BRYAN L.A.

LITERATURE NOTES BIOINFORMATICS

UNIVERSITY OF AMSTERDAM ∩ VRIJE UNIVERSITEIT

UNIVERSITY OF AMSTERDAM ∩ VRIJE UNIVERSITEIT

THIS FILE CONTAINS NOTES FROM THE LITERATURE PAPERS I READ THROUGHOUT MY BIOINFORMATICS INTERNSHIP, 2025

Last updated June 2025

Contents

June 11^{th}/2025 5

June 12^{th}/2025 13

June 13^{th}/2025 21

June 15^{th}/2025 27

June 16^{th}/2025 29

June 17^{th}/2025 35

Bibliography 37

June $11^{th}/2025$

Evaluation of an automated genome interpretation model for rare disease routinely used in a clinical genetic laboratory [1]

Variant prioritization: Process of filtering and ranking a large number of genetic variants identified from sequencing data (i.e., exome/genome) to produce manageable shortlist of plausible candidates that may be responsible for a specific disease for a specific disease of phenotype.

Emedgene aims to reduce bottleneck by automatically generating a shortlist of candidate variants. Emedgene was evaluated on $n = 180$ retrospective *accuracy* previously solved exome cases and $n = 334$ *prospective* production cohort of consecutive clinical cases.

Correlated features with higher rank: Rare familial segregation, known pathogenicity, functional severity.

Accuracy was reduced in some cases due to incomplete genetic data (uncalled copy number variants) or atypical patient phenotypes. The AI-augmented analysis once integrated into workflow, achieved *diagnostic rate* 28.7% *vs. comparable historical manual rates* but significantly reduced the overall time required for case analysis by enabling a single cycle of review by a geneticist instead of two

Methods:

- Supervised learning approach, trained on dataset of 10^3's of variants that had been manually curated
- Decision tree clustering. It creates model that - based on input - produces score for ranking variants
- **Features:** Integration of information from multiple sources
  - **Variant Level:** Allele freq/count and count of homozygotes in public (e.g., gnomAD) and internal databases
  - **Gene Level:** info about affected gene

[1] L Meng, R Attali, T Talmy, Y Regev, N Mizrahi, P Smirin-Yosef, L Vossaert, C Taborda, M Santana, I Machol, R Xiao, H Dai, C Eng, F Xia, and S Tzur. Evaluation of an automated genome interpretation model for rare disease routinely used in a clinical genetic laboratory. *Genet Med*, 25(6):100830, 2023. DOI: 10.1016/j.gim.2023.100830

- **Phenotypic Similarity:** Measure of match between patient's reported phenotypes (using Human Phenotype Ontology terms) and phenotypes associated with diseases linked to the variant's gene
- **family segregation:** Analysis of inheritance patterns (e.g., identifying de *novo* variants or assessing zygosity in context of recessive patterns in x family
- **Functional Effect:** Predicted impact of variant on protein (e.g., loss-of-function effects like frameshift or nonsense mutations)
- **Known Pathogenicity:** Variants previously reported as pathogenic or likely pathogenic in databases like ClinVar or internal laboratory databases are given a very high rank

Training and evaluation:

- **Training:** Trained on manually curated variants to learn the correlations between the input features and the likelihood of a variant being diagnostic
- **Validation:** Model's performance was tested using **CV** on different segments of the data
- **Performance Metrics:** Sensitivity and specificity evaluate final model on accuracy cohort. sensitivity of 95.3% and a specificity of 99.9% for identifying a variant as a "most likely" candidate
- **Ranking Accuracy:** Rank of true diagnostic variant in the model's prioritized list of candidates

Current version does not account for certain types of genetic variation including CNVs, STRs, mitochondrial DNA variants

---

DEEP LEARNING-BASED RANKING METHOD FOR SUBGROUP AND PREDICTIVE BIOMARKER IDENTIFICATION IN PATIENTS [2]

Biomarkers associated with treatment effect heterogeneity = Predictive biomarkers. ML + Causal inference for predictive biomarker identification and ITR exploration.

To consider: Meta-learning, Q-learning, D-learning, DNNs for handling *complex biomarker-treatment response relationship*.

[2] Zhen Liu, Yifan Gu, and Xiaoyang Huang. Deep learning-based ranking method for subgroup and predictive biomarker identification in patients. *Communications Medicine*, 5:221, 2025. DOI: 10.1038/s43856-025-00946-z. URL https://doi.org/10.1038/s43856-025-00946-z

DeepRAB

mathematical framework to model *treatment effect heterogeneity* and construct *individualized Treatment Rule (ITR)*. Formulated as **supervised** ML problem where objective is to predict how much benefit a patient is likely to receive from a treatment based on their specific characteristics (**biomarkers**).

CAE: Concrete Autoencoder. RElationship between covariates and disease outcomes instead of relationship between individual treatment effects.

Modeling relationship: between covariates and disease outcomes

Prognostic model whose goal is to predict patient's likely outcome based on their baseline characteristics (*covariates*), irrespective of any treatment they may receive.

Let $X = (x_1, x_2, ..., x_p)$ be vector of patient's baseline covariates (e.g., genetic biomarkers, age...)

Let Y be disease outcome (e.g., $Y = 1 \rightarrow$ disease progresses)

Goal: Learn f that models probability of outcome given covariates:

$$f(X) \approx P(Y = 1|X)$$

A standard logistic regression could model as:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

CAE would be used for feature selection. In multi dimensional space where $p >>>$, the goal is to identify small but relevant subset of original covariates $X_s \subset X$. Autoencoder learns to reconstruct input features using a *concrete selector* layer that is forced to choose only a few features. FINAL prognostic model is then built using ONLY selected subset:

$$f(X_s) \approx P(Y = 1|X_s)$$

Output: Model finds prognostic biomarkers that is features associated with the outcome itself.

Modeling relationship between individual treatment effects:

Predictive causal model whose goal is not just to predict outcome, *but to predict* how the outcome *changes* when a patient receives a specific treatment vs. control. (**Approach used by DeepRAB**).

Consider treatment variable W, where $W = 1$ for active treatment, $W = 0$ for control.

For patient with covariates X:

- Y^1: Outcome patient will experiences if they received treatment $W = 1$
- Y^0: Same but for control $W = 0$

Individual Treatment Effect (ITE): $ITE = Y^1 - Y^0$ can only ever observe one for a given patient.

