CAUSALBIND: Causal Modeling and Learning for Protein-Molecule 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

Protein-molecule virtual screening is increasingly cast as a problem of representation learning in a shared embedding space. Existing methods rely on dense holistic alignment, entangling invariant binding determinants with nuisance correlations and limiting transfer to new targets. It has been noted that binding in protein–molecule systems involves sparse cross-modality interactions: binding is governed by a small contact interface and a few decisive local interactions (e.g., hydrogen bonds, hydrophobic contacts, and salt bridges) rather than the global structures of the protein and molecule. We hypothesize that uncovering and leveraging sparse interaction patterns is critical for generalization beyond the training data, as these patterns are reusable and expected to improve performance across different scenarios. In this paper, we aim to identify and leverage sparse interaction patterns, and verify our hypothesis. Since the training data contain only observed binding pairs, we formalize this prior via a V-structure causal model under Heckman-style selection, and establish three theoretical results: (i) when the compatibility between concepts is bilinear, the latent concepts of interacting proteins and molecules are not identifiable, even if the interaction is sparse; (ii) with a non-bilinear compatibility function, these concepts and their sparse interaction support are component-wise identifiable under a sparse antichain condition; and (iii) a low-rank relaxation yields subspace identifiability of the interaction modes, and applies to bilinear compatibility as well. Inspired by these principles, we propose CAUSALBIND with three implementation variants. Extensive experiments on DUD-E and LIT-PCBA benchmarks show that all variants consistently outperform strong retrieval baselines, with the largest gains on LIT-PCBA early enrichment, and further generalize to target- and scaffold-level out-of-distribution splits. Code and scripts are freely available at `https://github.com/lokali/CausalBind`.

1 Introduction

Modern drug discovery faces a fundamental challenge: identifying promising drug candidates from libraries containing millions to billions of candidate compounds [Berdigaliyev and Aljofan, 2020]. Virtual screening [Walters and Wang, 2020], which prioritizes protein-molecule pairs before costly experimental validation, has become a major bottleneck in this pipeline and a central problem in AI for drug discovery [Jia et al., 2026]. An overview of the broader pipeline is illustrated in App. A1.

Existing virtual-screening methods broadly fall into three regimes: physics-based docking systems [Friesner et al., 2004, Trott and Olson, 2010] that score complexes via force fields; structure-aware machine-learning scoring models [McNutt et al., 2021, Shen et al., 2022, 2023, Cao et al., 2024, Moon et al., 2022, Lu et al., 2022, Brocidiacono et al., 2023] that learn to rank protein-molecule complexes from data; and retrieval-based representation learners [Gao et al., 2023, Han et al., 2025, Feng et al., 2025, Wang et al., 2026, He et al., 2026] that embed both modalities into a shared

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

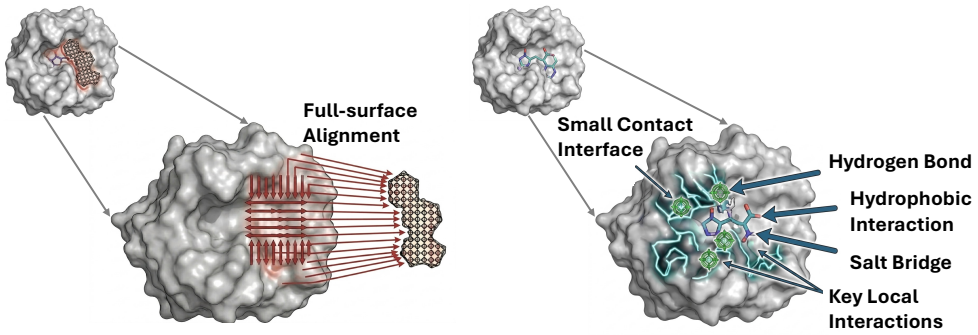


Figure 1: **Dense alignment versus sparse binding. (a) Holistic dense alignment.** Existing retrieval-based virtual screening methods compare every protein surface feature with every molecule feature in the shared embedding space; this "full-surface" alignment entangles binding-relevant signals with nuisance correlations. **(b) Biology-inspired sparse binding.** Protein-molecule binding is governed by a *small contact interface* and a handful of *key local interactions* (e.g., hydrogen bonds, hydrophobic contacts, salt bridges, and π-stacking), which together account for the majority of the binding affinity.

space. The first two reason explicitly about three-dimensional poses and are valuable for mechanistic analysis, but they are typically better suited to rescoring a small candidate set than to ultra-large-scale retrieval. Retrieval-based methods, by contrast, embed both molecules and proteins into a shared embedding space and screen by similarity, scaling naturally to libraries with millions of compounds and now consistently dominating the empirical leaderboard. See App. A2 for extended related work.

Despite this success, retrieval models still learn holistic representations: they align full molecule and protein embeddings without separating binding-relevant determinants from nuisance factors. This reflects a structural mismatch with the biological process, as illustrated in Figure 1, where the binding decision is governed by a small contact interface and a few decisive local interactions rather than the global structures of the protein and molecule. A classical example is the notion of "hot spots", where a small subset of interfacial contacts contributes the majority of binding free energy [Clackson and Wells, 1995, Bogan and Thorn, 1998]. As a result, holistic models are prone to capturing spurious global correlations rather than the local structural features that actually drive binding [Chen et al., 2019, Graber et al., 2025], which may hinder generalization to novel protein families and unseen chemical scaffolds. Additional details about biological observations are provided in App. A3.

These limitations motivate the core question to be answered in this paper: if protein-molecule binding is driven by a small number of local compatibilities, should one represent virtual-screening in terms of the sparse concept-level interactions, instead of potentially dense holistic alignment, and leverage this representation for better generalization ability? Our results strongly suggest that the answer is yes. To analyze this issue, we first develop a V-structure causal model of binding under Heckman-style selection [Schölkopf et al., 2021, Zhang et al., 2026], and establish three identifiability results for the sparse protein-molecule interaction: dense cross-modal interaction is generally non-identifiable, sparse antichain interaction is component-wise identifiable up to trivial equivalences, and under a low-rank condition the interaction modes are subspace-identifiable. Building on this theory, we propose CAUSALBIND as a flexible structure-constrained representation framework with three different implementations: the main CAUSALBIND$_{\mathrm{SP}}$ uses a *concept extractor* and a *learnable sparse concept-pair mask*, CAUSALBIND$_{\mathrm{LR}}$ uses a low-rank factorization of the same concept-axis mask, and CAUSALBIND$_{\mathrm{EMB}}$ applies the constrained mask after pooling in embedding space.

Our main technical contributions are threefold. (i) Theoretically, we provide an identifiability analysis of representation learning for protein-molecule binding under binder-only selection, showing that bilinear interactions are not identifiable even when they are sparse, whereas sparse antichain-structured interactions with a non-bilinear compatibility function are component-wise identifiable, and low-rank interactions are subspace-identifiable. (ii) Methodologically, we instantiate these principles through three CAUSALBIND implementations within a concept-based retrieval framework. CAUSALBIND$_{\mathrm{SP}}$ routes binding evidence through a learnable sparse concept-pair mask, CAUSALBIND$_{\mathrm{LR}}$ imposes an explicit low-rank concept-axis factorization, and CAUSALBIND$_{\mathrm{EMB}}$ applies the constraint in pooled embedding space. (iii) Empirically, on DUD-E and LIT-PCBA benchmarks, all three CAUSALBIND implementations outperform reproduced baseline methods, with the largest relative gains on the

property-debiased LIT-PCBA early enrichment. Additional analyses, including multi-seed ablations, hyperparameter studies, antichain diagnostics of the masks, out-of-distribution evaluation on DEKOIS 2.0, an interpretability case study, and affinity-ranking results, all support the efficacy of our method.

2 Theoretical Analysis: Sparse Binding under Heckman Selection

This section develops a causal perspective on protein-molecule binding under binder-only observation. In § 2.1, we model binding as a V-structure and study identifiability [Daunhawer et al., 2023] under Heckman-style selection [Heckman, 1979]. Our theorems in § 2.2 proceed in three phases, from non-identifiability to component-wise and subspace identifiability.

2.1 Causal Model: Binding as a Selected Collider

The causal model combines the biological sparsity motivation above with the binder-only nature of training data. In practice, we predominantly observe pairs with $B{=}1$; conditioning on this collider can make molecule and protein features statistically dependent even when they are marginally independent. Figure 2 summarizes this collider structure of sparse interactions.

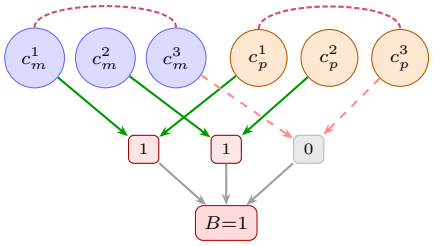


Figure 2: Binder-selected sparse-interaction view. Observing only binders ($B{=}1$) induces cross-modal dependence through the collider B. Green edges denote active local interactions. Red dashed edges mark blocked interactions. Purple dashed curves mark allowed within-modality dependence.

Let $x^M \in \mathcal{X}^M$ and $x^P \in \mathcal{X}^P$ denote observed molecule and protein, generated from latent concepts

$$x^M = g^M(\mathbf{c}^M, \varepsilon^M), \qquad x^P = g^P(\mathbf{c}^P, \varepsilon^P), \quad (1)$$

where $\mathbf{c}^M = (c_1^M, \ldots, c_{K_M}^M)$ and $\mathbf{c}^P = (c_1^P, \ldots, c_{K_P}^P)$ are modality-specific atomic features, and $\varepsilon^M, \varepsilon^P$ denote modality-specific observation noise or nuisance variation. Here K_M and K_P count the number of concepts, respectively. Binding arises through cross-modal interactions. Motivated by existing literature [Wang et al., 2026, Feng et al., 2025], we model binding as a Bernoulli collider:

$$B \sim \text{Bernoulli}\left(\sigma\left(\sum_{i=1}^{K_M}\sum_{j=1}^{K_P} \mathcal{M}_{ij}\,\phi(c_i^M, c_j^P)\right)\right), \quad x^M \leftarrow \mathbf{c}^M \rightarrow B \leftarrow \mathbf{c}^P \rightarrow x^P, \quad (2)$$

where $\mathcal{M} \in \mathbb{R}_{\geq 0}^{K_M \times K_P}$ is a (typically sparse) interaction mask; its scale is identifiable only jointly with ϕ. In the theory, each atomic concept is a scalar latent variable, i.e., $c_i^M, c_j^P \in \mathbb{R}$, and $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ maps one molecule concept and one protein concept to a scalar logit-level compatibility score; $\sigma(\cdot)$ is the logistic sigmoid. The nuisance variables affect only the observed modalities and are marginalized in $p(x^M, x^P \mid B{=}1)$; the binding variable depends on the recovered concepts. Before selection, $\mathbf{c}^M \perp\!\!\!\perp \mathbf{c}^P$. After conditioning on $B{=}1$, the two modalities become dependent. Here we denote the distribution of the observed training data as: $p^{\mathcal{D}}(x^M, x^P) = p(x^M, x^P \mid B = 1)$, which generally induces dependence between x^M and x^P. Furthermore, by Bayes' rule, we have

$$p^{\mathcal{D}}(\mathbf{c}^M, \mathbf{c}^P) \;\propto\; p(\mathbf{c}^M)\, p(\mathbf{c}^P)\, \sigma\Big(\textstyle\sum_{i=1}^{K_M}\sum_{j=1}^{K_P} \mathcal{M}_{ij}\phi(c_i^M, c_j^P)\Big). \quad (3)$$

This induced dependence is the only signal available for recovering $(\mathcal{M}, \phi)$. Below we give the definition of identifiability used in this paper, followed by some common assumptions.

Definition 1 (Identifiability)**.** *Identifiability is concerned with whether the latent concepts and interaction structure are uniquely determined by the observed binder-selected distribution $p^{\mathcal{D}}(x^M, x^P)$, modulo the unavoidable equivalences of the chosen representation. In the theory, the concept dimensions K_M and K_P are treated as given model dimensions. We use two levels of identifiability.*

1. ***Component-wise identifiability.*** *The concepts and sparse interaction support are component-wise identifiable if any two parameterizations $(g^M, g^P, \mathcal{M}, \phi)$ and $(\tilde{g}^M, \tilde{g}^P, \tilde{\mathcal{M}}, \tilde{\phi})$ inducing the same $p^{\mathcal{D}}(x^M, x^P)$ differ only by separate molecule-side and protein-side concept-index permutations, component-wise invertible transformations of individual concepts, and the trivial joint rescaling $(\mathcal{M}, \phi) \mapsto (a\mathcal{M}, a^{-1}\phi)$ for $a > 0$.*

2. ***Subspace identifiability.*** *Suppose the interaction kernel has a minimal rank-r factorization $S(\mathbf{c}^M, \mathbf{c}^P) = \mathbf{u}(\mathbf{c}^M)^\top \mathbf{v}(\mathbf{c}^P)$. The interaction is subspace-identifiable if every other minimal rank-r factorization $\tilde{\mathbf{u}}^\top \tilde{\mathbf{v}}$ inducing the same $p^{\mathcal{D}}(x^M, x^P)$ satisfies*

$$\mathrm{span}\{\tilde{u}_1, \ldots, \tilde{u}_r\} = \mathrm{span}\{u_1, \ldots, u_r\}, \qquad \mathrm{span}\{\tilde{v}_1, \ldots, \tilde{v}_r\} = \mathrm{span}\{v_1, \ldots, v_r\}.$$

Equivalently, the ordered mode basis may change by an invertible transformation in $\mathrm{GL}(r)$, but the molecule-side and protein-side interaction subspaces are fixed.

Assumption 1 (Concept recoverability)**.** *The latent concepts $\mathbf{c}^M$ and $\mathbf{c}^P$ can be written as functions of the corresponding observations $\mathbf{x}^M$ and $\mathbf{x}^P$, respectively.*

Assumption 2 (Prior independence)**.** *Before conditioning on B, the two modalities are independent, i.e., $p(\mathbf{c}^M, \mathbf{c}^P) = p(\mathbf{c}^M)p(\mathbf{c}^P)$.*

Assumption 3 (Positivity / coverage)**.** *We have $P(B{=}1) > 0$, the supports of $p(\mathbf{c}^M)$ and $p(\mathbf{c}^P)$ are open sets, and every atomic concept appears with positive probability under the binder-selected distribution $p^{\mathcal{D}}$.*

Assumption 4 (Log-linear selection)**.** *On the data support, the binding probability is an exponential tilt of the binding score: $P(B{=}1 \mid \mathbf{c}^M, \mathbf{c}^P) = \kappa \exp(S(\mathbf{c}^M, \mathbf{c}^P))$ for a constant $\kappa > 0$, where $S(\mathbf{c}^M, \mathbf{c}^P) := \sum_{i,j} \mathcal{M}_{ij} \phi(c_i^M, c_j^P)$.*

In Definition 1, component-wise identifiability is stronger because it recovers each concept, whereas subspace identifiability recovers only a low-dimensional subspace spanned by several concepts. Assumptions 1–3 are shared by all theorems. They ensure that (i) concept information is preserved from the observed modalities, (ii) cross-modal dependence arises only through selection, and (iii) sufficient coverage is available for recovery, respectively. Assumption 4, used by the two positive results, is the low-hit-rate form of the logistic model, under which our contrastive objective is well specified [Oord et al., 2018]. Each of the four assumptions is discussed in more detail in App. A4.2.

2.2 Identifiability Results

We next characterize, in a progressively structured way, when the latent interaction mechanism is identifiable from binder-selected data. First, binder-only selection by itself is insufficient: without suitable structural constraints, interacting protein and molecule concepts are not identifiable. Second, such concepts and sparse interactions are component-wise identifiable under proper structural sparsity conditions. Third, low-rank interaction structure yields subspace identifiability when only low-dimensional interaction subspaces can be determined. We give the three formal statements below.

Theorem 1 (Non-identifiability of bilinear interactions)**.** *Under the causal model in Section 2.1, suppose Assumptions 1–3 hold and the compatibility function is bilinear, $\phi(x, y) = xy$. Then for any training distribution $p^{\mathcal{D}}(x^M, x^P)$ there exist infinitely many distinct tuples $(g^M, g^P, \mathcal{M}, \phi)$ inducing the same $p^{\mathcal{D}}(x^M, x^P)$. Moreover, if $K_M \geq 2$ and $\mathcal{M} \neq 0$, some of these tuples keep the same mask, $\widetilde{\mathcal{M}} = \mathcal{M}$, while their concepts mix at least two of the original concepts. Hence the concepts are* non-identifiable *here, and constraining the mask, e.g. to be sparse with antichain supports, does not remove the ambiguity.*

Remark 1. Theorem 1 gives the baseline impossibility result: conditioning on $B{=}1$ induces cross-modal concept dependence, but this dependence alone does not determine which concepts or mask generated it. With a bilinear ϕ, infinitely many parameterizations explain the same distribution, and the second part shows this cannot be removed by constraining the mask. Theorems 2 and 3 specify what additional structure makes identification possible. The complete proof is provided in App. A4.4.

Theorem 2 (Component-wise identifiability under sparse interaction)**.** *Suppose Assumptions 1–4 hold, and let $(\tilde{g}^M, \tilde{g}^P, \widetilde{\mathcal{M}}, \tilde{\phi})$, with $\tilde{\phi} \in C^2$, be any other such tuple inducing the same $p^{\mathcal{D}}(x^M, x^P)$, with concepts $\tilde{\mathbf{c}}^M, \tilde{\mathbf{c}}^P$. Suppose further:*

2.i (Sparse antichain interaction) Letting

$$R_i := \mathrm{supp}(\mathcal{M}_{i,\cdot}) = \{j \in [K_P] : \mathcal{M}_{ij} \neq 0\},$$
$$C_j := \mathrm{supp}(\mathcal{M}_{\cdot,j}) = \{i \in [K_M] : \mathcal{M}_{ij} \neq 0\},$$

every row support and every column support is nonempty, and the row supports $\{R_i\}_{i=1}^{K_M}$ and column supports $\{C_j\}_{j=1}^{K_P}$ form antichains:

$$R_i \not\subseteq R_{i'} \quad \forall i \neq i', \qquad C_j \not\subseteq C_{j'} \quad \forall j \neq j'.$$

2.ii (Non-bilinear compatibility) ϕ is real-analytic on $\mathbb{R}^2$ and $\partial_1^2\partial_2^2\phi \not\equiv 0$; equivalently, the cross-derivative $\partial_1\partial_2\phi(x,y)$ is not of the form $\alpha(x)+\beta(y)$.

2.iii (Linear ambiguity and sparsity minimality) The two concept systems are related by invertible affine maps, $\tilde{\mathbf{c}}^M = T_M\mathbf{c}^M + \mathbf{t}_M$ and $\tilde{\mathbf{c}}^P = T_P\mathbf{c}^P + \mathbf{t}_P$ with $T_M \in \mathrm{GL}(K_M)$, $T_P \in \mathrm{GL}(K_P)$; and $\|\widetilde{\mathcal{M}}\|_0 \leq \|\mathcal{M}\|_0$.

Then there exist permutations π of $[K_M]$ and ς of $[K_P]$ such that each $\tilde{c}^M_{\pi(i)}$ is an affine function of c_i^M alone, each $\tilde{c}^P_{\varsigma(j)}$ is an affine function of c_j^P alone, and $\mathrm{supp}(\widetilde{\mathcal{M}}) = \{(\pi(i),\varsigma(j)) : (i,j) \in \mathrm{supp}(\mathcal{M})\}$; the concepts and sparse support of $\mathcal{M}$ are thus component-wise identifiable (Definition 1).

Remark 2. Theorem 2 resolves that ambiguity, and shows which ingredient does the work. Condition 2.iii restricts the competitors to linear reparameterizations, what contrastive learning already provides [Roeder et al., 2021]. Condition 2.ii makes $\mathcal{M}_{ij}\partial_1\partial_2\phi$ vary over the support, so matching the two distributions yields a *family* of matrix equations rather than the single one a bilinear ϕ gives. Condition 2.i gives each concept a distinct combinatorial footprint, and the minimality clause is the population counterpart of the sparsity penalty. App. A4.5 has the full proof and the counterexamples.

Theorem 3 (Subspace identifiability under low-rank interaction)**.** *Under the causal model above, if the observed binder-selected distribution matches the true joint distribution $p^{\mathcal{D}}(x^M, x^P)$, and Assumptions 1–4 hold, and:*

3.i (Low-rank interaction kernel) Let S be the binding score of Assumption 4, fix reference values $\mathbf{a}_0, \mathbf{b}_0$ in the supports of $\mathbf{c}^M, \mathbf{c}^P$, and let $S^\circ(\mathbf{c}^M,\mathbf{c}^P) := S(\mathbf{c}^M,\mathbf{c}^P) - S(\mathbf{c}^M,\mathbf{b}_0) - S(\mathbf{a}_0,\mathbf{c}^P) + S(\mathbf{a}_0,\mathbf{b}_0)$ be the centred kernel. There exists a minimal rank $r < \min\{K_M, K_P\}$ and linearly independent mode functions $u_1,\dots,u_r \in L^2(p^{\mathcal{D}}(\mathbf{c}^M))$ and $v_1,\dots,v_r \in L^2(p^{\mathcal{D}}(\mathbf{c}^P))$ such that

$$S^\circ(\mathbf{c}^M,\mathbf{c}^P) = \sum_{\ell=1}^{r} u_\ell(\mathbf{c}^M)v_\ell(\mathbf{c}^P) = \mathbf{u}(\mathbf{c}^M)^\top\mathbf{v}(\mathbf{c}^P),$$

where $\mathbf{u} = (u_1,\dots,u_r)^\top$ and $\mathbf{v} = (v_1,\dots,v_r)^\top$, and no decomposition with fewer than r separable terms exists.

3.ii (Low-rank sufficient variability) There exist probe values $\mathbf{a}_1,\dots,\mathbf{a}_{R_M} \in \mathrm{int}(\mathrm{supp}(\mathbf{c}^M))$ and $\mathbf{b}_1,\dots,\mathbf{b}_{R_P} \in \mathrm{int}(\mathrm{supp}(\mathbf{c}^P))$, with $R_M, R_P \geq r$, such that the mode-evaluation matrices

$$\mathbf{U}_A := \begin{bmatrix} u_1(\mathbf{a}_1) & \cdots & u_r(\mathbf{a}_1) \\ \vdots & \ddots & \vdots \\ u_1(\mathbf{a}_{R_M}) & \cdots & u_r(\mathbf{a}_{R_M}) \end{bmatrix}, \qquad \mathbf{V}_B := \begin{bmatrix} v_1(\mathbf{b}_1) & \cdots & v_r(\mathbf{b}_1) \\ \vdots & \ddots & \vdots \\ v_1(\mathbf{b}_{R_P}) & \cdots & v_r(\mathbf{b}_{R_P}) \end{bmatrix}$$

$$\mathrm{rank}(\mathbf{U}_A) = \mathrm{rank}(\mathbf{V}_B) = r.$$

Thus both modality-side mode families can be independently excited across the selected concept support.

Then the interaction subspaces

$$\mathcal{U}_M := \mathrm{span}\{u_1,\dots,u_r\} \subset L^2(p^{\mathcal{D}}(\mathbf{c}^M)), \qquad \mathcal{V}_P := \mathrm{span}\{v_1,\dots,v_r\} \subset L^2(p^{\mathcal{D}}(\mathbf{c}^P))$$

are subspace-identifiable from $p^{\mathcal{D}}(x^M, x^P)$ in the sense of Definition 1.

Remark 3. Under Condition 3.i the interaction kernel is low-rank, so binding is determined by rank-r interaction subspaces rather than by separately identifiable concept pairs. The kernel is centred because the selected distribution determines S only up to additive functions of a single modality. Unlike Theorem 2, this result does *not* need a non-bilinear ϕ: for $\phi(x,y) = xy$ the kernel has rank $\mathrm{rank}(\mathcal{M})$, so a low-rank mask suffices, the regime our implementation occupies. Proof in App. A4.6.

From theory to implementation. Identifiability matters here because these interaction patterns are reusable and expected to improve generalizability across different scenarios. These results pin down the inductive bias that our method should approximate: (i) concept encoders rich enough to preserve local binding factors, (ii) sparse interaction support or low-rank interaction subspaces, and (iii) exploitation of the collider-induced dependence under $B{=}1$. The next section will translate the proposed theoretical model into concrete neural architectures, with different implementations. The causal formulation enters the implementation through the choice and regularization of an identifiable representation class, rather than through a separate intervention-, propensity-, or counterfactual-based training loss: CAUSALBIND retains the contrastive/ranking supervision of retrieval-based methods, and replaces dense holistic interaction with the sparse-antichain or low-rank parameterizations identified above, together with the corresponding structural regularizer. Accordingly, the empirical gains arise from this theory-derived structural inductive bias, while the causal analysis establishes when these interaction classes are identifiable under binder-only selection; we do not claim that CAUSALBIND estimates interventional effects of individual concepts. Since our ϕ is an inner product, hence bilinear, the guarantee that applies is Theorem 3 rather than Theorem 2 (App. A6.8).

3 CAUSALBIND: Implementation

We implement the sparse causal principle of § 2. We develop three variants: CAUSALBIND$_{\mathrm{SP}}$, CAUSALBIND$_{\mathrm{LR}}$, and CAUSALBIND$_{\mathrm{EMB}}$. Due to space, we focus on CAUSALBIND$_{\mathrm{SP}}$, the sparse concept-pair implementation; supplements for CAUSALBIND$_{\mathrm{SP}}$, CAUSALBIND$_{\mathrm{LR}}$, and CAUSALBIND$_{\mathrm{EMB}}$ are given in App. A5.2, App. A5.3, and App. A5.4. Inspired by HypSeek [Wang et al., 2026], we use three modalities: molecule structure, pocket structure, and protein sequence, where the sequence branch provides complementary global context beyond local pocket geometry. Instead of aligning holistic embeddings, we first decompose each modality into atomic concepts, and then route protein-molecule matching through sparse concept-level interaction operators (see Figure 3).

Notation. Let $\mathbf{H}_{\mathrm{mol}} \in \mathbb{R}^{N_{\mathrm{mol}} \times d_{\mathrm{mol}}}$, $\mathbf{H}_{\mathrm{poc}} \in \mathbb{R}^{N_{\mathrm{poc}} \times d_{\mathrm{poc}}}$, and $\mathbf{H}_{\mathrm{seq}} \in \mathbb{R}^{N_{\mathrm{seq}} \times d_{\mathrm{seq}}}$ denote token representations from Uni-Mol [Zhou et al., 2023] (molecule and pocket) and ESM2 [Lin et al., 2023] (sequence). These project to a per-concept dimension d_c. The theory in § 2 uses scalar concepts, corresponding to the special case $d_c = 1$. In implementation, we allow vector-valued concept tokens with $d_c \geq 1$: for each modality $v \in \{\mathrm{mol}, \mathrm{poc}, \mathrm{seq}\}$, $\mathbf{C}_v \in \mathbb{R}^{K \times d_c}$ and the i-th row $\mathbf{C}_v[i,:]$ represents one concept group. The theory allows different concept counts across modalities (K_M and K_P); for simplicity, we use a common concept budget K for molecule, pocket, and sequence streams. To align notation with the theoretical mask $\mathcal{M}$, we denote the two implementation masks by $\mathcal{M}_{\mathrm{mol,poc}}$ and $\mathcal{M}_{\mathrm{mol,seq}}$. Their raw learnable parameters are $\widehat{\mathcal{M}}_{\mathrm{mol,poc}}$ and $\widehat{\mathcal{M}}_{\mathrm{mol,seq}}$, and the effective sparse masks $\mathcal{M}_{\mathrm{mol,poc}}, \mathcal{M}_{\mathrm{mol,seq}} \in \mathbb{R}^{K \times K}$ are obtained after activation and thresholding.

Concept Extractor. Each modality is decomposed into K latent concepts using a Perceiver-style cross-attention block [Jaegle et al., 2021]:

$$\begin{aligned} \widetilde{\mathbf{C}}_v^{(\ell+1)} &= \mathbf{C}_v^{(\ell)} + \mathrm{MHA}\big(\mathrm{LN}(\mathbf{C}_v^{(\ell)}),\, \mathrm{LN}(\widehat{\mathbf{H}}_v),\, \mathrm{LN}(\widehat{\mathbf{H}}_v)\big), \\ \mathbf{C}_v^{(\ell+1)} &= \widetilde{\mathbf{C}}_v^{(\ell+1)} + \mathrm{FFN}\big(\mathrm{LN}(\widetilde{\mathbf{C}}_v^{(\ell+1)})\big), \end{aligned} \tag{4}$$

where $\mathbf{C}_v^{(0)} = \mathbf{Q}_v \in \mathbb{R}^{K \times d_c}$ are learnable concept queries and $\widehat{\mathbf{H}}_v = \mathbf{H}_v W_v \in \mathbb{R}^{N_v \times d_c}$ is the projected token sequence. After L layers, we obtain $\mathbf{C}_v \in \mathbb{R}^{K \times d_c}$, which serves as the practical counterpart of the latent concepts in § 2. If a backbone exposes only a pooled vector, it corresponds to the special case $N_v = 1$; in the main structural branches, token-level features provide nontrivial cross-attention keys and values.

