A computational framework for estimating and testing specific parametric structures in generalized additive model

Yihe Yang and Xiaofeng Zhu
Department of Population and Quantitative Health Sciences,
Case Western Reserve University School of Medicine

July 14, 2025

Abstract

Before applying flexible nonparametric models such as a generalized additive model (GAM), it is natural to ask whether a simpler parametric form suffices to capture the underlying patterns in the data. To address this question, we develop TAPS (Test for Arbitrary Parametric Structure), a computational framework that integrates estimation and hypothesis testing to assess whether a prespecified parametric structure sufficiently captures a smooth effect in a generalized additive model. Notably, TAPS accommodates a wide range of structures, including linearity, piecewise linearity with slope changes, and discontinuities with jumps, and provides an implementation for estimation and testing, by building directly on the `mgcv`, the state-of-the-art implementation of a GAM. We demonstrate the utility of TAPS through two analyses based on the UK Biobank database, which includes hundreds of thousands of individuals and tens of thousands of phenotypic measurements. First, we apply TAPS to polygenic risk score analyses and uncover pervasive nonlinear effects for multiple complex traits, although the improvement in predictive accuracy over linear models is generally modest. Second, TAPS reveals complex gene-by-age interactions for these complex traits, some of which even extend beyond traditional linear interaction models.

Keywords: Generalized additive model; gene-environment interaction; hypothesis test; polygenic risk score; vary-coefficient model;

1 Introduction

Generalized additive model (GAM) (Wood, 2017), an extension of the generalized linear model (GLM) (Nelder and Wedderburn, 1972), is widely used to model the non-linear effects of covariates on an outcome. Since its first introduction by Hastie and Tibshirani (1990), GAM has undergone substantial methodological and computational development. For instance, the R package `mgcv` exemplifies these advances, which provides a broad class of smoothers (Wood, 2003, 2006) and automatic smoothing parameter selection using information criteria (Wood, 2004, 2011). Beyond methodological innovations, GAMs have also found wide applicability across scientific disciplines. For example, in epidemiology, GAMs have been used to smooth daily incidence data and to reveal trends in childhood type 1 diabetes during the COVID-19 pandemic (McKeigue et al., 2023). In ecology, GAMs have been employed to model the nonlinear effects of environmental factors in chaotic or near-chaotic ecological dynamic systems (Wood, 2010). In spatial transcriptomics, GAMs are used to characterize the spatial distribution of cell types across tissues (Shang et al., 2025).

Before applying a nonparametric model such as a GAM, it is natural to ask whether a simpler parametric model can already describe the variation in the data. This is because a parametric model, if correctly specified, is often of greater scientific interest due to its interpretability and simplicity. For example, the genetic effect of an exposure is often assumed to be in a linear relationship with an outcome (Choi et al., 2020), yet growing evidence suggests the presence of non-linear genetic effects (Sulc et al., 2022). There is an ongoing discussion as to whether linear additive genetic models are sufficient to capture the majority of heritability of human complex traits (Hill et al., 2008; Zuk et al., 2012). In addition, beyond interpretability and simplicity, some statistical models, such as causal inference, are based on specific parametric forms. For example, in regression discontinuity design (RDD), the effect of treatment is modeled as a sudden jump at a

cutoff point (Huntington-Klein, 2021). The causal interpretation relies on this specific parametric form. Nonparametric models may better capture complex associations, but do not, on their own, enable causal inference of treatment assignment in RDD.

Compared to the wide variety of estimation methods available for GAM, the tools for hypothesis testing remain relatively limited. Most existing hypothesis tests focus on assessing whether a smooth function is null (i.e., identically zero) (Wood, 2013), rather than whether it conforms to a specific parametric structure of scientific interest. To the best of our knowledge, among all parametric function forms with potential practical interpretations, existing methods only allow testing whether the effect of a covariate follows a polynomial structure (Cox et al., 1988; Zhang and Lin, 2003; Crainiceanu et al., 2005; Nummi et al., 2011). However, for other parametric structures, such as piecewise linearity or linear discontinuities, there are no established approaches for conducting corresponding hypothesis tests. Moreover, existing hypothesis tests are restricted to univariable nonparametric models, where only one nonlinear term is allowed to exist in the model. How to perform hypothesis testing for arbitrary parametric structures within a GAM remains an open question.

We develop the Test for Arbitrary Parametric Structure (TAPS), a computational framework that integrates estimation and formal hypothesis testing to assess whether a user-specified parametric structure adequately explains a smooth effect in a GAM. TAPS is not designed to discover the best-fitting structure from data, but rather to formally test whether a given structural form can sufficiently describe the variation of the data, for reasons of interpretability and generalizability. The methodological innovation of TAPS is providing a new hypothesis assessing whether a parametric form can adequately capture a target function within a GAM. Especially, we develop an R package, `mgcv.taps`, which integrates seamlessly with `mgcv`, one of the most powerful GAM software. This integration enables TAPS to be easily adopted without additional learning cost,

to support a wide range of outcome types including survival and ordinal responses (Wood et al., 2016), and to scale efficiently to biobank-scale datasets (Wood et al., 2017).

To demonstrate the practical utility of TAPS, we conducted two empirical analyses using data from the UK Biobank, which includes hundreds of thousands of individuals and tens of thousands of phenotypic measurements (Bycroft et al., 2018). First, we found that the relationship between a trait and its polygenic risk score (PRS) was often non-linear, but the prediction improvements were generally limited for most traits. Second, we employed TAPS to identify the complex interaction between PRSs and age beyond traditional linear interaction. To facilitate practical adoption, we present three examples in the **Supplementary Materials**, showcasing the end-to-end application of TAPS across different scenarios.

2 Method

2.1 Overview of TAPS

The model of a GAM is

$$
\begin{aligned}
&y_i \sim \mathcal{Y}(\mu_i, \phi_0), \\
&g(\mu_i) = \eta_i = \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + f_1(x_{i1}) + \cdots + f_J(x_{iJ}),
\end{aligned} \tag{1}
$$

where y_i is the response belonging to a certain distribution $\mathcal{Y}(\mu_i, \phi_0)$ with mean μ_i and dispersion parameter ϕ_0, $g(\cdot)$ is a known link function, η_i is a linear predictor, $\boldsymbol{Z}_i$ is the ith row of a parametric model matrix $\mathbf{Z}$, $\boldsymbol{\gamma}$ is a regression coefficient vector, and $f_j(\cdot)$ is a smooth function of covariate x_{ij}, $j = 1, \ldots, J$.

TAPS performs a hypothesis test on whether a target function in (1), such as $f_1(\cdot)$, adheres to

a parametric structure. Specifically, the null and alternative hypotheses that TAPS addresses are

$$\mathrm{H}_0 : f_1 \in \mathcal{H}_A \qquad \text{v.s.} \qquad \mathrm{H}_1 : f_1 \in \mathcal{H} \text{ and } f_1 \notin \mathcal{H}_A, \tag{2}$$

where $\mathcal{H}_A$ is the space of functions with the target parametric structure and $\mathcal{H}$ is a general functional space. In particular, $\mathcal{H}$ should be sufficiently large to accommodate the parametric functions in $\mathcal{H}_A$ (which may even be nonsmooth), but also capable of approximating smooth functions when $\mathcal{H}_A$ does not reflect the true structure. To achieve this, we propose a new mixed-effects representation of $f_1(x)$:

$$f_1(x_1) = \sum_{s=1}^{q} A_{1s}(x_1)\alpha_{1s} + b_1(x_1), \tag{3}$$

where $\boldsymbol{A}_1(x_1) = (A_{11}(x_1), \ldots, A_{1q}(x_1))^\top$ is a vector of the basis functions of the parametric structure, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1q})^\top$ is a fixed vector, and $b_1(x_1)$ is a non-parametric term orthogonal to $\sum_{s=1}^{q} A_{1s}(x_1)\alpha_{1s}$. According to the equivalence between reproducing kernel Hilbert space (RKHS) and Gaussian processes (Kanagawa et al., 2018), $b_1(x)$ can be regarded as a Gaussian process with a covariance function

$$\mathrm{cov}(b_1(x_{i1}), b_1(x_{j1})) = \phi_1 K_1(x_{i1}, x_{j1}), \tag{4}$$

where $K_1(x_i, x_j)$ is a known bivariate function, and $\phi_1 \geq 0$ is the variance of this Gaussian process. We call the equation (3) a mixed-effect representation of a function, as it shares the same form as the generalized linear mixed model (GLMM) (Breslow and Clayton, 1993). As a result, the hypothesis (2) reduces to

$$\mathrm{H}_0 : \phi_1 = 0 \qquad \text{v.s.} \qquad \mathrm{H}_1 : \phi_1 > 0. \tag{5}$$

Thus, the techniques of GLMM, including the Wald test (Wood, 2013) and score test (Zhang and Lin, 2003), can yield the p-value of (5).