Goal is to learn model that estimates **CATE** $\tau(x)$

DeepRAB approaches such problem by learning to estimate the outcome under both scenarios. it learns 2 functions (2 heads of NNs)

- $\hat{\mu}_1(x) = \mathbb{E}[Y|X = x, W = 1]$ (predicted outcome if treated)
- $\hat{\mu}_0(x) = \mathbb{E}[Y|X = x, W = 0]$ (predicted outcome if controlled)
- CATE is then estimated as difference between the two predictions:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

- The model is trained to minimize the prediction error on the observed outcomes (e.g., using data from a randomized clinical trial where some patients received treatment and others received control).

Core Mathematical Concepts:

Casual Inference Framework:

Neyman-Rubin potential outcome for causal inference

- Let X_i be the vector of baseline biomarkers for patient i, A_i be the treatment assignment ($A_i = 1$ for treatment, $A_i = -1$ for control), and Y_i be the observed outcome.
- The model assumes two potential outcomes for each patient: $Y_i(1)$ (outcome if treated) and $Y_i(-1)$ (outcome if on control). We only observe one of these.
- The conditional expectation of the outcome is modeled as:

$$\mathbb{E}[Y|A, X] = Z(X)A + H(X)$$

  where:

  - $H(X) = \frac{1}{2}[\mathbb{E}[Y|A = 1, X] + \mathbb{E}[Y|A = -1, X]]$ represents the **prognostic effect** of the biomarkers X.
  - $Z(X) = \frac{1}{2}[\mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = -1, X]]$ is the **contrast function** that reflects the **heterogeneous treatment effect** given biomarkers X. The goal of DeepRAB is to accurately estimate this function $Z(X)$.

DeepRAB Model Architecture

Deep Neural Network (DNN) composed of 3 main components designed to estimate a *personalized benefit score*, $f(x)$, which is a monotonic transformation of the treatment effect function $Z(X)$.

- **(Component 1) Encoder / Biomarker Selection Layer:** Input layer that performs feature selection using techniques from Concrete Autoencoders (CAE). It learns to select user-specified number k of most informative biomarkers from full set of p input biomarkers. Selection is achieved by learning weight vector $\beta_j^{(0)}$ for each of the k nodes in this layer. The weights are generated using Gumbel-Softmax reparameterization (allows differentiable approximation to sampling from categorical distribution)
  - Probability of selecting j-th biomarker for i-th node in this layer:

$$\beta_{ij}^{(0)} = \frac{\exp((\log \alpha_i + g_j)/T)}{\sum_{t=1}^{p} \exp((\log \alpha_t + g_t)/T)}$$

- α and g are learnable parameters, T is temperature parameter that is annealed towards 0 during training. As $T \rightarrow 0$, vector $\beta_j^{(0)}$ becomes a $one - hot$ vector selecting a single biomarker from original input features x. Output of layer is $z^{(1)}$ which is vector of k selected biomarkers
- **(Component 2) Decoder / Hidden Layers:** Selected biomarkers $z^{(1)}$ are fed into (standard) Multi-Layer Perceptron (MLP) of $h - 1$ hidden layers that model the potentially complex and non-linear relationships between selected biomarkers and treatment effect
  - Output of each hidden layer $d^{(j)}$:

$$d^{(1)} = \phi_1(W^{(1)} z^{(1)} + b^{(1)})$$
$$d^{(j)} = \phi_j(W^{(j-1)} d^{(j-1)} + b^{(j-1)})$$

  - W and b are standard weight matrices and bias vectors, ϕ is non-linear activation function
- **(Component 3) Output Layer and Loss Function:** It produces personalized benefit score $f(x)$. Model is trained by minimizing a specific loss function based on **A-learning** (Advantage-learning) which is designed to directly estimate optimal Individualized Treatment Rule (ITR) without needing to model prognostic function $H(X)$.
  - A-learning loss function defined as:

$$\mathcal{F}(\theta, x_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} M\{Y_i, (A_i - \pi(x_i)) f(x_i), \theta\}$$

- θ set all trainable parameters of network
- $\pi(X) = P(A = 1|X)$ is propensity score = probability of receiving treatment given covariates X. In 1:1 randomized trial $= \pi(X) = 0.5$
- $M(u, v)$ is loss function that depends on outcome type. For continuous, it is squared error loss $M(u, v) = (u - v)^2$; for binary, it is logistic loss $M(u, v) = u \log(1 + \exp(-v))$

Model Training / Evaluation

- **Training:** θ (including α FS parameters in encoder and W, b in decoder) are optimized by min. A-learning $\mathcal{F}$ (e.g., Adam optimizer)
- **Hyperparameter Tuning:** 10-fold CV on training via grid search
- **Evaluation Metric:** AUC. $\hat{f}(X)$ to rank patients. Vary cutoff of score to generate ROC by comparing predicted vs. true treatment (known in the simulations), then calculate AUC

Biomarker Identification:

One of the key goals of DeepRAB is to facilitate **predictive biomarker identification**. Thus after training $\rightarrow$ apply form of **model interpretability** analysis to determine which input features (biomarkers) *most strongly influence* model's prediction of high or low treatment effect $\hat{\tau}(x)$. e.g., methods:

- Gradient-based feature attribution
- Permutation feature importance
- SHAP values

Performance of mathematical framework is evaluated quantitatively using *simulated* and *real trial data*. Evaluation based on $DeepRAB$'s ability to identify patient subgroups with enhanced treatment responses $\rightarrow$ ranking by $\hat{\tau}(x)$ separates patients who $truly$ benefit from those who do not.

Mathematical Distinction

Prognostic Model (Covariates $\rightarrow$ Outcome): 'Prognostic' implies info about likely course of disease e.g., disease recurrence, progression, likelihood death. A **key attribute** of purely prognostic marker is its predictive value is *independent of the specific treatment being administered, that is, biomarker's ability to predict good/bad outcome is present in both treated and untreated*

- Y Outcome of interest (e.g., survival time, disease progression score)
- X Biomarker measurement (e.g., gene expression level, T_1 time)
- W Binary treatment indicator

$$\mathbb{E}[Y|X,W] = \beta_0 + \beta_x X + \beta_w W + \beta_{xw}(X \cdot W)$$

Conditions:

1. Must be associated with outcome: Coefficient for biomarker itself must be significant
$$\beta_X \neq 0$$
X provides information about Y even when $W = 0$
2. Effect must NOT depend on treatment: No significant interaction between biomarker and treatment
$$\beta_{XW} = 0$$
Effect of X on Y is same for both $W = 1$ and $W = 0$. Graphically, the lines representing the relationship between X and Y for the treated and control groups are **parallel**. Here, treatment effect β_W is constant $\forall x_i \in X$

- Models $P(Y|X)$
- *'Given x patient's biomarker, what is their likely prognosis?'*
- Biomarkers found: Prognostic. Such features predict outcome regardless of treatment

Predictive Model of Treatment Effect (Covariates $\rightarrow$ Treatment Effect):

- Models $\mathbb{E}[Y^{(1)} - Y^{(0)}|X = x]$
- *'Given patient's biomarkers, how much benefit will they get from $treatment$ vs. $control$?'*
- Biomarkers found: Predictive. Outcome prediction & Prediction of response difference to treatment. e.g., X biomarker might not have relationship with outcome in control but a strong relationship in treated

Summary

Subgroup identification and modeling treatment effect heterogeneity with predictive biomarker identification (feature selection method) being key component and outcome of process.

