

Predicting of Gene Functions by Leveraging Biological Insights with Mechanistic Machine Learning



Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

George G. Vega Yon, Ph.D.

george.vegayon@utah.edu

Division of Epidemiology @ University of Utah

Dec 8th, 2022 @ Quinlan Lab

Collaborators: Paul Thomas, Paul Marjoram, Huaiyu Mi, Christopher Williams (USC), Alun Thomas (UofU)

Table of Contents

Preliminaries

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

But first...

whoami



George G. Vega Yon's GitHub Stats

★ Total Stars Earned:	247
⌚ Total Commits (2022):	2.4k
🍴 Total PRs:	111
❗ Total Issues:	228
👤 Contributed to:	33



- ▶ Research Assistant Professor in Epidemiology since Nov. 2021.
- ▶ A methodologist researching Statistical Computing applied to Complex Systems Modeling.
- ▶ Got a Ph.D. in Biostats from USC and an M.Sc. in Economics from Caltech.
- ▶ Working with networks + programming since ~2010.
- ▶ Scientific software developer (R/C++)

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Table of Contents

Preliminaries

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Evolution of Gene Function

Mechanistic Machine Learning

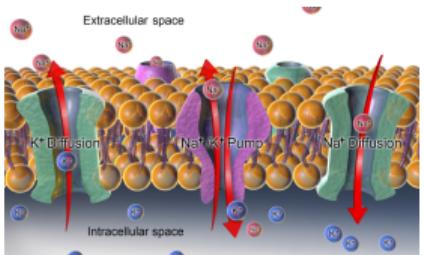
Proof of Concept

Gene Function

Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

Molecular function

Active transport
GO:0005215



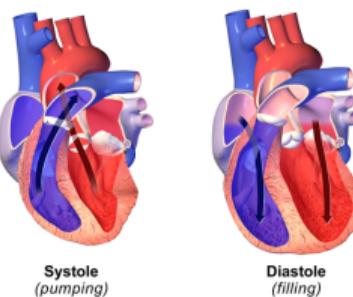
Cellular component

Mitochondria GO:0004016



Biological process

Heart contraction
GO:0060047



Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Gene Function: the Gene Ontology Project



GENEONTOLOGY
Unifying Biology

- ▶ The GO project has \sim 43,000 validated terms, \sim 7.4M annotations on \sim 5,200 species.
- ▶ About \sim 700,000 annotations are on human genes.
- ▶ Only half of the human gene annotations are based on experimental evidence.
- ▶ About \sim 173,000 publications have used the GO.

source: Statistics from <http://pantherdb.org/panther/summaryStats.jsp> and <http://geneontology.org/stats.html>

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Predicting Gene Function: Current methods

Sequences, phylogenomics, and ML.

- ▶ **BLAST**:² Prediction by sequence homology (~ 105,000 citations).
- ▶ **SIFTER**:^{6,5} An evolutionary model of gene function/loss using phylogenetics.
- ▶ **aphylo**¹⁵ (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- ▶ **Phylo-PFP**:⁹ A BLAST-based adding phylogenetic based distances.
- ▶ **DeepGOPlus**:¹¹ One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- ▶ **GOLabeler**:¹⁷ Top performing tool according to the CAFA challenge,¹⁸ is an ensemble of various simple ML methods, including K-means and logistic regression.
- ▶ **DeepFRI**:⁷ Uses Graph Convolutional Neural Networks (GCNs) to predict function based on protein structure and genetic sequence.

None of the ML-based methods relies on biological theory (mechanistic models).

Table of Contents

Preliminaries

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

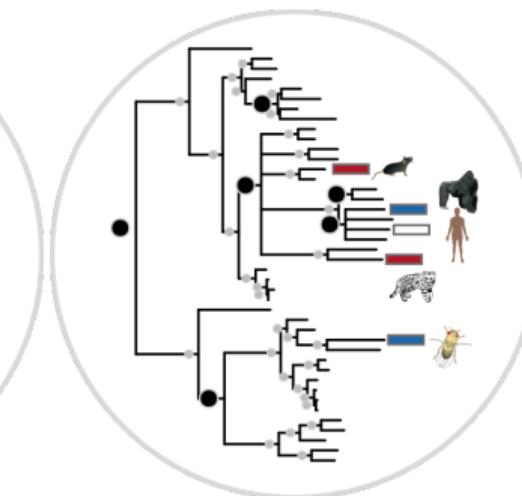
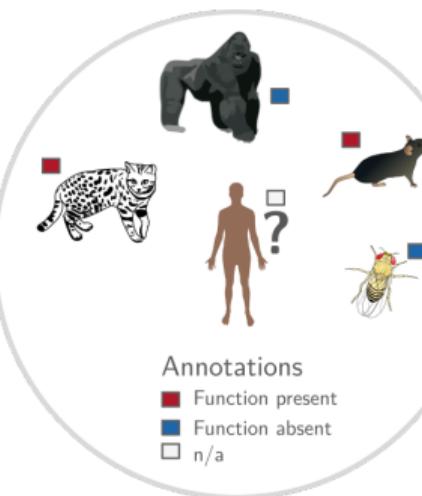
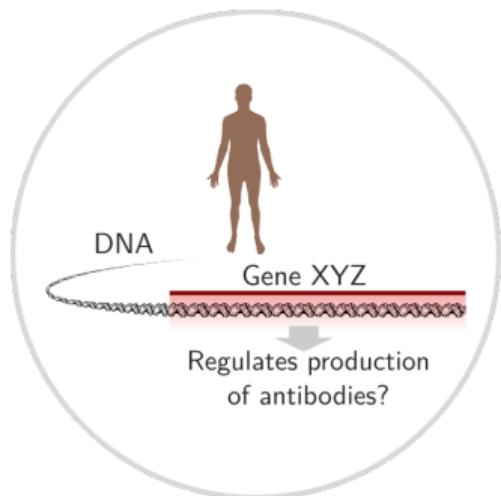
Proof of Concept

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Is gene XYZ involved in process ABC?



