Genome Biology

RESEARCH

Open Access

Genomic prediction with kinship-based multiple kernel learning produces hypothesis on the underlying inheritance mechanisms of phenotypic traits

Daniele Raimondi[1,2*], Nora Verplaetse[2], Antoine Passemiers[2], Deborah Sarah Jans[3], Isabelle Cleynen[3] and Yves Moreau[2]

*Correspondence:
daniele.raimondi@igmm.cnrs.fr

[1] Institut de Génétique Moléculaire de Montpellier (IGMM), CNRS-UMR5535, Université de Montpellier, Montpellier 34293, France
[2] ESAT-STADIUS, KU Leuven, Leuven 3001, Belgium
[3] Department of Human Genetics, KU Leuven, Leuven 3001, Belgium

Abstract

Background: Genomic prediction encompasses the techniques used in agricultural technology to predict the genetic merit of individuals towards valuable phenotypic traits. It is related to Genome Interpretation in humans, which models the individual risk of developing disease traits. Genomic prediction is dominated by linear mixed models, such as the Genomic Best Linear Unbiased Prediction (GBLUP), which computes kinship matrices from SNP array data, while Genome Interpretation applications to clinical genetics rely mainly on Polygenic Risk Scores.

Results: In this article, we exploit the positive semidefinite characteristics of the kinship matrices that are conventionally used in GBLUP to propose a novel Genomic Multiple Kernel Learning method (GMKL), in which the multiple kinship matrices corresponding to Additive, Dominant, and Epistatic Inheritance Mechanisms are used as kernels in support vector machines, and we apply it to both worlds. We benchmark GMKL on simulated cattle phenotypes, showing that it outperforms the classical GBLUP predictors for genomic prediction. Moreover, we show that GMKL ranks the kinship kernels representing different inheritance mechanisms according to their compatibility with the observed data, allowing it to produce hypotheses on the normally unknown inheritance mechanisms generating the target phenotypes. We then apply GMKL to the prediction of two inflammatory bowel disease cohorts with more than 6500 samples in total, consistently obtaining results suggesting that epistasis might have a relevant, although underestimated role in inflammatory bowel disease (IBD).

Conclusions: We show that GMKL performs similarly to GBLUP, but it can formulate biological hypotheses about inheritance mechanisms, such as suggesting that epistasis influences IBD.

Background

Genome Interpretation (GI) [1, 2] is an area of bioinformatics devoted to deciphering the functional and clinical significance of genetic information contained within an organism's genome [3, 4]. It focuses on the development of computational methods to model the relation between genotype and phenotype, with the goal of understanding the potential biological impact of variants and their relevance to traits, diseases, or other phenotypes [5, 6]. Outside of bioinformatics, the task of modeling the relationship between genotype and phenotype is called Genomic Prediction (GP) [7–9]. It originated in agriculture and animal breeding to predict the genetic merit of individuals towards valuable phenotypic traits [10, 11]. Modern approaches followed the input of [8, 12] and leverage information from genetic markers, rather than relying solely on the pedigrees derived from the observed phenotypic traits of the target individuals and their relatives, which was the traditional procedure before the advent of Single Nucleotide Polymorphisms (SNP) array technology [13, 14]. Since then, GP methods are using high-throughput genomic technologies, such as genotyping arrays [8, 15] or sequencing [10], to collect genetic information from individuals in a population, including in humans [16, 17]. This genomic data is then analyzed to combine the information from all markers into a single score, with the goal of estimating the individuals' breeding values, which consists in their genetic predisposition towards desirable traits. In the context of crop and animal breeding, these traits might for example relate to disease resistance, growth rate or milk production [12, 18].

Even though there has been some investigation of machine learning (ML) methods for GP [7, 19–21], including neural networks (NNs) [9, 22–27] and Reproducing Kernel Hilbert Spaces (RKHS) regression methods [28–30], such as support vector machines (SVMs) [19, 31, 32], in plant and animal breeding the most widely adopted approaches are variants of the Best Linear Unbiased Prediction (BLUP) Linear Mixed Model (LMM) [33, 34]. The Genomic BLUP (GBLUP)[8, 12, 35], in particular, uses marker-based relationship (kinship) matrices to specify the covariance structure between individuals without explicitly modeling the effects of the individual SNPs [36].

While in plant and animal breeding GP can be used to guide selection [37] by predicting the unobserved genetic value (i.e., the genetic propensity) towards agriculturally valuable traits, in human genetics the aim is to model the future phenotypic trajectory of an individual, assessing their *risk* towards developing disease traits [13]. These two fields have thus several aspects in common. Nonetheless, BLUP methods are significantly less popular in human genetics, which relies instead on linear models, called Polygenic Risk Scores (PRS), derived from array-based SNP association studies [38, 39]. Even though PRS are statistically less principled than LMMs/BLUP [40, 41], both approaches are linear models that aim at modeling additive genetic effects [42].

Most GBLUP methods use kinship matrices computed solely from additive inheritance effects to model the genetic similarity between individuals, but approaches to take into account dominance [43] and epistatic effects [44, 45] have been proposed and they can be integrated into GBLUP or other models [25]. Even if this aspect has—to the best of our knowledge—been ignored so far, the fact that kinship matrices are symmetric and positive definite makes them suitable to be used as kernel matrices into SVM models as well. SVMs [46] are supervised ML methods that use kernel

matrices to implicitly map data points into a higher-dimensional space, enabling the discovery of an optimal hyperplane for classification or regression. The choice of kernel in SVMs is crucial, since it determines which possibly highly nonlinear transformation is applied to the input data to facilitate the separability of the prediction classes. For example, one of the most popular choice is the Radial Basis functions (RBF) kernel, which corresponds to mapping the input data into an infinite dimensional space, and has also been used for marker-based GP [19, 32, 47].

In this article, we propose a novel approach to GP based on using kinship matrices describing Additive, Dominance, and Epistatic genetic effects as kernels in SVM models, instead of as covariance matrices in GBLUP models, as is conventionally done [34]. Depending on the Inheritance Mechanisms (IMs) involved in generating the predicted phenotype, we show that combining multiple kinship matrices encompassing different genetic effects can be beneficial for the predictions, and we therefore call our approach Genomic Multiple Kernel Learning (GMKL).

With respect to GBLUP and other RKHS methods used in quantitative genetics, GMKL provides two main advantages. First, the SVM algorithm guarantees that the optimal hyperplane separating the data is found during training [46]. Second, our GMKL approach can *rank* the kinship matrices representing different genetic inheritance effects in function of their *compatibility*[48, 49] with the target phenotypes. This means that our GMKL can produce *hypotheses* for the inheritance mechanisms (IMs) involved in generating the target phenotype, given the observed markers, and it can rank them (i.e., Additive, Dominant) in function of their prominence in the data.

To show the potential of GMKL for GP and GI in general, we tested it on 3 different datasets. First, we benchmarked it on nine synthetic phenotypes computed on cattle genotype data, showing that our approach improves the phenotypic predictions with respect to the conventional additive GBLUP, and performs on par with an hypothetical GBLUP(OP) model artificially provided with the (normally unknown) OP*timal* set of kinship matrices at each step [25]. Most importantly, we exploit the fact that the IMs generating these nine synthetic phenotypes are known, to show that the ranking provided by our MKL approach actually produces hypotheses on the IMs that are consistent with the ones used to generate these synthetic phenotypes.

We then extended our validation by benchmarking our GMKL against other methods on 3522 samples from an in-house case-control SNP array cohort of inflammatory bowel disease (IBD) patients, confirming that GMKL performs similarly to GBLUP in terms of prediction performance. In addition to GBLUP, our GMKL can formulate hypotheses on the most compatible IMs in this in-house cohort, which turned out to be mainly Dominant and Epistatic.

To further validate our approach, we performed the same GMKL analysis on a whole exome sequencing (WES) IBD cohort of 3798 samples (3318 cases, 480 controls) from dbGAP, obtaining the same ranking for the inheritance patterns, therefore confirming on different data that epistasis could have an important role in IBD, similarly to what we suggested in a previous study [50]. Additionally, we show that the GMKL predictions have performance similar to the most recent deep learning models developed on the same cohort [50].

Results

The Genomic Multiple Kernel Learning methods for Genomic Prediction

In Fig. 1, we benchmarked five variants of our GMKL approach for the regression of the nine real-valued synthetic phenotypes on the CATTLE dataset [25], which contains 1033 pure-bred Holstein samples genotyped with an Illumina BovineSNP50 Beadchip, leading to 26503 SNPs for each sample [21] (see the Methods "Simulated phenotypes on cattle SNP array data" section). We evaluated the prediction methods with a five-fold cross-validation, comparing them to state-of-the-art methods. Our GMKL method comes in different variants, called MEAN, FH, CKA, CKACLOSED, and GD. They differ in the strategy used to combine the kinship kernels (see the Methods "Kernel Learning methods" section for details). The number next to their name in each row of Fig. 1 indicates how many kinship matrices (from one to five) are considered each time and respectively describes Additive (A), Dominant (D), and three types of Epistatic effects.

MEAN is the simplest approach, since the combined $\mathbf{K}_{\text{MEAN}}$ kernel is just the arithmetic mean of the selected kinship matrices. For what concerns FH [48] and CKA [49], the combined kernels $\mathbf{K}_{\text{FH}}$ and $\mathbf{K}_{\text{CKA}}$ are the result of a linear combination of the available kinship matrices, weighted by their *compatibility* (i.e., correlation) with respect to the *ideal* kernel matrix produced by the phenotypic similarity between the training set samples. This *perfect* kernel matrix $\mathbf{K}_{\text{Y}}$ [49] is computed as the outer product $\mathbf{K}_{\text{Y}} = YY^{\top}$ between the vector Y containing the training set labels for a target phenotype. FH [48] and CKA [49] take their names from the two different ways to measure the compatibility between kernel matrices they use (see the "Methods" section for the details). CKACLOSED is a variation of the CKA algorithm in which the optimal weights for the combination of the kernels are computed jointly, in an analytical closed form [49]. The last GMKL approach we tested is GD, which stands for Gradient Descent. In this case, the combined kernel $\mathbf{K}_{\text{GD}}$ is obtained by optimizing the linear combination of the five kinship matrices trying to maximize its correlation with the perfect kernel $\mathbf{K}_{\text{Y}}$. Similarly to CKACLOSED, only a single GD score is present in Fig. 1 because we always provided *all* the kinship matrices to the optimizer, allowing it to select the optimal combination of weights, possibly *silencing* kinship matrices corresponding to irrelevant or detrimental inheritance mechanisms by assigning weights close to zero.

