REQVS: RELIABLE VIRTUAL SCREENING VIA INTERACTION ENTROPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-based virtual screening has emerged as a scalable approach for drug discovery by enabling compound prioritization directly from protein sequences without requiring 3D structures. However, its practical application remains limited by unreliable predictions on new targets and unexplored chemical space, largely due to the uncertainty introduced when inferring structure-dependent protein–ligand interactions from sequence information alone. Here, interaction uncertainty is quantified using an entropy-based estimator, while an ensemble mechanism integrates complementary information from heterogeneous sequence-based interaction predictors. Based on this, **ReqVS** is developed as a reliability-aware sequence-based virtual screening framework that prioritizes compounds maximizing predicted interaction scores while minimizing the uncertainty to produce robust compound rankings. Benchmark evaluations further show that ReqVS achieves performance comparable to structure-based methods while improving ranking robustness across targets. When applied to the challenging GPCR target HCAR1, ReqVS identified a high-affinity antagonist, representing the first experimentally validated HCAR1 antagonist discovered using sequence-based virtual screening. These results highlight reliability-aware sequence-based screening as a practical paradigm for structure-independent drug discovery.

1 INTRODUCTION

In drug discovery, virtual screening aims to identify promising compounds from large chemical libraries before experimental testing. The information available about the protein target determines how this prioritization can be performed. Structure-based virtual screening (SBVS) uses a three-dimensional representation of the target to evaluate candidate binding poses and protein–ligand interactions, typically through molecular docking and scoring McGibbon et al. (2023); Cao et al. (2024); Yang et al. (2023); Burley et al. (2021); He et al. (2025b); Borkakoti & Thornton (2023). Its application therefore depends on suitable receptor structures and binding-site information, which can be difficult to obtain for poorly characterized targets or particular functional states Zheng et al. (2018); Laurie & Jackson (2005); Kozakov et al. (2015); Tubiana et al. (2022). Sequence-based virtual screening (SeqBVS) offers an alternative: it predicts protein–ligand interactions directly from the target sequence and molecular representations, without requiring a target structure as an input Chen et al. (2023); Zhang et al. (2025); Wan et al. (2019); Li et al. (2022); Chatterjee et al. (2023); Huang et al. (2021); Lu et al. (2025).

Existing sequence-based predictors use molecular graphs, attention mechanisms, and pretrained representations to predict drug–target interactions (DTI) or drug–target affinities (DTA) Nguyen et al. (2021); Yang et al. (2022); Bian et al. (2023); Bai et al. (2023); Koh et al. (2024); Chen et al. (2025a); Zhao et al. (2025). Despite these advances, sequence-based screening approaches face two practical failure modes. First, compounds ranked among the top-K candidates may exhibit weak affinity or no measurable activity in subsequent wet-lab experiments, such that a high predicted score does not reliably translate into an experimental hit. Second, this mismatch can worsen when a predictor is transferred to a new target, particularly when interaction patterns or associated chemical space are poorly represented in the training data Chatterjee et al. (2023); Zhao et al. (2025); Chen et al. (2025a). These limitations are reflected in divergent rankings across predictors (Figure 1a) and unreliable top-ranked candidates (Figure 1b). A major source of these problems is the lack of direct structural information, as protein structure and interaction patterns shape ligand recognition

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and are directly linked to biological function Wang et al. (2026); Tang et al. (2026); Conflitti et al. (2025). Sequence-based predictors must therefore infer structure-dependent protein–ligand interactions from sequence alone. In SeqBVS workflows, a sequence-based predictor is first used to reduce a large chemical library, after which the top-ranked compounds may be subjected to molecular docking, structure-based rescoring, or interaction-based filtering Zhao et al. (2025); Shah et al. (2025); Kitchen et al. (2004); Ricci-Lopez et al. (2021); Chen et al. (2025b). However, this post-processing stage may fail when the target structure or relevant functional state is unknown. Therefore, how to improve the reliability of *sequence-only* virtual screening under epistemic uncertainty arising from unobserved structural information remains an important research question.

Our Work. In this paper, to quantify this uncertainty, we introduce **interaction entropy**, a statistical measure that characterizes the uncertainty of protein–ligand interaction under *only* sequence-based modeling. High entropy corresponds to high uncertainty and provides a natural estimator of uncertainty in sequence-based interaction inference. However, directly computing this entropy is nontrivial due to the absence of ground-truth annotations and structural information. To address this, the proposed interaction entropy estimates uncertainty across a heterogeneous ensemble of SeqBVS models. Intuitively, the proposed interaction entropy maps predictions from diverse sequence-based models for a given compound into a unified probabilistic space and is estimated through the entropy of predictions induced by heterogeneous SeqBVS models. By quantifying this uncertainty, compounds whose sequence-based interaction predictions are more confident and reliable can be systematically identified. Building on this measurement, the **Re**liability-aware S**eq**uence-based **V**irtual **S**creening (ReqVS) framework is proposed, which *prioritizes compounds that maximize predicted interaction scores while minimizing the uncertainty*. The framework flexibly integrates SeqBVS models of diverse architectures, leveraging their complementary strengths to enhance the robustness of compound ranking, reduce false positives, and facilitate more efficient prioritization of experimentally actionable molecules. The source code and implementation are available at `https://anonymous.4open.science/r/ReqVS-Anonymize-69F2`.

Our main contributions are:

- **Interaction entropy.** We introduce interaction entropy as a practical surrogate for uncertainty in sequence-based protein–ligand interaction prediction. Our latent-structure analysis relates uncertainty arising from unobserved binding configurations to conditional mutual information and characterizes the surrogate approximation error under explicit assumptions. This analysis motivates a ranking objective that jointly considers predicted interaction strength and interaction entropy for compound prioritization.
- **An extensible screening framework and pipeline.** ReqVS integrates heterogeneous DTI and DTA models through consensus weighting and entropy-regularized ranking. Evaluations on DUD-E Mysinger et al. (2012) and DEKOIS2.0 Bauer et al. (2013) show improved early enrichment over constituent models and competitive performance against reported structure-based baselines.
- **A prospective application to HCAR1-directed antagonist discovery.** We apply ReqVS to HCAR1, an emerging therapeutic target in colorectal cancer He et al. (2025a). Screening a commercial library of approximately ten million compounds with ReqVS, followed by expert-guided filtering, identified cpd60, which exhibited micromolar activity in cellular cAMP assays. Docking and molecular dynamics analyses suggest a possible inhibitory mechanism involving vestibular binding and altered receptor conformations.

2 REQVS

2.1 PROBLEM FORMULATION

Let $t \in \mathcal{T}$ denote a protein target and $c \in \mathcal{C}$ denote a compound in the candidate library. Let $s_t \in \mathcal{X}_p$ and $s_c \in \mathcal{X}_c$ denote the protein sequence and compound representation, respectively. Let $Y_{t,c} \in \mathcal{Y} \subset \mathbb{R}$ denote the biochemical interaction strength between target t and compound c. Sequence-based interaction models are trained on experimental interaction data to predict this quantity, $\hat{y}_{t,c} = f_\theta(s_t, s_c)$, leading to the conventional virtual screening objective

$$\tilde{c} = \arg\max_{c \in \mathcal{C}} \hat{y}_{t,c}. \tag{1}$$

For top-K screening, the $\arg\max(\cdot)$ operation is replaced by $\mathrm{TopK}(\cdot)$.

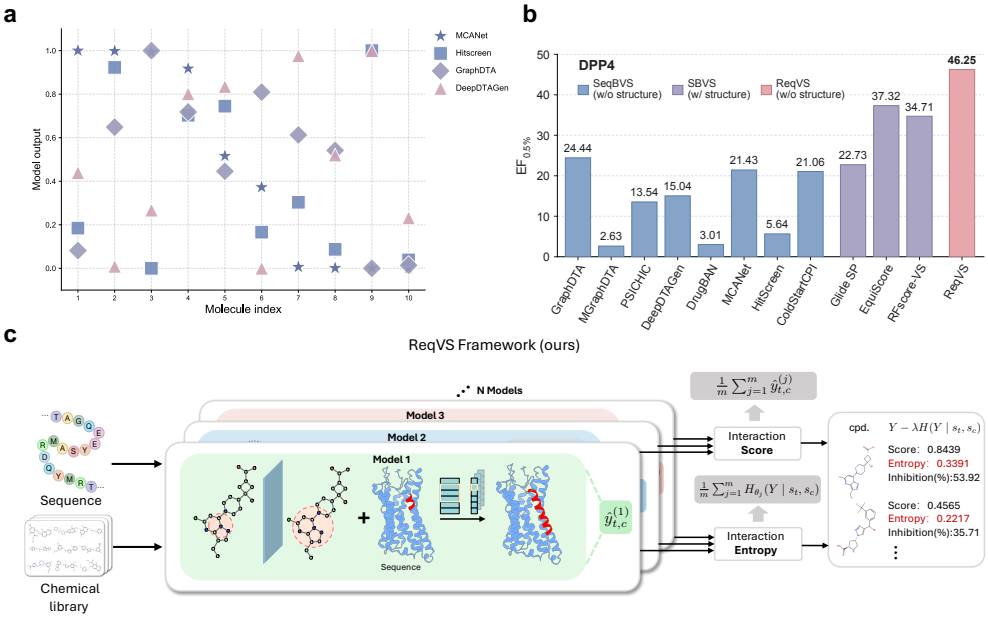


Figure 1: **ReqVS: reliability-aware sequence-based virtual screening. a** Predicted scores of five randomly selected active and inactive molecules from the DUD-E dataset for the AA2AR target. **b** Early enrichment ($\mathrm{EF}_{0.5\%}$) on the DUD-E DPP4 target. Eight individual SeqBVS models and ReqVS operate without a target structure as input (w/o structure), whereas the three SBVS baselines use target structures (w/structure). **c** Illustration of the proposed ReqVS framework. ReqVS integrates predictions from heterogeneous SeqBVS models and jointly estimates interaction scores and interaction entropy to produce reliability-aware compound ranking results.

To account for uncertainty in sequence-based predictions, ReqVS incorporates interaction entropy into the screening objective. The entropy term, $H(Y \mid s_t, s_c)$, is estimated using heterogeneous sequence-based interaction models. The reliability-aware screening objective is formulated as

$$\tilde{c} = \arg\max_{c \in \mathcal{C}} \Big(\hat{y}_{t,c} - \lambda H(Y \mid s_t, s_c) \Big), \tag{2}$$

where λ controls the trade-off between predicted interaction strength and interaction uncertainty. The objective in Equation 2 prioritizes compounds with high predicted interaction scores and low estimated uncertainty.

2.2 Interaction Entropy Estimation

Let $S \in \mathcal{S}$ denote the structural configuration (e.g., binding pose and conformation) of the protein–ligand complex. The biochemical interaction strength is primarily determined by the underlying structural configuration rather than the sequence itself.

Proposition 1 (Latent Structural Interaction Model)**.** *Given a protein sequence s_t and a compound sequence s_c, the true interaction outcome can be expressed as a marginal distribution over latent structural configurations:*

$$P(Y \mid s_t, s_c) = \int_{\mathcal{S}} P(Y \mid S)\, P(S \mid s_t, s_c)\, dS, \tag{3}$$

where $P(S \mid s_t, s_c)$ denotes the distribution over feasible structural configurations induced by the sequences, and $P(Y \mid S)$ characterizes the interaction outcome determined by a specific structure.

In practical sequence-based virtual screening, the structural configuration S is unobserved. As a result, models must implicitly marginalize over all possible structures compatible with the sequences, leading to inherent uncertainty in interaction prediction, which is defined as follows.

Proposition 2 (Sequence-induced Interaction Uncertainty)**.** *Interaction uncertainty is defined as the conditional mutual information between the interaction outcome and the structural configuration, given the sequences:*

$$I(Y; S \mid s_t, s_c) = H(Y \mid s_t, s_c) - H(Y \mid S), \tag{4}$$

where $H(\cdot)$ denotes Shannon entropy. This quantity measures the information about the interaction outcome that depends on the latent structural configuration and is therefore not captured by sequence information alone.

Interaction uncertainty can be interpreted as the dependence of the interaction outcome on latent structural configurations. A small interaction uncertainty indicates that the interaction behavior is largely determined by sequence information and is relatively insensitive to structural variations, which implies that sequence-only models can reliably capture the interaction. In contrast, a large interaction uncertainty suggests that the interaction depends on specific structural configurations and the associated conformational space, making sequence-only modeling inherently more difficult and less reliable.

In practice, the structural variable S is unobserved, making $H(Y \mid S)$ intractable. From the mutual information decomposition, the following relationship is obtained:

$$H(Y \mid s_t, s_c) = \underbrace{I(Y; S \mid s_t, s_c)}_{\text{Sequence-induced Interaction Uncertainty}} + \underbrace{H(Y \mid S)}_{\text{Intrinsic Noise}}, \tag{5}$$

where the first term captures uncertainty induced by missing structural information, and the second term reflects intrinsic variability such as experimental noise and conformational fluctuations.

This equation shows that the entropy under sequence-only modeling consists of two components: the uncertainty caused by missing structural information and the intrinsic biochemical variability. The latter mainly arises from experimental noise and conformational variability, and is expected to vary slowly across different compounds. Therefore, the variation of the entropy under sequence-only modeling is primarily dominated by the former term. As a result, the sequence-based interaction entropy is used as a practical **surrogate** for interaction uncertainty, i.e., $I(Y; S \mid s_t, s_c) \approx H(Y \mid s_t, s_c)$. The approximate error bound is presented in the Appendix A.

We next describe how predictive distributions from sequence-based models are used to construct the interaction entropy estimate.

