Published on Oct 30, 2025 by Arcadia Science **DOI**: 10.57844/arcadia-f7x4-dcgs

Equivalent linear mappings of deep networks are a promising path for biology

Deep networks make accurate predictions, but their nonlinearity makes them a black box, hiding what they have learned. Here, we look inside the black box and analyze the exact relationships they learn for UMAP embeddings and epistasis in a genotype–phenotype dataset.

Contributors (A-Z)

James R. Golden, George Sandler, Ryan York

Deep networks are increasingly popular in biology, but their fundamental nonlinear character makes it difficult to extract what input–output relationships they have actually learned. We describe a method for finding an equivalent linear mapping for a deep network given a specific input, and apply this to UMAP embeddings and epistasis in genotype–phenotype relationships.

The equivalent linear mapping method enables the use of powerful deep networks to accurately learn complex relationships, while also allowing for the straightforward interpretation of which gene features give rise to specific output representations.

This perspective piece is intended for both practitioners interested in interpretability for machine learning and biologists skeptical of the scientific utility of machine learning methods. We would be pleased to receive feedback from anyone who could make use of this approach for their own datasets, and especially whether it results in deeper insights into the structure of the data itself or the biological processes that underlie the system of interest.

- Check out **companion pubs** showing how to use equivalent linear mappings for interpreting globally nonlinear models for gene expression **[1]** and genotype–phenotype data **[2]**.

A fundamental challenge in machine learning is the black-box nature of nonlinear models. Models can be optimized to make accurate predictions across a wide range of datasets and problems, but we cannot always understand the relationship a model has learned between the input features and the output prediction. There has been constant progress on feature importance methods that capture approximate contributions, including Grad-CAM **[3]**, Integrated Gradients **[4]**, LIME **[5]**, and SHAP **[6]**. While useful, these methods always come with the “approximate” warning flag, as nonlinear functions can behave differently than their linear approximations.

In parallel, there have been less visible efforts toward local linear descriptions of deep-network models that capture the exact relationship between input features and output predictions as equivalent linear mappings (ELMs). “Analysis of deep neural networks with the extended data Jacobian matrix” **[7]** identified and explored this intriguing property. For simple deep networks consisting of only linear layers with zero bias and

ReLU activations, the Jacobian for a particular input, computed numerically with autograd, yields a linear system that exactly reproduces the output of the globally nonlinear deep network. The manifold of the output in the input space is piecewise linear. This method has been extended to convolutional networks with ReLU activations for image generation **[8][9]** and to large language models with Swish or GELU activations and softmax attention **[10]**, although these gated activations make the model "point-wise" linear as opposed to piecewise linear.

The ELM method leverages the full expressive power of deep networks but does not sacrifice quantitative interpretation of input features. If a linear model is held up as one ideal for interpretability (which itself is open to debate **[11]**), then the Jacobian method for a piecewise or pointwise linear network is a solid step in this direction. Instead of interpreting a nonlinear network, we must interpret a large collection of linear models for each input of interest. This is a shift from an extremely difficult mathematical problem to a challenging data problem.

We recently posted two new publications, "From black box to glass box: Making UMAP interpretable with exact feature contributions" and "A quantitative-genetic decomposition of a neural network". While there are many techniques for quantifying the contributions of genes in linear genotype–phenotype models and for computing approximations of this in nonlinear models, the ELM approach used in these publications maps the prediction of a deep network to an equivalent set of linear weights for a given input point.

We used standard deep networks with some simple constraints such that the network is locally linear with respect to the input features. By holding the bias terms of each linear layer to zero and using ReLU or leaky ReLU activations, the Jacobian matrix for a specific input (computed numerically with autograd) exactly reconstructs the network output, with reconstruction error approaching machine precision. This class of networks is linear at a given input point, but nonlinear between input points, which allows for both predictive power and clear feature interpretations from the Jacobian reconstruction. This is a technique that has not yet been widely used for genotype–phenotype prediction.

Exact feature contributions for UMAP

In “From black box to glass box: Making UMAP interpretable with exact feature contributions,” we used the Jacobian to quantify how a trained deep network transformed a cell’s normalized expression levels to a position in the embedding space. The conventional approach to quantifying feature contributions is to carry out differential expression on a given cluster formed from a UMAP embedding, but with the ELM method, we can directly measure what gene features the UMAP network uses to embed each cell’s expression vector. These features are computed for individual cells but can be averaged over labels to identify which genes drive the formation of clusters labeled with categories such as cell type.

ELMs and quantitative genetics

In “A quantitative-genetic decomposition of a neural network,” we used simulated data to show how the feature contributions captured by the Jacobian of a neural network through ELM can be used to estimate classical quantitative genetics parameters. We first demonstrated that by averaging the Jacobian over all test set points, we can accurately back out the ground truth additive effect sizes. This was encouraging, but given that we have a method that identifies different sets of linear features for each input point, could it also reveal pairwise epistatic interactions? We found that by iterating over pairs of loci and averaging the Jacobian over all possible genotypic combinations, we could infer epistatic interaction coefficients with high fidelity. This was true for phenotypes with a variety of genetic architectures and environmental noise levels, suggesting that this method should be broadly applicable to genotype–phenotype mapping.

By constraining a few aspects of the deep network architecture with no cost to its performance, we use the Jacobian to reveal equivalent linear relationships for each point in the dataset. This augmentation of the modeling pipeline enables a straightforward and principled approach to uncovering both locally linear and globally nonlinear relationships.

Beyond genotype–phenotype mappings, we hope to apply this technique to other areas of interest to Arcadia, including the interpretation of protein language models and the analysis of high-dimensional phenotypic data.

---

1 10.57844/arcadia-tnr4-7n9h

2 10.57844/arcadia-v4qf-vw3k

3 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. https://doi.org/10.48550/ARXIV.1610.02391

4 Sundararajan M, Taly A, Yan Q. (2017). Axiomatic Attribution for Deep Networks. https://doi.org/10.48550/ARXIV.1703.01365

5 Ribeiro MT, Singh S, Guestrin C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.48550/ARXIV.1602.04938

6 Lundberg S, Lee S-I. (2017). A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/ARXIV.1705.07874

7 Wang S, Mohamed A-R, Caruana R, Bilmes J, Philipose M, Richardson M, Geras K, Urban G, Aslan O. (2016). Analysis of deep neural networks with the extended data Jacobian matrix. https://proceedings.mlr.press/v48/wanga16.html

8 Mohan S, Kadkhodaie Z, Simoncelli EP, Fernandez-Granda C. (2019). Robust and interpretable blind image denoising via bias-free convolutional neural networks. https://doi.org/10.48550/ARXIV.1906.05478

9 Kadkhodaie Z, Guth F, Simoncelli EP, Mallat S. (2023). Generalization in diffusion models arises from geometry-adaptive harmonic representations. https://doi.org/10.48550/ARXIV.2310.02557

10 Golden JR. (2025). Equivalent Linear Mappings of Large Language Models. https://openreview.net/forum?id=oDWbJsIuEp

11 Lipton ZC. (2016). The Mythos of Model Interpretability. https://doi.org/10.48550/ARXIV.1606.03490