Applications of AI in the Chemical Industry: A Tutorial with Applications to Process Safety

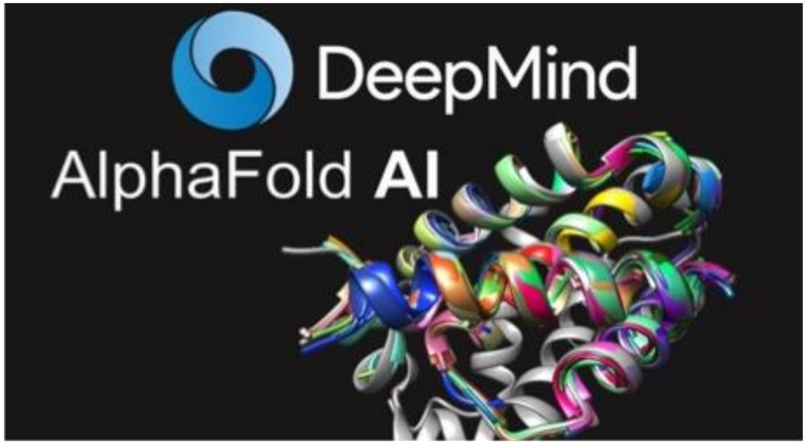
Can Li
Davidson School of Chemical Engineering

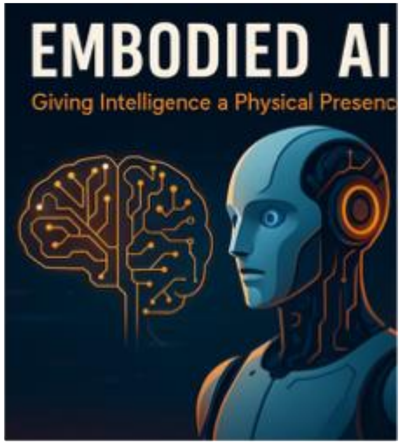
Outline

- ☐ Introduction to the basics of modern AI/machine learning
- ☐ Existing applications of AI in the chemical industry
- ☐ Limitations of existing approaches and future prospects

Recent Advances in AI/ML

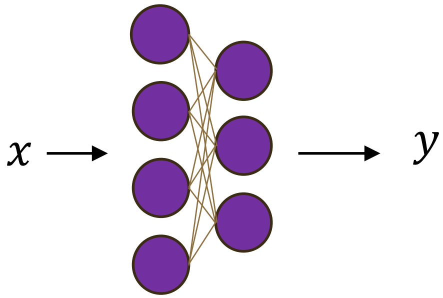
- Huge success in the past decade: vision, audio, science, generative AI, embodied AI.

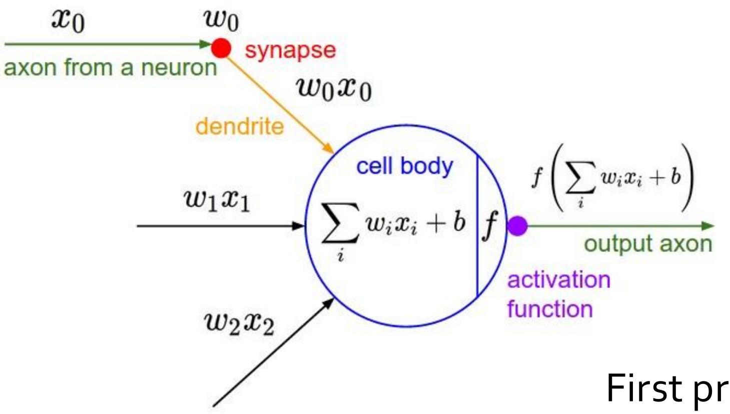




What’s Driving the Success

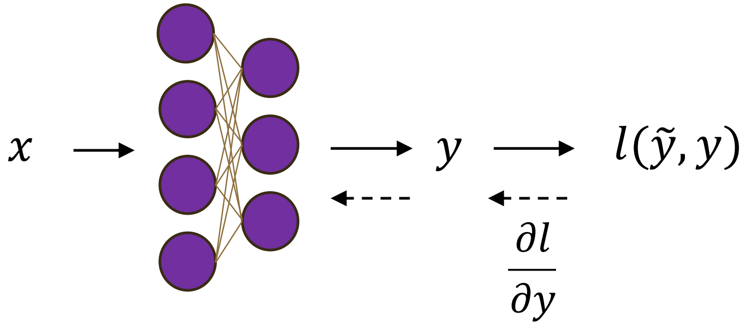
- **Models: deep neural networks**
  - Theoretically, they can represent any continuous nonlinear functions $y = f(x)$ arbitrarily well





First proposed by
McCulloch and Pitts in **1943**

- **Algorithms: backpropagation**
  - Update the weights by comparing with the correct answer



Popularized by Hinton and
co-workers in **1986**

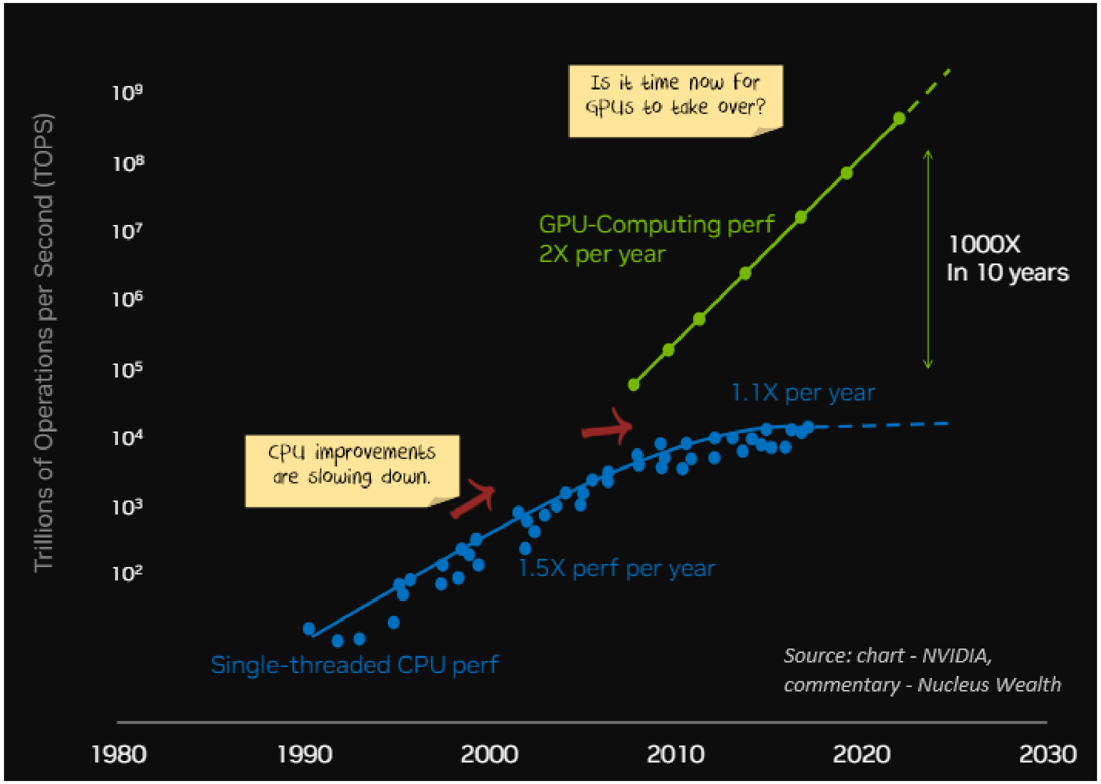
Geoffrey Hinton
Nobel prize 2024, Turing Award 2018

What's Driving the Success

- Abundance of data

  Large language models are trained on **trillions** of tokens of web scale data

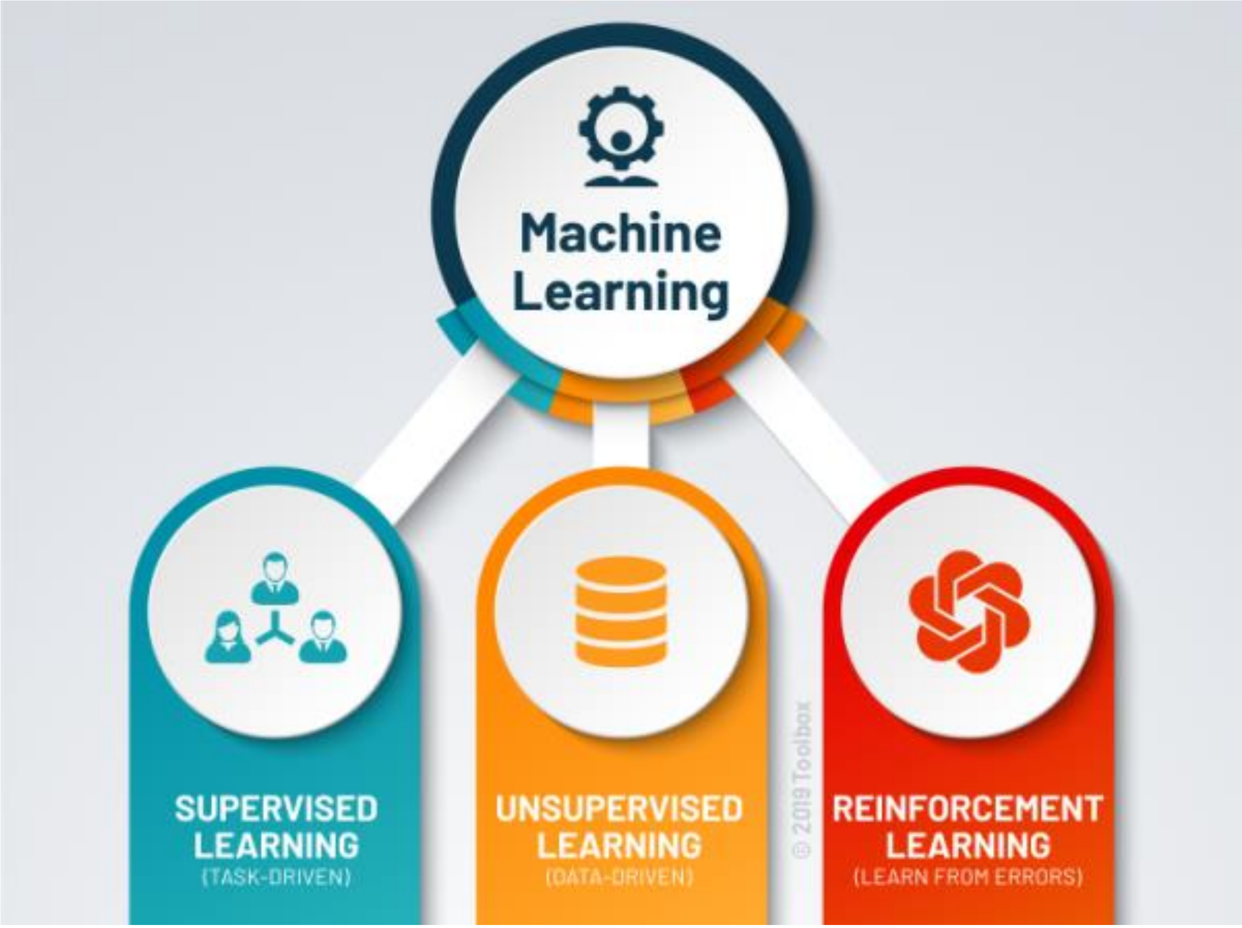
- Advances in hardware, particularly GPUs



Source: chart - NVIDIA, commentary - Nucleus Wealth

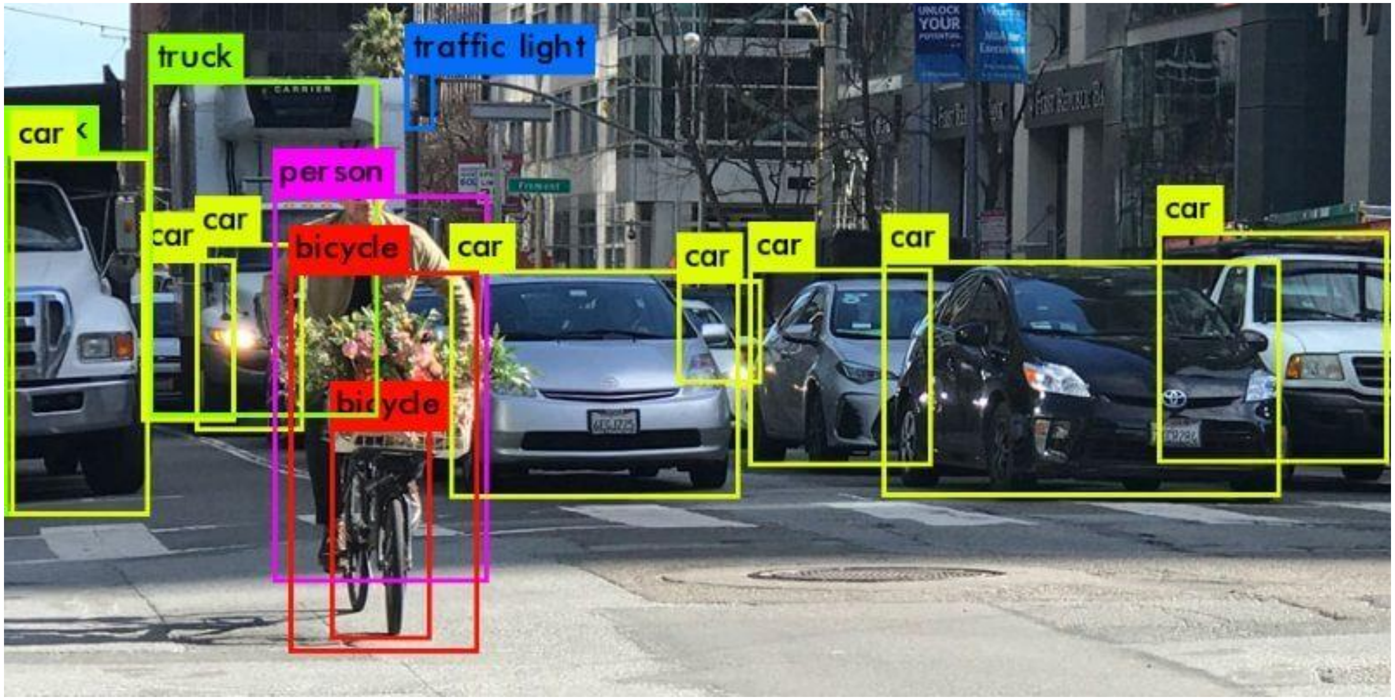
Types of Machine Learning

- Machine learning uses data to train a model and then use that model for prediction or decision-making.