Interaction Entropy Measurement. The deep learning model f_θ learns the mapping from protein and compound sequences to interaction strength, and can be used to approximate the entropy $H(Y \mid s_t, s_c)$ under sequence-only modeling, denoted as $H_\theta(Y \mid s_t, s_c)$. $P_\theta(Y \mid s_t, s_c)$ is used to represent the predictive distribution modeled by f_θ. In this paper, interaction entropy is measured using Shannon entropy as follows:

$$H_\theta(Y \mid s_t, s_c) = -\int_Y P_\theta(y \mid s_t, s_c) \log P_\theta(y \mid s_t, s_c)\, dy. \tag{6}$$

Probabilistic Output Modeling. For a general sequence model f_{θ^j}, the output is transformed into a probabilistic prediction over a binary interaction label space $\mathcal{Y} = \{0, 1\}$. Specifically, the probabilistic output is defined as $P_{\theta^j}(y = 1 \mid s_t, s_c) = \sigma(f_{\theta^j}(s_t, s_c))$, where $\sigma(\cdot)$ is a mapping function that converts model outputs into probabilities, as defined in Section B.2. For this binary predictive distribution, the interaction entropy is computed as:

$$\begin{aligned} H_{\theta^j}(Y \mid s_t, s_c) = & -\big(1 - \sigma(f_{\theta^j}(s_t, s_c))\big) \log\big(1 - \sigma(f_{\theta^j}(s_t, s_c))\big) \\ & - \sigma(f_{\theta^j}(s_t, s_c)) \log \sigma(f_{\theta^j}(s_t, s_c)). \end{aligned} \tag{7}$$

Finally, the entropy is approximated by averaging across models:

$$H(Y \mid s_t, s_c) \approx \frac{1}{M} \sum_{j=1}^{M} H_{\theta^j}(Y \mid s_t, s_c). \tag{8}$$

In the ReqVS setting, the DTI and DTA entropy estimates are first obtained separately and then combined using the same category-level weights as the interaction scores.

2.3 SCREENING WORKFLOW

Given a target protein sequence s_t and a candidate compound library $\mathcal{C}$, ReqVS performs reliability-aware virtual screening in five steps.

Step 1: heterogeneous DTI and DTA models generate predictions for each target–compound pair, and their outputs are mapped to a common probability scale. DTI scores are retained, whereas DTA scores are min–max normalized over the screening library.

Step 2: the models are evaluated using their agreement with the ensemble consensus, which provides a label-free estimate of model reliability.

Step 3: the mapped interaction scores are aggregated using the resulting model weights.

Step 4: the per-model interaction entropies are aggregated using the same weights.

Step 5: compounds are ranked according to the entropy-regularized score, and the top-K candidates are selected.

The complete screening pipeline, including mapping function, ensemble weighting, and compound ranking, is provided in Appendix B.2 and Appendix B.3. Algorithm 1 summarizes the complete procedure.

Algorithm 1 Reliability-aware sequence-based virtual screening

Require: Target sequence s_t, compound library $\mathcal{C}$, models $\{f_{\theta^j}\}_{j=1}^{M}$, λ, and K

Step 1: Sequence-based model inference

for $j = 1, \ldots, M$ **do**
  for all $c \in \mathcal{C}$ **do**
    $\hat{y}_c^j \leftarrow f_{\theta^j}(s_t, s_c)$
    $p_c^j \leftarrow \sigma(\hat{y}_c^j)$
    $h_c^j \leftarrow -p_c^j \log p_c^j - (1 - p_c^j)\log(1 - p_c^j)$
  end for
end for

Step 2: Consensus-based model evaluation

$\bar{p}_c \leftarrow M^{-1} \sum_{j=1}^{M} p_c^j$

$\boldsymbol{\omega} \leftarrow \textsc{ConsensusWeights}(\{p_c^j\}, \{\bar{p}_c\})$

Step 3–4: Score and entropy ensemble

for all $c \in \mathcal{C}$ **do**
  $\hat{Y}_c \leftarrow \sum_{j=1}^{M} \omega_j p_c^j$
  $H_c \leftarrow \sum_{j=1}^{M} \omega_j h_c^j$
end for

Step 5: Reliability-aware ranking

$R_c \leftarrow \hat{Y}_c - \lambda H_c$ for all $c \in \mathcal{C}$

$\mathcal{C}_K \leftarrow \textsc{TopK}_{c \in \mathcal{C}}(R_c)$

Ensure: Ranked candidate set $\mathcal{C}_K$

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

To evaluate the virtual-screening performance of ReqVS, the framework is compared with a diverse set of SBVS baselines using the results reported in EquiScore Cao et al. (2024). These baselines include traditional docking methods such as Glide Halgren et al. (2004), descriptor-based machine-learning methods such as RFScore Wójcikowski et al. (2017), DeltaVinaXGB Lu et al. (2019), and DeltaVinaRF Wang & Zhang (2017), as well as recent deep-learning methods including EquiScore, DeepDock Méndez-Lucio et al. (2021), and OnionNet Zheng et al. (2019). The methods are evaluated on two widely used benchmarks, DEKOIS2.0 Bauer et al. (2013) and DUD-E Mysinger et al. (2012). DEKOIS2.0 consists of 81 targets, each containing 40 active compounds and 1,200 decoys. DUD-E contains 102 targets and 22,886 active compounds, with approximately 220 actives

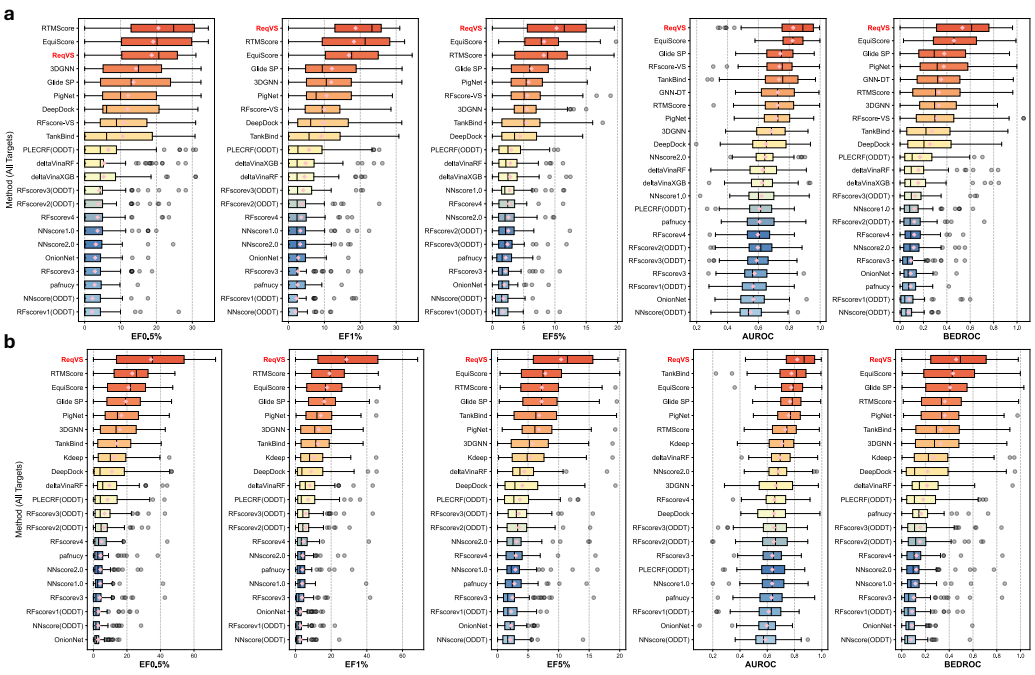


Figure 2: **Methods evaluation on the DEKOIS2.0 and DUD-E benchmarks.** Pink diamonds denote the mean value for each metric; a median line is included in each box-whisker plot. **a** Top row: Model performance on DEKOIS2.0 (data point $n = 81$) evaluated by EF0.5%, EF1%, EF5%, AUROC, and BEDROC(α=80.5). **b** Model performance on DUD-E (data point $n = 102$), evaluated by the same five metrics.

and 13,800 decoys per target. Because ReqVS does not rely on structural docking poses, all metrics are calculated without considering multiple poses. Detailed datasets, baselines, and ensemble configurations are provided in Appendix B.

3.2 Competitive Performance of ReqVS against Structure-based Virtual Screening

ReqVS is first evaluated on the DEKOIS2.0 benchmark. As shown in Figure 2a, the framework demonstrates strong performance across the full target set. ReqVS outperforms all comparative methods in EF1% (18.69; median 23.25), EF5% (10.27; median 11.50), AUROC (0.825; median 0.888), and BEDROC (0.533; median 0.612). Notably, ReqVS surpasses the second-best method by substantial margins of 2.6%, 24.1%, 0.5%, and 15.9% in these metrics, respectively. Although ranking third in mean EF0.5%, ReqVS achieves a higher median value (20.67) than Equiscore (20.17), indicating improved stability and fewer failure cases across heterogeneous targets.

On the larger-scale DUD-E benchmark, ReqVS demonstrates competitive performance (Figure 2b). It achieves the best results across all evaluated metrics. Compared to the second-best method, ReqVS improves mean performance by 47.8%, 50.0%, 32.9%, 5.3%, and 6.3% across these metrics, respectively.

Together, these results suggest that interaction uncertainty modeling contributes to improved ranking stability in sequence-based virtual screening. This consistent reliability is further corroborated in the evaluation of the subset illustrated in Appendix C.1, which consists of targets unseen by baselines trained on PDBbind.

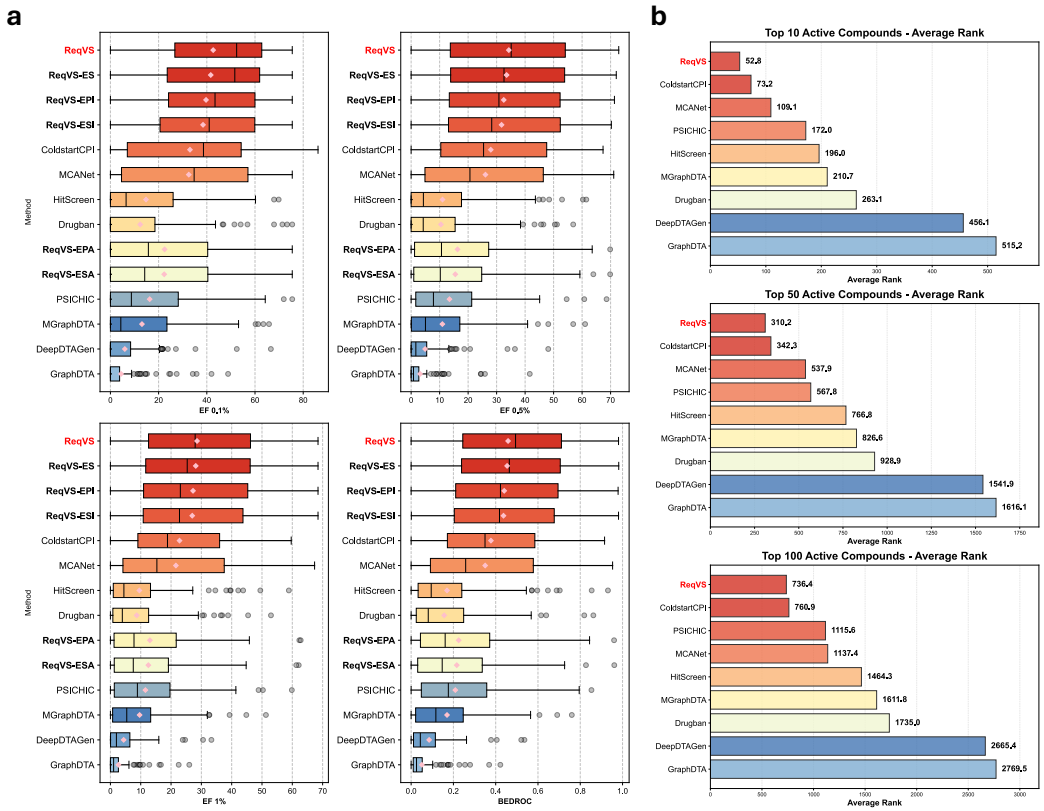


Figure 3: **Performance evaluation and ablation study of ReqVS on the DUD-E benchmark**. **a** Comparison of ReqVS against its framework variants and single constituent models across four key metrics: EF0.1%, EF0.5%, EF1%, and BEDROC. Pink diamonds denote the mean value for each metric; a median line is included in each box-whisker plot. **b** Comparison of the Average Rank for the top 10, 50, and 100 active compounds across all test targets. Lower values indicate a superior capability to prioritize active molecules at the absolute top of the screening list.

3.3 Ablation Studies and Ranking Analysis

In this section, the performance of ReqVS is first evaluated against its single-model components and framework variants on the DUD-E benchmark through an ablation study. The results for ReqVS, its variants, and individual models are presented in Figure 3. Different model configurations are evaluated, including DTI-only, DTA-only, and the full hybrid framework. The entropy-based variants for the DTI-only and DTA-only settings are denoted as ReqVS-EPI and ReqVS-EPA, respectively, while the full hybrid model corresponds to ReqVS. The ensemble weight configurations are summarized in Table 2. To further isolate the contribution of entropy-based reliability-aware screening, an ablation study without interaction entropy (i.e., with only interaction score) for screening is conducted, denoted as ReqVS-ESI (DTI-only), ReqVS-ESA (DTA-only), and ReqVS-ES (Hybrid).