GMKL positively compares against conventional GP methods on the CATTLE dataset

In Fig. 1, we compared our GMKL approaches to state-of-the-art methods. From top to bottom, the orange bars labeled RBF and linSVM show the results of conventional SVMs with radial basis function and linear kernels respectively, computed directly on

(See figure on next page.)

Fig. 1 Figure showing the comparison between the 9 methods we benchmarked on the CATTLE dataset. The MEAN, FH, and CKA GMKL approaches using 1 to 5 kernel matrices are respectively shown in shades of green, red and blue. The gray bars show the additive GBLUP model (light) and the optimal GBLUP model (dark), which always use the optimal set of kinship matrices. The phenotypes ranging from zero to four involve only A and D effects, with Pheno:0 being 100% additive, Pheno:2 being 50/50%, and Pheno:4 being purely Dominant. Phenotypes 5–9 include also epistatic effects. They are composed by a base 33% A and D components, plus a 34% epistatic component that is additive-additive (Pheno:5), additive-dominant (Pheno:6), and dominant-dominant (Pheno:7). Pheno:8 contains a mixture of all effects

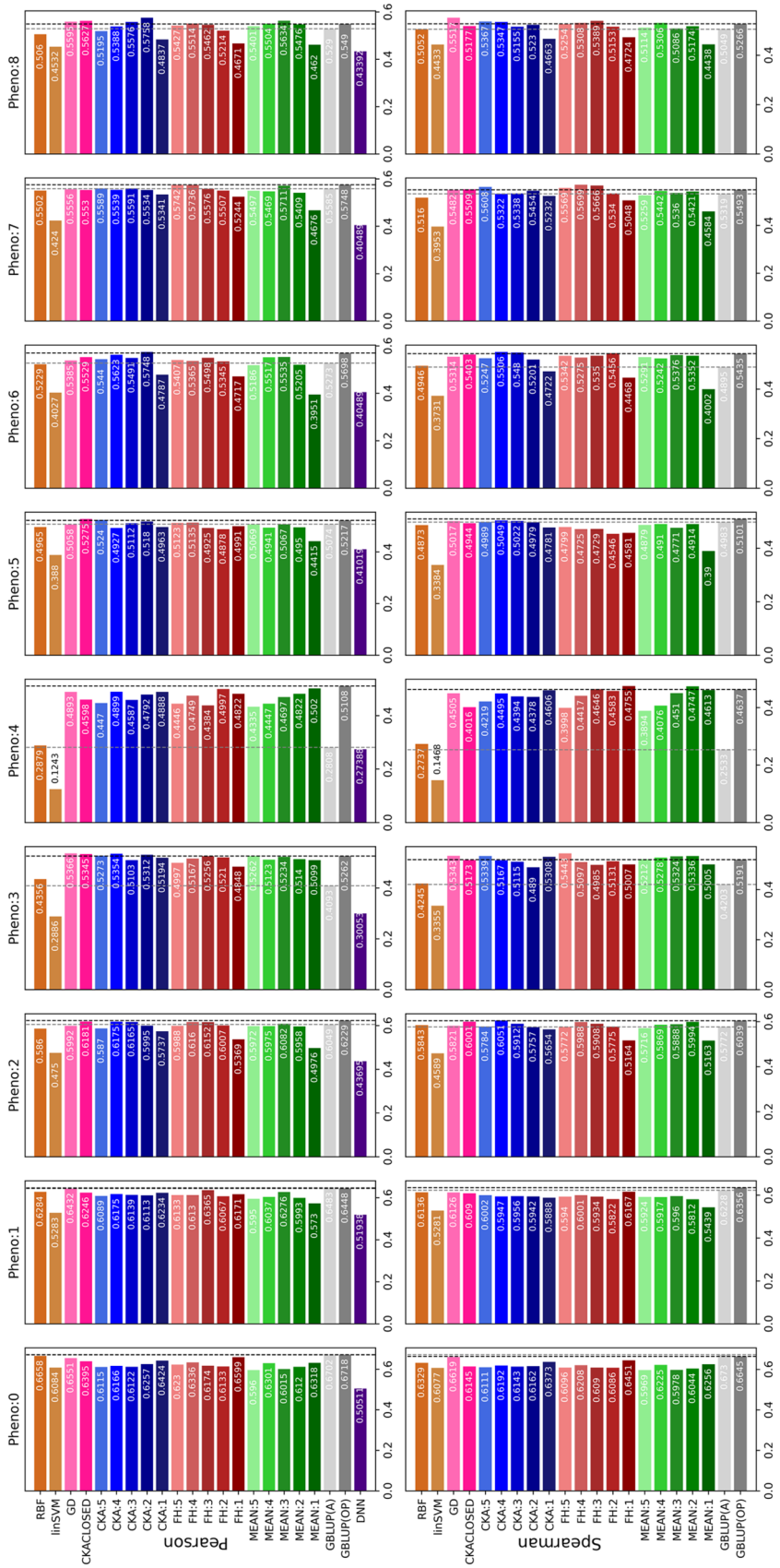

Fig. 1 (See legend on previous page.)

the genotype matrix, applying the same minor allele frequencies (MAF) thresholds we used to compute the kinship matrices to ensure a fair comparison. The light and dark gray bars show respectively the scores of the additive GBLUP(A) model and the optimal GBLUP(OP) model. The first uses as covariance matrix only the additive kinship matrix (standard GBLUP approach), while the latter always uses the *optimal* set of kinship matrices, corresponding to all the IMs involved in generating each synthetic phenotype [25] (see Table 4 for the full list). GBLUP(OP) therefore can be considered as an *upper bound* for the predictions on the CATTLE dataset, since it is not normally possible to know which IMs are involved in generating the target phenotype in real-life situations, where the phenotypes are not simulated. Finally, the violet bar shows the performance of the Convolution Neural Network (CNN) approach on kinship matrices proposed in [25]. The standard deviation of the GMKL prediction performances in terms of Pearson correlation across different phenotypes simulations is $\sigma < 0.1$ (see Additional file 1: Fig. S1).

GMKL uses the kernel alignment heuristics to rank which kinship kernels are optimal for the prediction of each phenotype

An important point to clarify is that, contrarily to GBLUP(OP), our GMKL approaches *do not* access the information concerning the optimal set of kinship matrices for each phenotype. MEAN, FH, CKA, CKACLOSED, and GD have access to the five kinship matrices, and use their *kernel alignment heuristics* (see the "Methods" section) to *rank* and reweight the kinship kernels in function of their compatibility (correlation) with the ideal kernel matrix $\mathbf{K}_Y$ derived from the labels available during training. Figure 1 shows in shades of green (MEAN), red (FH), and blue (CKA) the results obtained by adding each kinship kernel matrix one at a time (the numbers 1–5 next to each GMKL model), following the order (ranking) provided by the respective MKL heuristic. The CKACLOSED and GD GMKL methods (dark and light pink in Fig. 1) always take as input all the five kinship kernel matrices and the kernel combination is made by jointly optimizing their linear combination weights on the training data.

Therefore our GMKL approaches do not depend on external information when it comes to determine the relevance of the IMs associated to each kinship matrix, as the GBPLUP(OP) model does, but it relies solely on a heuristic assessment of the kinship kernel compatibility with the phenotype at hand. This procedure can be considered analogous to a *feature reweighting* in kernel space.

The inheritance mechanisms generating the phenotype influence the prediction performances of the GP methods

The 1033 Holstein samples in the CATTLE dataset used in the Fig. 1 are associated to nine polygenic synthetic phenotypes generated in [25]. They involve different mixtures of Additive (A), Dominant (D), and three types of Epistatic effects: additive-additive (E_{AA}), additive-dominant (E_{AD}), and dominant-dominant (E_{DD}). Each of these synthetic phenotypes has a polygenic nature, with 1000 randomly sampled causative SNPs and a broad sense heritability set to 0.7. We report the exact proportions of the IMs involved in generating each phenotype in the Methods "Simulated phenotypes on cattle SNP array data" section and in Table 4. In the following we summarize them while we discuss the prediction results.

Predicting the phenotypes in the Additive-Dominant spectrum

The phenotypes ranging from zero to four involve only A and D effects, with Pheno:0 being 100% additive, Pheno:2 being 50/50%, and Pheno:4 being purely Dominant. From Fig. 1, we can see that for the purely additive Pheno:0, GBLUP(A), and GBLUP(OP) have the same Pearson correlation, because they are indeed the same model. Our GMKL approaches FH and CKA achieve the best Pearson correlation when using only one kinship kernel matrix, which indeed correspond to the A kinship. As intuitively expected, adding kinships built from other IMs generally lowers the prediction performance because of the pure additive nature of Pheno:0. The SVM with RBF kernel outperforms the linear kernel SVM (orange bars), and slightly outperforms the other methods, showing once again the multipurpose nature of this popular kernel.

The second row of plots in Fig. 1 show the performance measured with the Spearman correlation, which evaluates the ability of the predictors to establish a reliable *rank* over the phenotypic values of the samples. This is a relevant metric for GP and even clinical genetics, since it assesses how well predictors can sort individuals by EBV (in agrotech) or by disease risk (in clinical genetics). From the Spearman scores in the second row of Fig. 1, we can see that FH:1 and GD are still the best performing GMKL methods, while the RFB SVM drops some points with respect to GBLUP(OP) and GBLUP(A).

When looking at Phenotypes 1–5, we can see that increasing the role of the Dominance effect (D) leads first to a situation in which using multiple kinship kernels (red, blue and green bars in Pheno:1,3) is beneficial with respect to the GMKL models using only one kernel, indicating that indeed GMKL *benefits* from combining multiple genetic effects, if multiple IMs are involved in the target phenotype. At the same time, the performance of GBLUP(A) and of the linear and RBF SVMs steadily decrease. Interestingly, when predicting the purely Dominant Phenotype 4, we see again a situation in which the best Pearson correlation is obtained when only one (MEAN) or two (FH) kinship kernels are used by our models, which indeed in both cases prioritizes the D kinship matrix over the others. The large divide between GBLUP(A) and GBLUP(OP) showcases the risks of using the standard additive kinship matrix when the IM generating the phenotype is purely D instead. From both the Pearson and the Spearman scores on Pheno:4 we can see that all our GMKL methods drastically outperform the Deep NN (DNN) method from [25] and the conventional SVM and GBLUP(A) approaches.

