PRSformer: Disease Prediction from Million-Scale Individual Genotypes

Payam Dibaeinia[1]
p.dibaeinia@gmail.com

Chris German[1]
chrisg@23andme.com

Suyash Shringarpure[1]
suyashss@gmail.com

Adam Auton[1]
aauton@23andme.com

Aly A. Khan[1,2]*
aakhan@uchicago.edu

[1]23andMe, Palo Alto, CA, USA
[2]University of Chicago, Chicago, IL, USA

Abstract

Predicting disease risk from DNA presents an unprecedented emerging challenge as biobanks approach population scale sizes ($N > 10^6$ individuals) with ultra-high-dimensional features ($L > 10^5$ genotypes). Current methods, often linear and reliant on summary statistics, fail to capture complex genetic interactions and discard valuable individual-level information. We introduce **PRSformer**, a scalable deep learning architecture designed for end-to-end, multitask disease prediction directly from million-scale individual genotypes. PRSformer employs neighborhood attention, achieving linear $O(L)$ complexity per layer, making Transformers tractable for genome-scale inputs. Crucially, PRSformer utilizes a stacking of these efficient attention layers, progressively increasing the effective receptive field to model local dependencies (e.g., within linkage disequilibrium blocks) before integrating information across wider genomic regions. This design, tailored for genomics, allows PRSformer to learn complex, potentially non-linear and long-range interactions directly from raw genotypes. We demonstrate PRSformer's effectiveness using a unique large private cohort ($N \approx 5$M) for predicting 18 autoimmune and inflammatory conditions using $L \approx 140$k variants. PRSformer significantly outperforms highly optimized linear models trained on the *same individual-level data* and state-of-the-art summary-statistic-based methods (LDPred2) derived from the *same cohort*, quantifying the benefits of non-linear modeling and multitask learning at scale. Furthermore, experiments reveal that the advantage of non-linearity emerges primarily at large sample sizes ($N > 1$M), and that a multi-ancestry trained model improves generalization, establishing PRSformer as a new framework for deep learning in population-scale genomics.

1 Introduction

Learning predictive models from high-dimensional, complex structured data is a fundamental machine learning challenge. This challenge is acutely relevant in modern genomics, where biobanks are rapidly scaling towards **million-sample sizes** ($N > 1,000,000$) and individual genomes are characterized by hundreds of thousands to millions of genetic variants (e.g., Single Nucleotide Polymorphisms, SNPs), yielding a regime of **ultra-high dimensionality** ($L > 100,000$). Effectively leveraging

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

individual-level genomic data at this unprecedented $N \times L$ scale is critical for unlocking deeper insights into complex trait genetics, such as predicting disease susceptibility [1, 2].

Current state-of-the-art methods for disease risk prediction primarily rely on Polygenic Risk Scores (PRS) derived from Genome-Wide Association Study (GWAS) summary statistics [3, 4, 5, 6]. While effective to a degree, these methods are predominantly linear, capturing additive genetic effects. Furthermore, by operating on summary statistics, they discard potentially valuable individual-level information and struggle to model non-additive genetic interactions (epistasis) [7, 8, 9, 10]. These limitations may cap predictive performance, especially as dataset sizes scale towards the million-sample regime where subtle interaction effects might become detectable.

Transformer architectures have revolutionized sequence modeling in other domains by capturing complex, long-range dependencies via self-attention [11]. We hypothesize that the attention mechanism provides a powerful inductive bias for genomics, enabling more effective modeling of pairwise linear and non-linear interactions between genetic loci compared to traditional architectures or inherently linear models. However, a critical barrier persists: the prohibitive $\mathcal{O}(L^2)$ computational complexity of standard self-attention renders its direct application to genome-scale sequences ($L > 100,000$) computationally infeasible. Addressing this scalability bottleneck is critical to harnessing the power of Transformers for large-scale genomics.

Here we introduce **PRSformer**, a novel Transformer-based architecture specifically engineered for scalable, end-to-end, multitask disease risk prediction directly from ultra-high-dimensional ($L > 100$k) individual-level genotypes ($N > 1$M). PRSformer's core innovation lies in its scalability, achieved by incorporating neighborhood attention (NA) [12], an efficient attention mechanism restricting computations to local genomic windows, resulting in $\mathcal{O}(L)$ complexity per layer. This design aligns with the biological structure of the genome, which we treat as a series of linked regions, called linkage disequilibrium (LD) blocks, where genetic variants are often inherited together. PRSformer stacks NA layers, which first model interactions within LD blocks, then progressively integrates information between neighboring LD blocks in deeper layers, capturing larger-scale genetic patterns influencing disease.

To evaluate PRSformer's ability to harness the shared genetics underlying multiple, related traits, we trained and validated it in a multitask setting across 18 autoimmune and inflammatory conditions. This trait set provides an ideal testbed for multitask learning due to immune-mediated inflammatory diseases frequently exhibiting shared inflammatory pathophysiology and overlapping genetic factors [13]. The multitask formulation allows us to test PRSformer's ability to exploit shared genetic associations while learning disease-specific patterns, aiming to enhance predictive performance through a shared representation.

We provide rigorous empirical validation using data from a unique large private cohort ($N \approx 3.8$M European-ancestry individuals) for training, validation, and evaluation across D = 18 autoimmune and inflammatory conditions using $L \approx 140$k variants. The scale of this cohort significantly exceeds current public biobanks [14], providing a critical real-world testbed for methods designed to handle genomic data of the million-scale magnitude. We conduct stringent comparisons against: (i) highly optimized linear models (regularized logistic regression with learnable embeddings) trained on the *exact same individual-level genotype data*, isolating the benefit of PRSformer's non-linear architecture; and (ii) state-of-the-art summary-statistic-based PRS methods (LDPred2 [5]) derived from the *exact same cohort*, enabling a direct comparison between end-to-end and summary-statistic-based approaches. Our experiments demonstrate statistically significant performance gains for PRSformer.

Our main contributions are:

- **Scalable deep learning for genomics at population scale:** We present an efficient multitask Transformer architecture applied to population-scale data ($N \approx 5$M) with ultra-high-dimensional features ($L \approx 140$K), establishing a blueprint for tackling other genome sequence prediction tasks.
- **Critical scaling law for non-linear models:** We empirically establish and quantify a key scaling law demonstrating that the predictive advantage of non-linear models over linear methods emerges primarily at the million-sample scale for complex immune-related conditions. Our analysis quantifies this effect, showing that performance gains grow consistently as the training set size increases beyond one million individuals.

- **Multitask learning improves genomic prediction:** We show that multitask training across related traits consistently outperforms the standard single-task paradigm, demonstrating the benefit of learning a shared genetic representation across complex immune-mediated inflammatory diseases.
- **Improved cross-ancestry generalization:** We show that training PRSformer on multi-ancestry data, including an additional ~ 1.1M non-European individuals, markedly improves prediction accuracy for held-out non-European individuals compared to a model trained only on European-ancestry data, offering a path toward more equitable genomic prediction.

2 Related work

This work is situated at the intersection of statistical genetics, genomics, and deep learning. We specifically advance upon prior work in three key areas: polygenic risk prediction, the application of deep learning to genomic data, and the development of efficient Transformer architectures for ultra-long sequences.

2.1 Polygenic risk score methods

Traditional complex trait prediction relies heavily on PRS derived from GWAS summary statistics [4, 6]. Early methods often involved simple thresholding and summing of SNP effects [3, 15]. More recent Bayesian approaches, such as LDpred2 [5] and PRS-CS [16], explicitly model LD patterns and utilize shrinkage priors to improve predictive accuracy. These methods represent the current state-of-the-art for prediction from summary statistics. However, PRS methods based on summary statistics are fundamentally limited in several ways.

First, by discarding individual-level genotype and haplotype information, these methods cannot capture LD structure and must instead rely on LD estimates that are typically imputed from external reference panels, which can introduce biases due to population mismatches [17, 18]. Second, they cannot capture variant-variant interactions such as epistasis as they are restricted to using marginal variant effects. Third, the use of precomputed summary statistics constrains these models to largely linear architectures, precluding the discovery of complex multi-locus or hierarchical genetic patterns. Recent summary-statistic approaches to leverage non-additive signal remain constrained by the lack of individual-level haplotype context [19]. Taken together, these limitations may cap predictive performance, particularly as biobank-scale datasets grow large enough to enable the detection of more subtle and nonlinear genetic effects.

Alternative individual-level approaches, such as BOLT-LMM [20] and GEMMA [21], estimate SNP effect sizes under a linear mixed model framework to account for population structure and polygenic background effects. However, these methods are computationally demanding at our study's scale (N=3.8M, D=18 traits): GEMMA's cubic complexity in N renders it intractable, while BOLT-LMM, though more scalable, operates on a single trait at a time, requiring 18 separate runs. Prior work has shown that LDPred2 achieves predictive performance comparable to BOLT-LMM across multiple traits [22, 23], supporting its use as a strong linear baseline for comparison.

2.2 Deep learning in genomics and trait prediction

Deep learning has been successfully applied to various supervised genomic prediction tasks. Much work has focused on modeling sequence-level information (DNA base pairs) to predict molecular phenotypes like transcription factor binding [24, 25, 26], chromatin accessibility [27], or gene expression [28, 29]. These approaches have predominantly utilized Convolutional Neural Networks (CNNs) or Transformers incorporating CNN-style tokenization, which are well-suited to capturing biologically meaningful motifs and local patterns at base-pair resolution. However, this paradigm is less intuitive when modeling the influence of genetic variants (e.g., SNPs) on complex traits, as causal variants can be spread across the genome and may interact over long distances, often without strong local sequence motifs defining their impact.

The application of deep learning to predict complex traits (like disease status) directly from *individual-level genotype data* (i.e., SNP arrays) remains relatively underexplored, particularly at the population scale addressed in this paper [30, 31]. This is largely due to the challenges of ultra-high dimensionality

(L) and, until recently, the limited statistical power of publicly available cohorts with both individual-level genotypes and phenotypes (N). Prior work in this specific domain has often relied on: (i) tree ensemble models such as gradient boosting or simple neural networks trained on reduced feature sets (e.g., using LASSO feature selection); (ii) smaller cohorts where complex interactions are difficult to detect; and (iii) models operating on precomputed PRS or summary statistics rather than raw genotype data [8, 32, 33, 34, 35, 36]. Recently, Phenformer [37] proposed a multi-scale Transformer that predicts disease risk from DNA sequences by linking genetic variation, gene expression, and phenotype through a pretrained sequence-to-expression backbone. While conceptually similar to our end-to-end genotype-to-phenotype goal, Phenformer operates on DNA sequences covering approximately $\approx 3\%$ of the genome and is trained on ~ 150K individuals, whereas our approach models variant-level genotype data and scales to millions of individuals, enabling systematic analysis of how nonlinearity interacts with data scale in complex trait prediction.

2.3 Efficient transformer architectures

Applying standard Transformers to genome-scale data ($L > 100$k) is computationally prohibitive due to the $\mathcal{O}(L^2)$ complexity of self-attention [11, 38]. A wide range of efficient attention mechanisms have been proposed to address this limitation, including sparse attention patterns (e.g., Longformer [39], BigBird[40]), low-rank approximations (e.g., Linformer [41]), and kernel-based methods (e.g., Performer [42]). Other architectures exploit locality through sliding windows or blockwise mechanisms (e.g., Swin Transformer [43]) to reduce complexity while capturing local dependencies. In our work, we adopt neighborhood attention [12], a variant of self-attention in which each query attends only to a fixed-size local window of neighboring tokens, rather than the full sequence. This inductive bias aligns well with the block-like correlation structure of genomic data driven LD. By limiting attention to a neighborhood of size $k \ll L$, NA reduces both computational and memory complexity to $\mathcal{O}(L \cdot k)$ -achieving linear scaling in sequence length. We employ the optimized GPU implementation provided by the NATTEN library [12, 44], which supports scalable training on long sequences while maintaining the expressiveness of content-based attention.

3 Methods

3.1 Problem definition

We aim to predict susceptibility to multiple ($D = 18$) autoimmune and inflammatory conditions from individual-level genotypes. Formally, given a dataset of N individuals, the input for individual i is their genotype profile $\mathbf{x}_i \in \{0, 1, 2, \text{UNKN}\}^L$, representing genotypes of L pre-selected genetic variants, where UNKN indicates missing data. The target output is a vector $\mathbf{y}_i \in \{0, 1, \text{UNKN}\}^D$, representing the binary status (case/control) for D diseases, where UNKN indicates unrecorded status. Our goal is to learn a multitask function $f : \mathbb{Z}^L \to [0, 1]^D$ that predicts the probability of each disease $\hat{\mathbf{y}}_i = f(\mathbf{x}_i)$. The primary challenges lie in the ultra-high dimensionality ($L \approx 140$k in this work), the need to capture potentially non-linear and long-range interactions, and leveraging the statistical power of population-scale datasets ($N \approx 5$M total used in this study). We propose a Transformer-based architecture adapted for this task, leveraging efficient attention mechanisms for scalability and multi-task learning for joint prediction across diseases.

