

Contrastive Learning for Bioactive Peptide Representation and Personalized Health Recommendation

Eetu Pekkarinen

Department of Computer Science, George Mason University, Fairfax, VA, USA.
pekkarineneetu@gmu.edu

Nela Pietraszkiewicz

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
nela516@binghamton.edu

Euan Toal

Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA.
euantoal00@unr.edu

Abstract

Bioactive peptides occupy a distinctive position in nutrition, pharmacology, and personalized health because their effects depend on sequence, structure, bioavailability, host physiology, and context of consumption. Computational systems that recommend peptides or peptide-rich interventions must therefore represent molecules, individuals, and environments in ways that support robust generalization rather than narrow pattern matching. This paper examines contrastive learning as a system-level strategy for bioactive peptide representation and personalized health recommendation. Rather than treating contrastive learning merely as an algorithmic technique, the discussion frames it as an infrastructural choice that shapes data collection, model governance, deployment, fairness, and clinical translation. The paper first reviews related work in self-supervised representation learning, protein language modeling, and health recommender systems. It then proposes a layered architecture that combines molecular encoders, individual context encoders, multimodal alignment, and human-in-the-loop recommendation services. Subsequent sections analyze governance challenges, including consent, provenance, privacy, federated learning, bias, and transparency. The paper also considers robustness under distribution shift, sustainability of large-scale training, edge deployment, and policy implications for regulated health products. The central argument is that contrastive learning can improve peptide representation by aligning heterogeneous views of bioactivity and individual response, but its value depends on governance mechanisms that make data, models, and recommendations accountable. A systems perspective is necessary because peptide recommendation is not only a prediction problem but also a sociotechnical infrastructure problem involving clinical validity, user trust, and equitable access.

Keywords

contrastive learning, bioactive peptides, personalized health, recommender systems, protein representation, data governance, fairness, clinical translation, federated learning, sustainability.

1. Introduction

Personalized health recommendation has moved from simple rule-based dietary advice toward computational systems that integrate biomarkers, dietary histories, wearable signals, genomic information, and clinical constraints. Within this movement, bioactive peptides are an especially difficult class of intervention because their effects are context-dependent, dose-sensitive, and mediated by digestion, absorption, microbiota, and host immune or metabolic states. A peptide that supports antioxidant activity in one physiological context may be neutral or even problematic in another. Consequently, recommendation systems for bioactive peptides require representations that capture molecular properties and individual susceptibility without reducing health to a single score or a static preference vector. Contrastive learning offers a promising way to learn such representations by pulling related molecular and physiological views together while separating unrelated ones. Its appeal is not only predictive performance but also flexibility: it can operate with weak supervision, align heterogeneous modalities, and support downstream tasks such as peptide function prediction, safety screening, and personalized ranking [1], [2], [3].

However, the translation of contrastive learning into health recommendation is not a straightforward matter of model selection. It raises architectural questions about how molecular views, individual views, and environmental views should be connected. It raises governance questions about who controls sensitive data and how consent is maintained across federated or distributed infrastructures. It raises fairness questions about populations that are underrepresented in peptide research, food composition databases, and clinical trials. It also raises sustainability questions about the energy and maintenance costs of large encoders that must remain current as new peptides, biomarkers, and clinical findings emerge. These issues are particularly salient for bioactive peptides because the evidence base is fragmented across food science, pharmacology, nutrigenomics, and clinical nutrition [10]. A system that recommends peptides must therefore be evaluated not only by accuracy but by traceability, contestability, safety, and long-term maintainability.

This paper argues that contrastive learning for bioactive peptides should be understood as a sociotechnical infrastructure rather than as a standalone machine learning method. The proposed perspective is system-level. It considers how data are collected and labeled, how encoders are pretrained and updated, how recommendations are served to clinicians and consumers, and how governance mechanisms shape the legitimacy of the entire pipeline. The discussion draws on contrastive representation learning, protein language models, health artificial intelligence, privacy engineering, and responsible machine learning. The goal is to outline a research and deployment agenda that recognizes both the technical promise and the institutional risks of personalized peptide recommendation.

