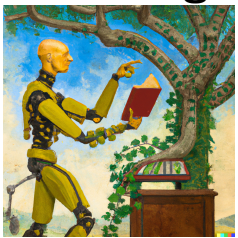
Predicting of Gene Functions by Leveraging Biological Insights with Mechanistic Machine Learning



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

May 3rd, 2023 @ USC IMAGE

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

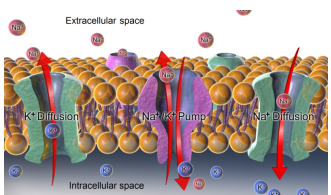
Table of Contents

You can download the slides from `https://ggv.cl/image2023`

Table of Contents

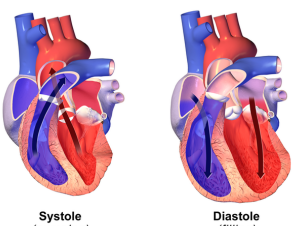
Encode the synthesis of genetic products that ultimately are related to a particular aspect of life, for example

Molecular function
Active transport GO:0005215



Cellular component
Mitochondria GO:0004016

Biological process
Heart contraction GO:0060047



- 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`

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[Jain and Kihara 2019] A BLAST-based adding phylogenetic based distances.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[Jain and Kihara 2019] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[Kulmanov and Hoehndorf 2019] One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[Jain and Kihara 2019] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[Kulmanov and Hoehndorf 2019] One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[You et al. 2018] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[Zhou et al. 2019a] is an ensemble of various simple ML methods, including K-means and logistic regression.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[Jain and Kihara 2019] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[Kulmanov and Hoehndorf 2019] One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[You et al. 2018] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[Zhou et al. 2019a] is an ensemble of various simple ML methods, including K-means and logistic regression.
- **DeepFRI**:[Gligorijević et al. 2021] Uses Graph Convolutional Neural Networks (GCNs) to predict function based on protein structure and genetic sequence.

Sequences, phylogenomics, and ML.

- **BLAST**:[Altschul et al. 1990] Prediction by sequence homology ($\sim$ 105,000 citations).
- **SIFTER**:[Engelhardt, Jordan, Muratore, et al. 2005],[Engelhardt, Jordan, Srouji, et al. 2011] An evolutionary model of gene function/loss using phylogenetics.
- **aphylo**[Vega Yon et al. 2021] (by yours truly): Another phylo-based method. Leverages negative annotations and pooled trees.
- **Phylo-PFP**:[Jain and Kihara 2019] A BLAST-based adding phylogenetic based distances.
- **DeepGOPlus**:[Kulmanov and Hoehndorf 2019] One of the top-performing models in the literature, uses a 2D convolutional neural network on gene sequences.
- **GOLabeler**:[You et al. 2018] Top performing tool according to the *Critical Assessment of Function Annotation* [CAFA] challenge,[Zhou et al. 2019a] is an ensemble of various simple ML methods, including K-means and logistic regression.
- **DeepFRI**:[Gligorijević et al. 2021] Uses Graph Convolutional Neural Networks (GCNs) to predict function based on protein structure and genetic sequence.

In the latest CAFA, **none** of the top-performing methods scored an AUC above 0.60, and most were outperformed by BLAST,[Zhou et al. 2019b] which annotates using homology based on sequence similarity.

Table of Contents

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

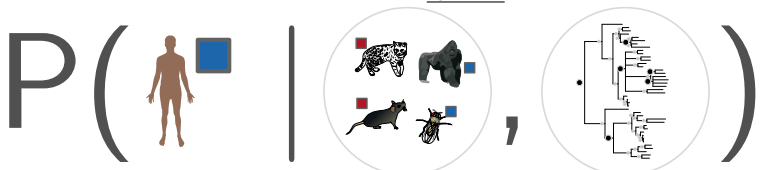
RESEARCH ARTICLE

Bayesian parameter estimation for automatic annotation of gene functions using observational data and phylogenetic trees

George G. Vega Yon, Duncan C. Thomas, John Morrison, Huaiyu Mi, Paul D. Thomas, Paul Marjoram

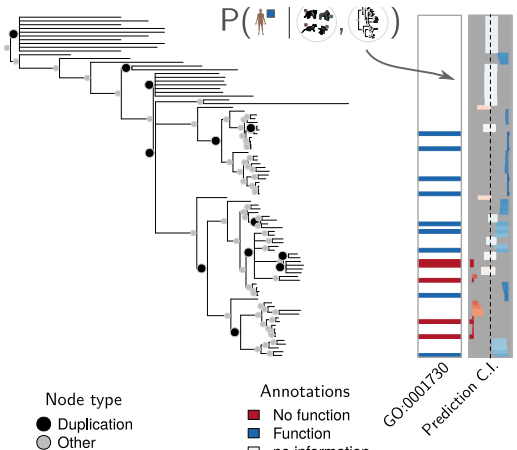
Version 2 Published: February 18, 2021 • https://doi.org/10.1371/journal.pcbi.1007948

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



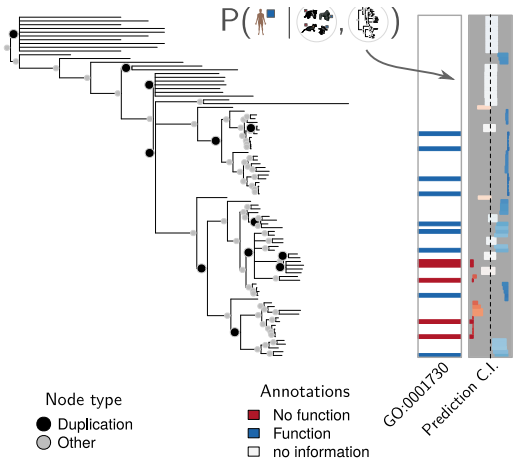
Built a big model (lots of trees and annotations) called aphylo:

- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).



