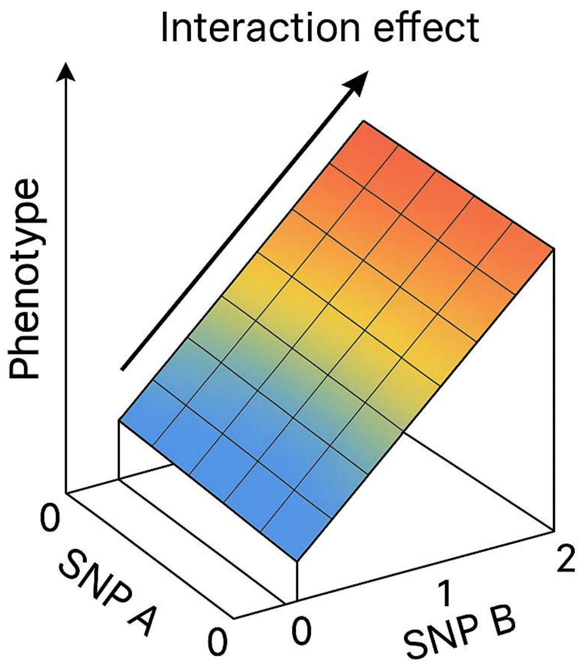
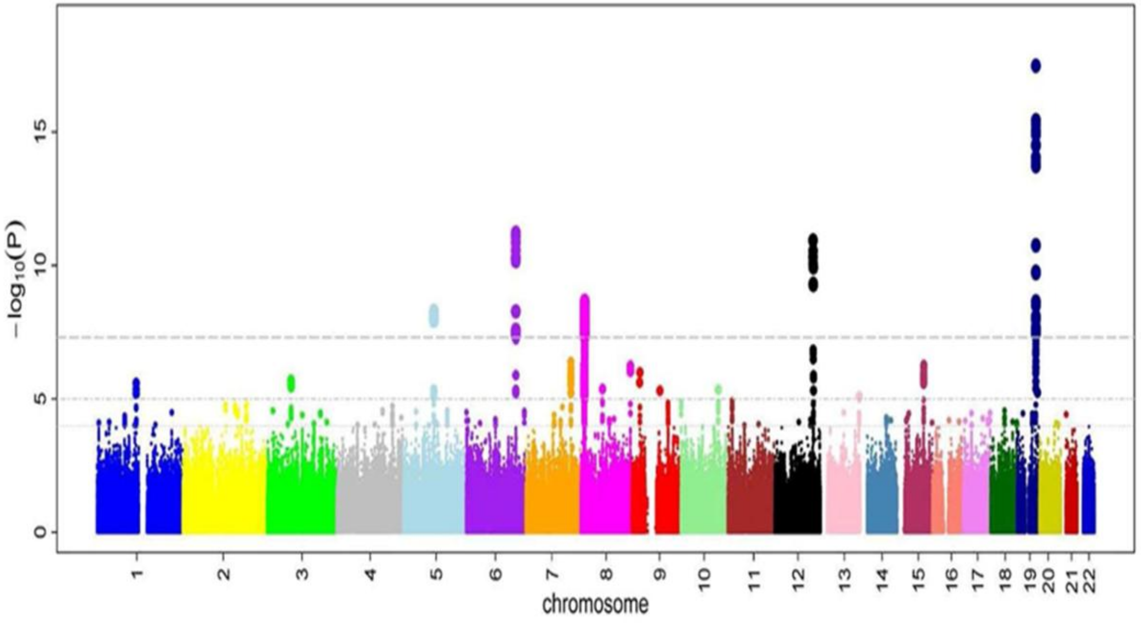
FIGHI: Fisher Information Guided Hyper-Interaction Inference

A scalable and interpretable framework for discovering high-order genomic interactions

Why FIGHI?

- Traditional GWAS misses interaction effects (epistasis)
- Traditional GWAS tests each SNP independently — one by one — for its association with a phenotype. That's fine for *additive* effects, but what if SNP A and SNP B only matter **together**?





The Challenge: The Combinatorial Explosion

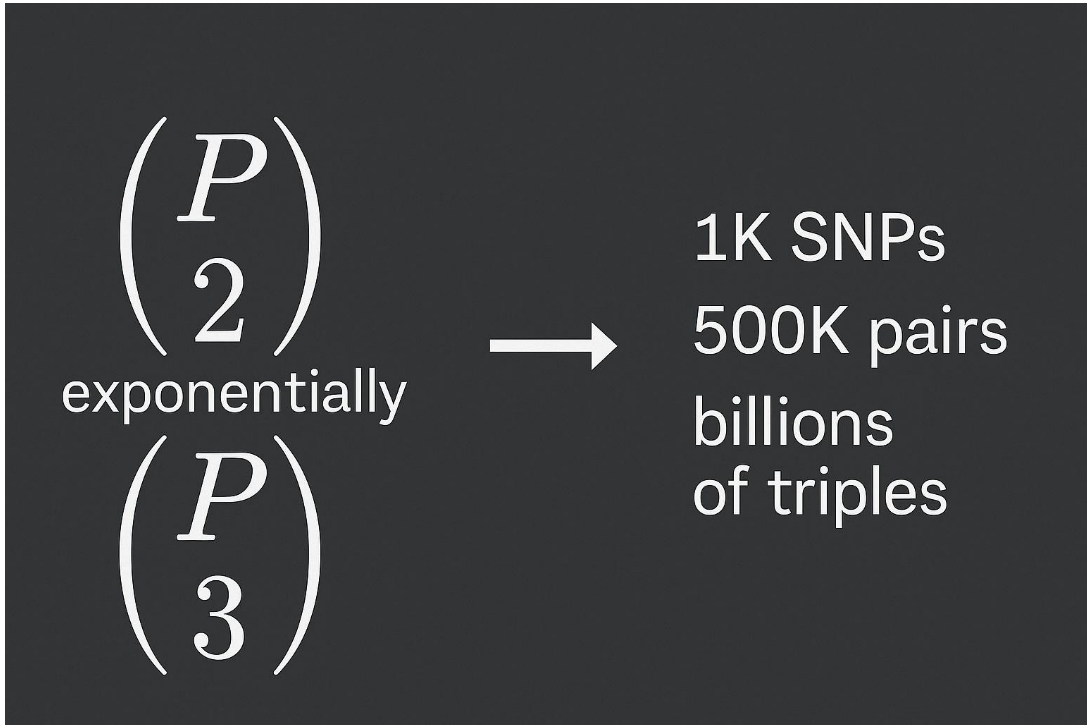
Problem:
P(y|X) may depend on interactions among SNPs.

