Pharmacophore

ISSN-2229-5402

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

CONTRASTIVE MOLECULAR LEARNING FOR ANTIVIRAL HIT PRIORITIZATION USING DOCKING, PROTEASE STRUCTURES, AND BIOACTIVITY DATA

Oliver Grant[1], David Clark[1]*, Sophia Nguyen[2]

1. ***Department of Pharmaceutical Data Analytics, Faculty of Science and Engineering, University of Glasgow, Glasgow, United Kingdom.***
2. ***Department of Intelligent Drug Systems, Faculty of Pharmacy, National University of Singapore, Singapore.***

ARTICLE INFO

Received:
28 February 2025
Received in revised form:
19 May 2025
Accepted:
28 May 2025
Available online:
28 June 2025

Keywords: Contrastive learning, Antiviral drug discovery, Molecular docking, Viral protease, Graph neural network, Bioactivity

ABSTRACT

Antiviral drug discovery targeting viral proteases often begins with structure-based virtual screening, but docking scores alone are unreliable predictors of true biochemical inhibition because they oversimplify complex binding, solvation, and conformational effects. Existing machine learning models typically treat molecular activity as a direct regression or classification target, which can overlook the relational structure among compounds, particularly when docked molecules share interaction patterns yet differ in measured bioactivity. To address these limitations, this article proposes a contrastive molecular learning framework for antiviral hit prioritization, designed to learn an embedding space in which active antiviral compounds cluster near structurally and interactionally similar inhibitors. The model integrates molecular graph encoders, docking-derived interaction fingerprints, protease pocket features, docking scores, and bioactivity labels, using a contrastive objective to bring together compounds with similar activity and interaction profiles while separating inactive or dissimilar molecules. Conceptually, this approach generates a ranked virtual library in which likely antiviral hits are distinguished from docking false positives, with latent space visualization and substructural attribution providing qualitative insight into compound prioritization. By combining binding plausibility with activity-consistent molecular representations, this contrastive framework could enhance the practical utility of virtual screening against viral proteases and reduce unnecessary biochemical testing.

To Cite This Article: Grant O, Clark D, Nguyen S. Contrastive Molecular Learning for Antiviral Hit Prioritization Using Docking, Protease Structures, and Bioactivity Data. Pharmacophore. 2025;16(3):22-31. https://doi.org/10.51847/5j6ZDADI4r

Introduction

The need for antiviral therapeutics remains pressing because emerging and persistent viral infections require rapid discovery pipelines that can respond before extensive clinical chemistry campaigns are feasible. Viral proteases are central drug targets because they process viral polyproteins into functional replication machinery, and their conserved catalytic sites can be interrogated structurally, as shown for SARS-CoV-2 main protease by crystallographic studies [1]. Structure-guided inhibitor design has also benefited from high-resolution protease complexes, including α-ketoamide inhibitor studies that clarified how antiviral ligands can be optimized within the active site [2]. Fragment-screening campaigns against viral proteases further demonstrate that structural information can reveal chemically tractable binding motifs before full inhibitor optimization [3].

A conventional antiviral virtual screening workflow docks large chemical libraries into a target protease pocket, ranks compounds by predicted affinity, and then sends a subset for biochemical validation. Ultra-large screening platforms have made this approach practical at chemical-library scale [4], while deep docking has shown how machine learning can accelerate exploration of very large SARS-CoV-2 main protease libraries [5]. Nevertheless, docking scores are imperfect proxies for inhibition because they can promote molecules with favorable scoring artifacts while missing ligands whose true binding depends on water networks, induced fit, or assay-specific context. Comparative scoring-function work has emphasized that docking scores require careful calibration and should not be treated as direct measures of experimental bioactivity [6].

Contrastive representation learning offers a model-oriented alternative because it learns by comparing related and unrelated examples rather than by fitting each molecule independently. Molecular contrastive learning has shown that graph neural

Corresponding Author: David Clark; Department of Pharmaceutical Data Analytics, Faculty of Science and Engineering, University of Glasgow, Glasgow, United Kingdom. E-mail: david.clark@gmail.com

networks can generate useful embeddings from augmented molecular views [7], while multi-representation approaches indicate that different molecular encodings can be aligned within self-supervised objectives [8]. Knowledge-enhanced and multi-level contrastive frameworks further suggest that molecular representations can absorb higher-level chemical and functional context beyond raw topology [9, 10]. However, applying contrastive learning specifically to antiviral hit prioritization requires integrating docking poses, protease interaction patterns, and bioactivity labels rather than treating the ligand as an isolated graph.

