CAUSALBIND: CAUSAL CONCEPT ALIGNMENT FOR PROTEIN-LIGAND VIRTUAL SCREENING

Loka Li[1], **Jin Tian**[1], **Kun Zhang**[1,2]
[1] Mohamed bin Zayed University of Artificial Intelligence
[2] Carnegie Mellon University

ABSTRACT

Drug discovery is a lengthy and costly process, with virtual screening serving as a critical computational step to identify promising drug candidates from vast compound libraries. While contrastive learning has emerged as a powerful paradigm for protein-ligand virtual screening by aligning molecular and protein pocket embeddings, existing methods directly align entire representations without distinguishing binding-relevant from binding-irrelevant features. This can lead to spurious correlations that limit generalization to novel drug targets. In this work, we propose a *plug-in causal concept extraction module* that decomposes entangled representations into disentangled atomic concepts using cross-attention and employs learnable sparse masking to identify causally relevant binding features. Experiments show that across both scratch and pretrained settings, our causal alignment consistently improves early enrichment (BEDROC and EF@1%), with the largest gains observed on the more realistic LIT-PCBA benchmark, indicating better prioritization of true binders despite marginal changes in overall AUC.

1 INTRODUCTION

Modern drug discovery faces a fundamental challenge: identifying promising drug candidates from libraries containing millions of compounds (Berdigaliyev & Aljofan, 2020). An overview of the drug discovery pipeline is given in Appendix A. Virtual screening addresses this bottleneck by computationally predicting protein-ligand binding affinity, enabling researchers to prioritize compounds for expensive experimental validation (Jia et al., 2026). Recent advances in deep learning, particularly contrastive learning methods, have achieved remarkable success in this domain by learning joint representations of molecules and protein binding pockets (Gao et al., 2023; Shen et al., 2023; Zhang et al., 2023; Cao et al., 2024; Feng et al., 2025; Han et al., 2025; Wang et al., 2026).

Despite their success, existing contrastive learning methods for virtual screening face a fundamental challenge: they directly align entire molecular and pocket representations without distinguishing between binding-relevant and binding-irrelevant features. While protein-ligand binding involves complex conformational adaptation (induced fit), certain structural features remain invariant determinants of binding affinity (Koshland Jr, 1958; Wermuth, 2006). See Appendix B for more discussions on biological foundation. Existing methods fail to explicitly identify these invariant features, instead learning entangled representations that may capture spurious correlations (Gao et al., 2023; Wang et al., 2026). For instance, the model might learn to associate molecular weight or general hydrophobicity patterns with pocket size, even when these correlations reflect dataset biases rather than true binding mechanisms. Such spurious correlations limit generalization to novel drug targets.

Related Work. Virtual screening methods span classical docking approaches like Glide (Friesner et al., 2004), AutoDock Vina (Trott & Olson, 2010), and Surflex (Spitzer & Jain, 2012), machine learning scoring functions such as Gnina (McNutt et al., 2021), RTMScore (Shen et al., 2022), GenScore (Shen et al., 2023), EquiScore (Cao et al., 2024), PIGNet (Moon et al., 2022), and PLANET (Zhang et al., 2023), as well as drug-target affinity (DTA) prediction methods like DeepDTA (Öztürk et al., 2018) and GraphDTA (Nguyen et al., 2021). Recent contrastive learning approaches, including DrugCLIP (Gao et al., 2023), DrugHash (Han et al., 2025), LigUnity (Feng et al., 2025), and HypSeek (Wang et al., 2026), have achieved state-of-the-art results by learning joint protein-ligand

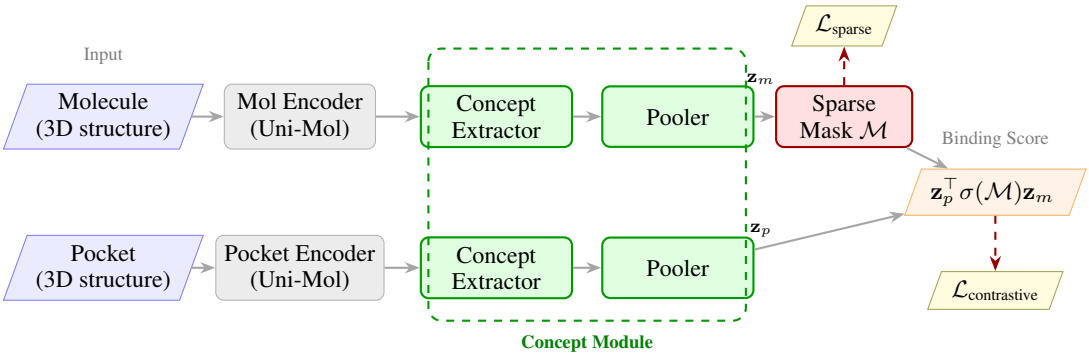

Figure 1: Overview of CAUSALBIND. Both branches use shared Concept Extractor and Pooler to obtain representations $\mathbf{z}_m$ and $\mathbf{z}_p$. The molecule branch will apply *SparseMask*: $\tilde{\mathbf{z}}_m = \sigma(\mathcal{M})\mathbf{z}_m$, where $\mathcal{M} \in \mathbb{R}^{d_c \times d_c}$ is a learnable mask matrix and $\sigma(\cdot) = \tanh(\cdot) + 1$ maps values to $[0, 2]$. The model is trained with contrastive loss on the binding score and L_1 sparsity loss on the mask.

representations, often leveraging pretrained molecular encoders such as Uni-Mol (Zhou et al., 2023) that capture rich 3D structural information. Structure-based methods like TANKBind (Lu et al., 2022) and BigBind (Brocidiacono et al., 2023) directly predict binding poses. However, none of these methods explicitly address spurious correlations in representation learning. Our work draws from causal representation learning (Schölkopf et al., 2021; Daunhawer et al., 2023; Xie et al., 2025; Sun et al., 2025b), which showed that sparse multimodal connections enable component-wise concept identification. A comprehensive review of related work is provided in Appendix C. Due to space constraints, theoretical analysis including identifiability conditions and proofs are in Appendix D.