But enumerating all possible interactions is **combinatorially explosive**:

- With 1M SNPs, even pairwise interactions = binom{10^6}{2} → impossible.
- We need a smarter way to decide *which* interactions are worth testing.

So, we reformulate the question as:

“Which combination of SNPs adds the most **Fisher Information** about the phenotype?”



$$\binom{P}{2}$$

exponentially

$$\binom{P}{3}$$

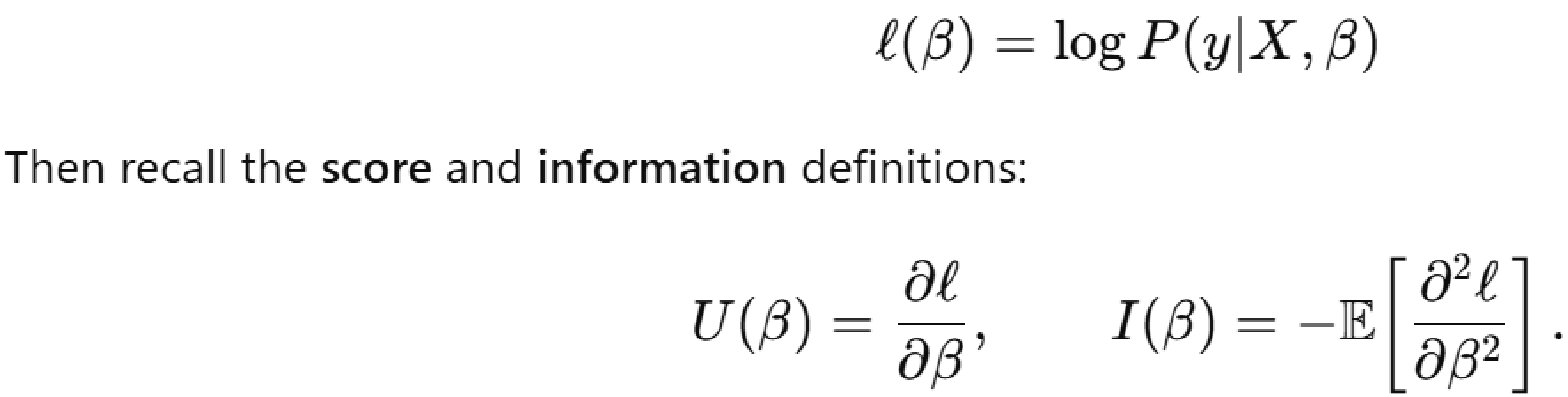
$\longrightarrow$

1K SNPs

500K pairs

billions of triples

From Likelihood to Fisher Information



$$\ell(\beta) = \log P(y|X, \beta)$$

Then recall the **score** and **information** definitions:

$$U(\beta) = \frac{\partial \ell}{\partial \beta}, \qquad I(\beta) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \beta^2}\right].$$

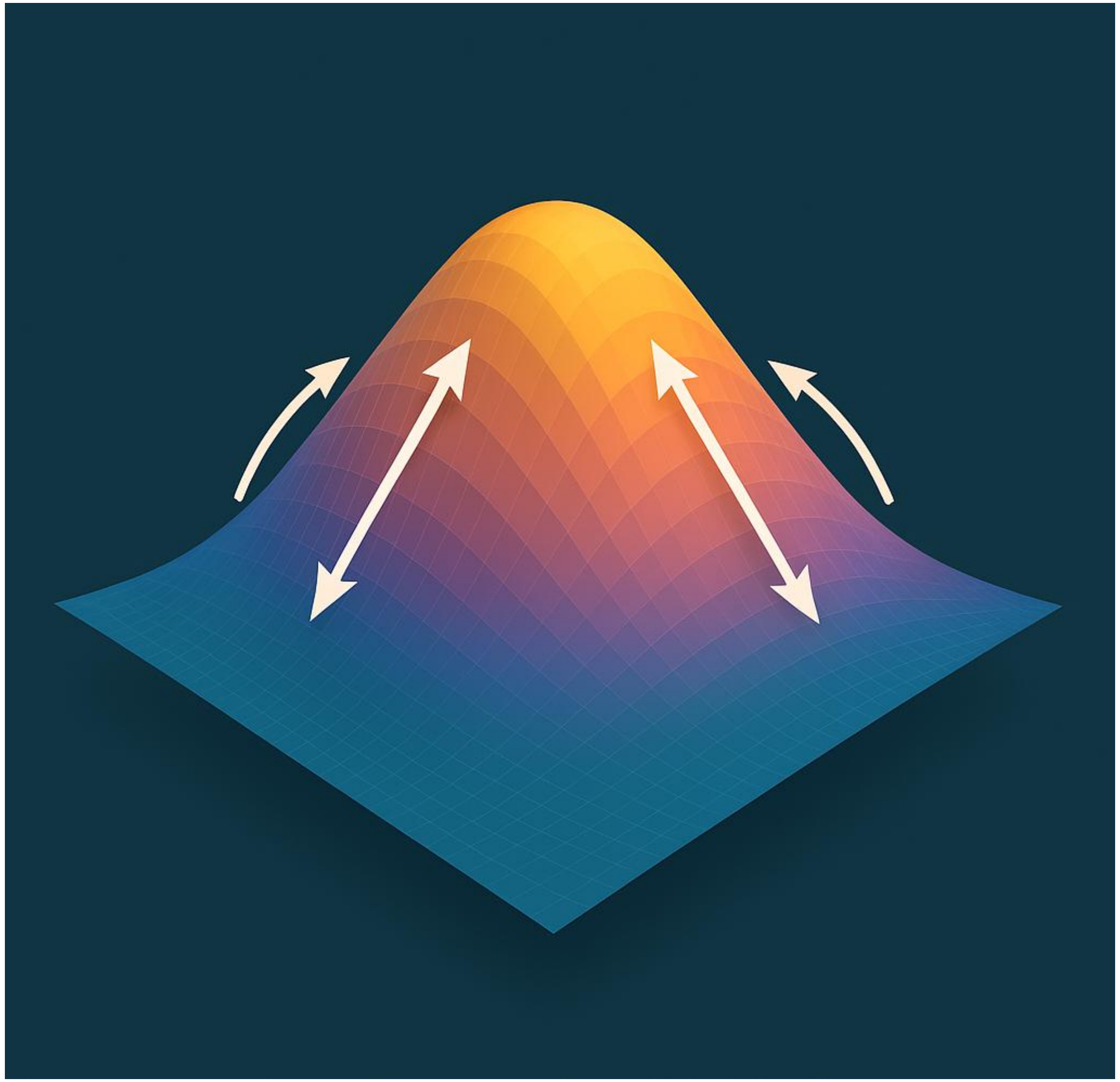
Interpretation:

- U(β) tells how the log-likelihood changes with β (slope).
- I(β) tells how *stable* that slope is; the curvature, or certainty.
- Large Fisher Information → steeper curvature → more certainty → stronger signal.

