EXPLORING HOW LLM-GUIDED SCIENTIFIC DISCOVERY CAN FIND SOLUTIONS TO LONG-STANDING CHALLENGES IN DISEASE RISK PREDICTION FROM GENETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting a person's risk of disease from their genome works best for people whose genetic ancestry is well represented in the data used to build the model, and worst for everyone else. Because data from underrepresented ancestry groups are scarce, genetic risk models in precision medicine routinely *borrow* data from other, better-represented groups. How much to borrow, and from whom, is decided by data-borrowing schemes that are designed by hand and apply one global rule to every genetic variant and every person. We ask whether LLM-guided scientific discovery can find a better scheme. An LLM writes candidate schemes as short Python functions that decide, for each person and each variant, how far through continuous genetic ancestry space to borrow labeled subjects. A fixed evaluator scores each candidate on real prostate cancer data, and the best-scoring scheme is locked before any validation data are touched. With East Asian ancestry as the underrepresented (focal) group and discovery performed on one cohort, the locked scheme (shared as a Python function at `https://anonymous.4open.science/r/iclr26535-code--0DD3/prio_func_disc_runs/LockedDiscoveredFunctionFocalEastAsian.py`) achieved the highest held-out discrimination (AUC) in three independent validation settings drawn from three different data sources, compared with four established borrowing schemes (Mixture Learning, Independent Learning, and two transfer-learning schemes). Pooling all 112 held-out focal subjects, one-sided paired tests favored the discovered scheme over every comparator ($p < 0.05$ for all four). An ablation shows that weakening the evaluator during discovery degrades the held-out performance of the resulting scheme, indicating that the evaluator, not the LLM alone, drives what is found.

1 INTRODUCTION

Many common diseases are strongly heritable. For prostate cancer, the disease studied here, twin studies attribute roughly 57% of risk to inherited factors (Mucci et al., 2016). This makes it possible to predict, from a person's germline DNA, who is at elevated risk and should be screened earlier or more intensively. Such genetic risk tools are entering clinical practice: in the BARCODE1 study, screening directed by a genetic risk score detected clinically significant prostate cancers that conventional screening pathways would have missed (McHugh et al., 2025). BARCODE1, however, enrolled only men of self-reported European ancestry and used a score built from European-ancestry data. This restriction reflects a general and long-standing problem: most genetic data come from people of European ancestry (Popejoy & Fullerton, 2016), and genetic risk scores are markedly less accurate in other groups (Martin et al., 2019; Kachuri et al., 2024). For prostate cancer this is consequential, because incidence in Black men is about 67% higher, and mortality about twice as high, as in White men (Kratzer et al., 2025).

A short primer on the terminology. For readers new to genetics, we briefly explain the terms used throughout. *Cases* are subjects diagnosed with the disease (here, prostate cancer) and *controls*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

are subjects without it; a good risk score assigns higher scores to cases than to controls. A *variant* is a specific position in the genome at which the DNA letter differs among people, and the alternative versions present at that position are its *alleles*; because each person carries two copies of every chromosome (one from each parent), a person carries 0, 1, or 2 copies of a given allele, which is called the *dosage* of that allele. *Neighboring variants* are variants located close to one another on the same chromosome; they tend to be inherited together, so their alleles are correlated, and the strength of this correlation (called linkage disequilibrium) differs between populations. *Ancestry groups* are broad genetic-ancestral backgrounds, such as African, East Asian, European, or Hispanic/Latino ancestry, reflecting the populations from which a person's recent ancestors descend; they are convenient summaries of a continuum rather than sharp categories, which is why we also use continuous ancestry coordinates. An ancestry group is *underrepresented* when it contributes far fewer subjects to genetic studies than its share of the population would warrant; most large genetic datasets are dominated by subjects of European ancestry, so models built from them have seen few subjects from other groups.

The task and the data. We study binary disease-risk prediction from genetic data. Each *subject* (one person) has a case/control label (prostate cancer or not), a genotype, and a description of their genetic ancestry. The genotype is summarized at p genetic *variants* (positions in the genome where people differ): for each variant, the *dosage* is the number of copies (0, 1, or 2; fractional if imputed) of the allele associated with disease. Genetic ancestry is summarized by a 16-dimensional continuous *ancestry coordinate*: following standard practice, each subject is represented by the top 16 principal components of genome-wide genotype data (Ding et al., 2023; Tsuo et al., 2026), so that genetically similar subjects are close in this space. The model's job is to output a risk score for a new *target subject* (the subject whose risk is being predicted). We measure this with the area under the receiver operating characteristic curve (AUC).

Why data borrowing is necessary. The association between a variant and disease is not constant across ancestry groups: allele frequencies, correlations between neighboring variants, and the variant's effect on risk can all differ (Kachuri et al., 2024; Rebbeck, 2017; Harlemon et al., 2020). Across 269 established prostate cancer variants, effect estimates from European-ancestry subjects correlated only moderately with those from African-, East Asian-, and Hispanic-ancestry subjects ($r = 0.45, 0.37, 0.51$) (Conti et al., 2021). We call this change in the variant–disease relationship and in genotype distributions across ancestry an *ancestry-dependent distribution shift*. Under this shift, a model trained only on the abundant group is miscalibrated for others, yet a model trained only on the underrepresented group has too few subjects to estimate effects reliably. The practical response, used widely across precision medicine and precision oncology, is to *borrow* information from other groups: pooling multi-ancestry data (Conti et al., 2021; Márquez-Luna et al., 2017), coupling effect estimates across populations (Ruan et al., 2022), transfer learning from a large source group to a small target group in cancer omics and clinico-genomic prediction (Gao & Cui, 2020; 2024; Wu et al., 2026), and hierarchical borrowing across subgroups of subjects in oncology trials (Berry et al., 2013). Synthetic tasks also play an important part in AI model development (Das et al., 2025), and synthetic datasets have been created to accommodate ancestry-dependent distribution shift (Das & Cui, 2026). We refer to any rule that decides which labeled subjects inform the model for a target subject as a *data-borrowing scheme*.

