Middle East Technical University
Informatics Institute

DEVELOPMENT OF AN INTEGRATED WEB INTERFACE FOR IN-SILICO PATHOGENICITY PREDICTION OF GENOMIC VARIANTS

Advisor Name: Assoc. Prof. Dr. Aybar Can ACAR
(METU)

Student Name: Ayşegül BALCI
(MI)

January 2026

TECHNICAL REPORT
METU/II-TR-2026-

Orta Doğu Teknik Üniversitesi
Enformatik Enstitüsü

GENOMİK VARYANTLARIN IN SİLİKO PATOJENİTE TAHMİNİ İÇİN BÜTÜNLEŞİK WEB ARAYÜZÜ GELİŞTİRİLMESİ

Danışman Adı: Dr. Öğr. Üyesi Aybar Can ACAR
(ODTÜ)

Öğrenci Adı: Ayşegül BALCI
(MI)

Ocak 2026

TEKNİK RAPOR
ODTÜ/II-TR-2026

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY (Internal Use)	**2. REPORT DATE** 16.01.2026

3. TITLE AND SUBTITLE

DEVELOPMENT OF AN INTEGRATED WEB INTERFACE FOR IN-SILICO PATHOGENICITY PREDICTION OF GENOMIC VARIANTS

4. AUTHOR (S) Ayşegül BALCI	**5. REPORT NUMBER (Internal Use)** **METU/II-TR-2026-**

6. SPONSORING/ MONITORING AGENCY NAME(S) AND SIGNATURE(S)

Informatics Online Master's Programme, Department of Information Systems, Informatics Institute, METU
Advisor: Assoc. Prof. Dr. Aybar Can ACAR Signature:

7. SUPPLEMENTARY NOTES

8. ABSTRACT (MAXIMUM 200 WORDS)

This project aimed to develop, a web-based clinical decision support system for genomic variant prioritization. The project focused on pointing the limitations of conventional variant data aggregation by introducing a context-aware prioritization approach that evaluates the functional impact of genetic variants. By integrating real-time genomic data through external APIs and avoiding reliance on large local databases, the project supports up-to-date and efficient variant analysis. Overall, this project contributes to improved clinical variant interpretation by providing a lightweight and biologically informed decision support framework. Project source code is available at: https://github.com/safiderum/thesis_project.git

9. SUBJECT TERMS	**10. NUMBER OF PAGES**
Genomic Variant Interpretation, In Silico Pathogenicity Prediction, Prediction Models	**25**

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

The completion of the Human Genome Project in 2003 marked a transformative milestone in biomedical research, revealing the approximately 3 billion base pairs that constitute human DNA and establishing a reference framework for understanding genetic variation (International Human Genome Sequencing Consortium, 2004). However, this achievement was almost the beginning of a far more complex challenge: deciphering the functional consequences of the millions of genetic variations that distinguish individuals and contribute to human disease. With the advent of next-generation sequencing (NGS) technologies over the past two decades, the cost of whole-genome and whole-exome sequencing has decreased exponentially, from nearly $100 million per genome in 2001 to less than $1,000 today (Goodwin et al., 2016). This technological revolution has democratized access to genetic information, enabling various opportunities for precision medicine, disease diagnosis, and therapeutic development.

Despite these advances, the interpretation of genetic variants remains a formidable bottleneck in translating genomic data into clinically actionable insights. A single human genome typically harbors 4 to 5 million single nucleotide variants (SNVs) compared to the reference genome, along with hundreds of thousands of small insertions and deletions (indels) and thousands of structural variants (Auton et al., 2015). The vast majority of these variations are benign polymorphisms that reflect normal diversity. However, there are rare pathogenic vairants within this genomic landscape, which responsible for Mendelian disorders, predisposition to complex diseases. The critical challenge facing clinical geneticists, researchers, and bioinformaticians is distinguishing the handful of disease-causing variants from the enormous background of neutral variation—a task analogous to finding needles in an ever-growing haystack.

The complexity of variant interpretation is compounded by the heterogeneity of genetic variation itself. Variants can affect coding sequences, splice sites, regulatory regions, or non-coding RNAs, each through distinct molecular mechanisms. Missense variants, which result in amino acid substitutions, represent the most common class of disease-associated variants but are also the most challenging to interpret, as their functional impact depends on factors such as evolutionary conservation, structural location within the protein, and biochemical properties of the substituted amino acid. Loss-of-function variants, including nonsense mutations and frameshift indels, generally have more predictable severe consequences, yet their clinical significance may be modulated by mechanisms such as nonsense-mediated decay or the existence of alternative transcripts. Furthermore, variants in non-coding regions—constituting approximately 98% of the genome—are increasingly recognized as contributors to disease through effects on gene regulation, yet remain particularly difficult to interpret due to our incomplete understanding of regulatory grammar (Maurano et al., 2012).

To address this interpretation challenge, the genomics community has developed a diverse ecosystem of computational tools designed to predict the functional and clinical impact of genetic variants. These *in silico* prediction tools leverage various sources of information, including evolutionary conservation across species, protein structural data, functional genomic annotations, and empirical observations of variant-disease associations. Early tools such as SIFT (Sorting Intolerant From Tolerant) and PolyPhen-2 (Polymorphism Phenotyping v2) pioneered the use of sequence conservation and protein structure to assess variant deleteriousness. More recent approaches, including BayesDel and REVEL (Rare Exome Variant Ensemble Learner), integrate dozens of diverse features using machine learning frameworks to generate comprehensive pathogenicity scores. Today, researchers and clinicians have access to more than 20 different prediction algorithms, each with distinct methodologies, training datasets, and performance characteristics.

CHAPTER 1: IN SILICO PREDICTION TOOLS

1.1 Overview of Variant Prediction Algorithms

The landscape of genetic variant interpretation has undergone a deep transformation, transitioning from labor-centered mostly manual curation to high-throughput computational frameworks driven by the need of quantifying the probability that a specific genomic substitution contributes to a disease phenotype. As Next-Generation Sequencing (NGS) continues to expand the volume of identified variants beyond human processing capacity, reliance on in silico prediction models has become inevitable for clinical diagnostics. This demand has lead to the development of more than 60 distinct computational tools over the past two decades, each of them uses unique methodological frameworks and diverse biological datasets (Ghosh et al., 2017). As the variant pathogenity coud arise through many different types of molecular distruptions, this increase is natural. According to Starita et al. (2017), genetic changes can compromise cellular homeostasis by altering amino acid biochemistry, destabilizing protein structures, interfering with protein-protein interactions, or affecting RNA splicing and gene expression. Because no single algorithm can capture these disparate mechanisms with uniform sensitivity, the field has evolved along parallel tracks, emphasizing different biological principles to address the limitations of individual predictive approaches.

