# My ~~37~~ 38-Year Journey with Neural Networks:

# Can They Further Unlock Pichia's Potential?


## Steven M. Muskal

Chief Executive Officer

Eidogen-Sertanty, Inc.

smuskal@eidogen-sertanty.com

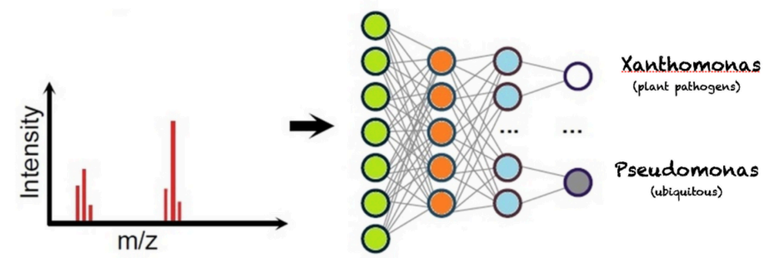
steven.muskal@gmail.com


3/26/2024

# My Journey Began in the Late 80's

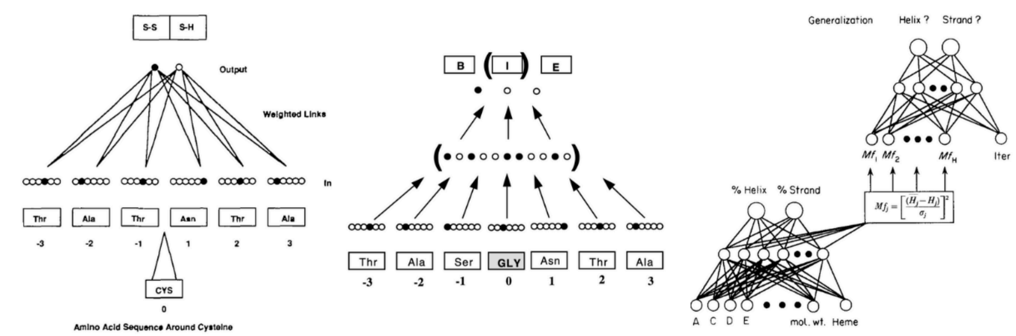
Mines

1986



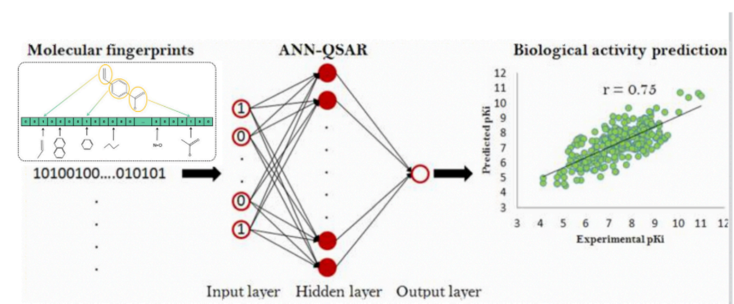

Berkeley

1990



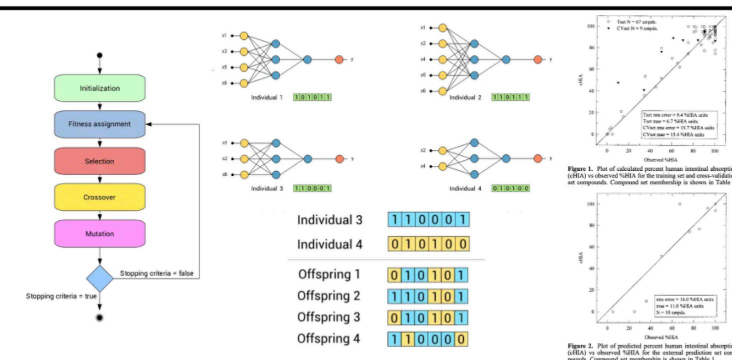

MDL

1993



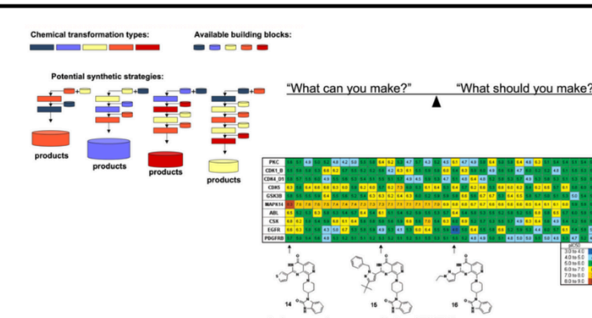

Affymax

1998

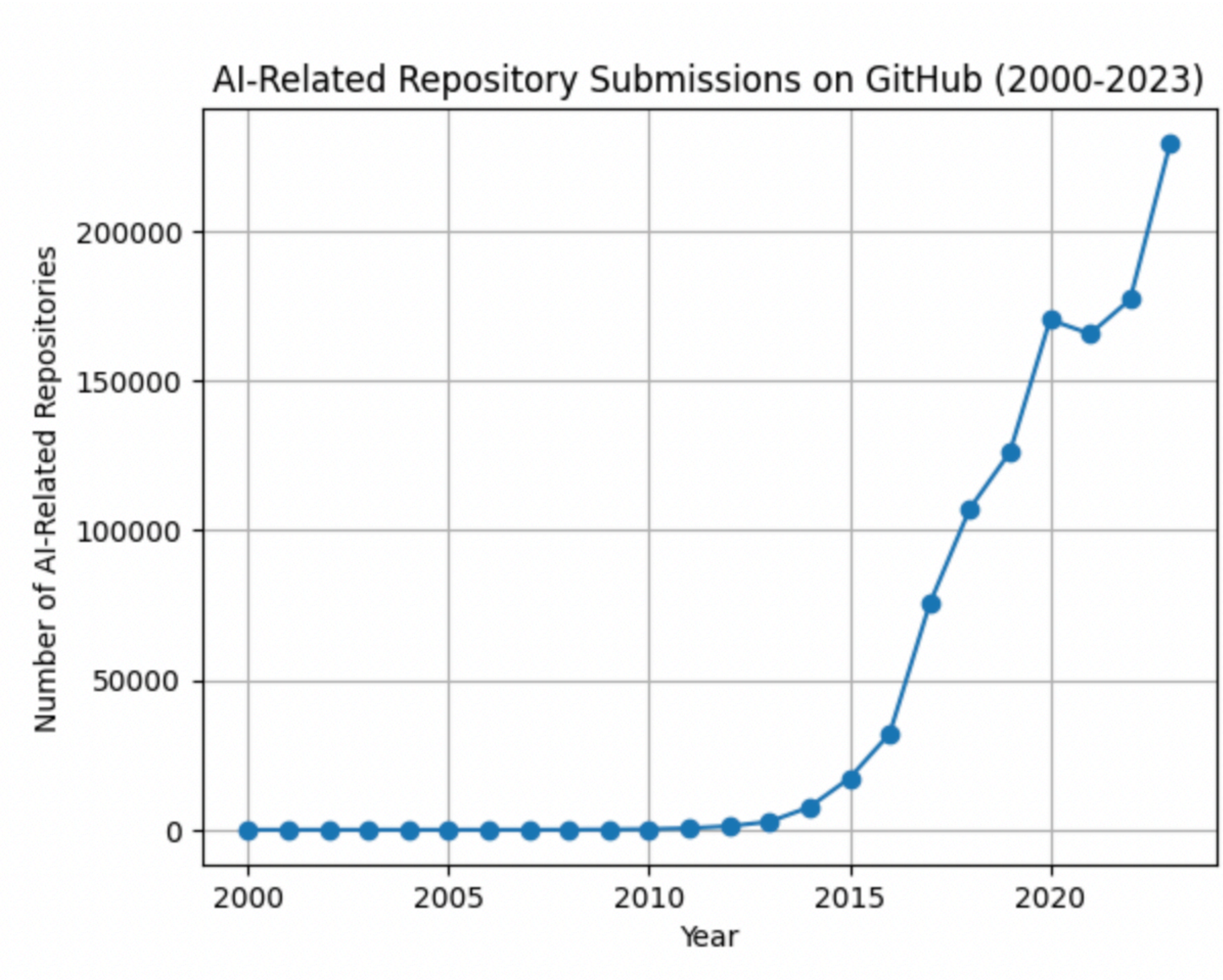



Eidogen

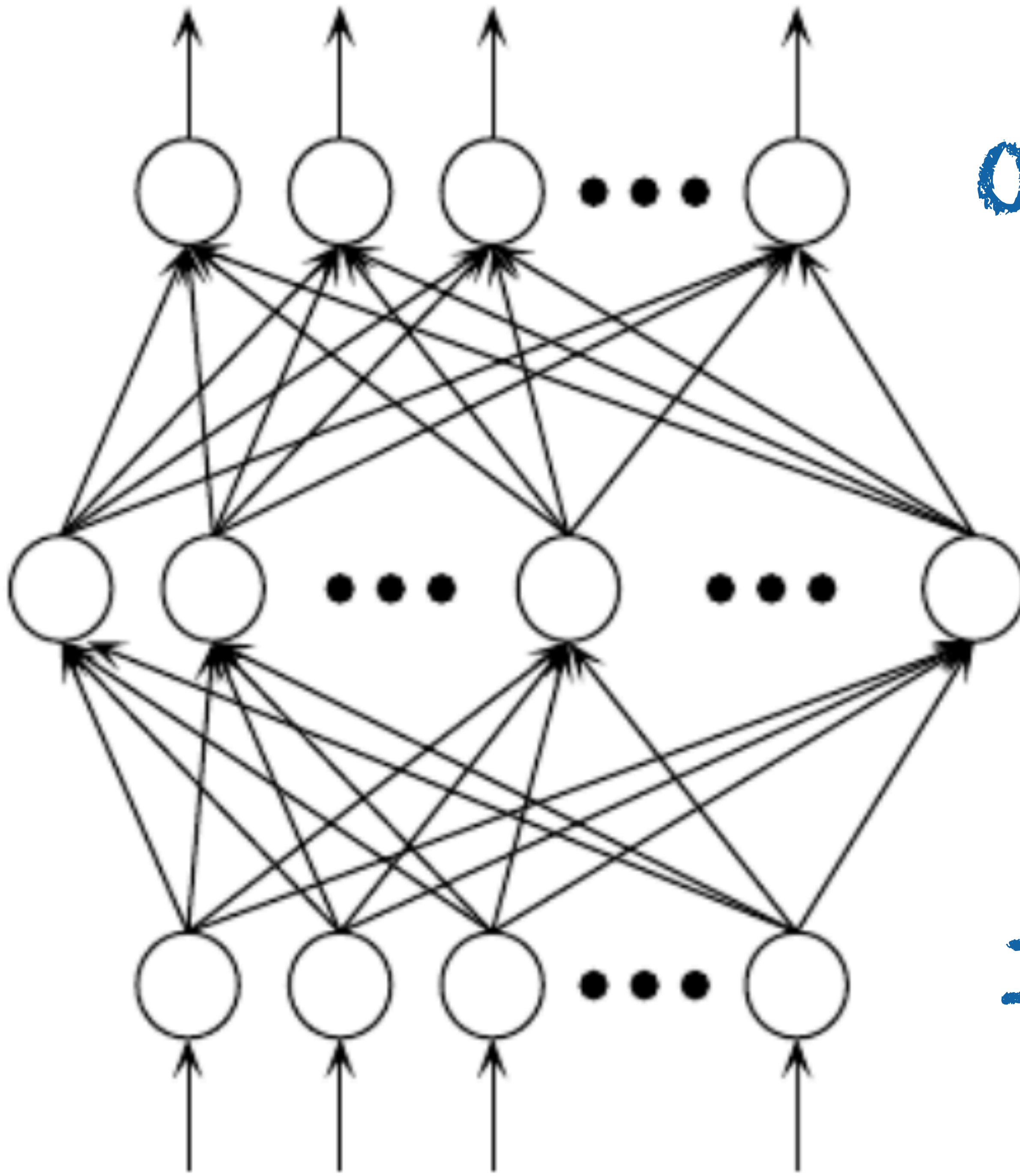
2004...



# Open Source Has Changed the World



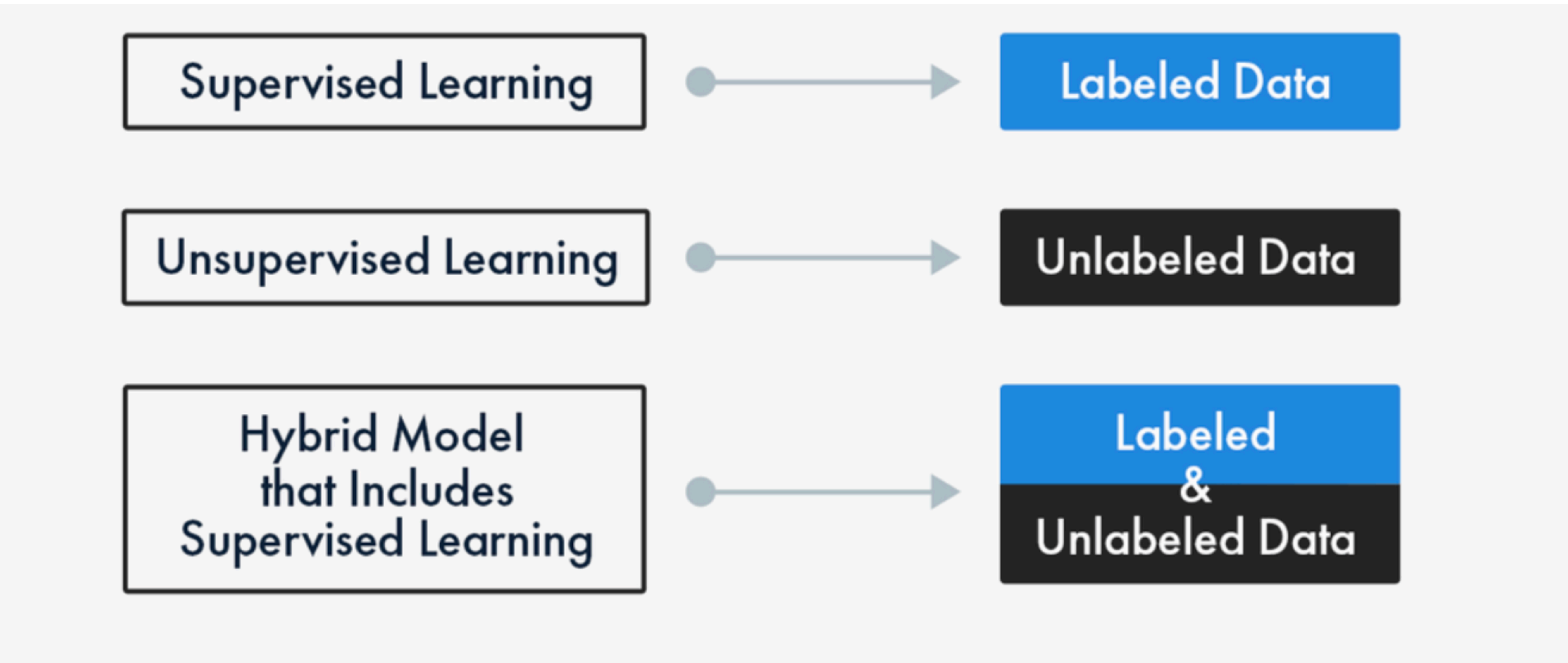
# Neural Network Basics:
# How They Learn and Predict



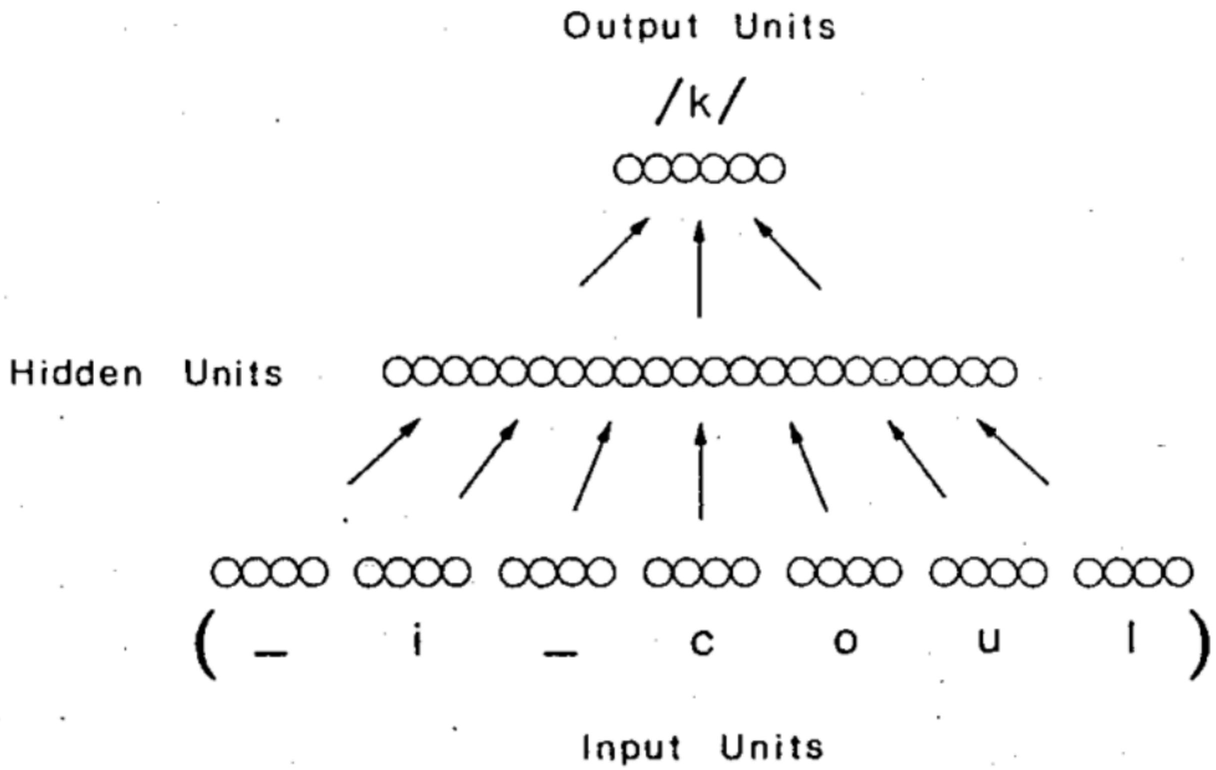
Output Patterns

Input Patterns

# Supervised vs Unsupervised Training



# An "Inspirational" Neural Network




| Symbol | Phoneme | Articulatory features |
|---|---|---|
| /a/ | *fa*ther | Low, Tensed, Central2 |
| /b/ | *b*et | Voiced, Labial, Stop |
| /c/ | b*ou*ght | Unvoiced, Velar, Medium |
| /d/ | *d*ebt | Voiced, Alveolar, Stop |
| /e/ | b*a*ke | Medium, Tensed, Front2 |
| /f/ | *f*in | Unvoiced, Labial, Fricative |
| /g/ | *g*uess | Voiced, Velar, Stop |
| /h/ | *h*ead | Unvoiced, Glottal, Glide |
| /i/ | P*e*te | High, Tensed, Front1 |
| /k/ | *K*en | Unvoiced, Velar, Stop |
| /l/ | *l*et | Voiced, Dental, Liquid |