The gap: existing schemes are heuristic and global. Existing schemes are designed by hand and apply the same borrowing rule to every variant and every target subject. Yet no single scheme dominates: pooling wins when effects are stable across groups and focal data are scarce, group-specific models win when effects differ and focal data are adequate, and transfer learning wins in between (Gao & Cui, 2020; 2024; Wu et al., 2026). Because the degree of cross-ancestry difference varies from variant to variant (Conti et al., 2021) and from person to person along the ancestry continuum (Ding et al., 2023), a global scheme is, by construction, one compromise imposed on a decision that is intrinsically local. The best rule would choose how far to borrow *per target subject and per variant*. The space of such rules is enormous, and no one has searched it.

LLM-guided scientific discovery. A recent line of work pairs a large language model (LLM) with an automatic evaluator: the LLM writes candidate programs, the evaluator scores them, and the best programs are fed back to the LLM to seed new candidates. This loop has found constructions

that surpass the best human-designed ones in combinatorics and algorithm design (Romera-Paredes et al., 2024; Novikov et al., 2025), and related systems have resolved open mathematical problems (Hubert et al., 2026; Tsoukalas et al., 2026; Howlett, 2026) and generated experimentally validated biomedical hypotheses (Gottweis et al., 2026; Ghareeb et al., 2026). The successes share three features: (i) the unknown can be written as a short executable program; (ii) any candidate can be scored automatically by a fixed, objective evaluator; and (iii) the space of plausible programs is far larger than the handful humans have written down. The choice of data-borrowing scheme satisfies all three.

Contributions.

- We cast the per-subject, per-variant data-borrowing decision as an executable program and define a fixed evaluator that scores any candidate scheme on real genetic data (Section 3).
- Using East Asian ancestry as the focal ancestry group, we run LLM-guided discovery on one prostate cancer cohort, lock the best scheme, and evaluate it on held-out subjects from three independent data sources never seen during discovery. The locked scheme attains the highest AUC in every setting (tied in one), and pooled paired tests favor it over all four established schemes (Section 4).
- An ablation that weakens the evaluator shows that held-out performance of the discovered scheme depends on evaluator strength (Section 4.3).

2 BACKGROUND

Established data-borrowing schemes. Suppose the *focal ancestry group* is the group whose predictive performance we want to improve and which is sparsely represented in the labeled data; the remaining groups are *non-focal*. Four schemes are standard. *Mixture Learning* pools all ancestry groups and fits one model. *Independent Learning* fits a model using only focal subjects. *Transfer learning* first learns from the abundant non-focal data and then adapts to the focal data (Gao & Cui, 2020; 2024). We use two transfer-learning schemes from Wu et al. (2026): *TL-GDES*, which starts from effect estimates learned on non-focal data and refines them on focal data by gradient descent with early stopping, and *TL-PR*, a penalized regression on focal data that shrinks the effect estimates toward those learned on non-focal data. Each scheme fixes one borrowing rule for the entire model: the same subjects (or the same degree of shrinkage) are used for every variant and every target subject, and focal membership is a discrete label.

Ancestry is continuous. Discrete ancestry labels hide substantial within-group variation. Ding et al. (2023) showed that the accuracy of genetic risk scores decays subject by subject with genetic distance from the training data, including within conventionally defined groups. This motivates describing each subject by a continuous ancestry coordinate and letting the scheme borrow from subjects that are *near* the target subject in that space, with a distance that may differ from variant to variant.

Program search with LLMs. Our search follows FunSearch (Romera-Paredes et al., 2024). Previously scored programs are stored in a database partitioned into *islands*, subpopulations that evolve independently to preserve diversity. A sampler selects two previously scored programs from one island and places them in the prompt together with their evaluator scores. Seeing a higher- and a lower-scoring program side by side, the LLM can infer in which direction the better program differs and extrapolate along that direction when writing its proposal. The LLM proposes a new program; the evaluator scores it; and the result is stored. Periodically, the weakest islands are reset and reseeded from strong programs in surviving islands. An attractive explanation from the LLM confers no advantage: only the evaluator's score determines what survives.

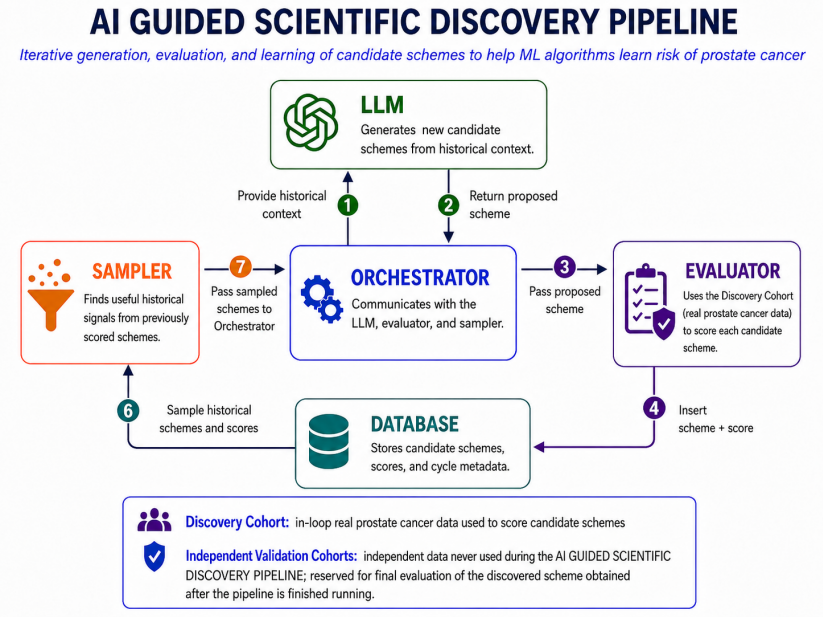


Figure 1: **Overview of LLM-guided discovery of a data-borrowing scheme.** The LLM proposes an executable candidate scheme (steps 1–2). The evaluator scores it using only the Discovery Cohort (step 3) and the scheme and its score are stored in the database (step 4). The sampler draws previously scored schemes, all generated earlier by the LLM itself, to provide context for the next proposal (steps 6–7). The Independent Validation Cohorts are never accessed by the evaluator or shown to the LLM; they are used only after the best scheme is locked.

