UNIVERSITY OF CAMBRIDGE

Unveil the unseen: uncover hidden information with machine learning

Gareth Conduit

Model sparse datasets

Exploit property-property relationships

Merge data, computer simulations, and physical laws

Exploit uncertainties to deliver most robust predictions

Model sparse datasets

Exploit property-property relationships

Merge data, computer simulations, and physical laws

Exploit uncertainties to deliver most robust predictions

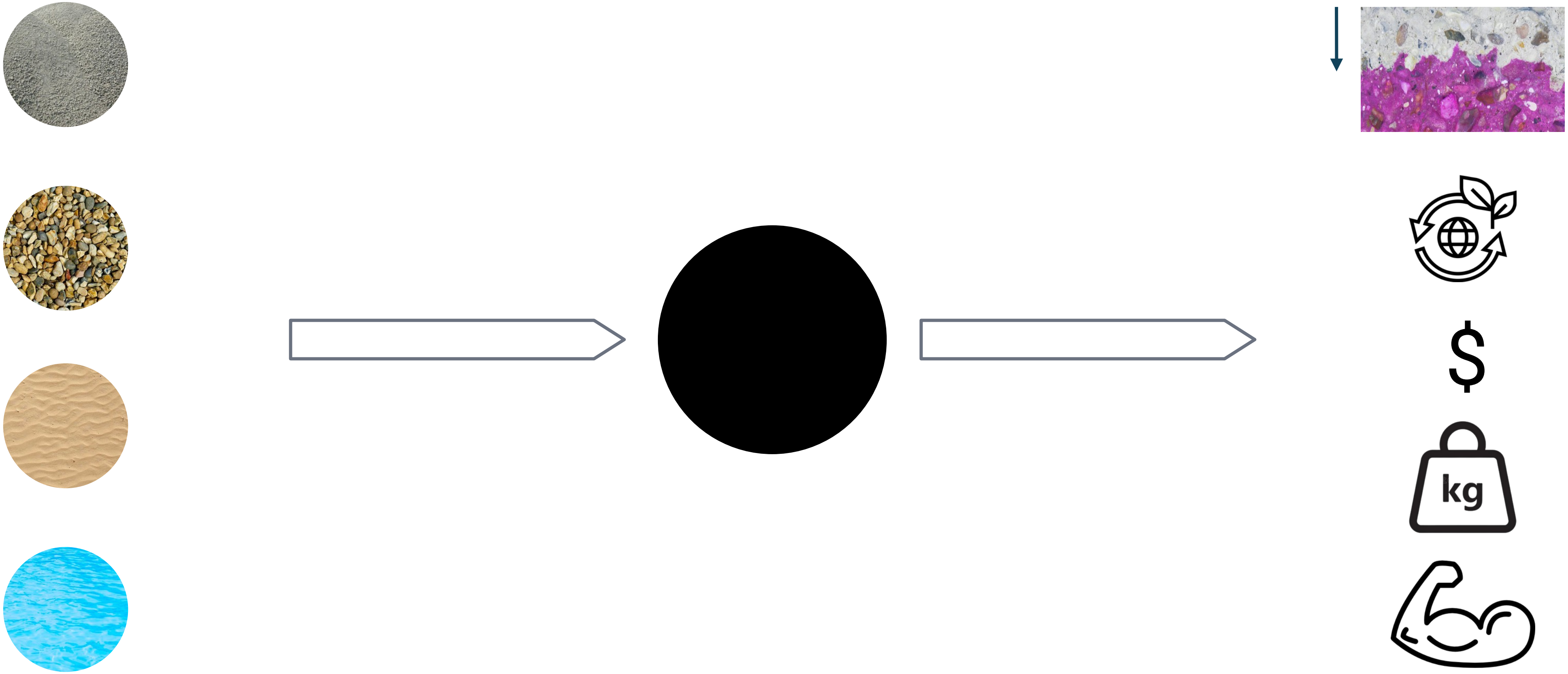
Extract information from noise itself

Machine learning architecture that can exploit uncertainty

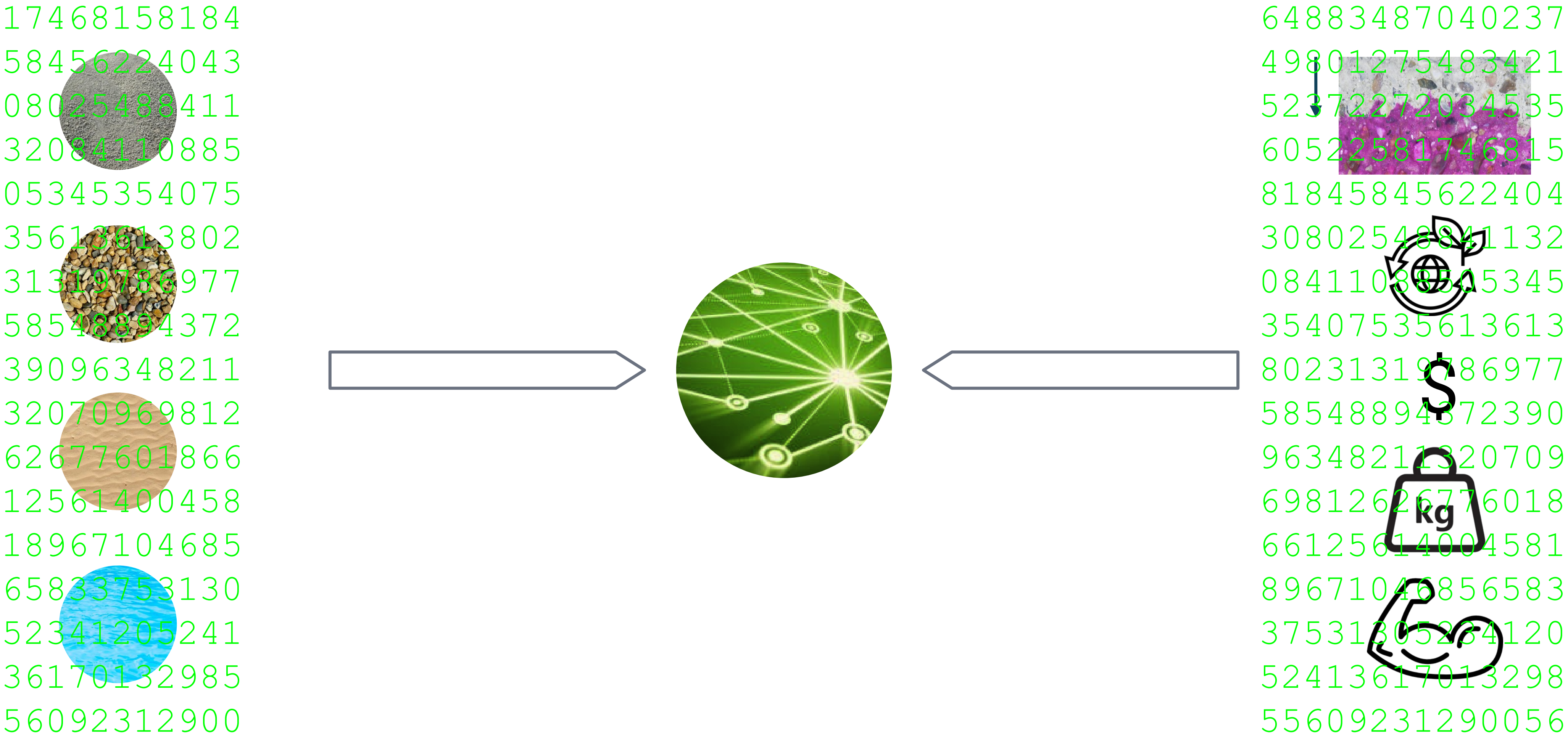
Bogdan Zviazhynski

Unveil the unseen: exploit information hidden in noise, BZ & GJC, Applied Intelligence (2022)

