




RESEARCH ARTICLE

Learning high-order interactions for polygenic risk prediction

Michela C. Massi ^{1,2}*, Nicola R. Franco ¹, Andrea Manzoni ¹, Anna Maria Paganoni ¹, Hanla A. Park ^{3,4}, Michael Hoffmeister ⁵, Hermann Brenner ^{5,6,7}, Jenny Chang-Claude ^{3,8}, Francesca Ieva ^{1,2}, Paolo Zunino ¹

1 MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy, **2** HDSC - Health Data Science Centre, Human Technopole, Milan, Italy, **3** Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, **4** Medical Faculty, University of Heidelberg, Heidelberg, Germany, **5** Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, **6** Division of Preventive Oncology, National Center for Tumor Diseases (NCT) and German Cancer Research Center (DKFZ), Heidelberg, Germany, **7** German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany, **8** Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

* These authors contributed equally to this work.

* michela.massi@fht.org



OPEN ACCESS

Citation: Massi MC, Franco NR, Manzoni A, Paganoni AM, Park HA, Hoffmeister M, et al. (2023) Learning high-order interactions for polygenic risk prediction. *PLoS ONE* 18(2): e0281618. <https://doi.org/10.1371/journal.pone.0281618>

Editor: Shuai Ren, Affiliated Hospital of Nanjing University of Chinese Medicine: Jiangsu Province Academy of Traditional Chinese Medicine, CHINA

Received: July 29, 2022

Accepted: January 27, 2023

Published: February 10, 2023

Copyright: © 2023 Massi et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The DACHS data that support the findings of this study are available on request. The data are not publicly available due to privacy or ethical restrictions. For further information about the DACHS study, such as collaborations and contact details, please refer to the official website: dachs.dkfz.org/dachs/. The code developed to reproduce the experiments carried out in this work, and to apply the proposed hiPRS to any other data, is freely distributed on a

Abstract

Within the framework of precision medicine, the stratification of individual genetic susceptibility based on inherited DNA variation has paramount relevance. However, one of the most relevant pitfalls of traditional Polygenic Risk Scores (PRS) approaches is their inability to model complex high-order non-linear SNP-SNP interactions and their effect on the phenotype (e.g. epistasis). Indeed, they incur in a computational challenge as the number of possible interactions grows exponentially with the number of SNPs considered, affecting the statistical reliability of the model parameters as well. In this work, we address this issue by proposing a novel PRS approach, called High-order Interactions-aware Polygenic Risk Score (hiPRS), that incorporates high-order interactions in modeling polygenic risk. The latter combines an interaction search routine based on frequent itemsets mining and a novel interaction selection algorithm based on Mutual Information, to construct a simple and interpretable weighted model of user-specified dimensionality that can predict a given binary phenotype. Compared to traditional PRSs methods, hiPRS does not rely on GWAS summary statistics nor any external information. Moreover, hiPRS differs from Machine Learning-based approaches that can include complex interactions in that it provides a readable and interpretable model and it is able to control overfitting, even on small samples. In the present work we demonstrate through a comprehensive simulation study the superior performance of hiPRS w.r.t. state of the art methods, both in terms of scoring performance and interpretability of the resulting model. We also test hiPRS against small sample size, class imbalance and the presence of noise, showcasing its robustness to extreme experimental settings. Finally, we apply hiPRS to a case study on real data from DACHS cohort, defining an interaction-aware scoring model to predict mortality of stage II-III Colon-Rectal Cancer patients treated with oxaliplatin.

GitHub repository that can be accessed through this link: github.com/NicolaRFranco/hiPRS.

Funding: NRF has received funding under the ERA-NET ERA PerMed / FRRB grant agreement No ERAPERMED2018-244, RADprecise - Personalized radiotherapy: incorporating cellular response to irradiation in personalized treatment planning to minimize radiation toxicity. DACHS study was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, BR 1704/6-6, CH 117/1-1, HO 5117/2-1, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1, BR 1704/17-1), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A, 01ER1505B, 01GL1712). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Within the framework of precision medicine, it is becoming more and more important to stratify individual genetic susceptibility based on inherited DNA variation, an approach that progressed together with the latest advances in human genetics [1]. A substantial interest in leveraging the massive amount of genome-wide data now available has emerged, with the purpose of maximizing the predictive power of risk prediction models by incorporating the effect of Single Nucleotide Polymorphisms (SNPs) on the outcome [2]. Accurate genomic risk prediction has two great potential aims: to prospectively identify individuals at increased risk of disease, thus informing early interventions, and to aid diagnosis for diseases where current diagnostic approaches are imperfect [3].

One of the most traditional approaches to model genetic risk is the Polygenic Risk Score (PRS). PRS exploit a fixed model approach to sum the contribution of a set of risk alleles to a specific complex disease [4]. Calculating PRS is a common practice because of its simplicity, its computational efficiency, and the straightforward interpretability of the model itself. Indeed, while polygenic scores are used to predict phenotypes, there are other interests beyond forecasting. For instance, model interpretability is often key for research purposes such as the discovery or validation of SNPs' role in disease risk. The schema in Fig 1 gives an overview on the strengths and weaknesses of the most common PRS approaches in the literature.

Standard weighted PRS estimation relies on Genome-Wide Association Study (GWAS) summary statistics obtained on one or more discovery cohorts modeling the independent

	GWAS-BASED PRSs	PENALIZED PRSs	MACHINE LEARNING
STRENGTHS	<ul style="list-style-type: none"> ✓ Easy to model ✓ Interpretable results 	<ul style="list-style-type: none"> ✓ Easy to model ✓ Interpretable results ✓ Generally, no need for external information 	<ul style="list-style-type: none"> ✓ Effective for multi-dimensional, complex data ✓ Naturally account for high-order interactions ✓ Assumptions-free non-parametric methods ✓ Generally, no need for external information Optimized for prediction performance
WEAKNESSES	<ul style="list-style-type: none"> ↑ Additive and independent predictor effects ↑ Normal distribution of underlying data ↑ No complex interactions ↑ Parameters' overestimation ↑ GWAS limitations 	<ul style="list-style-type: none"> ↑ Additive and independent predictor effects ↑ Normal distribution of underlying data ↑ No complex interactions (2nd order max) ↑ Parameters' overestimation 	<ul style="list-style-type: none"> ↑ Difficult to apply (algorithm and architecture design, hyperparameter tuning, etc.) ↑ Difficult to interpret ↑ Overfit on small samples Computationally expensive

✓ Strength enjoyed by *hiPRS* ↑ Weakness solved by *hiPRS*

Fig 1. Alternative approaches to polygenic risk scoring vs *hiPRS*. Strengths (first row) and Weaknesses (second row) of three main categories of PRS methods discussed in the Introduction. The green tick signals that the given point of strength applies to *hiPRS* as well. The blue arrow signals a point of weaknesses that *hiPRS* algorithm does not suffer, or some aspect that the algorithm was specifically designed to solve.

<https://doi.org/10.1371/journal.pone.0281618.g001>

effect of individual SNPs on the outcome. These PRSs (cf. Fig 1, first column, GWAS-based PRSs) exploit SNP-specific odds ratios or effect sizes to weight the contribution of the risk alleles on the disease risk or outcome [5, 6]. The set of SNPs to be included in this estimation may of course affect the score's predictive power significantly. Some approaches include all SNPs, with the risk of incorporating useless or redundant information, while others retain a subset of SNPs based on predefined criteria (e.g., those passing an arbitrary p-value threshold in the GWAS results [5]).

Despite their wide adoption and appreciated simplicity and interpretability, the performance and reliability of traditional PRSs have been largely discussed and several methodological concerns have been raised (see, e.g., [7] and references therein) as the approach presents some evident limitations.

In particular, (i) the GWAS studies exploited for weights estimation are oftentimes underpowered by multiple independent testing and their effect sizes overestimated because of winners' curse and biases [8–11]. Additionally, (ii) GWAS effect weights are traditionally computed considering the effect of single SNPs on the phenotype. By doing so, they do not account for the complex and nonlinear interactions between alleles in the genotype and their role in determining the phenotype, or the epistasis effect [4, 12–15]. Essentially, *epistasis* refers to departure from independence (or additivity, from a statistical point of view) of the effects of multiple loci in the way that they combine to cause disease [16]. In other words, interaction effects exist between loci, and their presence was reported to make major contributions to phenotypes [12, 13, 15, 17–20]. However, including SNP-SNP interactions in GWAS and risk scoring models is computationally challenging due to the high dimensions involved: in fact, the number of possible interactions grows exponentially with the number of SNPs considered. Additionally, this also increases dramatically the amount of independent tests to be carried out, thus strongly affecting their reliability. The authors in [21] attempted the inclusion of SNP-SNP interactions by filtering those that were relevant accordingly to some imposed biological criteria, but only managed to consider up to second-order interactions. Other methodological concerns of GWAS-based approaches are the fact that (iii) GWAS weights ignore the mediating role of clinical covariates when estimating SNPs effect on complex diseases [7], and that (iv) those models incorporate strict assumptions, e.g., they include additive and independent predictor effects, and assume that observations are uncorrelated [4, 14, 22]. These assumptions do not necessarily hold true when modeling complex polygenic diseases. For instance, Linkage Disequilibrium, i.e., the non-random association of alleles at two or more loci in a population [4], statistically translates into strong correlation between predictors. To account for LD some methods were developed to optimize SNPs *reweighting* (LD pruning and p-value thresholding, LDpred [23], and others).

Other well-recognized solutions to polygenic risk scoring discard GWAS weights and directly take genotype data as input, including various forms of penalization to restrict the pool of predictors (cf. Fig 1, Penalized PRSs). These include genomic BLUP [24], shrinkage methods (e.g. LASSO) or Generalized Linear Mixed Models (GLMMs). However, all these methods share with the most traditional PRSs the assumption of additive effect of single SNPs on the outcome, and may incur in the curse of dimensionality when trying to include all potential interaction terms, with the risk of overestimating effect sizes and obtaining unreliable models. This is particularly true if modeling high-order interactions. Nevertheless, these complex interactions were found useful in describing genotype-phenotype relationships for complex traits and common diseases both in humans and model organisms [25–29], highlighting the need for novel approaches able to account for high-order interactions.

In this respect, great attention has been recently devoted to polygenic risk prediction via Machine Learning (ML) algorithms (Fig 1, Machine Learning). These algorithms employ

multivariate, non-parametric methods that robustly recognize patterns from non-normally distributed and strongly correlated data [15, 20, 22]. Moreover, these methods are naturally capable of modeling highly interactive complex data structures, making them powerful tools for complex disease prediction [15, 20, 22]. However, most ML algorithms, s.a. Random Forests (RFs), Support Vector Machines (SVMs) or Neural Networks (NNs), demonstrate great predictive power but lack in model interpretability, leaving researchers with no information on the role and structure of the complex interactions influencing the phenotype prediction. Moreover, these algorithms require a lot of training samples to avoid overfitting, but the available cohorts in real world research settings are oftentimes quite small.

In this work, we aim at addressing polygenic risk prediction proposing a novel PRS approach, called High-order Interactions-aware Polygenic Risk Score (*hiPRS*), whose most remarkable feature is the capability of robustly and reliably incorporating high-order interactions in modeling polygenic risk, while constructing a simple and interpretable model.