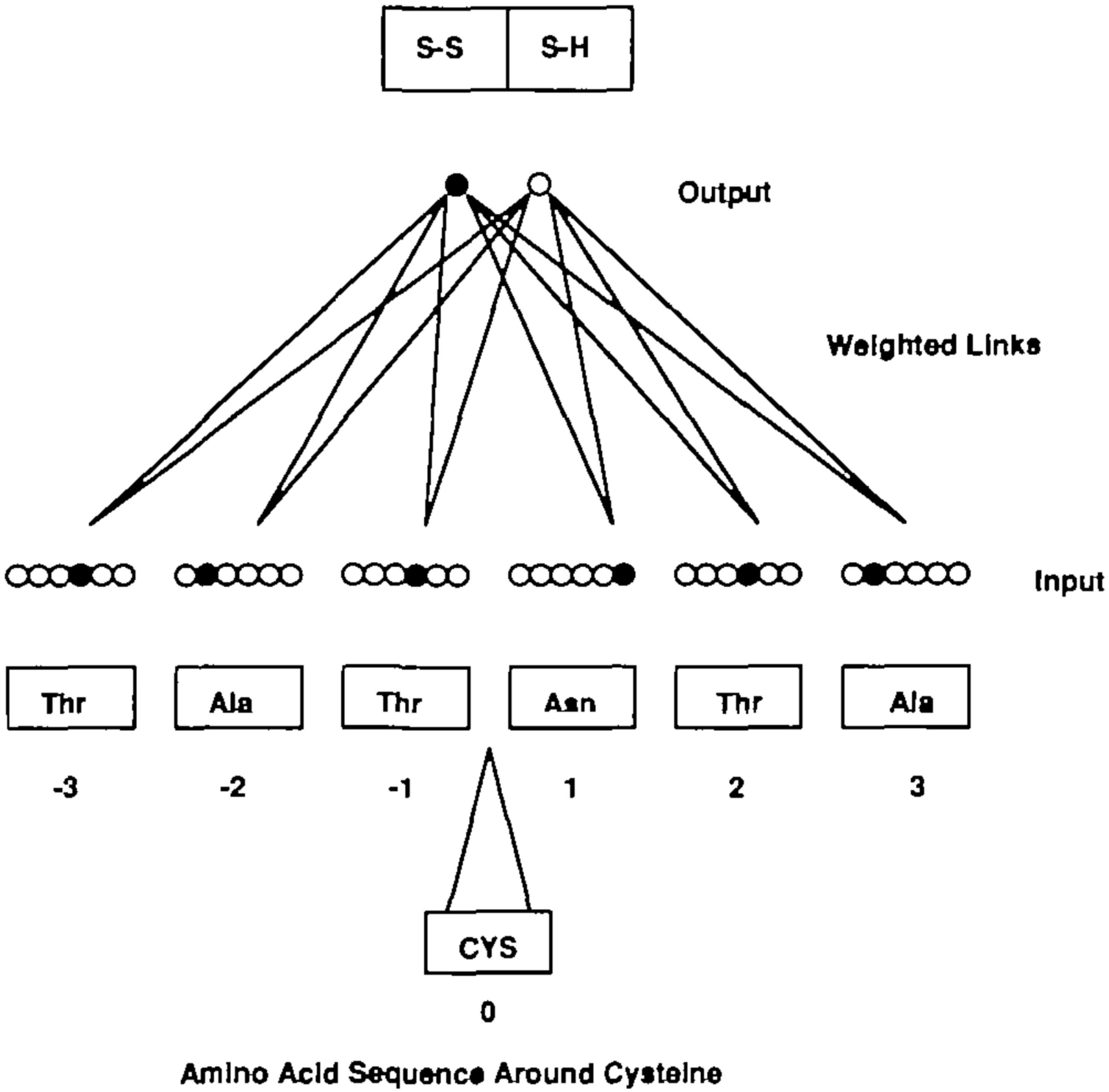
**(1986)**

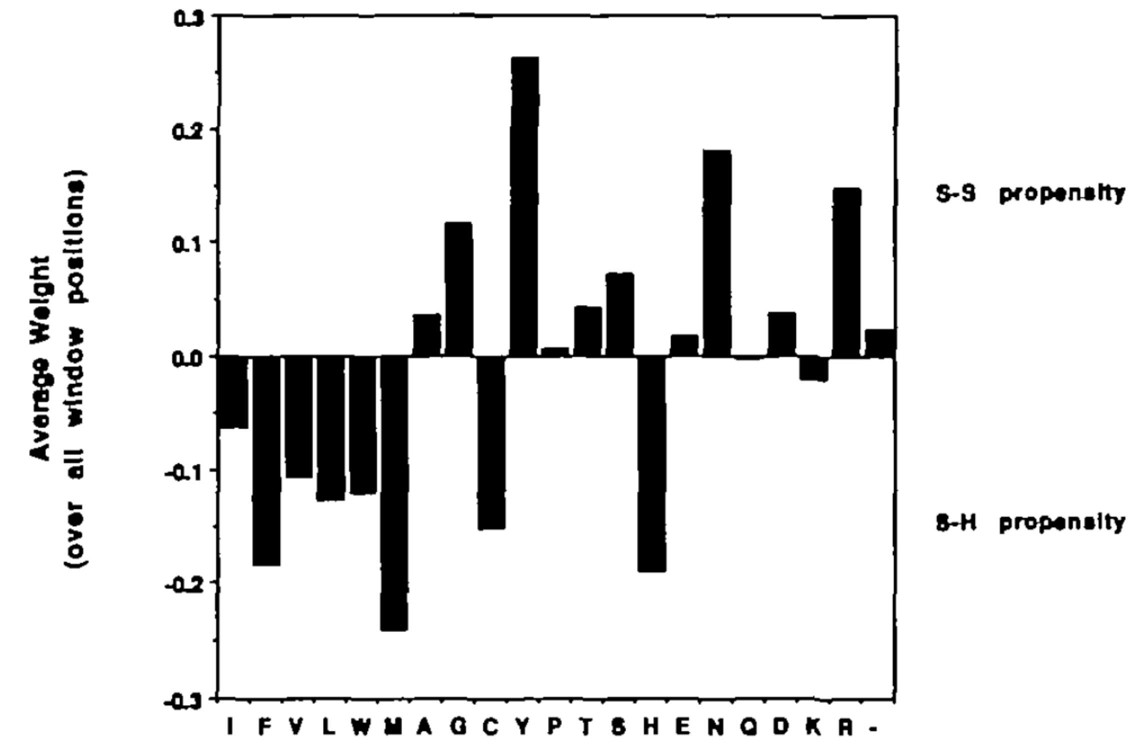
**Terrence J. Sejnowski and Charles R. Rosenberg**

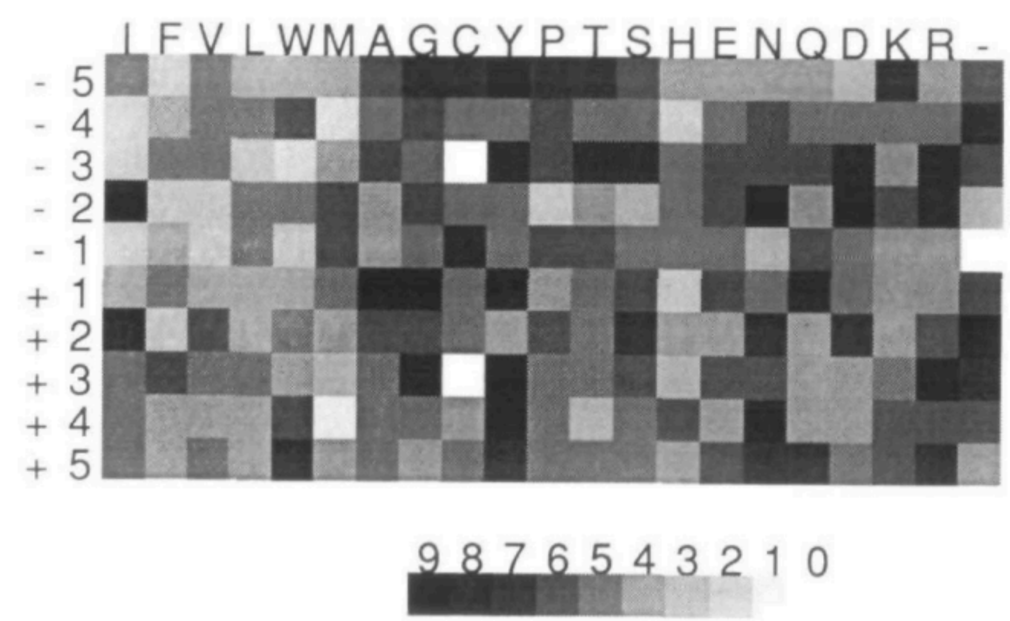
**NETtalk: a parallel network that learns to read aloud**

The Johns Hopkins University Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01, 32 pp.

# Predicting Cysteine's Disulfide-Bonding State
(1990)




https://pubmed.ncbi.nlm.nih.gov/2217140/

# "Attention Is All You Need"

(2017-present)

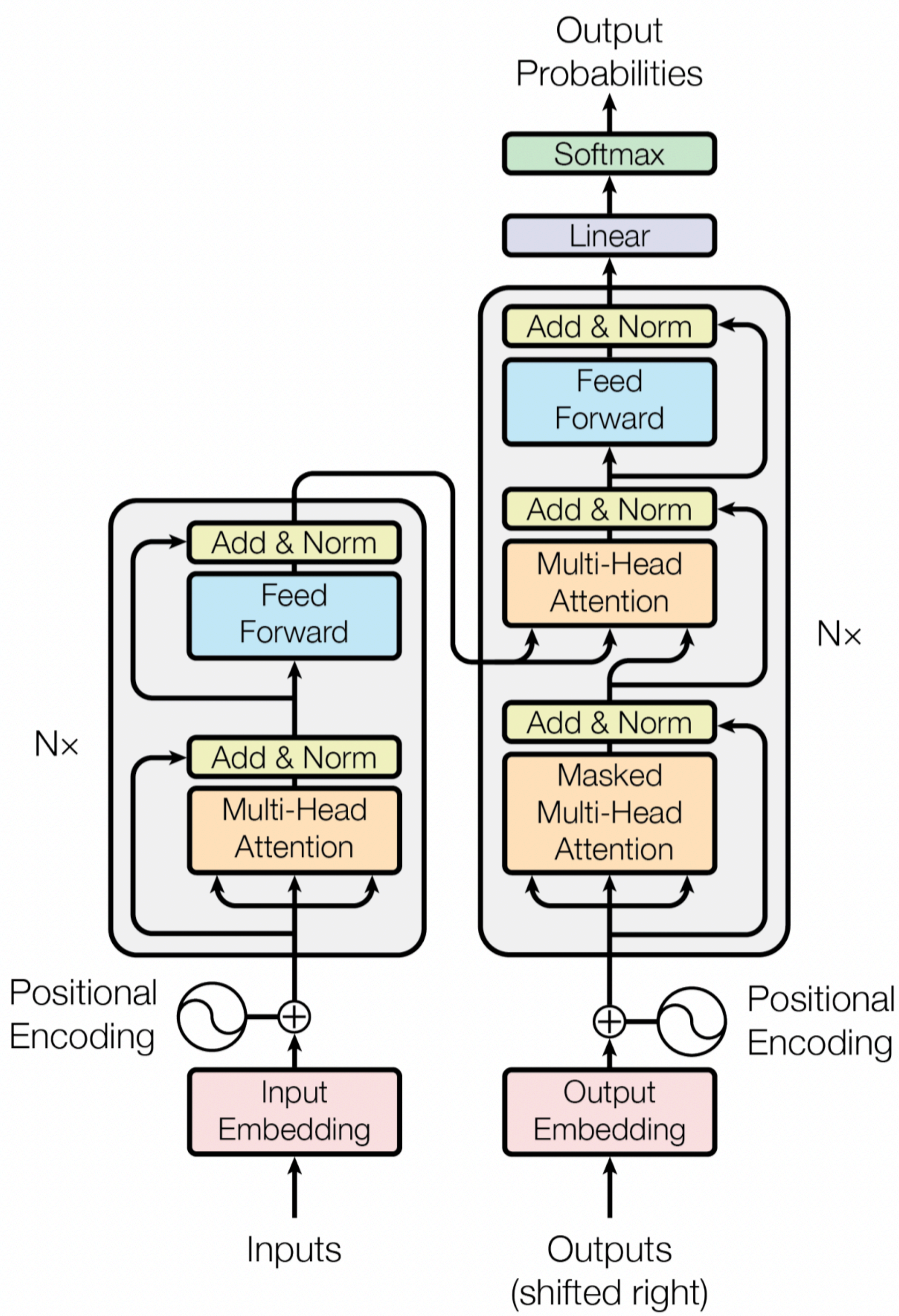



Figure 1: The Transformer - model architecture.

## Attention Is All You Need

**Ashish Vaswani*** Google Brain avaswani@google.com

**Noam Shazeer*** Google Brain noam@google.com

**Niki Parmar*** Google Research nikip@google.com

**Jakob Uszkoreit*** Google Research usz@google.com

**Llion Jones*** Google Research llion@google.com

**Aidan N. Gomez*** † University of Toronto aidan@cs.toronto.edu

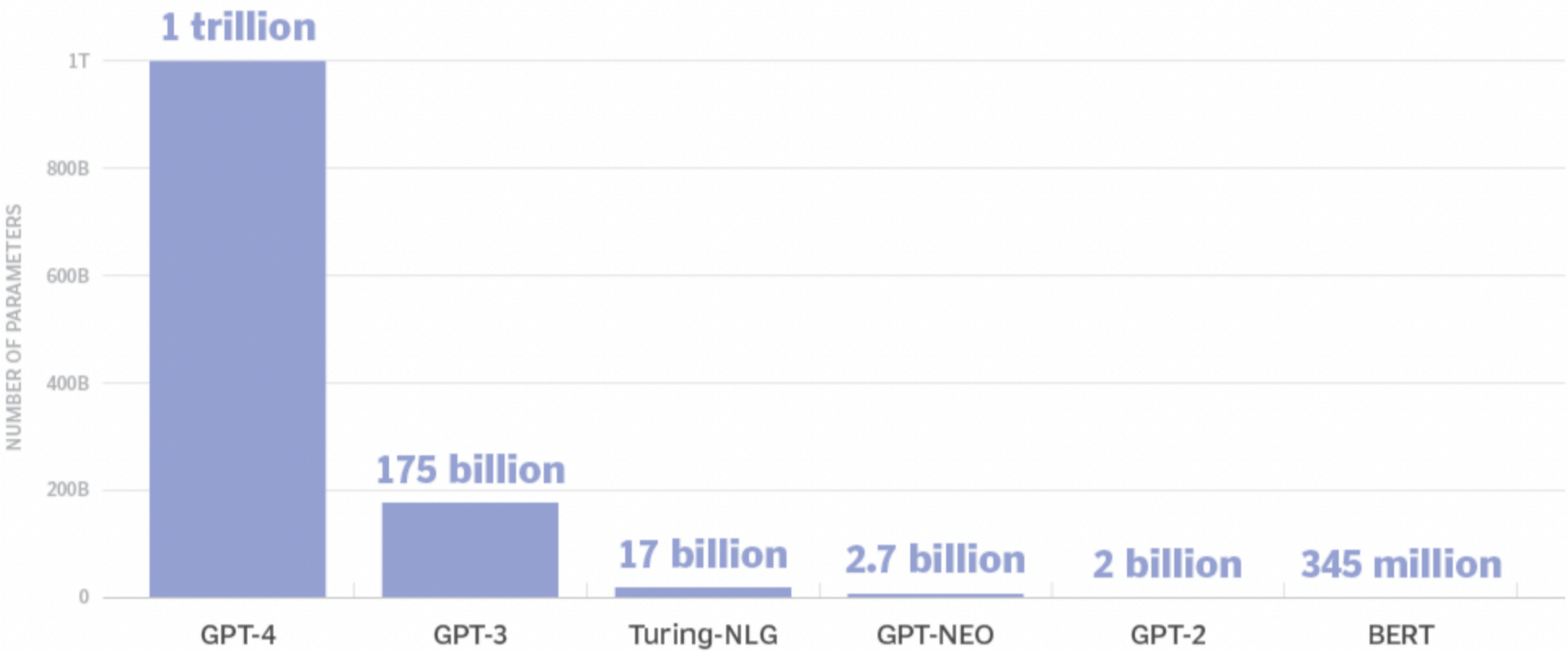
**Łukasz Kaiser*** Google Brain lukaszkaiser@google.com

**Illia Polosukhin*** ‡ illia.polosukhin@gmail.com

Jun2017 - https://arxiv.org/pdf/1706.03762.pdf

(last revised Aug2023)

# Parameters of transformer-based language models

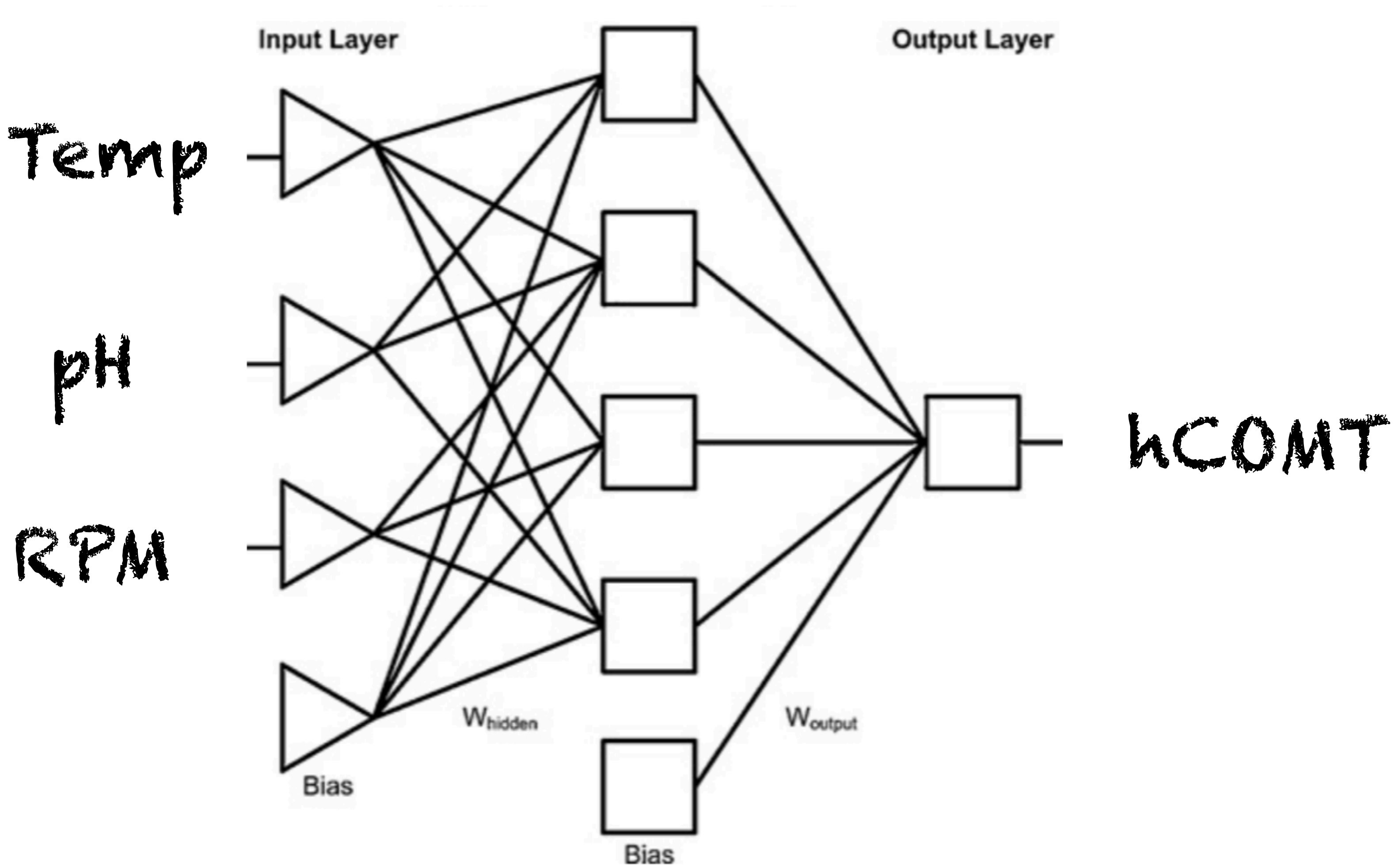


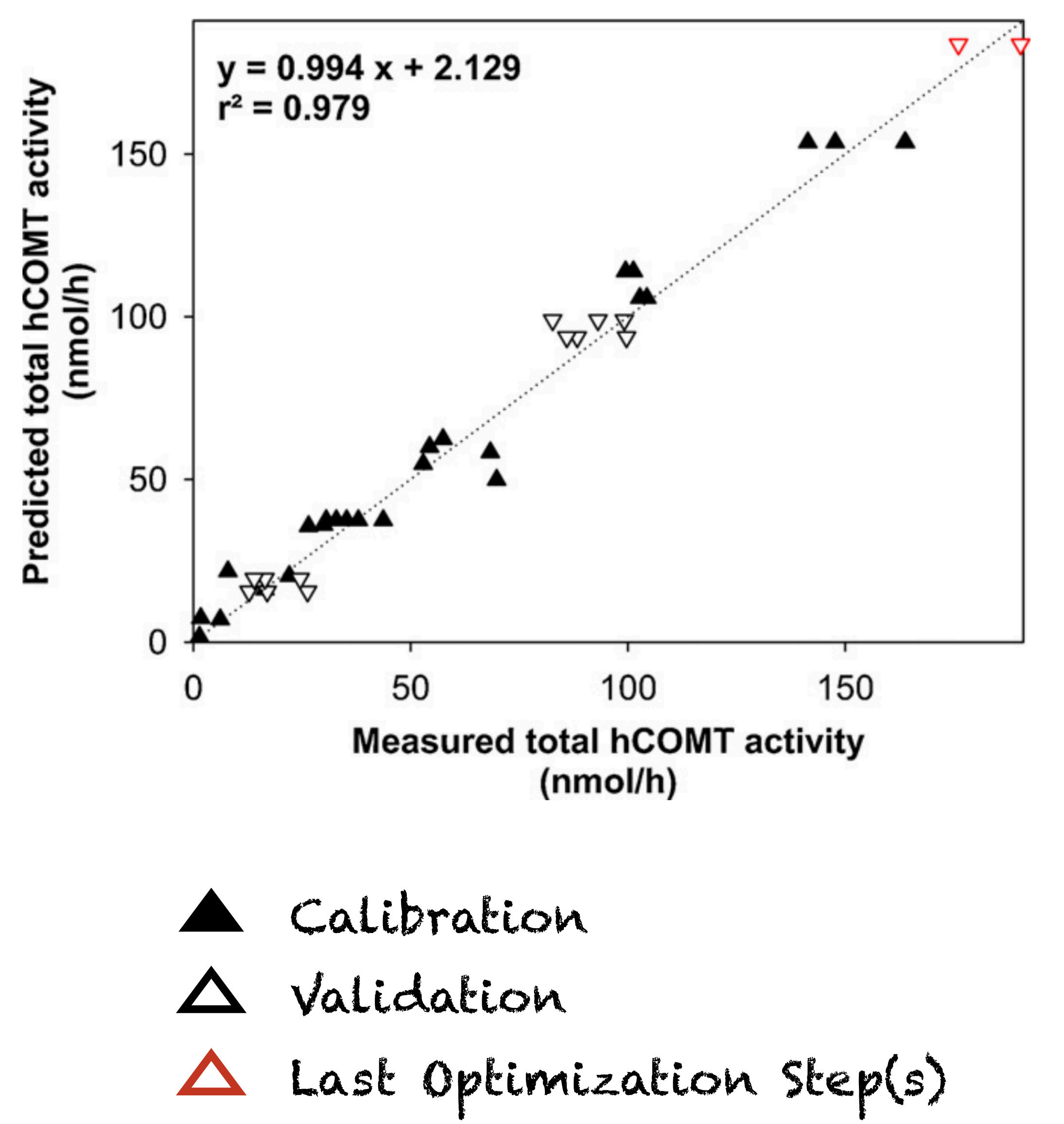

https://www.techtarget.com/whatis/definition/large-language-model-LLM

# Applicable Areas wrt Pichia

- Strain Engineering & Expression
- Optimal Expression Conditions
- Scalability
- Methanol Utilization Control
- Post-Translational Modifications
- Product Recovery and Purification
- Contamination Control
- Regulatory Compliance
- Cost-Effectiveness
- Advances in Bioreactor Design and Process Monitoring