The central thesis of this MDL article is that a contrastive molecular learning framework could refine structure-based antiviral screening by embedding compounds according to both molecular similarity and target-specific binding evidence. In this framework, docking poses provide interaction fingerprints, protease structures provide pocket context, and bioactivity labels define activity-aware neighborhoods in latent space. Protein–ligand contrastive and drug–target interaction studies support the broader feasibility of aligning ligand and target information in representation learning [11, 12]. The intended outcome is not to replace docking, but to re-rank docked libraries so that compounds with plausible poses and activity-consistent embeddings would be prioritized ahead of docking false positives.

Background

Viral Proteases as Drug Targets and Structure-Based Screening

Viral proteases such as SARS-CoV-2 main protease, HIV protease, and HCV protease are attractive antiviral targets because they perform essential catalytic processing steps required for replication. High-resolution structural biology has been especially important for SARS-CoV-2 main protease, where inhibitor-bound structures clarified the catalytic dyad, substrate-binding subsites, and ligandable regions of the active site [1, 2]. Crystallographic fragment screening expanded this view by identifying electrophilic and noncovalent fragments that occupy distinct protease subsites and suggest starting points for medicinal chemistry [3]. In a model-development context, these structures can define the receptor pocket used for docking, the spatial frame for interaction fingerprints, and the target-specific constraints imposed on representation learning.

Molecular Docking: Scoring, Poses, and Interaction Fingerprints

Molecular docking predicts candidate binding poses and estimates binding favorability through scoring functions, but its outputs are best interpreted as structured hypotheses rather than definitive activity measurements. Deep learning-enhanced docking platforms such as GNINA illustrate how pose evaluation can be improved by learning from protein–ligand complexes rather than relying only on handcrafted scoring terms [13]. Cross-docked structural datasets and three-dimensional convolutional models further show that binding poses can be transformed into learnable spatial representations for structure-based drug design [14]. For contrastive learning, docking-derived hydrogen bonds, hydrophobic contacts, salt bridges, and steric clashes can be encoded as interaction fingerprints that complement the molecular graph instead of being reduced to a single docking score.

Bioactivity Data in Antiviral Screening

Experimental bioactivity data, including IC_{50}, EC_{50}, percent inhibition, and active–inactive assay calls, provide the empirical signal needed to distinguish plausible binders from compounds that only score well computationally. Public resources such as PubChem BioAssay collect screening readouts across many biological assays [15], while DrugBank links approved and investigational compounds to targets and mechanisms relevant to antiviral repurposing. BindingDB curates protein–small molecule binding data that can support activity-aware pairing and benchmarking when target annotations and assay conditions are carefully harmonized. In contrastive model development, these data should be used conceptually to define positive and negative relationships among compounds rather than to claim direct prospective performance without validation. **Table 1** summarizes how protease structural information, docking outputs, and experimental bioactivity data can be converted into complementary model inputs for antiviral compound prioritization.

Table 1. Structure-Based and Bioactivity Inputs for Protease-Focused Antiviral Contrastive Learning

Input source	What it contributes to the model	Added value for antiviral screening
Protease crystal structures	Defines receptor pocket, catalytic residues, binding subsites, and ligandable regions	Grounds representation learning in biologically meaningful target geometry
Docking poses and scores	Provides candidate binding orientations and approximate binding favorability	Helps prioritize compounds for further comparison but avoids treating docking score as definitive activity
Interaction fingerprints	Encodes hydrogen bonds, hydrophobic contacts, salt bridges, and steric clashes	Adds target–compound contact information beyond the molecular graph alone
Bioactivity readouts	Supplies IC_{50}, EC_{50}, inhibition percentage, or active–inactive labels	Defines empirical positive and negative relationships for contrastive pairing
Assay and target harmonization	Aligns target identity, assay type, and experimental conditions	Reduces misleading comparisons across incompatible antiviral datasets

Contrastive Representation Learning for Molecules

Contrastive molecular learning trains encoders to bring related molecular views together while separating unrelated views, making it well suited to learning robust embeddings from noisy or incomplete supervision. Molecular graph contrastive

learning has demonstrated that graph augmentations can preserve chemical identity while forcing the encoder to learn invariant structural features [7]. Multi-representation contrastive approaches such as SMICLR show that SMILES, graph, and other molecular views can be aligned to improve representation quality [8]. For antiviral docking workflows, analogous augmentations could perturb nonessential graph features or interaction fingerprints while preserving the protease-binding signal that defines activity-relevant molecular neighborhoods.

Prior Multi-Modal Models for Hit Prioritization

Earlier drug–target affinity models established that ligand and protein information can be jointly learned for binding prediction, with sequence-based models such as DeepDTA providing an influential template for paired compound–protein learning [16]. GraphDTA extended this direction by representing compounds as molecular graphs while learning drug–target binding affinity from paired inputs [17]. Interpretable affinity models such as DeepAffinity also showed that sequence and compound features can be combined while retaining some explanatory structure [18]. However, simple concatenation of ligand, protein, and docking descriptors may not fully exploit relational structure among docked compounds, whereas contrastive objectives can explicitly organize compounds by shared activity and interaction patterns.

