Read but Don't Retrieve: Decomposing the Molecule-Invariance Failure in Biomedical RAG

Anonymous ACL submission

Abstract

Frontier LLMs identify the right PubMed abstracts for a molecule from its canonical SMILES, randomized SMILES, InChIKey, or Murcko scaffold with Recall@10 of 0.86–0.94 *when those abstracts are placed in the candidate pool*. In a realistic RAG setup over a 7,125-abstract corpus, the same models retrieve ≤ 0.014. We localise this 60× gap with MolRepBench, a controlled molecule-invariance benchmark of 312 drugs, five surface forms, and 1,555 query variants over fixed gold evidence. Three findings: (i) BM25, BGE-small, PubMedBERT, E5-small, and MedCPT all collapse from 0.55–0.62 Recall@10 on names to ≤ 0.005 on chemical strings, and four frontier LLM rerankers do not repair this because BM25's top-50 already contains no gold; (ii) an oracle-pool reranker that injects up to 10 missing gold abstracts into the same pool lifts chemical-rep Recall@10 to 0.86–0.94 for all four LLMs we test – GPT-4o-mini, Claude 3.5 Haiku, Llama-3.3-70B, Qwen-2.5-72B (paired Wilcoxon $p < 10^{-200}$ per model, chemical reps aggregate) – isolating the failure to candidate generation rather than chemical understanding; (iii) cheap query-side normalisation (BM25+QE, BGE+QE) already recovers much of the loss without document-side indexing. The takeaway for molecular RAG is operational: fix the upstream pool, not the reranker.

1 Introduction

Retrieval-augmented generation is now standard for biomedical question answering (Lewis et al., 2020; Karpukhin et al., 2020; Xian et al., 2025), but the query entity is not stable at the string level for chemistry. The same compound may appear as a common drug name, a canonical SMILES, a randomized SMILES, an InChIKey, or a Murcko scaffold (Weininger, 1988; Heller et al., 2015; Bemis and Murcko, 1996). An experienced chemist maps between these forms without hesitation; a retriever should too. If substituting, say, *terfenadine* with an equivalent SMILES changes the retrieved evidence, then any downstream claim such as "the LLM cannot reason over SMILES" is confounded: the model may simply have seen a different candidate pool.

Prior work documents that LLMs struggle with SMILES-centered reasoning (Sadeghi et al., 2024; Guo et al., 2023; Jang et al., 2025). Our question is upstream: *holding user intent and gold evidence fixed, does a biomedical retriever return the same documents when only the molecular surface form changes, and if not, where in the pipeline the failure may be located?* We answer this with MolRepBench, a controlled invariance benchmark, plus an oracle-pool control that cleanly separates two confounded hypotheses: "LLMs cannot read chemical strings" from "the candidate pool never contained the right documents".

Contributions. (1) We release MolRepBench[1]: 312 drugs, five surface forms, 1,555 query variants, 2,625 PubMed gold abstracts, and 4,500 distractors, with shared gold under the invariance hypothesis. (2) We show every standard text-first retriever and four frontier LLM rerankers collapse on chemical strings (≤ 0.014 Recall@10) while reaching 0.55–0.79 on names. (3) An oracle-pool reranker lifts chemical-rep Recall@10 to 0.86–0.94 for all four frontier LLMs we test, isolating the failure to candidate generation. (4) Deterministic query-side normalisation (BM25+QE, BGE+QE) recovers much of the gap without any LLM, identifying a cheap and deployable mitigation.

2 MolRepBench: A Controlled Invariance Benchmark

Molecules and representations. We start from a hand-curated set of 315 INN drugs spanning

[1] Code and data (anonymised for review): https://anonymous.4open.science/r/MolRepBench-308D

eleven therapeutic classes. For each we resolve a canonical PubChem CID and collect canonical SMILES, InChIKey, IUPAC name, and synonyms via PUG-REST (Kim et al., 2023), and compute randomized SMILES and Bemis–Murcko scaffold SMILES with RDKit (Landrum, 2013; Arús-Pous et al., 2019). Heparin lacks a unique CID and is removed; of the remaining 314, 312 have at least two usable gold abstracts. Each molecule is represented in up to five forms: NAME, canonical SMILES, randomized SMILES, INCHIKEY, and Murcko SCAFFOLD.

Evidence corpus. For each drug we issue a PubMed TIAB query using the INN and up to two common synonyms, take up to 12 results, and keep abstracts of ≥ 150 characters that explicitly mention the molecule under one of its known surface forms. This yields 2,625 gold abstracts (mean 8.6 per drug). We add 4,500 distractors sampled from twenty broad biomedical themes; the final corpus contains 7,125 title–abstract pairs.

