Unified Representation of Genomic and Biomedical Concepts through Multi-Task, Multi-Source Contrastive Learning

Hongyi Yuan [*1,2] **Suqi Liu** [*1,3] **Zongxin Yang** [*1] **Kelly Cho** [3,4,5] **Katherine Liao** [3,4,5] **Alexandre Pereira** [3,4,5] **Tianxi Cai** [1,3]

Abstract

Phenotype vocabularies and genomic studies use incompatible coding systems for biomedical concepts, hindering large biobanks from realizing their full potential for precision medicine. Existing biomedical language models (LMs) bypass code heterogeneity but cannot embed single-nucleotide polymorphisms (SNPs), while graph-based methods require brittle manual cross-walks. We introduce **GENEREL** (**GEN**omic **E**ncoding **RE**presentation with **L**anguage model), the first *ontology-agnostic, genetic-contextualized* framework that unifies diseases, drugs, pathways, genes, and 65,000 common SNPs in a single vector space. GENEREL encodes free-text concepts with a Transformer, embeds SNPs via a lightweight multilayer perceptron (MLP) with trainable embeddings, and aligns both domains through multi-task, weighted contrastive learning over UMLS synonyms, PrimeKG relations, GWAS/eQTL variant–trait links, and UK Biobank correlations. On four external benchmarks—DisGeNET, DrugBank, Million Veteran Program (MVP) and a held-out GWAS Catalog split—GENEREL surpasses specialized LMs and graph baselines, while its cosine similarity reliably tracks odds-ratio effect sizes. The resulting representation paves the way for cross-biobank retrieval, variant prioritization, and downstream integrative analyses.

[*]Equal contribution [1]Department of Biomedical Informatics, Harvard Medical School, Boston, USA [2]Department of Statistics and Data Science, Tsinghua University, Beijing, China [3]VA Boston Healthcare System, U.S. Department of Veteran Affairs, USA [4]Department of Medicine, Brigham and Women's Hospital, Boston, USA [5]Department of Medicine, Harvard Medical School, Boston, USA. Correspondence to: Tianxi Cai <tcai@hsph.harvard.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

1. Introduction

Large-scale biobanks (*e.g.*, UK Biobank [8], MVP [25], All of Us [1]) pair genome-wide variant information with rich electronic health records (EHRs), while traditional GWAS cohort studies, cataloged in resources like the GWAS Catalog [20], offer curated genetic associations across diverse traits. Together, these resources hold immense promise for advancing precision medicine. However, integration across them remains limited due to incompatible phenotype representations: EHR-linked studies use PheCodes, ICD, or SNOMED CT [22], while GWAS studies often rely on the Experimental Factor Ontology (EFO). These discrepancies in coding and variable definitions hinder cross-study harmonization, and existing manual mappings are brittle and difficult to scale.

Biomedical LMs (BioBERT [16], PubMedBERT [12], SapBERT [19]) offer a promising solution to interoperability challenges by learning from unstructured text and bypassing rigid code systems. However, these models lack detailed biological understanding, particularly of genetic variation. SNPs are essentially treated as out-of-vocabulary tokens, limiting the models' ability to reason over SNPs and clinical concepts jointly. As a result, semantically distinct diseases with different biological underpinnings—such as type 1 and type 2 diabetes—can receive nearly identical embeddings (e.g., cosine similarity > 0.99 with PubMedBERT), obscuring important mechanistic differences and reducing their utility for genetics-informed applications.

This paper proposes **GENEREL**, **GEN**omic **E**ncoding **RE**presentation with **L**anguage model, a unified embedding framework that bridges this divide. GENEREL (i) encodes any free-text biomedical concept using an LM, (ii) embeds 65 k common SNPs via a lightweight MLP, and (iii) aligns the two spaces through multi-task, weighted contrastive learning over four complementary resources: UMLS synonyms [5], PrimeKG biomedical relations [9], variant–trait links from the GWAS Catalog [7] and GTEx eQTL studies [21], and patient-level phenotype-genotype correlations from UK Biobank.

Our contributions are summarized as follows.

- *Ontology-agnostic harmonization* — by operating directly on natural language, GENEREL bypasses coding incompatibilities and integrates multi-source knowledge without manual mapping.
- *Genomic–clinical coupling* — joint training yields a single embedding space covering diseases, drugs, genes, pathways, and SNPs, enabling cross-domain reasoning.
- *State-of-the-art performance* — on external benchmarks from MVP, DisGeNET [23], DrugBank [15], and held-out GWAS studies, GENEREL outperforms specialized baselines and ranks SNP relevance in a manner consistent with reported odds ratio magnitudes.

