bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hypothesis-driven interpretable neural network for interactions between genes

Shuhui Wang[a,b], Alexandre Allauzen[a,b], Philippe Nghe[a], and Vaitea Opuu[a,‡]

[a]Laboratoire de Biophysique et Evolution, UMR CNRS-ESPCI 8231 Chimie Biologie Innovation, PSL University, Paris, France,
[b]LAMSADE, Universite Paris-Dauphine, PSL University, Paris, France
[‡]Contact email: vaitea.opuu@espci.fr

Abstract

Mechanistic models of genetic interactions are rarely feasible due to a lack of information and computational challenges. Alternatively, machine learning (ML) approaches may predict gene interactions if provided with enough data but they lack interpretability. Here, we propose an ML approach for interpretable genotype-fitness mapping, the Direct-Latent Interpretable Model (D-LIM). The neural network is built on a strong hypothesis: mutations in different genes cause independent effects in phenotypes, which then interact via non-linear relationships to determine fitness. D-LIM predicts genotype-fitness maps for combinations of mutations in multiple genes with state-of-the-art accuracy, showing the validity of the hypothesis in the case of a deep mutational scanning in a metabolic pathway. The hypothesis-driven structure of D-LIM offers interpretable features reminiscent of mechanistic models: the inference of phenotypes, fitness extrapolation outside of the data domain, and enhanced prediction in low-data regimes by the integration of prior knowledge.

1 Background

Genomic mutations induce changes in fitness or the degree of adaptation of one organism to its environment. Mutations first affect certain parameters, *e.g.* the stability of specific proteins. Then, the genes interact via metabolic or regulatory networks, ultimately determining the fitness of organisms. Typically, studies of genotype-to-fitness maps are done using large-scale experimental screens called deep mutational scans[11, 12], which consist of mutating genes and measuring the resulting fitness. With the progress of DNA-editing technologies combined with next-generation sequencing, these experiments can yield as many millions of fitness measurements [6]. In addition to its critical importance in the understanding of molecular adaptations, genotype-to-fitness maps are central to biomedical research in genetic diseases, the spread of infectious diseases, drug resistance, and drug target discovery.

Mathematical models of genotype-fitness maps are of the form $F(x)$, where the fitness F is a function of the mutations x [1]. Mechanistic models typically consider two distinct levels g and P so that $F = f(P(x))$. Indeed, mutations in a gene are first associated with parameters or phenotypes P, for example, the affinity of an enzyme for its substrate. These parameters are determined by how biomolecules fold and bind. They cannot yet be predicted from first principles, and ML approaches do not yet reliably predict the effects of mutations on biomolecular functions. Thus, in practice, P is inferred by fitting experimental measurements. Then f is chosen among a family of parametric models, based on our knowledge of detailed biochemical processes, namely metabolic networks, regulatory networks, or any other physical interaction that is already known [13].

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For instance, in the study by Kemble *et al* [5], the authors performed a mutational scan of 2 genes interacting via a metabolic pathway. They derived a mechanistic model $f(A, B)$ based on Michaelis-Menten kinetics and the structure of the pathway, where two enzymatic parameters, A and B, control the fitness, supposing that the gene coding for the first (resp. second) enzyme only affects the value of A (resp. B). Because the choice of model family dictates the role of each parameter, and these parameters are directly associated with a gene, this model is interpretable. However, to find the parameters as a function of mutations, the authors used 10^8 Monte Carlo optimization runs to find the most consistent mapping, under the constraint of the family f. While being highly interpretable, finding the correct set of hypotheses—which parametric family to use—and formalizing them into a mathematical model is challenging. Furthermore, there is no evident framework to optimize their parameters. Thus, this approach cannot be generalized to other similar experiments, especially in the context of larger or lesser characterized gene networks.

In contrast, Machine learning (ML) does not necessitate biological hypotheses to construct models. Modern approaches, such as neural networks (NNs), capitalize on the increasing volume of data, resulting in a significant enhancement in prediction accuracy, particularly in the context of experimental mutational scanning of proteins [11]. Indeed, NNs are universal approximators [4], which means that, instead of manually selecting the parametric family, the learning process autonomously determines the choice of $f(h)$, where h is a multi-dimensional latent vector representation representation of the genotype. Furthermore, parameter optimization is highly efficient. However, interpreting how the representation influences the predictions and its relationship with phenotypes is challenging. In ML, linear regression models are considered to be the most interpretable, but they offer limited predictive power [8], especially in the context of interactions between mutations, which are non-linear.

A recent trend has focused on making machine learning (ML) models more interpretable while maintaining their predictive power. In the context of deep mutational scan experiments, the LANTERN approach constrains the latent representation of mutations, represented by a vector h_i, to be additive, thereby facilitating the interpretation of the model parameters, where two vector representations pointing in the same direction should have similar effects, while the magnitude of these vectors informs on the strength of the mutation [15]. Nevertheless, the specific role of the dimensions composing the coordinates h_i remains unclear regarding predictions or phenotypes. From a biophysical perspective, Tareen et al. introduced MAVE-NN, which simplifies the development of mechanistic models for genotype-phenotype maps. In MAVE-NN [14], users specify the type of map—ranging from additive to NN-based—between the genotype and a latent phenotype. Subsequently, the inferred latent phenotype is transformed into fitness through a stochastic process. As a result, while the latent phenotype is never directly measured, it can be interpreted as true phenotypes. This approach presupposes that a single phenotype is influenced by all genotypes. Faure and Lehner further developed this line of reasoning by employing multiple latent phenotypes [2]. In a different approach, GenNet introduced an interpretable NN designed to study systems of genes or metabolic pathways [16]. Its architecture is informed by a priori biological knowledge, with neuron connectivity defined by gene annotations, pathways, cellular expression, and tissue expression. However, this method necessitates rich information about the system under study.

Here, we present a hypothesis-driven model for gene-gene interactions, which learns from genotype-to-fitness measurements and infers a genotype-to-phenotype and a phenotype-to-fitness map. The model uses a neural network f built upon the hypothesis $g_i \rightarrow \varphi_i$ that mutations g_i do not interact when determining the inferred phenotypes φ_i, but this phenotypes interact via a non-linear phenotype-to-fitness relationship $f(\varphi_1, \varphi_2, ...)$. Thus, our model encodes a strong hypothesis on the structure of mechanistic relationships, namely which gene affects which phenotype, but without requiring any further *a priori* knowledge, such as the values of the phenotypic parameters involved in these relationships. However, our inferred phenotypic parameters φ eventually get measured experimentally, allowing us to map our inference through a measurement process $\omega(\varphi) = \phi$, where ϕ is a true measured phenotype. The ability of the model to fit the data under such constraint informs us on the validity of our hypothesis while inferring the phenotype-to-fitness relationship, the phenotypic values, and the mutations corresponding to these

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

values.