Queries and invariance target. We wrap every representation in a fixed template: *"Find biomedical papers reporting pharmacological activity, targets, or adverse effects for {MOL}."* Only the molecular string changes. Across 312 drugs and five representations this yields 1,555 queries. Gold judgments are shared across representations: if two strings denote the same molecule, they inherit the same relevant abstract set. That shared gold set is the invariance hypothesis under test.

3 Methods

Retrievers. BM25 (Robertson and Zaragoza, 2009); three text-first dense retrievers (BGE-small-en-v1.5 (Xiao et al., 2024), PubMedBERT embeddings (Gu et al., 2021), E5-small (Wang et al., 2022)); a biomedical dual-encoder MedCPT (Jin et al., 2023); and one biomedical cross-encoder, MedCPT-Cross-Encoder, reranking BM25's top-50. We also evaluate four LLM listwise rerankers over the same top-50 pool (Sun et al., 2023): GPT-4o-mini (Achiam et al., 2023), Claude 3.5 Haiku (Anthropic, 2024), Llama-3.3-70B (Dubey et al., 2024), and Qwen-2.5-72B (Team, 2024).

Oracle-pool reranking. For each query we replace the lowest-BM25 items in the top-50 with up to 10 missing gold PMIDs (capped by the number of golds not already in the pool), shuffle to neutralise position priors, and rerun the same listwise prompt. Pool size and LLM context budget are unchanged; only pool membership moves. We run this for all four LLM rerankers (GPT-4o-mini, Claude 3.5 Haiku, Llama-3.3-70B, Qwen-2.5-72B).

Retriever	name	InChIK	c.SMI	r.SMI
BM25	0.550	0.002	0.004	0.002
BGE-small	0.600	0.002	0.003	0.004
PubMedBERT	0.237	0.004	0.002	0.005
E5-small	0.621	0.003	0.002	0.001
MedCPT	0.507	0.002	0.003	0.004
LLM rerankers	.77–.79	.013–.014	.007–.009	.010–.011
BM25+QE	0.738	0.451	0.082	0.139
BGE+QE	0.650	0.483	**0.436**	**0.441**
Hybrid	**0.822**	**0.671**	0.153	0.274

Table 1: Recall@10 across the four main representations. Full per-model results including SCAFFOLD, all $K \in \{10, 50, 100, 200\}$, and 95% CIs are in App. A. LLM rerankers add nothing on chemical strings because BM25's top-50 already has Recall@50 ≤ 0.015 there.

Query-side normalisation. A cheap baseline maps each query identifier to a PubChem CID and appends the top-3 synonyms and IUPAC name to the query before running BM25 or BGE-small (BM25+QE, BGE+QE). We also report a normalisation-aware hybrid that adds a binary indicator that the document mentions any synonym of the CID-resolved molecule (weight β=25 on top of BM25; details in App. B).

Evaluation. Recall@10, Recall@50, and nDCG@10 over the 312 drugs with ≥ 2 gold abstracts, with paired bootstrap 95% confidence intervals.

4 Results

4.1 Text-First Retrievers Collapse Outside Names

Table 1 gives the central baseline result. BM25, BGE-small, PubMedBERT, E5-small, and MedCPT all work on names (0.24–0.62 Recall@10) and collapse to ≤ 0.005 on every chemical representation. LLM rerankers help only on names (0.77–0.79); on chemical strings they stay at 0.007–0.014 across two closed-source vendors and two open-weight 70B models. The mechanism is mechanical: the reranker only sees BM25's top-50 pool, and for chemical strings that pool has Recall@50 ≤ 0.015, so there is no gold to surface. Deeper BM25 pools do not fix this either (App. C).

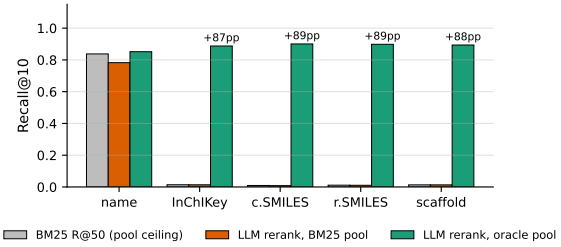

Figure 1: Oracle-pool decomposition. Replacing the lowest-BM25 items in the top-50 with up to 10 missing gold abstracts (pool size and LLM context budget unchanged) lifts chemical-representation Recall@10 from $\leq$ 0.014 to 0.86–0.94. Bars average four LLM rerankers (GPT-4o-mini, Claude 3.5 Haiku, Llama-3.3-70B, Qwen-2.5-72B); per-model values in Table 2. Paired Wilcoxon $p < 10^{-200}$ per model on the chemical-reps aggregate.