TAPS can also evaluate whether there is a complex interaction between two covariates beyond the simple linear interaction and other parametric forms. Specifically, for a bivariate function

$f_1(x_1, z_1)$, it is usually to simplify it as

$$f_1(x_1, z_1) = z_1 \cdot \beta_1(x_1), \tag{6}$$

where $\beta_1(x_1)$ is an unknown function. This model is known as the varying coefficient model Hastie and Tibshirani (1993). When $\beta_1(x_1) = x_1$, this model reduces to the traditional interaction model, which is commonly considered in current genome-wide environmental interaction (GWEI) studies (Bentley et al., 2019; Zhu et al., 2024). Similar to the hypothesis test (2), TAPS can test

$$\mathrm{H}_0 : \beta_1 \in \mathcal{H}_A \qquad \text{v.s.} \qquad \mathrm{H}_1 : \beta_1 \in \mathcal{H} \text{ and } \beta_1 \notin \mathcal{H}_A, \tag{7}$$

where $\mathcal{H}_A$ is a parametric functional space and $\mathcal{H}$ is a general functional space. Furthermore, we consider the same mixed-effects representation of $\beta_1(x_1)$ as (3):

$$\beta_1(x_1) = \sum_{s=1}^{q} A_{1s}(x_1)\alpha_{1s} + b_1(x_1), \tag{8}$$

and we evaluate if $\beta_1(x_1)$ can be sufficiently described by the parametric part $\sum_{s=1}^{q} A_{1s}(x_1)\alpha_{1s}$ by testing whether the variance of the random effect $b_1(x_1)$ is zero or not.

Figure 1A shows a classic example used in literature, the motorcycle acceleration data, which records head acceleration during a simulated crash (Wood et al., 2016). In this example, our goal is to test whether the time trend can be segmented into four phases, with changepoints marked by vertical dashed lines. The null hypothesis assumes that the function lies within a piecewise linear space $\mathcal{H}_A$, whose basis functions are shown in Figure 1C. To test for potential deviations from this structure, TAPS constructed $b_1(x)$ from a space $\mathcal{H}_B = \mathcal{H}/\mathcal{H}_A$, with basis functions adaptively learned from the data, as shown in Figure 1D. Figure 1B displays the fitted trend under the mixed representation. The first three phases aligned with the piecewise linear form, while the fourth showed a slightly non-linear effect. Figure 1E summarizes the overall TAPS workflow.

Methodologically, we address two key challenges in the implementation of TAPS. First, we

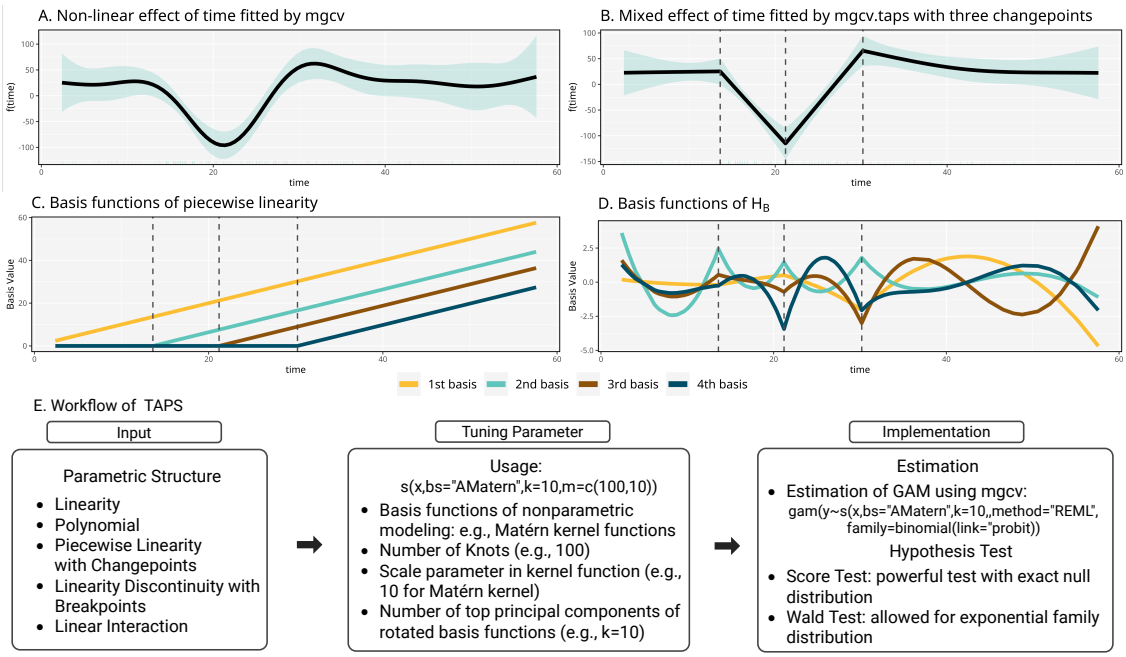

Figure 1: Overview of the TAPS framework. This figure uses the `mcycle` data, which records head acceleration over time in a simulated motorcycle crash experiment, to illustrate the workflow of TAPS. This data is provided by the R package `MASS` and is commonly used for benchmarking smooth curve fitting methods. **A**. The fitted smooth function of time using `mgcv` with default smooth terms. **B**. The fitted smooth function of time using `mgcv.taps` with three user-specified changepoints. **C**. Basis functions of the space of piecewise linearity space $\mathcal{H}_A$. **D**. Basis functions from the kernel-smoothed space $\mathcal{H}_B$, automatically learned via principal component decomposition. **E**. Summary of the TAPS components. Input allows users to specify parametric structure in the adaptive smooth term, such as linearity, polynomial trends, piecewise linearity, or discontinuities. Tuning Parameter panel outlines the usage of the adaptive smooth term with key options, including the number of knots and kernel parameters. Implementation involves model fitting using `mgcv` and hypothesis testing via either the score test or the Wald test.

show how to construct the random effect $b_1(x)$ that is adaptively orthogonal to $\boldsymbol{A}_1(x)$, guaranteeing an identifiable mixed-effect representation. We have developed the R package `mgcv.taps`, which enables `mgcv` to recognize the new mixed-effect representation (3). Second, we explore the methodology to conduct the hypothesis test (5) within a GAM model. We design two testing procedures for assessing model terms: a Wald test (Wood, 2013) and a score test (Zhang and Lin, 2003). Both tests control type I error rates well, but differ in applicability and power. The score test is generally more powerful, yet currently limited to outcomes from the exponential family. In contrast, the Wald test supports all outcome types implemented in `mgcv`. By offering both

options, our framework allows users to choose the method best suited to their analysis objectives.

2.2 Construction of Basis Functions

The construction of $b_1(x)$ starts from representing it using the RKHS (Wahba, 1990; Wainwright, 2019), under the alternative hypothesis that $b_1(x)$ is a general smooth function. According to the functional analysis theory, every RKHS $\mathcal{H}_K$ corresponds to a unique semi-positive definite kernel function $K(\cdot,\cdot)$ defined in the Cartesian product space $[0,1]\times[0,1]$. For any function $b_1 \in \mathcal{H}_K$,

$$\langle b_1, K(\cdot, x)\rangle_{\mathcal{H}_K} = b_1(x), \tag{9}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ is the inner product in the space $\mathcal{H}_K$. This property ensures that for n observation $\{x_{11},\ldots,x_{1n}\} \subset [0,1]$, there must exist an $(n\times n)$ matrix $\mathbf{K}_1$ with the (i,j)th element being $K(x_{1i}, x_{1j})$ and an $(n\times 1)$ vector $\boldsymbol{\theta}_1 = (\theta_{11},\ldots,\theta_{1n})^\top$ such that

$$\boldsymbol{b}_1 = \mathbf{K}_1\boldsymbol{\theta}_1, \tag{10}$$

where $\boldsymbol{b} = (b(x_1),\ldots,b(x_n))^\top$.

We construct the basis functions of $\mathcal{H}_B = \mathcal{H}_K/\mathcal{H}_A$ by finding an $((n+q)\times n)$ orthogonal matrix $\mathbf{Q}$ satisfying the following linear constraint

$$\mathbf{A}_1^\top\mathbf{F}_1\mathbf{Q}_1 = \mathbf{0}_{q\times n}, \tag{11}$$

where $\mathbf{F}_1 = (\mathbf{A}_1, \mathbf{K}_1)$ is the combination of $\mathbf{A}_1$ and $\mathbf{K}_1$. Here, the orthogonal matrix $\mathbf{Q}_1$ describes a special linear transformation that resorts the columns of $\mathbf{F}_1$, i.e., the basic functions of $\mathcal{H}$, into a new one perpendicular to $\mathbf{A}_1$. As a result, the columns of $\mathbf{F}_1\mathbf{Q}_1$ can be regarded as the basis functions of $\mathcal{H}_B$ and thus it can represent $\boldsymbol{b}_1 = \mathbf{F}_1\mathbf{Q}_1\boldsymbol{\pi}_1$ where $\boldsymbol{\pi}_1$ is an unconstrained vector in $\mathbb{R}^n$. The QR decomposition can yield a candidate of the orthogonal matrix $\mathbf{Q}_1$. Consider

the following QR decomposition:

$$\mathbf{F}_1^\top \mathbf{A}_1 = (\mathbf{Q}_{11}, \mathbf{Q}_{12}) \begin{pmatrix} \mathbf{R}_{11} \\ \mathbf{0}_{n\times q} \end{pmatrix}, \tag{12}$$

where $\mathbf{Q}_{11}$ is an $((n+q)\times q)$ orthogonal matrix, $\mathbf{Q}_{12}$ is an $((n+q)\times n)$ orthogonal matrix, and $\mathbf{R}_{11}$ is an $(q\times q)$ upper triangular matrix. Here $\mathbf{Q}_1$ is an orthogonal basis matrix of span$(\mathbf{F}_1^\top \mathbf{A}_1)$ and $\mathbf{Q}_{12}$ is an orthogonal basis matrix of $\overline{\text{span}}(\mathbf{F}_1^\top \mathbf{A}_1)$, thus we can choose $\mathbf{Q}_1 = \mathbf{Q}_{12}$.