2. Related Work

Graph-based concept embedding. Biomedical knowledge graphs represent biomedical entities, such as diseases, drugs, and genes, as nodes connected by curated or predicted relationships. Early work in link prediction applied matrix factorization to co-occurrence or adjacency matrices [3; 13; 11]; followed by more expressive techniques such as translational models (e.g. TransE, TransH, TransR) [6; 18], bilinear scoring (DistMult, SimplE) [27; 14], and graph neural networks (GNNs) [17]. While effective within a single coding system, these methods rely on explicit node identifiers, making cross-system interoperability dependent on fragile manually curated mappings — and rarely extend to SNP-level genetic variants. Efforts to embed SNP–phenotype via matrix factorization of genotype–phenotype correlation matrices [30] offer an alternative, but such approaches cannot incorporate textual evidence or generalize to unseen concepts, limiting their utility in complex, multi-modal biomedical settings.

Biomedical LMs. Domain-adapted Transformers such as BioBERT [16], ClinicalBERT [2], PubMedBERT [12], and SapBERT [19] learn rich semantics directly from text, alleviating code heterogeneity. Later models inject graph structure (CODER [28], KRISSBERT [29]) or improve retrieval (BGE [10]), yet all treat SNPs as out-of-vocabulary tokens and thus cannot reason jointly over genetic and clinical spaces. Our work departs from prior art by *simultaneously* embedding free-text biomedical concepts and 65 k common SNPs, and by aligning them through multi-task weighted contrastive learning over UMLS, PrimeKG, GWAS/eQTL, and UK Biobank, yielding the first ontology-agnostic, genetic-contextualized representation that transfers across databases.

3. GENEREL

Overview. A biomedical term (free text) or a SNP is fed into a specific encoder separately; multi-task, weighted contrastive learning then aligns all entities in one vector space.

Table 1. Training pairs after filtering.

Task	Source	#Pairs
Synonym	UMLS	246 k
Concept–Concept	PrimeKG	325 k
Concept–SNP	GWAS + eQTL	136 k
Concept–SNP	UK Biobank	467 k
Total		1.17 M

Unified Embedding Space. Given a biomedical concept c, we take the [CLS] hidden state from a pretrained LM $\mathcal{M}_\phi$ and project it to d dimensions:

$$c^e = W\,\mathcal{M}_\phi(c) + b.$$

For each SNP g (e.g. `rs2476601`) we look up a learnable vector from an embedding matrix $\psi \in \mathbb{R}^{M\times d}$ and refine it with a two-layer MLP: $g^e = \mathrm{MLP}(\psi_g)$. Both c^e and g^e inhabit the same $d = 768$ space, enabling direct comparison.

Multi-task Weighted Contrastive Objective. Across heterogeneous sources (§3) we collect positive pairs $\mathcal{S} = \{(h,t)\}$ and minimize a weighted InfoNCE loss:

$$\mathcal{L} = -\sum_{(h,t)\in\mathcal{S}} w_{h,t}\,\log\frac{\exp(\langle h^e, t^e\rangle/\tau)}{\sum_{\tilde{h}\in\mathcal{C}}\exp(\langle \tilde{h}^e, t^e\rangle/\tau)},$$

where τ is learnable as in CLIP [24]. Weights $w_{h,t} \in (0,2]$ reflect association strength—odds ratios or regression β—and default to 1 when unavailable.

We instantiate three tasks that share parameters and mini-batches: **i) Synonym identification** from UMLS [5]; **ii) Concept–concept relations** from PrimeKG [9]; **iii) Concept–SNP links** from the GWAS Catalog, GTEx eQTL [7; 21], and UK Biobank correlations.

Training Resources and Setup. Table 1 summarizes the four data sources. For GWAS/eQTL, we normalize study-specific statistics before clipping to $(0,2]$; UK Biobank pairs are selected by absolute correlation thresholding. GENEREL is trained for 25 epochs on a single L40S GPU using AdamW (learning rate: 2×10^{-5} for $\mathcal{M}_\phi$, 2×10^{-3} for ψ; batch size: 512).

4. Experiments

4.1. Benchmarks and Baselines

We assess two tasks: **(i) concept–concept relatedness** on DisGeNET and DrugBank; **(ii) concept–SNP association** on the GWAS test split and the MVP. Baselines include domain LMs (BioBERT, ClinicalBERT, PubMedBERT, SapBERT, CODER, KRISSBERT, BGE) [16; 2; 12; 19; 28; 29; 10], graph models (TransE/H/R, DistMult, SimplE) [6; 26; 18; 27; 14], and an SVD factorization of the UKB correlation matrix [30]. All embeddings use $d=768$.