Sparsity Mask. Following § 2, we model binding through sparse cross-modal interactions rather than dense alignment. For each branch $(a,b) \in \{(\mathrm{mol}, \mathrm{poc}), (\mathrm{mol}, \mathrm{seq})\}$, we compute

$$\begin{aligned} \mathbf{Z}_{a,b} &= \Gamma(\widehat{\mathcal{M}}_{a,b}), \qquad \mathcal{M}_{a,b} = H_\tau^{\mathrm{ST}}(\mathbf{Z}_{a,b})\,, \\ \mathbf{A}_{a,b} &= \phi(\mathbf{C}_a, \mathbf{C}_b), \qquad u_{a,b} = \mathrm{Agg}(\mathcal{M}_{a,b} \odot \mathbf{A}_{a,b}), \end{aligned} \tag{5}$$

where, under the common-K implementation, $\widehat{\mathcal{M}}_{a,b} \in \mathbb{R}^{K \times K}$ is an unconstrained parameter matrix and $\mathcal{M}_{a,b}$ is the effective sparse mask used for interaction. The activation $\Gamma(\widehat{\mathcal{M}}) = \tanh(\widehat{\mathcal{M}}) + 1$ first

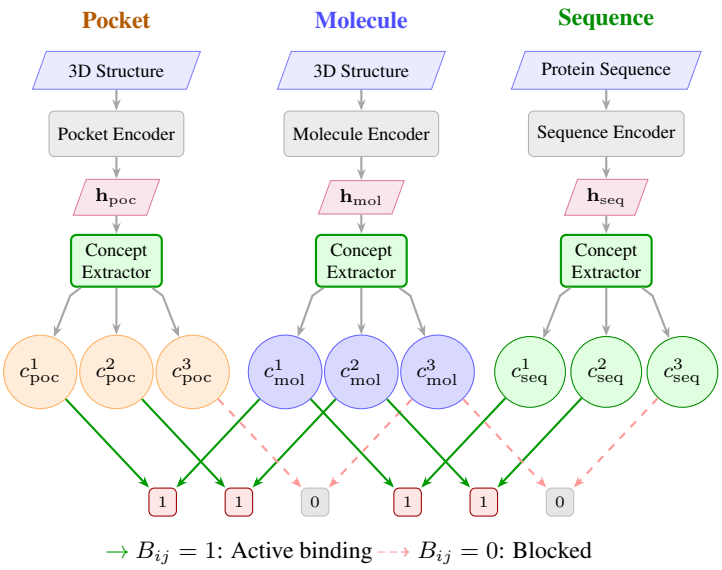


Figure 3: Implementation-level view of CAUSALBIND. Each modality-specific encoder maps its input to atomic concepts $\{c_v^k\}$, each with dimension $d_c \geq 1$. The indicators B_{ij} at the bottom visualize the branch-wise cross-modal mask $\mathcal{M}_{a,b}$ for $(a,b) \in \{(\text{mol},\text{poc}),(\text{mol},\text{seq})\}$: B_{ij}=1 when the (i,j) entry of $\mathcal{M}_{a,b}$ survives the forward threshold ($\tau = 0.5$) (green solid edges, active pair), and B_{ij}=0 otherwise (red dashed edges, blocked pair). Active pairs form the contrastive score, while $\mathcal{L}_{\text{spa}}$ encourages the interaction sparsity levels.

produces a continuous mask $\mathbf{Z}_{a,b} \in [0,2]^{K\times K}$, and the straight-through threshold operator H_τ^{ST} is applied element-wise with mask threshold $\tau = 0.5$: the forward pass uses $[H_\tau(Z)]_{ij} = Z_{ij}\mathbf{1}\{Z_{ij} > \tau\}$, while the backward pass sends gradients through the continuous Z_{ij}. Thus thresholding occurs before aggregation, and Agg receives only the thresholded masked compatibility matrix $\mathcal{M}_{a,b} \odot \mathbf{A}_{a,b}$. Each nonzero entry $\mathcal{M}_{a,b}[i,j]$ represents the strength of interaction between the i-th concept token of modality a and the j-th concept token of modality b. Here $\mathcal{M}_{a,b}$ is the implementation counterpart of the theoretical interaction mask $\mathcal{M}$ in § 2. For notational consistency, we use the same compatibility symbol ϕ: in the scalar theory $\phi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, while in the vector-valued implementation $\phi : \mathbb{R}^{d_c}\times\mathbb{R}^{d_c} \to \mathbb{R}$. Together, $\mathcal{M}_{a,b}$ and ϕ instantiate the gated interaction term $\mathcal{M}_{ij}\phi(c_i^M, c_j^P)$ at the token level. The matrix $\mathbf{A}_{a,b} \in \mathbb{R}^{K\times K}$ collects all concept-pair compatibilities, where each entry is computed by ϕ (we use a simple inner product between vector concept tokens). The operator $\odot$ denotes element-wise multiplication, and $\text{Agg}(\mathbf{B}) = K^{-1}\sum_{i,j} B_{ij}$ aggregates the retained interactions into a branch-level score in our implementation. Thus, the masks $\mathcal{M}_{\text{mol,poc}}$ and $\mathcal{M}_{\text{mol,seq}}$ act as cross-modal bottlenecks, enforcing the interaction sparsity.

Training Loss. We retain retrieval-based supervision applied after concept decomposition and sparse interaction. The scalar $u_{a,b}$ defines the pairwise score for a protein-molecule pair, and over a mini-batch of size N, these scores form matrices $\mathbf{U}^{a,b} \in \mathbb{R}^{N\times N}$. The contrastive loss is

$$\ell_{\text{cont}}(\mathbf{U}^{a,b}) = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(U_{ii}^{a,b}/\tau_{\text{c}})}{\sum_{j=1}^{N}\exp(U_{ij}^{a,b}/\tau_{\text{c}})} + \log\frac{\exp(U_{ii}^{a,b}/\tau_{\text{c}})}{\sum_{j=1}^{N}\exp(U_{ji}^{a,b}/\tau_{\text{c}})}\right], \tag{6}$$

where $\tau_{\text{c}} > 0$ is the contrastive temperature. The combined contrastive objective is

$$\mathcal{L}_{\text{cont}} = \ell_{\text{cont}}(\mathbf{U}^{\text{poc,mol}}) + \ell_{\text{cont}}(\mathbf{U}^{\text{seq,mol}}). \tag{7}$$

The theory models the binder-selected positive distribution; the off-diagonal batch entries here are retrieval negatives used by the contrastive surrogate, not additional observed B=0 labels.

We further add sparsity and geometric regularization terms [Wang et al., 2026]. The sparsity loss is the entry-wise L_1 penalty on the continuous activated masks, which preserves gradients to the raw mask parameters:

$$\mathcal{L}_{\text{spa}} = \|\mathbf{Z}_{\text{mol,poc}}\|_1 + \|\mathbf{Z}_{\text{mol,seq}}\|_1. \tag{8}$$

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cont}} + \lambda_{\text{spa}}\mathcal{L}_{\text{spa}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{9}$$

Here $\mathcal{L}_{\text{reg}}$ collects the geometric regularizers inherited from HypSeek [Wang et al., 2026].

Table 1: Virtual-screening comparison on two benchmarks between baselines and CAUSALBIND$_{\mathrm{SP}}$ (K=128, d_c=1024), CAUSALBIND$_{\mathrm{LR}}$ (K=64, r=1), and CAUSALBIND$_{\mathrm{EMB}}$ (K=64, d_c=256), and CAUSALBIND$_{\mathrm{ATOM}}$ (K=128, d_c=1024). †S^{2}Drug results are paper-reported [He et al., 2026].

Method	DUD-E (n=102)			LIT-PCBA (n=15)		
	AUROC	BEDROC$_{80.5}$	EF$_{1\%}$	AUROC	BEDROC$_{80.5}$	EF$_{1\%}$
Glide-SP [Friesner et al., 2004]	0.767	0.407	16.18	0.532	0.040	3.41
DeepDTA [Öztürk et al., 2018]	0.584	0.051	2.28	0.563	0.025	1.47
Gnina [McNutt et al., 2021]	0.782	0.299	17.73	0.609	0.054	4.63
RTMScore [Shen et al., 2022]	0.753	0.434	27.10	0.525	0.039	2.94
TankBind [Lu et al., 2022]	0.751	0.330	13.00	0.597	0.039	2.90
BigBind [Brocidiacono et al., 2023]	0.501	0.024	1.18	0.628	0.050	3.79
DrugCLIP [Gao et al., 2023]	0.809	0.505	31.89	0.572	0.062	5.51
GenScore [Shen et al., 2023]	0.816	0.473	28.53	0.596	0.065	5.14
EquiScore [Cao et al., 2024]	0.776	0.432	17.68	0.568	0.049	3.51
DrugHash [Han et al., 2025]	0.837	0.572	37.18	0.546	0.071	6.14
LigUnity [Feng et al., 2025]	0.897	0.674	44.20	0.599	0.075	6.50
HypSeek [Wang et al., 2026]	0.909	0.605	37.83	0.603	0.061	4.62
S^{2}Drug† [He et al., 2026]	0.925	**0.793**	43.06	0.582	0.087	7.38
CAUSALBIND$_{\mathrm{SP}}$ (ours)	0.935	0.744	47.69	**0.639**	0.097	8.56
CAUSALBIND$_{\mathrm{LR}}$ (ours)	0.935	0.704	44.87	0.632	**0.103**	8.84
CAUSALBIND$_{\mathrm{EMB}}$ (ours)	0.939	0.754	48.56	0.636	0.102	**9.60**
CAUSALBIND$_{\mathrm{ATOM}}$ (ours)	**0.943**	0.768	**49.66**	0.621	0.079	6.25

Instead of letting the contrastive objective operate on dense embeddings, we require it to act through sparse concept-level interactions. This enforces that binding-relevant dependence is captured through a small subset of cross-modal interactions. The other variants change only how the interaction is constrained: CAUSALBIND$_{\mathrm{LR}}$ factorizes the raw mask as $\widehat{\mathcal{M}}_{a,b} = \mathbf{U}_{a,b}\mathbf{V}_{a,b}^{\top}$ with $\mathbf{U}_{a,b}, \mathbf{V}_{a,b} \in \mathbb{R}^{K\times r}$ before the same activation and gate, CAUSALBIND$_{\mathrm{EMB}}$ applies the constrained mask to the pooled embedding instead of the concept axis, and CAUSALBIND$_{\mathrm{ATOM}}$ feeds the full per-atom Uni-Mol token sequences to the concept extractor. Further implementation details for the main CAUSALBIND$_{\mathrm{SP}}$ model, including the explicit form of the geometric regularizers in $\mathcal{L}_{\mathrm{reg}}$, and a side-by-side comparison of the three main implementations are provided in App. A5.2 and App. A5.5.

4 Experiments

Experimental setup. We evaluate on two complementary benchmarks: DUD-E [Mysinger et al., 2012], which covers broad target diversity, and LIT-PCBA [Tran-Nguyen et al., 2020, Mayr et al., 2018], which suppresses property-based shortcuts and is the stricter test of genuine binding discrimination. We report standard metrics: AUROC, BEDROC$_{80.5}$, and EF@1%, with the latter two emphasizing early enrichment. We adopt the same curated assay-level training dataset as LigUnity [Feng et al., 2025] and HypSeek [Wang et al., 2026], constructed from ChEMBL [Mendez et al., 2019], BindingDB [Gilson et al., 2016], and PDBBind [Liu et al., 2017]. As in these baselines, we remove every ChEMBL/BindingDB assay whose target UniProt ID appears in DUD-E, LIT-PCBA, or DEKOIS 2.0, reducing this portion from 26,729 to 18,316 assays, while the PDBBind pairs are not filtered by UniProt ID (App. A5). The evaluation is therefore a cross-dataset generalization protocol rather than a random split of the training pool, with about 476K training molecules and 130K pockets. All models are trained for 50 epochs on four NVIDIA A100 GPUs under the same pipeline for a fair comparison, and the main tables report evaluation results from the last checkpoint, which avoids tuning checkpoint selection to the held-out test benchmarks in the cross-dataset generalization setting.

Main Comparison on Virtual Screening. Table 1 compares CAUSALBIND with three families of virtual-screening methods: classical docking systems, structure-aware scoring models on protein-ligand complexes, and retrieval-based representation learners. Retrieval methods are strongest in this large-scale regime, so the key comparison is against recent retrieval baselines. CAUSALBIND$_{\mathrm{SP}}$ improves over LigUnity and HypSeek on all six metrics, with the largest gains on LIT-PCBA early enrichment, where property shortcuts are least useful: EF@1% rises from 6.50 to 8.56 over LigUnity and from 4.62 to 8.56 over HypSeek, while BEDROC$_{80.5}$ improves from 0.075 and 0.061 to 0.097. Against the paper-reported results of the concurrent S^{2}Drug [He et al., 2026], which uses a heavier two-stage sequence–structure pipeline, S^{2}Drug attains the best DUD-E BEDROC$_{80.5}$, while all three

Table 2: Ablation and diagnostic study on DUD-E and LIT-PCBA. CAUSALBIND$_{\mathrm{SP}}$ is the sparse mask, CAUSALBIND$_{\mathrm{LR}}$ is the low-rank mask, and CAUSALBIND$_{\mathrm{EMB}}$ is the pooled embedding-space mask. $-L_{\mathrm{spa}}$ removes the sparsity penalty from CAUSALBIND$_{\mathrm{SP}}$, while $+L_{\mathrm{spa}}$ tests an optional sparsity penalty on CAUSALBIND$_{\mathrm{LR}}$ without changing its rank-one parameterization. $\mathbf{M}_{\mathrm{emb}}^{r=1}$ replaces the pooled embedding mask with a rank-one mask. Rows report mean$_{\pm\,\mathrm{std}}$ over five runs.

Method	DUD-E (n=102)			LIT-PCBA (n=15)		
	AUROC	BEDROC$_{80.5}$	EF$_{1\%}$	AUROC	BEDROC$_{80.5}$	EF$_{1\%}$
LigUnity [Feng et al., 2025]	0.896 $\pm$ 0.005	0.660 $\pm$ 0.018	43.14 $\pm$ 1.42	0.586 $\pm$ 0.010	0.077 $\pm$ 0.001	6.46 $\pm$ 0.18
HypSeek [Wang et al., 2026]	0.914 $\pm$ 0.013	0.690 $\pm$ 0.081	44.12 $\pm$ 6.14	0.603 $\pm$ 0.009	0.067 $\pm$ 0.007	5.62 $\pm$ 0.71
CAUSALBIND$_{\mathrm{SP}}$	**0.939** $\pm$ 0.004	0.746 $\pm$ 0.004	47.99 $\pm$ 0.39	**0.635** $\pm$ 0.006	0.088 $\pm$ 0.007	7.55 $\pm$ 0.73
CAUSALBIND$_{\mathrm{SP}}$ $-L_{\mathrm{spa}}$	0.935 $\pm$ 0.004	0.732 $\pm$ 0.014	46.93 $\pm$ 1.07	0.623 $\pm$ 0.005	0.084 $\pm$ 0.004	7.17 $\pm$ 0.36
CAUSALBIND$_{\mathrm{LR}}$	0.937 $\pm$ 0.004	0.732 $\pm$ 0.011	46.87 $\pm$ 0.91	0.630 $\pm$ 0.006	0.085 $\pm$ 0.006	7.22 $\pm$ 0.31
CAUSALBIND$_{\mathrm{LR}}$ $+L_{\mathrm{spa}}$	0.935 $\pm$ 0.003	0.727 $\pm$ 0.007	46.39 $\pm$ 0.58	0.623 $\pm$ 0.007	0.083 $\pm$ 0.005	7.51 $\pm$ 0.94
CAUSALBIND$_{\mathrm{EMB}}$	0.938 $\pm$ 0.001	**0.752** $\pm$ 0.005	**48.47** $\pm$ 0.44	0.634 $\pm$ 0.008	**0.101** $\pm$ 0.007	**9.52** $\pm$ 0.96
CAUSALBIND$_{\mathrm{EMB}}$ $+\mathbf{M}_{\mathrm{emb}}^{r=1}$	0.723 $\pm$ 0.022	0.319 $\pm$ 0.013	20.96 $\pm$ 0.63	0.539 $\pm$ 0.003	0.018 $\pm$ 0.001	1.38 $\pm$ 0.10

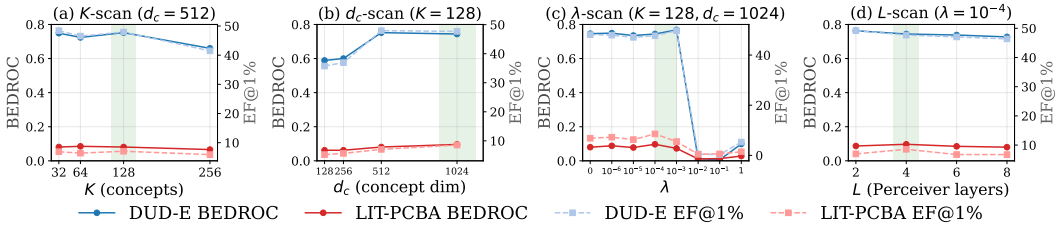


Figure 4: Hyperparameter sweeps for CAUSALBIND$_{\mathrm{SP}}$ on DUD-E and LIT-PCBA. Solid lines show BEDROC$_{80.5}$ and dashed lines show EF@1%; shaded bands mark the selected hyperparameters.

CAUSALBIND variants achieve higher DUD-E AUROC and EF@1% and higher values on all three LIT-PCBA metrics; a matched EF@0.5% comparison is given in App. A6.2. CAUSALBIND$_{\mathrm{LR}}$ and CAUSALBIND$_{\mathrm{EMB}}$ further show that the same constrained-interaction principle remains strong under explicit low-rank and embedding-level implementations. We also report CAUSALBIND$_{\mathrm{ATOM}}$, an atom-attentive version of CAUSALBIND$_{\mathrm{SP}}$ with the same configuration, whose concept extractor attends over the full per-atom token sequences of the Uni-Mol molecule and pocket encoders rather than their pooled tokens, so that concepts can be traced to individual atoms (App. A6.4). It attains the best DUD-E AUROC and EF@1%, but weaker LIT-PCBA early enrichment than the other variants. Overall, sparse or low-rank interaction helps when holistic alignment confronts nuisance correlations.

Ablation Diagnostics. Table 2 reports both component ablations and implementation diagnostics. For CAUSALBIND$_{\mathrm{SP}}$, removing L_{spa} weakens early enrichment, suggesting that the concept-pair mask is most useful when regularized toward a simple interaction structure. For CAUSALBIND$_{\mathrm{LR}}$, the rank-one factorization is retained in both rows; adding L_{spa} has limited effect, so the gain should be attributed to the explicit low-rank parameterization rather than to sparsity regularization. For CAUSALBIND$_{\mathrm{EMB}}$, imposing a rank-one pooled embedding mask sharply degrades performance, showing that rank compression must act directly on the concept axis rather than after pooling.

Hyperparameter Study. We sweep K (concept count), d_c (concept dimension), λ (sparsity weight λ_{spa}), and L (Perceiver depth). Figure 4 and Table 3 show robustness around our setting (K=128, d_c=1024, λ=10^{-4}, L=4), with failures: K=256 over-parameterizes, $d_c \leq 256$ underfits, and $\lambda \geq 10^{-2}$ collapses both benchmarks. In the useful regime, λ=10^{-4} favors LIT-PCBA and λ=10^{-3} favors DUD-E, so the sparsity level is benchmark-dependent. Overall, a soft interaction bottleneck helps, while hard-zero masking does not. The mask statistics in Table 3 describe the structure the model reaches: λ=10^{-4} has no hard zeros but concentrates its energy in one singular direction, the low-rank regime of Condition 3.i; λ=10^{-3} gives an exact sparse-antichain support; and λ=10^{-2} keeps the antichain property but over-sparsifies, so that property alone is insufficient. As ϕ is bilinear here, these are structural diagnostics rather than certificates of component-wise identifiability.

Out-of-distribution generalization and interpretability. We further evaluate zero-shot transfer on DEKOIS 2.0 [Bauer et al., 2013] under a target-level split over nine targets whose UniProt IDs are

Table 3: λ_{spa}-scan for CAUSALBIND$_{\text{SP}}$ with paired cross-modal mask statistics, averaged over the two branches. Small λ_{spa} preserves a low-rank regime, while large λ_{spa} drives hard-zero collapse. 'Anti. viol.' counts antichain violations, i.e., pairs of row (or column) supports in which one contains the other, over $32{,}512$ pairwise tests. Highlighted rows are the two operating points used in the text.

λ_{spa}	DUD-E			LIT-PCBA			Cross-modal mask statistics				
	AUC	BEDROC$_{80.5}$	EF@1%	AUC	BEDROC$_{80.5}$	EF@1%	L_1 avg	%zeros	r@95%	r@99%	Anti. viol.
0	0.938	0.746	48.14	0.617	0.080	6.83	0.93	0.0	1	25.0	32,512
10^{-6}	0.940	0.749	47.94	0.635	0.088	7.26	0.92	0.0	1	25.0	32,512
10^{-5}	0.938	0.735	47.09	0.607	0.078	6.29	0.88	0.0	1	26.0	32,512
10^{-4}	0.935	0.744	47.69	**0.639**	**0.097**	**8.56**	0.61	0.0	**1**	**34.0**	32,512
10^{-3}	**0.943**	**0.768**	**49.73**	0.614	0.073	5.49	**0.21**	**20.5**	45.5	79.5	**0**
10^{-2}	0.437	0.013	0.53	0.457	0.011	0.46	0.06	78.4	74.0	96.0	0
10^{-1}	0.297	0.006	0.30	0.437	0.013	0.59	0.02	92.0	75.5	97.5	1
1	0.725	0.099	5.30	0.595	0.028	1.44	0.01	96.0	75.5	96.5	324

absent from all training sources, and a joint target–scaffold split, where CAUSALBIND improves over HypSeek and LigUnity on AUROC and BEDROC$_{80.5}$ in both settings (App. A6.3). An interpretability case study on the Thyroid Hormone Receptor β (THRB) complex from DUD-E further shows that observed binding contacts concentrate within a learned concept pair, and that masking the associated atoms reduces the predicted binding score far more than size-matched random masks (App. A6.4).

Appendix roadmap. For readability, the main text reports only the core comparisons above. Implementation details and reproduction checks are in App. A5, including the three variants CAUSALBIND$_{\text{SP}}$, CAUSALBIND$_{\text{LR}}$, and CAUSALBIND$_{\text{EMB}}$. Additional protocol comparisons, including the `best_auc` and `best_bedroc` checkpoints of CAUSALBIND$_{\text{ATOM}}$, are in App. A6.1.2, and a longer discussion of the ablation study is in App. A6.1.3. The full hyperparameter tables for K/d_c, λ, and L are in App. A6.1.4, App. A6.1.5, and App. A6.1.6. The extended theory–practice analysis of the learned masks, including the antichain statistics of Table 3, is in App. A6.1.7. Supplementary DUD-E ROC-enrichment and Merck-FEP affinity-ranking analyses are in App. A6.1.8 and App. A6.1.9, low-rank and embedding-level variant results are in App. A6.5 and App. A6.6, and the full-sweep synthesis is in App. A6.7. The matched EF@0.5% comparison with S^{2}Drug, the DEKOIS 2.0 out-of-distribution evaluation, and the THRB interpretability case study are in App. A6.2, App. A6.3, and App. A6.4. A comprehensive discussion of the remaining theory–practice gap is in App. A6.8.

5 Conclusion

This paper revisits retrieval-based virtual screening from a structural perspective. Existing methods largely rely on dense shared-space alignment, although protein-molecule binding is typically driven by a small set of local compatibilities. We formalized this mismatch with a V-structure causal model under binder-only selection, showing that bilinear interaction is non-identifiable even when sparse, while sparse antichain interaction with a non-bilinear ϕ yields component-wise identifiability and low-rank interaction yields subspace identifiability. CAUSALBIND turns this principle into a concept-based retrieval framework with three structure-constrained implementations, and consistently improves over strong baselines on DUD-E and LIT-PCBA, with especially clear gains on LIT-PCBA early enrichment. Beyond the two main benchmarks, the learned interaction structures transfer to target- and scaffold-level out-of-distribution splits of DEKOIS 2.0, and the learned masks are strongly low-rank at our operating point, the regime our subspace-identifiability theorem covers. Overall, scalable virtual screening should learn not only better shared embeddings but also more reusable cross-modal interaction structures, together with the conditions under which they remain identifiable.

Acknowledgments

We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, MBZUAI-WIS Joint Program, and the Al Deira Causal Education project. Meanwhile, we also thank the reviewers for their constructive suggestions.

References

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.

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

Andrew A Bogan and Kurt S Thorn. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1):1–9, 1998.

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.

Tim Clackson and James A Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, 1995.

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*, pages 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, pages 17041–17049, 2025.

Bowei He, Bowen Gao, Yankai Chen, Yanyan Lan, Chen Ma, Philip S. Yu, Ya-Qin Zhang, and Wei-Ying Ma. S^2drug: Bridging protein sequence and 3d structure in contrastive representation learning for virtual screening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 354–362, 2026.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

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*, pages 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*, pages 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.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36:25112–25150, 2023.

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. In *International Conference on Learning Representations*, volume 2024, pages 10901–10934, 2024b.

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.

Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2): 302–309, 2017.

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.

Gongxu Luo, Haoyue Dai, Longkang Li, Chengqian Gao, Boyang Sun, and Kun Zhang. Gene regulatory network inference in the presence of selection bias and latent confounders. *Advances in Neural Information Processing Systems*, 38:18794–18826, 2026a.

Gongxu Luo, Loka Li, Guangyi Chen, Haoyue Dai, and Kun Zhang. Characterization and learning of causal graphs with latent confounders and post-treatment selection from interventional data. In *International Conference on Learning Representations*, volume 2026, pages 24716–24740, 2026b.

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.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 7, 1985.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

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*, pages arXiv–2411, 2025.

Zeyu Tang, Zhenhao Chen, Xiangchen Song, Loka Li, Yunlong Deng, Yifan Shen, Guangyi Chen, Peter Spirtes, and Kun Zhang. Reflection-window decoding: Text generation with selective refinement. *arXiv preprint arXiv:2502.03678*, 2025.

Viet-Khoa Tran-Nguyen, Célien Jacquemard, and Didier Rognan. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling*, 60(9): 4263–4273, 2020.

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.

W Patrick Walters and Renxiao Wang. New trends in virtual screening, 2020.

Jianhui Wang, Wenyu Zhu, Bowen Gao, Xin Hong, Ya-Qin Zhang, Wei-Ying Ma, and Yanyan Lan. Learning protein–ligand binding in hyperbolic space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 1150–1158, 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.

Kun Zhang, Jiaqi Sun, Yiqing Li, Ignavier Ng, Namrata Deka, and Shaoan Xie. Sedge: Structural extrapolated data generation. *arXiv preprint arXiv:2604.02482*, 2026.

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.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

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.

Klea Ziu, Slavomír Hanzely, Loka Li, Kun Zhang, Martin Takáč, and Dmitry Kamzolov. ψdag: Projected stochastic approximation iteration for DAG structure learning. *arXiv preprint arXiv:2410.23862*, 2024.

Appendix for

"CAUSALBIND: Causal Modeling and Learning for Protein-Molecule Virtual Screening"

Table of Contents:

A1 Drug Discovery Pipeline and Method Positioning 16

A2 Extended Related Work 18

A3 Biological Observations of Sparse Binding 21

A4 Theoretical Framework: Selection and Identifiability 22

A4.1 Structural Causal Model 23

A4.2 Discussion of Assumptions 24

A4.3 Toy Case: Two Concepts 25

A4.4 Theorem 1: Non-Identifiability Without Structural Constraints 26

A4.4.1 Proof Sketch 26

A4.4.2 Intuitive Example 27

A4.4.3 Full Proof of Theorem 1 27

A4.5 Theorem 2: Component-Wise Identifiability Under Sparse Interaction 28

A4.5.1 A Sparsity Lemma 28

A4.5.2 Full Proof of Theorem 2 29

A4.5.3 Beyond Linear Concept Ambiguity 30

A4.6 Theorem 3: Subspace Identifiability Under Low-Rank Interaction 31

A4.6.1 Proof Sketch 32

A4.6.2 Intuitive Example 32

A4.6.3 Full Proof of Theorem 3 32

A4.7 Summary and Discussion of the Theory 33

A5 Implementation Details 33

A5.1 Relation Between the Two-Modal Theory and the Three-Modal Implementation . 33

A5.2 CAUSALBIND$_{\mathrm{SP}}$: Sparse Concept-Pair Implementation 34

A5.2.1 Technical Supplement for the CAUSALBIND$_{\mathrm{SP}}$ Implementation 34

A5.2.2 Concrete Parameter Assignments and Experimental Setting 36

A5.3 CAUSALBIND$_{\mathrm{LR}}$: Low-Rank Concept-Basis Implementation 37

A5.3.1 Technical Supplement for the CAUSALBIND$_{\mathrm{LR}}$ Implementation 37

A5.3.2 Concrete Parameter Assignments and Experimental Setting 38

A5.4 CAUSALBIND$_{\mathrm{EMB}}$: Embedding-Level Masking 38

A5.4.1 Technical Supplement for the CAUSALBIND$_{\mathrm{EMB}}$ Implementation 38

A5.4.2 Concrete Parameter Assignments and Experimental Setting 39

A5.5 Comparison Across Three Implementations 39

A6 Experimental Analysis. 40
A6.1 CAUSALBIND$_{\mathrm{SP}}$: Sparse Concept-Pair Results 40
A6.1.1 Extended Analyses for the Main Experiments 40
A6.1.2 Main Comparison: Three-Protocol Evaluation 40
A6.1.3 Ablation and Diagnostic Study: Detailed Analysis 41
A6.1.4 Hyperparameter Study on K and d_c 42
A6.1.5 Hyperparameter Study on λ 42
A6.1.6 Hyperparameter Study on L 44
A6.1.7 Detailed Mask Structural Analysis 44
A6.1.8 ROC-Enrichment (RE) on DUD-E 46
A6.1.9 Affinity Ranking on Merck-FEP-8 47
A6.2 Comparison with S^{2}Drug 47
A6.3 Out-of-Distribution Evaluation on DEKOIS 2.0 48
A6.4 Interpretability Case Study on THRB. 48
A6.5 CAUSALBIND$_{\mathrm{LR}}$: Low-Rank Concept-Basis Results 49
A6.5.1 Main Rank Scan and Three-Protocol Results 49
A6.5.2 Multi-Seed Comparison and Ablation 50
A6.5.3 Hyperparameter Robustness 50
A6.5.4 Mask Analysis 51
A6.6 CAUSALBIND$_{\mathrm{EMB}}$: Experimental Results and Diagnostics 51
A6.6.1 Hyperparameter Study on K and d_c 51
A6.6.2 Experimental Results 52
A6.6.3 Theory Correspondence 53
A6.7 Overall Observations Across Three Implementations 54
A6.8 Discussion on Theory–Practice Gap 54
A7 Limitations, Broader Impact, and Usage of LLMs. 55
A8 Reproducibility and Asset Provenance. 56

A1 Drug Discovery Pipeline and Method Positioning

Figure A1 places CAUSALBIND within the broader drug-discovery workflow. The key point, consistent with the main text, is that our method is designed for the *high-throughput virtual-screening stage*: the part of the pipeline that must rapidly down-select a very large candidate library before more expensive docking and experimental follow-up. In other words, CAUSALBIND is not a pose-prediction or lead-optimization method; it is a retrieval-style early filter whose job is to improve the quality of the top-ranked compounds passed to later stages.