Mathematically, there is a well-known equivalence between a smooth function in an RKHS induced by a kernel function $K(\cdot,\cdot)$, and a Gaussian process defined on an L_2 space with covariance function $K(\cdot,\cdot)$ (Kanagawa et al., 2018). That is, for a smooth function $b_1(x) \in \mathcal{H}_K$, the n realizations of it can be regarded as a Gaussian random effect

$$\boldsymbol{b}_1 \sim \mathcal{N}(\mathbf{0}, \phi_1 \mathbf{K}_1). \tag{13}$$

However, $b_1(x)$ is not a free function in $\mathcal{H}_K$: it should be orthogonal to the fixed effect $\boldsymbol{A}_1^\top \alpha_1$. Hence, there is an additional step to construct the prior distribution of a constrained function $b_1(x)$. Specifically, since $\boldsymbol{b}$ also falls within $\mathcal{H} = \mathcal{H}_A \cup \mathcal{H}_K$, it can be alternatively represented as

$$\boldsymbol{b}_1 = \mathbf{F}_1 \tilde{\boldsymbol{\vartheta}}_1 = \mathbf{A}_1 \tilde{\boldsymbol{\alpha}}_1 + \mathbf{K}_1 \tilde{\boldsymbol{\theta}}_1, \tag{14}$$

where the vector $\tilde{\boldsymbol{\vartheta}}_1 = (\tilde{\boldsymbol{\alpha}}_1^\top, \tilde{\boldsymbol{\theta}}_1^\top)^\top$. The constraint to ensure $\boldsymbol{b}_1$ fully fall within $\mathcal{H}_B$ is

$$\mathbf{A}_1^\top \mathbf{F}_1 \tilde{\boldsymbol{\vartheta}}_1 = \mathbf{0}_{q\times 1}. \tag{15}$$

This constraint is similar to the boundary condition of the polynomial smoothing spline (Wainwright, 2019, Example 12.29), but it has a clearer meaning: the random effect $\boldsymbol{b}$ represented by $\mathbf{F}_1 \tilde{\boldsymbol{\vartheta}}_1$ is perpendicular to the fixed parametric effect $\mathbf{A}_1 \boldsymbol{\alpha}_1$. Based on this random effect representation of smooth functions in RKHS, the augmented vector $\tilde{\boldsymbol{\vartheta}}_1$ should follow $\mathcal{N}(\mathbf{0}_{(n+q)\times 1}, \phi \tilde{\mathbf{K}}_1^+)$ where $\tilde{\mathbf{K}}_1 = \text{diag}(\mathbf{0}_{q\times q}, \mathbf{K}_1)$ is an $((n+q)\times(n+q))$ block-diagonal matrix. The block corresponding to $\tilde{\boldsymbol{\alpha}}_1$ is a zero matrix because $\mathbf{A}_1 \tilde{\boldsymbol{\alpha}}_1$ is a fixed effect and hence this is no randomness

of $\tilde{\boldsymbol{\alpha}}_1$. Furthermore, since $\tilde{\boldsymbol{\vartheta}}_1 = \mathbf{Q}_1\boldsymbol{\pi}_1$, the prior distribution of $\boldsymbol{\pi}_1$ is then

$$\boldsymbol{\pi}_1 \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \phi_1(\mathbf{Q}_1^\top \tilde{\mathbf{K}}_1 \mathbf{Q}_1)^+). \tag{16}$$

Subsequently, since $\boldsymbol{b}_1 = \mathbf{F}_1\mathbf{Q}_1\boldsymbol{\pi}_1$, the prior distribution of $\boldsymbol{b}_1$ is

$$\boldsymbol{b}_1 \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \phi_1 \mathbf{F}_1\mathbf{Q}_1(\mathbf{Q}^\top \tilde{\mathbf{K}}_1 \mathbf{Q}_1)^+ \mathbf{Q}_1^\top \mathbf{F}_1^\top). \tag{17}$$

Thus, testing whether $\boldsymbol{f}_1 = \mathbf{A}_1\boldsymbol{\alpha}_1$ reduces to testing whether $\phi_1 = 0$.

2.3 Dimension Reduction

The construction of $\boldsymbol{b}_1$ requires the QR decomposition of the $((n+q)\times q)$ matrix $\mathbf{F}_1^\top\mathbf{A}_1$, which can be computationally expensive when n is large. To address this numerical challenge, we adopt two dimension-reduction techniques to simplify the representation of $\boldsymbol{b}_1$. First, we use a low-rank RKHS approximation (Kammann and Wand, 2003) to express the smooth component. Second, we apply principal component analysis (PCA) for further dimension reduction, following the strategy used by Wood (2003). See **Supplementary Materials** for details.

2.4 Score Test

We extend the score test proposed by Zhang and Lin (2003) to evaluate whether the smooth deviation term $b_1(x)$ is equal to zero. This test is based on the equivalence between GAM and GLMM. Specifically, the score test is constructed from the following GLMM model:

$$\boldsymbol{Y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{18}$$

where $\mathbf{A}$ is the design matrix for fixed effects $\boldsymbol{\alpha}$, and $\mathbf{B}$ is the design matrix for random effects $\boldsymbol{\beta}$. The noise term follows a Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \phi_0\mathbf{W}^{-1})$, where ϕ_0 is the dispersion parameter, $\mathbf{W} = \text{diag}(W_1, \ldots, W_n)$ and $W_i = 1/(V(\mu_i)g'(\mu_i)^2)$, with $V(\mu)$ denoting the variance function and g the canonical link function. For the GAM (1), $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_J, \mathbf{Z})$,

$\mathbf{B} = (\mathbf{F}_1\mathbf{Q}_1, \ldots, \mathbf{F}_J\mathbf{Q}_J)$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_J^\top, \boldsymbol{\gamma}^\top)^\top$, and $\boldsymbol{\beta} = (\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_J^\top)^\top$. Since both $\mathbf{B}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are random, it is reasonable to consider them as a new noise $\boldsymbol{\varepsilon} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, whose prior distribution is $\mathcal{N}(\mathbf{0}_{n\times 1}, \mathbf{V}(\boldsymbol{\phi}))$ with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)^\top$ and

$$\mathbf{V}(\boldsymbol{\phi}) = \phi_0\mathbf{W}^{-1} + \sum_{j=1}^{J} \phi_j \mathbf{F}_j\mathbf{Q}_j\mathbf{S}_j\mathbf{Q}_j^\top\mathbf{F}_j^\top, \tag{19}$$

where $\mathbf{S}_j$ is the penalty matrix for the jth smooth term (Wood, 2017). (Here $\mathbf{S}_1 = (\mathbf{Q}_1^\top\tilde{\mathbf{K}}_1\mathbf{Q}_1)^+$ For the target smooth term $\boldsymbol{f}_1$.) The restricted maximum likelihood (REML) function of this GLMM is:

$$\begin{aligned}\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}}) = &-\frac{1}{2}(\boldsymbol{Y} - \mathbf{A}\boldsymbol{\alpha})^\top\mathbf{V}(\boldsymbol{\phi})^{-1}(\boldsymbol{Y} - \mathbf{A}\boldsymbol{\alpha}) \\ &- \frac{1}{2}\log\det(\mathbf{V}(\boldsymbol{\phi})) - \frac{1}{2}\log\det\{\mathbf{A}^\top\mathbf{V}(\boldsymbol{\phi})^{-1}\mathbf{A}\},\end{aligned} \tag{20}$$