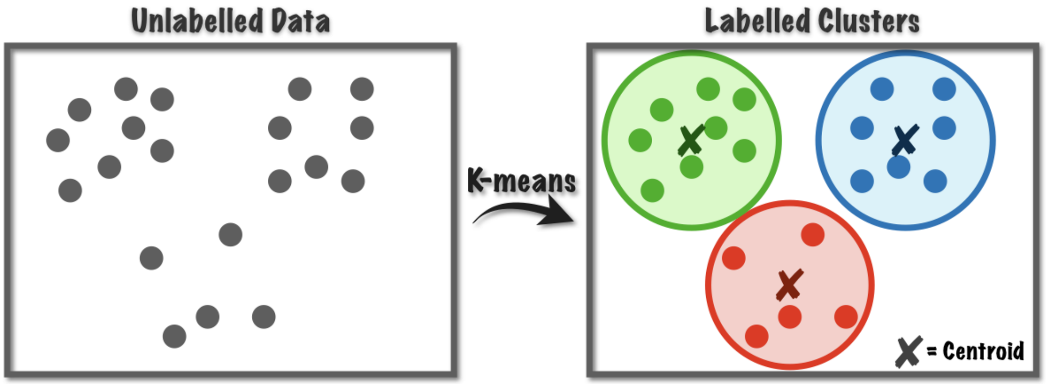
Supervised Learning

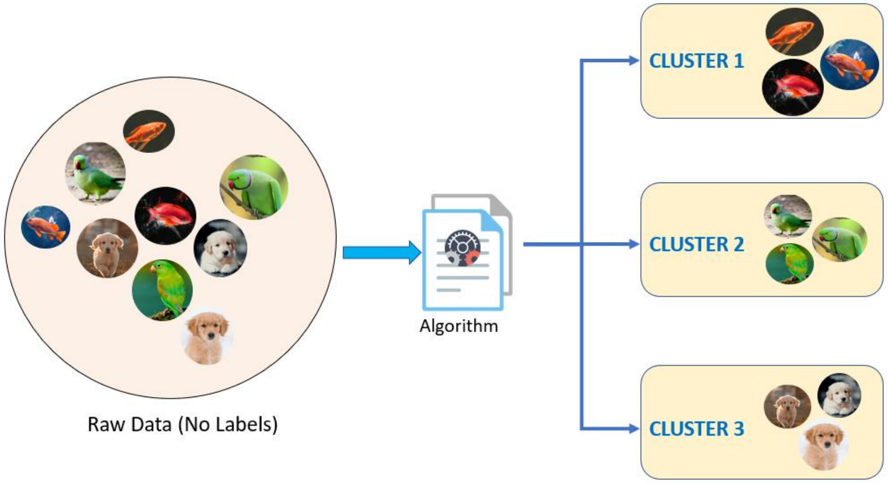
- Task: Given X predict y
- Given i.i.d. data $X_1, X_2, \dots, X_n$ and labels provided by an expert $y_1, y_2, \dots, y_n$
- Train a classifier f such that $f(X_i) \approx y_i$ (This is a mathematical optimization problem). Examples include neural networks, support vector machines.
- Test the model you have trained and use it in practice



Unsupervised learning

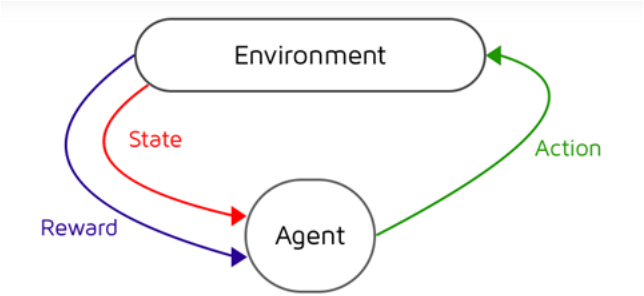
- Expert labels are unavailable. Usually used for identifying patterns from data
- Hidden Markov models, k-means, hierarchical clustering, and Gaussian mixture models

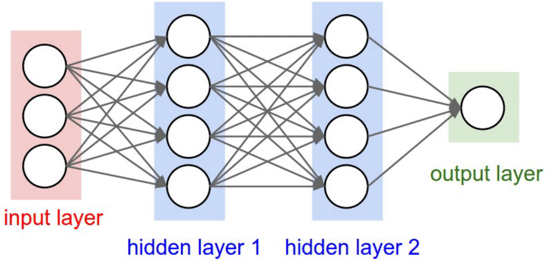


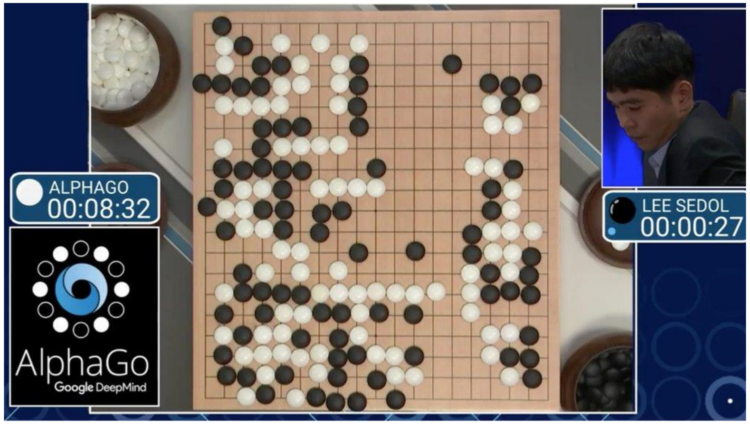


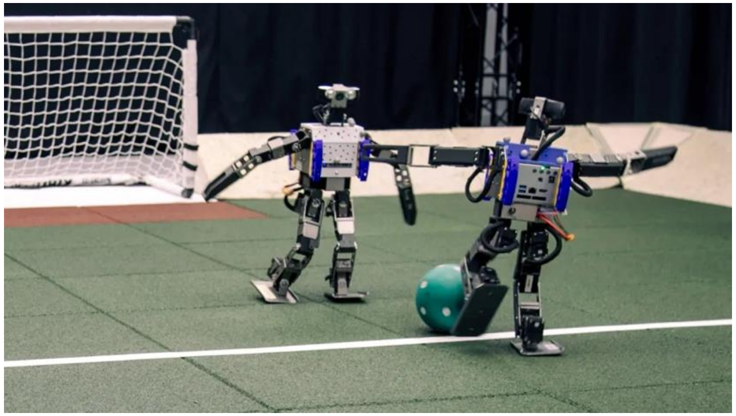
Reinforcement Learning

- Use AI to make decisions (take actions). Learning by doing.
- Exploration v.s. Exploitation. ϵ-greedy policy.



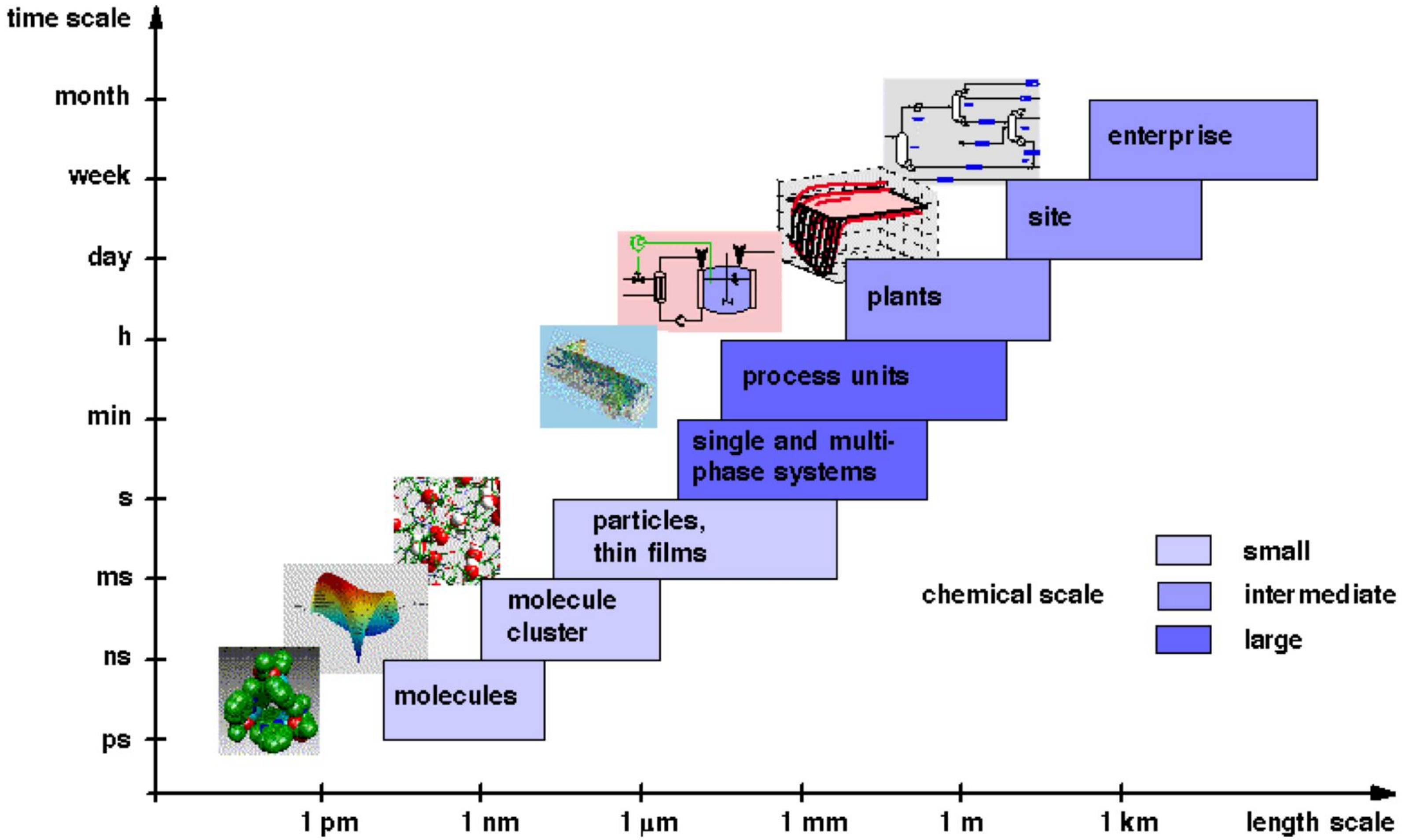






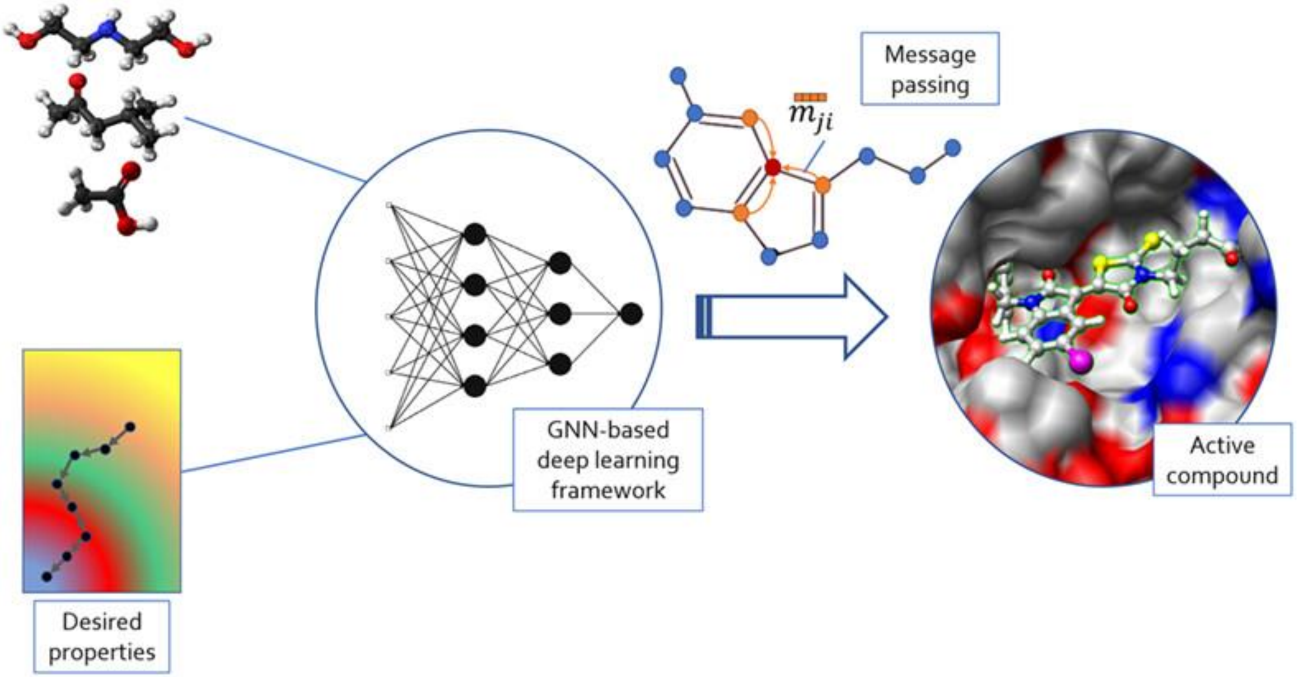
Existing Applications of AI in the Chemical Industry

Multi-Scale Applications



Predict Molecular Property

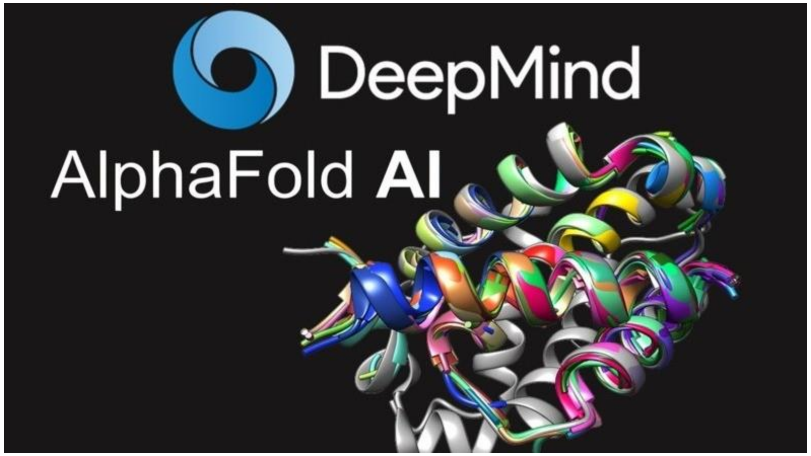
- Predict key molecular properties (e.g., solubility, toxicity, activity) from structure
- Accelerate materials and drug discovery through data-driven modeling
- Reduce need for costly experiments or simulations
- Enable inverse design of molecules with desired properties

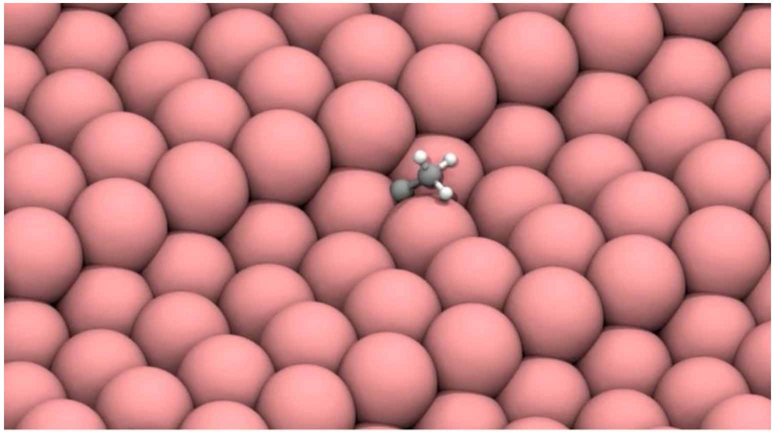


Source: Abate et al. 2023

Predict Property/Structure of Polymers/Catalysis

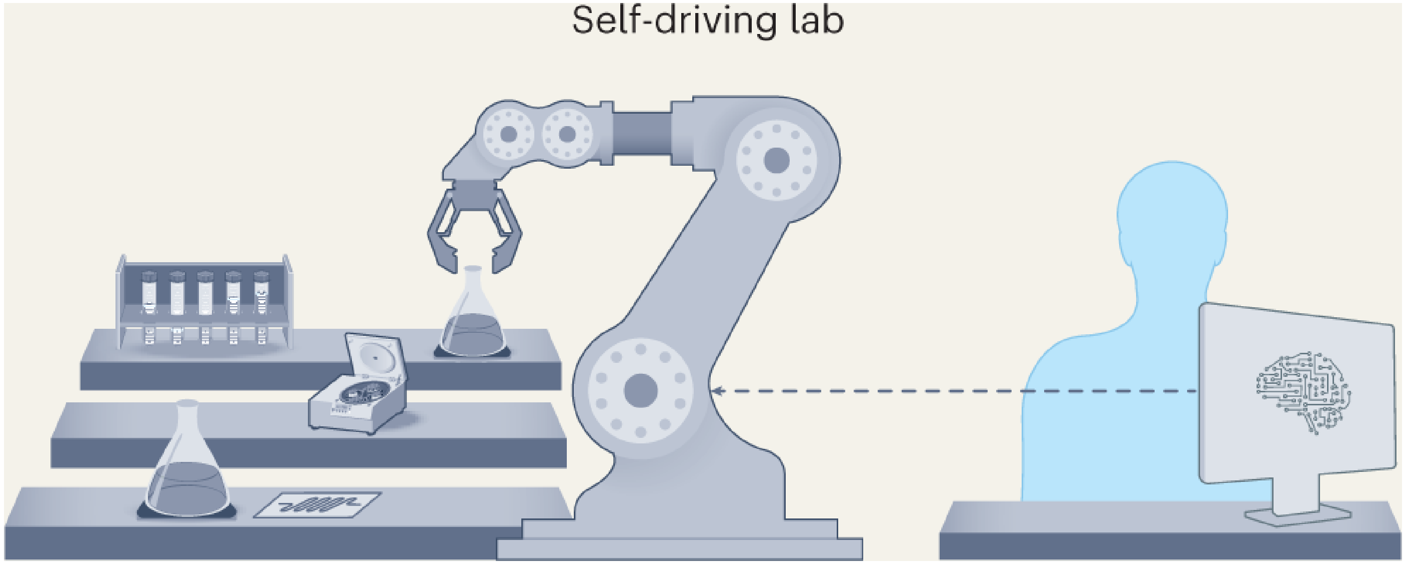
- AlphaFold: Nobel Prize in Chemistry (2024). Predict 3D protein structure from the sequence of amino acids. Replace expensive X-ray experiments
- Meta AI's Open Catalyst project: use AI for catalyst discovery. Replace DFT simulation and experiments.