In Figure 3a, ReqVS achieves the best performance across the mean and median values of all metrics (details are shown in Table 3). Specifically, in the critical early enrichment metric of EF0.1%, ReqVS (42.65, median 52.38) outperforms the best single model, ColdStartCPI (33.02, median 38.61), improving the mean value by 29.2% and the median value by 35.7%. This result suggests that ReqVS more effectively prioritizes active compounds toward the top of the ranking list. Notably, the single-group variants, ReqVS-EPA and ReqVS-EPI, both outperform their respective individual constituent models. Although the DTI-only setting yields stronger performance than the DTA-only setting, incorporating DTA models into the full framework generates a strong synergistic effect. For example, regarding EF0.1% and BEDROC, with the addition of DTA models, ReqVS shows a

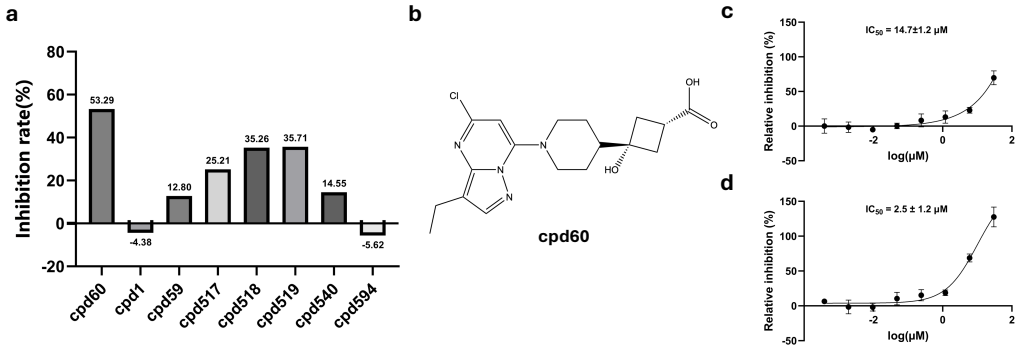


Figure 4: **Discovery of high affinity antagonist of HCAR1.** **a** Inhibition of forskolin-induced cAMP accumulation by eight candidate small molecules in HEK293 cells overexpressing HCAR1. **b** Chemical structure of cpd60. **c** Dose-response analysis of cpd60 using the GloSensor™ cAMP assay. IC_{50} values are presented as mean ± SEM from three technical replicates. **d** Dose-response analysis of cpd60 using the HTRF cAMP assay. IC_{50} values are presented as mean ± SEM from three technical replicates.

slight improvement in mean values (EF0.1% increased by 7.5%, BEDROC increased by 4.3%) and a significant improvement in median values (EF0.1% increased by 20.7%, BEDROC increased by 16.9%) compared with the DTI-only setting. Similar results are observed on the DEKOIS2.0 benchmark (Figure 6 and Table 4). This indicates that ReqVS effectively leverages the continuous affinity predictions provided by DTA models to further boost performance. Moreover, the entropy-based approach consistently outperforms ensemble-score-only variants, indicating that entropy-driven uncertainty estimation enhances the reliability of virtual screening decisions.

Second, an analysis of the screening ranks of ReqVS and different baseline models is presented. In practical drug discovery settings, the absolute ranking of active compounds directly influences screening efficiency and downstream experimental cost. Figure 3b reports the average rank of the top 10, 50, and 100 active molecules across all test targets for each method. ReqVS demonstrates a significant advantage across all cutoff thresholds. Specifically, for the top-10 active compounds, ReqVS achieves an average rank of 52.8, corresponding to an improvement of approximately 28% compared with the best single model, ColdStartCPI (73.2). Notably, ReqVS maintains robust ranking performance even when the majority of component models yield less favorable rankings. These results indicate that ReqVS can drastically narrow the scope of wet-lab screening, significantly reducing the costs associated with early-stage hit discovery.

Third, to further investigate the robustness of ReqVS, Figure 7 in Appendix visually depicts the distribution of EF0.1%, EF0.5%, and BEDROC performance across *all* DUD-E test targets. ReqVS consistently exhibits stronger and more stable performance across the majority of targets compared with individual baseline models. This target-wise analysis indicates that ReqVS does not rely solely on improvements in aggregated metrics, but maintains stable performance across diverse target classes. A similar trend is observed on the DEKOIS2.0 benchmark (see Figure 8), further supporting the generalization capability of the proposed framework.

3.4 Case study: Discovery of HCAR1 Antagonists Using ReqVS

To assess its practical applicability in drug discovery, ReqVS is applied to HCAR1, a Gi-coupled G protein–coupled receptor that recognizes lactate as its endogenous ligand Nezhady et al. (2023); Colucci et al. (2023). HCAR1 represents an emerging therapeutic target in cancer He et al. (2025a), but existing experimental data and structures predominantly concern agonists, with limited information on antagonism. These limitations make HCAR1 a challenging target for virtual screening (see Appendix D for additional background).

In this prospective screen, a six-model instantiation of ReqVS is deployed to screen a commercial chemical library from TiTan, comprising approximately 10 million compounds. Based on the

ReqVS scores, the top 1,000 molecules (the top 0.01% of the entire library) are selected. Expert-guided filtration is then applied to retain molecules containing both carboxyl and hydroxyl moieties. Only two candidates within the top 1,000 satisfy these constraints: the compounds originally ranked 552nd and 577th become the Top 1 and Top 2 candidates after filtration. The ranking within the library after this filtration is referred to as the filtered rank. Due to compound availability, only the Top 1 candidate, named cpd60, is subjected to experimental validation, together with seven manually selected compounds. The filtering rationale, candidate structures, and a retrospective audit using the eight-model configuration are provided in Appendix D.1.

The inhibitory effects of the eight candidates on HCAR1 are assessed in a forskolin-induced cAMP accumulation assay. Among the eight tested compounds, six exhibit inhibitory activity. Specifically, cpd60 achieves an inhibition rate of 53.29% (Figure 4a,b). Dose-response analysis using the GloSensor cAMP assay yields an IC_{50} of $14.7 \pm 1.2\,\mu$M (Figure 4c), while an HTRF-based assay yields an IC_{50} of $2.5 \pm 1.2\,\mu$M (Figure 4d). These results support further investigation of cpd60 as an HCAR1 antagonist. Detailed assay procedures are provided in Appendix E.

A retrospective analysis of the eight experimentally tested compounds reveals a negative correlation between the raw ReqVS ranks and experimental inhibition rates (Pearson $r = -0.774$, $p = 0.024$), with a similar trend for the filtered ranks (Appendix D.2). These results indicate that numerically lower ranks are associated with higher inhibitory activity within this tested set.

To investigate a possible structural basis for cpd60-mediated inhibition, molecular docking is combined with three independent 200 ns molecular dynamics (MD) simulations and MM/PBSA analysis. The predicted binding pose places cpd60 in an extracellular vestibule above the CHBA-binding site, spanning the entrance to the orthosteric pocket. The simulations support the stability of this pose, and MM/PBSA analysis yields an average estimated binding free energy of -25.15 kcal/mol. Compared with the CHBA-bound receptor, cpd60-bound HCAR1 exhibits distinct conformational rearrangements at the intracellular ends of TM5 and TM6. These results suggest that cpd60 may inhibit HCAR1 by obstructing agonist access to the orthosteric pocket and perturbing activation-associated conformational changes. Together, these findings support the practical utility of ReqVS-guided screening for identifying experimentally active candidates and provide cpd60 as a molecular starting point for HCAR1 antagonist development, with a testable hypothesis of inhibition through extracellular vestibular binding. Detailed methods and results are provided in Appendix F.

4 Discussion and Conclusion

Reliable prioritization of active compounds remains a major challenge in virtual screening, especially when only protein sequence information is available. Although recent SeqBVS methods have substantially expanded the searchable target space, their practical utility is often limited by unstable rankings and overconfident false positives. In this study, we addressed this limitation by introducing ReqVS, a reliability-aware SeqBVS framework that integrates heterogeneous DTI and DTA predictors through interaction uncertainty estimation and reliability-aware optimization. Our results suggest that this reliability-oriented design leads to meaningful practical gains. Across both DUD-E and DEKOIS2.0, ReqVS consistently improved early-recognition performance and reduced the absolute ranks of active compounds, indicating a stronger ability to place true hits at the very top of screening lists. Moreover, the successful identification of a novel HCAR1 antagonist highlights the practical utility of ReqVS as an early-stage screening tool. Future work will aim to expand ReqVS toward a more comprehensive drug discovery framework by incorporating additional predictive dimensions, such as ADMET properties. By unifying efficacy and developability-oriented models within a single framework, ReqVS has the potential to prioritize compounds that are not only active but also more likely to succeed in downstream development.

AI use statement

In this work, we used generative AI tools to assist with translation, implement methods through code, support qualitative or thematic data analysis, and interpret research results. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for mathematical proofs, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research method-

ology or experiments, or clean and reformat datasets. Additionally, we used generative AI tools for sourcing and searching for information, identifying relevant literature, and editing the manuscript to improve its language, clarity, and readability. We have reviewed all AI-assisted work. Specifically, we manually checked the literature identified with AI against the original publications, verified AI-assisted translations and language edits against the authors' intended meaning, executed and tested AI-assisted code, and independently checked the qualitative analyses and interpretations against the underlying data and experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

We have read and adhered to the ICLR Code of Ethics. Our study does not involve human subjects, sensitive personal data, or applications with direct negative societal consequences. We are committed to fairness, transparency, and responsible reporting of our results.

Reproducibility statement

To facilitate reproducibility, we document data preprocessing, model training, evaluation protocols, and experimental procedures in the Appendix. The benchmark datasets used in this study are publicly available. Our implementation is available at `https://anonymous.4open.science/r/ReqVS-Anonymize-69F2`.

References

Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.

Melissa F Adasme, Katja L Linnemann, Sarah Naomi Bolz, Florian Kaiser, Sebastian Salentin, V Joachim Haupt, and Michael Schroeder. Plip 2021: expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic acids research*, 49(W1):W530–W534, 2021.

Peizhen Bai, Filip Miljković, Bino John, and Haiping Lu. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence*, 5(2):126–136, 2023.

Matthias R Bauer, Tamer M Ibrahim, Simon M Vogel, and Frank M Boeckler. Evaluation and optimization of virtual screening workflows with dekois 2.0–a public library of challenging docking benchmark sets. *Journal of chemical information and modeling*, 53(6):1447–1462, 2013.

Jilong Bian, Xi Zhang, Xiying Zhang, Dali Xu, and Guohua Wang. Mcanet: shared-weight-based multiheadcrossattention network for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(2):bbad082, 2023.

Neera Borkakoti and Janet M Thornton. Alphafold2 protein structure prediction: Implications for drug discovery. *Current opinion in structural biology*, 78:102526, 2023.

Timothy P Brown and Vadivel Ganapathy. Lactate/gpr81 signaling and proton motive force in cancer: Role in angiogenesis, immune escape, nutrition, and warburg phenomenon. *Pharmacology & therapeutics*, 206:107451, 2020.

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.

Duanhua Cao, Geng Chen, Jiaxin Jiang, Jie Yu, Runze Zhang, Mingan Chen, Wei Zhang, Lifan Chen, Feisheng Zhong, Yingying Zhang, et al. Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nature Machine Intelligence*, 6(6):688–700, 2024.

Gabriella Cavallo, Pierangelo Metrangolo, Roberto Milani, Tullio Pilati, Arri Priimagi, Giuseppe Resnati, and Giancarlo Terraneo. The halogen bond. *Chemical reviews*, 116(4):2478–2601, 2016.

Michelangelo Certo, Alba Llibre, Wheeseong Lee, and Claudio Mauro. Understanding lactate sensing and signalling. *Trends in Endocrinology & Metabolism*, 33(10):722–735, 2022.

Ayan Chatterjee, Robin Walters, Zohair Shafi, Omair Shafi Ahmed, Michael Sebek, Deisy Gysi, Rose Yu, Tina Eliassi-Rad, Albert-László Barabási, and Giulia Menichetti. Improving the generalizability of protein-ligand binding predictions with ai-bind. *Nature communications*, 14(1): 1989, 2023.

Geng Chen, Jinbiao Liao, Yanzhen Yu, Kaixin Le, Hui Zhao, Yiyang Qin, Lvtao Cai, and Rong Sheng. Hitscreen: A sequence-based drug virtual screening approach using data augmentation and protein language models. *Journal of Chemical Information and Modeling*, 65(19):10152–10166, 2025a.

Lifan Chen, Zisheng Fan, Jie Chang, Ruirui Yang, Hui Hou, Hao Guo, Yinghui Zhang, Tianbiao Yang, Chenmao Zhou, Qibang Sui, et al. Sequence-based drug design as a concept in computational drug design. *Nature communications*, 14(1):4217, 2023.

Lin Chen, Vincent Blay, Pedro J Ballester, and Douglas R Houston. Scorch2: A generalized heterogeneous consensus model for high-enrichment interaction-based virtual screening. *Advanced Science*, 12(42):e08318, 2025b.

Anna Clara Machado Colucci, Isadora D'Ávila Tassinari, Eloísa da Silveira Loss, and Luciano Stürmer de Fraga. History and function of the lactate receptor gpr81/hcar1 in the brain: a putative therapeutic target for the treatment of cerebral ischemia. *Neuroscience*, 526:144–163, 2023.

Paolo Conflitti, Edward Lyman, Mark SP Sansom, Peter W Hildebrand, Hugo Gutiérrez-de Terán, Paolo Carloni, T Bertie Ansell, Shuguang Yuan, Patrick Barth, Anne S Robinson, et al. Functional dynamics of g protein-coupled receptors reveal new routes for drug discovery. *Nature Reviews Drug Discovery*, 24(4):251–275, 2025.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: new docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

Maksym Fedorchuk and Bart Lamiroy. Statistic metrics for evaluation of binary classifiers without ground-truth. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 1066–1071. IEEE, 2017.

Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

Jiacheng He, Xiaolei Chai, Qiansen Zhang, Yang Wang, Yijie Wang, Xinyu Yang, Jingbo Wu, Bo Feng, Jing Sun, Weiwei Rui, et al. The lactate receptor hcar1 drives the recruitment of immunosuppressive pmn-mdscs in colorectal cancer. *Nature Immunology*, 26(3):391–403, 2025a.

Xin-heng He, Jun-rui Li, Shi-yi Shen, and H Eric Xu. Alphafold3 versus experimental structures: assessment of the accuracy in ligand-bound g protein-coupled receptors. *Acta Pharmacologica Sinica*, 46(4):1111–1122, 2025b.

