

Insights into epistatic oligogenic traits using interpretable symbolic regression models

Daniele Raimondi ^{1*}

¹Institut de Génétique Moléculaire de Montpellier (IGMM), Université de Montpellier,
CNRS UMR 5535, 34090 Montpellier, France
* daniele.raimondi@igmm.cnrs.fr

Abstract

Epistasis, or the interaction between genetic loci, plays a crucial role in shaping oligogenic traits, yet its complexity often eludes traditional modeling approaches, which historically were more focused on detecting the additive contributions of loci towards a target phenotypes. While machine learning (ML) methods can capture nonlinear interactions between loci, their black-box nature limits their utility for uncovering the underlying mechanisms of epistatic traits. In this study, we propose Symbolic Regression (SR) as an interpretable alternative to model and analyze epistatic oligogenic traits. Using the Feyn Python library, we benchmarked SR against traditional linear models and nonlinear ML approaches on synthetic datasets and a real-world protein fitness landscape. Our results demonstrate that SR excels in scenarios with relatively few causative loci (< 50), providing interpretable, parsimonious models that remain robust under low sample sizes conditions, which are often common in quantitative genetics. SR models not only achieve competitive predictive accuracy but also reveal key epistatic interactions through the analysis of their mathematical expressions, enabling insights into higher-order genetic relationships. These findings underscore the potential of SR to advance our understanding of epistasis, offering a novel avenue for deciphering the complexity of genotype-phenotype relationships.

1 Introduction

Understanding the complex relationship between genotype and phenotype is a fundamental challenge in biology. Accurate modeling of this relationship could transform various fields, from genetics to medicine and agriculture, enabling groundbreaking advances. In clinical genetics, this understanding will be central to enable Precision Medicine [1], where treatments are tailored to individual genomes, offering personalized healthcare [2]. In agricultural technology, it could lead to the development of crops that are better adapted to the challenges of climate change and food security, addressing critical global concerns.

One of the primary barriers to achieving this goal lies in the complexity of the genome itself. Every individual carries millions of genetic variants [3], only a small subset of which influence observable traits or disease susceptibility [4]. This complexity, combined with the limitations of traditional approaches, has made it difficult to identify the variants responsible for oligo-to-polygenic traits. For example, while Genome-Wide Association Studies (GWAS) have been instrumental in linking many Mendelian genetic loci to traits [5], as a method it suffers from the inability to detect epistatic interactions between variants, since it selects the relevant loci based on univariate statistical tests [6]. GWAS therefore fails to capture the full extent of heritability [7] due to the fact that it is oblivious of any interaction between loci [8, 9, 10, 11].

Epistatic interactions, where the effect of one gene is influenced by the presence of others [9, 10, 12], are important for understanding oligogenic traits, which are influenced by the combined effect of a small number of loci (i.e. genes) [13]. Traditional modeling techniques derived from GWAS, like the Polygenic Risk Scores (PRS) [14, 15], which are commonly used in clinical genetics, are additive models and therefore ignore nonlinear interactions between the input variants, similarly to linear or logistic regressions.

The modeling of epistatic interactions is where Machine Learning (ML) offers a promising avenue over conventional statistical linear methods such as PRS [16]. By leveraging the capacity of ML to model complex, nonlinear interactions, we can ideally unravel the intricate web of genetic relationships that drive phenotypic outcomes, significantly improving our understanding of oligogenic traits [11, 16]. Unfortunately, the ability of nonlinear ML models to learn arbitrarily complex interaction patterns between loci comes at the cost of severely reduced interpretability and explainability of the decision process leading to their predictions. Although the predictions of a linear regression or a PRS model can be easily understood, since the output score is just the linear combination of the input variants with some predefined weights, interpreting a Support Vector Machine with nonlinear kernels or multi-layer Neural Networks (NN) is much more intricate, giving them the rightful reputation of being *black boxes* [17]. Although several attempts at interpreting nonlinear ML models such as NNs have been made, the Explainable Artificial Intelligence (XAI) field is relatively new, and often it is still unclear whether those approaches are reliable and in which conditions [17], as we also recently shown [18].

An underexplored alternative to this problem is provided by the Symbolic AI field, also called Symbolic Regression (SR) [19]. Instead of optimizing the weights of a parametric model with a fixed functional form, like linear regressions, SVM and NN do, SR methods aim at finding human-readable structures, such as mathematical expressions, that best describe a given dataset. To do so, SR methods use heuristic search algorithms [20], such as genetic programming, to explore the vast space of syntactically valid mathematical expressions, combining operators and variables to discover the underlying relationships between inputs and outputs. This procedure yields fully interpretable models, since the output is a valid mathematical expression [19, 21]. SR could thus be invaluable for uncovering hidden patterns or dynamics in complex data. Among its main drawbacks, SR is particularly useful when the number of features is small, and it is therefore used when the interpretation of the model and the analysis of the interactions between features are of primary interest [21, 22].

In this paper, we propose what is, to the best of our knowledge, the first application of SR to the modeling and interpretation of epistatic oligogenic traits. To do so, we use the Feyn python library [20, 22], and we applied it to different datasets, benchmarking the potential of SR against conventional ML methods and linear models used in quantitative genetics, both in terms of predictions and interpretability.

First, we generated 144 synthetic SNP-array datasets, testing the behavior of Feyn SR on 2160 unique synthetic phenotypes obtained with different combinations of heritability, number of causative variants (Quantitative Trait Loci, QTL), epistatic and dominance effects. Second, we generated two synthetic SNP-array datasets in which the phenotypes to be predicted are respectively simulated with a Gaussian and a Sigmoid crater model landscapes[23]. Third, we tested SR on a real world dataset from [24], in which the authors characterized the complete fitness landscape of 4 epistatic loci on protein GB1 characterizing the full $20^4 = 160000$ mutants.

Our synthetic and real-life benchmarks show that SR is beneficial when compared to conventional linear and nonlinear ML methods for the prediction of oligogenic epistatic phenotypes, namely when epistatic effects among relatively few (< 50) QTLs are involved. Moreover, the parsimonious models obtained from the SR are particularly robust at low sample sizes, and the automatic inspection of the SR equations allows the identification of the epistematically interacting pairs in a robust manner.

Finally, we test our SR approach on the GB1 protein fitness landscape, showing that an extremely parsimonious symbolic models (with less than 100 parameters) outperform linear ones and has just 26% lower Pearson correlation with respect to NN models with two orders of magnitude more parameters. At the same time, the SR models are able to consistently retrieve the loci that are interacting more tightly, and their explicit analytical nature allows them to be analysed to extract biological models from the equations, for example by detecting higher order epistatic interactions between loci by using mixed partial derivatives.

2 Results

2.1 Generating additive, dominant and epistatic synthetic quantitative traits datasets to evaluate Symbolic Regression for phenotype prediction

In this study, we used the SR library Feyn, which is based on Qlattice [22]. Feyn has shown to have state of the art performance in the most recent SR community benchmark in 2022 [25], and the well-documented python interface provides ease of use. To evaluate the suitability of SR methods for the regression of quantitative traits from genotypes, we started by creating several synthetic datasets as combinations of various parameters. We first randomly generated SNP-array datasets containing 2000 diploid samples with 2000 observed loci. Each loci has an assigned genotype encoded as 0,1,2 to indicate respectively homozygous wild-type alleles, heterozygous mutants and homozygous mutants. We then used the Simphe [26] library to generate synthetic phenotypes from this data, considering 144 combinations of several parameters regulating inheritance, including: epistatic and dominant components (0, 1), the heritability fraction $h = \{0.3, 0.6\}$ and the number of causative loci for the phenotype (QTLs), which ranged from 2 to 50. On top of this, we also varied the number of samples considered during each benchmark (500, 1000, 2000 samples), to simulate dataset under-determination, and the number of loci (0, 100, 500, 1000, 1900) that are unrelated to the phenotype (called *decoys*, since they act as confounders for the true signal). All these combinations of parameters produced a total of 2160 benchmark settings from the initial 144. See Methods for more details.