On Phenotypes 1–4, our GMKL always performed similarly to the GBLUP(OP) model, which has access to the knowledge of which IMs are involved in each phenotype and therefore represents the *upper bound* baseline that could be achieved with perfect information on this data.

Adding Epistatic effects to the benchmark showcases the benefit of using GMKL

Phenotypes 5–9 are more complex, since they involve Epistatic effects between the causative markers. They are composed by a *base* 33% A and D components, plus a 34% epistatic component that is additive-additive E_{AA} in Pheno:5, additive-dominant E_{AD} in Pheno:6, and dominant-dominant E_{DD} in Pheno:7. The last phenotype (Pheno:8) is produced by the base A and D components plus it divides its 34% epistatic component equally among all the 3 types $E_{AA,AD,DD}$, 11% each. See Table 4 for details.

From the green, red, and blue bars in the right half of Fig. 1, we can see that with these phenotypes in which multiple IMs are involved, the best GMKL model is always one in which at least three or four kernel kinship matrices are combined, highlighting the beneficial effects of our GMKL approach. We can also see that for the most complex phenotypes (i.e., Pheno:7,8), our GMKL models outperforms even GBLUP(OP), both in terms of Pearson and Spearman scores. The joint optimization of the kinship kernel combination provided by CKACLOSED and GD is among the top GMKL performers in multiple phenotypes.

The difference between GBLUP(A) and GBLUP(OP) is the largest for the E_{AD} effect in Pheno:6, while it appears that the other epistatic components result in a relatively high overall additive effect, as hypothesized in [51]. Both GMKL and GBLUP(OP) consistently outperform the conventional linear and RBF SVMs, as well as the DNN model.

GMKL can formulate hypothesis on the inheritance mechanisms involved in generating phenotypic traits

As already mentioned, an important difference between the GMKL and the GBLUP(OP) scores shown in Fig. 1 is that, while the GBLUP(OP) method always uses the optimal set of kinship matrices to predict each phenotype (therefore exploiting the information on the IMs generating each phenotype), our GMKL approaches use the FH or CKA heuristics to infer the most relevant kinship matrices for the prediction of each phenotype from the training data.

The ability of our GMKL model to gauge the *compatibility* of the IMs modeled by the kinship matrices with the true genetic effects generating the phenotypes is therefore crucial for its GP performance. In Table 1, we evaluate the ability of the FH, CKA, CKACLOSED, and GD MKL approaches we used when it comes to detect the IMs underlying the synthetic phenotypes of the CATTLE dataset, using the true genetic effects proportions described in Table 4 as ground truth. Table 1 shows the true IMs, the ranking provided by each method and the proportion B_* of correct assignments given in the highest ranking k kinship matrices, where k corresponds to the number of IMs truly involved in generating each phenotype. Pheno:8 is not shown, since all the IMs are involved in generating it, and obtaining a perfect B_* score would therefore be trivial.

Table 1 Table showing the comparison between the True IMs used to generate the synthetic phenotypes in the CATTLE dataset with the rankings of the IMs proposed and used by the FH, CKA, CKACLOSED (called CLSD here), and GD GMKL methods for their predictions. The B_* score counts how many of them are correctly predicted in the best k ranked IMs, where k is the number of TrueIMs involved in each phenotype

Pheno	TrueIM	FH	B_{FH}	CKA	B_{CKA}	CLSD	B_{CLSD}	GD	B_{GD}
0	A	A	1	A	1	A	1	A	1
1	A,D	A, D	1	A, D	1	A, D	1	A, E_{DD}	0.5
2	A,D	A, D	1	A, D	1	A, D	1	A, E_{AD}	0.5
3	D, A	D, A	1	D, A	1	A, D	1	D, A	1
4	D	A	1	A	1	D	1	D	1
5	A, D, E_{AA}	A, D, E_{AA}	1	A, D, E_{AA}	1	A, D, E_{DD}	0.66	A, E_{AD}, E_{DD}	0.33
6	A, D, E_{AD}	A, D, E_{AA}	0.66	A, D, E_{AA}	0.66	A, D, E_{DD}	0.66	A, D, E_{AD}	1
7	A, D, E_{DD}	A, D, E_{AA}	0.66	A, D, E_{AA}	0.66	A, D, E_{DD}	1	A, D, E_{DD}	1

From Table 1, we can see that Phenotypes 0–4, which are in the Additive-Dominant spectrum, are correctly identified by all methods except GD, which is not able to identify the D component in Pheno:1–2, mistaking it for epistasis. All the methods, except CKA-CLOSED, rank the dominant kinship matrix higher than the additive in Pheno:3, when indeed the D component becomes predominant (75% D, 25% A, see Table 4).

From Table 1, it appears that these kernel alignment methods generally find it more difficult to correctly identify the exact type of Epistatic effects involved in the phenotype. This might be due to the fact that epistatic effects might result in an additive component [51, 52], thereby reducing the apparent role of epistasis, consequently lowering the relevance of the corresponding kernels. While FH and CKA correctly identify the presence of E_{AA} in Pheno:5, they are not able to identify the E_{AD} and E_{DD} epistatic effects in Pheno:6–7. GD, by contrast, shows once again some confusion between the Dominant and Epistatic effects in Pheno:5.

Overall, this benchmark on the synthetic phenotypes of the CATTLE dataset shows that FH and CKA methods are generally able to correctly rank which IMs are involved in the phenotypes under analysis, in particular when it comes to effects in the Additive-Dominant spectrum, while the precise type of Epistatic effect involved is harder to detect correctly. Nonetheless, all the methods can detect the presence of *some form* of epistasis, when at least one type is present.

Using GMKL for case-control discrimination of inflammatory bowel disease

Besides the synthetic phenotypes generated on the CATTLE dataset, we extended the validation of our model to clinically relevant human data. To do so, we first employed SNP array data from an in-house case-control inflammatory bowel disease (IBD) dataset containing 3522 samples (2646 cases, 876 controls) and 156,500 SNPs [53]. In this study we will further refer to this dataset as IBDSNP.

IBD is a multifactorial disease where genetic as well as environmental factors impact the gut microbiome, the intestinal barrier and the immune response, eventually resulting in chronic inflammation of the gastrointestinal tract [54]. The important socio-economic impact of this incurable and chronic disease, together with its rising global prevalence [55], makes it a very relevant test case. In the last decades, over 300 associated genetic loci have been identified [56] through genome-wide association studies, with NOD2 being one of the most relevant genes [57]. However, the intricate process of how these risk variants and genes interact together to produce the heterogeneous group of IBD phenotypes remains largely unresolved [50].

In Fig. 2 we show the benchmark, obtained in a fivefolds crossvalidation, comparing the prediction performance of our GMKL methods with GBLUP and conventional ML methods such as a RBF SVM, a logistic regression and a ridge classifier [58].

From the AUC and AUPRC scores in Fig. 2 it appears that GBLUP results are very similar, regardless of the number of kinship matrices used to specify the model covariance structure.

On the other hand, the GKML MEAN, FH, and CKA models monotonically benefit from the combination of multiple kinship matrices, suggesting that multiple IMs could be at play. The joint kernel combinations provided by CKACLOSED and GD provide similar prediction scores as well.

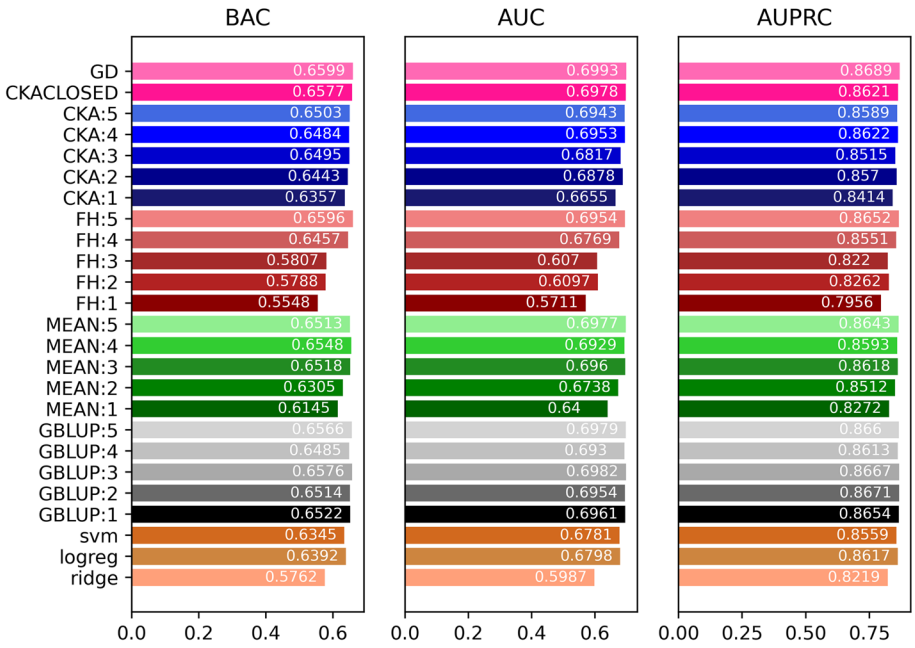

Fig. 2 Figure showing the balanced accuracy (BAC), area under the ROC curve (AUC), and area under the precision-recall curve (AUPRC) for the benchmark of our GMKL approaches with GBLUP and other ML methods on the IBDSNP dataset

Table 2 Table showing the rankings proposed by the FH, CKA, and GD methods for the IMs described by the A, D, E_{AA}, E_{AD}, and E_{DD} kinship kernels on the IBDSNP dataset

FH ranking	FH weights	CKA ranking	CKA weights	CKACLOSED ranking	CKACLOSED weights	GD ranking	GD weights
E_{AA}	0.364	D	0.027	D	1818.91	E_{AA}	0.343
E_{DD}	0.269	A	0.025	E_{DD}	1516.01	E_{DD}	0.323
E_{AD}	0.232	E_{DD}	0.018	A	1205.03	E_{AD}	0.221
D	0.070	E_{AD}	0.018	E_{AD}	917.83	D	0.056
A	0.062	E_{AA}	0.018	E_{AA}	558.59	A	0.055

GMKL can formulate hypothesis on the inheritance mechanisms underlying IBD

What is more interesting, is the way in which our GMKL methods rank the contribution of the 5 IMs under scrutiny. We summarize these rankings in Table 2. FH and CKA substantially disagree on the rankings of the IMs, with CKA placing Dominance in the first place, followed by Additivity and the Epistatic effects. The FH and GD rankings, on the other hand give more relevance to the epistatic effects, in the following order: E_{AA}, E_{DD}, and E_{AD}. The weights assigned to the A and D IMs are negligible for both FH and GD. The similarity between FH and GD can be explained by the fact that, as shown in Eq. 12, GD and FH both use the F-heuristic (see Eq. 8) in their computation, with the difference that GD jointly optimizes the linear combination of the kernels, while FH does it in a univariate way. CKACLOSED ranking is closer to CKA, except for the inversion between the E_{DD} and A effects. Since CKACLOSED performs a joint optimization of the kernel weights, including a whitening transformation to remove linear correlations between them, we consider its IM hypotheses to be the most mathematically sound (see

the Methods "Using kinship matrices for biological meaningful Genomic Multiple Kernel Learning" section and [49]).