Complex to directly assess

But we may know from other species

And we further know how these are *evolutionary* connected

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Is the human gene **XYZ** involved in process **ABC**, given what we know about that for
other related species?



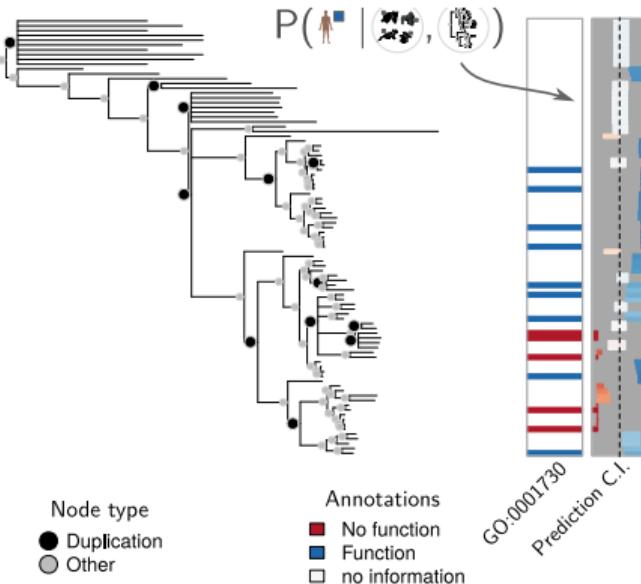
Annotations

- Function present
- Function absent
- n/a

Evolution of Gene function (of one function)

Built a big model (lots of trees and annotations) called aphylo:

- ▶ Only two sources of data:
Phylogenetic tree
(pantherdb.org) and functional annotations
(geneontology.org).
- ▶ Leverage negative annotation of GO terms (NOT).
- ▶ Use Felsenstein's tree pruning algorithm to compute tree likelihood.
- ▶ Fit pooled models featuring thousands of annotations in hundreds of trees (with split-second prediction capability).



... But what if we wanted to deal with multiple functions?

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Evolution of Gene function (multiple functions)

Tapping into Evol. Theory

- ▶ A fundamental part of Fun. Evol. is Duplication Events.
- ▶ Furthermore, knowing what happened to gene A (e.g., neofunctionalization) is highly informative to learn about the functional state of B.
- ▶ One way to model this is using a Markov Transition Model (as in SIFTER).

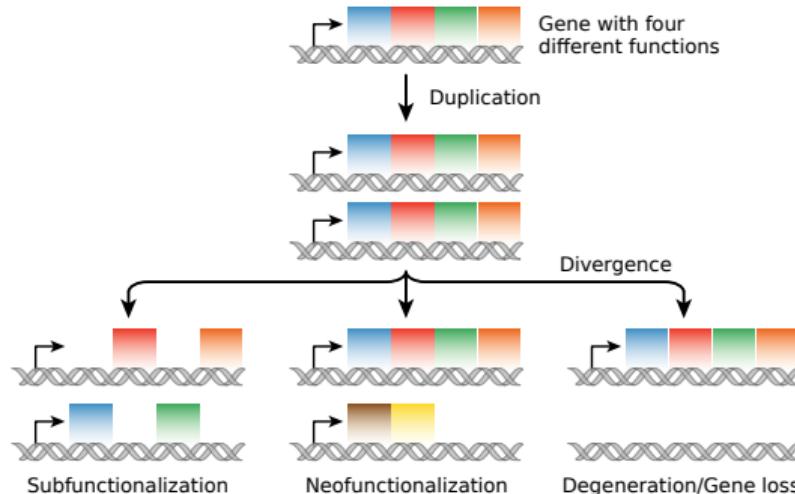


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Preliminaries

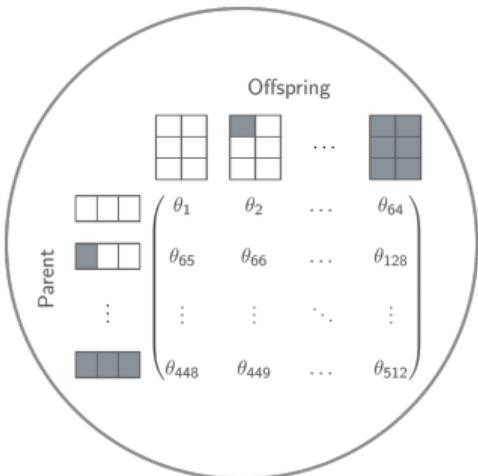
Evolution of Gene
FunctionMechanistic Machine
Learning

Proof of Concept

Evolution of Gene function (multiple functions)

If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



- ▶ 512 parameters
- ▶ Finding this many parameters is not easy.
- ▶ Even if you can, interpretation is awkward.

Social Network Analysis may help us...