The score test is based on the score equation $\mathbf{U}(\cdot)$ and Hessian matrix $\mathbf{H}(\cdot,\cdot)$ of the REML function (20):

$$\begin{aligned}\mathbf{U}(\phi_j) &= \left.\frac{\partial\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}})}{\partial\phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j} \\ &= \underbrace{\frac{1}{2}(\widehat{\boldsymbol{Y}} - \mathbf{A}\widehat{\boldsymbol{\alpha}})^\top\mathbf{V}(\boldsymbol{\phi})^{-1}\mathbf{G}_j\mathbf{V}(\boldsymbol{\phi})^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{A}\widehat{\boldsymbol{\alpha}})\}}_{\mathbf{u}(\phi_j)} - \underbrace{\frac{1}{2}\mathrm{tr}(\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_j)}_{\mathbf{e}(\phi_j)}\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j},\end{aligned} \tag{21}$$

$$\mathbf{H}(\phi_j, \phi_k) = \mathrm{E}\left\{\frac{-\partial^2\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}})}{\partial\phi_j\partial\phi_k}\right\}\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j} = \frac{1}{2}\mathrm{tr}(\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_j\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_k)\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j}, \tag{22}$$

where $\widehat{\boldsymbol{\phi}}_0^j = (\hat{\phi}_1, \ldots, \hat{\phi}_{j-1}, 0, \hat{\phi}_{j+1}, \ldots, \hat{\phi}_J)^\top$, $\mathbf{G}_j = \partial\mathbf{V}(\boldsymbol{\phi})/\partial\phi_j = \mathbf{F}_j\mathbf{Q}_j\mathbf{S}_j\mathbf{Q}_j^\top\mathbf{F}_j^\top$, and

$$\mathbf{P}(\hat{\boldsymbol{\phi}}) = \mathbf{V}(\hat{\boldsymbol{\phi}})^{-1} - \mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}. \tag{23}$$

Zhang and Lin (2003) used the moment matching method to show that $\mathbf{u}(\phi_1)$ in (22) asymptotically follows a scaled χ^2 distribution:

$$\mathbf{u}(\phi_1) \sim \kappa\chi_\nu^2, \tag{24}$$

where $\kappa = \mathbf{H}(\phi_1, \phi_1)/(2\mathbf{e}(\phi_1))$ and $\nu = 2\mathbf{e}(\phi_1)^2/\mathbf{H}(\phi_1, \phi_1)$. They thus used $\mathbf{u}(\phi_1)$ as the testing statistics and yielded the p-value by $p(\phi_1) = 1 - F_\chi(\mathbf{u}(\phi_1)/\kappa, \nu)$, where $F_\chi(\cdot, \nu)$ is the cumulative

distribution function of the χ^2_ν distribution.

When the sample size n is large, direct construction of the $(n \times n)$ precision matrix $\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}$ and the $(n \times n)$ projection matrix $\mathbf{P}(\hat{\boldsymbol{\phi}})$ is computationally infeasible. To address this, we develop an operator-based strategy to compute the score statistic $\mathbf{u}(\phi_1)$, its expectation $\mathbf{e}(\phi_1)$, and its variance component $\mathbf{H}(\phi_1, \phi_1)$, without explicitly forming these matrices. In particular, we construct matrix-vector product operators that evaluate the action of $\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}$ and $\mathbf{P}(\hat{\boldsymbol{\phi}})$ on arbitrary vectors, and use them to implement both residual projection and trace computation efficiently. For the trace terms, we exploit the low-rank structure of the form $\mathbf{G}_1 = \mathbf{F}_1\mathbf{Q}_1\mathbf{S}_1\mathbf{Q}_1^\top\mathbf{F}_1^\top$. These computational strategies make the score test both scalable and memory-efficient for large-scale datasets, and the same principles have been widely adopted in existing statistical genetics software such as GCTA (Yang et al., 2011) and SKAT (Wu et al., 2011).

2.5 Wald Test

We adopt the Wald test proposed by Wood (2013) to evaluate whether the smooth deviation term $b_1(x)$ is equal to zero. Specifically, let $\hat{\boldsymbol{\pi}}_1$ denote the estimate of $\boldsymbol{\pi}_1$ under the estimated variance component $\hat{\phi}_1$. The null hypothesis is H_0: $b_1(x) = 0$. The Wald test statistic is given by $\hat{\boldsymbol{b}}_1^\top \mathbf{S}_{b_1}^{-1}\hat{\boldsymbol{b}}_1$, where $\hat{\boldsymbol{b}}_1 = \mathbf{F}_1\mathbf{Q}_1\hat{\boldsymbol{\pi}}_1$ and $\mathbf{S}_{b_1} = \hat{\phi}_1\mathbf{F}_1\mathbf{Q}_1(\mathbf{Q}_1^\top\tilde{\mathbf{K}}_1\mathbf{Q}_1)^+\mathbf{Q}_1^\top\mathbf{F}_1^\top$. Although this statistic is theoretically assumed to follow a χ^2 distribution, Wood (2013) showed that it is a mixture of χ^2 distributions due to the penalization imposed on $b_1(x)$, which renders its effective degrees of freedom non-integer. To account for this, a numerical algorithm was developed to accurately approximate the null distribution of the test statistic.

3 Simulation

3.1 Simulation Settings

We consider the following GAM:

$$g(\mathrm{E}(y_i)) = \eta_i = \alpha_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}),$$

where y_i is the response variable following a Gaussian, Poisson, or binary distribution; α_0 is an intercept; and f_1 is the target function to be tested. We evaluate four structural scenarios for f_1:

- Linearity: $f_1(x_1) = \alpha_1 x_1$, representing a simple linear structure.
- Piecewise Linearity: $f_1(x_1) = \alpha_1 x_1 + \alpha_2(x_1 - 0.5)\mathrm{I}(x_1 > 0.5)$, a piecewise linear function with a change of slope at $\nu_1 = 0.5$.
- Linearity Discontinuity: $f_1(x_1) = \alpha_1 \mathrm{I}(x_1 > 0.5) + \alpha_2(x_1 - 0.5) + \alpha_3(x_1 - 0.5)\mathrm{I}(x_1 > 0.5)$, a discontinuous linear structure with a jump at $\nu_0 = 0.5$.
- Linear Interaction: $f_{12}(x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$. TAPS can test whether a multivariate function follows a specific parametric structure like linear interaction.

In addition, the functions f_3 and f_4 remain fixed, while f_2 is set to zero in the first three settings. The kernel function used to construct $b_1(x)$ is Matérn correlation function (Ruppert et al., 2003). Full expressions for f_1 through f_4 (and f_{12}), as well as additional simulation settings, and codes, are provided in the **Supplementary Materials**. Each scenario was replicated 1,000 times.

3.2 Simulation Results

We first evaluate whether the null distributions of the Wald and score test statistics are correctly specified. If so, the associated p-values should follow a uniform distribution on the interval

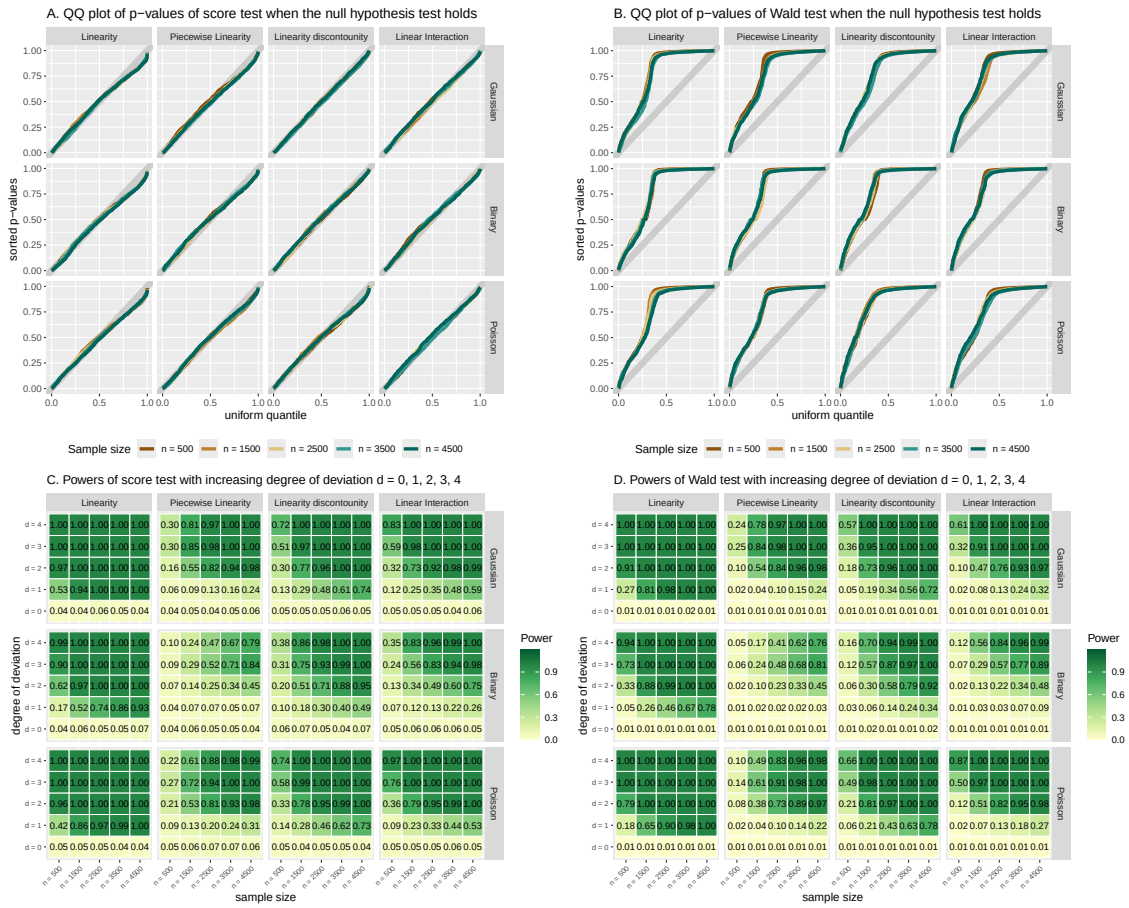

Figure 2: Performance of the Wald and score tests under the null and alternative hypotheses. **A**. QQ plots of p-values under the null hypothesis for the score test, and **B**. for the Wald test, across four structure types (columns) and three outcome distributions (rows), with sample sizes ranging from $n = 500$ to $n = 4500$. Each line compares the empirical distribution of p-values against the expected uniform distribution (gray dashed line). **C**. Heatmaps of power under the alternative hypothesis for the score test, and **D**. for the Wald test, with increasing deviation from the null controlled by parameter d (rows) and sample size (columns). Both color and overlaid numeric values indicate rejection frequencies (power).