We first show our model reaches state-of-the-art accuracy in epistasis prediction for mutational scanning experiments on multiple genes, despite imposing stronger constraints than existing models. Due to its architecture, the model allows the inference of phenotypic proxies, without requiring any sampling process. We then show that these values correspond to actual phenotypic values, via a non-linear 1-dimensional relationship. As the hypotheses built in our model should better constrain extrapolation, we test the prediction of novel genetic interactions outside of the range of mutations already explored. Finally, we propose a general strategy to incorporate a priori knowledge, even qualitative, to improve the model quality in the low data regimes.

2 Results

2.1 D-LIM architecture

The D-LIM architecture is a neural network (NN) designed to learn genotype-fitness maps from a list of genetic mutations and their associated fitness when distinct biological entities have been identified as meaningful. The hypothesis that drives the model architecture is based on this identification, as described below. From now on, for simplicity, we denote these entities as 'genes', although they could also be groups of genes or sectors within a gene (Fig 1A). The model comprises three levels of description: mutations, phenotypes, and fitness. Phenotypes are understood here in a very general sense as fitness-determining parameters, which could be low-level molecular properties (e.g. binding), as well as higher-level measurable properties (e.g. a flux). The space spanned by the phenotypic values is termed the phenotype space.

The model comprises a genotype-phenotype map and a phenotype-fitness map. The genotype-phenotype map encodes the hypothesis that mutations participating in different genes do not interact when determining phenotypes (note however that mutations affecting the same gene may interact while determining their associated phenotype). In the network, this corresponds to connecting each mutation to a single phenotypic value (Fig 1B). The genes, the ownership of mutations to genes, and the association between genes and phenotypes are the choices of the user. The phenotype-fitness map is automatically determined using an NN (Fig 1B), leading to a relationship of the form $\widehat{F} = f(\varphi_1, \varphi_2, ...)$. The fitting process consists of initializing phenotype values φ_i randomly, and then training the NN using the Adam optimizer [7] in the Pytorch deep learning framework [10]. To account for noise in the fitness measurement, we used a Negative Gaussian Log-likelihood loss function with an uncertainty level σ [7]. D-LIM determines the phenotypes during the training process while it learns to predict the fitness (Fig SI2). A phenotype φ inferred during this process should be considered as a proxy of the actual biological phenotype ϕ, ϕ being a smooth function of φ. As a result of the training, each mutation is assigned a specific value of its associated phenotype φ_i.

On the one hand, due to the gene-to-phenotype constraints, D-LIM is more rigid than existing ML approaches, but this trades for increased interpretability akin to mechanistic models. For instance, in LANTERN, mutations are represented in an unconstrained latent space. Thus, the effect of mutations may span multiple dimensions, without being linked to any specific biological entity (Fig 1C). There is also no direct interpretation of how phenotypes determine fitness. On the other hand, D-LIM uses a NN training approach, which facilitates model development as compared to mechanistic models. Indeed, there is no need to choose the phenotype-to-fitness map f in a family of parametric models, and knowledge of biochemical details is not required. Additionally, the inference of phenotypes results from the backpropagation process, which is highly efficient compared to the sampling process which would be required to fit a mechanistic parametric model (as in Kemble et al. [5]).

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2.2 Predictive constrained architecture

The foundational hypothesis of our model is that mutations in different genes do not interact while determining phenotypes. Being able to fit the model thus informs us of the validity of our hypothesis for the system under study. On the contrary, the inability to fit would reveal the presence of strong interactions between genes at the genotype-phenotype level.

To illustrate this, we applied D-LIM to a published mutation-fitness dataset [5]. This dataset is an ideal case study because the phenotypes are well-identified but their values have not been measured directly. This system involves two genes in the metabolic pathway of L-arabinose utilization: L-arabinose isomerase (AraA) and L-ribulokinase (AraB) (See Fig 2AA). These enzymes convert L-arabinose into L-ribulose-5-phosphate, which is critical for cellular growth. Mutations were systematically introduced in the promoter regions of both AraA and AraB genes, enabling comprehensive analysis of individual and combined genetic perturbations, covering a total of 111 mutations per gene promoter. The measurements were performed in three distinct environments (Env1, Env2, Env3) that differ by the concentration of inducers promoting the synthesis of both enzymes. Env1 had inducers for both genes, while Env2 and Env3 had respectively 5 ng and 200 ng of the araA inducer but without the araB inducer. The data consists of 1369 mutation combinations and the resulting fitness (measured as growth). The phenotypes are the activity of the enzymes AraA and AraB, which are not measured.

We used D-LIM to predict fitness using two phenotypes φ_1 and φ_2, which correspond to enzymes AraA and AraB, respectively. To assess the predictive accuracy of D-LIM, we trained the model on 70% of the data and computed Pearson's correlation with the remaining 30% of the data. In Env1, D-LIM predictions showed good agreement with measured fitness with Pearson's correlation $\rho = 0.98$ with pvalue 10^-5. The phenotype-to-fitness map fitted by D-LIM is smooth with a single maximum, where fitness partially correlates with the phenotypes taken alone:$\rho = -0.84$ for φ_1 and $\rho = 0.43$ for φ_2 with both pvalue 10^-5 (Fig 2Bb-d). This indicated that both genes influence the measured fitness. In Env2, the prediction agreed similarly well with the measurements ($\rho = 0.99$); however, while the landscape predicted was as smooth as in Env1, fitness only correlated with φ_1 ($\rho = 0.96$,pvalue 10^-5) but not φ_2 ($\rho = -0.07$, pvalue = 0.02), see Fig 2Be-h. Because no AraB inducer was used in Env2, we only expected basal effects which do not vary between mutants, which is consistent with our findings. Next, we used D-LIM to map mutations to epistasis. Our approach displayed similar performances ($\rho = 0.98$); however, the landscape spanned by the phenotype space was qualitatively different with non-monotonicity. In the landscape, epistasis varies along both inferred phenotypes (as fitness), but the correlations with the inferred phenotypes are lower ($\rho = 0.61$ and $\rho = 0.21$ for φ_1 and φ_2 respectively, with both pvalue 10^-5, see Fig 2C).