3 METHODS

3.1 OVERVIEW

Figure 1 summarizes the approach. Two data roles are strictly separated. The *Discovery Cohort* is real prostate cancer data used inside the loop to score candidate schemes. The *Independent Validation Cohorts* are never passed to the evaluator, summarized for the LLM, or used to rank or select schemes; they enter only after the best scheme is locked, so the final evaluation tests generalization beyond all data that shaped the search.

Discovery. For one focal ancestry group, the LLM writes candidate schemes as short Python functions. Given the labeled reference data, the ancestry coordinate of a target subject, and the identity of one variant, a candidate scheme returns a radius in ancestry space. Reference subjects within that radius are used to estimate the variant's effect for that target subject. Repeating this over variants produces the target subject's genomic features, which a fixed logistic regression turns into a risk score. The Discovery Cohort is deliberately structured so that its reference portion is dominated by non-focal subjects while its evaluation portion is enriched for focal subjects. A candidate therefore scores well only if it exploits abundant non-focal data without losing accuracy for the focal group. After the search ends, the highest-scoring scheme is locked.

Validation. The locked scheme is applied without modification to Independent Validation Cohorts and compared with the established schemes of Section 2. All schemes share the same data partitions and the same downstream logistic regression; only the borrowing scheme differs.

3.2 INPUTS

Let n be the number of labeled reference subjects and p the number of variants. Reference subject k has case/control label $y_k \in \{0, 1\}$, ancestry coordinate $a_k \in \mathbb{R}^{16}$, and dosage vector $g_k =$

$(G_{k1}, \ldots, G_{kp})^\top$. The labeled reference data are

$$\mathcal{R} = \{(y_k, a_k, g_k)\}_{k=1}^{n}, \qquad y_k \in \{0,1\}, \quad a_k \in \mathbb{R}^{16}, \quad g_k = (G_{k1}, \ldots, G_{kp})^\top. \tag{1}$$

A target subject i supplies the same inputs except the label: an ancestry coordinate a_i and dosages $G_{i1}, \ldots, G_{ip}$. For each variant j, a descriptor v_j records its identifier, chromosome and position, effect allele, and the matching dosage field in $\mathcal{R}$.

3.3 The data-borrowing decision

A candidate scheme is a function f that maps the reference data, the target subject's ancestry coordinate, and a variant descriptor to a non-negative radius. The *neighborhood* $\mathcal{N}_{ij}$ is the set of reference subjects whose ancestry coordinates lie within that radius of the target subject:

$$r_{ij} = f(\mathcal{R}, a_i, v_j) \ \geq 0, \qquad \mathcal{N}_{ij} = \big\{k \in \{1, \ldots, n\} : \ \|a_k - a_i\|_2 \leq r_{ij}\big\}. \tag{2}$$

Equation (2) is the *only* decision the LLM-written scheme makes: for each target-subject–variant pair, how far to borrow through ancestry space. A small radius borrows only from genetically similar subjects (close to Independent Learning when the target is focal); a very large radius borrows from everyone (Mixture Learning). The scheme may use any property of $\mathcal{R}$ and v_j, for example the local density of subjects or the variability of the variant, to choose a different radius for each variant. It never sees the target subject's label and cannot alter anything downstream.

3.4 Local variant effects and genomic features

For a label $y \in \{0,1\}$ and a real-valued log-odds η, let the logistic loss be

$$\ell(y, \eta) = \log\big(1 + e^{\eta}\big) - y\,\eta, \tag{3}$$

the negative log-likelihood of y under $\Pr(y=1) = \sigma(\eta) = 1/(1+e^{-\eta})$. This is the standard binary cross-entropy $-y \log \sigma(\eta) - (1-y)\log(1-\sigma(\eta))$, written in terms of the log-odds η. For target subject i and variant j, we fit a one-variant logistic regression using only the reference subjects in the neighborhood $\mathcal{N}_{ij}$:

$$\big(\hat{\beta}_0,\, b_j(a_i)\big) = \arg\min_{\beta_0,\, b} \sum_{k \in \mathcal{N}_{ij}} \ell\big(y_k,\ \beta_0 + b\, G_{kj}\big). \tag{4}$$

Here β_0 is a neighborhood-specific intercept and $b_j(a_i)$ is the *local effect* of variant j at the ancestry location a_i: the change in log-odds of disease per additional copy of the effect allele, estimated from the subjects the scheme chose to borrow. Because the neighborhood depends on a_i, two target subjects with the same dosage but different ancestry coordinates can receive different local effects; this is how the model adapts to ancestry-dependent distribution shift. The fit is attempted only when $\mathcal{N}_{ij}$ contains at least 25 subjects, both cases and controls, and non-constant dosage at variant j; otherwise $b_j(a_i) = 0$, so that a variant without reliable local information contributes nothing. The *genomic feature* contributed by variant j is the target subject's dosage weighted by its local effect,

$$X_{ij} = G_{ij}\, b_j(a_i), \tag{5}$$

and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is the target subject's genomic feature vector.

3.5 Evaluator: scoring a candidate scheme on the Discovery Cohort

The Discovery Cohort consists of a *reference portion* (the labeled data $\mathcal{R}$ from which effects are borrowed) and an *evaluation portion* (subjects on which the candidate is judged). The evaluation portion is deterministically split into H disjoint *scoring folds*. For scoring fold h, all remaining evaluation subjects form the *calibration fold* $\mathcal{C}_h$; the scoring fold is $\mathcal{S}_h$. The two folds have distinct purposes:

- The *calibration fold* learns how to combine the genomic features into a single risk score. Genomic features for calibration subjects are built by Eqs. (2)–(5) using only the reference portion.