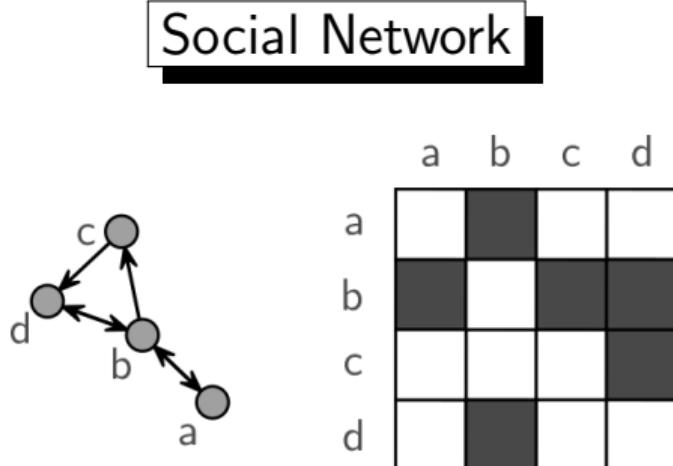
Exponential Random Graph Models (ERGMs)

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept



- ▶ Not about individual ties.
- ▶ Statistical inference on *motifs* (triangles, dyads, homophily, etc.)
- ▶ Literature about ERGMs is vast, a.k.a. a low-hanging fruit.

Ultimately...

ERGM ≡ **Modeling binary arrays**

Exponential Random Graph Models (ERGMs)

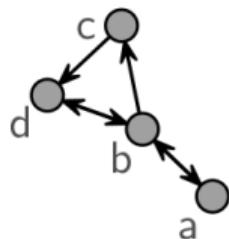
Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

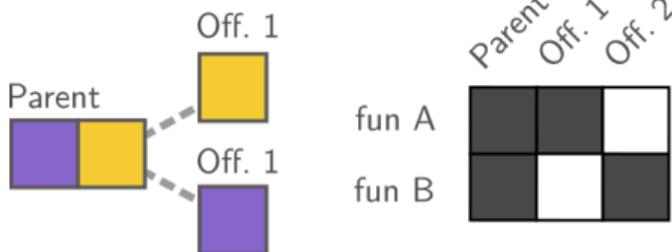
Proof of Concept

Social Network



	a	b	c	d
a				
b				
c				
d				

Evolutionary Event

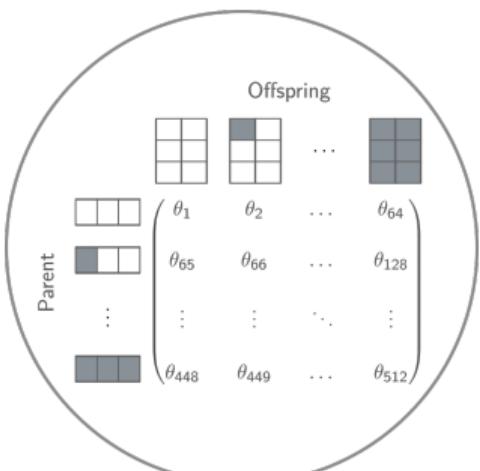


Social Networks are usually represented as **adjacency matrices**, and so can evolutionary events!

Evolution of Gene function (multiple functions) (cont.)

If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix

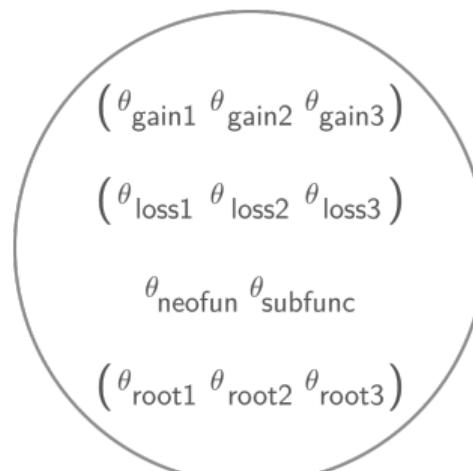


512 parameters

Sufficient statistics

$$\begin{pmatrix} \theta_{\text{gain}1} & \theta_{\text{gain}2} & \theta_{\text{gain}3} \\ \theta_{\text{loss}1} & \theta_{\text{loss}2} & \theta_{\text{loss}3} \\ \theta_{\text{neofun}} & \theta_{\text{subfunc}} \\ \theta_{\text{root}1} & \theta_{\text{root}2} & \theta_{\text{root}3} \end{pmatrix}$$

Easier to fit
Easier to interpret



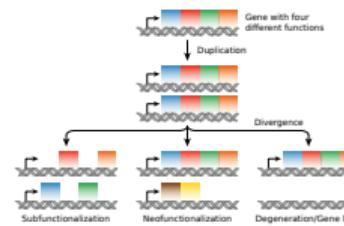
11 parameters (for example)

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept



Rep.	Description	Definition
	Gain of function	$(1 - x_p) \sum_{n:n \in Off} x_n$
	Loss of function	$x_p \sum_{n:n \in Off} (1 - x_n)$
	Subfunctionalization	$x_p^k x_p^j \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Neofunctionalization	$x_p^k (1 - x_p^j) \sum_{n \neq m} x_n^k (1 - x_n^j) (1 - x_m^k) x_m^j$
	Longest branch gains	$(1 - x_p^k) \mathbf{1} (x_m^k : m = \text{argmax}_n \text{blength}_n)$

Table: Example of sufficient statistics for evolutionary transitions.