2. Background and Related Work

Contrastive learning has become a central approach in self-supervised representation learning because it learns useful embeddings by comparing positive pairs against negative pairs. Early formulations such as contrastive predictive coding connected representation learning to predictive coding and established that useful features can emerge without explicit labels [3]. Later work in computer vision demonstrated that simple contrastive frameworks with strong augmentations can rival supervised pretraining [1], while momentum-based methods improved the stability of large negative queues and enabled effective unsupervised visual representations [2]. Multimodal contrastive learning extended the paradigm by aligning images and text in a shared space, showing that natural language supervision can produce transferable representations [4]. These developments are relevant to peptide health

recommendation because the same logic can align peptide sequences, structural descriptors, bioactivity assays, food sources, and individual response phenotypes.

In the biological domain, representation learning has advanced rapidly through protein language models and structure prediction systems. Transformer architectures provide the general sequence modeling backbone for many of these systems [9], and bidirectional pretraining has demonstrated the value of contextual representations for language and, by analogy, biological sequences [8]. Protein language models trained on large sequence databases have revealed that structural and functional properties can emerge from unsupervised learning at scale [6], while specialized transformer models have improved transfer across protein tasks [7]. AlphaFold demonstrated the power of deep learning for protein structure prediction, reinforcing the idea that molecular representation is a foundation for downstream biological inference [5]. For bioactive peptides, however, the challenge is not only structure prediction. Peptides are short, often flexible, and their bioactivity depends on enzymatic processing, transport, receptor interactions, and host factors. A representation that captures sequence syntax alone may miss the physiological context that determines whether a peptide is useful or risky.

Health recommendation research has explored machine learning for clinical decision support, patient risk stratification, and treatment personalization [12], [13]. Yet recommender systems in medicine face constraints that differ from consumer recommendation. They must handle sparse data, confounding by indication, safety constraints, and asymmetric costs of false positives and false negatives. Fairness research has shown that apparently neutral algorithms can encode structural inequities when trained on biased data or optimized for proxies that correlate with race, class, or access [15]. Methods for diagnosing and mitigating discriminatory classifiers have been proposed [14], and ethical frameworks for machine learning in health care emphasize transparency, oversight, and patient welfare [16]. Data stewardship principles such as FAIR data management provide guidance for making datasets findable, accessible, interoperable, and reusable [17], while privacy models such as k-anonymity and differential privacy offer different trade-offs between utility and disclosure risk [18], [19]. Model cards and datasheets have emerged as documentation practices for reporting model intent, performance, and limitations [20]. These strands form the governance backdrop for any contrastive peptide recommendation system.

The specific evidence base for bioactive peptides remains heterogeneous. A recent narrative review on animal-derived peptides and aging-related diseases illustrates the breadth of proposed benefits and the difficulty of comparing studies across sources, doses, and populations [10]. Such reviews are valuable but also reveal why automated recommendation is hazardous: the literature contains promising mechanisms, inconsistent endpoints, and varied preparation methods. Contrastive learning can help organize this heterogeneity by learning representations that separate stable bioactivity signatures from contextual confounders, but only if the training data and evaluation protocols are designed to preserve uncertainty. Without that design, a model may produce confident recommendations that overstate weak evidence. Thus, the related work points toward a hybrid approach: modern representation learning for molecular and individual modeling, combined with recommender system engineering and health governance.