Fig. 1 gives a visual illustration of how these synthetic phenotypes look like in a bi-dimensional setting involving two biallelic loci. The figure is arranged into two orthogonal axes, representing epistasis (horizontal x axis) and dominance (vertical y axis). The bottom left plot, in which both axes are 0, shows a perfectly additive phenotype.

We arranged epistasis and dominance as orthogonal axes because dominance can be understood as the strength of the interaction of alleles *at the same locus*, while epistasis is the interaction between alleles at *different loci*, and therefore are two separate effects. In this setting, pure additivity is not a separate inheritance mechanism, but just the absence of both dominance and epistasis between the target loci.

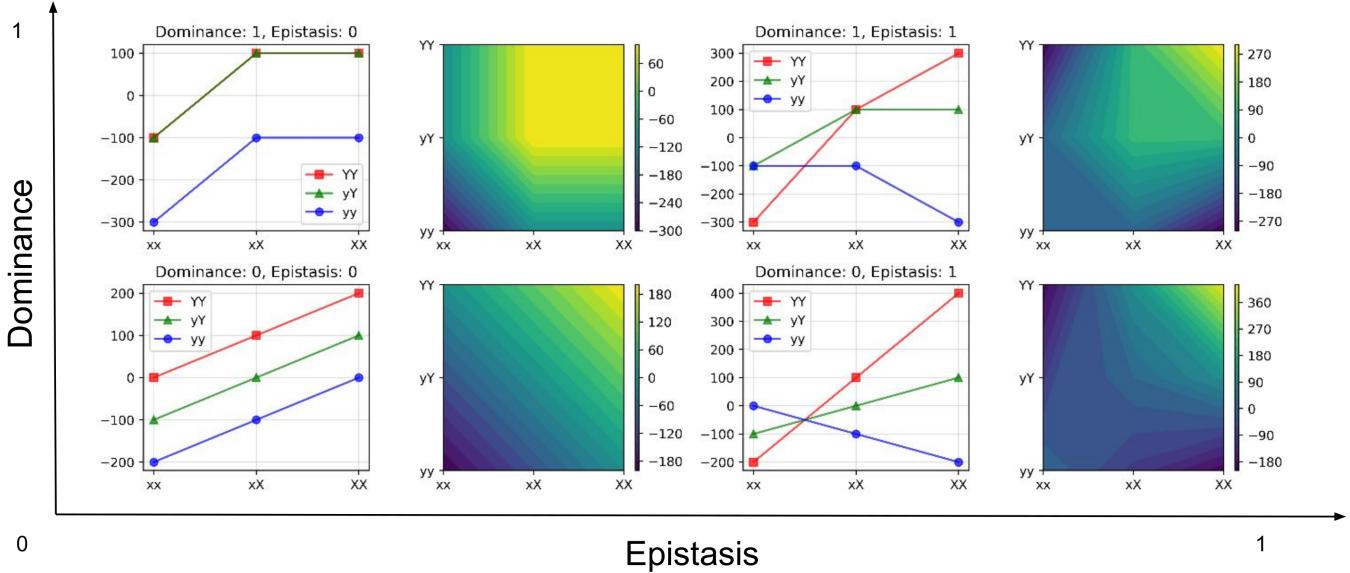


Figure 1: Figure

2.2 SR improves over conventional methods on the prediction of nonlinear oligogenic traits

We compared the Feyn SR approach with both linear and nonlinear conventional ML methods, to represent the approaches that are more commonly used in literature. As linear models, we used the Lasso and the Ridge regression methods, representing respectively L1 and L2-regularized linear regressions. As a nonlinear model, we used a standard fully-connected feed-forward Multi-Layer Perceptron Neural Network model (MLP).

As we mentioned earlier, differently from conventional ML methods, whose training searches the space of parameters of a fixed functional form, SR methods search the space of mathematical expressions, and therefore it could potentially have to deal with a much larger number of degrees of freedom. In this study, we are mainly interested in 1) the interpretability of the models we obtain and in 2) the detection of epistatic relations between the input variables (the loci). We therefore decided to artificially limit the size of the mathematical expression space that the SR algorithm needs to search, and to ensure the maximum interpretability of the resulting symbolic models, we restricted Feyn to consider only the *addition* and *multiplication* operators, which are indeed sufficient to model both additive and nonlinear interactions between features.

In Fig. 2A we show the comparison of the train/test prediction (70% train, 30% test) performances of the Feyn SR method with MLP, Lasso and Ridge. We measured the performance with the Pearson correlation coefficient, since we are comparing regression methods. The synthetic datasets shown in Fig. 2A have fixed values of 0.6 heritability for the phenotypes and 1000 decoys (irrelevant non-QTL loci in the input features). The number of samples tested is shown on the y axis of each heatmap, and the number of QTLs is shown on the x axis. The presence of epistatic and dominance effects is shown on the y axis as well.

The rightmost column of the heatmaps shows the comparison between the SR performance and the best of the other methods ($Feyn - \max(MLP, Lasso, Ridge)$) on each specific dataset. Red colors indicate situations in which Feyn outperforms all three methods, while blue colors indicate datasets in which *at least* one conventional method has a higher Pearson correlation with respect to the Feyn SR.

2.2.1 Lasso is optimal for the prediction of oligogenic additive traits, due to its linear sparse nature

From the first row in Fig. 2A we can see that in the simplest case, involving a purely additive phenotype, Lasso is the most consistent method. When very few QTLs (2-6) are involved, Feyn SR gives similar or marginally better predictions, but when more than 6 QTLs are involved, Lasso (a L1 linear regression) outperforms every other method. This is intuitively expected, since a sparse linear model is indeed the closest parametric form to the additive inheritance that has been used to generate this data. MLP is the most expressive conventional method and has the lowest performance, likely due to the small sample size with respect to the high number of trainable parameters (data underdetermination). On average, MLP performance increases with the available sample size, but remains lower than other approaches.

2.2.2 SR outperforms conventional methods for the prediction of epistatic oligogenic traits

The second and last rows of Fig. 2A show the predictions on datasets in which epistatic interactions between QTLs are involved in generating the phenotypes, corresponding to the rightmost column in the cartesian plane shown in Fig. 1. This introduces the necessity to model nonlinear interactions between the input features in order to achieve high Pearson scores. From the heatmaps in these two rows, which are summarized by the Comparison column, we can see that SR outperforms all the other predictors, up to when 26 QTLs are involved (second row), and up to around 30 QTLs when both epistatic and dominance effects are involved (last row). Linear methods (Ridge and Lasso) can only produce random predictions, since they are totally oblivious to epistatic interactions. On the other hand, the MLP, which is potentially the most expressive model in this benchmark, seems to struggle finding a meaningful convergence: the white squares indicate situations in which it was not possible to compute a meaningful Pearson correlation from its predictions, for example because the MLP was outputting the same value regardless of the input. This behavior is likely due to the high number of trainable MLP parameters ($101 + 100 \times num_{QTL}$) with respect to the relatively

