Sweden AI Factory

Explainable AI for Life Science

Turning Predictions into Biological Insights

Ivan Cucchi
AI Application Specialist in Life Science

What is Life Science: The study of life at every scale

Life science is the scientific study of living organisms, life processes, and their relationships with each other and the environment.

INCLUDES

- **Botany** — plants
- **Zoology** — animals
- **Microbiology** — microorganisms
- **Biochemistry** — chemical processes
- **Ecology** — ecosystems

Molecules

Areas
Structural biology, biochemistry, chemistry

Data
Sequences, 3D structures

Cells

Areas
Cell biology, genomics, omics

Data
Sequencing, microscopy

Organs

Areas
Pathology, physiology, neuroscience

Data
Histology, MRI, CT

Organisms

Areas
Clinical medicine, pharmacology, cardiology

Data
ECG, lab values, health records

Populations

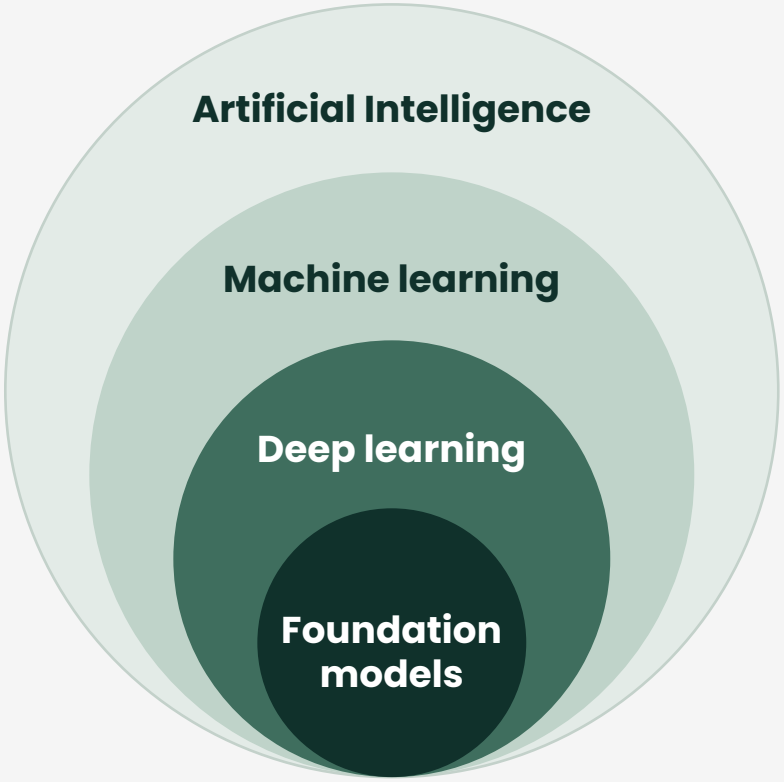
Areas
Epidemiology, public health, ecology

Data
Registers, cohorts, biobanks

Every scale of life now produces more data than researchers can analyse by hand.

Sweden AI Factory

What is AI: Models that learn patterns from data



Artificial intelligence
Any system performing tasks we associate with human intelligence: perceiving, reasoning, deciding.
e.g. rule-based clinical decision support

Machine learning
Learns the rules from labelled examples, instead of being programmed with them.
e.g. risk scores from health records

Deep learning
Many-layered neural networks that learn features directly from raw images, sequences and signals.
e.g. tumour detection on pathology slides

Foundation models
Very large models pretrained on massive data, then adapted to many downstream tasks.
e.g. protein language models, LLMs

The same learning recipe works on proteins, images, signals and records: one engine for every discipline.