Stage descriptions. The pipeline in Figure A1 is best read as a funnel in which computational cost increases while the candidate set shrinks. It begins with a *virtual compound library* of roughly 10^9–10^{10} commercially available or synthetically accessible molecules, such as those represented by ZINC or Enamine REAL. Basic drug-likeness and target-specific filters then reduce this space

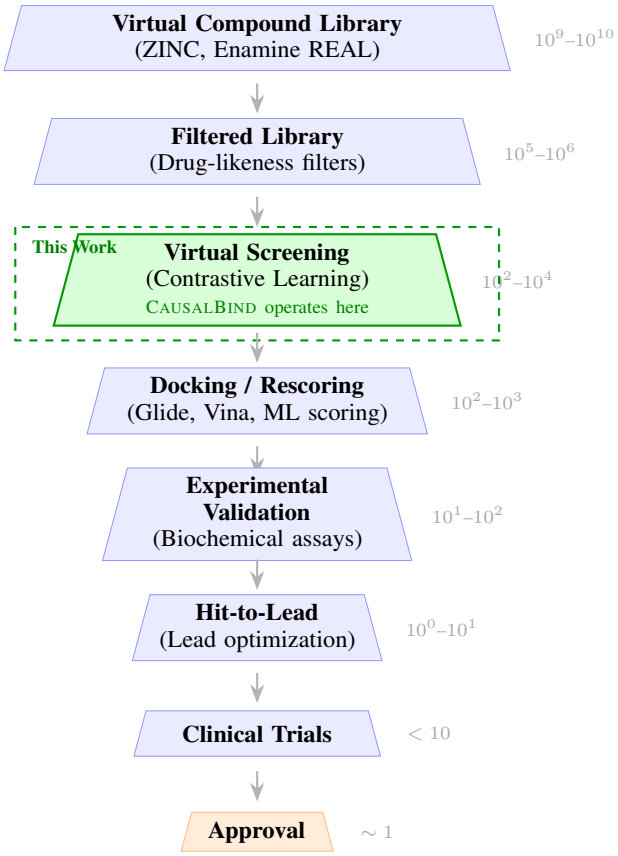


Figure A1: Drug discovery pipeline funnel. Each stage progressively filters compounds, with numbers indicating typical compound counts. Our CausalBind method (green box) operates at the virtual screening stage, improving the quality of top-ranked candidates before expensive downstream validation. By identifying invariant binding determinants and suppressing spurious correlations, our method aims to increase the hit rate in subsequent experimental testing.

to a *filtered library* of roughly 10^5–10^6 candidates. The next stage is *virtual screening*, where methods rank compounds by predicted binding relevance and reduce the candidate set to something tractable for follow-up; this is the stage where CausalBind operates. More expensive *docking or rescoring* is then applied to a much smaller set, typically 10^2–10^3 compounds, using physics-based or structure-aware scoring functions to reason in greater detail about poses and local interactions. Only after this reduction do compounds proceed to *experimental validation*, usually tens to hundreds of biochemical or cell-based assays, followed by *hit-to-lead* optimization, and eventually the even narrower stages of *clinical testing* and *approval*. The practical role of CausalBind is therefore upstream: it improves which candidates are allowed to reach these far more expensive downstream stages.

Why virtual screening matters. The virtual-screening stage is where the paper's evaluation metrics become practically meaningful. This stage is responsible for the largest candidate reduction before wet-lab testing, so the quality of the *very top* of the ranking matters far more than average prediction quality over a fixed benchmark. This is exactly why the main text emphasizes early-enrichment metrics such as $\text{BEDROC}_{80.5}$ and EF@1%: even a modest improvement there can substantially reduce the number of false positives advanced to docking or assay validation. Our focus on sparse binding determinants is intended to improve this high-value early-filtering regime.

Complementary to docking. This positioning also clarifies why the paper compares docking, structure-aware scoring, and retrieval as distinct regimes rather than as direct substitutes. Docking remains valuable because it reasons explicitly about poses and local geometry, but it is too expensive to serve as the first-pass search over very large libraries. Retrieval-style methods occupy a different

operating point: they trade mechanistic detail for scale and are therefore best used upstream of docking. CAUSALBIND fits this role. Its contribution is to improve the representation quality of that pre-filtering stage so that the compounds passed downstream are more likely to contain true binders.

A2 Extended Related Work

From docking to scalable screening. Classical virtual screening is dominated by docking pipelines such as Glide, AutoDock Vina, and Surflex, which explicitly search binding poses and rank compounds using physics-inspired scoring functions [Friesner et al., 2004, Trott and Olson, 2010, Spitzer and Jain, 2012]. Their strengths are interpretability and direct geometric reasoning about binding modes, but their cost grows quickly with library size and target flexibility. In practical workflows this often makes docking a *downstream* stage, applied after an initial coarse filter rather than as the first-pass search over millions of compounds. This is exactly the regime targeted in the main text: large-scale pre-filtering, where retrieval speed and early enrichment matter more than full pose resolution.

Learned scoring functions and pose-aware models. A large literature replaces hand-engineered docking scores with learned interaction models over protein-molecule complexes. Gnina introduced deep learning into docking-time scoring with 3D CNNs [McNutt et al., 2021]; RTMScore, GenScore, EquiScore, PIGNet, and PLANET further strengthened this line with graph, transformer, equivariant, and physics-informed architectures [Shen et al., 2022, 2023, Cao et al., 2024, Moon et al., 2022, Zhang et al., 2023]. Related structure-based models such as TankBind and BigBind learn from geometric complexes or large-scale structural supervision to improve affinity prediction and screening robustness [Lu et al., 2022, Brocidiacono et al., 2023]. These methods are highly relevant baselines because they capture detailed intermolecular geometry, but they usually assume a candidate complex, pose, or structure-conditioned interaction graph at inference time. By contrast, our focus is on *retrieval-style* virtual screening where compound scoring must remain fast enough for large libraries.

Affinity prediction beyond screening. Drug–target affinity (DTA) models such as DeepDTA and GraphDTA predict binding strength from sequence or graph representations without necessarily producing an efficient retrieval system [Öztürk et al., 2018, Nguyen et al., 2021]. This line clarified how learned molecular and protein embeddings can support affinity estimation, but it is typically evaluated on regression-style benchmarks rather than target-wise early enrichment. The distinction matters for our paper because the objective is not simply to estimate affinity well on average; it is to identify which compounds should be placed at the very top of the screening list.

Retrieval-based contrastive learning for virtual screening. The recent shift most relevant to our work is the move from explicit complex scoring to *shared embedding spaces*. DrugCLIP [Gao et al., 2023] framed protein-molecule virtual screening as dense retrieval, using Uni-Mol-based encoders and CLIP-style contrastive learning [Radford et al., 2021, Zhou et al., 2023]. This design substantially improved scalability and helped establish retrieval-based virtual screening as a viable alternative to docking, a direction later amplified by a genome-scale *Science* study built on DrugCLIP-style contrastive screening [Jia et al., 2026]. Subsequent methods enriched this paradigm in different ways: DrugHash compresses embeddings into binary codes for memory-efficient screening [Han et al., 2025]; LigUnity unifies virtual screening with hit-to-lead affinity ranking through scaffold discrimination and pharmacophore ranking in a shared pocket–ligand space [Feng et al., 2025]; HypSeek argues that hyperbolic geometry better captures hierarchical and fine-grained affinity variation than Euclidean retrieval alone [Wang et al., 2026]; and the concurrent S^2Drug bridges protein sequence and 3D structure through ChEMBL sequence pretraining and PDBBind fine-tuning with residue-level sequence–structure fusion and an auxiliary binding-site task [He et al., 2026]. These papers collectively show that the field has moved from asking whether retrieval can work at all to asking what kind of structural bias should govern the shared space. The present paper answers that question by shifting the focus from dense holistic alignment to sparse or low-rank cross-modal interaction.

Benchmark bias and generalization. A recurring concern in virtual screening is that benchmark artifacts can inflate apparent performance. Hidden biases in DUD-E and related datasets have long been known to create shortcuts based on simple physicochemical cues rather than genuine binding recognition [Chen et al., 2019]. More recent analyses argue that debiasing is essential for robust

binding-affinity generalization across targets and chemical series [Graber et al., 2025]. This motivates the paper's emphasis on LIT-PCBA and on early-enrichment gains under more realistic decoys: the goal is not merely to improve in-distribution ranking, but to learn representations that remain useful when easy shortcuts disappear.

Why three regimes? Core computational pattern. Although all three regimes ultimately rank protein-molecule pairs, their *inference-time computational pattern* differs sharply, and this is what determines their practical scaling behaviour and where their inductive bias has to live. The defining feature of retrieval-based methods is that protein and molecule embeddings are computed *independently* (a dual-tower architecture); the binding score then reduces to a similarity, typically a dot product, between two pre-computable vectors, exactly as in dense passage retrieval in NLP [Oord et al., 2018, Radford et al., 2021] or image–text retrieval. This pattern is what unlocks very large-scale screening, but it also constrains where cross-modal interaction can be expressed: the interaction has to be encoded into the representation or the similarity function, rather than being computed by a joint pair-input network at inference time. This is precisely where CAUSALBIND intervenes. Instead of leaving the retrieval score as a dense holistic similarity, we impose sparse concept-level interaction inside the retrieval regime itself. Table A1 makes the contrast explicit; Table A2 then drills down to representative methods.

Table A1: Core differences between the three virtual-screening regimes. The retrieval regime decouples encoders into two towers, enabling pre-computed library embeddings and millisecond-scale similarity scoring; the other two regimes condition on the (protein, ligand) pair at inference. CAUSALBIND$_{\mathrm{SP}}$ stays in the retrieval regime but replaces the dense holistic similarity with a sparse bipartite concept-pair interaction.

	Physics-based docking	**Structure-aware ML scoring**	**Retrieval-based learning**
Inference input	(protein, ligand) pair + pose search	(protein, ligand) pair, often pose-conditioned	protein and ligand encoded *independently*
Binding score	physics + empirical scoring on the pair	one forward pass through a joint network on the pair	dot product (or cosine) of two precomputed vectors
Pre-computable embeddings?	No (pose-coupled)	No (pair-coupled)	**Yes** (dual-tower)
Cost per (P, L) at inference	seconds–minutes (pose search dominates)	$\sim$10–100 ms (one full forward)	sub-millisecond (one dot product)
Practical library scale	10^4–10^5	10^5–10^6	$\mathbf{10^8}$–$\mathbf{10^9}$
Where the inductive bias lives	force field + sampling protocol	complex graph / 3D-voxel architecture	shape of the shared embedding space and the similarity used on it
Typical output	docked pose + affinity score	affinity score (sometimes a pose)	ranked retrieval list
Where CAUSALBIND$_{\mathrm{SP}}$ sits	N/A	N/A	retrieval regime, but with a *sparse bipartite* similarity over concept pairs instead of a dense one

Cross-regime comparison. Building on the regime-level distinctions of Table A1, Table A2 drills down to representative methods within each regime and highlights the design choice that distinguishes CAUSALBIND from prior retrieval baselines: not a new backbone family, but a different structural prior on cross-modal interaction.

Detailed positioning. The methods in Table 1 can be read as a progression in where the binding inductive bias is placed. Glide-SP is the classic physics-based reference point: it reasons explicitly about poses and local contacts, but at a computational cost that makes it unsuitable for first-pass screening over very large libraries. DeepDTA occupies a different corner of the design space: it is important historically because it showed that learned molecular and protein representations can support affinity prediction, but it is not a retrieval-oriented screening model and therefore is not optimized for the target-wise early-enrichment setting emphasized in this paper.

Table A2: Representative methods in the three virtual-screening regimes discussed in the main text. The table highlights the main design choice of each method and the role it plays in the comparison, with CAUSALBIND$_{SP}$ differing from prior retrieval baselines by imposing a sparse bipartite interaction prior rather than a dense holistic similarity.

Method	Core Design	Relevance Here
Physics-based docking systems		
Glide-SP [Friesner et al., 2004]	Hierarchical sampling with empirical scoring function	Industry standard; accurate on small candidate sets but expensive at library scale
AutoDock Vina [Trott and Olson, 2010]	Empirical scoring + iterated local search	Open-source, widely used; not designed for ultra-large libraries
Structure-aware machine-learning scoring models		
DeepDTA [Öztürk et al., 2018]	Sequence/graph-based affinity prediction without pairwise retrieval design	Useful DTA baseline; not optimized for target-wise early enrichment
Gnina [McNutt et al., 2021]	3D CNN scoring on docking-grid voxels	Strong pose-rescoring; depends on docked poses
RTMScore [Shen et al., 2022]	Residue-level GNN re-scoring of poses	Good DUD-E AUROC; needs upstream pose enumeration
GenScore [Shen et al., 2023]	Generalist pose generation + rescoring net	Unifies pose and score with one network
EquiScore [Cao et al., 2024]	SE(3)-equivariant GNN on the binding complex	Geometric inductive bias on the complex graph
PIGNet [Moon et al., 2022]	Physics-informed GNN combining empirical + learned terms	Hybrid physics/ML score; strong DUD-E early enrichment
TankBind [Lu et al., 2022]	Trigonometry-aware GNN, end-to-end pose+affinity	Joint binding-site prediction and scoring
BigBind [Brocidiacono et al., 2023]	GNN trained on large-scale curated binding corpus	Stresses the role of training-data scale and curation
Retrieval-based representation learning		
DrugCLIP [Gao et al., 2023]	CLIP-style contrastive on Uni-Mol encoders	Scalable; adopted at genome scale [Jia et al., 2026]
DrugHash [Han et al., 2025]	Binary-hash retrieval extension of DrugCLIP	Memory-efficient at billion-compound scale; weaker PCBA AUROC
LigUnity [Feng et al., 2025]	Unified screening + hit-to-lead ranking loss	Strong BEDROC/EF on both DUD-E and LIT-PCBA
HypSeek [Wang et al., 2026]	Hyperbolic geometry + cone hierarchy + ESM-2 sequence branch	Prior SOTA on DUD-E and LIT-PCBA; geometry-driven inductive bias
CAUSALBIND$_{SP}$ (ours)	**Sparse bipartite concept-pair interaction over Perceiver concepts**	**Theoretically grounded under Heckman selection; best LIT-PCBA early enrichment**

Among structure-aware learning-based methods, Gnina, RTMScore, GenScore, EquiScore, TankBind, and BigBind all retain a joint pair-input view of the protein-molecule problem. Gnina and RTMScore are closely tied to docked-pose scoring, inheriting much of the pose dependence of classical docking. GenScore and EquiScore strengthen this line with more expressive learned complex representations, while TankBind pushes further toward end-to-end geometric reasoning over binding sites and complexes. BigBind is especially notable for highlighting the importance of large-scale curated structural supervision. These methods are relevant because they capture explicit intermolecular geometry, but their inference-time cost and pair-conditioned architecture make them fundamentally different from the large-scale retrieval regime studied in our main experiments.

The retrieval baselines are the closest comparators. DrugCLIP established the modern dense-retrieval formulation for virtual screening and demonstrated that CLIP-style contrastive learning with Uni-Mol encoders can scale to very large libraries. DrugHash preserves that same dense-alignment

philosophy while prioritizing storage and retrieval efficiency through binary hashing, at the cost of some predictive fidelity. LigUnity extends dense shared-space retrieval by coupling virtual screening with hit-to-lead ranking signals, which helps its early-enrichment behavior but does not change the fact that the interaction remains holistic in the shared space. HypSeek is the strongest baseline in this family: it enriches retrieval with hyperbolic geometry, cone-hierarchy regularization, and a protein-sequence branch, yielding a powerful geometric inductive bias but still no explicit structural prior on which cross-modal interactions should matter.

This is the point at which CAUSALBIND differs. We stay in the retrieval regime and keep the scalability advantages of independently encoded molecule and protein representations, but we change the structural bottleneck through which binding information must pass. Compared with docking or pose-aware scoring, CAUSALBIND does not attempt to replace explicit structural modeling; instead it improves the *front-end screening* stage that decides which compounds deserve expensive downstream analysis. Compared with DrugCLIP, DrugHash, LigUnity, and HypSeek, we leave the high-level retrieval pipeline intact but modify the representation bottleneck: a shared *ConceptExtractor* decomposes each modality into atomic concepts, and a sparse cross-modal mask highlights which concept pairs are causally relevant for binding. Semantically, prior retrieval methods ask what inductive bias should structure the embedding *space* and answer with metric geometry, compression, or training-time objectives. CAUSALBIND instead asks what inductive bias should structure the *interaction between* molecule and protein representations, and answers with a sparse or low-rank bipartite kernel grounded in the V-structure causal model of Section 2. In that sense, CAUSALBIND is best understood not as a new backbone family, but as an orthogonal structural prior inside retrieval-based virtual screening itself.

Causal representation learning. The conceptual basis for our method comes from causal representation learning [Schölkopf et al., 2021, Hyvarinen et al., 2019], especially work showing that multimodal observations and sparse graphical structure can improve identifiability of latent variables [Daunhawer et al., 2023, Kivva et al., 2021]. ConceptAligner [Xie et al., 2025] operationalized this intuition in a multimodal generative setting, while Sun et al. [2025] extended related ideas to biomedical multimodal data with identifiability guarantees for the latent causal variables. We adapt this literature to protein-molecule modeling under a different causal graph: unlike text-to-image generation, neither modality generates the other; instead molecule and protein concepts jointly determine a binding event through a selected collider. This shift from a directional chain to a V-structure is what motivates the sparse bipartite mask at the center of CAUSALBIND.

Causal discovery and selection bias. Complementary to causal representation learning, causal discovery recovers causal graphs among observed variables, with recent advances that handle deterministic relations [Li et al., 2024a], decentralized heterogeneous data [Li et al., 2024b], and scalable DAG structure learning [Ziu et al., 2024]. A related line of work studies how selection distorts the observed distribution: causal graphs with latent confounders and post-treatment selection can be characterized and learned from interventional data [Luo et al., 2026b], gene regulatory networks can be inferred in the presence of selection bias [Luo et al., 2026a], and selective mechanisms have also been exploited in text generation [Tang et al., 2025]. Our setting shares this concern with selection: training data contain only observed binding pairs, and CAUSALBIND explicitly models this binder-only selection as conditioning on a collider when deriving which interaction structures remain identifiable.

A3 Biological Observations of Sparse Binding

Hot spots and sparse binding energy. A direct biological motivation for our sparse-interaction view comes from the classical hot-spot literature on molecular recognition. In the human growth hormone–receptor system, Clackson and Wells [1995] combined the crystal structure of the complex with alanine-scanning mutagenesis and found that many side chains participate in the physical interface, but the energetic contribution is highly concentrated: a central hydrophobic region on the receptor side, dominated by two tryptophan residues, accounts for more than three quarters of the binding free energy. This result separates the *structural epitope*, namely the broad set of contacts visible in the complex, from the *functional epitope*, namely the much smaller subset of contacts that actually maintains affinity. Bogan and Thorn [1998] then showed that this concentration of binding energy is a general property of protein interfaces rather than an isolated example: hot spots

are formed by a small subset of residues, are enriched in tryptophan, tyrosine, and arginine, and are often surrounded by less energetic residues that shield the hot spot from bulk solvent. These findings support the central modeling assumption of CAUSALBIND: the binding signal should not be represented as a uniformly dense agreement between whole protein and molecule embeddings. Instead, it is more natural to model binding through a limited number of cross-modal compatibilities, while treating much of the remaining surface or molecular context as supportive, redundant, or potentially nuisance-correlated.

Pharmacophores as invariant binding determinants. Even though induced fit allows conformational adjustment, medicinal chemistry has long emphasized that certain feature types remain consistently binding-relevant across conformational states. These are the *pharmacophoric* features [Wermuth, 2006]: hydrogen-bond donors and acceptors, hydrophobic centers, aromatic groups, charged moieties, and related steric/electronic motifs that repeatedly determine whether a ligand can engage a target pocket productively. The exact geometry of a bound complex may shift, but these feature types remain the stable determinants that make those local interactions possible. In the language of this paper, they are the biologically plausible candidates for the "atomic concepts" that should survive across proteins, ligands, and conformational variants.

Why static structures are still informative. A natural concern is that if induced fit is dynamic, then learning from static structures might miss the relevant mechanism. The key point is that our goal is not to recover every conformational degree of freedom; it is to recover the *interaction-relevant determinants* that persist across such variability. A hydroxyl group remains a hydrogen-bond donor; an aromatic ring remains capable of π-stacking; a hydrophobic patch remains nonpolar even if side chains rotate slightly around it. Benchmarks such as DUD-E and LIT-PCBA are built from static structures, yet successful virtual-screening methods already rely on the fact that these chemically meaningful determinants are more stable than the incidental conformational details. Our method does not remove this assumption; it makes it more explicit by organizing the representation around sparse cross-modal interactions rather than around dense global similarity.

From biology to the modeling choice in CAUSALBIND. This biological picture motivates the two central methodological choices in the paper. First, if binding depends on a small set of recurring local determinants, then each modality should be decomposed into smaller concept-like units rather than treated as a single holistic vector. Second, if only a subset of those units actually participates in binding for a given target family, then the cross-modal interaction should itself be sparse or low-rank rather than dense. This is exactly what the *ConceptExtractor* and *SparseMask* are designed to implement. The model is therefore not claiming that every learned concept is a textbook pharmacophore; rather, it is imposing an inductive bias that is much closer to the biological mechanism than dense shared-space alignment.

Relevance to generalization and drug discovery. This interpretation also clarifies why the largest gains tend to appear on the more bias-controlled LIT-PCBA benchmark. If a method relies on shortcut correlations, it may rank compounds well on easier datasets without truly isolating binding determinants. By contrast, a representation organized around invariant local compatibilities should transfer better to novel targets and novel scaffolds, because those compatibilities are closer to the actual mechanism medicinal chemists reason about during lead optimization. In this sense, the biological motivation of CAUSALBIND is not only explanatory but also practical: it supports the paper's broader claim that sparse concept-level interaction is a more plausible route to robust virtual screening than dense holistic alignment.

A4 Theoretical Framework: Selection and Identifiability

This appendix provides the complete theoretical framework for CAUSALBIND. We cast protein-molecule binding prediction as a *Heckman-style selection problem* on a V-structure causal graph: the training set consists only of binding pairs ($B = 1$), and our goal is to identify the latent concepts $\mathbf{c}^M, \mathbf{c}^P$ and the bipartite interaction mask $\mathcal{M}$ from this selection-biased observation. Following the didactic organization of SEDGE [Zhang et al., 2026] and ConceptAligner [Xie et al., 2025], we proceed from a careful SCM formulation (A4.1) through a worked $K_M = K_P = 2$ toy case (A4.3) to the general non-identifiability and identifiability results (A4.4–A4.5), a low-rank subspace-identifiability result (A4.6), and an explicit comparison to SEDGE (A4.7). Each theorem subsection

is organized into a proof sketch, an intuitive example, and a full proof, and a later discussion in the experimental appendix (A6.8) returns to how these identifiability results should be interpreted in light of the learned masks.

Our framework borrows the Heckman-selection machinery recently formalized for vision–language concept generation by Xie et al. [2025] and for compositional extrapolation by SEDGE, but differs in four crucial ways: (i) we have a *collider* causal structure (V-structure) rather than a chain; (ii) we have two *symmetric* observation sources (molecule and protein) rather than one feature vector; (iii) our binding kernel is bilinear-sparse rather than a parametric MLP head; and (iv) we replace the SEDGE-style "conservative" assumption ($P(S{=}0) \approx 0$) with a *low-rank binding kernel* assumption that is better suited to drug discovery, where most protein-molecule pairs do not bind.

A4.1 Structural Causal Model

Variables and generative process. Let $m \in \mathcal{M}$ and $p \in \mathcal{P}$ denote an observed molecule (3D structure) and protein (3D structure). We posit that each modality is generated from latent *concepts*:

$$\mathbf{c}^M = [c_1^M, \ldots, c_{K_M}^M] \in \mathbb{R}^{K_M}, \qquad \mathbf{c}^P = [c_1^P, \ldots, c_{K_P}^P] \in \mathbb{R}^{K_P}, \tag{10}$$

through observation maps with modality-specific nuisance noise

$$m = g^M(\mathbf{c}^M, \varepsilon^M), \qquad p = g^P(\mathbf{c}^P, \varepsilon^P), \tag{11}$$

where ε^M and ε^P capture observation noise, conformational variability, or other modality-specific nuisance factors that are independent of the cross-modal binding event given the concepts. We do not require g^M and g^P to be globally invertible on the full observation spaces. The recoverability assumption below only requires that the interaction-relevant concepts can be expressed as functions of their corresponding observations on the data support. Biologically, each c_i^M (resp. c_j^P) captures an *atomic* chemical/structural feature of the molecule (protein), analogous to a pharmacophoric element (hydrogen-bond donor, hydrophobic contact, aromatic ring, etc.).

Binding as a collider. The binding indicator $B \in \{0, 1\}$ is a *collider*: it is determined jointly by the two concept vectors through a sparse bipartite interaction,

$$B \sim \text{Bernoulli}\left(\sigma\left(\sum_{i=1}^{K_M}\sum_{j=1}^{K_P} \mathcal{M}_{ij}\, \phi(c_i^M, c_j^P)\right)\right), \tag{12}$$

where $\mathcal{M} \in \mathbb{R}_{\geq 0}^{K_M \times K_P}$ is the sparse *bipartite binding mask*; binary masks and $[0, 1]$ relaxations are special normalizations, while scale is shared with ϕ. Here K_M and K_P count concepts, and each atomic concept is scalar in this theoretical model, i.e., $c_i^M, c_j^P \in \mathbb{R}$. Thus $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ maps one molecule concept and one protein concept to a scalar logit-level compatibility score; σ is the logistic. The graph is

$$\varepsilon^M \to m \leftarrow \mathbf{c}^M \to B \leftarrow \mathbf{c}^P \to p \leftarrow \varepsilon^P, \tag{13}$$

exhibiting a V-structure with B as the collider (Figure 3 in the main text).

Bayesian network and conditional-independence structure. We treat $(\mathbf{c}^M, \mathbf{c}^P, \varepsilon^M, \varepsilon^P, B, m, p)$ as a Bayesian network [Pearl, 1985, Koller and Friedman, 2009] with respect to the DAG in Eq. (13). The local Markov property yields the following conditional independence (CI) relations, which we use repeatedly in proofs:

$$\mathbf{c}^M \perp\!\!\!\perp \mathbf{c}^P \quad \text{(marginal independence from V-structure)}, \tag{14}$$

$$m \perp\!\!\!\perp \{\mathbf{c}^P, \varepsilon^P, B, p\} \mid (\mathbf{c}^M, \varepsilon^M) \quad \text{(molecule view depends only on molecule concepts and noise)}, \tag{15}$$

$$p \perp\!\!\!\perp \{\mathbf{c}^M, \varepsilon^M, B, m\} \mid (\mathbf{c}^P, \varepsilon^P) \quad \text{(protein view depends only on protein concepts and noise)}, \tag{16}$$

$$B \perp\!\!\!\perp \{m, p, \varepsilon^M, \varepsilon^P\} \mid (\mathbf{c}^M, \mathbf{c}^P) \quad \text{(binding mediated entirely by concepts)}. \tag{17}$$

Heckman selection. The training corpus $\mathcal{D}$ contains only positive binding pairs, i.e., samples drawn from

$$p^{\mathcal{D}}(m, p, \mathbf{c}^M, \mathbf{c}^P) \;=\; p\big(m, p, \mathbf{c}^M, \mathbf{c}^P \mid B = 1\big)\,. \tag{18}$$

Equivalently, the full selected joint before marginalizing nuisance variables is

$$p^{\mathcal{D}}(m, p, \mathbf{c}^M, \mathbf{c}^P, \varepsilon^M, \varepsilon^P) = p\big(m, p, \mathbf{c}^M, \mathbf{c}^P, \varepsilon^M, \varepsilon^P \mid B = 1\big)\,. \tag{19}$$

In Heckman-selection terminology [Heckman, 1979], B plays the role of the selection variable S: we observe (x^M, x^P) if and only if $B = 1$. Because B is a collider in Eq. (13), conditioning on $B = 1$ induces the classical *explaining-away* dependence between $\mathbf{c}^M$ and $\mathbf{c}^P$ even though they are marginally independent. Explicitly, using Eq. (14) and Bayes' rule:

$$p^{\mathcal{D}}(\mathbf{c}^M, \mathbf{c}^P) \;:=\; p(\mathbf{c}^M, \mathbf{c}^P \mid B = 1) \;=\; \frac{p(\mathbf{c}^M)\, p(\mathbf{c}^P)\, P(B = 1 \mid \mathbf{c}^M, \mathbf{c}^P)}{P(B = 1)}. \tag{20}$$

This is the key handle: the observed joint distribution carries information about $(\mathcal{M}, \phi)$ through the explaining-away factor $P(B = 1 \mid \mathbf{c}^M, \mathbf{c}^P)$.

Notation. Following SEDGE's conventions [Zhang et al., 2026], uppercase letters denote random variables and lowercase their realizations. We write $p^{\mathcal{D}}(\cdot) := p(\cdot \mid B = 1)$ for distributions implied by the (selection-biased) training data, and reserve $p(\cdot)$ for population distributions (which may involve the unobserved $B = 0$ regime). We use $\propto$ for equality up to a positive multiplicative constant and $\mathbf{1}$ for an all-ones vector. Supports and densities are assumed to exist with respect to common dominating measures (typically Lebesgue in $\mathbb{R}^{K_M}$ and $\mathbb{R}^{K_P}$, or counting in the corresponding binary spaces).

A4.2 Discussion of Assumptions

The formal assumptions and theorem-specific conditions are stated in the main text, so we do not repeat them here. Instead, this subsection explains what each assumption or condition contributes to the proofs and how strong it should be read. We keep the main-text notation throughout: Assumptions 1–3 are shared by all theorems, Assumption 4 is used by the two positive results, Conditions 2.i–2.iii belong to Theorem 2, and Conditions 3.i–3.ii belong to Theorem 3.

- **Assumption 1 (concept recoverability).** This assumption says that the concepts entering the binding mechanism are statistically accessible from the observed molecule or protein. It should not be read as requiring a global inverse for the full observation maps g^M, g^P or for the neural encoders used in implementation. The proofs use it only to pull the binder-selected observation distribution $p^{\mathcal{D}}(x^M, x^P)$ back to concept coordinates on the data support.
- **Assumption 2 (prior independence).** This is the V-structure assumption: before conditioning on binding, molecule-side and protein-side concepts are independent objects. The observed dependence in $p^{\mathcal{D}}$ then comes from conditioning on the collider $B = 1$. Without this assumption, cross-modal dependence could already be present before selection, and the proof could not attribute it to the binding mechanism.
- **Assumption 3 (positivity / coverage).** This is the minimal support requirement. If a concept never appears in binder-selected data, it leaves no statistical trace and cannot be identified by any method. The assumption does not require uniform coverage of all concept combinations; it only rules out concepts that are completely absent from the selected support.
- **Assumption 4 (log-linear selection).** This assumption makes the binding score itself, rather than $\log \sigma$ of it, appear in the selected log-density, so that $\log p^{\mathcal{D}}$ equals S up to additive functions of a single modality. It matters for the proofs: under the logistic link, the cross-derivative of $\log \sigma(S)$ contains the term $-\sigma(S)(1 - \sigma(S))\, \partial_{c_i^M} S\, \partial_{c_j^P} S$, which is dense even when $\mathcal{M}$ is sparse, so the support of $\mathcal{M}$ would not be visible in the object the proof differentiates. The assumption is the low-hit-rate form of the logistic model, since $\sigma(s) = e^s(1 + O(e^s))$ as $s \to -\infty$ and most protein-molecule pairs do not bind, and it is also the model under which the contrastive objective of Section 3 is well specified: the optimal contrastive critic is the log density ratio up to terms depending on one modality [Oord et al., 2018], which under this assumption is exactly S up to such terms.