Self-Driving Lab

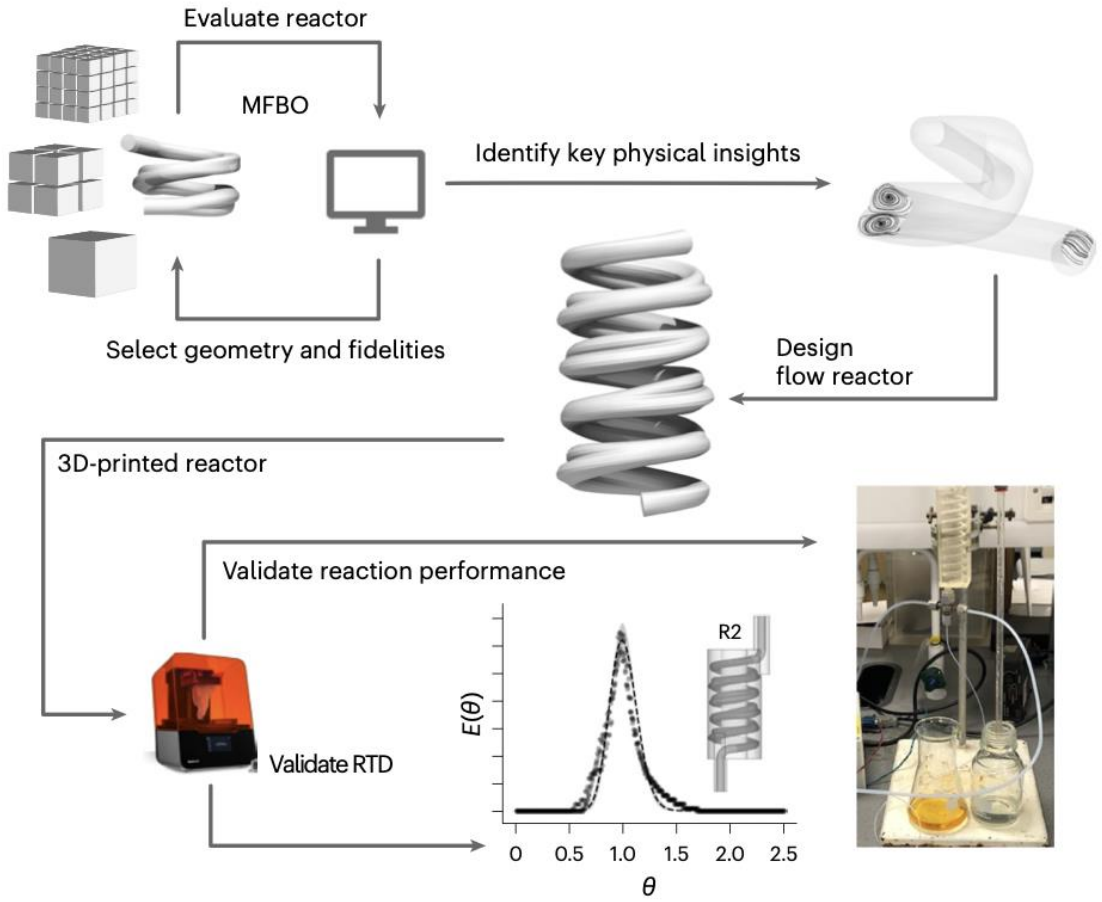
- Automate experimentation with AI-driven design–make–test–learn cycles
- Integrate robotics, machine learning, and real-time analytics for closed-loop discovery
- Optimize reaction conditions or material synthesis autonomously
- Accelerate discovery of catalysts, polymers, and battery materials
- Reduce human error and experimental cost while improving reproducibility



Source: Abolhasani & Kumacheva (2023)

Reactor Design

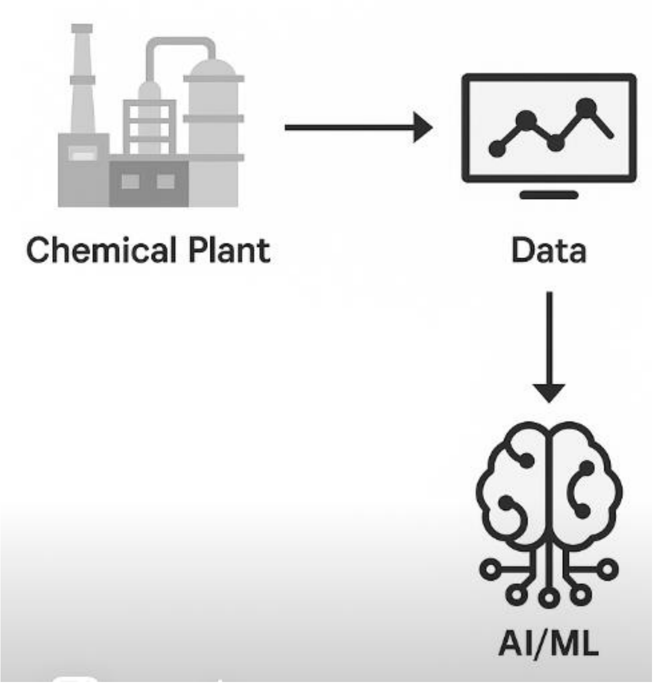
- Use Multi-Fidelity Bayesian Optimization (MFBO) to design efficient flow reactors
- Optimize reactor geometry by combining simulations and experimental data
- 3D-print and validate optimized designs through RTD and performance tests
- Iteratively refine models to uncover key physical insights and improve design

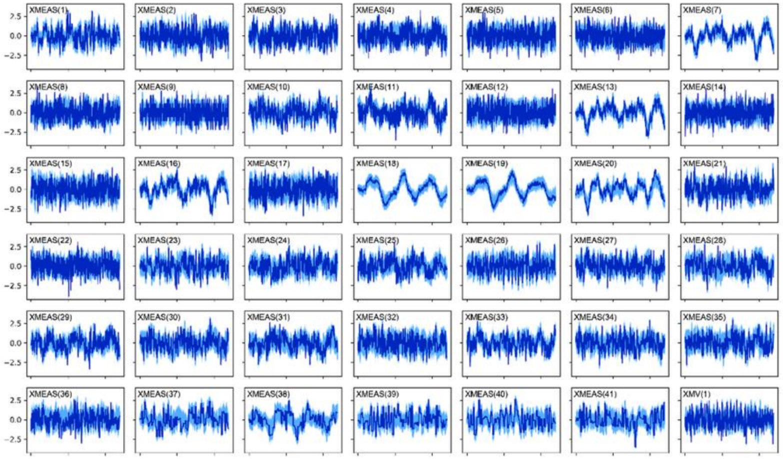


Source: Savage et al. 2024

Fault Detection and Diagnosis

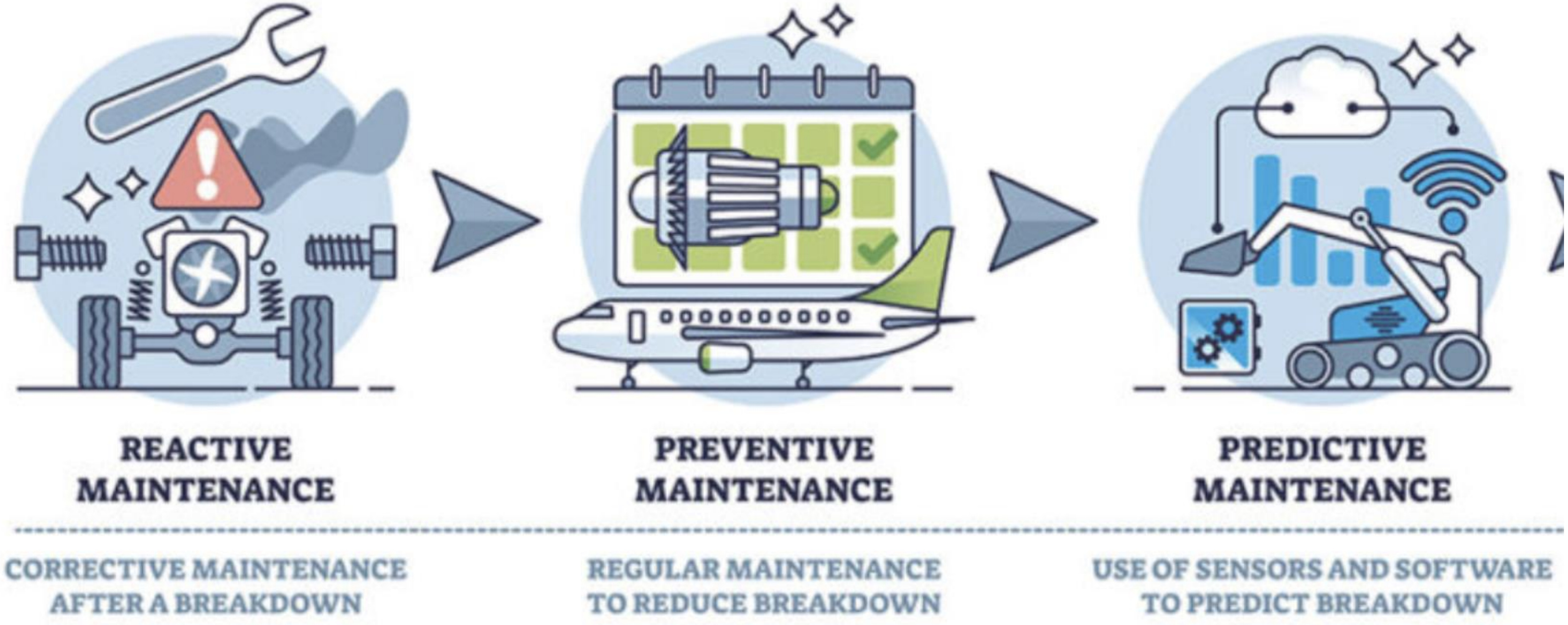
- Detect abnormal process behavior using sensor and operational data
- Diagnose root causes with machine learning and hybrid (physics + data) models
- Enable early intervention to prevent safety incidents and equipment damage





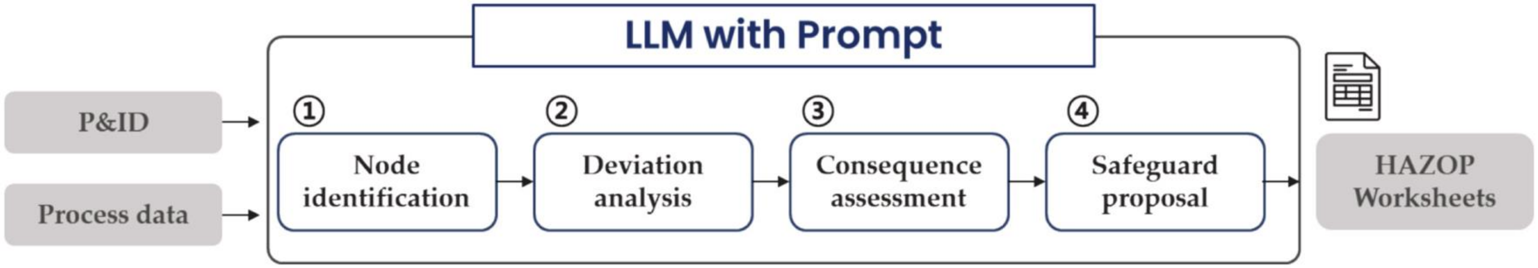
Predictive Maintenance

- **Reactive Maintenance:** Fix equipment only after a failure occurs — leads to unplanned downtime.
- **Preventive Maintenance:** Perform scheduled maintenance to avoid breakdowns — based on time or usage.
- **Predictive Maintenance:** Use sensors and AI to forecast failures before they happen — minimizes cost and disruption.



AI Assisted HAZOP Analysis

- Large Language Models (LLMs) can assist in automating the HAZOP (Hazard and Operability) study process.
- With proper prompts, LLMs can process P&IDs and process data to identify nodes, deviations, consequences, and safeguards.

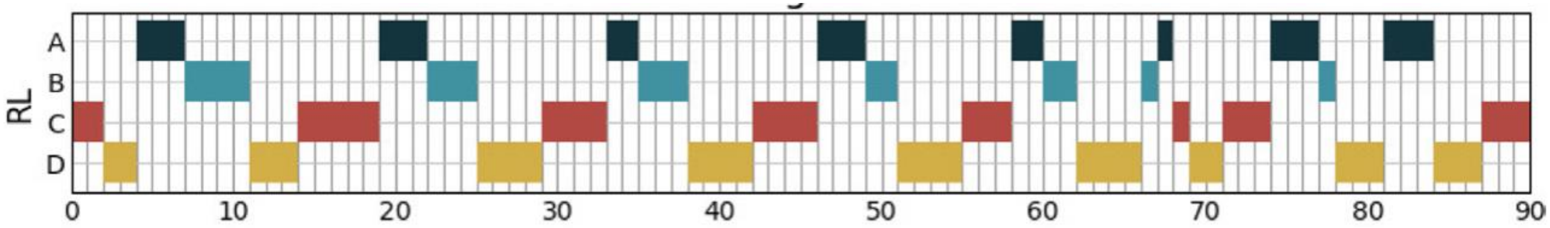


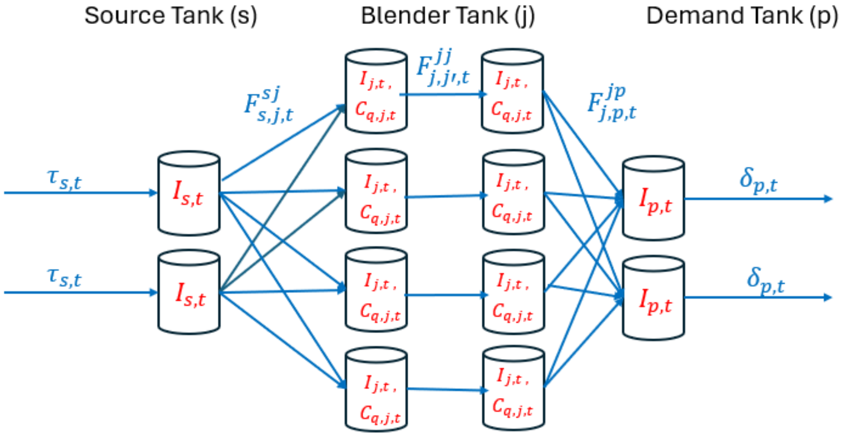
8:30-9:00	Ed Marszal (Kenexis): Converting P&IDs into an AI Ready Format
9:00-9:30	S. Nagavelu, A. Alagusundaram, L. Jiang, A. Duan, N. Taskiran, and V. Venkatasubramanian (Columbia): Ontology-Based HAZOP Automation with Domain-Informed Hybrid AI Model

Scheduling of Chemical Production

- AI can automatically make scheduling decisions—such as which product to produce and for how long—in real time.
- It learns through trial and error, gradually improving its decisions even under changing or uncertain conditions, like fluctuating demand.