Model Development Overview

High-Level Framework

The proposed framework begins with a contrastive pre-training stage in which pairs of docked compounds are formed using bioactivity class, interaction fingerprint similarity, and docking score agreement. Positive pairs would include active compounds that share protease-contact patterns, whereas negative pairs would include active–inactive contrasts or chemically similar decoys with divergent interaction evidence. Supervised graph co-contrastive learning for drug–target interaction prediction supports the idea that relational labels can be incorporated into contrastive objectives for binding-related tasks [19]. After pre-training, the encoder would map a docked virtual library into latent space, and compounds would be ranked by proximity to known active inhibitors and by consistency with target-specific interaction features.

Figure 1 illustrates the proposed contrastive molecular learning workflow in which molecular graphs, docking-derived protease interaction fingerprints, docking scores, and bioactivity labels are integrated into an activity-aware embedding space for antiviral hit prioritization.

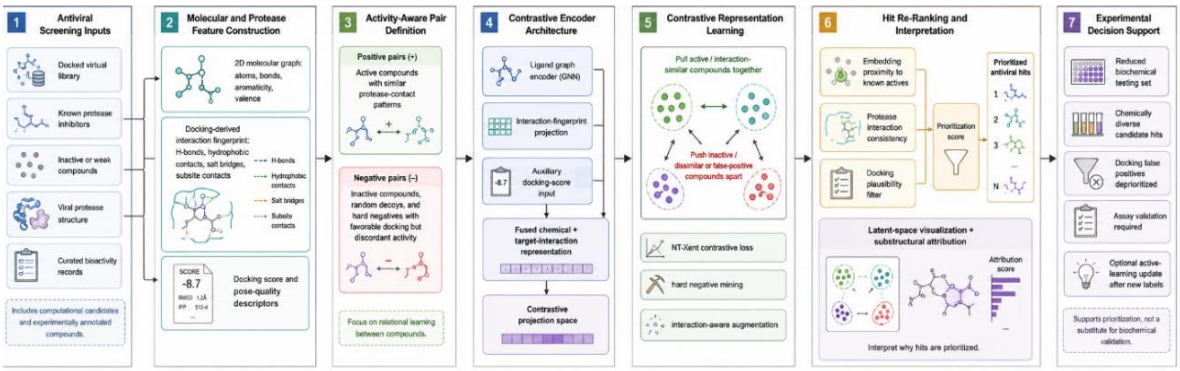

Figure 1. Contrastive Molecular Learning Workflow for Antiviral Hit Prioritization Using Docking, Protease Structures, and Bioactivity Data

Core Inputs

Each compound would be represented by a two-dimensional molecular graph, a docking-derived interaction fingerprint, a docking score, and, when available, an experimental bioactivity label. Molecular property prediction studies have shown that learned molecular representations can capture task-relevant chemical information from graph and descriptor inputs [20]. Large molecular benchmarks such as MoleculeNet provide a conceptual basis for organizing diverse molecular prediction tasks, although antiviral hit prioritization requires target-specific docking and assay context rather than generic property labels [21]. The model would therefore treat the ligand graph as the chemical identity layer, the interaction fingerprint as the structure-based target layer, and the bioactivity label as the empirical supervision layer.

Design Principles

The model design should encourage interaction-aware invariance rather than simple memorization of docking scores or scaffold identities. Self-supervised graph transformer work on molecular data suggests that pretraining can produce transferable chemical representations when the encoder is exposed to large and varied molecular structure signals [22]. Three-dimensional molecular representation learning, as in Uni-Mol, further motivates the inclusion of conformational or spatial context when the binding pose is informative [23]. The proposed encoder would ideally process new molecules after docking-derived features are generated, while remaining flexible enough to support inference workflows where ligand graph information and previously computed interaction descriptors are reused.

Data Sources and Preprocessing

Assembly of Training and Screening Sets

Training data would be assembled from known protease inhibitors, curated bioactivity records, and docked candidate molecules selected from screening libraries. DrugBank can provide approved, investigational, and repurposing-relevant compounds with annotated biological targets, while PubChem BioAssay can provide assay-level activity calls that help define active and inactive training examples [15]. BindingDB can contribute target-resolved binding measurements when assay metadata are compatible with the viral protease of interest.. For screening, the candidate set would be docked into the protease active site, following the general logic of large-scale virtual screening platforms and SARS-CoV-2 main protease deep docking campaigns [4, 5].

Molecular Graph and Interaction Fingerprint Encoding

Each ligand would be encoded as a molecular graph whose nodes represent atoms and whose edges represent bonds, with chemical features such as atom type, valence, aromaticity, and bond order used as input attributes. Graph-based affinity models demonstrate the practical value of representing compounds as graphs for drug–target binding tasks [17], while molecular graph pretraining studies show that such encoders can support downstream prediction after representation learning [22]. The protein–ligand interaction fingerprint would encode contact categories derived from the docked pose, including hydrogen bonds, hydrophobic contacts, ionic contacts, and proximity to protease subsites. This combined encoding allows the model to distinguish molecules with similar scaffolds but different target-contact patterns, which is essential for protease-focused hit prioritization.