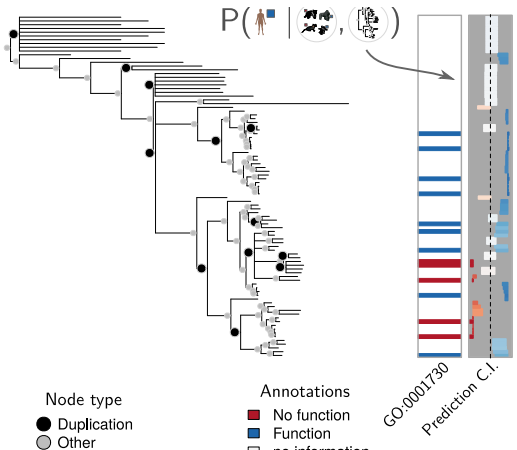
Built a big model (lots of trees and annotations) called aphylo:

- Only two sources of data: Phylogenetic tree (`pantherdb.org`) and functional annotations (`geneontology.org`).
- Leverage negative annotation of GO terms (NOT).



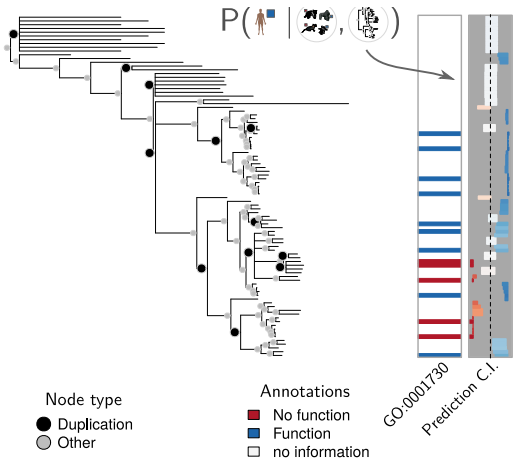
Built a big model (lots of trees and annotations) called aphylo:

- 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.



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?

Tapping into Evol. Theory (part of Proj. 3)

- A fundamental part of Fun. Evol. is Duplication Events.

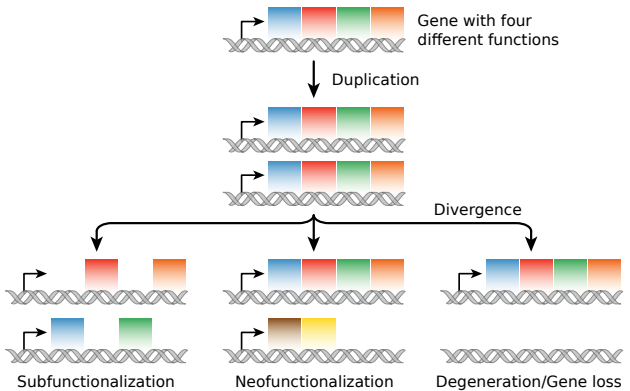


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Tapping into Evol. Theory (part of Proj. 3)

- 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.

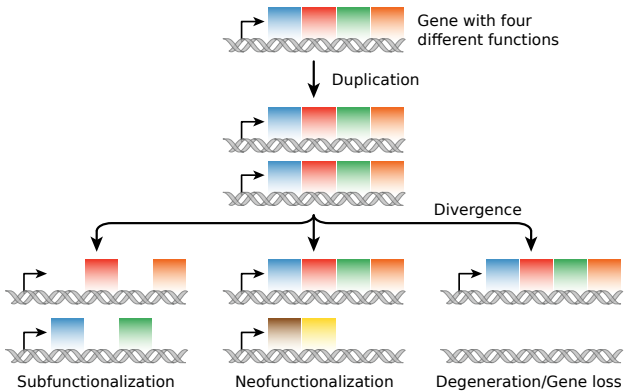


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

Tapping into Evol. Theory (part of Proj. 3)

- 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).

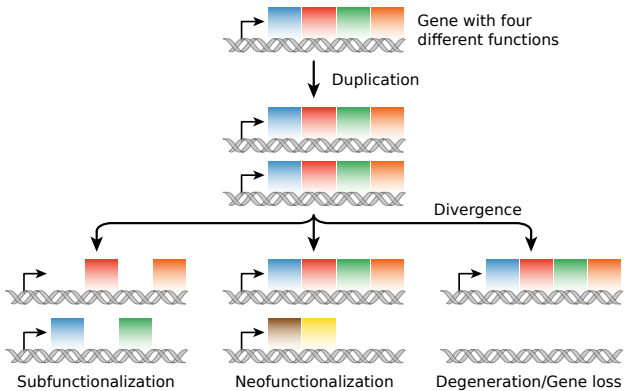
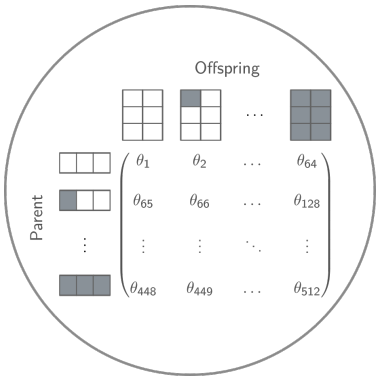


Figure: A key part of molecular innovation, gene duplication provides an opportunity for new functions to emerge (wikimedia)

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

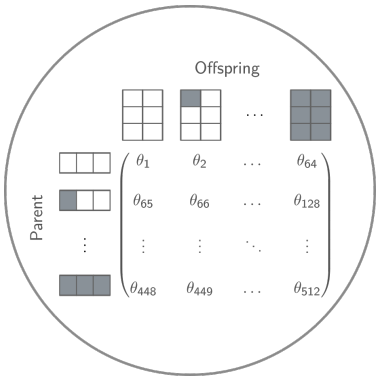
Full Markov Transition Matrix



512 parameters

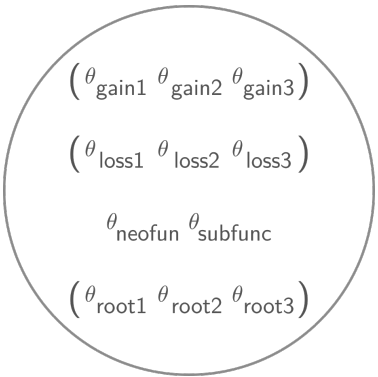
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