Source: Hubbs et al. 2020



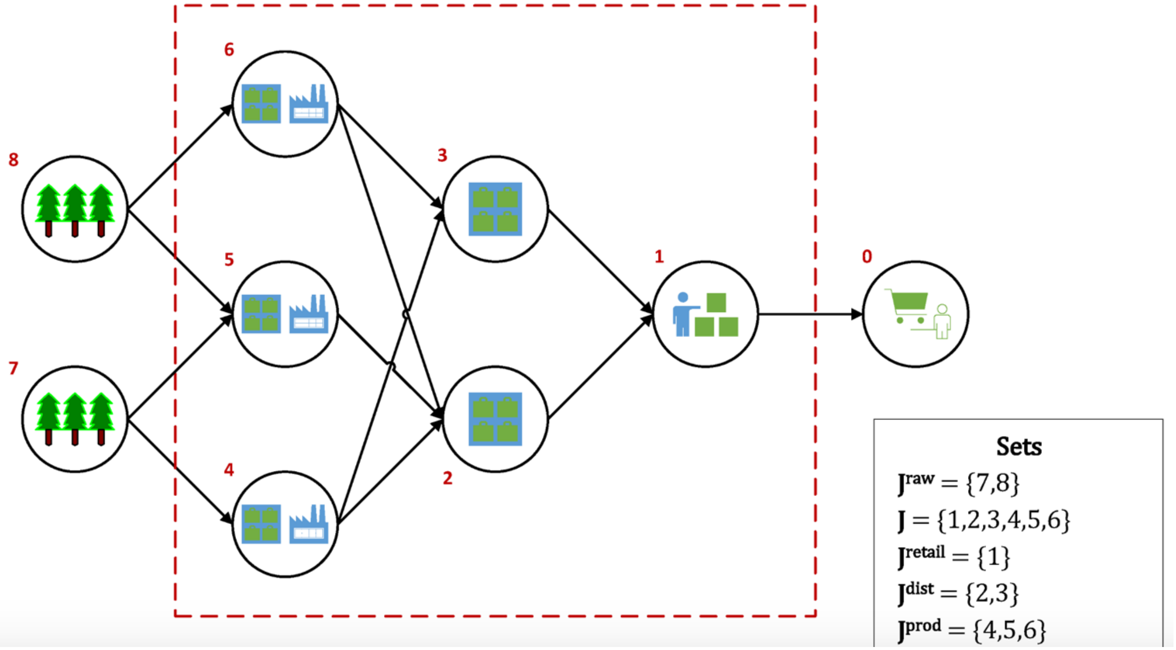


Our previous use reinforcement learning to schedule a reactor owned by Dow.

Ongoing work with ExxonMobil using AI for oil blending

Supply Chain Inventory Management

- Inventory management (IM) involves deciding how much to order, when to order, and from whom to keep products flowing efficiently.
- Key challenges include uncertain demand, variable lead times, and limited production or storage capacity.
- AI, especially reinforcement learning, can learn from experience to make smarter inventory decisions over time.
- AI can adapt to changes in demand or supply chain disruptions and find balanced policies that trade off profit, service level, and inventory cost.



Limitations and Future Directions

Limitations of Deep Neural Networks

- **Data efficiency**
  - Works well in a large data regime (image, text, audio).
  - Scientific and engineering data is scarce and expensive to obtain.

UCF101 200 k

City-scape 25 k

ImageNet 14 M

Common Voice 145 k

WikiText 100 k

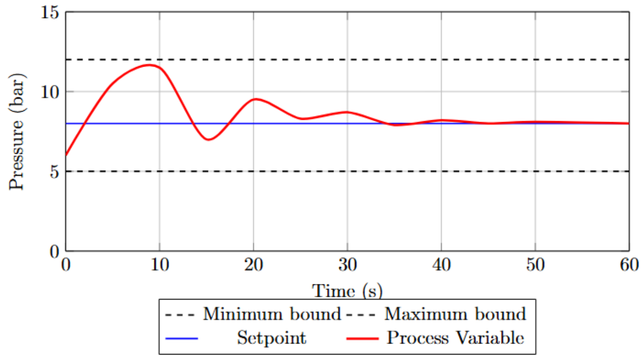
VGGround 200 k

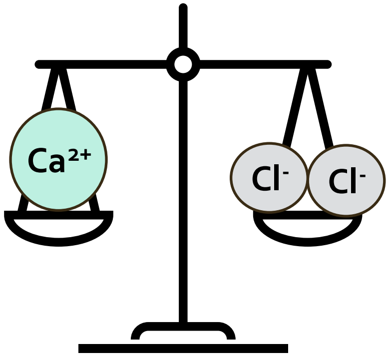
SQuAD 100 k

Generated with ChatGPT

Limitations of Deep Neural Networks

- **Data efficiency**
  - Works well in a large data regime (image, text, audio).
  - Scientific and engineering data is scarce and expensive to obtain.
- **Hard constraints**
  - No hard rules in language or image texts.
  - Hard constraints in engineering domains, e.g.: conservation laws, operational bounds.





Embedding Constraints in Neural Networks

- Improve performance in low-data regimes.
- Respect physical constraints.
- Incorporates prior scientific knowledge in data-driven predictions.

ODE

$$C_p \frac{dT}{dt} = \dot{Q}_{in} - \dot{Q}_{out} + r(T)\Delta\mathrm{H}$$

Algebraic

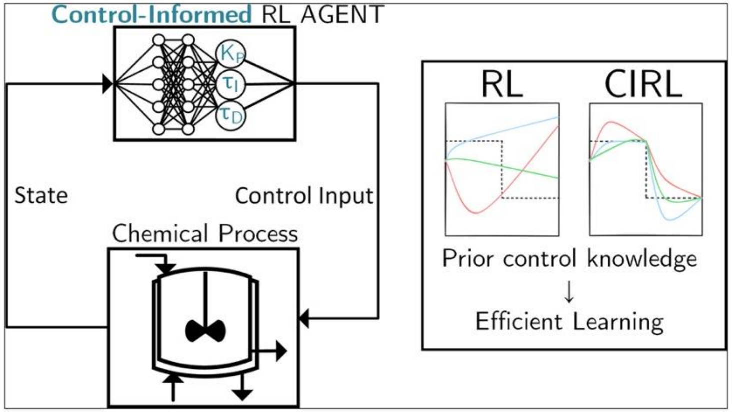
$$\dot{m}_{in} = \dot{m}_{out}$$

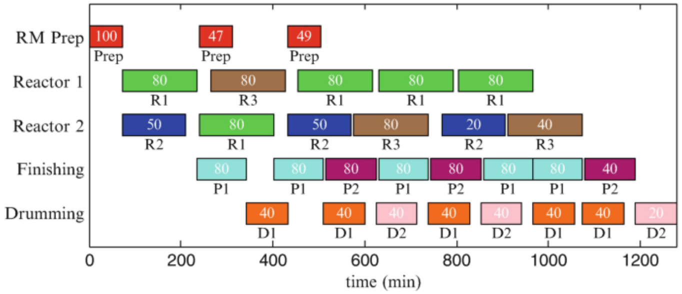
Logical

$$\text{If } s = \Sigma_i w_i g_i \geq \alpha \Rightarrow p_c \geq \beta$$

Limitations of Existing AI Methods

- **Interpretability**
  - Difficult to interpret. Lack of trust.
  - Why the model is predicting this?
- **Privacy**
  - Chemical engineering data are typically sensitive and confidential.
  - Equipment health status, drug properties, plant operating schedules





Satisfy Physical Constraints and Data Efficiency

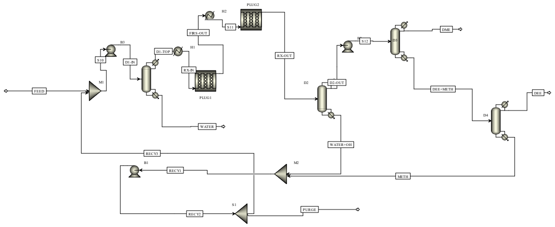
Outline of Proposed Methods

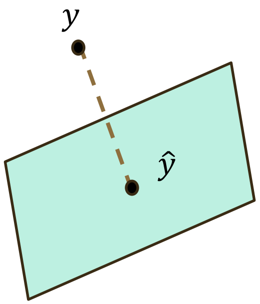
Constraint type **Application** **Geometric Intuition**

Linear equality

$$Ax + By = b$$

Mass energy balance



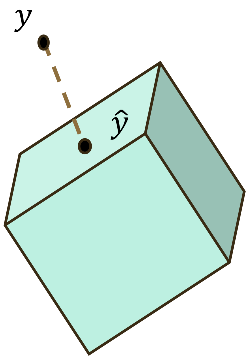


Input-dependent linear inequality

$$G(x)y = g(x)$$

$$H(x)y \le h(x)$$

Optimal power flow



Logic constraints

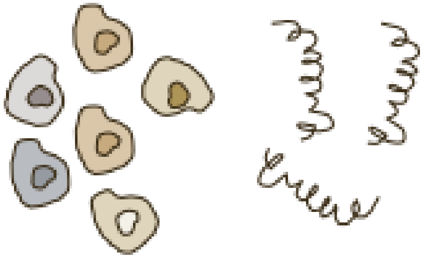
$$H_i(x)y \le h_i(x)$$

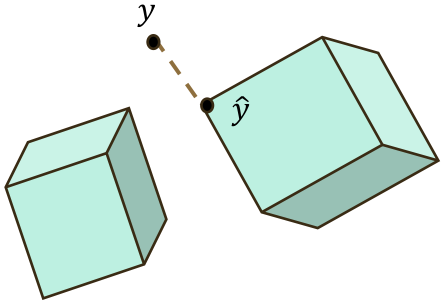
OR

$$H_i'(x)y \le h_i'(x)$$

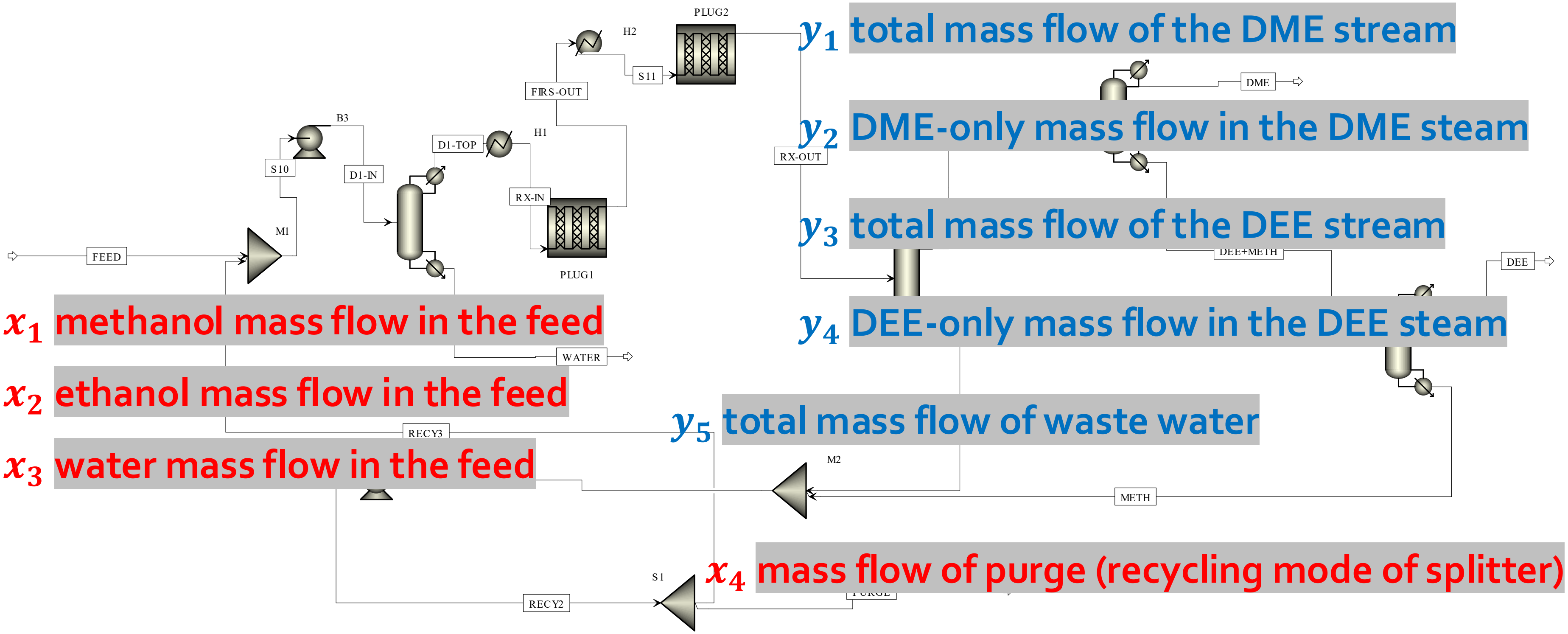
$$\forall i = 1, \dots, k$$

Cell annotation





Satisfy Mass Energy Blanace



y_1 total mass flow of the DME stream

y_2 DME-only mass flow in the DME steam

y_3 total mass flow of the DEE stream

y_4 DEE-only mass flow in the DEE steam

y_5 total mass flow of waste water

x_1 methanol mass flow in the feed

x_2 ethanol mass flow in the feed

x_3 water mass flow in the feed

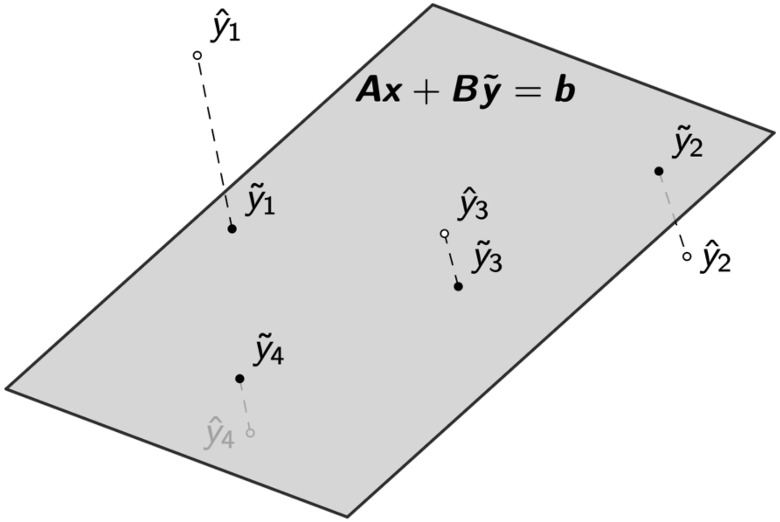
x_4 mass flow of purge (recycling mode of splitter)

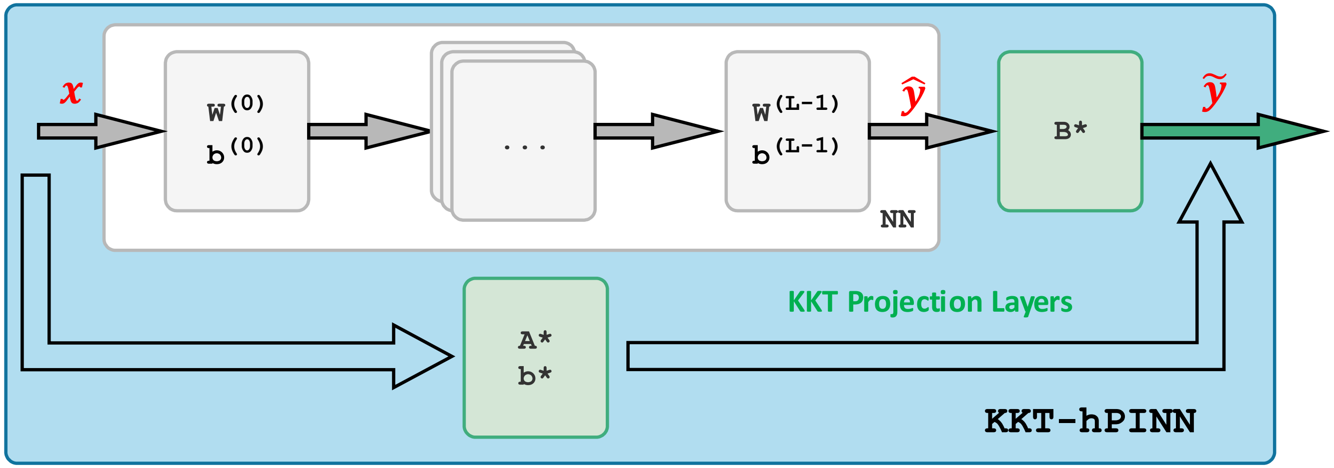
a chemical plant that uses methanol, ethanol, and water to manufacture dimethyl ether (DME) and diethyl ether (DEE)

- Surrogate model that predicts the outflow, given the inflow and recycling model
- Hard linear constraint embedded into KKT-hPINN: mass balance (flow-in = flow-out)

Orthogonal Projection

Geometric sense: **Orthogonal projection** of the old prediction, $\widehat{\boldsymbol{y}}$, onto a hyperplane where $\mathbf{Ax} + \mathbf{B}\widetilde{\boldsymbol{y}} = \boldsymbol{b}$ (linear equality constraints of interest are satisfied)





x

$W^{(0)}$
$b^{(0)}$

. . .