Table 2. AUCs for detecting the related biomedical concept pairs against randomly sampled negative pairs. The associations include disease-gene and pathway-gene pairs from DisGeNET and Indication-Drug and Indication-Gene pairs from DrugBank. The results are reported based on 5 independent runs.

Model	DisGeNET		DrugBank	
	Disease–Gene	Pathway–Gene	Indication–Drug	Indication–Gene
BioBERT	0.519 ± 0.013	0.568 ± 0.008	0.714 ± 0.010	0.579 ± 0.009
ClinicalBERT	0.483 ± 0.033	0.528 ± 0.011	0.636 ± 0.010	0.549 ± 0.009
PubmedBERT	0.528 ± 0.023	0.555 ± 0.011	0.711 ± 0.011	0.578 ± 0.011
SapBERT	0.627 ± 0.019	0.585 ± 0.011	0.667 ± 0.008	0.656 ± 0.006
CODER	0.564 ± 0.015	0.594 ± 0.013	0.811 ± 0.006	0.657 ± 0.006
KRISSBERT	0.623 ± 0.009	0.621 ± 0.010	0.753 ± 0.005	0.745 ± 0.012
BGE	0.640 ± 0.023	0.577 ± 0.014	0.763 ± 0.005	0.537 ± 0.015
GENEREL	$\mathbf{0.770} \pm 0.016$	$\mathbf{0.758} \pm 0.009$	$\mathbf{0.824} \pm 0.009$	$\mathbf{0.850} \pm 0.005$
#Pairs	1,366	778	4,207	6,148

Table 3. AUCs for detecting the related biomedical concepts and SNPs pairs on MVP and the GWAS test split. $-$Trait and $-$SNP indicates the anchors when randomly sampling negatives. Results are reported based on 5 independent runs.

	MVP-Trait	MVP-SNP	GWAS-Trait	GWAS-SNP
Cor.Mat.SVD	0.775 ± 0.009	$\mathbf{0.840} \pm 0.004$	-	-
TransE	0.543 ± 0.015	0.524 ± 0.008	0.693 ± 0.007	0.621 ± 0.003
TransH	0.531 ± 0.015	0.516 ± 0.004	0.655 ± 0.009	0.601 ± 0.003
TransR	0.578 ± 0.014	0.528 ± 0.014	0.767 ± 0.008	0.737 ± 0.008
DistMult	0.622 ± 0.009	0.761 ± 0.001	0.825 ± 0.008	0.893 ± 0.002
SimplE	0.636 ± 0.006	0.759 ± 0.004	0.813 ± 0.004	0.894 ± 0.001
GENEREL	$\mathbf{0.821} \pm 0.012$	0.810 ± 0.002	$\mathbf{0.942} \pm 0.003$	$\mathbf{0.939} \pm 0.002$

4.2. Results

Concept–Concept Relatedness. We use two held-out sources—DisGeNET (Disease–Gene, Pathway–Gene) and DrugBank (Indication–Drug, Indication–Gene); none of these pairs appears in PrimeKG after exact-match filtering. We compare GENEREL with domain LMs (BioBERT, ClinicalBERT, PubMedBERT, SapBERT, CODER, KRISSBERT [29]) plus the strong general model BGE [10]. AUC is computed against random negatives using cosine similarity. Table 2 shows GENEREL tops all four benchmarks, confirming it captures true biological relatedness beyond surface wording.

Concept–SNP Association. Table 3 shows that GENEREL outperforms every baseline on the GWAS split by a wide margin. On MVP it beats all graph learners and edges out SVD by 0.046 when traits are anchored, though it trails SVD by 0.030 when SNPs are anchored.

A key weakness of the graph learning and SVD baselines is their reliance on hard-coded IDs; discrepancies between PheCode, EFO and other vocabularies—and even simple synonymy (e.g. *reactive arthritis* vs. *Reiter's syndrome*)—prevent seamless integration across datasets. By encoding free text directly, GENEREL sidesteps these coding barriers and can harmonize information that would otherwise remain siloed.

4.3. Ablation Study

A key feature of the GENEREL framework is its incorporation of multi-task and multi-source training. To demonstrate the function of each training task, we conduct ablation experiments on different combinations of the training datasets. Besides the previous benchmarks, we also include COMETA [4], a dataset curated from public anonymous health discussions on Reddit, to evaluate the model performance on disambiguating synonyms in biomedical concepts. COMETA contains 20 k English biomedical mentions in various forms of daily languages. We pool the samples in the "general" and "specified" splits. We report the AUCs for synonym pairs and randomly sampled negative pairs to maintain consistency with other benchmarks. The results are listed in Table 4.