3.2 PRSformer architecture

PRSformer adapts the Transformer architecture for disease prediction from ultra-long ($L \approx 140$k) individual-level genotype sequences using the following key designs:

Scalability via Neighborhood Attention: Standard $\mathcal{O}(L^2)$ self-attention is computationally infeasible. We replace it with Neighborhood Attention (NA) [12], restricting each query token's attention to a symmetric local window of size k. This reduces complexity to $\mathcal{O}(L \cdot k)$, enabling efficient processing of the $L = 137,245$ input variants used in this study. We use $k = 385$, chosen via hyperparameter tuning (Section 3.5, Supplementary Table F8), which conceptually aligns with capturing dependencies within local LD blocks (Figure 1A) and corresponds to roughly ≈100 kilobases along the genome [45].

Genome-ordered input without explicit positional encodings: Input variants are ordered by their chromosomal position and then concatenated from Chr1 to Chr22. This fixed order is used for all

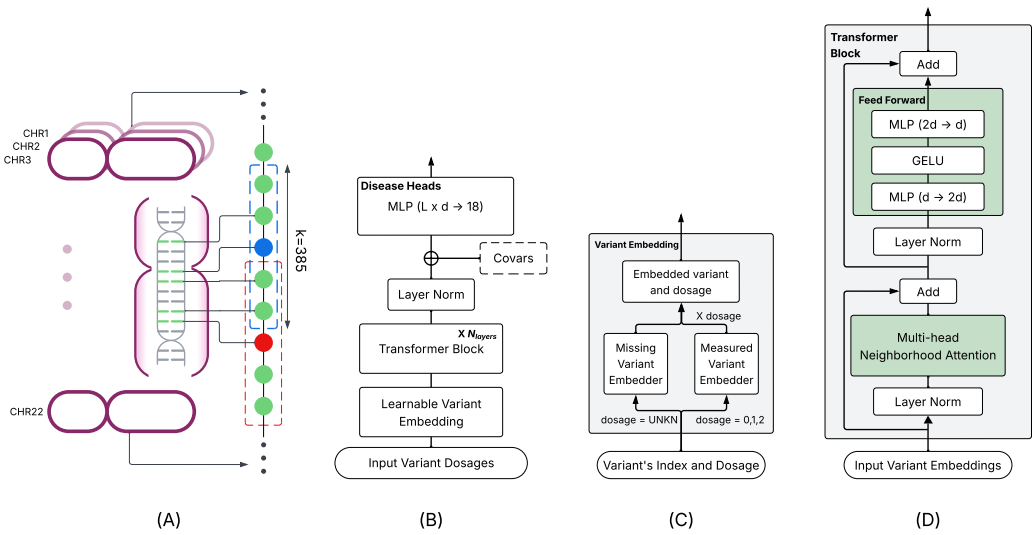

Figure 1: Schematic overview of PRSformer and its core components. (A) A preselected set of variants across the 22 chromosomes, sorted by genomic position, allows each query variant (e.g., blue or red) to attend within a local block ($k = 385$) via neighborhood attention. (B) Model architecture with input, Transformer blocks, and output layer. (C) Variant embedding layer, which encodes each variant and its corresponding genotype (0, 1, 2, or UNKN) into a 64-dimensional representation. (D) Transformer block with pre-layer normalization, neighborhood attention, and GELU activation.

individuals. We omit standard positional encodings (e.g., sinusoidal or learned absolute). The fixed genomic order provides implicit relative positional information that NA inherently leverages within its local attention windows. We also experimented with learned positional encodings, but did not observe measurable improvement in performance.

The overall data flow of PRSformer proceeds as follows (Figure 1B):

1. **Learnable variant embedding layer (Figure 1C):** Each variant in the input sequence of variant genotypes is mapped to a d_{model}-dimensional vector. For each variant j and individual i, an observed genotype $x_{ij} \in \{0, 1, 2\}$ is represented as $E_j \cdot x_{ij}$ using a learned variant-specific embedding E_j. A missing genotype ($x_{ij} = \text{UNKN}$) is represented by a separate learned variant-specific embedding M_j. This allows the model to distinguish missingness from observed genotypes distinctly for each variant. We use $d_{\text{model}} = 64$.

2. **Transformer blocks (Figure 1D):** The embedded sequence is processed by $N_{\text{layers}} = 2$ Transformer blocks. Each block applies pre-layer normalization [46] followed by multi-head NA ($N_{\text{heads}} = 4$) and a feed-forward network with GELU activation [47]. The choice of $N_{\text{layers}} = 2$ was based on validation performance, where deeper models did not show significant improvement for this task (Supplementary Table F7), suggesting two stacked NA layers provide a sufficient receptive field to capture interactions between adjacent LD blocks.

3. **Output layer:** The normalized sequence representation from the last block is flattened and optionally concatenated with covariates such as sex and age, and passed to a fully-connected layer generating $D = 18$ independent disease likelihood predictions. We also evaluated mean pooling and dedicated [CLS] tokens as alternatives to simple flattening of normalized representations, but neither outperformed the flattening-based design (see Supplementary Table F10).

Key architectural hyperparameters (d_{model}, N_{heads}, N_{layers}, NA window size k) were optimized based on validation set performance (Supplementary Tables F5-10).

3.3 Datasets

We utilized data from a large, private biobank consisting of individuals who consented to participate in research under an IRB-approved protocol. Starting from an internal GWAS data freeze timestamped 08-2021 (used to prevent information leakage, see Section 3.4), we identified individuals genotyped on the same platform and excluded all pairs of individuals related by more than 700 cM (i.e., first cousins or closer), thereby minimizing the risk of learning simple familial signals. Individuals included had recorded phenotypes (i.e., self-reported status) for at least one of the $D = 18$ autoimmune and inflammatory conditions considered. We also excluded individuals under age 10 who did not have a case diagnosis. This resulted in a training set of $N_{\text{train}} = 3,838,549$ individuals of European genetic ancestry (throughout this work, ancestry was determined via an internal classifier [48]).

We constructed temporally distinct validation and test sets using individuals who enrolled and consented after the 08-2021 data freeze date and up to 12-2024, applying the identical filtering criteria. The validation dataset was used for hyperparameter tuning, while the test dataset was used to report final performance metrics. This yielded $N_{\text{val}} = 525,448$ and $N_{\text{test}} = 494,265$ individuals. To assess whether models capture familial relationships versus causal genotype–phenotype associations, we constructed a kinship-controlled European test set ($N \approx 148$k) by subsetting test individuals related to any training sample by no more than 300 cM and ensuring that no pair within the subset is related by more than 700 cM. In total, the European dataset comprised $N = 4{,}858{,}302$ individuals across training, validation, and test sets, with an additional $N \approx 1.1$M non-European individuals included for cross-ancestry training (see Section 3.7). Case/control counts per disease and further details regarding cohort construction (including differences from the subset used for GWAS computations) are provided in Appendix A and Supplementary Tables F2-4.

3.4 Variant selection

To define the input feature space (L), we selected variants associated with at least one of the $D = 18$ diseases based on internal GWAS summary statistics (European ancestry cohort, computed prior to an 08-2021 data freeze to prevent information leakage into model training). For each disease, variants passing standard QC, located on autosomal chromosomes, had a genotyping rate ≥ 0.95, MAF > 0.001, and exhibited nominal association with the disease (GWAS p-value $< 1 \times 10^{-2}$). The final PRSformer input set was the union across all 18 diseases, resulting in $L = 137,245$ variants. Further details on GWAS procedures, exploration of variant's pruning by LD and per-disease variant counts are in Appendix B and Supplementary Table F9)

3.5 Training

Given the training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N_{\text{train}}}$, we trained PRSformer by minimizing the following loss, summed over individuals and their available (non-UNKN) disease labels $t \in T(i)$:

$$\mathcal{L} = -\sum_{i=1}^{N_{\text{train}}} \sum_{t \in T(i)} [y_{i,t} \log(\hat{y}_{i,t}) + (1 - y_{i,t}) \log(1 - \hat{y}_{i,t})]$$

where $T(i)$ denotes the set of recorded disease statuses for individual i. We also evaluated focal loss [49], task-uncertainty–weighted loss [50], and standard averaged cross-entropy, all of which were outperformed by the proposed loss function in terms of validation AUROC. We used the AdamW optimizer [51] ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay=0.05) with an initial learning rate of 5×10^{-4}, decreased via a Cosine Annealing scheduler. Training was performed efficiently on this large-scale dataset for 2 epochs consisting of $\approx$ 120,000 gradient updates in an effective batch size of 64 across four NVIDIA A100 GPUs, leveraging Distributed Data Parallel and mixed-precision (FP16) training (training duration was tuned based on validation AUROC across diseases (Table F5); most models showed signs of overfitting beyond two epochs). Hyperparameters, including architecture choices ($N_{\text{layers}}, d_{\text{model}}, N_{\text{heads}}, k$), were selected based on optimal AUROC on the validation set after extensive searches (e.g., See Supplementary Tables F5-10), following standard ML best practices to minimize overfitting to validation data and ensure that test performance provides an unbiased estimate of generalization. To isolate the contribution of genotype to model performance, all models presented in the main text were trained without including covariates such as sex or age.

3.6 Baseline models

To comprehensively evaluate PRSformer and validate our main claims regarding the utility of end-to-end non-linear modeling on large-scale individual-level genotypes, we established three rigorous baselines. These baselines are specifically designed to: (1) compare against the current state-of-the-art using conventional summary-statistic inputs (LDPred2), (2) benchmark against an enhanced version of this state-of-the-art (Stacked LDPred2), and (3) isolate the specific performance gains attributable to PRSformer's non-linear Transformer architecture via a carefully matched linear counterpart.

1. **LDPred2: state-of-the-art summary-statistic method.** We selected LDPred2 [5] due to its strong empirical performance and widespread adoption in the field [36, 52]. To ensure the most direct comparison possible, we configured LDPred2 meticulously:
    - **Matched data source:** LDPred2 was applied to GWAS summary statistics derived from the *same European training data freeze* used for PRSformer's data and variant selection.
    - **Cohort-specific LD:** An LD reference panel from our research cohort was used.
    - **Standard QC:** Input variants ($\sim$445K per disease) passed standard GWAS QC and LDPred2-specific filtering [18].
    - **Tuning:** LDPred2 hyperparameters (p, h^2) were extensively tuned (up to 100 models per disease) by maximizing AUROC on the *same validation set* used for PRSformer (details in Appendix C).
2. **Stacked LDPred2: enhanced summary-statistic baseline.** To create a stronger summary-statistic baseline, we ensembled the converged LDPred2 models from the hyperparameter search using elastic-net regression trained via cross-validation on the validation set (Appendix C).

    PRSformer+: Since Stacked LDPred2 uses the validation set for training ensemble weights, we develop and compare it against **PRSformer+**, which is the final PRSformer model retrained on the combined training and validation datasets, ensuring parity in total data usage (Supplementary Figure E2).
3. **Linear model: direct architectural ablation.** This crucial baseline isolates the contribution of PRSformer's non-linear Transformer architecture. It mirrors PRSformer precisely *except* for omitting the Transformer blocks:
    - **Identical data, inputs & training:** Uses the exact same $L \approx 140$k input variants and individual-level train/validation/test splits. Employs the same multitask framework ($D = 18$), loss function, AdamW optimizer, and training schedule (Section 3.5). Uses the same embedding layer (Figure 1C) for genotypes and missingness.
    - **Architecture difference:** The input embeddings are fed *directly* to the final linear output layer, bypassing the Transformer blocks (Figure 1D).

    This provides a multitask linear model on the same large-scale individual data, allowing direct assessment of the performance gain from PRSformer's non-linear processing.