# Optimization of Fermentation Conditions
## Human Soluble Catechol-O-methyltransferase (E. coli)
## (2012)

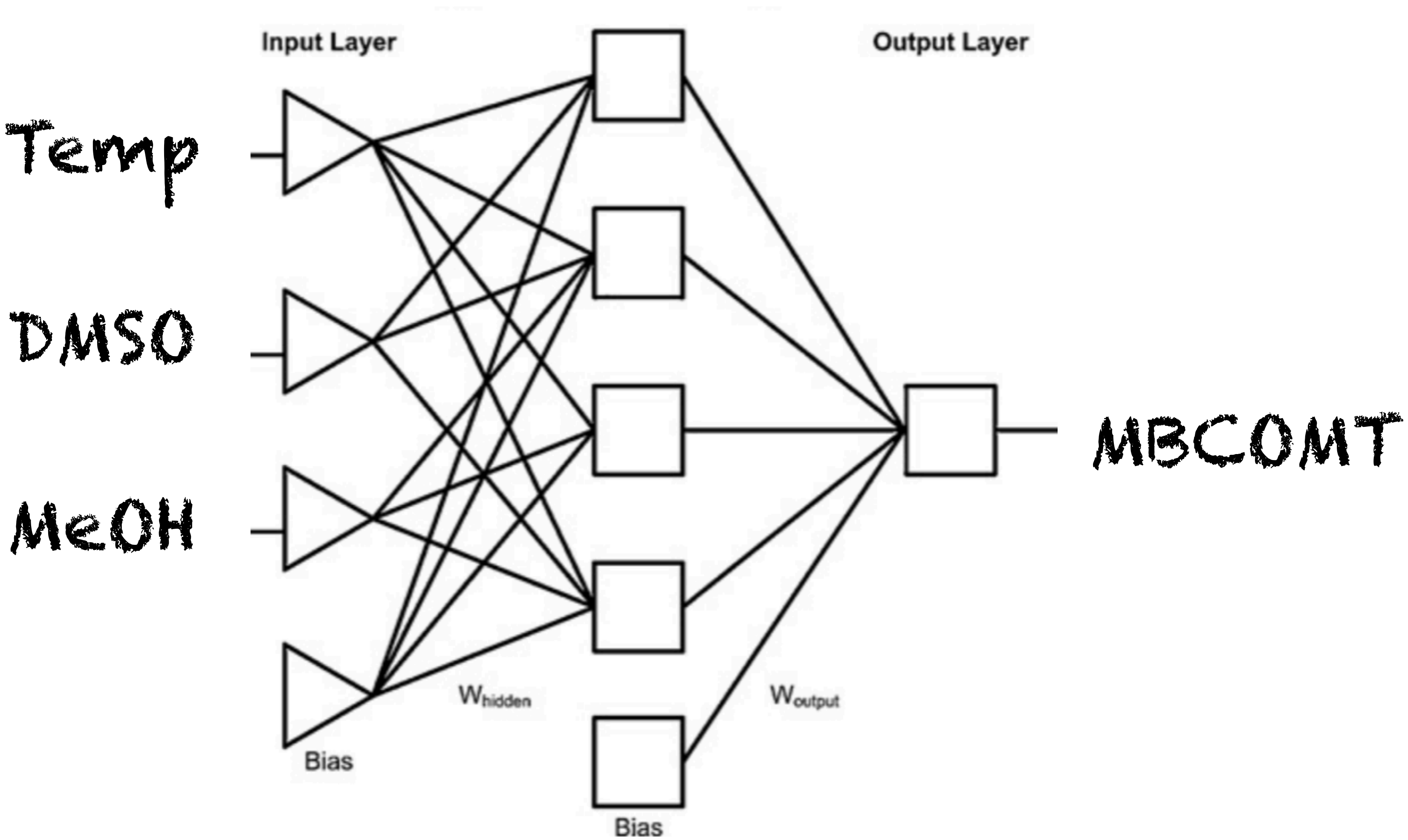


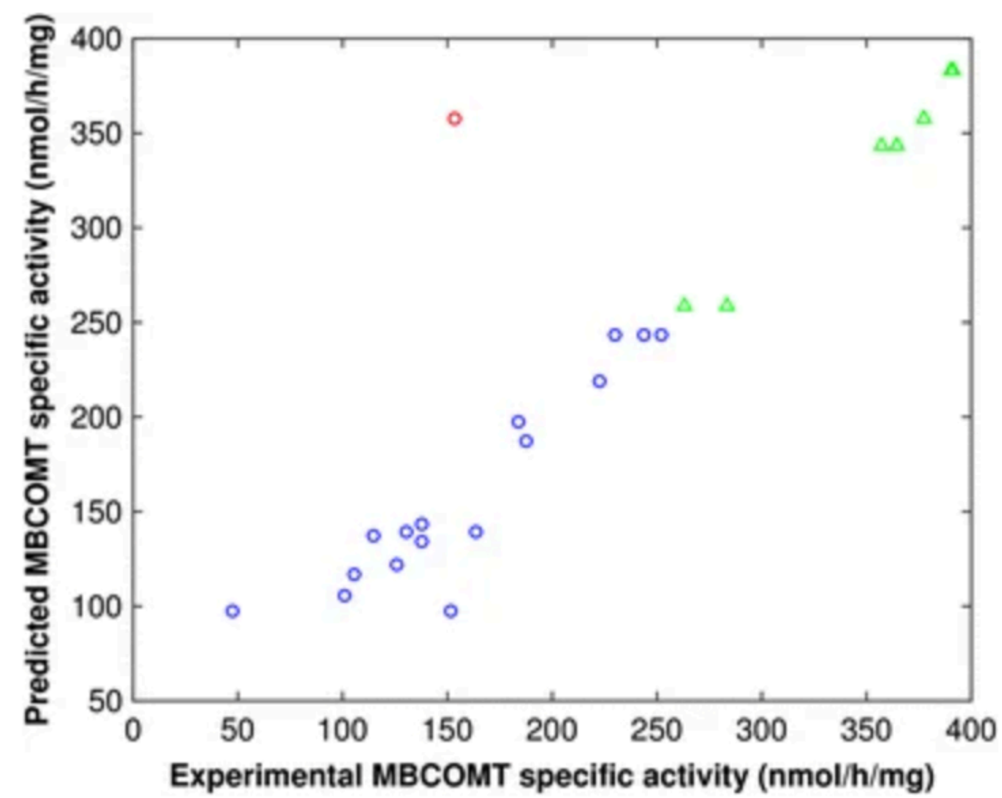




https://pubmed.ncbi.nlm.nih.gov/22498435/

# An Artificial Neural Network for Membrane-Bound Catechol-O-Methyltransferase Biosynthesis with Pichia Pastoris Methanol-Induced Cultures

(2015)




○ Central Composite Design

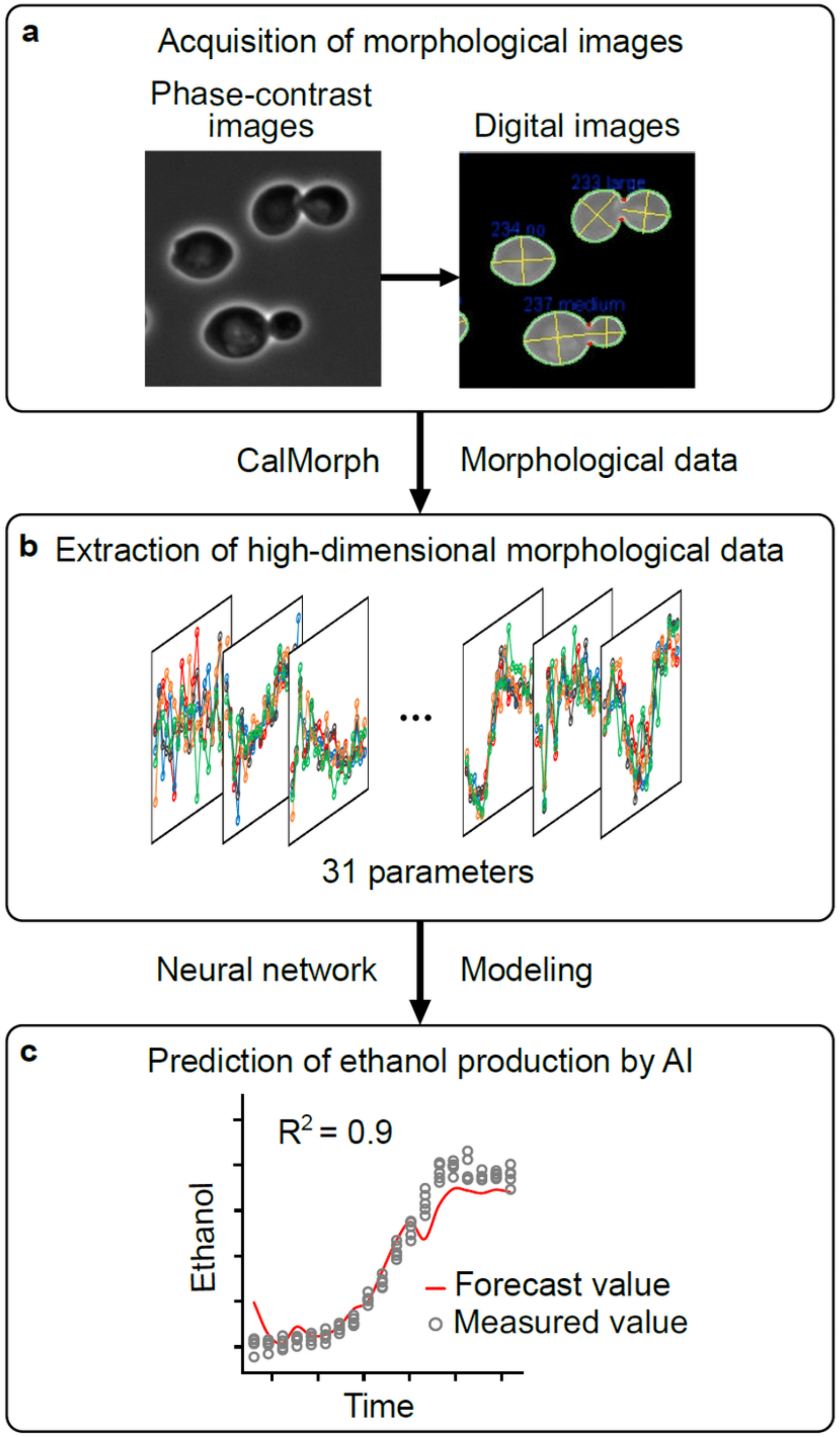
△ Model Optimization

○ Outlier(s)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4527236/

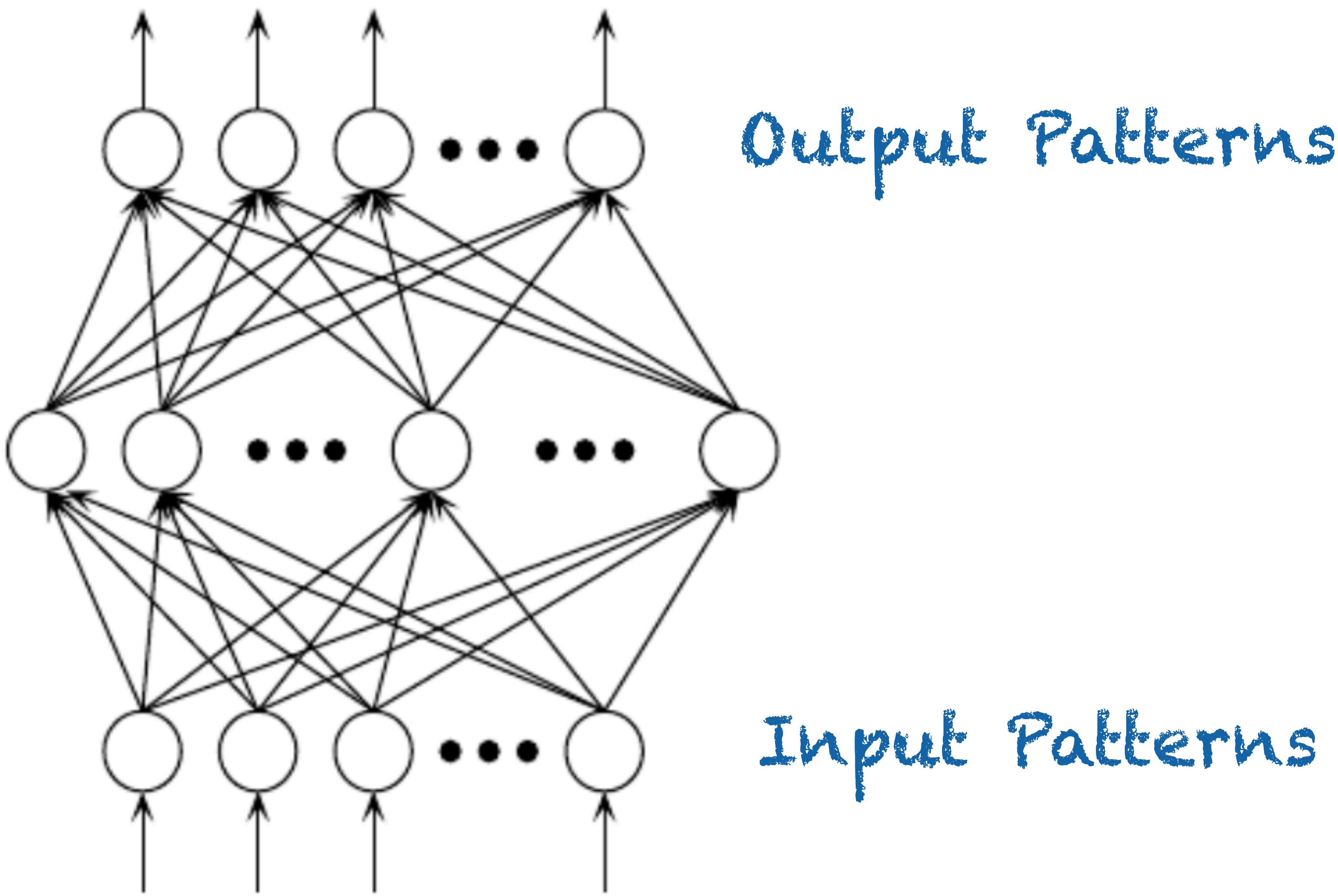
# AI-based Forecasting of Ethanol Fermentation Using Yeast Morphological Data

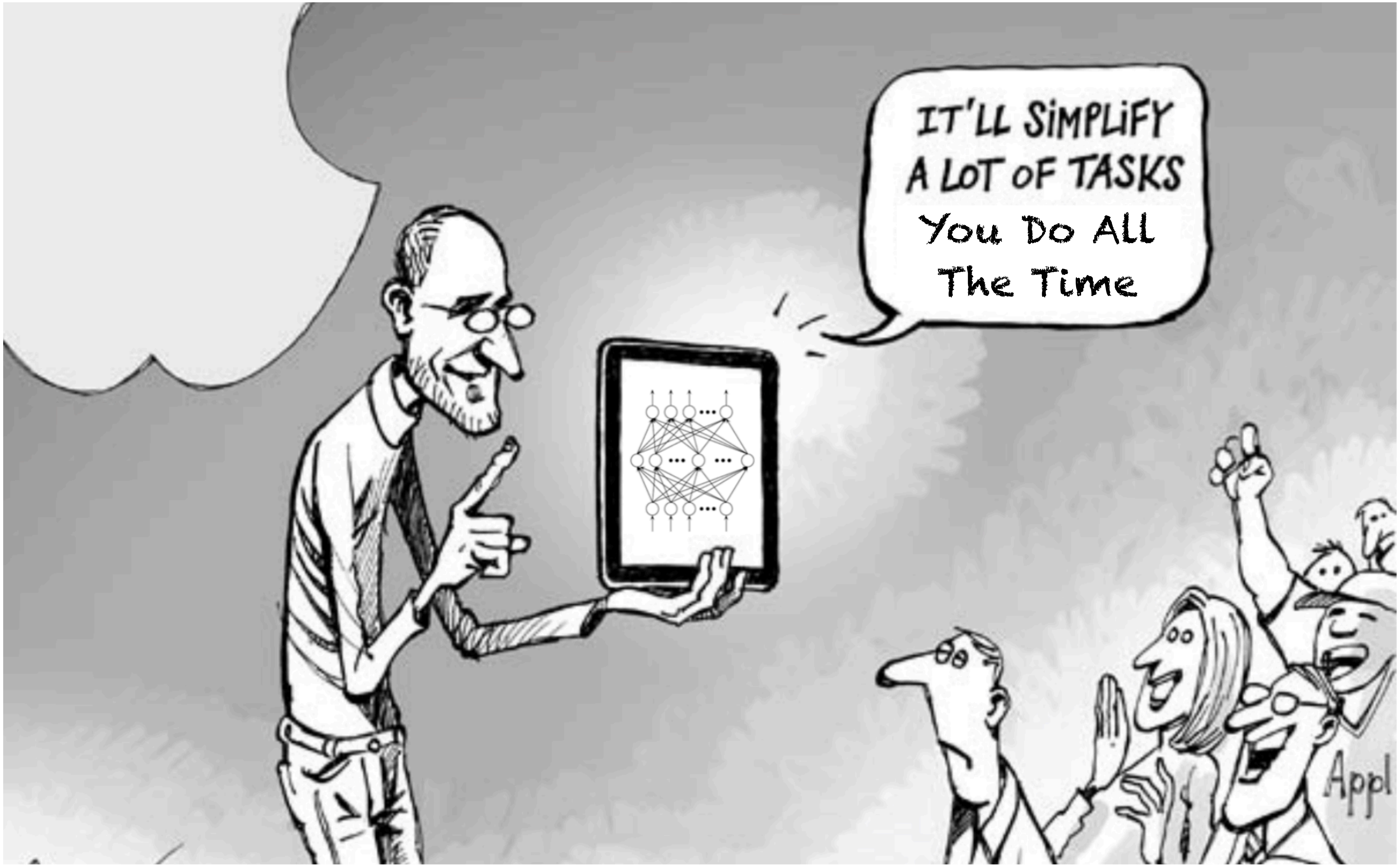
(2021)