Table 4. Ablation study (mean AUC). **U** = UMLS, **P** = PrimeKG, **G** = GWAS/eQTL. Columns: **DisG** = DisGeNET (Disease–Gene), **DrugB** = DrugBank (Indication–Drug), **COM** = COMETA.

Model	DisG	DrugB	MVP	GWAS	COM	Avg
Full (U+P+G)	0.764	0.837	0.816	0.941	0.977	**0.868**
–U	0.771	0.838	0.807	0.940	0.932	0.858
–U –P	0.683	0.737	0.815	0.950	0.944	0.826
–U –P –G	0.670	0.690	0.620	0.549	0.922	0.690

Without the UMLS training task, we observe a decline in performance on the COMETA benchmark, as the model's ability to disambiguate synonyms decreases due to the lack of synonym information in the other datasets. When fur-

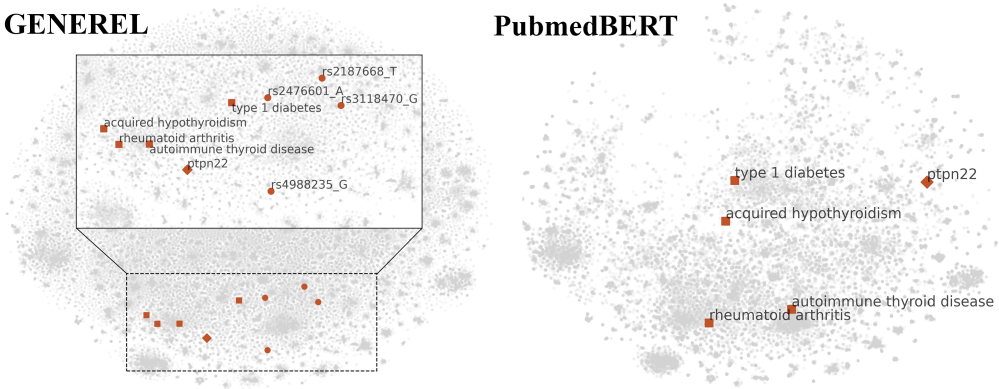

Figure 1. Embedding visualizations of GENEREL and PubMedBERT using t-SNE are shown. We highlight the autoimmune diseases, the related genes, and, additionally for GENEREL, the related SNPs.

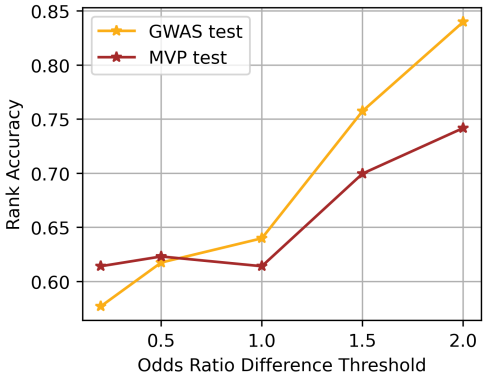

Figure 2. Evaluation of detecting the relative degree of relatedness. The plot depicts the accuracies of different sub-sample groups with various odds ratio differences.

ther excluding PrimeKG from training, the performance on DisGeNET and DrugBank drops by 0.081 and 0.100 respectively. The GWAS catalog and UK Biobank primarily focus on gene and trait concepts, lacking broader biomedical concepts such as pathways and drugs. PrimeKG enhances the model's learning by integrating this additional information. When trained only on UK Biobank, the model performs worse uniformly across the benchmarks, since GWAS covers a broader range of biomedical concepts and SNPs compared to UK Biobank. Overall, the ablation demonstrates the necessity and functionality of each training task,showing the benefits of the multi-task, multi-source training scheme.

5. Discussion

Interpretable distances. By weighting the InfoNCE loss with odds ratios, GENEREL aligns cosine similarity with biological effect sizes, allowing embedding distances to reflect the strength of genetic associations. Figure 2 shows that when the odds–ratio gap exceeds 2, the stronger SNP is ranked first in 84.7% of held-out GWAS pairs; Table 5 lists concrete high- vs. low-risk examples. *Such calibrated distances enable variant prioritization and causal follow-up studies.*

Table 5. AUCs of GENEREL detecting the concept-SNP relatedness against random negatives on the original concept phrases and the substituted synonyms on MVP and GWAS.