So, **Fisher Information quantifies how much certainty a parameter contributes.**

Intuitive Visualization

Geometric View of Fisher Information



Information Gain for a New Feature

Now suppose we’ve fit a base model with existing SNPs X, and we’re considering adding a new feature:

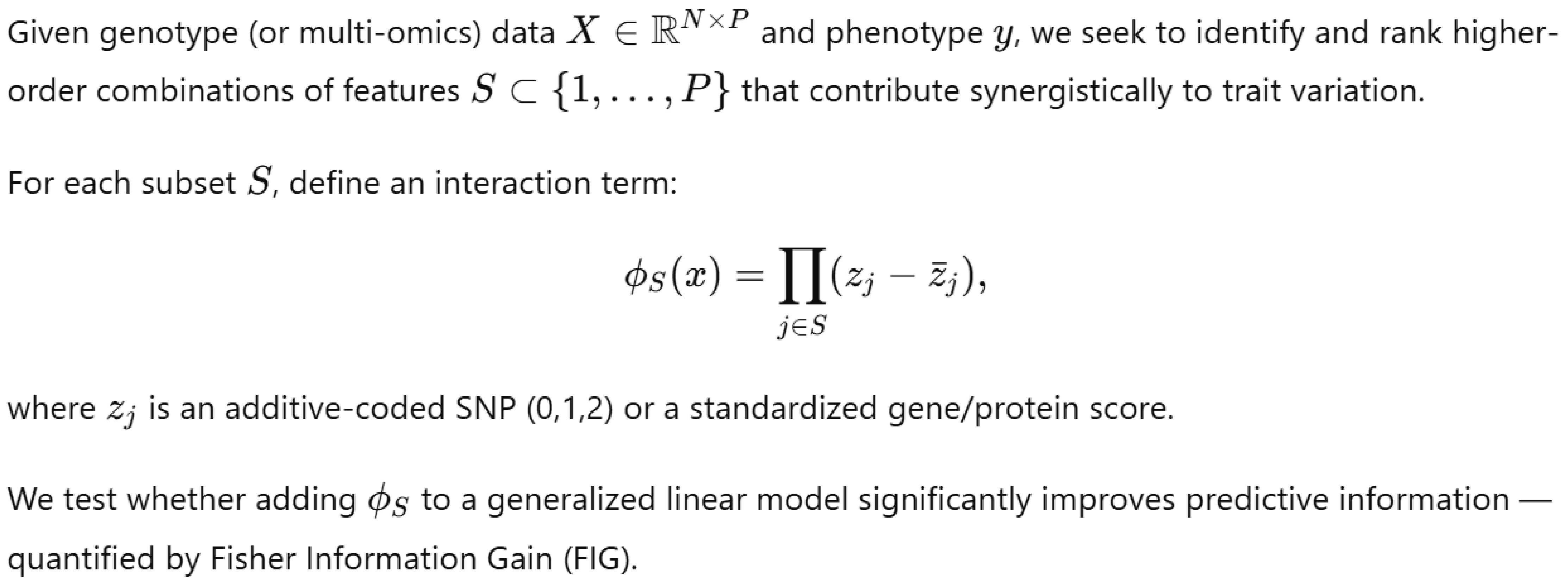
$$z = x_{j_1} \times x_{j_2} \times \cdots \times x_{j_K},$$

“Potential K-way interaction”

We don’t want to refit the whole model for every possible z. Instead, we use the **Score Test Approximation**.

Conceptual Overview

- Problem Statement



Given genotype (or multi-omics) data $X \in \mathbb{R}^{N \times P}$ and phenotype y, we seek to identify and rank higher-order combinations of features $S \subset \{1, \dots, P\}$ that contribute synergistically to trait variation.

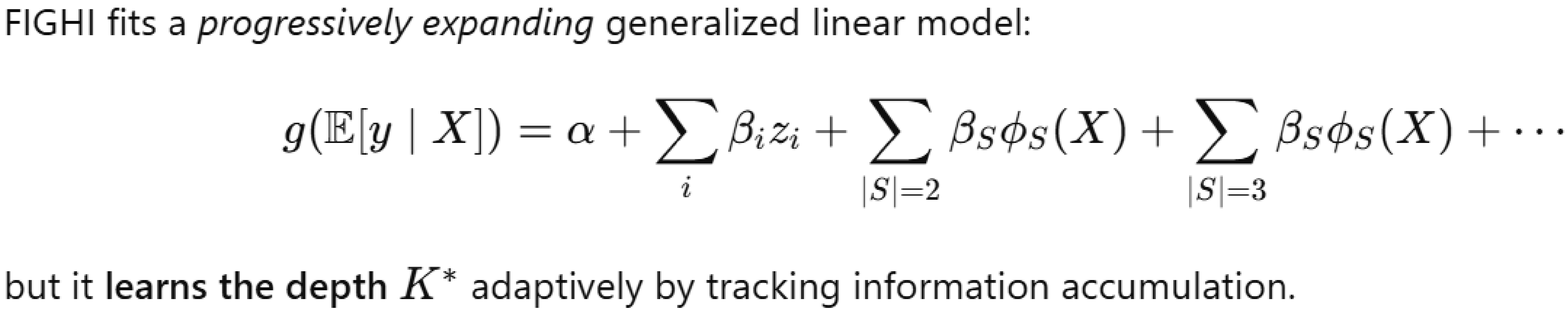
For each subset S, define an interaction term:

$$\phi_S(x) = \prod_{j \in S} (z_j - \bar{z}_j),$$

where z_j is an additive-coded SNP (0,1,2) or a standardized gene/protein score.

We test whether adding ϕ_S to a generalized linear model significantly improves predictive information — quantified by Fisher Information Gain (FIG).