To navigate this heterogeneous landscape, these computational methodologies are said to be organized into three foundational paradigms: conservation-based methods, functional impact predictors, and machine learning ensemble frameworks. Conservation-based tools, such as PhyloP, utilize cross-species genomic alignments to identify functionally vital sites subject to selective pressure, assuming that evolutionary constraint is a proxy for

biological importance. In contrast, functional impact predictors emphasize the mechanistic consequences of mutations, modeling disruptions in protein stability, hydrophobicity, and molecular folding. And there are machine learning ensemble frameworks, or meta-predictors, represent a more recent evolution. Tools like REVEL and BayesDel function as "aggregators," utilizing supervised learning algorithms to synthesize outputs from dozens of individual conservation and functional tools into a single, consolidated pathogenicity probability.

A significant technological divergence has recently emerged within this third category between traditional meta-scores and modern deep learning models, such as AlphaMissense or SpliceAI. While meta-scores depend on the pre-existing scores and inherent biases of their constituent algorithms, deep learning architectures function as "autonomous learners." These models utilize complex neural networks to extract high-resolution biological representations directly from raw genomic sequences or protein structures. By learning the "language" of protein folding or RNA splicing motifs without the need for secondary annotations, deep learning approaches offer a superior resolution of molecular perturbations, often capturing intricate pathogenic patterns that traditional ensemble methods might overlook.

1.2 Integrated In Silico Prediction Algorithms

1.2.1 REVEL

REVEL represented as “meta-prediction” that combines scores from 13 individual variant effect prediction tools using a random forest ensemble learning algorithm (Ioannidis et al., 2016). The tool was specifically optimized for rare missense variants with minor allele frequencies below 0.5%. REVEL generates scores ranging from 0 to 1, the higher values indicating the greater predicted pathogenicity. In comparative benchmarking, REVEL demonstrated superior performance with area under the curve (AUC) values 0.046-0.182 higher than individual predictors and outperformed seven other ensemble methods (Ioannidis et al., 2016). Clinical validation studies confirmed REVEL's utility, with Tian et al. (2019) reporting an overall performance probability (OPP) of 0.907, making it one

of the most reliable tools for ACMG/AMP PP3 criterion application in clinical variant classification workflows.

1.2.2 AlphaMissense

AlphaMissense applies deep learning architecture adapted from AlphaFold to predict the pathogenicity of missense variants across the entire human proteome (Cheng et al., 2023). This model was trained on human and primate variant population frequency data, combining structural context and evolutionary information simultaneously. AlphaMissense provides pathogenicity scores from 0 to 1, with a threshold of 0.564 distinguishing likely pathogenic from likely benign variants. The tool successfully classified 89% of the 71 million possible human missense variants as either likely benign or likely pathogenic, substantially reducing the proportion of variants of uncertain significance (Cheng et al., 2023). While AlphaMissense shows promise, recent evaluations suggest its performance may vary across different protein families and that integration with other evidence types remains essential for clinical interpretation.

1.2.3 BayesDel

BayesDel employs a Bayesian framework to integrate multiple genomic and functional annotations for predicting variant deleteriousness, with applicability to both coding and non-coding variants (Feng, 2017). The tool distinguishes itself from traditional “black-box” ensemble meta-predictors by employing a Naive Bayes classifier, which provides a transparent statistical framework rooted in conditional probability. By applying a probabilistic framework based on likelihood ratios, the tool weights this evidence to create a posterior probability score that ranging from -1.29 to 0.76 and quantifies the variant's pathogenic potential. Comparative studies demonstrated that BayesDel outperformed established tools including PolyPhen-2, SIFT, CADD, MetaLR, and MetaSVM across multiple benchmark datasets (Feng, 2017). In clinical validation, BayesDel achieved an OPP of 0.908, nearly identical to REVEL's performance, establishing it as a co-leader among meta-predictors for clinical variant classification (Tian et al., 2019).

1.2.4 SpliceAI

SpliceAI utilizes a deep neural network trained on 21,602 genes to predict splice-altering variants by analyzing up to 10,000 nucleotides of sequence context (Jaganathan et al., 2019). The tool outputs delta scores from 0 to 1, representing the predicted probability change in splice site use, with recommended thresholds of ≥0.2 for high sensitivity and ≥0.5 for high specificity applications. Benchmark studies using massively parallel splicing assays demonstrated that SpliceAI achieved the top overall performance among eight splice prediction tools tested, though performance was notably better for intronic variants compared to exonic variants (Dawes et al., 2023). The tool's ability to detect long-range splicing effects represents a significant advancement over traditional splice site prediction methods.

1.2.5 dbscSNV-ADA

The dbscSNV database provides ensemble splice variant predictions using both AdaBoost (ADA) and Random Forest (RF) algorithms that integrate eight individual splice prediction tools (Jian et al., 2014). The scores range from 0 to 1, with a recommended pathogenicity threshold of 0.6 for both ADA and RF variants. The method focuses on consensus splice regions (positions -3 to +8 at donor sites and -12 to +2 at acceptor sites). Performance evaluations demonstrated that dbscSNV-ADA showed particularly strong performance in exonic regions, and the AdaBoost algorithm retained the strongest signals of purifying selection, suggesting high accuracy in identifying functionally constrained positions (Liu et al., 2022). However, its limitation to consensus splice regions means it cannot evaluate deep intronic or exonic splicing regulatory elements that SpliceAI can assess.

1.3 Tool-Specific Strengths and Limitations

The effectiveness of prediction tools varies substantially depending on the type and location of genetic variants, requiring careful tool selection based on variant characteristics. For missense variants in protein-coding regions, REVEL and AlphaMissense demonstrate superior performance, with REVEL specifically optimized for rare variants (MAF <0.5%) commonly encountered in Mendelian disease diagnosis.