3. Contrastive Representation Architecture for Bioactive Peptides

A system-level architecture for bioactive peptide representation should begin with a layered view of data and computation. The lowest layer contains raw and curated inputs: peptide

sequences, known modifications, food matrix information, assay results, receptor binding data, absorption and bioavailability estimates, microbiome profiles, biomarkers, dietary preferences, allergies, medication lists, and clinical constraints. These inputs are heterogeneous in scale, reliability, and privacy sensitivity. The second layer contains encoders that transform each modality into a latent representation. Peptide encoders may use transformer or convolutional sequence models pretrained on large protein databases [6], [7]. Individual context encoders may use tabular models, temporal models for wearable data, or graph models for comorbidity and medication interactions. The third layer aligns these representations through contrastive objectives, while the fourth layer supports downstream tasks such as bioactivity classification, safety ranking, and personalized recommendation. The final layer exposes services to clinicians, dietitians, consumers, or researchers through application programming interfaces and user interfaces.

The contrastive alignment strategy must reflect the causal structure of peptide health effects. Positive pairs should not simply be arbitrary augmentations of the same peptide. They should represent biologically meaningful equivalence, such as the same peptide measured across different assays, the same bioactive motif in different food sources, or the same individual under stable physiological conditions. Negative pairs should represent meaningful dissimilarity, such as peptides with known opposing effects, individuals with incompatible clinical constraints, or contexts in which bioavailability is drastically different. This design choice matters because contrastive learning amplifies whatever invariance the designer specifies. If the system treats all sequence variants as equivalent, it may ignore critical potency differences. If it treats all individual differences as noise, it may recommend interventions that are unsafe for subpopulations. Therefore, the architecture should include explicit metadata channels and constraint encoders that preserve safety-relevant distinctions rather than forcing them into a single smooth embedding space.

A second architectural issue is multimodal alignment. Peptides do not act in isolation. Their effects are shaped by food matrices, processing methods, co-consumed nutrients, gut microbial enzymes, and host genetics. A contrastive framework can align peptide representations with physiological response representations, but it must avoid spurious correlations. For example, a peptide may appear beneficial in a population that also consumes high fiber or has better healthcare access. If the model aligns peptide embeddings with improved biomarkers without accounting for these confounders, recommendations may be biased toward peptides that merely co-occur with privilege. Causal inference techniques, stratification, and sensitivity analysis should complement contrastive pretraining. The architecture should also distinguish association from intervention. A representation learned from observational data can support hypothesis generation, but clinical recommendation requires evidence from trials, mechanistic studies, or carefully designed causal analyses.

A third issue is the relationship between pretraining and task adaptation. Contrastive pretraining can produce general peptide and individual embeddings, but downstream recommendation may require task-specific fine-tuning with small labeled datasets. This creates a trade-off between reuse and specialization. A large general encoder can amortize computation across many tasks, but it may be difficult to audit and update. A smaller task-specific model may be more interpretable and easier to deploy at the edge, but it may not generalize across populations. A practical architecture might therefore use a shared pretrained backbone with modular adapters for bioactivity, safety, and preference modeling. This modularity supports governance because each adapter can be documented, validated, and

rolled back independently. It also supports sustainability because the system can update one component without retraining the entire stack. However, modularity introduces integration complexity and potential inconsistency between modules, so system designers must define clear interfaces, versioning policies, and validation gates.

4. Personalized Recommendation Infrastructure

Personalized recommendation for bioactive peptides requires more than ranking a list of molecules. It requires translating molecular evidence into actionable dietary or clinical options, respecting individual constraints, and communicating uncertainty. The infrastructure should include a user or patient profile service, a knowledge graph of peptides and food sources, a constraint engine for allergies and drug interactions, a recommendation engine that combines predicted bioactivity with contextual relevance, and a feedback loop that captures outcomes and adverse events. Each component has different latency, privacy, and reliability requirements. A clinician-facing interface may need detailed provenance and evidence summaries, while a consumer-facing application may need simple explanations and safer defaults. The same underlying model can support both if the serving layer enforces different access controls and explanation granularity.