In brief (cf. Fig 2), *hiPRS* treats genotype-level data, and starts with a user pre-defined list of SNPs of interest. The algorithm exploits Frequent Itemset Mining (FIM) routines to build a list of candidate interactions of any desired order. This is achieved by scanning the observations in the positive class only, and by retaining those terms that have an empirical frequency above a given threshold δ . These candidates are subsequently ranked according to their relationship to the outcome in terms of Mutual Information (MI). From there, we select a

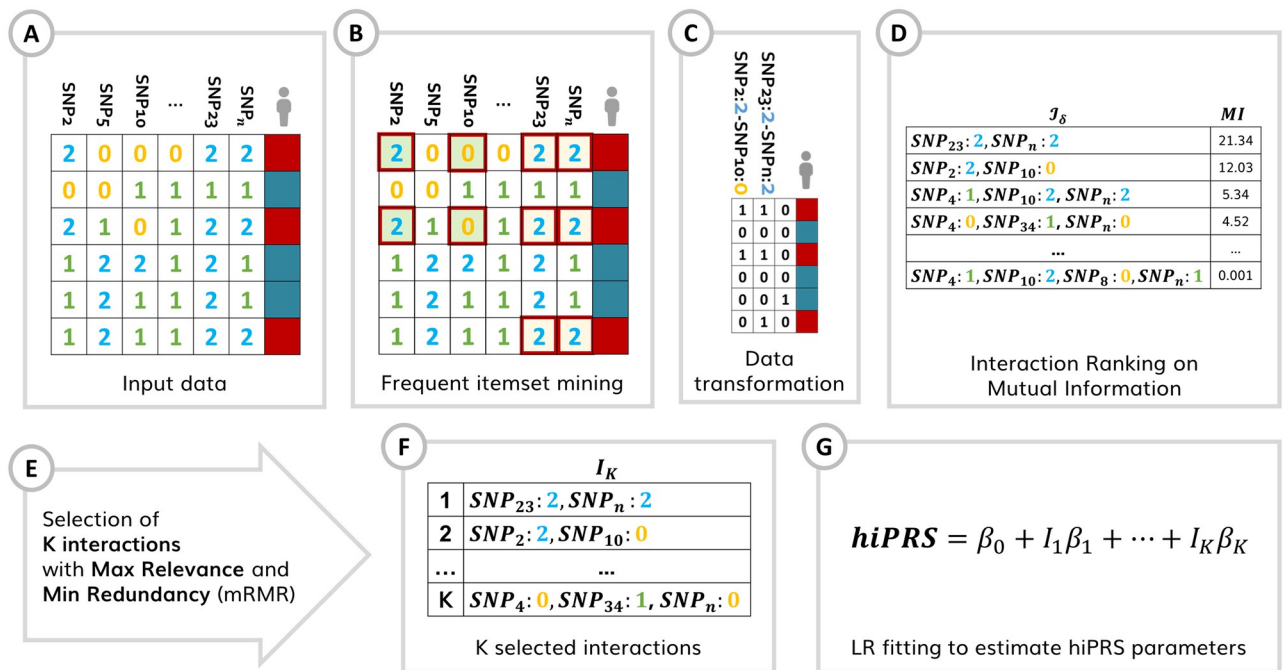


Fig 2. *hiPRS* algorithm process flow. (A) Input data is a list of genotype-level SNPs. (B) Focusing on the positive class only, the algorithm exploits FIM (*a priori* algorithm) to build a list of candidate interactions of any desired order, retaining those that have an empirical frequency above a given threshold δ . This leads to a filtered set of terms in the form of sequences of pairs of SNP and associated categorical level (i.e., allele frequency in this example). The sequences can include from a single SNP-allele pair up to a maximum number of pairs defined by the user (I_{max}). (C) The whole training data is then scanned, searching for these sequences and deriving a re-encoded dataset where interaction terms are binary features (i.e., 1 if sequence i is observed in j -th patient genotype, 0 otherwise). From this dataset we can compute the MI between each interaction and the outcome and (D) obtain a ranked list (I_δ) based on this metric. (E) Starting from the interaction at the top of I_δ , *hiPRS* constructs I_K , selecting K (where K is user-specified) terms through the greedy optimization of the ratio between MI (*relevance*) and a suitable measure of similarity for interactions (*redundancy*) (cf. Algorithm 1, [Materials and methods](#)). This leads to a set of predictive, yet diverse, interactions that (F) we use to define the score weighting their contribution by fitting a LR model and retaining the corresponding β coefficients.

<https://doi.org/10.1371/journal.pone.0281618.g002>

restricted pool of K interactions to include in the final PRS, where K is user-specified. To this end, *hiPRS* incorporates a novel interaction selection algorithm inspired by the minimum Redundancy Maximum Relevance (mRMR) literature. More precisely, the algorithm selects K terms through the greedy optimization of the ratio between MI (*relevance*) and a suitable measure of similarity for interactions (*redundancy*). This leads to a set of predictive, yet diverse, interactions that we use to define the score. In the end, the latter is built by weighting the contribution of each interaction term accordingly to the weights obtained when fitting a Logistic Regression (LR) model.

In general, to keep the interaction search computationally feasible, the algorithm has to start from a limited set of SNPs of interest. If a large number of SNPs is available, a preliminary feature selection may be desirable. The latter can be based on GWAS—with all the aforementioned limitations—, on biological rationale (s.a. [30]) or on more sophisticated multivariate methods accounting for interactions [31, 32]. This sort of prior selection is typical of validation studies [29, 33], and it is grounded on the idea that part of the missing variability can be later recovered when interaction effects are included, possibly bringing to novel discoveries in genotype-phenotype relationships.

Similarly to traditional methods, *hiPRS* is based on a simple and interpretable weighted model. However, as highlighted in the summarizing scheme in Fig 1, our proposal overcomes the limitations of both traditional PRSs and modern Machine Learning approaches. In particular,

1. Compared to PRSs methods, *hiPRS* can include high-order interactions in the model without affecting weights estimation. Furthermore, its model fitting is data-driven, not relying neither on GWAS summary statistics nor on any other type of external information.
2. *hiPRS* differs from most ML algorithms in that it provides an easily accessible, readable and interpretable model not meant for risk prediction only, but for discovery and validation of genetic susceptibility as well. Additionally, it is able to control overfitting, even on small samples.
3. On top of that, *hiPRS* is designed to be robust to class imbalance, which is oftentimes an issue when modeling rare complex diseases. Finally, we mention that, with very little effort, *hiPRS* can be generalized to include clinical covariates in the search for interactions, thus recovering their mediating role in determining SNPs effect on the outcome.

In the present work we demonstrate through a comprehensive simulation study the superior performance of *hiPRS* w.r.t. state of the art methods belonging either to Penalized PRSs or ML literature (cf. Fig 1), both in terms of scoring performance and interpretability of the resulting model. Moreover, we test *hiPRS* against complexities not seldom arising in real life research scenarios, such as small sample size, class imbalance and the presence of noise, showcasing its robustness to extreme experimental settings. Finally, we apply *hiPRS* to an interesting case study on real data from the DACHS cohort [34], defining an interaction-aware scoring model to predict mortality of stage II-III Colon-Rectal Cancer (CRC) patients treated with oxaliplatin.

Results

Simulation study

This section describes the results of a large set of experiments run on simulated data whose generative mechanism was designed to present complex non-linear dependencies between a binary outcome Y and high-order SNP-SNP interactions. In particular, the positive class in the

generated data, namely $\{Y = 1\}$, was defined in terms of the rules reported in Fig 3, which are ultimately high-order interactions. Additionally, each of the observed outcome values was mislabeled with probability ε (hereby called *random noise*), in an attempt to make the relationship between the outcome and the predictors less deterministic. More details on the generative algorithm can be found in Materials and Methods (cf. Simulated Data with Non-Linear Interaction Effect).

***hi*PRS outperforms benchmark algorithms in prediction performance.** To judge the performance of *hi*PRS, we run a first experiment on simulated data and we compared the results with those of a comprehensive set of benchmark algorithms coming both from the traditional and more recent literature on polygenic risk prediction. Note that, to ensure a fair comparison, we only included methods that did not rely on any type of external information (e.g. summary statistics) besides individual-level genotype. More precisely, we picked methods from the following two classes: PRSs taking raw SNP values as input (Penalized PRSs) and algorithms from the ML literature. In the first group we included Lasso [35], Ridge [36], Elastic-Net [37] and glinternet [38]. The first three are penalized LR-based PRS methods with additive main effects and no interactions, while the last one is a penalized LR with group-Lasso regularization that includes second-order interactions. Conversely, as ML methods, we included two scoring algorithms accounting for high-order interactions: the approach proposed by Behravan et al. [39], that exploits XGboost for interaction selection and SVM for classification, and the one by Badre et al. [40], that relies on a Deep Neural Network (DNN) model. Specific technical details on each of these approaches, together with their implementation and chosen hyperparameters, can be found in the Materials and Methods section (cf. Benchmark Algorithms). All of this was considered within an experimental setting of mild complexity (i.e. reasonable sample size and low noise), consisting of 30 independent simulations. Indeed, for each of these we generated 1000 observations for training and 500 observations for testing. The mislabeling *noise* probability was set to $\varepsilon = 0.01$.

The glinternet algorithm allows the user to define the number of interaction terms to include in the model, which we set to 3 and 8. Note that this method includes main effects of all considered interactions and estimates a parameter for each categorical level and their products. Therefore, 8 interaction terms actually correspond to more than $8 \times 3^2 = 72$ regressors in the LR model (in particular, in our experiments we obtained 93 of them). In light of this, we chose 8 as maximum number of interactions, as that reflected the amount of generating rules in the simulated data (cf. Fig 3), while the value of 3 is meant to test the performance of a simpler model. Differently from glinternet, in *hi*PRS the number of interaction terms K actually corresponds to the final dimension of the model. Here, we set $K = 10, 40$, resulting in two models of different complexity. In principle, both of them have enough degrees of freedom to discover the generative model, nevertheless, they both allow for handy readability and interpretability of the fitted model. We did not impose any limit on the order of the interactions, whereas we set the frequency threshold δ to 0.05.

In Fig 4 we report the boxplots of the performance metrics in the 30 independent trials. The three Penalized PRSs behave similarly, with an average AUC around 0.6 and an Average Precision (AP) slightly above 0.3. Their performance reflects the maximum predictive power achievable by additive PRSs in the presence of epistasis. SVM-Behravan has the worst performance among the classifiers on both metrics, probably due to the fact that despite the SNP selection step accounts for interactions, it is followed by a linear SVM classifier that considers additive effects only. The DNN is unsurprisingly extremely performant, but its black box nature does not allow us to identify the role of the predictors in scoring observations. Moreover, whilst still high, it has a slightly lower and more variable AP. This might be due to the tendency to overfit the small sample of positive observations during training, possibly

SNP 1, SNP 2, SNP 3	SNP 4	SNP 5	SNP 6	SNP 7	SNP 8	SNP 9	SNP 10	SNP 11 ... SNP N	
	2	2	2	2					A
	1	1	0	0					
	1	1	1	1					
	1	1	1	0					B
	1	1	1	1					
	1	1	0	1					
					0	0	0		C
					1	1	1		
					2	2	2		

Fig 3. Simulation data generating rules. Graphical representation of the three rules that determine the positive class. Cases are obtained when $A \vee B \vee C$. More details on the generative model are provided in the Materials and Methods Section.

<https://doi.org/10.1371/journal.pone.0281618.g003>

caused by the very large number of parameters typical of these models. Glinternet best AUC performance is achieved when including 8 interactions. However, the same model has a significant drop in AP, falling behind its simpler version with 36 terms: this goes to show that, despite the inclusion of more predictors, the identified interactions are not necessarily the most useful to identify the class of interest when the data is imbalanced. Moreover, the 93 parameters associated to the 8 interactions could easily overfit the underrepresented positive class. The best performer overall is *hi*PRS modeling 40 terms, showcasing the ability of our

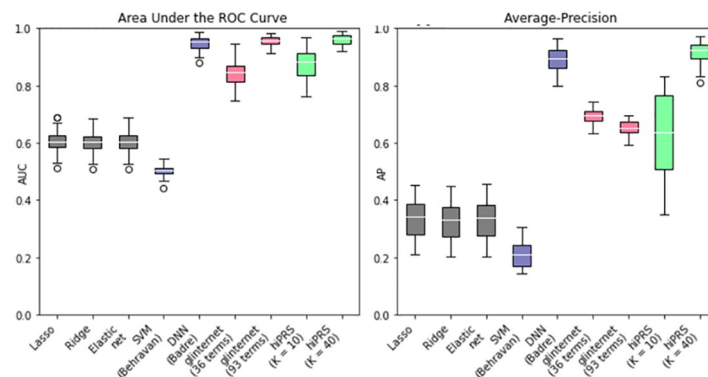


Fig 4. *hi*PRS results on risk prediction against benchmark PRSs and ML approaches. AUC (left) and AP (right) performance distributions of 30 independent trials. In grey the three traditional penalized PRSs approaches with additive effects only; in violet the two ML algorithms (SVM-Behravani and DNN-Badre); in pink glinternet algorithm for two model dimensions (3 interactions, i.e. 36 terms, and 8 interactions, i.e. 93 terms); in green *hi*PRS for $K = 10$ and $K = 40$.