- The *scoring fold* measures how well that risk score ranks new subjects. Genomic features for scoring subjects are built using the reference portion together with the calibration fold. A scoring subject's label is never used to build its own features or to fit the risk model.

On the calibration fold we fit a logistic regression with intercept α and weights ω on the genomic features:

$$\eta_i(\alpha, \omega) = \alpha + \omega^\top X_i, \qquad (\hat{\alpha}_h, \hat{\omega}_h) = \arg\min_{\alpha,\,\omega} \sum_{i \in \mathcal{C}_h} \ell\big(y_i, \eta_i(\alpha, \omega)\big). \tag{6}$$

Each scoring subject then receives the frozen risk score

$$s_{i,h} = \hat{\alpha}_h + \hat{\omega}_h^\top X_i, \qquad i \in \mathcal{S}_h. \tag{7}$$

The fold score F_h is the median AUC of the scores $s_{i,h}$ over 200 bootstrap resamples of $\mathcal{S}_h$, which makes the evaluator less sensitive to a few subjects. The candidate's performance score is the average over folds,

$$\bar{F} = \frac{1}{H} \sum_{h=1}^{H} F_h. \tag{8}$$

Candidates are ranked in their island by $\bar{F}$ plus a small bonus (at most ± 0.008) that favors shorter programs, which breaks near-ties in favor of simpler schemes (Appendix A.6). Only $\bar{F}$ and fold-level summaries, never subject-level data, are returned to the search.

3.6 Search and locking

The search follows Section 2. Four islands run in parallel; in each cycle, every island requests seven candidate schemes from the LLM, and each prompt contains up to two previously scored schemes from that island. Accepted candidates are registered immediately so later prompts can build on them. At the end of each cycle, the weaker two islands are reset and reseeded from strong schemes in the surviving islands. The search runs for at most 12 cycles and stops after one cycle without improvement. The scheme with the highest ranking score at stopping is locked. No Independent Validation Cohort is examined to choose the scheme, the stopping time, or anything else.

3.7 Post-lock validation

Each Independent Validation Cohort is split into three disjoint portions: a *reference portion*, a *calibration portion*, and a *held-out portion* containing only focal subjects. Mirroring the evaluator, genomic features for calibration subjects use the reference portion; the risk model of Eq. (6) is fit on the calibration portion; genomic features for held-out subjects use the reference and calibration portions together; and the fitted coefficients are applied without refitting. The reported AUC is computed on the held-out portion only; in every Independent Validation Cohort, held-out subjects are always from the focal ancestry group, so the AUC measures performance for exactly the group the scheme was discovered to help. Every comparison scheme receives the identical reference, calibration, and held-out portions, variants, and downstream logistic regression.

Subject-paired test. AUC summarizes a whole cohort, and with few held-out subjects it is noisy. We therefore also compare schemes within each held-out subject. Let $\hat{p}_i^{(m)} = \sigma(s_i^{(m)})$ be the risk probability that scheme m assigns to held-out subject i, and let the *correct-label probability* be $C_i^{(m)} = \hat{p}_i^{(m)}$ if $y_i = 1$ and $1 - \hat{p}_i^{(m)}$ if $y_i = 0$. For comparator scheme b (one of the four established schemes of Section 2, which are in common use and were not discovered by our search), the paired difference and hypotheses are

$$D_i^{(b)} = C_i^{(\mathrm{LLM})} - C_i^{(b)}, \qquad H_0 : \mathbb{E}\big[D_i^{(b)}\big] = 0 \quad \text{vs.} \quad H_A : \mathbb{E}\big[D_i^{(b)}\big] > 0, \tag{9}$$

tested with a one-sided paired t-test on held-out focal subjects pooled across all Independent Validation Cohorts. We report exact p-values rather than only a significance label (Amrhein et al., 2019).

4 RESULTS

4.1 EXPERIMENTAL SETUP

Focal ancestry group and data sources. We chose East Asian ancestry as the focal ancestry group because it is severely underrepresented in the available labeled data while held-out subjects are available from three independent prostate cancer data sources. The group is labeled Japanese American in the Multiethnic Cohort (MEC) (Kolonel et al., 2000; Haiman et al., 2007), Asian in OncoArray (Conti et al., 2021), and East Asian in All of Us (AoU) (All of Us Research Program Genomics Investigators, 2024). The non-focal groups in MEC are African American and Latino subjects; in OncoArray and AoU they comprise all other ancestry groups. Every data source was converted to the inputs of Eq. (1); cohort sizes are listed in Appendix Table 2.

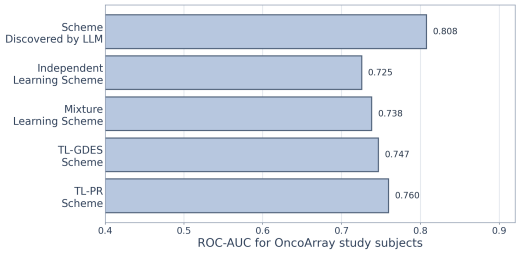
Discovery Cohort (MEC). The reference portion contained 1,922 subjects, of whom only 51 (2.7%) were focal. The evaluation portion contained 355 subjects, of whom 76 (21.4%) were focal. This enrichment is deliberate: the reference portion reproduces the scarcity faced by focal subjects, while the evaluation portion contains enough focal subjects for the evaluator to distinguish schemes that help them from schemes that do not. The LLM was GPT-5.5 (July 2026 version). The search ran two cycles; the second produced no improvement, triggering the stopping rule, and the best scheme was locked. We share the final locked discovered scheme at `https://anonymous.4open.science/r/iclr26535-code--0DD3/prio_func_disc_runs/LockedDiscoveredFunctionFocalEastAsian.py`.