Table 2 defines how ligand structure, protease docking evidence, docking scores, and bioactivity labels contribute distinct but complementary information to the proposed contrastive hit-prioritization framework.

Table 2. Activity-Aware Input Representation Strategy for Contrastive Antiviral Hit Prioritization

Representation Layer	Primary Data Elements	Conceptual Role in the Framework	Why It Adds Value Beyond Docking Score Alone	Preprocessing or Harmonization Need	Main Risk if Poorly Defined
Ligand molecular graph	Atom type, bond order, aromaticity, valence, formal charge, ring membership, subgraph motifs	Captures intrinsic chemical identity and scaffold-level relationships among candidate compounds	Allows the model to recognize chemical similarity, substructural motifs, and scaffold variants that may share antiviral relevance despite different docking scores	Standardize protonation states, tautomers, salts, stereochemistry, and graph featurization	The encoder may learn inconsistent chemical representations or overfit to trivial scaffold identity
Docking pose	Predicted ligand orientation within the viral protease pocket	Provides target-specific spatial context for each screened molecule	Distinguishes molecules that appear chemically similar but occupy protease subsites differently	Receptor preparation, pose selection, binding-site definition, and clash filtering	Incorrect poses may teach the model misleading structure–activity relationships
Protease interaction fingerprint	Hydrogen bonds, hydrophobic contacts, salt bridges, catalytic-site contacts, subsite occupancy, steric clashes	Converts docking output into interpretable interaction evidence	Retains mechanistic binding information that is lost when docking is reduced to a single score	Define contact thresholds, subsite labels, interaction categories, and pose-quality filters	Weak or noisy contacts may be mistaken for activity-relevant protease-binding patterns
Docking score	Predicted binding favorability or scoring-function output	Serves as an auxiliary plausibility filter and optional pair-weighting signal	Helps exclude implausible binders while preventing docking score from dominating hit selection	Calibrate score ranges across docking runs, receptor conformations, and compound libraries	The model may reproduce docking artifacts rather than learn activity-aware prioritization
Bioactivity label	Active/inactive assay call, IC_{50}, EC_{50}, percent inhibition, potency category	Defines empirical neighborhoods for positive and negative contrastive pair construction	Anchors the embedding space to experimental inhibition evidence instead of computational affinity alone	Harmonize assay type, target identity, concentration thresholds, activity cutoffs, and duplicate records	Inconsistent labels may collapse biologically different compounds into incorrect neighborhoods
Positive contrastive pair	Compounds sharing activity class and similar protease-contact evidence	Pulls activity-consistent and interaction-consistent	Makes the model learn relational similarity rather than	Define pair rules using both activity and interaction-fingerprint similarity	Overly broad positive pairs may blur meaningful potency or

		compounds closer in latent space	independent activity prediction		mechanism differences
Negative contrastive pair	Inactive compounds, random decoys, or hard negatives with favorable docking but discordant activity	Pushes false positives and biologically dissimilar compounds away from active neighborhoods	Directly targets the docking false-positive problem central to the manuscript	Balance easy negatives with hard negatives; prevent label leakage	Poor negative sampling may exaggerate performance or fail to separate realistic false positives
Virtual screening library	Predocked candidate compounds with graph and docking-derived features	Provides the inference set to be embedded and re-ranked	Converts the trained contrastive space into a practical prioritization tool	Ensure library compounds are processed using the same graph, docking, and fingerprint pipeline	Distribution shift may reduce ranking reliability for novel chemotypes

Defining Positive and Negative Pairs

Positive contrastive pairs would be compounds that share an activity label and exhibit similar interaction fingerprints within the same protease pocket. Negative pairs would include experimentally inactive compounds, random library molecules, or hard negatives that dock favorably but differ from known actives in bioactivity or key protease contacts. Semi-supervised heterogeneous graph contrastive learning for drug–target interaction prediction supports the value of combining labeled and relational information when direct labels are sparse or imbalanced [24]. This pair definition would allow the model to learn from antiviral assay imbalance by emphasizing relative similarity and dissimilarity rather than relying only on absolute class counts.

Contrastive Learning Architecture

Graph Encoder Backbone

The ligand encoder would use a graph neural network backbone, such as a graph isomorphism network or graph attention network, to transform the molecular graph into a fixed-length embedding. A separate multilayer perceptron would project the docking-derived interaction fingerprint into the same latent space, allowing additive fusion, concatenation, or gated integration with the molecular embedding. Protein–ligand contrastive learning and drug–target affinity studies suggest that paired molecular and target information can be aligned to support binding-related inference [11, 12]. The fused representation would therefore encode both intrinsic chemical structure and protease-specific interaction evidence before the contrastive projection head maps it into the training space.