In this paper, we propose CAUSALBIND, a plug-in causal concept extraction module that can be integrated into existing contrastive learning frameworks for virtual screening in drug discovery. Our contributions are threefold: (1) we identify the problem of spurious correlations in contrastive learning-based virtual screening and propose to address it through causal concept extraction, grounded in identifiability theory from causal representation learning; (2) we design a plug-in module consisting of a *ConceptExtractor* (Perceiver-style architecture) and a *SparseMask* (learnable sparse transformation) that can be seamlessly integrated into existing models (Figure 1); (3) we validate our approach on two complementary settings demonstrating consistent improvements on drug discovery tasks.

2 BACKGROUND

2.1 CONTRASTIVE LEARNING FOR VIRTUAL SCREENING

Existing contrastive learning methods for virtual screening (Gao et al., 2023; Han et al., 2025; Feng et al., 2025; Wang et al., 2026) follow a common framework. Given a dataset of protein-ligand binding pairs $\{(m_i, p_i)\}_{i=1}^N$, where m_i denotes a molecule and p_i denotes a protein pocket, the goal is to learn encoders that map molecules and pockets into a shared embedding space where binding pairs are close and non-binding pairs are far apart, enabling efficient similarity-based screening.

Specifically, let $f_m : \mathbb{M} \to \mathbb{R}^d$ and $f_p : \mathcal{P} \to \mathbb{R}^d$ denote the molecule and pocket encoders, respectively. The standard approach optimizes an InfoNCE-style contrastive loss:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2}\left(\log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ij}/\tau)} + \log \frac{\exp(s_{ii}/\tau)}{\sum_j \exp(s_{ji}/\tau)}\right), \tag{1}$$

where $s_{ij} = f_m(m_i)^\top f_p(p_j)$ is the similarity score between molecule i and pocket j, and τ is a temperature parameter. While this framework has achieved impressive results on standard benchmarks, it directly aligns the representations without distinguishing binding-relevant from binding-irrelevant features, potentially learning spurious correlations that limit generalization to novel targets.

2.2 PROBLEM: IDENTIFYING INVARIANT BINDING DETERMINANTS

Biological Context. Protein-ligand binding is governed by the *induced fit* mechanism (Koshland Jr, 1958), where both protein pockets and ligands undergo conformational changes during binding. Despite this flexibility, certain structural motifs (*pharmacophores*), such as hydrogen bond donors/acceptors and hydrophobic groups, consistently determine binding across diverse conformational states (Wermuth, 2006). Please refer to Appendix B for more detailed discussions.

The Challenge of Spurious Correlations. Molecular and pocket representations contain a mixture of invariant binding determinants (analogous to pharmacophoric features) and spurious correlations that arise from dataset biases (Chen et al., 2019; Graber et al., 2025). When contrastive learning directly aligns entire representations, it may exploit spurious correlations (e.g., large molecules bind to large pockets) rather than learning the structural features that causally drive binding. Such correlations fail to generalize to novel protein families, limiting the model's utility in real world.

3 METHODOLOGY

3.1 CAUSAL CONCEPT EXTRACTION MODULE

Notation. Let $\mathbf{h}_m, \mathbf{h}_p \in \mathbb{R}^d$ denote encoder outputs for molecule and pocket. The ConceptExtractor produces concept matrices $\mathbf{C}^M, \mathbf{C}^P \in \mathbb{R}^{K \times d_c}$ containing K atomic concepts $\mathbf{c}_m^k, \mathbf{c}_p^k \in \mathbb{R}^{d_c}$. Finally, $\mathbf{z}_m, \mathbf{z}_p \in \mathbb{R}^{d_c}$ are pooled representations (mean over concepts). Inspired by recent advances in causal representation learning (Xie et al., 2025; Sun et al., 2025b), we propose to decompose entangled representations into disentangled atomic concepts and learn sparse connections between them. Our plug-in module consists of two key components: *ConceptExtractor* and *SparseMask*.

Causal Structure: Binding as a Collider Variable. Unlike vision-language models where text concepts generate visual features (e.g., ConceptAligner (Xie et al., 2025) uses $\mathbf{c}^{\text{text}} \to \mathbf{c}^{\text{image}}$), protein-ligand binding involves *symmetric complementary matching*: neither molecule nor pocket "generates" the other. Instead, binding happens only when a specific pair of molecular and pocket concepts jointly activate the binding mechanism. We model this as a *V-structure*:

$$m \to \mathbf{c}_m^i \to B_{ij} \leftarrow \mathbf{c}_p^j \leftarrow p, \tag{2}$$

where molecule m and pocket p induce atomic concepts $\mathbf{c}_m^i$ and $\mathbf{c}_p^j$, respectively. The binding indicator B_{ij} acts as a collider variable that is activated only by specific pairs of molecular concept $\mathbf{c}_m^i$ and pocket concept $\mathbf{c}_p^j$. This structure reflects biological reality: a hydrogen bond donor in the molecule must match with an acceptor in the pocket to enable binding. Spurious concept pairs are explicitly blocked ($B_{ij} = 0$), while causally relevant pairs activate binding ($B_{ij} = 1$), as illustrated in Figure 2. The identifiability theory from Xie et al. (2025) remains applicable: by varying molecules (varying $\mathbf{C}^M$), we can identify pocket concepts $\mathbf{C}^P$ component-wise.