<https://doi.org/10.1371/journal.pone.0281618.g004>

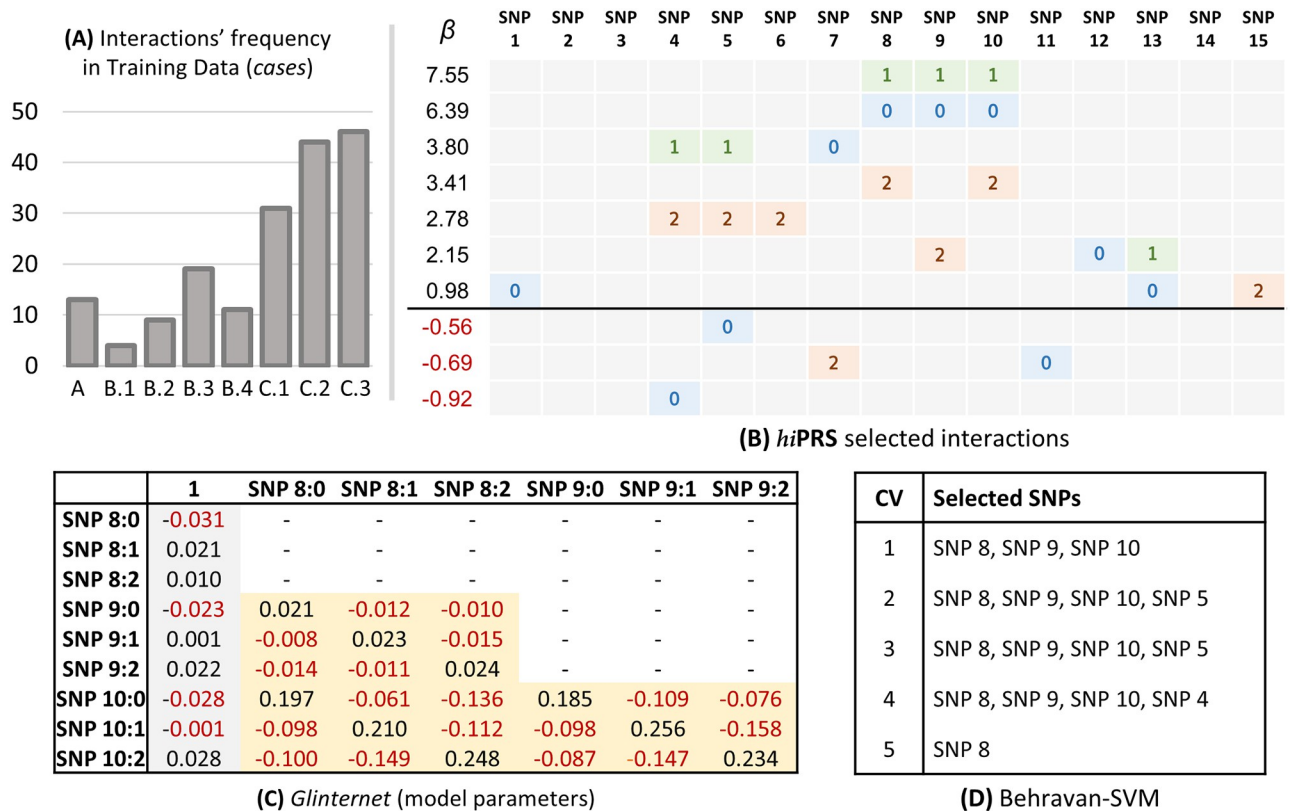


Fig 5. Interpretability analysis. (A) Absolute frequency of the generative rules in the training data, limited to the positive class. (B) Interactions selected by *hiPRS* with $K = 10$ and corresponding β coefficients. (C) Coefficients of the *glinternet* model with 3 interaction terms: main effects are in gray, interactions in yellow. (D) Lists of SNPs selected by SVM-Behravan during its five internal cross validations, cf. Benchmark Algorithms in the Materials and Methods Section. Note: reported results are limited to one simulation among the 30 randomly generated datasets.

<https://doi.org/10.1371/journal.pone.0281618.g005>

approach to identify predictive interactions that generalize well irrespectively of the under-representation of the positive class. The results obtained for $K = 10$ are also remarkable, especially if we consider that they correspond to the model having the least number of parameters.

***hiPRS* captures and explains interaction-based generative mechanisms better than benchmarks.** One of the added values of *hiPRS* is the capability of capturing dependencies among predictors, by selecting the most relevant to determine the phenotype and presenting them within a simple and interpretable model. In many research settings, a slight loss in prediction quality may be acceptable if it leads to a more meaningful interpretation of the predictors [41, 42]. To test the ability of *hiPRS* in capturing them, we focused on one of the previously mentioned datasets, and checked the 10 interactions selected by the simplest model, $K = 10$. We report the selected interactions in Fig 5. We mention that, to ensure a fair comparison, we picked the dataset where the worst performer, SVM-Behravan, achieved one of its highest AUC. This dataset contained 182 cases in the training set, i.e. 18.2% of the total. The bar chart in Fig 5 reports the absolute frequency of the generating rules available in the positive class.

hiPRS captures most of rule C, assigning very high positive betas to the first, second and fourth interaction. Combined together, the fourth and fifth selected interactions partially recover rule A, while the third one almost fully captures rule B.3. Note that, among positive samples in the training data, B.3 is the most frequent version of rule B (see Fig 5, left panel). It

is also interesting to note that *hi*PRS finds two protective terms, namely $\{\text{SNP}_4 = 0\}$ and $\{\text{SNP}_5 = 0\}$, that both nullify rules A and B.

Let us now briefly discuss the results obtained by the competitors. In panel (c) of Fig 5 we report the model parameters for glinternet when modeling 3 interactions. Here, we see that the good performance of glinternet is granted by the inclusion of three interaction terms in the model, which, despite being of second order only, are subsets of the true generating rule C. Moreover, the estimated effect sizes are correctly positive for couples of identical allele frequencies only (i.e., the diagonal of the matrices of coefficients associated to each interaction). However, due to extremely large number of parameters, it is very hard to inspect the model and drive suitable conclusions. Furthermore, main effect sizes are not truly relevant in determining the phenotype and the generative mechanism is only captured partially, as rule C is the only one that is actually identified. Within the same Figure, but in panel (d), we list the SNPs selected by SVM-Behruvan during its 5-fold cross-validation. Notably, SVM-Behruvan is able to recover some of the SNPs associated with the generating rules, however, we are left with no information on the structure of the interactions.

***hi*PRS can deal with extremely small sample size.** We tested *hi*PRS performance for small sample sizes. We believe this to be a relevant setting, as in most real research scenarios clinicians have to deal with individual-level data of significantly small cohorts. To this end, starting from the sample size that we considered in our previous experiments, we repeated our analysis on a sequence of four decreasing sample sizes, namely $n = 1000, 750, 500, 250$. That is, for each n , we run 30 independent simulations, and we registered *hi*PRS performance in terms of AUC and AP. To ensure comparable results, the noise level was fixed to $\epsilon = 0.01$ for all experiments, and the model was always evaluated on a test set of 500 instances. Once again, we tested *hi*PRS with $K = \{10, 40\}$, while $\delta = 0.05$ as before. Results are reported in Fig 6, panel A. Despite the unavoidable decrease in performance, we note that even for very small samples, 250 observations, *hi*PRS is able to provide insightful results, with AUC levels of 0.8 and ~ 0.7 respectively. AP is lower, ~ 0.5 for $K = 40$, but still significantly better than traditional penalized PRSs when trained on 1000 observations. Nonetheless, note that a training sample of 250 observations in total corresponds to less than 75 cases to learn from (cf. Simulated Data in Materials and methods), which is an extremely challenging setting. For 500 training samples (i.e. ~ 150 cases at most) *hi*PRS with $K = 40$ reaches almost 0.9 AUC and an AP of ~ 0.7 . These results testify in favour of *hi*PRS generalization potential, which is likely induced by the mRMR-based interaction selection algorithm. Indeed, the latter is optimized to select the most predictive features, while favouring the introduction of *diverse* information in the model. Minimizing redundancy in the selection makes *hi*PRS less prone to overfitting, irrespectively of the sample size or the number of cases.

***hi*PRS is robust to class imbalance.** To test *hi*PRS against extreme class imbalance, we had to modify slightly our procedure in the generation of the simulated data. This is because although our generative mechanism is based upon random generation of allele categories and random noise (ϵ), it also relies on a deterministic rule-based definition of the positive class. Therefore, the positive class will always appear in the data with an approximately fixed frequency depending on ϵ (cf. Materials and methods). We overcame this drawback via under-sampling. More precisely, let $0 \leq q \leq 1$ be any wished proportion. To obtain a training sample of 1000 instances and $1000q$ cases, we generate a larger dataset where observations are added iteratively until there are at least $1000q$ cases and $1000(1 - q)$ controls, then we discard all exceeding observations. We adopted this procedure for a varying proportion of cases, namely $q = 2.5\%, 5\%, 10\%, 15\%$ and 20% . For each of those values, we generated 30 independent training sets, and measured *hi*PRS performance over as many independent test sets of 500 instances. We mention that, while we used subsampling to generate the training data, we stuck

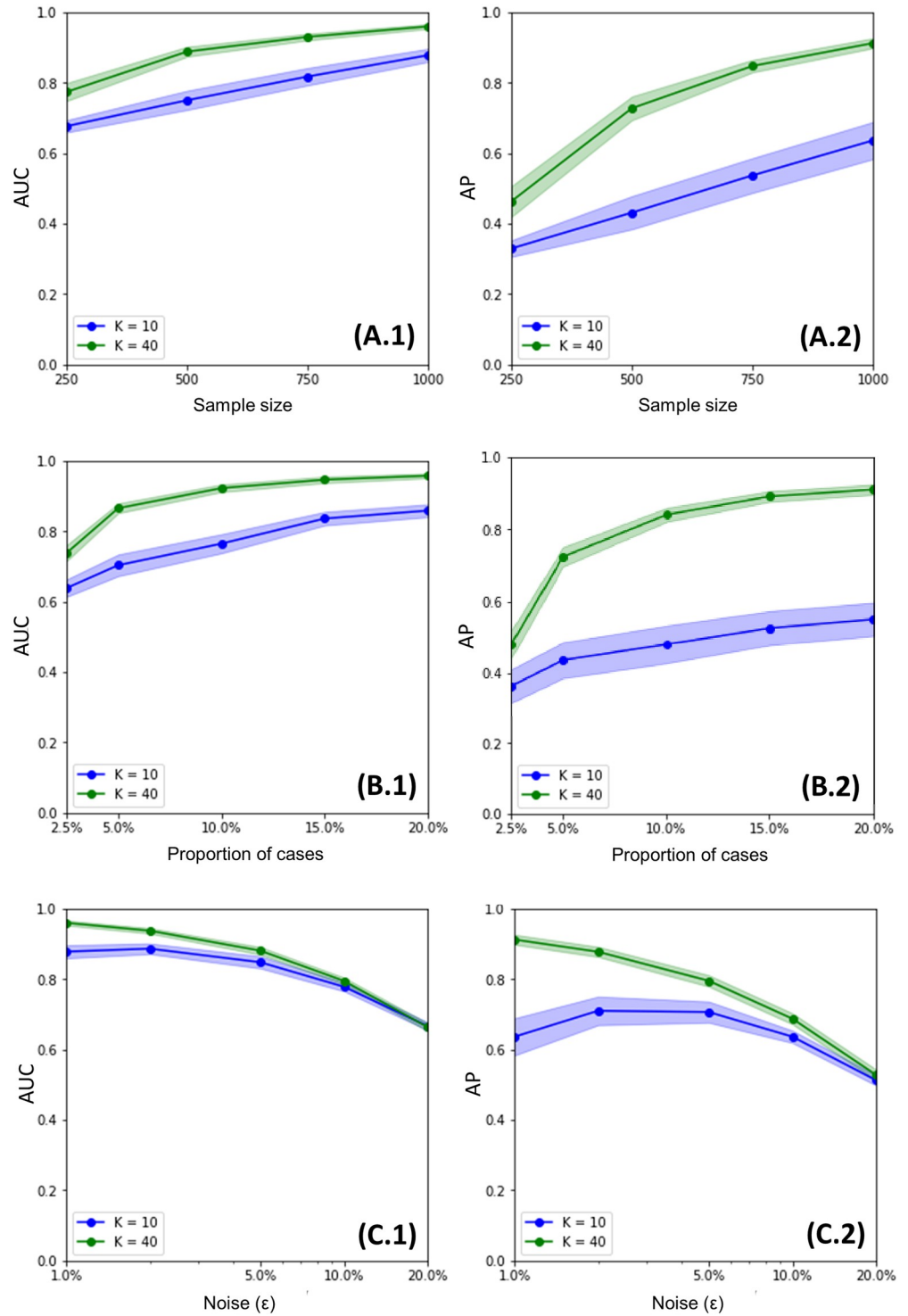


Fig 6. Sensitivity analysis results. Average performance of *hi*PRS in terms of AUC and AP for variable sample size (A.1 and A.2), class imbalance (B.1 and B.2) and missing heritability, i.e. noise (C.1 and C.2). Confidence bands are at the 95% level. The x-axis is in logarithmic scale for panels C.1 and C.2.