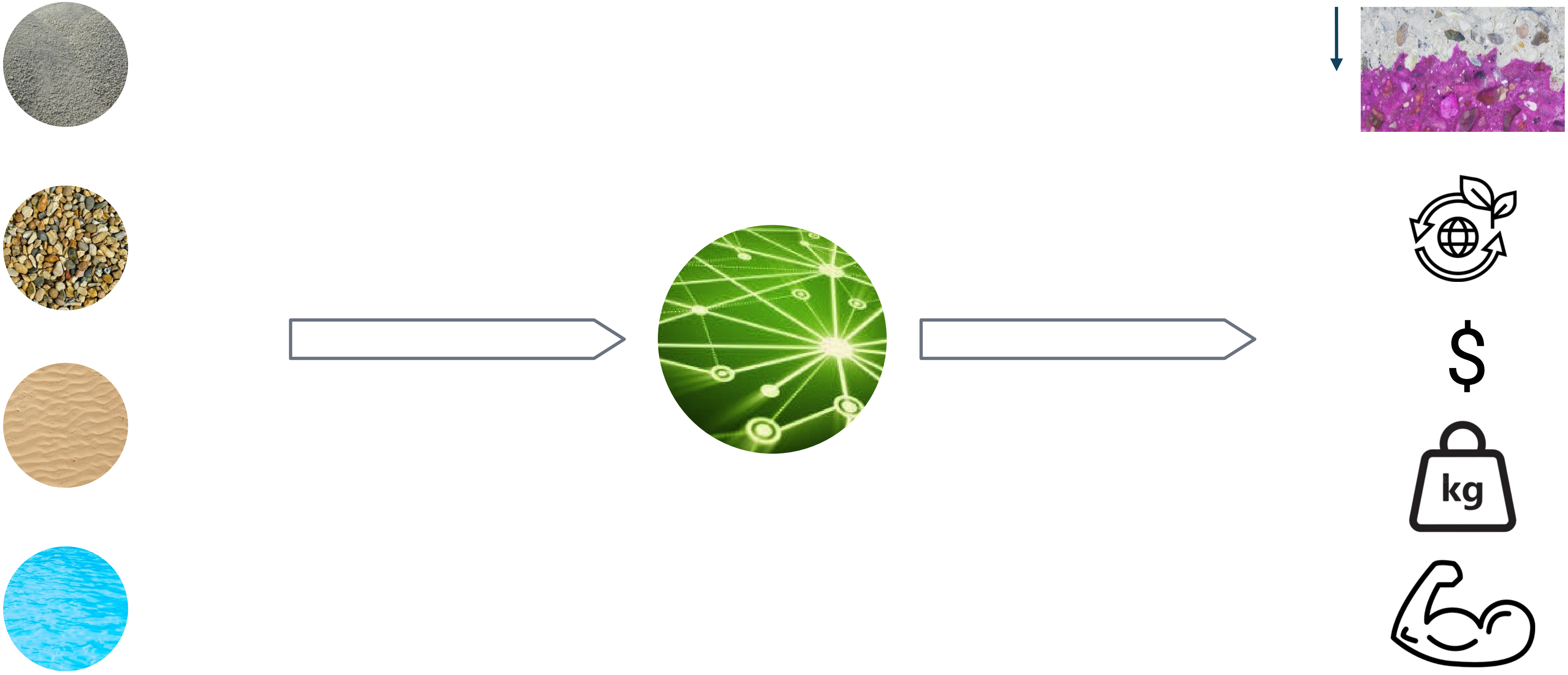
Black box machine learning for materials design