June $12^{th}/2025$

Variant Effect Predictors (VEPS)

Computational tools whose primary function is to assess potential functional impact of genetic variants (particularly **missense**) which *cause a change in AA sequence of a protein*. $\rightarrow$ crucial in addressing variants of unknown significance VUS.

Foundational Principles

Evolutionary Conservation, Sequence Homology, and Structural Information

- **SIFT (Sorting Intolerant From Tolerant:** Operates on principle that important AA will be conserved throughout evolution. It predicts whether AA substitution will impact protein function by analyzing conservation across multiple species.
- **PolyPhen-2 (Polymorphism Phenotyping** $v2$: Predictor takes multifaceted approach. Evaluates physicochemical differences between AA, position of substitution within protein's structure, proximity to functional domains on top of evolutionary conservation.
- **CADD (combined annotation dependent depletion):** Scores deleteriousness of variants by integrating multiple annotations. Trained by comparing a vast *set* of observed human variations (mostly neutral) against simulated mutations to learn how to distinguish between them. Such approach allows to score coding and non-coding variants

Modern AI-based VEPs use $transformers - based$ architecture which allows model to weigh importance of different parts of the sequence to **understand context**. Thus it aims to learn fundamental language and rules of protein structure and function *without being told* which variants are pathogenic.

Accurate proteome-wide missense variant effect prediction with AlphaMissense[3] [4]

[3] Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381(6664):eadg7492, 2023. DOI: 10.1126/science.adg7492. URL https://www.science.org/doi/abs/10.1126/science.adg7492

[4] J. Jumper, R. Evans, A. Pritzel, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873), 2021

α-Missense

By Google DeepMind. It predicts pathogenicity of missense variants. Architectures is inspired by $Evoformer$ block used in α-fold

Core idea is to iteratively refine 2 key representations:

- **MSA representation**: Captures evolutionary information
- **Pair representation:** Captures spatial and relational information between AA pairs

Input Representation:

- **Target Sequence:** Primary protein sequence of length L is typically $one - hot$ encoded into matrix $S_{target} \in \{0,1\}^{L\times 20}$, where 20 is AAs
- **MSA Representation (M):** N aligned sequences of length L represented a tensor $M \in \mathbb{R}^{N\times L\times d_{msa}}$. Each position (i, j) in tensor is an embedding for AA at residue j in sequence i of alignment
- **Pair Representation (P):** Tensor $P \in \mathbb{R}^{L\times L\times d_{pair}}$ is initialized to store information about pairs of residues (i, j). Can be initialized with info about relative positions of residues in the sequence, i.e., $j - i$.

Pair Representation: Tensor built and maintained by Evoformer. It stores and refine model's understanding of the relationship between every pair of residues in the protein.

Tensor P of dimensions $\mathrm{L} * \mathrm{L} * \mathrm{d_{pair}}$ where L length protein sequence and $\mathrm{d_{pair}}$ is # features the model stores for each pair. *Just like L x L matrix but instead of scalar in ij, a high dimensional vector of features describing relationship between ij*

Evoformer Block

The model consists of series of stacked Evoformer blocks. Each block takes M and P and outputs updated $\mathrm{M'}$ and $\mathrm{P'}$. Info is allowed to flow back and forth between MSA (evolutionary context) and implicit structural context (pair representation)

The Block:

MSA Representation Update (with Axial Attention MSA representation M is updated using axial attention which is applied independently along the rows and columns.

Row-wise Attention (within each sequence): For each sequence i in MSA i.e. each row of M, standard self-attention mechanism is applied across residues $j = 1, ..., L \to$ Model learns relationship between different residues within the same sequence. For a single row:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where $Q, K, V =$ linear projections of row's residue embeddings, augmented with information from the pair representation P. MSA representation is updated to M_{row}

Column-wise Attention (across sequences): A second attention mechanism is applied to each column j of intermediate MSA M_{row} Model then learns which sequences in the alignment are most informative for x residue position. Capable of identifying highly conserved positions and co-evolving mutations across different sequences (signal for functional importance).

Communication: MSA to Pair Representation

Information from updated MSA representation is used to update pair representation. Crucial step where *evolutionary information informs spatial relationships*.

Some context:

Note that pair representation has been used in 2 places within the single block up to this point. This is the core cyclical logic of the Evoformer architecture $\to$ **Iterative refinement**(reasoning cycle). Model uses *current* pair representation P to *bias/guide* attention mechanism. Thus instead of e.g., $Scores = QK^T$, it becomes $Scores = QK^T + Bias_{ij}$. A large $Bias_{ij} =$ model is forced to focus on relationship between residues i and j when updating MSA representation.

- Update is often achieved using $\otimes$-like operation on columns of MSA representation (correlation matrix). For a pair of residues (i, j) model takes corresponding columns from MSA embedding, $M_{:,i}$ and $M_{:,j}$, and combines them. Simplified view of update for pair (i, j):

$$P'_{ij} = \text{LinearLayer}\left(\sum_{k=1}^{N}(W_1 M_{ki}) \otimes \Big(W_2 M_{kj}\Big)\right) \tag{2}$$

$\text{M}_{:,1}$ is AA at **residue position** i across all the different sequences in the alignment. Embedding captures evolutionary variation at that specific site in the protein, thus operation computes covariance matrix over MSA embeddings for each pair of residues **columns** $\text{M}_{:,\text{i}}$ **and** $M_{:,j}$ capturing co-evolutionary signals

Pair Representation Update: Tensor P undergoes own refinement using series of convolutional layers or axial attention layers or axial attention layers applied over $L * L$ grid. Analogous to refining distance so model is allowed to enforce geometric consistency rules like e.g.,triangle inequality, on the relationship between residues. P' is fed back into MSA update in the next Evoformer block.