AlphaMissense offers broader proteome coverage with predictions for all possible missense substitutions, making it valuable for analyzing variants in less-studied proteins where empirical training data may be limited (Cheng et al., 2023). BayesDel provides a versatile alternative that extends beyond missense variants to include non-coding regions and insertion-deletion variants, offering comparable performance to REVEL for coding variants while maintaining utility across the genome (Feng, 2017). For splice-altering variants, the choice between tools depends critically on variant location. SpliceAI excels at detecting variants throughout exons and deep intronic regions by analyzing up to 10,000 nucleotides of sequence context, demonstrating particular strength for intronic variants that affect cryptic splice sites or splicing regulatory elements (Jaganathan et al., 2019; Dawes et al., 2023). Conversely, dbscSNV-ADA shows optimal performance for variants within canonical splice consensus regions, particularly in exonic positions immediately adjacent to splice junctions, where its ensemble approach combining eight prediction tools provides robust assessments (Jian et al., 2014; Liu et al., 2022).

To mention the balance between sensitivity and specificity, it represents a critical consideration when selecting and interpreting variant prediction tools, particularly in clinical diagnostic settings where false positives and false negatives carry different consequences. High-sensitivity tools minimize the risk of missing pathogenic variants but may generate more false positive predictions, potentially leading to unnecessary clinical interventions or patient anxiety. REVEL and BayesDel achieve balanced performance with sensitivities and specificities both exceeding 90% when using optimized thresholds, making them suitable for general clinical use (Tian et al., 2019). However, threshold selection significantly impacts this balance; for instance, SpliceAI recommends a delta score threshold of ≥ 0.2 for high-sensitivity applications such as research screening, but ≥ 0.5 for higher specificity needed in clinical reporting (Jaganathan et al., 2019). The 2015 ACMG/AMP guidelines recommend using multiple lines of evidence rather than relying solely on computational predictions, acknowledging that even the best tools achieve imperfect accuracy. Ghosh et al. (2017) demonstrated that tool performance varies when applied to the ACMG/AMP PP3 (supporting pathogenic) and BP4 (supporting benign) criteria, with some tools showing high specificity but insufficient sensitivity for confident benign classification. This asymmetry has important implications: while strong

computational evidence can support pathogenic classification when combined with other evidence, computational predictions alone rarely provide sufficient evidence for benign classification without additional functional or population data.

Integrating computational predictions into clinical variant interpretation requires understanding both the capabilities and limitations of prediction algorithms within established frameworks such as the ACMG/AMP guidelines. These guidelines assign varying weights to different types of evidence, with computational predictions typically contributing "supporting" rather than "strong" evidence for pathogenicity. Best practices recommend using predictions from multiple complementary tools rather than relying on a single algorithm, as concordance among independent methods increases confidence in the prediction. For missense variants, the combination of REVEL or BayesDel with functional domain information and population frequency data provides robust evidence, while discordant predictions should prompt additional investigation (Tian et al., 2019). Splice variants require special consideration, as high SpliceAI scores ($\geq$0.5) correlate strongly with aberrant splicing but cannot distinguish between splicing changes that cause disease and those that are functionally neutral; correlation with gene-specific loss-of-function mechanisms is essential (Jaganathan et al., 2019). Clinicians should also consider tool-specific biases and training data limitations, recognizing that prediction accuracy may be lower for variants in understudied genes, non-European populations, or proteins with unusual evolutionary patterns. The interpretation workflow should prioritize tools validated for the specific variant type, use appropriate thresholds based on clinical context, integrate computational predictions with experimental and clinical evidence, and document the rationale for tool selection and threshold choices in clinical reports.

CHAPTER 2: PROJECT IMPLEMENTATION AND METHODOLOGY

2.1 Project Objectives and Requirements

The aim of the project is to build a comprehensive variant prioritization platform (VarAI) that transcends the limitations of conventional data aggregators, functioning instead as a web-based clinical decision support system. The context-aware prioritization concept of the platform dynamically evaluates the functional class of a variant and promotes the most revelant *in silico* tools accordingly. Furthermore, the system is providing educational framework that informs the user of the biological rationale behind tool selection, thereby transforming a raw data query into an insightful and transparent decision-making process. By that means VarAI aims to mitigate the “black-box” nature of computational genomics.

To satisfy its core objectives, the VarAI system fulfills several critical functional requirements. The system interfaces with genomic APIs to facilitate automated variant effect prediction, which enables the automatic annotation of variants and their assignment to the most appropriate in silico tool based on their specific functional impact. Following this annotation process, the user interface dynamically reorganizes itself to prioritize and present the most relevant scores. By implementing this logic, the platform ensures that the data presented is correlated to the biological context of the variant, thereby streamlining the clinical decision-making process. This underlying logic also facilitates automated discordance management, ensuring that when conflicting scores arise from different computational tools, the platform prioritizes the evidence based on the tool's validated performance for that specific variant class.

The techinal requirements are based on the necessity for high data currency and architectrual simplicity. A core requirement is integration of genomic data in a context of eliminating the relieance on large local genomic databases to retrieve variant scores and annotation with real-time accuracy. The system processes individual genomic queries without the necessity of complex database management systems, as it utilizing a sateless architecture to handle transactions indepedently. By means of this VarAI envisions a lightweight analysis environment. Lastly, the frontend design maintains the presentation of these dynamic datasets through interactive visualization components. Responsivity of the presentation of genomic constraint and pathogeniticity distributions is preserved by lightweight, API-driven backend. This combined technical approach provides efficient and transparent environment for variant analysis, centered around architectual simplicity and data currency.

2.2 System Architecture and Design

The VarAI platform is engineered as a lightweight, web-based clinical decision support system designed to facilitate real-time genomic variant interpretation. The system architecture adheres to the model view template design pattern inherent to the Django framework but is specifically optimized for a stateless, API-centric operation. Unlike traditional bioinformatics pipelines that necessitate extensive local hardware and static SQL databases, VarAI functions as intelligent middleware. This architectural choice was driven by the need to resolve the "data silo" problem in genomics, where local databases quickly become obsolete due to the rapid pace of clinical discoveries. By dynamically retrieving data from external knowledge bases, the system ensures that clinical decisions are predicated on the most current annotations available while maintaining a zero-footprint approach to data privacy.