3.7 Cross-ancestry experiments

To assess generalization, we developed PRSformer-ME (Multi-Ethnic). We performed ancestry-specific GWAS (African American (AFR), European (EUR), Latino (LAT), East Asian (EAS), and South Asian (SAS); determined by internal classifier) using the same 08-2021 data freeze and variant selection criteria (Section 3.4, p-value $< 1 \times 10^{-2}$, QC) where sample sizes permitted (Supplementary Table F12). We defined an expanded input set ($L = 251,538$ variants) as the union of selected variants across all available disease-ancestry pairs (including Europeans). We constructed a multi-ancestry training set ($\sim$ 5M total individuals) by combining the European training set (Section 3.3) with $N = 1,136,746$M non-European individuals meeting the same filtering criteria. PRSformer-ME was trained on this combined dataset using the same architecture and hyperparameters as the European-only PRSformer, without additional ancestry-specific tuning. Evaluation was performed on a combined test set including the European test set and held-out non-European individuals processed identically (Supplementary Table F3).

4 Experiments and results

We present results evaluating PRSformer's performance against baselines, analyzing the impact of non-linearity and sample scale, assessing the benefit of multitask learning, and testing cross-ancestry generalization using AUROC as the primary metric unless otherwise stated.

4.1 PRSformer outperforms state-of-the-art baselines

We first benchmarked PRSformer against the highly optimized linear model and the state-of-the-art summary-statistic method, LDPred2, on the European test set ($N_{test} \approx 494$k). As shown in Figure 2, PRSformer consistently achieves higher AUROC scores than its linear counterpart across all 18 autoimmune and inflammatory conditions. This comparison, using identical data and training setups except for the Transformer blocks, directly quantifies the predictive benefit derived from PRSformer's non-linear architecture.

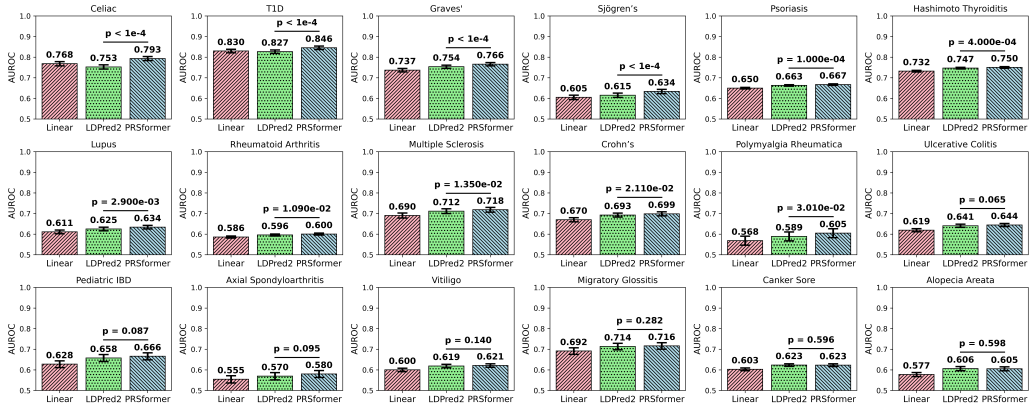

Figure 2: Benchmarking PRSformer against baseline methods using AUROC. PRSformer consistently outperforms the linear model (trained on identical individual-level data) and LDPred2 (state-of-the-art summary statistic method derived from the same cohort). Error bars: 95% CI (10k bootstraps). p-values: estimated using a one-sided paired bootstrap test (10,000 replicates), sampling with replacement from the test set and comparing AUROCs of PRSformer and LDPred2 on identical sample pairs. The p-value reflects the fraction of replicates where LDPred2's AUROC $\geq$ PRSformer's.

Crucially, PRSformer also significantly outperforms LDPred2 (using summary statistics derived from the *same* cohort) on 16 out of 18 diseases, with 11 differences being statistically significant ($p < 0.05$, one-sided paired bootstrap test). This demonstrates the advantage of end-to-end modeling on individual-level data compared to state-of-the-art methods relying on summary statistics. Consistent improvements were also observed in area under the precision–recall curve (Supplementary Figure E1) and explained variance (Supplementary Table F11), as well as when comparing against an enhanced Stacked LDPred2 baseline (PRSformer+, Supplementary Figure E2), confirming the robustness of PRSformer's advantage.

We also evaluated PRSformer and LDPred2 on the kinship-controlled test set, reproducing similar trends (Supplementary Figure E3): PRSformer outperformed LDPred2 in 14 of 18 diseases, maintaining its lead in 13 of the 16 and newly improving Alopecia Areata, with six remaining statistically significant ($p < 0.05$). The smaller number of significant improvements is expected given the reduced power of the kinship-controlled test set ($\sim$148k vs. $\sim$494k). These results confirm that PRSformer's advantage is not driven by familial confounding and persists under stringent kinship control, reinforcing the validity of our findings.

4.2 Benefit of non-linearity emerges at million-sample scale

To understand when the non-linear modeling capabilities of PRSformer become advantageous, we compared its performance against the linear baseline across varying training dataset sizes (down-sampling the $N \approx 3.8$M training set). Figure 3 reveals a critical insight: at smaller sample sizes, comparable to current large public cohorts like UK Biobank [53] (up to $N \approx 1$M), the performance

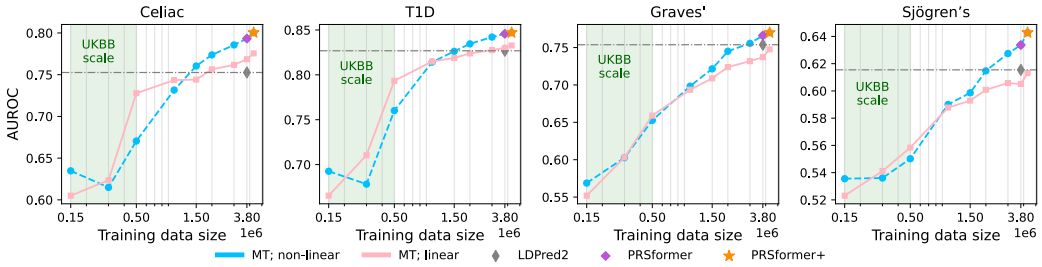

Figure 3: Impact of training scale on non-linear model advantage. Performance (AUROC) of PRSformer (non-linear) and the linear baseline across downsampled training sets (multitask setting). The benefit of non-linearity becomes apparent only at $N > 1$M scale.

of PRSformer is similar to the simpler linear model. However, as the training size exceeds one million individuals, a clear advantage for the non-linear PRSformer emerges and progressively widens. This trend holds across multiple diseases (Supplementary Figure E4) and persists even when using appropriately subsetted variant sets for smaller scales (Supplementary Figure E5). These results indicate a scaling law: the benefits of non-linear architectures like PRSformer manifest only when sample sizes are sufficient to resolve higher-order genetic interactions. Below this threshold, linear models remain competitive, whereas beyond the million-sample regime, PRSformer achieves measurable gains (although with higher computational cost in FLOPs per sample).

4.3 Multitask learning consistently improves performance

We investigated the benefit of PRSformer's multitask design by comparing it against single-task (ST) models trained independently for each disease. Figure 4 shows that multitask (MT) training consistently yields superior AUROC compared to ST training for both PRSformer and the linear baseline, across different data scales (see Supplementary Figure E6 for other diseases). This improvement was robust even when ST models used disease-specific optimized variant sets (Supplementary Figure E7). Thus, the gain stems from leveraging shared information across related immune-mediated inflammatory diseases allowing shared model components (variant embeddings and Transformer blocks in PRSformer, and variant embedding in the linear baseline) to be optimized more effectively. By training these shared layers across multiple related diseases, the model can capture generalizable representations that enhance performance beyond what is achievable with isolated, ST training.

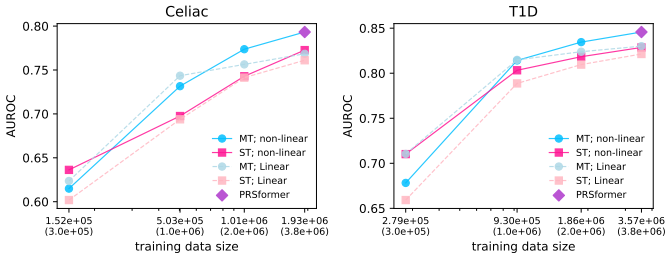

Figure 4: Multitask (MT) vs. Single-Task (ST) training for Celiac disease (left) and T1D (right). MT consistently outperforms ST for both non-linear (based on PRSformer) and the linear baseline across different scales (X-axis: ST sizes / (MT sizes)).

4.4 Improved cross-ancestry generalization via multitask multi-ancestry training

Recognizing the need for equitable genomic prediction [54], we trained PRSformer-ME on a combined multi-ancestry cohort ($\sim$ 5M individuals, including $\sim$ 1.1M non-Europeans) using an expanded variant set ($L \approx 252$k, see Section 3.7 Methods). We evaluated its performance on a held-out test set containing individuals from European (EUR), African American (AFR), Latino (LAT), East Asian (EAS), and South Asian (SAS) ancestries, comparing it to the original PRSformer trained only on EUR individuals.

As summarized in Table 1, PRSformer-ME demonstrates significantly improved generalization to non-EUR populations. It achieves substantially higher AUROC scores across most diseases in AFR, LAT, EAS, and SAS individuals compared to the EUR-only model. Importantly, this gain in non-EUR

populations is achieved with minimal to no degradation in performance on EUR individuals. These results indicate that training on diverse, aggregated individual-level data allows PRSformer-ME to capture both shared and ancestry-specific genetic signals, leading to more accurate and potentially more equitable predictions across populations compared to models trained on a single ancestry group (incorporating covariates such as sex and age further improves predictive performance, see Supplementary Table F14). This is an important finding since state-of-the-art methods generally rely on single-ancestry summary statistics, preventing them from jointly training on individual-level data across multiple ancestries and from leveraging shared cross-population genetic signals.

Table 1: AUROC of EUR-only PRSformer vs. multi-ancestry PRSformer-ME on the multi-ancestry test set. PRSformer-ME shows improved performance in non-EUR ancestries often without sacrificing EUR performance. Bold font denotes the higher AUROC in each pairwise comparison.

		Diseases																		
Ances.	Model	Celiac	T1D	Graves'	Sjögren's	Psoriasis	Hashimoto Thyroiditis	Lupus	Rheumatoid Arthritis	Multiple Sclerosis	Crohn's	Polymyalgia Rheumatica	Ulcerative Colitis	Pediatric IBD	Axial Spondyl.	Vitiligo	Migratory Glossitis	Canker Sore	Alopecia Areata
EUR	PRSformer	**0.7933**	**0.8457**	0.7661	0.6337	0.6669	**0.7504**	0.6338	0.6004	**0.7184**	**0.6985**	**0.6048**	0.6442	**0.6660**	0.5802	0.6212	**0.7164**	0.6227	0.6051
	PRSformer-ME	0.7867	0.8431	**0.7674**	**0.6404**	**0.6683**	0.7487	**0.6458**	**0.6069**	0.7157	0.6981	0.5995	**0.6447**	0.6654	**0.5818**	**0.6222**	0.7146	**0.6246**	**0.6125**
AFR	PRSformer	0.6219	0.6391	0.5676	0.5118	0.5559	0.6727	0.5404	0.5241	0.5447	0.5726	–	0.5401	0.5928	<0.5	0.5122	0.6008	0.5913	<0.5
	PRSformer-ME	**0.6269**	**0.6986**	**0.6997**	**0.6399**	**0.5822**	**0.6939**	**0.6140**	**0.5660**	**0.6397**	**0.6217**	–	**0.5902**	**0.6285**	**0.5410**	**0.5605**	**0.6550**	**0.6026**	**0.5844**
LAT	PRSformer	0.7437	0.7459	0.7060	0.6181	0.6449	0.7482	0.6063	0.5712	0.6832	0.6674	0.6526	0.6184	0.6216	0.5610	0.5871	0.6994	0.6283	0.5579
	PRSformer-ME	**0.7521**	**0.7732**	**0.7568**	**0.6685**	**0.6596**	**0.7581**	**0.6536**	**0.5979**	**0.7256**	**0.6842**	**0.6553**	**0.6291**	**0.6539**	**0.6038**	**0.6166**	**0.7270**	**0.6397**	**0.6346**
SAS	PRSformer	**0.8180**	0.6640	**0.7667**	0.6383	0.6209	0.6705	0.5656	**0.6052**	0.7341	0.6630	–	**0.6974**	0.6423	**0.5999**	0.5693	**0.5421**	**0.5999**	0.5754
	PRSformer-ME	0.8002	**0.6852**	0.7612	**0.6885**	**0.6269**	**0.6907**	**0.6196**	0.5974	**0.7503**	**0.7061**	–	.6969	**0.6858**	0.5828	**0.6009**	0.5146	0.5813	**0.5972**
EAS	PRSformer	–	0.6507	0.6937	0.5770	0.6167	0.6535	0.6144	0.5566	**0.5406**	0.6282	**0.7431**	0.6204	0.5902	0.5184	0.5463	0.6809	0.5731	0.5846
	PRSformer-ME	–	**0.7121**	**0.7487**	**0.6176**	**0.6355**	**0.6876**	**0.6573**	**0.6085**	0.5019	**0.6256**	0.6973	**0.6545**	**0.6518**	**0.5860**	**0.5602**	**0.7085**	**0.5929**	**0.6386**

5 Conclusion

We introduced PRSformer, a scalable Transformer architecture leveraging neighborhood attention to enable end-to-end, multitask disease prediction from population-scale individual genotypes ($N \approx 5$M, $L \approx 140$k). Our rigorous evaluation on a unique large private cohort, conducted under IRB and using consented research participant data, demonstrates that PRSformer significantly outperforms strong linear and state-of-the-art summary-statistic baselines (LDPred2) derived from the same cohort.