Validating the GMKL predictions on an independent whole exome sequencing IBD dataset

To further validate the GMKL ability to discriminate between IBD cases and controls, and the consistency of the IMs rankings produced across datasets, we benchmarked our GMKL methods on a completely different IBD case-controls Whole Exome Sequencing dataset from dbGAP. It contained 3798 samples (3318 cases vs. 480 controls), and it has been used in [50] to train and test end-to-end GI Neural Networks (NN) models for case-control *in silico* discrimination. We refer to this dataset as IBDWES in this study, to differentiate it from the previously used SNP array-based dataset. To adapt the WES data to the conventional genotype matrix format that GP methods take as input, we transformed the 2,121,171 variants in IBDWES in that standard format, representing each variant with its zygosity (0,1,2). We then filtered them by removing monomorphic variants and by keeping only variants with $0.01 < \text{MAF} < 0.2$, to keep the computation of the kinship matrices feasible by limiting the RAM used by the Sommer R package. Since this MAF filtering might be too stringent, removing the most variable loci in the dataset, we compared it with other approaches. They include (i) the random sampling of 50k variants, (ii) the selection of a non-redundant set of variants, ensuring that their all-against-all Pearson correlation is < 0.5, and (iii) the selection of just the variants with $0.2 < \text{MAF} < 0.5$. The results of this comparison are shown in Additional file 2. In 10 cases of the 17 total (59%), our original MAF filtering was the best performing in terms of AUC. The second best approach is the sampling of a non-redundant set of variants, which yields the highest AUC 29% of the times. See the "Methods" for the details about the IBDWES dataset and the processing steps.

Figure 3 shows the fivefold cross validation performance of our five GMKL methods, compared with GBLUP, two conventional ML methods (RF and logistic regression), and three GI NN models from [50]. On this data, our MEAN models slightly outperform the GBLUP models, but AUC scores are generally very similar. In general, the benefit of combining multiple kinship kernels is less pronounced than in the IBDSNP dataset, but still noticeable, particularly for MEAN.

Linear models could reach high performance by modeling the additive component of epistasis

On the IBDWES and IBDSNP data, the additive GBLUP:1 model reaches high performance, but at the same time shows little benefit from the introduction of additional kinship matrices. This behavior can be explained by the fact that real-life datasets generally have noisy and heterogeneous conditions that may confound or dilute the inheritance mechanisms signals. In these settings, the low complexity (high bias) of the linear models is beneficial as it translates into robustness [50].

From a biological perspective, the high performances of additive models such as GBLUP:1 can be further explained by the possibility to *approximate epistatic effects by their additive components*, as shown in [51]. Even purely epistatic effects might result in a noticeable additive component [52, 59–61], which could explain why nonepistatic mechanisms could be partially explanatory for the phenotype [52], in particular at

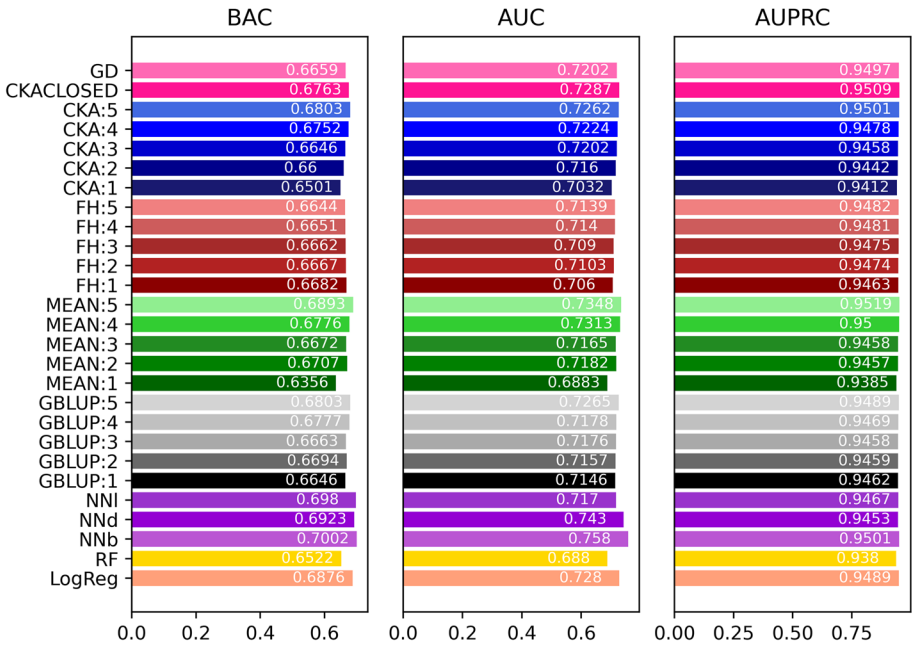

Fig. 3 Figure showing the comparison between the nine methods we benchmarked on the IBDWES dataset. The MEAN, FH, and CKA GMKL approaches using one to five kernel matrices are respectively shown in shades of green, red, and blue. The black to gray bars show the additive GBLUP models, while the bottom bars show the performance obtained by the Neural Networks methods proposed in [50]

relatively low sample size, where it is still difficult to reliably infer more complex patterns [50].

Deep learning GI models offer the best performance

The best performing models in terms of AUC are the NNd and NNb neural networks proposed in [50] (purple bars), while the BAC and AUPRC scores are more similar. Besides the fact that these end-to-end NNs do not filter out any of the WES variants as a preprocessing step, the main difference with GMKL lies in the way the genetic data is integrated into the model. In the kernel methods, the genetic data is used to compute a marker-based similarity between the samples, resulting in kinship matrices that no longer hold information on the individual variants and genes, because of the dataset underdetermination and the impossibility to compress information about individual variants in a small kinship matrix. While these matrices can take into account inheritance mechanisms, they are not specifically tailored to the phenotype and its causal variants.

Except for the newly introduced NNd and NNb e2e NN GI models, which have a slightly higher AUC, Fig. 3 shows that our GMKL approaches consistently perform on par with state-of-the-art methods and more conventional approaches on two unrelated IBD datasets, generated with completely different sequencing technologies and cohorts.

The inheritance mechanisms identified by GMKL are similar across IBD datasets

We then analyzed the ranking of the kinship kernel matrices produced by our GMKL methods on the IBDWES dataset, comparing them with the ones we previously obtained on the IBDSNP dataset. These results are shown in Table 3. We can see that the ranking

Table 3 Table showing the rankings proposed by the FH, CKA, and GD methods for the IMs described by the A, D, E_{AA}, E_{AD}, and E_{DD} kinship kernels on the IBDWES dataset

FH ranking	FH weights	CKA ranking	CKA weights	CKACLOSED ranking	CKACLOSED weights	GD ranking	GD weights
E_{AA}	0.378	D	0.028	E_{DD}	6600.12	E_{AA}	0.316
E_{AD}	0.309	A	0.026	D	3919.44	E_{AD}	0.300
E_{DD}	0.265	E_{AA}	0.019	E_{AD}	1676.01	E_{DD}	0.280
D	0.024	E_{AD}	0.019	A	1142.83	A	0.052
A	0.022	E_{DD}	0.019	E_{AA}	437.25	D	0.051

Table 4 Table summarizing the inheritance mechanisms involved in generating the nine synthetic phenotypes used in DS1. The numbers indicate the mean contribution of each type of inheritance towards the simulated phenotype. These values have been reported from [25] and shown here for clarity

Phenotype Number	Mean inheritance mechanism contribution					Description
	μ_A	μ_D	$\mu_{E_{AA}}$	$\mu_{E_{AD}}$	$\mu_{E_{DD}}$	
0	100	0	0	0	0	Purely additive phenotype
1	75	25	0	0	0	Predominantly additive
2	50	50	0	0	0	Mixture of additive and dominance
3	25	75	0	0	0	Predominantly dominant
4	0	100	0	0	0	Purely dominant phenotype
5	33	33	34	0	0	Mixture of additive, dominant and E_{AA}
6	33	33	0	34	0	Mixture of additive, dominant and E_{AD}
7	33	33	0	0	34	Mixture of additive, dominant and E_{DD}
8	33	33	11.3	11.3	11.3	Mixture of all the effects considered

provided by FH is quite similar to the one shown in Table 2, in the sense that the epistatic effects are deemed the most relevant, with once again negligible weights for the A and D IMs. The only difference is that the relevance of the E_{DD} and E_{AD} is swapped between the two datasets.

Also the CKA ranking is macroscopically identical, in the sense that D and A effects are indicated to be the most relevant, while the epistatic effects have 32% lower weights, but are still not negligible.

The ranking provided by the GD method is identical to the FH column in Table 3 and similar to the GD results shown in Table 2, in the sense that also in this case the epistatic effects are indicated to be the most relevant, with the same inversion between E_{DD} and E_{AD} shown in the FH ranking. Both these methods assign one order of magnitude lower weights to the A and D IMs. Finally, there is little agreement between the IM ranking provided by CKACLOSED between the IBDSNP and IBDWES datasets.

