AI-Driven Nutrigenomic Modeling of Gene–Diet Interactions in Obesity and Insulin Resistance Through Skeletal Muscle Splicing Signatures

Vaibhav Venkataraman
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
vaibhav.venkataraman774@ku.edu

Jean A. Harrison
Department of Computer Science, University of North Texas, Denton, TX, USA.
jeanharrison@unt.edu

Abstract

The rising prevalence of obesity and insulin resistance represents a global health crisis that demands integrative approaches capable of reconciling genetic predisposition with dietary intervention. Recent advances in transcriptomics have revealed that alternative splicing events in skeletal muscle tissue constitute a critical but underexplored layer of gene regulation that mediates metabolic responses to nutritional stimuli. This paper presents a system-level analysis of artificial intelligence-driven nutrigenomic modeling that leverages skeletal muscle splicing signatures to decode gene–diet interactions underlying obesity and insulin resistance. We argue that traditional linear models fail to capture the combinatorial complexity of splicing regulatory networks and the non-linear feedback loops between macronutrient composition, post-transcriptional modification, and metabolic phenotype. Artificial intelligence architectures, particularly deep learning and graph neural networks, offer the capacity to integrate multi-omic data with high-dimensional splicing profiles, while also enabling causal inference through counterfactual reasoning. However, the deployment of such models raises significant structural trade-offs involving robustness, generalizability, interpretability, and computational sustainability. The infrastructure required for federated learning across clinical cohorts introduces governance challenges regarding data sovereignty, algorithmic fairness, and equitable access to precision nutrition. We examine these dimensions through the lens of socio-technical infrastructure, drawing parallels with large-scale genomic data sharing initiatives. A synthesis of recent experimental evidence highlights that exercise and dietary weight loss interventions alter splicing patterns in a manner that is both gene-specific and polymorphism-dependent, further complicating predictive modeling. We conclude by outlining a framework for responsible AI deployment in nutrigenomics that prioritizes transparency, reproducibility, and policy alignment with population health objectives.

Keywords

nutrigenomics, alternative splicing, artificial intelligence, skeletal muscle, obesity, insulin resistance, gene–diet interactions, system architecture, fairness, governance.

1. Introduction

The interplay between genetic variation and dietary exposure constitutes a central axis of metabolic disease etiology that has historically been addressed through reductionist approaches such as candidate gene studies or genome-wide association scans. Yet the mechanistic link between genotype and phenotype in obesity and insulin resistance remains elusive, partly because regulatory processes like alternative splicing introduce substantial plasticity in gene expression that is not captured by static measures of DNA sequence variation. Skeletal muscle, as the primary site of glucose disposal and a major contributor to whole-body energy expenditure, presents a uniquely informative tissue for studying how diet modulates gene expression through splicing. The emergence of high-throughput RNA sequencing has uncovered thousands of splicing events that respond to nutritional interventions, but the dimensionality of these data far exceeds the capacity of conventional statistical methods [1]. Artificial intelligence offers a paradigm shift by enabling models that can learn complex, non-linear relationships from high-dimensional splicing signatures while incorporating prior biological knowledge. This paper adopts a systems engineering perspective to evaluate the architectural choices, governance challenges, and sustainability implications of deploying AI-driven nutrigenomic models in research and clinical settings. We do not focus on algorithm details but rather on the structural trade-offs that determine whether such models can be robust, fair, and scalable.

2. The Complexity of Gene–Diet Interactions in Skeletal Muscle

Nutrigenomics seeks to elucidate how dietary components influence gene expression and how genetic variants modulate responses to diet. In skeletal muscle, the interaction is particularly dynamic because this tissue exhibits high metabolic flexibility and undergoes constant remodeling in response to nutrient supply and physical activity. Alternative splicing adds a layer of regulatory complexity: a single gene can produce multiple protein isoforms with distinct functions in glucose uptake, insulin signaling, and lipid metabolism. For example, splicing of the insulin receptor gene generates isoforms that differ in their affinity for insulin, thereby modulating downstream signaling cascades [2]. Diet composition, especially the ratio of carbohydrates to fats and the presence of specific fatty acids, can shift splicing patterns in a time-dependent manner. Furthermore, polymorphisms in splice regulatory sequences or in genes encoding splicing factors can profoundly alter the splicing response to a given dietary challenge [3]. The resulting system is characterized by feedback loops: diet modifies splicing, which alters protein function, which in turn affects metabolic state and future dietary preferences. Traditional linear regression models cannot capture these non-linear dynamics, nor can they handle the combinatorial explosion of possible gene–diet–splice interactions. This complexity motivates the deployment of AI architectures that can approximate high-dimensional functions while remaining tractable for inference.