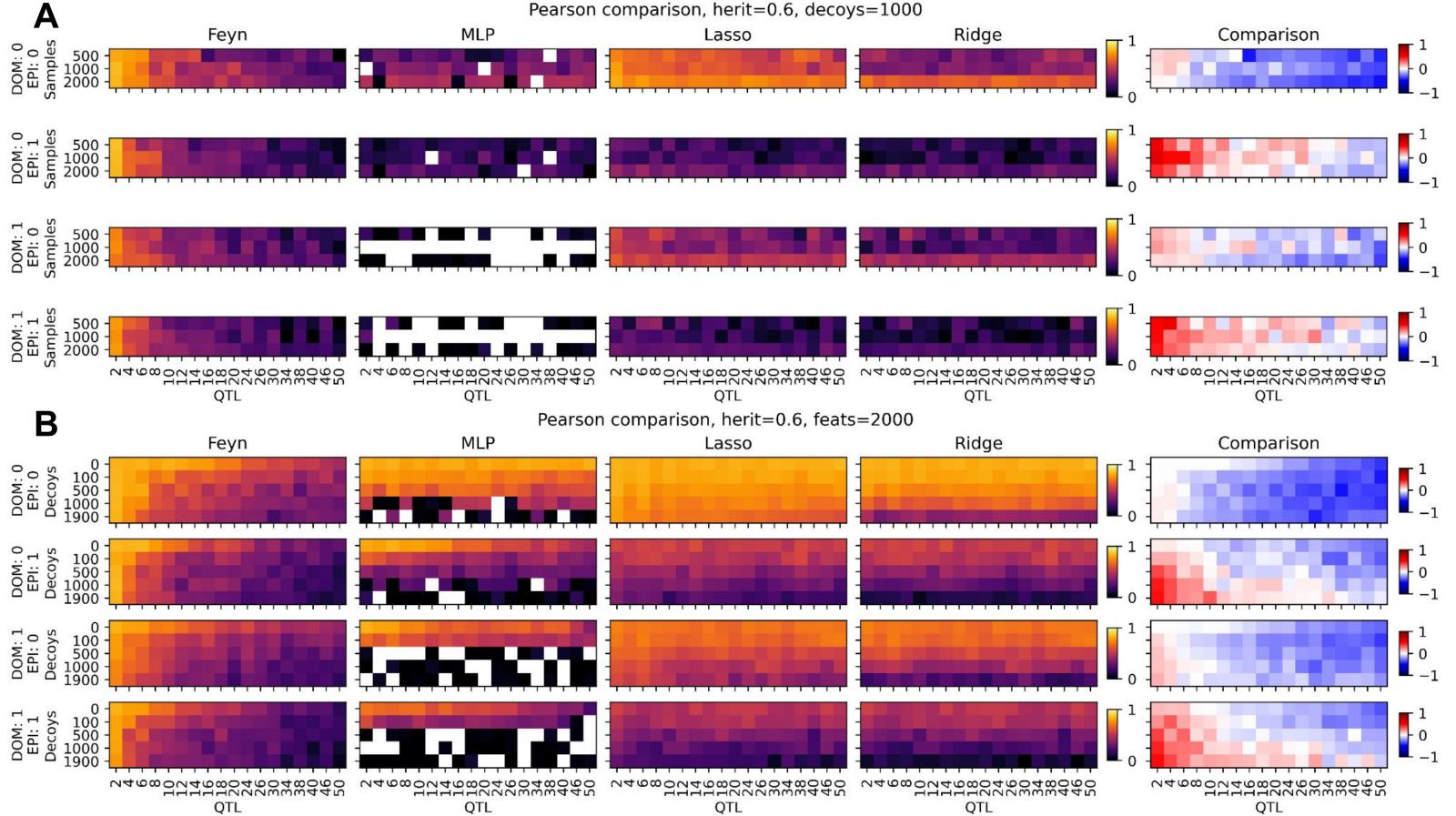


Figure 2: Figure showing the comparison of the train/test prediction performances of the Feyn SR with MLP, Lasso and Ridge. Performances are measured with Pearson correlation coefficient. Each cell in the heatmaps represents a synthetic datasets with a certain number of samples and QTLs. Each row in the A and B panels represents a different combination of Dominance and Epistasis used in generating the phenotypes. The rightmost column of the heatmaps shows the comparison between the SR performance and the best of the other methods ($Feyn - \max(MLP, Lasso, Ridge)$) on each specific dataset. Red colors indicate situations in which Feyn outperforms all three methods

few available sample. This unfavorable parameters-to-samples ratio leads to a severe under-determination scenario that is likely to prevent the MLP from reaching a proper convergence.

On the other hand, SR explores the space of possible mathematical expressions involving combinations of the addition and multiplication operators (see Methods) using a heuristic search [20]. The resulting models are ranked by their Bayesian Information Criterion (BIC) score, which balances both the model accuracy and the complexity necessary to achieve said accuracy. The BIC penalizes models with more parameters, discouraging overfitting and leading to simple models that appear to be robust to under-determined datasets.

2.2.3 Dominance effects are easier to predict than epistatic effects because they do not involve interactions between loci

In Fig. 1 we arranged epistasis and dominance as orthogonal axes because dominance indicates the interaction of alleles *at the same locus*, while epistasis is the interaction between alleles at *different loci*. As shown in the left column of Fig. 1, dominance produces a relatively slight deviation from the perfect linearity of a purely

additive phenotype, and therefore it can still be approximated by a linear model with somewhat acceptable error. Epistasis, on the other hand, 1) depends on the interactions between alleles at different loci, that can not be modeled by linear models, and 2) can have a potentially drastic effects, such as in the case of reciprocal sign epistasis (see the ($E = 1, D = 0$) plot in Fig. 1), where the change of alleles at locus Y reverses the effect of the alleles at locus X.

The third row of heatmaps in Fig. 2A shows the results obtained on purely dominant phenotypes. From the Comparison column, we can see that the SR slightly outperforms the conventional methods up to oligogenic traits involving 8 QTLs. Nevertheless, the performance of the linear models (Lasso, Ridge) are lower than in the additive case (first row), even though they exhibit a similar general pattern: Lasso performs better than Ridge and higher sample sizes leads to better performance. This is indeed likely due to the fact that a purely dominant behavior can partially be approximated by a linear model, and the additional nuances introduced by a more articulate SR model are not beneficial when more than 8 QTLs are involved in the phenotype, due to the general increase of the complexity of the task. Also in this case, the MLP struggles to learn anything, leading to various datasets on which Pearson could not be computed (white cells).

2.3 SR is more resistant to noise than conventional predictors on epistatic oligogenic traits

Fig. 2B shows the same benchmark on the 2160 combinations of parameters we used in our benchmark on the 144 synthetic phenotypes we generated with Simphe, except that this time we fixed the sample size to the maximum value (2000 samples), and we varied the number of *decoys* (non-causative loci/features) used as input. This means that the last row in each of the heatmaps in Fig. 2A corresponds to the "1000 decoys" row in each of the heatmaps in Fig. 2B. From this new figure, we can see that the significant presence of decoys in Fig. 2A was one of the reasons for the poor performance of the MLP. The MLP is in fact very dependent on the ratio between decoys and samples. It can tolerate up to < 50% decoys/samples when predicting a purely additive phenotype, but more than 5% (100/2000) is already enough to compromise the model when epistatic or dominant effects come into play.

Lasso is the best performing and more robust method in the additive case (first row in Fig. 2B), with Ridge close second, even if its performance drop when the ratio between decoys and samples reaches 95% (1900/2000).

Although SR is not very useful in the pure additive case, since Lasso and Ridge have already the optimal parametric form for this kind of problem, the Comparison heatmap (last column in the second and third rows of Fig. 2B) shows once again that, when epistatic effects come into play, there are regions of the space of possible datasets in which SR is the best performing method. In particular, this phenomenon takes the shape of a bi-dimensional phase-transition-like scenario in which SR is preferable on highly noisy data and relatively low numbers of causative QTLs.

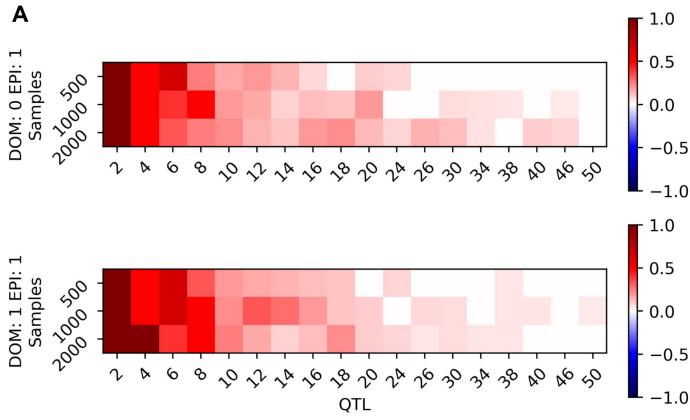
As we explained in the previous section, this effect is more prominent when epistatic effects are involved, and it is lower when only dominant inheritance is present, likely due to the univariate and quasi-linear nature of dominance.