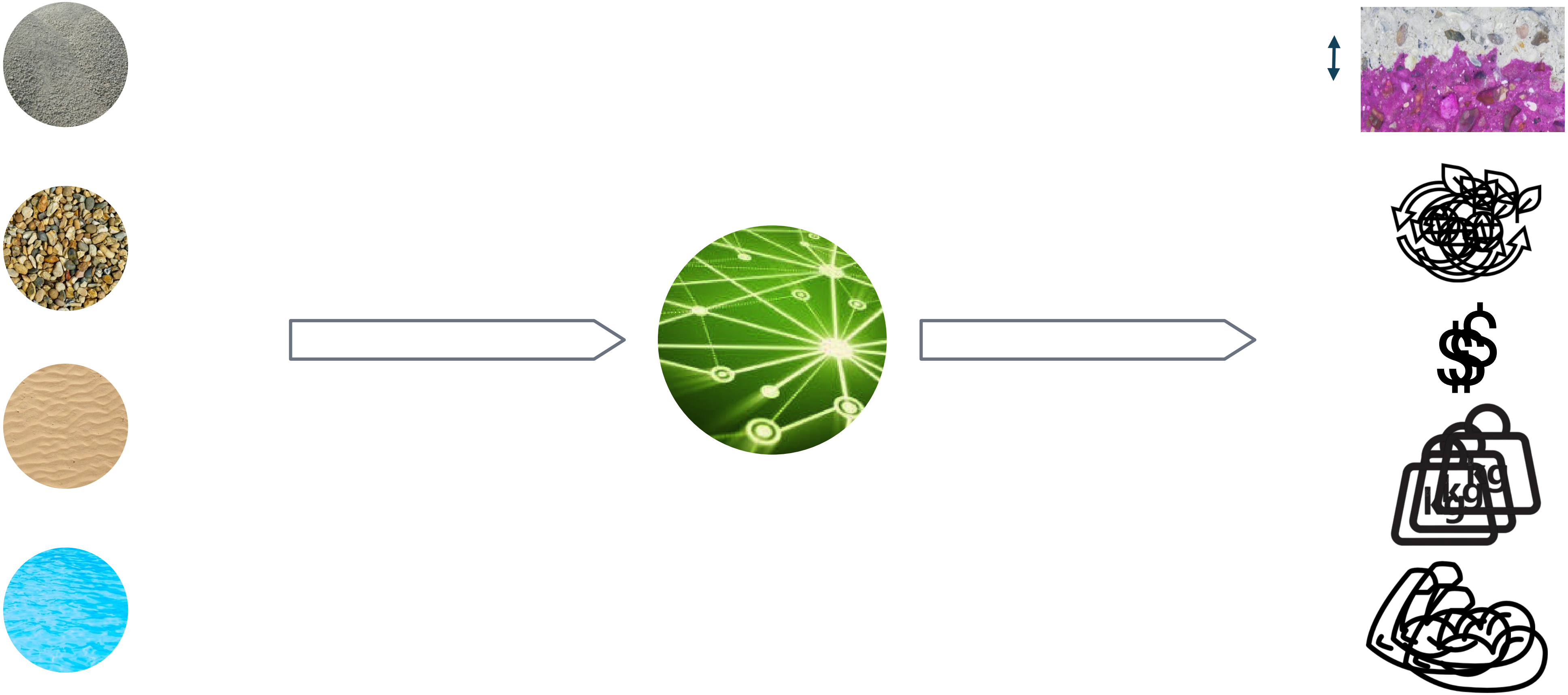
Train the machine learning



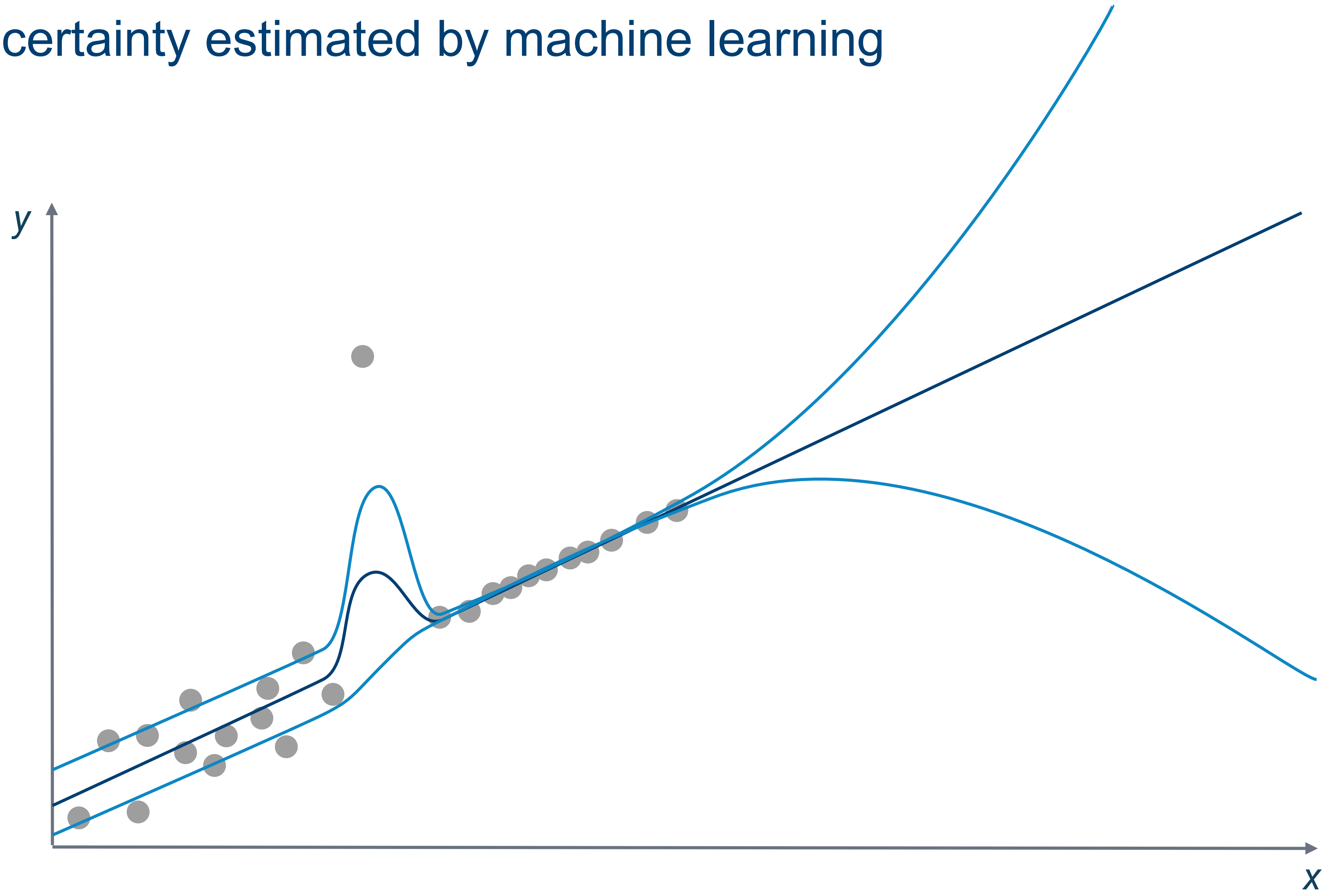
Machine learning predicts material properties