The selection of Django as primary backend framework was dictated by its highly efficient URL rooting system essential for implementing a data-driven orchestration layer. This framework enables for us to selectively extract and prioritize specific features from expansive genomic datasets, ensuring that the final presentation is dynamically arranged. Unlike micro-frameworks, Django’s native support for Python allows for the integration

of libraries such as Requests for API management and the implementation of custom conditional logic for data processing. By utilizing standard Python floating-point operations and hierarchical decision structures, the system dynamically ranks in silico scores according to the variant's biological context without the need for additional mathematical frameworks. This approach ensures that the platform maintains its architectural simplicity while providing precise, on-the-fly normalization of genomic scores.

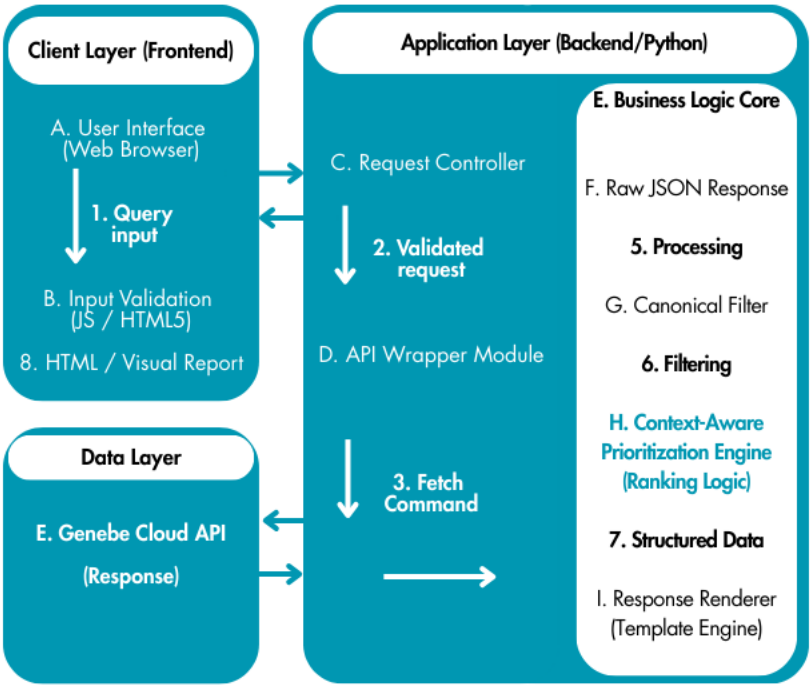

Figure 1: Software Architecture Schema

Expanding beyond its role as a standard data aggregator, the VarAI platform is engineered to function as an information-enriched decision support tool. While conventional platforms typically function as data aggregators that present raw scores without contextual weighting, VarAI is designed to guide the user through the clinical significance of each metric. This is achieved through an integrated interpretative layer that explains the biological relevance of specific scores through a simple ranking scheme based on the variant's genomic context. By doing so, the system addresses the common challenge in

clinical genomics where disparate scores may lead to conflicting interpretations, providing a structured framework for prioritizing evidence.

2.3 API Integration

The core data acquisition mechanism of the VarAI platform relies on a stateless integration with the Genebe Cloud REST API. Data acquisition is executed via synchronous HTTP GET requests utilizing the Python requests library(version 2.32.4). While the local development environment operates over standard HTTP, the VarAI architecture ensures end-to-end data encryption by executing all external genomic data requests via HTTPS/TLS 1.2+ protocols, preventing credential interception during API transactions. To facilitate retrievel the system transforms user-submitted genomic coordinates into a dynamic query structure. To fit the cretain API schema, input sanitization algorithms are implemented to normalize chromosome identifiers (e.g., removing the “chr” prefix) and standardize nucleotide inputs. To verify biological data integrity, the genome parameter hardcoded to the “hg38” (GRCH38) assembly version; this ensures the preventation of potential variant miscalling arising frome genome assembly conficts.

Upon receiving a response , the API returns a hierarchical JSON payload containing a list. That list includes all distinct mRNA transcripts map to the position. This multiplicity arises from alternative splicing wherby a single gene can produce multiple mRNA isoforms. The VarAI backend implements a specialized parsing engine to parse this structure. There is a risk of ambiguity; as a variant might be located in coding region in one transcript, suggesting pathogenity, while falling into non-coding region in another, suggesting a benign effect. To ensure clinical accuracy for this situation, the backend prioritize canonical transcripts. The backend algorithm filters the results to retain only the canonical annotated transcript, thereby presenting the user with the interpretation that is most clinically relevant and commonly used in standard practice.

To manage various risks associated with external dependencies, comprehensive expection handling meachanism is implemented in the integration layer. Critic scenarios such as network latency or data absence are managed instead of causing a system crash. While catching this exceptions,

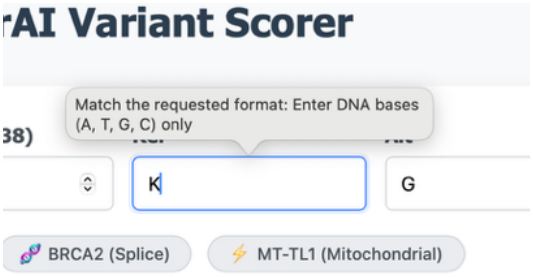

Figure 2: Real-time input validation mechanism

the system also translate them into user feedback. Additionally, to prevent submitting malformed queries to the API, a pre-request validation protocol is enforced. The system inputs against DNA nomenclature (as A, T, G, C, Figure 2) directly within the interface and blocks the submission of invalid queries beforehand. Therefore, typographic errors are caught instantly, allowing users to correct their data without ever triggering a server-side request.

2.4 Application Features

The user interface is engineered to prioritize workflow efficiency, diverging from the single-line input mechanism that is widely observed in traditional variant aggregators. Instead, VarAI has a structured query module supported by a proactive system that mentioned in previous part, that monitors user entries confrom to standard genomic nomeclature (Figure 3). Therefore, system is abstracting the technical complexities then allowing researchers to focus on biological data interpretation rather than navigating formatting ambiguities.