Independent Validation Cohorts. (1) *Held-out MEC:* 19 focal subjects excluded from discovery entirely; all 2,277 labeled Discovery Cohort subjects (127 focal) served as reference and calibration portions. Held-out subjects share the parent study, so this setting tests holding out labels while keeping measurement conditions fixed. (2) *OncoArray:* 63 held-out focal subjects; 28,786 labeled subjects (566 focal) for the reference and calibration portions. (3) *AoU:* 30 held-out focal subjects (15 cases, 15 controls); 6,606 labeled subjects (46 focal) for the reference and calibration portions. The variants also differ by source. The MEC and OncoArray variant sets were each drawn from prostate cancer scores in the PGS Catalog (Lambert et al., 2021), excluding PGS Catalog scores built with data from any of the sources used in the present study. The AoU variant set was selected differently: the 50 variants most associated with prostate cancer in an earlier AoU data release, after removing strongly correlated variants; the held-out subjects joined AoU in a later release. Because a scheme operates on whatever variant it is given, the locked scheme was applied to these different variant sets without change.

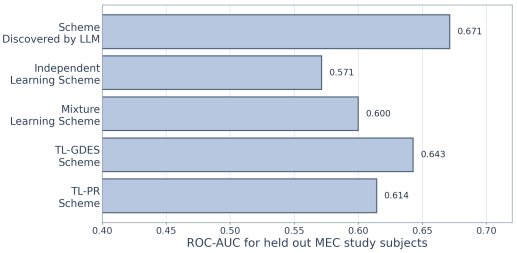
4.2 HELD-OUT PERFORMANCE

Figure 2 reports the main results. On OncoArray (Fig. 2a), a data source independent of discovery with the largest held-out portion, the discovered scheme achieved an AUC of 0.808, compared with 0.760 for the best established scheme (TL-PR) and 0.725–0.747 for the others. On held-out MEC (Fig. 2b), it achieved 0.671 versus 0.571–0.643. On AoU (Fig. 2c), it tied Mixture Learning at 0.733 and exceeded Independent Learning (0.707), TL-GDES (0.718), and TL-PR (0.700). No single established scheme was consistently second best: TL-PR ranked second on OncoArray, TL-GDES on MEC, and Mixture Learning (tied with the discovered scheme) on AoU. This is the behavior expected of global heuristics whose suitability depends on the setting, whereas the discovered scheme ranked first (or tied for first) in all three.

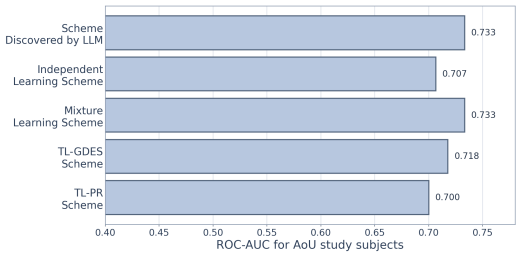
These three settings differ markedly in how scarce focal data are (from 46 to 566 labeled focal subjects), in data source, and in variant set, yet the scheme was locked before any of them was examined. Pooling the 112 held-out focal subjects, the subject-paired tests (Fig. 2d) favored the discovered scheme over every comparator: strongly over Independent Learning ($p = 0.0006$) and Mixture Learning ($p = 0.0001$), and more modestly over the two transfer-learning schemes ($p = 0.045$ for TL-GDES, $p = 0.049$ for TL-PR), which are themselves designed for this problem.



(a) OncoArray (63 held-out focal subjects).



(b) Held-out MEC (19 held-out focal subjects).



(c) AoU (30 held-out focal subjects).

Comparator scheme	One-sided p-value
Independent Learning	0.000591
Mixture Learning	0.000144
TL-GDES	0.044689
TL-PR	0.049001

(d) Subject-paired tests, Eq. (9), pooled over all 112 held-out focal subjects.

Figure 2: **Held-out performance of the locked scheme on three Independent Validation Cohorts.** (a–c) AUC on held-out focal (East Asian) subjects whose labels were never used during discovery, for the scheme discovered by the LLM and four established schemes; all five share identical data partitions and the same downstream logistic regression. (d) One-sided paired t-tests of H_A: the discovered scheme assigns higher correct-label probability than the comparator.

Table 1: **Discovery-Cohort ablation.** The evaluation portion of the Discovery Cohort was randomly undersampled (keeping all three MEC ancestry groups) while the reference portion was unchanged; discovery was rerun and the resulting scheme locked. Entries are AUC on the held-out focal subjects of each Independent Validation Cohort.

	Evaluation portion used during discovery		
	Full (main)	Moderate undersampling	Severe undersampling
Evaluation-portion subjects (focal)	355 (76)	200 (49)	100 (22)
Held-out MEC AUC ($n = 19$)	**0.671**	0.486	0.400
OncoArray AUC ($n = 63$)	**0.808**	0.765	0.746
AoU AUC ($n = 30$)	**0.733**	0.693	0.671

4.3 Ablation: the strength of the evaluator matters

If the LLM simply produced good schemes on its own, the evaluator's strength would not matter. To test this, we weakened the evaluator by randomly undersampling the evaluation portion of the Discovery Cohort from 355 subjects (76 focal) to 200 (49 focal) and 100 (22 focal), leaving the reference portion unchanged. Fewer scoring subjects make the evaluator noisier and less able to distinguish good candidates from lucky ones. Discovery was rerun from scratch at each level, and each resulting scheme was locked and evaluated on the same held-out subjects (Table 1).

Held-out AUC declined monotonically with evaluator strength on all three cohorts. The largest decline was on held-out MEC, the source that shares the discovery study; with only 19 held-out subjects, these AUCs are highly uncertain. On the two independent sources, the decline from the full to the severely undersampled evaluator was about 0.06 AUC, and even the scheme found with 22 focal scoring subjects reached AUC ≥ 0.67. Comparing with Figure 2, only the scheme found with the full evaluator matched or exceeded every established scheme on every cohort; with severe undersampling, the discovered scheme fell below at least one established scheme on each cohort. The quality of what is discovered is therefore driven by the evaluator.