A key finding of this work is that the benefit of PRSformer's non-linear modeling for complex immune-mediated inflammatory diseases emerges primarily at the million-sample scale ($N > 1$M). This advantage varies across diseases, with traits like celiac disease and type 1 diabetes benefiting substantially from non-linear modeling to explain disease risk variance [55]. This scaling law, alongside our findings that multitask training improves performance and multi-ancestry data enhances generalization, establishes a new framework for genomic prediction.

While PRSformer advances predictive accuracy, its gains come with higher computational demands that may limit immediate clinical scalability. Furthermore, future work is required to develop interpretation methods to understand the learned non-linear interactions, which is essential for biological hypothesis generation and experimental validation. A key future direction is to extend the framework beyond a single disease domain to a phenome-scale setting spanning thousands of traits. This approach is motivated by widespread genetic pleiotropy, where a single variant can influence multiple, seemingly disparate conditions. A unified model could therefore capture the shared genetic underpinnings linking diverse biological systems, such as the contribution of immune pathways to neurodegeneration and cancer.

Our research prioritizes fairness across diverse populations and the responsible deployment of genomic models. Recognizing the sensitivity of genomic data, we have taken steps to balance transparency with participant privacy: we provide detailed methodological descriptions and have released our implementation code at https://github.com/23andMe/PRSformer; however, the data and trained models are not publicly available.

Taken together, our results and scaling analyses position PRSformer as a foundation for phenome-scale genetic risk modeling that can fully leverage genetic pleiotropy to improve prediction and generalization at population scale.

Acknowledgments and Disclosure of Funding

The authors thank the past and present employees and research participants of 23andMe for making this work possible. We are grateful to Akele Reed, Teague Sterling, David Hinds, Steve Pitts, Wei Wang, Bertram Koelsch, Michael Holmes, Stella Aslibekyan, Cordell Blakkan and Barry Hicks for their valuable contributions and insightful comments on the manuscript, and to Ali Hassani for helpful discussions on employing Neighborhood Attention. The authors also gratefully acknowledge the support of AWS for providing GPU computing resources and credits. A. A. Khan is supported in part by a Chan Zuckerberg Investigator Award.

References

[1] Arnór I Sigurdsson, Ioannis Louloudis, Karina Banasik, David Westergaard, Ole Winther, Ole Lund, Sisse Rye Ostrowski, Christian Erikstrup, Ole Birger Vesterager Pedersen, Mette Nyegaard, DBDS Genomic Consortium, Søren Brunak, Bjarni J Vilhjálmsson, and Simon Rasmussen. Deep integrative models for large-scale human genomics. *Nucleic Acids Research*, 51(12):e67–e67, 05 2023.

[2] Erping Long, Peixing Wan, Qingyu Chen, Zhiyong Lu, and Jiyeon Choi. From function to translation: Decoding genetic susceptibility to human diseases via artificial intelligence. *Cell Genomics*, 3(6):100320, 2023.

[3] Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research*, 17(10):1520–1528, 2007.

[4] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018.

[5] Florian Privé, Julyan Arbel, and Bjarni J Vilhjálmsson. Ldpred2: better, faster, stronger. *Bioinformatics*, 36(22-23):5424–5431, 2020.

[6] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):44, 2020.

[7] Jana Schwarzerova, Martin Hurta, Vojtech Barton, Matej Lexa, Dirk Walther, Valentine Provaznik, and Wolfram Weckwerth. A perspective on genetic and polygenic risk scores—advances and limitations and overview of associated tools. *Briefings in bioinformatics*, 25(3):bbae240, 2024.

[8] Michael Elgart, Genevieve Lyons, Santiago Romero-Brufau, Nuzulul Kurniansyah, Jennifer A Brody, Xiuqing Guo, Henry J Lin, Laura Raffield, Yan Gao, Han Chen, et al. Non-linear machine learning models incorporating snps and prs improve polygenic prediction in diverse human populations. *Communications biology*, 5(1):856, 2022.

[9] Pankhuri Singhal, Yogasudha Veturi, Scott M Dudek, Anastasia Lucas, Alex Frase, Kristel Van Steen, Steven J Schrodi, David Fasel, Chunhua Weng, Rion Pendergrass, et al. Evidence of epistasis in regions of long-range linkage disequilibrium across five complex diseases in the uk biobank and emerge datasets. *The American Journal of Human Genetics*, 110(4):575–591, 2023.

[10] Juannan Zhou, Mandy S Wong, Wei-Chia Chen, Adrian R Krainer, Justin B Kinney, and David M McCandlish. Higher-order epistasis and phenotypic prediction. *Proceedings of the National Academy of Sciences*, 119(39):e2204233119, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[12] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194, 2023.

[13] Alexandra Zhernakova, Cleo C Van Diemen, and Cisca Wijmenga. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews Genetics*, 10(1):43–55, 2009.

[14] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[15] Florian Privé, Bjarni J Vilhjálmsson, Hugues Aschard, and Michael GB Blum. Making the most of clumping and thresholding for polygenic scores. *The American journal of human genetics*, 105(6):1213–1221, 2019.

[16] Tian Ge, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature communications*, 10(1):1776, 2019.

[17] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592, 2015.

[18] Florian Privé, Julyan Arbel, Hugues Aschard, and Bjarni J Vilhjálmsson. Identifying and correcting for misspecifications in gwas summary statistics and polygenic scores. *Human Genetics and Genomics Advances*, 3(4), 2022.

[19] Rikifumi Ohta, Yosuke Tanigawa, Yuta Suzuki, Manolis Kellis, and Shinichi Morishita. A polygenic score method boosted by non-additive models. *Nature Communications*, 15(1):4433, 2024.

[20] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[21] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[22] Guiyan Ni, Jian Zeng, Joana A Revez, Ying Wang, Zhili Zheng, Tian Ge, Restuadi Restuadi, Jacqueline Kiewa, Dale R Nyholt, Jonathan RI Coleman, et al. A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts. *Biological psychiatry*, 90(9):611–620, 2021.

[23] Ruilin Li, Christopher Chang, Yosuke Tanigawa, Balasubramanian Narasimhan, Trevor Hastie, Robert Tibshirani, and Manuel A Rivas. Fast numerical optimization for genome sequencing data in population biobanks. *Bioinformatics*, 37(22):4148–4155, 2021.

[24] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[25] Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107, 2016.

[26] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, 53(3):354–366, 2021.

[27] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[28] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[29] Johannes Linder, Divyanshi Srivastava, Han Yuan, Vikram Agarwal, and David R Kelley. Predicting rna-seq coverage from dna sequence as a unifying model of gene regulation. *Nature Genetics*, pages 1–13, 2025.

[30] Pau Bellot, Gustavo de Los Campos, and Miguel Pérez-Enciso. Can deep learning improve genomic prediction of complex human traits? *Genetics*, 210(3):809–819, 2018.

[31] Rostam Abdollahi-Arpanahi, Daniel Gianola, and Francisco Peñagaricano. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, 52:1–15, 2020.

[32] Arnór I Sigurdsson, Ioannis Louloudis, Karina Banasik, David Westergaard, Ole Winther, Ole Lund, Sisse Rye Ostrowski, Christian Erikstrup, Ole Birger Vesterager Pedersen, Mette Nyegaard, et al. Deep integrative models for large-scale human genomics. *Nucleic Acids Research*, 51(12):e67–e67, 2023.

[33] Adrien Badré, Li Zhang, Wellington Muchero, Justin C Reynolds, and Chongle Pan. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics*, 66(4):359–369, 2021.

[34] Clara Albiñana, Zhihong Zhu, Andrew J Schork, Andrés Ingason, Hugues Aschard, Isabell Brikell, Cynthia M Bulik, Liselotte V Petersen, Esben Agerbo, Jakob Grove, et al. Multi-pgs enhances polygenic prediction by combining 937 polygenic scores. *Nature communications*, 14(1):4702, 2023.

[35] Han Li, Jianyang Zeng, Michael P Snyder, and Sai Zhang. Prs-net: Interpretable polygenic risk scores via geometric learning. In *International Conference on Research in Computational Molecular Biology*, pages 377–380. Springer, 2024.

[36] Zijie Zhao, Tim Gruenloh, Meiyi Yan, Yixuan Wu, Zhongxuan Sun, Jiacheng Miao, Yuchang Wu, Jie Song, and Qiongshi Lu. Optimizing and benchmarking polygenic risk scores with gwas summary statistics. *Genome Biology*, 25(1):260, 2024.

[37] Frederik Träuble, Lachlan Stuart, Andreas Georgiou, Pascal Notin, Arash Mehrjou, Ron Schwessinger, Mathieu Chevalley, Kim Branson, Bernhard Schölkopf, Cornelia van Duijn, et al. Multi-megabase scale genome interpretation with genetic language models. *arXiv preprint arXiv:2501.07737*, 2025.

[38] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

[39] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[40] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

[41] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[42] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[44] Ali Hassani, Wen-mei Hwu, and Humphrey Shi. Faster neighborhood attention: Reducing the o(nˆ2) cost of self attention at the threadblock level. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 64717–64734. Curran Associates, Inc., 2024.

[45] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2):283, 2015.

[46] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[48] Eric Y Durand, Chuong B Do, Peter R Wilton, Joanna L Mountain, Adam Auton, G David Poznik, and J Michael Macpherson. A scalable pipeline for local ancestry inference using tens of thousands of reference haplotypes. *bioRxiv*, pages 2021–01, 2021.

[49] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[50] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[51] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[52] Oliver Pain, Kylie P Glanville, Saskia P Hagenaars, Saskia Selzam, Anna E Fürtjes, Héléna A Gaspar, Jonathan RI Coleman, Kaili Rimfeld, Gerome Breen, Robert Plomin, et al. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS genetics*, 17(5):e1009021, 2021.

[53] Hannah Taylor, Melissa Lewins, M George B Foody, Oliver Gray, Jelena Bešević, Megan C Conroy, Rory Collins, Ben Lacey, Naomi Allen, and Lucy Burkitt-Gray. Uk biobank—a unique resource for discovery and translation research on genetics and neurologic disease. *Neurology: Genetics*, 11(1):e200226, 2025.

[54] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[55] Tobias L Lenz, Aaron J Deutsch, Buhm Han, Xinli Hu, Yukinori Okada, Stephen Eyre, Michael Knapp, Alexandra Zhernakova, Tom WJ Huizinga, Goncalo Abecasis, et al. Widespread non-additive and interaction effects within hla loci modulate the risk of autoimmune diseases. *Nature genetics*, 47(9):1085–1090, 2015.

[56] The 23andMe Research Team. 23andme technical white paper: Overview of 23andMe GWAS release: r8_g1. Technical report, 23andMe, Inc., 2023.

[57] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

A Details of the GWAS runs

The internal ancestry classifier assigns individuals to one of five major genetic ancestry groups - African American (AFR), European (EUR), East Asian (EAS), South Asian (SAS), or Latino (LAT) - based on local ancestry inference [48]. To reduce confounding introduced by population structure, GWAS analyses were stratified by these genetically inferred ancestry groups.

Principal component analysis (PCA) was conducted separately within each ancestry group using a subset of <100,000 high-quality genotyped variants shared across all internal platforms. A randomly selected subset of individuals was used for each group: 513K for AFR, 398K for EAS, 1M for EUR, 1M for LAT, and 111K for SAS [48].