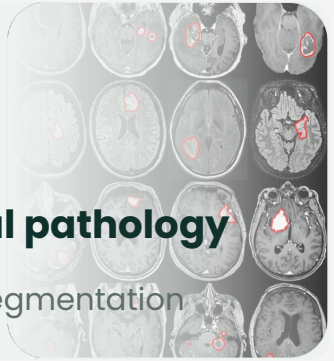
Sweden AI Factory

The new standard: AI is redefining speed and scale in Life Science



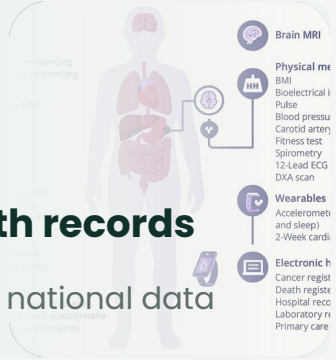
Structure prediction

Folding and complex prediction



Imaging & digital pathology

Detection, grading, segmentation



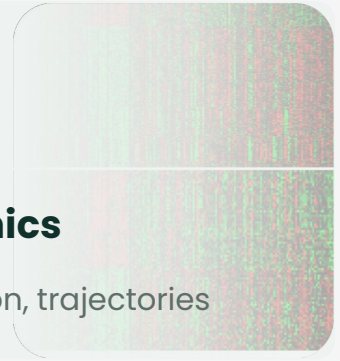
Registers & health records

Risk models on linked national data



Cardiac risk prediction

ECG and imaging signals flag disease early



Single-cell & omics

Cell typing, integration, trajectories



Drug discovery

Screening and design before the bench

These advances have established AI as a cornerstone of modern biomedical research and clinical innovation.

Every discipline → AI ← one shared engine

Sweden AI Factory

The new standard: AI is redefining speed and scale in Life Science

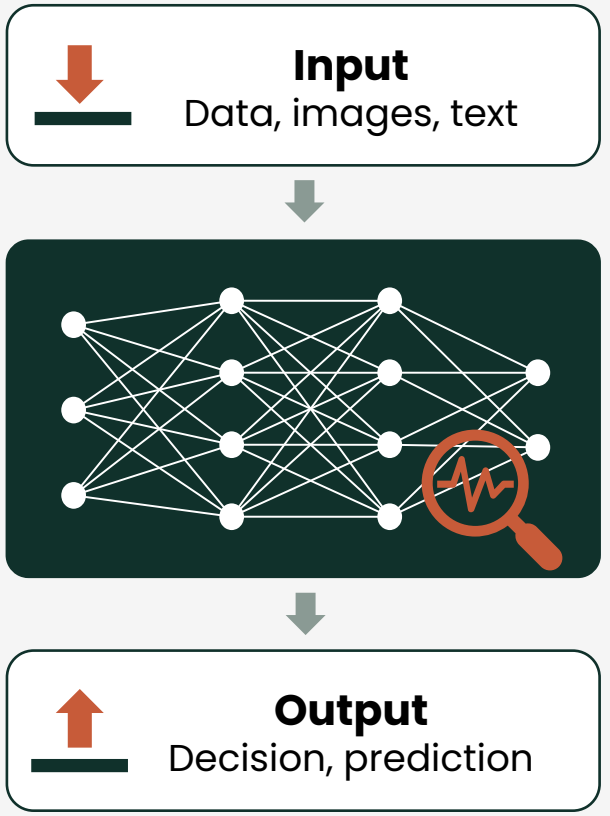


AI is not magic

Sweden AI Factory

The black box problem: High predictive power, zero insight

- **Hidden Logic:** Billions of parameters create internal representations that humans cannot audit or trace.
- **Silent Failure Modes:** High benchmark accuracy can mask reliance on artifacts, bias, or spurious correlations.
- **The Adoption Bottleneck:** High-stakes fields, such as *healthcare and biomedicine*, cannot act on predictions without understanding the clinical rationale.



XAI has emerged as a critical field aimed at improving transparency without sacrificing predictive performance.

Sweden AI Factory

What is XAI: Methods that make a model's reasoning visible to the people who rely on it

DEFINITION

Explainable AI (XAI) is the set of methods that make an AI model's outputs understandable to humans.

It answers two questions: why did the model decide this, and should I trust it?

Interpretable by design vs. post-hoc

Some models can be read as built (linear models, decision trees); post-hoc methods explain a trained black box afterwards.

Global vs. local

Global explanations describe the model's overall logic; local ones explain a single prediction, such as one patient.