512 parameters

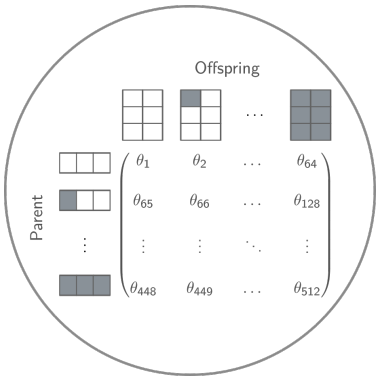
Sufficient statistics



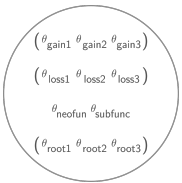
11 parameters (for example)

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

Full Markov Transition Matrix

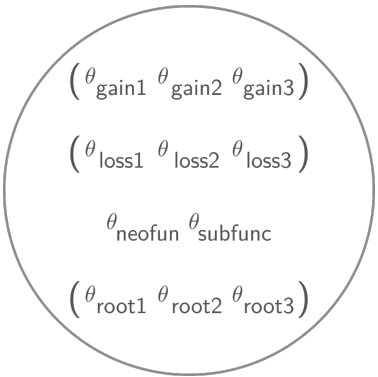


512 parameters

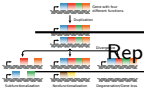


Easier to fit
Easier to interpret

Sufficient statistics



11 parameters (for example)



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	$(1 - x_p^k)\mathbf{1}\left(x_m^k : m = \text{argmax}_n \text{blength}_n\right)$

Table: *Example of sufficient statistics for evolutionary transitions.* $x_n^i \in \{0, 1\}$, equal to 1 if the function i is present in gene n. The p subscript denotes parent gene.

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?

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.

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.

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.*, only one offspring changes)

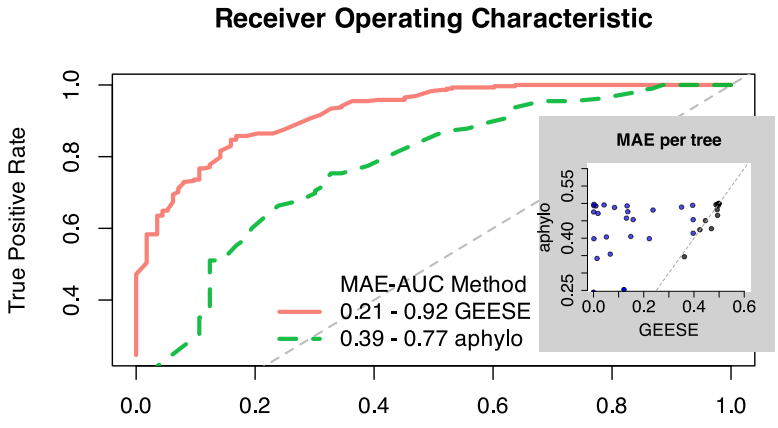
I implemented what I just described in a C++ library with a companion R package called geese. The question is: How much do we earn by using these motifs?

- 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.*, only one offspring changes)
- Fitted both of them using MCMC.

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.*, only one offspring changes)
- Fitted both of them using MCMC.
- Used LOO cross-validation to compute aggregated AUCs and MAE.

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



Phylogenetic models predicting 401 GO annotations.

Table of Contents

After all the data pouring, attention to causal inference and mechanistic models is coming back[Baker et al. 2018],[Pearl 2019]

Mechanistic Models

- Inference-driven (causality).
- Great for small datasets.
- Knowledge beyond the observed data.

Machine Learning Models

- Data-driven (prediction).
- Great for big data.
- Finds hidden knowledge in observed data.

Ways in which it's been applied:

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]
- Using pathway-networks to add "external knowledge" as features.[Al taweraqi and King 2022]

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]
- Using pathway-networks to add "external knowledge" as features.[Al taweraqi and King 2022]
- Creating a loss-function with a mechanistic penalty for modeling tumor cell-density[Gaw et al. 2019]

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]
- Using pathway-networks to add "external knowledge" as features.[Al taweraqi and King 2022]
- Creating a loss-function with a mechanistic penalty for modeling tumor cell-density[Gaw et al. 2019]
- and more...[Jia et al. 2021],[Jorner et al. 2021],[von Rueden et al. 2023],[Willard et al. 2022]

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]
- Using pathway-networks to add "external knowledge" as features.[Al taweraqi and King 2022]
- Creating a loss-function with a mechanistic penalty for modeling tumor cell-density[Gaw et al. 2019]
- and more...[Jia et al. 2021],[Jorner et al. 2021],[von Rueden et al. 2023],[Willard et al. 2022]

Important: Mechanistic Machine Learning **is not** domain-knowledge aided feature engineering. You need a whole other model to complement the ML algorithm.

Ways in which it's been applied:

- Adjusting errors in mechanistic-based prediction models (like ABMs).[Compagni et al. 2022]
- Incorporating mechanistically inferred data as additional -omics layer.[Zampieri et al. 2019]
- Using pathway-networks to add "external knowledge" as features.[Al taweraqi and King 2022]
- Creating a loss-function with a mechanistic penalty for modeling tumor cell-density[Gaw et al. 2019]
- and more...[Jia et al. 2021],[Jorner et al. 2021],[von Rueden et al. 2023],[Willard et al. 2022]

Important: Mechanistic Machine Learning **is not** domain-knowledge aided feature engineering. You need a whole other model to complement the ML algorithm.