Machine learning estimates uncertainty



Uncertainty estimated by machine learning



Handling uncertainty

Design robust formulations

Outlier detection

Design of experiments

Unveil the unseen

Exploit information hidden in noise

Exemplar information extracted from noise

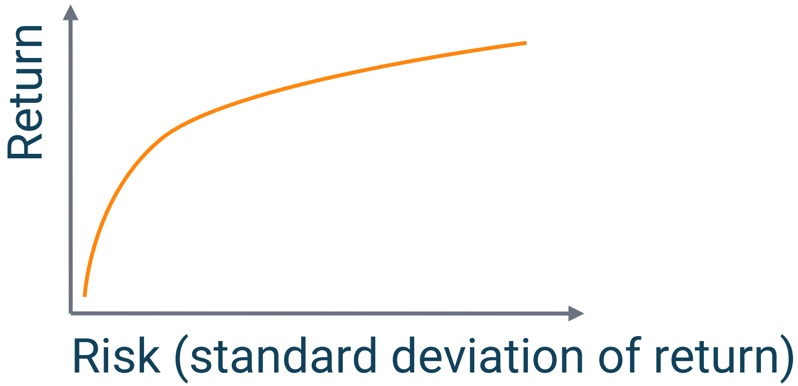
Renormalization group theory

applied to phase transitions

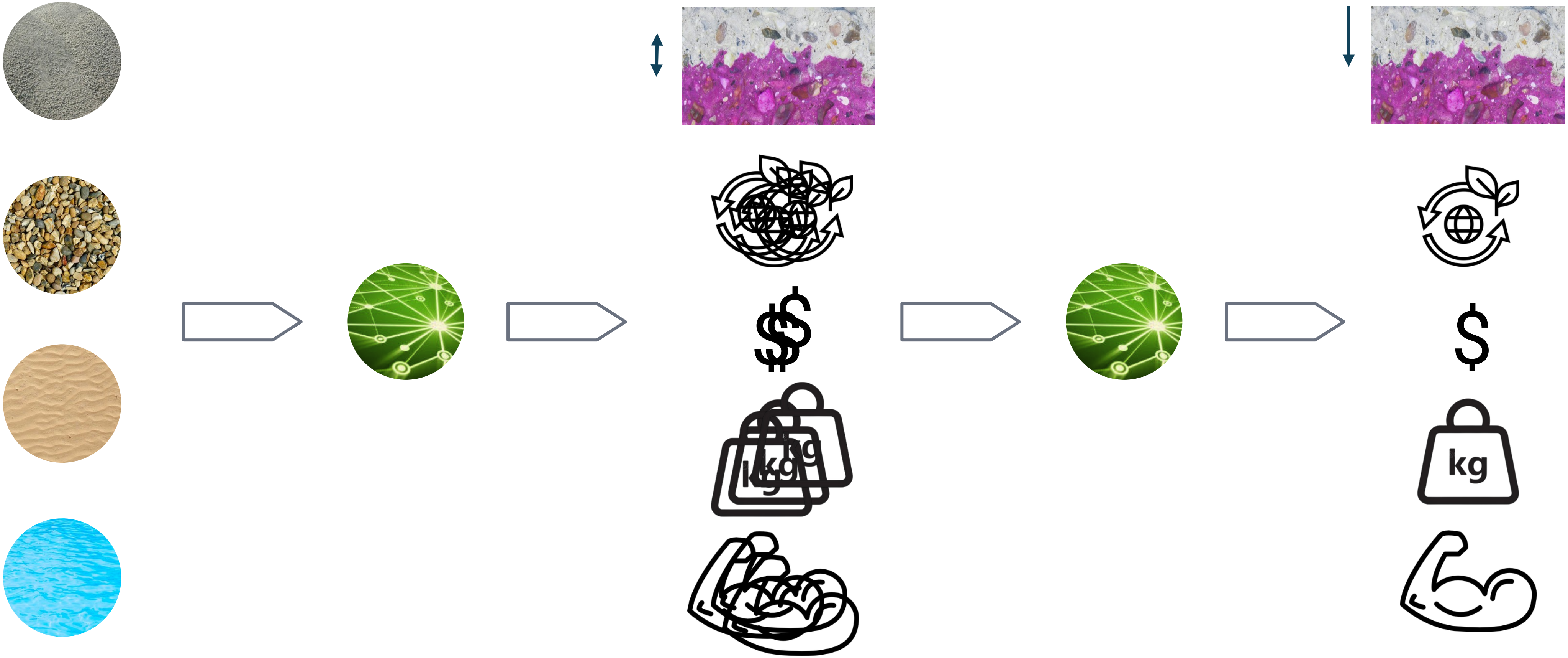
1982 Nobel Prize in Physics

Markowitz model

1990 Nobel Memorial Prize