https://academic.oup.com/bbb/article/86/1/125/6424286

# What Challenges Are You Facing?

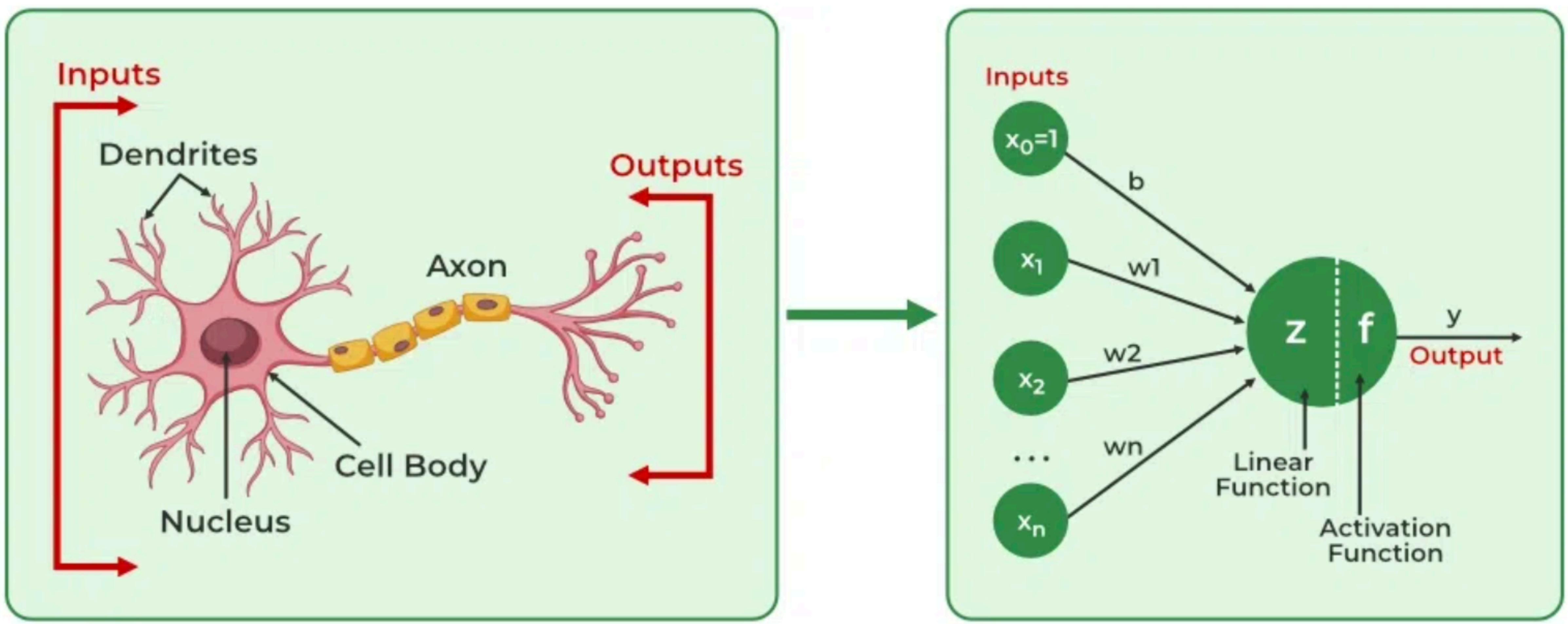





smuskal@eidogen-sertanty.com
steven.muskal@gmail.com

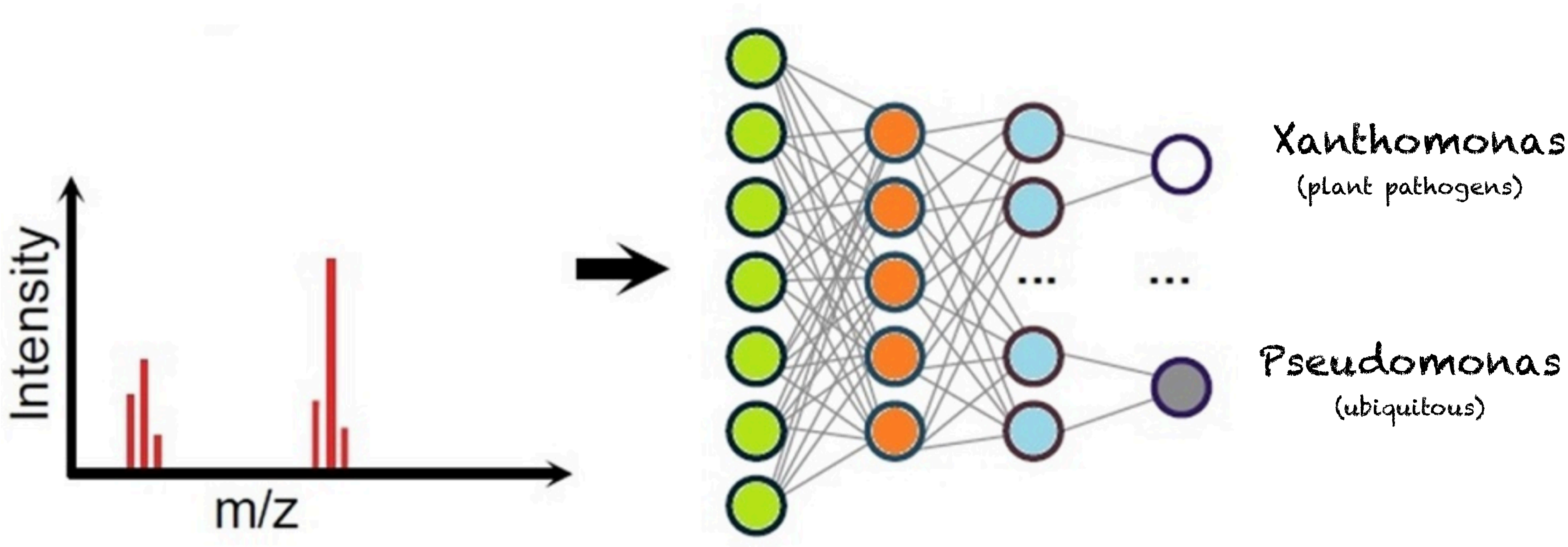
# Supplemental Slides

# The Ultimate Decision Maker...

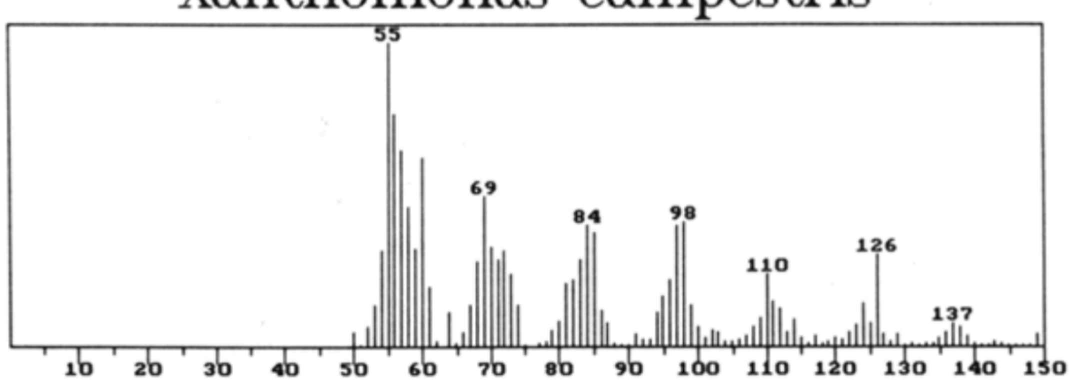


# Classifying Complex Samples: Pyrolysis MassSpec

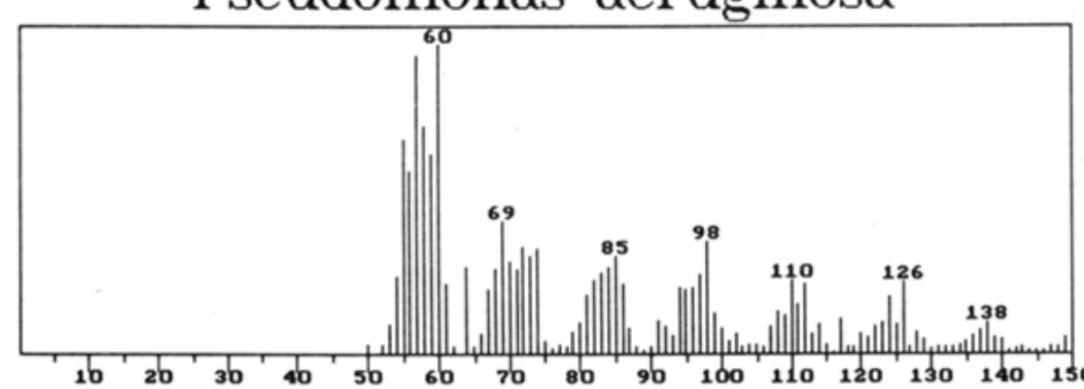
## (circa 1986)




Xanthomonas campestris

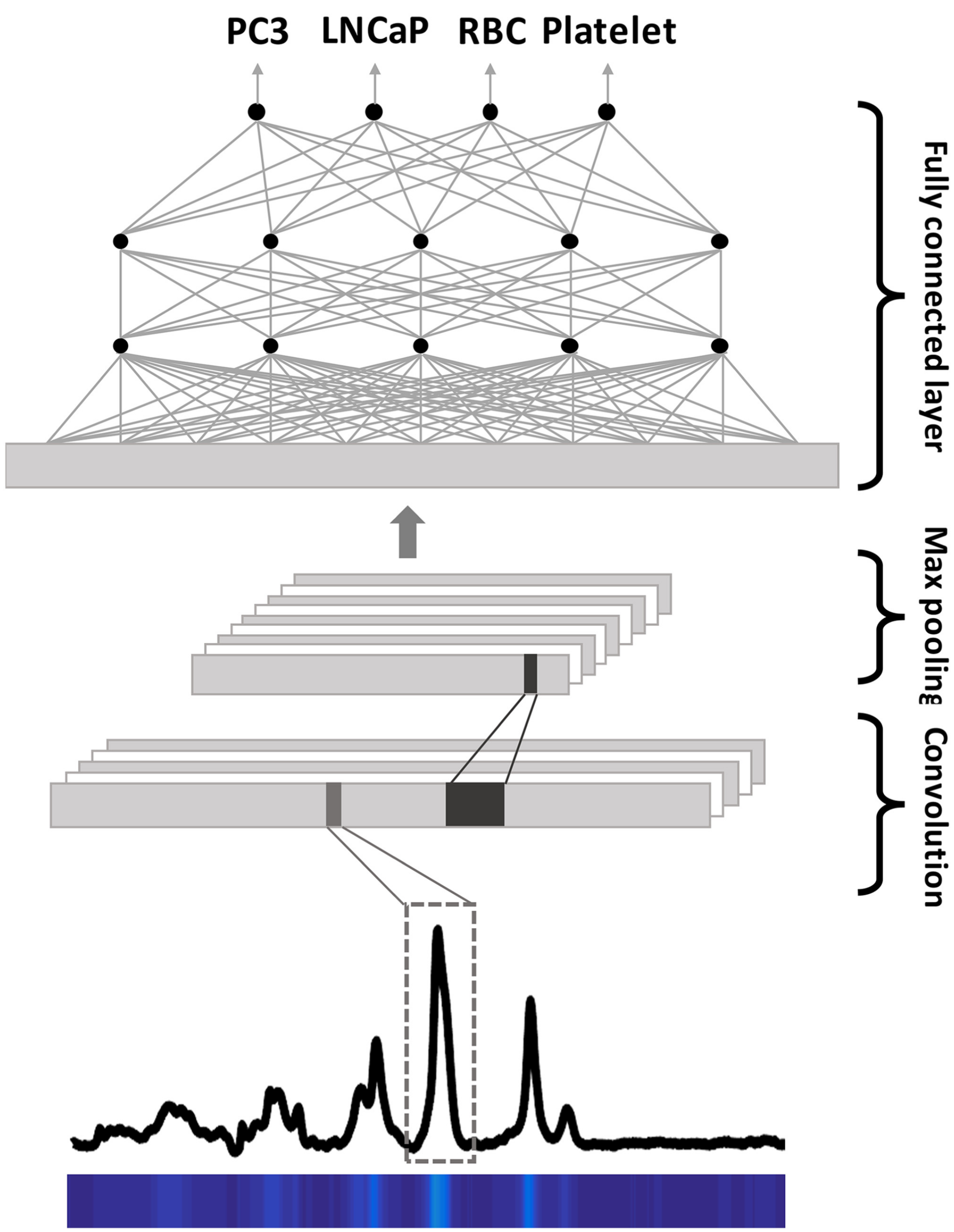



Pseudomonas aeruginosa



# Classifying Raman Spectra of Extracellular Vesicles based on Convolutional Neural Networks for Prostate Cancer Detection

(2019)



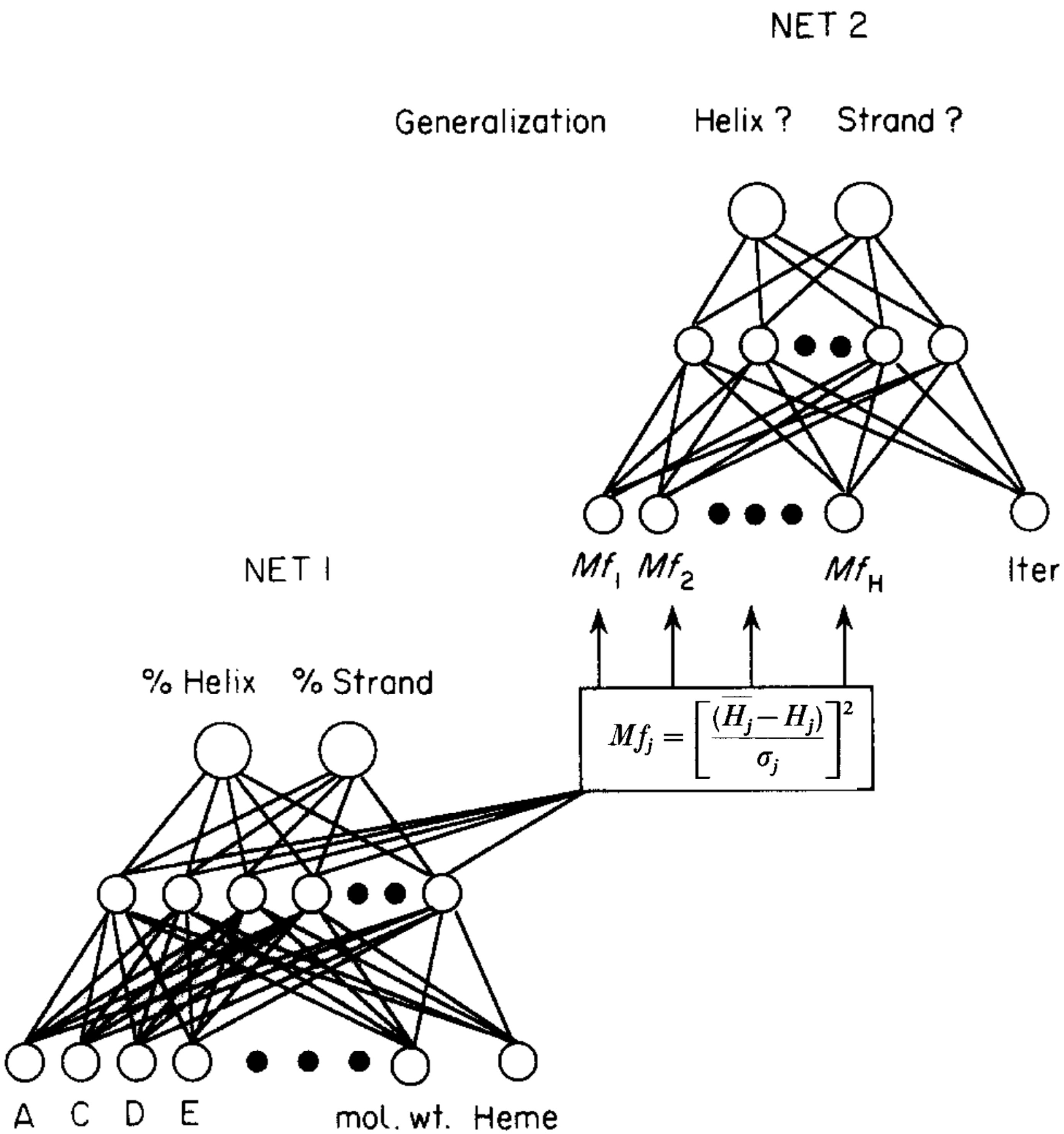

Classifies EVs w/accuracy of >90%

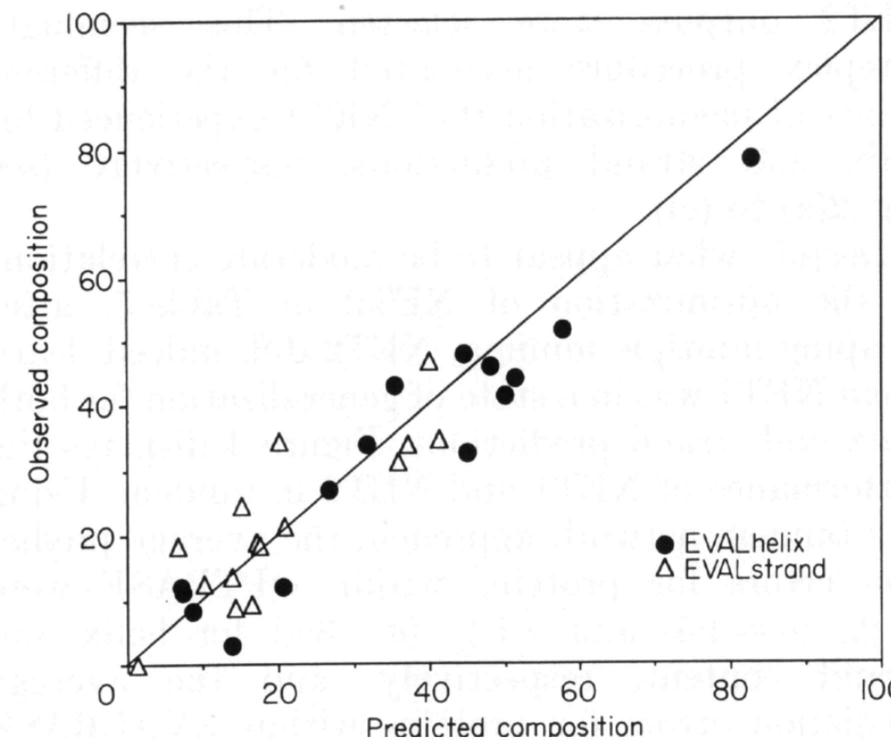
Prostate cancer cell line [PC3]
Lymph node carcinoma of the prostate [LNCaP]

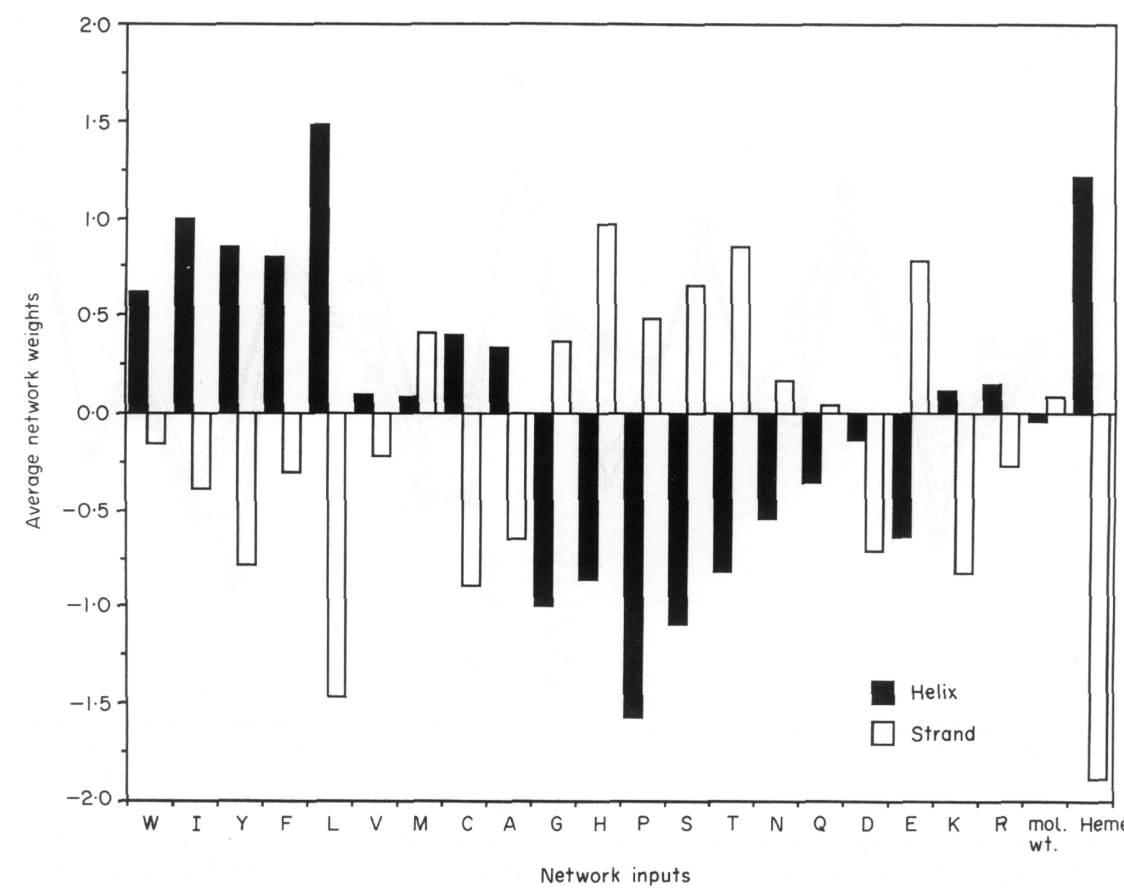
https://doi.org/10.1002/jrs.5770

# Predicting Protein Secondary Structure Content

(1992)