Important 2: This isn't just ML-ensemble, you need to have an ML and a Mech 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)

a. **ML Correction**: Use machine learning to learn the errors of a mechanistic 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)

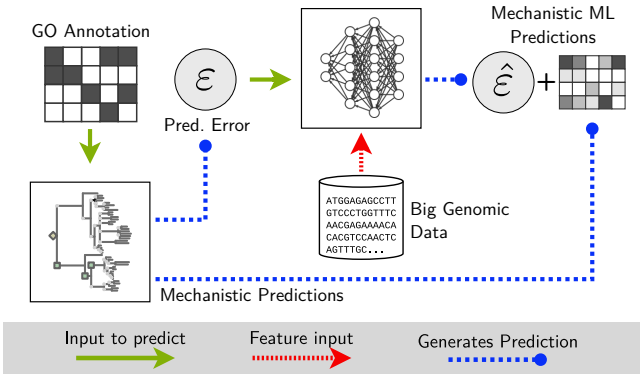
a. **ML Correction**: Use machine learning to learn the errors of a mechanistic model.
b. **Mechanistic Feature**: Add mechanistic predictions as a feature of a machine learning model.

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)

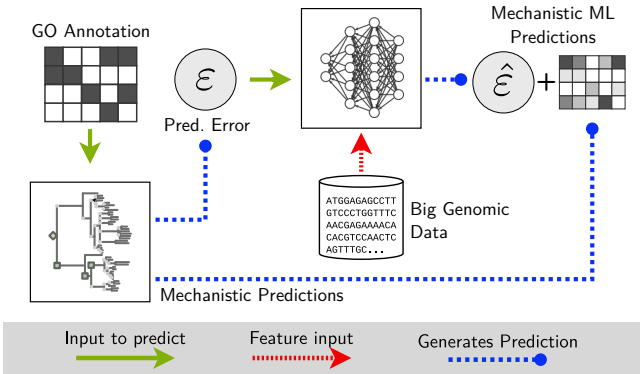
a. **ML Correction**: Use machine learning to learn the errors of a mechanistic model.
b. **Mechanistic Feature**: Add mechanistic predictions as a feature of a machine learning model.
c. **Mechanistic Penalty**: Add constraints to the ML algorithm based on a mechanistic model.

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)

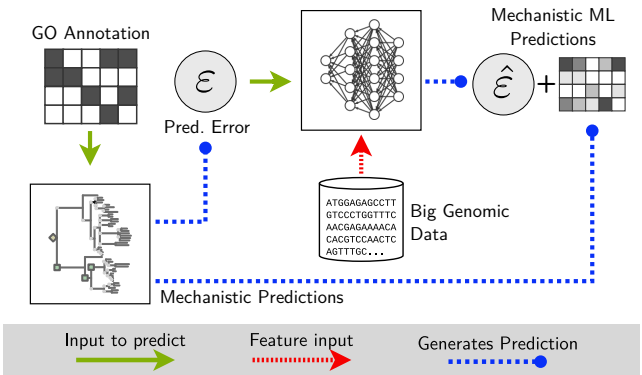
1. Fit the mechanistic model using GEESE



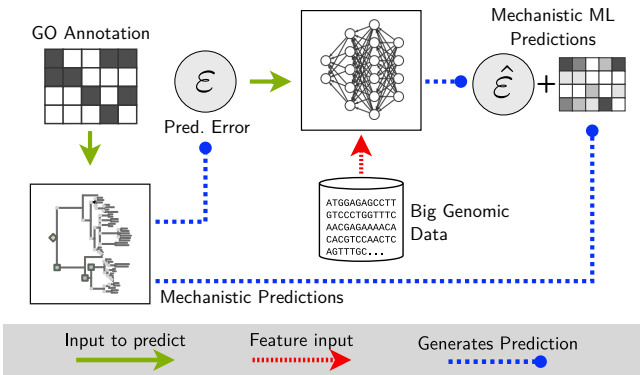
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,



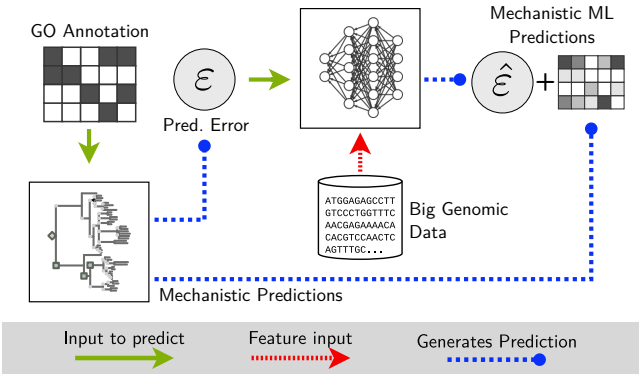
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,



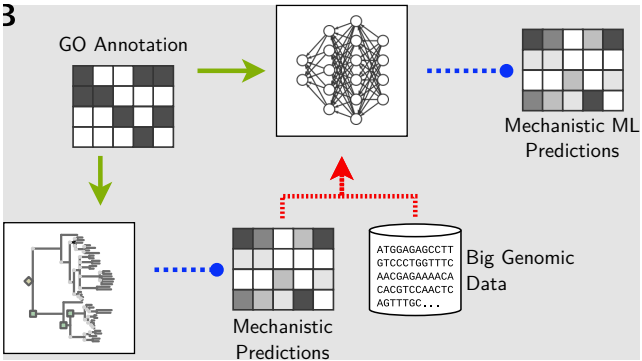
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,
4. generate the predictions of $\hat{\varepsilon}$, and



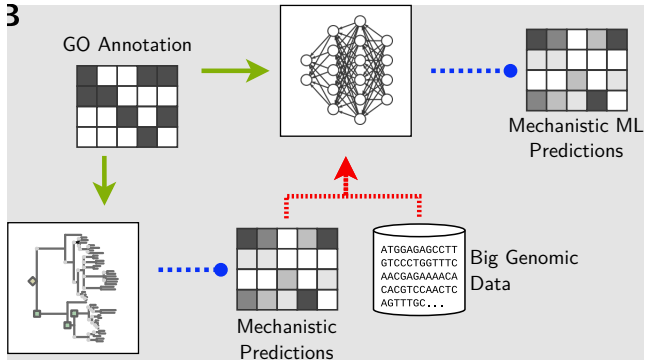
1. Fit the mechanistic model using GEESE
2. Generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model $f(X, \Omega)$ to predict $\varepsilon \equiv (\mathbf{y} - \hat{\mathbf{y}}^{GEESE})$,
4. generate the predictions of $\hat{\varepsilon}$, and
5. Compute the Mechanistic-ML predictions as $\hat{\mathbf{y}}^{MML1} \equiv \hat{\mathbf{y}}^{GEESE} + \hat{\varepsilon}$