Model-agnostic vs. model-specific

Agnostic methods (LIME, SHAP) work on any model; specific ones (Grad-CAM) use a network's internals.

The right explanation depends on who is asking and which decision it supports.

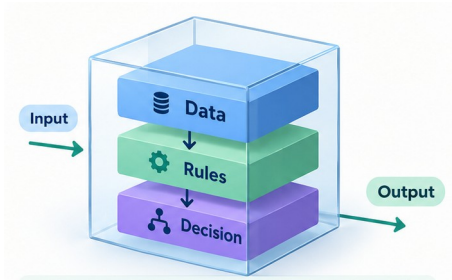
Sweden AI Factory

Interpretability vs explainability: Understanding the model, or explaining one output

INTERPRETABILITY

How does the model work?

- **What it is:** the model is transparent by design, so a person can follow its logic end to end
- **Examples:** linear risk scores, short decision trees, rule lists
- **Trade-off:** often less accurate on images, omics and other complex data

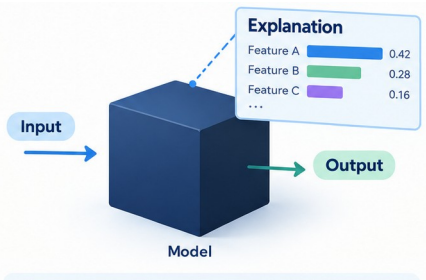


VS

EXPLAINABILITY

Why did it produce this output?

- **What it is:** a second method explains a complex model after training
- **Examples:** SHAP values, LIME, Grad-CAM heatmaps
- **Trade-off:** the explanation approximates the model and can be wrong



Use an interpretable model where you can; explain a complex model where you must.

Sweden AI Factory

Types of XAI: Different methods for different question

FEATURE ATTRIBUTION

Which inputs drove this prediction?

Methods
SHAP, LIME, Integrated Gradients

Life-science example
Lab values behind a sepsis risk score

SALIENCY MAPS

Where in the image did the model look?

Methods
Grad-CAM, occlusion, attention maps

Life-science example
Tumour regions on a pathology slide

EXAMPLE-BASED

What would change the outcome?

Methods
Counterfactuals, prototypes

Life-science example
The molecular change that removes toxicity

INTERPRETABLE MODELS

Can we read the model directly?

Methods
Decision trees, rule lists, GAMs

Life-science example
A bedside risk score clinicians can check

Explanations approximate the model: validate them like any other result.

Sweden AI Factory

Right prediction, wrong reason: The *wolf vs. husky* failure mode



Convincing at first glance

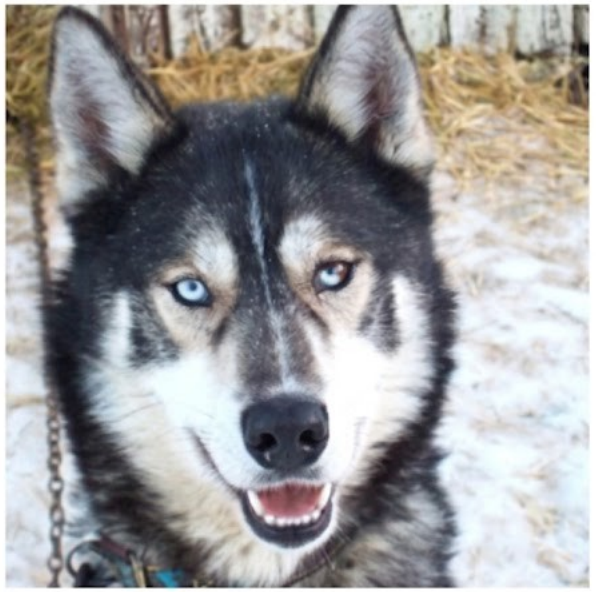
The classifier got 8 of 10 test images right.

Shown only its predictions, 10 of 27 ML-trained users said they would trust it.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sweden AI Factory

Right prediction, wrong reason: The *wolf vs. husky* failure mode



(a) Husky classified as wolf



(b) Explanation

The Hidden Shortcut

Explanation techniques revealed the model ignored biological features and classified based entirely on snow in the background.