Discussion

Epistatic effects may be important genetic drivers of IBD

IBD is a multifactorial disease where environmental and genetic factors produce perturbations on different pathways that are relevant for the homeostasis between the gut microbiome, the intestinal barrier, and the host immune system, eventually leading to

the dysregulated immune response and chronic inflammation that characterizes the disease. Multiple identified risk and protective variants can be mapped on different disease pathways, together leading to the complex spectrum of IBD symptoms and phenotypes [62]. It is therefore not surprising that variants and genes affecting such an intricate system would exhibit complex interaction patterns, such as epistasis. Our previous work on the case-control prediction of IBD showed empirical evidence for the presence of epistasis between genes [50], by demonstrating a predictive advantage of including such interactions in the model. This has already been hypothesized for Crohn's disease [52], and it has been shown for severe very-early onset IBD, which exhibits Mendelian inheritance patterns with casual rare genetic variants [62], and where interactions between multiple genetic factors can modulate its severity [57, 63].

Since our GMKL method uses heuristics to evaluate the compatibility between kinship kernels corresponding to different inheritance mechanisms (IMs), it can be used also to empirically formulate hypothesis on the IMs involved in generating the predicted phenotypes. In Table 1, we benchmarked the reliability of the identified IMs on the CATTLE dataset synthetic phenotypes, showing that identifying phenotypes in the A-D spectrum is relatively easy (in particular for FH and CKA), while discerning the specific types of Epistasis is more difficult. Nonetheless, the methods appear to be reliable when they are just asked to identify the presence or absence of *any* form of epistasis.

In Tables 2, 3, we used our GMKL to *rank* the kinship kernels corresponding to the 5 IMs hypothesis we consider, on two different IBD cohorts. While the results are not identical, much is preserved between the two tables, even if the two datasets are completely different in terms of samples and sequencing technology (SNP array and WES). In particular, if we collapse the three types of epistasis into a single category E_*, we see that all methods assign a nonnegligible relevance to it. FH and GD rank it the highest, CKACLOSED assigns the highest ranking to a form of epistasis in both datasets, and the weights assigned to E_* kernels by CKA is just around 30% lower than A and D, and we therefore cannot consider it as an attempt to completely discard it, like the A and D matrices for FH and GD (see Tables 2 and 3).

To empirically verify the relevance of the epistatic effects E_* for the IBD prediction, thereby possibly corroborating our previous findings [50], we compared the prediction performance obtained with the GMKL models when respectively considering only A, D or A, D, E_* effects on both datasets. On the IBDSNP dataset, adding the three E_* kernels on top of the A and D ones increases AUC performances significantly for all the GMKL methods (DeLong [64] p-values: 1.7×10^{-11}, 7.9×10^{-05}, 5.2×10^{-08}, 6.4×10^{-15}, and 3.3×10^{-05} for MEAN, FH, CKA, CKACLOSED, and GD, respectively). We see similar results on the IBDWES dataset, with the AUC improvements because of the inclusion of the E_* kernels being significant for the MEAN (DeLong [64] $p = 0.0003$), CKA ($p = 0.04$), and CKACLOSED ($p = 0.0003$) GMKL methods.

Similarly to our previous work [50], here we provide empirical evidence for the importance of epistasis in the genotype-to-phenotype modeling of IBD, although through a completely different approach and data. In both studies, the incorporation of epistasis, although through two conceptually different computational approaches, enhances prediction. Moreover, in contrast to most existing evidence for epistasis in IBD, typically investigating the interaction between only a limited set of susceptibility loci [65–69], our

GI approaches manage to model epistasis without the need to explicitly predefine the interacting variants or genes.

Using the kernel alignment heuristics to rank inheritance mechanisms

While the kernel alignment heuristics used in MKL can improve the predictive signal and lower the contribution from kernel matrices that poorly correlate with the target labels, the exact weights used to combine these kernel matrices have limited impact on the end results. Indeed, lowering the contribution of a kernel matrix only reduces the variance associated with the features in the corresponding high-dimensional space, but these features can, under certain circumstances, still be exploited by the ML model for prediction. More specifically, lowering the weight of a kernel has the effect of decreasing the scale of its corresponding features in the high-dimensional space, and therefore shifting them toward lower principal components. However, the number of dimensions visible to the ML model is bounded by the size of the training set. Indeed, any positive-definite kernel matrix of size $n \times n$ can be decomposed into n eigenvector-eigenvalue pairs, therefore limiting the number of dimensions to n in practice. Because the latter is finite, the consequence is that there is always a point at which reducing the weight of a kernel ensures that the corresponding features are (almost) no longer reflected in the n principal components, and can no longer be used by the ML model for prediction. Therefore, the differences in weights that should be highlighted when using our GMKL to hypothesize the IMs underlying the phenotype under prediction are the most drastic drops, such as the ones assigned by FH to D and A IMs in Tables 2, 3, instead of more gradually decreasing rankings (i.e., CKA weights in Table 3).

How to choose a MKL heuristic for genomic prediction

In this article, we used two kernel alignment heuristics (FH [48] and CKA [49]) to build and benchmark five GMKL methods (MEAN, FH, CKA, CKACLOSED, and GD) for GP, showing that these methods can both predict the phenotypes *and* formulate hypotheses on the IMs underlying them. But which of the GMKL approaches we tested should be used in practice, on unseen data?

In terms of predictions, these GMKL methods perform similarly, without a clear winner, as shown in Figs. 1, 2, 3. In terms of the detection of the IMs underlying the predicted phenotype, there is a disagreement between the FH and CKA alignment methods, as shown in Tables 2 and 3. This disagreement is due to the algorithmic differences between these two approaches: while they all build on the idea of kernel alignment, in CKA and CKACLOSED, the alignment is computed after centering the kernel matrices [49], therefore turning the alignment function into something closer to an actual correlation metric, while this step is not considered in FH, MEAN, and GD [48] (see the Methods "Using kinship matrices for biological meaningful Genomic Multiple Kernel Learning" section). The centering of the kernels is a theoretical improvement of CKA/CKACLOSED over FH/GD, as explained in [49]. Moreover, CKACLOSED performs a whitening transformation on the kernel alignments to remove the linear correlations between them. Because the CKACLOSED algorithm appears to be the most mathematically sound kernel alignment method currently available in literature, we recommend the users to rely on this approach for their analyses. This might change in the future,

if novel approaches are proposed, for example allowing a nonlinear combination of the kernels.

Conclusion

In this article, we propose novel Genomic Multiple Kernel Learning (GMKL) methods for Genomic Prediction and Genome Interpretation (GI). GMKL revolves around using the kinship matrices commonly used to determine the similarity between samples in GBLUP models as biologically meaningful kernels in support vector machines (SVM) models, using them for the regression or binary classification of quantitative and phenotypic dichotomic traits. Several kinship matrices, mirroring different inheritance mechanisms (IMs), such as Additive, Dominant, and different Epistasic models, can be combined with kernel alignment techniques, obtaining GMKL models with a complete view of the IMs underlying the phenotype under study.

We show that our GMKL approach positively compares with conventional Machine Learning (ML) approaches and GP methods, such as GBLUP, evaluating it both on synthetic data and two inflammatory bowel disease (IBD) cohorts.

More importantly, we show that our GMKL approach is can *rank* the IMs generating the phenotype by evaluating the compatibility of the kinship kernels corresponding to each IM with the training set data. This allows our approach to effectively produce hypotheses on the IMs generating the target phenotypes, opening new possibilities for the understanding of the relationship between genomic data and the observed phenotypes, and taking a step towards Explainable ML.

Methods

Datasets

Simulated phenotypes on cattle SNP array data

We retrieved from [25] an Illumina BovineSNP50 Beadchip dataset involving 1033 pure-bred Holstein Friesian samples genotyped [21]. After quality control based on call-rate, MAF and the removal of SNPs with unknown map position or mapped on the sex chromosomes [21, 25], the final dataset contains 26,503 SNPs. We also retrieved nine simulated phenotypes computed on this data [25]. We refer to the combination of these genotypes and the corresponding simulated phenotypes as CATTLE dataset in this paper.

We summarize here the procedure used in [25] to generate them. These phenotypes have been generated with the Simphe R package [70] with different mixtures of Additive (A), Dominant (D), and Epistatic effects. Three types of epistasis have been considered: additive-additive (E_{AA}), additive-dominant (E_{AD}), and dominant-dominant (E_{DD}) [25].

To mimick the polygenicity of the traits, the simulated phenotypes are functions of 1000 randomly-sampled SNPs. We refer to these SNPs as Quantitative Traits Markers (QTMs), since they are artificially *causative* for the nine phenotypes. To each QTM, Simphe assigned an effect sampled from Gaussian distributions with inheritance mechanism-specific means $\mu_A, \mu_D, \mu_{E_{AA}}, \mu_{E_{AD}}, \mu_{E_{DD}}$, and variance σ^2_* equal to 10% of the corresponding mean μ_*. The sign of the effect of each QTM is randomly generated, with equal probabilities of positive and negative contributions to the final phenotype to remove biases toward reference/alternative alleles[25]. Additional Gaussian noise has

been added to simulate a broad-sense heritability of 0.7. The mixtures of A, D, and E_* effects used to generate these phenotypes are summarized in Table 4. See [25] for more details.

The inflammatory bowel disorder SNP array data

SNP array data of an in-house case-control dataset including 2646 IBD cases and 876 controls was used. Genotyping was performed using Immunochip (Illumina). Immunochip is a high-throughput genotyping chip based on the Illumina Infinium chip including approximately 240,000 SNPs [71]. The majority of these (196,524 SNPs) are based on genome-wide association studies of 12 autoimmune and inflammatory diseases, including Crohn's disease (CD) and ulcerative colitis (UC). The remaining approximately 25,000 SNPs are from other diseases and included as control (null-SNPs). The main puproses of the Immunochip were finemapping of known loci, and replication of suggestive associations. SNPs on the Immunochip were mapped on the GRCh37/hg19 build of the human genome.

Quality control (QC) on the genotype data was performed according to [53] with missingness per person < 0.02, heterozygosity rate within 95% interval, missingness per SNP < 0.02 and Hardy-Weinberg equilibrium p-value (computed on controls) $> 10^{-10}$. We then applied a final number of SNPs after QC is 156,500. We refer to this dataset as IBDSNP.