Prediction Head and Pathogenicity Score

To make a prediction for a x missense variant (e.g., wiild type AA a_{wt} at i is replaced by a_{mut} a prediction head is used.

- Extract final embedding for residue i from target sequence's representation $h_i \in \mathbb{R}^{d_{model}}$
- During self-supervised pre-training, a classification head (e.g., linear layer followed by $softmax$) is used to predict probability distribution over all 20 AAs for x position

$$\text{Logits} = W_{pretrain} h_i + b_{pretrain} \tag{3}$$

$$P(\text{amino acid}_j|\text{context}) = \text{softmax}(\text{Logits})_j = \frac{e^{\text{logit}_j}}{\sum_{k=1}^{20} e^{\text{logit}_k}} \tag{4}$$

For final pathogenicity prediction, fine-tuning process trains simpler head. Head takese final representation h_i (which contains information about a_{wt} and a_{mut} context) and *projects it to a single scalar value*. This is a **binary classification task**.

$$\alpha = \sigma(W_{patho} h_i + b_{patho}) \tag{5}$$

α is final pathogenicity score

SIFT: PREDICTING AMINO ACID CHANGES THAT AFFECT PROTEIN FUNCTION [5]

---

[5] P. C. Ng and S. Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. DOI: 10.1093/nar/gkg509

SIFT Algorithm

Evolutionary conservation. It says that AA positions critical for a protein's structure/function will be conserved across homologous sequences from different species. Therefore, a substitution at a highly conserved position is likely to be deleterious, whereas a substitution at a highly variable position is more likely to be tolerated without significant functional consequence. SIFT quantifies this by calculating a score based on the probability of observing a particular A at a specific position, derived from a multiple sequence alignment (MSA).

Step 1: Multiple Sequence Alignment (MSA) Generation

Given a query protein sequence ,the first step is to gather set of homologous sequences from large protein database (e.g., Swiss-Prot/TrEMBL) using an algorithm like $PSI - BLAST$ (Position-Specific Iterated Basic Local Alignment Search Tool). Result is an MSA where related sequences are aligned, revealing *patterns of conservation and variation at each position.*

Step 2: Position-Specific Probability Matrix (PSPM) Calculation

Core calculation in SIFT. For each position i in the alignment, the algorithm computes a **probability distribution** over the 20 AAs

Let P_{ij} be probability of AA j occurring at position i. Probabilities are calculated from the **observed frequencies** of AAs at that position in the MSA. To handle sampling bias (e.g., over-representation of very similar sequences) and zero-frequency events (amino acids not seen at a position), a Bayesian approach using a Dirichlet mixture as prior probabilities is used

Simplified representation using pseudocounts:

$$P_{ij} = \frac{n_{ij} + b_j}{N_i + B} \tag{6}$$

- n_{ij} is weighted count of sequences in the MSA having AA j at position i. Sequence weights are used to down-weight redundant, highly similar sequences
- $N_i = \sum_{j=1}^{20} n_{ij}$ is total weighted count of sequences at position i
- b_j is the pseudocount for AA j. Derived from a prior probability distribution (substitution matrix e.g., BLOSUMXX)
- $B = \sum_{j=1}^{20} b_j$ is the total number of pseudocounts

The set of these probabilities for a given position i, $\{P_{i1}, P_{i2}, \ldots, P_{i,20}\}$, forms the Position-Specific Probability Matrix (a vector for position i)

and satisfies:

$$\sum_{j=1}^{20} P_{ij} = 1 \tag{7}$$

Step 3: SIFT Score Calculation and Classification

The SIFT score for a given substitution from the wild-type AA (aa_{wt}) to a mutant AA (aa_{mut}) at position i is the probability of observing that mutant AA at that position, as derived from the PSPM.

$$\text{SIFT Score}(i, aa_{wt} \to aa_{mut}) = P_{i,aa_{mut}} \tag{8}$$

This score is a *measure of tolerance*. A high score (high probability) indicates that the substitution is commonly observed in homologs and is therefore predicted to be tolerated. A low score indicates the substitution is rare and likely not tolerated (deleterious).

The final classification is made by applying a threshold, typically 0.05:

$$\text{Prediction} = \begin{cases} \textbf{Deleterious} & \text{if SIFT Score} < 0.05 \\ \textbf{Tolerated} & \text{if SIFT Score} \geq 0.05 \end{cases} \tag{9}$$

Step 4: Conservation Index

SIFT's predictions are more reliable for positions that are *highly conserved*. **To quantify**, a conservation index is calculated for each position i, which is derived from the information content or negative of Shannon's Entropy of the probability distribution P_i.

First, the Shannon's Entropy H_i for position i is calculated:

$$H_i = -\sum_{j=1}^{20} P_{ij} \log_2(P_{ij}) \tag{10}$$

The entropy H_i is a measure of uncertainty/variability at position i. It ranges from 0 (perfect conservation, only 1 AA is possible) to $\log_2(20)$ (all 20 AA equally likely).

The Conservation Index C_i is then the information content: difference between the maximum possible entropy and the observed entropy:

$$C_i = \log_2(20) - H_i \tag{11}$$

A high conservation index ($C_i \to \log_2(20)$) indicates low entropy and high conservation, making SIFT prediction at that site more reliable. Likewise $C_i \to 0$ indicates high variability and predictions at such sites are considered less certain.

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome [6]

Methods to improve statistical power of gene-level association tests by partitioning rare variants into K functional categories $(S_1, \ldots, S_K)$. Methods outperform standard tests that treat all variants equally especially when different functional categories have different magnitudes.

[6] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 02 2021. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083

Method 1: Omnibus SKAT (oSKAT)

Performs separate Sequence Kernel Association Test (SKAT) for each of K variant sets yielding K p-values $(p_1, \ldots, p_K)$. p-values are combined using *Simes' method* to produce single gene-level p-value (then adjust for correlation between tests)

Simes' p-value:

$$p_{\text{Simes}} = \min_{i=1,\ldots,K} \frac{K \cdot p_{(i)}}{i} \tag{12}$$

where $p_{(i)}$ is i-th smallest p-value.