$[0, 1]$. When the sorted p-values systematically lie above the diagonal in a quantile-quantile (QQ) plot, it indicates inflated Type II error rates (i.e., the test is overly conservative). Conversely, when the sorted p-values lie below the diagonal, it suggests inflated Type I error rates (i.e., the test is overly liberal). Figure 2A-2B shows the QQ plots of p-values under the null across the four testing scenarios. We observe that the empirical quantiles of the score test closely follow the uniform distribution, regardless of the sample size or whether the response distribution is continuous or discrete. These results confirm that the score test controls both Type I and Type

II error rates well. In contrast, the p-values from the Wald test tend to deviate upward from the uniform quantiles, indicating that the Wald test is more conservative and prone to inflated Type II error. This conservativeness arises from the penalization of nonparametric terms in `mgcv`, which affects the estimation of their effective degrees of freedom (edf). Since the null distribution of the Wald statistic depends on numerical approximation (Wood, 2013), it may become inaccurate when the nonparametric component has near-zero edf, thus leading to inflated Type II errors.

We then investigate the power of the two testing procedures. To simulate deviations from the null hypothesis, we introduce a deviation parameter d, where larger values of d correspond to stronger departures of f_1 from the parametric structure under the null (**Supplementary Materials**). As shown in Figure 2C-2D. The score test consistently achieves high power in detecting deviations from the null across all scenarios under Gaussian and Poisson outcomes, with slightly lower power under the binary distribution. As the deviation parameter d increases, power quickly approaches 1, especially when the sample size exceeds 1500. The Wald test is generally less powerful than the score test but still shows a clear upward trend and converges to high power with moderate sample sizes. Across the four scenarios, the linearity and interaction tests are particularly sensitive to deviations, with both tests attaining high power rapidly. In contrast, the piecewise linearity and discontinuity scenarios require larger sample sizes or stronger deviations (larger d) to achieve comparable power. Overall, both tests become increasingly effective as the deviation from the null grows and the available data become more informative.

Despite its strong performance, the score test is limited to outcomes from the exponential family, whereas the Wald test, though slightly less powerful, is not subject to this restriction. Especially, the Wald test can be applied to quantile generalized additive models (Fasiolo et al., 2021), such as the median GAM (quantile = 0.5). In the **Supplementary Materials**, we further examine the performance of the Wald test when the outcome is a survival time or an ordinal

categorical variable, as well as its robustness to outliers in GAM and median GAM (Figure S6-S8). Similar simulation studies for evaluating the performance of the varying-coefficient model are presented in Figure S9.

4 Real Data Analysis

4.1 UK Biobank database

The UK Biobank is a large-scale, population-based prospective database that enrolled approximately 500,000 individuals from across the United Kingdom between 2006 and 2010 (Bycroft et al., 2018). It integrates extensive phenotypic profiling, including biomarker assays, lifestyle questionnaires, clinical outcomes, physical examinations, and multimodal imaging, with comprehensive genetic data, including array-based genotyping and imputed variants, whole-exome sequencing, and whole-genome sequencing. Genetic data have undergone rigorous centralized quality control, phasing, and imputation, resulting in a high-resolution variant dataset encompassing more than 96 million single-nucleotide polymorphisms (SNPs) and short indels (Bycroft et al., 2018). The breadth and depth of this resource make it exceptionally well-suited for dissecting the genetic architecture of complex traits, developing polygenic risk models, and exploring gene–environment interplay. Owing to its scale, data quality, and open-access model, the UK Biobank has become a foundational platform for genetic epidemiology, contributing to a wide range of studies in cardiovascular, metabolic, neurological, and psychiatric research (Abdellaoui et al., 2023).

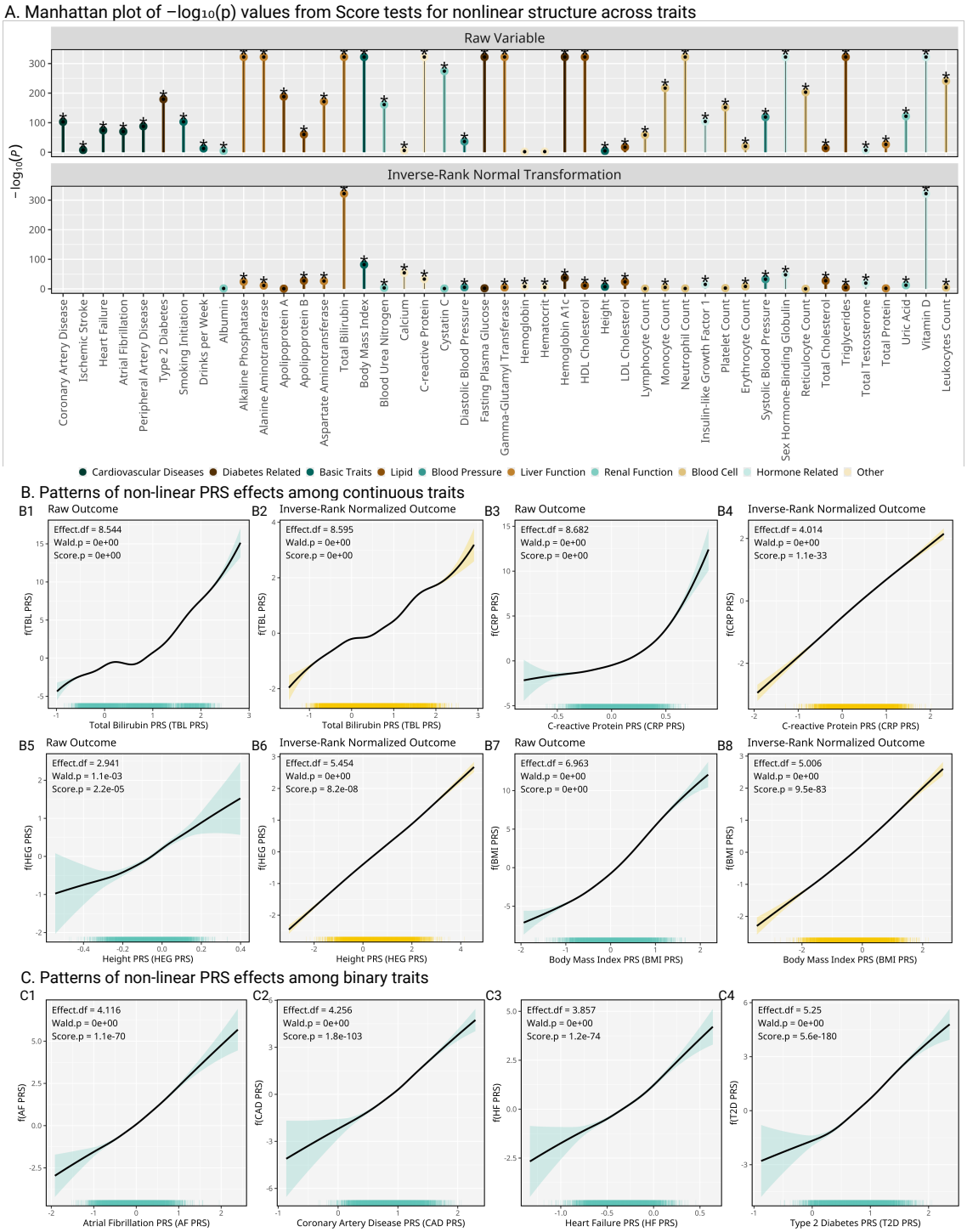

Figure 3: Linearity tests and nonlinear effect patterns of PRSs. **A**. Manhattan plot of $-\log_{10}(p)$-values from score tests evaluating the linearity of PRS effects across a range of traits, respectively to IRNT and raw outcomes, and there is INRT for binary outcomes. Each point represents one trait, with higher values indicating stronger evidence of deviation from linearity. Asterisk * indicates score test p-value $< 0.05/45$. **B**. Estimated nonlinear effect functions for continuous traits. Green curves correspond to raw traits, and yellow curves correspond to IRNT outcomes. **C**. Estimated nonlinear effect functions for selected binary traits. In both **B** and **C**, shaded bands indicate 95% confidence intervals, and tick marks on the x-axis show the distribution of PRS values. Each panel also reports the edf, the p-value from the score test, and the p-value from the Wald test for nonlinearity.