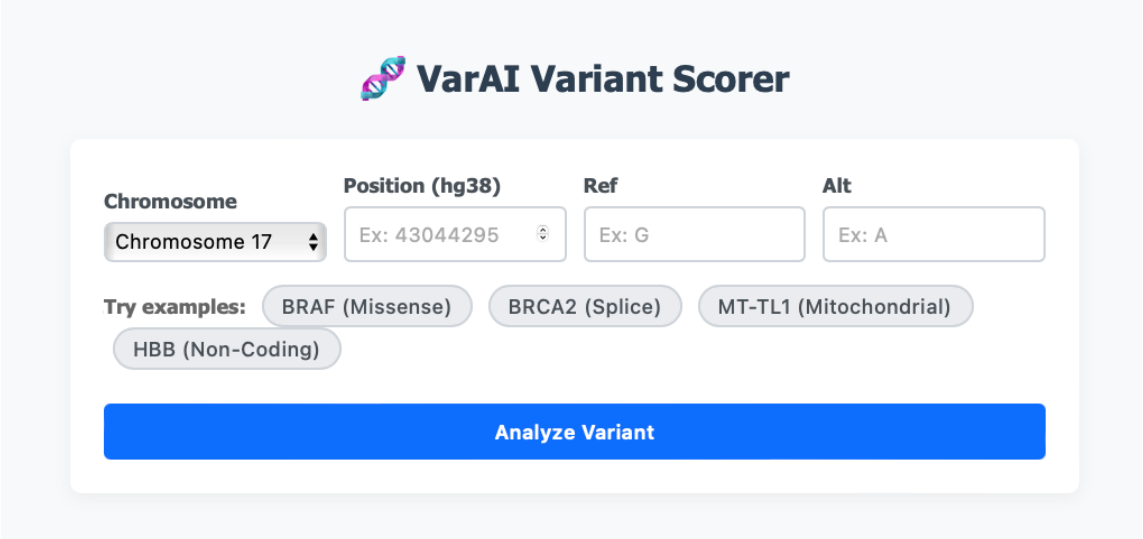

Figure 3: The VarAI homepage

Furthermore, the system addresses the critical challenge of interpretation, uncertainity arising from conflicting computational evidence. In clinical genomics, differences between general scoring algortihms and specilized predictors can confuse which metric should be prioritized in clinical decision-making. To solve this, VarAI implicates context-aware data organization.The system dynamically organizing the interface based on biological context that aligns with peer-reviewed benchmarking studies. For instance, for splice variants after SpliceAI, BayesDel prioritized over REVEL as it funcitons as meta-predictor designed for all variant types, whereas REVEL is strictly specilizied for missense variants (Figure 4). By ordering the data to specify the most validated tool for the specific variant, the platform prevents users from bieng confused or misled from random sorting of the metric, ensuring an accurate classification process.

In Silico Prediction Results

Tool	Score	Prediction
SpliceAI HIGHEST PRIORITY	0.98	**Pathogenic**
dbscSNV ADA	0.999938422912956	**Pathogenic**
BayesDel	0.25	**Pathogenic**
REVEL	None	None
AlphaMissense	None	None
MitoTip	-	-

Figure 4: Output dashboard showing prioritized pathogenicity scores

CHAPTER 3: CONCLUSION AND FUTURE WORK

3.1 Summary

This platform successfully indicates the practicality and clinical utility of a lightweight, stateless architecture for variant interpretation. As the system is directly integrating Genebe Cloud API it is delivering both resource-efficient and continuously updated solution to genomic data warehousing and its computational overhead. The development process prioritized clinical decision support over solely data aggregation; this is accomplished by user-centric interface that enforces data integrity via proactive input validation.

Crucially, the platform is distinguished by its context-aware engine which dynamically reorganizes specialized scoring methods interface based on variant with current scientific validation standards including prioritizing MitoTip for mitochondiral variants and SpliceAI for splicing events. This ordered evaluation approach effectively resolves the suspence in cınflicting computational evidence. In conclusion, VarAI found as validated model for next-generation clinical genomic tools, demonstrading that guided, context-aware interfaces can enhence the accuracy of clinical variant interpretation.

3.2 Contributions of This Project

This project contributes a functional framework for next-generation clinical decision support, demonstrating that robust variant analysis can be obtained through a resource-efficient, stateless architecture. By successfully decoupling application logic from data

persistence via Genebe Cloud integration, the platform eliminates the computational overhead of local warehousing while ensuring real-time access to evolving genomic annotations. Beyond architectural optimization, the study advances the field by implementing an evidence-based prioritization logic that addresses the challenge of interpretation uncertainity. Unlike traditional aggregators, the system operationalizes findings from recent validation literature to dynamically rank specialized predictors over generalized metrics based on biological context. This intelligent data organization, combined with proactive input validation and canonical transcript filtering, significantly reduces cognitive load, effectively transforming the interpretation process from a passive data-mining task into a streamlined, guided clinical assessment.

3.3 Limitations

While the VarAI platform effectively serves as a clinical decision support tool, several limitations are inherent to its current scope. Primarily, the enforcement of canonical transcript filtering, eventhough it is a necessity for standardization, introduces a potential blind spot regarding isoform-specific pathogenicity in non-canonical transcripts. In terms of workflow, the system is designed for single-variant verification rather than high-throughput discovery, thus it lacks support for batch processing or VCF file integration. Finally, the strategic choice of a stateless architecture precludes local caching. This results in redundant API calls for repeated queries and creates a critical external dependency risk, rendering the platform's operational stability directly dependent the uptime and network latency of the Genebe Cloud service.

3.4 Future Directions

Future development phases aim to enable VarAI to assume an educational role and transfer it to a fully explanatory clinical assistant. To enhance the interpretability of computational metrics, the interface could be augmented with dynamic contextual guides. Instead of presenting raw numerical scores in isolation, the system could incorporate interactive tooltips that explain the specific biological implications of each value. This

"Explainable AI" approach ensures that clinicians understand not just what the score is, but why it matters.

Furthermore, this context-aware data arrangment could be expanded to protein-level structural analysis. In future, integration with protein domain databases could specify protein topology then flag whether variant disturpts a critical functional element (e.g., conserved catalytic domain, an active site residue, known pathogenic hotspot). Multidimensional risk assessment could be offered by combining this structural data with existing pathogenity predictions.