The recommendation engine should treat personalization as constrained optimization rather than unconstrained preference matching. In consumer systems, engagement is often optimized because clicks and purchases are measurable. In health systems, the objective must include safety, efficacy, adherence, and equity. A peptide recommendation may be biologically plausible but impractical because of taste, cost, preparation burden, or cultural acceptability. Conversely, a highly acceptable food may contain insufficient bioactive peptide concentrations to produce the desired effect. The system should therefore model feasibility and adherence alongside bioactivity. Contrastive representations can help by aligning peptide properties with individual tolerability and lifestyle embeddings, but the recommendation layer must include hard constraints that cannot be overridden by a high similarity score. For example, a peptide derived from a common allergen should be excluded regardless of predicted benefit if the user has a severe allergy.

Feedback loops are essential but dangerous. If the system recommends a peptide and the user reports improved wellbeing, that signal may be influenced by placebo, concurrent lifestyle changes, or seasonal variation. If the system recommends a peptide and the user discontinues it, that signal may reflect cost rather than inefficacy. A robust infrastructure should distinguish explicit clinical outcomes from implicit behavioral signals and should weight them accordingly. It should also support active learning in which uncertain cases are escalated to human experts rather than automatically resolved. Federated learning can allow multiple institutions to improve the model without centralizing sensitive data, but it introduces challenges in heterogeneity, communication cost, and adversarial participation [21]. Differential privacy can reduce disclosure risk, but it may disproportionately affect rare conditions and small subpopulations [19]. The infrastructure must therefore make trade-offs explicit and configurable by governance bodies.

Deployment models will vary by setting. A hospital system may run the recommendation service on-premises or in a regulated cloud with strict audit logging. A consumer wellness application may run lighter models on a smartphone, syncing only aggregated updates. A research consortium may use a centralized data enclave with controlled access. These models have different implications for latency, cost, and equity. On-device inference can protect privacy and reduce connectivity requirements, but it may limit model size and update

frequency. Cloud inference can support large multimodal encoders, but it raises concerns about data transfer and vendor lock-in. A hybrid architecture may be most practical, with sensitive personalization occurring locally and global knowledge updated through privacy-preserving aggregation. The key is that deployment choices are not merely technical. They determine who can use the system, under what conditions, and with what recourse.

5. Data Governance, Fairness, and Accountability

Data governance is the foundation of trustworthy peptide recommendation. The system depends on data that are often sensitive, proprietary, or fragmented across institutions. Consent must be meaningful and dynamic, allowing individuals to understand how their biomarkers, dietary records, and genetic data will be used. Provenance must be recorded for every peptide, assay, and outcome so that recommendations can be traced back to evidence. FAIR principles provide a useful starting point for data management [17], but they do not by themselves resolve questions of power, access, and benefit sharing. A peptide database built from Indigenous knowledge or community food practices requires particular care to avoid extraction without reciprocity. Governance should include community representation, data use agreements, and mechanisms for benefit sharing.

Fairness in peptide recommendation has multiple dimensions. Representational fairness concerns whose data are included in training and whose health outcomes are predicted accurately. Allocation fairness concerns whether beneficial recommendations reach populations that need them. Procedural fairness concerns whether affected communities can contest model decisions. Bias can enter through food composition databases that underrepresent regional diets, clinical trials that exclude older adults or people with disabilities, and wearable datasets that skew toward affluent users. Methods for detecting discriminatory classifiers are necessary but insufficient [14]. The system should monitor performance across subgroups, including intersectional groups, and should evaluate whether recommendations reinforce existing health disparities. For example, if a model recommends expensive supplements because they correlate with better outcomes in a privileged cohort, it may widen inequality even if its predictions are statistically accurate [15].

Accountability requires documentation, auditability, and redress. Model cards can describe intended use, performance metrics, and limitations [20]. Datasheets can document dataset composition, collection methods, and known biases [22]. Audit logs should record which model version produced a recommendation, what data were used, and what constraints were applied. When harm occurs, there must be a path for reporting, investigation, and correction. In clinical settings, this path may involve institutional review boards, safety committees, and regulatory reporting. In consumer settings, it may involve user support, transparency reports, and independent audits. The governance architecture should also address secondary use. Data collected for one purpose may be valuable for another, but reuse should require renewed consent or a legitimate governance decision. Without such controls, peptide recommendation systems can become surveillance infrastructures that monetize health data under the guise of personalization.