For each disease and ancestry group, GWAS was performed using logistic regression with additive allelic effects as predictors. Covariates included age, sex, genotype platform (to adjust for batch effects), and top principal components - specifically, the top 5 PCs for EUR, EAS, and SAS; the top 6 for AFR; and the top 9 for LAT. Association p-values were derived using a likelihood ratio test, comparing a reduced model fitted using covariates only to a full model fitted with both additive genetic effects and covariates [56].

B Exploration of LD-based variant pruning

We additionally experimented with training PRSformer on a subset of variants that had been LD-pruned using *PLINK 2.0* [57]. LD pruning removes highly correlated SNPs to retain approximately independent markers. In this procedure, a sliding window is moved across the genome, pairwise linkage disequilibrium (r^2) is computed among variants, and SNPs exceeding a specified correlation threshold with nearby variants are iteratively removed until no pair within each window remains above that threshold. Starting from a union variant set constructed similarly to that in Section 3.4 (but with slightly adjusted filtering thresholds), we applied *PLINK 2.0* with a window size of 6,000 kb (6 Mb), a step size of one variant, and an r^2 threshold of 0.5. Supplementary Table F11 compares two models from the hyperparameter tuning round trained with and without LD-based variant pruning. Interestingly, despite reducing multicollinearity among variant features, LD pruning led to lower validation performance, suggesting that PRSformer benefits from leveraging the local correlation structure within LD blocks to capture causal signals more effectively.

C Details of LDpred2 runs

For each disease we used the Gibbs sampler LDpred2 software [5] on the summary statistics with an internal LD panel. LD matrix computation included variants with minor allele frequency greater than 0.1%, and genotype call rate greater than 90%. Variants greater than 5cM apart were assumed to be independent. Summary statistics were filtered to keep variants that had a minor allele frequency greater than 0.1% and had a genotype call rate greater than 95%. This consisted of variant sets with roughly 445,000 variants. We estimated posterior SNP-effect sizes using the grid option with a set of 100 combinations of hyperparameters, leading to up to 100 sets of polygenic risk scores (PRS) per disease (depending on convergence). The hyperparameters that LDpred2 takes are an estimate for the proportion of causal variants, p, and trait heritability, h^2. We used LD score regression to estimate h^2, then used a grid of the h^2 estimate multiplied by 0.6, 0.8, 1, 1.2, and 1.4. For p, we used a sequence of values equally spaced on a logarithmic scale from 10^{-5} to 1. The best hyperparameters for each disease were selected based on validation AUROC leading to the final LDPred2 PRS models. For Stacked LDPred2, however, we ensembled all of the converged PRSs per disease (up to 100) by training elastic net on the validation data using 5-fold cross validation.

D Subsetting variants for down-sampled experiments

When training on smaller datasets, we may not have access to the same high-powered variant selection as in the full-data setting. To account for this, we repeated variant selection using GWAS summary statistics adjusted to reflect the reduced sample size of each downsampled dataset. For each downsampled dataset, we estimated GWAS p-values under the reduced sample size using the original

GWAS summary statistics:

$$Z = \frac{\beta}{SE}; \quad Z_{ds} = Z \times \sqrt{\frac{N_{ds}}{N}}; \quad p_{ds} = 2\Phi(-|Z_{ds}|)$$

where β and SE denote the effect size and standard error from the original GWAS, N is the original sample size, N_{ds} is the downsampled sample size, and Φ is the standard normal cumulative distribution function. Subsequently, variant selection was performed independently for each disease and each downsampled dataset using the estimated p-values, applying a threshold of $p < 1e{-}2$. Multi-task model training was then conducted using the union of the selected variant sets across diseases at each downsampled scale, following the same procedure as in the full-data experiments. Additional details on variant sets and data sizes and model configurations are provided in Supplementary Tables F1,F2 and F13.

E Supplementary figures

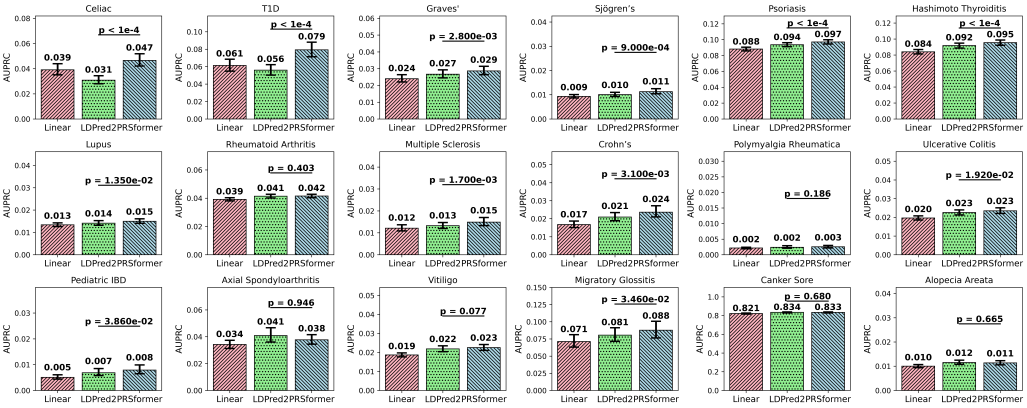

Figure E1: Benchmarking PRSformer against baseline methods using AUPRC as the evaluation metric. Numbers above the bars indicate test set AUPRC values; error bars denote 95% confidence intervals estimated via bootstrapped test samples. The reported p-values reflect the one-sided statistical significance of PRSformer outperforming LDPred2.

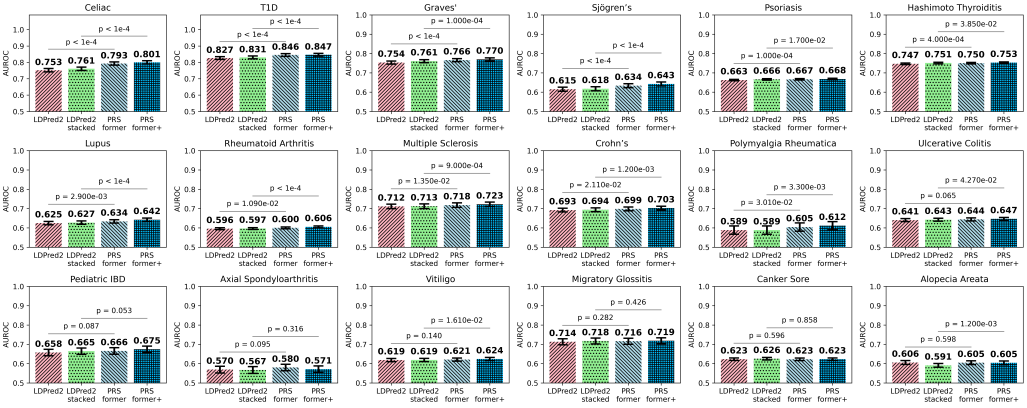

Figure E2: Benchmarking of models trained on the combined training and validation datasets. Numbers above the bars indicate test set AUROC; error bars represent 95% confidence intervals computed via bootstrapped test samples. The two sets of p-values reflect the one-sided statistical significance of PRSformer+ outperforming stacked LDPred2, and PRSformer outperforming non-stacked LDPred2—the latter being the same as those reported in Figure 2 of the main text.

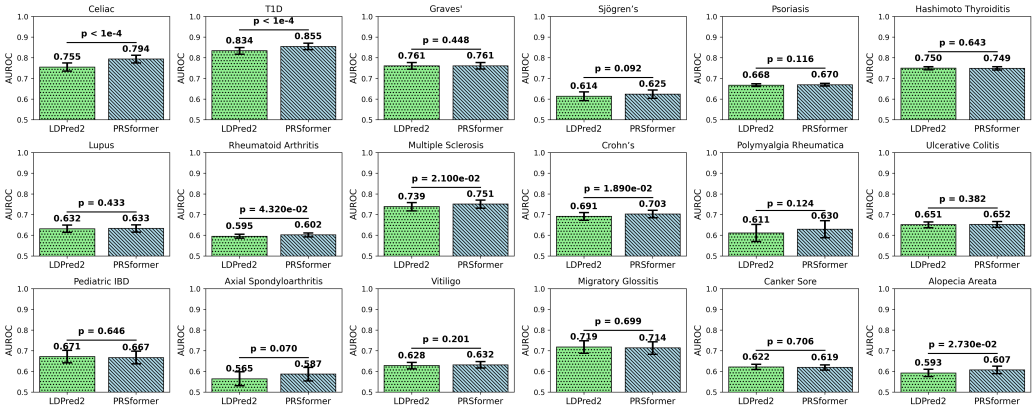

Figure E3: Comparison of PRSformer and LDPred2 on the kinship-controlled European test set, evaluated by AUROC. Numbers above the bars indicate AUROC values, and error bars represent 95% confidence intervals estimated from 10,000 bootstrapped samples. The reported p-values reflect the one-sided statistical significance of PRSformer outperforming LDPred2.

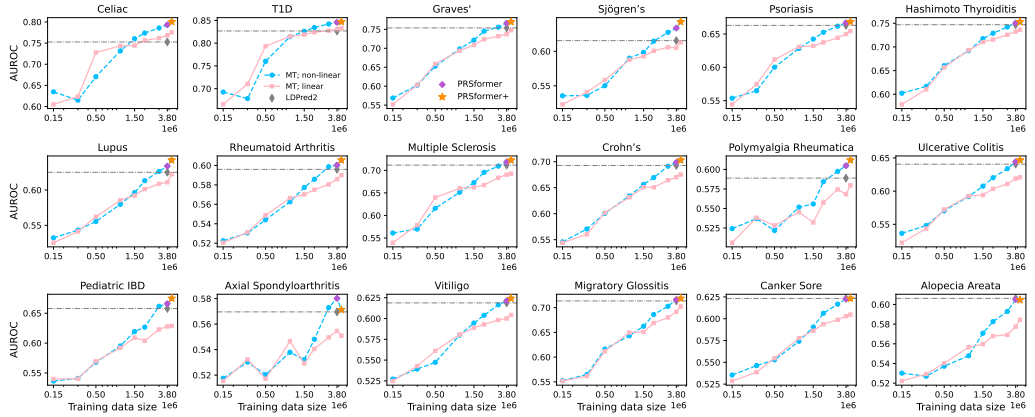

Figure E4: Prediction performance across different training scales for the non-linear model and linear baseline, both trained in a multitask (MT) setting using the same input variant set as PRSFormer. For most diseases, performance improves with more training data, with the non-linear model surpassing the linear baseline at larger scales. Fluctuations in performance for Polymyalgia Rheumatica and Axial Spondyloarthritis likely stem from the former's rarity and the latter's relatively small training size (see Supplementary Table F2).

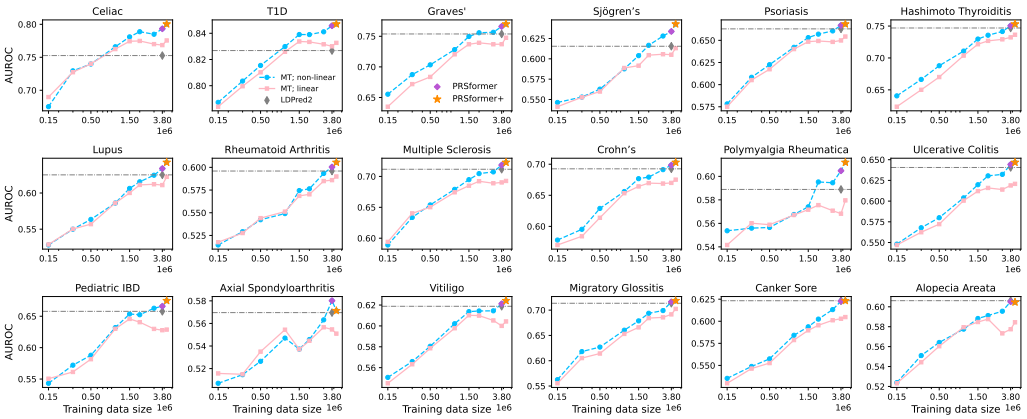

Figure E5: Prediction performance across different training scales for the non-linear model and linear baseline, both trained in a multitask (MT) setting using subsetted variant sets at each down-sampled scale (see Supplementary section D and Table F13). For most diseases, performance improves with increasing training data, with the non-linear model outperforming the linear baseline at larger scales. These trends are consistent with those observed using a fixed input variant set across scales (see Supplementary Figure E4).