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

GEESE: GEne functional Evolution using SuficiEncy

I implemented what I just described in a C++ library with a companion R package called `geese`. The question is: How much do we earn by using these motifs?

- ▶ Using 37 phylogenetic trees featuring 401 go annotations.
- ▶ `aphylo`: Fitted a *simple gain/loss* of function model.
- ▶ **GEESE**: Fitted an evolutionary model controlling for *functional preservation* (i.e., like neofun or subfun.)
- ▶ Fitted both of them using MCMC.
- ▶ Used LOO cross-validation to compute aggregated AUCs and MAE.

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

GEESE for predicting gene function (cont.)

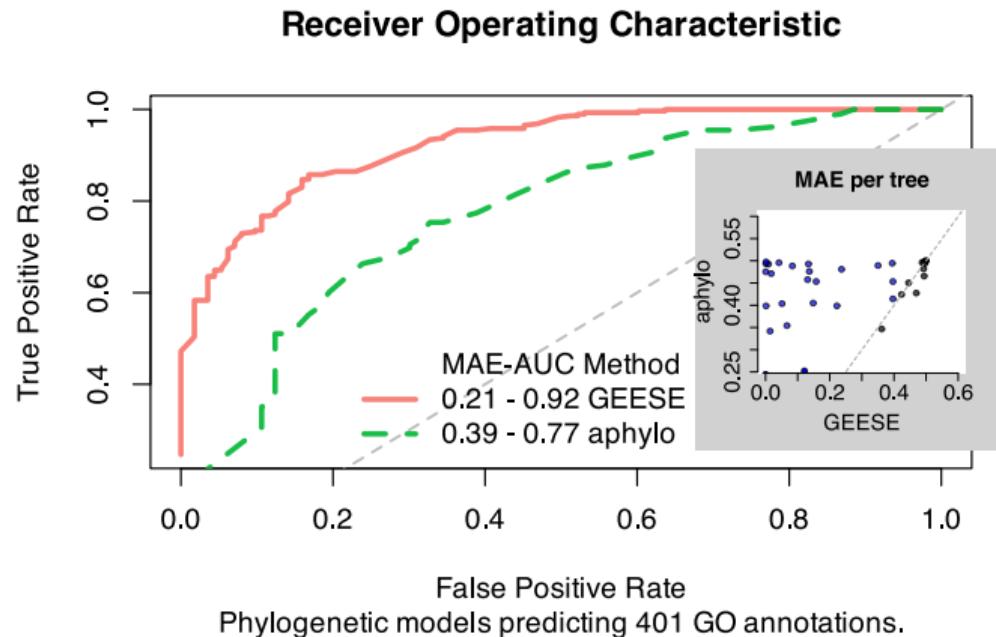
How much can we gain from a joint dist. model?

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept



Just controlling for preservation (having only one duplicate changing) significantly improves our predictions.

Table of Contents

Preliminaries

Preliminaries

Evolution of Gene
Function

Evolution of Gene Function

Mechanistic Machine
Learning

Mechanistic Machine Learning

Proof of Concept

Proof of Concept

Mechanistic Machine Learning: State-of-the-art

- ▶ After all the data pouring, attention to causal inference and mechanistic models is coming back^{3,13}
- ▶ Applications in Physics, Chemistry, and Biology^{16,10,1} show the benefits of combining the two approaches.

Mechanistic Models

- ▶ Inference driven (causality)
- ▶ Great for small datasets
- ▶ Not the most accurate

Machine Learning Models

- ▶ Data driven (prediction)
- ▶ Lots of points to “learn”
- ▶ Great for big data

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Potential Strategies

- ▶ Use machine learning to learn the errors of a mechanistic model.
- ▶ Add constraints to the ML algorithm based on a mechanistic model.
- ▶ Add mechanistic predictions as a feature of a machine learning model.

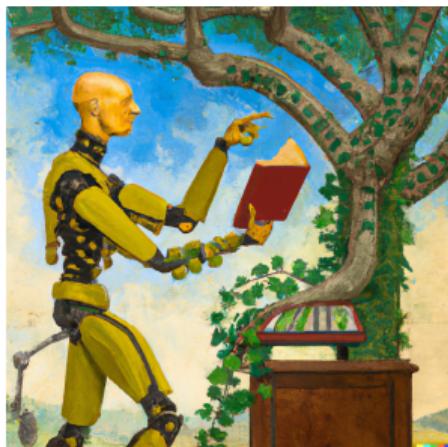


Figure: “A van Gogh-style painting of an android holding a large biology book in one hand and a computer in another, examining an evolutionary tree that, instead of leaves, have genes.”—DALL-E’s interpretation of my description ([link](#))