FIGHI Model Structure

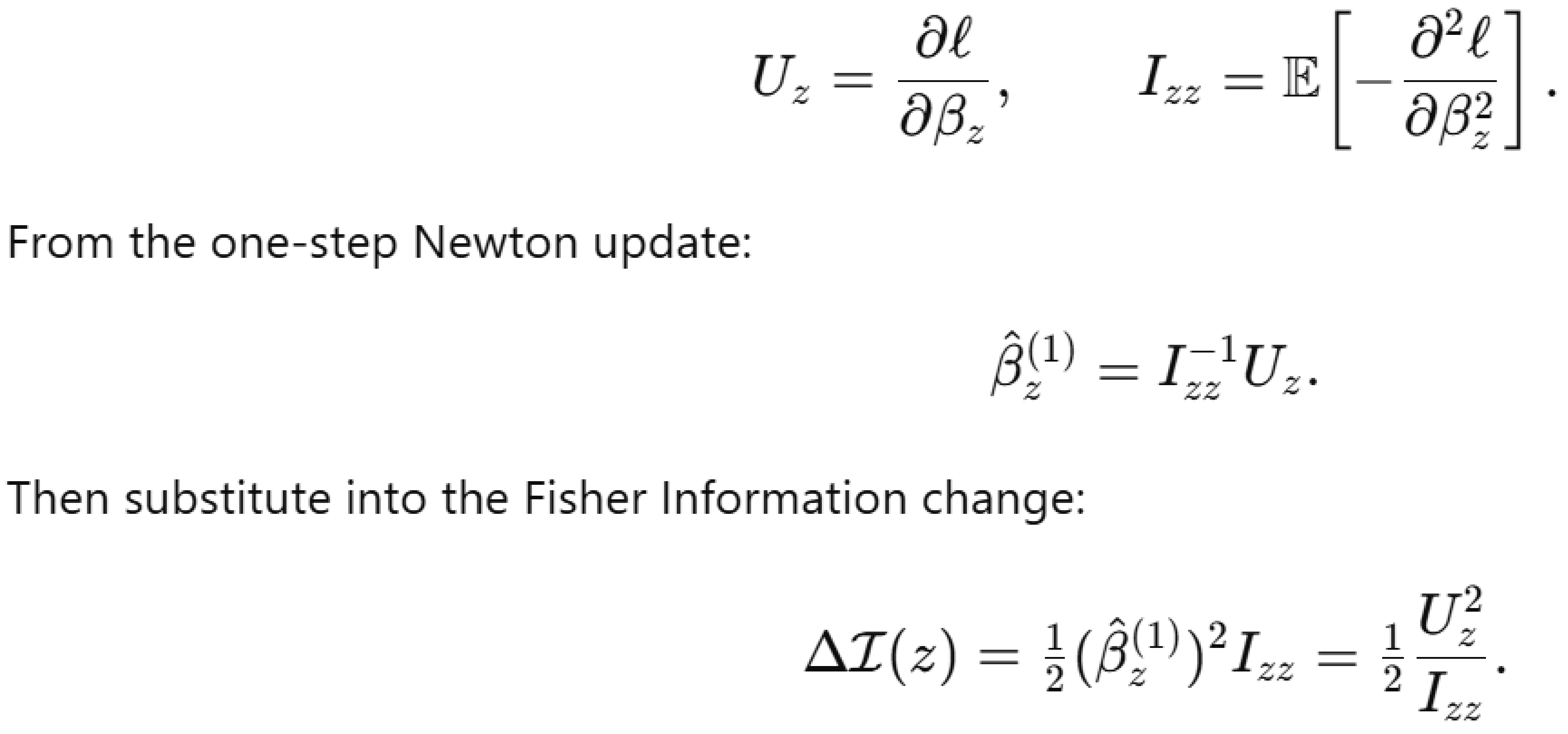


FIGHI fits a *progressively expanding* generalized linear model:

$$g(\mathbb{E}[y \mid X]) = \alpha + \sum_i \beta_i z_i + \sum_{|S|=2} \beta_S \phi_S(X) + \sum_{|S|=3} \beta_S \phi_S(X) + \cdots$$

but it **learns the depth** K^* adaptively by tracking information accumulation.

Score Test Derivation



$$U_z = \frac{\partial \ell}{\partial \beta_z}, \qquad I_{zz} = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \beta_z^2}\right].$$

From the one-step Newton update:

$$\hat{\beta}_z^{(1)} = I_{zz}^{-1} U_z.$$

Then substitute into the Fisher Information change:

$$\Delta\mathcal{I}(z) = \tfrac{1}{2}(\hat{\beta}_z^{(1)})^2 I_{zz} = \tfrac{1}{2}\frac{U_z^2}{I_{zz}}.$$

This is the **core equation of FIGHI**.