Does imposing the constraint of independent phenotypes result in a trade-off with predictive accuracy? To explore this, we compared D-LIM to three analogous models that do not impose constraints on their latent spaces: a) LANTERN, which employs a Gaussian process for fitness prediction, and b) the Additive Latent Model (ALM) utilizing a neural network to predict fitness, and c) MAVE-NN that uses biophysics hypotheses to perform predictions (see Si4.0.2 for details). The fitness is largely additive (see Fig SI1), which can be directly seen from the linear regression ($\rho = 0.80 \pm 0.11$ with 94/403 (23%) as training/validation ratio); so the challenge is to predict epistasis. As expected, all other models performed significantly better than the regression for epistasis prediction (Fig 3A), except MAVE-NN. For the latter, we choose to illustrate the additive genotype-phenotype map offered in MAVE-NN, which essentially corresponds to a regression. For the NN-based models, we did not observe any striking loss of fitting performances when comparing D-LIM with its version with the unconstrained latent space (called ALM here).

Finally, we determined how many data points are required to achieve high prediction accuracy. For fitness prediction, D-LIM reached a performance of $\rho = 0.9$ using only 15% of the data (144/942, Fig 3B), similar to the other methods. For epistasis prediction, neural network-based methods achieved $\rho = 0.97$ at 50% dataset whereas the linear regression converged at most at only $\rho = Y$. However, D-LIM performances converged rapidly with 23% of the training set whereas the unconstrained version required more data points.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Our results showed that D-LIM performances are on par with state-of-the-art approaches for fitness and epistasis prediction. The high performances in fitness and epistasis prediction suggest that our independence of phenotype hypothesis is largely satisfied for this system.

2.3 Phenotype inference

In our genotype-phenotype map, genes are assumed to act independently in determining the phenotypic value. While our phenotypes φ are not measured but inferred, they are thought to be proxies for measurable biological phenotypes. We now show on simulated data that using our hypothesis infers phenotypes related to the true underlying phenotypes.

To perform our test, we would need experimental data such that both phenotypes and fitness are measured independently, which is not the case for high-throughput experiments. Instead, we simulated a training dataset based on the mechanistic model corresponding to the phenotype-fitness map of the former section [5], where fitness depends on two genes AraA, and AraB (see methods section 4.0.4). For AraA and AraB, denoted respectively X and Y, we simulated a set of 30 mutations to which we assigned phenotypic values randomly drawn in $[0, 5]$, as shown in Fig 4A. This resulted in a dataset of 900 mutation pairs for which we assigned a fitness value from the mechanistic model, adding a Gaussian noise to measurement errors.

We trained D-LIM with half of the samples and observed a good agreement with the fitness ($\rho = 0.96$). While the fitness landscape spanned by the trained phenotype space is different from the original one (Fig 4), the inferred phenotypes and the corresponding measured phenotype (enzyme activity) were strongly correlated ($\rho \approx 1$). If the underlying landscape is not additive and non-monotonous, the inferred phenotypes are found to be a smooth and continuous non-linear function of the actual phenotypes (Fig SI3).

2.4 Phenotypic measurement process for fitness extrapolation

Extrapolation of the prediction outside of the data domain is typically not expected for NN models. In contrast, mechanistic models can be fitted within a certain data domain, so that in turn, the model parameters can be used for prediction outside of the initial domain. Given that the D-LIM retains some of the constraints of mechanistic approaches, we surmised that its hypothesis-driven architecture may be used for extrapolation within a certain range beyond the training data. The extrapolation strategy is based on the relationship between the inferred and actual phenotypes, requiring the measurement of some of the latter for single genes. Indeed, when experimental measurements of the phenotypes are available for some of the mutants, the inverse relationship $\omega^{-1}(\phi) = \varphi$ between inferred and actual phenotypes can be fitted by a smooth function. Fitness predictions can then be extended to new mutants by measuring their actual phenotypic values φ associated with single mutations, then deducing their latent phenotypic value φ via ω^{-1}.

To illustrate this approach, we simulated an experiment with the Kemble model, where fitness is controlled by two phenotypes X and Y—enzyme activities by AraA and AraB. We restricted the training data to pairs of mutations yielding low fitness (black square in Fig 5A) whereas mutation combinations outside of this domain were left for validation. For the sake of comparison, we first tested a naive extrapolation strategy, directly using the model to predict fitness for combinations involving never-observed mutations. The prediction displayed poor accuracy (Fig 5C). Then, we applied our strategy using a third-order polynomial to fit the relationship between φ_1, φ_2, and their respective phenotypic values ϕ_1, ϕ_2, as shown in Fig 5B. Using the resulting phenotypes as an input for the NN yielded high fitness prediction accuracy (MSE=0.19), as shown in Fig 5C. This extrapolation technique also worked well for non-monotonous fitness functions (see Fig SI3).

Overall, the independence hypothesis allows D-LIM to extrapolate fitness predictions beyond the training data using phenotypic measurements, which would not have been straightforward to implement with

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

other ML approaches. Although this extrapolation strategy requires additional measurements, it may be of practical interest because it relies on single mutation data instead of combinations of mutations.

2.5 Integrating heterogeneous data sources

Typically, in experimental fitness screening campaigns, the measurements never cover the complete space of possible mutations. This means that with respect to the studied system, we are often in a low data regime. For example, in [5], the promoters of the studied genes have been mutated up to only a single nucleotide at the time. However, more information may be available, from the literature but also from other experiments which are not necessarily directly related to fitness or not even quantitative. We show now that such additional information can be incorporated in the learning process, which improved significantly the performances in fitness prediction in low data regimes.

This strategy starts with the assumption that some mutations yield similar fitness effects because they have similar phenotypic values. Therefore, the idea is to encourage pairs of mutations known as similar to have close inferred phenotype values during the learning process. This is done by first defining pairs of mutations (of the same gene) that are similar enough. We record these relations in a knowledge graph where edges are form between similar mutations (see Fig 6A). Then, during the training of the model, the genotype-phenotype map is constrained such that these pairs of mutations are associated with very close phenotypic values. We implement this using harmonic constraints on the genotype-phenotype map in addition to the loss function (see Fig 7B).