<https://doi.org/10.1371/journal.pone.0281618.g006>

to our usual approach for the test data. Indeed, the complexity lies in training models on imbalanced data; furthermore, the distribution of classes in the test set has no impact on the metrics that we use for evaluation, i.e. AUC and AP.

In Fig 6, panel B, we report the results for this simulation study. Notably, even for the smallest proportion of observations in the positive class (2.5%), *hi*PRS succeeds in learning a sufficiently generalizable set of interactions, with AUC values above 0.6 when $K = 10$, and above 0.75 when $K = 40$. Moreover, the larger model goes beyond 0.7 on both AUC and AP for a proportion of cases of 5%, meaning only 50 observations to learn from.

***hi*PRS is robust to variability (missing heritability).** Missing heritability is the proportion of variance in the phenotype that cannot be explained by genotype information [43]. This variability can be induced by several factors [43]: one is the need to account for SNP-SNP interactions when modeling phenotypic traits [44], but it can be nonetheless induced by factors that cannot be modeled with the data at hand. In our simulation setting, we reproduce the effect of missing heritability via the *noise* parameter, as the two can be easily linked together. The explicit way by which the missing heritability depends on ε is detailed in Materials and Methods Section, Eq (9). To test *hi*PRS robustness to variability in the training data, we generate 30 training and test set, respectively with 1000 and 500 observations, for each of the following noise levels, $\varepsilon = 0.01, 0.02, 0.05, 0.1$ and 0.2 . Results are in Fig 6, panel C. Note that, up to a noise level of 10% (i.e. ~ 100 mislabeled observations in the training set, and ~ 50 in the test set), both models maintain an average AUC near 0.8. This is considerably interesting if we consider that a noise of $\varepsilon = 0.1$ corresponds to a missing heritability of 51%, i.e. a setting where the SNPs themselves can only explain at most 49% of the target variability.

Simulated real biological model

In this section we present the results about a second simulation study based on a simpler but biologically grounded model. Indeed, while the previous experiments featured complex gene-gene interactions, the underlying generative mechanism was somewhat artificial, possibly resulting in complexities that deviate from biological epistasis.

In light of this, we repeated our analysis on a different case study concerning genetic variants and susceptibility to atrial fibrillation, a context in which epistatic effects have already been reported [45, 46]. For the sake of our analysis, we considered a simple setting in which three SNPs are used to predict the risk of atrial fibrillation. In particular, we relied on the dataset available in [46], which we exploited to generate 30 independent bootstrap samples, each counting 5000 observations for training and 1000 for testing. We refer to the Materials and Methods section for a schematic summary of the exploited dataset (Fig 9), a detailed description about the biological mechanisms and the data generation.

We fit *hi*PRS with a frequency threshold of $\delta = 2\%$, while we set $K = 13$. In order to grant additional insights, we also compare the performance of *hi*PRS with that of the benchmark algorithms considered in the previous section: a purely additive PRS (this time without penalization terms, given the small number of SNPs involved), DNN-Badre and glinternet. For DNN-Badre, we employed the same neural network architecture as before. For glinternet, we set the number of interactions to include in the model to 3, allowing all the possible combinations of the three SNPs. As for SVM-Behavran, we decided not to include the model in this comparison due to the extremely scarce performance achieved in the previous experiment.

Results are in Fig 7 and Table 1. Fig 7 shows the performance of *hi*PRS and its competitors across the 30 independent trials. In terms of AUC, the three models including interactions perform comparably well, with a small but consistent gain with respect to the purely additive PRS. DNN-Badre has the most volatile performance, which is most likely due to the large number

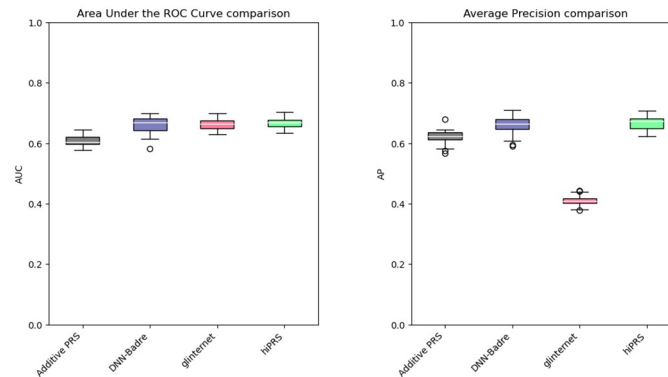


Fig 7. Results and comparisons for a real biological setting. AUC (left) and AP (right) performance distributions of 30 independent trials, sampled according to a real biological mechanism where SNPs regulate factors associated to atrial fibrillation. In grey, a traditional PRS with additive effects only; in violet the ML based algorithm, DNN-Badre; in pink glinternet (31 logistic terms); in green *hiPRS* ($K = 13$).

<https://doi.org/10.1371/journal.pone.0281618.g007>

Table 1. Fitting times in a real biological setting. Fitting times for *hiPRS* and the benchmarks algorithms, averaged across 30 independent simulations.

PRS approach	Average fitting time (s)
Additive PRS	0.02
DNN-Badre	66.59
glinetnet	0.31
<i>hiPRS</i>	0.19

<https://doi.org/10.1371/journal.pone.0281618.t001>

of parameters in the DNN architecture and their randomized initialization. Conversely, glinternet and *hiPRS* show very similar results. However, as for our previous case study, the performance of the former drops significantly when we consider the AP metric (cf. Fig 4). Overall, *hiPRS* appears to be the most valuable alternative, both in terms of prediction power and computational cost (see Table 1). Nonetheless, it should be noted that all models have a limited capability in predicting the outcome, with AUC and AP nearly always below 0.7. This, however, is not surprising: atrial fibrillation is a complex phenomenon that is not fully determined by genetic factors (missing heritability), instead it is also affected by external factors such as life-style and other comorbidities [47, 48].

Case study on real genotype data

Oxaliplatin-based chemotherapy is a standard treatment to treat colorectal cancer patients, including stage III and selected stage II patients. However, response rates to oxaliplatin vary substantially between patients and some patients suffer from serious side effects caused by the treatment. Therefore, additional consideration of genetic markers to identify patients who are more likely to benefit from oxaliplatin would be highly desirable [33]. Here, we exploited *hiPRS* to build a polygenic risk score that could predict the short-term survival of oxaliplatin treated patients. The analysis included 349 stage II-III patients who received oxaliplatin from an ongoing population-based case-control study (DACHS, colorectal cancer: chances for prevention through screening). We refer to the Material and Methods Section for a detailed description of the data. We restricted our attention to those patients that either survived (0) or died within three years (1), and we constructed the PRS on the base of nine recently validated SNPs.

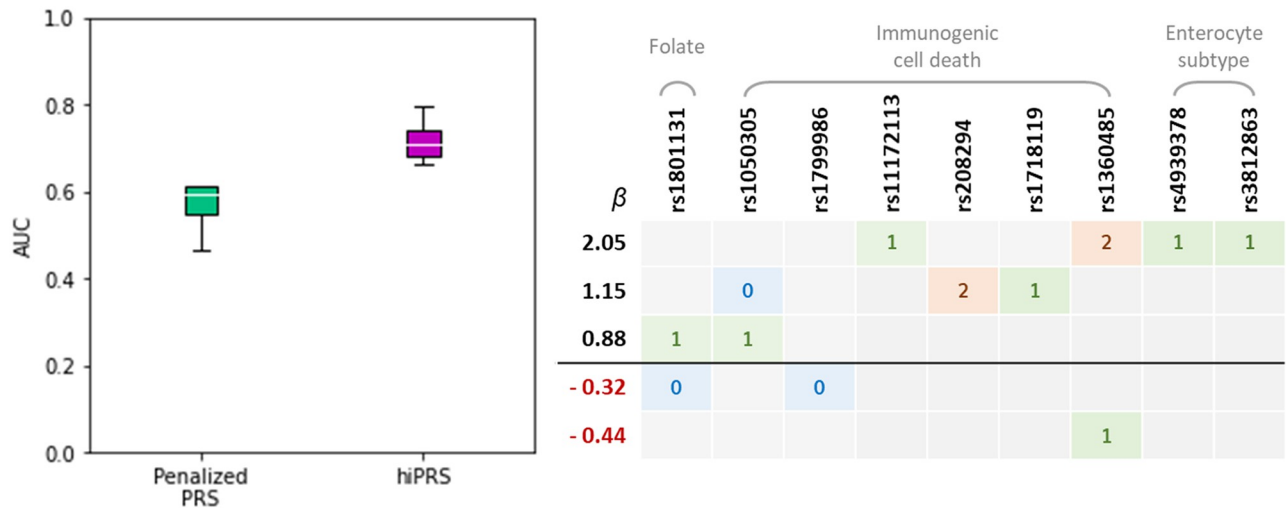


Fig 8. DACHS case study results. Left panel: AUCs obtained by *hiPRS* and a benchmark model during cross-validation (four folds). Average AUCs are 0.72 and 0.57 respectively for *hiPRS* and the benchmark model. Right panel: interactions selected by *hiPRS*, and corresponding effect-sizes, when fitting the model on the whole dataset. In grey are reported the pathways each SNP belongs to (cf. [Materials and methods](#)).

<https://doi.org/10.1371/journal.pone.0281618.g008>

To account for the small sample size, the analysis was carried out via cross-validation on 4 folds. This means that the whole procedure required to build the *hiPRS* was repeated on each fold separately, with a common choice of the hyperparameters δ and K . The values of the two were optimized via grid-search and resulted in $\delta = 15.5\%$ and $K = 5$. To highlight the improved performance derived from the inclusion of high-order interactions in the model, a Penalized PRS with Lasso regularization (i.e. modeling additively SNP's main effects only) was also implemented.

Results are shown in [Fig 8](#). The boxplots report the AUC values obtained by the two PRSs across the four folds of the cross-validation procedure. Despite the small number of training data, it appears that *hiPRS* is able to find interactions that generalize well and that are far more predictive with respect to single SNPs. Indeed, *hiPRS* reports an average AUC of 0.72 and uses as little as 6 terms, counting the intercept. In contrast, the Penalized PRS introduces $9 \times 3 + 1 = 28$ parameters but only manages to obtain an average AUC of 0.57. This result supports the idea that SNPs interactions may play a role in quantifying the response to oxaliplatin treatments.

At this regard, it is of interest to know which interactions were selected by the *hiPRS* routine. We have reported in [Fig 8](#), right panel, the five interactions identified when fitting *hiPRS* on the whole dataset. Of note, we mention that one of the longest interactions, namely

$$\{\text{rs1172113} = 1, \text{rs4939378} = 1, \text{rs3812863} = 1, \text{rs1360485} = 2\},$$

was also the most robust. Indeed, that was the only interaction to be always selected during the cross-validation routine.

Discussion

In this work we presented *hiPRS*, a novel approach to tackle polygenic risk scoring that captures and models the effect on the phenotype of single SNPs and SNP-SNP interactions of potentially very high order. The algorithm takes individual-level genotype data as an input, overcoming potential biases of GWAS information and can work on any predefined set of SNPs to provide an easily interpretable tool not only for accurate prediction, but also to

perform solid inference and SNP-SNP interaction discovery or validation. *hiPRS* allows the user to define the size of the model and the maximum order of the interactions to search for, which allows for reliable parameter estimations, especially for small samples, and convenient inspection by domain practitioners.

We have tested *hiPRS* against similar benchmark methods that rely on individual-level data, demonstrating its superior performance with respect to traditional PRSs and more complex ML-based methods. Indeed, *hiPRS* performs significantly better than any model accounting for additive effects only (i.e. Lasso, Ridge, Elastic-Net and SVM-Behrvan) when the true generative mechanism incorporates SNP-SNP interactions, i.e., in the presence of *epistatic* effects. Moreover, its results are on par with state-of-the-art methods accounting for interactions, such as *glnet*, a more advanced penalized PRSs, and DNN-Badre, a nonlinear model based on artificial neural networks. Nevertheless, with respect to *glnet*, *hiPRS* allows to include interactions of higher order with a much smaller set of parameters to estimate. ML-based approaches like DNN-Badre can instead model interactions of extremely high order and perform very well in patients' scoring, as shown in our benchmark experiment, but tracing back the effect of all predictors or reconstructing the fundamental generating interactions from those complex black-box models is almost impossible. Even by applying to these models the wide variety of Explanation algorithms available, such as SHAP or LIME, one may at most quantify the role of each covariate alone, without however being able to retrieve any information on their interaction patterns.