Score Test Derivation

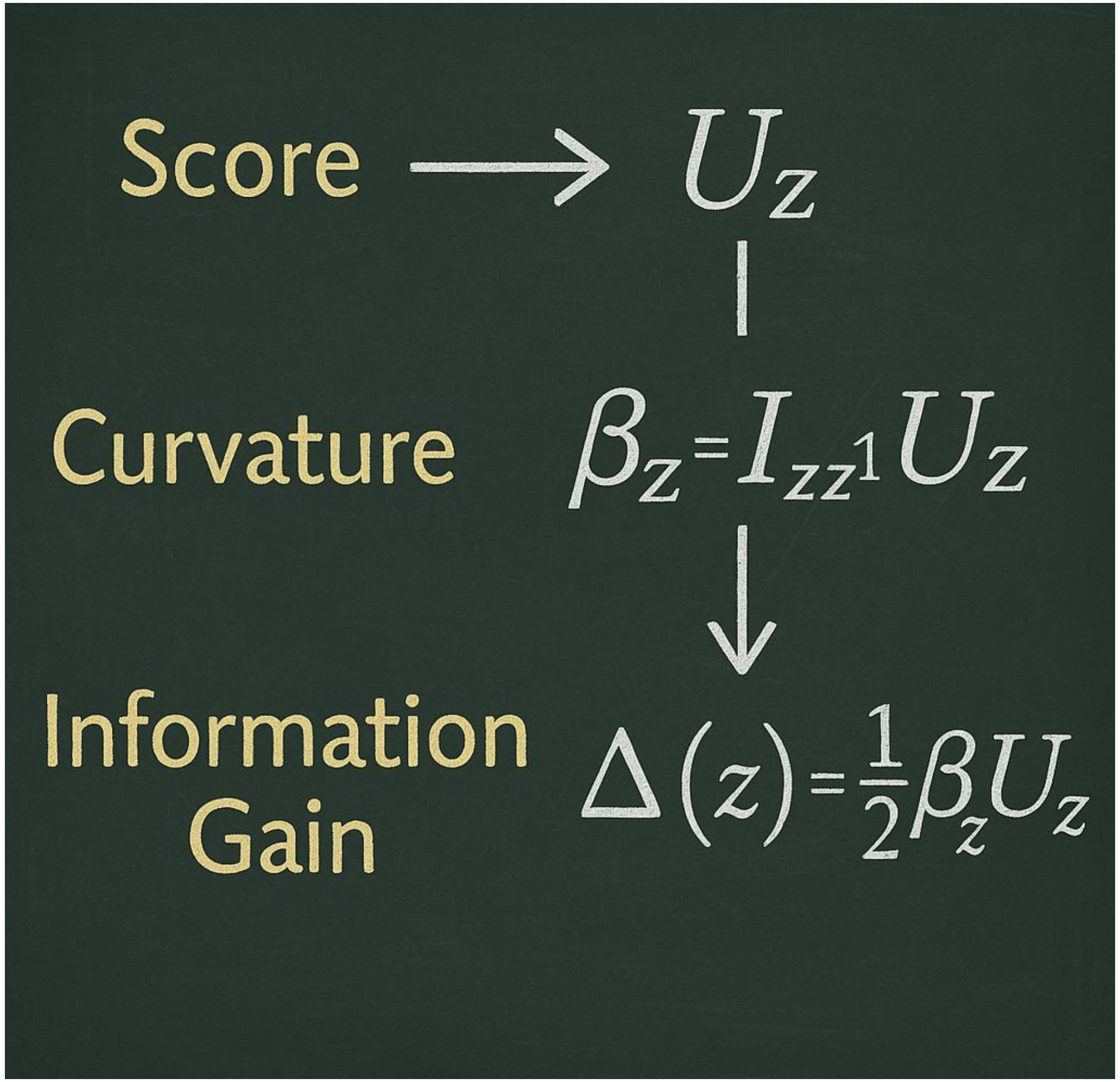
Interpretation

- Uz: how correlated the new interaction is with the residuals
- Izz: how stable (non-collinear) that interaction is
- ΔI(z) how much extra certainty this interaction adds

So, we can **rank all possible candidate interactions** by ΔI(z) without refitting the model for each.

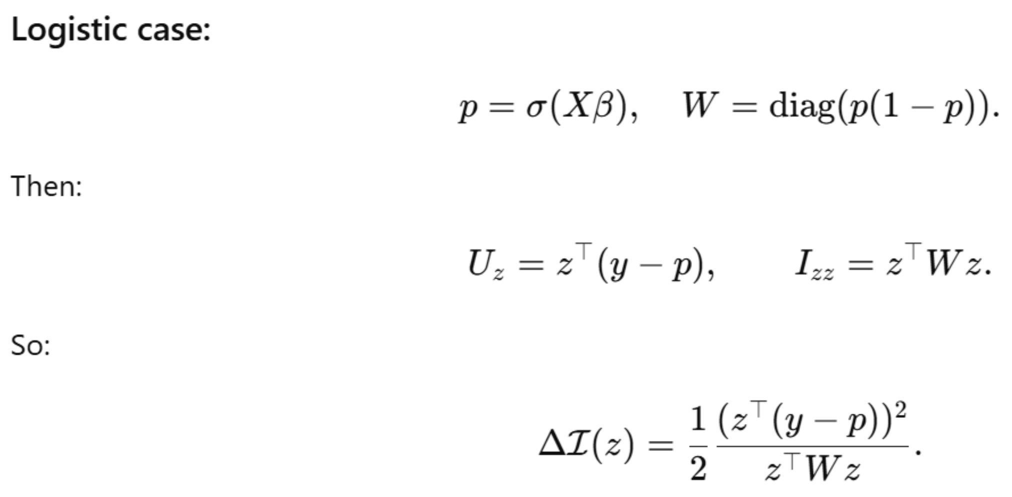
Mathematical Consistency

Score-Test Derivation & Computational Simplicity


Score → U_z
Curvature $\beta_z = I_{zz}^{1} U_z$
Information Gain $\Delta(z) = \frac{1}{2}\beta_z U_z$

Fisher Information Gain for Logistic and Linear Models

We can derive Uz and Izz explicitly.



Logistic case:

$$p = \sigma(X\beta), \quad W = \mathrm{diag}(p(1-p)).$$

Then:

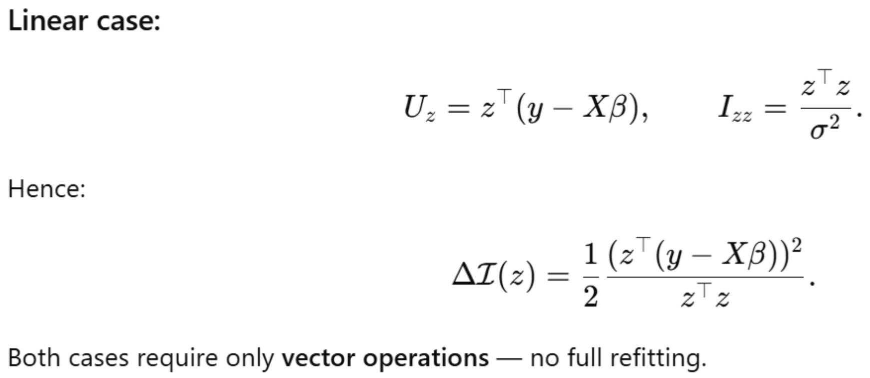
$$U_z = z^\top (y - p), \qquad I_{zz} = z^\top W z.$$

So:

$$\Delta \mathcal{I}(z) = \frac{1}{2} \frac{(z^\top (y-p))^2}{z^\top W z}.$$

Fisher Information Gain for Logistic and Linear Models

We can derive Uz and Izz explicitly.



Linear case:

$$U_z = z^\top (y - X\beta), \qquad I_{zz} = \frac{z^\top z}{\sigma^2}.$$

Hence:

$$\Delta\mathcal{I}(z) = \frac{1}{2}\frac{(z^\top (y - X\beta))^2}{z^\top z}.$$

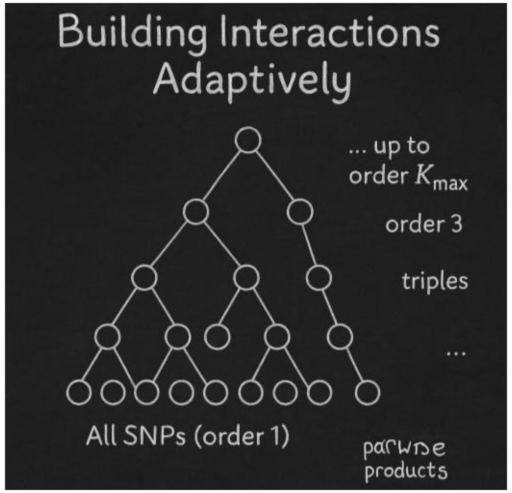
Both cases require only **vector operations** — no full refitting.

Building Interactions Adaptively

But we don't explore everything.
We keep only the interactions that **add significant Fisher Information**.

FIGHI **adapts its order depth** — if information saturates early, it stops at 2- or 3-way.

That's why even if we set max_order=4, you may only see 2-way interactions — because that's where Fisher Information stops growing.



