framework, the report would not claim validation, but would provide a defensible rationale for why selected compounds are suitable for follow-up consideration.

Continuous and Adaptive Screening

As experimental feedback becomes available, the agent could incorporate activity labels, inactivity labels, and expert annotations into a new screening cycle. Regression-based active learning for ultra-large library docking illustrates how iterative model updating can guide selection in large chemical spaces [26], while pretraining-informed active learning suggests that prior molecular representations may improve the efficiency of adaptive virtual screening strategies [27]. The workflow could therefore move from a one-time screen to a continuous process that proposes new compounds, receives feedback, and revises its prioritization criteria. The human expert would remain responsible for deciding when the model has enough project-specific evidence to justify narrower or broader exploration.

Evaluation Strategy

Retrospective Enrichment and Hit-Rate Simulation

The proposed system should be evaluated retrospectively before any prospective deployment, using known actives, decoys, and benchmark workflows to compare agent-assisted prioritization with docking-only ranking. LIT-PCBA provides an example of a benchmark designed to support machine-learning and virtual screening assessment while reducing bias in evaluation [28]. DOCKSTRING also illustrates how docking tasks can be standardized for method comparison across ligand design settings [14]. In this article, such evaluation would be framed conceptually: the goal is to assess whether the agent's combined ranking behaves plausibly and transparently, not to report performance numbers.

End-to-End Workflow Efficiency

Workflow efficiency should be evaluated by comparing the structure and continuity of the agent-managed process with a manual screening campaign. Deep docking protocols show that scalable virtual screening requires disciplined handling of library preparation, model-guided selection, docking, and downstream analysis [20]. Open-source ultra-large screening platforms similarly demonstrate that infrastructure and reproducibility are central to practical virtual screening at scale [2]. An evaluation plan could therefore examine whether the agent reduces manual file handling, improves traceability, and shortens the path from input library to expert-ready hit list without claiming quantified gains in this conceptual article.

Expert Satisfaction and Decision Quality

The human-facing dimension should be evaluated through structured expert review rather than through computational benchmarks alone. Medicinal chemists could compare agent-curated lists with conventional workflow outputs and assess whether the proposed rationales, ADMET summaries, and pose visualizations are actionable. Reviews of AI-driven drug discovery emphasize that adoption depends not only on model capability, but also on user trust, interpretability, and alignment with real project decisions [21, 24]. Such evaluation should focus on whether the system improves clarity and consistency of triage while preserving expert authority.

Limitations

Dependency on External Predictive Model Quality

The proposed workflow depends heavily on the quality, applicability domain, and calibration of its docking, re-scoring, and ADMET models. Learned binding predictors such as KDEEP, OnionNet, GNINA, and PotentialNet demonstrate valuable modeling directions, but each model may carry biases from its training data, representation choices, and target coverage [4–6, 11]. ADMET tools likewise provide useful screening signals, yet poor calibration or domain mismatch could promote false confidence in unsuitable compounds [7-9]. The agent must therefore expose uncertainty, preserve manual override, and avoid treating model outputs as experimentally validated facts.

Scalability and Computational Cost

Scalability remains a major limitation because ultra-large virtual screening can require substantial computational resources, careful batching, and disciplined prioritization. Deep docking, VirtualFlow, and synthon-based screening studies demonstrate how large chemical spaces can be approached computationally, but they also imply the need for infrastructure-aware workflow design [1, 2, 13]. An agentic system would need to respect cluster policies, monitor resource usage, and adapt screening depth to project constraints. Without such controls, automation could increase computational burden rather than improve practical discovery operations.

Conclusion

The proposed AIF describes an agentic AI workflow that consolidates docking, ADMET filtering, re-scoring, compound prioritization, and human hit triage into a single orchestrated system. Its purpose is not to claim experimental discovery or to replace medicinal chemistry expertise. Instead, it frames virtual screening as a coordinated workflow in which automation performs repetitive execution while humans retain final judgment.

The main strength of this design is its ability to connect fragmented computational tasks into a traceable and reviewable pipeline. A screening campaign could move from target and library inputs to an expert-ready hit list with consistent documentation of assumptions, model outputs, decision rules, and human overrides. The modular architecture would also allow new docking engines, scoring functions, ADMET predictors, and visualization tools to be substituted as the field advances.