ConceptExtractor. We employ a Perceiver-style architecture (Jaegle et al., 2021) to extract atomic concepts from encoder outputs. Given an encoder output $\mathbf{h} \in \mathbb{R}^{d_{\text{enc}}}$ (e.g., the CLS token from a transformer encoder), the *ConceptExtractor* produces a set of concept representations:

$$\mathbf{C} = \textit{ConceptExtractor}(\mathbf{h}) \in \mathbb{R}^{K \times d_c}, \tag{3}$$

where K is the number of concepts and d_c is the concept dimension.

The *ConceptExtractor* uses learnable concept queries $\mathbf{Q} \in \mathbb{R}^{K \times d_c}$ that attend to the input through cross-attention layers:

$$\mathbf{C}^{(l+1)} = \mathbf{C}^{(l)} + \textit{CrossAttn}(\mathbf{C}^{(l)}, \textit{Proj}(\mathbf{h})) + \textit{FFN}(\mathbf{C}^{(l)}), \tag{4}$$

where $\mathbf{C}^{(0)} = \mathbf{Q}$ and $\textit{Proj}(\cdot)$ projects the encoder output to the concept dimension. The concept representations are then aggregated via mean pooling:

$$\mathbf{z} = \frac{1}{d_c} \sum_{t=1}^{d_c} \mathbf{C}_t \in \mathbb{R}^K. \tag{5}$$

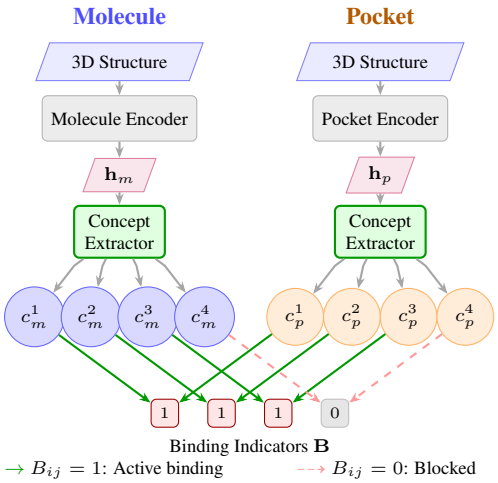

Figure 2: Causal structure for protein-ligand binding. Both molecule and pocket are processed through encoders and ConceptExtractor to produce atomic concepts. Unlike text-to-image where causality is unidirectional, binding involves **symmetric matching** modeled as a V-structure: binding indicator B_{ij} is activated only when concept pair (c_m^i, c_p^j) jointly determines binding (e.g., H-bond donor matches acceptor). Spurious pairs are blocked ($B_{ij} = 0$). The sparse connectivity enables identifiability. Please refer to Appendix D for the detailed analysis.

SparseMask. To identify binding-relevant concepts, we introduce a learnable sparse mask that applies a linear transformation to the aggregated concept representation:

$$\tilde{\mathbf{z}} = \sigma(\mathcal{M})\mathbf{z}, \tag{6}$$

where $\mathcal{M} \in \mathbb{R}^{d_c \times d_c}$ is a learnable mask matrix applied to the molecule branch. Here $\sigma(\cdot) = \tanh(\cdot)+1$ is applied element-wise to $\mathcal{M}$, mapping each entry to $[0, 2]$, and $\sigma(\mathcal{M})\mathbf{z}$ denotes standard matrix-vector multiplication. To encourage sparsity, we add an L_1 regularization term:

$$\mathcal{L}_{\text{sparse}} = \|\sigma(\mathcal{M})\|_1. \tag{7}$$

3.2 Why Sparse Masking Helps

The key insight is that *sparsity enables identifiability*. Standard contrastive learning allows dense connections where every molecular feature can influence every pocket feature, enabling spurious correlations (e.g., aligning "large molecules" with "large pockets"). In contrast, L_1 regularization forces the model to select only critical connections—for example, connecting "H-bond donors" specifically to "H-bond acceptors" while suppressing irrelevant links. This sparse structure makes concepts identifiable: we can determine which molecular features causally determine binding. Recent theory (Xie et al., 2025; Daunhawer et al., 2023) confirms that sparse graphical structures enable component-wise identifiability of latent concepts. See Theorem 1 in Appendix D for more details.

Practical Implementation. As a plug-in module, our method augments any base loss $\mathcal{L}_{\text{base}}$ with sparsity regularization:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{sparse}}\|\sigma(\mathcal{M})\|_1, \tag{8}$$

where $\mathcal{M}$ is the learnable mask matrix, $\sigma(\cdot) = \tanh(\cdot) + 1$ maps values to $[0, 2]$, and λ_{sparse} controls the sparsity strength. For example, $\mathcal{L}_{\text{base}}$ can be InfoNCE in DrugCLIP (Gao et al., 2023), or the combination of hyperbolic contrastive and cone hierarchy loss in HypSeek (Wang et al., 2026).

4 Experiments

Experimental Setup. We evaluate on two complementary benchmarks: DUD-E (Mysinger et al., 2012) (102 targets, 22,886 actives, 1.3M decoys) and LIT-PCBA (Mayr et al., 2018) (15 targets with more challenging, experimentally-derived decoys). DUD-E provides comprehensive coverage across diverse protein families, while LIT-PCBA offers more realistic evaluation with decoys selected to minimize property-based biases. We report AUC (overall ranking quality), BEDROC with $\alpha = 20$ (early enrichment), and EF@1% (enrichment factor at top 1%). For drug discovery, early enrichment metrics are particularly important since only top-ranked compounds proceed to experimental validation. Implementation details including hyperparameters are provided in Appendix F.

Table 1: Results of DrugCLIP trained from scratch. Our module significantly improves all metrics on both benchmarks, with particularly large gains on the more challenging LIT-PCBA benchmark.