Contrastive Loss and Augmentation

The training objective would use a normalized temperature-scaled contrastive loss to pull together related compound embeddings and push apart unrelated or activity-discordant embeddings. Molecular augmentations such as atom masking, bond perturbation, and subgraph sampling are motivated by graph contrastive learning, where augmented molecular views encourage robust representation learning [7]. Multi-level molecular contrastive pretraining further supports using more than one similarity level, such as scaffold similarity, functional similarity, and activity similarity, when constructing training pairs [10]. For antiviral docking, interaction fingerprint augmentation could mask weak or uncertain contacts while preserving catalytically meaningful contacts, encouraging the model to focus on stable protease-binding signals.

Incorporating Docking Score as an Auxiliary Signal

Docking score would be incorporated as an auxiliary signal rather than as the sole target of the model, because scoring functions can rank plausible poses but may not reliably identify true inhibitors. Comparative scoring-function studies show that docking scores should be interpreted cautiously and evaluated against experimental binding data rather than accepted as direct activity estimates [6]. Deep learning docking systems such as GNINA indicate that learned pose assessment can complement classical docking, but they also reinforce the need to integrate structural evidence with broader molecular context [13]. In the proposed architecture, docking score agreement could weight contrastive pairs or inform a secondary ranking head, while bioactivity and interaction similarity would remain central to the learned antiviral hit-prioritization space.

Integrating Docking Scores and Bioactivity Labels

Docking Score as a Baseline and Filter

Docking score should function as a baseline structural plausibility filter rather than as the dominant predictor of antiviral activity. Structure-based virtual screening studies against SARS-CoV-2 main protease illustrate how docking can rapidly identify plausible candidates, but such lists still require additional prioritization because favorable poses do not necessarily imply biochemical inhibition [25]. Covalent docking workflows for main protease inhibitors further show that docking can be especially useful when the binding mechanism and warhead geometry are explicitly modeled [26]. In the proposed framework,

a meta-classifier would combine the contrastive embedding with docking score so that molecules with severe steric clashes or implausible poses are not promoted solely because they resemble known actives in graph space.

Using Quantitative Bioactivity for Ranking

When quantitative activity values such as IC_{50} are available, they can guide the geometry of the embedding space by placing potent inhibitors closer to one another than weak or inactive analogues. Read-across affinity modeling, as illustrated by SimBoost, supports the idea that similarity relationships among compounds and targets can inform binding prediction without treating every compound independently [27]. DeepAffinity similarly motivates the use of interpretable compound–protein representations when transforming molecular and target information into affinity-relevant predictions [18]. In this conceptual MDL framework, embedding distance to known potent inhibitors could be used as a local ranking signal, while the manuscript avoids claiming numerical potency prediction without prospective validation.

Handling Imbalanced Bioactivity Data

Antiviral screening datasets are often imbalanced because confirmed inhibitors are much rarer than inactive or untested compounds, making direct classification vulnerable to majority-class bias. Contrastive learning can address this issue conceptually by focusing on pairwise and neighborhood relationships rather than only on global class frequencies. Semi-supervised contrastive drug–target models indicate that labeled and unlabeled relationships can be combined when direct activity annotation is incomplete [24], while graph contrastive drug–target affinity work suggests that molecular semantics can improve discrimination among related compounds [28]. Hard negative mining would be especially important because inactive compounds with good docking scores are precisely the false positives the framework is designed to deprioritize.

Model Interpretability and Hit Prioritization

Visualization of the Contrastive Latent Space

The contrastive latent space would be inspected using dimensionality-reduction methods that project compound embeddings into a two-dimensional visualization colored by activity label, docking score category, or protease-contact pattern. Such visualization should be interpreted qualitatively, because the purpose is to determine whether known actives occupy coherent neighborhoods rather than to claim a numerical screening result. Molecular contrastive learning studies show that learned embeddings can organize compounds in chemically meaningful ways [7], and knowledge graph-enhanced molecular contrastive learning further suggests that functional context can sharpen representation structure [9]. In antiviral hit prioritization, compounds that dock well but lie far from active neighborhoods would be treated as possible docking false positives.

Substructural Attribution

Substructural attribution would help chemists understand whether the model prioritizes compounds because of plausible pharmacophoric features rather than irrelevant scaffold correlations. Graph neural network models for molecular prediction can support atom- or bond-level interpretation through attention, gradient, or perturbation analyses, and learned molecular representation studies have emphasized the need to inspect what chemical features models encode [20]. Three-dimensional binding models such as K_DEEP show that spatial protein–ligand information can be learned from complex structures, motivating attribution methods that connect ligand substructures to pocket interactions [29]. For viral proteases, attribution should highlight groups responsible for catalytic-site recognition, subsite occupancy, or interaction-fingerprint similarity to known inhibitors.