2.4 SR can be used to detect epistatic interactions between loci

SR is an unconventional approach to ML in which instead of optimizing the trainable weights of a certain parametric form (i.e. linear combination, or MLP), heuristic searches (a genetic algorithm in the case of feyn [22]) are used to explore the space of mathematical expressions, looking for the one that best suits the problem at hand. This, coupled with a parsimony principle such as the Bayesian Information Criterion (BIC) applied as prioritization criterion in the final selection of these expressions, privileges simple and interpretable models, that can be understood by looking at their analytical equation.

These expressions can be analysed both visually and automatically. In the case of the synthetic datasets we generated with Simphe, we exploited this potential by automatically visiting the parsing tree of the

Jaccard score for epistatic pairs identification, herit=0.6, decoys=100



Jaccard score for epistatic pairs identification, herit=0.6, feats=2000

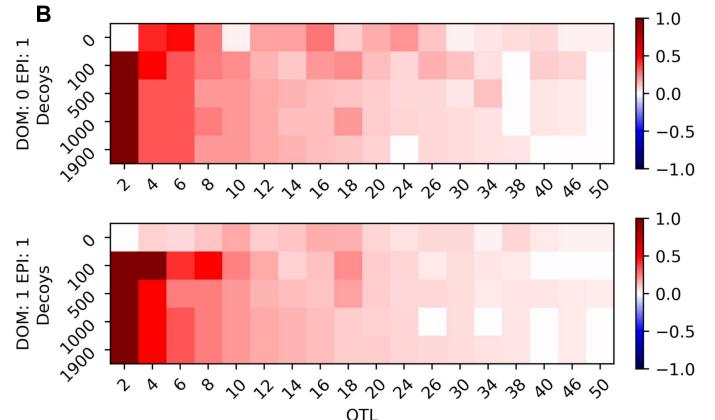


Figure 3: Figure showing that the automatic investigation of the parsing trees of the SR equations can lead to the correct identification of epistatically interacting pairs. Panel A shows these results while varying the sample size, Panel B shows the results while varying the number of decoys. We left the top left cell of the heatmaps in Fig. B blank, since it represents the trivial case in which there are only two interacting loci, and no decoys.

mathematical expressions corresponding to the models generated by Feyn, detecting every time two input loci are joined by a multiplication operator, since it indicates that the model assumes an interaction between the two. In Fig. 3 we then compared the Jaccard score of the epistatic pairs selected by Feyn with respect to the true epistatic pairs in each synthetic phenotype, with the Jaccard scores obtained by repeated random selections. In the heatmaps in Fig. 3, the value of each cell is thus calculated as $Jaccard(Feyn) - MeanJaccard(100RandomSelections)$, and red colors indicate when the ability of the SR models to identify which QTLs epistatically interact is higher than what could be expected by chance.

Fig. 3A shows the SR ability to identify epistatic interactions between loci as a function of the number of samples, while in Fig. 3B the number of samples is fixed to 2000 and what varies is the number of decoys. From the left plot we can see that increasing the sample size generally improves the detection ability. When only 2 epistatic QTLs are present, SR is always able to properly detect their interaction regardless of the number of decoys. This performance degrades with the increase of QTLs, due to the higher difficulty of the problem. When the number of epistatic QTLs reaches half the number of decoys (100), the SR ability to detect epistatic pairs is indistinguishable from random, regardless of the sample size available.

In Fig. 3B we see that at low number of QTLs (≤ 10), SR detects epistatic interactions in a robust way regardless of the number of decoys (loci not related to the phenotype). In general, the addition of decoys causes only a mild degradation of the detection ability, and it appears that the sheer number of QTLs involved is the major driver of the lower detection.

2.5 Benchmarking Symbolic Regression and conventional predictors on the Crater and Mesa fitness landscapes

So far we benchmarked SR against conventional methods on synthetic phenotypes that have been generated following clear inheritance mechanisms (see Fig. 1) generated with Simphe [26]. Another option used in the literature to capture the effect of higher order interactions between loci, as they could be observed in real data, is to assume that the observed phenotype is a nonlinear function of some underlying non-epistatic trait [27]. This approach is generally called non-specific [28] or global [29, 30] epistasis, and has, for example, been used in [23, 27].

Here, we focused on two biophysically inspired global epistasis fitness landscapes that were originally

designed to describe transcriptional regulation [23]. They are called *crater* or *mesa* landscapes due to the characteristic shape of their fitness curves.

First, we randomly generated a SNP array-like dataset containing 5000 samples with 5000 biallelic loci each. We then produced the crater and mesa fitness landscapes by selecting a random sample x_R as *reference* sequence, and assigning the fitness to all the other sample x_i as $f(d(x_R, x_i))$, where d is a distance function between the QTLs of two samples x_R and x_i , and f is a nonlinear function of this distance (see Methods for more details). To obtain the "crater" and the "mesa" models [23], we respectively used a Gaussian and a Sigmoid as nonlinear functions f , and we will indeed call these synthetic phenotypes Gaussian and Sigmoid in the rest of the paper, for clarity.

Fig. 4 shows the prediction performance in terms of Pearson correlation of SR and the MLP, Ridge and Lasso models. Each square in the heatmaps shows a different train/test setting, in which both the number of available samples (from 100 to 4000) and of causative QTLs (from 2 to 100) are varied.

As we mentioned earlier, since in this study we are investigating the ability of SR to model epistatic interactions between loci in oligogenic traits, we limited the space of the possible mathematical expressions that SR should visit to the ones composed by arrangements of the addition (to model additivity) and multiplication (to model interactions between loci) operators. To investigate to which extent the use of these two simple operators is limiting the performance of SR, when dealing with more complicated functions, like the Gaussian and the Sigmoid, in the leftmost column of Fig. 4 we also allowed Feyn (FeynGauss) to use the bivariate Gaussian and the Exponential operators for the prediction of the corresponding landscapes.

2.5.1 SR is the best predictor on the Gaussian fitness landscape

From the first row of Fig. 4 we can see that the Gaussian fitness landscape (called "crater landscape" in [23, 27]) is generally non-trivial to predict. For what concerns the Ridge and Lasso models, this is likely due to the fact that, due to its non-monotonicity, the Gaussian landscape is the regression analogous of the archetypal nonlinear separable XOR problem, and it cannot be approximated by a linear model. The MLP is able to express this function, but, likely due to the difficulty of discerning clear patterns in the data, performance degrades rapidly when more than 4 QTLs are involved. This result is in line with the findings of our recent benchmark on the ability of NN methods to identify non-linear patterns in noisy and underdetermined data, which indeed found these methods lacking [18].

SR appears to be the most robust method in this problem, as shown in the Comparison column in Fig. 4. In particular, it appears that allowing Feyn to use the Gaussian function among the operators it can use to compose symbolic expressions produces higher performance. This is due to the fact that inevitably, approximating a Gaussian with just compositions of additions and multiplications inevitably leads to higher errors, as shown in the first row of plots in Fig. 5.

2.5.2 The Sigmoid landscape can be reasonably approximated by L2 linear regression

The second row of Fig. 4 shows the analogous benchmark on the Sigmoid fitness landscape (called "mesa landscape" in [23]). In this case, performances are generally higher, likely due to the simpler, monotonic function that needs to be predicted. Interestingly, the Ridge (L2-norm regularized linear regression) and the MLP models appear to be the best performing. In particular, MLP performs best at high sample sizes and < 50 QTLs, while the Ridge is more robust even at lower samples sizes and high number of QTLs. This is an illustration of the bias-variance trade-off in ML. If enough data are available to robustly learn complex patterns, highly expressive models like MLP can benefit from their flexibility. On the other hand, in noisy conditions or data under-determination, the simplicity of linear models turns into robustness [16].