Privacy engineering must be integrated rather than added after deployment. k -anonymity can protect against re-identification in released datasets, but it is vulnerable to linkage attacks and does not provide a formal guarantee [18]. Differential privacy offers a stronger mathematical framework, but it requires careful parameter selection and can reduce utility for small groups [19]. Federated learning can keep raw data local, but it does not eliminate privacy risk because model updates can leak information [21]. Secure aggregation, homomorphic

encryption, and trusted execution environments can strengthen protections, but they add complexity and cost. The appropriate mix depends on the sensitivity of the data, the threat model, and the regulatory environment. A peptide recommendation system that handles genetic and microbiome data should adopt a defense-in-depth approach, with clear data minimization, purpose limitation, and retention policies.

6. Robustness, Sustainability, and Deployment

Robustness is a central concern because peptide recommendation operates in a changing biological and social environment. New peptides are discovered, food processing methods change, clinical guidelines are updated, and individual physiology shifts over time. A model trained on historical data may fail when confronted with distribution shift. Contrastive representations can improve robustness by learning invariances to irrelevant variation, but they can also create brittle invariances if the augmentation strategy is mismatched to biology. Evaluation should therefore include stress tests for new food sources, underrepresented populations, noisy self-reported data, and adversarial manipulation. Adversarial actors might attempt to promote ineffective or harmful peptides by poisoning public databases or submitting misleading reviews. The system should include anomaly detection, source reliability scoring, and human review for high-stakes recommendations.

Sustainability has computational, environmental, and institutional dimensions. Large protein language models require substantial energy for pretraining and fine-tuning, and repeated retraining can be carbon-intensive [23]. The system should consider model reuse, distillation, pruning, and parameter-efficient adaptation to reduce energy consumption. Edge deployment can lower data transfer costs but may increase the number of devices that need updates. Institutional sustainability concerns maintenance: models degrade as data distributions shift, and documentation becomes outdated. A peptide recommendation service requires ongoing curation of peptide databases, monitoring of model performance, and periodic revalidation. If this maintenance is underfunded, the system may quietly become unsafe. Sustainability planning should therefore include total cost of ownership, staffing, and governance capacity, not only training compute.

Deployment in clinical and consumer contexts requires different validation thresholds. A consumer wellness application may be allowed to provide general educational information but should not make medical claims without evidence. A clinical decision support tool may require regulatory clearance, prospective trials, and integration into electronic health records. The distinction between wellness and medical claims is often blurred in practice, and regulators are increasingly attentive to software that recommends interventions. The system should include clear labeling of evidence strength, uncertainty, and contraindications. It should also support override by clinicians and users. A recommendation that cannot be questioned is not a recommendation but a command. Human-in-the-loop design is therefore both an ethical requirement and a robustness strategy, because human judgment can catch errors that automated metrics miss.

Monitoring should continue after deployment. The system should track recommendation uptake, adverse events, subgroup performance, and model drift. It should also track the provenance of new data and the reasons for model updates. When a model is updated, the change should be documented and, where necessary, revalidated. Rollback mechanisms should be available if a new version performs poorly or causes harm. These operational requirements are often neglected in research prototypes, but they determine whether a system can be trusted at scale. Contrastive learning does not remove the need for engineering

discipline. It increases the importance of data versioning, reproducible training, and continuous evaluation because the representation space can change subtly with new data.

7. Policy, Clinical Translation, and Future Directions

Policy frameworks for artificial intelligence in health are evolving, but they often lag behind technical capabilities. For peptide recommendation, policy must address data protection, medical device regulation, advertising claims, and liability. If a system recommends a peptide supplement and a user experiences harm, who is responsible? The developer, the clinician, the platform, or the data provider? Clear accountability requires contracts, audit trails, and insurance mechanisms. Regulatory agencies may classify some recommendation systems as medical devices, especially if they diagnose, treat, or prevent disease. This classification brings requirements for clinical evidence, quality management, and post-market surveillance. Developers should engage regulators early rather than assuming that a wellness label avoids scrutiny.