Method 2: Functional SKAT (F-SKAT)

F-SKAT = unified variance component test within single mixed model. Overall genetic effect for an individual, γ_i:

$$\gamma_i = \sum_{k=1}^{K} \gamma_{ik}, \quad \text{where} \quad \gamma_{ik} = \sum_{j \in S_k} w_j \beta_j G_{ij} \tag{13}$$

Assumes variant effects β_j for each category k are random variables drawn from distribution with a category-specific variance component, τ_k:

$$\beta_j \sim N(0, \tau_k) \quad \text{for } j \in S_k \tag{14}$$

H_0 is 'all variance components are zero' = no genetic effect from any category:

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_K = 0 \tag{15}$$

The F-SKAT score test statistic is an optimal linear combination of the individual SKAT statistics (Q_k) for each category:

$$Q_F = \sum_{k=1}^{K} \lambda_k Q_k \tag{16}$$

This statistic follows mixture of χ^2 distributions from which p-value can be derived.

Simulations and real data analyses show F-SKAT is generally the most powerful and flexible approach.

June $13^{th}/2025$

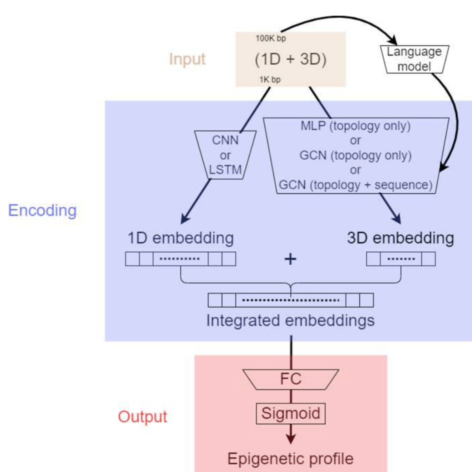
Multimodal learning of noncoding variant effects using genome sequence and chromatin structure [7]

▶ DL framework to predict the effects of noncoding genetic variants by integrating $1D$ local genome sequence and $3D$ global chromatin structure. Mathematical architecture overcomes challenges of multimodal data integration (e.g., significance difference in data resolution)

1D Sequence and Epigenetic data from DeepSEA (already processed). $5.2 * 10^6$ samples where each sample is $1kb$ DNA sequence. Each sequence is labeled with 919 binary values corresponding to epigenetic effects across 148 cell lines

3D Structure data was sourced from Hi-C experiments on the ENCODE portal which provides genome-wide chromatin interaction frequency matrices for cell lines GM12878, IMR90, and K562. Matrices reoersent 3D proximity of different genomic regions (100kb resolution).

Mathematical Framework



[7] Wuwei Tan and Yang Shen. Multimodal learning of noncoding variant effects using genome sequence and chromatin structure. *Bioinformatics*, 39 (9):btad541, 09 2023. ISSN 1367-4811. DOI: 10.1093/bioinformatics/btad541. URL https://doi.org/10.1093/bioinformatics/btad541

1D Sequence Encoding: CNNs and RNNs ((biLSTM)

For a $1D$ input sequence S and kernel K of size k, the output feature map C at position i:

$$C_i = f(\sum_{j=1}^{k} K_j * S_{i+j-1} + b)$$

f is non-linear

and,

$$h_{t-1} : h_t = \tanh\left(W_{hh}H_{t-1} + W_{xh}x_t + b_h\right)$$

3D Structure Encoding: GNNs

- GCNs capture complex, non-grid-like topology of chromatin interactions
- Feature vector (embedding) h_i^{l+1} for node i at layer $l+1$

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)})$$

where $H^{(l)}$ is matrix of node features at layer l, $\hat{D}$ is diagonal degree matrix of $\hat{A}$ and $\hat{A} = A + I$ (adjacency matrix).

Pre-trained DNA language model used:m **DNABERT**.

Training | Prediction

Loss: Binary cross-entropy. For single data sample:

$$L = -\sum_{i=1}^{919} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Regularization: $L1$ and $L2$

$$\text{L1 penalty: } \lambda_1 \sum |w|$$
$$\text{L2 penalty: } \lambda_2 \sum w^2$$

Results

Sequence-only models: DeepSEA, DanQ, Sei-based model.

Incorporating $3D$ chromatin structure into model outperforms sequence-only. GCN + DNABERT for structure embedding → Improved AUPRC by over 10%.

Performance Variant Effect Prediction:

- **eQTL pred.:** Structure-informed models as feature generators yielded better prediction of noncoding variant effects on gene expression vs. sequence-only

- **Pathogenicity pred.:** in *Unsupervised* setting, AUROC $\approx$ 0.75 and AUPRC $\approx$ 0.15. In *Supervised* setting (even with 100 labeled training samples) AUROC $>$ 0.8 and AUPRC $>$ 0.25

Findings

- 3*D* chromatin structure helps explain disparity between DNA sequence similarity and epigenetic profile similarity. DNA regions far apart in the 1D sequence *but have* similar epigenetic profiles tend to *be closer in 3D space*
- Models corrected bias of sequence-only predictors (where sequence similarity was poor indicator of epigenetic similarity)
- Using GCNs to embed 3D structure data was **key strategy**. Chromatin interaction modeled as graphs.
- Structure-informed models identified 2 motifs related to long-range interactions - POU3F1 and TFDP1) which were missed by DanQ BUT at the same time they missed 7 motifs that DanQ found because they lacked **long-range interaction patters**
- Little to no feature engineering. applicable to multiple types of mutations (insertions, deletions ...)
- DID NOT outperformed CADD in their OWN but when combined

▶ 7 motifs missed that were found by sequence-only DanQ likely because motifs lacked long-range interactions that Hi-C data captures. *Using higher-resolution chromatin structure data e.g., from **Micro-C** experiments, may help.*

Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology [8]

▶ Framework for standardizing interpretation and classification of genetic variants for Mendelian diseases.

Goal: To create robust, evidence-based system for consistent variant classification in a clinical context.

5-Tier Classification System

- **Pathogenic**

[8] S Richards, N Aziz, S Bale, D Bick, S Das, J Gastier-Foster, W W Grody, M Hegde, E Lyon, E Spector, K Voelkerding, H L Rehm, and ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424, 2015. DOI: 10.1038/gim.2015.30

- **Likely Pathogenic****
- **Uncertain Significance (VUS)**
- **Likely Benign****
- **Benign**

** Represent $> 90\%$

A Variant of Uncertain Significance (VUS) should not be used in clinical decision-making and efforts should be made to resolve its classification

Evidence Framework

For combining different types of evidence to arrive at x classification. Authors chose a rule-based system for combining evidence codes instead of numerical scoring system. Exact numerical relationships cannot be proved yet. Given that the impact of a piece of evidence is often context-dependent they would fail to capture context.