After seeing the explanations, only 3 of 27 still trusted it.

Without explainability, a "great snow detector" passed as a wolf classifier.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sweden AI Factory

The illusion of robustness: Unfamiliar inputs still get confident answers

Out-of-Distribution Blind Spots
When fed entirely unfamiliar inputs, closed-world models do not output *"unknown"*.
Instead, they force a prediction into existing classes with deceptively high statistical confidence.

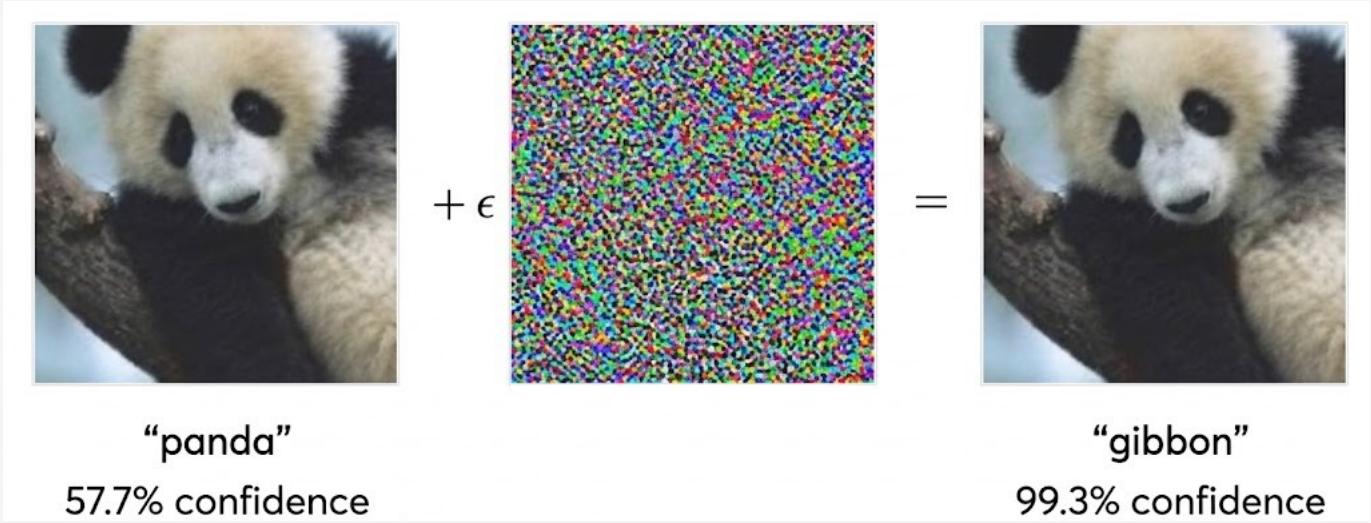


Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. 5th International Conference on Learning Representations (ICLR 2017). https://doi.org/10.48550/arXiv.1610.02136

Sweden AI Factory

The illusion of robustness: Imperceptible noise flips confident predictions

Perturbation Fragility
Adding imperceptible pixel-level noise flips a confident prediction.
It highlights that models rely on high-dimensional mathematical artifacts rather than true semantic features.



Neural networks do not necessarily know when and what they don’t know

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1412.6572

Sweden AI Factory

The new mandate: Why explainability is essential for Life Science

CLINICAL

Clinicians must verify the reasoning

- No diagnosis or treatment on a prediction nobody can explain
- Exposes shortcuts, like scanner tags or text markers, before they reach patients

ETHICAL

Decisions must be fair and contestable

- Reveals reliance on sex, ethnicity or hospital site instead of biology
- Gives patients and clinicians a reason they can challenge

REGULATORY

The law requires transparency

- **EU AI Act:** medical AI is high-risk (Art. 13 transparency, Art. 14 oversight)
- **GDPR:** meaningful information about the logic involved
- **MDR / FDA:** transparency principles for ML devices

SCIENTIFIC

Explanations generate new hypotheses

- Feature attributions point to candidate biomarkers and mechanisms
- Turns a predictive model into a tool for discovery

In biomedicine, explainability is not an academic convenience, it is clinical, ethical, and regulatory necessity.