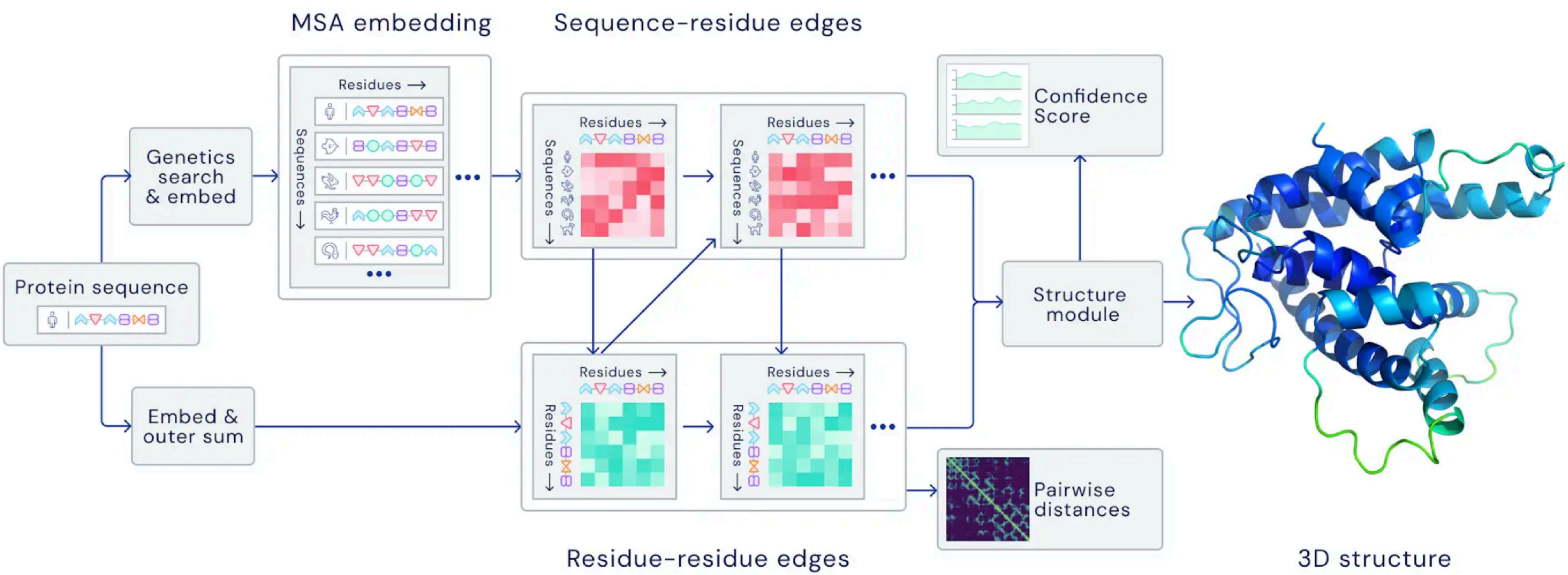




https://pubmed.ncbi.nlm.nih.gov/1602478/

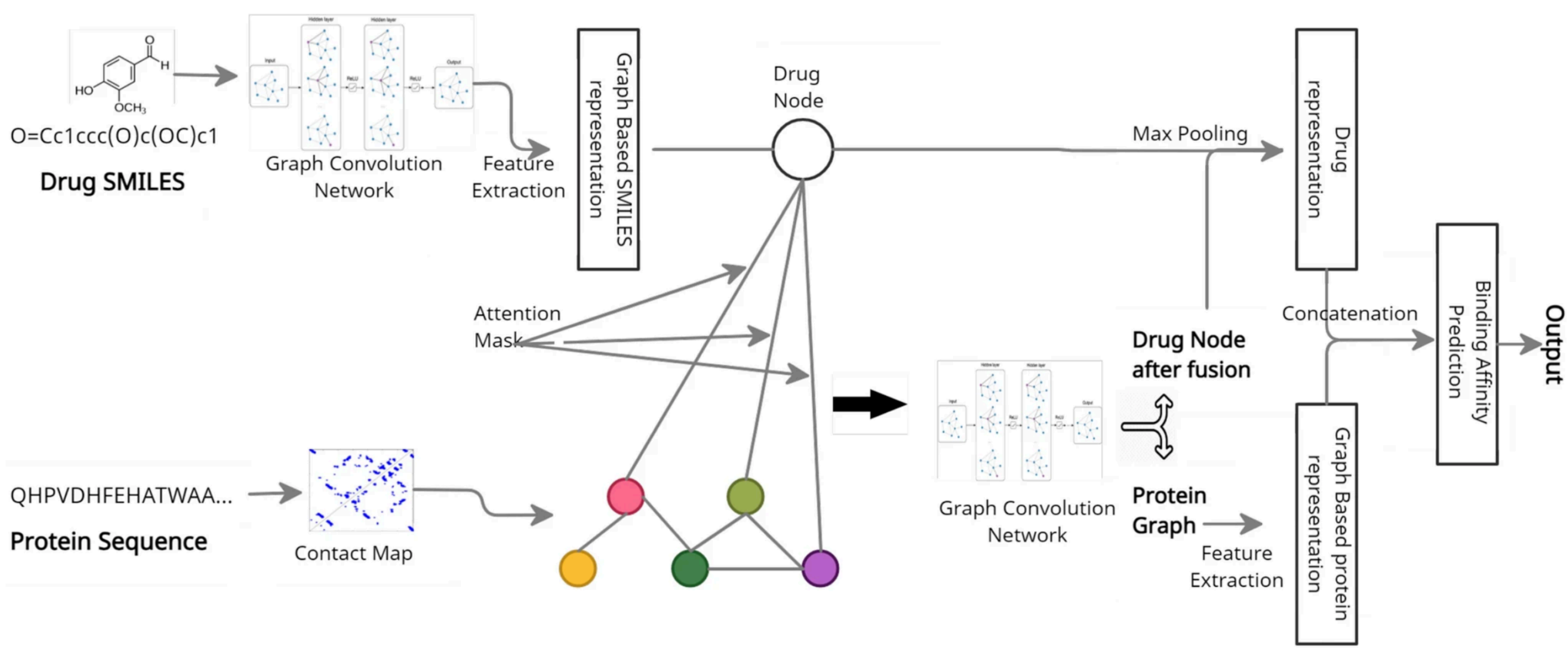
# Highly accurate protein structure prediction with AlphaFold

(2021)




https://www.nature.com/articles/s41586-021-03819-2

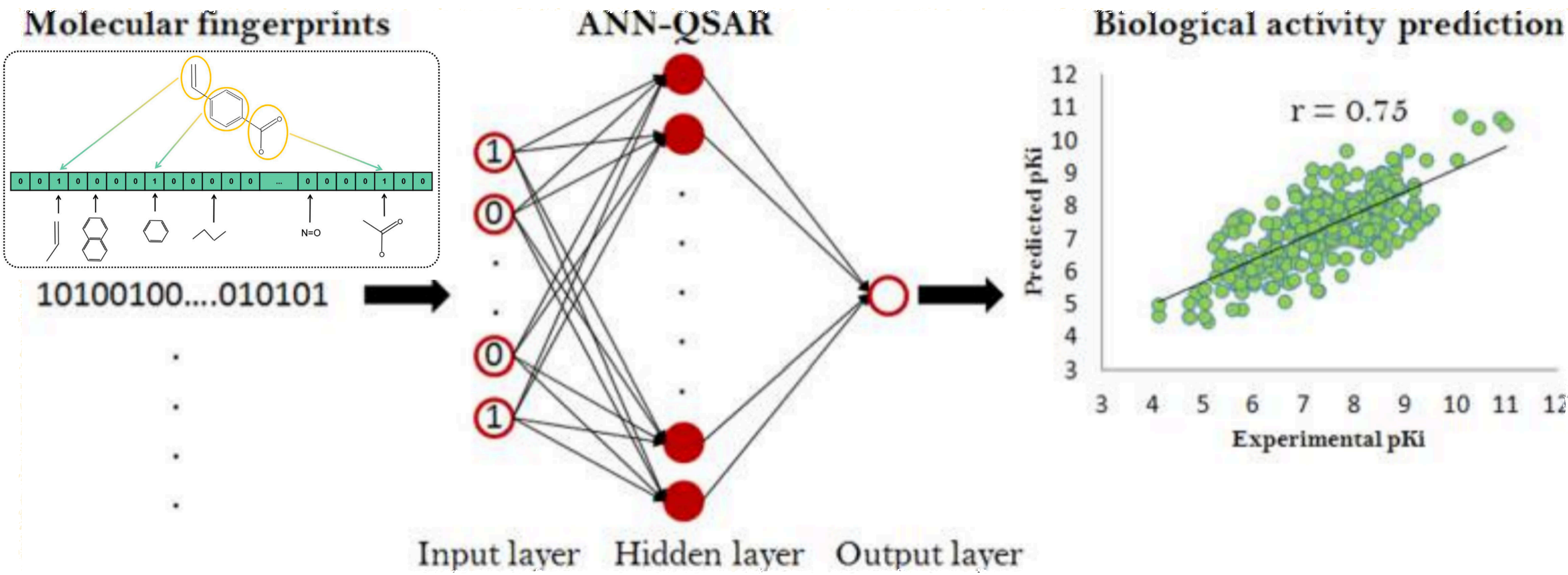
# Generating novel molecule for target protein (SARS-CoV-2) using drug-target interaction based on graph neural network (2022)




https://link.springer.com/article/10.1007/s13721-021-00351-1

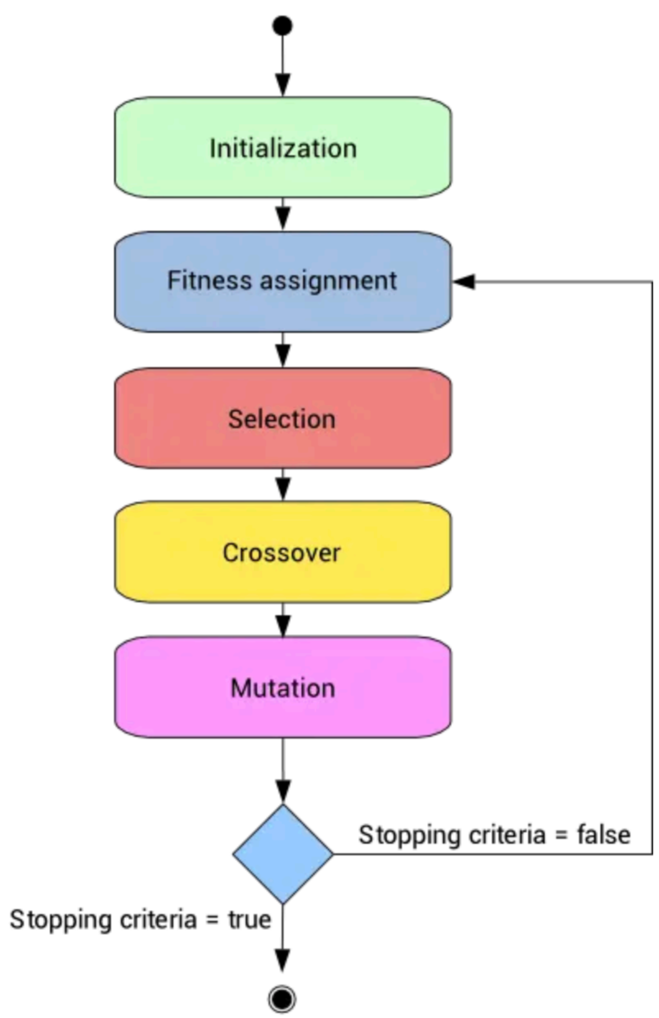
# MDL Keys and QSAR

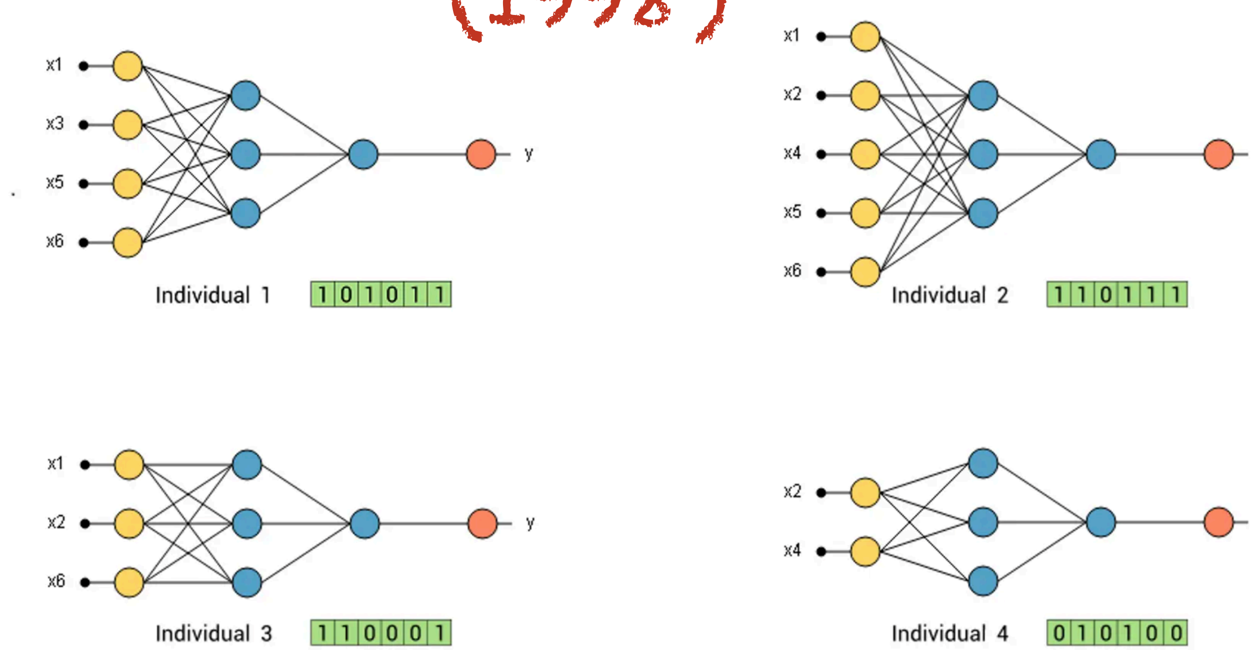
(circa 1993)

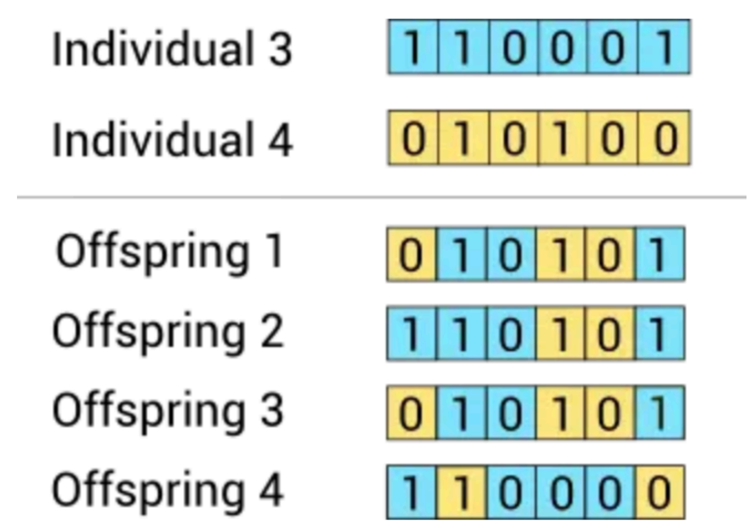


# Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure

(1998)



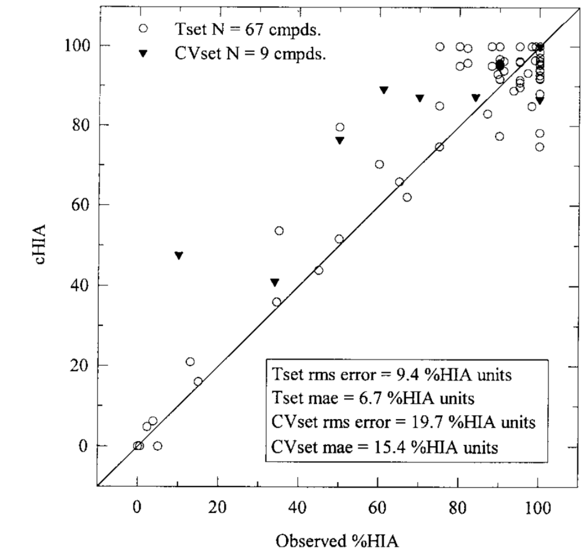




**Figure 1.** Plot of calculated percent human intestinal absorption (cHIA) vs observed %HIA for the training set and cross-validation set compounds. Compound set membership is shown in Table 1.

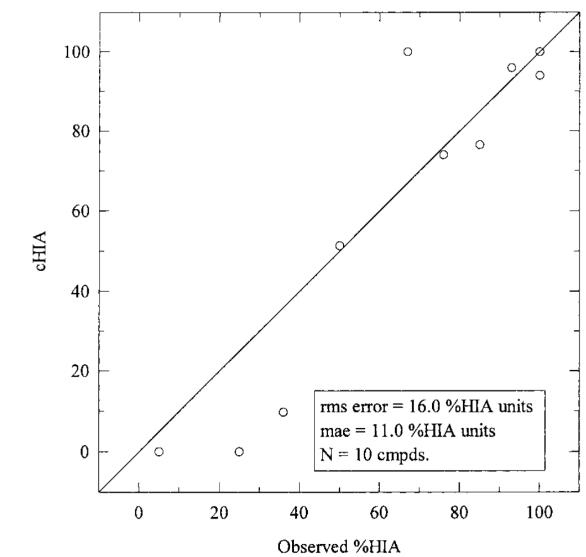



**Figure 2.** Plot of predicted percent human intestinal absorption (cHIA) vs observed %HIA for the external prediction set compounds. Compound set membership is shown in Table 1.

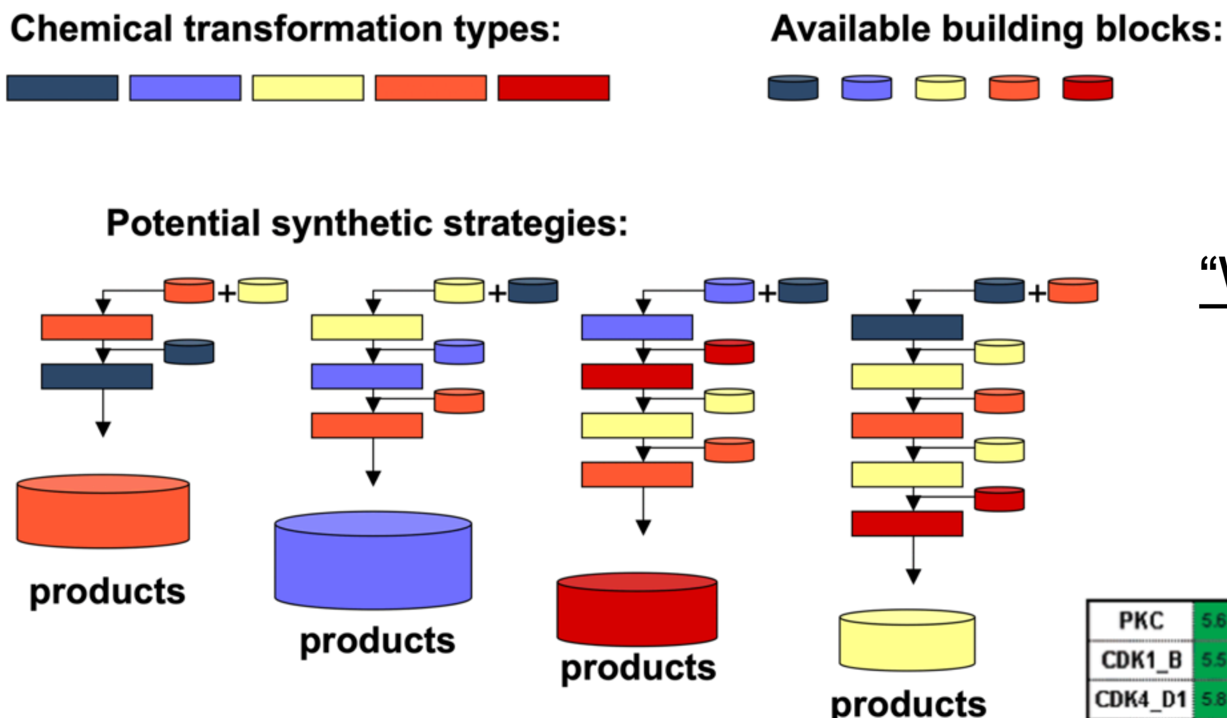
**Table 2.** The Six Descriptors in the Neural Network Model for cHIA Estimation

| Descriptor Label - Definition |
|---|
| NSB - number of single bonds |
| SHDW-6 - normalized 2D projection of molecule on YZ plane |
| CHDH-1 - charge on donatable hydrogen atoms |
| SAAA-2 - surface area of hydrogen bond acceptor atoms/number of hydrogen bond acceptor atoms |
| SCAA-2 - surface area × charge of hydrogen bond acceptor atoms/number of hydrogen bond acceptor atoms |
| GRAV-3 - Cube root of gravitational index |