5 DISCUSSION

Data-borrowing schemes are a long-standing, central tool for making genetic risk prediction work for underrepresented groups, and they have been designed by hand for decades. Our results suggest that this design step can itself be automated. By writing the borrowing decision as a short program that returns a per-subject, per-variant radius in continuous ancestry space, we turned a methodological question into a search problem with an objective evaluator, the setting in which LLM-guided discovery has succeeded elsewhere. The locked scheme outperformed hand-designed schemes, including transfer-learning schemes built for exactly this problem, on held-out subjects from three data sources, with different variant sets, none of which influenced the search.

Three design choices appear important. First, the separation between the Discovery Cohort and the Independent Validation Cohorts prevents the LLM from adapting, directly or indirectly, to validation data; without it, apparent gains could reflect selection on test data. Second, the Discovery Cohort was structured to create selection pressure: a reference portion with very few focal subjects and an evaluation portion enriched for them. The ablation (Table 1) shows that weakening this pressure yields schemes that transfer less well, supporting the view that the evaluator, not the LLM, is the scientific instrument, while the LLM supplies a proposal distribution far richer than the few schemes humans have written. Third, the output is inspectable code, and scoring a new subject requires no LLM.

The discovered scheme is not tied to the discovery variants. The discovered scheme is a program, not a vector of fitted coefficients. It acts on one variant at a time, using only the current reference data, the target subject's ancestry coordinate, and that variant's descriptor (Eq. (2)). The only requirement is that the reference data and the target subject both have dosages for the variant; whether the variant appeared during discovery is irrelevant. Consequently, the Independent Validation Cohorts did not need to share the Discovery Cohort's variants: each used its own variant set (drawn from the PGS Catalog for MEC and OncoArray, and from a separate association analysis for AoU), and the same locked scheme ran on all of them unchanged. In other words, the discovered Python function works on variants that it never encountered during the discovery process, which is what allows a scheme discovered once to be reused as new variants and new cohorts become available. Although we studied prostate cancer and one focal ancestry group, nothing in the formulation is specific to either: the same evaluator applies to any binary trait with genotype data and continuous ancestry coordinates.

6 LIMITATIONS

Several considerations define the scope of this study. The present work demonstrates the feasibility of LLM-guided discovery of adaptive data-borrowing strategies using genetic data. The number of held-out focal subjects was modest, reflecting the underrepresentation that motivates this work. Larger cohorts will allow more precise estimates of the size of the improvement. Further studies across additional diseases, genomic architectures, populations, and data sources will help establish the broader generalizability of the framework.

More generally, data-borrowing methods can improve how existing information is used under ancestry-dependent distribution shift, but they cannot substitute for continued expansion of genomic and phenotypic data across the genetic ancestry continuum. As larger and more heterogeneous genomic resources become available, they will provide opportunities to further evaluate and refine discovered borrowing strategies. Continuous ancestry coordinates provide a useful representation of genetic similarity, but they do not capture all sources of biological, environmental, or clinical heterogeneity that may influence disease risk and model transportability.

As with other automated discovery approaches, the solutions identified by the system are shaped by the program representation, search procedure, and evaluator used to guide discovery. Alternative formulations may therefore uncover additional effective strategies. Finally, although the discovery process involves stochastic LLM generation, the selected scheme is a fixed executable program that can be applied reproducibly once locked.

AI USE STATEMENT

(1) *Research execution:* an LLM (GPT-5.5) is a component of the method itself; it proposed the candidate data-borrowing schemes that were scored by the fixed evaluator (Section 3). Only evaluator scores determined which scheme was selected, and all validation was performed by the authors' fixed code after locking. (2) AI was used to generate code; all code was verified by the authors. (3) AI was used to draft the LaTeX source of this paper and to polish its writing. (4) Figure 1 was created with the help of an LLM. (5) LLMs were used to search for relevant literature in the scientific discovery space. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data were accessed through approved controlled-access mechanisms (dbGaP for MEC and OncoArray; the All of Us Researcher Workbench Controlled Tier), and all analyses were performed within the permitted environments. No subject-level genotype, phenotype, or ancestry data were sent to the LLM; candidate schemes were executed locally, and the LLM received only previously generated schemes and aggregate evaluator scores. Ancestry group labels are the descriptors used by each data source; we use continuous ancestry coordinates to avoid treating these labels as complete descriptions of genetic similarity. The work aims to reduce, not widen, disparities in the accuracy of genetic risk prediction.

REPRODUCIBILITY STATEMENT

Section 3 specifies the inputs, the scheme interface (Eq. (2)), the evaluator (Eqs. (4)–(8)), the search, and the post-lock validation. Appendix A gives the full model with covariates, regularization, multiple analysis conditions, and the ranking rule; Appendix B lists cohort sizes and variant-selection procedures. The code for the search and validation loops is available at `https://anonymous.4open.science/r/iclr26535-code--0DD3`; it can be used by any interested party with access to the OncoArray and MEC data via dbGaP and to the Controlled Tier of the All of Us platform. The discovered scheme for East Asian subjects, with prostate cancer as the endpoint, is available at `https://anonymous.4open.science/r/iclr26535-code--0DD3/prio_func_disc_runs/LockedDiscoveredFunctionFocalEastAsian.py`.

REFERENCES

All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature*, 627:340–346, 2024.

V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance. *Nature*, 567:305–307, 2019.

S. M. Berry, K. R. Broglio, S. Groshen, et al. Bayesian hierarchical modeling of patient subpopulations: efficient designs of Phase II oncology clinical trials. *Clinical Trials*, 10(5):720–734, 2013.

D. V. Conti, B. F. Darst, L. C. Moss, et al. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature Genetics*, 53(1):65–75, 2021.

A. Das and Y. Cui. Bridging ancestry gaps in genomic risk prediction with tabular foundation models. *Bioinformatics*, 42(Supplement_1):btag217, 2026.

A. Das, M. I. Khalid, R. Peñaloza, et al. When no paths lead to Rome: Benchmarking systematic neural relational reasoning. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025), Datasets and Benchmarks Track*, 2025.

Y. Ding, K. Hou, Z. Xu, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature*, 618:774–781, 2023.

Y. Gao and Y. Cui. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nature Communications*, 11:5131, 2020.

Y. Gao and Y. Cui. Optimizing clinico-genomic disease prediction across ancestries: a machine learning strategy with Pareto improvement. *Genome Medicine*, 16:76, 2024.