There are some relevant facts about *hiPRS* that are worth mentioning. One lies in the fact that the algorithm explicitly defines and selects interaction terms as sequences of feature-level pairs. This aspect allows to gather fundamental insights on the true generating mechanism of the phenotype, especially in the presence of epistatic effects. For instance, there may be situations in which the same SNPs can have either a risk or a protective effect on the phenotype depending on the alleles' configuration, as in the simulated example for which we reported *hiPRS* selected interactions (cf. SNP₄ and SNP₅ in Fig 5). In that case, only *hiPRS* consistently assigned a proper effect to each level of the two features, by estimating a coefficient for each interaction that included one of the predictive SNP-allele pairs. While a method like *glnet* might capture the effect of specific genotypes, this comes at the expense of estimating a β parameter for each of the possible SNP-allele combinations. Conversely, *hiPRS* can directly select the predictive patterns only, and estimate their unique effect on the disease.

Besides all the relevant aspects mentioned above, in the present work we wished to demonstrate *hiPRS*' capability to tackle the complexities of real research scenarios. To this end, we validated our method through an extensive set of simulations, showcasing its ability to deal with noise, strong class imbalance and even extremely small sample sizes without overfitting on the training data. This result was made possible thanks to the combination of the frequency threshold δ , which reduces the number of candidate interactions to those appearing in a sufficiently large portion of the positive class, and the mRMR-based interaction selection algorithm. Indeed, minimizing redundancy (while maximizing relevance) forces *hiPRS* to select the most diverse predictive interactions, which allows the model to generalize and capture the broader picture of the generative mechanism. This aspect was confirmed inspecting the 10 interactions selected in Fig 5. While all competitors picked the SNPs defining the most frequent rules in training data only (i.e. rule C in that example), *hiPRS* managed to include instances of other less frequent generative rules, and assign these terms a large positive coefficient.

Besides testing *hiPRS* on simulated settings, we applied the algorithm to a Case Study on real genotype data. Results show a significant improvement in performance w.r.t. a traditional Penalized PRS. Notably, the two terms in the resulting model (cf. Fig 8, right panel) with the

highest positive effect on the phenotype are third- and second-order interactions, which would require an extremely large number of predictors for their effect to be captured by traditional PRSs. Conversely, *hi*PRS estimates their effect together with other three terms only, allowing for an accessible inspection of the lightweight resulting model. The results presented in this case study were only validated internally via cross-validation, as further investigations were not feasible due to the lack of external data. Nevertheless, the experiment was meant to showcase the potential of *hi*PRS and its applicability on real data. Since the obtained results are promising, we plan to deepen the clinical interpretation and to validate the discoveries in future works, as soon as a comparable external dataset is available. Of note, a previous study [49] conducted a systematic review on prediction models incorporating multiple SNPs for CRC risk prediction. The reviewed studies reach comparable performance to *hi*PRS only when including both genomic and clinical covariates in their models. There, the clinical information captures most of phenotype variability, while the independent effect of SNPs -represented via traditional PRSs- add little power to the model. In this sense, it appears that *hi*PRS can exploit much better the predictive power intrinsic in the genome, potentially raising AUC values if coupled with clinical information.

At the moment, one major limitation of *hi*PRS lies in the scalability of the algorithm, especially when the number of SNPs grows dramatically. In fact, the search of high-order interactions is intrinsically expensive, and the computational cost of the mRMR selection routine grows quadratically with the number of candidate interactions. For this reason, our algorithm is better suited for those contexts where the number of SNPs is limited, as in clinical studies that start from literature-validated SNPs. There, the computational cost can be almost negligible: for instance, fitting *hi*PRS always took less than 5 seconds in our experiments.

Nevertheless, the computational cost of *hi*PRS can be reduced in several ways, e.g. by increasing the frequency threshold δ or decreasing the maximum interaction length l_{\max} (for further insights, see Fig 10 in the Appendix). Furthermore, *hi*PRS can easily accommodate modern approaches to Frequent Itemset Mining, such as GPU-accelerated versions of *Apriori* [50] or single-scan algorithms [51, 52], to boost the preliminary search of interactions. Similarly, *hi*PRS can be integrated on top of those ML algorithms that are designed to return (long) lists of predictive high-order interactions, without actually building a score, see e.g. [53, 54].

Conclusions

In conclusion, in this work we presented *hi*PRS and demonstrated how this simple yet effective novel PRS approach can represent a valuable tool for the analysis of genotype data in a precision medicine framework. Indeed, it can effectively be exploited in real life experimental settings with potentially very small sample sizes to model either common or rare phenotypes with the inclusion of high-order SNP-SNP interactions effects. We largely discussed potentials and limitations of our proposed approach, but another remarkable aspect to note, is that the method can flexibly accommodate any kind of categorical data, even though *hi*PRS was presented here to process SNPs data with allele frequency categories. For instance, it might be effortlessly applied to data describing genetic variants or epigenetic mutations encoded as binary variables, single and multi-level categorical clinical information, or their combination. This makes the *hi*PRS approach particularly interesting to tackle multi-omic studies, or to model the recognized interactions between genotype and environmental factors, or to investigate the mediating role of clinical conditions in determining the genotype effect on complex traits, widening the scope of applicability and relevance of our proposal.

Materials and methods

Simulated data with non-linear interaction effect

The synthetic data was built as follows. We considered a pool of $p = 15$ independent SNPs, S_1, \dots, S_p , and a binary outcome variable Y . We assume each of the SNPs to follow a uniform distribution over $\{0, 1, 2\}$. In order to impose a nonlinear dependence between the two, we considered a model of the following form

$$Y := (1 - F)\tilde{Y} + F(1 - \tilde{Y}), \tag{1}$$

where \tilde{Y} is univocally determined by the fifteen SNPs via the relation below,

$$\begin{aligned} \tilde{Y} = 1 \Leftrightarrow \{S_4 = 2, S_5 = 2, S_6 = 2, S_7 = 2\} \vee \\ \{S_4 = 1, S_5 = 1, S_6 \neq 2, S_7 \neq 2\} \vee \{S_8 = S_9 = S_{10}\}, \end{aligned} \tag{2}$$

whereas $F \sim B(\epsilon)$ is a binary random variable, independent on the SNPs, that we use to model either noise in the data or unexplained variability. Note in fact that if $F = 0$, then \tilde{Y} and Y coincide, otherwise the labels are flipped. It is worth noting the following facts.

1. The generative model features SNPs interactions up to order 7. To see this, let

$$\begin{aligned} I_1 &:= \mathbf{1}_{\{2\}}(S_4)\mathbf{1}_{\{2\}}(S_5)\mathbf{1}_{\{2\}}(S_6)\mathbf{1}_{\{2\}}(S_7), \\ I_2 &:= \mathbf{1}_{\{1\}}(S_4)\mathbf{1}_{\{1\}}(S_5)(1 - \mathbf{1}_{\{2\}}(S_6)\mathbf{1}_{\{2\}}(S_7)), \\ I_3 &:= \sum_{l=0}^2 \mathbf{1}_{\{l\}}(S_8)\mathbf{1}_{\{l\}}(S_9)\mathbf{1}_{\{l\}}(S_{10}), \end{aligned} \tag{3}$$

where $\mathbf{1}_A$ denotes the indicator function of the set A . Then, the relationship described in (2) can be rephrased in terms of interactions as

$$\tilde{Y} = 1 - (1 - I_1)(1 - I_2)(1 - I_3). \tag{4}$$

Since $I_1 \cdot I_2 \equiv 0$, it is easily seen that the highest order interactions in the above are of order $4 + 3 = 7$.

2. Let $\tilde{p}_Y := \mathbb{P}(\tilde{Y} = 1)$ and $p_Y := \mathbb{P}(Y = 1)$. Upto some basic calculations, one can show that

$$\tilde{p}_Y = 121/729 \approx 16.6\% \quad \text{and} \quad p_Y = \tilde{p}_Y(1 - \epsilon) + (1 - \tilde{p}_Y)\epsilon. \tag{5}$$

To see this, define the events $A := \{S_4 = 2, S_5 = 2, S_6 = 2, S_7 = 2\}$, $B := \{S_4 = 1, S_5 = 1, S_6 \neq 2, S_7 \neq 2\}$, $C := \{S_8 = S_9 = S_{10}\}$. Since we assumed the SNPs to be independent and uniformly distributed, it follows that

$$\mathbb{P}(A) = (1/3)^4, \quad \mathbb{P}(B) = (1/3)^2 \cdot (2/3)^2, \quad \mathbb{P}(C) = 3 \cdot (1/3)^3. \tag{6}$$

Also, since A and B are disjoint, $\mathbb{P}(A \cup B) = 1/81 + 4/81 = 5/81$. Finally, since $A \cup B$ and C are independent, we have

$$\begin{aligned} \tilde{p}_Y &= \mathbb{P}(A \cup B \cup C) = \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}((A \cup B) \cap C) = \\ &= \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}(A \cup B)\mathbb{P}(C) = \\ &= 5/81 + 1/9 - (5/81)(1/9) = 121/729, \end{aligned} \tag{7}$$

thus proving the first statement in (5). Conversely, the second one is obvious as the

independence of \tilde{Y} and F implies

$$p_Y = \mathbb{E}[Y] = \mathbb{E}[(1 - F)\tilde{Y}] + \mathbb{E}[F(1 - \tilde{Y})] = \mathbb{E}[1 - F]\mathbb{E}[\tilde{Y}] + \mathbb{E}[F]\mathbb{E}[1 - \tilde{Y}]. \tag{8}$$

In particular, we note that for $\epsilon \leq 0.2$ we have $p_Y \leq 30\%$, resulting in a class imbalance where cases are the minority.

3. We can explicitly quantify the amount of variability in Y that cannot be explained by the SNPs only. Indeed,

$$1 - \frac{\text{Var}(\mathbb{E}[Y|S_1, \dots, S_p])}{\text{Var}(Y)} = 1 - \frac{(1 - 2\epsilon)^2 \tilde{p}_Y(1 - \tilde{p}_Y)}{p_Y(1 - p_Y)}, \tag{9}$$

since $\mathbb{E}[\tilde{Y}|S_1, \dots, S_p] = \tilde{Y}$ and thus $\mathbb{E}[Y|S_1, \dots, S_p] = (1 - 2\epsilon)\tilde{Y} + \epsilon$, by exploiting independence and classical properties of conditional expectations. We refer to (9) as to *missing heritability*. In our final experiment, we use the latter to provide better insights on the *hiPRS* performance for different values of ϵ .

4. For any $m \in \{1, \dots, p\}$ and $l \in \{0, 1, 2\}$, let $\mathbf{1}_{\{l\}}(S_m)$ be the indicator function of the event $\{S_m = l\}$, i.e. the l th dummy variable associated to the SNP S_m . The best model f_* for predicting Y with the given SNPs can be thought as the solution to the following minimization problem

$$f_* = \underset{f}{\operatorname{argmin}} \mathbb{E}|Y - f(S_1, \dots, S_p)|^2 \tag{10}$$

where f is any map from $\{0, 1, 2\}^p \rightarrow \mathbb{R}$. It is not hard to prove that there exists interaction terms $I_k = \prod_{m \in \mathcal{M}_k} \mathbf{1}_{\{l_m\}}(S_m)$, where $\mathcal{M}_k \subseteq \{1, \dots, p\}$, and coefficients β_k such that

$$f_*(S_1, \dots, S_p) = \beta_0 + \sum_k \beta_k I_k, \tag{11}$$

meaning that the best possible model has precisely the form hypothesized by *hiPRS*, see Eq (13) in the Proposed methodology. We recall that the solution to (10) is given by the conditional expectation, that is, $f_*(S_1, \dots, S_p) = \mathbb{E}[Y|S_1, \dots, S_p]$. However, as we noted previously, the latter equals $(1 - 2\epsilon)\tilde{Y} + \epsilon$. In particular, for our claim to hold, it is sufficient to prove that \tilde{Y} can be written in terms of sums of products of dummy variables: however, as we showed in (4), this is trivially true. More in general, this follows from what is known as the *canonical disjunctive normal form* of a boolean function [55].