Setup	name	InChIK	c.SMI	r.SMI	scaf
BM25 R@50 (ceiling)	0.838	0.014	0.009	0.011	0.013
Baseline reranker (BM25 top-50 pool)					
GPT-4o-mini	0.792	0.014	0.009	0.011	0.013
Claude-Haiku	0.785	0.013	0.007	0.010	0.013
Llama-3.3-70B	0.779	0.014	0.007	0.010	0.013
Qwen-2.5-72B	0.774	0.014	0.009	0.011	0.013
Oracle-pool reranker (same prompt, gold injected)					
GPT-4o-mini	**0.872**	**0.876**	**0.914**	**0.910**	**0.886**
Claude-Haiku	**0.859**	**0.931**	**0.935**	**0.932**	**0.935**
Llama-3.3-70B	**0.820**	**0.866**	**0.868**	**0.863**	**0.873**
Qwen-2.5-72B	**0.855**	**0.877**	**0.885**	**0.889**	**0.881**

Table 2: Per-model oracle-pool numbers across four LLM rerankers. Top row is BM25 R@50 (the pool ceiling). Middle block is the baseline reranker on that pool; bottom block is the same reranker on the oracle pool. Lift oracle-vs-baseline is ≥ 0.85 on every chemical representation for every model (paired Wilcoxon $p < 10^{-50}$ per cell, $p < 10^{-200}$ in the chemical-reps aggregate per model).

Cheap query-side normalisation already changes the picture. BM25+QE deterministically resolves the identifier and appends synonyms, pushing INCHIKEY to 0.451 and randomised SMILES to 0.139 – already above every LLM reranker. Pairing the same normalisation with a dense encoder is stronger still: BGE+QE reaches 0.436–0.483 across all chemical representations. Hybrid is best on NAME and INCHIKEY where exact document-side indexing pays, but is dominated by BGE+QE on SMILES. Top-10 Jaccard between NAME and canonical SMILES is 0.011 for BM25 and 0.022 for BGE-small without normalisation, rising to 0.083 (BM25+QE) and 0.369 (BGE+QE); the systems are not returning the same documents in a different order, they are returning different evidence sets entirely (App. D).

4.2 Oracle Pool Isolates the LLM-Reranker Failure to Candidate Generation

The baseline LLM result above is consistent with two hypotheses we cannot otherwise distinguish: (i) the LLM cannot identify documents about a molecule from a SMILES or InChIKey, or (ii) the LLM never sees a relevant document because BM25's top-50 contains none. Figure 1 and Table 2 rule out (i). When up to 10 missing gold PMIDs are injected into the pool and shuffled, oracle Recall@10 on chemical representations reaches 0.91–0.94 for GPT-4o-mini and Claude 3.5 Haiku, and 0.86–0.89 for Llama-3.3-70B and Qwen-2.5-72B. The lift over the BM25-pool baseline is ≥ 85 percentage points on every chemical representation for every model, with paired Wilcoxon $p < 10^{-50}$ per rep, $p < 10^{-200}$ in the chemical-reps aggregate per model. Conditioned on the injected golds alone (those BM25 missed), GPT-4o-mini lifts 88–92% of them into the top-10 across chemical representations.

A small additional observation: chemical-rep oracle Recall@10 is at least as high as name oracle Recall@10 for every model. This is consistent with pool composition. BM25 chemical-rep pools are essentially uniform-random biomedical noise, so injected gold stands out clearly; name pools contain semantically similar but not gold-marked abstracts in the same therapeutic area, which act as harder distractors. The conclusion is unchanged across two closed-source vendors and two open-weight 70B-class models: frontier LLMs can match an abstract to a SMILES, randomised SMILES, InChIKey, or Murcko scaffold; what they cannot do is conjure gold abstracts into a pool that BM25 never surfaced.