4.2 Linearity Test of Polygenic Risk Scores

PRS is a score of an individual's genetic liability to a trait or disease, calculated based on genotype profiles and genome-wide association study (GWAS) results (Choi et al., 2020). PRS has long been assumed to have a linear relationship with phenotypic outcomes (Hill et al., 2008). However, recent studies suggest that this assumption may be overly simplistic, and PRS may exert nonlinear effects. For example, He et al. (2024) developed a deep learning method called DeLIVR, which identified additional genes associated with HDL cholesterol (HDL-C) and LDL cholesterol (LDL-C) via nonlinear patterns that traditional linear approaches would miss. Motivated by these emerging observations of potential nonlinearity in PRS effects, we applied TAPS to UK Biobank data to formally test whether linear models of PRS are sufficient to capture the genetic variation underlying several complex traits.

Specifically, we analyzed 37 continuous traits and 8 binary traits, including markers of blood pressure, lipids, glucose, liver function, kidney function, type 2 diabetes, and cardiovascular diseases. Our analysis is based on the genetically inferred European population in the UK Biobank (Privé et al., 2022). The sample sizes range from 340,864 (peripheral artery disease) to 501,286 (alcohol drinking status), with a median of 392,102. We observed that many traits exhibited skewed distributions. Therefore, we performed analyses using both the raw trait values and their inverse rank normal transformation (IRNT) counterparts, and used the corresponding GWAS summary statistics from the Neale Lab (Lab, 2018) for each setting. In addition, PRSs were estimated using SBayesRC (Zheng et al., 2024) and predicted for individuals of European ancestry in the UK Biobank. Details are provided in **Appendix** and **Supplementary Materials**.

We used the score test to evaluate whether the effect of a PRS on each trait could be adequately modeled as linear. As shown in Figure 3A, the negative $\log_{10}(p)$ values from the score tests varied substantially across traits, yet the vast majority are far below 0.05/45, a threshold

based on Bonferroni correction, indicating that the linearity assumption rarely holds. To visualize these effects, Figure 3B presents example continuous traits under raw and IRNT traits, and Figure 3C shows example binary traits with estimated nonlinear PRS effects. For instance, total bilirubin (TBL) and C-reactive protein (CRP) exhibit strong statistical evidence of nonlinearity, while height and body mass index (BMI) showed weaker but still significantly non-linear effects. However, when analyzing the IRNT traits, we observed a systematic reduction in the edf for the estimated smooth functions, suggesting weaker nonlinearity after transformation (Table S2). For example, the edf for gamma-glutamyltransferase (GGT) decreased from 8.18 in the raw trait to 3.44 after IRNT, while the score test p-value increased from effectively zero to 1.87E-5, indicating a substantial attenuation of the nonlinear signal. Similar patterns were observed for fasting glucose (FPG) (edf: 7.61 to 2.24, p-values: 0 to 7.92E-3) and glycated hemoglobin (HBA1C) (edf: 7.78 to 6.32, p-values: 0 to 6.49E-38). This may occur because IRNT reduces skewness and long tails, which are often where nonlinear effects are most prominent. These comparisons highlight that data transformation can significantly alter the apparent degree of nonlinearity, potentially obscuring interpretable biological signals in some cases.

We compared the PVE between linear and nonlinear PRS models to quantify how nonlinear modeling improves predictive performance. Table S3 presents the full results for raw traits and Table S4 for IRNT traits. Across all 45 traits, the median gain in PVE under raw modeling was 0.34%, with the largest improvement observed for TBL (21.6%). Other traits with notable increases include CRP (12.4%) and gamma-glutamyl transferase (GGT) (6.3%). Consistent with the score tests of linearity, the analysis-of-variance (ANOVA) (i.e., F or likelihood-ratio tests) supports the presence of nonlinear effects, even though the predictive gain appears modest. Given the large sample size of the UK Biobank, even subtle nonlinear patterns can lead to statistically significant results. Moreover, when we repeated the comparison under IRNT outcomes to reduce

the influence of outliers and skewness, we found that the nonlinear signal became much weaker, with a median of gain in PVE being 0.03%. Only TBL and vitamin D (VTD) retained PVE gains above 1% after transformation, suggesting that for most traits, deviations from linearity are sensitive to distributional features (Table S3-S4). This underscores the importance of modeling scale and transformation when assessing nonlinear genetic effects.

4.3 Linear Interaction Test of Gene-by-Age Interaction

The identification of gene-gene (GxG) and gene-environment (GxE) interactions in complex human traits and diseases has long attracted considerable attention, yet progress has often fallen short of expectations (Crow, 2010). One key limitation, as highlighted by theoretical studies, is that even when phenotypic variation arises purely through interaction effects, a substantial proportion of variance can still be captured by an additive genetic model (Aschard, 2016; Zhu et al., 2024). This explains why traditional interaction analyses, those that test for interactions conditional on additive main effects, often suffer from low statistical power. To address this challenge, Aschard (2016) proposed a more powerful alternative, which tests for interactions between a scaling environmental variable and the PRS of a trait, which aggregates genetic information across loci. In this analysis, we adopt this PRS-based approach to investigate whether age modifies the effect of PRSs across a range of complex traits. Specifically, we aim to determine (1) whether a significant age–PRS interaction exists, and (2) whether this interaction significantly departs from the traditional linear interaction structure.

Specifically, we considered a varying-coefficient model defined as $f(\text{Age}, \text{PRS}) = \text{PRS} \cdot \beta(\text{Age})$. In this framework, the PRS itself entered the model linearly, reflecting prior findings that linear PRS models adequately explained most of the phenotypic variance for a wide range of traits. In the mixed-effect representation (8), the fixed effect was the linear effect, corresponding