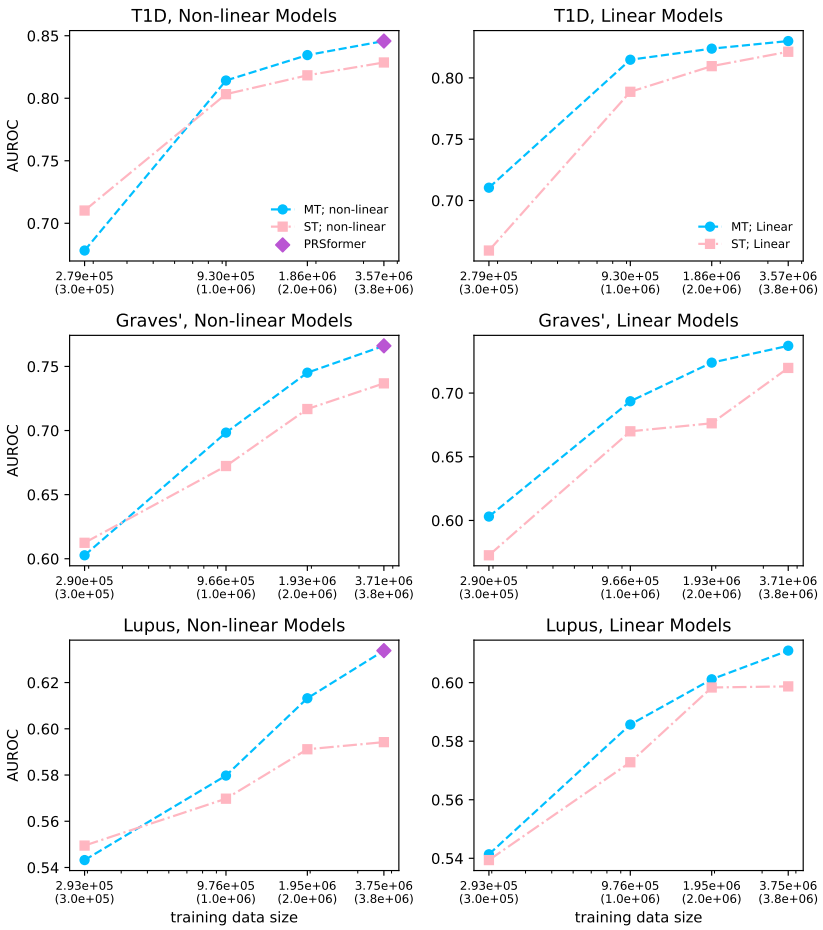

Figure E6: Comparison of multitask (MT) versus single-task (ST) training for three additional diseases across different training scales, all using the same input variant set as PRSformer. X-axis values outside parentheses indicate ST training sizes, and those inside indicate corresponding MT training sizes. Across all tested diseases, MT training outperforms ST training for both the non-linear model (left) and the linear baseline (right).

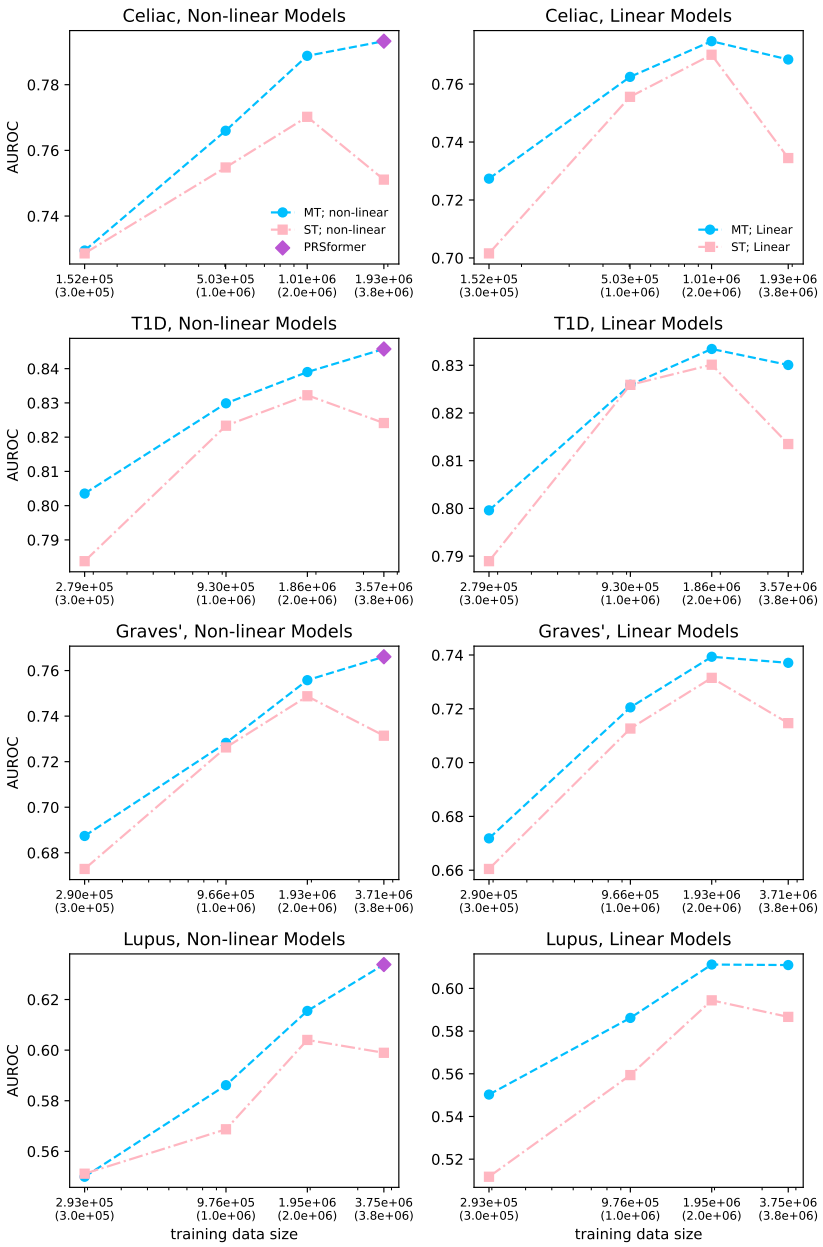

Figure E7: Comparison of multitask (MT) versus single-task (ST) training for four tested diseases across different training scales. Both MT and ST models were trained on downsampled datasets using subsetted variants; additionally, ST models used disease-specific variant sets (see Supplementary section D and Table F13). X-axis values outside parentheses indicate ST training sizes, while those inside indicate the corresponding MT sizes. Across all tested diseases, MT training outperforms ST training for both the non-linear model (left) and the linear baseline (right). These results are consistent with those observed using a fixed input variant set for both MT and ST across scales (see Supplementary Figure E6).

F Supplementary tables

Table F1: Number of selected variants per disease across European datasets

Disease	EUR dataset	3M	2M	1.5M	1M	0.5M	0.3M	0.15M
Celiac	14291	9595	6461	5576	4923	3906	3380	2722
T1D	16176	12607	7865	6424	5323	4033	3371	2530
Graves'	17350	14104	8380	6614	5231	3725	2984	2096
Sjögren's	11949	9381	4692	3502	2633	1954	1439	845
Psoriasis	24685	19621	11943	9088	6770	4529	3478	2148
Hashimoto Thyroiditis	29806	25103	15141	11314	7869	4677	3256	2030
Lupus	13618	10252	4840	3305	2384	1512	943	494
Rheumatoid Arthritis	18426	13683	6303	3980	2497	1484	891	373
Multiple Sclerosis	13347	10124	5587	4309	3301	2328	1683	1226
Crohn's	14140	10444	5126	3397	2099	739	302	90
Polymyalgia Rheumatica	8113	5818	2440	1674	1135	558	267	78
Ulcerative Colitis	14138	10381	5081	3487	2300	1042	547	164
Pediatric IBD	8677	5679	1954	1095	650	204	72	24
Axial Spondyloarthritis	6352	2215	668	316	122	20	4	0
Vitiligo	12980	9720	5231	3926	2900	1695	1092	525
Migratory Glossitis	14309	8692	4516	3327	2394	1457	1019	610
Canker Sore	18418	6181	3204	2126	1322	485	206	34
Alopecia Areata	9741	7117	3160	2087	1325	680	312	59
Union	137245	94176	38018	23397	15060	9181	6829	4863

Table F2: Training sample sizes (case / control) per disease across European and multi-ancestry datasets

Disease	Multi-Ancestry Dataset	Full Dataset	3M	2M	1.5M	1M	0.5M	0.3M	0.15M
Celiac	14851 / 2444838	13304 / 1933687	10451 / 1511685	6828 / 1007277	5228 / 755823	3458 / 503346	1747 / 251838	1052 / 151689	555 / 75772
T1D	25815 / 4600428	21108 / 3568017	16519 / 2788978	11126 / 1859420	8246 / 1393969	5525 / 929582	2714 / 464434	1626 / 278712	808 / 139611
Graves'	34377 / 4767527	26982 / 3708583	21132 / 2898733	14099 / 1932335	10573 / 1449024	7154 / 966394	3518 / 482997	2131 / 289748	1043 / 144919
Sjögren's	23480 / 4791439	19544 / 3718113	15259 / 2906128	10245 / 1937041	7562 / 1452756	5032 / 968626	2588 / 484220	1497 / 290448	735 / 145343
Psoriasis	181096 / 4404301	151767 / 3560128	118647 / 2782686	79073 / 1854947	59461 / 1391012	39630 / 927636	19622 / 463485	11922 / 278024	5934 / 139104
Hashimoto Thyroiditis	124104 / 4677800	106911 / 3708583	83363 / 2898733	55523 / 1932335	41881 / 1449024	27653 / 966394	14022 / 482997	8314 / 289748	4172 / 144919
Lupus	35926 / 4816513	27795 / 3747933	21623 / 2929376	14499 / 1952708	10807 / 1464450	7293 / 976354	3661 / 488102	2144 / 292815	1092 / 146446
Rheumatoid Arthritis	128990 / 4532187	104481 / 3611921	81737 / 2823204	54452 / 1882141	40607 / 1411177	27066 / 941019	13610 / 470158	8174 / 282097	3970 / 141143
Multiple Sclerosis	20230 / 4812927	17015 / 3735898	13384 / 2920007	8962 / 1946527	6689 / 1459824	4390 / 973437	2170 / 486450	1329 / 291870	639 / 145979
Crohn's	27434 / 4667812	23802 / 3637333	18571 / 2842879	12475 / 1895460	9301 / 1421090	5975 / 947547	3210 / 473495	1952 / 284047	992 / 142142
Polymyalgia Rheumatica	7469 / 4620931	6940 / 3588209	5392 / 2804710	3597 / 1869745	2742 / 1401959	1802 / 934872	880 / 467030	557 / 280181	274 / 140198
Ulcerative Colitis	53021 / 4636460	44568 / 3632871	34799 / 2839378	23012 / 1893078	17467 / 1419395	11663 / 946417	5880 / 472898	3456 / 283722	1784 / 141974
Pediatric IBD	10014 / 4613960	8311 / 3577312	6470 / 2796009	4303 / 1864258	3316 / 1397699	2145 / 932056	1143 / 465543	639 / 279336	337 / 139724
Axial Spondyloarthritis	7786 / 336188	6843 / 292915	5377 / 229150	3527 / 153102	2682 / 114299	1812 / 76600	901 / 37712	532 / 22931	258 / 11477
Vitiligo	38209 / 4534376	29309 / 3550146	22855 / 2774956	15261 / 1849633	11479 / 1387067	7725 / 925064	3682 / 462197	2295 / 277303	1159 / 138722
Migratory Glossitis	32394 / 1217541	27982 / 1031326	21764 / 805875	14687 / 537379	10837 / 402928	7404 / 268314	3627 / 134668	2176 / 80804	1122 / 40505
Canker Sore	578838 / 194610	493951 / 643681	385900 / 502693	257526 / 335273	193006 / 251409	128495 / 167596	64644 / 84032	38620 / 50224	19452 / 25453
Alopecia Areata	35533 / 4498787	22506 / 3522691	17705 / 2753494	11820 / 1835462	8740 / 1376318	5895 / 917773	2885 / 458592	1725 / 275141	809 / 137684

Table F3: Test data sample sizes (case / control) per disease across ancestry groups