Table 3 summarizes how latent-space visualization and substructural attribution can be interpreted as qualitative checks for chemically meaningful antiviral hit prioritization.

Table 3. Interpretation Checks for Contrastive Latent-Space Visualization and Substructural Attribution

Interpretation layer	What should be inspected	Added research value
Contrastive latent-space neighborhoods	Whether known active compounds cluster in coherent embedding regions when colored by activity label, docking category, or protease-contact pattern	Helps distinguish chemically plausible active-like candidates from isolated docking hits that may represent false positives
Docking–embedding agreement	Whether high-scoring docked compounds are positioned near active neighborhoods rather than distant from them	Connects structure-based docking evidence with representation-learning evidence instead of relying on docking score alone
Substructural attribution	Whether highlighted atoms, bonds, or fragments correspond to catalytic-site recognition, subsite occupancy, or known protease-interaction patterns	Supports chemical interpretability by showing whether the model focuses on pharmacophoric features rather than irrelevant scaffold correlations
Hit-prioritization decision use	Whether a compound has both favorable embedding proximity and interpretable substructure–pocket relevance	Provides a practical qualitative filter for selecting candidates for experimental follow-up

Integration Into the Antiviral Discovery Pipeline

Deploying the Trained Model as a Scoring Function

After pretraining, the contrastive encoder and ranking head could be wrapped as a scoring function that re-ranks a pre-docked antiviral screening library. High-throughput virtual screening workflows demonstrate that computational triage is most useful when it can reduce large candidate pools into chemically plausible subsets for purchase or synthesis [4]. Deep docking against SARS-CoV-2 main protease further illustrates how machine learning can be inserted into virtual screening to accelerate prioritization before experimental testing [5]. In the proposed workflow, the final score would combine docking plausibility, embedding proximity to known actives, and interaction-fingerprint consistency with the viral protease pocket.

Iterative Active Learning

The framework could be embedded in an iterative active-learning loop in which newly tested compounds provide updated bioactivity labels for contrastive fine-tuning. A prospective screening campaign that validates a main protease noncovalent inhibitor illustrates how computational prioritization and biochemical testing can be linked in a discovery cycle [30]. The model would use new active and inactive labels to refine positive and negative pair definitions, especially around chemically similar compounds that docking alone cannot separate. Over successive rounds, the latent space would be expected to become more target-specific, although each update should be evaluated prospectively rather than assumed to improve discovery outcomes.

Evaluation Strategy

Retrospective Enrichment Metrics

Retrospective evaluation should compare contrastive re-ranking against docking-only ranking, graph-only prediction, and conventional affinity-learning baselines using held-out active and inactive compounds. Benchmarking principles from MoleculeNet support standardized evaluation across molecular prediction tasks, although antiviral protease prioritization requires target-specific splits and assay-aware labeling [21]. Deep learning affinity models such as DeepDTA and GraphDTA provide relevant baselines because they learn compound–target relationships from paired molecular and protein inputs [16, 17]. Metrics such as enrichment factors, BEDROC, and precision–recall curves could be calculated conceptually, but this MDL article does not report numerical outcomes.

Table 4 organizes the proposed evaluation strategy around whether the contrastive framework improves antiviral hit prioritization beyond docking-only ranking, graph-only learning, and conventional affinity-prediction baselines.

Table 4. Evaluation and Decision-Use Framework for Contrastive Molecular Hit Prioritization

Evaluation Dimension	Core Question Addressed	Recommended Comparator	Appropriate Evidence or Metric	Interpretation for Antiviral Hit Prioritization	Key Methodological Safeguard
Docking-only baseline comparison	Does contrastive re-ranking improve over conventional docking-score prioritization?	Ranked compounds by docking score alone	Enrichment factor, BEDROC, early precision, precision–recall curve	Improvement would suggest that activity-aware embeddings reduce reliance on noisy scoring-function artifacts	Use identical docked library and receptor preparation across methods
Graph-only molecular baseline	Does protease-interaction information add value beyond ligand chemistry alone?	Molecular graph classifier or graph-based ranking model without docking fingerprints	AUROC, AUPRC, scaffold-split performance, early enrichment	Better contrastive performance would indicate that target-specific docking interactions contribute meaningful prioritization signal	Use scaffold-aware splits to avoid overstating generalization
Drug–target affinity model baseline	Does contrastive relational learning outperform direct affinity-style prediction?	DeepDTA-like or GraphDTA-like paired compound–target model	Ranking performance, calibration, enrichment among held-out actives	Advantage would support the manuscript's claim that pairwise activity/interactions provide structure beyond direct regression or classification	Keep target annotations and assay labels harmonized across models
Hard negative discrimination	Can the model identify molecules that dock well but lack activity-consistent evidence?	Docking-score ranking and standard supervised classifier	False-positive reduction rate among high-scoring docked compounds	Strong performance would address the central practical problem of docking false positives	Construct hard negatives from compounds with favorable docking but inactive or discordant bioactivity
Latent-space organization	Do known active inhibitors form coherent neighborhoods in	Untrained embeddings, graph-only embeddings, or	Qualitative UMAP/t-SNE visualization, cluster purity,	Coherent active neighborhoods would support the	Treat visualization as qualitative support, not proof of