Clinical translation also requires evidence standards. Observational data and mechanistic studies can generate hypotheses, but randomized trials or rigorous quasi-experimental designs are needed to establish efficacy. Peptide interventions are difficult to study because blinding is hard, adherence varies, and effects may be modest. The system should support trial recruitment, endpoint tracking, and subgroup analysis. It should also be transparent about the difference between statistical association and causal effect. A contrastive representation may suggest that a peptide is promising, but that is not the same as proving that it improves health. Clinicians need summaries that distinguish these levels of evidence. Patients need explanations that are accurate without being overwhelming.

Future directions include causal contrastive learning, where positive and negative pairs are defined by interventions rather than mere co-occurrence. This would help align representations with treatment effects. Another direction is continual learning that updates encoders as new evidence emerges without catastrophic forgetting. A third is participatory design with patients, dietitians, and communities to ensure that recommendations reflect lived experience. A fourth is standard benchmarks for peptide bioactivity and personalized recommendation that include safety, fairness, and robustness metrics. Without shared benchmarks, progress will remain fragmented and difficult to compare. A fifth is regulatory science that develops methods for auditing adaptive systems. If a model changes over time, regulators need ways to evaluate the change without requiring full re-approval for every update.

International coordination will be important because peptide databases, food supplies, and health data cross borders. Different countries have different privacy laws, cultural food practices, and regulatory pathways. A system that works in one context may fail in another. Governance should therefore be federated in a political sense as well as a technical one. Local communities should be able to set constraints and priorities while contributing to shared knowledge. This is difficult but necessary. The alternative is a monolithic system that imposes one population's assumptions on others. Contrastive learning can support pluralism if it is designed to align multiple views without erasing difference. It can also undermine pluralism if it collapses all contexts into a single universal embedding. The choice is architectural and political.

8. Conclusion

Contrastive learning offers a powerful approach to bioactive peptide representation and personalized health recommendation because it can integrate heterogeneous data and learn relationships without requiring exhaustive labels. It can align peptide sequences with bioactivity, individual contexts with responses, and molecular evidence with dietary feasibility. However, the value of contrastive learning depends on how it is embedded in a broader system. The representation architecture must respect biological causality and safety constraints. The recommendation infrastructure must support human oversight, uncertainty communication, and equitable access. The governance framework must ensure privacy, accountability, and community participation. The deployment strategy must address robustness, sustainability, and regulatory compliance. These are not secondary concerns. They determine whether a technically sophisticated model becomes a trustworthy health service or a source of new harms.

This paper has argued for a systems perspective that treats peptide recommendation as a sociotechnical infrastructure. The central challenge is not merely to learn better embeddings but to build institutions and incentives that make those embeddings useful, contestable, and fair. Future research should develop causal contrastive methods, participatory governance models, and regulatory sandboxes for adaptive health recommendation. It should also invest in data curation for underrepresented populations and food systems. If these efforts succeed, contrastive learning can help personalize peptide-based health interventions in ways that respect biological complexity and social context. If they are neglected, the result may be another opaque algorithmic system that reinforces existing inequities. The choice is ours to design.

References

1. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119, 1597-1607.
2. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729-9738.
3. Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-8763.
5. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
6. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118.

7. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 7112-7127.
8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
10. Liu, T., Song, X., Dai, X., Ma, H., Muhammad, A., He, Z., & Zhao, K. (2025). Effects of animal-derived peptides on aging-related diseases: a narrative review. *Aging Advances*, 2(4), 160-165.
11. Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), 931-934.
12. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.
13. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358.
14. Chen, I. Y., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 31, 3539-3550.
15. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
16. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981-983.
17. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
18. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.
19. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 211-407.
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
21. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.

- B., ... Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), 1-210.
22. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.
23. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J. F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503-2511.