- **Condition 2.i (sparse antichain interaction).** Sparsity alone is not enough for component-wise identifiability: two concepts can still be confounded if one concept's interaction support is absorbed by another's. The antichain requirement gives each molecule-side and protein-side concept a distinct combinatorial footprint in the bipartite interaction graph. It is the combinatorial input of Lemma 1, and Example 1(b) shows that it cannot be dropped.
- **Condition 2.ii (non-bilinear compatibility).** This condition makes the cross-derivative of each interaction term vary over the support, so that matching the two selected distributions yields a family of matrix equations spanning all matrices supported on $\mathrm{supp}(\mathcal{M})$, rather than the single equation available in the bilinear case. It is a generic property of ϕ alone and does not depend on the mask: $\phi(x, y) = \tanh(xy)$ satisfies it, whereas $\phi(x, y) = xy$ and $\phi(x, y) = -(x - y)^2$ do not. Without it, Theorem 1 provides counterexamples even for sparse antichain masks, which is why the inner-product compatibility used in our implementation falls under Theorem 3 rather than Theorem 2.
- **Condition 2.iii (linear ambiguity and sparsity minimality).** The first clause restricts the competing parameterizations to the linear reparameterizations of Theorem 1. Linear identifiability is the usual first stage of identifiability analyses and holds for representations learned with contrastive objectives [Roeder et al., 2021]; Proposition 1 removes it at the price of a stronger, pointwise nonlinearity condition. The second clause requires that the data not be explained equally well by a denser mask. It is the population counterpart of the sparsity penalty on the mask in the training objective, and Example 1(c) shows that it cannot be dropped.
- **Condition 3.i (low-rank interaction kernel).** This is the structural condition for the subspace theorem. It replaces exact concept-pair support recovery with the assumption that the binding score has a minimal rank-r separable factorization. Thus the target of identification changes: we no longer seek a unique entry-wise mask, but the molecule-side and protein-side subspaces spanned by the low-rank interaction functions.
- **Condition 3.ii (low-rank sufficient variability).** This is the statistical richness condition for the subspace theorem, not another low-rank constraint. It requires the mode-evaluation matrices to have full rank at finitely many molecule-side and protein-side probes. This guarantees that the low-rank modes specified by Condition 3.i are sufficiently excited on the selected support, so their spans can be recovered from the observed distribution.

A4.3 Toy Case: Two Concepts

Before the general theorems, we work out the simplest non-trivial case explicitly, following SEDGE's didactic strategy. Assume $K_M = K_P = 2$ with continuous scalar concepts, binding score $S(\mathbf{c}^M, \mathbf{c}^P) = \sum_{i,j=1}^{2} \mathcal{M}_{ij}\phi(c_i^M, c_j^P)$, and log-linear selection (Assumption 4), so that two parameterizations induce the same $p^{\mathcal{D}}$ exactly when their binding scores agree up to additive functions of a single modality. Write $\mathbf{c}^M = (c_1, c_2)$ and $\mathbf{c}^P = (c_1', c_2')$.

Case (a): Diagonal mask with bilinear compatibility (not identifiable). Let $\mathcal{M} = I_2$ and $\phi(x, y) = xy$, so $S = c_1c_1' + c_2c_2'$. The row supports are $\{1\}$ and $\{2\}$, an antichain, and the mask is as sparse as it can be while leaving every concept active. Nevertheless, with

$$T_M = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad T_P = T_M^{-\top} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix},$$

the concepts $\tilde{\mathbf{c}}^M = T_M\mathbf{c}^M = (c_1 + c_2,\, c_2)$ and $\tilde{\mathbf{c}}^P = T_P\mathbf{c}^P = (c_1',\, c_2' - c_1')$ satisfy $\tilde{\mathbf{c}}^{M\top} I_2\, \tilde{\mathbf{c}}^P = \mathbf{c}^{M\top} T_M^\top T_M^{-\top} \mathbf{c}^P = S$. The same mask I_2 therefore explains the data with mixed concepts: sparsity and the antichain property alone do not identify the concepts. This is the smallest instance of the second part of Theorem 1.

Case (b): Diagonal mask with non-bilinear compatibility (identifiable). Keep $\mathcal{M} = I_2$ but let $\phi(x, y) = \tanh(xy)$, which satisfies Condition 2.ii. The cross-derivative of channel i, namely $\partial_1\partial_2\phi(c_i, c_i')$, now varies with (c_i, c_i'), so the two channels generate a family of diagonal cross-derivative matrices spanning all diagonal matrices. The mixing of Case (a) then fails: the transformed score $\tanh((c_1 + c_2)c_1') + \tanh(c_2(c_2' - c_1'))$ has a nonzero cross-derivative in (c_2, c_1'), which no score with mask I_2 has in the original coordinates. Theorem 2 shows that every affine mixing is excluded in this way, leaving only permutations and rescalings of the two channels.

Case (c): Nested supports. Let $\mathcal{M} = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$, so $R_1 = \{1\} \subseteq R_2 = \{1, 2\}$ and Condition 2.i fails. Then $P = \left[\begin{smallmatrix} 1 & 0 \\ a & 1 \end{smallmatrix}\right]$ with $a \neq 0$ maps every matrix supported on $\mathrm{supp}(\mathcal{M})$ to a matrix with the same support, so the combinatorial step of the proof (Lemma 1) cannot exclude adding a multiple of the first concept's cross-derivatives to the second's.

Case (d): Dense mask with bilinear compatibility. Suppose $\mathcal{M}$ is the 2×2 all-ones matrix and $\phi(x, y) = xy$. Then $S = (c_1 + c_2)(c'_1 + c'_2)$, which depends *only* on the sums $s^M := c_1 + c_2$ and $s^P := c'_1 + c'_2$. Consequently, any reparameterization that preserves these sums yields the same observed $p^{\mathcal{D}}$; for instance $(\tilde{c}_1, \tilde{c}_2) = (c_1 + ac_2,\ (1 - a)c_2)$ with $a \neq 1$, keeping the all-ones mask.

Takeaway. The toy case isolates the two ingredients of component-wise identifiability: a compatibility function whose cross-derivative varies, which turns a single matrix equation into a spanning family, and a support pattern in which no concept's footprint is contained in another's, fitted with no more interactions than necessary. Case (a) shows that the second ingredient without the first is not enough, which is why our bilinear implementation is covered by the subspace result rather than the component-wise one.

A4.4 Theorem 1: Non-Identifiability Without Structural Constraints

Before establishing positive identifiability, we show that structural constraints on the interaction mask are indispensable: without them, the latent concepts and mask are not identifiable from the training distribution $p^{\mathcal{D}}(x^M, x^P)$, even under Assumptions 1–3.

Theorem 1 (Non-identifiability of bilinear interactions)**.** *Suppose Assumptions 1–3 hold and $\phi(x, y) = xy$. Then for any training distribution $p^{\mathcal{D}}(x^M, x^P)$ there exist infinitely many distinct tuples $(g^M, g^P, \mathcal{M}, \phi)$ all inducing the same $p^{\mathcal{D}}(x^M, x^P)$. Moreover, if $K_M \geq 2$ and $\mathcal{M} \neq 0$, some of these tuples satisfy $\widetilde{\mathcal{M}} = \mathcal{M}$ while their concepts mix at least two of the original concepts. Hence, in the bilinear subclass the concepts are non-identifiable, and constraining the mask, for example to be sparse with antichain supports, does not remove the ambiguity.*

Discussion. This theorem is the appendix counterpart of Theorem 1 in the main text. Without structural restrictions on $\mathcal{M}$, the latent coordinate systems remain unconstrained. The resulting ambiguity is exactly the large $\mathrm{GL}(K_M) \times \mathrm{GL}(K_P)$ reparameterization orbit discussed in Section 2.2: the same training distribution can be explained by infinitely many equivalent latent coordinate systems.

A4.4.1 Proof Sketch

Fix any valid tuple $(g^M, g^P, \mathcal{M}, \phi)$ consistent with $p^{\mathcal{D}}$. In the proof only, let e^M and e^P denote auxiliary concept-recovery functions witnessing Assumption 1. For any pair of invertible linear maps $T_M \in \mathrm{GL}(K_M)$ and $T_P \in \mathrm{GL}(K_P)$, set $\tilde{\mathbf{c}}^M = T_M \mathbf{c}^M$ and $\tilde{\mathbf{c}}^P = T_P \mathbf{c}^P$, and define

$$\begin{aligned} \tilde{g}^M(\mathbf{c}, \varepsilon^M) &:= g^M(T_M^{-1}\mathbf{c}, \varepsilon^M), \quad & \tilde{e}^M(x^M) &:= T_M e^M(x^M), \\ \tilde{g}^P(\mathbf{c}, \varepsilon^P) &:= g^P(T_P^{-1}\mathbf{c}, \varepsilon^P), & \tilde{e}^P(x^P) &:= T_P e^P(x^P), \\ \tilde{\mathcal{M}} &:= T_M^{-T} \mathcal{M}\, T_P^{-1}, & \tilde{\phi} &:= \phi. \end{aligned}$$

Then the binding score in (12) is preserved:

$$\sum_{i,j} \tilde{\mathcal{M}}_{ij} \phi(\tilde{c}_i^M, \tilde{c}_j^P) = \sum_{i,j} \mathcal{M}_{ij} \phi(c_i^M, c_j^P),$$

so the selection factor is unchanged. The transformed recovery functions show that Assumption 1 remains satisfied, and the nuisance variables enter only through the reparameterized observation maps. After integrating those variables out, the reparameterized tuple induces the same training distribution $p^{\mathcal{D}}(x^M, x^P)$. Since $\mathrm{GL}(K_M) \times \mathrm{GL}(K_P)$ is an infinite group, infinitely many distinct transformations yield the same $p^{\mathcal{D}}$. Critically, these transformed masks generally violate the structural restrictions required for positive recovery, so the ambiguity is ruled out only after such restrictions are imposed.

A4.4.2 Intuitive Example

The dense case in App. A4.3 is the smallest example. With a 2×2 all-ones mask and $\phi(x, y) = xy$, the binding score becomes

$$S(\mathbf{c}^M, \mathbf{c}^P) = (c_1^M + c_2^M)(c_1^P + c_2^P).$$

The selected distribution therefore sees only the molecule-side sum and the protein-side sum, not the two coordinates separately. Different latent coordinates that preserve these sums, or more generally invertible reparameterizations absorbed by a transformed mask, produce the same binder-selected observations. This is why dense holistic interaction leaves the concepts and mask underdetermined.

A4.4.3 Full Proof of Theorem 1

Proof. We construct an explicit infinite family of tuples all producing the same $p^{\mathcal{D}}$.

Fix any reference tuple $(g^M, g^P, \mathcal{M}, \phi)$ compatible with $p^{\mathcal{D}}$, and let e^M, e^P be auxiliary functions witnessing Assumption 1 on the data support. For clarity, consider the bilinear case $\phi(x, y) = xy$; the argument extends to compatibility functions that are locally linear in each argument on the relevant support.

Pick any $(T_M, T_P) \in \mathrm{GL}(K_M) \times \mathrm{GL}(K_P)$. Set $\tilde{\mathbf{c}}^M = T_M \mathbf{c}^M$ and $\tilde{\mathbf{c}}^P = T_P \mathbf{c}^P$, and define the reparameterized tuple:

$$\tilde{g}^M(\tilde{\mathbf{c}}^M, \varepsilon^M) := g^M(T_M^{-1}\tilde{\mathbf{c}}^M, \varepsilon^M), \quad \tilde{g}^P(\tilde{\mathbf{c}}^P, \varepsilon^P) := g^P(T_P^{-1}\tilde{\mathbf{c}}^P, \varepsilon^P), \quad \widetilde{\mathcal{M}} := T_M^{-\top}\, \mathcal{M}\, T_P^{-1}. \tag{21}$$

The corresponding auxiliary recovery functions are $\tilde{e}^M(x^M) := T_M e^M(x^M)$ and $\tilde{e}^P(x^P) := T_P e^P(x^P)$, so the reparameterized tuple also satisfies Assumption 1. Setting $\tilde{\mathbf{c}}^M = T_M \mathbf{c}^M$ and $\tilde{\mathbf{c}}^P = T_P \mathbf{c}^P$:

$$\sum_{i,j} \widetilde{\mathcal{M}}_{ij} \tilde{c}_i^M \tilde{c}_j^P = \tilde{\mathbf{c}}^{M\top} \widetilde{\mathcal{M}} \tilde{\mathbf{c}}^P = (T_M \mathbf{c}^M)^\top T_M^{-\top} \mathcal{M} T_P^{-1} (T_P \mathbf{c}^P) = \mathbf{c}^{M\top} \mathcal{M} \mathbf{c}^P.$$

Hence the binding probability $P(B = 1 \mid \mathbf{c}^M, \mathbf{c}^P)$ is identical for the two tuples. Therefore the selected joint over observations, with nuisance variables integrated out, is identical:

$$p^{\mathcal{D}}(x^M, x^P) = \iint p(x^M \mid \mathbf{c}^M) p(x^P \mid \mathbf{c}^P) p^{\mathcal{D}}(\mathbf{c}^M, \mathbf{c}^P)\, d\mathbf{c}^M\, d\mathbf{c}^P,$$

where $p(x^M \mid \mathbf{c}^M) = \int p(x^M \mid \mathbf{c}^M, \varepsilon^M) p(\varepsilon^M)\, d\varepsilon^M$ and analogously for x^P. The same expression is obtained under the transformed tuple, so $\tilde{p}^{\mathcal{D}}(x^M, x^P) = p^{\mathcal{D}}(x^M, x^P)$.

Since $\mathrm{GL}(K_M) \times \mathrm{GL}(K_P)$ is a connected Lie group of dimension $K_M^2 + K_P^2$, this parameterizes an uncountable family of equivalent tuples, which proves the first part.

For the second part, let $K_M \geq 2$ and $\mathcal{M} \neq 0$. It suffices to find $A \in \mathrm{GL}(K_M)$ that is *not* a permutation matrix times a diagonal matrix, together with $B \in \mathrm{GL}(K_P)$, such that $A^\top \mathcal{M} B = \mathcal{M}$: setting $T_M := A^{-1}$ and $T_P := B^{-1}$ then gives $\widetilde{\mathcal{M}} = T_M^{-\top} \mathcal{M} T_P^{-1} = A^\top \mathcal{M} B = \mathcal{M}$, while $\tilde{\mathbf{c}}^M = A^{-1}\mathbf{c}^M$ mixes at least two concepts, because the inverse of a permutation times a diagonal matrix is again of that form.

- If $\mathrm{rank}(\mathcal{M}) < K_M$, pick $\mathbf{w} \neq 0$ with $\mathbf{w}^\top \mathcal{M} = 0$, an index k with $w_k \neq 0$, and any $m \neq k$, and set $\mathbf{u} := w_m \mathbf{e}_k - w_k \mathbf{e}_m \neq 0$, so that $\mathbf{w}^\top \mathbf{u} = 0$. Let $A^\top := I + \varepsilon\, \mathbf{u}\mathbf{w}^\top$ and $B := I$. Then $A^\top \mathcal{M} = \mathcal{M}$; moreover $(\mathbf{u}\mathbf{w}^\top)^2 = \mathbf{u}(\mathbf{w}^\top \mathbf{u})\mathbf{w}^\top = 0$, so $A^\top$ is invertible; and for small $\varepsilon > 0$ its column k has both the diagonal entry $1 + \varepsilon u_k w_k$ and the off-diagonal entry $-\varepsilon w_k^2$ in row m nonzero.
- If $\mathrm{rank}(\mathcal{M}) = K_M$, take any $A \in \mathrm{GL}(K_M)$ that is not a permutation times a diagonal matrix, let $\mathcal{M}^+$ be the Moore–Penrose pseudoinverse, and set $B := \mathcal{M}^+ A^{-\top} \mathcal{M} + (I - \mathcal{M}^+ \mathcal{M})$. Since $\mathcal{M}\mathcal{M}^+ = I$, we get $\mathcal{M}B = A^{-\top}\mathcal{M}$, i.e. $A^\top \mathcal{M} B = \mathcal{M}$. Moreover B is the identity on $\ker \mathcal{M}$ and, on the row space $\mathcal{R} = \mathrm{range}(\mathcal{M}^\top)$, equals the composition of the bijections $\mathcal{M}|_{\mathcal{R}} : \mathcal{R} \to \mathbb{R}^{K_M}$, $A^{-\top}$, and $\mathcal{M}^+ : \mathbb{R}^{K_M} \to \mathcal{R}$; as $\mathbb{R}^{K_P} = \mathcal{R} \oplus \ker \mathcal{M}$, B is invertible.

In both cases the mask, hence its support, sparsity level and antichain property, is unchanged. The example $\mathcal{M} = I_2$, $A^{-1} = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, $B = A^{-\top}$ is Case (a) of App. A4.3.

Thus, with a bilinear compatibility function the reparameterizations cannot be distinguished by $p^{\mathcal{D}}$, and no constraint on $\mathcal{M}$ alone removes the ambiguity; a non-bilinear ϕ (Condition 2.ii) is what Theorem 2 adds. □

A4.5 Theorem 2: Component-Wise Identifiability Under Sparse Interaction

We now prove the central identifiability result. The argument has two parts. A cross-derivative step turns the equality of the two selected distributions into a family of matrix equations, one for each point of the concept support. A combinatorial step (Lemma 1) then shows that, for a sparse antichain support, this family admits only permutations and rescalings as solutions. The combinatorial step follows the sparsity-based identifiability arguments of Lachapelle et al. [2022] and Zheng et al. [2022].

Theorem 2 (Component-wise identifiability under sparse interaction)**.** *Suppose Assumptions 1–4 hold for the true tuple $(g^M, g^P, \mathcal{M}, \phi)$ and for an alternative tuple $(\tilde{g}^M, \tilde{g}^P, \widetilde{\mathcal{M}}, \tilde{\phi})$ with $\tilde{\phi} \in C^2$ inducing the same $p^{\mathcal{D}}(x^M, x^P)$, and that Conditions 2.i–2.iii of Theorem 2 hold. Then there exist permutations π of $[K_M]$ and ς of $[K_P]$ such that each $\tilde{c}^M_{\pi(i)}$ is an affine function of c^M_i alone, each $\tilde{c}^P_{\varsigma(j)}$ is an affine function of c^P_j alone, and* $\operatorname{supp}(\widetilde{\mathcal{M}}) = \{(\pi(i), \varsigma(j)) : (i,j) \in \operatorname{supp}(\mathcal{M})\}$.

Discussion. This is the appendix counterpart of Theorem 2 and the main positive result of the paper. Condition 2.iii reduces the problem to the linear ambiguity of Theorem 1. Condition 2.ii supplies variation: it turns the single matrix equation available for a bilinear compatibility function into a family of equations spanning all matrices supported on $\operatorname{supp}(\mathcal{M})$. Condition 2.i and the minimality clause supply the combinatorial constraint that this family can satisfy only through permutations and rescalings. A single matrix equation is never enough, whatever the support of the mask (Example 1(a)).

A4.5.1 A Sparsity Lemma

For a set $S \subseteq [K_M] \times [K_P]$, let $\mathbb{R}^S$ denote the space of $K_M \times K_P$ matrices supported in S, and write $R_k(S) = \{l : (k,l) \in S\}$ and $C_l(S) = \{k : (k,l) \in S\}$ for its row and column supports.

Lemma 1 (Antichain recovery)**.** *Let $S \subseteq [K_M] \times [K_P]$ have nonempty row supports forming an antichain and nonempty column supports forming an antichain. Let $P \in \mathrm{GL}(K_M)$, $Q \in \mathrm{GL}(K_P)$, and $\widetilde{S} \subseteq [K_M] \times [K_P]$ with $|\widetilde{S}| \leq |S|$. Suppose there is a family $\{T_t\}_{t \in \mathcal{T}} \subseteq \mathbb{R}^S$ that spans $\mathbb{R}^S$ and satisfies $\operatorname{supp}(PT_tQ) \subseteq \widetilde{S}$ for every $t \in \mathcal{T}$. Then there exist permutations π of $[K_M]$ and ς of $[K_P]$ such that $P_{ik} \neq 0$ if and only if $i = \pi(k)$, and $Q_{lj} \neq 0$ if and only if $j = \varsigma(l)$; that is, P and Q are each a permutation matrix times an invertible diagonal matrix. Moreover $\widetilde{S} = \{(\pi(k), \varsigma(l)) : (k,l) \in S\}$.*

Proof. The linear map $L(X) = PXQ$ is injective. Since the T_t span $\mathbb{R}^S$ and $L(T_t) \in \mathbb{R}^{\widetilde{S}}$ for every t, linearity gives $L(\mathbb{R}^S) \subseteq \mathbb{R}^{\widetilde{S}}$. Comparing dimensions, $|S| \leq |\widetilde{S}| \leq |S|$, so $|\widetilde{S}| = |S|$ and $L(\mathbb{R}^S) = \mathbb{R}^{\widetilde{S}}$.

For $(k,l) \in S$ the matrix unit E_{kl} lies in $\mathbb{R}^S$, and $L(E_{kl}) = \mathbf{p}_k \mathbf{q}_l^\top$, where $\mathbf{p}_k$ is the k-th column of P and $\mathbf{q}_l^\top$ the l-th row of Q. Hence

$$\operatorname{supp}(\mathbf{p}_k) \times \operatorname{supp}(\mathbf{q}_l) \subseteq \widetilde{S} \qquad \text{for all } (k,l) \in S. \tag{22}$$

Since $\det P = \sum_\pi \operatorname{sgn}(\pi) \prod_k P_{\pi(k)k} \neq 0$, some permutation π has $P_{\pi(k)k} \neq 0$ for all k; likewise some ς has $Q_{l\varsigma(l)} \neq 0$ for all l. By (22), $\Gamma := \{(\pi(k), \varsigma(l)) : (k,l) \in S\} \subseteq \widetilde{S}$. As $(k,l) \mapsto (\pi(k), \varsigma(l))$ is injective, $|\Gamma| = |S| = |\widetilde{S}|$, so $\widetilde{S} = \Gamma$.

Suppose $P_{ik} \neq 0$ for some $i \neq \pi(k)$, and put $k' = \pi^{-1}(i) \neq k$. For every $l \in R_k(S)$, (22) gives $(i, \varsigma(l)) \in \widetilde{S} = \Gamma$, hence $(k', l) \in S$, i.e. $l \in R_{k'}(S)$. Thus $R_k(S) \subseteq R_{k'}(S)$, contradicting the row

antichain; so column k of P has the single nonzero entry $P_{\pi(k)k}$. Symmetrically, if $Q_{lj} \neq 0$ for some $j \neq \varsigma(l)$, put $l' = \varsigma^{-1}(j) \neq l$; for every $k \in C_l(S)$, (22) gives $(\pi(k), j) \in \Gamma$, hence $(k, l') \in S$, so $C_l(S) \subseteq C_{l'}(S)$, contradicting the column antichain. □

Example 1 (Each hypothesis of Lemma 1 is needed)**.** *Let $K_M = K_P = 2$.*

(a) Spanning family. *Let $S = \{(1,1),(2,2)\}$, $P = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ and $Q = P^{-1}$. The single matrix $T = I_2$ gives $PTQ = I_2$, supported in $\widetilde{S} = S$, although P is not a permutation times a diagonal matrix. The family $\{I_2\}$ does not span $\mathbb{R}^S$: for $T = \mathrm{diag}(2,5)$ one gets $PTQ = \left[\begin{smallmatrix} 2 & 3 \\ 0 & 5 \end{smallmatrix}\right]$, which leaves $\widetilde{S}$. This is exactly the bilinear situation, where the cross-derivative matrix is the constant $\mathcal{M}$.*

(b) Antichain. *Let $S = \{(1,1),(2,1),(2,2)\}$, so $R_1 = \{1\} \subseteq R_2 = \{1,2\}$, and let $P = \left[\begin{smallmatrix} 1 & 0 \\ a & 1 \end{smallmatrix}\right]$ with $a \neq 0$, $Q = I_2$. Then $PTQ \in \mathbb{R}^S$ for every $T \in \mathbb{R}^S$.*

(c) Sparsity minimality. *If $\widetilde{S} = [K_M] \times [K_P]$, every P and Q satisfy the support condition.*

A4.5.2 Full Proof of Theorem 2

Proof. Write $S(\mathbf{c}^M, \mathbf{c}^P) = \sum_{k,l} \mathcal{M}_{kl} \phi(c_k^M, c_l^P)$ and $\widetilde{S}(\tilde{\mathbf{c}}^M, \tilde{\mathbf{c}}^P) = \sum_{k,l} \widetilde{\mathcal{M}}_{kl} \tilde{\phi}(\tilde{c}_k^M, \tilde{c}_l^P)$, and let e^M, e^P and $\tilde{e}^M, \tilde{e}^P$ be the recovery maps of Assumption 1 for the two tuples.

Step 1: Pull-back to concept coordinates. By Eq. (17) and Assumption 1, B depends on the observations only through $\mathbf{c}^M = e^M(x^M)$ and $\mathbf{c}^P = e^P(x^P)$. By Assumption 2 and the DAG in Eq. (13), $(\mathbf{c}^M, \varepsilon^M)$ is independent of $(\mathbf{c}^P, \varepsilon^P)$, so $p(x^M, x^P) = p(x^M)p(x^P)$. Bayes' rule and Assumption 4 then give

$$\log p^{\mathcal{D}}(x^M, x^P) = \log p(x^M) + \log p(x^P) + S\big(e^M(x^M), e^P(x^P)\big) + \log \kappa - \log P(B{=}1). \quad (23)$$

The same identity holds for the alternative tuple with its own marginals, constant, and recovery maps. Subtracting,

$$S\big(e^M(x^M), e^P(x^P)\big) - \widetilde{S}\big(\tilde{e}^M(x^M), \tilde{e}^P(x^P)\big) = \alpha(x^M) + \beta(x^P)$$

for some α, β. By Condition 2.iii, $\tilde{e}^M = T_M e^M + \mathbf{t}_M$ and $\tilde{e}^P = T_P e^P + \mathbf{t}_P$, so the left-hand side depends on x^M only through $e^M(x^M)$; fixing x^P shows the same for α, and symmetrically for β. Hence, on the concept support,

$$S(\mathbf{c}^M, \mathbf{c}^P) - \widetilde{S}\big(T_M \mathbf{c}^M + \mathbf{t}_M,\, T_P \mathbf{c}^P + \mathbf{t}_P\big) = a(\mathbf{c}^M) + b(\mathbf{c}^P). \quad (24)$$

Under the logistic link, S in (23) would be replaced by $\log \sigma(S)$, whose cross-derivatives are dense; Assumption 4 is what lets the support of $\mathcal{M}$ survive the next step.

Step 2: Cross-derivatives. By Assumptions 2 and 3 the concept support is a product of open sets, so we may differentiate (24). Apply $\partial^2 / \partial c_i^M \partial c_j^P$. The right-hand side vanishes. In S only the term (i,j) depends on both c_i^M and c_j^P, so $\partial^2 S / \partial c_i^M \partial c_j^P = \mathcal{M}_{ij}\, \partial_1 \partial_2 \phi(c_i^M, c_j^P)$; by the chain rule the second term contributes $\sum_{k,l} (T_M)_{ki} (T_P)_{lj} \widetilde{\mathcal{M}}_{kl}\, \partial_1 \partial_2 \tilde{\phi}(\tilde{c}_k^M, \tilde{c}_l^P)$. Define

$$T(\mathbf{c})_{ij} := \mathcal{M}_{ij}\, \partial_1 \partial_2 \phi(c_i^M, c_j^P), \qquad \widetilde{T}(\mathbf{c})_{kl} := \widetilde{\mathcal{M}}_{kl}\, \partial_1 \partial_2 \tilde{\phi}(\tilde{c}_k^M, \tilde{c}_l^P),$$

with $\mathbf{c} = (\mathbf{c}^M, \mathbf{c}^P)$. Then $T(\mathbf{c}) \in \mathbb{R}^S$ for $S := \mathrm{supp}(\mathcal{M})$, $\mathrm{supp}(\widetilde{T}(\mathbf{c})) \subseteq \widetilde{S} := \mathrm{supp}(\widetilde{\mathcal{M}})$, and $T(\mathbf{c}) = T_M^\top \widetilde{T}(\mathbf{c}) T_P$, that is,

$$\widetilde{T}(\mathbf{c}) = P\, T(\mathbf{c})\, Q, \qquad P := T_M^{-\top}, \quad Q := T_P^{-1}, \quad (25)$$

for every $\mathbf{c}$ in the support, with P and Q independent of $\mathbf{c}$.

Step 3: The cross-derivative matrices span $\mathbb{R}^S$. Suppose not. Then some nonzero $W \in \mathbb{R}^S$ satisfies $\sum_{(i,j) \in S} W_{ij} \mathcal{M}_{ij}\, \partial_1 \partial_2 \phi(c_i^M, c_j^P) = 0$ for all $\mathbf{c}$ in the open support. Pick $(i,j) \in S$ with $W_{ij} \neq 0$ and apply $\partial^2 / \partial c_i^M \partial c_j^P$; again only the term (i,j) survives, so $W_{ij} \mathcal{M}_{ij}\, \partial_1^2 \partial_2^2 \phi(c_i^M, c_j^P) = 0$ on a nonempty open set. By Condition 2.ii, $\partial_1^2 \partial_2^2 \phi$ is real-analytic and not identically zero, so it cannot vanish on a nonempty open set; since $W_{ij} \mathcal{M}_{ij} \neq 0$, this is a contradiction.

Step 4: Combinatorial step. By Condition 2.i, S has nonempty row and column supports forming antichains, and by the minimality clause of Condition 2.iii, $|\widetilde{S}| \le |S|$. By Steps 2 and 3 the family $\{T(\mathbf{c})\}$ spans $\mathbb{R}^S$ and satisfies $\operatorname{supp}(PT(\mathbf{c})Q) \subseteq \widetilde{S}$. Lemma 1 therefore yields permutations π, ς with $P_{ik} \neq 0 \iff i = \pi(k)$, $Q_{lj} \neq 0 \iff j = \varsigma(l)$, and $\widetilde{S} = \{(\pi(k), \varsigma(l)) : (k,l) \in S\}$.