Another interesting aspect is the striking difference in performance between the Ridge and the Lasso models, which is likely due to their different regularization. The L2-norm is indeed more suitable for the prediction of a global/non-specific epistatic trait, since it indeed shrinks the weights assigned to each feature gradually, without forcing any to zero, thereby potentially retaining many features. On the other hand, the L1 norm privileges sparse solutions, in which few features are used for predictions and most of the others are silenced. While Lasso was outperforming Ridge in the sparser phenotypes generated by Simphe (See

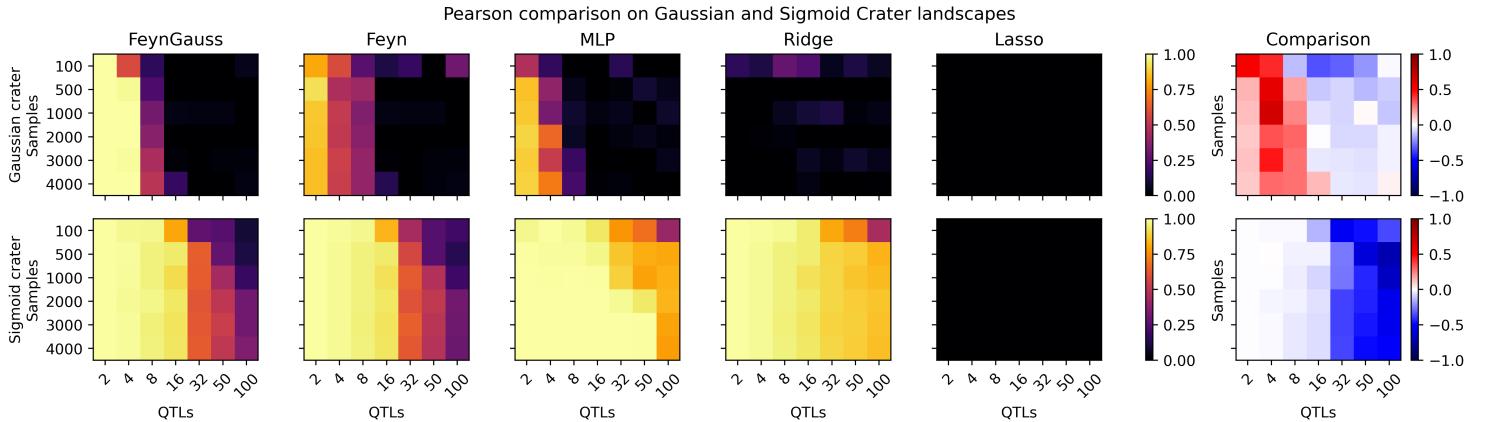


Figure 4: Figure showing the prediction performance (Pearson correlation) of FeynGauss, Feyn, MLP, Ridge, and Lasso models on the crater and mesa landscapes (Gaussian and Sigmoid). Each heatmap cell represent a train/test configurations with varying sample sizes and causative QTL counts. The rightmost column of plots shows the comparison between the best SR model and the others.

Section 2.5.1), in which it was important to detect the specific causative QTLs and silence the unrelated loci (decoys), here the phenotype is a nonlinear function of the "average similarity" of the QTLs, and the direct relationship between QTLs and phenotype is masked. As previously observed, the global epistasis models result in dense interactions between all QTLs, instead of sparse interactions between specific pairs of loci [27] as in the previous synthetic datasets.

2.6 Visualizing the global epistasis landscapes and the predictions

The global epistasis fitness landscapes allow for an intuitive 2D visual evaluation of the performance and behavior of the models [27]. In Fig. 5 we show how the Gaussian (first row of plots) and Sigmoid (second row of plots) landscapes look like. The black line corresponds to the target fitness function, while the colored lines indicate the predictions provided by the different models.

In the first two Gaussian landscape plots of Fig. 5A, we can see that the Feyn model using the Gaussian and exponential functions as well (Feyn($+^*N$), in dark green) perfectly matches the target function, while the Feyn($+$) with just addition and multiplication has a higher error. Ridge and Lasso always produce a flat line, while the MLP struggles to match the exact shape of the Gaussian curve. When higher numbers of QTLs are involved, no method can infer the correct pattern. The plots show that the MLP (blue line) tends to overfit wildly (rightmost plots of Fig. 5AI, while the SR is more conservative, producing a straight line, similarly to the linear models. This is clearly due to the BIC parsimony principle in the SR models' selection.

The second row of plots (Fig. 5B) visually shows why the Sigmoid landscape is generally easier to predict (see Fig. 4). When few QTLs are involved (2 or 4), the function is very close to a straight line. From 8 QTLs on, the sigmoid shape appears, but only the MLP is truly able to approximate it. When many QTLs are involved (> 50), all methods fail, including the MLP. In all cases, the SR fails to find the correct mathematical expression to describe a sigmoid, even when Feyn is allowed to use the exponential operator.

2.7 Using SR for the interpretable prediction of epistatic relations in the fitness landscape of Streptococcal G protein

As a last benchmark for the SR approach we propose, we tested it on the prediction of a real fitness landscape. To do so, we used the combinatorial experimental exploration [31] of an epistatic region of the protein G domain B1 (PDB: 1PGA), which is an immunoglobulin-binding protein expressed in Streptococcal

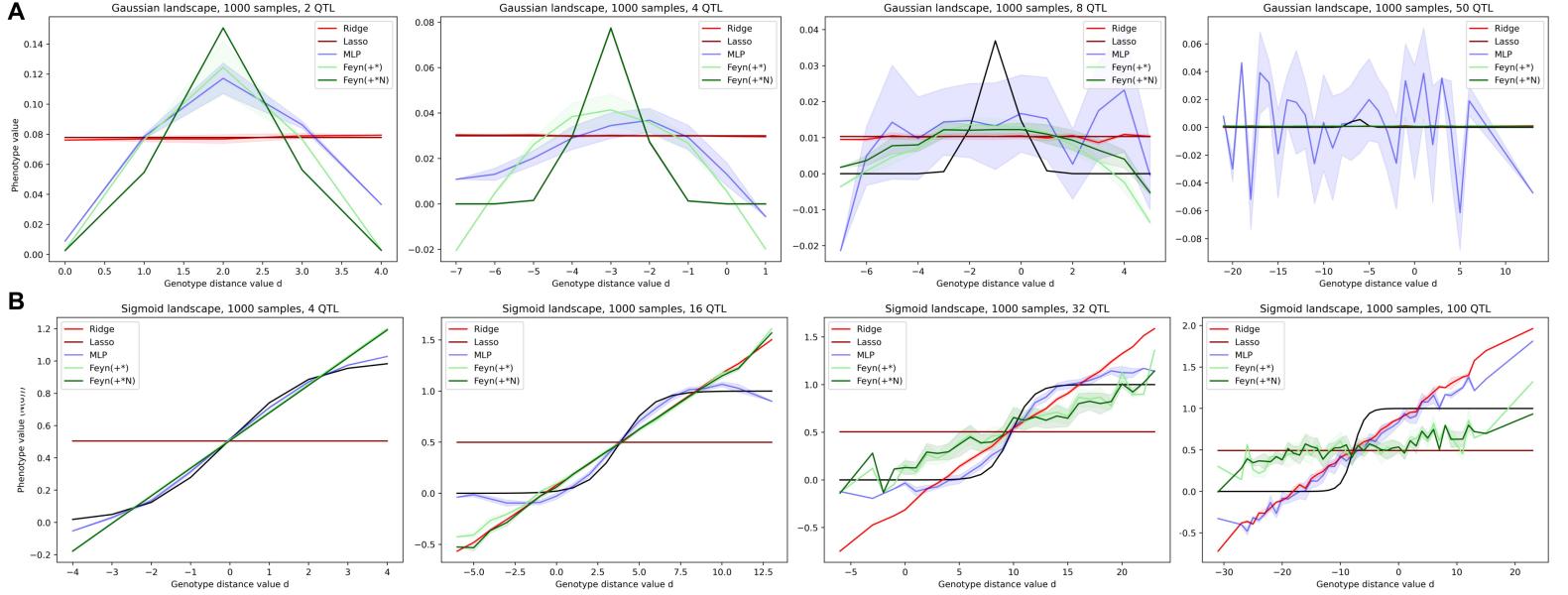


Figure 5: Figure showing an intuitive 2D visualization of the behavior of the models on the Gaussian/crater (first row of plots) and Sigmoid/mesa (second row of plots) landscapes. The black line corresponds to the target fitness function, while the colored lines indicate the predictions provided by the different models.