Model	DUD-E			LIT-PCBA		
	AUC	BEDROC	EF@1%	AUC	BEDROC	EF@1%
DrugCLIP (Baseline)	0.496	0.037	1.64	0.504	0.013	0.60
+ CausalBind (Ours)	**0.523**	**0.049**	**2.42**	**0.532**	**0.035**	**2.49**
Δ Improvement	+5.4%	+32.4%	+47.6%	+5.6%	+169%	+315%

Table 2: Results of HypSeek with pretrained Uni-Mol encoders. Our causal module provides consistent improvements in early enrichment metrics, with larger gains on the LIT-PCBA benchmark.

Model	DUD-E			LIT-PCBA		
	AUC	BEDROC	EF@1%	AUC	BEDROC	EF@1%
HypSeek (Baseline)	0.928	0.694	44.10	0.597	0.056	4.97
+ CausalBind (Ours)	0.924	**0.705**	**45.02**	**0.617**	**0.071**	**5.53**
Δ Improvement	-0.4%	+1.6%	+2.1%	**+3.4%**	**+26.8%**	**+11.3%**

Exp 1: DrugCLIP from Scratch. We select DrugCLIP (Gao et al., 2023) as our base model since it is an early representative work applying contrastive learning to virtual screening, widely recognized and recently adopted by a *Science* study (Jia et al., 2026). To isolate our module's effect without confounding from pretrained representations, we train DrugCLIP from scratch using randomly initialized Uni-Mol encoders. This controlled setting reveals how much our causal concept extraction contributes to representation learning itself. As shown in Table 1, our module provides substantial gains across all metrics: on DUD-E, AUC improves by +5.4%, BEDROC by +32.4%, and EF@1% by +47.6%; on the more challenging LIT-PCBA, improvements are even more pronounced with AUC +5.6%, BEDROC +169%, and notably EF@1% +315%. The larger gains on LIT-PCBA suggest that our sparse masking mechanism is particularly effective at suppressing property-based spurious correlations that LIT-PCBA's more realistic decoys are specifically designed to exploit.

Exp 2: HypSeek with Pretraining. We integrate our module into HypSeek (Wang et al., 2026), a state-of-the-art method that combines hyperbolic embeddings with pretrained Uni-Mol (molecular) and ESM-2 (protein sequence) encoders, trained via a three-stage pipeline with hyperbolic contrastive and cone hierarchy losses. Table 2 shows that even on top of this highly optimized baseline, our module provides consistent improvements: on DUD-E, BEDROC improves by +1.6% and EF@1% by +2.1% while maintaining competitive AUC; on LIT-PCBA, we observe larger gains with AUC +3.4%, BEDROC +26.8%, and EF@1% +11.3%. The larger improvements on LIT-PCBA (which has more realistic decoys) compared to DUD-E suggest our causal concept extraction helps the model generalize beyond dataset-specific biases. The slight AUC decrease on DUD-E (-0.4%) accompanied by improved early enrichment indicates that our module trades off ranking of low-confidence predictions for better identification of high-confidence true binders, a favorable trade-off for practical drug discovery where only top candidates matter. See Appendix G for details.

5 Conclusion

We proposed CausalBind, a plug-in causal concept extraction module for protein-ligand virtual screening that can be seamlessly integrated into existing contrastive learning frameworks. By decomposing representations into atomic concepts and learning sparse connections, our module helps identify binding-relevant features while suppressing spurious correlations. Experiments on two complementary settings, DrugCLIP from scratch and HypSeek with pretraining, demonstrate that our module provides consistent improvements, particularly on early enrichment metrics that are most relevant for practical drug discovery. Limitations and broader impact are discussed in Appendix H.

REFERENCES

Nurken Berdigaliyev and Mohamad Aljofan. An overview of drug discovery and development. *Future medicinal chemistry*, 12(10):939–947, 2020.

Michael Brocidiacono, Paul Francoeur, Rishal Aggarwal, Konstantin I Popov, David Ryan Koes, and Alexander Tropsha. Bigbind: learning from nonstructural data for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 64(7):2488–2495, 2023.

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.

Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J Dickson, Jose S Duca, Viktor Hornak, David R Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, 14(8):e0220113, 2019.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. *arXiv preprint arXiv:2303.09166*, 2023.

Bin Feng, Zijing Liu, Hao Li, Mingjun Yang, Junjie Zou, He Cao, Yu Li, Lei Zhang, and Sheng Wang. Hierarchical affinity landscape navigation through learning a shared pocket-ligand space. *Patterns*, 6(10), 2025.

Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36:44595–44614, 2023.

Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, pp. 1–13, 2025.

Jin Han, Yun Hong, and Wu-Jun Li. Drughash: Hashing based contrastive learning for virtual screening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17041–17049, 2025.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 859–868. PMLR, 2019.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Yinjun Jia, Bowen Gao, Jiaxin Tan, Jiqing Zheng, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Liping Tan, Hongyi Cai, et al. Deep contrastive learning enables genome-wide virtual screening. *Science*, 391(6781):eads9530, 2026.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.

Daniel E Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences*, 44(2):98–104, 1958.

Loka Li, Haoyue Dai, Hanin Al Ghothani, Biwei Huang, Jiji Zhang, Shahar Harel, Isaac Bentwich, Guangyi Chen, and Kun Zhang. On causal discovery in the presence of deterministic relations. *Advances in Neural Information Processing Systems*, 37:130920–130952, 2024a.

Loka Li, Ignavier Ng, Gongxu Luo, Biwei Huang, Guangyi Chen, Tongliang Liu, Bin Gu, and Kun Zhang. Federated causal discovery from heterogeneous data. *arXiv preprint arXiv:2402.13241*, 2024b.

Loka Li, Wong Yu Kang, Minghao Fu, Guangyi Chen, Zhenhao Chen, Gongxu Luo, Yuewen Sun, Salman Khan, Peter Spirtes, and Kun Zhang. Personax: Multimodal datasets with llm-inferred behavior traits. *arXiv preprint arXiv:2509.11362*, 2025.

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical science*, 9(24):5441–5451, 2018.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Seokhyun Moon, Wonho Zhung, Soojung Yang, Jaechang Lim, and Woo Youn Kim. Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022.