Step 5: Conclusion. Since $P = T_M^{-\top}$ has the single nonzero entry $P_{\pi(k)k}$ in each column k, the matrix $T_M = P^{-\top}$ has the single nonzero entry $(T_M)_{\pi(k)k} = 1/P_{\pi(k)k}$ in each column k; hence $\tilde{c}^M_{\pi(k)} = (T_M)_{\pi(k)k} c^M_k + (\mathbf{t}_M)_{\pi(k)}$ depends on c^M_k alone. Likewise $T_P = Q^{-1}$ gives $\tilde{c}^P_{\varsigma(l)}$ as an affine function of c^P_l alone. Together with the support identity of Step 4, the two tuples differ only by concept permutations, component-wise invertible maps, and the scale shared between mask and compatibility function, which are exactly the equivalences allowed in Definition 1. □

Remark 1 (Which conditions are needed)**.** *Condition 2.ii cannot be dropped: for $\phi(x,y) = xy$ the second part of Theorem 1 gives counterexamples satisfying Conditions 2.i and 2.iii, the smallest being Case (a) of App. A4.3. Condition 2.i and the minimality clause cannot be dropped from Lemma 1 (Example 1). They are, however, not necessary for identifiability as such: Proposition 1 obtains it without any sparsity, under a stronger nonlinearity condition that depends on the mask and must hold at every point of the support. What the sparse antichain structure buys is that the analytic requirement, Condition 2.ii, is a generic property of ϕ alone, while the combinatorial requirement can be checked directly on a learned mask.*

A4.5.3 Beyond Linear Concept Ambiguity

Condition 2.iii restricts the competing parameterizations to affine reparameterizations of the concepts. The pairwise form of the binding score also constrains general nonlinear reparameterizations, by an argument similar to the one for additive decoders [Lachapelle et al., 2023].

Proposition 1 (Nonlinear reparameterizations)**.** *Suppose Assumptions 1–4 hold for both tuples, the support of $p(\mathbf{c}^M)$ is open and convex, $\phi \in C^3$, $\tilde{\phi} \in C^1$, and the concepts of the two tuples are related by C^2 diffeomorphisms $\tilde{\mathbf{c}}^M = F(\mathbf{c}^M)$ and $\tilde{\mathbf{c}}^P = G(\mathbf{c}^P)$. Fix protein-side probes $\mathbf{b}_1, \dots, \mathbf{b}_R$ in the support of $\mathbf{c}^P$, write $b_{r,j}$ for the j-th coordinate of $\mathbf{b}_r$, and define for each $i \in [K_M]$*

$$\mathbf{h}_i(x) := \big(\mathcal{M}_{ij}\, \partial_2 \phi(x, b_{r,j})\big)_{r \in [R],\, j \in [K_P]} \in \mathbb{R}^{RK_P}.$$

If, for every $\mathbf{c}^M$ in the support, the $2K_M$ vectors $\{\mathbf{h}_i'(c_i^M), \mathbf{h}_i''(c_i^M)\}_{i=1}^{K_M}$ are linearly independent, then there is a permutation π of $[K_M]$ such that each $F_{\pi(i)}$ depends on c_i^M alone. The symmetric condition with molecule-side probes gives the same conclusion for G.

Proof. As in Step 1 above, with F, G in place of the affine maps, $S(\mathbf{c}^M, \mathbf{c}^P) - \widetilde{S}(F(\mathbf{c}^M), G(\mathbf{c}^P)) = a(\mathbf{c}^M) + b(\mathbf{c}^P)$. Differentiate with respect to c_j^P and evaluate at $\mathbf{c}^P = \mathbf{b}_r$. On the left, $\partial S / \partial c_j^P(\mathbf{c}^M, \mathbf{b}_r) = \sum_i \mathcal{M}_{ij} \partial_2 \phi(c_i^M, b_{r,j})$, so stacking over (r,j) gives $\mathbf{h}(\mathbf{c}^M) := \sum_i \mathbf{h}_i(c_i^M)$. On the right, the corresponding derivative equals $\sum_l \partial_j G_l(\mathbf{b}_r) \sum_k \widetilde{\mathcal{M}}_{kl} \partial_2 \tilde{\phi}(F_k(\mathbf{c}^M), G_l(\mathbf{b}_r))$, a sum of functions of the single coordinates $F_k(\mathbf{c}^M)$, while $\partial b / \partial c_j^P(\mathbf{b}_r)$ is constant. Hence $\mathbf{h} = \widetilde{\mathbf{H}} \circ F$ for some $\widetilde{\mathbf{H}}(\mathbf{u}) = \sum_k \widetilde{\mathbf{H}}_k(u_k)$ additive in $\mathbf{u}$.

Let $V := F^{-1}$. Then $\widetilde{\mathbf{H}} = \mathbf{h} \circ V$ is C^2 and additive, so for $k \neq k'$,

$$0 = \frac{\partial^2}{\partial u_k \partial u_{k'}} \sum_i \mathbf{h}_i(V_i(\mathbf{u})) = \sum_i \Big[\mathbf{h}_i''(V_i)\, \partial_k V_i\, \partial_{k'} V_i + \mathbf{h}_i'(V_i)\, \partial^2_{kk'} V_i\Big].$$

By the linear independence assumption, $\partial_k V_i\, \partial_{k'} V_i = 0$ for all i and $k \neq k'$, so each row of $J_V(\mathbf{u})$ has at most one nonzero entry; as J_V is invertible, it is a permutation times a diagonal matrix, and so is $J_F(\mathbf{c}^M) = J_V(F(\mathbf{c}^M))^{-1}$. The pattern is locally constant by continuity, hence constant on the connected support, so there is π with $\partial F_{\pi(i)} / \partial c_k^M \equiv 0$ for $k \neq i$; convexity of the support then makes $F_{\pi(i)}$ a function of c_i^M alone. The argument for G is symmetric. □

The linear independence condition requires $RK_P \ge 2K_M$. It fails for a bilinear ϕ, for which $\mathbf{h}_i'' \equiv 0$, and it fails wherever two concepts whose rows of $\mathcal{M}$ are proportional take the same value, since their

vectors then coincide up to scale; distinct row supports, as in Condition 2.i, remove this degeneracy. If in addition $\|\widetilde{\mathcal{M}}\|_0 \leq \|\mathcal{M}\|_0$ and $\partial_1\partial_2\phi \neq 0$ on the support, the cross-derivative identity of Step 2 shows that $\mathrm{supp}(\widetilde{\mathcal{M}})$ is the permuted support of $\mathcal{M}$, as in Theorem 2.

A4.6 Theorem 3: Subspace Identifiability Under Low-Rank Interaction

Condition 2.i requires an exactly sparse antichain support and Condition 2.ii a non-bilinear compatibility function, both of which may be too strong in practice. For protein-molecule binding, we argue that the subspace-identifiability alternative in Condition 3.i is also natural empirically: the binding kernel has low rank.

Motivation. Protein–molecule binding in drug discovery is dominated by a small number of pharmacophoric motifs: hydrogen bonds, π-stacking, hydrophobic packing, and electrostatic complementarity. It is therefore physically reasonable that the complete binding response can be decomposed into r independent *interaction modes*, each characterized by a pair (u_k, v_k) describing how each modality contributes to mode k. We expect r to be small; typical estimates from medicinal chemistry place r at 5–20 for a given target family.

Why this replaces SEDGE's conservative assumption. SEDGE's identifiability argument relies on $P(S{=}0) \approx 0$, i.e., "unseen configurations are rare." This is false for binding: most protein-molecule pairs do *not* bind, so $P(B{=}0) \gg P(B{=}1)$. Condition 3.i replaces this with a physics-motivated structural constraint on the kernel itself, preserving identifiability without requiring the pathological SEDGE-style coverage condition.

Architectural implication. Under Condition 3.i, the bilinear binding score can be implemented with a rank-r factorized mask $\mathcal{M} = UV^\top$ where $U \in \mathbb{R}^{K_M \times r}$ and $V \in \mathbb{R}^{K_P \times r}$, reducing parameters from $K_M K_P$ to $r(K_M + K_P)$ and naturally enforcing the low-rank prior. This is the structural basis for our low-rank concept-basis implementation in App. A5.3; the main concept-pair implementation in App. A5.2 can also reach this regime empirically through regularization-induced low rank.

Subspace identifiability under low-rank interaction. When Conditions 2.i–2.iii are replaced by Conditions 3.i–3.ii, the full concept-pair mask need not be identifiable. However, an analog of Theorem 2 still holds at the level of the rank-r *interaction subspaces*:

Theorem 3 (Subspace identifiability under low-rank interaction)**.** *Suppose Assumptions 1–4 and Conditions 3.i–3.ii of Theorem 3 hold, so the centred binding score S° of Condition 3.i admits a minimal rank-r separable factorization*

$$S^\circ(\mathbf{c}^M, \mathbf{c}^P) = \sum_{\ell=1}^{r} u_\ell(\mathbf{c}^M) v_\ell(\mathbf{c}^P) = \mathbf{u}(\mathbf{c}^M)^\top \mathbf{v}(\mathbf{c}^P), \qquad r < \min\{K_M, K_P\},$$

with linearly independent mode families $\{u_\ell\}_{\ell=1}^r \subset L^2(p^{\mathcal{D}}(\mathbf{c}^M))$ and $\{v_\ell\}_{\ell=1}^r \subset L^2(p^{\mathcal{D}}(\mathbf{c}^P))$. Assume the low-rank sufficient-variability condition: there exist probe values $\mathbf{a}_1, \ldots, \mathbf{a}_{R_M} \in \mathrm{int}(\mathrm{supp}(\mathbf{c}^M))$ and $\mathbf{b}_1, \ldots, \mathbf{b}_{R_P} \in \mathrm{int}(\mathrm{supp}(\mathbf{c}^P))$, with $R_M, R_P \geq r$, such that

$$\mathbf{U}_A := \begin{bmatrix} u_1(\mathbf{a}_1) & \cdots & u_r(\mathbf{a}_1) \\ \vdots & \ddots & \vdots \\ u_1(\mathbf{a}_{R_M}) & \cdots & u_r(\mathbf{a}_{R_M}) \end{bmatrix} \quad \text{and} \quad \mathbf{V}_B := \begin{bmatrix} v_1(\mathbf{b}_1) & \cdots & v_r(\mathbf{b}_1) \\ \vdots & \ddots & \vdots \\ v_1(\mathbf{b}_{R_P}) & \cdots & v_r(\mathbf{b}_{R_P}) \end{bmatrix}$$

$$\mathrm{rank}(\mathbf{U}_A) = \mathrm{rank}(\mathbf{V}_B) = r.$$

Then the molecule-side and protein-side interaction subspaces

$$\mathcal{U}_M := \mathrm{span}\{u_1, \ldots, u_r\} \subset L^2(p^{\mathcal{D}}(\mathbf{c}^M)), \qquad \mathcal{V}_P := \mathrm{span}\{v_1, \ldots, v_r\} \subset L^2(p^{\mathcal{D}}(\mathbf{c}^P))$$

are subspace-identifiable from $p^{\mathcal{D}}(x^M, x^P)$ in the sense of Definition 1.

Discussion. This is the appendix counterpart of Theorem 3. The two theorem-specific conditions play different roles:

- **Condition 3.i** is the structural low-rank condition. It weakens exact entry-wise sparsity into a mode-level prior, so the identifiable object is the pair of low-dimensional interaction subspaces rather than each individual concept pair.
- **Condition 3.ii** is the sufficient-variability condition for those low-rank modes. It ensures that the selected data excite the mode-evaluation matrices with full rank, which makes the subspaces statistically recoverable.

Together, these conditions define the subspace-identifiability regime relevant to our experiments, where the learned masks are empirically low-rank rather than exactly sparse.

A4.6.1 Proof Sketch

As in the preceding proofs, nuisance variables $\varepsilon^M, \varepsilon^P$ are marginalized into the observation kernels and Assumption 1 pulls the selected observation distribution back to recovered concept coordinates. Since the selected distribution determines the binding score only up to additive functions of a single modality, the argument operates on the rank-r *centred* kernel $S^\circ(\mathbf{c}^M, \mathbf{c}^P)$. If two minimal factorizations induce the same centred kernel, then evaluating it at the full-rank protein-side probes from Condition 3.ii expresses each molecule-side mode basis as a linear combination of the same finite set of kernel sections $S^\circ(\cdot, \mathbf{b}_q)$. Hence the two molecule-side mode families span the same subspace. The full-rank molecule-side probes give the symmetric statement for the protein-side modes. Thus the identifiable object is the pair of interaction subspaces, not an ordered basis of individual modes.

Intuitively, the rank-r subspaces are invariants of the low-rank interaction kernel, even when concept-wise support is not identifiable. This is exactly the subspace-identifiability analog we seek: we identify interaction *subspaces* rather than individual concept-pair entries. The full derivation is given below.

A4.6.2 Intuitive Example

A low-rank interaction can be dense. For example, with

$$\mathbf{u} = \begin{bmatrix} 0.2 \\ 0.8 \\ 1.5 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} 1.0 \\ 0.5 \\ 2.0 \end{bmatrix}, \qquad \mathcal{M} = \mathbf{u}\mathbf{v}^\top = \begin{bmatrix} 0.2 & 0.1 & 0.4 \\ 0.8 & 0.4 & 1.6 \\ 1.5 & 0.75 & 3.0 \end{bmatrix},$$

all nine concept-pair entries are nonzero, so sparse edge-level recovery is not the right target. Nevertheless, every entry is generated by one shared interaction mode: $M_{ij} = u_i v_j$. A larger block mask with three dense 3×3 all-ones blocks is similarly rank 3: it is dense within blocks, but generated by three interaction modes. Theorem 3 captures exactly this situation: the identifiable object is the span of the dominant modes, not the individual visual block labels or every pairwise edge.

A4.6.3 Full Proof of Theorem 3

Proof. As in Step 1 of the proof of Theorem 2 (Eq. (23)), two tuples inducing the same $p^{\mathcal{D}}$ have binding scores that agree, as functions of the observations, up to $\alpha(x^M) + \beta(x^P)$; centring at the reference observations corresponding to $\mathbf{a}_0, \mathbf{b}_0$ cancels these terms, so the two centred kernels agree as functions of the observations. We therefore compare mode functions of the two tuples as functions of the observations, $u_\ell \circ e^M$ and $\tilde{u}_\ell \circ \tilde{e}^M$, suppressing the composition below. The remaining question is whether the rank-r interaction subspaces are uniquely determined by the kernel.

Suppose two minimal rank-r factorizations induce the same centred kernel on the data support:

$$S^\circ(\mathbf{c}^M, \mathbf{c}^P) = \mathbf{u}(\mathbf{c}^M)^\top \mathbf{v}(\mathbf{c}^P) = \tilde{\mathbf{u}}(\mathbf{c}^M)^\top \tilde{\mathbf{v}}(\mathbf{c}^P).$$

Let $\mathcal{U} = \operatorname{span}\{u_1, \ldots, u_r\}$ and $\tilde{\mathcal{U}} = \operatorname{span}\{\tilde{u}_1, \ldots, \tilde{u}_r\}$, and define $\mathcal{V}, \tilde{\mathcal{V}}$ analogously. For each protein-side probe $\mathbf{b}_q$, the kernel section

$$S^\circ(\cdot, \mathbf{b}_q) = \sum_{\ell=1}^{r} v_\ell(\mathbf{b}_q) u_\ell(\cdot) = \sum_{\ell=1}^{r} \tilde{v}_\ell(\mathbf{b}_q) \tilde{u}_\ell(\cdot)$$

belongs to both $\mathcal{U}$ and $\tilde{\mathcal{U}}$. Because $\operatorname{rank}(\mathbf{V}_B) = r$, these sections span $\mathcal{U}$. Since the same r-dimensional span is contained in $\tilde{\mathcal{U}}$ and the alternative factorization is also minimal rank-r, $\dim(\tilde{\mathcal{U}}) = r$; hence $\mathcal{U} = \tilde{\mathcal{U}}$.

The same argument with molecule-side probes $\mathbf{a}_m$ and $\mathrm{rank}(\mathbf{U}_A) = r$ shows $\mathcal{V} = \tilde{\mathcal{V}}$. Since each pair of mode functions is a basis for the same r-dimensional subspace, there exists an invertible matrix $A \in \mathrm{GL}(r)$ such that, on the data support, $\tilde{\mathbf{u}} = A^\top \mathbf{u}$ and $\tilde{\mathbf{v}} = A^{-1}\mathbf{v}$. This change of basis leaves $\mathbf{u}^\top \mathbf{v}$ unchanged but does not change the identified subspaces. Therefore the pair $(\mathcal{U}_M, \mathcal{V}_P)$ is identifiable, while the ordered mode basis is not. □

A4.7 Summary and Discussion of the Theory

The results above establish a coherent theoretical picture for CAUSALBIND. First, the structural causal model clarifies why binder-only training should be treated as a Heckman-style selection problem on a V-structure rather than as ordinary supervised learning. Second, Theorem 1 shows that a bilinear interaction is non-identifiable under this selection mechanism even when the mask is sparse. Third, Theorem 2 shows that a sparse antichain interaction with a non-bilinear compatibility function restores identifiability up to the standard ambiguities of permutation, component-wise reparameterization, and joint rescaling, and Proposition 1 extends this to nonlinear reparameterizations under a stronger nonlinearity condition. Finally, Theorem 3 gives a subspace-identifiability result when exact sparse support is too strong: even if entry-wise mask identification fails, the dominant interaction subspace can still remain identifiable.

What the three results establish together. The theoretical development above should be read as a progression from impossibility, to component-wise identifiability under strong sparse structure, to subspace identifiability under weaker but more realistic low-rank structure. Theorem 1 shows that without additional constraints, binder-only observations do not determine a unique latent interaction mechanism: many dense factorizations can explain the same selected distribution. Theorem 2 then shows that component-wise identifiability becomes possible when the interaction mask is sparse in a combinatorially distinctive way and the selected conditional family varies sufficiently across molecule-side concepts. Finally, Theorem 3 shows that even when exact element-wise sparsity is relaxed, the dominant interaction modes remain identifiable at the subspace level under a low-rank kernel assumption.

Why the causal structure matters. The key source of information is the V-structure

$$\mathbf{c}^M \to B \leftarrow \mathbf{c}^P,$$

not an externally supplied domain index or an auxiliary supervision signal. Before conditioning on binding, the molecule-side and protein-side concepts are independent under Assumption 2; after restricting to $B{=}1$, they become statistically coupled through the collider. The identifiability problem is therefore not solved by adding a second dataset or a separate view, but by exploiting the dependence structure already induced by binder-only selection. The sparse-support theorem uses that dependence to establish component-wise identifiability up to standard equivariances, while the low-rank theorem uses it to establish subspace identifiability of the leading interaction modes when exact support identification is too strong a requirement.

A5 Implementation Details

Architecture overview. For completeness, Figure A2 summarizes the implementation-level view of CAUSALBIND. We place it in the appendix to keep the main text focused on the theory-to-method narrative, while still providing a compact diagram of the encoder, concept-factorization, sparse-interaction, and retrieval-loss pipeline.

A5.1 Relation Between the Two-Modal Theory and the Three-Modal Implementation

The theory in Sections 2 and A4 is deliberately written for the simplest protein-molecule setting with two modalities: one molecule view and one protein view. The implementation, however, uses three modalities: molecule structure, pocket structure, and protein sequence. This does not contradict the theory; rather, it instantiates the same two-modal idea twice. Concretely, the model contains two retrieval branches, molecule-pocket and molecule-sequence, with score matrices $\mathbf{U}^{\mathrm{poc,mol}}$ and $\mathbf{U}^{\mathrm{seq,mol}}$. Each branch is a protein-molecule pairing of exactly the kind analyzed in the theory, and each branch is equipped with its own constrained interaction bottleneck. The final objective simply sums the two pairwise retrieval losses and their associated regularization terms.

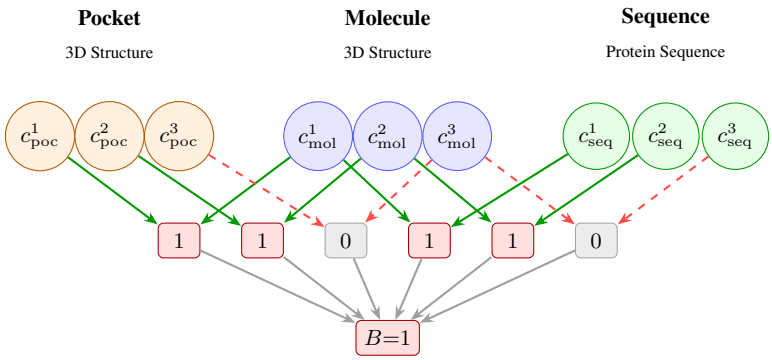


Figure A2: Three-view concept-level schematic of CAUSALBIND. The inputs are pocket structure, molecule structure, and protein sequence. Each view is decomposed into modality-specific concepts c_{poc}, c_{mol}, and c_{seq}. Local binary indicators summarize whether a concept pair is active (1) or blocked (0). We show two interaction branches, pocket-molecule and molecule-sequence, which are then aggregated into the final pair-level binding event B. Green solid arrows denote active local interactions, red dashed arrows denote blocked interactions, and gray arrows denote aggregation to the overall binding variable.

For this reason, the main theoretical claims still apply at the level at which they were intended: they describe why dense holistic cross-modal interaction is non-identifiable, why sparse interaction can be component-wise identifiable, and why low-rank interaction can be subspace-identifiable for each protein-molecule branch individually. The three-stream architecture should therefore be read as a multi-view extension of the same pairwise principle rather than as a different causal problem. The pocket branch and the sequence branch provide two complementary protein views, but neither changes the core structural role of the protein-molecule interaction analyzed in the theory. Put differently, the implementation is richer than the theory in terms of modalities, but not different in its fundamental interaction pattern: it is still a sum of two protein-molecule pairing problems, each governed by the same sparse/low-rank inductive bias.

A5.2 CAUSALBIND$_{\text{SP}}$: Sparse Concept-Pair Implementation

CAUSALBIND$_{\text{SP}}$ is the primary implementation used in the main text. It keeps an explicit $K \times K$ concept-pair mask for each protein-molecule branch and therefore most directly matches the sparse interaction form in Theorem 2. The learned dense mask can also enter the low-rank regime analyzed in Theorem 3, as diagnosed by the singular-value analyses in App. A6.1.7. Unless stated otherwise, CAUSALBIND in the main paper refers to CAUSALBIND$_{\text{SP}}$.

A5.2.1 Technical Supplement for the CAUSALBIND$_{\text{SP}}$ Implementation

This subsection complements Section 3 at the design level. The emphasis here is on how the abstract ingredients in the main text are instantiated in practice: how concept extraction is implemented, how the sparse mask is parameterized and applied, and how the inherited geometric losses fit into the final objective.

***ConceptExtractor* and *SparseMask*.** For the *main sparse concept-pair operating point* used in the paper, we set K=128, d_c=1024, L=4 Perceiver layers, and 8 attention heads per layer; the strongest alternative concept-level settings in the sweeps vary one of these quantities at a time. The default mask type is `tanh_plus_1`, the hard-gating threshold is 0.5, and the effective mask is

$$\mathbf{Z} = \tanh(\widehat{\mathcal{M}}) + 1,$$

followed by hard thresholding so entries at or below the threshold are set to zero exactly in the forward pass. Gradients through the hard gate use a straight-through convention, and the sparsity penalty is applied to the continuous activated masks $\mathbf{Z}$ rather than to the raw mask logits. The main-text operating point uses $\lambda_{\text{spa}} = 10^{-4}$, while the sweep additionally studies $\lambda \in \{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. In all cases, the branchwise penalties for the molecule-pocket and molecule-sequence masks are summed and then scaled by the chosen λ_{spa}.

Key implementation equations. For each branch $(a, b) \in \{(\mathrm{mol}, \mathrm{poc}), (\mathrm{mol}, \mathrm{seq})\}$, the sparse concept-pair implementation can be summarized as

$$\begin{aligned} \mathbf{C}_v &= F_v(x_v) \in \mathbb{R}^{K \times d_c}, \\ \mathbf{Z}^{\mathrm{SP}}_{a,b} &= \tanh(\widehat{\mathcal{M}}_{a,b}) + 1, \qquad \mathcal{M}^{\mathrm{SP}}_{a,b} = H^{\mathrm{ST}}_{\tau}\big(\mathbf{Z}^{\mathrm{SP}}_{a,b}\big), \qquad \widehat{\mathcal{M}}_{a,b} \in \mathbb{R}^{K \times K}, \\ \mathbf{A}_{a,b}[i,j] &= \phi(\mathbf{C}_a[i,:], \mathbf{C}_b[j,:]) = \langle \mathbf{C}_a[i,:], \mathbf{C}_b[j,:] \rangle, \\ u^{\mathrm{SP}}_{a,b} &= \mathrm{Agg}\big(\mathcal{M}^{\mathrm{SP}}_{a,b} \odot \mathbf{A}_{a,b}\big). \end{aligned} \tag{26}$$

Here F_v denotes the backbone-plus-concept-extractor map for modality v, $\tau = 0.5$ is the forward hard-gating threshold, $\mathrm{Agg}(\mathbf{B}) = K^{-1} \sum_{i,j} B_{ij}$ in the implementation, and $u^{\mathrm{SP}}_{a,b}$ is the branch-level score entering the contrastive objective.

Detailed form of $\mathcal{L}_{\mathrm{cone}}$ and $\mathcal{L}_{\mathrm{reg}}$. In the main text we keep the HypSeek-inherited geometric terms abstract to avoid overloading Section 3. Here we make them explicit. Consider assay i with pocket embedding $\mathbf{h}_{\mathrm{poc},i}$ and candidate molecule embeddings $\{\mathbf{h}_{\mathrm{mol},ij}\}_{j \in \mathcal{L}_i}$, where $\mathcal{L}_i$ is the set of molecules paired with assay i in the training batch. Let $N_\ell = \sum_i |\mathcal{L}_i|$ be the total number of pocket-molecule pairs. Following HypSeek [Wang et al., 2026], the Lorentz distance between pocket i and molecule j is denoted d_{ij}, the exterior angle at the pocket is ϕ_{ij}, and the pocket half-aperture is ω_i. Affinity values are discretized into buckets b_{ij}, from which the bucket-specific radial and angular limits are derived as

$$r_{ij} = r_0 + b_{ij}\Delta r, \qquad \eta_{ij} = \eta_0 - b_{ij}\Delta\eta.$$

The cone-hierarchy loss then penalizes violations in both radial distance and angular admissibility:

$$\begin{aligned} L_{\mathrm{rad}} &= \frac{1}{\sqrt{N_\ell}} \sum_{i,j} \max\big(d_{ij} - r_{ij},\, 0\big), \\ L_{\mathrm{ang}} &= \frac{1}{\sqrt{N_\ell}} \sum_{i,j} \max\big(\phi_{ij} - \eta_{ij}\omega_i,\, 0\big), \\ \mathcal{L}_{\mathrm{cone}} &= \lambda_{\mathrm{rad}} L_{\mathrm{rad}} + \lambda_{\mathrm{ang}} L_{\mathrm{ang}}. \end{aligned} \tag{27}$$

The $\sqrt{N_\ell}$ normalization is inherited unchanged from HypSeek; it is a code-level loss scaling convention rather than a new statistical average introduced by CAUSALBIND.

The additional regularization term in our implementation inherits two further components from HypSeek. Concretely, the $\mathcal{L}_{\mathrm{reg}}$ appearing in the main text is exactly the weighted sum of the two terms below:

$$\mathcal{L}_{\mathrm{reg}} = \lambda_{\mathrm{angu}} R_{\mathrm{ang}} + \lambda_{\mathrm{het}} R_{\mathrm{het}}.$$

The first term is an *angular margin regularizer*. It prevents trivial angular collapse by requiring each molecule to remain at least a fixed margin $m_{\mathrm{margin}} > 0$ beyond its cone boundary:

$$R_{\mathrm{ang}} = \frac{1}{\sqrt{N_\ell}} \sum_{i,j} \max\big(\phi_{ij} - \eta_{ij}\omega_i + m_{\mathrm{margin}},\, 0\big). \tag{28}$$

Here the same ϕ_{ij}, η_{ij}, and ω_i as above are reused; the only new notation is the constant angular margin m_{margin}, which is set to 0.15 in the current code.

The second term is an *intra-assay heterogeneity regularizer*. Its role is to encourage the model to assign larger probability mass to the stronger active molecules within the same assay, rather than treating all actives as equally informative. For assay i, let $L_i = |\mathcal{L}_i|$ be the number of paired molecules, and let $d_i = (d_{i1}, \ldots, d_{iL_i})$ be their pocket-molecule distances. We define the within-assay rank of molecule j by

$$\mathrm{rank}_{ij} = 1 + \sum_{j'=1}^{L_i} \mathbf{1}[d_{ij'} < d_{ij}],$$

so smaller distance implies better rank. HypSeek then assigns a rank-based weight

$$w_{ij} = \exp\Big(-\beta\, \frac{\mathrm{rank}_{ij} - 1}{L_i}\Big),$$

where β controls how strongly higher-ranked molecules are emphasized (the implementation uses $\beta = 80.5$). Let ℓ_{ij} denote the molecule-retrieval logit for assay i and molecule j within that assay-specific subset, and define the within-assay softmax probability

$$p_{ij} = \frac{\exp(\ell_{ij})}{\sum_{j'=1}^{L_i} \exp(\ell_{ij'})}.$$

Let v_{ij} be the activity or affinity annotation of molecule j in assay i, and let v_{th} be the HypSeek preprocessing threshold used to mark active compounds. The implementation keeps only molecules with $v_{ij} < v_{\mathrm{th}}$ (equivalently, the positive subset under the assay's activity convention). Writing C for the number of assays that contain at least one such active molecule, the heterogeneity regularizer is

$$R_{\mathrm{het}} = \frac{1}{\max(C, 1)} \sum_{i} \sum_{\{j:\, v_{ij} < v_{\mathrm{th}}\}} - w_{ij} \log p_{ij}. \tag{29}$$

Intuitively, R_{het} is a rank-weighted within-assay negative log-likelihood: stronger actives receive larger weights, so the model is encouraged to separate fine-grained affinity variation inside each pocket-specific candidate set.

Relation to our Section 3 notation. In the current paper, the retrieval loss in Section 3 is written in terms of batch-level score matrices $\mathbf{U}^{\mathrm{poc,mol}}$ and $\mathbf{U}^{\mathrm{seq,mol}}$. The HypSeek terms above operate on the same training pairs after they have been lifted into the hyperbolic branch-specific geometry. In other words, $\mathcal{L}_{\mathrm{cone}}$ and $\mathcal{L}_{\mathrm{reg}}$ do not replace the retrieval objective; they act as geometric regularizers on top of it. This is why the main text groups them separately from $\mathcal{L}_{\mathrm{cont}}$, while the appendix spells out the exact radial, angular, and within-assay ranking quantities used in the inherited implementation.

A5.2.2 Concrete Parameter Assignments and Experimental Setting

We now turn to the code-level side of reproducibility for CausalBind$_{\mathrm{SP}}$: what values are assigned to the key hyperparameters, how the inherited loss weights are instantiated in code, and what training/evaluation settings are shared across the reported runs.