3. AI Architectures for Nutrigenomic Modeling

The design of an AI system for nutrigenomic modeling involves several architectural decisions that directly impact its ability to generalize across populations and interventions. Deep neural networks, particularly those with attention mechanisms or transformer layers, have shown promise in integrating splicing quantitative trait loci with dietary exposure variables [4]. Graph neural networks are especially suited for modeling regulatory networks because they can represent genes, splicing factors, and metabolites as nodes with edges encoding known or inferred interactions. By propagating information through the graph, such models can capture indirect effects that are critical for understanding how a dietary change propagates through the splicing regulatory network to influence insulin resistance. However,

these architectures introduce a trade-off between expressive power and interpretability. A highly parameterized model may achieve excellent predictive accuracy on a training cohort but fail to generalize to new genetic backgrounds or dietary patterns due to overfitting to cohort-specific confounders [5]. Regularization techniques, dropout, and Bayesian inference can improve robustness but at the cost of increased computational demand. Another structural consideration is the choice between supervised and semi-supervised learning. Given that labeled clinical outcomes (e.g., insulin sensitivity indices) are often scarce, while unlabeled splicing data are abundant, models that leverage self-supervised pretraining on large transcriptomic databases can capture general splicing patterns before fine-tuning on specific outcomes. This approach has been successful in other genomic domains [6] and offers a pathway toward more sustainable model development by reducing the need for expensive human annotation.

4. Splicing Signatures as Functional Biomarkers

Skeletal muscle splicing signatures are not merely passive reporters of genetic variation; they actively mediate the functional response to diet and exercise. Recent experimental studies have demonstrated that thousands of exons are differentially spliced in response to a high-fat diet or a weight loss intervention, and that these changes correlate with improvements in insulin resistance independently of overall gene expression levels [7]. This suggests that splicing signatures could serve as early biomarkers for dietary efficacy or as predictors of metabolic trajectory. In the context of AI-driven modeling, splicing features present both an opportunity and a challenge. On one hand, they offer a high-dimensional representation that is rich in biological signal; on the other, they are influenced by technical artifacts such as library preparation, sequencing depth, and batch effects. Robust AI models must therefore incorporate preprocessing pipelines that normalize splicing metrics across heterogeneous data sources. The integration of splicing signatures into a predictive model for insulin resistance requires careful feature selection, as many splicing events may be redundant or noise-driven. Sparse autoencoders or variational inference methods can distill the most informative splicing patterns while preserving the underlying regulatory structure [8]. A particularly compelling result comes from a recent large-scale cohort study that examined the impact of polymorphisms on gene expression and splicing in response to exercise and diet-induced weight loss in human skeletal muscle, highlighting that the same polymorphism can have divergent effects on splicing depending on the dietary context [9]. This finding underscores the need for AI models that condition on both genetic and environmental variables simultaneously, rather than treating them as additive factors.

5. System-Level Trade-offs: Robustness vs. Generalizability

In any AI-driven system deployed across multiple clinical sites or population groups, the tension between robustness and generalizability must be carefully managed. Robustness refers to the model's ability to maintain performance under distributional shifts that arise from differences in sequencing platforms, dietary recall methods, or measurement error. Generalizability, by contrast, pertains to the model's accuracy when applied to new populations with distinct allele frequencies, cultural dietary habits, or comorbidities. In nutrigenomic modeling, these two objectives often conflict. A model that is heavily regularized to perform well on a specific cohort may exhibit low variance but high bias, failing to capture true biological variation across groups [10]. Conversely, a highly flexible model may fit the training data well but produce erratic predictions on new samples. Ensemble methods that combine predictions from multiple architectures can mitigate this