Machine learning exploits uncertainty



Exploit uncertainty to design concrete

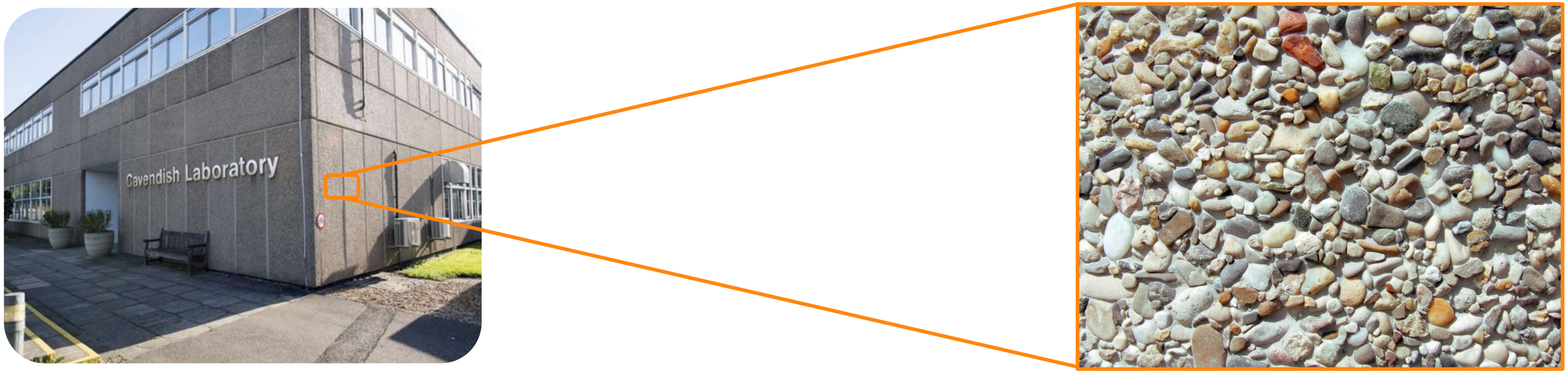
Bogdan Zviazhynski

Jess Forsdyke

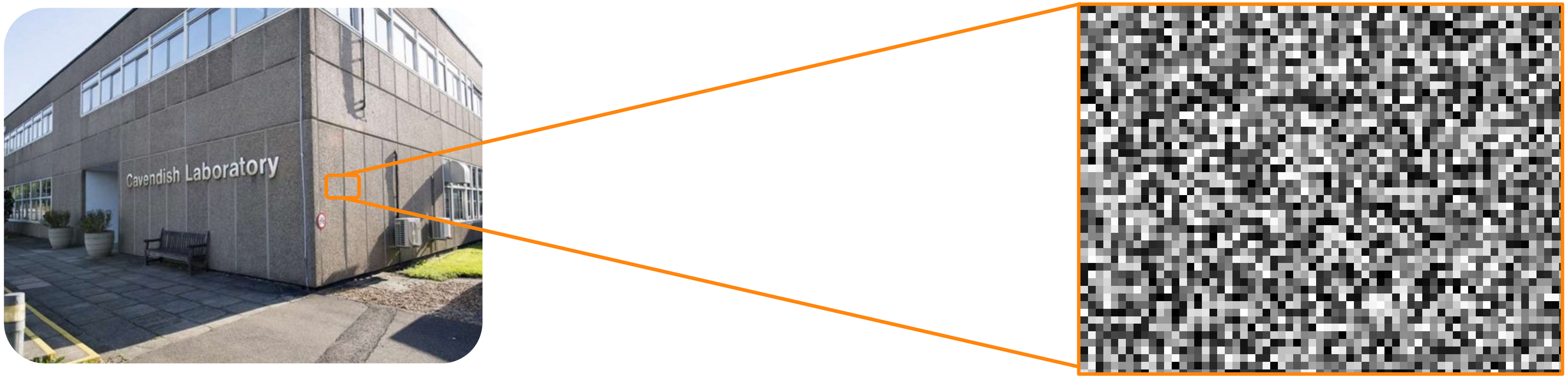
Professor Janet Lees

Concrete in construction

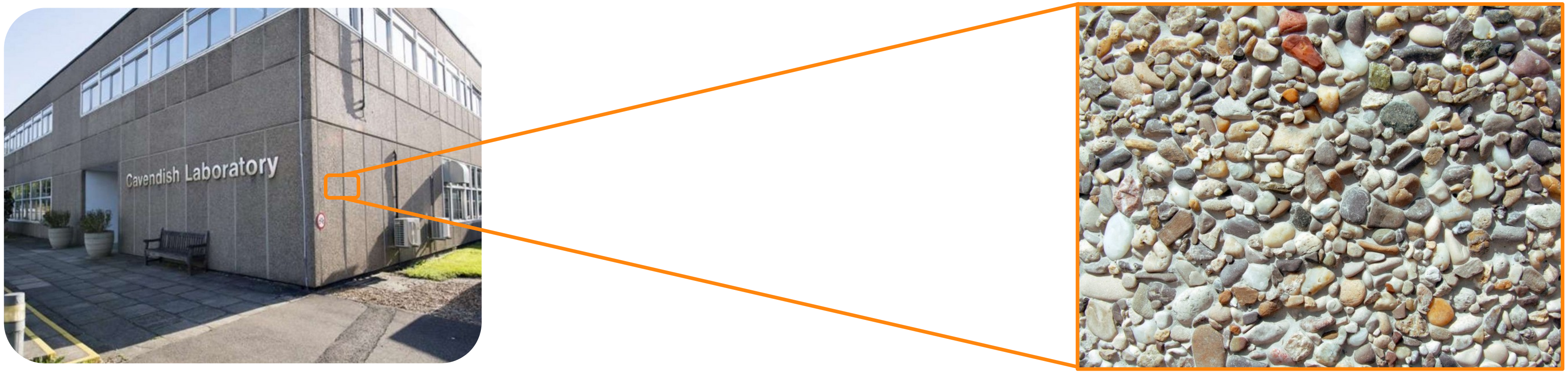
Cement & aggregate look like noise



Cement & aggregate look like noise



Mission



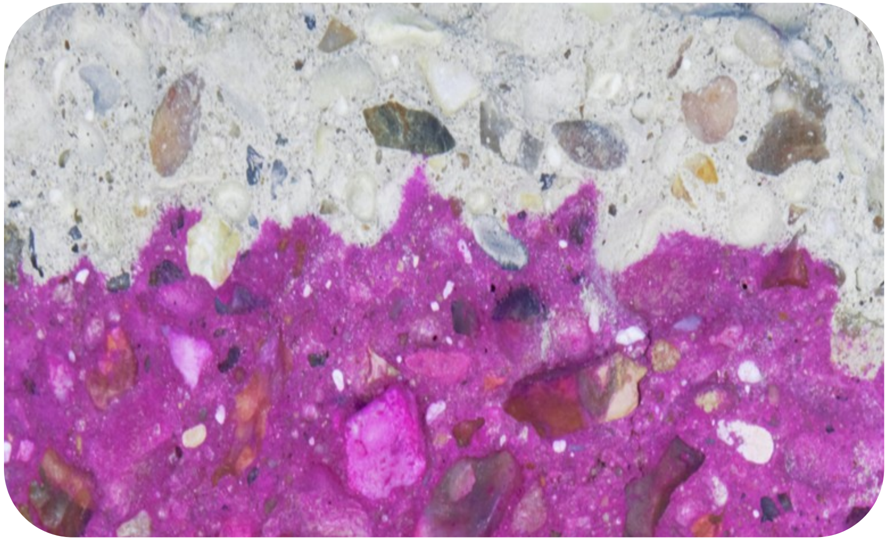
Design a concrete that is robust and environmentally friendly