APPENDIX: VARIANT ANALYSIS CASE STUDIES

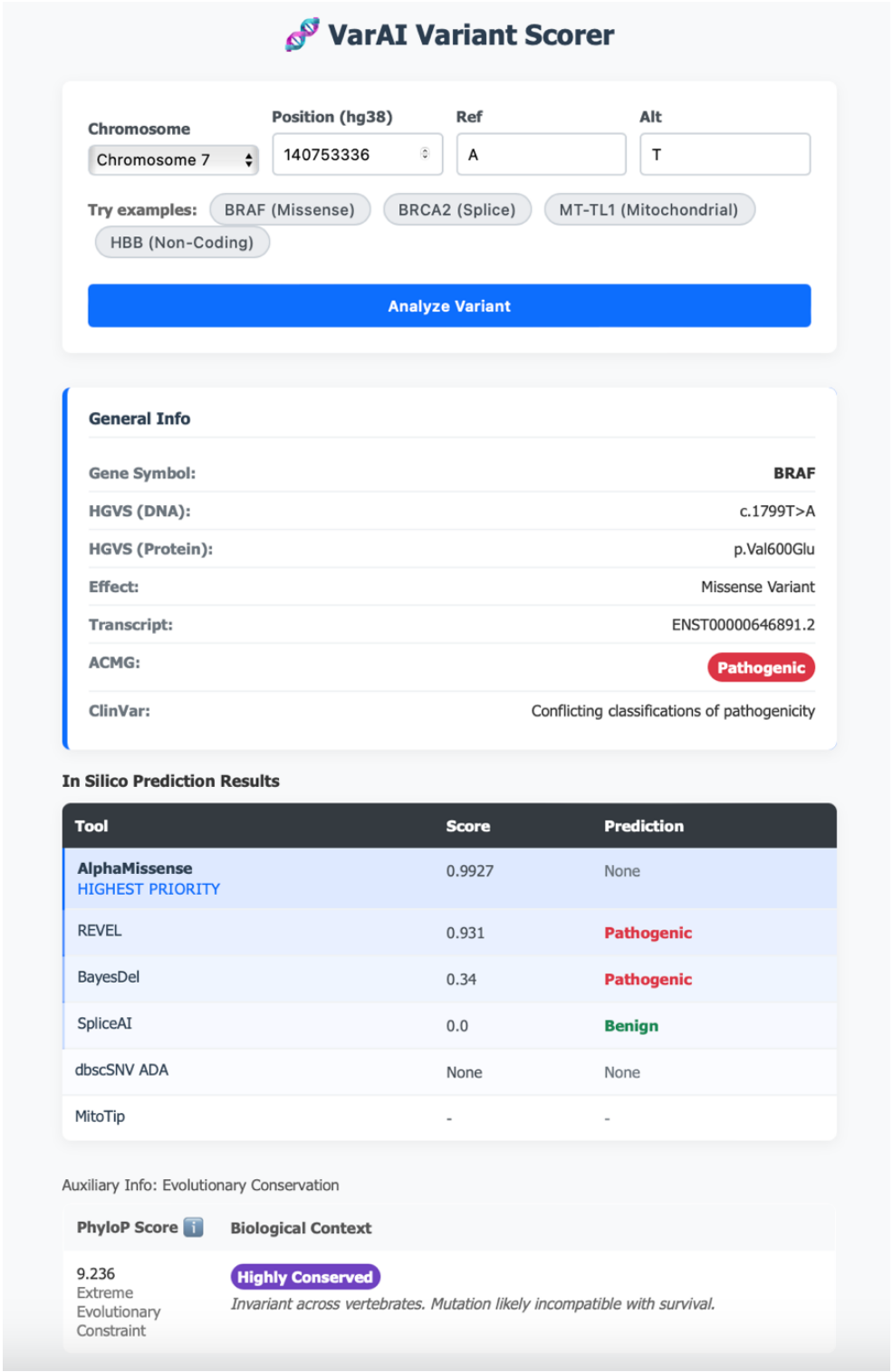

Figure 5: Full-page visualization of the VarAI diagnostic interface for a missense variant analysis