Krejcar, O., Abdullah, J., & Namazi, H. (2026). Implementing XAI in life sciences: Key challenges and pathways to solutions. *Artificial Intelligence in the Life Sciences*, *6*. https://doi.org/10.1016/j.ailsci.2026.100153

Sweden AI Factory

Clinical risk prediction: Real-time explanations helped anaesthesiologists anticipate low oxygen

THE MODEL

Prescience forecasts hypoxaemia

- Gradient-boosted model trained on records from over 50000 surgeries
- Forecasts the risk of low blood oxygen minutes ahead

WHAT XAI REVEALED

The drivers behind each risk score

- SHAP values show which factors raise or lower risk for this patient, right now
- Drivers include BMI, oxygen saturation trend and ventilation settings

WHY IT MATTERS

Clinicians made better calls

- Anaesthesiologists anticipated hypoxaemia more accurately with explanations
- Each driver can be checked against the patient on the table

Explanations turn a risk score into something a clinician can act on.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, *2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Sweden AI Factory

Drug discovery: Explainable deep learning found a new structural class of antibiotics

THE MODEL

Graph neural networks screen compounds

- Trained on over 39000 compounds tested against S. aureus
- Also predicted toxicity to human cells

WHAT XAI REVEALED

The substructures behind activity

- A search over each molecule found the chemical fragments driving the prediction
- Chemists got a rationale, not just a ranked list of hits

WHY IT MATTERS

New antibiotics against MRSA

- The rationales pointed to a structural class active against MRSA and VRE
- Lead compounds worked in mouse infection models

Explanations can turn a prediction into a hypothesis scientists can test.

Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., Manson, A. L., Friedrichs, J., Helbig, R., Hajian, B., Fiejtek, D. K., Wagner, F. F., Soutter, H. H., Earl, A. M., Stokes, J. M., … Collins, J. J. (2024). Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, *626*(7997), 177–185. https://doi.org/10.1038/s41586-023-06887-8

Sweden AI Factory

Medical imaging: Explanations showed X-ray models reading the hospital, not the lungs

THE MODEL

Deep networks detect pneumonia and COVID-19

- Trained on chest X-rays from a few hospitals
- High accuracy on test images from the same sources

WHAT XAI REVEALED

The models learned shortcuts

- Saliency maps lit up text markers, laterality tokens and image borders
- Models could tell which hospital an image came from

WHY IT MATTERS

Accuracy did not transfer

- Performance fell on X-rays from other hospitals
- Accuracy metrics alone had not exposed the flaw

A high test score is not evidence that the model learned medicine.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002683

DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, *3*(7), 610–619. https://doi.org/10.1038/s42256-021-00338-7

Sweden AI Factory

Protein language models: Explanations revealed the biology a foundation model learned

THE MODEL

ESM-2 learns proteins from sequence alone

- Foundation model pretrained on millions of protein sequences
- Single neurons mix many concepts, so they are hard to read

WHAT XAI REVEALED

Thousands of readable biological features

- Sparse autoencoders split each layer into thousands of features
- Features map to binding sites, motifs, domains and unannotated concepts

WHY IT MATTERS

From black box to discovery tool

- Features flag missing annotations in protein databases
- Dialling a feature up or down steers sequence generation

Explanations can show what a model has learned about biology, not only why it made one prediction.

Simon, E., & Zou, J. (2025). InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, *22*, 2107–2117. https://doi.org/10.1038/s41592-025-02836-7

Sweden AI Factory

Take-home message: Three things to remember

- **AI is everywhere:** it is now part of every scale of life science, from molecules to populations.
- **Accuracy is not understanding:** a model can score very well for the wrong reasons, and the numbers will not warn you.
- **Explainability closes the gap:** clinicians can check predictions, regulators can hold models accountable, and scientists can turn predictions into hypotheses they can test.

Next time a model gives you a great result, ask it: "Why should I trust you?"

Sweden AI Factory