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

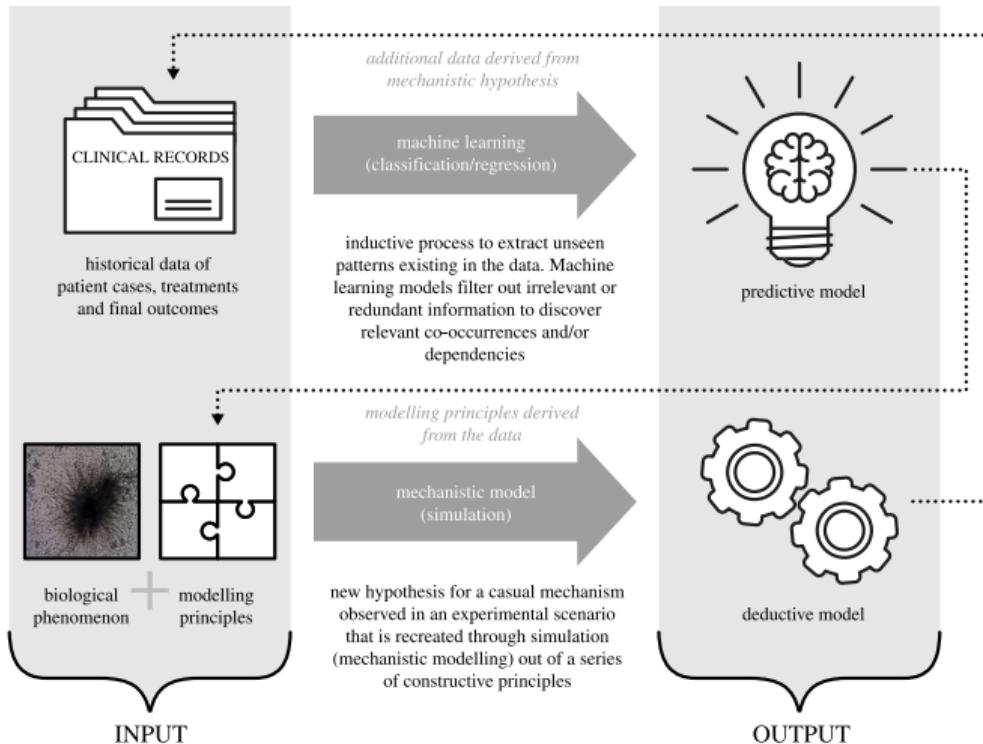


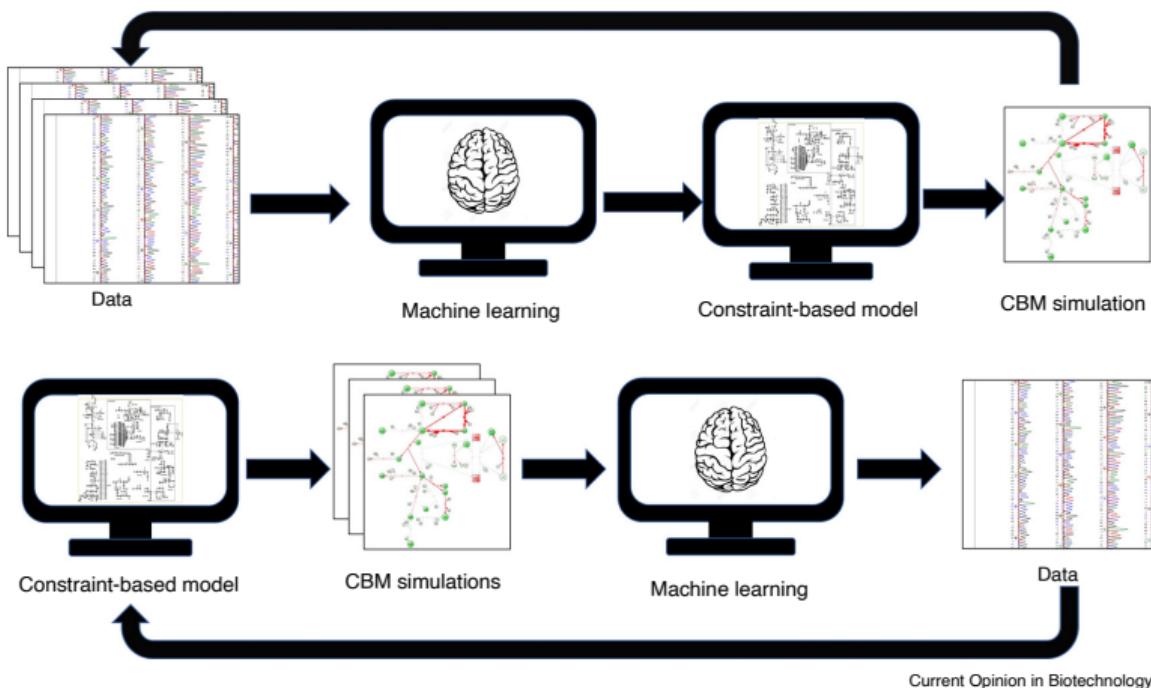
Figure: “The inputs and outputs from machine learning and mechanistic modelling approaches, and the potential for synergy between the two.” (Figure 1 reproduced directly from Baker et al. [3].