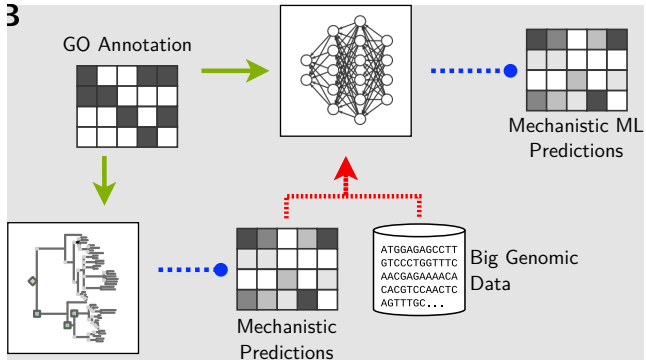
1. Fit the mechanistic model using GEESE,



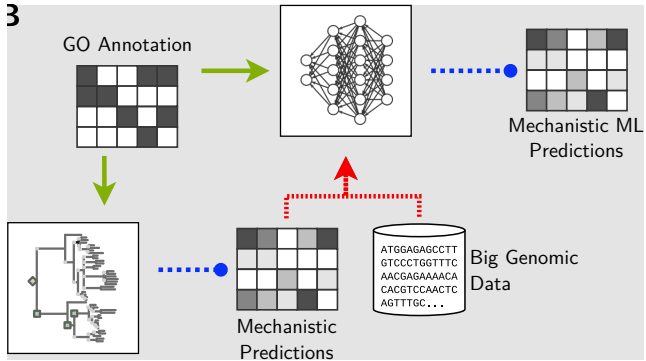
1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,



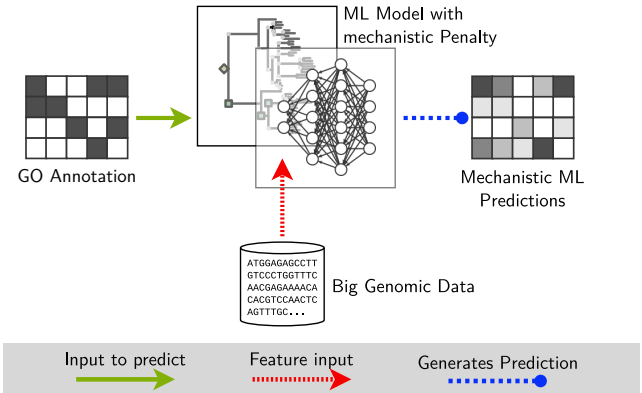
1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model that uses the mechanistic predictions as features, $f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$, and



1. Fit the mechanistic model using GEESE,
2. generate the mechanistic-based predictions, $\hat{\mathbf{y}}^{GEESE}$,
3. fit an ML model that uses the mechanistic predictions as features, $f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$, and
4. Compute the Mechanistic-ML predictions as $\hat{\mathbf{y}}^{MML2} \equiv f(X, \Omega, \hat{\mathbf{y}}^{GEESE})$



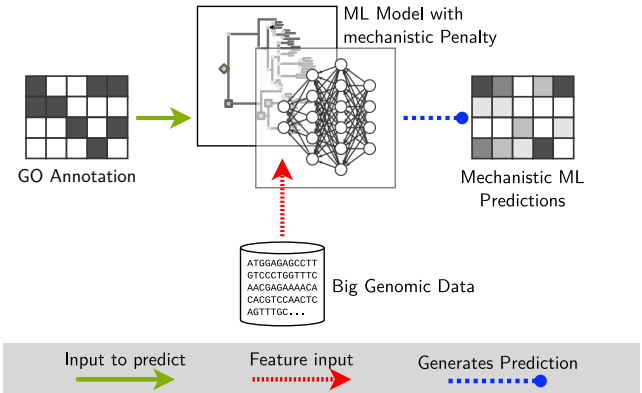
1. Fit the mechanistic model using GEESE and store the parameter estimates $\hat{\theta}$,



1. Fit the mechanistic model using GEESE and store the parameter estimates $\hat{\theta}$,
2. minimize the following loss function:

$$L(y_{,}^{obs}X,\Omega) - \mathcal{L}\left(f(y_{,}^{obs}X,\Omega)\right)_{GEESE},$$

where $\mathcal{L}(\cdot)$ is the likelihood function under GEESE.



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?'[.]" – Bastian et al. (2021)

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

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?'[.]" – Bastian et al. (2021)

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

Divergence across species in gene expression levels has been linked to evolutionary events,[Hodgins-Davis and Townsend 2009],[Nabholz, Ellegren, and Wolf 2013] *i.e.*, expression levels clustered phylogenies.

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.

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 $2^{3\times 10} \sim$1 billion 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**

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 $2^{3\times 10}$ $\sim$1 billion 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, Cercocebus, 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, Sus, Xenopus

We are comparing three models:

GEESE

Phylogenetic based predictions
(evolution of gene function)

We are comparing three models:

GEESE

Phylogenetic based predictions
(evolution of gene function)

Bgee

Linear Prob. model using
expression as predictors.

We are comparing three models:

GEESE

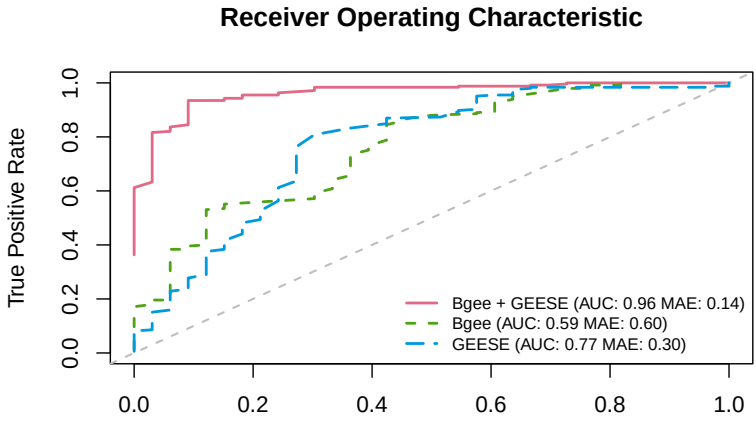
Phylogenetic based predictions (evolution of gene function)

Bgee

Linear Prob. model using expression as predictors.