Scott A Hollingsworth, Brendan Kelly, Celine Valant, Jordan Arthur Michaelis, Olivia Mastromihalis, Geoff Thompson, AJ Venkatakrishnan, Samuel Hertig, Peter J Scammells, Patrick M Sexton, et al. Cryptic pocket formation underlies allosteric modulator selectivity at muscarinic gpcrs. *Nature Communications*, 10(1):3289, 2019.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. Charmm-gui: a web-based graphical user interface for charmm. *Journal of computational chemistry*, 29(11):1859–1865, 2008.

Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, 6(6):673–687, 2024.

Dima Kozakov, Laurie E Grove, David R Hall, Tanggis Bohnuud, Scott E Mottarella, Lingqi Luo, Bing Xia, Dmitri Beglov, and Sandor Vajda. The ftmap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature protocols*, 10(5):733–755, 2015.

Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

Alasdair TR Laurie and Richard M Jackson. Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.

Jumin Lee, Xi Cheng, Sunhwan Jo, Alexander D MacKerell, Jeffery B Klauda, and Wonpil Im. Charmm-gui input generator for namd, gromacs, amber, openmm, and charmm/openmm simulations using the charmm36 additive force field. *Biophysical journal*, 110(3):641a, 2016.

Yang Li, Guanyu Qiao, Keqi Wang, and Guohua Wang. Drug–target interaction predication via multi-channel graph neural networks. *Briefings in Bioinformatics*, 23(1):bbab346, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Jianing Lu, Xuben Hou, Cheng Wang, and Yingkai Zhang. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *Journal of chemical information and modeling*, 59(11):4540–4549, 2019.

Zhangli Lu, Guoqiang Song, Huimin Zhu, Chuqi Lei, Xinliang Sun, Kaili Wang, Libo Qin, Yafei Chen, Jing Tang, and Min Li. Dtiam: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nature Communications*, 16(1):2548, 2025.

Miles McGibbon, Sam Money-Kyrle, Vincent Blay, and Douglas R Houston. Scorch: Improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation. *Journal of Advanced Research*, 46:135–147, 2023.

Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023.

Bill R Miller III, T Dwight McGee Jr, Jason M Swails, Nadine Homeyer, Holger Gohlke, and Adrian E Roitberg. Mmpbsa. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation*, 8(9):3314–3321, 2012.

Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

Mohammad Ali Mohammad Nezhady, Monir Modaresinejad, Aliabbas Zia, and Sylvain Chemtob. Versatile lactate signaling via hcar1: a multifaceted gpcr involved in many biological processes. *American Journal of Physiology-Cell Physiology*, 2023.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

Xin Pan, Fang Ye, Peiruo Ning, Yiping Yu, Zhiyi Zhang, Jingxuan Wang, Geng Chen, Zhangsong Wu, Chen Qiu, Jiancheng Li, et al. Structures of g-protein coupled receptor hcar1 in complex with gi1 protein reveal the mechanistic basis for ligand recognition and agonist selectivity. *PLoS biology*, 23(4):e3003126, 2025.

Joel Ricci-Lopez, Sergio A Aguila, Michael K Gilson, and Carlos A Brizuela. Improving structure-based virtual screening with ensemble docking and machine learning. *Journal of chemical information and modeling*, 61(11):5362–5376, 2021.

Pir Masoom Shah, Huimin Zhu, Zhangli Lu, Kaili Wang, Jing Tang, and Min Li. Deepdtagen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nature Communications*, 16(1):5021, 2025.

Chenyu Tang, Mayank Prakash Pandey, Cheng Giuseppe Chen, Alberto Megías, François Dehez, and Christophe Chipot. Breaking timescales with generative sampling of conformational transitions. *Nature*, pp. 1–6, 2026.

Jérôme Tubiana, Dina Schneidman-Duhovny, and Haim J Wolfson. Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature methods*, 19(6):730–739, 2022.

Mario S Valdés-Tresanco, Mario E Valdés-Tresanco, Pedro A Valiente, and Ernesto Moreno. gmx_mmpbsa: a new tool to perform end-state free energy calculations with gromacs. *Journal of chemical theory and computation*, 17(10):6281–6291, 2021.

Fangping Wan, Yue Zhu, Hailin Hu, Antao Dai, Xiaoqing Cai, Ligong Chen, Haipeng Gong, Tian Xia, Dehua Yang, Ming-Wei Wang, et al. Deepcpi: a deep learning-based framework for large-scale in silico drug screening. *Genomics, proteomics & bioinformatics*, 17(5):478–495, 2019.

Baoli Wang, Chenglin Wang, Jingyang Chen, Danlin Liu, Changzhi Sun, Jie Zhang, Kai Zhang, and Honglin Li. Conditional diffusion with locality-aware modal alignment for generating diverse protein conformational ensembles. *Nature Machine Intelligence*, pp. 1–20, 2026.

Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry*, 38(3):169–177, 2017.

Rainer Wilcken, Markus O Zimmermann, Andreas Lange, Andreas C Joerger, and Frank M Boeckler. Principles and applications of halogen bonding in medicinal chemistry and chemical biology. *Journal of medicinal chemistry*, 56(4):1363–1388, 2013.

Maciej Wójcikowski, Pedro J Ballester, and Pawel Siedlecki. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 7(1):46710, 2017.

Zhenyu Yang, Xiaoxi Zeng, Yi Zhao, and Runsheng Chen. Alphafold2 and its applications in the fields of biology and medicine. *Signal Transduction and Targeted Therapy*, 8(1):115, 2023.

Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science*, 13(3): 816–833, 2022.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Shengyu Zhang, Donghui Huo, Robert I Horne, Yumeng Qi, Sebastian Pujalte Ojeda, Aixia Yan, and Michele Vendruscolo. Sequence-based virtual screening using transformers. *Nature Communications*, 16(1):6925, 2025.

Qichang Zhao, Haochen Zhao, Linyuan Guo, Kai Zheng, Yajie Li, Qiao Ling, Jing Tang, Yaohang Li, and Jianxin Wang. Coldstartcpi: Induced-fit theory-guided dti predictive model with improved generalization performance. *Nature Communications*, 16(1):6436, 2025.

Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4(14):15956–15965, 2019.

Mingyue Zheng, Jihui Zhao, Chen Cui, Zunyun Fu, Xutong Li, Xiaohong Liu, Xiaoyu Ding, Xiaoqin Tan, Fei Li, Xiaomin Luo, et al. Computational chemical biology and drug design: Facilitating protein structure, function, and modulation studies. *Medicinal research reviews*, 38(3):914–950, 2018.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*.

A Latent-Structure Formulation and Entropy Approximation

We propose the following proposition to derive the approximation error bound when using sequence-based entropy as a practical **surrogate** for interaction uncertainty, i.e.,

$$I(Y; S \mid s_t, s_c) \approx H(Y \mid s_t, s_c).$$

Proposition 3 (Surrogate Approximation Error)**.** *The approximation error of using sequence-based entropy as a surrogate for interaction uncertainty satisfies*

$$H(Y \mid s_t, s_c) - I(Y; S \mid s_t, s_c) = H(Y \mid S). \tag{9}$$

Therefore, the approximation error is bounded by

$$0 \leq H(Y \mid s_t, s_c) - I(Y; S \mid s_t, s_c) = H(Y \mid S) \leq H(Y). \tag{10}$$

Under the assumption that $Y \in \{0, 1\}$, the entropy of a binary random variable is bounded by

$$H(Y) \leq \log_2 2 = 1. \tag{11}$$

Therefore, the surrogate approximation error satisfies

$$0 \leq H(Y \mid s_t, s_c) - I(Y; S \mid s_t, s_c) \leq 1. \tag{12}$$

This result shows that the worst-case approximation error is bounded by one bit, which corresponds to the maximum uncertainty of a binary interaction outcome.

B Detailed Experimental Setup

B.1 Data Curation and Model Training

B.1.1 Data Sets Curation and Preprocessing

ChEMBL (version 36) Zdrazil et al. (2024) is selected as training dataset due to its comprehensive collection of protein-ligand interaction data. Before model training, any entries containing ligands or proteins present in the DUD-E and DEKOIS 2.0 benchmarks are excluded. This procedure removed 192,183 of 854,953 classification records (22.48%) and 86,386 of 439,980 regression

records (19.63%). Data quality is ensured by retaining protein-ligand pairs with exact relationships ("="), and nM units, while excluding molecules with > 70 heavy atoms or proteins exceeding 2,000 residues. Activity values are log-transformed ($Y_A = -\log_{10}\left(\frac{\text{value}}{10^9}\right)$). To address the distinct requirements of training sequence-interaction models, specific data cleaning protocols are implemented for classification and regression models, respectively. For the classification model, EC50, IC50, K_d, and K_i measurements are utilized; samples are binarized into positives ($Y_A > 7$, < 100 nM) and negatives ($Y_A < 5$, $> 10,000$ nM), with intermediate values discarded. Conversely, the regression model is restricted to K_i and K_d data to ensure consistency, using Y_A directly as the label. Duplicate entries are curated for consistency: classification pairs with conflicting labels are discarded, whereas regression pairs are averaged only if the value deviation is within 1.0. The final curated datasets comprise 660k entries for classification and 353k entries for regression.

Table 1: Statistics of the curated ChEMBL v36 activity datasets used in this study.

Dataset version	Task	# Activities	# Targets	# Molecules	# Mixed-label molecules
Unaugmented	Classification	662,770	6,608	514,612	12,140
Unaugmented	Regression	353,594	4,975	196,623	/
Augmented	Classification	1,170,590	6,608	514,612	403,623
Augmented	Regression	521,566	4,975	196,623	/

Note: Mixed-label molecules refer to the number of unique molecules that are associated with both positive and negative labels across different targets or assays.

B.1.2 Sequence-based Interaction Model Selection

According to previously related works, existing sequence models can be categorized into 2 groups: Drug-Target Interaction (DTI) and Drug-Target Affinity (DTA) models, optimize different objectives—answering the fundamental questions of "Does it bind?" and "How strongly does it bind?", and their outputs reside in distinct spaces: DTI models learn the binding probability ($\hat{y} \in [0, 1]$), whereas DTA models learn the binding affinity ($\hat{y} \in \mathbb{R}$).

Eight heterogeneous models are selected, comprising four DTI models and four DTA models. These models are selected to provide complementary views of protein-ligand interactions by employing diverse architectural designs and feature-extraction strategies. To verify the concept of reliable virtual screening, the following models are included:

- *DTI Models: MCANet Bian et al. (2023), DrugBAN Bai et al. (2023), HitScreen Chen et al. (2025a) and ColdStartCPI Zhao et al. (2025).* MCANet and DrugBAN place greater emphasis on interaction modeling in latent space. MCANet adopts parallel convolutional encoders for drug and protein sequences and introduces a multi-head cross-attention mechanism to capture cross-modal interaction features, whereas DrugBAN combines molecular graph encoding with protein sequence encoding and employs a bilinear attention module to explicitly model local pairwise interactions between compounds and targets.
- *DTA Models: GraphDTA Nguyen et al. (2021), MGraphDTA Yang et al. (2022), DeepDTAGen Shah et al. (2025), and PSICHIC Koh et al. (2024).* GraphDTA and MGraphDTA encode compounds as molecular graphs and proteins as 1D sequences, combining graph neural networks for ligands with convolutional encoders for proteins. While GraphDTA provides a graph-based baseline for affinity prediction, MGraphDTA further enhances ligand representation through a multiscale graph neural network, enabling more effective extraction of both local chemical patterns and global molecular topology

Besides, these models further enrich the ensemble by incorporating pretrained biological knowledge, physicochemical priors, or auxiliary learning objectives, which can be leveraged to model latent structure information in the framework. DeepDTAGen jointly learns affinity prediction and target-aware molecule generation, thereby introducing a synthesizable and task-informed latent space for ligands. PSICHIC, ColdStartCPI, and HitScreen all leverage pretrained protein language models to extract protein features that contain evolutionary information and structure-related semantics learned

from large-scale protein corpora. Among them, PSICHIC integrates physicochemical embeddings and clustering-based partitioning of ligand functional groups and protein regions to characterize binding patterns under explicit physicochemical constraints. ColdStartCPI models protein-ligand interactions through a transformer-based architecture inspired by the induced-fit mechanism, allowing the model to capture interaction patterns under molecular flexibility. HitScreen integrates multiple pretrained models, including protein language models such as ESM-2 Lin et al. (2023) and Ankh Elnaggar et al. (2023), as well as the 3D molecular foundation model Uni-Mol Zhou et al., thereby introducing richer protein semantics and ligand spatial information into sequence-based screening. By integrating these heterogeneous models, the ensemble captures complementary information ranging from sequence semantics and molecular topology to local cross-modal interactions, physicochemical constraints, and structural priors, which together provide a more robust basis for uncertainty estimation and reliability-aware ranking.

B.1.3 Training Strategy for Sequence-based Models

A data augmentation strategy inspired by the related study Chen et al. (2025a) is implemented for both classification and regression tasks. Proteins in the curated dataset are first clustered using CD-HIT Fu et al. (2012) at a sequence similarity threshold of 70%, and 2048-bit Morgan fingerprints are computed for all compounds using RDKit Landrum et al. (2013). For classification models, augmentation is performed at the protein-cluster level. For each cluster, all positive protein-ligand pairs associated with that cluster are treated as active samples, and candidate ligands are drawn from positive samples belonging to other clusters. A candidate ligand is accepted only if its Tanimoto similarity to all positive ligands in the current cluster is below 0.3. The accepted ligand is then re-paired with a target protein from the current cluster to form a synthetic negative sample. To increase the proportion of molecular scaffolds appearing in both positive and negative contexts while avoiding excessive duplication, each externally sourced positive ligand is allowed to be used as a negative sample only once. For regression models, the same cluster-level cross-cluster sampling strategy is applied, with the source pool restricted to high-affinity samples ($Y_A > 7$, < 100 nM). Synthetic negative pairs are constructed by re-pairing target proteins with ligands sampled from high-affinity samples in other clusters, again under the constraint that the sampled ligand must have Tanimoto similarity below 0.3 to all high-affinity ligands in the current cluster. The affinity labels of these synthetic negatives are then generated by subtracting 4 from the original Y_A value ($Y_{\text{new}} = Y_{\text{original}} - 4$), thereby shifting strong binders into the low-affinity regime and providing explicit supervision on non-binders while preserving a realistic affinity distribution. Following augmentation, 403,623 of the 514,612 unique molecules (78.43%) in the classification data are assigned both positive and negative labels across different target contexts, compared with 12,140 (2.36%) before augmentation (Table 1). Overall, cross-cluster re-pairing diversifies target-dependent labels in both tasks and helps reduce ligand-driven model bias by weakening the association between ligand identity and a fixed activity class or affinity range.