bacteria. This domain consists of 56 aminoacids, and in [31] the authors experimentally investigated the fitness landscape of the V39, D40, G41 and V54 residues, which are known to show epistatic interactions in GB1 (see Fig. 6A). Unlike previous studies, the authors of [31] probed the *complete* combinatorial mutational space involving these residues, for a total of $20^4 = 160000$ combinations of variants. The fitness measurements of the mutants were determined by considering protein stability (i.e. the ability to properly fold) and their binding affinity to IgG-Fc (i.e. function), by using a high-throughput screening coupling mRNA display with Illumina sequencing [31]. The fitness measurements were obtained in comparison with the wildtype sequence (VDGV).

We thus used this data to train and test our SR approach. We randomly split the data into a 70% training set and a 30% test set, and we trained Feyn for 40 epochs (39925 models explored), allowing only mathematical expressions containing addition and multiplication operators. We obtained a Pearson correlation of $r = 0.71$ and a R^2 of 0.5 on the test set. This is respectively 61% and 163% higher than the r and R^2 obtained by the best linear model (Ridge), but it is 26% lower than the Pearson correlation obtained by MLP on the same train/test splits. This indicates that a strong non-linear (i.e. epistatic) component is indeed present in this data, and that a purely linear model cannot fully model this fitness landscape.

MLP obtains very good performance for the regression of the fitness, but at the expense of 1) a higher number of parameters (8201) and 2) the lack of interpretability of the final model. On the other hand, the SR models produced by Feyn 1) have at most 87 parameters (20 dimensional embeddings for the aminoacids plus the weights associated to the maximum of 7 allowed edges in the expressions' parse tree), and 2) it is fully interpretable due to the explicit nature of the analytical equations.

The best performing SR equation is as simple as $y = p_0 p_2 (p_2 + p_3)$, where y is the regression output (the fitness) and $p_0 \dots p_3$ are the variables indicating the four residues at positions 39, 40, 41, and 54 of the target protein. Other proposed models are $p_2^2 p_0 p_3$ and $p_2 p_0 p_3^2$. The prevalence of multiplication operations already highlights the predominance of epistasis in this landscape. We show the visual representation of two of these equations in 6E. The full list of the 25 final models obtained from Feyn is available in Suppl. Mat. File S1.

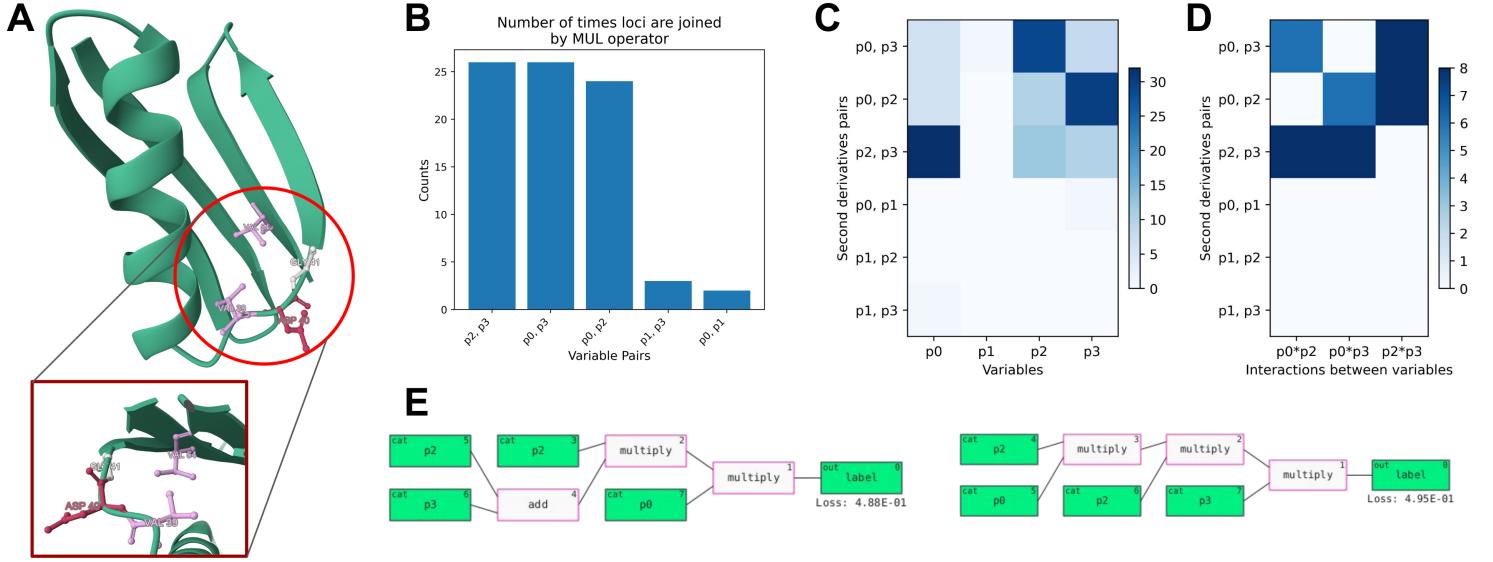


Figure 6: **A:** Illustration of the Protein G Domain B1 chain (PDB:1PGA), and the positioning of the 4 loci investigated in [31]. **B:** Plot showing the tendency of the pairs of loci to be joined by a multiplication in the parse trees of the top performing 25 models produced by Feyn for the prediction of the fitness landscape. **C,D:** Heatmaps showing the occurrences of the variables (x axis) in the mixed partial derivative equation between all the possible pairs of loci (y axis), indicating the possibility of higher-order interactions **E:** Visual depiction of the parse trees of two of the best performing models produced by Feyn.

2.7.1 Interpreting SR equations by analysing their parse tree

We then exploited the explicit nature of these equations to perform further analyses, complementing the results already presented in [31]. For example, experimental results show that the epistatic interactions are enriched among specific amino acid combinations of site 39, 41 and 54 [31]. This is also in line with what is visible from the protein structure (see Fig. 6A), where the side chains of residues 39, 41, and 54 can physically interact with each other [31]. We therefore analysed whether our models were able to pick up this. During training, we allowed only addition and multiplication operators, in order to ensure the complete interpretability of the predictions. Addition is used to model additive effects, while the epistatic interactions are modeled by the multiplication. We used the Sympy [32] library to automatically parse the 25 best SR models provided by Feyn, counting how often the variables representing the 4 loci were arguments of the same multiplication. The barchart in Fig. 6B shows that p_2, p_3 and p_0 are indeed vastly more likely to be joined by multiplications than pairs involving p_1 (residue 40), indicating that the models are able to detect the different levels of epistatic interactions among loci from the data.