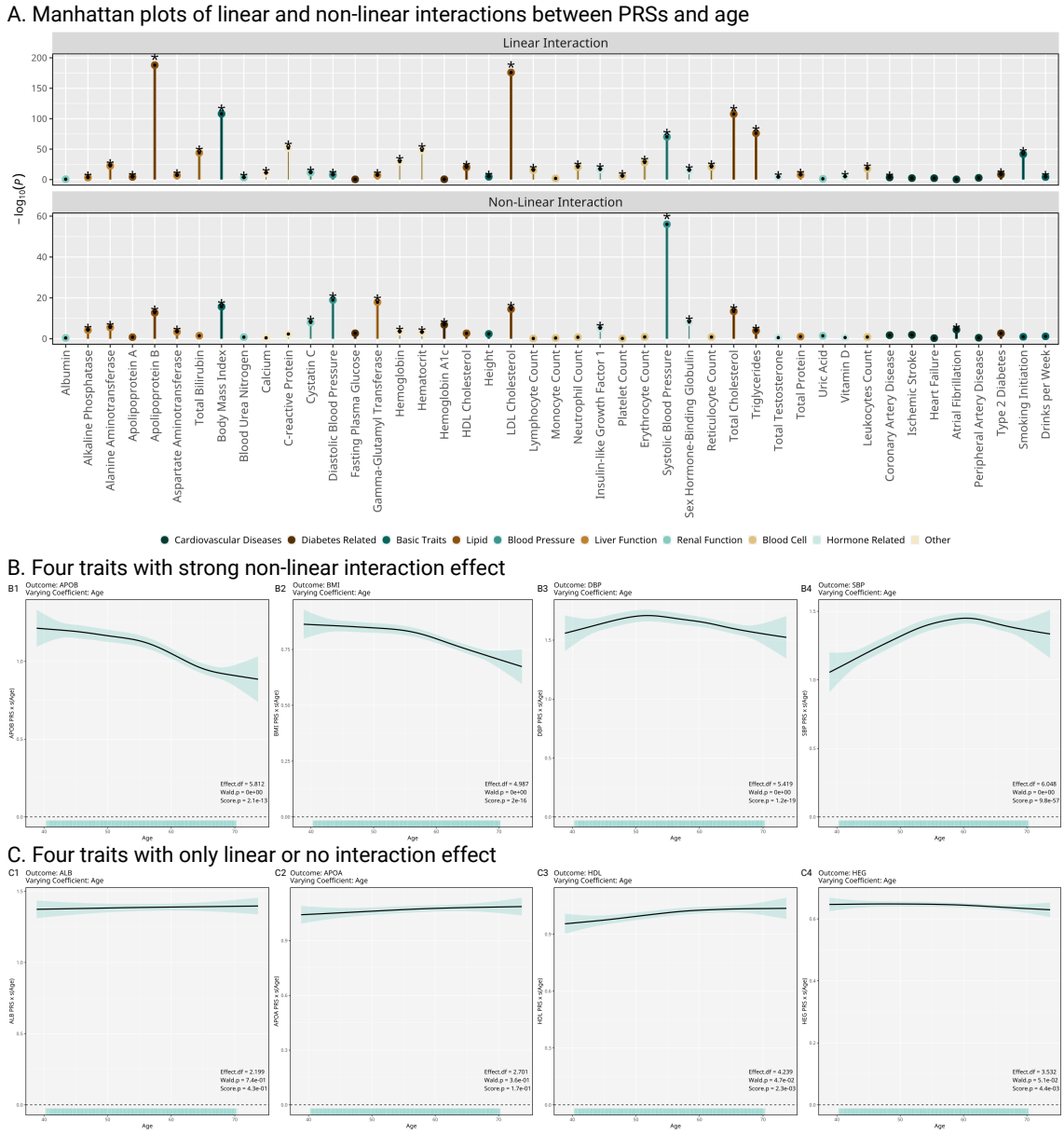

Figure 4: Age-varying coefficient analysis of PRS effects across traits. **A**. Manhattan plots of interaction tests evaluating whether polygenic risk score (PRS) effects vary with age. The top panel shows $-\log_{10}(p)$-values from Wald tests for linear age–PRS interactions, while the bottom panel shows $-\log_{10}(p)$-values from score tests for non-linear age-varying effects, orthogonal to the linear interaction terms. Asterisk * indicates score test p-value $< 0.05/45$. **B**. Estimated age-varying effect functions for selected traits with strong evidence of nonlinearity. **C**. Estimated age-varying effect functions for traits with linear or no detectable interaction. Each curve represents the fitted age-dependent PRS effect; shaded bands indicate 95% confidence intervals. Each panel reports the effective degrees of freedom (edf), the score test p-value for nonlinearity, and the Wald test p-value for linear interaction.

to the traditional linear interaction model. When the null hypothesis holded, $f(\text{Age}, \text{PRS})$ reduces to $f(\text{Age}, \text{PRS}) = \alpha_0 \text{PRS} + \alpha_1(\text{Age} \cdot \text{PRS})$, which reflects the linear age-PRS interaction.

(Note that Age was also considered in the model (**Appendix**).) Consistent with our previous analyses, we focused on 37 continuous traits and 8 binary traits in individuals of genetically inferred European ancestry from the UK Biobank. For each trait, we conducted analyses using its INRT counterpart for a continuous trait, with GWAS summary statistics obtained from the Neale Lab. In addition, we simultaneously assessed both linear and nonlinear age–PRS interactions. We used a univariate Wald test to evaluate the significance of the linear interaction component and a score test to assess whether the PRS effect deviates nonlinearly with age beyond the linear age–PRS interaction.

As shown in Figure 4A, the top panel presents the Manhattan plot of $-\log_{10}(p)$ values from the Wald tests for linear age–PRS interactions, while the bottom panel reports $-\log_{10}(p)$ values from the score tests evaluating non-linear age-varying effects orthogonal to the linear interaction term. Full results are shown in Table S5. In contrast to the results in Figure 3A, where most traits exhibited strong evidence of nonlinearity in the PRS effect itself, the majority of traits had significant linear age–PRS interactions, whereas only about half showed evidence of additional non-linear interactions. Notably, the most significant score test p-value for non-linear age-varying effects was observed for systolic blood pressure (SBP) (p-value = 9.84E-57), which is markedly weaker than the strongest nonlinear PRS effects reported in Figure 3.

Figure 4B showed estimated age-dependent PRS effect curves for four representative traits with significant non-linear interaction effects: apolipoprotein B (APOB), BMI, diastolic blood pressure (DBP), and SBP. A consistent pattern emerged across these examples: the age-varying effect appeared approximately piecewise linear, with inflection points typically occurring between ages 50 and 60. In particular, both DBP and SBP exhibited a rise-and-fall structure: the PRS effect increased during middle age and then declined later in life, with the changepoint for DBP occurring earlier than for SBP. Figure 4C presents traits with either purely linear or

no significant age-varying genetic effects. For instance, apolipoprotein AI (APOA), HDL, and height showed significant linear interaction effects without evidence of non-linearity, while albumin (ALB) had no detectable age-dependent genetic influence. These traits served as negative controls, demonstrating that our score test was not overly liberal: it successfully distinguished between linear and non-linear patterns and yielded appropriately conservative results when no age-dependent variation was present. In addition, we assessed whether incorporating non-linear age-varying effects could improve the PVE of the linear PRS models (Table S6). However, we only observed smoking initiation (SMK) for which the inclusion of non-linear age–PRS interactions led to a PVE gain greater than 1% (1.34%), and the median gain in PVE was 0.04%.

5 Discussion

In this paper, we propose TAPS, a new framework for estimating and testing whether a target function in a GAM can be fully described by a parametric structure. Existing hypothesis tests are typically limited to testing whether a smooth function is zero, rather than whether it follows a scientifically meaningful parametric form. TAPS addresses this gap by formulating the null hypothesis as the function belonging to a given parametric structure, and the alternative as a general smooth function. This formulation allows us to directly answer the question: does the function follow the specified structure? Methodologically, TAPS ensures model identifiability by constructing an orthogonal smooth deviation component and provides both Wald and score tests for hypothesis testing. The Wald test can be applied to all families of outcomes which `mgcv` supports, while the score test generally offers greater statistical power. Given its superior control of Type I error and higher power, the score test is recommended as the default option for TAPS when applicable. The Wald test serves as a practical alternative for non-exponential-family outcomes. In addition, we develop the R package `mgcv.taps`, which is tightly integrated with

`mgcv`, one of the most widely used GAM software, allowing users to apply TAPS with minimal additional learning, supports a wide range of outcome types, and scales efficiently to biobank-scale data.

Our two real data analyses also provide novel insights. First, the linearity test of PRSs across 45 traits revealed that the null hypothesis of linearity was rejected for nearly all traits, while the improvement in predictive performance from nonlinear modeling was modest for most traits (median of gain in PVE = 0.34%). These results confirm that for highly polygenic traits such as height, linear additive models remain sufficient: modeling height nonlinearly yielded a negligible gain in PVE (0.03%), consistent with prior findings that linear models suffice for many common traits (Hill et al., 2008). However, some traits, including TBL and CRP, showed clear departures from linearity, with PVE gains exceeding 10%, suggesting that linear models were inadequate for these outcomes. Moreover, we found that the presence and impact of nonlinear effects were influenced by data transformation. After applying IRNT, the score test continued to detect significant nonlinear patterns for most traits (median p-values 5E-12); however, the gain in PVE from nonlinear modeling substantially decreased, with a median improvement of only 0.03%. This underscores that the predictive contribution of non-linear PRSs is often limited, especially after transforming traits to reduce skewness and outliers. Our findings align with Albiñana et al. (2023), who reported that improvements from non-linear models (e.g., XGBoost (Chen and Guestrin, 2016)) primarily stem from capturing non-linear interactions between covariates, such as sex and age.

In our age-varying coefficient analysis, we detected several traits, most prominently APOB, BMI, DBP, and SBP, whose genetic effects vary with age in a manner that exceeds the traditional linear age–PRS interaction. The estimated curves are well-approximated by piece-wise linear segments with changepoints clustered between 50 and 60 years of age. This rise-then-fall pattern

is consistent with twin data showing that SBP heritability is 0.5$\sim$0.6 in mid-life but falls to 0.1 after age 65, a decline attributed to widespread antihypertensive treatment and increasing arterial stiffening (Hong et al., 1994; Hottenga et al., 2005). Despite these clear age-dependent signals, adding the non-linear age–PRS component to the conventional linear PRS model conferred only a marginal improvement in predictive performance: across all traits, the largest gain in phenotypic variance explained was 0.04%. This accords with earlier theoretical and empirical work showing that gene–environment interactions, such as age-PRS effects, typically account for an infinitesimal share of heritability beyond the additive model and are therefore intrinsically challenging to detect in population-scale data (Aschard, 2016; Zhu et al., 2024).

Several limitations of TAPS warrant further investigation. First, TAPS requires users to specify the parametric structure to be tested, including hyperparameters such as the location of a jump or slope change. In applications like RDD, these locations correspond to cutoffs of treatment assignments and are not subject to tuning (Huntington-Klein, 2021). However, identifying appropriate structures remains challenging, and data-driven methods for selecting or validating structural forms, such as the number and position of breakpoints, are still underdeveloped (Muggeo, 2003; Zhang and Li, 2023). Second, the Wald test relies on numerical approximations and is consistently less powerful than the score test, especially when the smooth term has low degrees of freedom. Whereas, the score test currently applies only to outcomes from exponential family distributions. For other types, such as survival and ordinal outcomes, only the Wald test is available. We therefore recommend using the score test as the first choice when applicable, and applying the Wald test as a practical alternative when the score test is unavailable.