B.2 Design of Mapping Function $\sigma(\cdot)$.

To map heterogeneous outputs of selected models into probability space, a mapping function $\sigma(\cdot)$ is designed as follows:

$$\sigma(\hat{y}) := \begin{cases} \hat{y} & \text{for DTI models} \\ \frac{\hat{y} - \hat{y}_{\text{min}}}{\hat{y}_{\text{max}} - \hat{y}_{\text{min}}} & \text{for DTA models} \end{cases} \tag{13}$$

where $\hat{y}_{\text{min}}$ and $\hat{y}_{\text{max}}$ are the minimum and maximum values of DTA predictions on the chemical library $\mathcal{C}$, respectively. For DTI models, the probabilistic outputs are naturally bounded and are strictly preserved. For DTA models, the normalization relies on the distribution of affinity scores within the current screening run. This normalization ensures that all model outputs are on a comparable scale, facilitating subsequent consensus calculations. Additionally, as a linear transformation, it preserves not only the ranking order but also the relative magnitude intervals between candidates, strictly maintaining the underlying distribution structure of the predictions.

B.3 TECHNICAL SCREENING PIPELINE

Given the candidate compound library $\mathcal{C}$, the screening pipeline for optimizing the reliability-aware objective in Equation 2 is described as follows.

- *Step-1: Sequence-based model inference.* For each molecule in the chemical library, i.e., $c \in \mathcal{C}$,
  (a) The inference results from heterogeneous sequence-based models are obtained as $\{\hat{y}^j_{s_t,s_c}\}_{j=1}^M$, where j denotes the index of each sequence-based model described in Section B.1.2.
  (b) The interaction probability $P_{\theta^j}(y = 1 \mid s_t, s_c)$ is computed using $\sigma(\hat{y}^j_{s_t,s_c})$, which is denoted as $P^j_{s_t,s_c}$.
  (c) The entropy H_{θ^j} for the j-th sequence-based model is then calculated according to Equation 7.
- *Step-2: Uncertainty-based model evaluation.* Uncertainty-based evaluation metrics are further proposed to assess the performance of each sequence-based model on a new target t when no test data are available. The main idea is to replace the unavailable ground truth with a consensus metric derived from the outputs of heterogeneous sequence-based models, following related studies Fedorchuk & Lamiroy (2017). The approximated ground truth is defined as $\bar{P}_{s_t,s_c} = \frac{1}{M}\sum_{j=1}^M P^j_{s_t,s_c}$, and the precision, recall, and F1 score for the j-th sequence-based model can be approximated as:

$$\text{Precision} = \frac{\sum_{c\in\mathcal{C}} \bar{P}_{s_t,s_c} \cdot P^j_{s_t,s_c}}{\sum_{c\in\mathcal{C}} P^j_{s_t,s_c}}, \tag{14}$$

$$\text{Recall} = \frac{\sum_{c\in\mathcal{C}} \bar{P}_{s_t,s_c} \cdot P^j_{s_t,s_c}}{\sum_{c\in\mathcal{C}} \bar{P}_{s_t,s_c}}, \tag{15}$$

$$\text{F1} = \frac{2 \cdot \text{Precision}^j \cdot \text{Recall}^j}{\text{Precision}^j + \text{Recall}^j}. \tag{16}$$

  These approximated metrics are subsequently used as weights in the ensemble process.
- *Step-3: Interaction score ensemble.* According to the objective in Equation 2, for each molecule $c \in \mathcal{C}$, the interaction score $\hat{Y}_{s_t,s_c}$ is obtained by ensembling the probabilistic predictions of sequence-based models as $\hat{Y}_{s_t,s_c} := \sum_{j=1}^M \omega^j P^j_{s_t,s_c}$, which helps alleviate the epistemic uncertainty of deep models. Here, ω^j denotes the aggregation weight of the j-th sequence-based model. To evaluate the capabilities of heterogeneous sequence-based models, three ensemble strategies are designed for DTI-only, DTA-only, and hybrid model settings. The aggregation weights ω^j are configured differently under each strategy, as summarized in Table 2. Notably, different ensemble weight configurations are used for hybrid and single-model settings because different model categories have different capabilities in modeling protein–ligand interactions. For example, regression tasks in DTA are typically more difficult than classification tasks in DTI; therefore, using uniform ensemble weights without considering model performance may introduce bias into the ensemble predictions. For single-category ensembles, all models in the corresponding category share the same weight. For the hybrid ensemble, DTI models and DTA models share separate category-level weights, which are applied consistently to both interaction scores and entropies.
- *Step-4: Interaction entropy ensemble.* For the interaction entropy term in Equation 2, the entropy $H_{\theta^j}(Y \mid s_t, s_c)$ is similarly aggregated across models as $H_\theta(Y \mid s_t, s_c) := \sum_{j=1}^M \omega^j H_{\theta^j}(Y \mid s_t, s_c)$. The same ensemble strategies used in Step-3 are applied here, and the weights are shared with the score ensemble.
- *Step-5: Reliability-aware screening.* The final screening score for each molecule $c \in \mathcal{C}$ is calculated as $\hat{Y}_{s_t,s_c} - \lambda H_\theta(Y \mid s_t, s_c)$, where the interaction score and entropy are obtained from Step-3 and Step-4, respectively. Here, $\lambda = 0.2$ is used for all ensemble strategies and screening tasks based on empirical knowledge. All molecules in the chemical library $\mathcal{C}$ are then ranked according to the final scores in descending order. Finally, given a candidate list length K, the top-K molecules are selected as the virtual screening results.

Table 2: ReqVS and its ablation variants under different ensemble strategies.

Name	Model Types	Ensemble Weights ω^j
ReqVS	Hybrid	$\frac{1}{M_{\mathrm{DTA}}^2}\sum_{j=1}^{M_{\mathrm{DTA}}}\mathrm{F1}_j$, $\frac{1}{M_{\mathrm{DTI}}^2}\sum_{j=1}^{M_{\mathrm{DTI}}}\mathrm{F1}_j$ for DTA and DTI model categories, respectively
ReqVS-EPA	DTA only	$\frac{1}{M_{\mathrm{DTA}}}$
ReqVS-EPI	DTI only	$\frac{1}{M_{\mathrm{DTI}}}$

B.4 Evaluation Metrics

Virtual screening performance is assessed using standard metrics that emphasize early enrichment, a capability critical for practical drug discovery applications. Enrichment Factors (EF) at multiple thresholds ($\mathrm{EF}_{0.1\%}$, $\mathrm{EF}_{0.5\%}$, $\mathrm{EF}_{1\%}$, $\mathrm{EF}_{5\%}$) are employed to measure the concentration of active compounds in the top-ranked fraction of the screening library. The EF at a specific threshold γ is defined as:

$$\mathrm{EF}_{\gamma} = \frac{N_{\text{hits}}^{(\gamma)}/N_{\text{sampled}}^{(\gamma)}}{N_{\text{actives}}/N_{\text{total}}} \tag{17}$$

where $N_{\text{hits}}^{(\gamma)}$ denotes the number of active compounds identified within the top γ subset (with size $N_{\text{sampled}}^{(\gamma)}$), while N_{actives} and N_{total} represent the total number of active compounds and the total size of the screening library, respectively.

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUROC) is reported for global ranking assessment, together with the Boltzmann-Enhanced Discrimination of ROC (BEDROC) with $\alpha = 80.5$ to emphasize early enrichment. For comprehensive evaluation, the Average Rank of the top N active compounds (where $N \in \{10, 50, 100\}$) is also computed to directly quantify practical screening efficiency.

C Additional and detailed Benchmark Results

C.1 Comparisons with Structure-Based Methods

We further evaluate ReqVS on the predefined subsets of DEKOIS2.0 and DUD-E, containing 11 and 12 targets, respectively (Figure 5). On the DEKOIS2.0 subset, ReqVS achieves the highest median performance across all evaluated metrics (Figure 5a). On the DUD-E subset, ReqVS ranks second to Glide SP in EF0.5% and EF1% despite using sequence inputs alone (Figure 5b). These results further support the competitive early-recognition performance of ReqVS across the evaluated benchmark subsets.

C.2 Constituent Models and Ensemble Variants

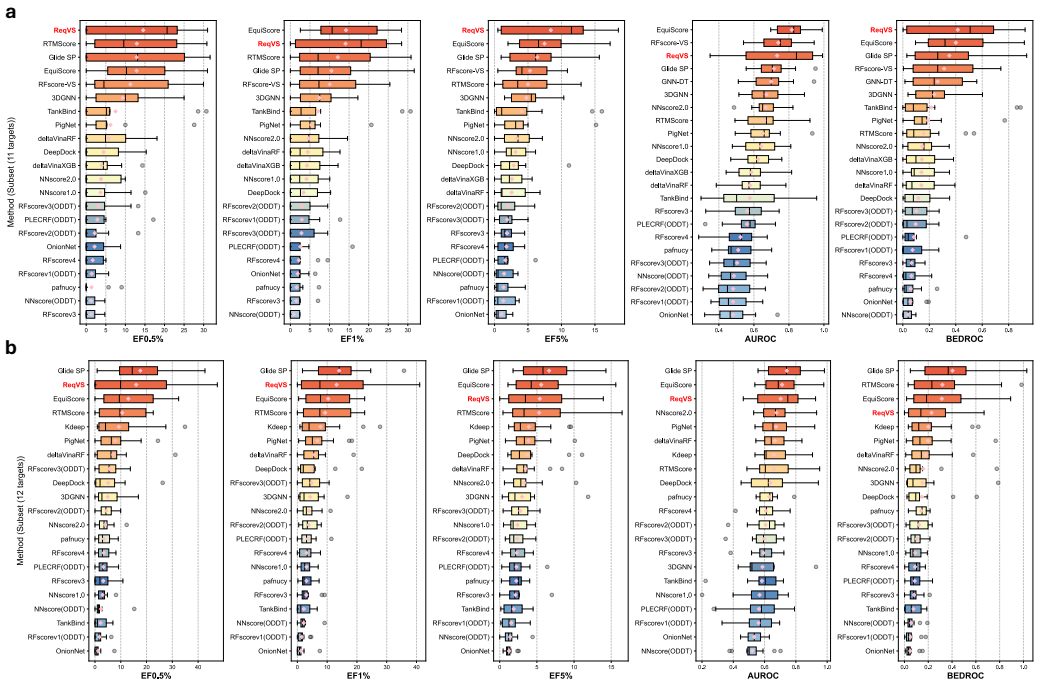


Figure 5: **Methods evaluation on the subsets of the DUD-E and DEKOIS2.0 benchmarks.** Pink diamonds denote the mean value for each metric; a median line is included in each box-whisker plot. **a** Model performance on the subset of DEKOIS2.0 (data point $n = 11$) evaluated by EF0.5%, EF1%, EF5%, AUROC, and BEDROC(α=80.5). **b** Model performance on the subset of DUD-E (data point $n = 12$) evaluated by the same five metrics.

Table 3: Performance comparison on the benchmark DUD-E. Values are reported as mean/median across targets. (All models are retrained from scratch)

Method	EF0.1%		EF0.5%		EF1%		EF5%		AUROC		BEDROC	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
GraphDTA Nguyen et al. (2021)	4.42	0.00	3.15	0.84	2.64	1.06	1.86	1.23	0.537	0.627	0.051	0.026
MGraphDTA Yang et al. (2022)	13.09	4.32	10.96	5.03	9.63	5.39	5.12	3.68	0.672	0.671	0.170	0.117
PSICHIC Koh et al. (2024)	16.29	8.70	13.45	7.83	11.54	8.90	6.44	5.89	0.758	0.795	0.209	0.176
DeepDTAGen Shah et al. (2025)	5.98	0.00	4.91	1.66	4.37	1.98	3.05	2.00	0.590	0.605	0.085	0.044
DrugBAN Bai et al. (2023)	12.37	0.00	10.45	4.25	8.68	3.93	4.61	3.49	0.682	0.723	0.157	0.081
MCANet Bian et al. (2023)	32.52	34.74	26.08	20.64	21.60	15.33	8.33	7.35	0.742	0.782	0.350	0.258
HitScreen Chen et al. (2025a)	14.77	6.51	11.02	4.30	9.55	4.45	4.95	3.62	0.708	0.722	0.170	0.095
ColdstartCPI Zhao et al. (2025)	33.02	38.61	27.98	25.42	22.79	18.61	9.20	9.27	0.813	0.865	0.377	0.350
ReqVS-ESA	22.33	14.23	15.50	10.23	12.52	7.54	5.92	5.18	0.724	0.750	0.216	0.147
ReqVS-EPA	22.48	15.75	16.28	10.65	13.02	7.78	6.16	5.47	0.727	0.754	0.225	0.161
ReqVS-ESI	38.43	41.00	31.74	28.28	26.99	22.79	10.19	**11.58**	0.806	0.866	0.437	0.418
ReqVS-EPI	39.68	43.41	32.51	30.81	27.25	23.07	10.04	11.49	0.798	0.857	0.440	0.423
ReqVS-ES	41.64	51.62	33.51	32.59	28.16	25.33	**10.43**	11.40	**0.821**	0.866	0.454	0.465
ReqVS	**42.65**	**52.38**	**34.29**	**35.12**	**28.58**	**27.95**	10.40	11.44	0.820	**0.871**	**0.459**	**0.494**

Note: Bold values indicate the best performance in each metric.