2.7.2 A partial mixed derivatives-based analysis to identify epistatic interactions between loci

Moreover, the analytical nature of the SR models allows also more advanced forms of analysis. For example, given a SR model f representing a fitness landscape, a way to quantify the degree of epistatic interaction between loci i, j that has been proposed in [10] consists of computing the mixed partial derivative of f with respect to i, j : $D_{ij}^2 = \frac{\partial^2 f}{\partial i \partial j}$. This value can then be normalized by the partial derivatives $D_i = \frac{\partial f}{\partial i}$ of the loci involved, which represent their independent contributions to the phenotype change [10]. The mixed partial derivative D_{ij}^2 measures how the rate of change of f with respect to i depends on j variable. In other words, it quantifies the interaction between two variables in influencing the function's behavior: if f describes a

surface, the mixed partial derivative describes the curvature of the surface in a direction influenced by both i and j . In the context of fitness landscapes or genetic interactions, a mixed partial derivative can represent epistasis: how the fitness effect of a mutation at one locus depends on the state of another locus [10]. For example, a non-zero D_{ij}^2 indicates that the two loci interact epistatically, while a zero mixed partial derivative implies that the loci contribute additively (independently) to fitness.

Moreover, also the analysis of the analytical form that D_{ij}^2 assumes can be interesting to extract biological knowledge from the SR models. For example, if $f(i, j, z) = zij$, the $D_{ij}^2 = z$, indicating that 1) z directly modulates the interaction strength between i, j and the direction of the epistasis as well: $z > 0$, we have positive epistasis, and vice versa. If the learned model has the form $f(i, j, z) = ij + zl^2j$ then $D_{ij}^2 = 1 + 2zx_i$, showing a more indirect influence of z . In this case, if $z = 0$ (e.g. z corresponds to the wildtype residue), the interaction between z and (i, j) is purely additive (while i, j are still in epistatic interaction, since $D_{ij}^2 > 0$). If $z \neq 0$, the interaction between i, j becomes also dependent on zi , indicating a possible higher-order interaction between these 3 loci.

In the case of the $y = p_0p_2(p_2 + p_3)$ model we learned to represent the fitness landscape of the protein G domain B1, we obtained the following mixed partial derivatives between the p_0, p_2 and p_3 loci: $D_{p_0,p_3}^2 = 10 - p_2$, $D_{p_0,p_2}^2 = -p_2 - p_3 + 4$ and $D_{p_2,p_3}^2 = 10 - p_2$. In these equations, we omitted the coefficients of the variables (except the sign) to avoid cluttering and help readability. Following the example above, in all three cases we observe that there is positive epistasis between the differentiated variables (due to the fixed constant present in their derivative), with often the remaining locus that is present as a third variable that acts as modifier for their interaction.

If we consider the second best performing model, $y = pop_2^2p_3$, the mixed derivatives are the following: $D_{p_0,p_3}^2 = -80p_2^2 - 247p_2 - 21$, $D_{p_2,p_3}^2 = -p_0p_2 - 7p_0 - 7p_2 - 0.71$ and $D_{p_0,p_2}^2 = -80p_2p_3 + 5.6p_2 - 7p_3 + 0.5$. In this case, we kept also the coefficients, with some rounding. We see that the epistatic interaction D_{p_0,p_3}^2 is indeed indirectly but nonlinearly modulated by the residue at position 41 ($-p_2^2$), and that in general a third variable is again often involved at least as a linear modifier for each pair's interaction. The full list of mixed partial derivatives we computed is available in Suppl. Mat. File S2.

Following this reasoning further, we used Sympy to compute the mixed partial derivatives for each of the $4 \times 3/2 = 6$ possible pairs of loci investigated on protein G1, and for each of the final 25 models produced by Feyn. In order to obtain a more general understanding from these equations, we then performed two types of analysis, which are shown in Fig. 6C and D. First, we parsed the $D_{p_ip_j}^2$ equations to identify which "third locus" played the role of the above-mentioned z variable given each pair of i, j loci, thus potentially indicating higher-order epistatic interactions between them. Fig. 6C shows that p_1 is never involved in such derivatives, while p_0 is the most frequent term in the derivatives describing the relation between p_2 and p_3 . In general p_0, p_2 or p_3 always appear as the third locus when the derivative of the other two is computed, while p_1 is never present, indicating that the first 3 loci are indeed more tightly interconnected epistatically [31].

In Fig. 6C, instead of looking at single variables (loci), we look at the most frequent multiplicative terms involving at least two loci in the same $D_{p_ip_j}^2$ derivatives, indicating *how* the residues outside the differentiated pair modulates their relation.

3 Conclusions

The goal of this study is to demonstrate the potential of Symbolic Regression (SR) for modeling and interpreting epistatic oligogenic traits, in comparison with more conventional approaches. Using the Feyn Python library, which allows an out-of-the-box access to SR, we applied this uncommon technique to both synthetic and real-world datasets, benchmarking its performance against traditional linear models and nonlinear machine learning methods. Our results highlight that SR excels in scenarios with relatively few causative loci (< 50) and even at low sample sizes, providing parsimonious models that are both robust and interpretable. Unlike black-box ML models, SR generates explicit mathematical expressions that not only achieve competitive predictive performance but also enable detailed analysis of epistatic interactions, such as detecting higher-order interactions through mixed partial derivatives.

Furthermore, our application to the GB1 protein fitness landscape demonstrated that SR models, despite their simplicity, retain substantial predictive power while offering significant insights into underlying biological mechanisms. These findings underscore the value of SR as a powerful tool for unraveling complex genotype-phenotype relationships.

4 Methods

4.1 Datasets

To assess the effectiveness of SR methods in regressing quantitative traits from genotypes, we generated two collections of synthetic datasets, incorporating combinations of parameters modulating the inheritance mechanisms, the number of samples and the level of noise in these datasets. In total, we used two collections of synthetic datasets, and an experimental one, and we describe them in the following section.

4.1.1 Generating synthetic SNP array datasets and associate phenotypes with Simphe

We randomly generated SNP-array-like data containing 2000 diploid samples with 2000 observed loci. Each loci has an assigned genotype encoded as 0,1,2 to indicate respectively homozygous wild-type alleles, heterozygous mutants and homozygous mutants. We then used the Simphe [26] library to generate synthetic phenotypes from this data. To do so, we considered 144 combinations of parameters regulating 1) the type of inheritance, divided into epistatic and dominant components (possible values {0,1}), 2) the heritability fraction $h = \{0.3, 0.6\}$ and 3) the number of causative loci for the phenotype (QTLs), which ranged from 2 to 50.

Additionally, we varied the number of samples used in each benchmark (500, 1000, 2000 samples) to simulate dataset under-determination and adjusted the number of loci unrelated to the phenotype (0, 100, 500, 1000, 1900), referred to as decoys since they act as confounders for the true signal. These parameter combinations resulted in a total of 2160 benchmark settings derived from the original 144.

4.1.2 Generating the crater and mesa fitness landscapes benchmarks

We randomly generated a SNP array-like dataset containing 5000 samples with 5000 biallelic loci each, using the same {0,1,2} encoding. We generated the crater and mesa fitness landscapes by selecting a random sample x_R as the *reference* sequence and assigning fitness to each other sample x_i as $f(d(x_R, x_i))$, where d represents a distance function between the QTLs of x_R and x_i , and f is a nonlinear function of this distance. Since in our case both x_R and x_i are vectors representing the causative loci, we opted for the simplest possible distance function, namely the element-wise difference between them. This is computed as $d = \sum_i^L x_R^{(i)} - x_i^{(i)}$, where i is the index of each position and L is the total number of causative loci. In the case of the Gaussian (crater) landscape, $f(x_i|x_R) = \exp(-(d(x_i, x_R) - \hat{d})^2)/5\sigma_{\hat{d}}^2$, where \hat{d} and $\sigma_{\hat{d}}^2$ are respectively the mean and the variance of the distances obtained on the entire dataset. In the case of the Sigmoid (mesa) landscape, $f(x_i|x_R) = 1/(1 + \exp(\hat{d} - d(x_i, x_R)))$.

We computed several synthetic datasets by varying the number of samples ($\{100, 500, 1000, 2000, 3000, 4000\}$) and the number of QTLs ($\{2, 4, 8, 16, 32, 50, 100\}$), for a total of 42 combinations of parameters for each type of landscape (84 datasets in total).