To establish a proof of principle, we used the mutational scanning experiment of the two-gene system in [5]. First, we built two knowledge graphs representing the similarity between mutations in each gene. To determine which pairs of mutations were similar enough, we first computed Pearson's correlation between fitnesses of pairs of mutations in same backgrounds, and paired the 10% most correlated mutations. If two mutations tend to yield similar fitness perturbations in same genetic backgrounds, we assume that they should have close phenotypic value. We determined the pairing directly based on the fitness to ensure that the choice of pairing is relevant. Indeed, the outcome of this strategy depends strongly on the quality of the knowledge graph.

We tested this approach across several training data size, with 30 random cross validation, where we measure the performance over a validation set that was kept constant (30% of the full data). Our results showed that improved significantly (with t-test pvalue < 0.05) the performance of D-LIM in a very low data regime (10^2 training data points), see Fig 6C. What happening is the information contain in one data point is used to train the D-LIM over more than strictly the mutations involve in the one data point.

3 Discussion

Genotype-to-fitness maps are central to understanding and predicting evolutionary adaptation. These maps can be fitted with mechanistic models, which are expected to be predictive and interpretable. However, they often require prior knowledge which may be unavailable, especially in biological systems. In contrast, ML models can be predictive without knowledge of the mechanisms, but at the cost of interpretability. In this work, we proposed D-LIM, a model that encodes biological hypotheses in the ML architecture: a relationship between mutations and genes, and the constraint that one gene affects only one phenotype. D-LIM thus mimics the mathematical form of a mechanistic model, but where the phenotype-to-fitness relationship is found automatically by neural network instead of being fitted with a parametric family of biochemical functions.

D-LIM is interpretable in the classical sense of hypothesis testing, where the fitting quality assesses the hypotheses underlying the model. Specifically, using data from mutants in a metabolic pathway, we showed that D-LIM predicts epistasis between genes with the accuracy of mechanistic and pure ML models, thus validating the hypothesis that mutations in each metabolic gene mostly affect the properties of the enzyme

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

encoded by this gene, and not other enzymes. This contrasts with mechanistic models, which may validate additional hypotheses (e.g. the functional form of enzymatic responses), and with ML models, which would predict fitness without proving or disproving any property of the metabolic pathway.

In contrast with ML, mechanistic models typically have virtues additional to pure prediction, due to their interpretability: the inference of biological parameters, the extrapolation beyond the data range, and the improvement of the integration of prior information. As D-LIM is partly constrained by the biological hypotheses used in mechanistic models, we showed that these virtues can be partly recovered.

First, D-LIM infers phenotypes that are formally the coordinates of a latent space between mutations and fitness. Although the inferred phenotypes are not *a priori* related to biological phenotypes, we demonstrated that the former could be interpreted as proxies of the latter up to a non-linear relationship. In particular, when the true phenotype-to-fitness relationship is monotonous, there is a univocal and monotonous relationship between each inferred phenotype and each biological one. This contrasts with other ML methods using latent spaces, where a mutation may span multiple dimensions. In LANTERN, the proximity between mutations in the latent space indicates a similarity between their effects. However, the coordinates of mutations cannot be interpreted as a biological property. In MAVE-NN, there is a single latent interpretable phenotype, but the approach is not suited for several phenotypes. MoCHI addresses the case of several phenotypes but does not deconvolve the effect of separate genes. In GenNet, nodes of the NN are connected according to the underlying gene network. Although this could improve interpretability, it requires extensive *a priori* biological knowledge, and the relationship with the phenotypes remained elusive.

Second, if provided with some phenotype values, D-LIM can perform extrapolation beyond the initial training domain. Extrapolation was performed in the latent space, by fitting the relationship between the inferred versus actual phenotypes with a polynomial. Measuring the phenotype of new mutants allows one to project them into the latent space, and then compute their fitness using the NN. This may be of practical interest, as it requires measuring phenotypes for single mutations instead of combinations thereof.

Finally, D-LIM can use prior knowledge to improve predictions in the low data regime. Indeed, sparse data is a common situation in biology due to the combinatorial explosion of possible mutations and conditions. We proposed a generic framework to integrate variable sources of measurement, even when this information is qualitative. This approach relies on a general encoding scheme where pairs of mutations are associated in a knowledge graph if they are expected to have similar fitness effects. These associations then encourage D-LIM to infer similar phenotype values for similar mutations.

Overall, our results show that introducing constraints on the form of the model enhanced the interpretability of neural networks while retaining most of their predictive power. At this stage, the hypotheses need to be determined by the user. However, it may become possible to automatically generate the mutations-to-genes mapping, ultimately allowing us to devise interpretable ML for large biological systems.

4 Methods

4.0.1 D-LIM hyperparameter and loss function

To train D-LIM, we used the Negative Gaussian loglikelihood:

$$\mathcal{L} = \frac{1}{2}\left[\log(\sigma) + \frac{(\hat{F} - F)^2}{\sigma}\right], \tag{1}$$

where $\hat{F}$ and σ are D-LIM predictions respectively for the fitness and its uncertainty. To choose the hyperparameters for D-LIM, we performed a grid search (for the dataset in [5]), where we varied the number of hidden layers in $[1, 2, 3, 4]$, and the number of neurons per layer in $[2, 4, 6, 8, 16, 32]$. Based on this search, we selected the architecture with 1 hidden layer of 32 neurons for D-LIM.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: **CPU time for different algorithms in fitness prediction in Env1 of [5].** Each model was run five times and the results were average (in seconds)

Models	CPU time (s)
D-LIM	11.54 ± 1.64
LR	4.04 ± 0.01
ALM	14.40 ± 1.93
Lantern	**133.90** $\pm$ 135.91
MAVE-NN	21.99 ± 2.63

We trained D-LIM on a machine equipped with 104 Intel(R) Xeon(R) Gold 6230R CPUs at frequency 2.1HZ, where the learning typically finished in 11.54 seconds for 300 learning steps, see comparisons in table 1. We implemented D-LIM using the deep learning backend PyTorch [9], while the analyses were performed using the routines implemented in the package Numpy [3].