dilemma but introduce additional computational and infrastructure overhead. Another key trade-off involves the choice between population-level inference and individualized predictions. Most current models aim to predict the average effect of a diet on a splicing signature, but precision nutrition demands personalized recommendations. Individual-level predictions require modeling interactions between thousands of rare variants and dietary components, which is statistically challenging even with large sample sizes. AI models that incorporate hierarchical Bayesian structures can partially address this by sharing information across individuals while allowing for patient-specific parameters [11]. The deployment of such models in clinical decision support systems must also account for the uncertainty in predictions, providing confidence intervals or prediction intervals rather than point estimates, to avoid overreliance on spurious results.

6. Infrastructure and Deployment Considerations

The practical implementation of AI-driven nutrigenomic modeling requires an infrastructure that is both scalable and secure. Splicing data derived from RNA sequencing are typically large (tens of gigabytes per sample), necessitating high-performance computing clusters or cloud-based storage with efficient data transfer protocols. Federated learning has emerged as a promising architecture for training models across multiple institutions without centralizing sensitive genomic data, thereby addressing privacy concerns and regulatory compliance [12]. In a federated setting, each site trains a local model on its own data and only shares model updates (gradients or weights) with a central aggregator. However, heterogeneity in data distributions across sites can lead to suboptimal global models if not addressed through techniques like weighted aggregation or personalization. Moreover, the communication overhead of federated learning for high-dimensional splicing features can be substantial, requiring careful optimization of model compression and quantization. Another infrastructure dimension is the integration of AI models with electronic health records and dietary tracking platforms. Interoperability standards such as FHIR (Fast Healthcare Interoperability Resources) must be extended to accommodate genomic and transcriptomic data, while ensuring that splicing signatures can be computed in real time from raw sequencing reads. The deployment of such systems also raises the question of sustainability: the carbon footprint of training large neural networks for genomics is non-trivial, and researchers should consider strategies such as model distillation or sparse computation to reduce energy consumption [13]. In low-resource settings, infrastructure constraints may limit the adoption of complex models, reinforcing inequities in access to precision nutrition. Therefore, infrastructure design must prioritize lightweight models that can run on modest hardware without sacrificing accuracy.

7. Governance, Fairness, and Policy Implications

The governance of AI-driven nutrigenomic models encompasses issues of data ownership, algorithmic bias, informed consent, and regulatory oversight. Genomic and dietary data are inherently sensitive, and participants must have clear control over how their data are used, particularly when it is aggregated for model training. The risk of re-identification from splicing data, though lower than from whole-genome sequences, is not negligible, and data sharing agreements must incorporate robust de-identification protocols and audit trails [14]. Fairness is a critical concern because splicing signatures may differ systematically across ethnic groups due to population-specific allele frequencies and environmental exposures. If AI models are trained predominantly on cohorts of European ancestry, they may yield biased predictions for other populations, leading to unequal effectiveness of dietary recommendations [15]. Algorithmic fairness metrics, such as demographic parity or equalized

odds, should be evaluated during model validation and used to inform retraining or recalibration. Policy mechanisms such as the European Union's Artificial Intelligence Act may classify nutrigenomic models as high-risk applications requiring conformity assessments and human oversight. Additionally, the use of AI to predict individual responses to diet raises ethical questions about medicalization of lifestyle, potential stigmatization, and the displacement of holistic dietary counseling by algorithmic recommendations [16]. A policy framework should promote transparency by requiring that model outputs be accompanied by explanations of their limitations and uncertainty intervals. It should also incentivize the inclusion of diverse populations in training data through equitable funding mechanisms for community-based participatory research.