Disease	EUR	AFR	LAT	EAS	SAS
Celiac	2160 / 243495	80 / 46973	630 / 172120	–	30 / 8072
T1D	2720 / 457259	615 / 105176	1391 / 309332	77 / 55295	48 / 17309
Graves'	3388 / 469283	914 / 110197	1766 / 317074	422 / 55642	45 / 17946
Sjögren's	2816 / 472870	520 / 111322	1310 / 319124	157 / 56119	50 / 18187
Psoriasis	21015 / 431120	2521 / 101166	9615 / 292205	1530 / 50308	569 / 16048
Hashimoto Thyroiditis	14186 / 458485	804 / 110307	6093 / 312747	367 / 55697	317 / 17674
Lupus	3936 / 473772	1137 / 111165	2442 / 319080	225 / 56313	72 / 18237
Rheumatoid Arthritis	13184 / 443140	3280 / 101623	7071 / 297862	507 / 52176	188 / 16702
Multiple Sclerosis	2025 / 446886	475 / 101646	780 / 297888	36 / 51393	29 / 16457
Crohn's	3005 / 455308	513 / 105068	1033 / 305040	67 / 53528	66 / 16929
Polymyalgia Rheumatica	708 / 435628	–	124 / 287923	10 / 49222	–
Ulcerative Colitis	5298 / 452158	796 / 104667	2518 / 303129	192 / 52713	147 / 16845
Pediatric IBD	1147 / 449546	191 / 104216	546 / 302110	33 / 52611	37 / 16782
Axial Spondyloarthritis	1086 / 36301	47 / 3144	363 / 16630	53 / 2073	22 / 503
Vitiligo	5109 / 446859	1246 / 102158	3199 / 298421	298 / 51520	214 / 16422
Migratory Glossitis	1092 / 36473	108 / 4993	358 / 18007	39 / 2133	11 / 565
Canker Sore	28275 / 8842	2442 / 2463	11662 / 6178	1579 / 596	284 / 272
Alopecia Areata	3263 / 445400	2198 / 100221	3527 / 294859	606 / 50802	358 / 15913

Table F4: Validation dataset sample sizes (case/control) per disease (only European individuals)

Disease	EUR Validation Size
Celiac	2096 / 262416
T1D	3088 / 488880
Graves'	3469 / 506279
Sjögren's	2964 / 509623
Psoriasis	22883 / 467419
Hashimoto Thyroiditis	14555 / 495193
Lupus	4173 / 511037
Rheumatoid Arthritis	14315 / 479237
Multiple Sclerosis	2203 / 484295
Crohn's	3449 / 493246
Polymyalgia Rheumatica	662 / 472910
Ulcerative Colitis	5966 / 489957
Pediatric IBD	1285 / 487101
Axial Spondyloarthritis	1153 / 39350
Vitiligo	5320 / 484047
Migratory Glossitis	1374 / 46592
Canker Sore	36835 / 11355
Alopecia Areata	3567 / 482771

Table F5: Tuning of the training steps based on validation AUROC

Model	Epoch	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D
model-t059	2.0	**0.7732**	**0.6843**	**0.7621**	**0.7412**	**0.6295**	**0.7165**	**0.6593**	**0.5911**	**0.6401**	**0.8392**
model-t059	3.0	0.7368	0.6592	0.7323	0.7225	0.6087	0.6829	0.6415	0.5780	0.6164	0.8162
model-t15	1.8	0.7775	0.6869	0.7623	0.7417	0.6279	0.7137	0.6581	0.5901	0.6357	0.8398
model-t16	1.9	**0.7785**	0.6895	0.7632	0.7435	**0.6345**	0.7173	0.6600	0.5924	0.6371	0.8404
model-t17	2.0	0.7784	**0.6904**	**0.7648**	**0.7446**	0.6341	**0.7218**	**0.6613**	**0.5939**	0.6405	**0.8424**
model-t18	2.1	0.7463	0.6717	0.7386	0.7235	0.6147	0.6905	0.6410	0.5782	0.6164	0.8173
model-t19	2.2	0.7430	0.6666	0.7341	0.7193	0.6099	0.6829	0.6376	0.5758	0.6132	0.8132

Note. Most models (e.g., model-t059) showed signs of overfitting beyond 2 training epochs. We also explored training durations around this point (1.8–2.2 epochs) and selected 2 epochs as the final configuration.

Table F6: Tuning of attention heads and model dimension based on validation AUROC

Model	#Atten. Head	#d_model	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D	Mean
Model-t17	4	64	**0.7784**	0.6904	0.7648	0.7446	0.6341	**0.7218**	0.6613	0.5939	0.6405	**0.8424**	**0.7072**
Model-t20	4	32	0.7649	0.6794	0.7568	0.7418	0.6214	0.7081	0.6584	0.5902	0.6343	0.8371	0.6992
Model-t21	4	96	0.7754	0.6892	**0.7656**	**0.7464**	0.6334	0.7208	**0.6628**	**0.5954**	**0.6427**	**0.8424**	**0.7074**
Model-t22	3	48	0.7783	**0.6909**	0.7640	0.7449	**0.6357**	0.7192	0.6612	0.5927	0.6408	**0.8424**	0.7070
Model-t23	3	24	0.7714	0.6819	0.7527	0.7404	0.6259	0.7136	0.6558	0.5897	0.6315	0.8392	0.7002
Model-t24	3	72	0.7667	0.6865	0.7649	0.7448	0.6289	0.7131	0.6611	0.5932	0.6403	0.8397	0.7039
Model-t25	5	80	0.7659	0.6871	0.7643	0.7446	0.6278	0.7127	0.6613	0.5941	0.6406	0.8405	0.7039
Model-t26	5	40	0.7741	0.6874	0.7602	0.7426	0.6309	0.7145	0.6591	0.5912	0.6377	0.8401	0.7038

Note. We selected 4 attention heads with $d_{\text{model}} = 64$ for their competitive performance despite a smaller model dimension compared to Model-t21.

Table F7: Tuning of attention dilation and number of transformer blocks based on validation AUROC

Model	#Transformer Blocks	NA Dilation	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D	Mean
Model-t9	2	(1-1)	0.7807	**0.6938**	**0.7772**	**0.7508**	**0.6422**	**0.7297**	0.6654	0.5987	**0.6501**	0.8425	0.7131
Model-t12	2	(1-2)	0.7806	0.6904	0.7766	0.7507	0.6401	0.7292	**0.6659**	**0.5991**	0.6497	**0.8435**	0.7126
Model-t13	3	(1-2-3)	**0.7808**	0.6921	0.7749	0.7498	0.6398	0.7291	0.6650	0.5982	0.6459	0.8427	0.7118

Note. Increasing the number of transformer blocks and applying dilated attention (e.g., (1–2), (1–2–3)) led to mild overfitting. We selected 2 transformer blocks without dilation (1–1) as the final configuration.

Table F8: Tuning of the window size of neighborhood attention based on validation AUROC

Model	#Atten. Head	#NA Window	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D
Model-t75	4	385	0.7806	0.6944	**0.7700**	**0.7415**	0.6308	**0.7265**	**0.6625**	**0.5975**	0.6311	**0.8418**
Model-t127	4	129	0.7796	**0.6955**	0.7699	0.7410	**0.6310**	**0.7265**	0.6618	0.5963	0.6322	0.8412
Model-t128	4	257	**0.7821**	0.6922	0.7679	0.7403	0.6299	0.7246	0.6610	0.5963	0.6303	0.8410
Model-t129	4	513	0.7798	0.6945	0.7684	0.7412	0.6292	0.7252	0.6619	0.5969	0.6305	0.8413
Model-t130	4	641	0.7787	0.6940	0.7691	0.7411	0.6308	0.7233	**0.6625**	0.5959	**0.6342**	0.8405
Model-t131	2	129	0.7799	0.6927	0.7670	0.7398	0.6298	0.7252	0.6615	0.5954	0.6297	0.8403
Model-t132	2	257	0.7785	0.6936	0.7676	0.7408	0.6303	0.7234	0.6617	0.5962	0.6315	0.8404
Model-t133	2	513	0.7799	0.6946	0.7694	0.7413	0.6297	0.7256	0.6619	0.5961	0.6319	0.8410
Model-t134	2	641	0.7807	0.6923	0.7668	0.7404	0.6286	0.7247	0.6618	0.5948	0.6307	0.8407

Note. We selected 4 attention heads with Neighborhood Attention's window size of 385 as the final configuration.

Table F9: Exploring the impact of LD-based variant set pruning ($r^2 = 0.5$) on validation AUROC.

Model	No. variants	LD-pruned	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D
model-t031	∼344K	No	**0.7640**	**0.6103**	**0.7334**	**0.7335**	**0.6192**	**0.6717**	**0.6572**	**0.5755**	**0.6340**	**0.8328**
model-t034	∼232K	Yes	0.7564	0.5951	0.7154	0.7179	0.6143	0.6595	0.6466	0.5732	0.6300	0.8238

Note. Interestingly, removing correlated variants via LD pruning lowers AUROCs, indicating that PRSformer benefits from the underlying LD structure when identifying causal signals.

Table F10: Exploring different output heads based on validation AUROC.

Model	Output Head	Celiac	Crohn's	Graves'	Hashimoto Thyroiditis	Lupus	Multiple Sclerosis	Psoriasis	Rheumatoid Arthritis	Sjögren's	T1D
model-t5	Flatten+FC	**0.7799**	**0.6896**	**0.7773**	**0.7497**	**0.6385**	**0.7301**	**0.6655**	**0.5981**	**0.6491**	**0.8448**
model-t10	10-[CLS]-tokens+Flatten+FC	0.7343	0.5678	0.6656	0.6682	0.5840	0.6410	0.6036	0.5590	0.6007	0.8004
model-t11	Global-Avg-Pool+FC	0.5820	0.5348	0.5955	0.5822	0.5516	0.5187	0.5504	0.5292	0.5601	0.5996

Note. In model-t10, we introduced ten [CLS] tokens into the input sequence and vocabulary, each with learnable embeddings and full attention over all variant tokens. The flattened, normalized representations of these [CLS] tokens were passed to a linear layer producing an 18-dimensional output. In model-t11, an average-pooling layer was applied to the normalized representations from the last transformer block (reducing tensors from $B \times L \times d$ to $B \times d$), followed by a linear layer mapping d to the 18-dimensional output.

Table F11: Comparison of explained variance between PRSformer and LDpred2 across diseases on the European test set

Model	Celiac	T1D	Graves'	Sjögren's	Psoriasis	Hashimoto Thyroiditis	Lupus	Rheumatoid Arthritis	Multiple Sclerosis	Crohn's	Polymyalgia Rheumatica	Ulcerative Colitis	Pediatric IBD	Axial Spondyloarthritis	Vitiligo	Migratory Glossitis	Canker Sore	Alopecia Areata
PRSformer	**0.1269**	**0.2189**	**0.0885**	**0.0224**	**0.0570**	**0.1039**	0.0231	**0.0163**	**0.0868**	**0.0562**	**0.0226**	**0.0328**	0.0364	**0.0179**	**0.0289**	0.0898	0.0555	**0.0165**
LDpred2	0.0870	0.1837	0.0874	0.0194	0.0545	0.1038	**0.0236**	0.0143	0.0687	0.0513	0.0178	0.0324	**0.0401**	0.0146	0.0275	**0.0927**	**0.0591**	0.0133

Note. Predicted probabilities from both models were calibrated on the test data.

Table F12: Number of selected variants per disease across non-European ancestries

Disease	AFR	LAT	EAS	SAS
Celiac	0	7515	0	0
T1D	7067	8546	0	0
Graves'	6964	8889	6611	0
Sjögren's	5926	6645	0	0
Psoriasis	6211	10545	6524	6041
Hashimoto Thyroiditis	6342	11037	5780	0
Lupus	6172	7413	0	0
Rheumatoid Arthritis	6342	8104	4453	0
Multiple Sclerosis	5936	6730	0	0
Crohn's	5177	4959	0	0
Polymyalgia Rheumatica	0	0	0	0
Ulcerative Colitis	5254	5743	0	0
Pediatric IBD	0	4862	0	0
Axial Spondyloarthritis	0	0	0	0
Vitiligo	5599	7017	0	0
Migratory Glossitis	0	6927	0	0
Canker Sore	5732	6812	4296	5021
Alopecia Areata	5705	7521	4626	0

Table F13: Characteristics of multitask models trained on down-sampled datasets with subsetted input variants

Model	# Input Variants	Train Data Size	# Model Parameters
Non-linear_downsampled	94176	3000000	120.61M
Non-linear_downsampled	38018	2000000	48.73M
Non-linear_downsampled	23397	1500000	30.01M
Non-linear_downsampled	15060	1000000	19.34M
Non-linear_downsampled	9181	500000	11.82M
Non-linear_downsampled	6829	300000	8.81M
Non-linear_downsampled	4863	150000	6.29M
Linear_downsampled	94176	3000000	120.55M
Linear_downsampled	38018	2000000	48.66M
Linear_downsampled	23397	1500000	29.95M
Linear_downsampled	15060	1000000	19.28M
Linear_downsampled	9181	500000	11.75M
Linear_downsampled	6829	300000	8.74M
Linear_downsampled	4863	150000	6.22M

Table F14: AUROC comparison across diseases under different covariate settings (sex and age). For each disease and ancestry group, bold font denotes the best performance.