	the learned embedding space?	docking-score-only grouping	nearest-neighbor activity consistency	interpretability of contrastive organization	prospective performance
Substructural attribution	Are prioritized hits supported by plausible antiviral pharmacophoric or protease-contact features?	Attribution from graph-only or docking-only models	Atom-level attribution, contact-level attribution, medicinal chemistry review	Useful attribution would help chemists judge whether rankings reflect meaningful binding motifs	Require expert review to avoid overinterpreting noisy attention or gradient maps
Timestamp-based prospective simulation	Would the model retrieve later-discovered inhibitors when trained only on earlier records?	Docking-only, graph-only, and supervised molecular baselines under the same timestamp split	Retrieval of later actives, early enrichment, precision among top-ranked candidates	Strong performance would better approximate real discovery use than random retrospective splits	Enforce strict temporal separation to reduce information leakage
Experimental validation readiness	Which ranked compounds should be sent first for biochemical testing?	Chemically diverse top-ranked set from docking-only prioritization	Hit rate, biochemical inhibition, potency confirmation, structural follow-up when available	The model is useful only if it improves the quality of compounds selected for real assays	Maintain biochemical validation as the final decision standard
Transfer to new viral protease	Can the learned ligand encoder support new protease targets after recalibration?	Target-specific model trained from scratch or docking-only screening	Performance after fine-tuning, target-specific enrichment, interaction-fingerprint recalibration	Transferability would support broader antiviral utility but should not be assumed	Rebuild pocket-specific interaction fingerprints and pair definitions for each protease
Active-learning update value	Do new assay labels improve future ranking rounds?	Static model without fine-tuning	Change in enrichment, reduction in false positives, diversity of newly prioritized hits	Improvement would support iterative discovery deployment	Evaluate each update prospectively rather than assuming all new labels improve performance

Prospective Virtual Screening Simulation

A prospective virtual screening simulation should use a timestamp split, where the model is trained on earlier bioactivity records and evaluated on later-discovered inhibitors. This design reduces information leakage and better approximates how the model would behave in a real antiviral campaign. Public binding resources such as BindingDB can support time-aware construction of target-specific activity records when metadata are sufficiently curated, while PubChem BioAssay can provide complementary assay-level annotations for activity calls [15]. The contrastive model should be compared with docking-only selection and supervised molecular baselines to assess whether activity-aware embedding neighborhoods retrieve newly reported actives more effectively in principle.

Prospective Experimental Validation

Prospective validation would require selecting a small, chemically diverse set of predicted hits for biochemical testing against the target viral protease. SARS-CoV-2 main protease inhibitor discovery studies show that structure-guided screening becomes most meaningful when computational predictions are connected to experimental inhibition assays and structural follow-up [1, 3]. Covalent and noncovalent main protease campaigns also demonstrate that mechanism, binding pose, and assay confirmation must be interpreted together when judging candidate hits [26, 30]. In the proposed evaluation, the model's role would be to justify which compounds should be tested first, not to replace biochemical validation.

Limitations

Dependence on Docking Accuracy and Target Structure

The framework is limited by the quality of docking poses and the structural accuracy of the protease model used to generate interaction fingerprints. If the binding site is incorrectly defined, if the protein conformation is inappropriate, or if docking fails to capture water-mediated interactions, the contrastive model may learn misleading structural relationships. Cross-docked data and three-dimensional convolutional studies show that pose and receptor-context variation are central challenges in structure-based learning [14], while docking score assessments caution against overinterpreting scoring outputs as activity

measurements [6]. Therefore, the model should be treated as a prioritization aid whose reliability depends on receptor preparation, pose generation, and assay-consistent validation.

Generalization to New Protease Targets

A model trained on one viral protease may not transfer directly to another protease because pocket topology, catalytic geometry, substrate-recognition motifs, and ligand interaction patterns can differ substantially. Three-dimensional molecular pretraining frameworks such as Uni-Mol suggest that broader spatial representation learning may improve transferability [23], but target-specific fine-tuning would still be needed when the binding pocket changes. Self-supervised molecular graph transformers also motivate pretraining on broad chemical structure before adapting to specialized antiviral tasks [22]. For new protease targets, the safest strategy would be to reuse the ligand encoder while recalibrating interaction-fingerprint projections and pair definitions with target-specific docking and bioactivity evidence.