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept



Current Opinion in Biotechnology

Figure: “**Top:** Machine learning is applied to the input data to identify the important features for constructing reduced order constraint-based models; the CBM simulations can be iteratively matched with input data for convergence until the proper set of features are identified. **Bottom:** Machine learning is iteratively applied to CBM simulations to reconcile with experimental data. Interplay between the Top and Bottom parts can iteratively lead to convergence between CBM simulations, experimental data and machine learning based predictions.” (Figure 2 reproduced directly from Rana et al. [14])

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

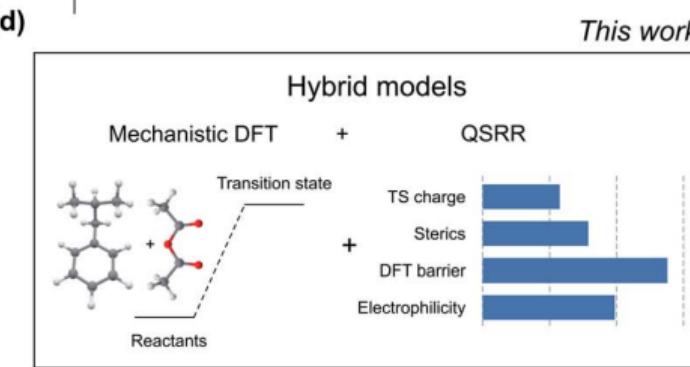
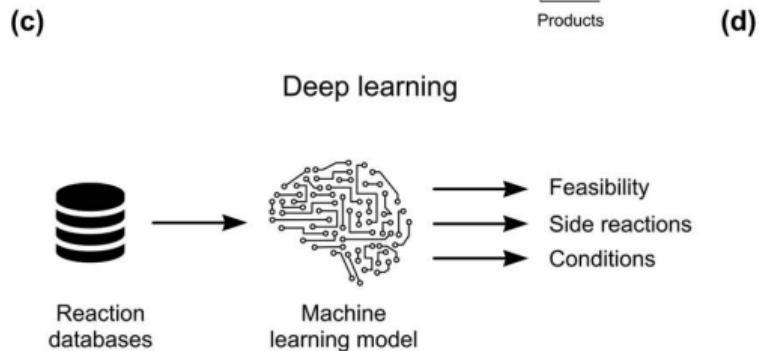
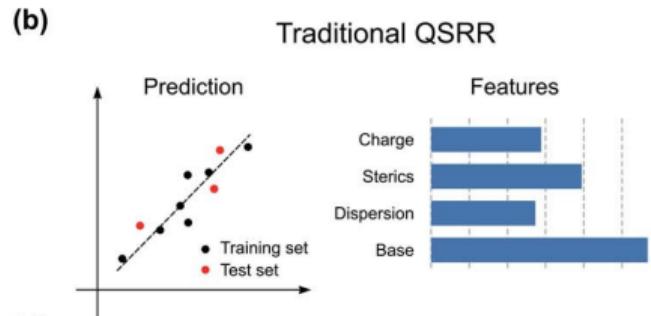
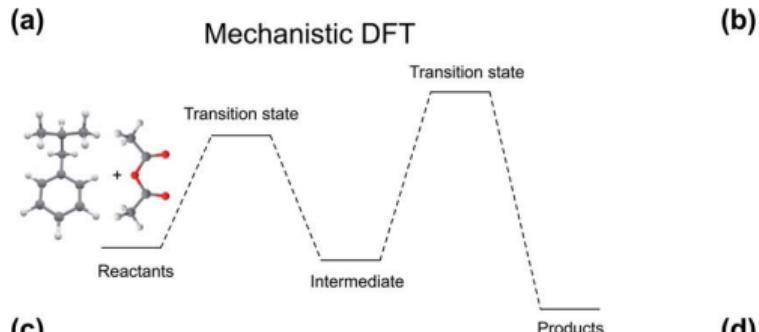


Figure: “Different types of quantitative reaction prediction approaches. Mechanistic DFT (a) and QSRR (b) are the current gold standard methods. Deep learning models (c) are emerging as an alternative. Hybrid models (d) combine mechanistic DFT modelling with traditional QSRR” (Figure 2 reproduced directly from Jorner et al. [10])

Table of Contents

Preliminaries

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Beyond GO and Trees... Bgee

The **Bgee** project “is a **database** for retrieval and **comparison of gene expression** patterns **across multiple animal species**. It provides an intuitive answer to the question ‘where is a gene expressed?’ and supports research in cancer and agriculture as well as evolutionary biology.” – Bastian et al. [4]

- ▶ Raw expression annotations.
- ▶ Standardized expression scores (so can compare across species/tissues).
- ▶ And also yes/no expression annotations based on the standardized scores.

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Beyond GO and Trees... Bgee (cont.)

Divergence across species in gene expression levels has been linked to evolutionary events,^{12,8} i.e., expression levels clustered phylogenies.

Thinking of different ways to use it as:

- ▶ As an additional feature for our model: “Given the phylo, observed annotations, **and expression levels in n tissues**, . . . ”
- ▶ As 0/1 variable (expression is present/absent) to predict in our model: “Model the evolution of gene function **and expression**. ”
- ▶ As part of a prediction model in, say, a Machine Learning Model.

What went into the blender

Data Feats

- ▶ Bgee 15 dataset: approx 7 billion annotations for 1.5 million genes.
- ▶ Our dataset: 1,484 predictions for 1,318 genes.
- ▶ Search by Gene name: 9,923,427 Bgee annotations.