5 Discussion and Conclusion

The methodological takeaway is a decomposition. MolRepBench plus the oracle-pool control separates two confounded layers of molecular RAG: *candidate generation*, where BM25 and standard dense retrievers fail completely on non-name representations, and *reranker understanding*, where frontier LLMs are competent once they are shown the relevant documents. Without this decomposition, a reasoning-style failure on SMILES is indis-

tinguishable from a retrieval failure on SMILES. With it, the practical surface for making an essential improvement is well-defined: fix the upstream pool. Cheap query-side normalisation (BM25+QE, BGE+QE) is the most deployable instance of this fix and already closes a substantial fraction of the gap; an embedding analysis on six encoder spaces (App. E) shows why even biomedical dense retrievers do not solve it for free.

Conclusion. Replacing a molecule's name with a chemically equivalent string costs every text retriever and every frontier LLM reranker we tested more than 97% of its top-10 recall. An oracle-pool control shows that the same four LLMs – two closed-source and two open-weight 70B – reach 0.86–0.94 Recall@10 on chemical strings once gold is in the pool. Molecular RAG over PubMed therefore fails upstream, in candidate generation, not in the LLM's ability to read chemical strings – and the most deployable mitigation is cheap query-side normalisation, not a more capable reranker.

6 Limitations

Our corpus is small relative to full PubMed, and gold abstracts are built from name-and-synonym lexical matching, which favours NAME. That is the invariance test, not a bug: an invariant retriever should recover the same name-anchored evidence from any equivalent identifier. The query-expansion baseline resolves identifiers perfectly inside our cached PubChem snapshot, which overestimates open-world coverage. Scaffold queries remain mostly unresolved because scaffolds do not map one-to-one to CIDs.

References

Josh Achiam, Steven Adler, and 1 others. 2023. GPT-4 technical report. In *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 haiku. *Model card*.

Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. 2019. Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1):1–13.

Guy W Bemis and Mark A Murcko. 1996. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893.

Abhimanyu Dubey and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *NeurIPS Datasets and Benchmarks Track*.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. 2015. InChI, the IUPAC international chemical identifier. *Journal of Cheminformatics*, 7(1):1–34.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2025. Improving chemical understanding of LLMs via SMILES parsing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2023. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380.

Greg Landrum. 2013. RDKit: Open-source cheminformatics. http://www.rdkit.org.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Shaghayegh Sadeghi, Alan Bui, Ali Forooghi, Jianguo Lu, and Alioune Ngom. 2024. Can large language models understand molecules? *BMC Bioinformatics*, 25(1):225.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. *EMNLP*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

David Weininger. 1988. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. MolRAG: Unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2024. C-Pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*.

A Full Retrieval Matrix

Table 3 gives Recall@10 for every system across all five representations, grouped by retriever family. Within each group the chemical-rep collapse is uniform; only query-side normalisation and the oracle-pool reranker escape it.

Retriever	name	InChIKey	c.SMI	r.SMI	scaf
Sparse / lexical					
BM25	0.550	0.002	0.004	0.002	0.003
Text-first dense retrievers					
BGE-small	0.600	0.002	0.003	0.004	0.004
PubMedBERT	0.237	0.004	0.002	0.005	0.003
E5-small	0.621	0.003	0.002	0.001	0.001
Biomedical dense & cross-encoder					
MedCPT	0.507	0.002	0.003	0.004	0.002
LLM listwise rerank over BM25 top-50					
GPT-4o-mini	0.792	0.014	0.009	0.011	0.013
Claude-Haiku	0.785	0.013	0.007	0.010	0.013
Llama-3.3-70B	0.779	0.014	0.007	0.010	0.013
Qwen-2.5-72B	0.774	0.014	0.009	0.011	0.013
LLM listwise rerank over oracle pool					
GPT-4o-mini-oracle	0.872	0.876	0.914	0.910	0.886
Claude-Haiku-oracle	0.859	0.931	0.935	0.932	0.935
Llama-3.3-70B-oracle	0.820	0.866	0.868	0.863	0.873
Qwen-2.5-72B-oracle	0.855	0.877	0.885	0.889	0.881
Query-side normalisation					
BM25+QE-real	0.738	0.451	0.082	0.139	0.005
BM25+QE-oracle	0.738	0.451	0.082	0.138	0.376
BGE+QE-real	0.650	0.483	0.436	0.441	0.007
BGE+QE-oracle	0.650	0.483	0.436	0.442	0.488
Document-side normalisation					
Hybrid	**0.822**	**0.671**	0.153	0.274	0.009

Table 3: Full retrieval matrix for all completed systems, Recall@10 across five representations. Suffixes `-real` / `-oracle` on QE rows denote query-resolution provenance (realistic API lookup vs. cached canonicalisation).