4.0.2 Models comparison

- **Additive latent model (ALM)**: This model used an unconstrained additive latent space, similar to the one proposed in LANTERN. However, here, we used a neural network to predict fitness out of the latent space. Here, we used an 8-dimension latent space, 2 layer of 32-units neural network for fitness prediction. We trained the model for 300 steps, using ADAM and a learning rate 0.01 and a regularization strength of 0.001. The batch size used here is 64.
- **LANTERN**: This model used an additive latent space, where the importance of each dimension is controlled explicitly. The fitness is predicted using Gaussian process. Here, we used K=4 latent dimension and a neighborhood number of 200, optimized for 10^3 steps using ADAM and a learning rate of 0.001.
- **MAVE-NN**: this model introduced a biophysical interpretation using a latent phenotype and a deterministic genotype to phenotype map, combined with a stochastic measurement process. Here, we used an additive genotype-phenotype map along with the skewed-t noise model for the measurement process. The model has been trained for 10^3 steps with a learning rate of 0.001 and batch size of 64.
- **Linear regression (LR)**: the linear regression is an additive map we implemented using Pytorch. We optimized the model using Adam for 300 steps and a learning rate of 0.01 and batch size 64.

4.0.3 Cross validation and statistical analyses

All models were evaluated with 30 random fold cross validation where 30% of the original dataset (Env1 in [5]) was left for validation across the whole experiment. The accuracy across folds was quantified with Pearson's correlation; whereas, we used the Mean Square Error (MSE) to estimate the gain in predictive power in the context of the extrapolation. Then, to measure the significance of Pearson's correlation sample across random samples and methods, we used the student statistical test (t-test).

To test the generalization of each model, we varied the size of the training set in $[99, 152, 232, 354, 540, 824, 999]$ data points, where we trained 30 model (from random parameters initialization), while the validation set was kept fixed at 30% of the original data.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4.0.4 Kemble *et al* dataset and mechanistic model

The data published in [5] tested all mutation combinations on the promoter of 2 adjacent genes from the linear Escherichia coli l-arabinose pathway. Mutations are restricted to 12 positions in both promoters. There are 72 single mutants and 1296 double mutants, in total 1369 variants. The fitness has been measured using barcoding tracking in three environments (defined by inducers concentrations): Env1: $20\ ng/ml$ aTc and $30\ \mu M$ IPTG; Env2: $5\ ng/ml$ aTc and no IPTG; Env3: $200\ ng/ml$ aTc and no IPTG). Hence, there are 1369 data samples in each environment.

The mechanistic model proposed in [5] is as follow:

$$F(X,Y) = \left(w + \mu\varphi - \frac{\nu}{1/\eta - \varphi}\right)(1 - \theta_X X - \theta_Y Y)\,, \tag{2}$$

where , $\varphi = \frac{1}{1/X+1/Y+\eta}$ denotes for flux, η is the inverse of the maximal flux φ, θ_X and θ_Y represent the cost of increasing cellular enzyme activity, w describes the growth rate, μ and ν are two variables related to downstream enzyme properties.

5 Availability of data and material

Code and data generated as well as fetch from other studies are available in the Github repository https://github.com/LBiophyEvo/D-LIM-model.git. D-LIM is provided in the format of a python (version 3.10) package. The training time of the model scales linearly with the number of data points, which in the case of [5] is a couple of seconds.

References

[1] Chou, H.-H., Delaney, N. F., Draghi, J. A., and Marx, C. J. Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. *PLoS genetics 10*, 2 (2014), e1004149.

[2] Faure, A. J., and Lehner, B. Mochi: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis and allostery from deep mutational scanning data. *bioRxiv* (2024), 2024–01.

[3] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature 585*, 7825 (2020), 357–362.

[4] Hornik, K., Stinchcombe, M., and White, H. Multilayer feedforward networks are universal approximators. *Neural networks 2*, 5 (1989), 359–366.

[5] Kemble, H., Eisenhauer, C., Couce, A., Chapron, A., Magnan, M., Gautier, G., Le Nagard, H., Nghe, P., and Tenaillon, O. Flux, toxicity, and expression costs generate complex genetic interactions in a metabolic pathway. *Science Advances 6*, 23 (2020), eabb2236.

[6] Kinney, J. B., and McCandlish, D. M. Massively parallel assays and quantitative sequence–function relationships. *Annual review of genomics and human genetics 20* (2019), 99–127.

[7] Nix, D. A., and Weigend, A. S. Estimating the mean and variance of the target probability distribution. 55–60.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[8] OTWINOWSKI, J., AND PLOTKIN, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proceedings of the National Academy of Sciences 111*, 22 (2014), E2301–E2309.

[9] PASZKE, A., GROSS, S., CHINTALA, S., CHANAN, G., YANG, E., DEVITO, Z., LIN, Z., DESMAISON, A., ANTIGA, L., AND LERER, A. Automatic differentiation in pytorch.

[10] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., ET AL. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (2019).

[11] SARKISYAN, K. S., BOLOTIN, D. A., MEER, M. V., USMANOVA, D. R., MISHIN, A. S., SHARONOV, G. V., IVANKOV, D. N., BOZHANOVA, N. G., BARANOV, M. S., SOYLEMEZ, O., ET AL. Local fitness landscape of the green fluorescent protein. *Nature 533*, 7603 (2016), 397–401.

[12] STARR, T. N., GREANEY, A. J., HILTON, S. K., ELLIS, D., CRAWFORD, K. H., DINGENS, A. S., NAVARRO, M. J., BOWEN, J. E., TORTORICI, M. A., WALLS, A. C., ET AL. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *cell 182*, 5 (2020), 1295–1310.

[13] STARR, T. N., AND THORNTON, J. W. Epistasis in protein evolution. *Protein science 25*, 7 (2016), 1204–1218.

[14] TAREEN, A., KOOSHKBAGHI, M., POSFAI, A., IRELAND, W. T., MCCANDLISH, D. M., AND KINNEY, J. B. Mave-nn: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome biology 23*, 1 (2022), 98.

[15] TONNER, P. D., PRESSMAN, A., AND ROSS, D. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. *Proceedings of the National Academy of Sciences 119*, 26 (2022), e2114021119.