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.

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.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan, Tingjun Hou, and Yu Kang. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *Journal of Medicinal Chemistry*, 65(15):10691–10706, 2022.

Chao Shen, Xujun Zhang, Chang-Yu Hsieh, Yafeng Deng, Dong Wang, Lei Xu, Jian Wu, Dan Li, Yu Kang, Tingjun Hou, et al. A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chemical Science*, 14(30):8129–8146, 2023.

Russell Spitzer and Ajay N Jain. Surflex-dock: Docking benchmarks and real-world application. *Journal of computer-aided molecular design*, 26(6):687–699, 2012.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multi-modal biomedical observations. *ArXiv*, pp. arXiv–2411, 2025a.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multimodal biomedical observations. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Jianhui Wang, Wenyu Zhu, Bowen Gao, Xin Hong, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Learning protein-ligand binding in hyperbolic space. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Camille G Wermuth. Pharmacophores: historical perspective and viewpoint from a medicinal chemist. *Pharmacophores and pharmacophore searches*, 32:1–13, 2006.

Shaoan Xie, Lingjing Kong, Yujia Zheng, Zeyu Tang, Eric P Xing, Guangyi Chen, and Kun Zhang. Learning vision and language concepts for controllable image generation. In *Forty-second International Conference on Machine Learning*, 2025.

Xiangying Zhang, Haotian Gao, Haojie Wang, Zhihang Chen, Zhe Zhang, Xinchong Chen, Yan Li, Yifei Qi, and Renxiao Wang. Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 64(7):2205–2220, 2023.

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*, 2023.

Appendix

Table of Contents

A. Drug Discovery Pipeline and Method Positioning
B. Biological Foundation: Induced Fit and Invariant Features
C. Extended Related Work
D. Identifiability Theory: Formal Statement
E. Proof of Theorem 1
F. Implementation Details
G. Experimental Analysis
H. Limitations and Broader Impact

A Drug Discovery Pipeline and Method Positioning

Figure 3 illustrates the typical drug discovery pipeline and positions our method within this workflow. Virtual screening serves as a critical early-stage filter, reducing compound libraries from 10^5–10^6 to 10^2–10^3 candidates for experimental validation. Our CausalBind module operates at this stage, improving the quality of contrastive learning-based virtual screening methods.

Stage Descriptions. We provide detailed descriptions of each stage in the drug discovery pipeline:

- **Virtual Compound Library** (10^9–10^{10}): Large-scale databases of commercially available or synthetically accessible compounds. Examples include ZINC (purchasable compounds) and Enamine REAL (make-on-demand via combinatorial chemistry). These libraries represent the theoretical search space for drug discovery.
- **Filtered Library** (10^5–10^6): Compounds passing basic drug-likeness filters such as Lipinski's Rule of Five (molecular weight $<$500, logP $<$5, H-bond donors $<$5, acceptors $<$10), PAINS filters (removing pan-assay interference compounds), and target-specific property constraints. This is the practical starting point for most virtual screening campaigns.

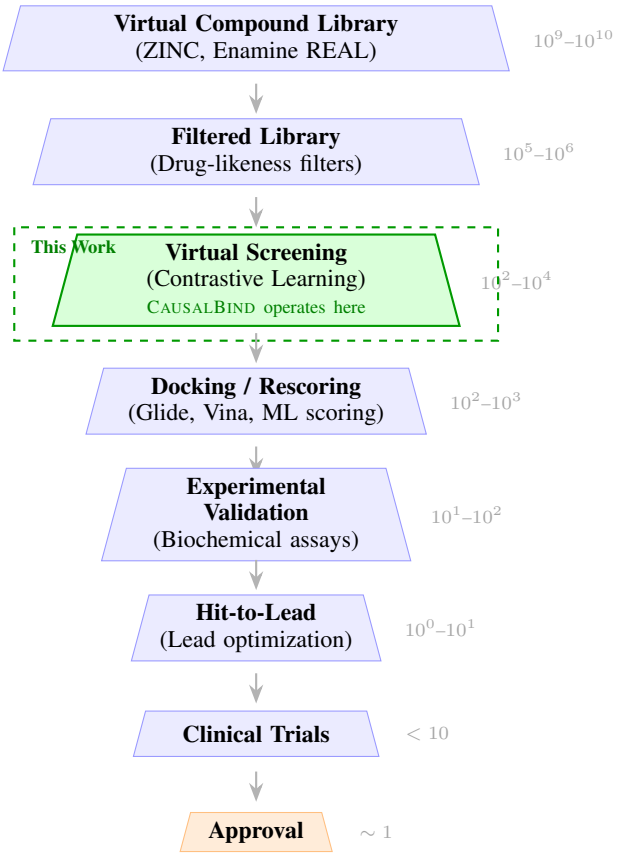

Figure 3: Drug discovery pipeline funnel. Each stage progressively filters compounds, with numbers indicating typical compound counts. Our CausalBind method (green box) operates at the virtual screening stage, improving the quality of top-ranked candidates before expensive downstream validation. By identifying invariant binding determinants and suppressing spurious correlations, our method aims to increase the hit rate in subsequent experimental testing.