Pathogenic Evidence Strengths:

- **PVS1 (very strong)** Predicted Null variant (e.g., nonsense, frameshift) in a gene where loss of function is a known mechanism of
- **PS1-PS4 (strong)** Includes evidence like a variant being previously established pathogenic missense variant, confirmed de novo variant in patient with consistent phenotype or variant showing statistically significant increased prevalence in affected individuals vs. controls
- **PM1-PM6 (Moderate)** Includes evidence e.g., being located in a mutational hotspot, being absent from controls in population databases (e.g., ExAC, 1000 Genomes) or being detected *in trans* with another pathogenic variant for a recessive disorder
- **PP1-PP5 (Supporting)** Evidence like co-segregatuin with disease in a faimily, multiple lines of a computational evidence (tools like SIFT, PolyPhen-2) supporting a damaging effect or a patient's phenotype being highly specific for the gene

Benign Evidence Strengths

- **BA1(Stand-Alone)**: Allele frequency is $> 5\%$ in a large population database

- **BS1-BS4 (Strong)**: Includes evidence like having an allele frequency greater than expected for x disorder, being observed in a healthy adult for a fully penetrant childhood disease, well established functional studies showing no damaging effect.
- **Supporting**: Includes evidence like being a missense variant in a gene where only truncating variants are known to cause disease or multiple computational tools suggesting no impact

Classification Rules | 'Algorithm Classifier'

Table 1: Rules for Combining Criteria to Classify Sequence Variants

Pathogenic

1. 1 Very Strong (PVS1) **AND** one of the following:
   (a) $\geq$1 Strong (PS1–PS4)
   (b) $\geq$2 Moderate (PM1–PM6)
   (c) 1 Moderate (PM1–PM6) **AND** 1 Supporting (PP1–PP5)
   (d) $\geq$2 Supporting (PP1–PP5)
2. $\geq$2 Strong (PS1–PS4)
3. 1 Strong (PS1–PS4) **AND** one of the following:
   (a) $\geq$3 Moderate (PM1–PM6)
   (b) 2 Moderate (PM1–PM6) **AND** $\geq$2 Supporting (PP1–PP5)
   (c) 1 Moderate (PM1–PM6) **AND** $\geq$4 Supporting (PP1–PP5)

Likely Pathogenic

1. 1 Very Strong (PVS1) **AND** 1 Moderate (PM1–PM6)
2. 1 Strong (PS1–PS4) **AND** 1–2 Moderate (PM1–PM6)
3. 1 Strong (PS1–PS4) **AND** $\geq$2 Supporting (PP1–PP5)
4. $\geq$3 Moderate (PM1–PM6)
5. 2 Moderate (PM1–PM6) **AND** $\geq$2 Supporting (PP1–PP5)
6. 1 Moderate (PM1–PM6) **AND** $\geq$4 Supporting (PP1–PP5)

Benign

1. 1 Stand-Alone (BA1) **OR**
2. $\geq$2 Strong (BS1–BS4)

Likely Benign

1. 1 Strong (BS1–BS4) **AND** 1 Supporting (BP1–BP7) **OR**
2. $\geq$2 Supporting (BP1–BP7)

Variants should be classified as Uncertain Significance if other criteria are unmet or the criteria for benign and pathogenic are contradictory.

June 15^{th}/2025

► ACMG/AMP provides standardized system for variant classification but it is **not** a rigid algorithm. THe guidelines are built with flexibility and state that expert judgement is required.

► Exceptions to guidelines may be cases where context, gene-specific knowledge, whether quality of evidence allows a rule to be applied differently or not.

Guidelines instead of rigid point-based system

Authors state:

> 'that the assignment of specific points for each criterion implied a level of quantitative understanding...that is currently not supported scientifically and does not take into account the complexity of interpreting genetic evidence'

Entire framework is thus built on the premise of *allowing expert curation and judgement* to upgrade/downgrade/ignore *x* piece of evidence based on context.

Exceptions and Context-Dependent Rule [9]

- **Context-Dependent Evidence Strength:** The weight of several criteria can be adjusted based on the available data. For example:
  - The evidence for a variant co-segregating with disease in a family (**PP1**) can be upgraded from 'Supporting' to 'Moderate' or 'Strong' if data from multiple large families is available
  - Observing a variant *in trans* with a known pathogenic variant for a recessive disorder (**PM3**) can be upgraded from 'Moderate' to 'Strong' if this is observed multiple times
- **Gene-Specific Disease Mechanisms:** The applicability of certain rules depends entirely on the known biology of the gene in question.

[9] S Richards, N Aziz, S Bale, D Bick, S Das, J Gastier-Foster, W W Grody, M Hegde, E Lyon, E Spector, K Voelkerding, H L Rehm, and ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424, 2015. DOI: 10.1038/gim.2015.30

  - The **PVS1** (predicted null variant) criterion is considered 'Very Strong' pathogenic evidence, but only for genes where loss-of-function is a known disease mechanism. For many cardiovascular genes like *MYH7*, where missense variants are the primary cause of disease, a heterozygous null variant is not pathogenic, and PVS1 does not apply

- **Evolving Guidelines:** The framework is intended to evolve as new data becomes available.

  - The **PM2** criterion (variant is absent from controls in population databases) was originally weighted as 'Moderate'. However, subsequent analysis by the ClinGen consortium recommended downgrading its strength to 'Supporting' on its own, acknowledging that most very rare or novel variants are benign

June 16^{th}/2025

Cardiovascular Diseases and the Role of Missense Variants

Missense Variant

single nucleotide change in DNA resulting in another AA incorporated in x protein. Possible outcome $\rightarrow$ *benign* or *severly disruptive* which often leads to alteration of protein structure and thus function.

In cardiogenetics, missense variants are a common in mechanism of disease particularly for *inherited heart muscle and rhythm disorder*.

Key examples:

Hypertrophic Cardiomyopathy (HCM): Thickening of the heart muscle. **MYH7** amd **MYBPC3** are the most common involved genes which encode proteins of the sarcomere (heart's contractile unit). The variants often lead to an altered protein that gets incorporated into the sarcomere and disrupts its function often through a **Dominant-Negative Effect** where the abnormal protein interferes with the function of the normal protein from the other allele.

▶ As noted in ACMG/AMP, simple loss-of-function (null) variants in many of such genes are more likely to be benign. Therefore, presence of x faulty component (missense variant) in the machine contributes to its abnormal behaviour *not* the lack of such component.