B Hybrid Scoring Detail

A query-side canonicaliser f_q maps names, SMILES, and InChIKeys to a PubChem CID via a synonym table, RDKit canonicalisation, and exact InChIKey lookup. A lightweight document index $D(c)$ maps each CID to the PMIDs of abstracts that mention that molecule under a known name, synonym, or InChIKey. The final ranking score is $\alpha\,\mathrm{BM25}(q,d) + \beta\,\mathbf{1}[d \in D(f_q(q))]$ with α=1 and β=25. The contribution is the invariance layer, not a new dense retriever.

C Candidate-Depth Ablation

Retriever	Rep	R@10	R@50	R@100	R@200
BM25	name	0.550	0.838	0.879	0.879
BM25	c.SMI	0.004	0.009	0.017	0.017
BM25	r.SMI	0.002	0.011	0.021	0.021
BM25	InChIKey	0.002	0.014	0.027	0.027
MedCPT	name	0.507	0.795	0.867	0.867
MedCPT	c.SMI	0.003	0.010	0.020	0.020
MedCPT	r.SMI	0.004	0.009	0.015	0.015
MedCPT	InChIKey	0.002	0.011	0.021	0.021

Table 4: Candidate-depth ablation for BM25 and MedCPT. Increasing retrieval depth helps NAME substantially but leaves chemical-string recall near zero, confirming that the LLM-reranker failure on chemical strings is not due to a slightly-too-shallow pool.

D Top-10 Jaccard Across Representations

E Embedding-Space Probe

The retrieval result has a corresponding signature in encoder geometry (Table 6, Figure 2). On six encoders – MedCPT, BGE-small, E5-small, PubMedBERT-MS-MARCO, ChemBERTa-zinc, and ChemBERTa-77M-MTR – the mean cosine similarity between two representations of the same drug (Δ_{within}) is lower than the mean cosine similarity between two *different* drugs written in the same representation ($\Delta_{\text{between}}(r)$) for every representation r we test. For MedCPT, Δ_{within}=0.544 versus Δ_{between}(c.SMI)=0.856. For BGE-small, 0.673 versus 0.921. The ordering survives a token-

Figure 2: t-SNE of the 1,555 bare molecular identifiers, coloured by representation type. Across biomedical query encoders, general-purpose text retrievers, and chemistry-pretrained encoders, the dominant structure is representation type rather than molecular identity.

Retriever	name vs c.SMI	name vs InChIK
BM25	0.011	0.291
BGE-small	0.022	0.075
BM25+QE	0.083	0.537
BGE+QE	0.369	0.486
Hybrid	0.133	0.685

Table 5: Top-10 Jaccard overlap for the same drug expressed in two representations. Without normalisation the retrievers return mostly different evidence sets for NAME vs. canonical SMILES; query-side normalisation closes the gap. Canonical and randomized SMILES tend to agree more with each other than with names, i.e. the retrievers are consistent on the *wrong subspace*: similar chemistry-looking documents for chemistry-looking strings, but not the documents tied to the queried molecular identity.

length-matched control. The t-SNE in Figure 2 shows the qualitative version of the same effect: the dominant clusters are string types, not drugs. This is consistent with why BGE-like dense retrieval fails so catastrophically on chemical strings (§4), but is not the core diagnostic of this paper – the oracle-pool decomposition pins down where the failure lives within an actual retrieval pipeline.

		$\Delta_{\text{between}}(r)$, different drugs, same rep				length-matched	
Encoder	Δ_{within}	name	c.SMI	r.SMI	InChIKey	Δ_{within}	c.SMI
MedCPT	0.544	0.447	0.856	0.829	0.763	0.660	0.862
BGE-small	0.673	0.614	0.921	0.908	0.767	0.783	0.929
E5-small	0.856	0.814	0.960	0.960	0.906	0.903	0.965
PubMedBERT-MS-MARCO	0.879	0.857	0.972	0.966	0.892	0.910	0.975
ChemBERTa-zinc	0.530	0.896	0.679	0.587	0.895	0.565	0.718
ChemBERTa-77M-MTR	0.275	0.882	0.324	0.120	0.751	0.527	0.381

Table 6: Embedding-space invariance probe. Δ_{within}: mean cosine similarity between different representations of the same drug. $\Delta_{\text{between}}(r)$: mean cosine similarity between different drugs written in the same representation r. Right block repeats the within/between comparison on a token-length-matched subset. An invariant encoder should satisfy $\Delta_{\text{within}} > \Delta_{\text{between}}(r)$ for every r – no encoder we test does.