Whole exome sequencing IBD data

From the inflammatory bowel disease (IBD) Exome Sequencing Study (dbGaP phs001076.v1.p1) [72], we retrieved 3318 IBD cases and 480 controls. Similarly to IBDSNP, the 3318 cases consist of the two main subtypes of IBD: 2036 Crohn's disease (CD) patients and 1215 ulcerative colitis (UC) patients[50]. For 67 cases, the IBD subtype is unknown (indeterminate colitis). In the control group, 39.4% of the participants are male compared to 46.7% of the cases. The data is provided as a VCF file listing the observed variants. From this, a total superset of 2,121,171 biallelic variants was extracted, describing them with a 0,1,2 value indicating their zygosity in each sample. We then filtered out the monomorphic variants, and we applied a MAF filtering. This time we used 0.2 as maximum MAF threshold instead of the 0.5 used on the previous datasets because the `sommer` [44] R package used to compute the kinship matrices was requiring excessive RAM memory. This resulted in a total of 118001 variants. We refer to this dataset as IBDWES.

Building kinship matrices

Genotypes coming from SNP array data are traditionally represented as a $N \times M$ genotype matrix $\mathbf{M}$ where N is the number of samples, M the number of observed SNPs and each position $M_{i,j}$ contains the simplified allele count 0, 1, 2 for biallelic loci. This representation is also called marker count matrix or minimal allele count matrix [34] in literature, and it is used as input features for ML [47] and statistical methods such as MBLUP [36], linear regressions [13, 39], and SVMs with conventional kernels [19, 31, 47].

Kinship matrices (also called Genomic Relationship Matrices) are generally computed from the genotype matrix $\mathbf{M}$ with some variation of a scaled and centered inner product

$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}^\top}{m}$ [8, 36, 43] and therefore are Positive Definite symmetric matrices. They can be built to measure the similarity between genomic samples from the point of view of several genetic inheritance mechanisms, including additive (A), dominant (D), additive-additive (E_{AA}), additive-dominant (E_{AD}) and dominant-dominant (E_{DD}) epistasic effects[25, 43, 45].

In this article we compute the A, D, E_{AA}, E_{AD}, and E_{DD} kinship matrices with the R package `sommer` [44]. We first filtered the input genotype matrices M to include only markers with Minor Allele Frequency (MAF) between 0.01 and 0.5.

Linear Mixed Models and RKHS Regression for Genomic Prediction

The Genomic Best Linear Unbiased Prediction (GBLUP) is a member of the Linear Mixed Model (LMM) family [33, 36] that has been adapted for GP[35, 36]. It leverages genomic information, typically represented by SNP array data, to estimate the genetic merit of individuals for agriculturally relevant traits [13]. The model can be written as

$$\mathbf{y} = \mu + \mathbf{u} + \mathbf{e}, \tag{1}$$

where μ is the dataset mean for the trait $\mathbf{y}$, $\mathbf{u}$ is the vector of random genetic effects, assumed to follow a multivariate normal distribution $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}\sigma_a^2)$, $\mathbf{G}$ is a genomic relationship (kinship) matrix and $\mathbf{e}$ is the vector of residual errors, assumed to follow a normal distribution $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, I\sigma_e^2)$ [25]. GBLUP derives the genetic information from the kinship matrix $\mathbf{G}$, which captures the pairwise genetic relationships among individuals, usually in terms of additive effects. Reproducing Kernel Hilbert Space (RKHS) regression extends previous model by allowing G to be constructed from pairwise evaluations of a reproducing kernel, therefore implicitly constructing a Hilbert space, and enabling the nonlinear modeling of genomic information [28, 30].

Also, the GBLUP model in Eq. 1 can be extended to include multiple kinship matrices at the same time [25]:

$$\mathbf{y} = \mu + \sum_{i \in IM} \mathbf{u_i} + \mathbf{e} \tag{2}$$

where each random effect $\mathbf{u}_i$ is sampled from a different random variable $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_i\sigma_i^2)$ whose covariance matrix $\mathbf{G}_i$ is defined by the kinship matrix computed considering different inheritance mechanisms $IM = \{A, D, E_{AA}, E_{AD}, E_{DD}\}$, corresponding to additive, dominant or epistatic effects [43–45]. Usually it is suboptimal to include kinship matrices corresponding to inheritance mechanisms that are not involved in the phenotype under study [25], and the standard GBLUP model shown in Eq. 1 is preferred, equipped with just the additive kinship matrix $\mathbf{G}_A$.

In this article, we use two versions of the GBLUP model as baseline. We call the first GBLUP(A) because it corresponds to Eq. 1 and uses only an additive kinship matrix. The second is based on Eq. 2 and it is called GBLUP *Optimal* (GBLUP(OP)) because on the CATTLE dataset it always has access to the optimal set of random effects that correspond to all the known inheritance mechanisms involved in generating the nine synthetic phenotpes, and it is therefore an approximation of the optimal predictions that can be obtained on this data. See Table 4 for the details of the inheritance mechanisms involved in each synthetic phenotype. We implemented these GBLUP models with the

BGLR R package [29]. We provide a Python wrapper to this package here: https://bitbucket.org/eddiewrc/gmkl/src/main.

Kernel learning methods

Kernel methods are a type of instance-based learners. Instead of learning the parameters $\mathbf{w}$ associated to the feature representation $\phi(\mathbf{x}_i)$ for each target object $\mathbf{x}_i$, they learn a weight a_i for each training sample $(\mathbf{x}_i, y_i) \in \mathcal{T}$. The prediction of unseen samples $\mathbf{x}_j \notin \mathcal{T}$ is computed as the weighted sum of a similarity function $k(\mathbf{x}_j, \mathbf{x}_i)$ between the target sample $\mathbf{x}_j$ and the training samples $\mathbf{x}_i \in T$.

This function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric positive definite function commonly called *kernel function*. It corresponds to the inner product $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ of the samples $\mathbf{x}$ in the feature space produced by $\phi(\mathbf{x})$. The representer theorem [73] states that a function f (i.e., a ML model) minimizing an empirical risk function (i.e., an objective function used for training f) can be equivalently be expressed as the (1) linear combination of the weights $\mathbf{w}$ and the features $\phi(\mathbf{x})$ or (2) as an instance-based model obtained by the linear combination of a parameter a_i for each training sample $\mathbf{x}_i$ and the similarities between samples provided by K:

$$f(\mathbf{x}') = \langle \mathbf{w}, \phi(\mathbf{x}') \rangle + b = \sum_{i=1}^{N} a_i K(\mathbf{x}', \mathbf{x}_i) + b. \tag{3}$$

The right hand side of Eq. 3 formula is called *kernel expansion* and provides an equivalent *dual formulation* that allows expressing the relationships between data points $\mathbf{x}_i$ in a transformed space without explicitly computing the transformations $\phi(\mathbf{x}_i)$, enabling working in high-dimensional or even infinite-dimensional spaces, as in the case of Radial Basis Functions (RBF) kernels.

Support vector machines

Support vector machines (SVMs) are a specific instance of kernel learning methods that aim at finding the best hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ separating the two classes, namely the one with the largest distance $2/\|\mathbf{w}\|$ from any of the samples [46]. Here, we briefly recap the SVM algorithm for the binary classification case, see [74] an in-depth explanation.

Given a dataset $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ with n samples $\mathbf{x}_i \in \mathbb{R}^m$ and prediction labels $y_i \in \{-1, 1\}$, SVMs aim at finding the optimal hyperplane separating the two classes. Hyperplane $\mathbf{w}$ is found by solving the following constrained optimization problem:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \mathbf{w}^T\mathbf{w} + C \sum_{i=1}^{n} \zeta_i, \\ &\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \zeta_i \qquad \forall i \in \{1, \ldots, n\}, \\ &\qquad\qquad \zeta_i \geq 0 \qquad\qquad\qquad\qquad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{4}$$

where C is the penalty term for misclassified samples (soft-margin SVM). Contrarily to the kernel expansion in Eq. 3, the SVM algorithm does not consider *all* the training samples when computing the decision function. The decision hyperplane is instead based on a subset of the data, called *support vectors*, for which the constraints in the second term of Eq. 3 are active, meaning that they lie inside the margin: $\langle \mathbf{w}, \mathbf{x}_i \rangle + b = y_i$. SVMs are therefore also called *sparse* kernel machines. The Lagrangian of this problem is given by

$$\mathcal{L}(w, b, \zeta, a, r) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\zeta_i - \sum_{i=1}^{n} a_i\Big(y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) - 1 + \zeta_i\Big) - \sum_{i=1}^{n} r_i\zeta_{ic}, \tag{5}$$

where r, a are Lagrange multipliers.

After solving the Karush-Kuhn-Tucker conditions, the SVM objective function becomes

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i\alpha_i k(\mathbf{x}_i, \mathbf{x}_j) y_j\alpha_j \\ \text{subject to} \quad & C \geq \alpha_i \geq 0 \qquad\qquad \forall i \in \{1, \ldots, n\}, \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \qquad\qquad . \end{aligned} \tag{6}$$

In the dual formulation (Eq. 6), the kernel $k(\mathbf{x}_i, \mathbf{x}_j)$ emerges, allowing the SVM to learn a linear decision function (hyperplane) over an implicit space that is a possibly highly nonlinear transformation of the original data space wherein each $\mathbf{x}_i$ point lies. In Additional file 1: Section S1 we discuss the connection between SVMs and the RKHS regression methods used in GP.

Multiple kernel learning

Various kernels have been designed to operate on specific types of data, including documents, strings[49], text[75], graphs, trees [76], and biological sequences [77, 78].

Selecting the appropriate kernel K_i for the data at hand is indeed crucial [49, 79], and ways to combine P kernels $K_P(\mathbf{x}_i, \mathbf{x}_j) = f_\nu\Big(\{K_p(\mathbf{x}_i^p, \mathbf{x}_j^p)\}_{p=1}^{P}\Big)$ have been developed [79]. They can be used to incorporate in the SVM model different notions of similarity on the same data, or to combine heterogeneous sources of information, evaluating each with a data-specific kernel [79, 80].

These techniques fall under the term of multiple kernel learning (MKL) [79]. Initial MKL approaches focused on modifying the SVM optimization algorithm to jointly learn the SVM parameters and the parameters ν of the function f_ν used to combine the kernels (one-step approches) [79, 81], but the added computational complexity and the generally disappointing results [49, 82] pushed researchers towards *two-steps* approaches in which the combined kernel $K_P(\mathbf{x}_i, \mathbf{x}_j)$ is first devised *offline*, for example using heuristic techniques, and it is then used in a conventional SVM [49].