8. Sustainability and Future Directions

The long-term sustainability of AI-driven nutrigenomic modeling depends on the continuous curation of high-quality splicing datasets, the refinement of architectures to reduce computational demands, and the development of standardized evaluation benchmarks. As new omics technologies such as single-cell RNA sequencing and long-read sequencing become more widespread, splicing signatures will be resolved at higher resolution, revealing cell-type-specific and isoform-specific responses to diet [17]. AI models must evolve to incorporate these multi-resolution data while maintaining computational efficiency. One promising direction is the use of meta-learning or few-shot learning to adapt pre-trained splicing models to new dietary interventions with limited labeled data. Another is the integration of causality-based frameworks that go beyond correlation to identify which splicing events are causal drivers of metabolic improvement, thereby enabling targeted nutritional strategies [18]. From a systems perspective, future research should focus on building digital twins of skeletal muscle metabolism that incorporate splicing dynamics, dietary inputs, and feedback from exercise. Such digital twins could simulate the long-term effects of dietary changes on insulin sensitivity before they are implemented in humans, reducing the cost and risk of clinical trials. Sustainability also demands that open science principles be applied to model sharing, with repositories hosting trained models, preprocessing pipelines, and benchmark datasets under permissive licenses. The academic community should establish standards for reporting model performance across diverse populations, similar to the STREAMS (Standardized Reporting of Exposures and Molecular Signatures) initiative [19].

9. Conclusion

AI-driven nutrigenomic modeling that leverages skeletal muscle splicing signatures holds transformative potential for understanding and intervening in obesity and insulin resistance. This paper has argued that the key challenges are not merely algorithmic but systemic, involving trade-offs between robustness and generalizability, infrastructure scalability, data governance, and fairness. The complexity of gene–diet interactions, mediated through alternative splicing, demands architectures that can capture non-linear dynamics and high-dimensional interactions while remaining interpretable and reproducible. Federated learning and graph neural networks offer promising technical pathways, but their deployment must be accompanied by careful attention to data sovereignty and algorithmic bias. As the field moves toward clinical translation, policy frameworks must evolve to ensure equitable access and ethical oversight. Future work should prioritize sustainability through computational efficiency, causal inference, and the inclusion of diverse populations. Ultimately, the success

of this endeavor will depend on interdisciplinary collaboration that bridges genomics, nutrition science, artificial intelligence, and social science.

References

1. Corella, D., & Ordovas, J. M. (2014). Nutrigenomics in cardiovascular medicine. Circulation: Cardiovascular Genetics, 7(2), 175–185.

2. Nutter, C. A., & Bregman, A. (2018). Alternative splicing of the insulin receptor: Implications for metabolic disease. Trends in Endocrinology & Metabolism, 29(6), 397–408.

3. Zhang, L., & Kwan, T. (2020). Genetic regulation of alternative splicing in human tissues. Nature Reviews Genetics, 21(2), 122–139.

4. Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. Molecular Systems Biology, 12(7), 878.

5. Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100–1122.

6. Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics, 20(7), 389–403.

7. Keller, M., & Häring, H. U. (2021). Skeletal muscle splicing in response to dietary interventions. Diabetes, 70(Suppl 1), A198.

8. Rampášek, L., & Goldenberg, A. (2020). Variational autoencoders for sparse genomic data. Proceedings of the 2020 ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 23–32.

9. Wang, W., Liew, W. L., Huang, S., Chan, E., Tan, A. L. M., Tian, C., ... & Liu, B. (2025). Impact of polymorphisms on gene expression and splicing in response to exercise and diet-induced weight loss in human skeletal muscle tissues. Cell Genomics, 5(9).

10. Domínguez-Villanueva, D., & Loza, E. (2022). Robustness vs. generalization in genomic deep learning: A survey. Briefings in Bioinformatics, 23(4), bbac261.

11. Calderón, A. A., & Zaitlen, N. (2018). Hierarchical Bayesian models for precision medicine. Annual Review of Statistics and Its Application, 5, 179–200.

12. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

13. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.

14. Berger, B., & Cho, H. (2020). Survey of data privacy in genomics. Nature Biotechnology, 38(7), 790–799.

15. Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., & Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics, 51(4), 584–591.

16. Vayena, E., & Blasimme, A. (2017). Biomedical big data: New models of control over access, use and governance. Journal of Bioethical Inquiry, 14(4), 543–554.

17. Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., ... & Yosef, N. (2017). The Human Cell Atlas. eLife, 6, e27041.

18. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54–60.

19. Bouchard, C., & Ordovas, J. M. (2022). Standardized reporting of exposures and molecular signatures in nutrigenomics. Nutrients, 14(11), 2298.