4.1.3 Experimentally determined fitness landscape of 4 epistatic loci of Streptococcal g protein

To evaluated the SR models we propose on a real-life dataset, we used the data produced in [31], in which the authors conducted a combinatorial experimental exploration of an epistatic region of the protein G domain B1 (PDB: 1PGA), which is an immunoglobulin-binding protein expressed in Streptococcal bacteria.

This domain consists of 56 amino acids, and in [31] the authors experimentally investigated the fitness landscape of the V39, D40, G41, and V54 residues. These 4 residues contain 12 of the strongest 20 epistatic

interactions in GB1 (see Fig. 6A). This dataset contains the complete combinatorial mutational space involving these residues, for a total of $20^4 = 160000$ combinations of variants. The fitness measurements of the mutants were determined by considering protein stability (i.e., the ability to properly fold) and their binding affinity to IgG-Fc (i.e., function), using a high-throughput screening coupling mRNA display with Illumina sequencing [31]. The fitness measurements were obtained in comparison with the wildtype sequence (VDGV). More details can be found in [31].

4.1.4 Implementation details of the Machine Learning models, the performance evaluation and the equations' processing

The implementations of the Multi Layer Perceptron, Ridge regression and Lasso models were taken from the scikit-learn library [33]. We used the default parameters for the Ridge and Lasso models. In the MLP, we used the default bnumber of hidden layers (1) and of hidden neurons (100). To improve the performance on regression, we used Tanh activations instead of ReLU, since the former is more suitable for regression. We also used an adaptive learning rate, 400 iterations and initial learning rate of 0.02.

As SR method, we used the Feyn [20, 21], which is based on Qlattice [22]. To train our SR models, we used the `q1.auto_run` function, specifying "regression" as prediction type. To ensure the feasibility of the computations, we used 10 epochs on the 2160 datasets generated with Simphe, 25 epochs on 84 the crater datasets, and 40 epochs on the GB1 dataset. We set the maximum complexity of the Feyn expression to $c = 2n - 1$, where n is the maximum number of QTLs ($c = 99$ for Simphe, $c = 199$ for the crater data, and $c = 7$ on GB1). We ran Feyn on an 11th gen Intel i9-11950H CPU, with 16 cores and 2.6 Ghz. We used the already implemented Bayesian Information Criterion (BIC) for the model selection, and we allowed only addition and multiplication as operators, except where is specified otherwise.

We used the Sympy [32] library to automatically parse the SR equations we wanted to analyse. We used the Pearson correlation as main evaluation criterion, since we are dealing with regressions.

Acknowledgements

DR is grateful to Anna Laura Mascagni for the constructive discussion. DR is grateful to N. Nazzicari and F. Biscarini for the help understanding the code and the data associated to their article [34], and for sharing additional scripts. DR is grateful to Kevin Broløs for the help understanding the details of the Feyn library.

Funding

DR is funded by a CNRS Chaire de Professeur Junior fellowship.

Competing interests

The authors declare no competing interests.

Data availability

The code and the publicly available data used in this article are available at <https://bitbucket.org/eddiewrc/srepistasis/src/main>. The GB1 fitness landscape experimental data are available from [31].

Author contributions

DR conceived the experiments. DR developed the methods. DR wrote the manuscript.

References

- [1] Emidio Capriotti, Kivilcim Ozturk, and Hannah Carter. Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, page e1443, 2018.
- [2] Holger Fröhlich, Rudi Balling, Niko Beerewinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H Maathuis, Yves Moreau, Susan A Murphy, Teresa M Przytycka, et al. From hype to reality: data science enabling personalized medicine. *BMC medicine*, 16(1):1–15, 2018.
- [3] Pauline C Ng, Samuel Levy, Jiaqi Huang, Timothy B Stockwell, Brian P Walenz, Kelvin Li, Nelson Axelrod, Dana A Busam, Robert L Strausberg, and J Craig Venter. Genetic variation in an individual human exome. *PLoS genetics*, 4(8):e1000160, 2008.
- [4] Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681, 2013.
- [5] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nat Rev Methods Primers*, 1(1), 2021.
- [6] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.
- [7] Greg Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145, 2012.
- [8] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.
- [9] Örjan Carlborg and Chris S Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5(8):618–625, 2004.
- [10] Arnaud Le Rouzic. Estimating directional epistasis. *Frontiers in genetics*, 5:198, 2014.
- [11] Ciaran Michael Kelly and Russell Lewis McLaughlin. Comparison of machine learning methods for genomic prediction of selected arabidopsis thaliana traits. *Plos one*, 19(8):e0308962, 2024.
- [12] Trudy FC Mackay. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics*, 15(1):22–33, 2014.
- [13] Marleen Balvert, Johnathan Cooper-Knock, Julian Stamp, Ross P Byrne, Soufiane Mourragui, Juami van Gils, Stefania Benonisdottir, Johannes Schlüter, Kevin Kenna, Sanne Abeln, et al. Considerations in the search for epistasis. *Genome Biology*, 25(1):296, 2024.
- [14] Nicholas J Wald and Robert Old. The illusion of polygenic disease risk prediction. *Genetics in Medicine*, 21(8):1705–1707, 2019.
- [15] Cathryn M Lewis and Evangelos Vassos. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):1–11, 2020.
- [16] Nora Verplaetse, Antoine Passemiers, Adam Arany, Yves Moreau, and Daniele Raimondi. Large sample size and nonlinear sparse models outline epistatic effects in inflammatory bowel disease. *Genome Biology*, 24(1):224, 2023.

- [17] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.
- [18] Antoine Passemiers, Pietro Folco, Daniele Raimondi, Giovanni Birolo, Yves Moreau, and Piero Fariselli. A quantitative benchmark of neural network feature selection methods for detecting nonlinear signals. *Scientific Reports*, 14(1):31180, 2024.
- [19] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1):2, 2024.
- [20] Kevin René Broløs, Meera Vieira Machado, Chris Cave, Jaan Kasak, Valdemar Stentoft-Hansen, Victor Galindo Batanero, Tom Jelen, and Casper Wilstrup. An approach to symbolic regression using feyn. *arXiv preprint arXiv:2104.05417*, 2021.
- [21] Niels Johan Christensen, Samuel Demharter, Meera Machado, Lykke Pedersen, Marco Salvatore, Valdemar Stentoft-Hansen, and Miquel Triana Iglesias. Identifying interactions in omics data for clinical biomarker discovery using symbolic regression. *Bioinformatics*, 38(15):3749–3758, 2022.
- [22] Casper Wilstrup and Jaan Kasak. Symbolic regression outperforms other models for small data sets. *arXiv preprint arXiv:2103.15147*, 2021.
- [23] Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC evolutionary biology*, 4:1–12, 2004.
- [24] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.
- [25] Fabrício Olivetti de França, Marco Virgolin, Michael Kommenda, Maimuna S Majumder, M Cranmer, Guilherme Espada, Leon Ingelse, Alcides Fonseca, Mikel Landajuela, B Petersen, et al. Interpretable symbolic regression for data science: analysis of the 2022 competition. *arXiv preprint arXiv:2304.01117*, 2023.
- [26] B. Jiang, B. Pütz. Simphe: Tools to simulate phenotype(s) with epistatic interaction. <https://github.com/beibeij/SimPhe>, 2018.
- [27] Juannan Zhou and David M McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nature communications*, 11(1):1782, 2020.
- [28] Zachary R Sailer and Michael J Harms. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics*, 205(3):1079–1088, 2017.
- [29] Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, 2018.
- [30] Sergey Kryazhimskiy, Daniel P Rice, Elizabeth R Jerison, and Michael M Desai. Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science*, 344(6191):1519–1522, 2014.
- [31] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife*, 5:e16965, 2016.
- [32] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.

- [33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- [34] Nelson Nazzicari and Filippo Biscarini. Stacked kinship cnn vs. gblup for genomic predictions of additive and complex continuous phenotypes. *Scientific Reports*, 12(1):19889, 2022.