Final model

- ▶ 10 GO terms (in a full-Markov model, this is 1 MM params).
- ▶ 278 annotations for 256 genes.
- ▶ 10 GEESE predictions for each gene.
- ▶ 46 Bgee score for gene expression computed as **mean expression score by gene-genus**

GO terms: GO:0004672, GO:0004713, GO:0004867, GO:0005730, GO:0005829, GO:0005886, GO:0006468, GO:0009408, GO:0015020, GO:0060070

Genus: Anguilla, Anolis, Astatotilapia, Astyanax, Bos, Branchiostoma, Caenorhabditis, Callithrix, Canis, Capra, Cavia, Coccocebus, Chlorocebus, Danio, Drosophila, Equus, Esox, Felis, Gadus, Gallus, Gasterosteus, Gorilla, Heterocephalus, Homo, Latimeria, Lepisosteus, Macaca, Manis, Meleagris, Microcebus, Monodelphis, Mus, Neolamprologus, Nothobranchius, Ornithorhynchus, Oryctolagus, Oryzias, Ovis, Pan, Papio, Poecilia, Rattus, Salmo, Scophthalmus,

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

Mechanistic ML

We are comparing three models:

GEESE

Bgee

GEESE + Bgee

Phylogenetic based
predictions (evolution of
gene function)

Linear Prob. model using
expression as predictors.

Linear Prob. model using
expression as predictors
and predictions made by
GEESE.

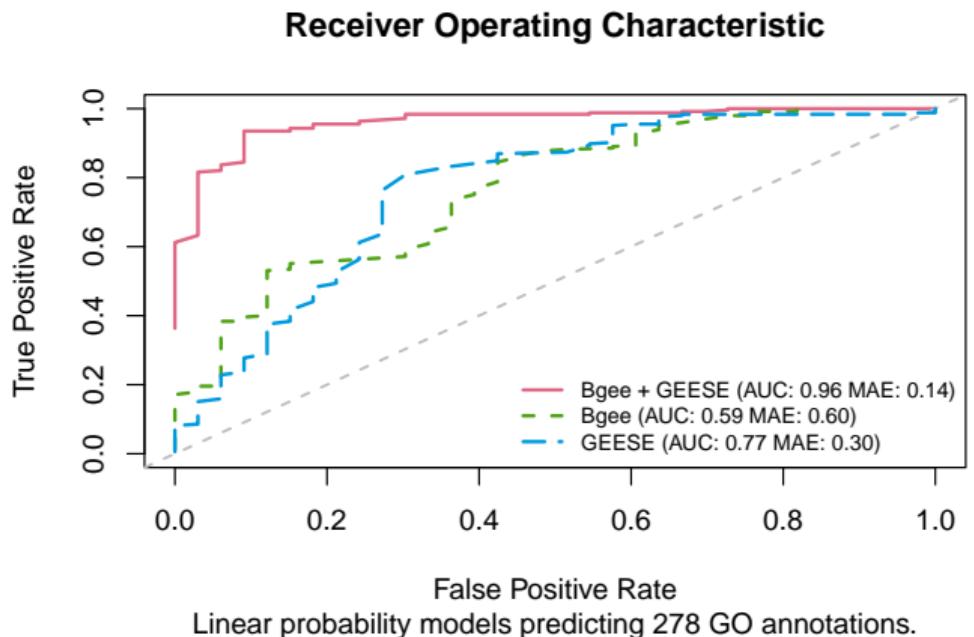
Mechanistic ML (prelim res.)

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept



Both AUC and MAE were computed only using predictions for which we knew the true value.

Discussion

Gene function

- ▶ We are racing to discover what genes do.
- ▶ Experimental assessment is expensive (money and time,) → automatic annotations.
- ▶ Many ways to do it (seq. homology, evolutionary theory, ML, etc.)
- ▶ The best methods use ML (pattern discovery)... but none (AFAIK) are based on bio. theory.

Evol. Model

- ▶ We proposed an Evolutionary model of Gene Function.
- ▶ This new model, GEESE, uses sufficiency to reduce “Markov complexity.”
- ▶ We showed it really helps.

Mechanistic ML

- ▶ Mechanistic Machine Learning (mixing theory-based models with ML) promises improved predictions.
- ▶ I showed an application using gene expression (Bgee).
- ▶ Adding our mechanistic predictions (based on GEESE) boosted quality

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

Discussion (cont.)

Preliminaries

Evolution of Gene
Function

Mechanistic Machine
Learning

Proof of Concept

- ▶ This is the core of an R01 first submitted in Feb 2022.
- ▶ The original version did not feature any ML, just the mechanistic part.
- ▶ Two important critiques: “Not leveraging large genomic data” and “Not using Neural Networks”
- ▶ We believe we are addressing both using gene expression data (Bgee) and Mechanistic ML.
- ▶ ...your thoughts?

Thank you!

Predicting of Gene Functions by Leveraging Biological Insights with Mechanistic Machine Learning



George G. Vega Yon, Ph.D.
george.vegayon@utah.edu
Division of Epidemiology @ University of Utah

Dec 8th, 2022 @ Quinlan Lab

Preliminaries

Evolution of Gene Function

Mechanistic Machine Learning

Proof of Concept

- [1] Nada Al taweraqi and Ross D. King. "Improved Prediction of Gene Expression through Integrating Cell Signalling Models with Machine Learning". In: BMC Bioinformatics 23.1 (Aug. 6, 2022), p. 323. ISSN: 1471-2105. DOI: 10.1186/s12859-022-04787-8. URL: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04787-8> (visited on 12/07/2022).
- [2] Stephen F. Altschul et al. "Basic Local Alignment Search Tool". In: Journal of Molecular Biology 215.3 (Oct. 1990), pp. 403–410. ISSN: 00222836. DOI: 10.1016/S0022-2836(05)80360-2. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602> (visited on 12/08/2022).
- [3] Ruth E. Baker et al. "Mechanistic Models versus Machine Learning, a Fight Worth Fighting for the Biological Community?" In: Biology Letters 14.5 (May 2018), p. 20170660. ISSN: 1744-9561, 1744-957X. DOI: 10.1098/rsbl.2017.0660. URL: <https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0660> (visited on 12/01/2022).