Important challenges remain before such a workflow could be adopted in routine discovery settings. Predictive model fidelity, uncertainty handling, computational scaling, and data governance would all need careful validation. Equally, teams would need to develop trust in an agent that coordinates scientific work while accepting that the agent's recommendations remain advisory rather than authoritative.

Open-source implementations and collaborative pilots would be valuable next steps for evaluating this framework in realistic discovery environments. Such pilots should emphasize transparency, reproducibility, user experience, and responsible human oversight. If developed carefully, agent-driven virtual screening could make advanced computational triage more accessible, adaptive, and scientifically auditable.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, Norinder U, et al. Deep docking: a deep learning platform for augmentation of structure based drug discovery. ACS Cent Sci. 2020;6(6):939-49.
2. Gorgulla C, Boeszoermenyi A, Wang ZF, Fischer PD, Coote PW, Padmanabha Das KM, et al. An open-source drug discovery platform enables ultra-large virtual screens. Nature. 2020;580(7805):663-8.
3. 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.
4. Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. J Chem Inf Model. 2018;58(2):287-96.
5. Zheng L, Fan J, Mu Y. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS Omega. 2019;4(14):15956-65.
6. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. J Cheminformatics. 2021;13(1):43.

7. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7(1):42717.
8. Fu L, Shi S, Yi J, Wang N, He Y, Wu Z, et al. ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res. 2024;52(W1):W422-31.
9. Swanson K, Walther P, Leitz J, Mukherjee S, Wu JC, Shivnaraine RV, et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. Bioinformatics. 2024;40(7):btae416.
10. 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.
11. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, et al. PotentialNet for molecular property prediction. ACS Cent Sci. 2018;4(11):1520-30.
12. Graff DE, Shakhnovich EI, Coley CW. Accelerating high-throughput virtual screening through molecular pool-based active learning. Chem Sci. 2021;12(22):7866-81.
13. Sadybekov AA, Sadybekov AV, Liu Y, Iliopoulos-Tsoutsouvas C, Huang XP, Pickett J, et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature. 2022;601(7893):452-9.
14. García-Ortegón M, Simm GN, Tripp AJ, Hernández-Lobato JM, Bender A, Bacallado S. DOCKSTRING: easy molecular docking yields better benchmarks for ligand design. J Chem Inf Model. 2022;62(15):3486-502.
15. Sunseri J, Koes DR. Virtual screening with Gnina 1.0. Molecules. 2021;26(23):7369.
16. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77.
17. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64.
18. Gorgulla C, Çınaroğlu SS, Fischer PD, Fackeldey K, Wagner G, Arthanari H. VirtualFlow ants-ultra-large virtual screenings with artificial intelligence driven docking algorithm based on ant colony optimization. Int J Mol Sci. 2021;22(11):5807.
19. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics. 2018;34(21):3666-74.
20. Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton AT, Ban F, et al. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nat Protoc. 2022;17(3):672-97.
21. Mak KK, Wong YH, Pichika MR. Artificial intelligence in drug discovery and development. In: Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays. 2024:1461-98.
22. Fromer JC, Graff DE, Coley CW. Pareto optimization to accelerate multi-objective virtual screening. Digital Discovery. 2024;3(3):467-81.
23. Graff DE, Aldeghi M, Morrone JA, Jordan KE, Pyzer-Knapp EO, Coley CW. Self-focusing virtual screening with active design space pruning. J Chem Inf Model. 2022;62(16):3854-62.
24. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G. Artificial intelligence in drug discovery: recent advances and future perspectives. Expert Opin Drug Discov. 2021;16(9):949-59.
25. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A deep learning approach to antibiotic discovery. Cell. 2020;180(4):688-702.
26. Marin E, Kovaleva M, Kadukova M, Mustafin K, Khorn P, Rogachev A, et al. Regression-based active learning for accessible acceleration of ultra-large library docking. J Chem Inf Model. 2023;64(7):2612-23.
27. Cao Z, Sciabola S, Wang Y. Large-scale pretraining improves sample efficiency of active learning-based virtual screening. J Chem Inf Model. 2024;64(6):1882-91.
28. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. J Chem Inf Model. 2020;60(9):4263-73.