Therefore, by selecting the correct group of interactions, *hiPRS* may actually recover the best model completely. Nonetheless, we remark that this is not our main priority: first of all, we aim at building a PRS that is sufficiently predictive of the phenotype; then, we may as well investigate whether the generative mechanism is being captured or not.

Simulated atrial fibrillation data

For our second simulation study we considered the case of genetic risk prediction for atrial fibrillation, a clinical setting in which interaction effects are known to take place (see e.g. [45]). In short, atrial fibrillation is an abnormal heart rhythm characterized by irregular and rapid beats that can lead to blood clots in the heart. At the genetic level, researchers have linked atrial fibrillation to a biological pathway known as the Renin–Angiotensin System (RAS). In fact, through a complicated mechanism that is yet to be understood completely, RAS genes can influence blood pressure values, and, indirectly, impact on the susceptibility to atrial

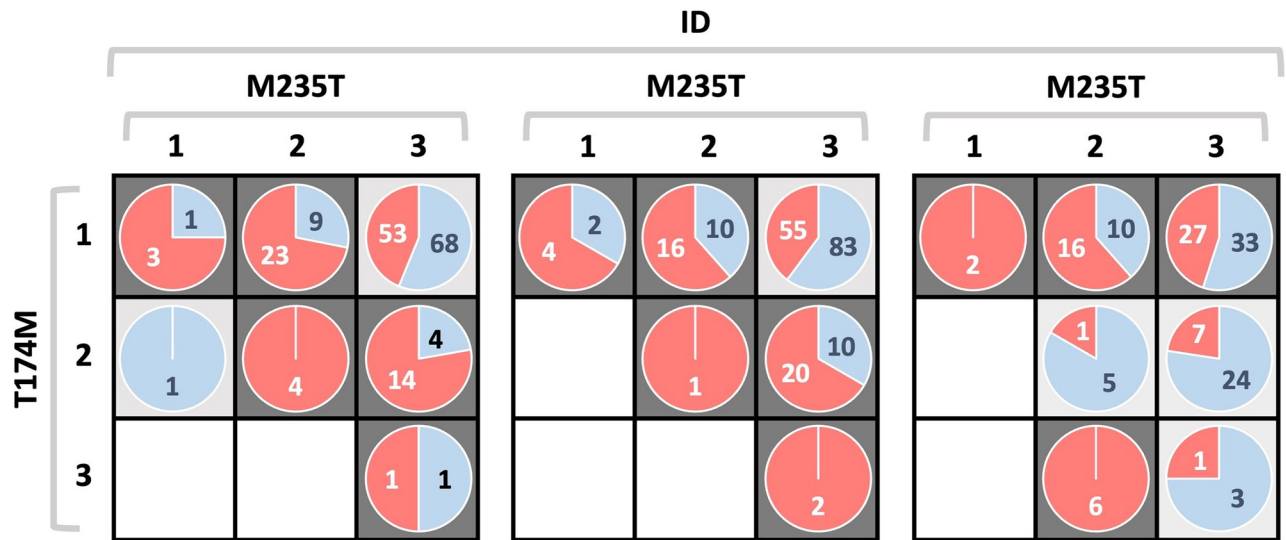


Fig 9. Simulated atrial fibrillation data. Schema of the distribution of cases (red section of the pie-plots) and controls (light blue section) in the case-control study described in [46], for each multilocus genotype combination of M235, T174M and ID. Dark gray cells are associated to higher risk of atrial fibrillation, while light gray cells are associated to protective combinations. The schema proposed here is based on the image reported in [45].

<https://doi.org/10.1371/journal.pone.0281618.g009>

fibrillation [56]. In Tsai *et al.* [46], the authors identify three main polymorphisms in the RAS pathway as associated to atrial fibrillation. The first two, M235 and T174M, are related to the *angiotensinogen gene*. This means they provide instructions for the production of a protein called angiotensinogen, a polipeptide that can become a vasoconstrictor when interacting with suitable enzymes. The third one, instead, is the so-called Insertion/Deletion polymorphism (ID), a mutation of the Angiotensin Converting Enzyme (ACE).

The case-control study in [46] counts 512 patients, half of which are cases. In their work, the authors report in full detail the joint distribution of M235, T174M and ID across the two classes, as summarized in Fig 9 (based on the related figure in [45]). Here, we use that empirical distribution to generate our own datasets. Equivalently: we apply bootstrap resampling to the cohort considered by Tsai *et al.* in [46].

DACHS data

For our case study we considered genotype-level data and clinical information coming from an ongoing population-based case-control study (DACHS, colorectal cancer: chances for prevention through screening). Details of the study have been previously described by Brenner *et al.* in [34, 57]. The dataset originally included patients recruited between 2003 and 2015. Genotype and complete follow-up data (for either 3, 5, or 10 years) were available for a total of 3689 histologically confirmed cases diagnosed between 2003 and 2014.

We excluded patients who have metastatic disease, who had not received adjuvant chemotherapy, received neoadjuvant chemotherapy, had an unknown start date of chemotherapy, or died within 30 days of the start of chemotherapy. Of patients treated with first-line adjuvant chemotherapy, we included stage II-III patients who received at least four cycles of oxaliplatin-based treatments. Finally, in order to study the short-term survival, we excluded all patients that died after 3 years. This resulted in a subsample of 349 patients. The main characteristics of the study population are shown in Table 2. Details about genotyping and imputation for the DACHS population have been described in detail somewhere else, see [33]. Among the available SNPs, we focused on the nine reported in Table 3. These are of particular interest as they

Table 2. Patient characteristics of the subsample analysed in the DACHS case study.

	Number (%)
Age (years ≤ 50)	52 (14.9%)
Sex (Female)	130 (37.2%)
Stage (Stage II)	28 (8.0%)
Grade (grade 3-4)	109 (31.9%)
CRC site (Rectum)	87 (24.9%)
Death	50 (14.3%)

<https://doi.org/10.1371/journal.pone.0281618.t002>

Table 3. SNPs considered in the DACHS case study.

Pathway	SNPs
Folate	rs1801131
Immunogenic cell death	rs1050305
	rs1799986
	rs11172113
	rs208294
	rs1718119
	rs1360485
Enterocyte subtype-related genes	rs4939378
	rs3812863

<https://doi.org/10.1371/journal.pone.0281618.t003>

were only recently validated as associated with the efficacy of oxaliplatin-based treatment in CRC patients [33].

Ethics statement and informed consent

All patients gave their written informed consent. The DACHS study was approved by the ethics committees of the Medical Faculty of the University of Heidelberg and the State Medical Boards of Baden-Wuerttemberg and Rhineland-Palatinate.

Proposed methodology

Preliminaries and problem statement. In order to describe our approach, we first introduce some notation. Let S_1, \dots, S_p be p SNPs and let Y be a random variable denoting a binary outcome. We consider each SNP to be a categorical variable taking values in $\{0, 1, 2\}$, where labels read as follows. We denote major allele homozygosis by 0, heterozygosis by 1, and minor allele homozygosis by 2. To model SNP-alleles interactions, we use products of indicator functions. For instance, $I := \mathbf{1}_{\{0\}}(S_1)\mathbf{1}_{\{1\}}(S_2)$ encodes the interaction between major allele homozygosis in S_1 and heterozygosis in S_2 . We may then define the collection of all SNP-allele interactions as

$$\mathcal{I} := \left\{ \prod_{m \in \mathcal{M}} \mathbf{1}_{\{l_m\}}(S_m) \text{ such that } \mathcal{M} \subseteq \{1, \dots, p\} \text{ and } l_m \in \{0, 1, 2\} \right\}. \tag{12}$$

We note that, with little abuse of notation, the set \mathcal{I} also contains the dummy-variables associated to the SNP-alleles, as those are obtained when \mathcal{M} is a singleton.

We are given a dataset $\{s_1^{(j)}, \dots, s_p^{(j)}, y^{(j)}\}_{j=1}^N$ consisting of N i.i.d. realizations of the SNPs and the outcome Y . Starting from these, we aim to construct a PRS of the form

$$hiPRS = \beta_0 + \beta_1 I_1 + \dots + \beta_K I_K \tag{13}$$

where $\{I_k\}_{k=1}^K \subset \mathcal{I}$. To this end, we propose a novel scoring method, *hiPRS*, where K is user-specified and a data-driven algorithm returns the list of interactions with the corresponding weights. We detail the whole idea in the following section.

hiPRS algorithm. For any $I \in \mathcal{I}$, let $\{i^{(j)}\}_{j=1}^N$ be the corresponding observations in the dataset. We define the collection of all cases and its complement as

$$O := \{j \mid y^{(j)} = 1\}, \quad Z := \{j \mid y^{(j)} = 0\}, \tag{14}$$

so that $O \subseteq \{1, \dots, N\}$ contains the indexes of those observations associated to the event $Y = 1$, whereas Z refers to those patients for which $Y = 0$. We make the assumption that $|O| \ll |Z|$, which is the typical scenario of a rare outcome. We refer to O and Z respectively as the minority and majority class. As a first step, we scan the data relative to the minority class, and we search for those interactions $\mathcal{I}_{\delta, l_{\max}}$ that appear with an empirical frequency above a given threshold $\delta > 0$, and have length at most l_{\max} . More precisely, we define

$$\mathcal{I}_{\delta, l_{\max}} := \left\{ I \in \mathcal{I} \text{ such that } \frac{1}{|O|} \sum_{j \in O} i^{(j)} > \delta \text{ and } \text{Length}(I) \leq l_{\max} \right\}, \tag{15}$$

where the $\text{Length}(I)$ is the number of SNPs involved in the definition of I . In principle, computing $\mathcal{I}_{\delta, l_{\max}}$ can be very demanding since $|\mathcal{I}| = 4^p$. However, this drawback is mitigated by two key factors. First of all, we note that each interaction uniquely corresponds to a *pattern of alleles*. For instance, let

$$I = \mathbf{I}_{\{1\}}(S_2) \mathbf{I}_{\{0\}}(S_3) \mathbf{I}_{\{1\}}(S_5) = \mathbf{I}_{\{1\} \times \{0\} \times \{1\}}(S_2, S_3, S_5). \tag{16}$$

Then $i^{(j)} = 1$ if and only if the pattern $\{S_2 = 1, S_3 = 0, S_5 = 1\}$ is observed in the j th patient. This duality between interactions and patterns allows us to reframe (15) in the context of frequent itemsets mining, where we can relay on a multitude of algorithms such as *Apriori* and *FP-Growth*. Additionally, the computational cost is alleviated by the fact that we limit the search of candidates to the minority class O .

The next step is to extract a suitable sublist $\{I_k\}_{k=1}^K \subset \mathcal{I}_{\delta, l_{\max}}$ to be used in (13). While this problem can be framed in the context of feature selection, finding an optimal solution can be very hard due to the large number of candidates. To overcome this drawback, we introduce a filtering technique based on the so-called Minimum Redundancy—Maximum Relevance approaches, mRMR for short. The idea goes as follows. First, we introduce a relevance measure based on the (empirical) mutual information, that is

$$\mathbb{I}(I, Y) := \sum_{\substack{j=1, \dots, N \\ a=0,1 \\ b=0,1}} \frac{1}{N} |\{i^{(j)} = a, y^{(j)} = b\}| \log \left(\frac{|\{i^{(j)} = a, y^{(j)} = b\}|/N}{|\{i^{(j)} = a\}|/N \cdot |\{y^{(j)} = b\}|/N} \right). \tag{17}$$

By definition, $\mathbb{I}(I, Y) \geq 0$ and larger values are obtained when I and Y are informative with respect to each other. Parallel to this, we introduce a redundancy measure $\mathbb{S} : \mathcal{I} \times \mathcal{I} \rightarrow \mathbb{N}$ that quantifies the similarity between two given interactions. More precisely, we define $\mathbb{S}(I, T)$ to be the number of common alleles within the two patterns. We then construct the set $\{I_k\}_{k=1}^K$ along the lines of mRMR methods, that is through the greedy optimization of the ratio between relevance and redundancy. We report a detailed pseudocode in Algorithm 1.

We remark that K is a hyperparameter chosen by the user. In particular, one may as well optimize its value accordingly to some grid-search algorithm of choice. Indeed, the most computationally expensive parts in the *hiPRS* pipeline are the candidates search and the

computation of relevance/redundancy measures. Once these steps have been carried out, multiple values of K can be tested and the user may choose the one considered to be optimal (e.g., the one yielding the model with the highest AUC).