Define:

$$r_K = \frac{\sum_{k=1}^{K} \Delta\mathcal{I}_k}{\sum_{k=1}^{K_{max}^{theor}} \Delta\mathcal{I}_k}.$$

When rK > threshold (say 0.95), stop.

Computational Implementation

Memory-efficient pipeline.

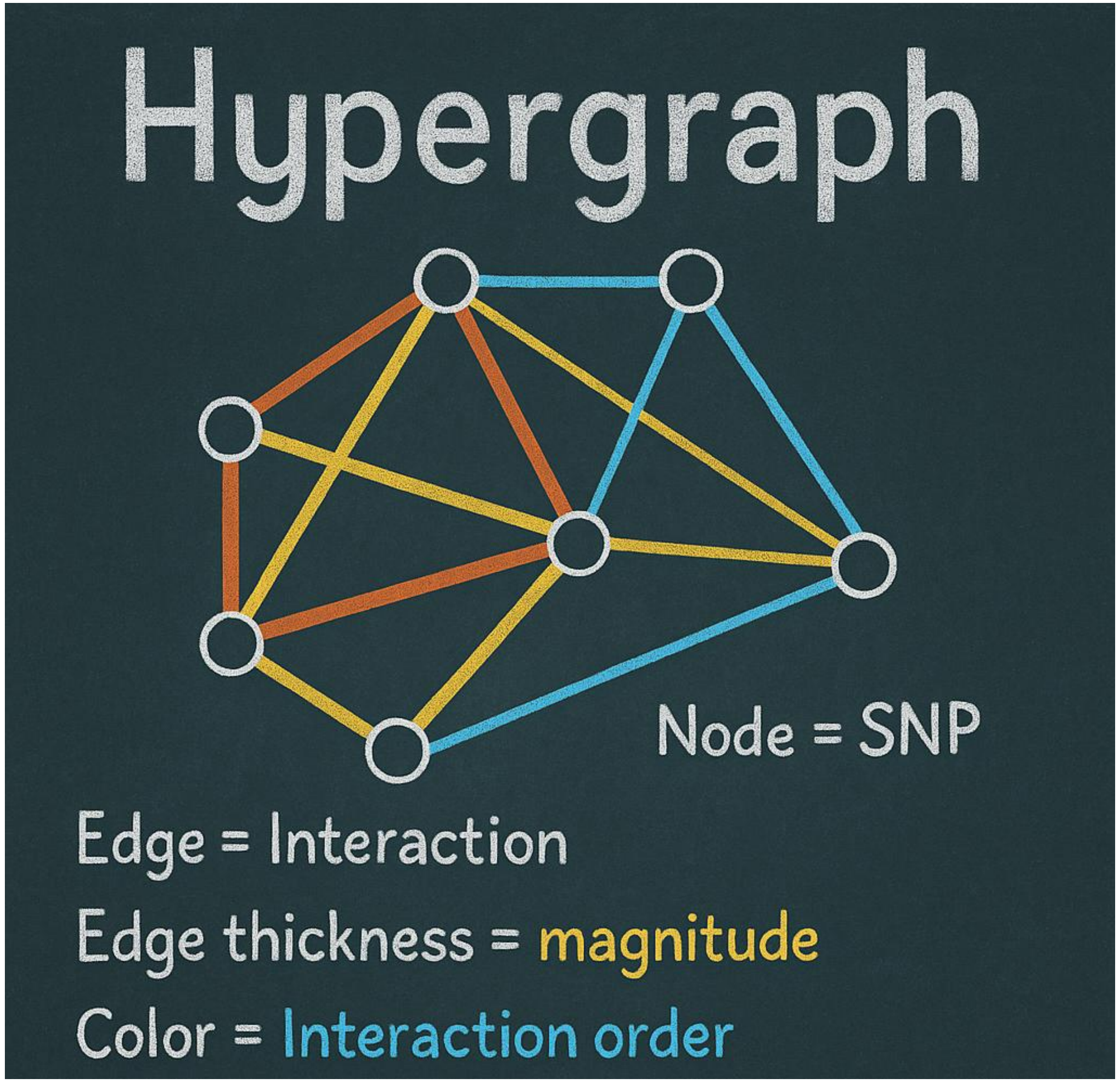
- Reads huge genotype CSVs in **chunks** (read_csv(chunksize=...))
- Uses **prescreening**: keep only top M SNPs by correlation with phenotype
- Streams data blockwise, computes $\Delta I(z)$ incrementally

Hypergraph Representation

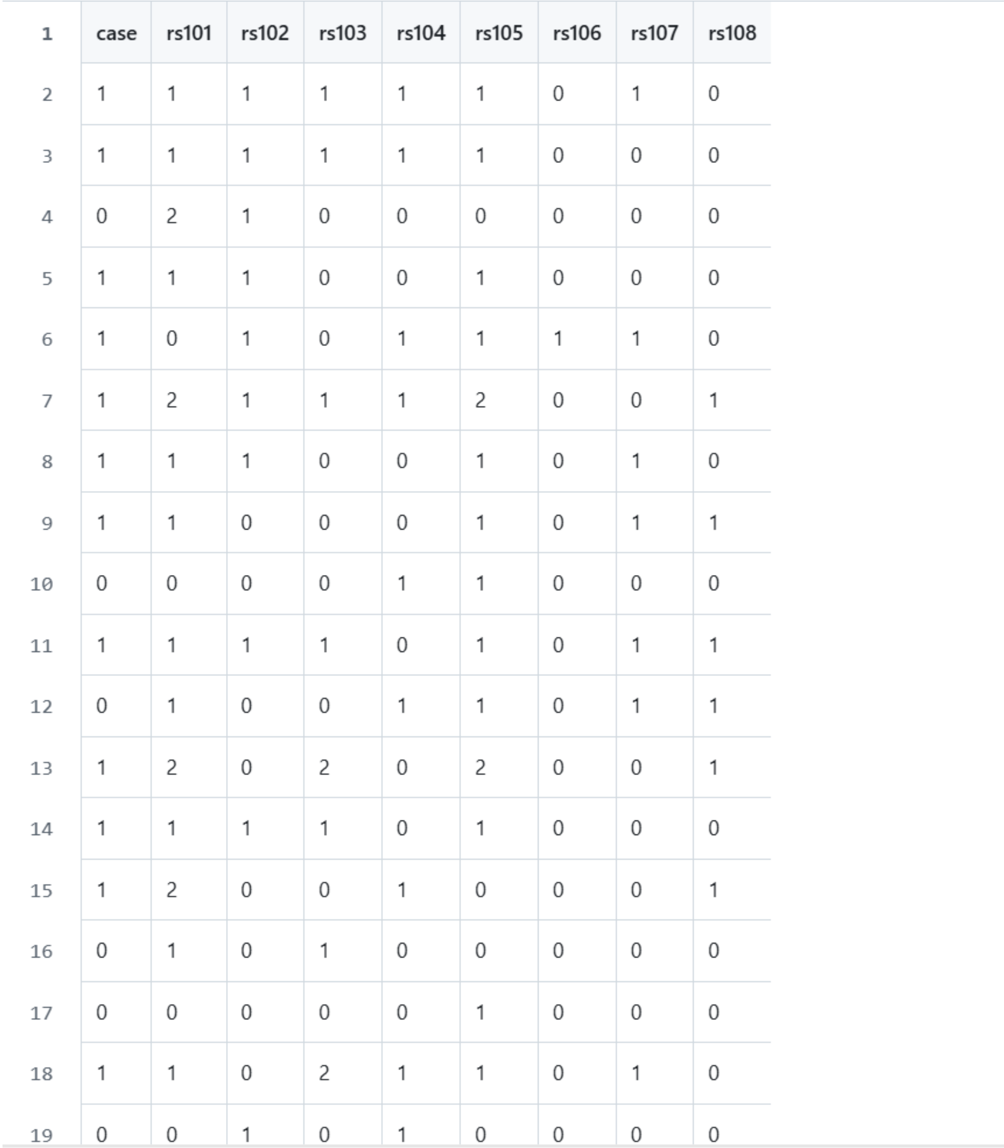
Say:

- Each SNP = node
- Each discovered interaction = hyperedge
- Edge weight = ΔI(e)

So, the **hypergraph encodes multi-level cooperation** among SNPs.

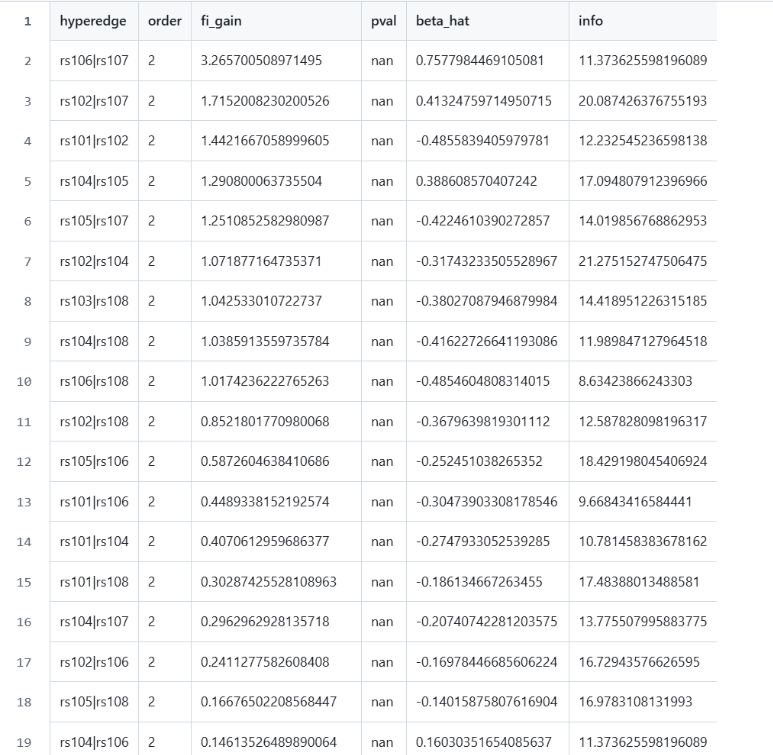


Simulation

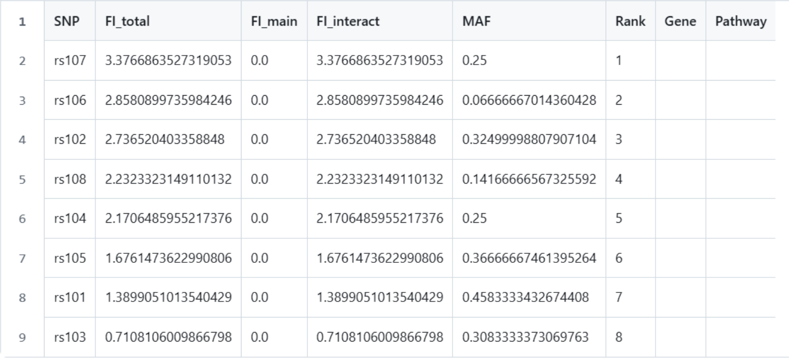


case	rs101	rs102	rs103	rs104	rs105	rs106	rs107	rs108
1	1	1	1	1	1	0	1	0
1	1	1	1	1	1	0	0	0
0	2	1	0	0	0	0	0	0
1	1	1	0	0	1	0	0	0
1	0	1	0	1	1	1	1	0
1	2	1	1	1	2	0	0	1
1	1	1	0	0	1	0	1	0
1	1	0	0	0	1	0	1	1
0	0	0	0	1	1	0	0	0
1	1	1	1	0	1	0	1	1
0	1	0	0	1	1	0	1	1
1	2	0	2	0	2	0	0	1
1	1	1	1	0	1	0	0	0
1	2	0	0	1	0	0	0	1
0	1	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0
1	1	0	2	1	1	0	1	0
0	0	1	0	1	0	0	0	0