Using kinship matrices for biological meaningful Genomic Multiple Kernel Learning

In this article, we compare five heuristic two-steps MKL methods to combine kinship kernels for GP, and we describe them here. For each dataset used in this article, we computed the additive (A), dominant (D), additive-additive (E_{AA}), additive-dominant (E_{AD}), and dominant-dominant (E_{DD}) epistasic effects[25, 43, 45] with `sommer` [44]. We refer to these kinship kernels as *base kernels* in the following text, to avoid loss of generality since the MKL methods described here are valid for any choice of input kernels.

These methods assume that the theoretically optimal kernel is the ideal one derived from the outer product of the training labels $\mathbf{K}_Y = \mathbf{Y}\mathbf{Y}^\top$, and that each kernel K_i can be *ranked* in function of its *alignment* with $\mathbf{K}_Y$ [83]. We use two notions of alignment between kernels,

the F-heuristic proposed in [48] and the Centered Kernel Alignment (CKA) proposed in [49]. The F-heuristic measures the correlation between two matrices and it is defined as

$$F(\mathbf{K}_i, \mathbf{K}_Y) = \frac{\langle \mathbf{K}_i, \mathbf{K}_Y \rangle_F}{\sqrt{\langle \mathbf{K}_i, \mathbf{K}_Y \rangle_F \langle \mathbf{K}_Y, \mathbf{K}_Y \rangle_F}} \tag{7}$$

where $\langle \cdot, \cdot \rangle_F$ indicates the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle_F = \sum_{i=1}^{L} \sum_{j=1}^{L} a_{ij} b_{ij}$ for two matrices $L \times L$ [48].

FH method: The F-heuristic computes the Pearson correlation between kernels, and these values are then used to define a linear combination $\mathbf{K}_{\text{FH}}$ of the kernels $\mathbf{K}_i$ based on these alignment weights:

$$\mathbf{K}_{\text{FH}} = \sum_{i}^{P} \mu_i \mathbf{K}_i \quad \text{where} \quad \mu_i = \frac{F(\mathbf{K}_i, \mathbf{K}_Y)}{\sum_h F(\mathbf{K}_h, \mathbf{K}_y)}, \tag{8}$$

we refer to this approach as FH in the results section of the article.

MEAN, CKA, and CKACLOSED methods: The CKA method extends this approach by centering the kernels in the feature space first [49], and then computes the F-heuristic alignment on them. The authors of [49] claim that without centering, there is no guarantee that the alignment provided by F (Eq. 7) truly correlates with the prediction performance. To combine the centered base kernels, the authors of [49] propose three approaches. The simplest one, that we call MEAN, is to compute the combined kernel as the arithmetic average of the centered kinship base kernels:

$$\mathbf{K}_{\text{MEAN}} = \frac{1}{P} \sum_{i=1}^{P} \mathbf{K}_i^c, \tag{9}$$

where we used the superscript ^{c} to denote the centering described in [49].

The second approach, which we call CKA, computes the combined kernel:

$$\mathbf{K}_{\text{CKA}} = \sum_{i=1}^{P} F(\mathbf{K}_i^c, \mathbf{K}_Y) \mathbf{K}_i^c \tag{10}$$

as a linear combination of the centered kinship kernels $\mathbf{K}_i^c$, weighted proportionally to their F alignment with the perfect kernel $\mathbf{K}_Y$.

The weighted linear combination above computes the kernel weights independently from each other, meaning that it ignores the possible correlation between the base kernel matrices [49]. To overcome this issue, in [49] they proposed a method to determine the kernel mixture weights μ_i jointly. We call this approach CKACLOSED, since it can be computed in the following closed form:

$$\mathbf{K}_{\text{CKACLOSED}} = \sum_{i=1}^{P} \mu_i \mathbf{K}_i^c \quad \text{where} \quad \boldsymbol{\mu} = \frac{\mathbf{M}^{-1}\mathbf{a}}{\|\mathbf{M}^{-1}\mathbf{a}\|_2}, \tag{11}$$

where M is the matrix defined as $\mathbf{M}_{ij} = \langle \mathbf{K}_i^c, \mathbf{K}_j^c \rangle_F \forall i, j \in [1, P]$ and $\mathbf{a}$ is the vector $\mathbf{a} = (\langle \mathbf{K}_i^c, \mathbf{K}_Y \rangle_F, ..., \langle \mathbf{K}_P^c, \mathbf{K}_Y \rangle_F)^\top$ [49].

GD method: The last method is called GD because it uses `pytorch` [84] Gradient Descent optimization to learn the optimal weights $\mathbf{w}$ to combine the base kernels to maximize the alignment $F(\mathbf{K}_{\text{GD}}, \mathbf{K}_Y)$, with the following maximization:

$$\max_{\mathbf{w}} F(\sum_{i=1}^{P} \mathbf{K}_i^c \ \text{Softmax}(\mathbf{w})_i, K_Y). \tag{12}$$

We applied a softmax to the weights $\mathbf{w}$ to ensure that each $w_i \geq 0$, therefore ensuring that the resulting combined matrix $\mathbf{K}_{\text{GD}}$ is PSD, since the sum of kernels is a kernel.

Evaluation of the performance

We evaluated the prediction performance using a fivefold cross-validation on all the datasets used. The metrics used are specific to each dataset. The synthetic phenotypes in the CATTLE dataset are real-valued, and therefore we evaluated them with the Pearson correlation and the Spearman correlation which evaluates the ability of the predictors to establish a reliable *rank* over the samples. This is a relevant metric in GP, since it measures how reliably these methods can be used to select the samples with the highest breeding value, assuming that higher phenotypic values are desirable [25].

The two IBD datasets provide a binary classification problem (cases vs. controls), and therefore we used the area under the receiver operating characteristic (ROC) curve (AUC), the area under the precision-recall curve (AUPRC), and the balanced accuracy, which is the mean between sensitivity and specificity.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03544-3.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements
DR is grateful to Anna Laura Mascagni for the constructive discussion. DR is grateful to N. Nazzicari and F. Biscarini for the help understanding the code and the data associated to their article [25], and for sharing additional scripts. DR is grateful to P. Soerensen for his great didactic material on quantitative genetics, BLUPs and LMMs (available at https://psoerensen.github.io/qgnotes/).

Review history
The review history is available as Additional file 3.

Peer review information
Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
DR, AP, NV, and YM conceived the experiments. DR, AP, and NV developed the methods. IC, DSJ, and NV provided the data. DR, NV, AP, IC, and YM wrote the manuscript.

Funding
DR is funded by a FWO senior postdoctoral fellowship (grant number 12Y5623N). Antoine Passemiers is funded by a FWO doctoral fellowship (1SB2721N). Research Council KU Leuven: Symbiosis 4 (C14/22/125), Symbiosis3 (C14/18/092); CELSA - Active Learning (CELSA/21/019) Flemish Government: FWO SBO S005024N and S003422N, Elixir Belgium I002819N. This research received funding from the Flemish Government (AI Research Program). Yves Moreau is affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

Data availability

The original CATTLE data used in [25] is available from https://zenodo.org/records/6602439#.YpofCHVBxhE (the kinship matrices) and from their repo (https://github.com/filippob/paper_deep_learning_vs_gblup/tree/main/data (the SNPs and the simulated phenotypes).
The IBDSNP is an in-house dataset for which ethical approval has been already obtained (secondary use).
The IBDWES dataset is available from dbGaP (Inflammatory Bowel Disease Exome Sequencing Study, dbGaP Study Accession: phs001076.v1.p1). Access to the data can be requested through dbGaP.

Code availability

The code and the publicly available data used in this article are available at https://bitbucket.org/eddiewrc/gmkl/src/main/ and from Zenodo (DOI: 10.5281/zenodo.15007226).

Declarations

Ethics approval and consent to participate

Concerning the IBDSNP dataset, IBD cases and non-IBD controls from the in-house dataset were collected as part of the IBD genetics study (CCare), initiated by the IBD unit at University Hospitals Leuven. All participants provided written informed consent, and the study received ethical approval from the Ethics Committee Research UZ/KU Leuven (S53684). Samples and data are stored in a coded, anonymized biobank and database. The IBDWES dataset is available from dbGaP (Inflammatory Bowel Disease Exome Sequencing Study, dbGaP Study Accession: phs001076.v1.p1). Access to the data can be requested through dbGaP.

Consent for publication

All authors gave their consent to the publication.

Competing interests

The authors declare no competing interests.

Received: 23 February 2024 Accepted: 17 March 2025
Published online: 04 April 2025

References

1. Raimondi D, Corso M, Fariselli P, Moreau Y. From genotype to phenotype in Arabidopsis thaliana: in-silico genome interpretation predicts 288 phenotypes from sequencing data. Nucleic Acids Res. 2022;50(3):e16–e16.
2. Raimondi D, Orlando G, Verplaetse N, Fariselli P, Moreau Y. Towards genome interpretation: Computational methods to model the genotype-phenotype relationship. Front Bioinforma. 2022;2:1098941.
3. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. Genome Biol. 2024;25(1):53.
4. Andreoletti G, Pal LR, Moult J, Brenner SE. Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. Hum Mutat. 2019;40(9):1197–201.
5. Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat. 2017;38(9):1182–92.
6. Raimondi D, Tanyalcin I, Ferté J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201–6.
7. Ubbens J, Parkin I, Eynck C, Stavness I, Sharpe AG. Deep neural networks for genomic prediction do not estimate marker effects. Plant Genome. 2021;14(3):e20147.
8. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91(11):4414–23.
9. Ma W, Qiu Z, Song J, Li J, Cheng Q, Zhai J, et al. A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta. 2018;248:1307–18.
10. Zhu D, Zhao Y, Zhang R, Wu H, Cai G, Wu Z, et al. Genomic prediction based on selective linkage disequilibrium pruning of low-coverage whole-genome sequence variants in a pure Duroc population. Genet Sel Evol. 2023;55(1):72.
11. Heffner EL, Lorenz AJ, Jannink JL, Sorrells ME. Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. 2010;50(5):1681–90.
12. Meuwissen TH, Hayes BJ, Goddard M. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819–29.
13. Wray NR, Kemper KE, Hayes BJ, Goddard ME, Visscher PM. Complex trait prediction from genome data: contrasting EBV in livestock to PRS in humans: genomic prediction. Genetics. 2019;211(4):1131–41.
14. Crossa J, Campos Gdl, Pérez P, Gianola D, Burgueno J, Araus JL, et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186(2):713–24.
15. Lee SH, Van Der Werf JH, Hayes BJ, Goddard ME, Visscher PM. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. PLoS Genet. 2008;4(10):e1000231.
16. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE. 2008;3(10):e3395.
17. De Los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet. 2010;11(12):880–6.

18. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, et al. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22(11):961–75.
19. Zhao W, Lai X, Liu D, Zhang Z, Ma P, Wang Q, et al. Applications of support vector machine in genomic prediction in pig and maize populations. Front Genet. 2020;11:598318.
20. Ubbens J, Stavness I. Sharpe AG. GPFN: Prior-Data Fitted Networks for Genomic Prediction. bioRxiv; 2023. p. 2023–09.
21. van den Berg S, Calus MP, Meuwissen TH, Wientjes YC. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. BMC Genet. 2015;16:1–12.
22. Bellot P, de Los Campos G, Pérez-Enciso M. Can deep learning improve genomic prediction of complex human traits? Genetics. 2018;210(3):809–19.
23. Khaki S, Wang L, Archontoulis SV. A cnn-rnn framework for crop yield prediction. Front Plant Sci. 2020;10:1750.
24. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H. DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant. 2023;16(1):279–93.
25. Nazzicari N, Biscarini F. Stacked kinship CNN vs. GBLUP for genomic predictions of additive and complex continuous phenotypes. Sci Rep. 2022;12(1):19889.
26. Sandhu K, Patil SS, Pumphrey M, Carter A. Multitrait machine-and deep-learning models for genomic selection using spectral information in a wheat breeding program. Plant Genome. 2021;14(3):e20119.
27. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JW, Fajardo-Flores SB, et al. A review of deep learning applications for genomic selection. BMC Genomics. 2021;22:1–23.
28. Gianola D, Van Kaam JB. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008;178(4):2289–303.
29. Pérez P, de Los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 2014;198(2):483–95.
30. de los Campos G, Gianola D, Rosa GJ. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci. 2009;87(6):1883–7.
31. Montesinos-López OA, Martín-Vallejo J, Crossa J, Gianola D, Hernández-Suárez CM, Montesinos-López A, et al. A benchmarking between deep learning, support vector machine and Bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding. G3: Genes Genomes Genet. 2019;9(2):601–18.
32. Yadav S, Ross EM, Wei X, Liu S, Nguyen LT, Powell O, Hickey LT, Deomano E, Atkin F, Voss-Fels KP, Hayes BJ. Use of continuous genotypes for genomic prediction in sugarcane. Plant Genome. 2024 Mar;17(1):e20417. https://doi.org/10.1002/tpg2.20417. Epub 2023 Dec 8. PMID: 38066702.
33. Henderson CR. Estimation of changes in herd environment. J Dairy Sci. 1949;32(8):706.
34. Høj-Edwards SM, Sørensen P. Linear Mixed Models used in Quantitative Genomics. 2022. https://psoerensen.github.io/qgnotes/LMM.pdf. Accessed Jan 2024.
35. Habier D, Fernando RL, Garrick DJ. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics. 2013;194(3):597–607.
36. Høj-Edwards SM, Sørensen P. Best Linear Unbiased Prediction used in Quantitative Genomics. 2022. https://psoerensen.github.io/qgnotes/BLUP.pdf. Accessed Jan 2024.
37. Biscarini F, Nazzicari N, Bink M, Arús P, Aranzana MJ, Verde I, et al. Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies. BMC Genomics. 2017;18:1–15.
38. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12(1):1–11.
39. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–90.
40. Wald NJ, Old R. The illusion of polygenic disease risk prediction. Genet Med. 2019;21(8):1705–7.
41. Francisco M, Bustamante CD. Polygenic risk scores: a biased prediction? Genome Med. 2018;10(1):1–3.
42. Calus MP. Genomic breeding value prediction: methods and procedures. Animal. 2010;4(2):157–64.
43. Nishio M, Satoh M. Including dominance effects in the genomic BLUP method for genomic evaluation. PLoS ONE. 2014;9(1):e85792.
44. Covarrubias-Pazaran G. Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE. 2016;11(6):e0156744.
45. Zhao W, Qadri QR, Zhang Z, Wang Z, Pan Y, Wang Q, et al. PyAGH: a python package to fast construct kinship matrices based on different levels of omic data. BMC Bioinformatics. 2023;24(1):153.
46. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
47. Grinberg NF, Orhobor OI, King RD. An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. Mach Learn. 2020;109:251–77.
48. Qiu S, Lane T. A framework for multiple kernel support vector regression and its applications to siRNA efficacy prediction. IEEE/ACM Trans Comput Biol Bioinforma. 2008;6(2):190–9.
49. Corinna C, Mehryar M, Afshin R. Two-stage learning kernel algorithms. In Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10). Omnipress, Madison, WI, USA. 2010:239–46.
50. Verplaetse N, Passemiers A, Arany A, Moreau Y, Raimondi D. Large sample size and nonlinear sparse models outline epistatic effects in inflammatory bowel disease. Genome Biol. 2023;24(1):224.
51. Mackay TF. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014;15(1):22–33.
52. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci. 2012;109(4):1193–8.
53. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119–24.
54. Chang JT. Pathophysiology of inflammatory bowel diseases. N Engl J Med. 2020;383(27):2652–64.
55. Alatab S, Sepanlou SG, Ikuta K, Vahedi H, Bisignano C, Safiri S, et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol. 2020;5(1):17–30.

56. Liu Z, Liu R, Gao H, et al. Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. Nat Genet. 2023;55:796–806. https://doi.org/10.1038/s41588-023-01384-0.
57. De Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet. 2017;49(2):256–61.
58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.
59. Kelly CM, McLaughlin RL. Comparison of machine learning methods for genomic prediction of selected Arabidopsis thaliana traits. PLoS ONE. 2024;19(8):e0308962.
60. Carlborg Ö, Haley CS. Epistasis: too often neglected in complex trait studies? Nat Rev Genet. 2004;5(8):618–25.
61. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nat Rev Genet. 2013;14(1):1–2.
62. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature. 2020;578(7796):527–39.
63. Mortlock S, Lord A, Montgomery G, Zakrzewski M, Simms LA, Krishnaprasad K, et al. An extremes of phenotype approach confirms significant genetic heterogeneity in patients with ulcerative colitis. J Crohn's Colitis. 2023;17(2):277–88.
64. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–45. PMID: 3203132.
65. Zhang J, Wei Z, Cardinale CJ, Gusareva ES, Van Steen K, Sleiman P, et al. Multiple epistasis interactions within MHC are associated with ulcerative colitis. Front Genet. 2019;10:257.
66. Lin Z, Wang Z, Hegarty JP, Lin TR, Wang Y, Deiling S, et al. Genetic association and epistatic interaction of the interleukin-10 signaling pathway in pediatric inflammatory bowel disease. World J Gastroenterol. 2017;23(27):4897.
67. Török HP, Glas J, Endres I, Tonenchi L, Teshome MY, Wetzke M, et al. Epistasis between Toll-like receptor-9 polymorphisms and variants in NOD2 and IL23R modulates susceptibility to Crohn's disease. Off J Am Coll Gastroenterol. 2009;104(7):1723–33.
68. van Heel DA, Dechairo BM, Dawson G, McGovern DP, Negoro K, Carey AH, et al. The IBD6 Crohn's disease locus demonstrates complex interactions with CARD15 and IBD5 disease-associated variants. Hum Mol Genet. 2003;12(20):2569–75.
69. Vermeire S, Rutgeerts P, Van Steen K, Joossens S, Claessens G, Pierik M, et al. Genome wide scan in a Flemish inflammatory bowel disease population: support for the IBD4 locus, population heterogeneity, and epistasis. Gut. 2004;53(7):980–6.
70. Jiang BB, Pütz. SimPhe: Tools to Simulate Phenotype(s) with Epistatic Interaction. GitHub; 2018. https://github.com/beibeiJ/SimPhe. Accessed Jan 2024.
71. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther. 2011;13:1–3.
72. DbGaP. Inflammatory Bowel Disease Exome Sequencing Study. 2017. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001076.v1.p1. Accessed Jan 2024.
73. Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: International conference on computational learning theory. Springer; 2001. pp. 416–426.
74. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning, vol 4. Springer; 2006.
75. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. J Mach Learn Res. 2002;2(Feb):419–44.
76. Scornet E. Random forests and kernel methods. IEEE Trans Inf Theory. 2016;62(3):1485–500.
77. Eskin E, Weston J, Noble W, Leslie C. Mismatch string kernels for SVM protein classification. Becker S, Thrun S, Obermayer K. Adv Neural Inf Process Syst. MIT Press; 2002;15. https://proceedings.neurips.cc/paper_files/paper/2002/file/12b1e42dc0746f22cf361267de07073f-Paper.pdf.
78. Leslie C, Eskin E, Noble WS. The spectrum kernel: a string kernel for SVM protein classification. In: Biocomputing 2002. World Scientific; 2001. pp. 564–575.
79. Gönen M, Alpaydın E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.
80. De Bie T, Tranchevent LC, Van Oeffelen LM, Moreau Y. Kernel-based data fusion for gene prioritization. Bioinformatics. 2007;23(13):i125–32.
81. Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semidefinite programming. J Mach Learn Res. 2004;5(Jan):27–72.
82. Cortes C. Can learning kernels help performance. In: Invited talk at International Conference on Machine Learning (ICML 2009). Montréal; 2009.
83. Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. Adv Neural Inf Process Syst. Dietterich T, Becker S, Ghahramani Z. MIT Press; 2001;14. https://proceedings.neurips.cc/paper_files/paper/2001/file/1f71e393b3809197ed66df836fe833e5-Paper.pdf.
84. Adam P, Sam G, Soumith C, Gregory C, Edward Y, Zachary D, Zeming L, Alban D, Luca A, Adam L. Automatic differentiation in pytorch. In NIPS Workshop. 2017.

Publisher's Note