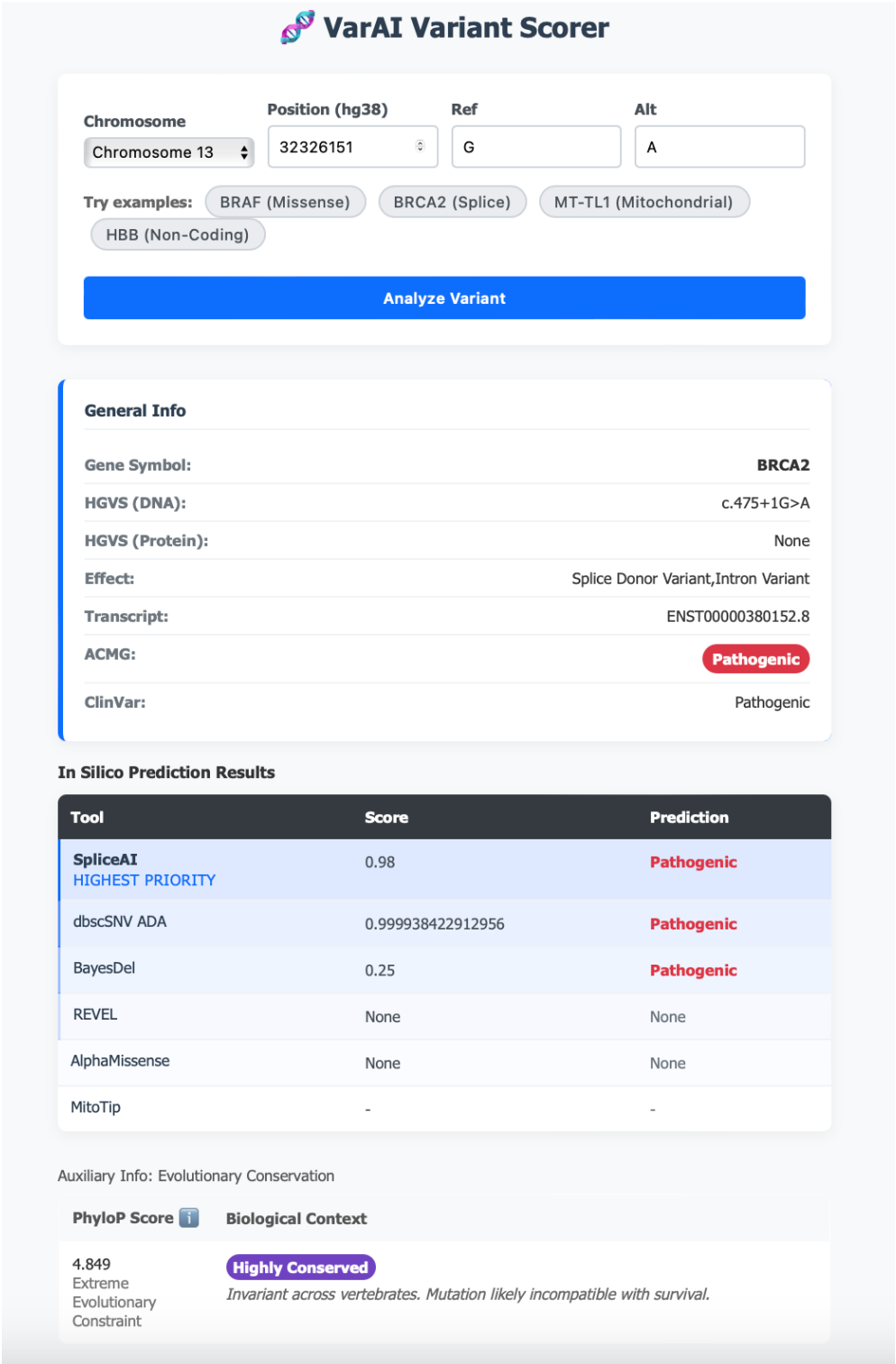

Figure 6: Full-page visualization of the VarAI diagnostic interface for a splice-donor variant analysis

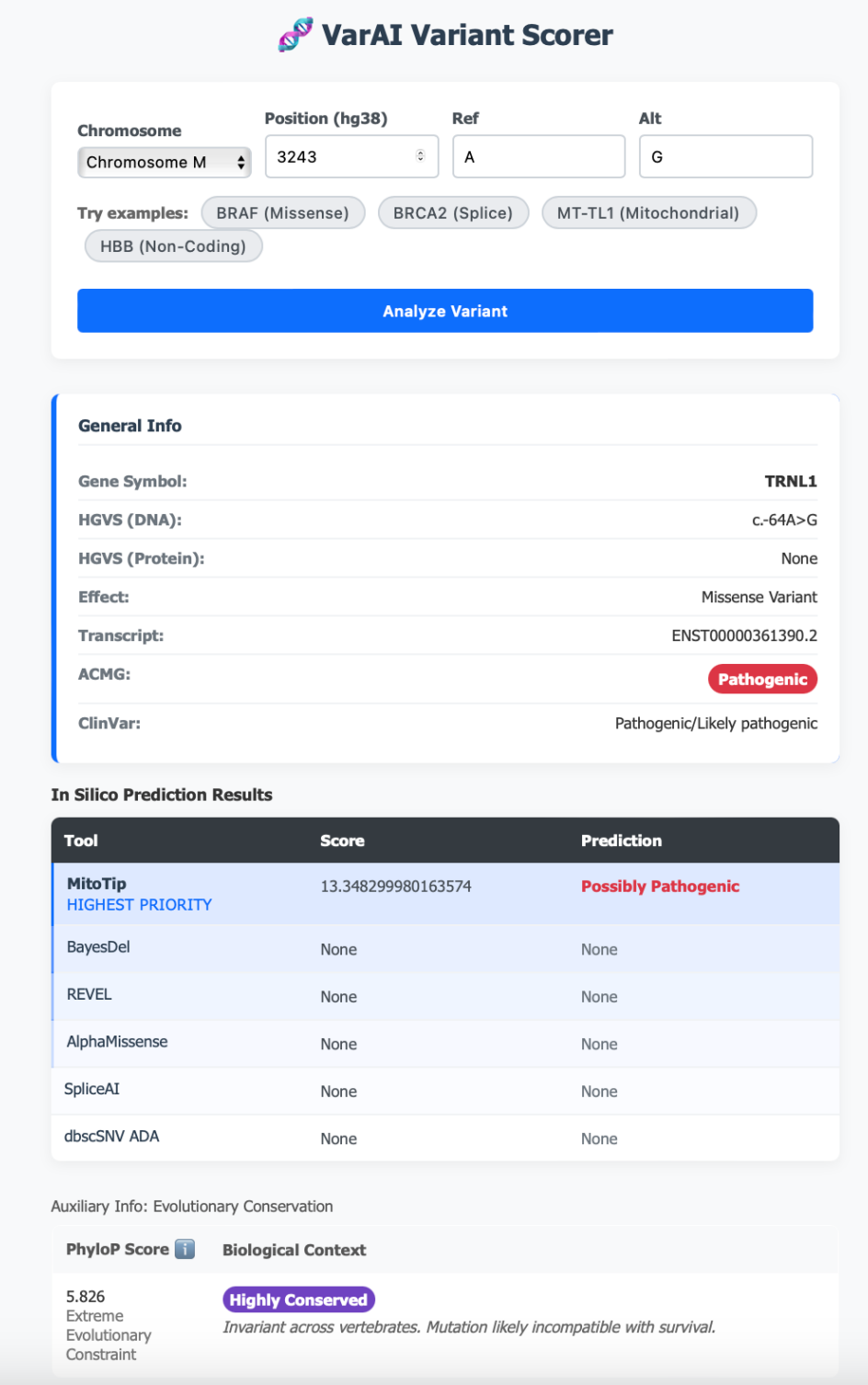

Figure 7: Full-page visualization of the VarAI diagnostic interface for a mitochondrial variant analysis