$W^{(L-1)}$
$b^{(L-1)}$

$\widehat{y}$

NN

B*

$\widetilde{y}$

A*
b*

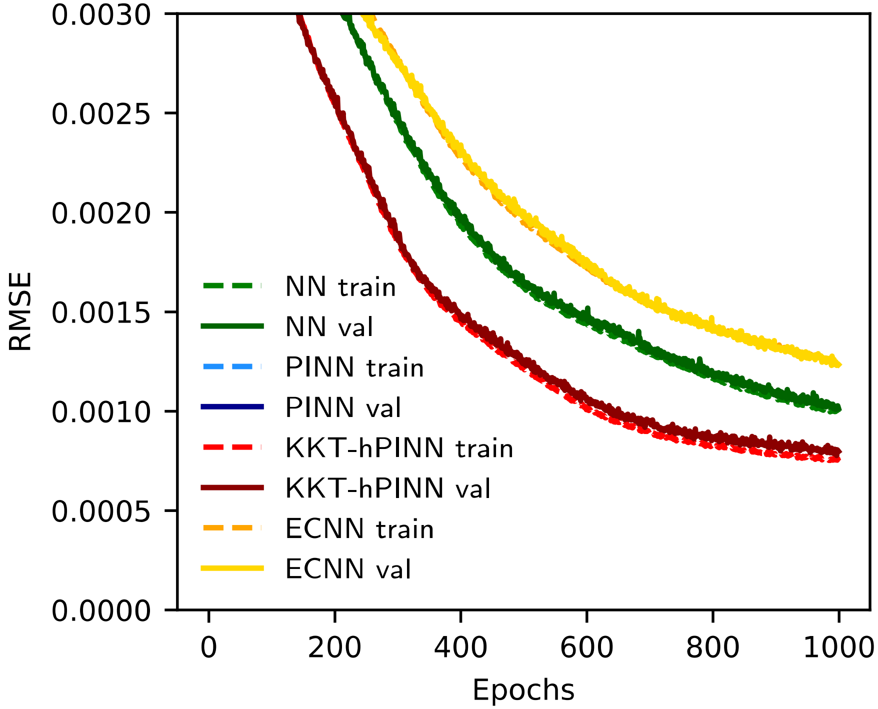
KKT Projection Layers

KKT-hPINN

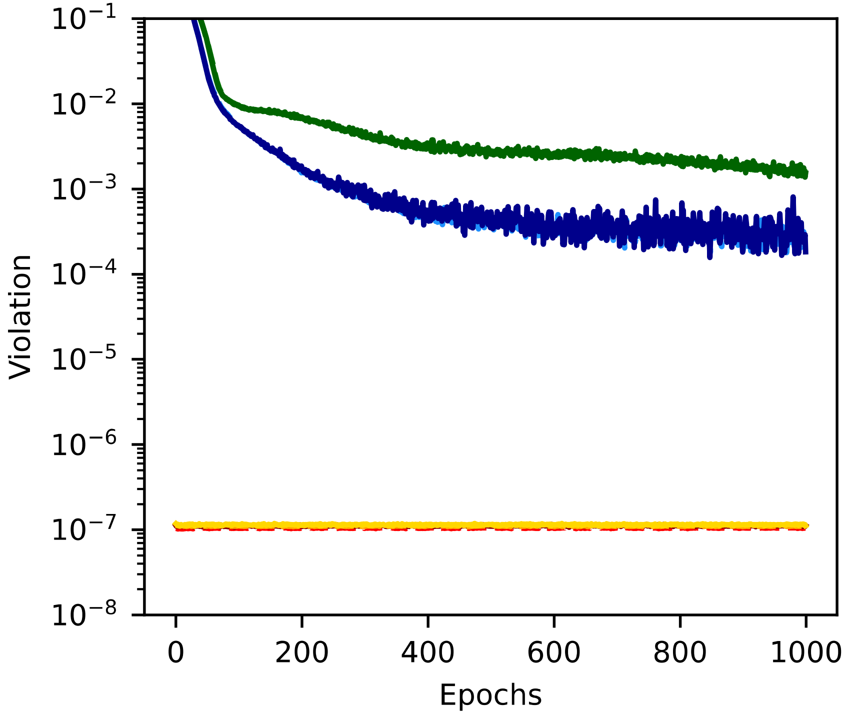
Model Performance: Learning Curves

- ✓ **Higher accuracy:** outperforms neural network with a significant margin
- ✓ **Generalizability:** no overfitting and can be generalized well to unseen dataset
- ✓ **Inviolable hard constraints:** strictly satisfies the mass balance constraint

RMSE: root mean square error
between prediction and ground truth

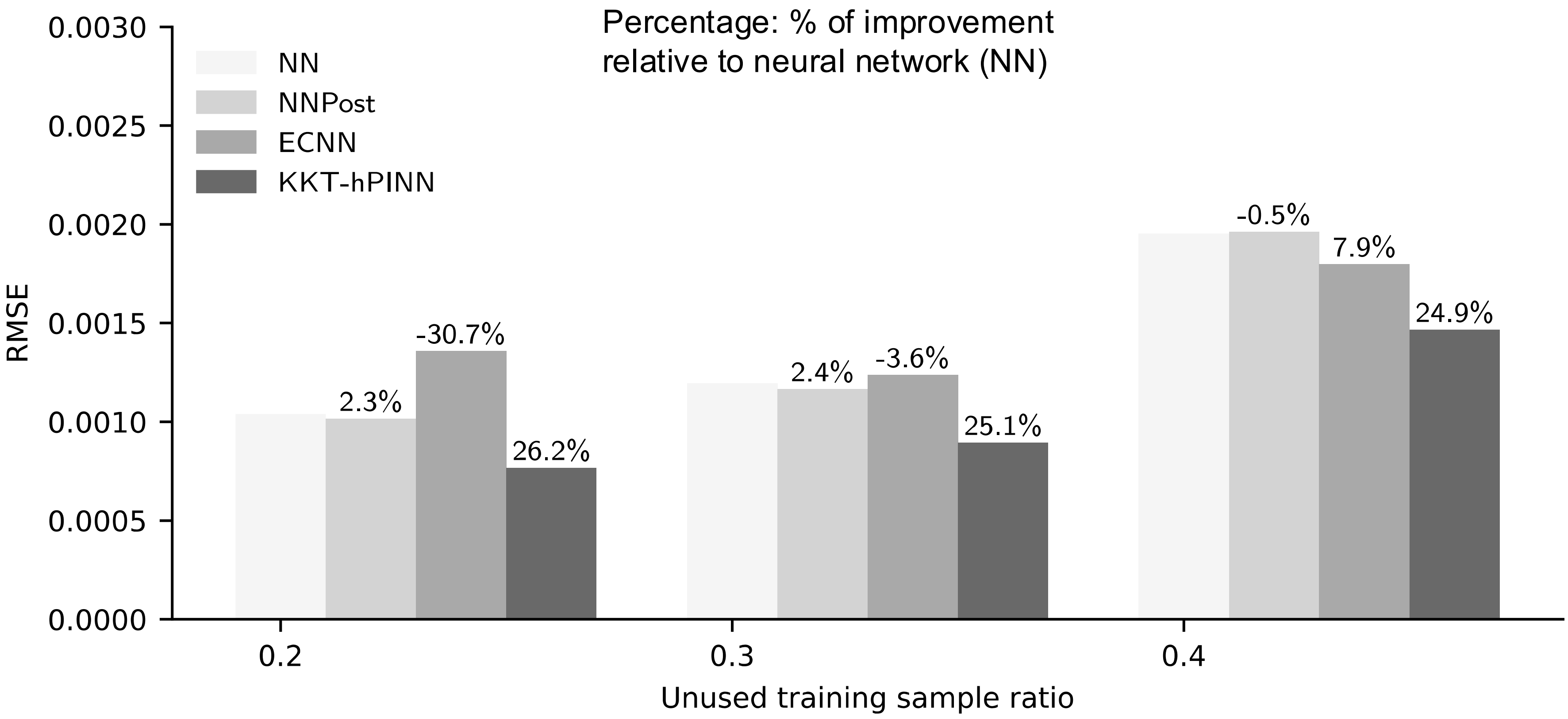


extent of mass balance violation
$|\text{inflow} - \text{outflow}|$



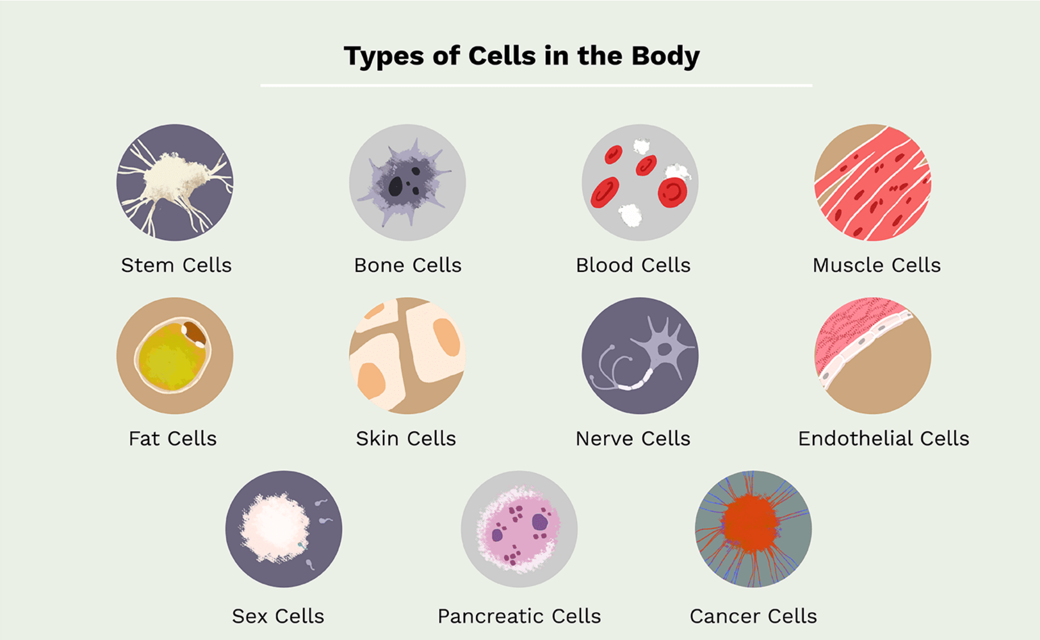
Model Performance vs. Training Samples

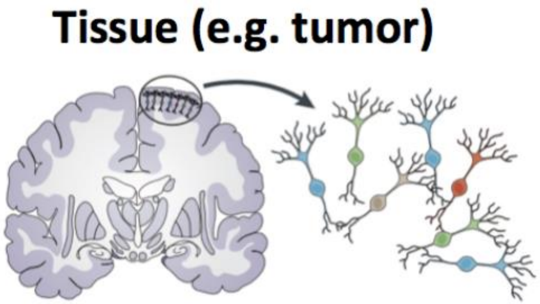
✓ **Minimal data requirement:** relatively good performance with less data



Cell Classification

- **Motivating application:** inspect the concentration of cancer cells before and after treatment.

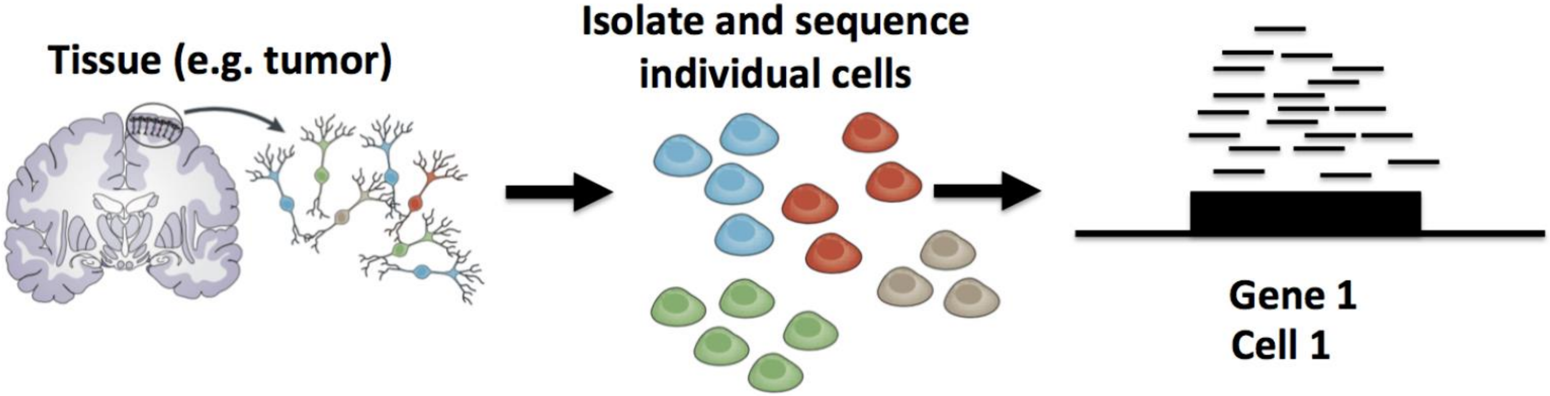




https://www.thoughtco.com/types-of-cells-in-the-body-373388
https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

Single Cell RNA Sequencing

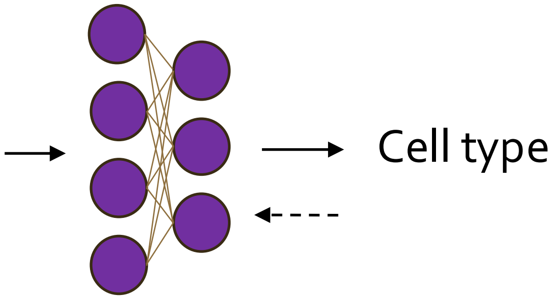
- scRNA-seq: experimental methods to measure the number of messenger RNA in every single cell.
- Gene expression is inferred by measuring the number of messenger RNA molecules corresponding to each gene.



Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

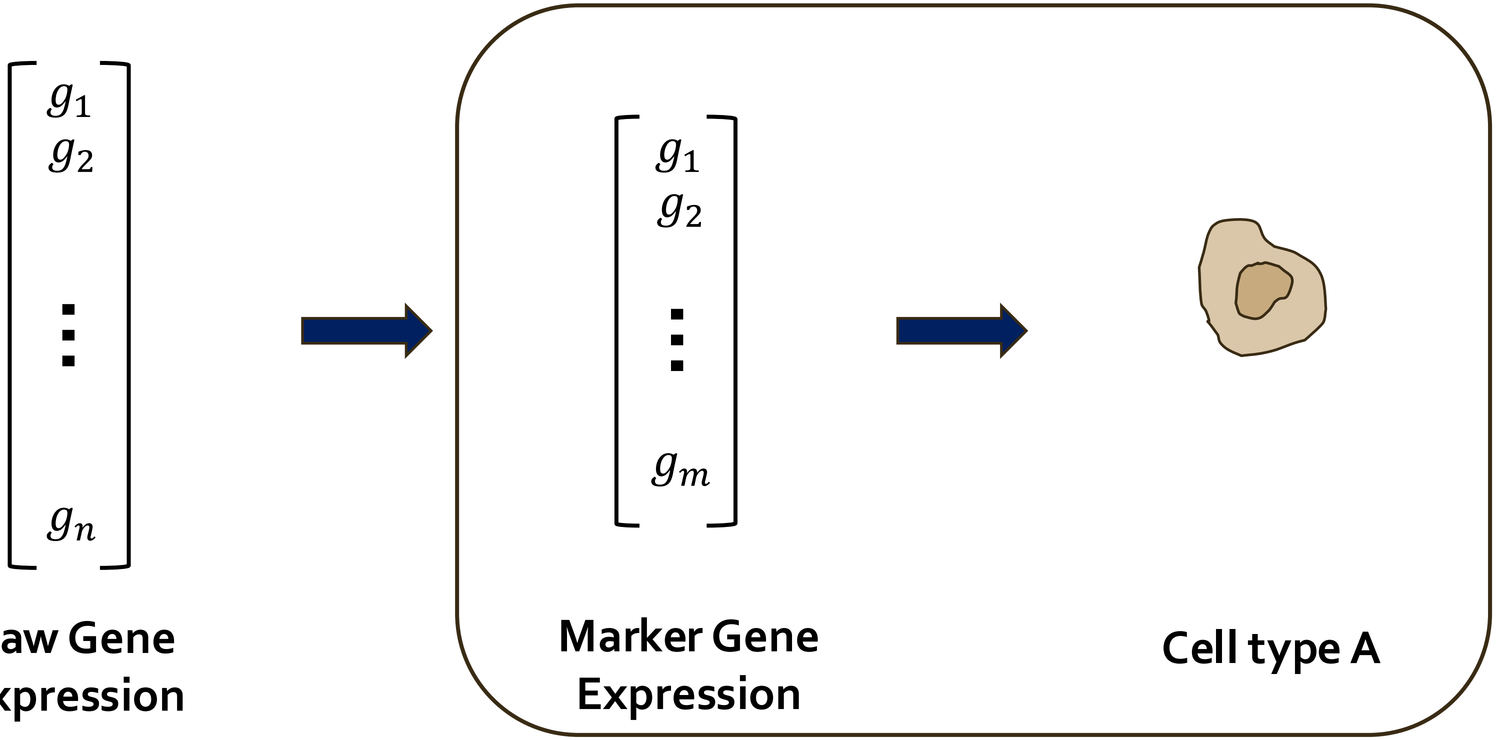
$$\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$$



https://learn.gencore.bio.nyu.edu/single-cell-rnaseq/

Cell Annotation

- Data are expensive to collect: each sample costs around $3,000.
- Prior knowledge: marker genes are the highly expressed genes for a given cell type



$$\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix} \rightarrow \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix} \rightarrow$$

Raw Gene Expression

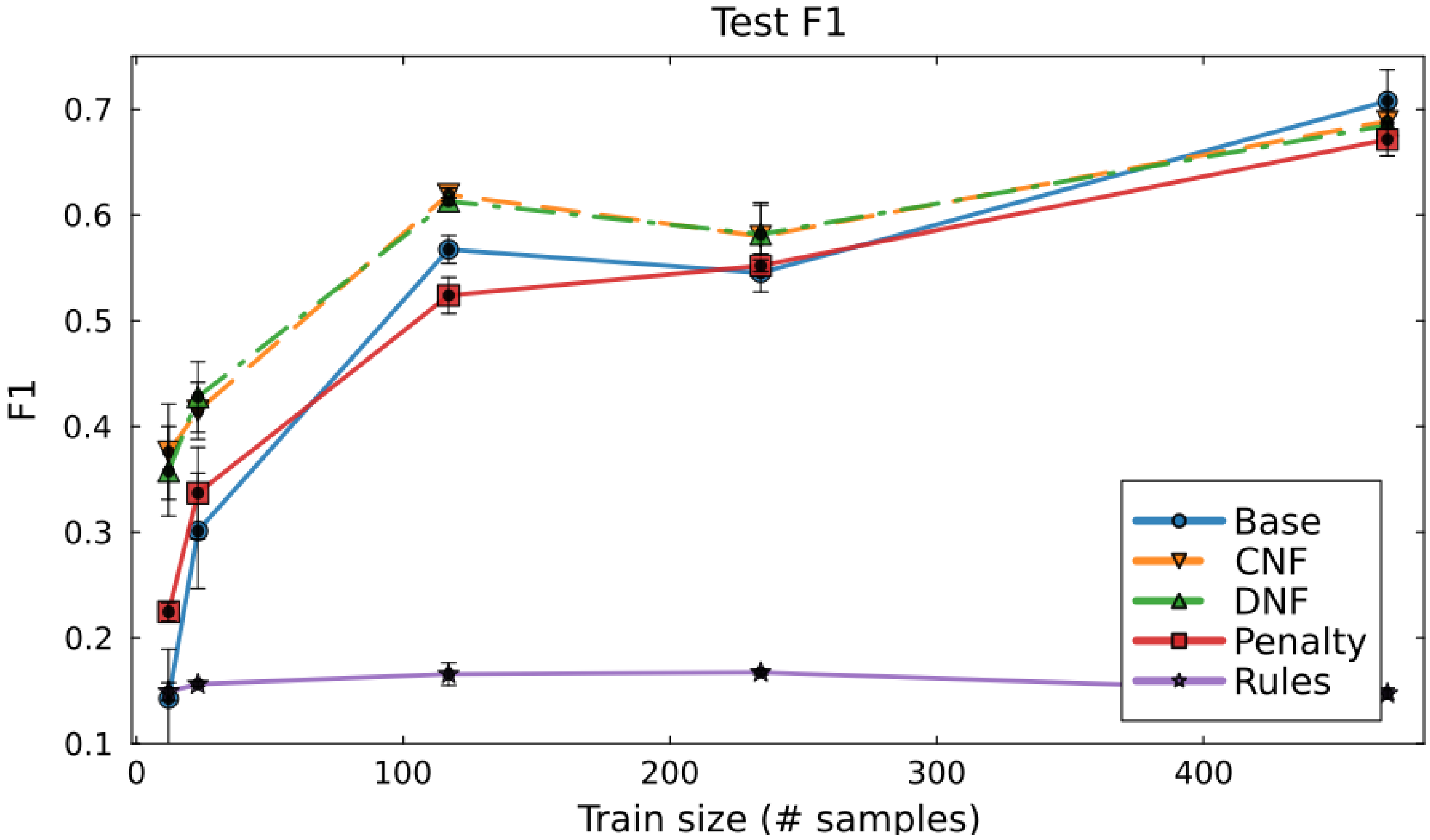
Marker Gene Expression

Cell type A

$$G_r(u) \geq \tau_r \;\Rightarrow\; \bigvee_{c \in \mathcal{C}_r} y_c \geq \rho_r,$$

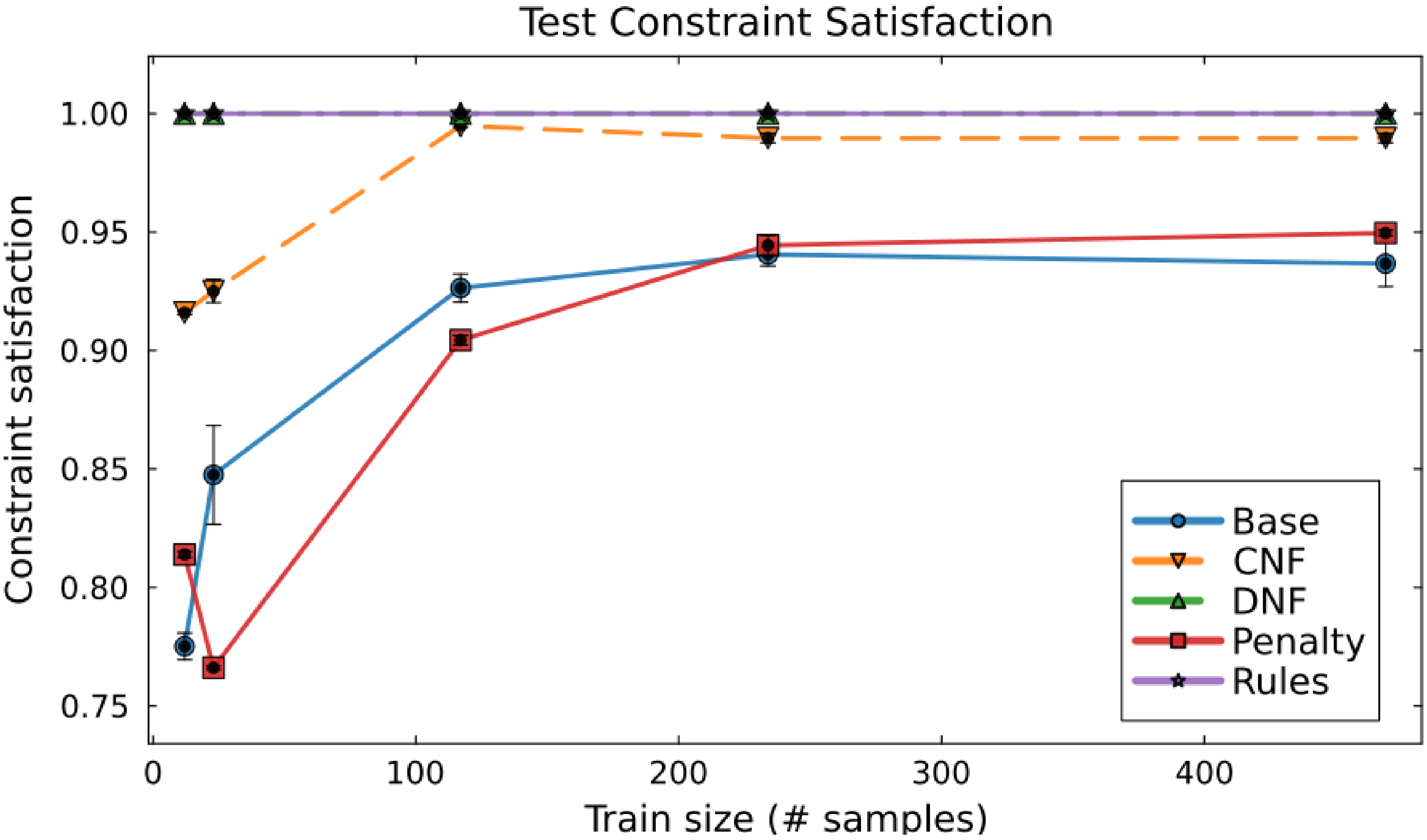
Test Performance

- Enforcing rules via projection improves performance in low-data regimes.
- As the dataset size increases, the unconstrained model also improves.



Constraint Satisfaction

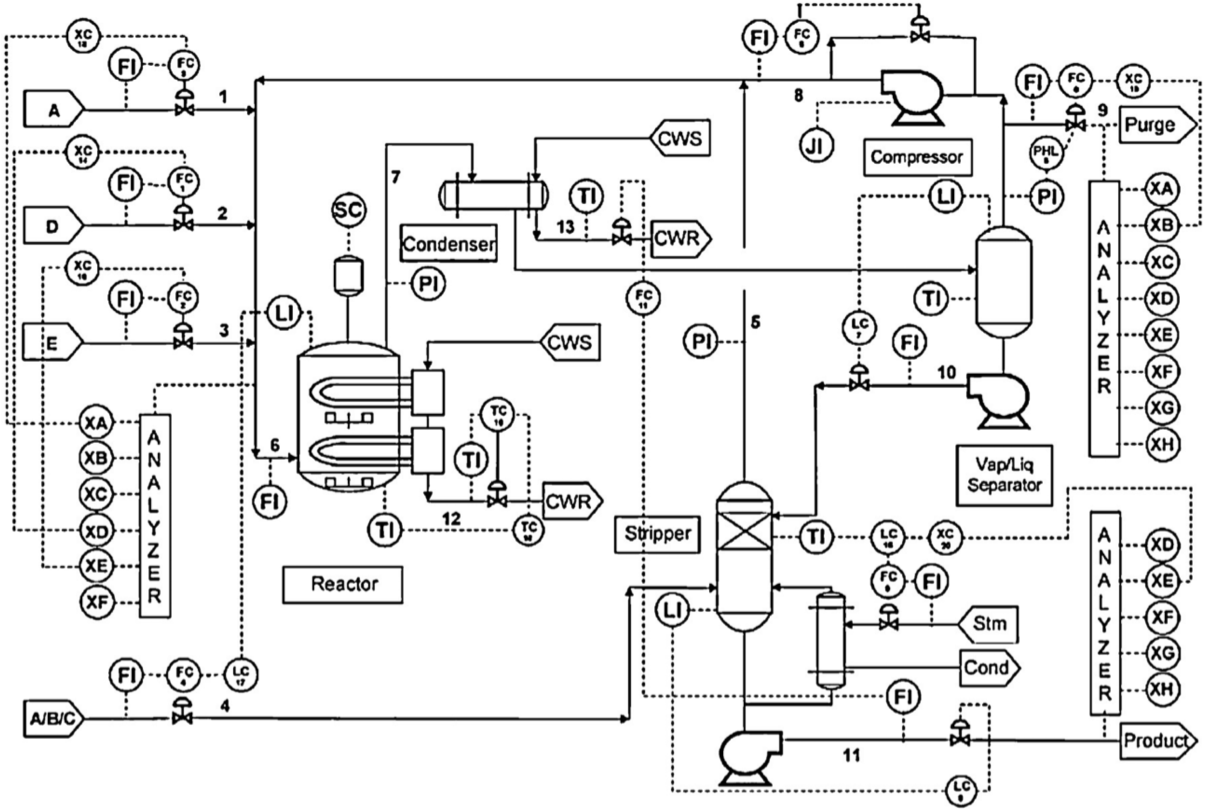
- Tighter DNF-based projections can trade a small amount of F1 for stronger rule adherence.
- The unconstrained model can gradually learn the rules as the dataset size increases.



Improve Interpretability

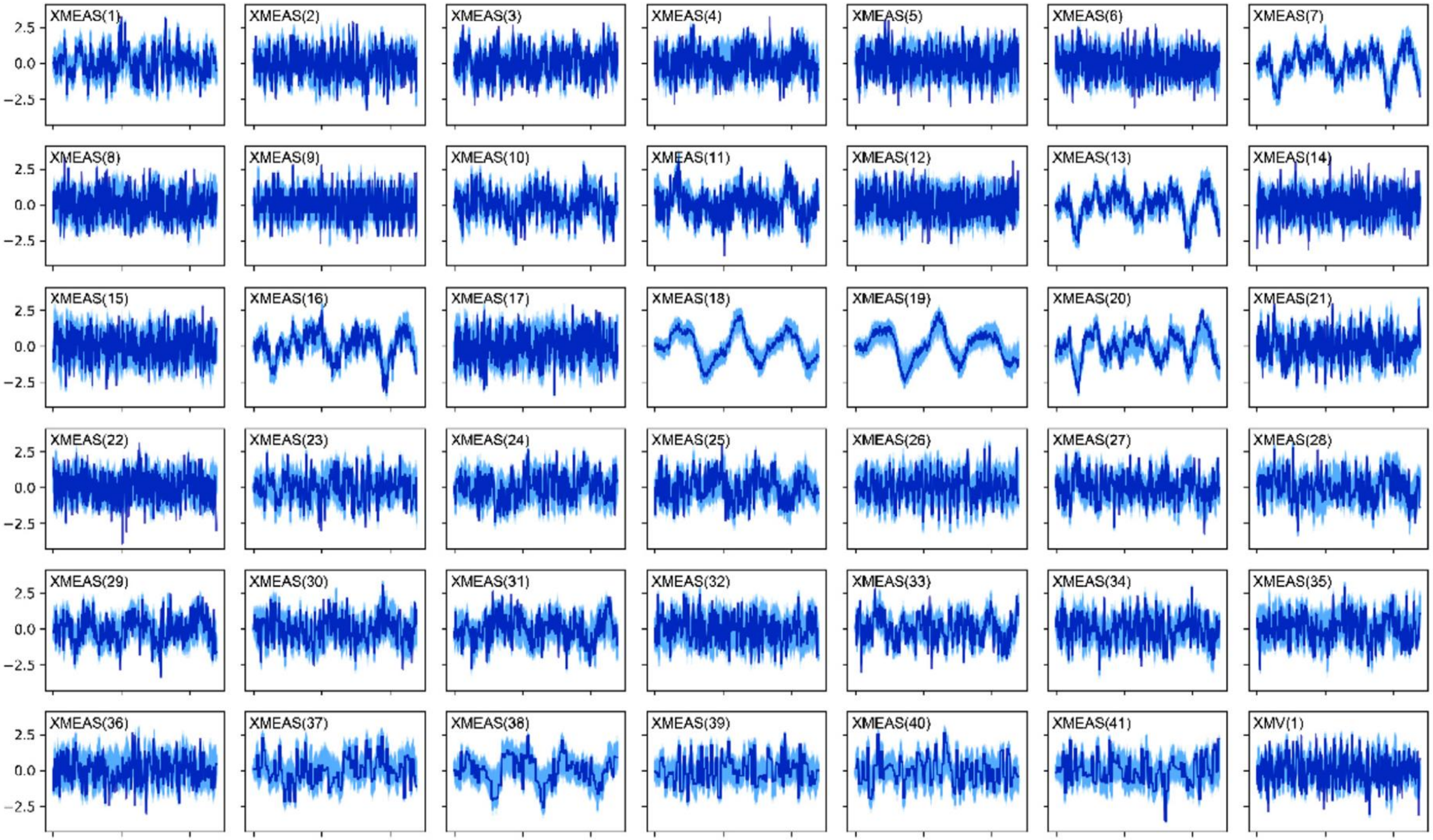
Tennessee Eastman Process

- TEP is an open-source simulator written in Fortran that resembles a real chemical process by Eastman
- Time series data can be collected from over 40 sensors that measure the state variables.
- Task: From measured state variables, perform fault detection using ML/AI