Conclusion

The proposed MDL framework presents contrastive molecular learning as a conceptual strategy for antiviral hit prioritization. It integrates molecular graph structure, docking-derived protease interactions, docking scores, and experimental bioactivity labels into a shared embedding space. In this space, compounds would be prioritized not only because they dock favorably, but because they resemble known inhibitors in activity-aware and interaction-aware ways.

A key strength of the framework is its ability to treat docking outputs as structured evidence rather than as final answers. By learning relationships among compounds, it could reduce the influence of noisy docking scores and support more coherent ranking of antiviral candidates. The approach also provides a natural route for interpreting predictions through latent-space organization and substructural attribution.

Important challenges remain before such a framework could be considered reliable in practice. Its usefulness would depend on accurate receptor structures, meaningful docking poses, harmonized bioactivity labels, and prospective testing against the intended viral protease. Transfer to new viral targets would require careful adaptation because different proteases may reward different interaction patterns and chemical motifs.

Future work should emphasize open pretrained models, transparent benchmark datasets, and evaluation protocols that distinguish retrospective fit from prospective utility. Shared antiviral protease datasets containing docked poses, interaction fingerprints, and harmonized bioactivity labels would make contrastive learning methods easier to compare. Such resources could accelerate adoption of representation learning in antiviral discovery while keeping biochemical validation at the center of decision-making.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020;582(7811):289-93.
2. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science. 2020;368(6489):409-12.
3. Douangamath A, Fearon D, Gehrtz P, Krojer T, Lukacik P, Owen CD, et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nat Commun. 2020;11(1):5047.
4. 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.
5. Ton AT, Gentile F, Hsing M, Ban F, Cherkasov A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. Mol Inform. 2020;39(8):2000028.
6. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. J Chem Inf Model. 2019;59(2):895-913.
7. Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell. 2022;4(3):279-87.
8. Pinheiro GA, Da Silva JL, Quiles MG. SMICLR: contrastive learning on multiple molecular representations for semisupervised and unsupervised representation learning. J Chem Inf Model. 2022;62(17):3948-60.
9. Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell. 2023;5(5):542-53.

10. Zhang X, Xu Y, Jiang C, Shen L, Liu X. MoleMCL: a multi-level contrastive learning framework for molecular pre-training. Bioinformatics. 2024;40(4):btae164.
11. Zhang Y, Huang C, Wang Y, Li S, Sun S. CL-GNN: contrastive learning and graph neural network for protein-ligand binding affinity prediction. J Chem Inf Model. 2025;65(4):1724-35.
12. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proc Natl Acad Sci U S A. 2023;120(24):e2220778120.
13. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. J Cheminform. 2021;13(1):43.
14. Francoeur PG, Masuda T, Sunseri J, Jia A, Iovanisci RB, Snyder I, et al. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. J Chem Inf Model. 2020;60(9):4200-15.
15. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373-80.
16. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17):i821-9.
17. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics. 2021;37(8):1140-7.
18. Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics. 2019;35(18):3329-38.
19. Li Y, Qiao G, Gao X, Wang G. Supervised graph co-contrastive learning for drug-target interaction prediction. Bioinformatics. 2022;38(10):2847-54.
20. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model. 2019;59(8):3370-88.
21. 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.
22. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, et al. Self-supervised graph transformer on large-scale molecular data. Adv Neural Inf Process Syst. 2020;33:12559-71.
23. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-Mol: a universal 3D molecular representation learning framework. In: Int Conf Learn Represent. 2023.
24. Yao K, Wang X, Li W, Zhu H, Jiang Y, Li Y, et al. Semi-supervised heterogeneous graph contrastive learning for drug-target interaction prediction. Comput Biol Med. 2023;163:107199.
25. Gahlawat A, Kumar N, Kumar R, Sandhu H, Singh IP, Singh S, et al. Structure-based virtual screening to discover potential lead molecules for the SARS-CoV-2 main protease. J Chem Inf Model. 2020;60(12):5781-93.
26. Amendola G, Ettari R, Previti S, Di Chio C, Messere A, Di Maro S, et al. Lead discovery of SARS-CoV-2 main protease inhibitors through covalent docking-based virtual screening. J Chem Inf Model. 2021;61(4):2062-73.
27. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform. 2017;9(1):24.
28. Yang X, Yang G, Chu J. GraphCL-DTA: a graph contrastive learning with molecular semantics for drug-target binding affinity prediction. IEEE J Biomed Health Inf. 2024;28(8):4544-52.
29. Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J Chem Inf Model. 2018;58(2):287-96.
30. Clyde A, Galanie S, Kneller DW, Ma H, Babuji Y, Blaiszik B, et al. High-throughput virtual screening and validation of a SARS-CoV-2 main protease noncovalent inhibitor. J Chem Inf Model. 2021;62(1):116-28.