Dilated Cardiomyopathy (DCM): Enlarged and weakened left ventricle. Missense variants in genes encoding proteins of the sarcomere (e.g., TTN), cytoskeleton, nuclear envelop are major the major cause *but* loss-of-function variants can also cause DCM. These variants can compromise the structural integrity and force-generating capacity of the heart muscle cells

Arrythmogenic Cardiomyopathy (ACM): Tissue is replaced by fatty and fibrous tissue which lead to arrhythmias. Often caused by missense variants in genes that encode **desmosomal** protein (e.g., PKP2, DSG2, DSP) which are essential for holding heart cells together. A single AA change can disrupt cell-cell junctions → cell death / disease.

Cardiac Channelopathies: Group of disorders caused by mutations in genes encoding cardiac ion channels that control the heart's electrical activity. Their primary mechanism is often missense variants.

Genetics of myocardial interstitial fibrosis in the human heart and association with disease [10]

[10] V Nauffal, P Di Achille, M D R Klarqvist, and et al. Genetics of myocardial interstitial fibrosis in the human heart and association with disease. *Nature Genetics*, 55:777–786, 2023. DOI: 10.1038/s41588-023-01371-5

Myocardial Interstitial Fibrosis:

Pathological scarring of heart muscle. Excessive deposition of extracellular matrix proteins (e.g., collagen) that lead to cardiac stiffness, impaired function, and ultimately, heart failure.

Monogenetic Causes:

These are typically rare variants with a large effect size

Sarcomeric Genes in Hypertrophic Cardiomyopathy (HCM) Most common genetic heart disease. A major cause are mutations that code for the sarcomere (contractile unit).

Frequently implicated genes:

- **MYH7 (Myosin Heavy Chain 7):** Encodes β-myosin heavy chain (motor protein for muscle contraction)
- **MYBPC3 (Myosin Binding Protein C, Cardiac):** Encodes protein that regulate the contraction and relaxation of the heart muscle

Other sarcomere-related genes are TNNT2, TNNI3. These mutations are thought to initiate a cascade that results in fibrosis. The presence of such mutations is strongly associated with a greater extent of myocardial fibrosis

Polygenic Risk: Cumulative Effect

For a large portion of the population the rick of developing myocardial fibrosis is not tied to a single faulty gene but rather to the combined influence of many common genetic variants. GWAS have identified loci associated with an increased risk of fribrosis. Such variants are linked to a variety of biological processes that when altered can promote a pro-fibrotic environment in the heart.

- **Glucose Transpose:** Variants in **SLC2A12** may alter energy metabolism within the heart
- **Iron Homeostasis: HFE** and **TMPRSS6** suggest a link between iron regulation and cardiac scarring
- **Oxidative Stress: CAMK2D** variant is involved in calcium signaling which is critical for cardiomyocyte function and survival
- **Tissue Repair and Remodelings: ADAMTSL1** and **VEGFC** play roles in how the extracellular matrix is maintained and repaired
- **Chromatin Remodeling: SMARCB1** variants have been linked to an increased risk of fibrosis. Decreased expression of it can lead to an exaggerated response to fibrotic stimuli.

► Regardless of the trigger (monogenic or polygenic), multiple signaling pathways are implicated in downstream process of fibrosis. Genetic variations tend to converge on such pathways → Amplification Fibrotic Response.

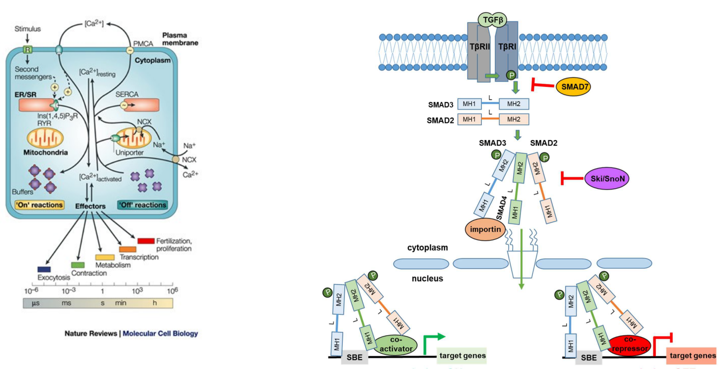


Figure 1: Signaling pathways: Calcium and TGF

Authors **goal** is to overcome limitation in understanding genetic basis of **myocardial interstitial fibrosis** using ML to quantify fibrosis from *cardiac magnetic resonance imaging* ***(CMRI)*** (UK biobank cohort) then identify novel genetic pathways in the disease.

U-Net with DenseNet-121 encoder

Symmetric encoder-decoder structure. Designed to capture context and localization

DenseNet-121 Encoder

Input image $\rightarrow$ hierarchical feature representations at different spatial scales.

- **Cmmposite Layer Function (H_l)**

  Let x_{l-1} output preceding layer: $x'_l = Conv(ReLU(BN(x_{l-1})))$

  - **Batch Normalization (BN)** normalizes input z, scales, shifts it. For input feature map z:

    $$BN(z) = \gamma(\frac{z - \mu_{batch}}{\sqrt{\sigma^2_{batch} + \epsilon}}) + \beta$$

    where μ and σ^2 pertain to mini-batch. $\beta$is shifting parameter, γ is learning rate

- **Dense Connectivity**

  Output of l is $x_l = H_l([x_0, x_1, ..., x_{l-1}])$ composite function H_l is applied to **concatenation** of feature maps along channel dimension (e.g., $k_0 + (l+1) * k$ channels for l)

- **Transition Layers**

  [Between dense blocks] Used to control complexity and *downsample spatial dimensions of feature maps*.

  - $1 * 1$ convolution to reduce number of feature maps
  - $2 * 2$ avg. pooling layer to halve height/width of feature map

$$x_{trans} = AvgPool(Conv_{1*1}(x))$$

U-Net Decoder and Skip Connections

Overall framework for segmentation. Input is output from encoder.