- [4] Frederic B Bastian et al. "The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals". In: Nucleic Acids Research 49.D1 (Jan. 8, 2021), pp. D831–D847. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa793. URL: <https://academic.oup.com/nar/article/49/D1/D831/5920517> (visited on 12/06/2022).
- [5] B.E. Engelhardt et al. "Genome-scale phylogenetic function annotation of large and diverse protein families". In: Genome research 21.11 (2011), pp. 1969–1980. DOI: 10.1101/gr.104687.109.
- [6] B.E. Engelhardt et al. "Protein Molecular Function Prediction by Bayesian Phylogenomics". In: PLOS Computational Biology 1.5 (2005). DOI: 10.1371/journal.pcbi.0010045.

- [7] Vladimir Gligorijević et al. "Structure-Based Protein Function Prediction Using Graph Convolutional Networks". In: Nature Communications 12.1 (May 26, 2021), p. 3168. ISSN: 2041-1723. DOI: 10.1038/s41467-021-23303-9. URL: <https://www.nature.com/articles/s41467-021-23303-9> (visited on 12/08/2022).
- [8] Andrea Hodgins-Davis and Jeffrey P. Townsend. "Evolving Gene Expression: From G to E to GxE". In: Trends in Ecology & Evolution 24.12 (Dec. 2009), pp. 649–658. ISSN: 01695347. DOI: 10.1016/j.tree.2009.06.011. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0169534709002213> (visited on 11/16/2022).
- [9] Aashish Jain and Daisuke Kihara. "Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences". In: Bioinformatics 35 (5 Mar. 2019), pp. 753–759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty704.

References IV

- [10] Kjell Jorner et al. "Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies". In: Chemical Science 12.3 (2021), pp. 1163–1175. ISSN: 2041-6520, 2041-6539. DOI: 10.1039/DOSC04896H. URL: <http://xlink.rsc.org/?DOI=DOSC04896H> (visited on 12/07/2022).
- [11] Maxat Kulmanov and Robert Hohendorf. "DeepGOPlus: Improved Protein Function Prediction from Sequence". In: Bioinformatics (July 27, 2019). Ed. by Lenore Cowen, btz595. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz595. URL: <https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz595/5539866> (visited on 11/17/2022).

References V

- [12] B. Nabholz, H. Ellegren, and J. B. W. Wolf. "High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes". In: Molecular Biology and Evolution 30.2 (Feb. 1, 2013), pp. 272–284. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/mss238. URL: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss238> (visited on 11/16/2022).
- [13] Judea Pearl. "The Seven Tools of Causal Inference, with Reflections on Machine Learning". In: Communications of the ACM 62.3 (Feb. 21, 2019), pp. 54–60. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3241036. URL: <https://dl.acm.org/doi/10.1145/3241036> (visited on 11/18/2022).
- [14] Pratip Rana et al. "Recent Advances on Constraint-Based Models by Integrating Machine Learning". In: Current Opinion in Biotechnology 64 (Aug. 2020), pp. 85–91. ISSN: 09581669. DOI: 10.1016/j.copbio.2019.11.007. URL: <https://linkinghub.elsevier.com/retrieve/pii/S095816691930117X> (visited on 12/06/2022).

References VI

- [15] G.G. Vega Yon et al. "Bayesian parameter estimation for automatic annotation of gene functions using observational data and phylogenetic trees". In: PLoS Comput Biol 17 (2021), e1007948. DOI: [10.1371/journal.pcbi.1007948](https://doi.org/10.1371/journal.pcbi.1007948).
- [16] Jared Willard et al. "Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems". In: ACM Computing Surveys (Mar. 25, 2022), p. 3514228. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3514228. URL: <https://dl.acm.org/doi/10.1145/3514228> (visited on 11/11/2022).
- [17] Ronghui You et al. "GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank". In: Bioinformatics 34.14 (July 15, 2018). Ed. by Jonathan Wren, pp. 2465–2473. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bty130. URL: <https://academic.oup.com/bioinformatics/article/34/14/2465/4924212> (visited on 11/17/2022).

References VII

- [18] Naihui Zhou et al. “The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens”. In: Genome Biology 20 (1 2019), pp. 1–23. ISSN: 1474760X. DOI: [10.1186/s13059-019-1835-8](https://doi.org/10.1186/s13059-019-1835-8).

Tree likelihoods: Felsenstein's Pruning algorithm

All possible transitions from x_n → Transition Probability (ERGM)

$$\mathbb{P}(\tilde{D}_n \mid x_n, \Theta) = \sum_x \mathbb{P}(x \mid x_n) \prod_{m \in O(n)} \mathbb{P}(\tilde{D}_m \mid x_m)$$

Model Parameters → Vector of Sufficient Statistics

$$\mathbb{P}(x \mid x_n) = \frac{\exp\{\Theta^t s(x, x_n)\}}{\sum_{x'} \exp\{\Theta^t s(x', x_n)\}}$$

Normalizing Constant → the *lingua franca* of SNA

... I implemented this (and more) on **barry**

Some computational features of barry