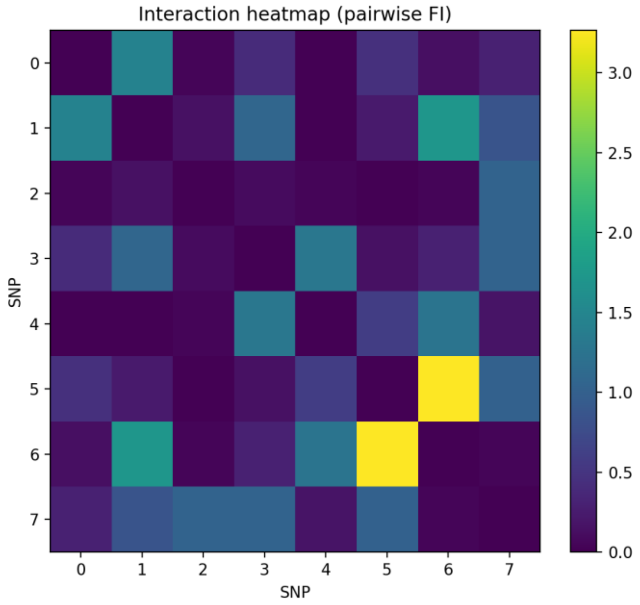
Simulation Example



1	hyperedge	order	fi_gain	pval	beta_hat	info
2	rs106|rs107	2	3.265700508971495	nan	0.7577984469105081	11.373625598196089
3	rs102|rs107	2	1.7152008230200526	nan	0.41324759714950715	20.087426376755193
4	rs101|rs102	2	1.4421667058999605	nan	-0.4855839405979781	12.232545236598138
5	rs104|rs105	2	1.290800063735504	nan	0.388608570407242	17.094807912396966
6	rs105|rs107	2	1.2510852582980987	nan	-0.4224610390272857	14.019856768862953
7	rs102|rs104	2	1.071877164735371	nan	-0.31743233505528967	21.275152747506475
8	rs103|rs108	2	1.042533010722737	nan	-0.38027087946879984	14.418951226315185
9	rs104|rs108	2	1.0385913559735784	nan	-0.41622726641193086	11.989847127964518
10	rs106|rs108	2	1.0174236222765263	nan	-0.4854604808314015	8.63423866243303
11	rs102|rs108	2	0.8521801770980068	nan	-0.3679639819301112	12.587828098196317
12	rs105|rs106	2	0.5872604638410686	nan	-0.252451038265352	18.429198045406924
13	rs101|rs106	2	0.4489338152192574	nan	-0.30473903308178546	9.66843416584441
14	rs101|rs104	2	0.4070612959686377	nan	-0.2747933052539285	10.781458383678162
15	rs101|rs108	2	0.30287425528108963	nan	-0.186134667263455	17.48388013488581
16	rs104|rs107	2	0.2962962928135718	nan	-0.20740742281203575	13.775507995883775
17	rs102|rs106	2	0.2411277582608408	nan	-0.16978446685606224	16.72943576626595
18	rs105|rs108	2	0.16676502208568447	nan	-0.14015875807616904	16.9783108131993
19	rs104|rs106	2	0.14613526489890064	nan	0.16030351654085637	11.373625598196089

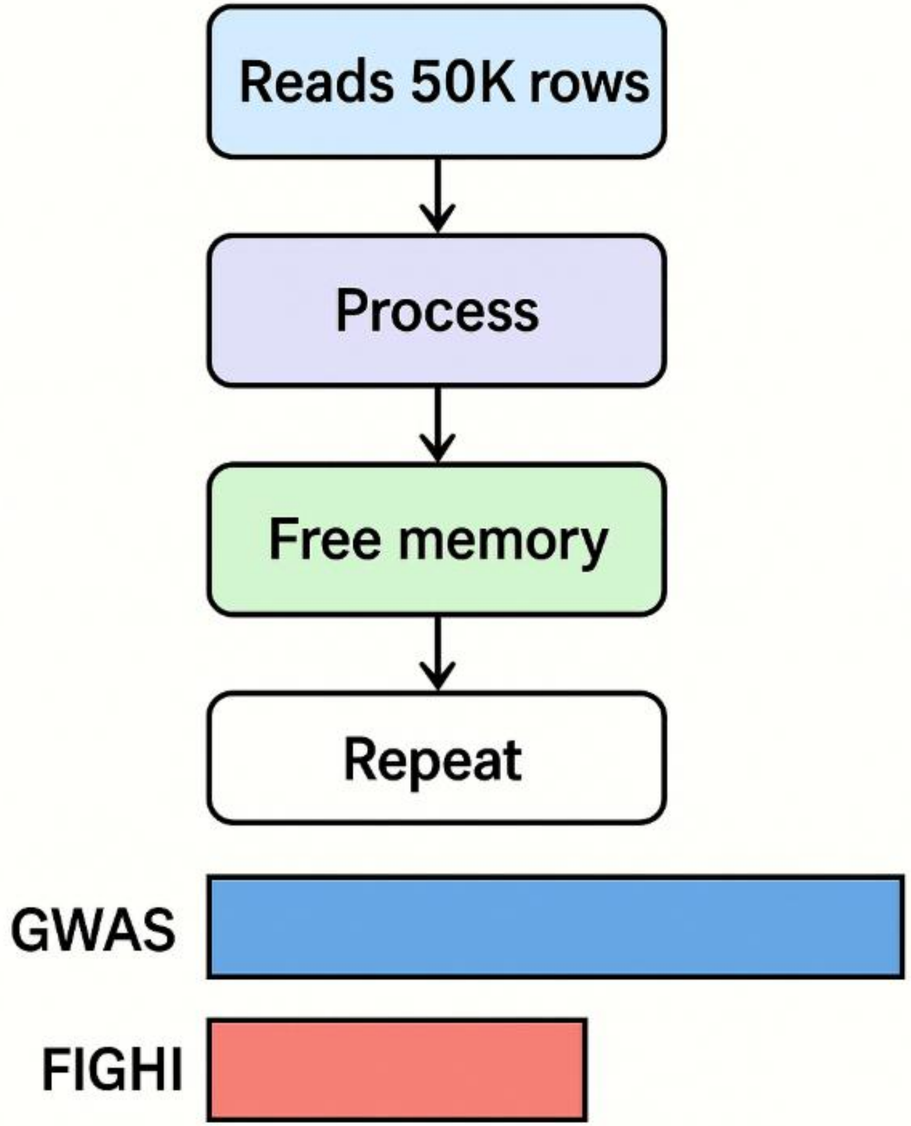


1	SNP	FI_total	FI_main	FI_interact	MAF	Rank	Gene	Pathway
2	rs107	3.3766863527319053	0.0	3.3766863527319053	0.25	1		
3	rs106	2.8580899735984246	0.0	2.8580899735984246	0.06666667014360428	2		
4	rs102	2.736520403358848	0.0	2.736520403358848	0.32499998807907104	3		
5	rs108	2.2323323149110132	0.0	2.2323323149110132	0.14166666567325592	4		
6	rs104	2.1706485955217376	0.0	2.1706485955217376	0.25	5		
7	rs105	1.6761473622990806	0.0	1.6761473622990806	0.36666667461395264	6		
8	rs101	1.3899051013540429	0.0	1.3899051013540429	0.4583333432674408	7		
9	rs103	0.7108106009866798	0.0	0.7108106009866798	0.3083333373069763	8		



Computational Efficiency

Memory-efficient pipeline.



Biological and Theoretical Takeaways

- **Biological** – Captures epistasis efficiently, beyond GWAS main effects.
- **Mathematical** – Based on score test and Fisher Information curvature.
- **Computational** – Scales to large datasets via streaming and pruning.

$$\boxed{\Delta\mathcal{I}(z) = \frac{1}{2}\frac{U_z^2}{I_{zz}}}$$

THANK YOU!

Full FIGHI documentation and code:

https://github.com/1234-Ariel-code/fighi