Ancestry	Model	Covariates	Celiac	T1D	Graves'	Multiple Sclerosis	Pediatric IBD	Rheumatoid Arthritis	Vitiligo	Polymyalgia Rheumatica	Lupus	Axial Spondyloarthritis	Crohn's	Hashimoto Thyroiditis	Psoriasis	Canker Sore	Alopecia Areata	Ulcerative Colitis	Sjögren's	Migratory Glossitis
EUR	PRSformer	none	0.7933	0.8457	0.7661	0.7184	0.6660	0.6004	0.6212	0.6048	0.6338	0.5802	0.6985	0.7504	0.6669	0.6227	0.6051	0.6442	0.6337	0.7164
EUR	PRSformer	sex	**0.8051**	0.8469	0.7997	0.7392	0.6707	0.6197	0.6264	0.6139	0.7199	0.5697	**0.7007**	0.8063	0.6684	0.6266	0.6262	0.6493	0.7292	**0.7248**
EUR	PRSformer	sex+age	0.8037	**0.8478**	**0.8224**	**0.7575**	0.7098	0.7208	0.6556	0.8536	0.7328	0.6174	0.7000	**0.8123**	0.6730	0.6274	0.6352	0.6757	0.7711	0.7213
EUR	PRSformer-ME	none	0.7867	0.8431	0.7674	0.7157	0.6654	0.6069	0.6222	0.5995	0.6458	0.5818	0.6981	0.7487	0.6683	0.6246	0.6125	0.6447	0.6404	0.7146
EUR	PRSformer-ME	sex	0.7972	0.8418	0.7996	0.7330	0.6651	0.6249	0.6254	0.6088	0.7259	0.5768	0.6991	0.8044	0.6687	0.6296	0.6347	0.6476	0.7302	0.7197
EUR	PRSformer-ME	sex+age	0.7993	0.8420	**0.8224**	0.7549	**0.7124**	**0.7265**	**0.6668**	**0.8565**	**0.7425**	**0.6271**	0.7002	0.8120	**0.6758**	**0.6315**	**0.6433**	**0.6763**	**0.7766**	0.7097
AFR	PRSformer	none	0.6219	0.6391	0.5676	0.5447	0.5928	0.5241	0.5122	–	0.5404	<0.5	0.5726	0.6727	0.5559	0.5913	<0.5	0.5401	0.5118	0.6008
AFR	PRSformer	sex	0.6346	0.6496	0.6153	0.5709	0.5983	0.5630	0.5207	–	0.6500	<0.5	0.5755	0.7473	0.5674	0.5932	<0.5	0.5554	0.6068	0.6054
AFR	PRSformer	sex+age	0.6412	0.6589	0.6964	0.5997	0.6161	0.7445	0.5699	–	0.6917	<0.5	0.5790	0.7695	0.5797	0.5968	0.5322	0.5971	0.6923	0.6110
AFR	PRSformer-ME	none	0.6269	0.6986	0.6997	0.6397	0.6285	0.5660	0.5605	–	0.6140	0.5410	**0.6217**	0.6939	0.5822	0.6026	0.5844	0.5902	0.6398	0.6550
AFR	PRSformer-ME	sex	**0.6576**	0.7046	0.7431	0.6834	0.6215	0.5971	0.5779	–	0.7007	0.5460	0.6145	0.7559	0.5872	0.6070	0.6243	0.6021	0.7026	**0.6577**
AFR	PRSformer-ME	sex+age	0.6446	**0.7092**	**0.7893**	**0.7150**	**0.6647**	**0.7662**	**0.6350**	–	**0.7346**	**0.6374**	**0.6217**	**0.7782**	**0.5972**	**0.6108**	**0.6563**	**0.6494**	**0.7688**	0.6463
LAT	PRSformer	none	0.7437	0.7459	0.7060	0.6832	0.6216	0.5712	0.5871	0.6526	0.6063	0.5610	0.6674	0.7482	0.6449	0.6283	0.5579	0.6184	0.6181	0.6994
LAT	PRSformer	sex	0.7599	0.7491	0.7486	0.7055	0.6261	0.6042	0.5882	0.6641	0.6930	0.5689	0.6674	0.8000	0.6512	0.6299	0.5751	0.6262	0.6982	0.7088
LAT	PRSformer	sex+age	0.7573	0.7580	0.7875	0.7316	0.6490	0.7516	0.6226	0.8390	0.7225	0.6270	0.6697	0.8135	0.6601	0.6299	0.5985	0.6661	0.7613	0.7083
LAT	PRSformer-ME	none	0.7521	0.7732	0.7568	0.7256	0.6539	0.5979	0.6166	0.6553	0.6536	0.6038	0.6842	0.7581	0.6596	0.6397	0.6346	0.6291	0.6685	0.7270
LAT	PRSformer-ME	sex	**0.7649**	0.7762	0.7915	0.7427	0.6548	0.6300	0.6210	0.6756	0.7335	0.5962	0.6827	0.8081	0.6601	0.6417	0.6387	0.6294	0.7382	**0.7279**
LAT	PRSformer-ME	sex+age	0.7642	**0.7792**	**0.8170**	**0.7658**	**0.6819**	**0.7671**	**0.6601**	**0.8588**	**0.7616**	**0.6700**	**0.6900**	**0.8222**	**0.6713**	**0.6426**	**0.6522**	**0.6736**	**0.7962**	0.7241
SAS	PRSformer	none	0.8180	0.6640	0.7667	0.7341	0.6423	0.6052	0.5693	–	0.5656	0.5999	0.6630	0.6705	0.6209	0.5999	0.5754	**0.6974**	0.6383	0.5421
SAS	PRSformer	sex	**0.8183**	0.6550	0.7756	0.7496	0.6330	0.6455	0.5805	–	0.6754	0.5857	0.6618	0.7377	0.6224	0.6053	0.5843	0.6865	0.7393	**0.5915**
SAS	PRSformer	sex+age	0.8042	0.6685	0.8185	0.7536	0.7044	0.7479	0.5993	–	0.6860	0.6193	0.6684	0.7479	0.6186	**0.6146**	0.5867	0.6862	0.7593	0.5706
SAS	PRSformer-ME	none	0.8002	**0.6852**	0.7612	0.7503	0.6858	0.5974	0.6009	–	0.6196	0.5828	**0.7061**	0.6907	0.6269	0.5813	0.5972	0.6969	0.6885	0.5146
SAS	PRSformer-ME	sex	0.8108	0.6743	0.7784	**0.7813**	0.6663	0.6666	0.5935	–	0.7053	0.6221	0.6732	0.7452	0.6230	0.6069	0.6042	0.6653	0.7423	0.5492
SAS	PRSformer-ME	sex+age	0.8096	0.6796	**0.8231**	0.7737	**0.7561**	**0.7632**	**0.6104**	–	**0.7097**	**0.6559**	0.6748	**0.7580**	**0.6325**	0.6032	**0.6041**	0.6861	**0.7701**	0.5348
EAS	PRSformer	none	–	0.6507	0.6937	0.5406	0.5902	0.5566	0.5463	0.7431	0.6144	0.5184	**0.6282**	0.6535	0.6167	0.5731	0.5846	0.6204	0.5770	0.6809
EAS	PRSformer	sex	–	0.6357	0.7357	0.6136	0.6106	0.5808	0.5425	0.7208	0.6893	<0.5	0.5928	0.7335	0.6182	0.5642	0.5986	0.6027	0.6541	0.6659
EAS	PRSformer	sex+age	–	0.6588	0.7585	**0.6612**	0.6153	0.7759	0.5902	**0.8856**	0.7164	0.5640	0.5932	0.7491	0.6298	0.5773	0.6154	0.6607	0.7467	0.6860
EAS	PRSformer-ME	none	–	0.7121	0.7487	0.5019	**0.6518**	0.6085	0.5602	0.6973	0.6573	0.5860	0.6256	0.6876	0.6355	0.5929	0.6386	0.6545	0.6176	**0.7085**
EAS	PRSformer-ME	sex	–	0.7001	0.7828	0.5988	0.6385	0.6380	0.5725	0.6981	0.7295	0.6081	0.5960	0.7518	0.6331	**0.5973**	0.6379	0.6345	0.6918	0.6938
EAS	PRSformer-ME	sex+age	–	**0.7157**	**0.7964**	0.6455	0.6222	**0.7978**	**0.6193**	0.8587	**0.7450**	**0.6586**	0.6001	**0.7729**	**0.6449**	**0.5972**	**0.6662**	**0.6980**	**0.7742**	0.7041

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim the development of a scalable Transformer (PRSformer) for disease prediction from large-scale genotypes, outperforming linear and summary-statistic methods, demonstrating benefits of non-linearity at scale and multitask learning, and improving cross-ancestry generalization. These claims are supported by the experimental results presented in Section 4 and associated figures/tables.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 explicitly discusses limitations, namely the inability to share the private dataset and the need for future work on interpretability methods.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper presents an empirical study based on a novel architecture and its experimental validation. It does not include theoretical results, theorems, or formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides substantial detail on the model architecture (Section 3.2, Fig 1), datasets and filtering (Section 3.3, Appendix A), variant selection (Section 3.4, Appendix A-B, and D), training procedure (Section 3.5), baseline implementations (Section 3.6, Appendix C), and hyperparameter selection (Section 3.5, and Appendix F, Supplementary Tables F5-F10). The GitHub repository containing the model architecture is also referenced in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper states the dataset is from a large, private biobank and cannot be shared publicly (Section 3.3, Section 5). The implementation code has been released with the accepted version (Section 5), although reproducing the main experimental results remains infeasible without access to the private data.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 provides details on data splits (3.3), variant selection (3.4), training procedure including optimizer and scheduler (3.5), and model architecture (3.2). Appendix F and associated supplementary tables provide extensive details on hyperparameter tuning and selection.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Figures 2 and E1-E3 report performance metrics (AUROC, AUPRC) with error bars explicitly defined as 95% confidence intervals calculated via 10k bootstraps. P-values for significance testing (one-sided paired bootstrap test) comparing PRSformer to LDPred2 are also provided (Section 4.1, Figure 2 caption).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.5 mentions training was performed on four NVIDIA A100 GPUs using Distributed Data Parallel and FP16 for 2 epochs ($\approx 120,000$ gradient updates). While total wall-clock time isn't explicitly stated, the key hardware and training duration metrics are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully conforms to the NeurIPS Code of Ethics. It uses consented data under IRB approval (Section 3.3, 5) and discusses ethical considerations like data privacy, potential misuse, and fairness (Section 5).

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 discusses potential positive impacts (disease prediction/discovery) and negative impacts/risks (genomic data misuse, discrimination). It also touches upon fairness across populations and responsible deployment.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper uses sensitive genomic data. Safeguards mentioned include using data from a private biobank under IRB approval (Section 3.3, Section 5) and explicitly stating that the trained models are withheld to mitigate misuse potential (Section 5).

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The primary asset is the private biobank data obtained under IRB protocol. The methods use standard libraries/techniques (e.g., PyTorch, AdamW) or cited methods (e.g., LDPred2 [5], NATTEN [12, 44]) for which explicit license discussion within the paper text is not standard practice or necessary. No other major external assets requiring specific license attribution appear to be used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new model architecture (PRSformer), which is well-documented (Section 3.2, Figure 1, and Appendix F). Furthermore, the code is available at GitHub (Section 5). However, no new publicly released datasets are provided with the submission (data is private and trained models are withheld).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research uses existing data from a biobank where individuals previously consented to participate in research (Section 3.3 and 5). The study did not involve new recruitment, direct interaction with human subjects, or crowdsourcing.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper explicitly states that the data was utilized under an IRB-approved protocol (Sections 3.3) and references the use of consented research participant data (Section 5). Given it's secondary analysis of existing, likely de-identified data under IRB oversight, specific risks incurred by this study are minimal and typically covered by the initial consent/IRB review; the paper focuses on broader ethical considerations of genomic data (Section 5).

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methodology involves a custom Transformer architecture (PRSformer) with neighborhood attention, trained directly on genotype data. No LLMs were used as a core component of the method development or experiments.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.