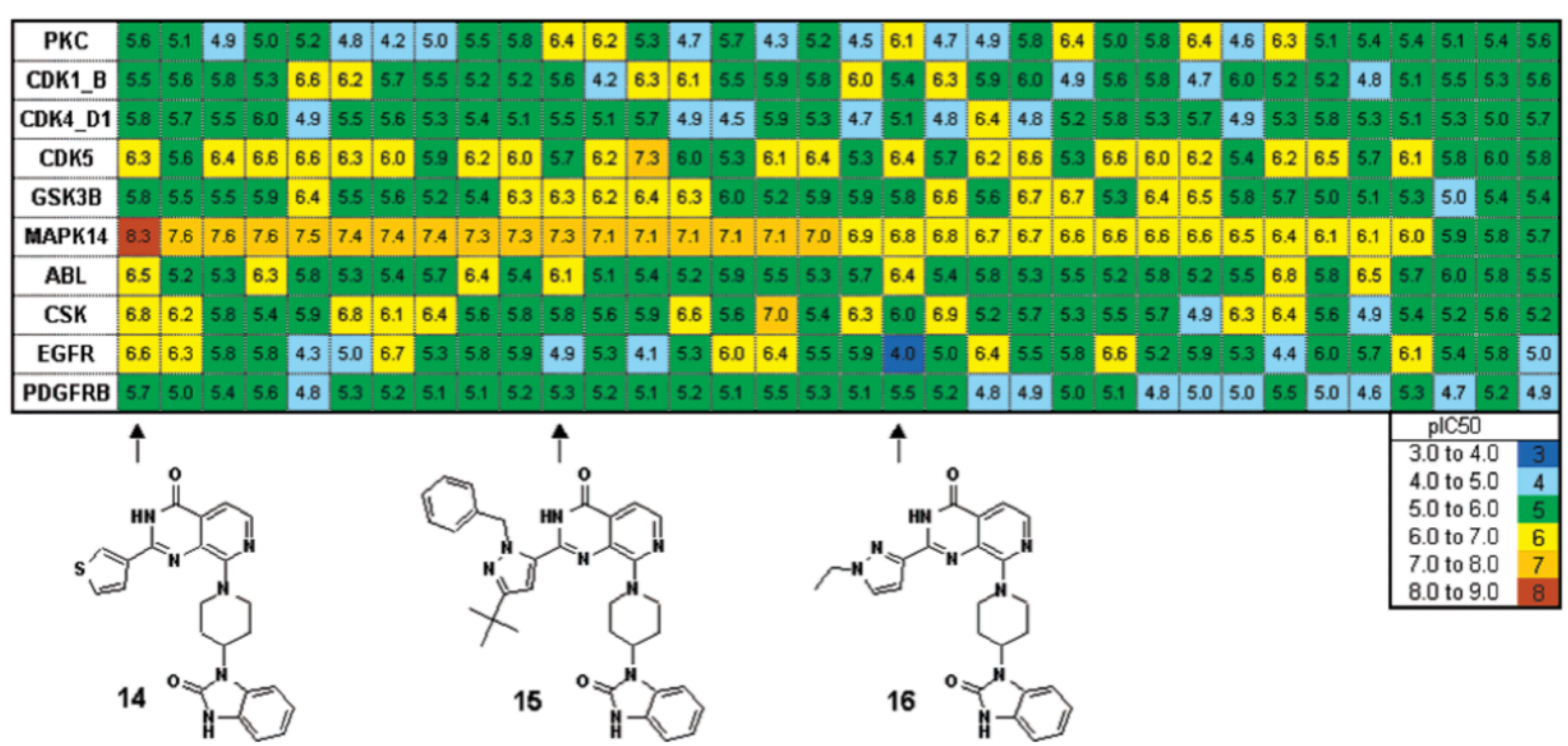
J. Chem. Inf. Comput. Sci. 1998, 38, 726-735

# Prospective Exploration of Synthetically Feasible, Medicinally Relevant Chemical Space

(2004)




"What can you make?" "What should you make?"

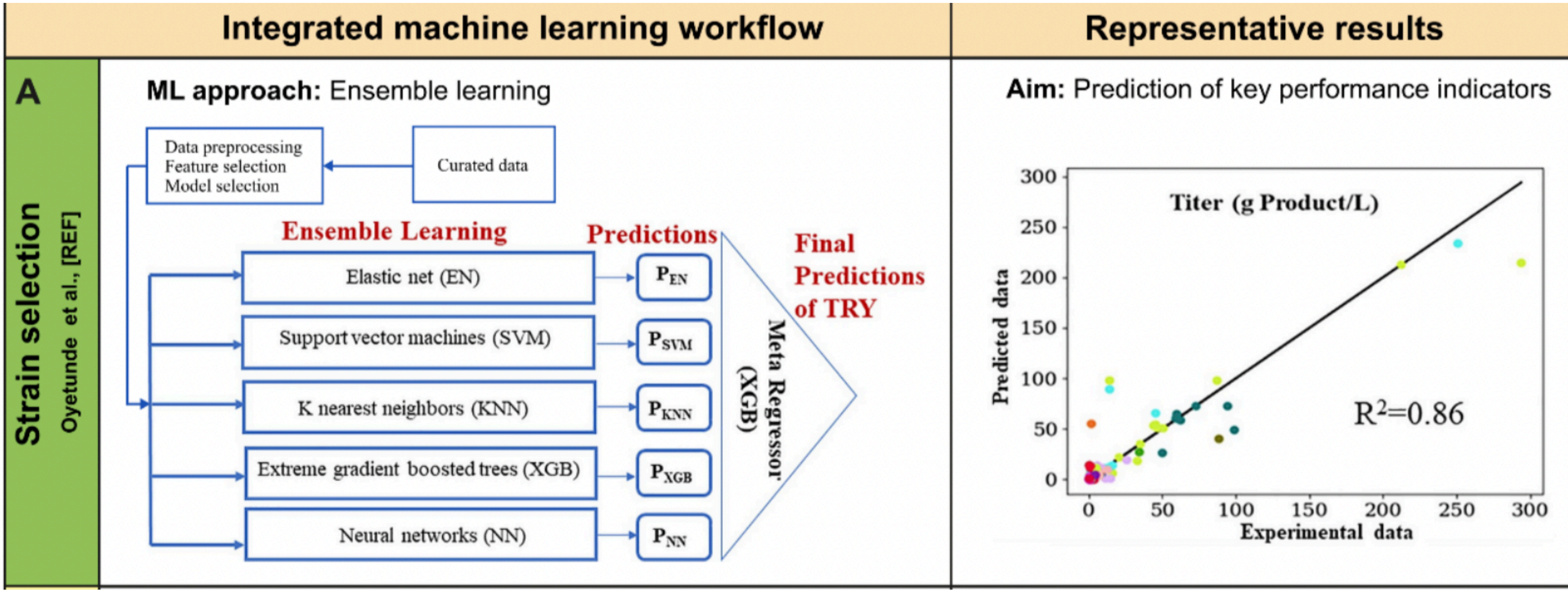



https://pubmed.ncbi.nlm.nih.gov/15807484/

- **Strain Engineering & Expression:**
  - ML analyzes genes to identify modifications for improved yield/quality.
  - Predicts effects of genetic changes on protein expression in Pichia.
- **Optimal Expression Conditions:**
  - Neural networks model & predict best growth conditions (temp, pH, nutrients).
  - Analyze past fermentations to optimize Pichia for maximum yield.
- **Scalability:**
  - Machine learning predicts how scaling affects fermentation outcomes.
  - Simulates large-scale Pichia processes based on small-scale data.
- **Methanol Utilization Control:**
  - Neural networks optimize methanol feeding in Pichia pastoris fermentations.
  - Predict optimal feeding rate for maximum protein expression without toxicity.
- **Post-Translational Modifications:**
  - AI analyzes patterns to ensure consistent, high-quality Pichia products.
  - Crucial for therapeutic proteins where modifications affect efficacy/safety.
- **Product Recovery and Purification:**
  - Machine learning optimizes downstream processing for Pichia-produced proteins.
  - Predicts most efficient purification methods and conditions.
- **Contamination Control:**
  - Neural networks monitor Pichia fermentations for real-time contamination detection.
  - Design processes that minimize contamination risk using predictive models.
- **Regulatory Compliance:**
  - AI monitors and documents Pichia production parameters for regulatory compliance.
  - Ensures processes adhere to relevant guidelines.
- **Cost-Effectiveness:**
  - Machine learning models optimize the overall Pichia production process for cost.
  - Balances factors like raw materials, energy use, and yield.
- **Advances in Bioreactor Design and Process Monitoring:**
  - Neural networks design advanced bioreactors and process monitoring systems for Pichia.
  - Analyze complex data sets to improve control strategies and reactor designs.

# Strain Selection

(2019)




Oyetunde, T. et al. (2019) Machine learning framework for assessment of microbial factory performance. PLoS One 14, e0210558. 10.1371/journal.pone.0210558

https://pubmed.ncbi.nlm.nih.gov/36456404/

# Strain Engineering

(2022)

Received: 30 June 2022 | Revised: 1 September 2022 | Accepted: 5 September 2022

Engineering Biology

DOI: 10.1049/enb2.12025

The Institution of Engineering and Technology WILEY

INDUSTRY ARTICLE

# Prediction of strain engineerings that amplify recombinant protein secretion through the machine learning approach MaLPHAS

**Evgenia A. Markova** | **Rachel E. Shaw** | **Christopher R. Reynolds**

Eden Bio Ltd, Scale Space, London, UK

**Correspondence**

Evgenia A. Markova, Eden Bio Ltd, Scale Space, 58 Wood Lane, London W12 7RZ, UK.
Email: evgenia@eden.bio

**Funding information**

Eden Bio Ltd

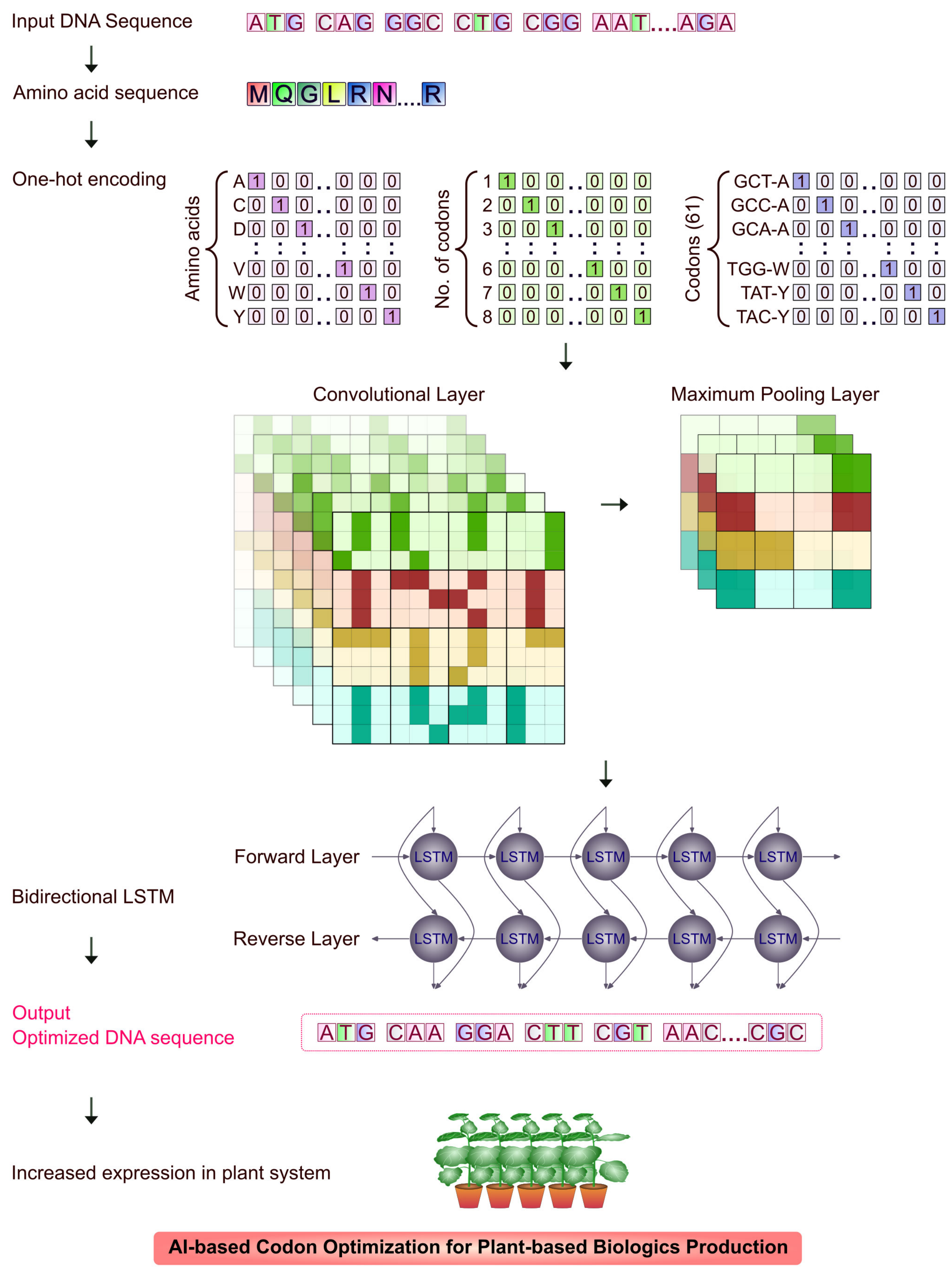
**Abstract**

This article presents a discussion of the process of precision fermentation (PF), describing the history of the space, the expected 70% growth over the next 5 years, various applications of precision fermented products, and the markets available to be disrupted by the technology. A range of prokaryotic and eukaryotic host organisms used for PF are described, with the advantages, disadvantages and applications of each. The process of setting up PF and strain engineering is described, as well as various ways that computational analysis and design techniques can be employed to assist PF engineering. The article then describes the design and implementation of a machine learning method, machine learning predictions having amplified secretion (MaLPHAS) to predict strain engineerings, which optimise the secretion of a recombinant protein. This approach showed an in silico cross-validated $R^2$ accuracy on the training data of up to 46.6% and in an in vitro test on a *Komagataella phaffii* strain, identified one gene engineering out of five predicted, which was shown to double the secretion of a heterologous protein and outperform three of the best-known edits from the literature for improving secretion in *K. phaffii*.

https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/enb2.12025

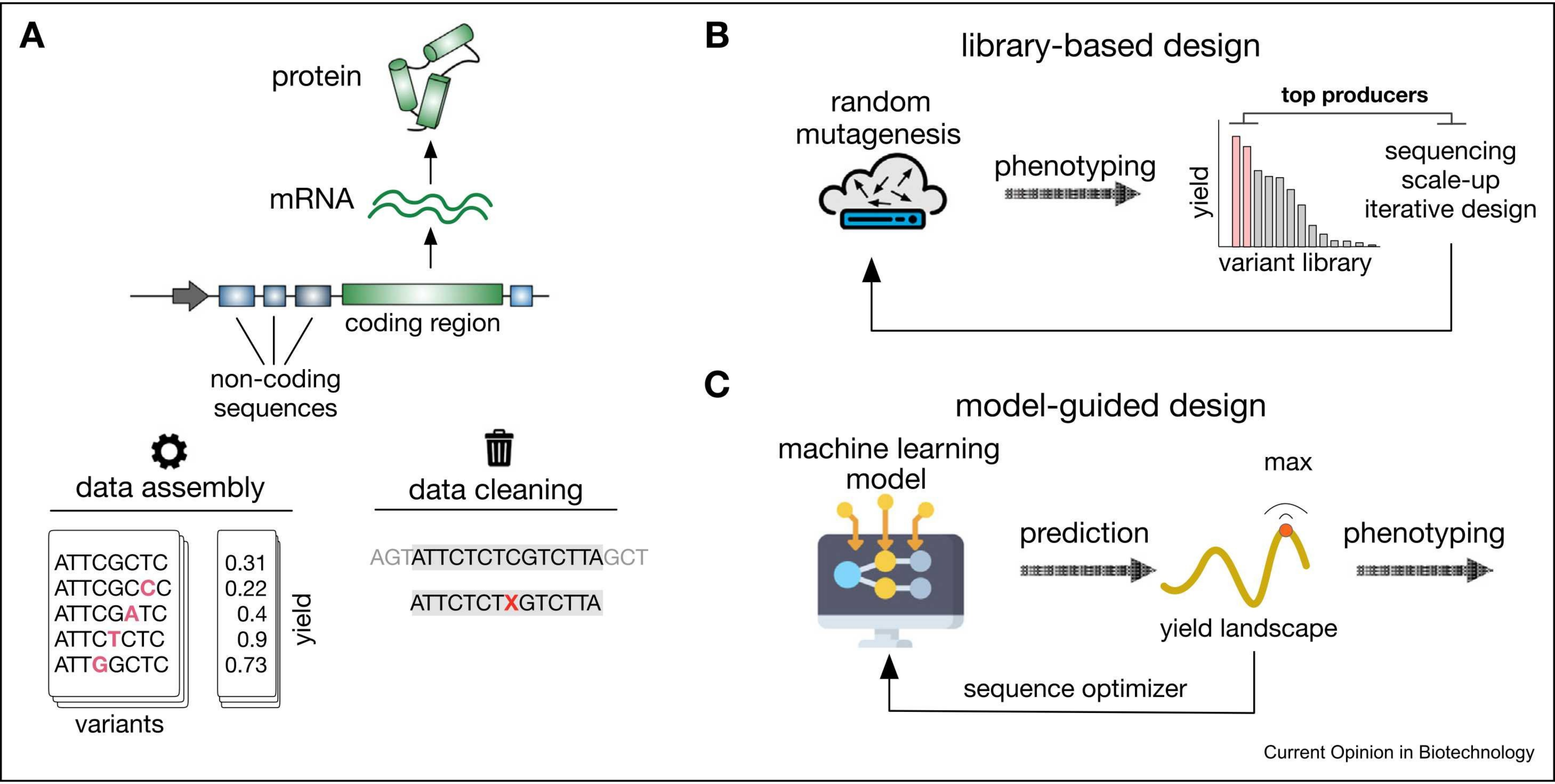
# Codon optimization for plant expression system

(2023)




https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1252166/full

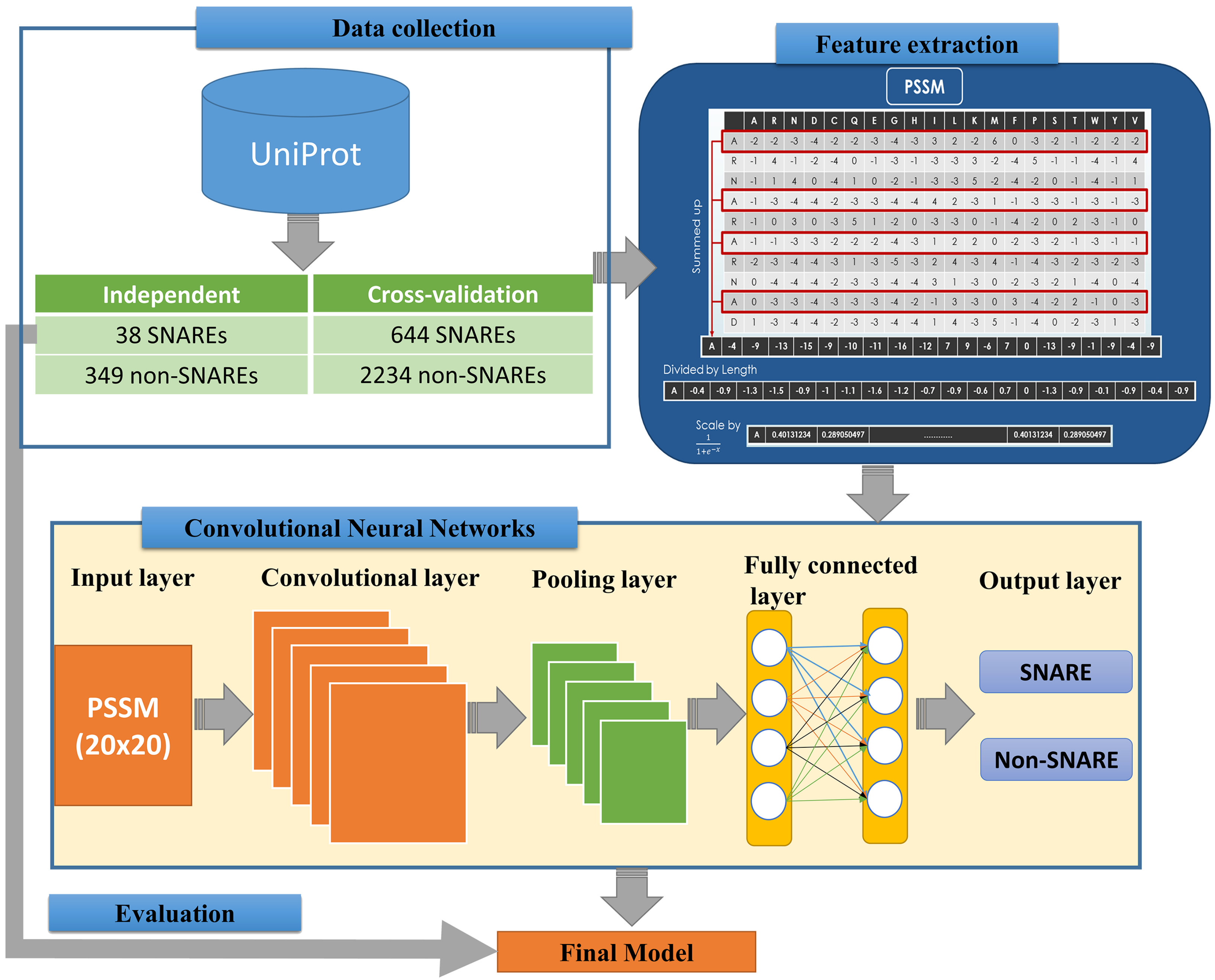
# Deep learning for optimization of protein expression (2023)




https://pubmed.ncbi.nlm.nih.gov/37087839/

# SNARE-CNN: a 2D Convolutional Neural Network Architecture to Identify SNARE Proteins from High-Throughput Sequencing Data