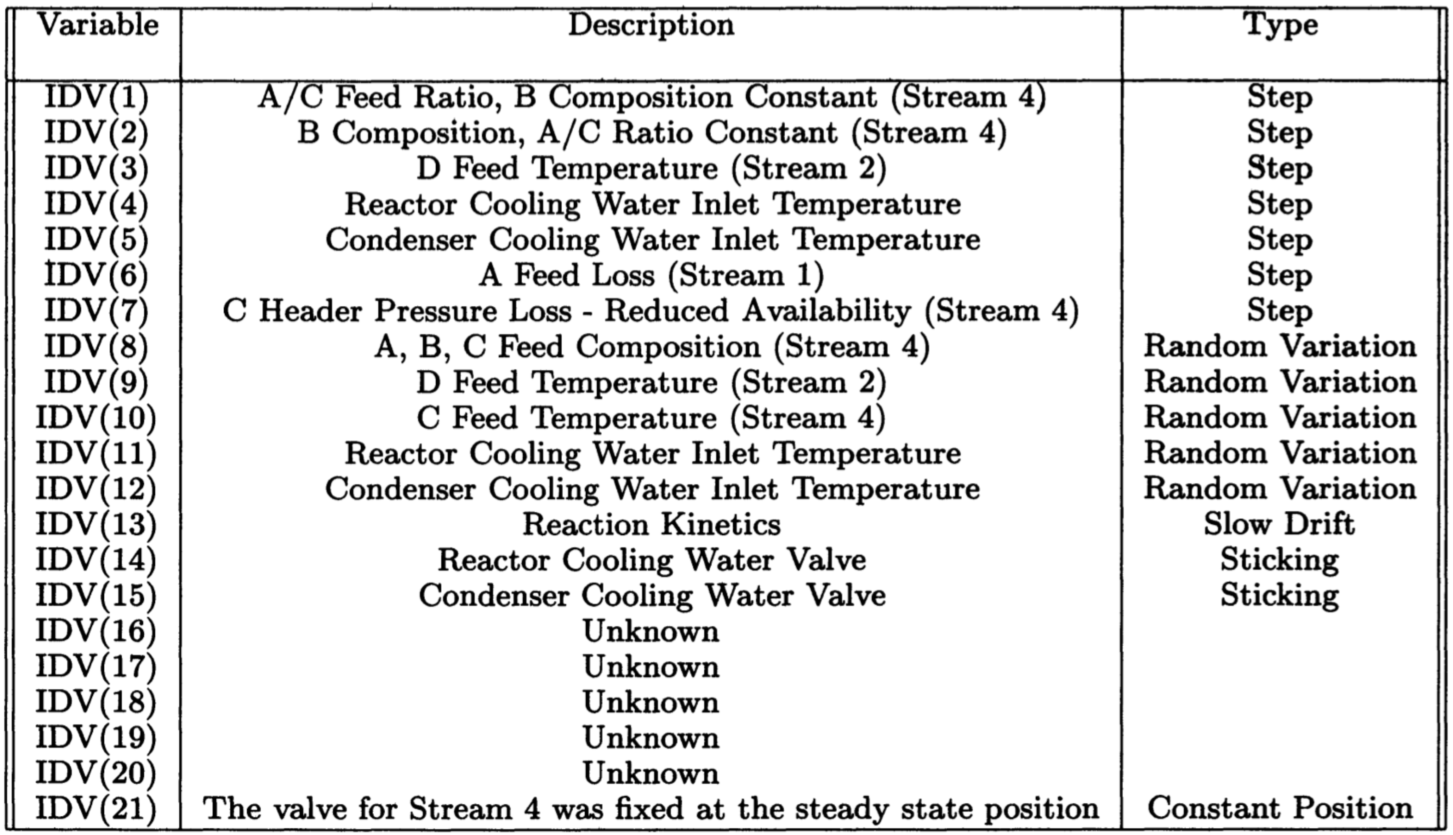
Examples of State Variables with Sensor Data

- Examples include feed flow rates, temperatures, pressures



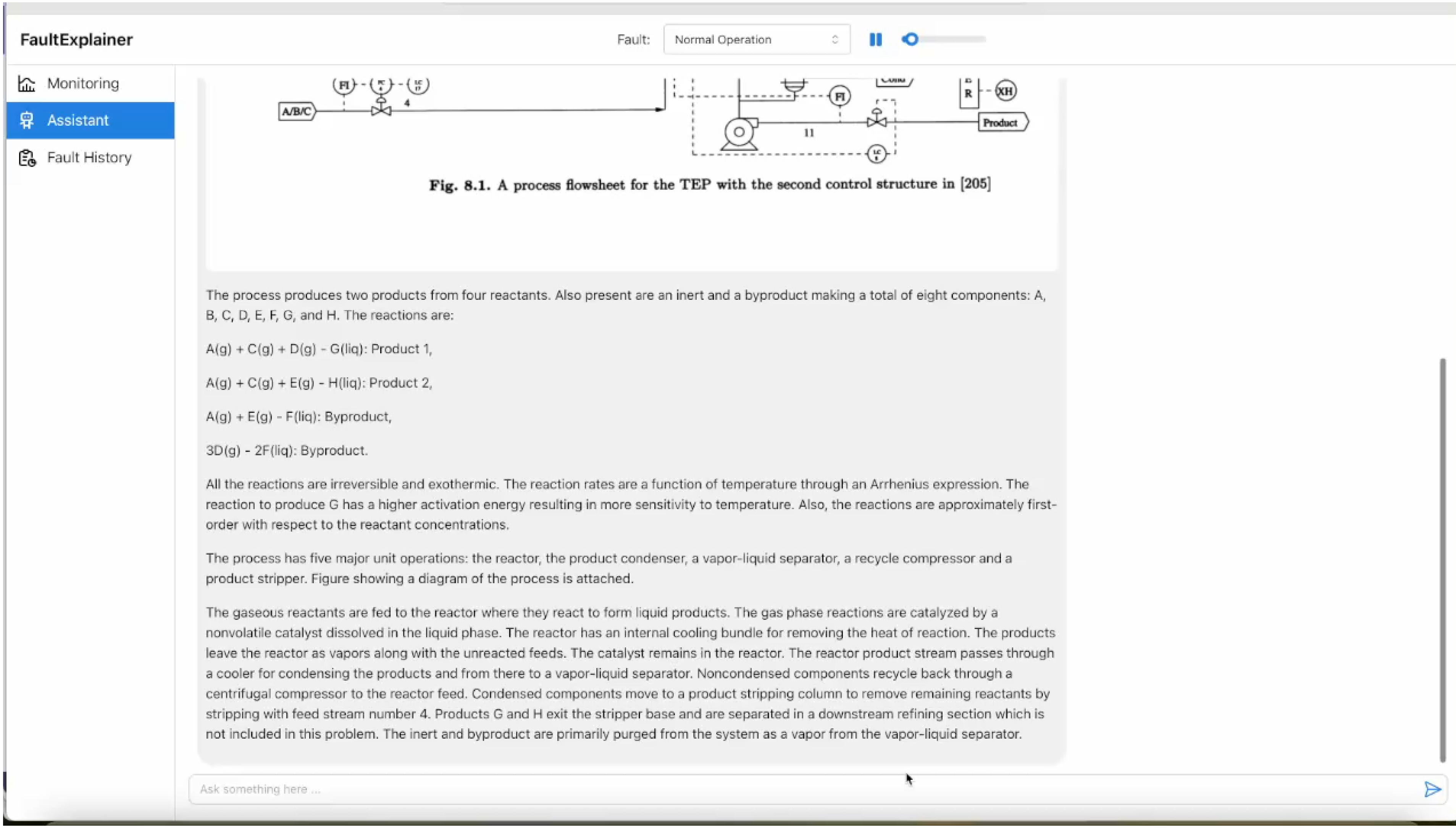
List of Potential Faults

- The following “faults” are created synthetically by the simulator
- These faults will cause the measured state variables to change from their normal operating conditions which further cause safety hazards.



Variable	Description	Type
IDV(1)	A/C Feed Ratio, B Composition Constant (Stream 4)	Step
IDV(2)	B Composition, A/C Ratio Constant (Stream 4)	Step
IDV(3)	D Feed Temperature (Stream 2)	Step
IDV(4)	Reactor Cooling Water Inlet Temperature	Step
IDV(5)	Condenser Cooling Water Inlet Temperature	Step
IDV(6)	A Feed Loss (Stream 1)	Step
IDV(7)	C Header Pressure Loss - Reduced Availability (Stream 4)	Step
IDV(8)	A, B, C Feed Composition (Stream 4)	Random Variation
IDV(9)	D Feed Temperature (Stream 2)	Random Variation
IDV(10)	C Feed Temperature (Stream 4)	Random Variation
IDV(11)	Reactor Cooling Water Inlet Temperature	Random Variation
IDV(12)	Condenser Cooling Water Inlet Temperature	Random Variation
IDV(13)	Reaction Kinetics	Slow Drift
IDV(14)	Reactor Cooling Water Valve	Sticking
IDV(15)	Condenser Cooling Water Valve	Sticking
IDV(16)	Unknown	
IDV(17)	Unknown	
IDV(18)	Unknown	
IDV(19)	Unknown	
IDV(20)	Unknown	
IDV(21)	The valve for Stream 4 was fixed at the steady state position	Constant Position

FaultExplainer Demo



Fault: Normal Operation

Fig. 8.1. A process flowsheet for the TEP with the second control structure in [205]

The process produces two products from four reactants. Also present are an inert and a byproduct making a total of eight components: A, B, C, D, E, F, G, and H. The reactions are:

A(g) + C(g) + D(g) - G(liq): Product 1,

A(g) + C(g) + E(g) - H(liq): Product 2,

A(g) + E(g) - F(liq): Byproduct,

3D(g) - 2F(liq): Byproduct.

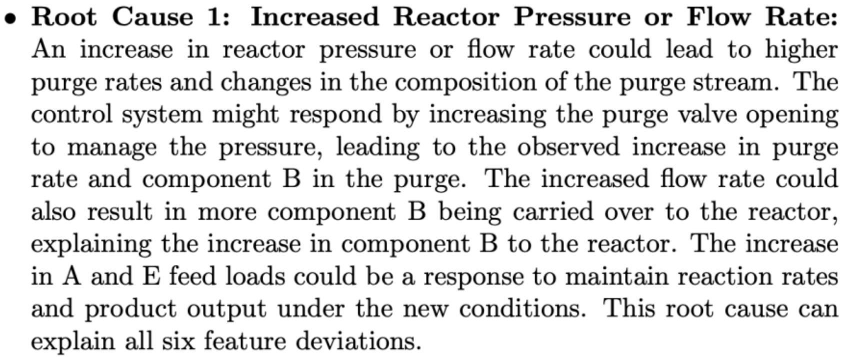
All the reactions are irreversible and exothermic. The reaction rates are a function of temperature through an Arrhenius expression. The reaction to produce G has a higher activation energy resulting in more sensitivity to temperature. Also, the reactions are approximately first-order with respect to the reactant concentrations.

The process has five major unit operations: the reactor, the product condenser, a vapor-liquid separator, a recycle compressor and a product stripper. Figure showing a diagram of the process is attached.

The gaseous reactants are fed to the reactor where they react to form liquid products. The gas phase reactions are catalyzed by a nonvolatile catalyst dissolved in the liquid phase. The reactor has an internal cooling bundle for removing the heat of reaction. The products leave the reactor as vapors along with the unreacted feeds. The catalyst remains in the reactor. The reactor product stream passes through a cooler for condensing the products and from there to a vapor-liquid separator. Noncondensed components recycle back through a centrifugal compressor to the reactor feed. Condensed components move to a product stripping column to remove remaining reactants by stripping with feed stream number 4. Products G and H exit the stripper base and are separated in a downstream refining section which is not included in this problem. The inert and byproduct are primarily purged from the system as a vapor from the vapor-liquid separator.

Qualitative Analysis of GPT-4o and o1 Models

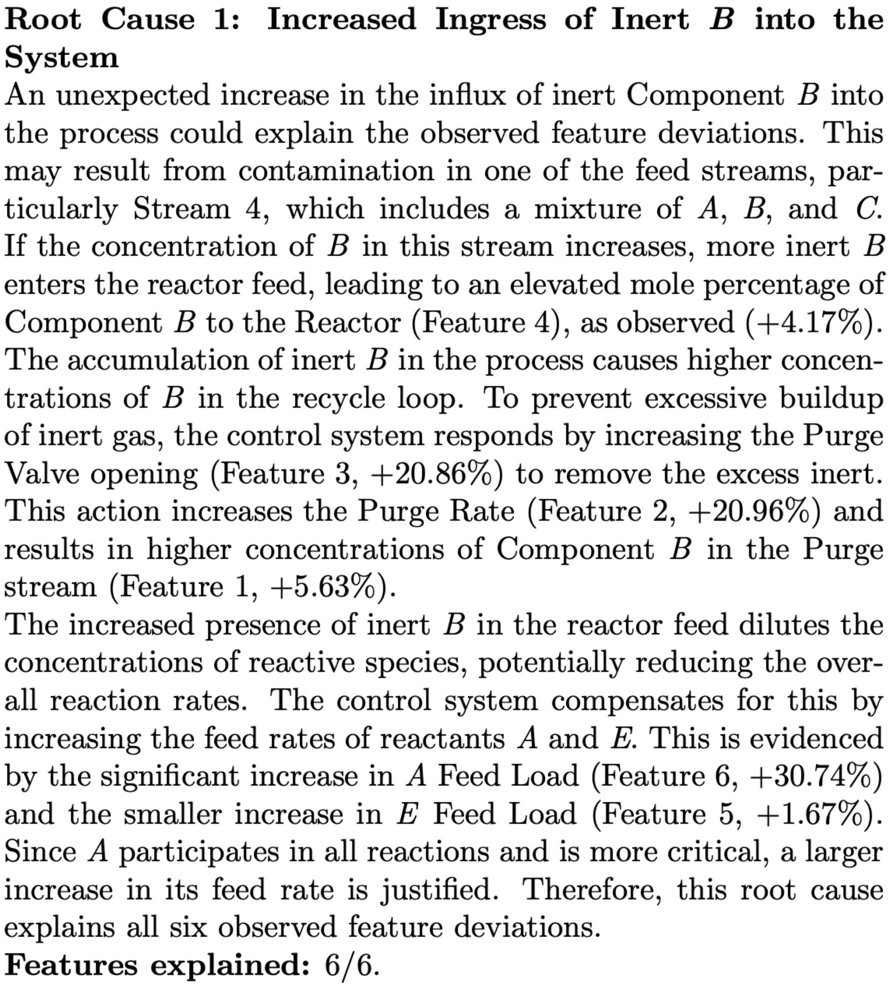
- Both models are prompted to diagnose the step change in the feed composition of inert B based on the top six changes in the sensor measurement.



- **Root Cause 1: Increased Reactor Pressure or Flow Rate:** An increase in reactor pressure or flow rate could lead to higher purge rates and changes in the composition of the purge stream. The control system might respond by increasing the purge valve opening to manage the pressure, leading to the observed increase in purge rate and component B in the purge. The increased flow rate could also result in more component B being carried over to the reactor, explaining the increase in component B to the reactor. The increase in A and E feed loads could be a response to maintain reaction rates and product output under the new conditions. This root cause can explain all six feature deviations.

GPT-4o

O1 is able to provide a more logical and quantitative reasoning of the fault propagation, including how the control system reacted to the fault.