[16] VAN HILTEN, A., KUSHNER, S. A., KAYSER, M., IKRAM, M. A., ADAMS, H. H., KLAVER, C. C., NIESSEN, W. J., AND ROSHCHUPKIN, G. V. Gennet framework: interpretable deep learning for predicting phenotypes from genetic data. *Communications biology 4*, 1 (2021), 1094.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

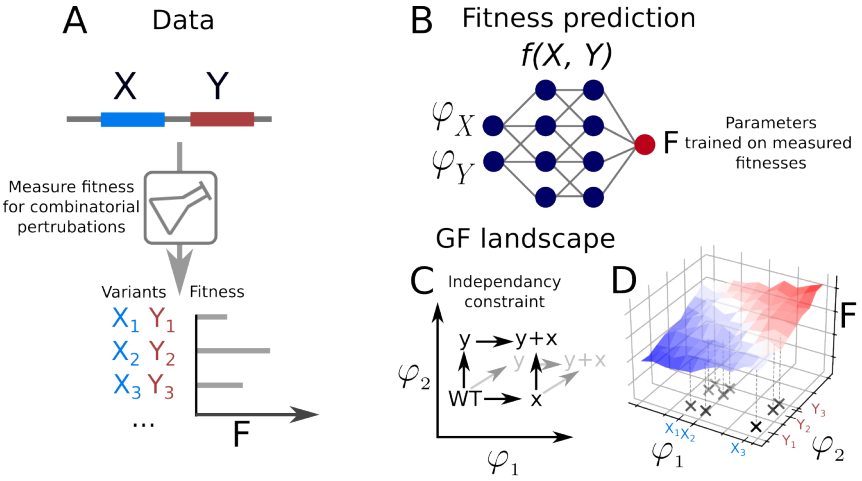


Figure 1: D-LIM framework. A) An example of data acquisition for a pair of genes X and Y, for which fitnesses are measured for a combination of mutations. B) D-LIM architecture. D-LIM takes one value φ per gene variant and processes it through a feed-forward neural network to predict the fitness of these two genes variants. C) One gene phenotype per dimension in the inferred phenotype space. The coordinates of gene variants along the axis are obtained during the training of D-LIM, where each inferred phenotype φ axis represents a single gene. Consequently, mutations in such a constrained space follow are independent (represented by the dark arrows). If mutations were not independent, they would vary along all dimensions at once (as represented by the grey arrows). D) The constrained phenotype space. The fitness is mapped to the constrained phenotype space where combination of gene variants are organized in a grid.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

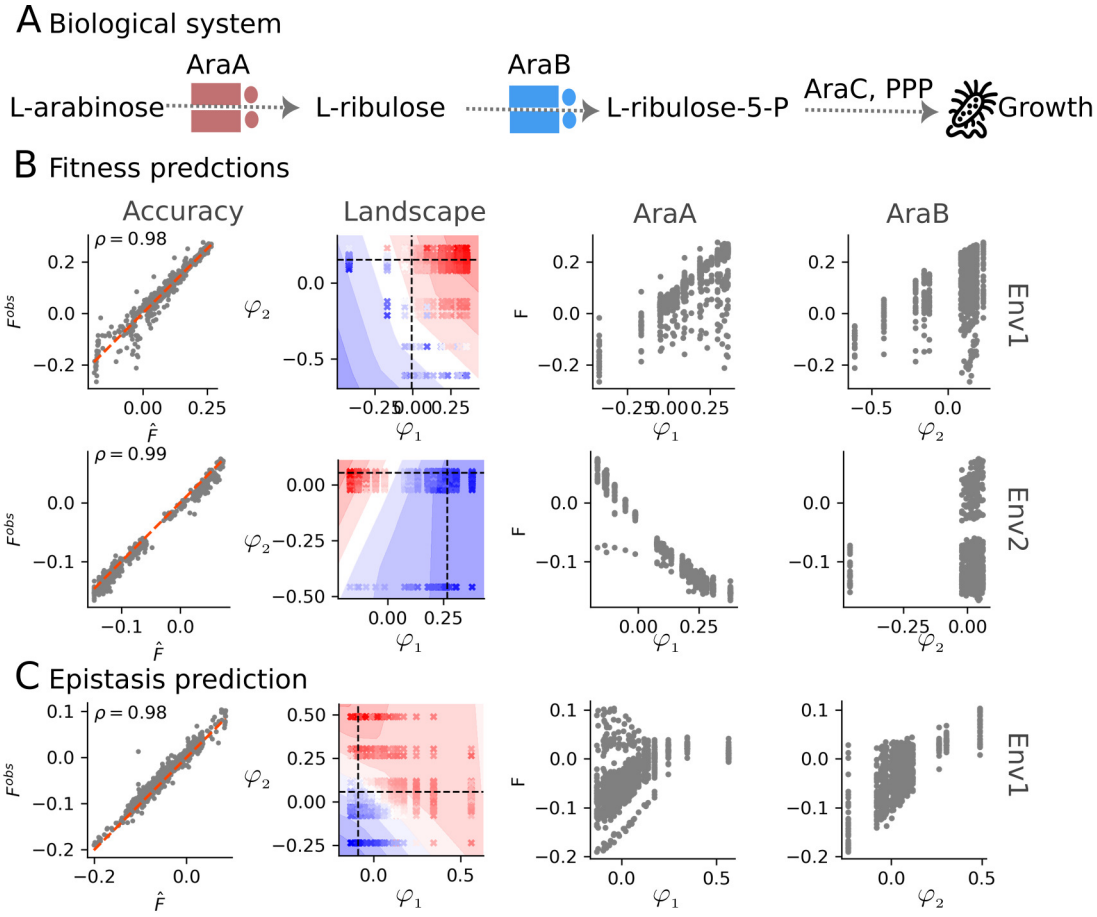


Figure 2: Gene-gene interaction pathway test case. A) L-Arabinose pathway of E. coli. B) Fitness prediction in two environments. (From left to right) We show the accuracy of the fitness predictions and the computed Pearson correlation. The predicted phenotype space is represented by a blue to red color gradient, where φ_1 and φ_2 are the inferred phenotypes for mutations respectively on genes AraA, and AraB, and the color is the measured fitness for each mutation. The color of the contour is the predicted fitness by D-LIM. Next, we show the relation between the measured fitness and two inferred phenotypes. The first row represents the analyses in Env1 whereas the second row is in Env2. C) Epistasis prediction. The row displays the accuracy, landscape, and correlation between inferred phenotypes and epistasis in Env1.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