Mission

Design a concrete that is robust and environmentally friendly

Experimentally validate the concrete

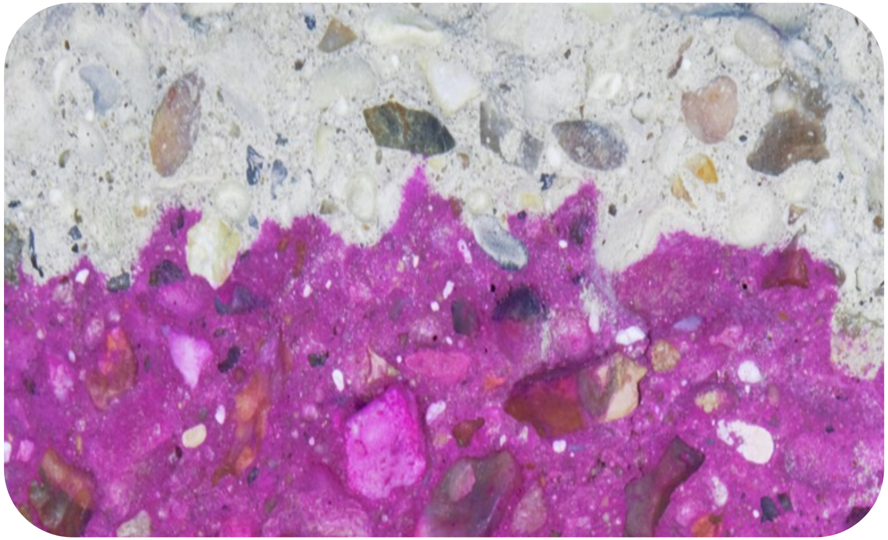
Carbonation is the probe of noise



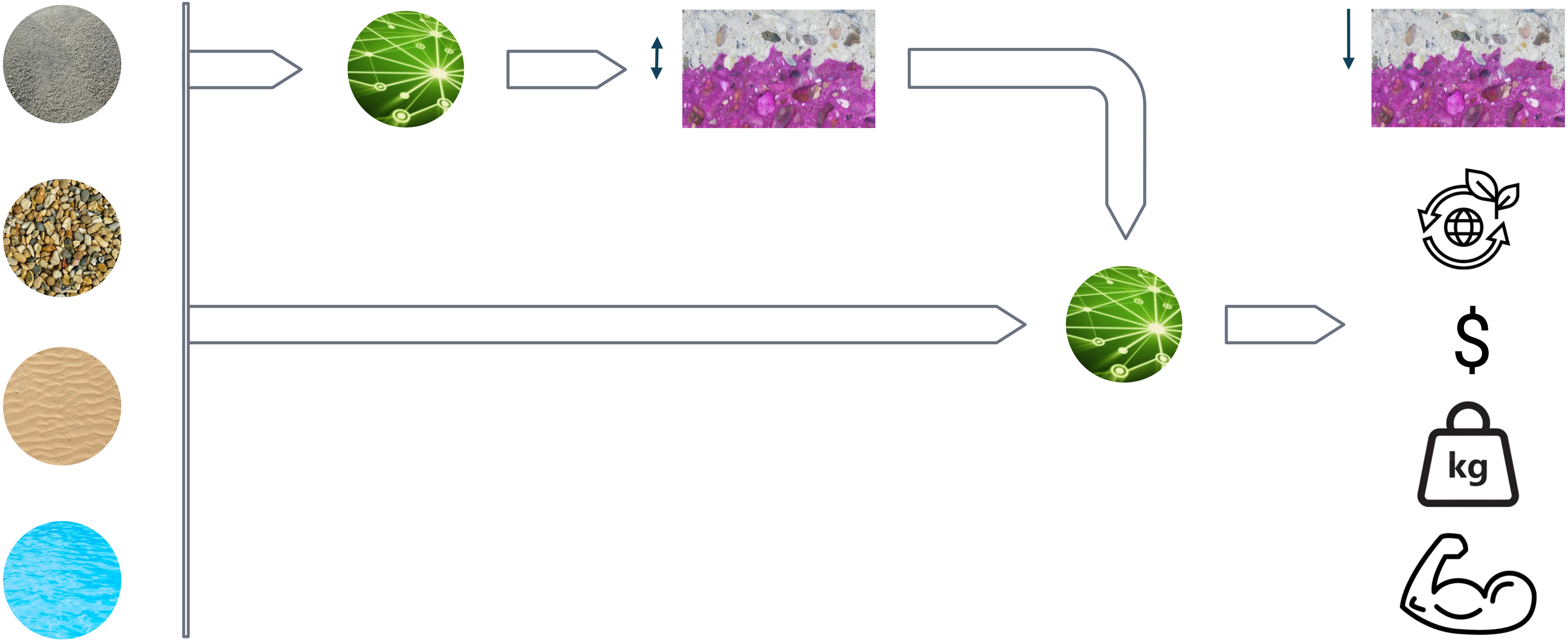
Depth and uncertainty in carbonation

Depth

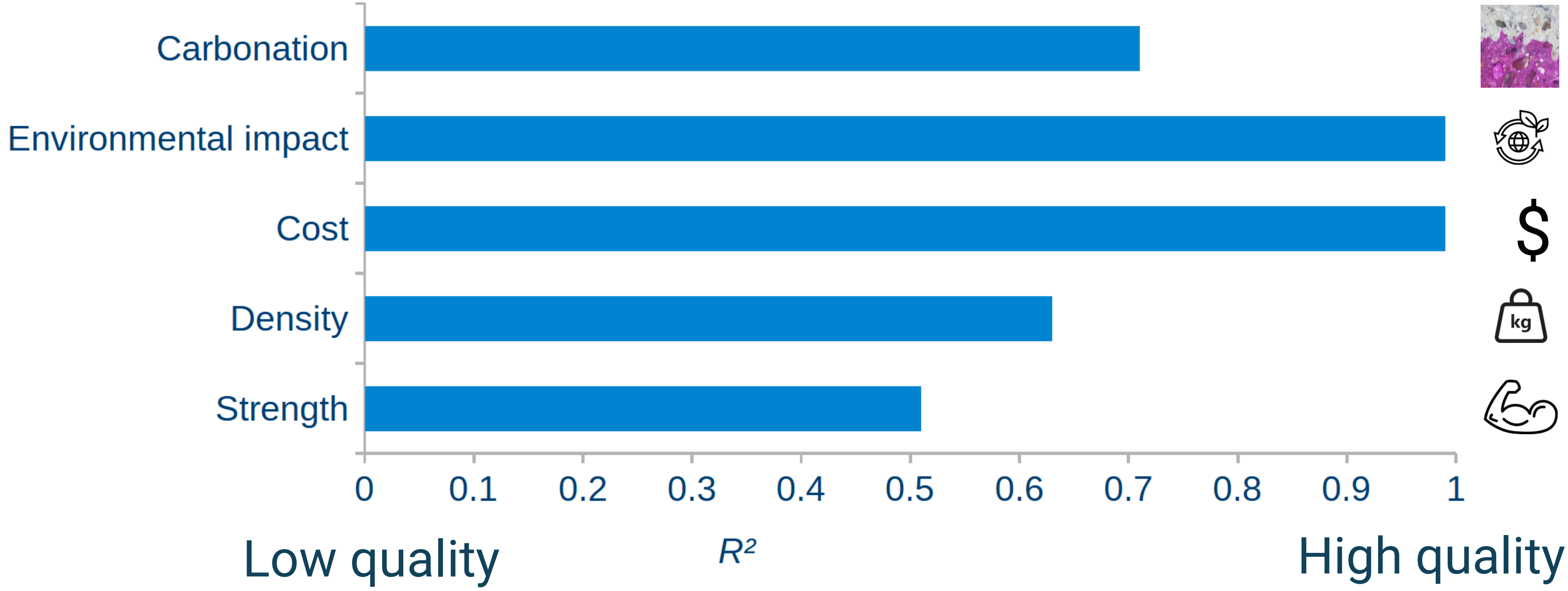
Uncertainty



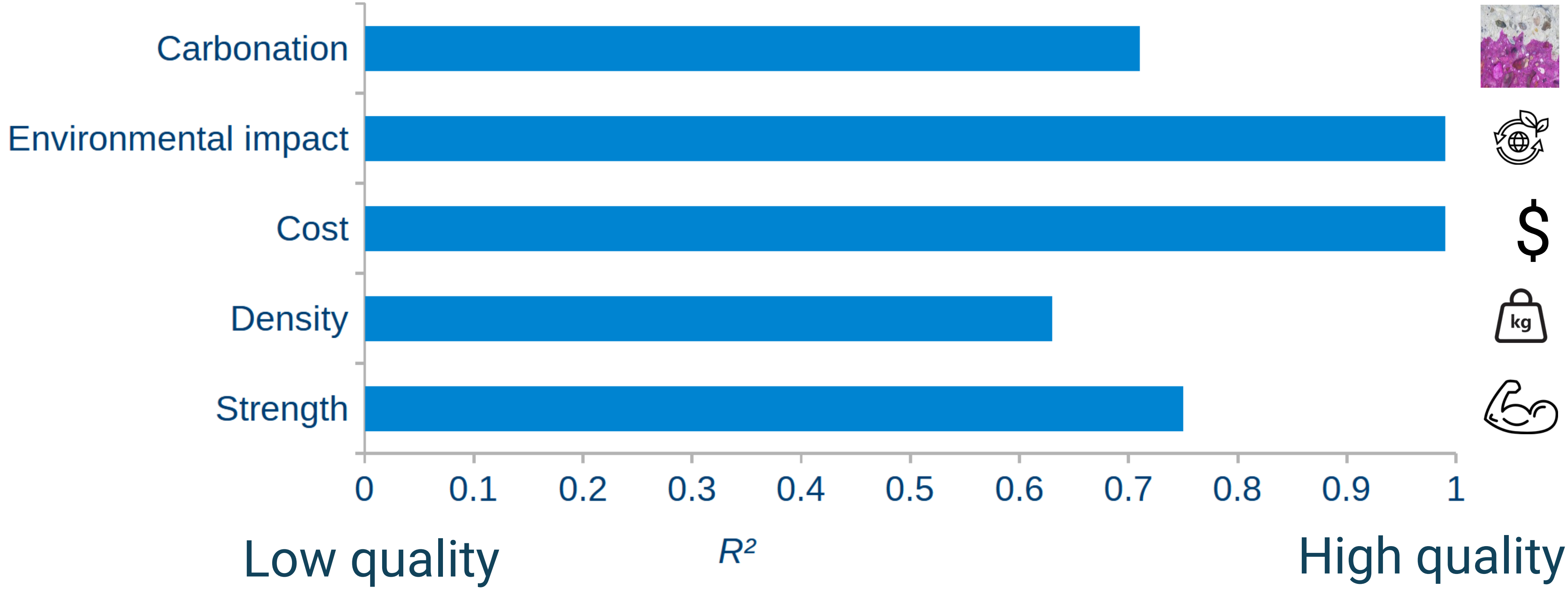
Machine learning exploits uncertainty