- **Virtual Screening** (10^2–10^4): *Our method operates here.* Computational methods rank compounds by predicted binding affinity. This includes ligand-based methods (similarity search, pharmacophore matching), structure-based methods (docking), and learning-based methods (contrastive learning, as in DrugCLIP and HypSeek). The goal is to identify a manageable set of candidates for experimental testing.
- **Docking / Rescoring** (10^2–10^3): Detailed binding pose prediction and affinity estimation using physics-based (Glide, AutoDock Vina) or ML-based (Gnina, RTMScore) scoring functions. This step provides atomic-level binding hypotheses and further filters candidates.
- **Experimental Validation** (10^1–10^2): Biochemical assays (e.g., fluorescence polarization, surface plasmon resonance) and cell-based assays measure actual binding activity. This is the first "wet lab" validation, typically costing \$100–\$1000 per compound.
- **Hit-to-Lead** (10^0–10^1): Medicinal chemistry optimization of confirmed hits. Chemists modify molecular scaffolds to improve potency, selectivity, and drug-like properties (ADMET: absorption, distribution, metabolism, excretion, toxicity).
- **Clinical Trials** ($<$10): Regulatory-approved testing in humans, progressing through Phase I (safety), Phase II (efficacy), and Phase III (large-scale efficacy). Each phase has high attrition rates.
- **Approval** ($\sim$1): Regulatory approval (FDA, EMA) for market release. The entire pipeline typically spans 10–15 years and costs \$1–2 billion per approved drug.

Why Virtual Screening Matters. The virtual screening stage is critical because it represents the largest reduction in compound count (from 10^5–10^6 to 10^2–10^4) before expensive experimental validation. The quality of this filtering directly impacts the success rate of downstream stages: a 10% improvement in early enrichment (e.g., BEDROC) can translate to significant cost savings by reducing the number of false positives sent to biochemical assays. Our method's focus on identifying causal binding determinants is particularly valuable at this stage, as it helps ensure that top-ranked compounds are selected based on true binding mechanisms rather than spurious correlations.

Complementary to Docking. Our contrastive learning approach is complementary to, not a replacement for, molecular docking. While docking methods like Glide and AutoDock Vina provide detailed binding pose predictions, they are computationally expensive ($\sim$seconds per compound). Contrastive learning methods can rapidly score millions of compounds ($\sim$milliseconds each), serving as an efficient pre-filter before docking. The improved representations learned by CAUSALBIND can enhance this pre-filtering, ensuring that compounds advanced to docking are more likely to be true binders.

B BIOLOGICAL FOUNDATION: INDUCED FIT AND INVARIANT FEATURES

The Induced Fit Theory. The *induced fit* mechanism, first proposed by Koshland in 1958 (Koshland Jr, 1958), revolutionized our understanding of protein-ligand binding. Unlike the earlier "lock-and-key" model which assumed rigid complementarity, induced fit recognizes that both the protein binding pocket and the ligand undergo conformational changes upon binding. This dynamic adaptation allows proteins to optimize interactions with ligands through mutual structural adjustment, enhancing binding affinity and specificity. Induced fit is now recognized as a fundamental principle governing molecular recognition in biological systems, from enzyme-substrate interactions to drug-target binding.

Pharmacophore: Invariant Structural Features. Despite the conformational flexibility inherent in induced fit, medicinal chemistry has established that certain structural features—termed *pharmacophores*—remain essential determinants of binding (Wermuth, 2006). A pharmacophore is defined as the ensemble of steric and electronic features necessary for optimal molecular interactions with a specific biological target. These include hydrogen bond donors and acceptors, hydrophobic centers, aromatic rings, and charged groups. Crucially, pharmacophoric features represent **invariant properties** that determine binding across conformational states: while the exact geometry may adjust through induced fit, the presence and spatial arrangement of these features remain causally relevant.

Connection to Causal Concept Learning. Our causal concept extraction approach is grounded in this biological insight. We hypothesize that deep learning representations of molecules and pockets contain both: (1) **invariant binding determinants** analogous to pharmacophoric features, which causally determine binding across diverse protein-ligand pairs and conformational states; and (2) **spurious correlations** arising from dataset biases (e.g., correlations between molecular weight and pocket size), which do not reflect true binding mechanisms. By decomposing representations into atomic concepts and learning sparse connections between modalities, our *SparseMask* effectively identifies which features correspond to invariant binding determinants.

Static Structures and Conformational Flexibility. A natural question arises: if binding involves dynamic conformational changes, how can learning from static structures identify invariant features? The key insight is that pharmacophoric features—the targets of our causal concept extraction—are themselves **conformationally robust**. While induced fit involves adjustments in bond angles and side-chain rotations, the fundamental chemical properties that enable binding (e.g., a hydroxyl group's ability to donate hydrogen bonds, a hydrophobic patch's interaction with nonpolar residues) remain invariant. Virtual screening benchmarks like DUD-E and LIT-PCBA use static crystal or docked structures, yet successful methods implicitly learn features that generalize across conformational variability. Our approach makes this implicit learning explicit through causal disentanglement.

Relevance to Drug Discovery. This biological grounding strengthens the practical relevance of our method. In drug discovery, identifying pharmacophoric features guides lead optimization:

medicinal chemists modify scaffolds while preserving key binding determinants. Our causal concepts, learned through sparse masking, offer a data-driven approach to discovering these determinants automatically. Moreover, by suppressing spurious correlations, our method should generalize better to novel targets—a critical requirement when screening compounds against previously unexplored proteins, where dataset biases from training data may not hold.

C Extended Related Work

Molecular Docking. Classical docking methods estimate binding affinity through physics-based scoring functions. Glide (Friesner et al., 2004) uses hierarchical filtering with optimized scoring, AutoDock Vina (Trott & Olson, 2010) employs empirical free energy scoring with fast search algorithms, and Surflex (Spitzer & Jain, 2012) combines molecular similarity with docking. These methods remain widely used but are computationally expensive and have limited accuracy for flexible binding sites.

ML-based Scoring Functions. Machine learning scoring functions have emerged as powerful alternatives to physics-based approaches. Gnina (McNutt et al., 2021) pioneered CNN-based scoring for docking, RTMScore (Shen et al., 2022) uses equivariant transformers for pose scoring, GenScore (Shen et al., 2023) leverages graph neural networks with molecular fingerprints, EquiScore (Cao et al., 2024) incorporates SE(3)-equivariant message passing, PIGNet (Moon et al., 2022) integrates physics-informed constraints, and PLANET (Zhang et al., 2023) uses multi-objective optimization. BigBind (Brocidiacono et al., 2023) extends these approaches to learn from nonstructural data. These methods typically require known binding poses and focus on rescoring rather than virtual screening.