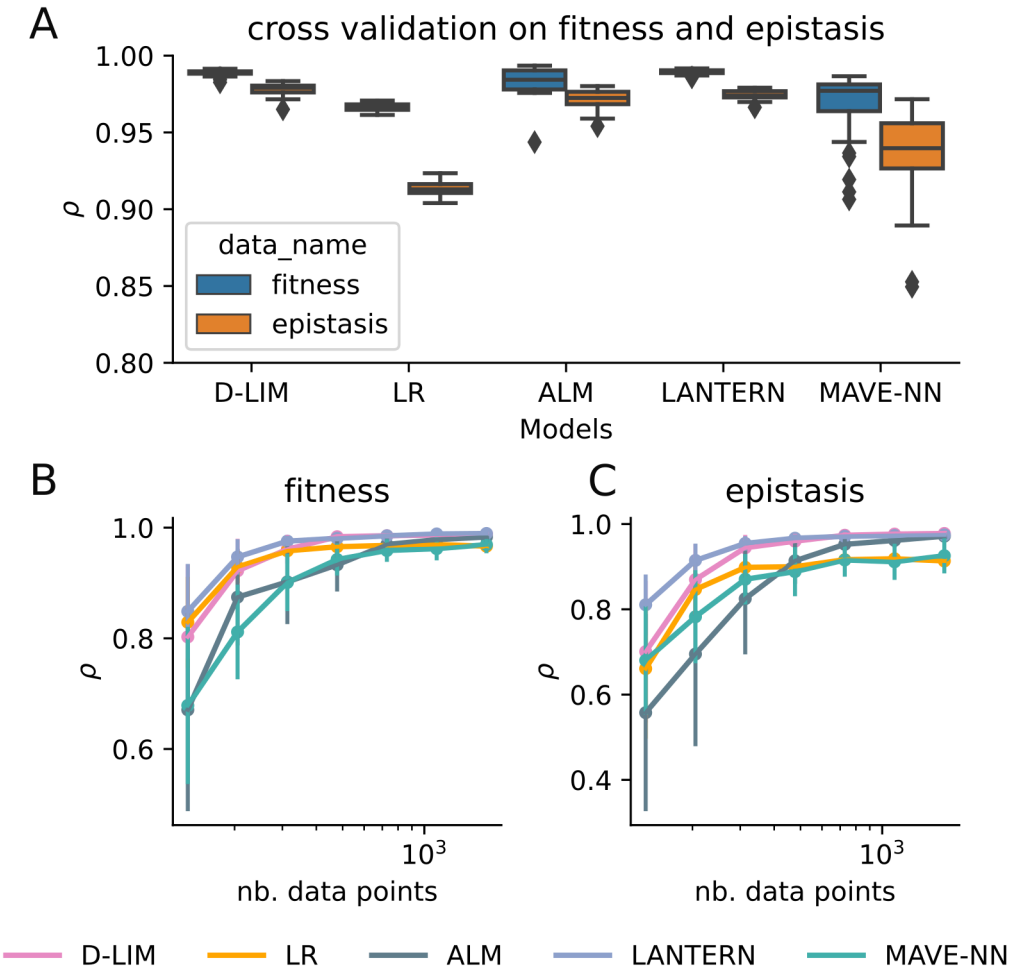


Figure 3: Comparing prediction across models. A) Boxplots of Pearson's correlations by random cros validation obtained for validation sets corresponding to 30% (257 data samples) of the dataset for fitness or epistasis. B-C) Prediction quality for fitness or epistasis as a function of training set size. We varied the training set size in [0.1 , 0.15 , 0.23, 0.35, 0.54, 0.82, 1.0]) while using the same validation (30% of the original dataset). The bars represent the standard deviation across the 30 replicates per training size and method.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

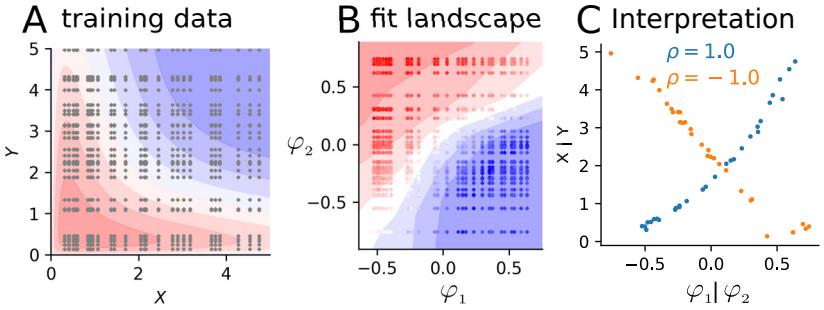


Figure 4: Improved interpretation. A) Artificial landscape and training data. The color scale represents the fitness computed from the mechanistic model detailed in Equ 2 $F(X, Y)$ [5], whereas the grey points correspond to the samples which we split into 50%/50% training validation sets. B) D-LIM fitted landscape. φ_1 and φ_2 represents respectively the mutations in X and Y, whereas the color scale now represents the predicted fitness. C) Correlation between the underlying true phenotypes X (blue) and Y (orange), and the inferred phenotypes φ_1 and φ_2.

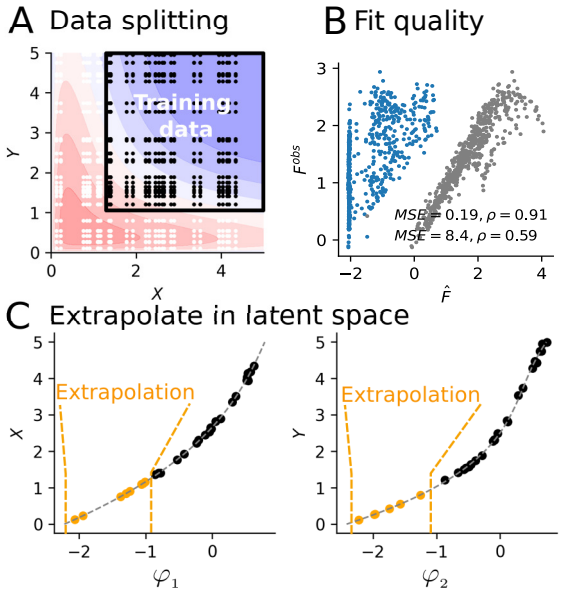


Figure 5: Extrapolation with un-seen mutations and their measured phenotypes. A) The landscape represented is obtained from the mechanistic model of [5]. Black points are the training whereas the white ones are the validation points. B) Fit quality. After converting the true phenotypic values in inferred phenotypes, used by D-LIM to extrapolate: for blue points, we did not used the additional true phenotype values whereas for grey points, we did. C) Extrapolation using X (left), Y (right) phenotypic measurements. Black points are (X, φ_1) pairs used to parametrize the polynomial fit in a dotted grey line. Orange points are the converted X and Y only seen in the validation, which are converted into inferred phenotypes using the polynomial fit.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