Training setup and optimization. We build on the HypSeek code base [Wang et al., 2026] with pretrained Uni-Mol encoders for molecules and pockets and an ESM-2 encoder for protein sequence. Training follows the same three-branch retrieval pipeline on the `blended ChEMBL+BindingDB` corpus (18,316 assays after removing assays that overlap the FEP benchmarks and every assay whose target UniProt ID appears in DUD-E, LIT-PCBA, or DEKOIS 2.0; as in the LigUnity and HypSeek protocol, the PDBBind structural pairs are not filtered by UniProt ID, so some PDBBind complexes share targets with these benchmarks), with CASF used for validation when validation-selected protocols are needed. Unless noted otherwise, all reported runs train for 50 epochs on 4×A100-80GB GPUs with per-GPU batch size 24 (distributed world size 4, effective global batch size 96), Adam optimizer, learning rate 10^{-4}, polynomial decay, warmup ratio 0.06, weight decay 0, gradient clipping at 1.0, and fp16 mixed precision. In our cluster logs and planning estimates, a typical 50-epoch run takes several hours on 4×A100, so full sweeps scale roughly linearly with the number of variants. The default seed is 1; additional seeds are only introduced in the dedicated robustness checks. All compared retrieval runs (HypSeek, concept-level ablations, concept-level sweeps, and embedding-level variants) share this optimization setup unless a subsection explicitly states otherwise.

Inference and storage. CausalBind remains a retrieval-style dual-tower model: molecule-side and protein-side concept features are computed independently, and the fixed branch masks are applied before producing score matrices. Thus library molecules can still be precomputed rather than passed through a joint protein-molecule network for every target. The explicit $K \times K$ CausalBind$_{\mathrm{SP}}$ masks add a small fixed branch computation, while CausalBind$_{\mathrm{LR}}$ reduces mask parameters from K^2 to $2Kr$ per branch. For billion-scale deployment one would typically combine these representations with approximate-nearest-neighbor indexing or distillation; the present experiments evaluate ranking quality rather than an optimized production index.

Loss weights and the relation between paper notation and code notation. In Section 3, we write the training objective using a compact notation with a sparsity weight λ_{spa} and a grouped geometric term $\lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}$. In the code, however, the inherited HypSeek regularization is implemented through

several explicit coefficients rather than through a single scalar λ_{reg}. Concretely, the cone-hierarchy term uses $\gamma_{\text{chl}} = 0.1$ together with $\lambda_{\text{rad}} = 0.5$ and $\lambda_{\text{ang}} = 0.5$, while the additional regularization bundle uses $\lambda_{\text{angu}} = 0.1$ and $\lambda_{\text{het}} = 0.1$. Thus, the paper's $\mathcal{L}_{\text{reg}}$ should be read as a compact umbrella notation for the inherited angular and heterogeneity regularizers, while λ_{reg} is not a single independently tuned scalar in the current implementation. The only new coefficient introduced by CAUSALBIND itself is λ_{spa}, whose values are reported explicitly in the corresponding experiment tables.

Evaluation protocol. We evaluate on DUD-E (102 targets) and LIT-PCBA (15 targets). For each target we rank all molecules in the test set and compute AUROC, BEDROC ($\alpha = 80.5$), and EF@1%. Per-target metrics are macro-averaged across targets.

Evaluation metrics. **AUROC** measures overall ranking quality across the full ranked list: it is the probability that a randomly chosen active is ranked above a randomly chosen inactive, so it reflects global discrimination rather than early retrieval. **BEDROC**$_{80.5}$ is an early-enrichment metric that exponentially upweights the top of the ranked list; the parameter $\alpha = 80.5$ is the standard virtual-screening setting that places roughly 80% of the weight within the top $\sim 2\%$ of ranks, making the metric highly sensitive to whether true actives appear near the front. **EF@1%** (enrichment factor at the top 1%) compares the number of actives retrieved in the top 1% of the ranked list against the number expected from random ranking. In practical screening, BEDROC$_{80.5}$ and EF@1% are often more informative than AUROC because only a small top-ranked subset is typically advanced to experimental follow-up.

A5.3 CAUSALBIND$_{\text{LR}}$: Low-Rank Concept-Basis Implementation

CAUSALBIND$_{\text{LR}}$ keeps the same concept extractor, retrieval branches, and losses as CAUSALBIND$_{\text{SP}}$, but replaces the unconstrained $K \times K$ mask logits with an explicit rank-r factorization in the concept basis. It is therefore the direct implementation of the low-rank interaction structure in Theorem 3.

A5.3.1 Technical Supplement for the CAUSALBIND$_{\text{LR}}$ Implementation

For each branch $(a, b) \in \{(\text{mol}, \text{poc}), (\text{mol}, \text{seq})\}$, CAUSALBIND$_{\text{LR}}$ parameterizes the raw mask as

$$\widehat{\mathcal{M}}_{a,b} = \mathbf{U}_{a,b}\mathbf{V}_{a,b}^{\top}, \qquad \mathbf{U}_{a,b}, \mathbf{V}_{a,b} \in \mathbb{R}^{K \times r},$$

and then applies the same activation and hard gate as CAUSALBIND$_{\text{SP}}$:

$$\mathbf{Z}_{a,b}^{\text{LR}} = \tanh(\mathbf{U}_{a,b}\mathbf{V}_{a,b}^{\top}) + 1, \qquad \mathcal{M}_{a,b}^{\text{LR}} = H_{\tau}^{\text{ST}}\big(\mathbf{Z}_{a,b}^{\text{LR}}\big)\,.$$

The raw mask logit has rank at most r and remains a concept-axis operator: it modulates the K concept tokens while preserving each token's d_c-dimensional feature vector. The monotone activation and hard gate can change exact algebraic rank, but the trainable interaction directions are still controlled by the $2Kr$ low-rank factors. This is the key distinction from an embedding-space rank constraint, which would project pooled features into a low-dimensional feature subspace and can create a much stronger information bottleneck.

Key implementation equations. Equivalently, the full CAUSALBIND$_{\text{LR}}$ branch computation is

$$\begin{aligned}
\widehat{\mathcal{M}}_{a,b}^{\text{LR}} &= \mathbf{U}_{a,b}\mathbf{V}_{a,b}^{\top} = \sum_{q=1}^{r} \mathbf{u}_{a,b}^{(q)} (\mathbf{v}_{a,b}^{(q)})^{\top}, \\
\mathbf{Z}_{a,b}^{\text{LR}} &= \tanh(\widehat{\mathcal{M}}_{a,b}^{\text{LR}}) + 1, \\
\mathcal{M}_{a,b}^{\text{LR}} &= H_{\tau}^{\text{ST}}\big(\mathbf{Z}_{a,b}^{\text{LR}}\big)\,, \\
\mathbf{A}_{a,b}[i,j] &= \phi(\mathbf{C}_a[i,:], \mathbf{C}_b[j,:]), \\
u_{a,b}^{\text{LR}} &= \text{Agg}\big(\mathcal{M}_{a,b}^{\text{LR}} \odot \mathbf{A}_{a,b}\big)\,.
\end{aligned} \tag{30}$$

Thus CAUSALBIND$_{\text{LR}}$ differs from CAUSALBIND$_{\text{SP}}$ only in the parameterization of the same concept-axis mask: the dense mask logits $\widehat{\mathcal{M}}_{a,b}$ are replaced by a rank-r factorization. The defining constraint in CAUSALBIND$_{\text{LR}}$ is the rank constraint itself. The default CAUSALBIND$_{\text{LR}}$ setting uses $\lambda_{\text{spa}}^{\text{LR}} = 0$, while the optional $+L_{\text{spa}}$ diagnostic adds an L_1 penalty on the continuous activated low-rank masks without changing the rank-r logit factorization.

Executed loss. The training loss used for the default CAUSALBIND$_{\mathrm{LR}}$ is the same code-level objective as the main implementation, with the low-rank masks substituted into the forward pass and $\lambda_{\mathrm{spa}}^{\mathrm{LR}} = 0$. Concretely, the code applies the HypSeek branch retrieval loss to the molecule-pocket and molecule-sequence score matrices and then adds the inherited geometric terms; the optional $+L_{\mathrm{spa}}$ diagnostic additionally includes the mask penalty:

$$\begin{aligned}
\mathcal{L}_{\mathrm{exec}}^{\mathrm{LR}} &= \alpha_{\mathrm{poc}}\,\mathcal{L}_{\mathrm{hcc}}\big(\mathbf{U}^{\mathrm{poc,mol,LR}}\big) \;+\; \alpha_{\mathrm{prot}}\,\mathcal{L}_{\mathrm{hcc}}\big(\mathbf{U}^{\mathrm{seq,mol,LR}}\big) \\
&\quad + \gamma_{\mathrm{chl}}\mathcal{L}_{\mathrm{cone}} + \lambda_{\mathrm{angu}}R_{\mathrm{ang}} + \lambda_{\mathrm{het}}R_{\mathrm{het}} + \lambda_{\mathrm{spa}}^{\mathrm{LR}}\mathcal{L}_{\mathrm{spa}}^{\mathrm{LR}}, \qquad \lambda_{\mathrm{spa}}^{\mathrm{LR}} = 0 \text{ by default,} \\
\mathcal{L}_{\mathrm{hcc}}(\mathbf{U}) &= \mathcal{L}_{\mathrm{pocket}}(\mathbf{U}) + \mathcal{L}_{\mathrm{mol}}(\mathbf{U}) + \mathcal{L}_{\mathrm{rank}}(\mathbf{U}), \\
\mathcal{L}_{\mathrm{spa}}^{\mathrm{LR}} &= \|\mathbf{Z}_{\mathrm{mol}}^{\mathrm{LR}}\|_1 + \|\mathbf{Z}_{\mathrm{poc}}^{\mathrm{LR}}\|_1 + \|\mathbf{Z}_{\mathrm{seq}}^{\mathrm{LR}}\|_1.
\end{aligned} \tag{31}$$

Here $\mathcal{L}_{\mathrm{hcc}}$ is the actual branch loss returned by `compute_hcc_pair_official_style`: a pocket-to-molecule term, a molecule-to-pocket term, and an intra-assay ranking term. The terms $\mathcal{L}_{\mathrm{cone}}$, R_{ang}, and R_{het} are exactly the geometric regularizers defined above for CAUSALBIND$_{\mathrm{SP}}$. At the code level, `SparseMask` modules are attached to the molecule, pocket, and sequence streams; the branch-level notation in Eq. (30) is the paper-level interaction view induced by these stream masks.

A5.3.2 Concrete Parameter Assignments and Experimental Setting

Unless otherwise stated, CAUSALBIND$_{\mathrm{LR}}$ uses the same backbone, optimizer, training set, evaluation protocol, and concept-extractor configuration as the main CAUSALBIND$_{\mathrm{SP}}$ run: K=128, d_c=1024, L=4, 8 attention heads, and threshold $\tau = 0.5$. The representative direct low-rank setting uses $r = 1$ and $\lambda_{\mathrm{spa}}^{\mathrm{LR}} = 0$; rank scans additionally use $r \in \{1, 4, 8, 16, 32, 64\}$. The optional $+L_{\mathrm{spa}}$ diagnostic uses $\lambda_{\mathrm{spa}}^{\mathrm{LR}} = 10^{-4}$. For K=128, the rank-1 mask uses $2Kr = 256$ mask parameters per branch, compared with $K^2 = 16{,}384$ for the dense concept-pair mask.

A5.4 CAUSALBIND$_{\mathrm{EMB}}$: Embedding-Level Masking

CAUSALBIND$_{\mathrm{EMB}}$ is an embedding-level alternative that applies the sparse/low-rank interaction idea after pooling. It is useful as an empirical robustness check, but it is not the primary theory-aligned method because its mask entries do not correspond to explicit concept pairs.

A5.4.1 Technical Supplement for the CAUSALBIND$_{\mathrm{EMB}}$ Implementation

The embedding-level variant places the learnable interaction mask after mean-pooling, rather than directly on the concept axis. Concretely, CAUSALBIND$_{\mathrm{SP}}$ first extracts K causal concepts per modality, applies the mask on the concept axis, and then pools; by contrast, CAUSALBIND$_{\mathrm{EMB}}$ first obtains pooled modality embeddings and then applies a mask $\mathcal{M} \in \mathbb{R}^{d_c \times d_c}$ in the shared embedding space. The two variants therefore share the same high-level retrieval pipeline and the same sparsity-controlled interaction idea, but differ in where the bottleneck is imposed: before pooling for CAUSALBIND$_{\mathrm{SP}}$, versus after pooling for CAUSALBIND$_{\mathrm{EMB}}$.

This difference matters conceptually. CAUSALBIND$_{\mathrm{SP}}$ is more directly aligned with the theory in Section 2, because each entry of the mask can be interpreted as a relation between two concepts. CAUSALBIND$_{\mathrm{EMB}}$ is less interpretable in that sense, but it is architecturally simpler and still tests a related hypothesis: dense holistic interaction is unnecessary, while a constrained pooled-space interaction can already improve retrieval.

Key implementation equations. The embedding-level variant uses the same concept extractor, but moves the mask to the pooled feature axis:

$$\begin{aligned}
\mathbf{C}_v &= F_v(x_v) \in \mathbb{R}^{K \times d_c}, \qquad \bar{\mathbf{h}}_v = \mathrm{Pool}(\mathbf{C}_v) = \frac{1}{K}\sum_{i=1}^{K}\mathbf{C}_v[i,:], \\
\mathbf{Z}_v^{\mathrm{EMB}} &= \tanh(\widehat{\mathcal{M}}_v) + 1, \qquad \mathcal{M}_v^{\mathrm{EMB}} = H_\tau^{\mathrm{ST}}\big(\mathbf{Z}_v^{\mathrm{EMB}}\big), \qquad \widehat{\mathcal{M}}_v \in \mathbb{R}^{d_c \times d_c}, \\
\widetilde{\mathbf{h}}_v &= \bar{\mathbf{h}}_v \mathcal{M}_v^{\mathrm{EMB}}, \qquad \mathbf{z}_v = P_v(\widetilde{\mathbf{h}}_v), \\
u_{a,b}^{\mathrm{EMB}} &= \mathrm{sim}(\mathbf{z}_a, \mathbf{z}_b).
\end{aligned} \tag{32}$$

Here P_v is the modality-specific projection head and $\mathrm{sim}(\cdot,\cdot)$ denotes the branch retrieval similarity. Unlike CAUSALBIND$_{\mathrm{SP}}$ and CAUSALBIND$_{\mathrm{LR}}$, the mask in CAUSALBIND$_{\mathrm{EMB}}$ acts on pooled d_c-dimensional features, so its entries are not concept-pair interactions.

Executed loss. The embedding-level implementation uses the same executed objective, but its sparsity term is applied to the pooled embedding-space masks rather than to concept-axis masks. The score matrices are formed from the embedding-level branch scores $u_{a,b}^{\mathrm{EMB}}$ in Eq. (32):

$$\begin{aligned}\mathcal{L}_{\mathrm{exec}}^{\mathrm{EMB}} &= \alpha_{\mathrm{poc}}\,\mathcal{L}_{\mathrm{hcc}}\big(\mathbf{U}^{\mathrm{poc,mol,EMB}}\big) \;+\; \alpha_{\mathrm{prot}}\,\mathcal{L}_{\mathrm{hcc}}\big(\mathbf{U}^{\mathrm{seq,mol,EMB}}\big)\\ &\quad + \gamma_{\mathrm{chl}}\mathcal{L}_{\mathrm{cone}} + \lambda_{\mathrm{angu}}R_{\mathrm{ang}} + \lambda_{\mathrm{het}}R_{\mathrm{het}} + \lambda_{\mathrm{spa}}\mathcal{L}_{\mathrm{spa}}^{\mathrm{EMB}},\\ \mathcal{L}_{\mathrm{hcc}}(\mathbf{U}) &= \mathcal{L}_{\mathrm{pocket}}(\mathbf{U}) + \mathcal{L}_{\mathrm{mol}}(\mathbf{U}) + \mathcal{L}_{\mathrm{rank}}(\mathbf{U}),\\ \mathcal{L}_{\mathrm{spa}}^{\mathrm{EMB}} &= \|\mathbf{Z}_{\mathrm{mol}}^{\mathrm{EMB}}\|_1 + \|\mathbf{Z}_{\mathrm{poc}}^{\mathrm{EMB}}\|_1 + \|\mathbf{Z}_{\mathrm{seq}}^{\mathrm{EMB}}\|_1.\end{aligned} \tag{33}$$

Thus the difference between CAUSALBIND$_{\mathrm{EMB}}$ and the concept-basis variants is not the supervision signal: all three use the same retrieval, cone, angular, and heterogeneity terms. The difference is where the sparsity-controlled mask enters the forward computation and therefore what object $\mathcal{L}_{\mathrm{spa}}$ regularizes.

A5.4.2 Concrete Parameter Assignments and Experimental Setting

The representative CAUSALBIND$_{\mathrm{EMB}}$ setting uses K=64, d_c=256, L=4, 8 attention heads, mask type `tanh_plus_1`, threshold $\tau = 0.5$, and $\lambda_{\mathrm{spa}} = 10^{-2}$. It shares the same training corpus, optimizer, GPU setup, checkpoint protocols, and evaluation metrics as CAUSALBIND$_{\mathrm{SP}}$. In this variant the mask has size $d_c \times d_c$ rather than $K \times K$; for the representative setting, this gives $256^2 = 65{,}536$ mask parameters per stream.

A5.5 Comparison Across Three Implementations

Table A3 summarizes the relationship among the three implementations. The main text prioritizes CAUSALBIND$_{\mathrm{SP}}$ because it gives the cleanest sparse concept-pair story and keeps the method section compact. CAUSALBIND$_{\mathrm{LR}}$ is the most direct architectural instantiation of the low-rank theorem, while CAUSALBIND$_{\mathrm{EMB}}$ serves as a pooled-basis alternative that is empirically useful but theoretically less direct.

Table A3: Comparison of the three CAUSALBIND implementations. The main paper presents CAUSALBIND$_{\mathrm{SP}}$; CAUSALBIND$_{\mathrm{LR}}$ and CAUSALBIND$_{\mathrm{EMB}}$ are appendix variants.

Variant	Mask object	Forward bottleneck	Closest theory link	Role in the paper
CAUSALBIND$_{\mathrm{SP}}$	Dense $K \times K$ concept-pair mask with L_1 sparsity and hard thresholding.	Concept axis before pooling; each active entry selects a protein-molecule concept pair.	Direct sparse-support form of Theorem 2; learned masks can also enter the low-rank regime of Theorem 3.	Primary method in the main text.
CAUSALBIND$_{\mathrm{LR}}$	Rank-r concept-basis mask $\mathbf{U}\mathbf{V}^\top$ with $\mathbf{U},\mathbf{V} \in \mathbb{R}^{K\times r}$.	Concept axis before pooling; rank controls the number of interaction subspaces.	Direct implementation of the low-rank subspace-identifiability structure in Theorem 3.	Appendix variant for explicit low-rank implementation.
CAUSALBIND$_{\mathrm{EMB}}$	Dense $d_c \times d_c$ mask in pooled embedding space.	After mean-pooling; the mask acts on pooled features rather than concept pairs.	Compatible with a compressed interaction view, but not a direct instantiation of concept-pair sparse support.	Appendix alternative and robustness check.

The key implementation distinction is therefore the axis on which the constraint acts. CAUSALBIND$_{\mathrm{SP}}$ and CAUSALBIND$_{\mathrm{LR}}$ constrain the concept axis and preserve per-concept feature vectors, whereas CAUSALBIND$_{\mathrm{EMB}}$ constrains the pooled feature axis after the K concepts have already been averaged. This is why CAUSALBIND$_{\mathrm{SP}}$ and CAUSALBIND$_{\mathrm{LR}}$ have a cleaner theory-method correspondence, while CAUSALBIND$_{\mathrm{EMB}}$ is best read as an auxiliary empirical variant.

A6 Experimental Analysis

This section revisits the main-text experiments with more explanation and details. It should be read as an extension of the main experimental section rather than as a separate experimental theme. The goal here is to unpack the evidence behind the main claims, make the evaluation protocol fully explicit, and separate several empirical questions that are necessarily compressed in the main paper.

The subsection layout follows the three implementations introduced in App. A5. App. A6.1 collects the results for CAUSALBIND$_{\mathrm{SP}}$, the sparse concept-pair implementation used in the main text. App. A6.5 reports CAUSALBIND$_{\mathrm{LR}}$, the explicit low-rank concept-basis implementation. App. A6.6 reports CAUSALBIND$_{\mathrm{EMB}}$, the embedding-level alternative. We then synthesize the three implementations and close with the theory–practice discussion.

A6.1 CAUSALBIND$_{\mathrm{SP}}$: Sparse Concept-Pair Results

A6.1.1 Extended Analyses for the Main Experiments

Why does causal concept extraction help? The main paper argues that dense holistic alignment is too coarse a structural prior for protein-molecule binding, because binding is typically driven by a small number of local compatibilities rather than by uniform all-to-all agreement. The role of causal concept extraction is to operationalize this inductive bias. The concept extractor encourages each modality to decompose into a smaller set of binding-relevant factors, and the sparse interaction mask then decides which cross-modal pairs matter for prediction. In the scratch-training setting, this provides a strong bias toward disentangled structure. In the pretrained setting considered in our main experiments, it plays a more selective role: rather than learning chemistry and protein semantics from nothing, the module filters a pretrained representation and suppresses features that are predictive but not genuinely binding-relevant.

Why do the gains concentrate on early enrichment? Across both experimental settings, the improvements in BEDROC$_{80.5}$ and EF@1% are often larger than the improvements in AUROC. This pattern is important rather than incidental. AUROC measures global ranking quality over the entire candidate list, whereas BEDROC$_{80.5}$ and EF@1% emphasize whether true actives are pushed into the very top of the ranked list, which is the practically important regime in virtual screening. Our interpretation is that the sparse interaction prior is especially helpful when the model must decide which few candidates deserve the highest confidence, rather than when it is merely asked to produce a smooth global ordering.

Why is LIT-PCBA the stricter test? The larger improvements on LIT-PCBA than on DUD-E suggest that the causal prior is most useful when shortcut features become unreliable. DUD-E remains a useful benchmark because of its scale and broad target coverage, but LIT-PCBA is deliberately constructed to reduce property-based biases and therefore makes it harder for a model to succeed by exploiting easy decoy artifacts. When improvements persist on LIT-PCBA, especially in early-enrichment metrics, they are more plausibly reflecting better binding discrimination rather than better exploitation of dataset quirks. This is why the appendix repeatedly treats LIT-PCBA as the main stress test for whether the sparse interaction prior is doing the intended job.

A6.1.2 Main Comparison: Three-Protocol Evaluation

The main text reports the `last` checkpoint for all methods because, in this cross-dataset generalization setting, it is the cleanest fixed-endpoint protocol and avoids tuning checkpoint selection to the test benchmark. Here we additionally report two validation-selected alternatives, `best_auc` and `best_bedroc`, to show how much the main conclusions depend on checkpoint choice. To keep the comparison unambiguous, Table A4 uses a single seed throughout and does not mix in additional exploratory variants. The three methods shown are the two strongest retrieval baselines under our pipeline, HypSeek and LigUnity, together with our main concept-level configuration (K=128, d_c=1024, λ=10^{-4}, L=4)and its atom-attentive version CAUSALBIND$_{\mathrm{ATOM}}$ under the same configuration.

Two points matter for interpretation. First, the protocol choice does not reverse the main qualitative ranking between HypSeek and CAUSALBIND$_{\mathrm{SP}}$: under all reported checkpoints, our concept-level model remains clearly stronger on LIT-PCBA, which is the more realistic benchmark in

Table A4: Three-protocol evaluation for main comparisons. The main text reports the fixed-endpoint `last` protocol; `best_auc` and `best_bedroc` are complementary validation-selected references. All rows are seed-1 results.

Method	Protocol	DUD-E			LIT-PCBA		
		AUROC	$\text{BEDROC}_{80.5}$	EF@1%	AUROC	$\text{BEDROC}_{80.5}$	EF@1%
LigUnity	last	0.897	0.674	44.20	0.599	0.075	6.50
	best_auc	0.915	0.716	46.42	0.581	0.089	8.15
	best_bedroc	0.896	0.671	44.00	0.599	0.076	6.38
HypSeek	last	0.909	0.605	37.83	0.603	0.061	4.62
	best_auc	0.913	0.629	39.63	0.608	0.063	5.13
	best_bedroc	0.923	0.676	42.71	0.607	0.063	4.56
CAUSALBIND$_{\text{SP}}$	last	0.935	0.744	47.69	0.639	**0.097**	**8.56**
	best_auc	0.934	0.722	46.41	0.628	0.074	5.84
	best_bedroc	0.935	0.747	48.24	**0.643**	0.094	8.42
CAUSALBIND$_{\text{ATOM}}$	last	0.943	**0.768**	**49.66**	0.621	0.079	6.25
	best_auc	**0.949**	0.749	47.51	0.623	0.070	6.02
	best_bedroc	0.943	0.768	49.55	0.621	0.084	6.75

our evaluation. The largest gap appears under the fixed-endpoint `last` protocol used in the main text, where LIT-PCBA $\text{BEDROC}_{80.5}$ improves from 0.061 to 0.097 and EF@1% improves from 4.62 to 8.56. Second, the three protocols emphasize different aspects of the same model. For CAUSALBIND$_{\text{SP}}$, `best_bedroc` slightly improves DUD-E and raises LIT-PCBA AUROC to 0.643, but `last` gives the strongest LIT-PCBA early-enrichment profile, which is precisely the regime we care about in practical screening. The `best_auc` checkpoint is noticeably weaker on LIT-PCBA for our model, showing that checkpoint selection by a validation AUC criterion does not necessarily align with the practical objective of top-ranked hit discovery.

LigUnity shows a useful contrast. On DUD-E, validation-selected checkpoints improve over its `last` endpoint, especially under `best_auc`. On LIT-PCBA, however, the picture is mixed: `best_auc` improves early enrichment over its `last` result but lowers AUROC, while `best_bedroc` stays close to the fixed-endpoint baseline. This reinforces the same lesson as for CAUSALBIND$_{\text{SP}}$: checkpoint selection changes the metric emphasis, but it does not overturn the main comparison. Across all three protocols, CAUSALBIND$_{\text{SP}}$ remains stronger than LigUnity on the LIT-PCBA metrics emphasized in the main text. CAUSALBIND$_{\text{ATOM}}$ gives the strongest DUD-E results under every protocol, whereas its LIT-PCBA early enrichment stays below that of CAUSALBIND$_{\text{SP}}$ and varies with checkpoint choice (6.02–6.75 EF@1%), mirroring the pattern of the main comparison.

A6.1.3 Ablation and Diagnostic Study: Detailed Analysis

The ablation and diagnostic table in the main text is intended to answer a mechanistic question rather than to produce a second leaderboard: *which structural biases are responsible for the improvement, and how should we interpret the fact that different constraints matter differently across benchmarks?* This distinction is important because a component ablation need not improve every metric monotonically in order to be scientifically informative. In particular, the LIT-PCBA rows should be read as evidence about *which inductive bias is doing what*, not as evidence that each added module must dominate under every protocol and every metric.

Table 2 should be read in two parts: the CAUSALBIND$_{\text{SP}}$ rows are direct component ablations, whereas the CAUSALBIND$_{\text{LR}}$ and CAUSALBIND$_{\text{EMB}}$ rows are diagnostics for where and how rank compression is imposed. Reading the rows from top to bottom reveals the following:

(1) HypSeek and LigUnity establish the dense-alignment baseline. These two rows show what current retrieval systems can do without the explicit concept-level interaction structure introduced in CAUSALBIND. They provide the reference point for asking whether constrained concept-level interaction improves over dense holistic retrieval.

(2) The full model (CAUSALBIND$_{\text{SP}}$) gives the strongest overall balance. This row is the best summary of our main claim. On DUD-E it is the clear winner on all three metrics. On LIT-PCBA it gives the best AUROC and remains strong on $\text{BEDROC}_{80.5}$ and EF@1%, so the full architecture gives the best cross-benchmark tradeoff.

Table A5: CAUSALBIND$_{\mathrm{SP}}$ concept-level K-scan (fix d_c=512), seed=1, all three protocols. K=256 underperforms consistently.

K	Protocol	DUD-E			LIT-PCBA		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
32	last	0.938	0.749	48.41	0.620	0.081	6.90
	best_auc	0.939	0.750	48.38	0.625	0.083	6.85
	best_bedroc	0.938	0.750	48.57	0.621	0.081	6.90
64	last	0.930	0.724	46.55	0.637	0.085	6.41
	best_auc	0.933	0.715	45.90	0.638	0.088	7.88
	best_bedroc	0.935	0.732	47.08	0.635	0.079	5.96
128	last	0.943	0.752	47.93	0.633	0.081	7.06
	best_auc	0.943	0.755	48.33	0.634	0.082	7.04
	best_bedroc	0.943	0.752	47.93	0.632	0.082	7.06
256	last	0.916	0.660	41.50	0.610	0.066	5.90
	best_auc	0.917	0.662	41.82	0.609	0.065	5.46
	best_bedroc	0.916	0.664	42.09	0.612	0.069	6.02

(3) Removing only L_{spa} isolates an unguided learned mask. The cross-modal mask is still present, but it is no longer coupled to an explicit sparsity objective. The resulting drop suggests that the mask is most useful when regularized toward a simple interaction structure.

(4) The LR and EMB diagnostics separate two kinds of compression. In CAUSALBIND$_{\mathrm{LR}}$, both rows keep the rank-one concept-axis factorization; adding the optional L_{spa} penalty changes little, indicating that the low-rank parameterization itself is the operative constraint. In CAUSALBIND$_{\mathrm{EMB}}$, however, the $\mathbf{M}_{\mathrm{emb}}^{r=1}$ row imposes rank one after pooling and therefore becomes a much stronger information bottleneck.

Main takeaway from the ablation. The cleanest reading is that the constrained concept-level contrastive route is the essential structural change. The concept extractor and mask help because they regularize how contrastive learning sees the two modalities, and the comparison with CAUSALBIND$_{\mathrm{LR}}$ shows that explicit low-rank concept-axis compression can provide a similar structural bias.

Connection to the later theory–practice analysis. This reading is consistent with the mask-structure results. The useful empirical regime is not aggressively hard-sparse, but moderately regularized and low-rank. The ablation therefore supports the same calibrated view as the theory–practice analysis: structural constraints matter, but their benefit comes from shaping the interaction into a simple reusable form rather than from adding parameters.

A6.1.4 Hyperparameter Study on K and d_c

The next question is whether the selected concept count K and concept dimension d_c are robust design choices or just narrow local optima. We therefore report the full protocol-resolved scans for both quantities. These tables are more informative than the single operating-point curves in the main text because they make two things visible at once: where performance is consistently strong, and where apparent improvements depend strongly on the checkpoint-selection rule.

Interpretation. The concept-level sweep shows a broad stable region rather than a single fragile optimum. In Table A5, K=128 is the most consistent choice across all three checkpoint protocols, while K=256 degrades on both DUD-E and LIT-PCBA. In Table A6, increasing d_c from 512 to 1024 trades a small amount of DUD-E margin for a clear gain on LIT-PCBA early enrichment, which is why the main text uses the higher-capacity setting.