- **Encoder-Decoder Symmetry** The U-Net has 2 paths:

  - **Encoder:** DenseNet-121 **output** from different levels: e_1, e_2, e_3, e_4
  - The **Decoder** symmetrically reconstructs the spatial resolution to produce segmentation map

- **Upsampling (Transposed Convolution (or deconvolution)**

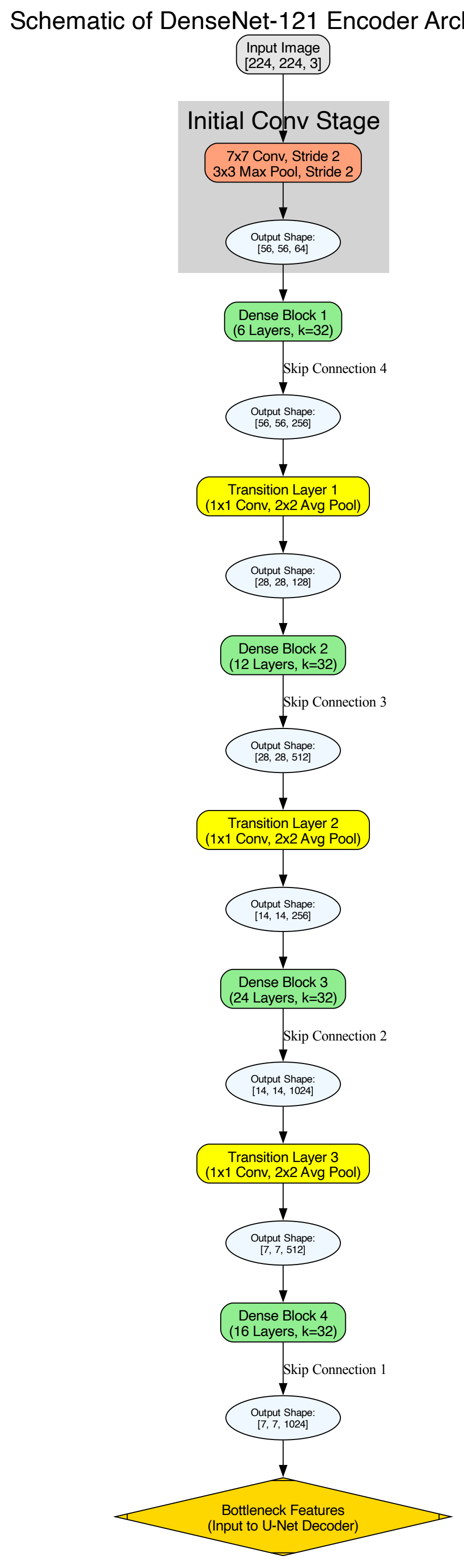
  let d_i be the feature map at level i in the decoder. The unsampled feature map u_i: $u_i = TransConv(d_i)$

$$L(p, g) = \alpha \cdot L_{\text{focal}} + (1 - \alpha) \cdot L_{\text{dice}} \tag{17}$$

$$
\begin{aligned}
L_{\text{focal}} &= -(1-p_t)^{\gamma}\log(p_t) && (18)\\
p_t &= \begin{cases} p & \text{if } g = 1 \\ 1-p & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
L_{\text{dice}} = 1 - \frac{2\sum_i (p_i \cdot g_i) + \epsilon}{\sum_i p_i^2 + \sum_i g_i^2 + \epsilon} \quad (19)
$$

Figure 2: Model structure.



June 17^{th} / 2025

ClinVar

[11] [12]

[11] M J Landrum, J M Lee, M Benson, G Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, J Hoover, W Jang, K Katz, M Ovetsky, G Riley, A Sethi, R Tully, R Villamarin-Salomon, W Rubinstein, and D R Maglott. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–8, 2016. DOI: 10.1093/nar/gkv1222

[12] M J Landrum, J M Lee, M Benson, G R Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, W Jang, K Karapetyan, K Katz, C Liu, Z Maddipatla, A Malheiro, K McDaniel, M Ovetsky, G Riley, G Zhou, J B Holmes, B L Kattman, and D R Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018. DOI: 10.1093/nar/gkx1153

Bibliography

Jun Cheng, Guido Novati, Joshua Pan, Clare Bycroft, Akvilė Žemgulytė, Taylor Applebaum, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli, and Žiga Avsec. Accurate proteome-wide missense variant effect prediction with alphamissense. *Science*, 381 (6664):eadg7492, 2023. DOI: 10.1126/science.adg7492. URL https://www.science.org/doi/abs/10.1126/science.adg7492.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 02 2021. ISSN 1367-4803. DOI: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

J. Jumper, R. Evans, A. Pritzel, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873), 2021.

M J Landrum, J M Lee, M Benson, G Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, J Hoover, W Jang, K Katz, M Ovetsky, G Riley, A Sethi, R Tully, R Villamarin-Salomon, W Rubinstein, and D R Maglott. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–8, 2016. DOI: 10.1093/nar/gkv1222.

M J Landrum, J M Lee, M Benson, G R Brown, C Chao, S Chitipiralla, B Gu, J Hart, D Hoffman, W Jang, K Karapetyan, K Katz, C Liu, Z Maddipatla, A Malheiro, K McDaniel, M Ovetsky, G Riley, G Zhou, J B Holmes, B L Kattman, and D R Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018. DOI: 10.1093/nar/gkx1153.

Zhen Liu, Yifan Gu, and Xiaoyang Huang. Deep learning-based ranking method for subgroup and predictive biomarker identification in patients. *Communications Medicine*, 5:221, 2025. DOI:

10.1038/s43856-025-00946-z. URL https://doi.org/10.1038/s43856-025-00946-z.

L Meng, R Attali, T Talmy, Y Regev, N Mizrahi, P Smirin-Yosef, L Vossaert, C Taborda, M Santana, I Machol, R Xiao, H Dai, C Eng, F Xia, and S Tzur. Evaluation of an automated genome interpretation model for rare disease routinely used in a clinical genetic laboratory. *Genet Med*, 25(6):100830, 2023. doi: 10.1016/j.gim.2023.100830.

V Nauffal, P Di Achille, M D R Klarqvist, and et al. Genetics of myocardial interstitial fibrosis in the human heart and association with disease. *Nature Genetics*, 55:777–786, 2023. doi: 10.1038/s41588-023-01371-5.

P. C. Ng and S. Henikoff. Sift: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003. doi: 10.1093/nar/gkg509.

S Richards, N Aziz, S Bale, D Bick, S Das, J Gastier-Foster, W W Grody, M Hegde, E Lyon, E Spector, K Voelkerding, H L Rehm, and ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5):405–424, 2015. doi: 10.1038/gim.2015.30.

Wuwei Tan and Yang Shen. Multimodal learning of noncoding variant effects using genome sequence and chromatin structure. *Bioinformatics*, 39(9):btad541, 09 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad541. URL https://doi.org/10.1093/bioinformatics/btad541.