Table 4: Performance comparison on the benchmark DEKOIS2.0. Values are reported as mean/median across targets. EF0.1% is not reported because the limited number of compounds per target in DEKOIS2.0 makes this metric statistically meaningless. (All models are retrained from scratch)

Method	EF0.5%		EF1%		EF5%		AUROC		BEDROC	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
GraphDTA Nguyen et al. (2021)	2.68	0.00	2.26	0.00	1.67	1.50	0.558	0.558	0.069	0.040
MGraphDTA Yang et al. (2022)	7.72	5.17	7.40	5.17	4.50	4.00	0.699	0.735	0.215	0.154
PSICHIC Koh et al. (2024)	10.01	5.17	9.25	7.75	5.56	5.00	0.750	0.805	0.267	0.237
DeepDTAGen Shah et al. (2025)	3.32	0.00	3.25	2.58	2.22	1.50	0.560	0.559	0.099	0.066
DrugBAN Bai et al. (2023)	7.97	5.17	7.24	5.16	4.29	3.50	0.678	0.710	0.206	0.110
MCANet Bian et al. (2023)	15.95	15.50	14.64	15.50	8.27	8.00	0.763	0.851	0.421	0.414
HitScreen Chen et al. (2025a)	6.57	5.17	6.25	5.17	4.46	3.50	0.710	0.744	0.192	0.141
ColdstartCPI Zhao et al. (2025)	16.90	15.50	15.72	15.50	8.53	9.00	0.814	0.880	0.442	0.448
ReqVS-ESA	10.84	10.33	9.41	7.75	5.49	5.50	0.735	0.784	0.271	0.267
ReqVS-EPA	11.16	10.33	9.82	10.33	5.57	5.00	0.736	0.788	0.280	0.266
ReqVS-ESI	18.37	**20.67**	17.51	20.67	10.01	10.50	0.817	0.891	0.509	0.586
ReqVS-EPI	18.56	**20.67**	17.64	20.67	9.93	10.50	0.814	0.886	0.513	0.586
ReqVS-ES	18.24	**20.67**	18.12	20.67	10.24	11.00	**0.828**	**0.890**	0.527	0.603
ReqVS	**18.69**	**20.67**	**18.69**	**23.25**	**10.27**	**11.50**	0.825	0.888	**0.533**	**0.612**

Note: Bold values indicate the best performance in each metric.

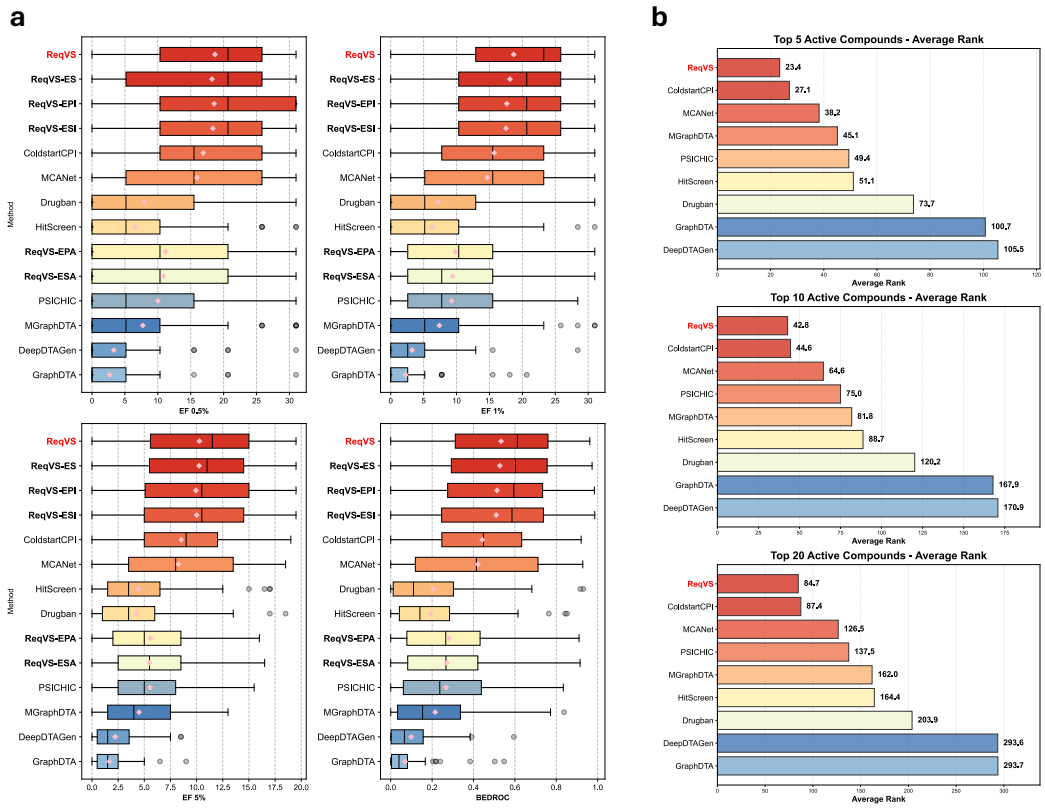


Figure 6: **Performance evaluation and ablation study of ReqVS on the DEKOIS2.0 benchmark.a** Comparison of ReqVS against its framework variants and single constituent models across four key metrics: EF0.5%, EF1%, EF5%, and BEDROC. EF0.1% is not reported because the limited number of compounds per target in DEKOIS2.0 makes this metric statistically meaningless. Pink diamonds denote the mean value for each metric; a median line is included in each box-whisker plot. **b** Comparison of the Average Rank for the top 5, 10, and 20 active compounds across all test targets. Lower values indicate a superior capability to prioritize active molecules at the absolute top of the screening list.

C.3 TARGET-WISE RESULTS

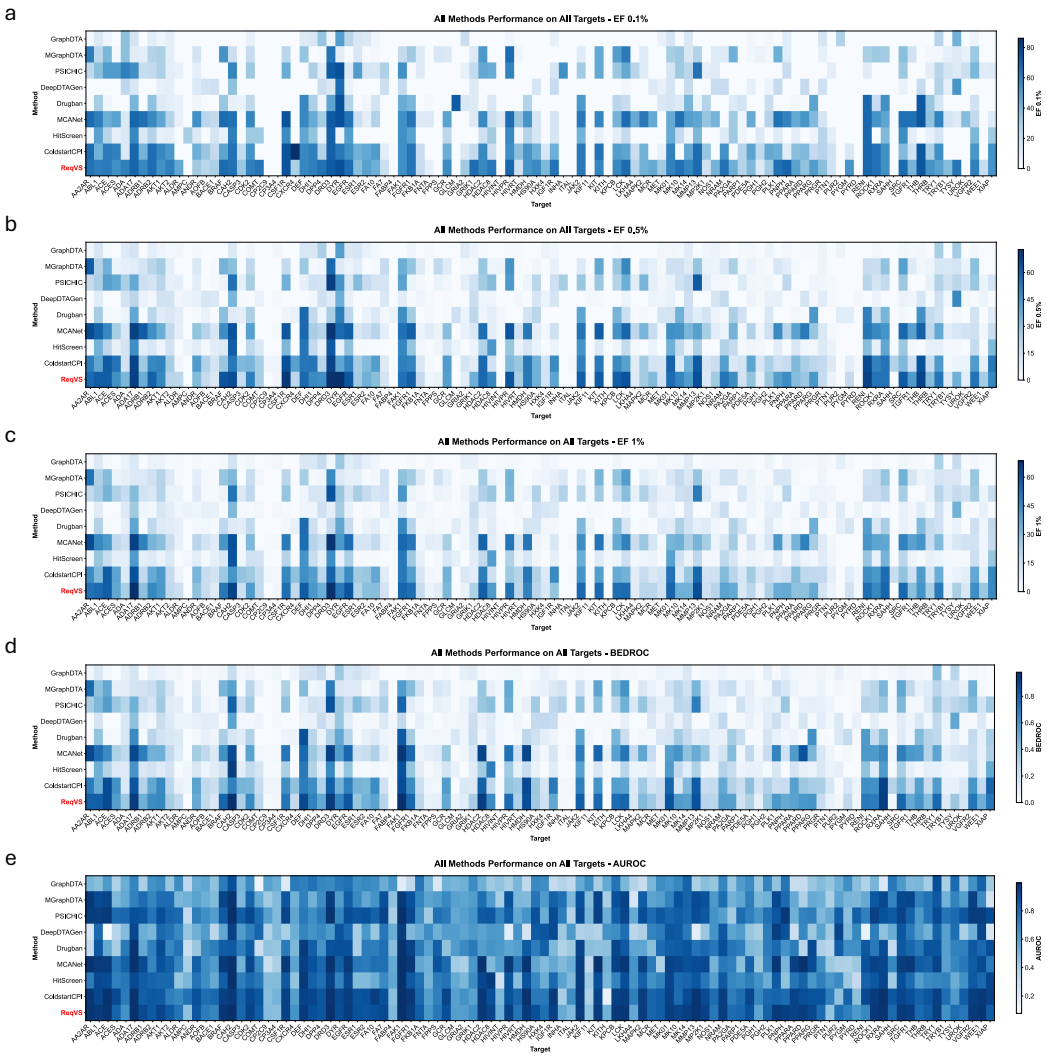


Figure 7: **Target-wise performance heatmaps of ReqVS and 8 contained methods across all targets in the DUD-E dataset.** The heatmaps across targets in terms of **a** EF0.1%, **b** EF0.5%, **c** EF1%, and **d** AUROC and **e** BEDROC, respectively.

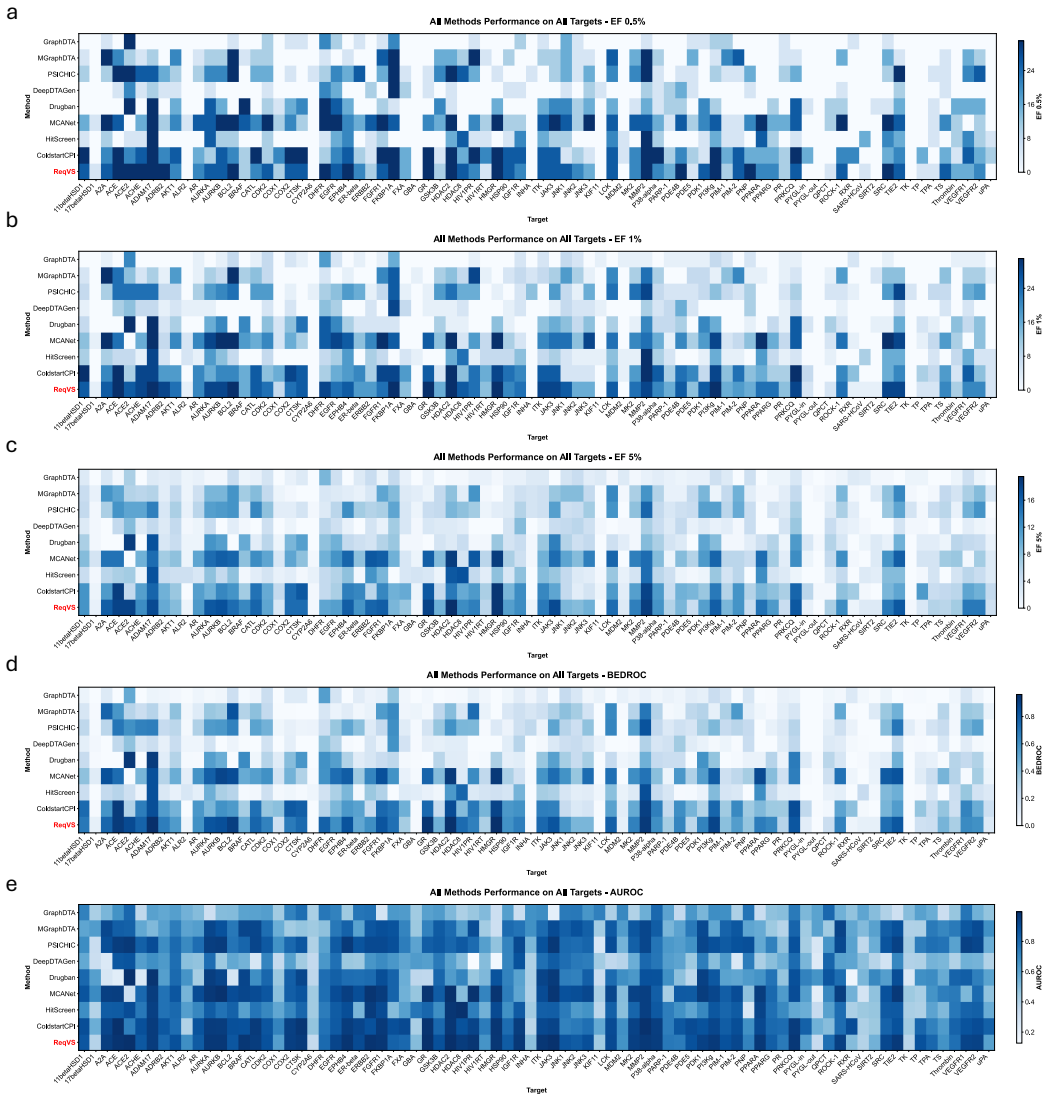


Figure 8: **Target-wise performance heatmaps of ReqVS and 8 contained methods across all targets in the DEKOIS2.0 dataset.** The heatmaps across targets in terms of **a** EF0.5% , **b** EF1% , **c** EF5% , **d** BEDROC, and **e** AUROC, respectively.