A6.1.5 Hyperparameter Study on λ

This sweep asks a more specific question than the K/d_c tables above: once the concept budget is fixed at the main operating point (K=128, d_c=1024, L=4), what range of sparsity-loss weights produces useful behavior, and where does the model collapse? We therefore vary the L_1 sparsity weight λ across eight orders of magnitude and report all three checkpoint protocols. The resulting

Table A6: CAUSALBIND$_{\mathrm{SP}}$ concept-level d_c-scan (fix K=128), seed=1, all three protocols. d_c=1024 gives the strongest LIT-PCBA results.

d_c	Protocol	DUD-E AUROC	DUD-E BEDROC$_{80.5}$	DUD-E EF@1%	LIT-PCBA AUROC	LIT-PCBA BEDROC$_{80.5}$	LIT-PCBA EF@1%
128	last	0.919	0.589	35.75	0.615	0.062	5.25
	best_auc	0.922	0.629	39.43	0.630	0.060	4.75
	best_bedroc	0.920	0.592	36.03	0.620	0.062	5.35
256	last	0.912	0.600	36.93	0.614	0.062	5.68
	best_auc	0.922	0.683	43.07	0.620	0.067	4.73
	best_bedroc	0.918	0.642	40.05	0.621	0.070	6.09
512	last	0.943	0.752	47.93	0.633	0.081	7.06
	best_auc	0.943	0.755	48.33	0.634	0.082	7.04
	best_bedroc	0.943	0.752	47.93	0.632	0.082	7.06
1024	last	0.935	0.744	47.69	**0.639**	**0.097**	**8.56**
	best_auc	0.934	0.722	46.41	0.628	0.074	5.84
	best_bedroc	0.935	0.747	48.24	0.643	0.094	8.42

pattern is useful for interpreting both the main-text hyperparameter plot and the theory–practice discussion that follows in Section A6.1.7.

Table A7: CAUSALBIND$_{\mathrm{SP}}$ concept-level λ-scan, seed=1, all three protocols. K=128, d_c=1024, L=4. Rows below the dotted rule show mask collapse under `last` and `best_bedroc`.

λ	Protocol	DUD-E AUROC	DUD-E BEDROC$_{80.5}$	DUD-E EF@1%	LIT-PCBA AUROC	LIT-PCBA BEDROC$_{80.5}$	LIT-PCBA EF@1%
0	last	0.938	0.746	48.14	0.617	0.080	6.83
	best_auc	0.943	0.744	47.68	0.629	0.073	6.26
	best_bedroc	0.934	0.732	47.03	0.613	0.075	6.10
10^{-6}	last	0.940	0.749	47.94	0.635	0.088	7.26
	best_auc	0.941	0.750	47.75	0.628	0.086	7.41
	best_bedroc	0.941	0.755	48.38	0.631	0.086	7.67
10^{-5}	last	0.938	0.735	47.09	0.607	0.078	6.29
	best_auc	0.940	0.732	47.08	0.628	0.082	6.41
	best_bedroc	0.940	0.735	47.15	0.613	0.070	5.72
$\mathbf{10^{-4}}$	last	0.935	0.744	47.69	**0.639**	**0.097**	**8.56**
	best_auc	0.934	0.722	46.41	0.628	0.074	5.84
	best_bedroc	0.935	0.747	48.24	0.643	0.094	8.42
10^{-3}	last	**0.943**	**0.768**	**49.73**	0.614	0.073	5.49
	best_auc	0.936	0.712	45.34	0.620	0.072	5.89
	best_bedroc	0.936	0.712	45.34	0.620	0.072	5.89
—- *mask collapses below (`last` and `best_bedroc` protocols)* —-							
10^{-2}	last	0.437	0.013	0.53	0.457	0.011	0.46
	best_auc	0.921	0.699	44.71	0.622	0.074	5.83
	best_bedroc	0.442	0.014	0.64	0.460	0.011	0.54
10^{-1}	last	0.297	0.006	0.30	0.437	0.013	0.59
	best_auc	0.930	0.696	44.53	0.619	0.078	6.73
	best_bedroc	0.930	0.696	44.53	0.619	0.078	6.73
1	last	0.725	0.099	5.30	0.595	0.028	1.44
	best_auc	0.936	0.721	46.22	0.642	0.072	5.87
	best_bedroc	0.930	0.697	43.99	0.635	0.064	4.92

Interpretation. Two regimes are practically useful. λ=10^{-4} and λ=10^{-3} both give strong results, but they emphasize different benchmarks: the former is clearly better on LIT-PCBA, while the latter is strongest on DUD-E. By contrast, $\lambda \leq 10^{-6}$ is almost indistinguishable from turning the sparsity loss off, and $\lambda \geq 10^{-2}$ pushes the hard gate into collapse under the fixed-endpoint and BEDROC-selected

protocols. The `best_auc` rows remain more resilient in that high-λ region because they often pick an earlier checkpoint, before the mask has fully saturated.

A6.1.6 Hyperparameter Study on L

This scan isolates a different modeling choice: once the concept count, channel width, and sparsity weight are fixed, how much depth is actually useful in the concept extractor itself? Since the main text argues that the key inductive bias comes from structured cross-modal interaction rather than depth for its own sake, this table is mainly a check that deeper extractors do not hide an unreported gain.

Table A8: CAUSALBIND$_{\text{SP}}$ concept-level Perceiver-depth (L) scan, seed=1, all three protocols. K=128, d_c=1024, λ=10^{-4}.

L	Protocol	DUD-E			LIT-PCBA		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
2	last	0.940	0.763	49.20	0.625	0.087	6.98
	best_auc	**0.946**	**0.776**	**49.99**	0.616	0.082	6.53
	best_bedroc	0.942	0.757	48.48	0.622	0.085	7.95
4 (default)	last	0.935	0.744	47.69	**0.639**	**0.097**	**8.56**
	best_auc	0.934	0.722	46.41	0.628	0.074	5.84
	best_bedroc	0.935	0.747	48.24	0.643	0.094	8.42
6	last	0.936	0.738	47.05	0.633	0.085	6.75
	best_auc	0.937	0.729	46.70	0.634	0.086	6.89
	best_bedroc	0.937	0.733	46.79	0.629	0.086	6.93
8	last	0.934	0.727	46.38	0.630	0.080	6.73
	best_auc	0.938	0.724	46.32	0.622	0.072	6.05
	best_bedroc	0.931	0.702	45.03	0.638	0.079	6.38

Interpretation. The depth sweep is simple: modest depth is enough. L=2 is strongest on DUD-E under `best_auc`, while L=4 gives the best LIT-PCBA early enrichment under the two protocols that matter most for our main narrative, `last` and `best_bedroc`. Increasing depth further to L=6 or 8 never helps. This is consistent with the paper's central claim that the structural bias comes mainly from the concept factorization and cross-modal mask, not from stacking a deeper extractor.

A6.1.7 Detailed Mask Structural Analysis

The compact λ sweep in Table 3 motivates a more detailed question: does the learned interaction become element-wise sparse, spectrally low-rank, or both? We therefore measure both element-wise sparsity and spectral low-rank structure directly on the learned concept-level masks. Table A9 reports code-level stream diagnostics for the molecule, pocket, and sequence mask modules (K=128), while the main-text Table 3 reports compact paired-branch statistics for the two protein-molecule retrieval branches. In the notation of Section 2, Condition 2.i corresponds to element-wise sparse support, whereas Condition 3.i corresponds to low-rank or mode-compressed interaction.

Measurement. Let $\mathcal{M}_{\text{eff}} = \sigma(\mathcal{M}) = \tanh(\mathcal{M}_{\text{raw}}) + 1 \in (0, 2)$. The concept-level forward pass applies a hard gate: entries with $\mathcal{M}_{\text{eff}} \leq 0.5$ are zeroed exactly. The L_1 mass and spectral ranks below are measured on the continuous pre-gate $\mathcal{M}_{\text{eff}}$, while the zero counts are measured after the forward threshold. We report:

- **Element-wise sparsity** (corresponds to Condition 2.i): fraction of entries with $\mathcal{M}_{\text{eff}} < 0.5$, i.e. the number of hard zeros after the forward gate.
- **Spectral effective rank** (corresponds to Condition 3.i): the smallest r such that the top-r singular values explain at least 95% or 99% of $\|\mathcal{M}_{\text{eff}}\|_F^2$.
- Summary scalars: the average mask value $\overline{\sigma(\mathcal{M})}$ and the total mask mass $L_1 = \sum_{ij} \sigma(\mathcal{M})_{ij}$.
- **Antichain violations** (corresponds to the antichain clause of Condition 2.i): on the hard support $\mathbf{1}[\mathcal{M}_{\text{eff}} > 0.5]$ of each cross-modal mask (molecule–pocket and molecule–sequence), we compare

every pair of distinct row supports and every pair of distinct column supports, and count a violation whenever one support is contained in the other (equal supports also count). Each branch contributes $\binom{128}{2} = 8{,}128$ row pairs and $8{,}128$ column pairs, giving $32{,}512$ tests in total. This statistic is reported in the last column of Table 3.

The antichain statistic sharpens the two-regime interpretation. With no hard zeros ($\lambda \leq 10^{-4}$), all row and column supports equal the full index set, so every pair violates the antichain condition; these operating points should be interpreted only through Condition 3.i and Theorem 3. At $\lambda{=}10^{-3}$, every one of the 128 row and 128 column supports is unique in both branches, with zero containment relations, so the antichain clause is exactly satisfied by the learned hard support. At $\lambda{=}10^{-2}$ the masks remain exact antichains but retrieval collapses, so the antichain property alone does not explain performance. We note that the $\lambda{=}10^{-3}$ mask remains only moderately sparse (79.5% active entries) and that Condition 2.ii is not directly measured; the diagnostic validates the antichain subcondition rather than every premise of Theorem 2. All statistics are computed exactly on the same single-seed `last` checkpoints as Table 3.

This subsection uses the per-stream quantities $\overline{\sigma(\mathcal{M})}$ and $L_1 = \sum_{ij} \sigma(\mathcal{M})_{ij}$ because Table A9 reports stream-level mask structure. By contrast, the main-text Table 3 reports the aggregated statistic 'L_1 avg', i.e., the average absolute mask entry after averaging over all entries and then over the two paired retrieval branches. The two views are consistent but not the same object: one is a stream-level total-mass diagnostic, while the other is a normalized paired-branch summary.

Operational criteria for "sparse" and "low-rank". The theory itself does not specify a single numerical threshold such as "10% zeros" or "rank below 5"; the conditions are structural. Still, for the purpose of interpreting the experiments, it is useful to make the comparison quantitative. Here $K{=}128$, so Condition 3.i and the low-rank regime of Theorem 3 are most plausible when the effective rank needed to explain the mask energy is much smaller than 128. By that standard, the evidence for low-rank structure is strong: for $\lambda{=}10^{-4}$ we have $r@95\%{=}1$ on all three stream masks and $r@99\% \in \{9, 12, 14\}$, i.e., only about 7%–11% of the full concept dimension is needed to explain 99% of the squared singular-value energy. For $\lambda{\geq}10^{-3}$ the evidence is even stronger in rank terms, since both $r@95\%$ and $r@99\%$ collapse to 1.

For Condition 2.i and the sparse-support regime of Theorem 2, the practically relevant operating point is $\lambda{=}10^{-3}$. Condition 2.i requires a genuinely sparse support pattern together with the antichain-style separation discussed in the theory. At $\lambda{=}10^{-3}$, the three stream masks show about 9.7%–13.4% exact zeros while still maintaining strong retrieval performance, which is exactly the qualitative behavior expected from the sparse-support theorem. At $\lambda{=}10^{-4}$, by contrast, there are no hard zeros at all, so this setting should be interpreted through the low-rank theorem rather than the sparse one. In this sense, the experiments support two different theorem-aligned regimes: $\lambda{=}10^{-3}$ for sparse-support identifiability and $\lambda{=}10^{-4}$ for low-rank subspace identifiability.

Findings. Three patterns matter most. First, the effective masks are low-rank throughout the sweep: $r@95\%$ is always 1, so the dominant interaction mode is already highly compressed even without explicit sparsity. Second, $\lambda{=}10^{-4}$ does *not* create hard zeros, but it does increase the spectral tail to $r@99\% = 9$–14, which is exactly the regime where LIT-PCBA performance is strongest. Third, once λ reaches 10^{-3} and above, hard zeros appear abruptly and the masks collapse back to rank-1 even at the 99% level. In other words, stronger sparsity regularization does not gradually sharpen a useful structure; it eventually removes the richer low-rank structure that seems to matter most in practice.

How the conditions appear in practice. The current measurements support both theorem regimes, but through different diagnostics. Condition 3.i is fundamentally about dimensionality reduction of the interaction kernel, and singular-value diagnostics measure exactly that, so $\lambda{=}10^{-4}$ gives direct quantitative support for the low-rank theorem. Condition 2.i is about sparse support structure, and the hard-zero statistics at $\lambda{=}10^{-3}$ show that the model has indeed entered that sparse-support regime while retaining strong retrieval quality. The table therefore supports a practical two-regime reading rather than a single-regime one.

Best-effort conclusion about the assumptions. The most useful reading is therefore complementary rather than asymmetric. For $\lambda{=}10^{-3}$, the experiments support the sparse-support theorem in the regime where meaningful exact zeros appear and performance remains strong. For $\lambda{=}10^{-4}$, the experiments support the low-rank theorem in the regime where the mask remains dense entry-wise

Table A9: Mask structural analysis across λ for the concept-level model. Per-stream numbers are reported at $K{=}128$ under the `last` protocol, seed=1. L_1 and rank statistics use the continuous pre-gate masks; zeros use the hard-gated forward masks. $\lambda{=}10^{-4}$ remains element-wise dense but has a richer spectral tail, while $\lambda{\geq}10^{-3}$ produces hard zeros and rank-1 masks.

λ	branch	$\overline{\sigma(\mathcal{M})}$	max $\sigma(\mathcal{M})$	L_1 (sum)	zeros	%zeros	r@95%	r@99%
0	mol	0.920	1.311	15068	0	0.0%	1	5
	poc	0.993	1.398	16267	0	0.0%	1	1
	seq	1.023	1.374	16758	0	0.0%	1	1
10^{-6}	mol	0.910	1.301	14914	0	0.0%	1	5
	poc	0.991	1.395	16228	0	0.0%	1	1
	seq	1.023	1.374	16755	0	0.0%	1	1
10^{-5}	mol	0.894	1.286	14648	0	0.0%	1	6
	poc	0.968	1.377	15863	0	0.0%	1	2
	seq	1.007	1.361	16505	0	0.0%	1	1
$\mathbf{10^{-4}}$	mol	0.756	1.148	12383	0	0.0%	1	**14**
	poc	0.781	1.200	12793	0	0.0%	1	**12**
	seq	0.841	1.206	13774	0	0.0%	1	**9**
$\mathbf{10^{-3}}$	mol	0.512	0.773	8383	1645	10.0%	1	1
	poc	0.512	0.797	8390	2198	13.4%	1	1
	seq	0.513	0.751	8404	1593	9.7%	1	1
10^{-2}	mol	0.501	0.656	8206	8677	53.0%	1	1
	poc	0.502	0.708	8222	8448	51.6%	1	1
	seq	0.502	0.659	8222	9199	56.1%	1	1
10^{-1}	mol	0.500	0.599	8195	11796	72.0%	1	1
	poc	0.500	0.653	8199	11484	70.1%	1	1
	seq	0.500	0.610	8199	11790	72.0%	1	1
1	mol	0.500	0.571	8193	13233	80.8%	1	1
	poc	0.500	0.613	8194	12994	79.3%	1	1
	seq	0.500	0.561	8194	13075	79.8%	1	1

but is strongly spectrally compressed. The two theorems are therefore both realized in practice, but at different operating points of the same model.

Implication for theory. The correspondence to the theory is therefore not "Theorem 2 fails and only Theorem 3 matters." It is that the model exhibits two neighboring theorem-aligned regimes. Around $\lambda{=}10^{-3}$, the learned mask behaves in the sparse-support sense emphasized by Theorem 2. Around $\lambda{=}10^{-4}$, it behaves in the low-rank sense emphasized by Theorem 3. Since both settings remain empirically strong, the theory is useful in exactly the way we want: it explains two distinct but valid operating points of the learned interaction.

A6.1.8 ROC-Enrichment (RE) on DUD-E

ROC-enrichment (RE) is another early-recognition metric, complementary to $\text{BEDROC}_{80.5}$ and EF@1%. At a fixed false-positive-rate threshold α, it is defined as $\text{RE@}\alpha = \text{TPR}(\alpha)/\alpha$, so larger values mean that the model retrieves actives much more efficiently than random ranking in the very early part of the list. Following prior reporting practice [Wang et al., 2026], we report RE at 0.5%, 1%, 2%, and 5% false-positive-rate thresholds on DUD-E.

We report RE only on DUD-E because it is the more standard setting in prior work and provides a more stable target-averaged estimate than LIT-PCBA. Concretely, DUD-E contains 102 targets, whereas LIT-PCBA contains only 15. Table A10 therefore augments the main-text DUD-E comparison with RE, while keeping AUROC / $\text{BEDROC}_{80.5}$ / EF@1% for context. To stay aligned with the main text, all rows use the same fixed-endpoint `last` checkpoint protocol after the full training budget.

Detailed analysis. Table A10 supports three conclusions. First, $\text{CAUSALBIND}_{\text{SP}}$ improves very-early retrieval not only relative to the HypSeek reproduction, but also relative to the broader retrieval

Table A10: ROC-enrichment (RE) on DUD-E under the `last` checkpoint protocol. Values are 102-target means. AUROC / $\text{BEDROC}_{80.5}$ / EF@1% are reproduced from the main DUD-E comparison for context.

Configuration	Primary metrics			ROC-enrichment (RE)			
	AUROC	$\text{BEDROC}_{80.5}$	EF@1%	@0.5%	@1%	@2%	@5%
DrugCLIP [Gao et al., 2023]	0.809	0.505	31.89	73.97	41.79	23.68	11.16
LigUnity [Feng et al., 2025]	0.897	0.674	44.20	104.69	57.47	33.76	13.88
HypSeek [Wang et al., 2026]	0.909	0.605	37.83	95.20	54.49	30.74	14.01
CAUSALBIND$_{\text{SP}}$ (K=128, d_c=512, L=4)	0.943	0.752	47.93	127.56	69.94	37.83	**16.39**
CAUSALBIND$_{\text{SP}}$ (K=128, d_c=1024, L=4)	0.935	0.744	47.69	128.06	69.26	37.07	15.99
CAUSALBIND$_{\text{SP}}$ (K=128, d_c=1024, L=2)	0.940	0.763	49.20	**131.55**	**71.17**	**38.04**	16.30

Table A11: Merck-FEP-8 affinity-ranking correlations. Higher Pearson r and Spearman ρ are better. CAUSALBIND$_{\text{SP}}$ uses the main K=128, d_c=1024 configuration under seed 1 and the `last` checkpoint.

Method	Pearson r	Spearman ρ
LigUnity	0.532	0.499
HypSeek	0.423	0.390
CAUSALBIND$_{\text{SP}}$	**0.598**	**0.548**

baselines included in the main comparison. All three reported CAUSALBIND$_{\text{SP}}$ settings are clearly stronger than DrugCLIP and LigUnity on AUROC, $\text{BEDROC}_{80.5}$, EF@1%, and all four RE thresholds. This is useful because RE asks a slightly different question from $\text{BEDROC}_{80.5}$ and EF@1%: instead of weighting the top of the ranking or counting actives in the top percentile, it measures how efficiently true actives are recovered at extremely small false-positive rates. The consistent improvement across all of these views strengthens the claim that CAUSALBIND$_{\text{SP}}$ genuinely improves early retrieval on DUD-E.

How to read the CAUSALBIND$_{\text{SP}}$ rows. Second, the ordering among the three CAUSALBIND$_{\text{SP}}$ settings is informative. The main-text operating point (K=128, d_c=1024, L=4) is already much stronger than the prior baselines, but the shallower L=2 variant is slightly better on the RE thresholds and on DUD-E $\text{BEDROC}_{80.5}$ / EF@1%. This matches the broader hyperparameter-study pattern that, on DUD-E alone, a slightly shallower interaction depth can be favorable for very-early retrieval. Third, this does not overturn the model choice in the main text. The paper uses L=4 because it gives the better cross-benchmark balance once LIT-PCBA is taken into account, whereas RE is reported only on DUD-E as a supplementary diagnostic. We therefore read Table A10 as supporting evidence for the DUD-E early-recognition story, not as a new primary criterion for model selection.

A6.1.9 Affinity Ranking on Merck-FEP-8

We additionally evaluate whether the main CAUSALBIND$_{\text{SP}}$ checkpoint transfers from virtual screening to affinity ranking. We use HypSeek's `test_fep()` pipeline on the 16 FEP benchmark targets, consisting of JACS-8 and Merck-FEP-8; because the current comparison focuses on structurally diverse affinity-ranking targets, Table A11 reports only Merck-FEP-8. The evaluation uses the single-seed `last` checkpoint protocol for CAUSALBIND$_{\text{SP}}$ and the reproduced HypSeek baseline. Missing `bench/` CSV files only affect ACC/AUC outputs, which are skipped; Pearson r and Spearman ρ are unaffected.

On Merck-FEP-8, CAUSALBIND$_{\text{SP}}$ improves over HypSeek by $+0.175$ Pearson r and $+0.158$ Spearman ρ, and also exceeds LigUnity by $+0.066$ and $+0.049$, respectively. This suggests that the sparse concept-pair interaction learned for retrieval also transfers to affinity ranking on diverse ligand sets.

A6.2 Comparison with S^{2}Drug

S^{2}Drug [He et al., 2026] is a concurrent retrieval-based virtual-screening method. Because its implementation was not publicly available, we could not reproduce it under our pipeline; Table 1 therefore reports its paper-reported results, and this comparison should be read as paper-to-paper

rather than controlled. S^2Drug reports EF@0.5% as its primary LIT-PCBA enrichment metric, whereas our main table follows HypSeek and reports EF@1%. Table A12 provides the matched EF@0.5% comparison, computed by the same evaluator from the same checkpoints as Table 1. At the matched cutoff, CAUSALBIND$_{\mathrm{LR}}$ and CAUSALBIND$_{\mathrm{EMB}}$ exceed the paper-reported S^2Drug result, while CAUSALBIND$_{\mathrm{SP}}$ is comparable; at EF@1%, all three variants exceed it.

Table A12: LIT-PCBA EF@0.5% comparison with the paper-reported S^2Drug result.

Method	EF@0.5%	EF@1%
S^2Drug (paper-reported)	11.44	7.38
CAUSALBIND$_{\mathrm{SP}}$	10.88	8.56
CAUSALBIND$_{\mathrm{LR}}$	11.59	8.84
CAUSALBIND$_{\mathrm{EMB}}$	**12.20**	**9.60**

A6.3 Out-of-Distribution Evaluation on DEKOIS 2.0

To test generalization beyond DUD-E and LIT-PCBA, we perform zero-shot evaluation on the established DEKOIS 2.0 benchmark [Bauer et al., 2013] under two complementary settings, both fixed before running any model. **Target-level OOD** retains the nine DEKOIS targets whose UniProt IDs are absent from all training sources, including PDBBind (11BETAHSD1, ACE2, ADRB2, ALR2, COX1, CYP2A6, INHA, MMP2, and TK). **Scaffold-level OOD** further retains only test pairs whose Bemis–Murcko scaffolds are absent from all $439{,}639$ unique training molecules; requiring at least five such actives per target leaves six targets, $5{,}385$ pairs, and 98 actives. Metrics are macro-averaged over targets.

Table A13: Zero-shot out-of-distribution evaluation on DEKOIS 2.0. CAUSALBIND$_{\mathrm{ATOM}}$ is the atom-attentive version of CAUSALBIND$_{\mathrm{SP}}$ (Section 4).

Method	Target-level OOD (9 targets)			Scaffold-level OOD (6 targets)		
	AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
HypSeek [Wang et al., 2026]	0.819	0.593	20.06	0.880	0.630	44.47
LigUnity [Feng et al., 2025]	0.859	0.570	19.82	0.915	0.612	43.11
CAUSALBIND$_{\mathrm{SP}}$	0.860	0.658	**23.00**	0.945	0.712	43.66
CAUSALBIND$_{\mathrm{LR}}$	**0.867**	0.653	22.71	**0.949**	0.716	44.70
CAUSALBIND$_{\mathrm{ATOM}}$	0.837	**0.666**	22.94	0.941	**0.744**	**50.77**

Table A13 shows that the structured-interaction models improve AUROC and BEDROC$_{80.5}$ over both retrieval baselines under both shifts, while CAUSALBIND$_{\mathrm{SP}}$ remains comparable to HypSeek at the strictest EF@1% cutoff of the scaffold-level split. Together with the cross-dataset protocol of the main experiments, these results indicate that the learned interaction structures transfer under explicit target-level and scaffold-level distribution shifts. Under CD-HIT sequence clustering at 40% identity, only one of the nine targets (ACE2) has its entire cluster absent from training, so these results should not be read as evidence of broad protein-family generalization.

A6.4 Interpretability Case Study on THRB

To connect the learned concepts with biologically meaningful binding patterns, we analyze the aligned protein–ligand crystal complex of the Human Thyroid Hormone Receptor β (THRB) from DUD-E, following the evidence chain from local structural contacts to learned concepts, to the predicted binding score, and finally to a random-control validation (Table A14, Figure A3). Observed ligand–pocket contacts are heavy-atom pairs within 4 Å. Concepts are selected solely by their attention coverage of these contacts, using mean-over-heads attention from the final Perceiver layer, without using the downstream score change. Among the contacting atoms, the four ligand atoms and eight pocket atoms most attended by the selected concepts are then excluded from the concept bottleneck while the Uni-Mol encoder outputs are kept fixed, and the final pocket–ligand score is recomputed. As a control, we repeat this procedure for 64 matched random sets of four ligand and eight pocket atoms.

Table A14: Local-to-global attribution on the DUD-E THRB complex.

Stage	Evidence	Observation	Purpose
Local structure	Heavy-atom contacts (≤ 4 Å)	98 contacts involving 27 ligand and 47 pocket atoms	Identify the local binding interface
Learned concept	Contact concentration in the most contact-aligned concept pair	0.0505 vs. 0.00881 uniform baseline (5.73× enrichment)	Test whether contacts concentrate in concepts
Predicted score	Mask the associated 4 ligand and 8 pocket atoms (encoder fixed)	Score decreases $1.878 \rightarrow 0.851$ (-54.7%)	Test contribution to the prediction
Random control	Mask 64 matched random sets of 4 ligand and 8 pocket atoms	Mean decrease 0.026 ± 0.073; guided decrease 1.027 (99.2nd percentile)	Exclude an arbitrary masking effect

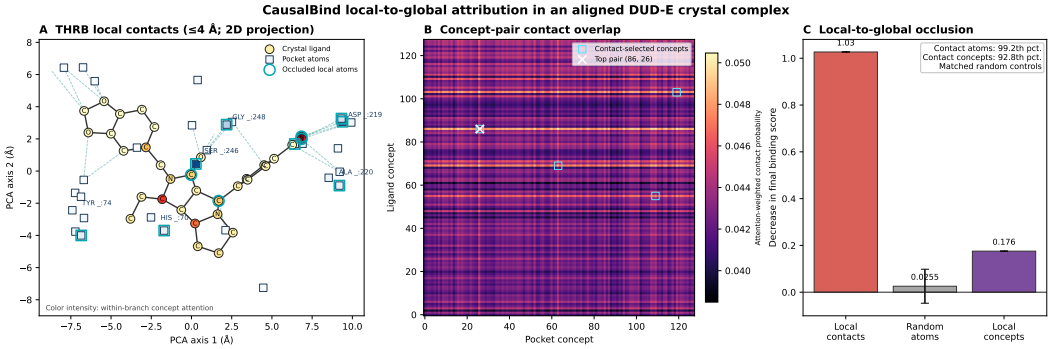


Figure A3: Local-to-global attribution on the DUD-E THRB complex. (A) Ligand–pocket heavy-atom contacts within 4 Å in a 2D projection; highlighted atoms are those occluded at the concept bottleneck. (B) Attention-weighted contact probability over ligand–pocket concept pairs, with the contact-selected concepts and the top pair marked. (C) Decrease in the final binding score after masking the local contact atoms, matched random atoms, or the contact-selected concepts.

The highest-contributing contacts are chemically plausible close polar contacts: ligand N–Asp219 OD1 (2.71 Å), N–Asp219 OD2 (2.78 Å), N–Gly250 O (2.88 Å), and O–Gly248 N (3.04 Å). These results show that biologically meaningful local contacts concentrate within a learned concept pair, and that concept-guided masking causes a much larger score decrease than random masking, supporting the intended local-structure-to-concept-to-score mechanism rather than an arbitrary masking effect. This post-hoc diagnostic uses CausalBind$_{\text{ATOM}}$, which exposes the full Uni-Mol atom sequence to the Perceiver; it does not claim that individual concepts carry one-to-one biochemical meanings. Attention and occlusion provide model-level attribution evidence rather than proof of a physical causal mechanism.

A6.5 CausalBind$_{\text{LR}}$: Low-Rank Concept-Basis Results

This subsection reports the explicit low-rank concept-basis implementation, denoted CausalBind$_{\text{LR}}$ in the paper and recorded as V3 in the experiment log. It uses the same concept extractor and retrieval loss as CausalBind$_{\text{SP}}$, but replaces the dense $K \times K$ concept-pair mask with a hard rank-r factorization. Thus it is the most direct empirical test of the subspace-identifiability view in Theorem 3.

A6.5.1 Main Rank Scan and Three-Protocol Results

For each branch, CausalBind$_{\text{LR}}$ parameterizes the concept-basis mask as

$$\sigma(\mathcal{M}_{\text{LR}}) = \tanh(\mathbf{U}\mathbf{V}^{\top}) + 1, \qquad \mathcal{M}_{\text{eff}} = H_{0.5}(\sigma(\mathcal{M}_{\text{LR}})), \qquad \mathbf{U}, \mathbf{V} \in \mathbb{R}^{K \times r},$$

where $H_{0.5}$ denotes the same threshold gate used by CausalBind$_{\text{SP}}$. Unless otherwise stated, the rank scan uses K=128, d_c=1024, L=4, seed 1, and λ=10^{-4}. The mask-parameter count per branch is $2Kr$, compared with K^2 for the dense CausalBind$_{\text{SP}}$ mask.