A Appendix

A.1 Linearity Test of PRS

For continuous traits, we consider the following model:

$$\text{Trait}_i = f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + \epsilon_i, \tag{25}$$

and for binary traits, we adopt the following generalized additive model:

$$\Pr(\text{Trait}_i = 1) = \frac{\exp(f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma})}{1 + \exp(f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma})}, \tag{26}$$

where $\boldsymbol{Z}_i = (1, \text{Sex}_i, \text{PC}_{i1}, \ldots, \text{PC}_{i10})^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{12})^\top$. For the term $f_1(\text{Trait PRS})$, we use the mixed-effect representation (3) with a linear parametric component to test whether the PRS effect can be sufficiently modeled as linear. For computational efficiency, the age effect $f_2(\text{Age})$ is modeled using cubic regression splines. Smoothing parameters were estimated via REML as implemented in `mgcv`. Note that if a participant was on blood pressure medication, we added 15 mmHg and 10 mmHg to their systolic and diastolic blood pressure measurements, respectively. The cholesterol-lowering medication status was included as a covariate when analyzing LDL-C and total cholesterol, following the procedure of Sinnott-Armstrong et al. (2021).

To compute the PVE, we first fit a fully parametric linear model for $f_1(\text{Trait PRS})$ by replacing the smooth term with a linear effect. For continuous outcomes, PVE is defined as the standard coefficient of determination:

$$R^2_{\text{continuous}} = 1 - \frac{\text{var}(\hat{\epsilon}_i)}{\text{var}(\text{Trait}_i)}, \tag{27}$$

where $\hat{\epsilon}$ denotes the residuals from the linear model. For binary outcomes, we report the pseudo R^2 on the logit liability scale (Lee et al., 2012), which approximates the variance explained on

an underlying continuous liability. Specifically, the transformation is:

$$R^2_{\text{binary}} = \frac{\text{var}(\hat{f}_1(\text{Trait PRS}_i) + \hat{f}_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}})}{\text{var}(\hat{f}_1(\text{Trait PRS}_i) + \hat{f}_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}}) + 3.29}. \tag{28}$$

This formulation accounts for the inherent noise in binary traits and allows more interpretable comparisons across traits with different prevalences.

To reduce the influence of outliers and skewed distributions, we applied the following IRNT to all continuous outcomes:

$$\text{IRNT}(y_i) = \Phi^{-1}(n^{-1} r_i - 0.5), \tag{29}$$

where r_i is the rank of y_i among all y values in ascending order, and $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. This transformation ensures that the transformed variable approximately follows a standard normal distribution and improves the robustness and comparability of model fitting. Moreover, to mitigate the influence of outliers and skewed distributions in the raw trait analyses, we excluded individuals falling in the top and bottom 1% of the trait distribution.

A.2 Linearity Interaction Test of Age-PRS Interaction

For continuous traits, we consider the following model:

$$\text{Trait}_i = (\text{Trait PRS}_i) \cdot \beta_1(\text{Age}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + \epsilon_i, \tag{30}$$

and for binary traits, we adopt the following generalized additive model:

$$\Pr(\text{Trait}_i = 1) = \text{expit}((\text{Trait PRS}_i) \cdot \beta_1(\text{Age}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}), \tag{31}$$

where $\boldsymbol{Z}_i = (1, \text{Sex}_i, \text{PC}_{i1}, \ldots, \text{PC}_{i10})^\top$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{12})^\top$, and expit(x)=exp(x)/(1+exp(x)). For the term $f_1(\text{Trait PRS})$, we use the mixed-effect representation (8) with a linear parametric component to test whether the age-varying effect can be sufficiently modeled as linear. Other settings are the same as the linearity test of PRS.

Acknowledgments

We would like to thank the editors and the anonymous referees for their very careful reviews and constructive suggestions. The first author thanks Dr. Jianxin Pan for multiple insightful discussions. We also thank Dr. Simon Wood for providing the flexible `mgcv` interface, which facilitates user-defined smooth terms and enables seamless methodological extensions. This work was supported by grants HG011052 and HG011052-03S1 (to X.Z.) from the National Human Genome Research Institute (NHGRI), USA.

Data Availability

All GWAS summary statistics used in this study are publicly available, with their sources listed in Table S1. Given the large number of traits analyzed, we have implemented an interactive Shiny web interface (`https://ovlwff-yihe-yang.shinyapps.io/visualization_of_taps/`) to visualize all plots. The individual-level data from the UK Biobank used for replication analyses are available under Application ID: 81097.

Code Availability

The R package `mgcv.taps`, which implements the TAPS method, is available at `https://github.com/harryyiheyang/mgcv.taps`. All analysis code used in this paper is available at `https://github.com/harryyiheyang/mgcv.taps.code`.

SUPPLEMENTARY MATERIAL

Supplementary Materials: Supplementary Materials contain the supplementary introduction to TPAS, the supplementary simulations, and the supplementary real data analysis.

Supplementary Tables: Supplementary Tables contain multiple tables recording the results of real data analysis.

References

Abdellaoui, A., Yengo, L., Verweij, K. J., and Visscher, P. M. (2023). 15 years of gwas discovery: realizing the promise. *The American Journal of Human Genetics* **110,** 179–194.

Albiñana, C., Zhu, Z., Schork, A. J., Ingason, A., Aschard, H., Brikell, I., Bulik, C. M., Petersen, L. V., Agerbo, E., Grove, J., et al. (2023). Multi-pgs enhances polygenic prediction by combining 937 polygenic scores. *Nature communications* **14,** 4702.

Aschard, H. (2016). A perspective on interaction effects in genetic association studies. *Genetic epidemiology* **40,** 678–688.

Bentley, A. R., Sung, Y. J., Brown, M. R., Winkler, T. W., Kraja, A. T., Ntalla, I., Schwander, K., Chasman, D. I., Lim, E., Deng, X., et al. (2019). Multi-ancestry genome-wide gene–smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nature Genetics* **51,** 636–648.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88,** 9–25.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Choi, S. W., Mak, T. S.-H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols* **15,** 2759–2772.

Cox, D., Koh, E., Wahba, G., and Yandell, B. S. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics* **16,** 113–119.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika* **92,** 91–103.

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365,** 1241–1244.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., and Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association* **116,** 1402–1412.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **55,** 757–779.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. CRC Press.

He, R., Liu, M., Lin, Z., Zhuang, Z., Shen, X., and Pan, W. (2024). Delivr: a deep learning approach to iv regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics* **25,** 468–485.

Hill, W. G., Goddard, M. E., and Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4,** e1000008.

Hong, Y., De Faire, U., Heller, D. A., McClearn, G. E., and Pedersen, N. (1994). Genetic and environmental influences on blood pressure in elderly twins. *Hypertension* **24,** 663–670.

Hottenga, J.-J., Boomsma, D. I., Kupper, N., Posthuma, D., Snieder, H., Willemsen, G., and de Geus, E. J. (2005). Heritability and stability of resting blood pressure. *Twin Research and Human Genetics* **8,** 499–508.

Huntington-Klein, N. (2021). *The effect: An introduction to research design and causality*. Chapman and Hall/CRC.

Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52,** 1–18.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582* .

Lab, N. (2018). UK Biobank GWAS Round 2 Results. Accessed: 2025-04-06.

Lee, S. H., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2012). A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology* **36,** 214–224.

McKeigue, P. M., McGurnaghan, S., Blackbourn, L., Bath, L. E., McAllister, D. A., Caparrotta, T. M., et al. (2023). Relation of incident type 1 diabetes to recent covid-19 infection: cohort study using e-health record linkage in scotland. *Diabetes Care* **46,** 921–928.

Muggeo, V. M. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine* **22,** 3055–3071.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384.

Nummi, T., Pan, J., Siren, T., and Liu, K. (2011). Testing for cubic smoothing splines under dependent data. *Biometrics* **67,** 871–875.

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P. F., and Vilhjálmsson, B. J. (2022). Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109,** 12–23.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Shang, L., Wu, P., and Zhou, X. (2025). Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics. *Nature Communications* **16,** 1059.

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G. R., Wainberg, M., Ollila, H. M., Kiiskinen, T., et al. (2021). Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature genetics* **53,** 185–194.

Sulc, J., Sjaarda, J., and Kutalik, Z. (2022). Polynomial mendelian randomization reveals non-linear causal effects for obesity-related traits. *Human Genetics and Genomics Advances* **3,**.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Wood, S. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62,** 1025–1036.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99,** 673–686.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466,** 1102–1104.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 3–36.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika* **100,** 221–228.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC Press.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). Generalized additive models for gigadata: modeling the uk black smoke network daily data. *Journal of the American Statistical Association* **112,** 1199–1210.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111,** 1548–1563.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89,** 82–93.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88,** 76–82.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4,** 57–74.

Zhang, F. and Li, Q. (2023). Segmented correspondence curve regression for quantifying covariate effects on the reproducibility of high-throughput experiments. *Biometrics* **79,** 2272–2285.

Zheng, Z., Liu, S., Sidorenko, J., Wang, Y., Lin, T., Yengo, L., Turley, P., Ani, A., Wang, R., Nolte, I. M., et al. (2024). Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics* **56,** 767–777.

Zhu, X., Yang, Y., Lorincz-Comi, N., Li, G., Bentley, A. R., de Vries, P. S., Brown, M., Morrison, A. C., Rotimi, C. N., Gauderman, W. J., et al. (2024). An approach to identify gene-environment interactions and reveal new biological insight in complex traits. *Nature Communications* **15,** 3385.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109,** 1193–1198.