	MVP	GWAS
Original	$0.798_{\pm 0.008}$	$0.901_{\pm 0.004}$
Synonyms	$0.786_{\pm 0.005}$	$0.836_{\pm 0.005}$

Robustness to lexical variation. Substituting traits with UMLS synonyms only reduces GWAS AUC from 0.901 to 0.836 (Table 5), far smaller than graph-based baselines—evidence that text-level encoding mitigates coding noise common in EHRs.

Qualitative structure. The t-SNE plot in Figure 1 clusters autoimmune diseases, their genes, and associated SNPs tightly in GENEREL, whereas PubMedBERT scatters them. *This illustrates the benefit of coupling genomic and clinical signals within one embedding space.*

Limitations and future work. Current vocabularies cover 65 k common SNPs but omit rare or structural variants; training depends on summary statistics rather than raw genotypes. Future directions include (i) genetic-contextualized tokenization for rare variants, (ii) multi-modal contrastive heads for imaging or time-series data, and (iii) instruction fine-tuning for zero-shot genomic QA.

6. Conclusion

GENEREL presents the first *ontology-agnostic, genetic-contextualized* embedding that jointly positions diseases, drugs, pathways, genes, and SNPs in a unified vector space—enabling the construction of a biomedical knowledge graph that transcends the limitations of incompatible coding systems. Across four external benchmarks, it outperforms specialized LMs and graph baselines, while its similarity scores faithfully reflect biological effect size. We believe this unified representation will serve as a foundation for cross-biobank discovery and precision-medicine applications.

References

[1] All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[2] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. Publicly available clinical BERT embeddings. In Rumshisky, A., Roberts, K., Bethard, S., and Naumann, T. (eds.), *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[3] Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[4] Basaldella, M., Liu, F., Shareghi, E., and Collier, N. COMETA: A corpus for medical entity linking in the social media. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3122–3137, Online, November 2020. Association for Computational Linguistics.

[5] Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70, 2004.

[6] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.

[7] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[8] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N. E., Donnelly, P., and Marchini, J. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203 – 209, 2018.

[9] Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10, 2022.

[10] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2023.

[11] Gan, Z., Zhou, D., Rush, E., Panickan, V. A., Ho, Y.-L., Ostrouchov, G., Xu, Z., Shen, S., Xiong, X., Greco, K. F., et al. Arch: Large-scale knowledge graph via aggregated narrative codified health records analysis. *medRxiv*, 2023.

[12] Gu, Y., Tinn, R., Cheng, H., Lucas, M. R., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23, 2020.

[13] Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V. M., Schubert, P., Panickan, V. A., Cai, T., et al. Clinical knowledge extraction via sparse embedding regression (keser) with multi-center large scale electronic health record data. *NPJ digital medicine*, 4(1):151, 2021.

[14] Kazemi, S. M. and Poole, D. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

[15] Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E., Strawbridge, S. A., et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52 (D1):D1265–D1275, 2024.

[16] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2019.

[17] Li, M. M., Huang, K., and Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering*, 6(12):1353–1369, 2022.

[18] Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[19] Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. Self-alignment pretraining for biomedical entity representations. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pp. 4228–4238, Online, June 2021. Association for Computational Linguistics.

[20] Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., and Parkinson, H. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26 (8):1112–1118, 03 2010. ISSN 1367-4803.

[21] Nica, A. C. and Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368, 2013.

[22] Organization, W. H. et al. Icd-10: international statistical classification of diseases and related health problems: tenth revision. *World Health Organization*, 2004.

[23] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, pp. gkw943, 2016.

[24] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

[25] Verma, A., Huffman, J. E., Rodriguez, A., Conery, M., Liu, M., Ho, Y.-L., Kim, Y., Heise, D. A., Guare, L., Panickan, V. A., et al. Diversity and scale: Genetic architecture of 2068 traits in the va million veteran program. *Science*, 385(6706):eadj1182, 2024.

[26] Wang, Z., Zhang, J., Feng, J., and Chen, Z. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

[27] Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[28] Yuan, Z., Zhao, Z., Sun, H., Li, J., Wang, F., and Yu, S. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of Biomedical Informatics*, 126:103983, 2022. ISSN 1532-0464.

[29] Zhang, S., Cheng, H., Vashishth, S., Wong, C., Xiao, J., Liu, X., Naumann, T., Gao, J., and Poon, H. Knowledge-rich self-supervision for biomedical entity linking. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 868–880, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[30] Zhao, Y., Cai, H., Zhang, Z., Tang, J., and Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nature communications*, 12(1):5261, 2021.