Original model accuracy



Uncertainty improves the model accuracy



Concrete specification

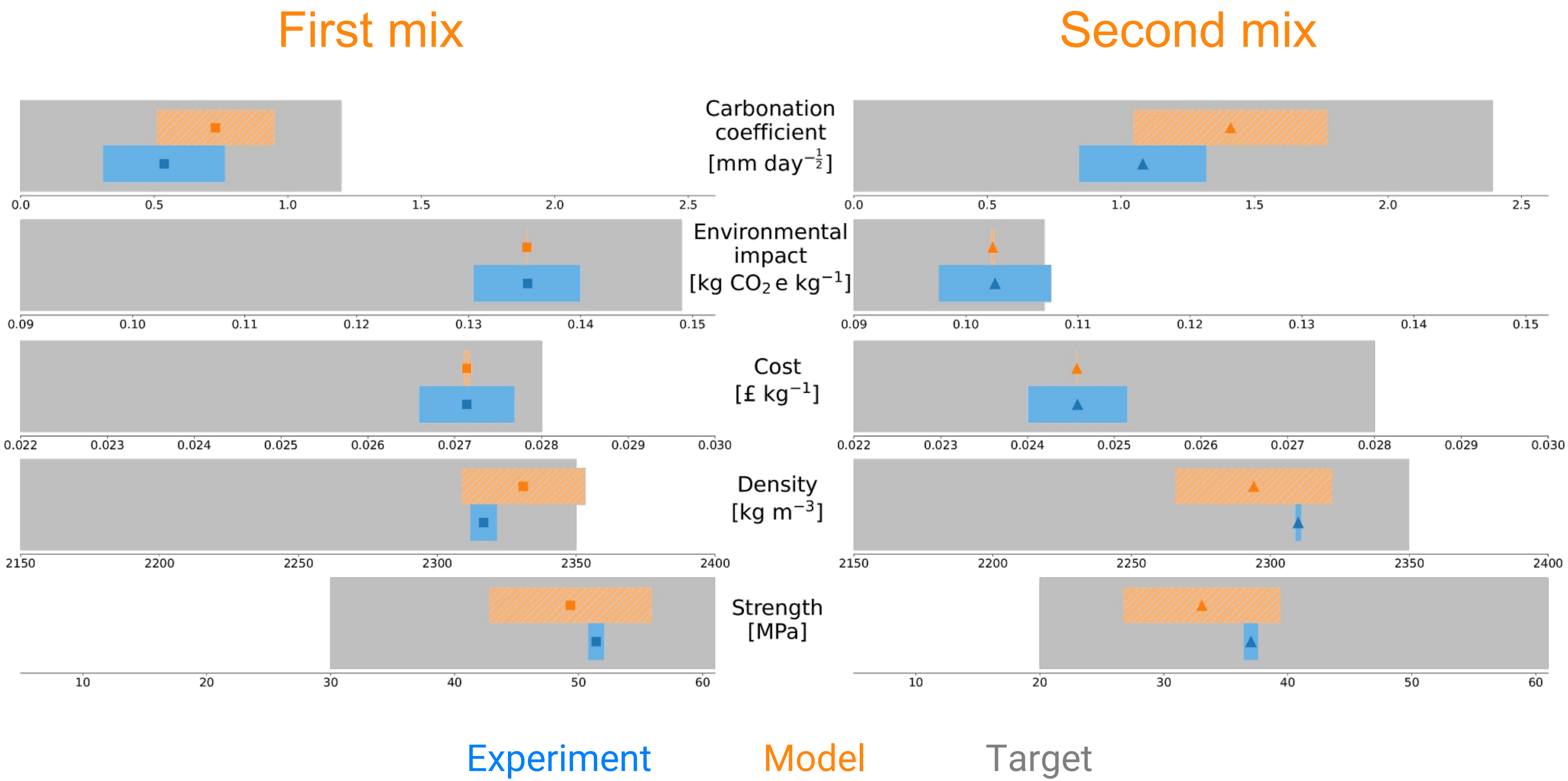
First mix		Second mix
⬇ carbonation		✔ carbonation
✔ environmental impact		⬇ environmental impact
✔ cost		✔ cost
✔ density		✔ density
✔ strength		✔ strength

Concrete design

First mix		Second mix
14.2% cement		10.5% cement
48.9% gravel		48.4% gravel
28.4% sand		32.6% sand
8.5% water		8.5% water