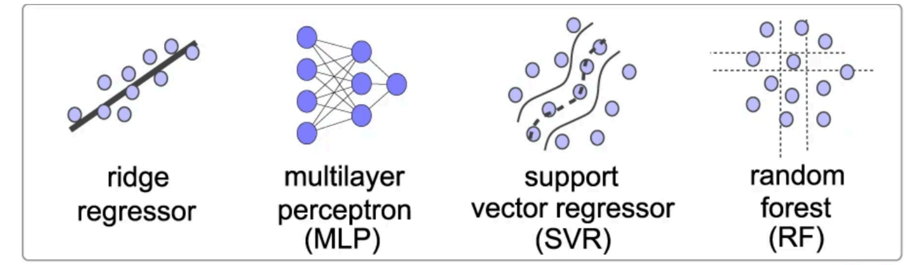
(2019)

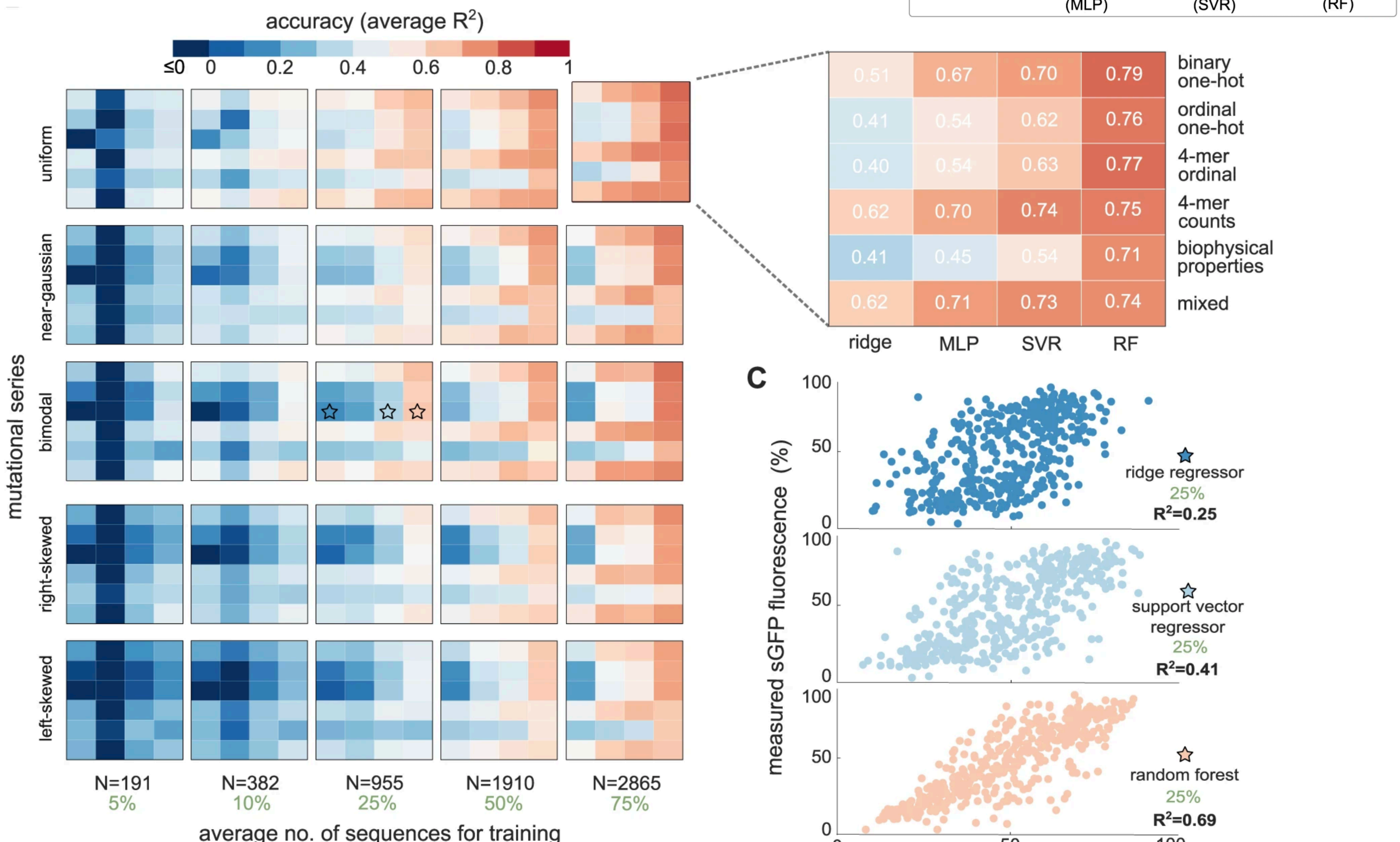



https://peerj.com/articles/cs-177/

# Accuracy and Data Efficiency in Deep Learning Models of Protein Expression

(2022)

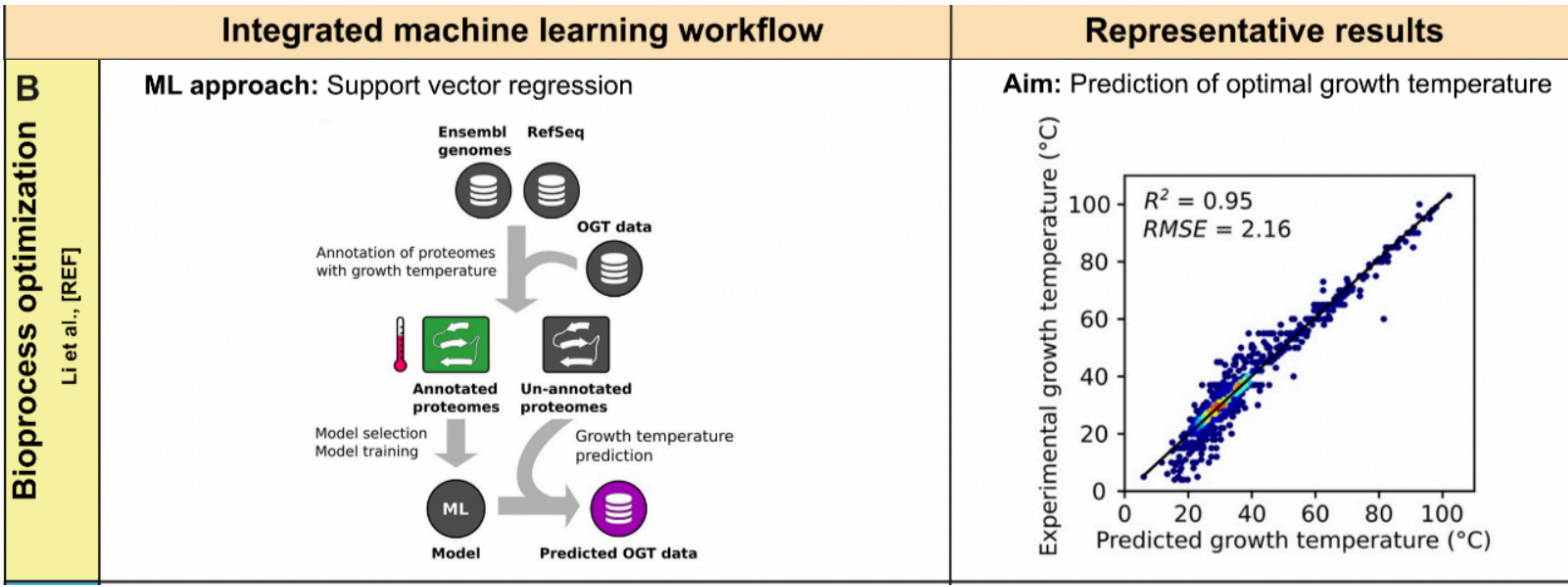



https://www.nature.com/articles/s41467-022-34902-5

# Bioprocess Optimization

(2019)




Li, G. et al. (2019) Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. ACS Synth Biol 8, 1411-1420. 10.1021/acssynbio.9b00099

https://pubmed.ncbi.nlm.nih.gov/36456404/

# Process optimization

(2019)

ELSEVIER

Biochemical Engineering Journal

Volume 144, 15 April 2019, Pages 18-27

Regular article

## A robust feeding control strategy adjusted and optimized by a neural network for enhancing of alpha 1-antitrypsin production in *Pichia pastoris*

Tina Tavasoli [a] [1], Sareh Arjmand [b] [1], Seyed Omid Ranaei Siadat [b], Seyed Abbas Shojaosadati [a], Abbas Sahebghadam Lotfi [c]

[a] Biotechnology Group, Department of Chemical Engineering, Tarbiat Modares University, Tehran, Iran

[b] Protein Research Center, Shahid Beheshti University, G.C., Tehran, Iran

[c] Department of Clinical Biochemistry, Faculty of Medical Science, Tarbiat Modares University, Tehran, Iran

Received 19 June 2018, Revised 12 December 2018, Accepted 2 January 2019, Available online 3 January 2019, Version of Record 16 January 2019.

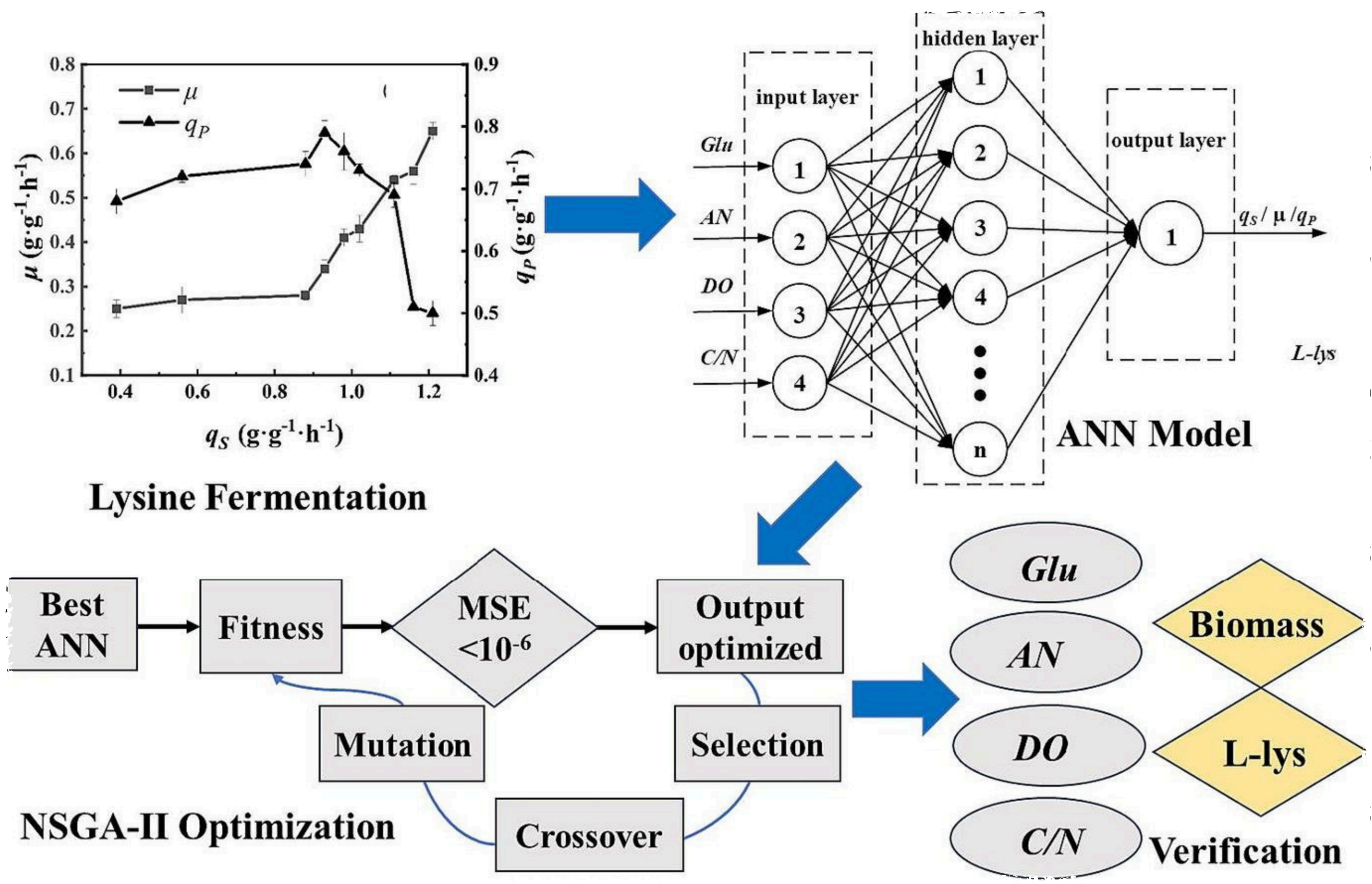
### Highlights

- Novel on-line μ-stat approach is presented for regulating methanol feeding rate in fermenter.
- Methanol feeding was controlled in fed-batch based on the on-line ammonia consumption rate.
- MLP3 neural network was used to reconstruct the controller.
- The designed controller was used for A1AT production process control in *P. pastoris*.
- Control and maintaining μ at the optimal level led to increase in target protein production.

https://www.sciencedirect.com/science/article/abs/pii/S1369703X19300051

# Artificial Neural Network and Genetic Algorithm Coupled Fermentation Kinetics to Regulate L-Lysine Fermentation

(2024)



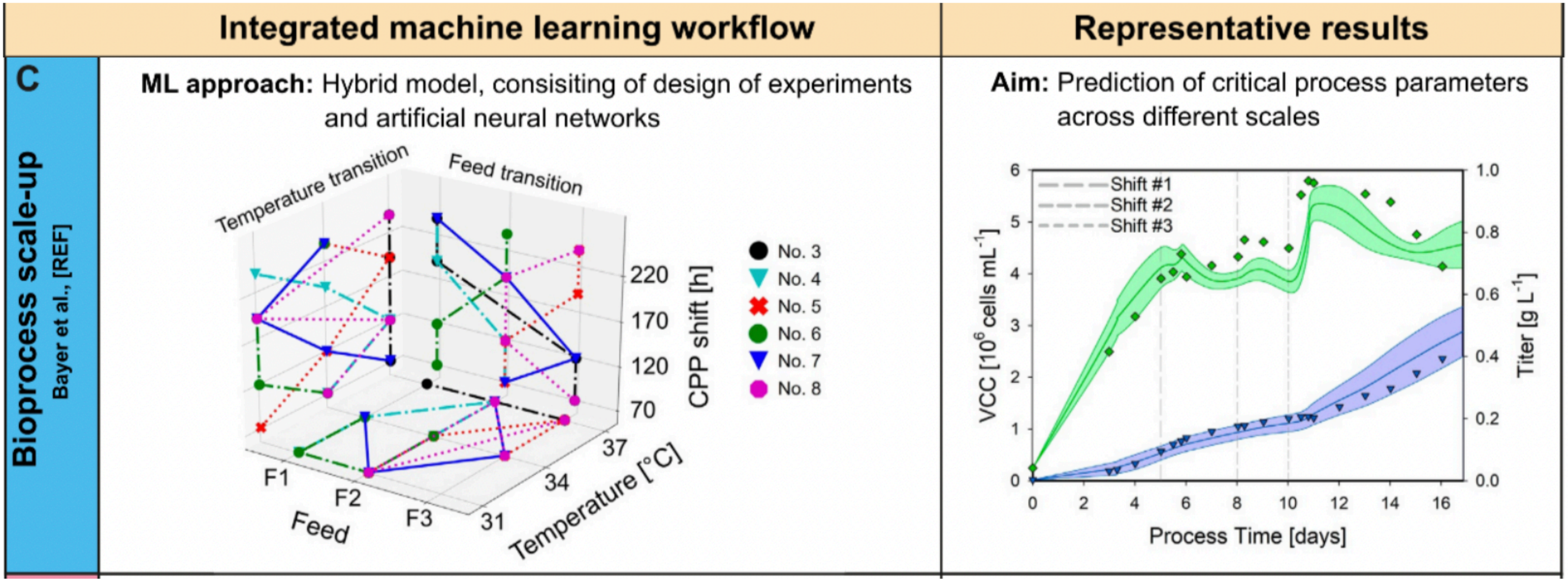

## Highlights

- Artificial neural network (ANN) coupled genetic algorithm (GA) was established.
- ANN-GA was utilized to establish fermentation control strategy.
- ANN-GA coupled fermentation kinetics was employed to regulate lysine fermentation.
- Optimal parameters were achieved using ANN-GA coupled fermentation kinetics.
- The ANN-GA optimized model showed significantly enhanced lysine yield.

https://doi.org/10.1016/j.biortech.2023.130151

# Bioprocess Scale-Up

(2021)

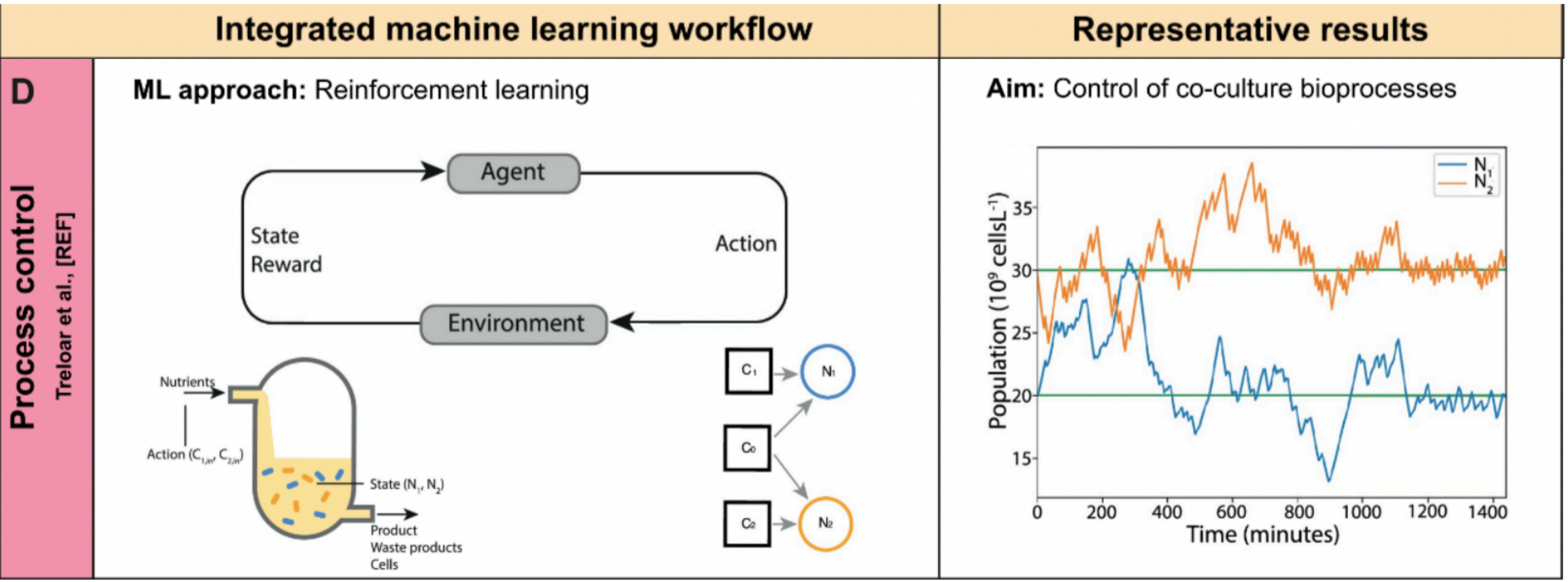



Bayer, B. et al. (2021) Model Transferability and Reduced Experimental Burden in Cell Culture Process Development Facilitated by Hybrid Modeling and Intensified Design of Experiments. Front Bioeng Biotechnol 9, 740215. 10.3389/fbioe.2021.740215

https://pubmed.ncbi.nlm.nih.gov/36456404/

# Process Control

(2020)