GEESE + Bgee

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



Linear probability models predicting 278 GO annotations.

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

This talk has been submitted as an R01 to the National Human Genome Research Institute [NHGRI]

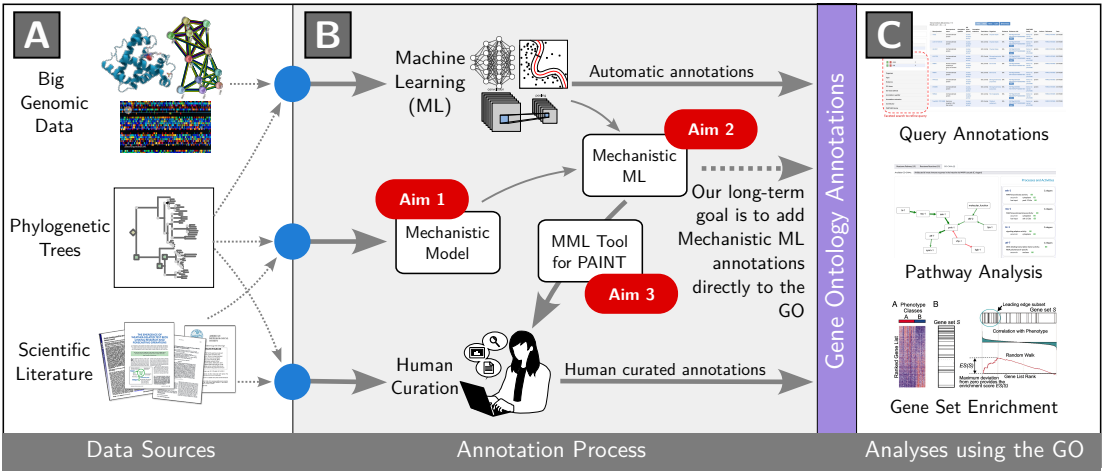


Figure: Building a Novel Prediction Framework Leveraging Biological Insights to Boost Machine Learning Algorithms for Annotating Gene Function

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) $\rightarrow$ 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.

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) $\rightarrow$ 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.

Gene function

- We are racing to discover what genes do.
- Experimental assessment is expensive (money and time,) $\rightarrow$ 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

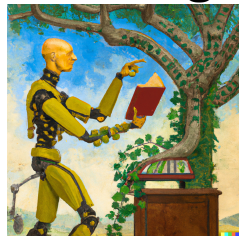
- We proposed an Evolutionary model of Gene Function.
- This new model, GEESE, uses sufficiency to reduce "Markov complexity."
- We showed it really helps.

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

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

- Al taweraqi, Nada and Ross D. King (Aug. 6, 2022). "Improved Prediction of Gene Expression through Integrating Cell Signalling Models with Machine Learning". In: BMC Bioinformatics 23.1, 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).
- Altschul, Stephen F. et al. (Oct. 1990). "Basic Local Alignment Search Tool". In: Journal of Molecular Biology 215.3, 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).
- Baker, Ruth E. et al. (May 2018). "Mechanistic Models versus Machine Learning, a Fight Worth Fighting for the Biological Community?" In: Biology Letters 14.5, 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).

- Bastian, Frederic B et al. (Jan. 8, 2021). "The Bgee Suite: Integrated Curated Expression Atlas and Comparative Transcriptomics in Animals". In: Nucleic Acids Research 49.D1, 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).
- Compagni, Riccardo Delli et al. (Mar. 2022). "A Hybrid Neural Network-SEIR Model for Forecasting Intensive Care Occupancy in Switzerland during COVID-19 Epidemics". In: PLOS ONE 17.3, e0263789. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0263789. (Visited on 05/04/2023).
- Engelhardt, B.E., M.I. Jordan, K.E. Muratore, et al. (2005). "Protein Molecular Function Prediction by Bayesian Phylogenomics". In: PLOS Computational Biology 1.5. DOI: 10.1371/journal.pcbi.0010045.
- Engelhardt, B.E., M.I. Jordan, J.R. Srouji, et al. (2011). "Genome-scale phylogenetic function annotation of large and diverse protein families". In: Genome research 21.11, pp. 1969–1980. DOI: 10.1101/gr.104687.109.

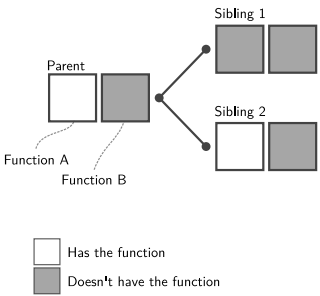
- Gaw, Nathan et al. (Dec. 2019). "Integration of Machine Learning and Mechanistic Models Accurately Predicts Variation in Cell Density of Glioblastoma Using Multiparametric MRI". In: Scientific Reports 9.1, p. 10063. ISSN: 2045-2322. DOI: 10.1038/s41598-019-46296-4. URL: http://www.nature.com/articles/s41598-019-46296-4 (visited on 12/09/2022).
- Gligorijević, Vladimir et al. (May 26, 2021). "Structure-Based Protein Function Prediction Using Graph Convolutional Networks". In: Nature Communications 12.1, 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).
- Hodgins-Davis, Andrea and Jeffrey P. Townsend (Dec. 2009). "Evolving Gene Expression: From G to E to G×E". In: Trends in Ecology & Evolution 24.12, 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).
- Jain, Aashish and Daisuke Kihara (Mar. 2019). "Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences". In: Bioinformatics 35 (5), pp. 753–759. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty704.