Drug-Target Affinity Prediction. DTA methods predict binding affinity from sequence and structure information. DeepDTA (Öztürk et al., 2018) pioneered deep learning for DTA using CNNs on SMILES and protein sequences. GraphDTA (Nguyen et al., 2021) improved upon this with graph neural networks for molecular representation. These methods typically do not produce ranked compound libraries suitable for high-throughput screening.

Contrastive Learning for Virtual Screening. Recent work has applied contrastive learning to learn joint protein-ligand representations. DrugCLIP (Gao et al., 2023) adapts CLIP-style training with Uni-Mol encoders, DrugHash (Han et al., 2025) introduces efficient molecular hashing for large-scale screening, LigUnity (Feng et al., 2025) improves pocket representation with multi-scale features, and HypSeek (Wang et al., 2026) combines hyperbolic geometry with multi-modal encoders including ESM-2 for protein sequences. Our work complements these approaches by providing a plug-in module that addresses spurious correlations in the learned representations.

Causal Representation Learning. Recent theoretical advances (Schölkopf et al., 2021; Hyvarinen et al., 2019) have established conditions under which latent causal variables can be identified from observational data (Li et al., 2024a;b). Multimodal settings provide additional identifiability guarantees (Daunhawer et al., 2023; Sun et al., 2025a; Li et al., 2025), and sparse graphical structures enable component-wise identification (Kivva et al., 2021). The ConceptAligner framework (Xie et al., 2025) operationalized these insights for vision-language models. We adapt this framework to the protein-ligand domain.

D Identifiability Theory: Formal Statement

This appendix provides the formal identifiability theorem referenced in Section 3.1. The theorem establishes conditions under which molecular and pocket concepts can be recovered from observations under our V-structure (collider) model.

D.1 Data-Generating Process

We formalize the assumed data-generating process based on the V-structure described in Section 3.1. Let $\mathbf{c}^M = [c_1^M, \ldots, c_K^M]$ denote the latent molecular concepts and $\mathbf{c}^P = [c_1^P, \ldots, c_K^P]$ denote

the latent pocket concepts. The observed molecule m and pocket p are generated through invertible functions:

$$m = g^M(\mathbf{c}^M), \quad p = g^P(\mathbf{c}^P). \tag{9}$$

Crucially, $\mathbf{c}^M$ and $\mathbf{c}^P$ are **marginally independent**—neither generates the other. Instead, they jointly determine the binding indicator B through a sparse interaction structure:

$$B \sim p(B|\mathbf{c}^M, \mathbf{c}^P), \quad \text{where } B = \mathbf{1}\left[\sum_{i,j} \mathcal{M}_{ij} \cdot f(c_i^M, c_j^P) > \tau\right]. \tag{10}$$

Here $\mathcal{M}_{ij} \in \{0, 1\}$ is a sparse mask indicating whether concept pair (i, j) contributes to binding, and $f(\cdot, \cdot)$ measures compatibility. This corresponds to the V-structure: $\mathbf{c}^M \to B \leftarrow \mathbf{c}^P$.

Selection effect (Collider Bias): While $\mathbf{c}^M$ and $\mathbf{c}^P$ are *marginally independent*, they become *conditionally dependent given the binding indicator* B. This is the classic "explaining away" effect in causal inference: knowing that binding occurred ($B = 1$) and observing one set of concepts provides information about the other. In our training, we only observe positive binding pairs ($B = 1$), so this conditional dependence is what the model learns to capture. The sparse structure of $\mathcal{M}$ determines which concept pairs become correlated under this selection.

D.2 Identifiability Theorem

Theorem 1 (Concept Identifiability under Selection)**.** *Assume the following conditions hold:*

- *(i)* ***Invertibility****: The generating functions g^M and g^P are invertible and smooth.*
- *(ii)* ***Marginal Independence****: The molecular and pocket concepts are marginally independent: $p(\mathbf{c}^M, \mathbf{c}^P) = p(\mathbf{c}^M)p(\mathbf{c}^P)$.*
- *(iii)* ***Sufficient Variability****: The training data contains diverse molecule-pocket pairs such that different concept dimensions are activated across samples.*
- *(iv)* ***Sparse Interaction****: The binding mask $\mathcal{M}$ is sparse, with each molecular concept c_i^M interacting with a distinct (non-subset) set of pocket concepts.*

Then the molecular and pocket concepts $\mathbf{c}^M$ and $\mathbf{c}^P$ are ***component-wise identifiable*** *up to permutation and invertible element-wise transformations.*

D.3 Discussion of Conditions

Condition (i) - Invertibility. This requires that concepts uniquely determine observations. Neural encoders with sufficient capacity can approximate invertible mappings, satisfying this condition by design.

Condition (ii) - Marginal Independence. Unlike ConceptAligner (Xie et al., 2025) where text causally generates image features, our V-structure assumes molecular and pocket concepts are *marginally independent*: $p(\mathbf{c}^M, \mathbf{c}^P) = p(\mathbf{c}^M)p(\mathbf{c}^P)$. However, *given the binding indicator*, they become conditionally dependent: $p(\mathbf{c}^M, \mathbf{c}^P|B) \neq p(\mathbf{c}^M|B)p(\mathbf{c}^P|B)$. This conditional dependence arises from the collider structure and is exactly what enables learning: by observing which molecule-pocket pairs bind, the model discovers the sparse interaction patterns encoded in $\mathcal{M}$. Biologically, this reflects that molecules and pockets exist independently, but their compatibility for binding creates structured correlations in the observed data.