D ADDITIONAL HCAR1 SCREENING DETAILS

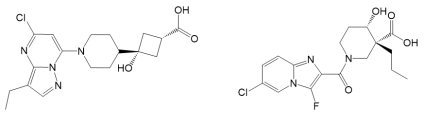
Recent studies indicate that aberrant lactate accumulation in the tumor microenvironment is not only a metabolic hallmark but can also modulate tumor-associated immune responses through HCAR1-mediated signaling, contributing to the establishment of an immunosuppressive milieu Brown & Ganapathy (2020); Certo et al. (2022). In colorectal cancer, HCAR1 signaling in tumor cells has been shown to induce chemokine expression, including CCL2 and CCL7, which promote the recruitment of immunosuppressive $CCR2^{+}$ polymorphonuclear myeloid-derived suppressor cells (PMN-MDSCs) and attenuate antitumor immune responses He et al. (2025a). Additionally, structure prediction methods such as AlphaFold remain limited in accurately modeling GPCR–ligand complexes due to the intrinsic conformational flexibility of GPCRs He et al. (2025b).

D.1 CANDIDATE STRUCTURES AND RANKS

HCAR1 is known to activate signaling pathways by recognizing ligands such as lactate and CHBA. Crystallographic analysis of HCAR1 with CHBA reveals that the carboxyl and hydroxyl groups are important for ligand binding and receptor activation. Specifically, the carboxyl group establishes salt bridges and hydrogen bonds with residues R99 and Y268, while the hydroxyl group engages in additional hydrogen bonds with E166 and R71, thereby stabilizing the ligand within the orthosteric binding pocket (OBP) Pan et al. (2025). Based on this structural insight, molecules containing both carboxyl and hydroxyl moieties are selected.

Following this expert-guided filtration, only two candidates within the top 1,000 satisfy the constraints (Figure 9). These two compounds, originally ranked 552nd and 577th based on the raw ReqVS ranking, emerge as the final Top 1 and Top 2 candidates. Due to logistical constraints regarding compound synthesis and procurement cycles, only the Top 1 candidate, cpd60, is subjected to experimental validation. Seven manually selected compounds are also tested. The chemical structures and raw and filtered ranks of all eight tested compounds are shown in Figure 10.

This prospective screen represents an early-stage deployment of ReqVS using a six-model instantiation. A retrospective audit using the eight-model configuration shows that cpd60 remains within the top 34 by filtered rank.



Candidate 1/cpd60
Rank_in_filtered : 1
Rank : 552

Candidate 2
Rank_in_filtered : 2
Rank : 577

Figure 9: Chemical structures and virtual screening rankings of the top 2 candidate compounds.

D.2 RANK CORRELATION ANALYSIS

A retrospective analysis is conducted to examine the relationship between the computational rankings and experimental inhibition rates of the eight tested compounds.

Based on the filtered rankings, the most active compound, cpd60, is ranked 1st with an inhibition rate of 53.29% (raw library rank: 552nd). In order of filtered rank, this is followed by cpd540 (filtered rank: 1,957th; raw rank: 473,233rd; inhibition rate: 14.55%), cpd519 (filtered rank: 6,029th; raw rank: 1,881,221st; inhibition rate: 35.71%), cpd517 (filtered rank: 6,932nd; raw rank: 2,229,469th; inhibition rate: 25.21%), cpd59 (filtered rank: 12,444th; raw rank: 4,416,547th; inhibition rate: 12.80%), and cpd518 (filtered rank: 12,870th; raw rank: 4,584,575th; inhibition rate: 35.26%). The two inactive compounds, cpd1 and cpd594, exhibit negative inhibition rates of -4.38% and -5.62%, corresponding to filtered ranks of 19,819th and 22,222nd (raw ranks: 7,311,514th and 8,226,619th), respectively.

Statistical analysis reveals a negative correlation between the raw ReqVS ranks and experimental inhibition rates (Pearson $r = -0.774$, $p = 0.024$; Figure 11a). This correlation is maintained when

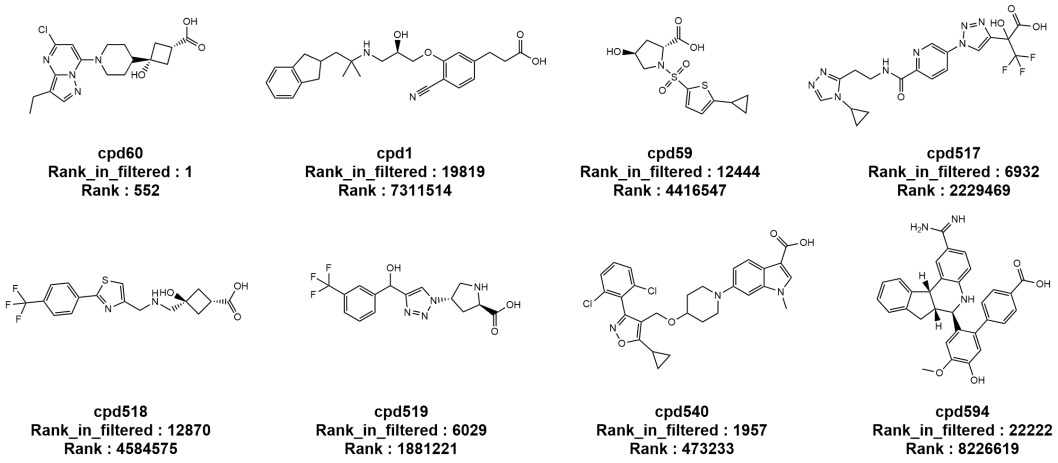


Figure 10: Chemical structures and virtual screening rankings of the eight candidate compounds.

evaluating the filtered ranks (Pearson $r = -0.779$, $p = 0.023$; Figure 11b). These associations are observed within the eight experimentally tested compounds.

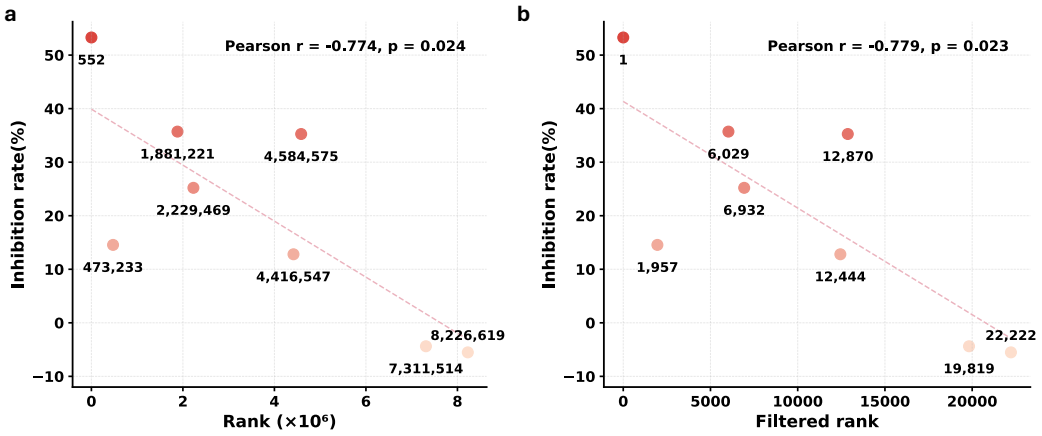


Figure 11: **Correlation between ReqVS predicted ranks and inhibition rate of HCAR1. a** Scatter plot illustrating the relationship between the raw virtual screening rank ($\times 10^6$) of the eight candidate compounds and their experimental inhibition (%). A significant negative correlation is observed, with a Pearson $r = -0.774$, $p = 0.024$. **b** Scatter plot displaying the relationship between the candidate compounds' inhibition rate and their filtered rank after applying expert-guided structural constraints. The correlation is maintained with Pearson $r = -0.779$, $p = 0.023$.

E cAMP Inhibition Assay

The inhibitory effects of test compounds on forskolin-induced cAMP accumulation were evaluated in HEK293 cells overexpressing human HCAR1 using both the GloSensor™ cAMP assay (Promega) and the cAMP–Gi HTRF assay (Cisbio). HEK293 cells (*Homo sapiens*; ATCC CRL-1573; RRID:CVCL_0045; female fetal origin) were obtained from the American Type Culture Collection (ATCC). The cells were confirmed to be free of mycoplasma contamination. For the GloSensor™ cAMP assay, HEK293 cells were cultured in DMEM supplemented with 10% fetal bovine serum and seeded into white 96-well plates, followed by overnight incubation at 37 °C with

5% CO_2. Cells were transiently co-transfected with human HCAR1 and the pGloSensor-22F cAMP reporter plasmids using FuGENE® HD and incubated for 24 h. The culture medium was then removed and replaced with an equilibration solution containing the GloSensor™ reagent, and plates were incubated at room temperature for 2 h to stabilize the luminescence baseline. Test compounds were prepared as DMSO stocks and diluted to eight final concentrations spanning 0.000384–30 μM with 5-fold spacing, and the final DMSO concentration was matched across wells. After incubation with compounds for 10 min at room temperature, forskolin was added, and plates were incubated for 20 min at room temperature. Luminescence was recorded using a microplate reader, and signals were normalized to the pre-compound baseline for each well. For the cAMP–Gi HTRF assay, HEK293 cells were dispensed into white 384-well plates at a density of 5×10^5 cells/mL with 5 μL per well. Test compounds were prepared as DMSO stocks and diluted to eight final concentrations spanning 0.000384–30 μM with 5-fold spacing, and the final DMSO concentration was matched across wells. Cells and compounds were incubated at 37 °C for 30 min, followed by addition of 1 μL forskolin per well together with IBMX to a final concentration of 0.5 mM. After incubation, 5 μL cAMP-Cryptate working solution per well and 5 μL Anti-cAMP-d2 working solution per well were sequentially added, and both reagents were prepared in Lysis & Detection Buffer. Plates were incubated for 1 h at room temperature in the dark, and HTRF signals were measured on an HTRF-compatible reader by recording emissions at 665 nm and 620 nm. The 665/620 nm fluorescence ratio was used for downstream analysis and comparison across concentrations. For both assays, concentration–response curves were fitted by nonlinear regression using a four-parameter logistic model in GraphPad Prism 10.0 to obtain IC_{50} values. IC_{50} values are presented as mean ± SEM from three technical replicates.

F Molecular Docking and Dynamics

F.1 Structural Insights into cpd60 Binding to HCAR1

Superposition of the cpd60–HCAR1 complex with the CHBA-bound HCAR1 crystal structure reveals that cpd60 adopts a binding mode distinct from that of the synthetic agonist CHBA (Figure 12a). Rather than deeply inserting into the buried OBP, cpd60 is positioned in an extracellular vestibule directly above the CHBA-binding site, where it spans the pocket entrance and forms a cap-like pose over the orthosteric cavity (Figure 12a). Notably, cpd60 binds to the neighborhood of the N-terminus and ECL2 at the extracellular entrance of HCAR1, instead of the canonical orthosteric pocket. Occupancy of this upper vestibule suggests that cpd60 may attenuate receptor activation by blocking agonist entering into the buried binding pocket Hollingsworth et al. (2019). Interaction analysis indicates that cpd60 is stabilized by a contact network centered on the N-terminus/ECL2 interface at the extracellular entrance of HCAR1 (Figure 12b). In particular, cpd60 forms hydrogen bonds with Cys7 and Arg8, and salt bridges with His257$^{7.31}$ and His261$^{7.35}$, remaining in close contact with the E166$^{45.51}$/S167$^{45.52}$ region. Synergistically, such interactions facilitate cpd60 anchoring within the extracellular vestibule in place of the canonical orthosteric cavity. Critically, the proximity of cpd60 to the E166$^{45.51}$/S167$^{45.52}$ region suggests that its inhibitory effect on HCAR1 may not only be limited to steric occlusion of the pocket entrance, but also involve local perturbation of the extracellular interaction condition which is associated with agonist recognition and receptor activation. Besides, given the reported effective role of E166$^{45.51}$ in ligand recognition of HCAR1 Pan et al. (2025), cpd60 positioning close to the E166$^{45.51}$/S167$^{45.52}$ region further supports its advantages in hindering the native agonist lactic acid from reaching its OBP region and activating the receptor.