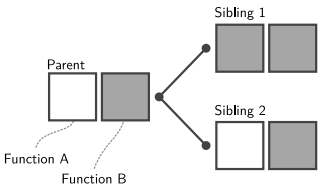
- Jia, Xiaowei et al. (May 17, 2021). "Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles". In: ACM/IMS Transactions on Data Science 2.3, pp. 1–26. ISSN: 2691-1922. DOI: 10.1145/3447814. URL: https://dl.acm.org/doi/10.1145/3447814 (visited on 11/11/2022).
- Jorner, Kjell et al. (2021). "Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies". In: Chemical Science 12.3, pp. 1163–1175. ISSN: 2041-6520, 2041-6539. DOI: 10.1039/D0SC04896H. URL: http://xlink.rsc.org/?DOI=D0SC04896H (visited on 12/07/2022).
- Kulmanov, Maxat and Robert Hoehndorf (July 27, 2019). "DeepGOPlus: Improved Protein Function Prediction from Sequence". In: Bioinformatics. 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).

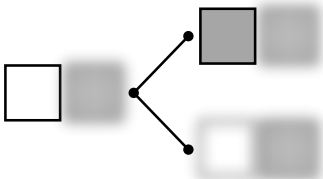
- Nabholz, B., H. Ellegren, and J. B. W. Wolf (Feb. 1, 2013). "High Levels of Gene Expression Explain the Strong Evolutionary Constraint of Mitochondrial Protein-Coding Genes". In: Molecular Biology and Evolution 30.2, pp. 272–284. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/mss238. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss238 (visited on 11/16/2022).
- Pearl, Judea (Feb. 21, 2019). "The Seven Tools of Causal Inference, with Reflections on Machine Learning". In: Communications of the ACM 62.3, 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).
- Vega Yon, G.G. et al. (2021). "Bayesian parameter estimation for automatic annotation of gene functions using observational data and phylogenetic trees". In: PLoS Comput Biol 17, e1007948. DOI: 10.1371/journal.pcbi.1007948.
- von Rueden, Laura et al. (Jan. 2023). "Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems". In: IEEE Transactions on Knowledge and Data Engineering 35.1, pp. 614–633. ISSN: 1558-2191. DOI: 10.1109/TKDE.2021.3079836.

- Willard, Jared et al. (Mar. 25, 2022). "Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems". In: ACM Computing Surveys, 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).
- You, Ronghui et al. (July 15, 2018). "GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank". In: Bioinformatics 34.14. 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).
- Zampieri, Guido et al. (July 2019). "Machine and Deep Learning Meet Genome-Scale Metabolic Modeling". In: PLOS Computational Biology 15.7, e1007084. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007084. (Visited on 05/04/2023).
- Zhou, Naihui et al. (2019a). "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: Genome Biology 20 (1), pp. 1–23. ISSN: 1474760X. DOI: 10.1186/s13059-019-1835-8.

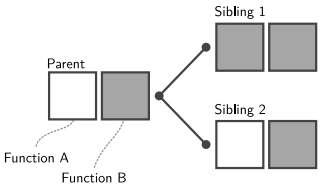
 Zhou, Naihui et al. (2019b). "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: Genome Biology 20 (1), pp. 1–23. ISSN: 1474760X. DOI: 10.1186/s13059-019-1835-8.

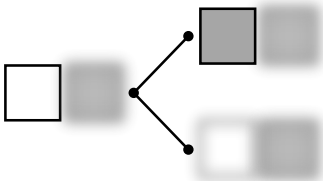

Sibling 1
Parent
Sibling 2
Function A
Function B
Has the function
Doesn't have the function



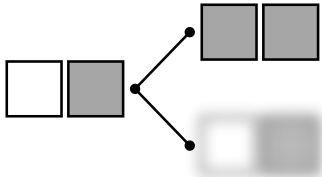


(a) Sibling and Function Conditional Independence

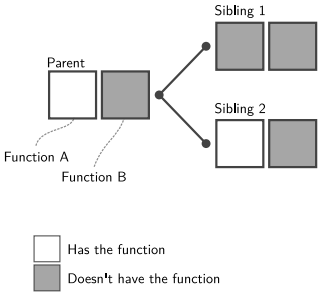


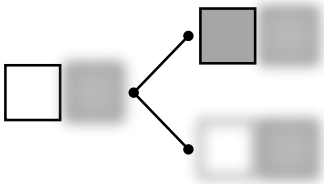


(a) Sibling and Function Conditional Independence

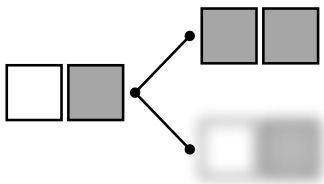


(b) Sibling Conditional Independence

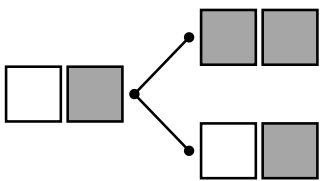




(a) Sibling and Function Conditional Independence



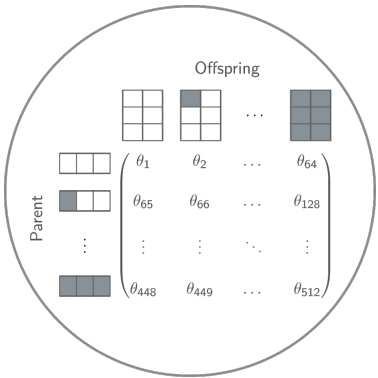
(b) Sibling Conditional Independence



(c) No conditional independence

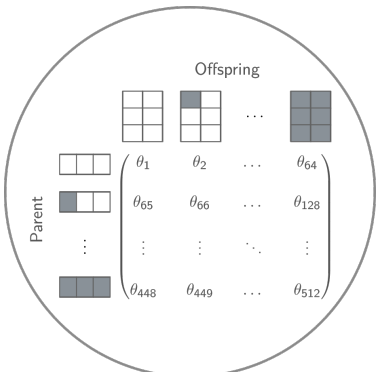
If we wanted to build a model with 3 functions, we would need to estimate...