Root Cause 1: Increased Ingress of Inert B into the System

An unexpected increase in the influx of inert Component B into the process could explain the observed feature deviations. This may result from contamination in one of the feed streams, particularly Stream 4, which includes a mixture of A, B, and C. If the concentration of B in this stream increases, more inert B enters the reactor feed, leading to an elevated mole percentage of Component B to the Reactor (Feature 4), as observed (+4.17%). The accumulation of inert B in the process causes higher concentrations of B in the recycle loop. To prevent excessive buildup of inert gas, the control system responds by increasing the Purge Valve opening (Feature 3, +20.86%) to remove the excess inert. This action increases the Purge Rate (Feature 2, +20.96%) and results in higher concentrations of Component B in the Purge stream (Feature 1, +5.63%).

The increased presence of inert B in the reactor feed dilutes the concentrations of reactive species, potentially reducing the overall reaction rates. The control system compensates for this by increasing the feed rates of reactants A and E. This is evidenced by the significant increase in A Feed Load (Feature 6, +30.74%) and the smaller increase in E Feed Load (Feature 5, +1.67%). Since A participates in all reactions and is more critical, a larger increase in its feed rate is justified. Therefore, this root cause explains all six observed feature deviations.

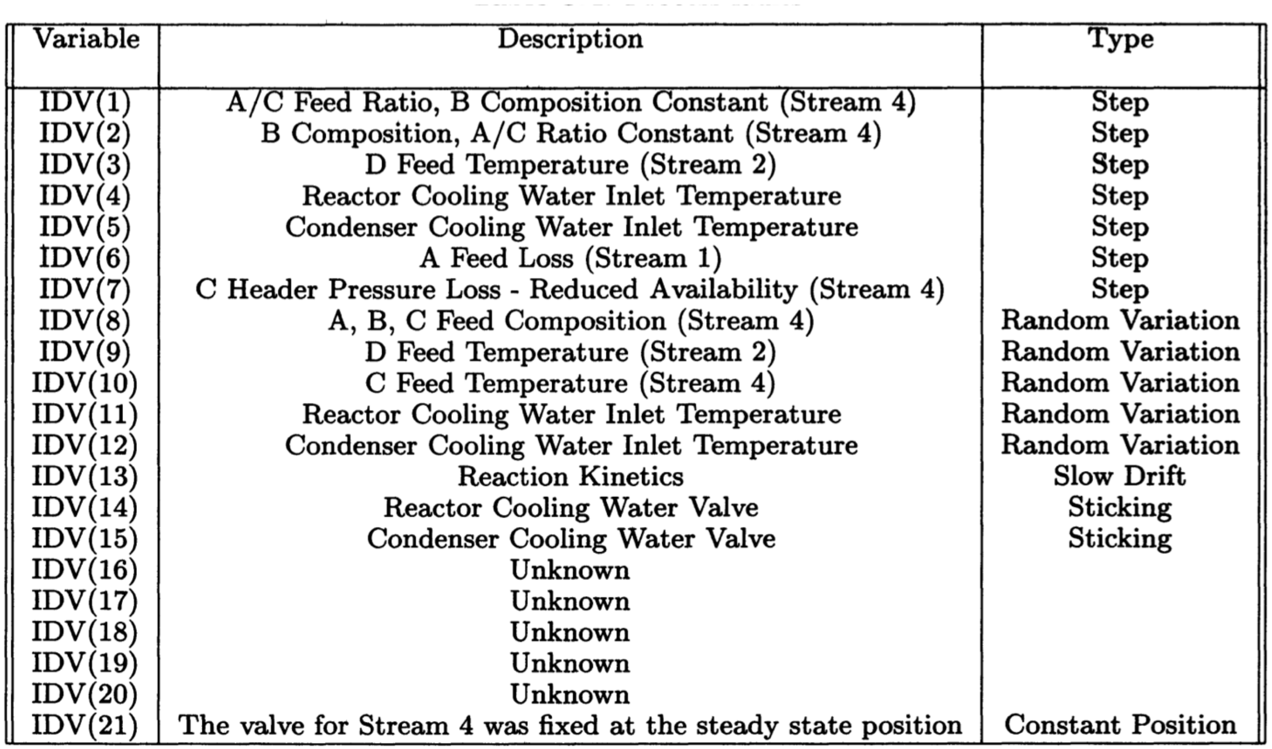
Features explained: 6/6.

O1

Quantitative Results

- We prompt the LLMs to give top 3 possible causes of the fault based on the top 6 contributing features
- PCA can identify 11/15 of the faults
- The success rate of identifying the root cause of the fault

GPT-4o: 6/11 o1: 9/11

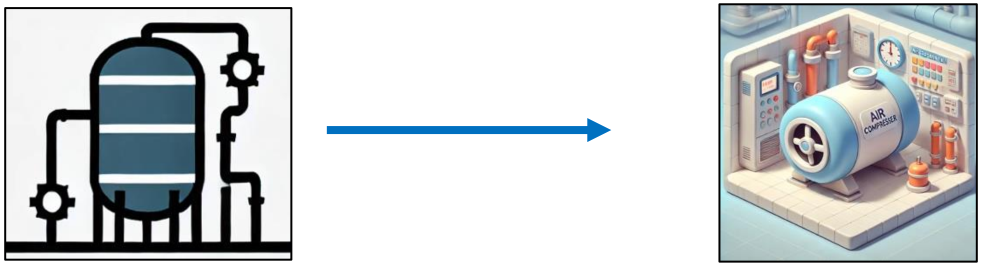


Variable	Description	Type
IDV(1)	A/C Feed Ratio, B Composition Constant (Stream 4)	Step
IDV(2)	B Composition, A/C Ratio Constant (Stream 4)	Step
IDV(3)	D Feed Temperature (Stream 2)	Step
IDV(4)	Reactor Cooling Water Inlet Temperature	Step
IDV(5)	Condenser Cooling Water Inlet Temperature	Step
IDV(6)	A Feed Loss (Stream 1)	Step
IDV(7)	C Header Pressure Loss - Reduced Availability (Stream 4)	Step
IDV(8)	A, B, C Feed Composition (Stream 4)	Random Variation
IDV(9)	D Feed Temperature (Stream 2)	Random Variation
IDV(10)	C Feed Temperature (Stream 4)	Random Variation
IDV(11)	Reactor Cooling Water Inlet Temperature	Random Variation
IDV(12)	Condenser Cooling Water Inlet Temperature	Random Variation
IDV(13)	Reaction Kinetics	Slow Drift
IDV(14)	Reactor Cooling Water Valve	Sticking
IDV(15)	Condenser Cooling Water Valve	Sticking
IDV(16)	Unknown	
IDV(17)	Unknown	
IDV(18)	Unknown	
IDV(19)	Unknown	
IDV(20)	Unknown	
IDV(21)	The valve for Stream 4 was fixed at the steady state position	Constant Position

Protect Privacy

Forecasting Maintenance

- Predictive maintenance of equipment



Air separation unit

Compressor in the plant

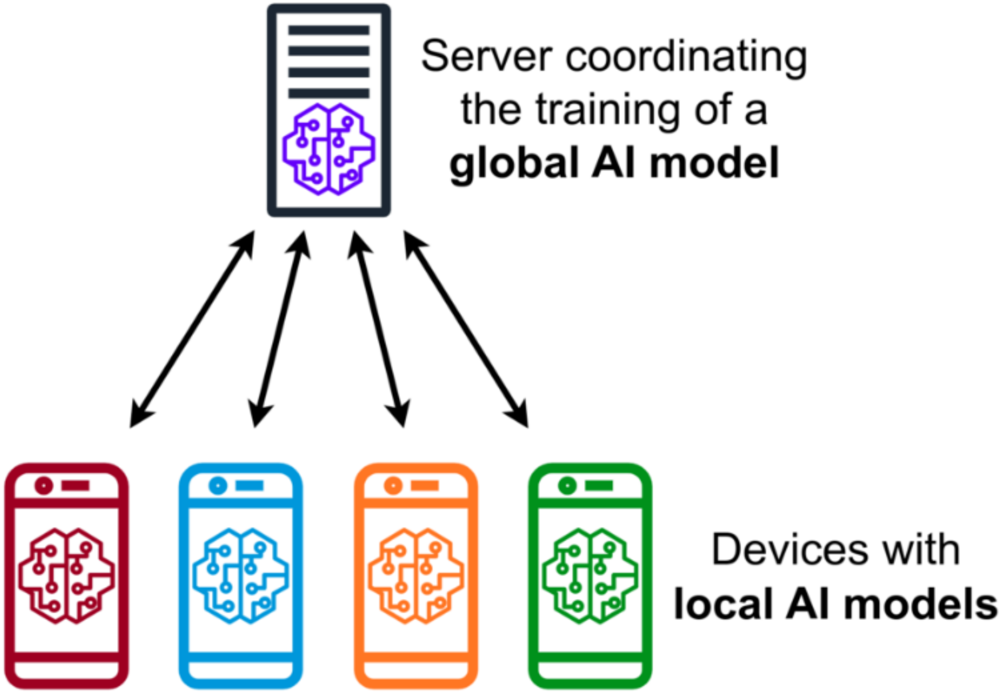
Sensor

S. No.	Historical Data Points	State of Compressor
1	….	Healthy
2	….	Healthy
3	….	About to get Faulty
4	….	Faulty
	….	….

- ○ Observed disruption in the plant!
- ○ Plant aims to forecast such disruption using ML
- ○ ***Challenge***: Insufficient data
- ○ **Solution**: Federated Learning

Federated Learning

- **Federated Learning**: Extensively used by Big-Tech companies to trainn prediction models, such as ‘next letter that you will type’



Personal devices

Differential Privacy

- **Differential Privacy**: Technique to provide guarantees that private data won't be sacrificed
- Routinely used by Big-Tech companies and for high stakes government projects

APPLE

Apple will not see your data

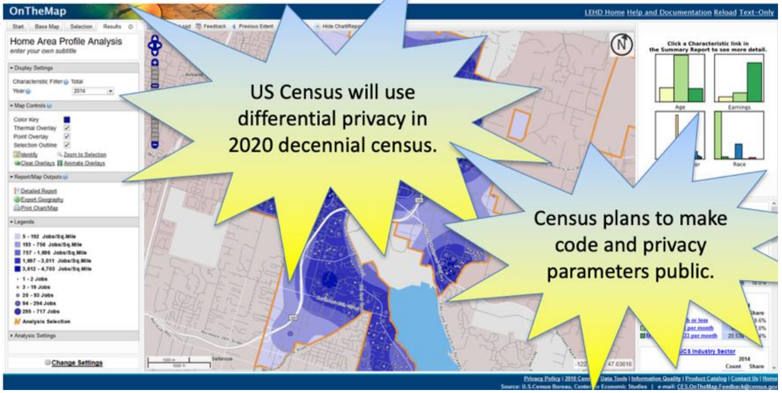
ANDY GREENBERG SECURITY 06.13.16 07:02 PM

APPLE'S 'DIFFERENTIAL PRIVACY' IS ABOUT COLLECTING YOUR DATA—BUT NOT *YOUR* DATA

*

U.S. CENSUS BUREAU

http://onthemap.ces.census.gov



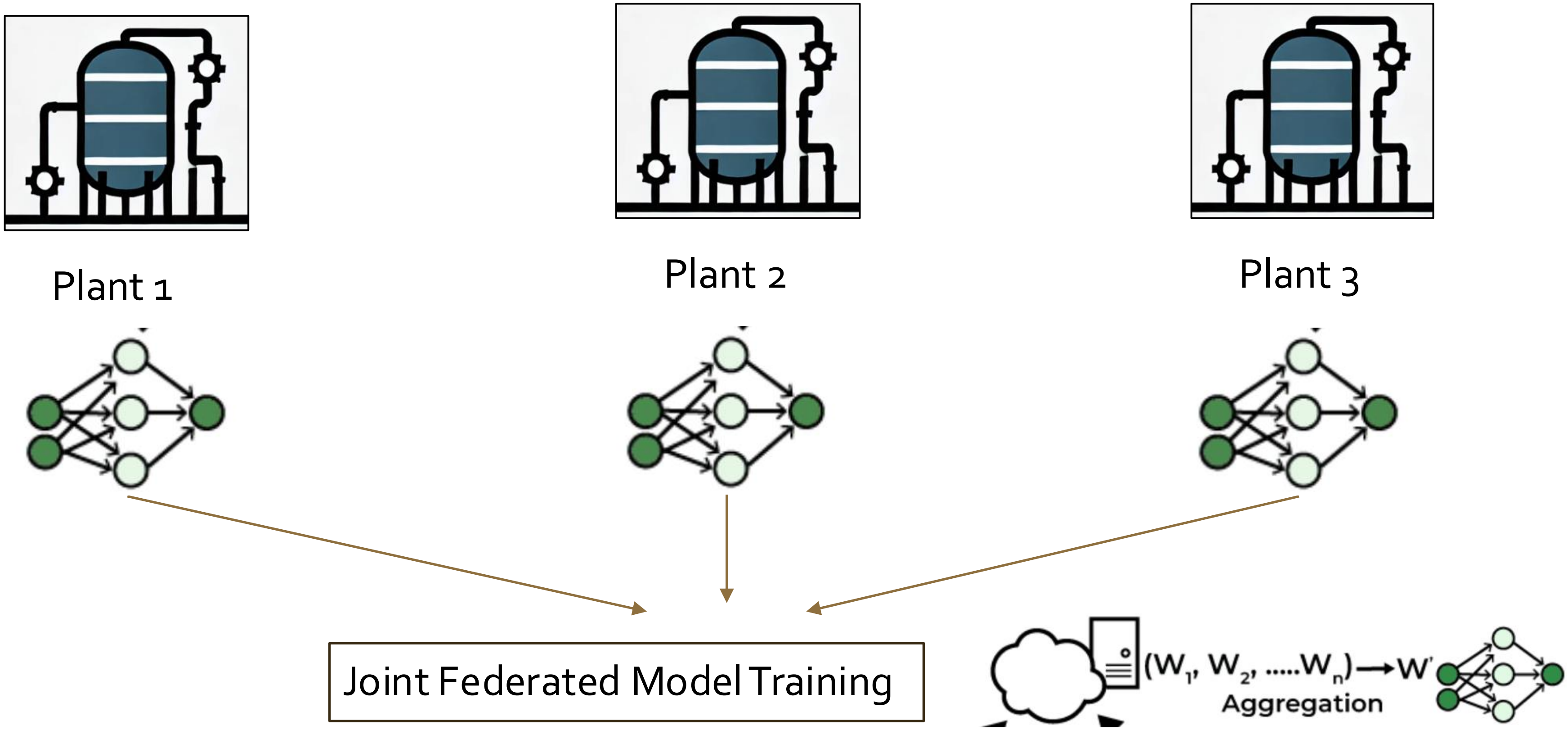
US Census will use differential privacy in 2020 decennial census.

Census plans to make code and privacy parameters public.

2008 AD

Federated Forecasting Model

- Federated predictive maintenance of equipment



Joint Federated Model Training

- Issues in joint training: Privacy of data points
- Solution: ***Differential privacy*** based → Differentially private federated learning

Acknowledgement

P2SAC
Purdue Process Safety & Assurance Center

Air Liquide

ExxonMobil

ORACLE

amazon

U.S. National
Science
Foundation

American
Cancer
Society

NIH
National Institutes
of Health

Office of Naval Research
Science & Technology

ACS
Chemistry for Life

References

- Flores, G. E. C., Chen, H., & Li, C. Enforcing Hard Linear Constraints in Deep Learning Models with Decision Rules. In The Thirty-ninth Annual Conference on Neural Information Processing Systems. 2025
- Li, C. (2025). Breaking data silos in drug discovery with federated learning. Nature Chemical Engineering, 2(5), 288-289.
- Khan, A., Nahar, R., Chen, H., Flores, G. E. C., & Li, C. (2025). FaultExplainer: Leveraging large language models for interpretable fault detection and diagnosis. Computers & Chemical Engineering, 109152.
- Chen, H., Flores, G. E. C., & Li, C. (2024). Physics-informed neural networks with hard linear equality constraints. Computers & Chemical Engineering, 189, 108764.
- Pal, S & Li, C (2026). DisjunctiveNet: Neural Symbolic Learning via Differentiable Convexified Optimization Layers. Forty-third International Conference on Machine Learning (ICML 2026)