To assess whether the docked pose of cpd60 could be maintained under dynamic conditions, three independent molecular dynamics (MD) simulations are performed for the cpd60–HCAR1 complex. In the later stage of the simulations, the ligand RMSD remains within a limited fluctuation range of 1.5–2.2 Å across all three trajectories, indicating the overall stability of the bound pose (Figure 12c). The number of intermolecular hydrogen bonds between cpd60 and HCAR1 is also maintained at around three during the simulations (Figure 12d). Consistent with these observations, MM/PBSA analysis yields an average binding free energy of -25.15 kcal/mol across the three simulations (Figure 12e and Figure 13a). Together, these results support the stable association of cpd60 with HCAR1 under dynamic conditions. Consistent with the overall stability of the complex, structural snapshots extracted at 0, 100, and 200 ns also show that cpd60 remains in a similar binding region throughout the trajectory despite local rearrangements of side-chain contacts (Figure 12f). Detailed interaction

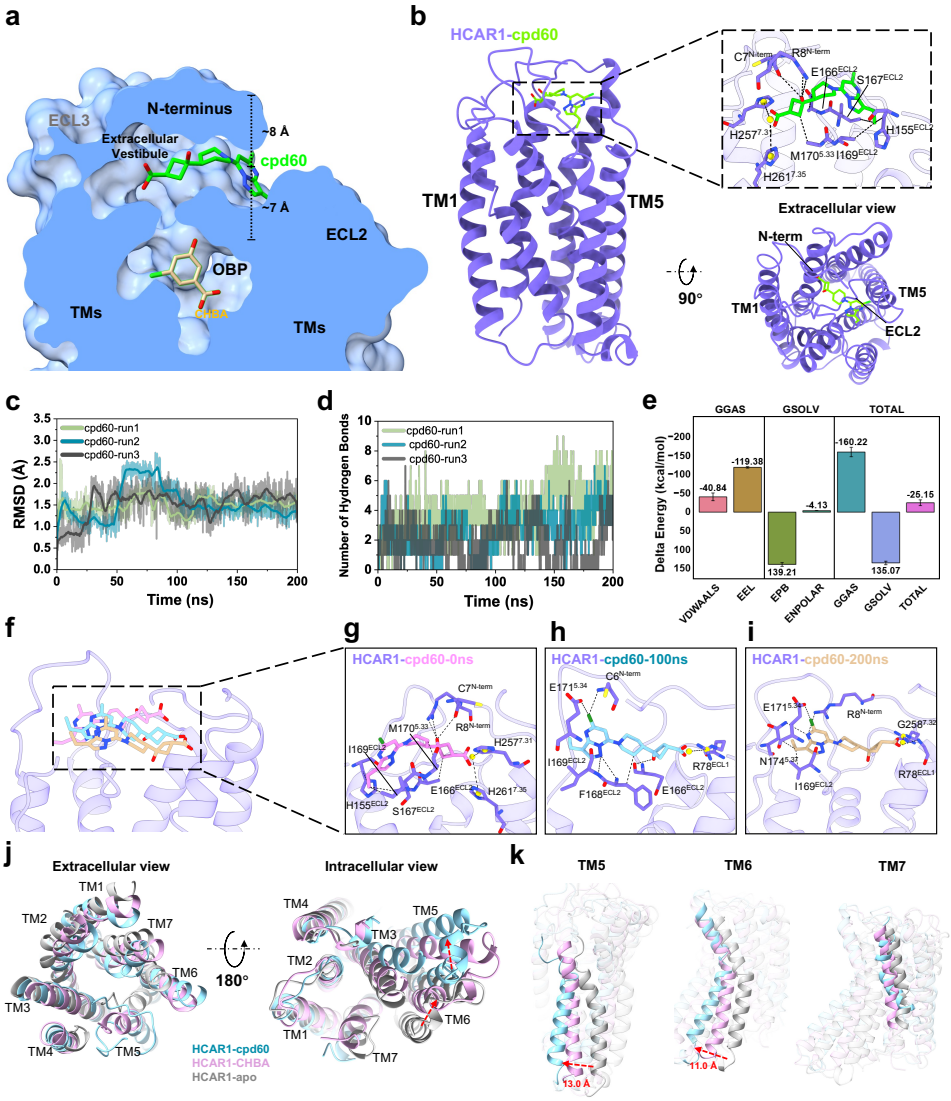


Figure 12: **Stable vestibular binding of cpd60 with HCAR1 and its conformational effects on HCAR1**. **a** Cross-sectional view of the superposition between the cpd60-HCAR1 docking model and the CHBA-bound HCAR1 crystal structure (PDB ID: 9IZD), showing that cpd60 occupies the extracellular vestibule above the OBP region. **b** Docking pose of cpd60 in HCAR1 and detailed interactions with the surrounding residues. Dashed lines indicate non-covalent interactions. **c** Ligand RMSD of cpd60 in three independent MD simulations. **d** Number of intermolecular hydrogen bonds between cpd60 and HCAR1. **e** MM/PBSA energy decomposition and average binding free energy of the cpd60–HCAR1 complex from the three simulations. **f** Superposition of cpd60-bound conformations at 0, 100, and 200 ns. **g-i** Detailed interactions of cpd60 with the surrounding residues at 0, 100, and 200 ns in MD simulation. Dashed lines indicate non-covalent interactions. **j** Structural superposition of apo-, CHBA-bound, and cpd60-bound HCAR1 conformations at 200 ns after alignment to the TM1–4 core, shown from extracellular and intracellular views. **k** Comparison of the conformational changes of TM5–TM7 among apo-, CHBA-bound, and cpd60-bound HCAR1 at 200 ns.

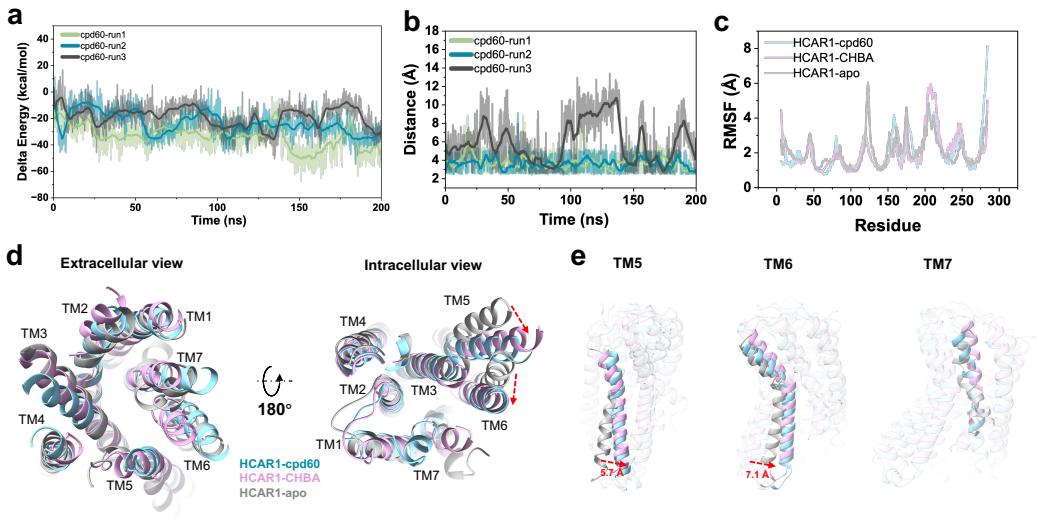


Figure 13: **Additional MD analyses of the cpd60-HCAR1.** **a** Binding free energy of the cpd60–HCAR1 complex during three independent MD simulations. **b** Distance between the chlorine atom of cpd60 and the OE2 atom of Glu171. **c** RMSF of HCAR1 in the cpd60-bound, CHBA-bound, and apo states. **d** Structural superposition of apo-, CHBA-bound, and cpd60-bound HCAR1 conformations at 100 ns after alignment to the TM1–4 core, shown from extracellular and intracellular views. **e** Comparison of the conformational changes of TM5–TM7 among apo-, CHBA-bound, and cpd60-bound HCAR1 at 100 ns.

views further show that the ligand maintained persistent contacts with residues in the extracellular entrance region, including His261 and Ile169, although the specific side-chain interaction patterns varied over time (Figure 12g–i). Significantly, the chlorine substituent of cpd60 forms halogen-bond interactions with the OE2 atom of Glu171 during the simulations. The distance between the chlorine atom of cpd60 and the OE2 atom of Glu171 remained at $\sim$2.5$-$4 Å (see Figure 13), which is favorable for halogen-bond formation Wilcken et al. (2013). Halogen atoms can serve as directional interaction sites through anisotropic charge distribution, thereby helping to stabilize cpd60 binding Cavallo et al. (2016).

To further compare the conformational effects on HCAR1 of cpd60 binding, the structure of cpd60-HCAR1 complex is superposed with those of CHBA-HCAR1 and the HCAR1 apo state, and structural rearrangements of HCAR1 are examined across the three systems. By 200 ns, pronounced differences in HCAR1 structure are observed between the cpd60-bound and CHBA-bound conformations, particularly at the intracellular ends of TM5 and TM6 (Figure 12j). At this stage, the intracellular ends of TM5 and TM6 in the cpd60-bound receptor are shifted by $\sim$13.0 Å and $\sim$11.0 Å, respectively, relative to the CHBA-bound state (Figure 12k). TM7 also underwent a smaller but detectable accompanying rearrangement. These findings indicate that, by the end of the simulation, cpd60 binding changes the transmembrane conformation of HCAR1, which is distinct from that of the CHBA-bound receptor. Notably, such divergence has already emerged at an earlier stage of the simulation. At 100 ns, both the HCAR1 conformations of cpd60-bound and CHBA-bound has already deviated from the apo state, with detectable rearrangements at the intracellular ends of TM5 and TM6 (see Figure 13d-e). At this time point, the displacement between the cpd60-bound and CHBA-bound conformations has reached $\sim$5.7 Å for TM5 and $\sim$7.1 Å for TM6, indicating that the HCAR1 conformational divergence between the two ligand-bound states is already evident during the intermediate phase of the simulation. The difference in displacement patterns between 100 and 200 ns further suggests that the observed motions do not simply reflect a rigid-body shift, but rather gradual repacking and reorientation of the intracellular helical bundle over time. This behavior is consistent with the conformational plasticity reported for HCAR1, in which receptor activation is accompanied by coordinated rearrangements of TM5–TM7, particularly the TM6 at the

cytoplasmic side Pan et al. (2025). In this context, the larger displacement of intracellular TM5 and TM6 in the cpd60-bound simulation indicates that cpd60 does not simply preserve the CHBA-like helical arrangement, but shifts HCAR1 toward a distinct transmembrane conformational ensemble. Although these data do not define a fully inactive state, they suggest that vestibular binding of cpd60 may perturb the transmembrane signaling architecture linking the extracellular entrance region to the intracellular activation interface. Collectively, these findings suggest that targeting the extracellular vestibule may represent a feasible inhibitory mechanism for HCAR1, with cpd60 providing a structural scaffold for the development of entrance-blocking antagonists.

F.2 Molecular Docking

Molecular docking is performed using the HCAR1 structure (PDB ID: 9J8Z) as the receptor model. Docking calculations are carried out with AutoDock Vina Eberhardt et al. (2021), with the grid box centered at 128.5, 143.4, and 95.9, and the box dimensions set to 22.5 × 22.5 × 22.5 Å to encompass the ligand-accessible pocket region. Protein–ligand interactions in the resulting coupling poses are characterized using PLIP Adasme et al. (2021), and representative binding modes are rendered and visualized in ChimeraX 1.11.1 Meng et al. (2023).

F.3 Molecular Dynamics Simulations

The membrane builder module of the CHARMM-GUI server Jo et al. (2008) is used to prepare the simulation system, including a POPC lipid bilayer, TIP3P water molecules, and 0.15 M Na^+/Cl^- ions. The system is described using the CHARMM36m force field, and ligand parameters are generated using the CGenFF force field Lee et al. (2016). All MD simulations are performed using GROMACS 2024.2 Abraham et al. (2015). The MD simulation protocol consisted of energy minimization, equilibration and production phases. The system is constructed using a rectangular box. The temperature is controlled through NVT equilibration (tcoupl = v-rescale), while pressure is regulated via NPT equilibration (pcoupl = c-rescale). Hydrogen-containing bonds are constrained (constraints = h-bonds), and all constraints are handled using the LINCS algorithm. Nonbonded interactions are treated using the Verlet cutoff scheme, with cutoff distances of 1.2 nm for the neighbor list, van der Waals interactions, and short-range electrostatic interactions (rlist = 1.2, rvdw = 1.2, rcoulomb = 1.2). Long-range electrostatic interactions are calculated using the particle mesh Ewald (PME) method, with the default PME grid spacing of 0.12 nm. Periodic boundary conditions are applied in all three dimensions (pbc = xyz), and default GROMACS settings are used for PME-related parameters. All systems are first energy-minimized for 5000 steps using the steepest descent algorithm, with positional restraints of 1000 kJ·(mol·nm^2)$^{-1}$ applied to solute atoms. The equilibration process is then performed in six steps to gradually relax the system at the target temperature of 303.15 K and pressure of 1.0 bar. Backbone and ligand atoms are restrained with force constants of 4000, 2000, 1000, 500, 200, and 50 kJ·(mol·nm^2)$^{-1}$, respectively; side-chain atoms are restrained with 2000, 1000, 500, 200, 50, and 0 kJ·(mol·nm^2)$^{-1}$; and lipid atoms were restrained with 1000, 400, 400, 200, 40, and 0 kJ·(mol·nm^2)$^{-1}$. The six equilibration steps are conducted for 125, 125, 125, 500, 500, and 500 ps, respectively. During the production phase, simulations are carried out without any additional restraints using a time step of 2 fs for 100,000,000 steps, corresponding to a total simulation time of 200 ns. In this phase, temperature is maintained using the Nosé–Hoover thermostat (tcoupl = Nose-Hoover), and pressure is controlled using the Parrinello–Rahman barostat (pcoupl = Parrinello-Rahman). Three independent production runs are performed for each system. The binding free energy of the ligand-protein complex is calculated by gmx_MMPBSA Valdés-Tresanco et al. (2021); Miller III et al. (2012). The binding free energy is calculated as follows:

$$\Delta G_{\text{total}} = \Delta G_{\text{gas}} + \Delta G_{\text{solv}} = \Delta G_{\text{vdw}} + \Delta G_{\text{ele}} + \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}}, \tag{18}$$

where ΔG_{gas} and ΔG_{solv} represent the molecular mechanics term (energy in the gas phase) and solvation energy, respectively. ΔG_{vdw}, ΔG_{ele}, ΔG_{polar}, and $\Delta G_{nonpolar}$ represent the van der Waals energy term, electrostatic energy term, polar solvation energy, and nonpolar solvation energy, respectively.

F.4 Statistical Analysis

Details regarding data preprocessing, model evaluation, cAMP assays, and MD simulations are provided in Sections B.1.1– B.1.3, B.4, F.2 and F.3, respectively. For virtual-screening benchmarks, re-

sults are reported using mean and median values, together with EF, AUROC, BEDROC ($\alpha = 80.5$), and Average Rank. Pearson correlation is used to assess the association between ReqVS ranks and inhibition rates. Concentration–response curves are fitted using a four-parameter logistic model, and IC_{50} values are reported as mean $\pm$ SEM from three technical replicates. MD results are summarized across three independent simulations. Statistical analysis and visualization are performed using Python and GraphPad Prism 10.0.