Full Markov Transition Matrix



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

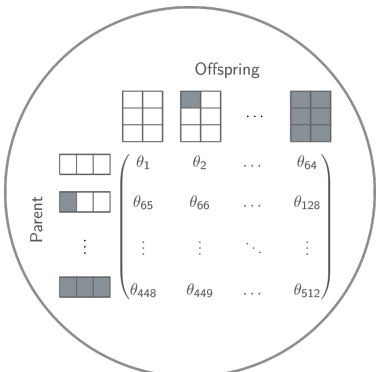
Full Markov Transition Matrix



- 512 parameters

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

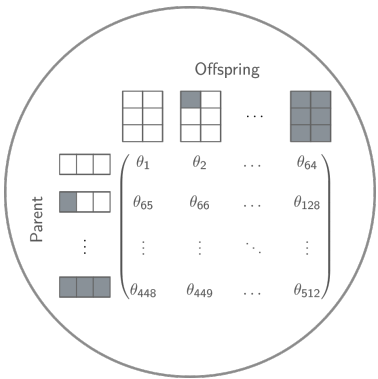
Full Markov Transition Matrix



- 512 parameters
- Finding this many parameters is not easy.

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

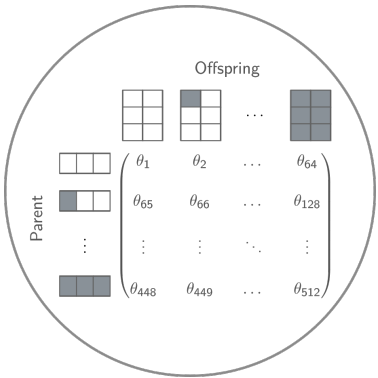
Full Markov Transition Matrix



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

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

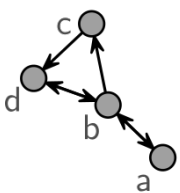
Full Markov Transition Matrix



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

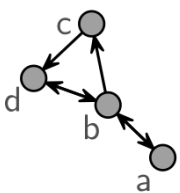
Social Network Analysis may help us...

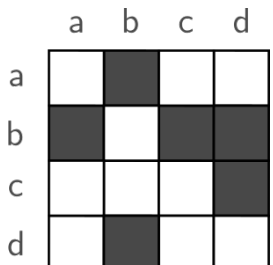
Social Network



	a	b	c	d
a		■		
b	■		■	■
c				■
d		■		

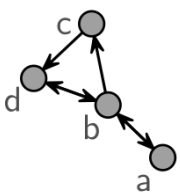
Social Network

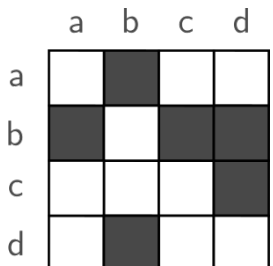




- Not about individual ties.

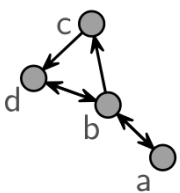
Social Network

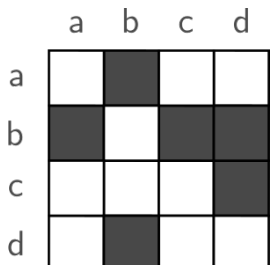




- Not about individual ties.
- Statistical inference on *motifs* (triangles, dyads, homophily, etc.)

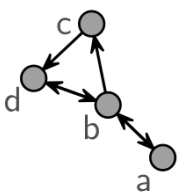
Social Network

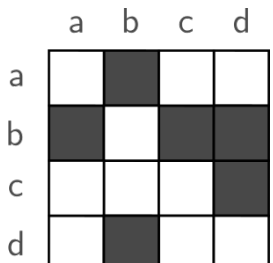




- Not about individual ties.
- Statistical inference on *motifs* (triangles, dyads, homophily, etc.)
- Literature about ERGMs is vast, a.k.a. a low-hanging fruit.

Social Network



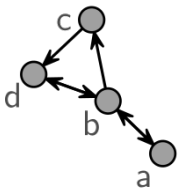


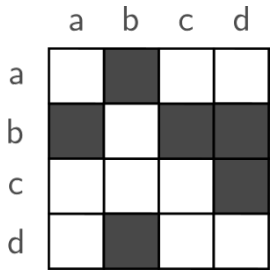
- Not about individual ties.
- Statistical inference on *motifs* (triangles, dyads, homophily, etc.)
- Literature about ERGMs is vast, a.k.a. a low-hanging fruit.

Ultimately...

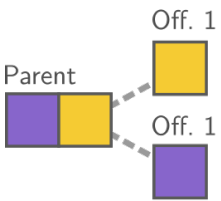
ERGM $\equiv$ Modeling binary arrays

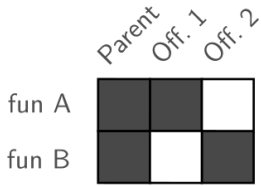
Social Network





Evolutionary Event





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

All possible transitions from $\mathbf{x}_n$

Transition Probability (ERGM)

$$\mathbb{P}\left(\tilde{D}_n \,\middle|\, \mathbf{x}_n, \Theta\right) = \sum_{\mathbf{x}} \mathbb{P}\left(\mathbf{x} \mid \mathbf{x}_n\right) \prod_{m \in \mathbf{O}(n)} \mathbb{P}\left(\tilde{D}_m \,\middle|\, \mathbf{x}_m\right)$$

All possible transitions from $\mathbf{x}_n$

Transition Probability (ERGM)

$$\mathbb{P}\left(\tilde{D}_n \,\middle|\, \mathbf{x}_n, \Theta\right) = \sum_{\mathbf{x}} \mathbb{P}(\mathbf{x} \mid \mathbf{x}_n) \prod_{m \in \mathbf{O}(n)} \mathbb{P}\left(\tilde{D}_m \,\middle|\, \mathbf{x}_m\right)$$

Model Parameters

Vector of Sufficient Statistics

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

Normalizing Constant

the ***lingua franca*** of SNA

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

Some computational features of barry