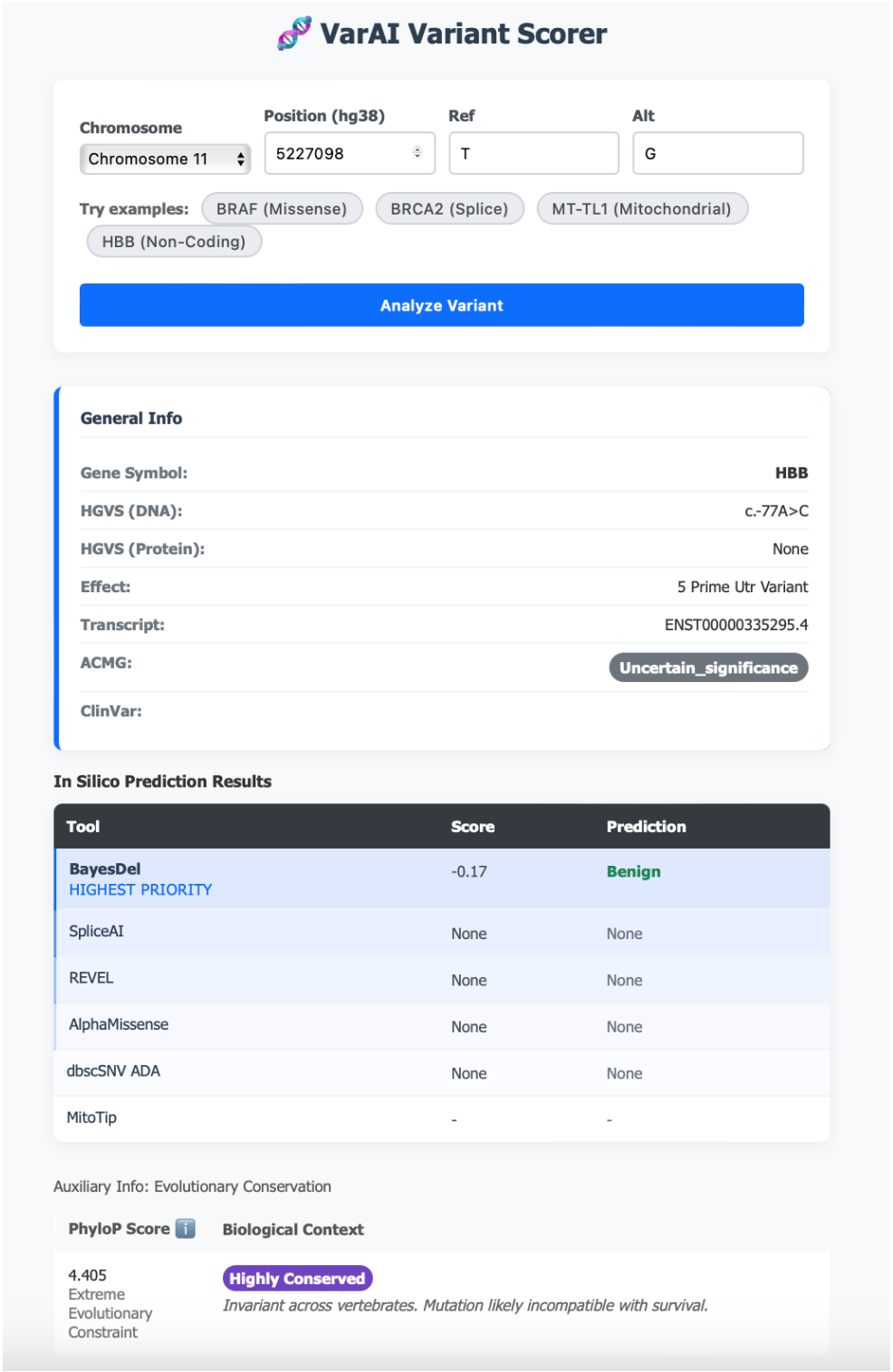

Figure 8: Full-page visualization of the VarAI diagnostic interface for a non-coding variant analysis

References

1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393. PMID: 26432245; PMCID: PMC4750478.

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, Schneider RG, Senior AW, Jumper J, Hassabis D, Kohli P, Avsec Ž. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023 Sep 22;381(6664):eadg7492. doi: 10.1126/science.adg7492. Epub 2023 Sep 22. PMID: 37733863.

Dawes, R., et al. (2023). "Benchmarking splice variant prediction algorithms using massively parallel splicing assays." *Genome Biology*, 24(1), 284.

Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017 Mar;38(3):243-251. doi: 10.1002/humu.23158. Epub 2017 Jan 28. PMID: 27995669; PMCID: PMC5299048.

Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017 Nov 28;18(1):225. doi: 10.1186/s13059-017-1353-5. PMID: 29179779; PMCID: PMC5704597.

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016 May 17;17(6):333-51. doi: 10.1038/nrg.2016.49. PMID: 27184599; PMCID: PMC10373632.

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016

Oct 6;99(4):877-885. doi: 10.1016/j.ajhg.2016.08.016. Epub 2016 Sep 22. PMID: 27666373; PMCID: PMC5065685.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KK. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan 24;176(3):535-548.e24. doi: 10.1016/j.cell.2018.12.015. Epub 2019 Jan 17. PMID: 30661751.

Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014 Dec 16;42(22):13534-44. doi: 10.1093/nar/gku1206. PMID: 25416802; PMCID: PMC4267638.

Liu H, Dai J, Li K, Sun Y, Wei H, Wang H, Zhao C, Wang DW. Performance evaluation of computational methods for splice-disrupting variants and improving the performance using the machine learning-based framework. Brief Bioinform. 2022 Sep 20;23(5):bbac334. doi: 10.1093/bib/bbac334. PMID: 35976049.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190-5. doi: 10.1126/science.1222794. Epub 2012 Sep 5. PMID: 22955828; PMCID: PMC3771521.

Smith, C., Kitzman, J.O. Benchmarking splice variant prediction algorithms using massively parallel splicing assays.*Genome Biol* **24**, 294 (2023). https://doi.org/10.1186/s13059-023-03144-z

Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM. Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet. 2017 Sep 7;101(3):315-325. doi: 10.1016/j.ajhg.2017.07.014. PMID: 28886340; PMCID: PMC5590843.

Tian, Y., Pesaran, T., Chamberlin, A. C., et al. (2019). "REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification." *Scientific Reports*, 9(1), 12304.