A. E. Ghareeb, B. Chang, L. Mitchener, et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026.

J. Gottweis, W.-H. Weng, A. Daryin, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026.

C. A. Haiman, N. Patterson, M. L. Freedman, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics*, 39:638–644, 2007.

M. Harlemon, O. Ajayi, P. Kachambwa, et al. A custom genotyping array reveals population-level heterogeneity for the genetic risks of prostate cancer and other cancers in Africa. *Cancer Research*, 80(13):2956–2966, 2020.

J. Howlett. AI just solved an 80-year-old 'Erdős problem,' and mathematicians are amazed. *Scientific American*, 21 May 2026. https://www.scientificamerican.com/article/ai-just-solved-an-80-year-old-erdos-problem-and-mathematicians-are-amazed/.

T. Hubert, R. Mehta, L. Sartran, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651:607–613, 2026.

L. Kachuri, N. Chatterjee, J. Hirbo, et al. Principles and methods for transferring polygenic risk scores across global populations. *Nature Reviews Genetics*, 25(1):8–25, 2024.

L. N. Kolonel, B. E. Henderson, J. H. Hankin, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *American Journal of Epidemiology*, 151(4):346–357, 2000.

T. B. Kratzer, N. Mazzitelli, J. Star, et al. Prostate cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 2025.

S. A. Lambert, L. Gil, S. Jupp, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nature Genetics*, 53:420–425, 2021.

C. Márquez-Luna, P.-R. Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, et al. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41 (8):811–823, 2017.

A. R. Martin, M. Kanai, Y. Kamatani, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584–591, 2019.

J. K. McHugh, E. K. Bancroft, E. Saunders, et al. Assessment of a polygenic risk score in screening for prostate cancer. *New England Journal of Medicine*, 392(14):1406–1417, 2025.

L. A. Mucci, J. B. Hjelmborg, J. R. Harris, et al. Familial risk and heritability of cancer among twins in Nordic countries. *JAMA*, 315(1):68–76, 2016.

A. Novikov, N. Vũ, M. Eisenberger, et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

A. B. Popejoy and S. M. Fullerton. Genomics is failing on diversity. *Nature*, 538:161–164, 2016.

T. R. Rebbeck. Prostate cancer genetics: variation by race, ethnicity, and geography. *Seminars in Radiation Oncology*, 27(1):3–10, 2017.

B. Romera-Paredes, M. Barekatain, A. Novikov, et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Y. Ruan, Y.-F. Lin, Y.-C. A. Feng, et al. Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics*, 54:573–580, 2022.

G. Tsoukalas, A. Kovsharov, S. Shirobokov, et al. Advancing mathematics research with AI-driven formal proof search. *arXiv preprint arXiv:2605.22763*, 2026.

K. Tsuo, Z. Shi, T. Ge, et al. All of Us diversity and scale yield context-dependent improvements in polygenic prediction. *Nature Genetics*, pp. 1–9, 2026.

H. Wu, P. Pérez-Rodríguez, M. Boehnke, et al. Improving polygenic score prediction for underrepresented groups through transfer learning. *Nature Communications*, 17:1973, 2026.

Appendix: Supplementary Material

The following supplementary material is provided as an appendix to the main paper. Appendix A gives the full model and evaluator; Appendix B gives cohort sizes and variant-selection procedures.

A Full model and evaluator

This appendix restates the method of Section 3 in full generality: with optional covariates, with the fixed regularization used in the implementation, and with several analysis conditions. Setting $d = 0$ and $Q = 1$ and removing the regularizers recovers the equations of the main text.

A.1 Inputs with covariates

Reference subject k additionally carries a prespecified covariate vector $c_k \in \mathbb{R}^d$ ($d = 0$ when no covariates are used):

$$\mathcal{R} = \{(y_k, a_k, g_k, c_k)\}_{k=1}^{n}, \qquad y_k \in \{0,1\},\ a_k \in \mathbb{R}^{16},\ g_k = (G_{k1}, \ldots, G_{kp})^\top,\ c_k \in \mathbb{R}^d. \tag{A1}$$

A target subject i supplies a_i, $g_i = (G_{i1}, \ldots, G_{ip})^\top$, and c_i, but no label. The ancestry coordinates are the 16-dimensional continuous coordinates of the type distributed with the All of Us genomic release; they are computed upstream and are not part of the search. The variant descriptor v_j contains the stable variant identifier, chromosome and position when available, effect allele, and the key identifying the matching dosage field in the current reference data.

A.2 Scheme interface

As in Eq. (2), $r_{ij} = f(\mathcal{R}, a_i, v_j) \geq 0$ and $\mathcal{N}_{ij} = \{k : \|a_k - a_i\|_2 \leq r_{ij}\}$. The candidate scheme must return a finite non-negative radius. It never receives the target subject's label and cannot change the label definition, the local effect estimator, the calibration model, the fold assignment, or the evaluation metric.

A.3 Local effects with covariates

Let $z_{kj} = (1, G_{kj}, c_k^\top)^\top$ and $\theta = (\beta_0, b, \theta_c^\top)^\top$, and let $\theta_{-0} = (b, \theta_c^\top)^\top$ denote the non-intercept coefficients. With a small fixed ridge constant $\lambda_0 > 0$ for numerical stability,

$$\hat{\theta}_{ij} = \arg\min_{\theta} \Big[\sum_{k \in \mathcal{N}_{ij}} \ell\big(y_k, z_{kj}^\top \theta\big) + \frac{\lambda_0}{2} \|\theta_{-0}\|_2^2 \Big]. \tag{A2}$$

The dosage coordinate of $\hat{\theta}_{ij}$ is the local effect $b_j(a_i)$. The same identifiability conditions as in the main text apply (at least 25 subjects, both classes, non-constant dosage; otherwise $b_j(a_i) = 0$), and $X_{ij} = G_{ij}\, b_j(a_i)$.