Algorithm 1: Extraction of the *hi*PRS allele-patterns, starting from a list of candidates and raw data.

```

Input:  $K, \mathcal{I}_{\delta, I_{\max}}$  and  $\{(i_1^{(j)}, \dots, i_{|\mathcal{I}_{\delta, I_{\max}}|}^{(j)})\}_{j=1}^N$ .
Output:  $\{I_k\}_{k=1}^K \subset \mathcal{I}_{\delta, I_{\max}}$ .
/* Select most relevant interaction */
 $I_1 \leftarrow \operatorname{argmax}_{I \in \mathcal{I}_{\delta, I_{\max}}} \mathbb{I}(I, Y)$ 
 $\mathcal{I}_{\text{picked}} \leftarrow \{I_1\}$ 
 $\mathcal{I}_{\text{left}} \leftarrow \mathcal{I}_{\delta, I_{\max}} \setminus \mathcal{I}_{\text{picked}}$ 
/* Add patterns iteratively */
while  $|\mathcal{I}_{\text{picked}}| < K$  do
  for  $I \in \mathcal{I}_{\text{left}}$  do
     $V_I \leftarrow \mathbb{I}(I, Y)$  // Relevance
     $W_I \leftarrow \frac{1}{|\mathcal{I}_{\text{picked}}|} \sum_{T \in \mathcal{I}_{\text{picked}}} \mathbb{S}(I, T)$  // Average redundancy
  end
  if  $\min_{I \in \mathcal{I}_{\text{left}}} W_I = 0$  then
    /* If no redundancy, select most relevant */
     $I^* \leftarrow \operatorname{argmax}_{I \in \mathcal{I}_{\text{left}}} V_I$ 
  else
    /* Otherwise, pick best compromise */
     $I^* \leftarrow \operatorname{argmax}_{I \in \mathcal{I}_{\text{left}}} (V_I / W_I)$ 
  end
   $\mathcal{I}_{\text{picked}} \leftarrow \mathcal{I}_{\text{picked}} \cup \{I^*\}$ 
   $\mathcal{I}_{\text{left}} \leftarrow \mathcal{I}_{\delta, I_{\max}} \setminus \mathcal{I}_{\text{picked}}$ 
end
return  $\mathcal{I}_{\text{picked}}$ 

```

Once the set $\mathcal{I}_{\text{picked}} = \{I_k\}_{k=1}^K$ has been identified, we compute the weights β_0, \dots, β_K by fitting the Logistic Regression model below

$$\operatorname{logit}\mathbb{P}(Y = 1) = \beta_0 + \beta_1 I_1 + \dots + \beta_K I_K. \tag{18}$$

Finally, we define the *hi*PRS accordingly to (13), meaning that the score is actually the (affine) linear predictor in (18).

Benchmark algorithms

We report below a short description and the benchmark algorithms together with their implementation details.

Penalized PRSs with additive effects. From this class of methods we chose three of *hi*PRS competitors, namely Lasso [35], Ridge [36] and Elastic-Net [37]. These algorithms are traditional penalized LR models that only account for the additive main effect of the predictors on the target variable, while imposing an ℓ_2 bound of the form $\|\beta\|_2^2 \leq s$ (Ridge) or an ℓ_1 bound $\|\beta\|_1 \leq s$ (Lasso), or the combination of the two (Elastic-Net) on the coefficients. Note that ℓ_1 penalizations also perform feature selection by shrinking coefficients to zero, making these approaches popular to model polygenic risk from large genotype data. Fundamentally, the penalization terms enter the loss function during the training phase, and their contribution is weighted by some hyperparameter, e.g. λ for Lasso and Ridge, λ_1 and λ_2 for Elastic-Net. To run these algorithms in our experiments, we relied on the implementation available in the

Python library scikit-learn, with default values for λ , λ_1 and λ_2 . All the code was written in Python 3.7.

Glinternet. Glinternet is a method for learning pairwise interactions in a LR satisfying strong hierarchy: whenever an interaction is estimated to be nonzero, both its associated main effects are also included in the model. The idea of the algorithm is based on a variant of Lasso, namely group-Lasso [58], that sets groups of predictors to zero. glinternet sets up main effects and first-order (i.e., pairwise) interactions via groups of variables and then intuitively selects those that have a strong overall contribution from all their levels toward explaining the response. For more formal definitions we refer the reader to [38].

The amount of regularization is controlled by λ , with larger values corresponding to stronger shrinkage and less interactions included. The freely distributed R implementation of glinternet allows the user to define the number of pairwise interactions (n_int) to find and the size of a grid of λ values of decreasing strength to fit the model with. This grid is built automatically by splitting equally from λ_{max} , that is data-derived as the value for which all coefficients are zero, to the minimum value, computed as $\lambda_{min} = \lambda_{max}/0.01$. The algorithm will fit this path of values and stop when n_int is reached. For this experiment we set $n_int = [3, 8]$, defining a grid of 100 λ s. All other hyperparameters were left to default settings.

DNN-Badre. DNN-Badre [40] constructs a Deep Feed Forward Neural Network (DNN) to predict a binary outcome. To implement their approach, we built a DNN architecture with the same specifics provided by the authors, i.e. number of layers, neurons and activation functions. We trained the models via stochastic gradient descent, with mini-batches of batch size 10 and for a total of 200 epochs. The optimization was carried out using the Adam optimizer, with a learning rate of 10^{-3} . All the code was written in Python 3.7, mostly using the Pytorch library.

SVM-Behravan. SVM-Behravan is a multi-step ML-based algorithm recently presented as a state-of-the-art approach to polygenic risk scoring [39, 59]. We will provide here an intuitive description of the algorithm with the needed details to understand our settings for the present work, while for more technical specifications we refer the reader to the seminal work in [39]. The algorithm presented in Behravan et al. (2018) is composed of a so called *first module*, where an XGBoost is used to evaluate the importance of SNPs on a risk prediction task by providing an initial list of candidate predictive SNPs. The authors use the average of feature importances (a.k.a. “gain”) provided by the gradient tree boosting method, as the contribution of each SNP to the risk. Then, in the *second module*, the candidate SNPs are used for an adaptive iterative search to capture the optimal ways of combining candidate SNPs to achieve high risk prediction accuracy on validation data. In particular, top M and bottom M SNPs from the candidate list are ranked separately based on accuracy, then top and bottom N SNPs are switched between the two lists. The process is repeated with M (i.e., *window size*) increased of W at each iteration, until the two sublists overlap and the optimal ranking is achieved. Finally, an SVM is trained to distinguish cases (positive samples) and controls (negative samples) using the S top-ranked SNPs in the optimal ranking as feature vectors and a linear kernel. In the original paper the performance of the algorithm is averaged across 5-fold CV, meaning that the pipeline from first module to SVM is repeated 5 times on different folds of the training and test set. As the author included this step to overcome the problem of small samples to train high-performance risk prediction models, we followed their instructions and for each of the 30 simulated dataset we performed the 5-fold CV. This lead to the 5 different subsets of SNPs selected as shown in the right panel of Fig 5. However, to keep the running time of this computationally very expensive method reasonable and for a fair comparison with all the other methods for which we did not perform extensive optimal hyperparameter search, we avoided the very long optimization of the

XGBoost model by setting manually the optimal hyperparameters described in the paper. In particular, we set $M = 2$, $W = 1$, $N = 1$.

Performance metrics

In imbalanced settings, the Accuracy of a classifier (i.e., the fraction of correctly classified observations) can be a misleading metric to assess a model performance. Therefore, to evaluate *hiPRS* and the benchmark algorithms in predicting the outcome, we exploited two metrics that combined provide a full picture of the real performance of these methods, especially in predicting the most interesting class (i.e. the positive class), which is oftentimes the underrepresented one in real data. Therefore, for our experiments we chose the Area Under the receiving operating characteristic Curve (AUC) and the Average Precision (AP).

The two are computed as follows. For any fixed discrimination threshold, let TP and FP be the true and false positives, respectively. Similarly, let TN and FN be the true and false negatives, respectively. First, we derive the following quantities,

$$TPR = \frac{TP}{TP + FP}, \quad FPR = \frac{FP}{FP + FN}, \quad Recall = \frac{TP}{TP + FN}. \quad (19)$$

that is, the True Positive Rate (TPR), the False Positive Rate (FPR) and the Recall. TPR is the fraction of true positives out of the positives, also known as *sensitivity* or *precision*. FPR, or *specificity*, is the fraction of false positives out of the negatives. Finally, the Recall quantifies the ability of the classifier to recognize positive samples. By plotting TPR against FPR for various discrimination threshold levels, we obtain the ROC curve. Conversely, the pair ($Recall$, TPR) yields the precision-recall curve. AUC summarizes the performance of a binary classifier by computing the area under the ROC curve, while AP considers the area under the precision-recall curve. In practice, the two areas are estimated using quadrature rules. In our implementation, we employ the trapezoidal rule for AUC, while we use the rectangular rule for AP.

Appendix

Computational cost and time complexity of *hiPRS*

In this section, we provide some additional insights about the computational cost entailed by *hiPRS*. To this end, we recall the notation used in the Materials and Methods section: here, p , δ , K are the number of SNPs, the frequency threshold over minority class, and the final number of interactions included in the model, respectively.

The major source of computational cost, is given by the preliminary search of candidate interactions. In fact, even though we rely on efficient data mining algorithms, the identification of such candidates may still require $\propto e^{p^2}$ operations. This computational burden is not affected by the choice of K , as that only impacts the final (cheaper) selection, while it is related to the thresholding value δ . Fig 10 shows the overall time complexity of *hiPRS*, when dealing with increasingly large datasets, $p = 15, 20, \dots, 100$, for different choices of δ . Here, we fixed $K = 40$ and $l_{\max} = +\infty$ for all the experiments, and we generated the data following the same procedure adopted in our first simulation study (total number of observations $N = 1000$). We mention that the picture is reported with a logarithmic scale on the y -axis.

For large thresholds, $\delta \geq 0.2$, the algorithm can be safely applied to datasets featuring hundreds of SNPs, with a total fitting time of few seconds. For $0.1 \leq \delta \leq 0.2$, the situation remains manageable, with *hiPRS* taking at most 1-5 minutes to process 100 SNPs. Conversely, smaller values of δ can make the algorithm prohibitive if too many SNPs are considered ($\delta \leq 0.05$ and $p \geq 40$). In general, the overall computational cost always increases exponentially with p , but the rate at which this happens is modulated by δ . This means that

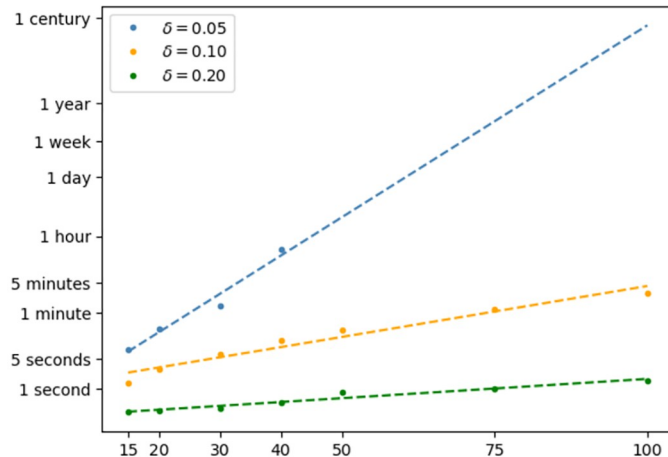


Fig 10. Time complexity of *hiPRS*. Fitting times of *hiPRS* for different values of δ and different numbers of SNPs (x-axis). For better readability, the y-axis is reported in logarithmic scale. Dashed-lines are obtained via least-squares.

<https://doi.org/10.1371/journal.pone.0281618.g010>

the algorithm can easily scale up when analyzing high-order interactions between common genetic variants with high penetrance in the cases' population, which allows the user to set higher δ thresholds without the risk of missing informative patterns. On the other hand, when seeking interaction patterns that include rare or ultra-rare variants, to keep computational time manageable, a strong initial candidate SNP selection (exploiting prior knowledge and/or model-based approaches) is recommended.

Author Contributions

Conceptualization: Michela C. Massi, Nicola R. Franco, Anna Maria Paganoni.

Data curation: Nicola R. Franco, Hanla A. Park, Michael Hoffmeister, Hermann Brenner, Jenny Chang-Claude.

Formal analysis: Michela C. Massi, Nicola R. Franco.

Methodology: Michela C. Massi, Nicola R. Franco.

Project administration: Paolo Zunino.