Table A15: CAUSALBIND$_{\mathrm{LR}}$ rank scan, seed=1, all three checkpoint protocols. All rows use K=128, d_c=1024, L=4, and λ=10^{-4}.

r	Params/branch	Protocol	DUD-E			LIT-PCBA		
			AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
1	256	last	0.935	0.720	45.86	0.630	0.087	8.41
		best_auc	**0.946**	**0.756**	**48.78**	0.636	0.079	6.53
		best_bedroc	0.935	0.724	46.28	0.632	**0.089**	**8.56**
4	1,024	last	0.934	0.731	47.14	0.632	0.085	7.45
		best_auc	0.933	0.710	45.14	0.633	0.080	7.67
		best_bedroc	0.938	0.744	47.95	0.632	0.078	6.52
8	2,048	last	0.941	**0.756**	48.67	0.626	0.085	7.60
		best_auc	0.941	**0.756**	48.38	0.634	0.079	6.61
		best_bedroc	0.940	0.750	48.29	0.627	0.083	7.32
16	4,096	last	0.933	0.725	46.39	0.630	0.075	6.27
		best_auc	0.928	0.693	44.07	0.616	0.075	6.01
		best_bedroc	0.932	0.717	45.82	0.630	0.067	5.55
32	8,192	last	0.939	0.738	47.13	0.622	0.085	7.25
		best_auc	0.941	0.741	47.44	0.624	0.083	6.75
		best_bedroc	0.933	0.705	45.11	0.611	0.075	5.88
64	16,384	last	0.937	0.748	48.37	0.625	0.081	6.83
		best_auc	0.936	0.731	46.88	0.622	0.074	5.58
		best_bedroc	0.936	0.743	47.79	0.625	0.075	6.58

Table A16: Multi-seed `last`-checkpoint comparison for CAUSALBIND$_{\mathrm{LR}}$ and the main CAUSALBIND$_{\mathrm{SP}}$ reference.

Method	n	DUD-E			LIT-PCBA		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
CAUSALBIND$_{\mathrm{LR}}$ (r=1)	3	0.937±0.004	0.732±0.011	46.87±0.91	0.630±0.006	0.085±0.006	7.22±0.31
CAUSALBIND$_{\mathrm{LR}}$ (r=1) +L_{spa}	3	0.935±0.003	0.727±0.007	46.39±0.58	0.623±0.007	0.083±0.005	7.51±0.94
CAUSALBIND$_{\mathrm{LR}}$ (r=8) +L_{spa}	3	0.940±0.001	0.750±0.011	48.30±0.97	0.621±0.005	0.087±0.003	7.73±0.40
CAUSALBIND$_{\mathrm{SP}}$ main	4	0.938±0.004	0.746±0.004	47.99±0.39	0.631±0.006	0.088±0.007	7.55±0.73

Table A15 shows that very small ranks are already competitive. Rank r=1 is the most parameter-efficient setting and reaches the best LIT-PCBA EF@1% under `best_bedroc`, while r=8 gives the strongest DUD-E early-enrichment numbers. Increasing the nominal rank beyond this range does not provide a consistent gain, suggesting that the useful interaction signal is much lower-dimensional than the full $K \times K$ mask.

A6.5.2 Multi-Seed Comparison and Ablation

We next compare the most relevant CAUSALBIND$_{\mathrm{LR}}$ settings across three seeds under the fixed-endpoint `last` protocol. The default row uses r=1 with $\lambda_{\mathrm{spa}}^{\mathrm{LR}} = 0$; the $+L_{\mathrm{spa}}$ row adds the optional mask penalty. The table also includes the main CAUSALBIND$_{\mathrm{SP}}$ multi-seed reference so that the low-rank implementation can be read against the main model directly.

The key observation is that CAUSALBIND$_{\mathrm{LR}}$ with r=1 essentially matches the main dense-mask CAUSALBIND$_{\mathrm{SP}}$ within one standard deviation while using only 256 mask parameters per branch instead of 16,384. Adding L_{spa} to the rank-one model changes little, indicating that the hard rank constraint already provides the structural compression that the L_1 term induces only indirectly in the dense-mask version. Rank r=8 improves DUD-E, but the LIT-PCBA early-enrichment metrics are already close to saturated at rank one.

A6.5.3 Hyperparameter Robustness

The rank-one implementation also changes how the model behaves as the concept count grows. In the dense CAUSALBIND$_{\mathrm{SP}}$ mask, increasing K expands the mask as K^2 and can over-parameterize the interaction layer. In CAUSALBIND$_{\mathrm{LR}}$ with r=1, the mask has only $2K$ parameters per branch, so increasing K mainly changes the number of available concepts rather than the rank of their interaction.

Table A17: CAUSALBIND$_{\mathrm{LR}}$ K-scan with r=1, seed=1, all three protocols. All rows use d_c=1024, L=4, and λ=10^{-4}.

K	Params/branch	Protocol	DUD-E			LIT-PCBA		
			AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
64	128	last	0.935	0.704	44.87	0.632	**0.103**	8.84
		best_auc	0.934	0.702	44.63	**0.638**	0.099	**9.03**
		best_bedroc	0.936	0.716	45.92	0.631	0.095	6.54
128	256	last	0.935	0.720	45.86	0.630	0.087	8.41
		best_auc	**0.946**	**0.756**	**48.78**	0.636	0.079	6.53
		best_bedroc	0.935	0.724	46.28	0.632	0.089	8.56
256	512	last	0.935	0.734	47.16	0.622	0.074	5.82
		best_auc	0.936	0.744	47.91	0.617	0.077	6.79
		best_bedroc	0.938	0.740	47.77	0.629	0.066	5.72

Table A18: Mask diagnostics for the CAUSALBIND$_{\mathrm{LR}}$ rank scan. The reported branch uses the same protocol as Table A9.

r capacity	$\overline{\sigma(\mathcal{M})}$	%zeros	r@95%	r@99%	Interpretation
1	0.86	0.0%	1	1	rank-one by construction
4	0.63	0.0%	1	1	trained solution collapses to rank one
8	0.54	8.9%	1	1	some thresholded zeros, still rank one
16	0.51	23.3%	1	1	more zeros, same effective rank
32	0.51	34.3%	1	1	unused rank capacity
64	0.50	39.7%	1	1	unused rank capacity
CAUSALBIND$_{\mathrm{SP}}$ dense ref.	0.76	0.0%	1	9–14	rank-one at 95%, richer 99% tail

Table A17 shows that the rank-one constraint protects the model from the large-K collapse observed in the dense concept-pair sweep: K=256 remains strong on DUD-E instead of failing under an oversized K^2 mask. At the other end, K=64, r=1 gives the best LIT-PCBA early-enrichment numbers in this scan, suggesting that the property-debiased benchmark can be served by a very small interaction subspace.

A6.5.4 Mask Analysis

Finally, we analyze whether the explicit rank capacity is actually used by the optimizer. The diagnostics use the same thresholded effective mask and singular-value energy criteria as App. A6.1.7.

Across all nominal capacities $r \in \{1, 4, 8, 16, 32, 64\}$, the trained CAUSALBIND$_{\mathrm{LR}}$ mask has r@95%=r@99%=1. This provides a useful check on the theory–practice story: the low-rank behavior observed in the dense CAUSALBIND$_{\mathrm{SP}}$ mask is not merely an artifact of the $\tanh + L_1$ parameterization, because an explicitly rank-constrained model also converges to a rank-one interaction.

A6.6 CAUSALBIND$_{\mathrm{EMB}}$: Experimental Results and Diagnostics

The implementation and parameter settings of CAUSALBIND$_{\mathrm{EMB}}$ are given in App. A5.4, and the three-way architectural comparison is summarized in App. A5.5. Here we only report its retrieval behavior and mask diagnostics. This keeps A5 focused on implementation and keeps the present A6 section focused on evidence.

A6.6.1 Hyperparameter Study on K and d_c

We first report the embedding-level K/d_c sweep that determines the representative CAUSALBIND$_{\mathrm{EMB}}$ configuration used below. Unlike CAUSALBIND$_{\mathrm{SP}}$ and CAUSALBIND$_{\mathrm{LR}}$, this implementation applies the mask after concept pooling, so these scans should be read as pooled-space sensitivity analyses rather than concept-pair support analyses.

The embedding-level sweeps are smoother than the concept-pair sweeps, but the best operating point is still concentrated around moderate concept count and intermediate channel width. We therefore use K=64, d_c=256, and λ=10^{-2} as the representative CAUSALBIND$_{\mathrm{EMB}}$ configuration in Table A21.

Table A19: CAUSALBIND$_{\text{EMB}}$ K-scan (fix d_c=256), seed=1, all three protocols.

K	Protocol	DUD-E			LIT-PCBA		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
32	last	0.942	0.743	47.35	0.638	0.090	7.48
	best_auc	0.942	0.747	47.84	0.640	0.085	6.86
	best_bedroc	0.937	0.716	45.79	0.643	0.066	4.38
64	last	0.939	0.754	48.56	0.636	**0.102**	**9.60**
	best_auc	0.941	0.765	49.47	0.635	0.097	8.18
	best_bedroc	0.939	0.751	48.33	0.636	0.101	9.25
128	last	0.935	0.715	45.62	0.623	0.082	6.50
	best_auc	0.938	0.713	45.29	0.630	0.084	7.50
	best_bedroc	0.938	0.701	44.49	0.630	0.080	7.01
256	last	0.936	0.737	47.55	0.624	0.080	7.08
	best_auc	0.937	0.739	47.69	0.623	0.080	7.39
	best_bedroc	0.934	0.728	46.92	0.624	0.081	7.04

Table A20: CAUSALBIND$_{\text{EMB}}$ d_c-scan (fix K=64), seed=1, all three protocols. d_c=256 is strongest on LIT-PCBA early enrichment.

d_c	Protocol	DUD-E			LIT-PCBA		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
128	last	0.943	0.732	47.09	0.606	0.074	5.88
	best_auc	0.944	0.726	46.59	0.617	0.075	6.97
	best_bedroc	0.943	0.738	47.49	0.602	0.071	6.03
256	last	0.939	0.754	48.56	0.636	**0.102**	**9.60**
	best_auc	0.941	0.765	49.47	0.635	0.097	8.18
	best_bedroc	0.939	0.751	48.33	0.636	0.101	9.25
512	last	0.942	0.750	48.32	0.624	0.079	6.45
	best_auc	0.945	0.747	48.11	0.625	0.078	6.68
	best_bedroc	0.943	0.753	48.53	0.619	0.079	6.43
1024	last	0.940	0.745	48.01	0.622	0.088	8.12
	best_auc	0.943	0.760	48.82	0.628	0.090	8.47
	best_bedroc	0.935	0.716	45.73	0.622	0.077	6.92

A6.6.2 Experimental Results

Main comparison. Table A21 now puts the three relevant models under the same three checkpoint-selection protocols, so the comparison is directly aligned with the earlier protocol analyses in this appendix. The two CAUSALBIND variant rows use the representative configurations selected earlier in their respective sweeps: CAUSALBIND$_{\text{SP}}$ is the paper's main operating point, while CAUSALBIND$_{\text{EMB}}$ is the strongest embedding-level setting identified by the K/d_c scan. Two patterns stand out. First, both CAUSALBIND variants remain clearly stronger than HypSeek across all three protocols, especially on LIT-PCBA early-enrichment metrics. This confirms that the gain is not tied to one specific checkpoint rule. Second, the sparse concept-pair and embedding-level variants emphasize slightly different strengths: CAUSALBIND$_{\text{SP}}$ gives the best LIT-PCBA AUROC under `last`/`best_bedroc`, whereas CAUSALBIND$_{\text{EMB}}$ is especially strong on LIT-PCBA BEDROC$_{80.5}$ / EF@1% under `last`, and also gives the best DUD-E numbers under `best_auc`.

What this means empirically. The purpose of this section is not to argue that CAUSALBIND$_{\text{EMB}}$ should replace the main model, but to show that the sparse/low-rank interaction prior is effective beyond a single implementation. CAUSALBIND$_{\text{SP}}$ remains the paper's primary architecture because it offers a cleaner causal interpretation and aligns more directly with the concept-pair theory. CAUSALBIND$_{\text{EMB}}$, however, demonstrates that a simpler pooled-space realization can also produce strong retrieval, which makes the overall design principle more credible.

Table A21: Three-protocol comparison among HypSeek, CAUSALBIND$_{\text{SP}}$, and CAUSALBIND$_{\text{EMB}}$. CAUSALBIND$_{\text{SP}}$ uses K=128, d_c=1024, λ=10^{-4}, L=4, which is the main operating point analyzed in the paper. CAUSALBIND$_{\text{EMB}}$ uses K=64, d_c=256, λ=10^{-2}, selected as the strongest setting in the embedding-level sweep. All rows are seed-1 results.

Model	Protocol	**DUD-E**			**LIT-PCBA**		
		AUROC	BEDROC$_{80.5}$	EF@1%	AUROC	BEDROC$_{80.5}$	EF@1%
HypSeek [Wang et al., 2026]	last	0.909	0.605	37.83	0.603	0.061	4.62
	best_auc	0.913	0.629	39.63	0.608	0.063	5.13
	best_bedroc	0.923	0.676	42.71	0.607	0.063	4.56
CAUSALBIND$_{\text{SP}}$	last	0.935	0.744	47.69	**0.639**	0.097	8.56
	best_auc	0.934	0.722	46.41	0.628	0.074	5.84
	best_bedroc	0.935	0.747	48.24	**0.643**	0.094	8.42
CAUSALBIND$_{\text{EMB}}$	last	0.939	0.754	48.56	0.636	**0.102**	**9.60**
	best_auc	**0.941**	**0.765**	**49.47**	0.635	0.097	8.18
	best_bedroc	0.939	0.751	48.33	0.636	0.101	9.25

A6.6.3 Theory Correspondence

The theory correspondence of the embedding-level variant is simpler, but also narrower, than for the main concept-level model. Because the mask acts on pooled embeddings rather than on explicit concept pairs, it does not naturally match the sparse-support picture behind Condition 2.i and Theorem 2. In particular, there is no direct analogue of concept-pair support geometry or the antichain-style condition discussed for the concept-level mask. For that reason, this implementation should not be presented as evidence for the sparse identifiability theorem.

By contrast, the connection to the low-rank theorem is much more natural. The embedding-level mask is exactly a constrained interaction operator in a d_c-dimensional shared space, so asking whether useful solutions are low-rank is directly meaningful here. In that sense, the variant is much closer in spirit to Condition 3.i and Theorem 3: it tests whether a small number of interaction modes in the pooled cross-modal space is already sufficient for strong retrieval.

Table A22: Spectral diagnostics for the embedding-level mask (K=64, d_c=256, λ=10^{-2}, seed 1). We report the two branchwise product masks used in the paper's interaction view, evaluated under the `last` checkpoint.

Branch	%zeros	r@95%	r@99%
mol–pocket	0.0%	1	46
mol–sequence	0.0%	1	46

Table A22 makes the interpretation more concrete. The embedding-level mask exhibits no exact zeros after gating, so it provides no support for the sparse-support assumptions behind Theorem 2. At the same time, its spectral profile is clearly compressed: for both branches, a single dominant mode already explains 95% of the squared singular-value energy, and only 46 modes are needed to explain 99% out of the full ambient dimension d_c=256. This is not a strict low-rank matrix in the algebraic sense, but it is a strongly *effectively* low-rank interaction operator. We also checked nearby late checkpoints (epochs 40, 45, and 50), and the same r@95%=1, r@99%=46 pattern persists, so this is not a single-checkpoint artifact.

Across all scanned configurations, the embedding-level variant is also somewhat smoother than the concept-level one: its DUD-E AUROC varies within a relatively narrow band (0.935–0.943, n=7), whereas the concept-level sweeps in Tables A5 and A6 span a wider range. This difference is plausible architecturally. The embedding-level mask acts after pooling and therefore modulates a coarser representation; the concept-level mask, in contrast, directly shapes the fine-grained interaction structure before pooling. Taken together, these observations suggest a clean division of roles: the concept-level model is the better vehicle for the paper's causal and sparse-support story, while the embedding-level alternative provides additional empirical support for the low-rank interaction view.

A6.7 Overall Observations Across Three Implementations

What remains stable across the sweep. The most important observation is that the main qualitative conclusion does not depend on a single hyperparameter choice, checkpoint protocol, or implementation variant. Across the main CAUSALBIND$_{\mathrm{SP}}$ comparison (Table A4), the explicit low-rank results (Tables A15–A18), and the embedding-level alternative (Tables A19–A21), constrained cross-modal interaction consistently improves over dense holistic alignment on the retrieval metrics emphasized in the main text, especially early enrichment on LIT-PCBA.

How the three implementations differ. CAUSALBIND$_{\mathrm{SP}}$ is the main paper model: it keeps explicit concept-pair structure and is therefore the cleanest implementation for studying sparse support, mask calibration, and concept-level interpretation. CAUSALBIND$_{\mathrm{LR}}$ is the direct low-rank concept-basis implementation: it matches the main dense-mask model with far fewer mask parameters and shows that rank-one or near-rank-one interaction is often sufficient. CAUSALBIND$_{\mathrm{EMB}}$ is a pooled-space alternative: it can be very strong empirically, especially on LIT-PCBA early enrichment, but its mask acts after concept pooling and therefore has a weaker connection to component-wise concept identifiability.

What the hyperparameter studies collectively suggest. The CAUSALBIND$_{\mathrm{SP}}$ $K/d_c/\lambda/L$ scans point to a coherent operating regime: moderate concept granularity, adequate channel width, and moderate interaction depth work best, while overly large dense concept-pair masks can become brittle. The CAUSALBIND$_{\mathrm{LR}}$ scans sharpen this lesson by showing that explicit rank control protects against the large-K collapse and that increasing rank beyond a small value rarely helps. The CAUSALBIND$_{\mathrm{EMB}}$ scans are smoother, but the strongest setting is still a moderate pooled representation (K=64, d_c=256).

How the sweep connects back to theory. Taken together, the three implementations separate form from empirical regime. CAUSALBIND$_{\mathrm{SP}}$ has the sparse concept-pair form of Theorem 2, but its strongest λ=10^{-4} setting behaves more like a low-rank solution; at λ=10^{-3} it moves closer to sparse support. CAUSALBIND$_{\mathrm{LR}}$ directly realizes the subspace-identifiability form of Theorem 3 and confirms that the learned interaction can be rank one in practice. CAUSALBIND$_{\mathrm{EMB}}$ supports the broader low-rank interaction principle but should be interpreted as a pooled-space variant rather than a component-wise concept-identifiability result.

A6.8 Discussion on Theory–Practice Gap

The theoretical results in Section 2 are intentionally sharp: Theorem 2 describes what becomes identifiable under the sparse-support assumptions, while Theorem 3 gives a subspace-identifiability statement under low-rank interaction. The experiments in App. A6 show that these two theorems are both useful for interpretation, but not in the same way. This subsection makes that distinction explicit by connecting the compact mask statistics in Table 3 with the branchwise diagnostics in Table A9.

Main takeaway. The central lesson is not that the learned interaction becomes exactly element-wise sparse. Rather, the causal prior pushes training toward structurally simpler cross-modal interactions: stronger sparsification moves CAUSALBIND$_{\mathrm{SP}}$ toward the sparse-support picture, while milder sparsification and explicit CAUSALBIND$_{\mathrm{LR}}$ reveal a low-rank regime. Thus low-rank subspace identifiability is not merely a secondary theoretical case, but an important description of the regime that modern neural training naturally reaches.

What the experiments do support. The experiments support both theorem regimes, but through different implementations and operating points. In Table 3, 'L_1 avg' is the average absolute mask entry, %zeros is the fraction of entries thresholded to exact zero, and $r@95\%$/$r@99\%$ are effective ranks explaining 95%/99% of squared singular-value energy. For CAUSALBIND$_{\mathrm{SP}}$, λ=10^{-3} yields strong DUD-E retrieval together with meaningful exact sparsity (20.5% zeros in the compact table), matching the sparse-support picture of Theorem 2; the default λ=10^{-4} setting remains dense entry-wise (0% zeros) but is strongly low-rank in the dominant energy component ($r@95\% = 1$), matching Theorem 3. For $\lambda \geq 10^{-2}$, the mask enters an excessive hard-zero regime and performance degrades sharply, so sparsity alone is not the desired behavior. In CAUSALBIND$_{\mathrm{LR}}$, the rank constraint is imposed directly and the learned masks still collapse to effective rank one, which independently

supports the low-rank/subspace interpretation. In CAUSALBIND$_{\mathrm{EMB}}$, the same low-rank tendency appears after pooling, but this evidence is architectural rather than component-wise.

Why this is not a contradiction. The theorem statements are intentionally sharp, whereas optimization can realize more than one useful structural regime. A dense $K \times K$ mask with an L_1 penalty can behave as sparse support under stronger regularization or as a low-rank operator under milder regularization; an explicit rank-factorized mask enforces the latter by construction. The empirical message is therefore not that one theorem replaces the other, but that the two theorem regimes identify different useful limits of the same causal interaction prior.

Which theorem covers the implementation. One gap between the theory and the implementation deserves to be stated plainly. For simplicity, all three variants use an inner product between concept tokens as the compatibility function ϕ, which is *bilinear*. Condition 2.ii therefore fails, and by Theorem 1 the latent concepts are not component-wise identifiable in this model class, whatever structure is imposed on $\mathcal{M}$: there exist reparameterizations that mix concepts while leaving both the mask and the selected distribution unchanged. Theorem 2 consequently does not certify that the concepts CAUSALBIND learns are the true ones, and an antichain support observed in a learned mask is a structural diagnostic rather than a certificate.

The result that does apply is Theorem 3, which places no non-bilinearity requirement on ϕ: for $\phi(x, y) = xy$ the centred kernel has rank $\mathrm{rank}(\mathcal{M})$, so a low-rank mask satisfies Condition 3.i, and the identified objects are the molecule-side and protein-side interaction subspaces. This matches what we observe: at the default $\lambda_{\mathrm{spa}}{=}10^{-4}$ the learned masks have no hard zeros but $r@95\%{=}1$ (Table 3), and CAUSALBIND$_{\mathrm{LR}}$ imposes the same structure by construction. A second, smaller mismatch remains even there: the theory treats scalar concepts ($d_c{=}1$), whereas the implementation uses $d_c{>}1$ concept tokens, which raises the kernel rank to roughly $\mathrm{rank}(\mathcal{M}) \cdot d_c$, so Condition 3.i holds for the theoretical model rather than verbatim for the implemented one.

We therefore position CAUSALBIND as a causally motivated structural prior rather than as an estimator with a component-wise identifiability certificate. The theory says why unconstrained bilinear interaction cannot identify concepts, which structural restrictions would make them identifiable, and which weaker object (the interaction subspace) remains identifiable in the class we implement. Closing the gap is a concrete next step: replacing the inner product by a non-bilinear compatibility function, for instance $\tanh(\langle \mathbf{c}_i, \mathbf{c}_j \rangle)$ or a small shared network, would place the model inside the class of Theorem 2, and we leave this to future work.

A7 Limitations, Broader Impact, and Usage of LLMs.

Summary (moved from the main text). Our evidence is centered primarily on retrieval-based virtual screening under cross-dataset generalization, especially the two main benchmarks, DUD-E and LIT-PCBA, with an additional Merck-FEP-8 affinity-ranking transfer analysis in App. A6.1.9 and target- and scaffold-level out-of-distribution evaluation on DEKOIS 2.0 in App. A6.3. This is the right, deliberately focused scope for the main question of the paper, but it does not yet establish uniform gains across sequence-family or time-stratified splits or larger benchmark collections, and our interpretability analysis provides model-level attribution rather than proof of a physical causal mechanism (see App. A7). As with any virtual-screening method, our model could be misused to prioritize ligands for harmful targets. Although it is a scoring tool rather than a generator, downstream deployment should still be paired with safety and experimental validation filters before any wet-lab follow-up. All source code is publicly available in our GitHub repository to support reproducibility.

Limitation on scope of the current evidence. This paper should be read primarily as a study of *retrieval-based virtual screening under cross-dataset generalization*, not as a complete evaluation of protein-molecule modeling in every downstream regime. Accordingly, our main evidence primarily comes from two retrieval benchmarks, DUD-E and LIT-PCBA, while App. A6.1.9 provides a supplementary Merck-FEP-8 affinity-ranking transfer check; the strongest claims still concern early enrichment rather than absolute affinity calibration. This scope is deliberate: the central question of the paper is whether replacing dense holistic alignment with sparse or low-rank cross-modal interaction improves the high-throughput screening stage. It also means that broader claims should

be made cautiously. The current experiments do not yet establish how uniformly the gain transfers across protein families, scaffold splits, time-stratified settings, or much larger benchmark collections.

Broader impact and misuse clarification. Better virtual screening can reduce the cost of early discovery by improving which candidates are advanced to docking and experimental follow-up, potentially accelerating the identification of useful therapeutics. At the same time, any method that improves prioritization of protein-molecule interactions could in principle be misused to search for ligands against harmful targets. Our model is less directly dual-use than a generative molecular-design system, because it is a scoring and ranking tool rather than a molecule generator, but this does not remove the need for safeguards. In practical use, its predictions should be treated as one stage in a controlled pipeline and paired with downstream safety, ADMET, toxicity, and experimental validation filters before any wet-lab deployment.

Usage of LLMs. LLMs were used to assist with writing and editing, but they are not the source of any scientific claims or empirical results in this paper. All technical content is derived from our own theoretical analysis and experimental work, and all key numbers are directly traceable to our codebase and training logs. We did not use LLMs to generate scientific hypotheses. Their role was limited to language polishing and formatting assistance, which we carefully reviewed and edited for accuracy and clarity.

A8 Reproducibility and Asset Provenance

Code basis and development path. The implementation of CAUSALBIND was developed by extending the retrieval pipeline used in HypSeek and cross-checking evaluation behavior against LigUnity. Concretely, our code keeps the same broad three-stream retrieval setting as HypSeek, while replacing the dense shared-space interaction with the concept extractor and sparse/low-rank cross-modal masking modules described in Section 3. The implementation is thus a method-development layer on top of existing retrieval infrastructure, not a from-scratch rewrite.

Main software assets. The most important inherited software assets are the HypSeek code base, the LigUnity code base used for comparison and protocol alignment, the Uni-Mol/Uni-Core stack used by the molecular and pocket branches, and the ESM-2 protein language model used by the sequence branch. Many inherited Uni-Mol-style source files carry MIT license headers, and the HypSeek repository is MIT-licensed. The LigUnity repository releases code under Apache 2.0 and its data under CC BY-NC 4.0. These upstream assets are credited throughout the paper via citations to HypSeek, LigUnity, Uni-Mol, Uni-Core, and ESM-2.

Data assets. The training and evaluation data are not newly created by this paper. As described in Section 4 and Appendix A5, training follows the same assay-level setup used by LigUnity and HypSeek, drawing from ChEMBL, BindingDB, and PDBBind, where ChEMBL/BindingDB assays whose target UniProt IDs appear in the DUD-E, LIT-PCBA, or DEKOIS 2.0 benchmarks are removed. The main evaluation benchmarks are DUD-E and LIT-PCBA, and the appendix additionally reports transfer results on Merck-FEP-8. All are existing public resources and remain subject to the terms and redistribution rules of their original providers; we do not re-license them.

What is provided for reproducibility. The code is publicly available at `https://github.com/lokali/CausalBind`. The repository contains the CAUSALBIND implementation (the core model, loss, and task files for all variants), scripts that download the public training data, benchmarks, and pretrained backbones from their original sources, training and evaluation entry points with the exact configurations reported in the paper, and the tools for the mask-structure analysis, the DEKOIS 2.0 out-of-distribution evaluation, and the interpretability case study. We also release the trained checkpoints of the reported models together with a one-command script that reproduces the DUD-E and LIT-PCBA results. Processed datasets are not redistributed; they remain subject to the terms of their original providers.

Scope of the current license discussion. This section makes the provenance of the main assets explicit and clarifies which parts are inherited versus newly added by CAUSALBIND. We do not re-publish upstream assets under a unified license: the new code is layered on existing open-source assets, and each inherited component remains under its original license.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the same three contribution pillars developed in the paper: the causal formulation, the constrained interaction design, and the empirical validation on retrieval benchmarks. The scope is also kept narrow and consistent throughout, with the strongest claims restricted to retrieval-based virtual screening and cross-dataset generalization.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A7 discusses the main limitations, including the benchmark scope, the focus on retrieval rather than full affinity calibration, and the fact that broader transfer settings remain to be tested. It also states that the implemented bilinear compatibility function is not covered by the component-wise identifiability theorem. The discussion also clarifies the intended use of the method and where stronger claims would require additional evidence.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 2 states the main assumptions and theorem statements, and Appendix A4 provides the complete assumptions and full proofs in the corresponding theorem subsections. Theorems, propositions, assumptions, and proof dependencies are explicitly numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 gives the main evaluation protocol, datasets, and model-selection rule, while Appendices A5–A6 provide implementation details, concrete parameter assignments, protocol variants, and the hyperparameter scans behind the reported operating points. Together these sections disclose the information needed to reproduce the main comparisons and analyses.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide detailed methodological and experimental information in Section 4 and Appendices A5–A6 to support reproducibility. All training and evaluation data are publicly available from their original providers, and our source code, trained checkpoints, and scripts for obtaining the data and reproducing the main results are publicly released at `https://github.com/lokali/CausalBind`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 summarizes the training data, evaluation benchmarks, and main protocol, and Appendix A5 gives the concrete optimization settings, loss weights, and implementation details. Appendix A6 further documents how the main operating points were selected through protocol and hyperparameter analyses.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 2 reports mean $\pm$ standard deviation across multiple seeds for the key ablation comparisons, and Appendix A6 further provides multi-seed and multi-protocol robustness analyses for the strongest methods in the main experimental thread. We do not run the full Table 1 baseline block with multiple seeds, because of the number of baselines and limited compute, but we do report variance for the most important comparisons, including the SOTA baselines LigUnity and HypSeek in Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The main paper and Appendix A5 explicitly state the key compute information needed to reproduce the experiments, namely the use of four NVIDIA A100 GPUs together with the training setup and optimization details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and, to the best of our knowledge, the research conforms to it. The paper also includes explicit discussion of limitations, misuse risks, and responsible deployment considerations in Section 5 and Appendix A7.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 5 and Appendix A7 discuss both the positive impact of better prioritization in early-stage drug discovery and the negative or dual-use risk of improving ligand prioritization for harmful targets. They also note that the method should be used within a controlled downstream validation pipeline.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This submission does not release a high-risk generative model, scraped dataset, or comparable asset requiring a dedicated controlled-release mechanism. The method studied here is a ranking model for virtual screening, and the paper discusses misuse considerations in Section 5 and Appendix A7.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the main inherited assets used in this work, including HypSeek, LigUnity, Uni-Mol/Uni-Core, ESM-2, and the public datasets and benchmarks such as ChEMBL, BindingDB, PDBBind, DUD-E, and LIT-PCBA. Appendix A8 summarizes their provenance and license or terms-of-use information, and clarifies that inherited components remain under their original licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the CAUSALBIND source code and trained checkpoints at `https://github.com/lokali/CausalBind`. The repository documents installation, data preparation, training, evaluation, and the analysis tools, and states the licenses: our code is released under the MIT License, and inherited third-party components remain under their original licenses as listed in the repository. No new dataset is introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects. All experiments are computational and use existing molecular and protein datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects or participant studies. Accordingly, IRB review and participant-risk disclosures are not applicable here.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, theory, or experiments in this work, so declaration is not required under the checklist question. For transparency, Appendix A7 still notes that LLMs were used only for writing and editing assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.