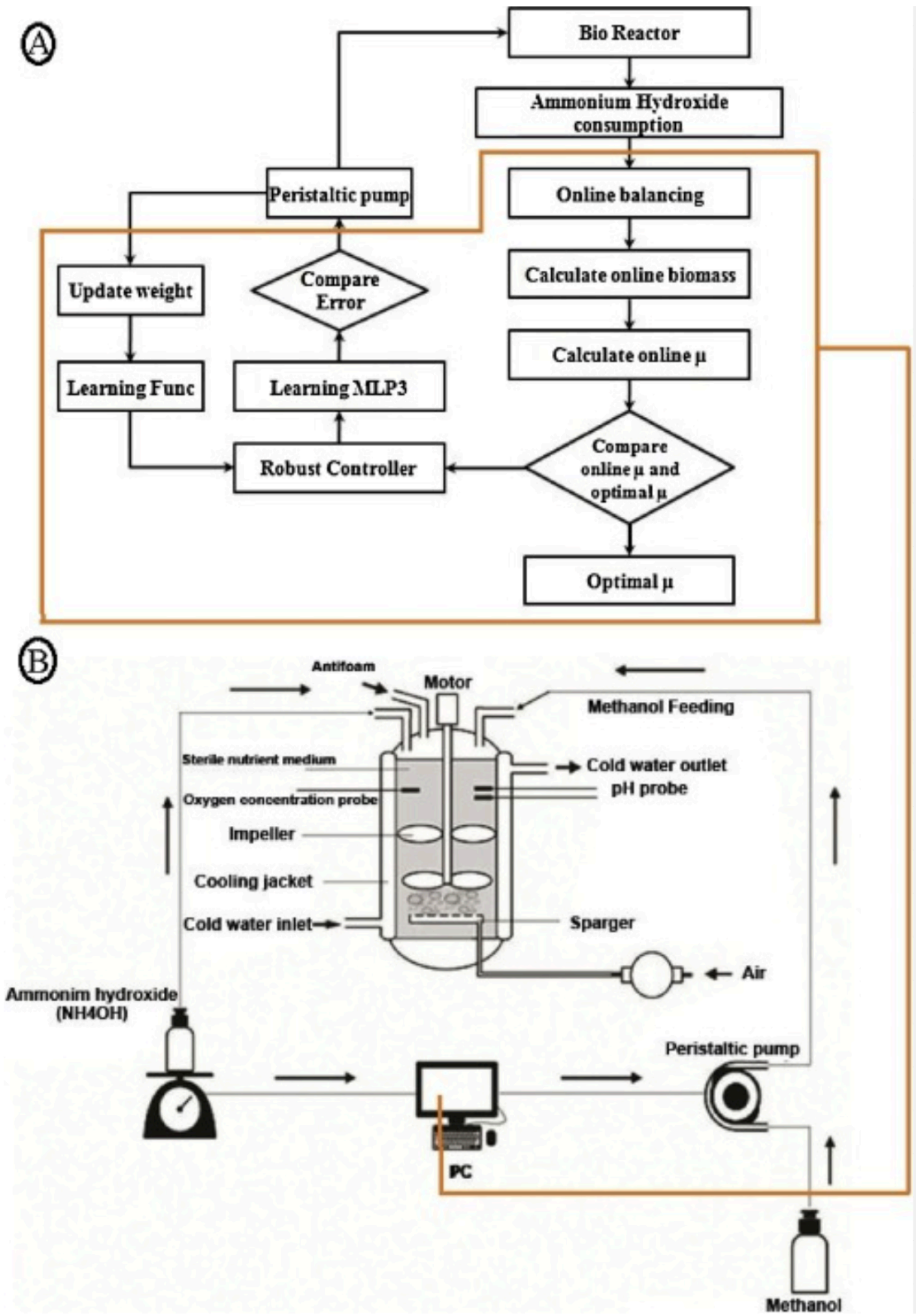

Treloar, N.J. et al. (2020) Deep reinforcement learning for the control of microbial co-cultures in bioreactors. PLoS Comput Biol 16, e1007783. 10.1371/journal.pcbi.1007783

https://pubmed.ncbi.nlm.nih.gov/36456404/

# MeOH Feeding Control Strategy Adjusted/ Optimized by a Neural Network

## alpha 1-antitrypsin production in Pichia pastoris

(2019)

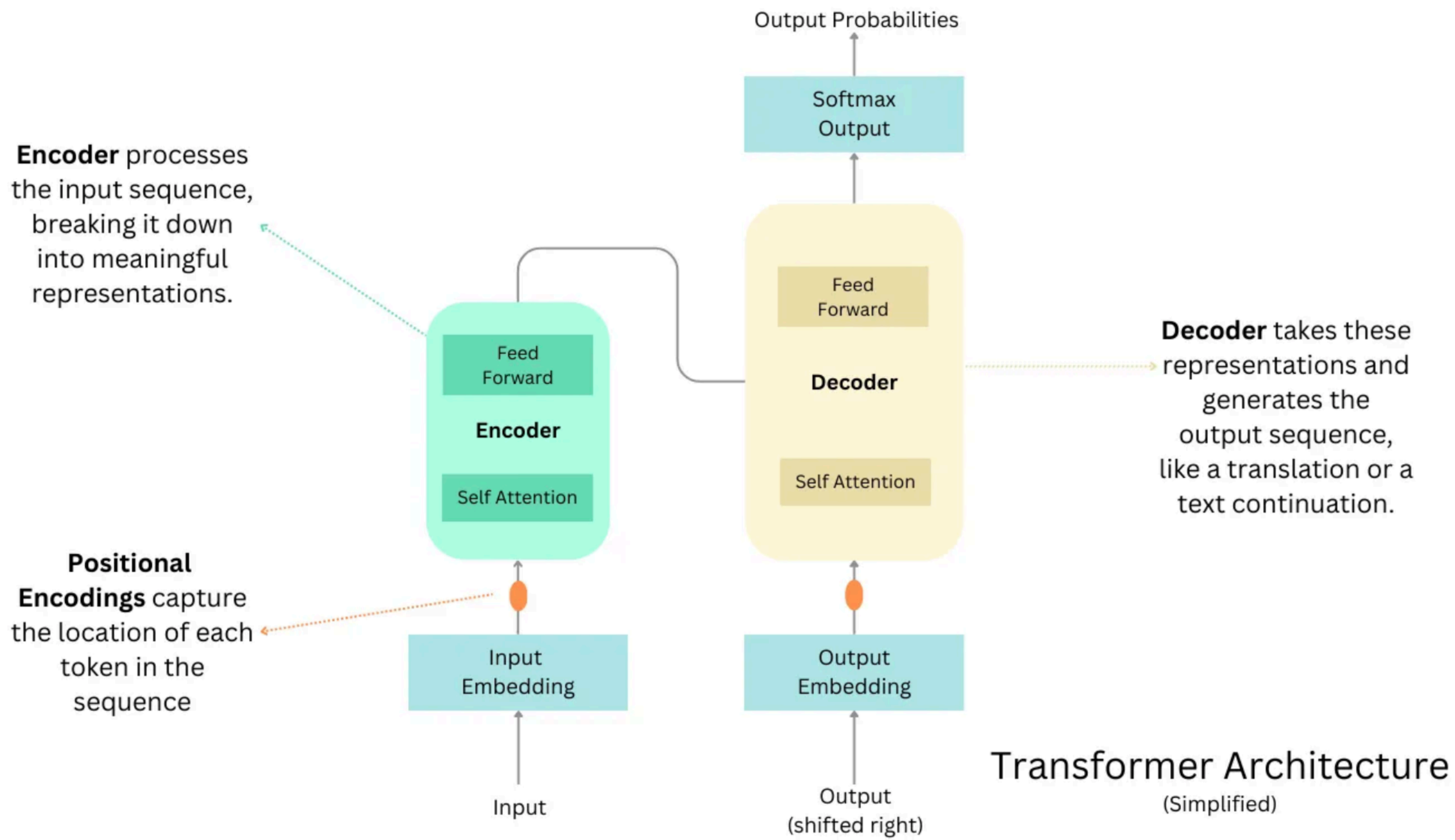



Highlights

- Novel on-line μ-stat approach is presented for regulating methanol feeding rate in fermenter.
- Methanol feeding was controlled in fed-batch based on the on-line ammonia consumption rate.
- MLP3 neural network was used to reconstruct the controller.
- The designed controller was used for A1AT production process control in *P. pastoris*.
- Control and maintaining μ at the optimal level led to increase in target protein production.

https://www.sciencedirect.com/science/article/abs/pii/S1369703X19300051

# Transformer Architecture Simplified




https://medium.com/@tech-gumptions/transformer-architecture-simplified-3fb501d461c8 https://medium.com/machine-intelligence-and-deep-learning-lab/transformer-the-self-attention-mechanism-d7d853c2c621

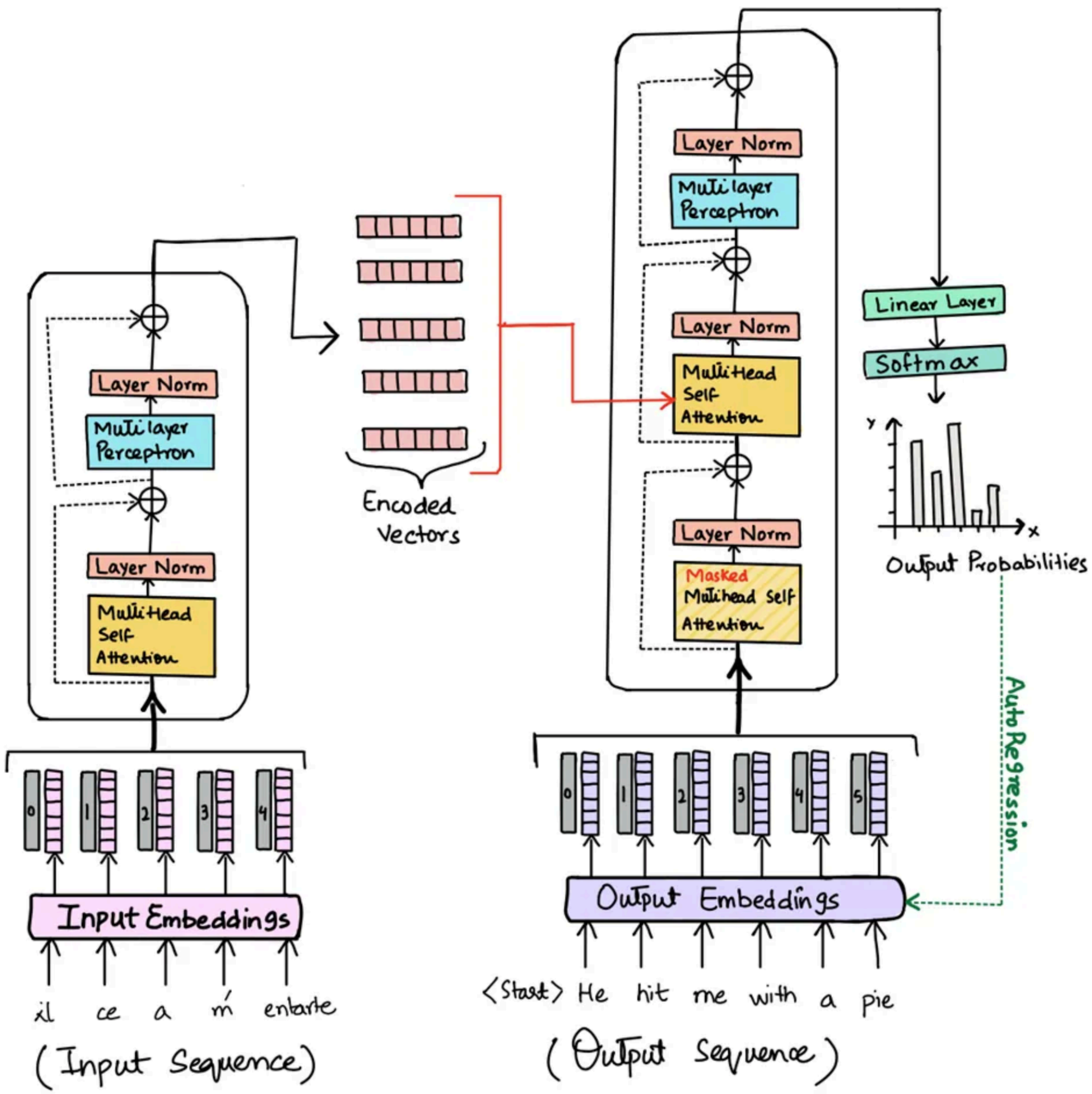



Figure 1. Transformer Architecture

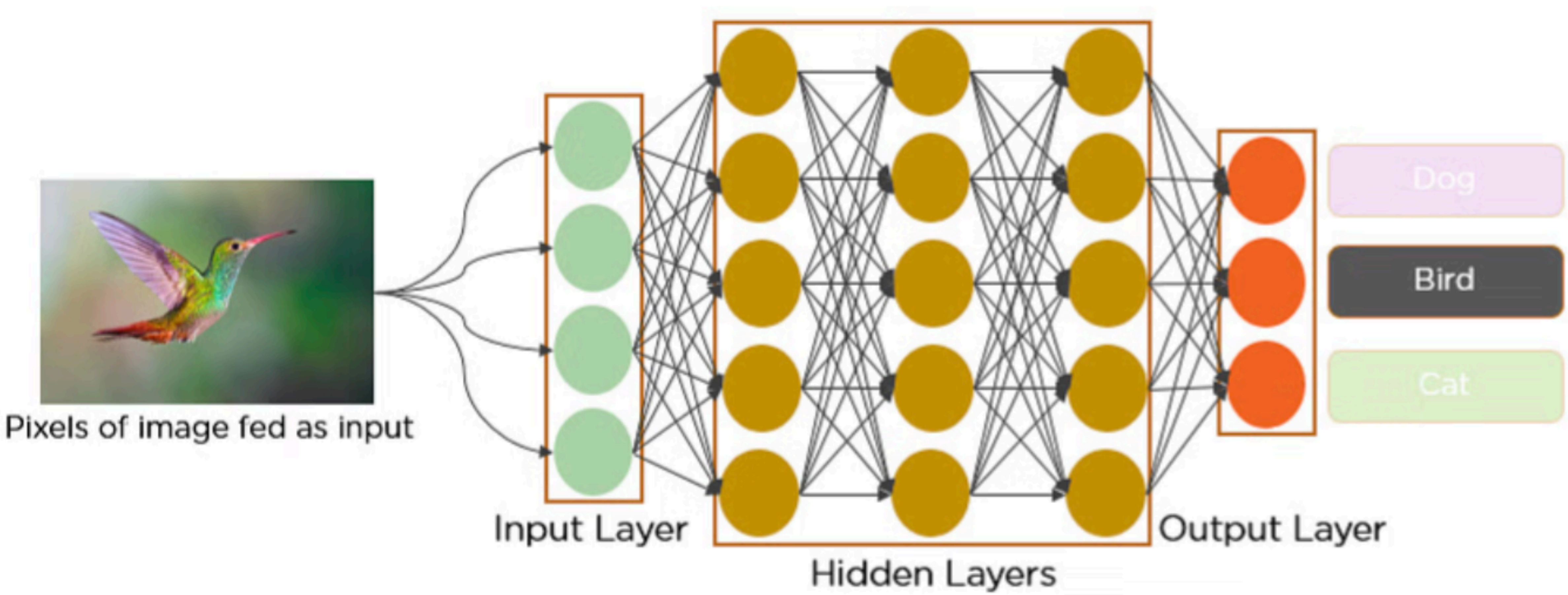
Here's a breakdown of the diagram:

- **Encoder-decoder architecture:** The model is divided into two parts: an encoder and a decoder. The encoder's role is to receive the input sequence and transform it into a contextual representation. The decoder's role is to generate the output sequence based on the encoded representation.
- **Layers:** Both the encoder and decoder consist of multiple layers stacked on top of each other. Each layer refines the previous layer's output.
- **Self-attention mechanism:** This is a core component of the Transformer model. It allows the model to attend to relevant parts of the input sequence when processing a word. In the image, this is depicted by the "Multi-head Self Attention" block.
- **Layer normalization:** This is another essential building block that helps stabilize the training process of the model.
- **Input and output embeddings:** These are dense vector representations of words in the input and output sequences. The embedding layer maps words from the vocabulary space into a continuous vector space.

Overall, the Transformer model takes an input sequence, encodes it, and then decodes it to generate an output sequence. The self-attention mechanism allows the model to focus on important parts of the input sequence when generating the output.

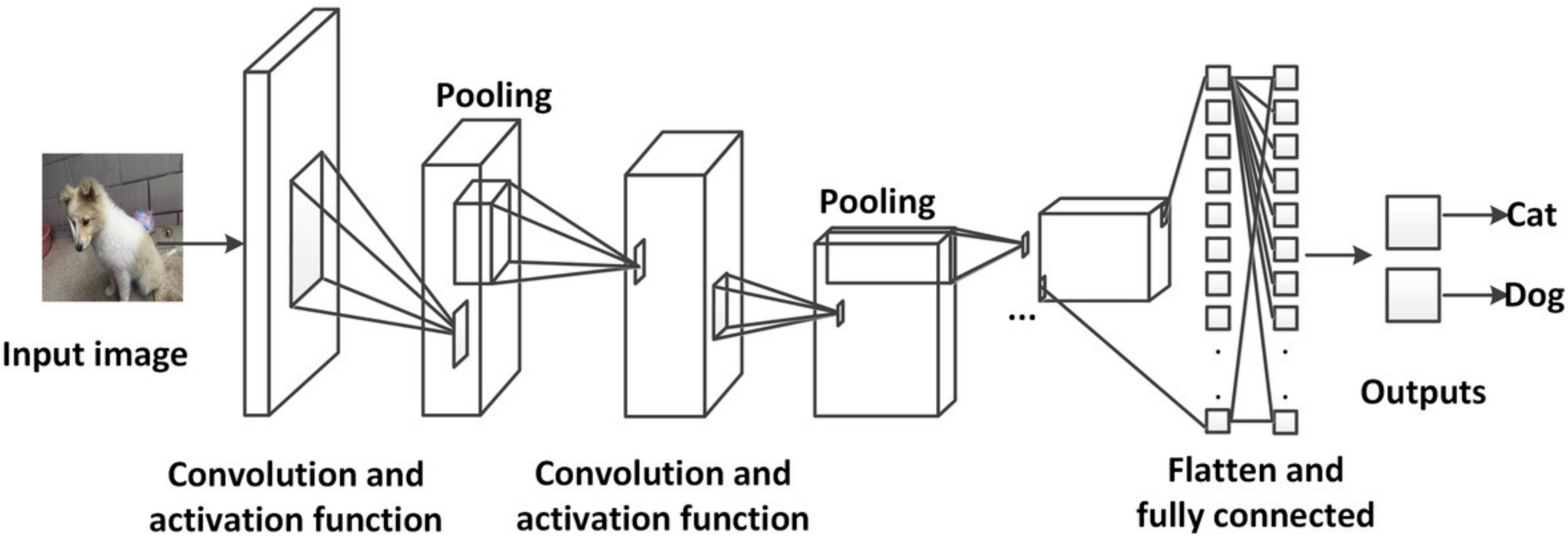
# Explaining Convolutional Neural Networks

https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network



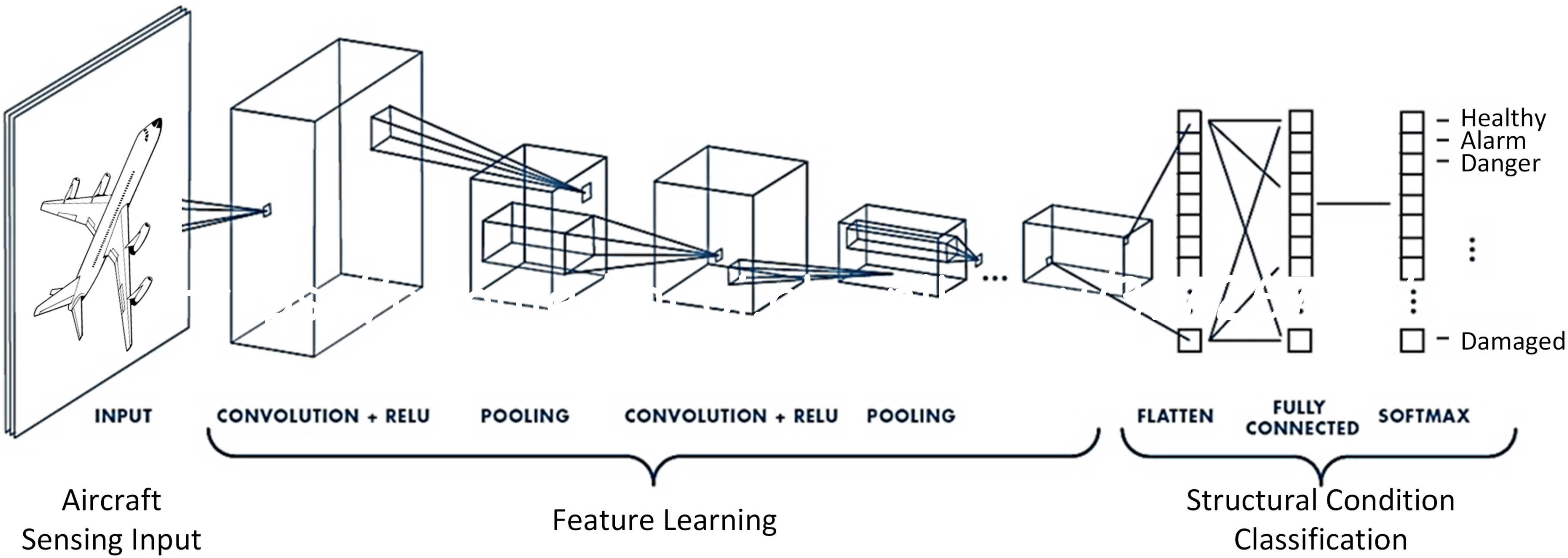

https://poloclub.github.io/cnn-explainer/

# Convolutional Neural Network: Image Classification




https://peerj.com/articles/cs-773/

# Convolutional Neural Network: Structural Health Monitoring




https://pubmed.ncbi.nlm.nih.gov/31726762/