Supervision: Andrea Manzoni, Anna Maria Paganoni, Francesca Ieva, Paolo Zunino.

Writing – original draft: Michela C. Massi, Nicola R. Franco.

Writing – review & editing: Andrea Manzoni, Anna Maria Paganoni, Hanla A. Park, Jenny Chang-Claude, Francesca Ieva, Paolo Zunino.

References

1. Konuma T, Okada Y. Statistical genetics and polygenic risk score for precision medicine. *Inflammation and Regeneration*. 2021; 41(1):1–5. <https://doi.org/10.1186/s41232-021-00172-9> PMID: 34140035
2. Song S, Jiang W, Hou L, Zhao H. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS computational biology*. 2020; 16(2):e1007565. <https://doi.org/10.1371/journal.pcbi.1007565> PMID: 32045423
3. Abraham G, Inouye M. Genomic risk prediction of complex human disease and its clinical application. *Current opinion in genetics & development*. 2015; 33:10–16. <https://doi.org/10.1016/j.gde.2015.06.005> PMID: 26210231

4. Che R, Motsinger-Reif A. Evaluation of genetic risk score models in the presence of interaction and linkage disequilibrium. *Frontiers in genetics*. 2013; 4:138. <https://doi.org/10.3389/fgene.2013.00138> PMID: 23888168
5. Chasioti D, Yan J, Nho K, Saykin AJ. Progress in polygenic composite scores in Alzheimer's and other complex diseases. *Trends in Genetics*. 2019; 35(5):371–382. <https://doi.org/10.1016/j.tig.2019.02.005> PMID: 30922659
6. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*. 2020; 15(9):2759–2772. <https://doi.org/10.1038/s41596-020-0353-1> PMID: 32709988
7. Janssens ACJ. Validity of polygenic risk scores: are we measuring what we think we are? *Human molecular genetics*. 2019; 28(R2):R143–R150. <https://doi.org/10.1093/hmg/ddz205> PMID: 31504522
8. Shi J, Park JH, Duan J, Berndt ST, Moy W, Yu K, et al. Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS genetics*. 2016; 12(12):e1006493. <https://doi.org/10.1371/journal.pgen.1006493> PMID: 28036406
9. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. *Elife*. 2019; 8:e39725. <https://doi.org/10.7554/eLife.39725> PMID: 30895923
10. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *Elife*. 2019; 8:e39702. <https://doi.org/10.7554/eLife.39702> PMID: 30895926
11. Multhaup ML, Kita R, Krock B, Eriksson N, Fontanillas P, Aslibekyan S, et al. The science behind 23andMe's Type 2 Diabetes report. *Sunnyvale (CA): 23andMe*. 2019; p. 23–19.
12. Lehner B. Modelling genotype–phenotype relationships and human disease with genetic interaction networks. *Journal of Experimental Biology*. 2007; 210(9):1559–1566. <https://doi.org/10.1242/jeb.002311> PMID: 17449820
13. Lehner B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*. 2011; 27(8):323–331. <https://doi.org/10.1016/j.tig.2011.05.007> PMID: 21684621
14. Abraham G, Kowalczyk A, Zobel J, Inouye M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic epidemiology*. 2013; 37(2):184–195. <https://doi.org/10.1002/gepi.21698> PMID: 23203348
15. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. *PLoS genetics*. 2014; 10(11):e1004754. <https://doi.org/10.1371/journal.pgen.1004754> PMID: 25393026
16. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*. 2002; 11(20):2463–2468. <https://doi.org/10.1093/hmg/11.20.2463> PMID: 12351582
17. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *The American Journal of Human Genetics*. 2009; 85(3):309–320. <https://doi.org/10.1016/j.ajhg.2009.08.006> PMID: 19733727
18. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. *Frontiers in genetics*. 2019; 10:267. <https://doi.org/10.3389/fgene.2019.00267> PMID: 30972108
19. Vivian-Griffiths T, Baker E, Schmidt KM, Bracher-Smith M, Walters J, Artemiou A, et al. Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2019; 180(1):80–85. <https://doi.org/10.1002/ajmg.b.32705> PMID: 30516002
20. Silver M, Chen P, Li R, Cheng CY, Wong TY, Tai ES, et al. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. *PLoS genetics*. 2013; 9(11):e1003939. <https://doi.org/10.1371/journal.pgen.1003939> PMID: 24278029
21. Lee KY, Leung KS, Ma SL, So HC, Huang D, Tang NLS, et al. Genome-Wide Search for SNP Interactions in GWAS Data: Algorithm, Feasibility, Replication Using Schizophrenia Datasets. *Frontiers in genetics*. 2020; 11. <https://doi.org/10.3389/fgene.2020.01003> PMID: 33133133
22. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS genetics*. 2009; 5(10):e1000678. <https://doi.org/10.1371/journal.pgen.1000678> PMID: 19816555
23. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet*. 2015; 97(4):576–592. <https://doi.org/10.1016/j.ajhg.2015.09.001> PMID: 26430803
24. Habier D, Fernando RL, Garrick DJ. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. *Genetics*. 2013; 194:597–607. <https://doi.org/10.1534/genetics.113.152207> PMID: 23640517

25. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*. 2001; 69(1):138–147. <https://doi.org/10.1086/321276> PMID: 11404819
26. Collins RL, Hu T, Wejse C, Sirugo G, Williams SM, Moore JH. Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. *Bio-Data mining*. 2013; 6(1):1–5. <https://doi.org/10.1186/1756-0381-6-4> PMID: 23418869
27. Taylor MB, Ehrenreich IM. Higher-order genetic interactions and their contribution to complex traits. *Trends in genetics*. 2015; 31(1):34–40. <https://doi.org/10.1016/j.tig.2014.09.001> PMID: 25284288
28. Guerrero RF, Scarpino SV, Rodrigues JV, Hartl DL, Ogbunugafor CB. Proteostasis environment shapes higher-order epistasis operating on antibiotic resistance. *Genetics*. 2019; 212(2):565–575. <https://doi.org/10.1534/genetics.119.302138> PMID: 31015194
29. Franco NR, Massi MC, Ieva F, Manzoni A, Paganoni AM, Zunino P, et al. Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. *Radiotherapy and Oncology*. 2021; 159:241–248. <https://doi.org/10.1016/j.radonc.2021.03.024> PMID: 33838170
30. Manduchi E, Le TT, Fu W, Moore JH. Genetic analysis of coronary artery disease using tree-based automated machine learning informed by biology-based feature selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2021; 19(3):1379–1386. <https://doi.org/10.1109/TCBB.2021.3099068>
31. Massi MC, Gasperoni F, Ieva F, Paganoni AM, Zunino P, Manzoni A, et al. A deep learning approach validates genetic risk factors for late toxicity after prostate cancer radiotherapy in a REQUITE multinational cohort. *Frontiers in oncology*. 2020; 10:541281. <https://doi.org/10.3389/fonc.2020.541281> PMID: 33178576
32. Mahendran N, PM DRV. A deep learning framework with an embedded-based feature selection approach for the early detection of the Alzheimer's disease. *Computers in Biology and Medicine*. 2022; 141:105056. <https://doi.org/10.1016/j.compbiomed.2021.105056> PMID: 34839903
33. Park HA, Seibold P, Edelmann D, Benner A, Canzian F, Alwers E, et al. Validation of genetic markers associated with survival in colorectal cancer patients treated with oxaliplatin-based chemotherapy. *Cancer Epidemiology and Prevention Biomarkers*. 2022; 31(2):352–361. <https://doi.org/10.1158/1055-9965.EPI-21-0814> PMID: 34862210
34. Brenner H, Chang-Claude J, Seiler ea Christoph M. Colonoscopy Prevents Colorectal Cancer in Both the Right and Left Colon. *Gastroenterology*. 2011; 141(1):393–396. <https://doi.org/10.1053/j.gastro.2011.05.015>
35. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996; 58(1):267–288.
36. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 1970; 12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>
37. Zou H, Hastie T. Regression shrinkage and selection via the elastic net, with applications to microarrays. *JR Stat Soc Ser B*. 2003; 67:301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
38. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*. 2015; 24(3):627–654. <https://doi.org/10.1080/10618600.2014.938812> PMID: 26759522
39. Behravan H, Hartikainen JM, Tengström M, Pyrkäs K, Winqvist R, Kosma VM, et al. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific reports*. 2018; 8(1):1–13. <https://doi.org/10.1038/s41598-018-31573-5>
40. Badré A, Zhang L, Muchero W, Reynolds JC, Pan C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics*. 2021; 66(4):359–369. <https://doi.org/10.1038/s10038-020-00832-7> PMID: 33009504
41. Cecile A, Janssens J, Joyner MJ. Polygenic risk scores that predict common diseases using millions of single nucleotide polymorphisms: is more, better? *Clinical chemistry*. 2019; 65(5):609–611. <https://doi.org/10.1373/clinchem.2018.296103>
42. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. *Epigenetics*. 2020; 15(1-2):1–11. <https://doi.org/10.1080/15592294.2019.1644879> PMID: 31318318
43. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. <https://doi.org/10.1038/nature08494> PMID: 19812666
44. Mackay TF, Moore JH. Why epistasis is important for tackling complex human disease genetics. *Genome medicine*. 2014; 6(6):1–3. <https://doi.org/10.1186/gm561> PMID: 25031624

45. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*. 2005; 27(6):637–646. <https://doi.org/10.1002/bies.20236> PMID: 15892116
46. Tsai CT, Lai LP, Lin JL, Chiang FT, Hwang JJ, Ritchie MD, et al. Renin-angiotensin system gene polymorphisms and atrial fibrillation. *Circulation*. 2004; 109(13):1640–1646. <https://doi.org/10.1161/01.CIR.0000124487.36586.26> PMID: 15023884
47. Voskoboinik A, Prabhu S, Ling Lh, Kalman JM, Kistler PM. Alcohol and atrial fibrillation: a sobering review. *Journal of the American College of Cardiology*. 2016; 68(23):2567–2576. <https://doi.org/10.1016/j.jacc.2016.08.074> PMID: 27931615
48. Staerk L, Sherer JA, Ko D, Benjamin EJ, Helm RH. Atrial fibrillation: epidemiology, pathophysiology, and clinical outcomes. *Circulation research*. 2017; 120(9):1501–1517. <https://doi.org/10.1161/CIRCRESAHA.117.309732> PMID: 28450367
49. Sassano M, Mariani M, Quaranta G, Pastorino R, Boccia S. Polygenic risk prediction models for colorectal cancer: a systematic review. *BMC cancer*. 2022; 22(1):1–21. <https://doi.org/10.1186/s12885-021-09143-2> PMID: 35030997
50. Zhang F, Zhang Y, Bakos J. Gp priori: Gpu-accelerated frequent itemset mining. In: 2011 IEEE International Conference on Cluster Computing. IEEE; 2011. p. 590–594.
51. Djenouri Y, Djenouri D, Lin JCW, Belhadi A. Frequent Itemset Mining in Big Data With Effective Single Scan Algorithms. *IEEE Access*. 2018; 6:68013–68026. <https://doi.org/10.1109/ACCESS.2018.2880275>
52. Djenouri Y, Djenouri D, Belhadi A, Cano A. Exploiting GPU and cluster parallelism in single scan frequent itemset mining. *Information Sciences*. 2019; 496:363–377. <https://doi.org/10.1016/j.ins.2018.07.020>
53. Basu S, Kumbier K, Brown JB, Yu B. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*. 2018; 115(8), 1943–1948. <https://doi.org/10.1073/pnas.1711236115> PMID: 29351989
54. Fang G, Haznadar M, Wang W, Yu H, Steinbach M, Church TR, et al. High-order SNP combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PloS one*. 2012; 7(4), e33531. <https://doi.org/10.1371/journal.pone.0033531> PMID: 22536319
55. Pahl PJ, Damrath R. *Mathematical foundations of computational engineering: a handbook*. Springer Science & Business Media; 2001.
56. Takahashi N, Smithies O. Human genetics, animal models and computer simulations for studying hypertension. *TRENDS in Genetics*. 2004; 20(3):136–145. <https://doi.org/10.1016/j.tig.2004.01.004> PMID: 15036807
57. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med*. 2014; 154(1):22–30. <https://doi.org/10.7326/0003-4819-154-1-201101040-00004>
58. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68(1):49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x>
59. Behravan H, Hartikainen JM, Tengström M, Kosma VM, Mannermaa A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific Reports*. 2020; 10(1):1–16. <https://doi.org/10.1038/s41598-020-66907-9> PMID: 32632202