Condition (iii) - Sufficient Variability. Training on diverse molecule-pocket pairs from datasets like DUD-E and LIT-PCBA (spanning different protein families and chemical scaffolds) ensures that different concept dimensions are activated across samples, satisfying this condition.

Condition (iv) - Sparse Interaction. This is the **key condition we explicitly enforce** through L_1 regularization: $\lambda_{\text{sparse}}\|\sigma(\mathcal{M})\|_1$. The sparse mask $\mathcal{M}$ ensures that each molecular concept interacts with only a subset of pocket concepts, and these subsets are distinguishable across different molecular concepts. This sparsity is essential for identifiability: it allows us to disentangle which concept pairs are causally relevant for binding versus spuriously correlated due to dataset biases.

E PROOF OF THEOREM 1

We provide a proof sketch adapted from the framework in Xie et al. (2025) and Kivva et al. (2021), modified for our V-structure (collider) setting.

Step 1: Selection-Induced Dependence. Under our V-structure $\mathbf{c}^M \to B \leftarrow \mathbf{c}^P$, conditioning on binding ($B = 1$) induces a dependence between molecular and pocket concepts. Let $p(\mathbf{c}^M, \mathbf{c}^P | B = 1)$ denote the distribution of concepts in binding pairs. By Bayes' rule:

$$p(\mathbf{c}^M, \mathbf{c}^P | B = 1) \propto p(B = 1 | \mathbf{c}^M, \mathbf{c}^P) \cdot p(\mathbf{c}^M) \cdot p(\mathbf{c}^P). \tag{11}$$

The binding probability $p(B = 1 | \mathbf{c}^M, \mathbf{c}^P)$ depends on concept pairs through the sparse mask $\mathcal{M}$, inducing structured correlations.

Step 2: Identifiability via Sparse Structure. The sparse interaction structure (condition iv) ensures that different molecular concepts contribute to binding through distinct patterns of pocket concept interactions. Formally, if $\mathcal{M}_{i\cdot}$ denotes the i-th row of the mask (interactions of c_i^M), condition (iv) requires $\mathcal{M}_{i\cdot} \not\subseteq \mathcal{M}_{j\cdot}$ for $i \neq j$. Combined with sufficient variability (condition iii), this allows us to disentangle concepts through the induced correlation structure (Kivva et al., 2021).

Connection to *SparseMask*. The L_1 regularization $\|\sigma(\mathcal{M})\|_1$ directly enforces sparse interactions. By learning which concept pairs contribute to binding score $\mathbf{z}_p^\top \sigma(\mathcal{M}) \mathbf{z}_m$, the model identifies causally relevant features. Dimensions with near-zero mask values correspond to spurious correlations that do not contribute to binding.

F IMPLEMENTATION DETAILS

For the *ConceptExtractor*, we use $K = 64$ concepts with dimension $d_c = 256$ and 4 Perceiver layers with 8 attention heads. The sparsity weight $\lambda_{\text{sparse}} = 0.001$. All models are trained for 50 epochs with Adam optimizer and learning rate 10^{-4}.

For DrugCLIP experiments, we use the same encoder architecture as the original paper but without pretrained weights. For HypSeek experiments, we use the pretrained Uni-Mol and ESM-2 encoders and follow the three-stage training pipeline described in the original work.

G EXPERIMENTAL ANALYSIS

Why does causal concept extraction help? We hypothesize that the sparse masking mechanism helps identify binding-relevant features while suppressing spurious correlations. In the scratch training setting (Experiment 1), where the model has no prior knowledge, the causal module provides a strong inductive bias for learning disentangled representations. In the pretrained setting (Experiment 2), the module helps refine already-learned representations by identifying which features are most relevant for binding prediction.

Early enrichment improvements. Across both experiments, we observe that improvements in BEDROC and EF@1% are often larger than AUC improvements. This is particularly valuable for drug discovery, where identifying true actives among the top-ranked compounds is more important than overall ranking quality. The sparse masking mechanism appears to especially help with identifying high-confidence predictions.

Generalization to realistic benchmarks. The larger improvements on LIT-PCBA compared to DUD-E across both experimental settings suggest that our causal concept extraction is particularly effective when facing more realistic decoy distributions. LIT-PCBA's decoys are specifically designed to minimize property-based biases, making it harder for models to exploit spurious correlations. Our module's consistent gains on this challenging benchmark indicate that the learned sparse masks successfully filter out non-causal features.

H LIMITATIONS AND BROADER IMPACT

Limitations. Our current evaluation is limited to two benchmark datasets (DUD-E and LIT-PCBA). The effectiveness of causal concept extraction may vary depending on the specific characteristics of the target proteins and compound libraries. Additional benchmarks such as BindingDB (Gilson et al., 2016) and ChEMBL could provide further validation. Future work should also explore interpretability of the learned concepts and their correspondence to known binding determinants.

Future Directions. To extend this work for a full conference submission, several directions warrant further investigation: (1) **Comprehensive benchmarking**: Evaluation on additional datasets such as BindingDB, ChEMBL, and cross-dataset generalization tests to validate robustness across diverse chemical spaces. (2) **Ablation studies**: Systematic analysis of the number of concepts K, sparsity weight λ_{sparse}, and the relative contributions of ConceptExtractor vs. SparseMask components. (3) **Concept interpretability**: Visualization and analysis of learned concepts to verify whether they correspond to known binding determinants such as hydrogen bonding, hydrophobic interactions, and electrostatic complementarity. (4) **Integration with molecular generation**: Exploring how identified causal concepts can guide structure-based drug design and lead optimization. (5) **Computational efficiency**: Profiling the overhead introduced by our module and potential optimizations for large-scale industrial screening campaigns.

Broader Impact. Improved virtual screening methods can accelerate drug discovery, potentially benefiting patients by reducing the time and cost of developing new treatments. However, as with all computational drug discovery tools, predictions should be validated experimentally before clinical use.