A.4 Analysis conditions, folds, and calibration

Let Q be the total number of prespecified analysis conditions, indexed by $q = 1, \ldots, Q$, and let H_q be the number of scoring folds for condition q. In our experiments $Q = 2$: one condition without additional covariates ($d = 0$) and one with additional prespecified covariates; each used $H_q = 2$ scoring folds. For condition q, the evaluation portion is deterministically partitioned into disjoint scoring folds $\mathcal{S}_{qh}$; the remaining evaluation subjects form the calibration fold $\mathcal{C}_{qh}$. Genomic features of calibration subjects use only the condition-specific reference portion in Eq. (A2); genomic features of scoring subjects use the reference portion together with $\mathcal{C}_{qh}$. On the calibration fold, with a fixed regularization value λ^*,

$$\eta_i(\alpha, \gamma, \omega) = \alpha + \gamma^\top c_i + \omega^\top X_i, \qquad (\hat{\alpha}_{qh}, \hat{\gamma}_{qh}, \hat{\omega}_{qh}) = \arg\min_{\alpha,\gamma,\omega} \Big[\sum_{i \in \mathcal{C}_{qh}} \ell\big(y_i, \eta_i\big) + \frac{\lambda^*}{2} \|\omega\|_2^2 \Big]. \tag{A3}$$

The same λ^* is used for every condition and fold, in discovery and in validation. For $i \in \mathcal{S}_{qh}$, the frozen score is $s_{i,qh} = \hat{\alpha}_{qh} + \hat{\gamma}_{qh}^\top c_i + \hat{\omega}_{qh}^\top X_i$, computed without y_i; $\sigma(s_{i,qh})$ is the corresponding risk probability.

A.5 Fold score and performance score

Let A_{qh} be the AUC of the scores $s_{i,qh}$ on $\mathcal{S}_{qh}$. The fold score F_{qh} is the median AUC across $B = 200$ bootstrap resamples of $\mathcal{S}_{qh}$; resamples containing a single class are discarded, and if none remain, $F_{qh} = A_{qh}$. The performance score is

$$\bar{F} = \Big(\sum_{q=1}^{Q} H_q\Big)^{-1} \sum_{q=1}^{Q} \sum_{h=1}^{H_q} F_{qh}, \tag{A4}$$

which in our experiments averages four fold scores.

A.6 Ranking rule

Let s_{simp} be the negative number of abstract-syntax-tree nodes in the candidate's code. When a candidate is registered, s_{simp} is compared with $M \le 100$ schemes sampled from the destination island. If $M > 0$, let r be the number of those schemes with strictly greater simplicity plus one half the number tied; the simplicity bonus is $B_{\mathrm{simp}} = 0.008\,(1 - 2r/M) \in [-0.008, 0.008]$, and $B_{\mathrm{simp}} = 0$ if $M = 0$. The ranking score is

$$S_{\mathrm{rank}} = \bar{F} + B_{\mathrm{simp}}. \tag{A5}$$

The globally best candidate by S_{rank} at the stopping point is locked.

A.7 Search configuration

Four islands; seven candidate schemes requested per island per cycle; up to two previously scored schemes per prompt; at the end of each cycle the weaker two islands are reset and reseeded from strong schemes in the surviving islands; at most 12 cycles, with early stopping after one cycle without improvement. Every evaluated candidate, its fold-level scores, $\bar{F}$, island, cycle, prompt, and LLM completion are stored in the database. In the main experiment the search stopped after two cycles.

A.8 Post-lock validation with covariates

Validation follows Section 3.7 with Eqs. (A2)–(A3) in place of Eqs. (4) and (6), using the same λ_0 and λ^*. The held-out label of a subject is used only to compute the final AUC and the paired statistics of Eq. (9); it is never used to choose a radius, estimate a local effect, select a scheme, or fit the calibration model.

B Data details

Table 2: Cohort composition. Each cohort is split into disjoint portions. The labeled portions supply the data from which effects are borrowed and the risk model is fit (reference and evaluation portions for the Discovery Cohort; reference and calibration portions for each Independent Validation Cohort). Held-out portions contain focal subjects only and were never used during discovery.

Role	Data source	Portion	Focal subjects	Non-focal subjects
Discovery	MEC	reference	51	1,871
		evaluation	76	279
Validation 1	MEC	reference + calibration	127	2,150
		held-out	19	0
Validation 2	OncoArray	reference + calibration	566	28,220
		held-out	63	0
Validation 3	AoU	reference + calibration	46	6,560
		held-out	30	0

Ancestry-group composition of the Discovery Cohort. Reference portion: 1,165 African American, 51 Japanese American (focal), and 706 Latino subjects. Evaluation portion: 165 African American (84 controls / 81 cases), 76 Japanese American (31 / 45), and 114 Latino (64 / 50) subjects. The ablation of Section 4.3 undersampled the evaluation portion to 200 subjects (African American 44/44, Japanese American 21/28, Latino 40/23; controls/cases) and to 100 subjects (African American 28/18, Japanese American 9/13, Latino 15/17).

Variant selection. *MEC:* variants from prostate cancer scores in the PGS Catalog (Lambert et al., 2021), excluding PGS Catalog scores whose metadata indicated use of data from any of the sources used in the present study (MEC, OncoArray, or AoU). *OncoArray:* variants independently drawn from prostate cancer scores in the PGS Catalog, with the same exclusion; the variants may therefore differ from those used for MEC. *AoU:* the 50 variants most strongly associated with prostate cancer in the version 7 All by All association analysis, followed by linkage-disequilibrium pruning (removal of variants strongly correlated with a retained variant). Subjects from version 7 provided the labeled data; the held-out subjects joined in version 8 and were not part of the version 7 analysis used for variant selection.

Harmonization. Each source required its own upstream processing (aligning subjects to labels, harmonizing variant identifiers and effect alleles, converting genotypes to effect-allele dosages, and quality control), after which all sources were represented in the common format of Eq. (A1). The locked scheme and evaluator operate only on this common format.