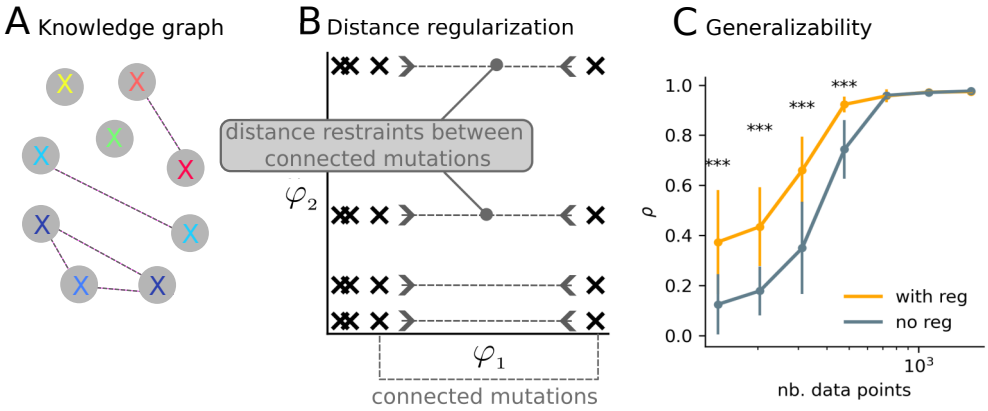


Figure 6: Integrative modelling. A) The nodes of the knowledge graph represent genetic variants that are connected if they are assumed to have similar effects on the fitness. B) Distance regularization. Connected variants/mutations are restrained during the learning so that similar variants are closer. C) The curves represent the Pearson correlation between measured and predicted fitness evolve with the increment of training dataset size. The grey curve and the yellow curve display the performance of D-LIM with and without adding the knowledge graph as a regularization constraint.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6 Supplementary Information: Biology-driven constraint for neural network genotype-fitness maps

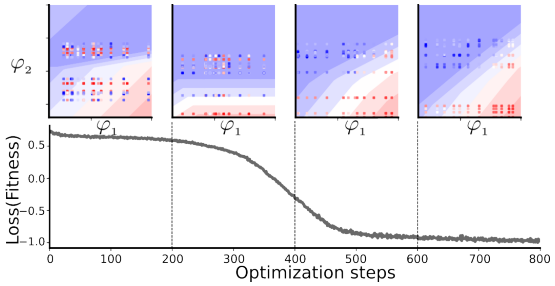


Fig. SI1: **Simultaneous fitting of Fitness and inference of phenotypes.** For the mechanistic model [5], we created artificial mutations that were randomly initialized in the inferred phenotype space of D-LIM (first of the four top sub-panels). The real fitness values of each mutant (represented by the colour of the dots) did not aligned with the landscape predicted by D-LIM (the blue-red gradient colour); because neither of the parameters of the fitness predictor, nor the inferred phenotype were yet optimized. Then, we monitor across three time points (after 200 steps of ADAM optimization) the evolution of the fitness prediction quality (Loss(Fitness)), and the inferred phenotypes. Along this optimization, the parameters of the fitness predictor and the coordinates in the latent space were optimized.

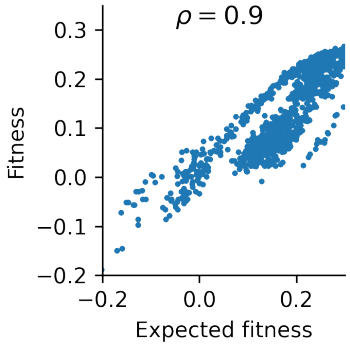


Fig. SI2: **The fitness measured in [5] is largely additive.** We compared the double mutant fitness measured F(A, B) with the sum of single mutant fitnesses F(A)+F(B).

6.0.1 Extrapolation by using latent variables

We tested here the extrapolation strategy using a non-monotonous fitness function. First, we sampled from the uniform distribution: $X_i, Y_i \sim U_{[0,5]}, i \in \{1, 2 \cdots, 36\}$, representing the true phenotype values. To validate the trained models, we extracted the data points whose phenotype values lied in $X_i \in [1.5, 2.7]$ and $Y_i \in [1.5, 2.7]$. The other data points were used for training, as shown in Fig SI3.

bioRxiv preprint doi: https://doi.org/10.1101/2024.04.09.588719; this version posted April 13, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

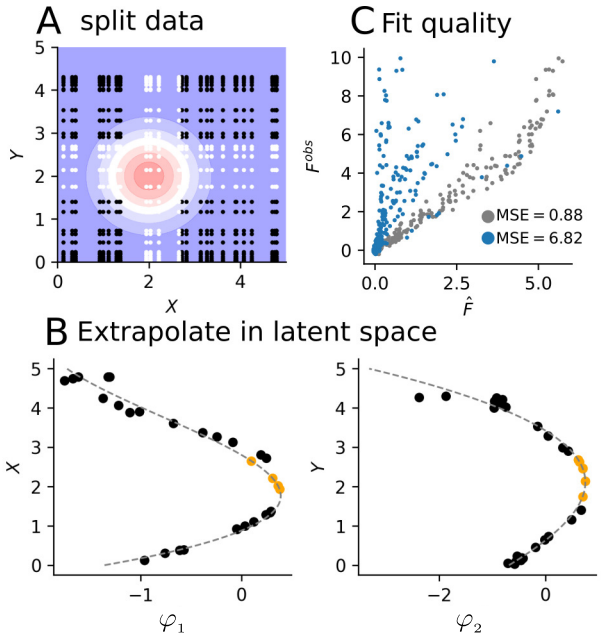


Fig. SI3: Extrapolation of a non monotonic fitness landscape. A) Gaussian-type fitness landscape centered at $X = 2$ and $Y = 2$. Black points are the training points whereas the white ones are the validation points. B) Extrapolation using new X (left), Y (right) phenotypic measurements. Black points are (X, φ_1) pairs used to parametrize the polynomial fit in a dotted grey line. Orange points are the converted X and Y only seen in the validation, which are converted into latent variables using the polynomial fit. C) Fit quality. After converting the phenotypic values only seen in the validation to latent parameters: blue points are the conversion whereas grey ones are after.