Concrete manufacture

Experimental validation of the proposed mixes



Exploit uncertainty to predict cancer

Bogdan Zviazhynski

Adriana Fonseca

Dr Jamie Blundell

Screening for disease

Entire population

Accurate
but expensive

Screening

Cheap but
low fidelity

Accurate
but expensive

Entire
population

0.1% of
population

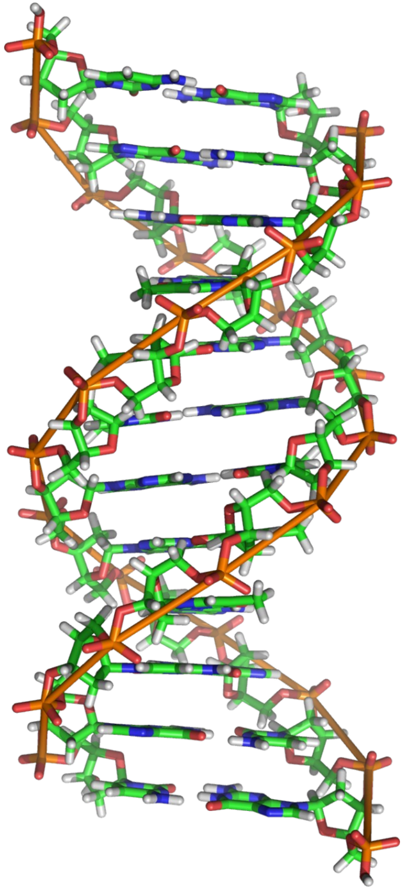
Lateral flow 78%

Prostate-specific antigen 25%

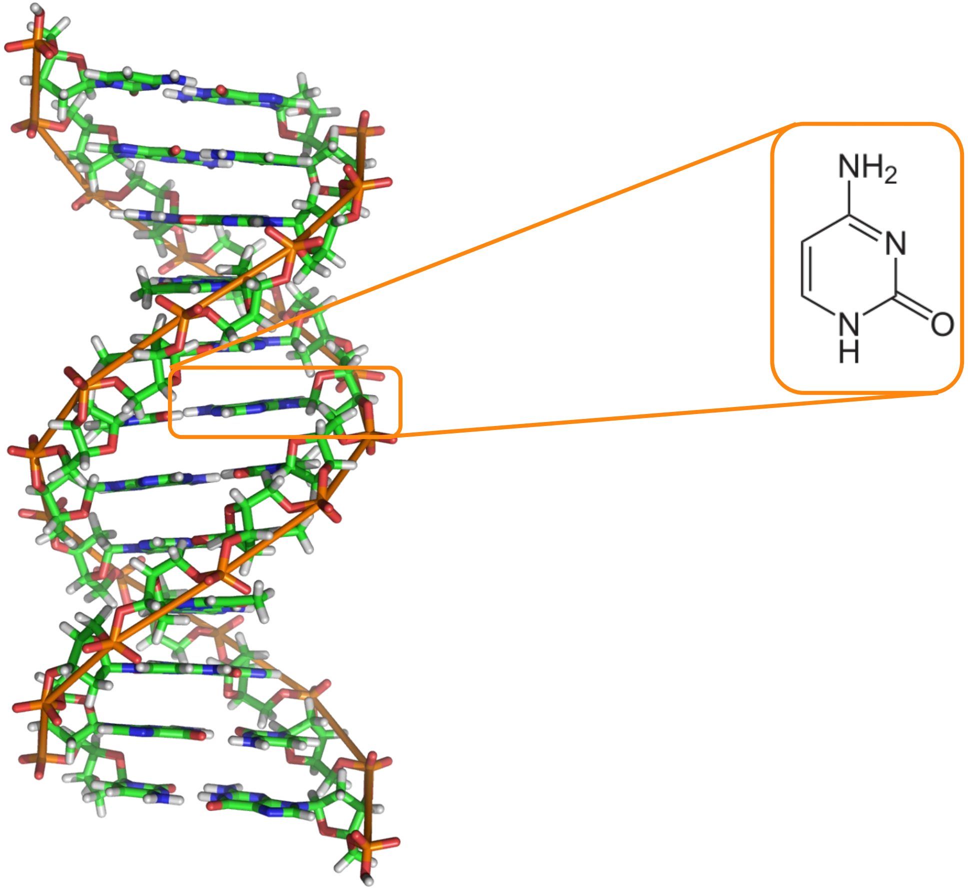
Mammogram 87%

Glucose Screening for diabetes 70%

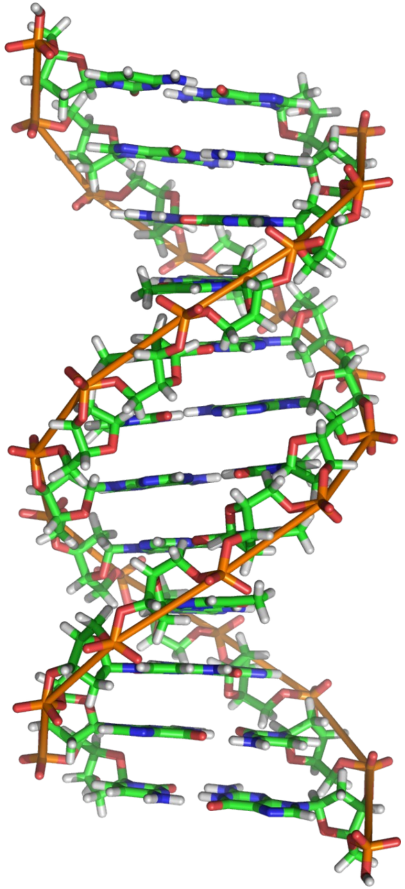
DNA

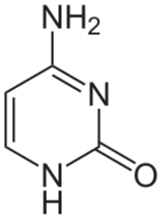


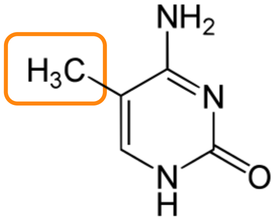
Cytosine nucleobase



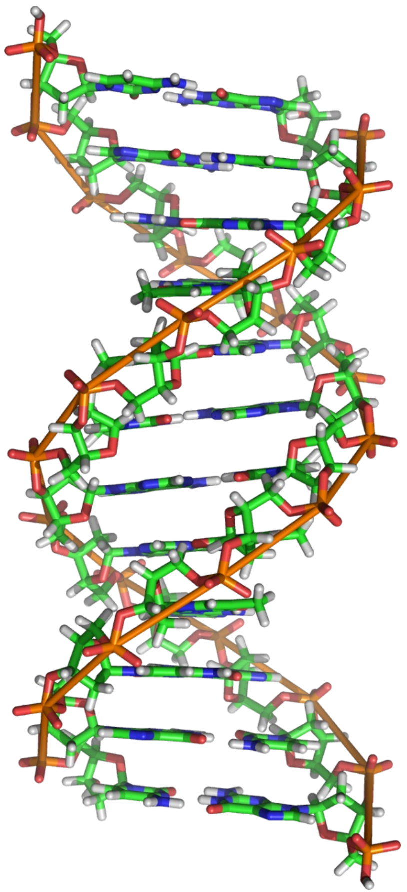
Methylation of cytosine

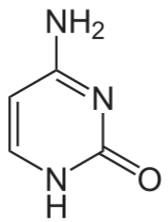


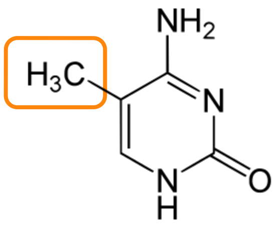




Methylation of cytosine







Required process in mammals, can repress genes

Associated with cancer

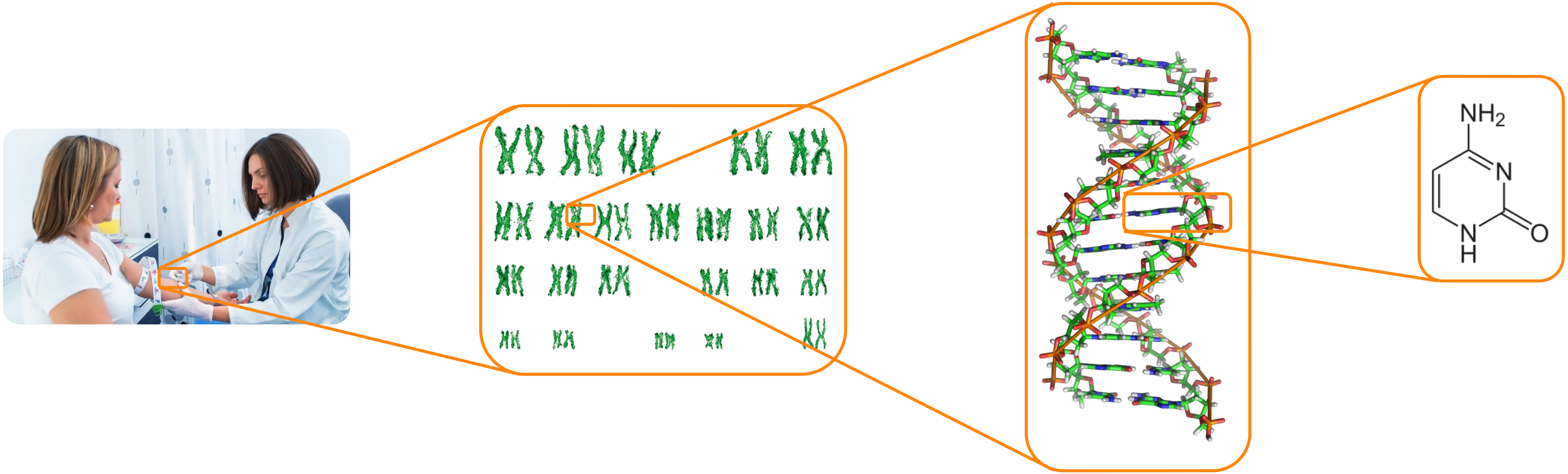
Rogue biology

causes chaotic methylation

resulting in cancer

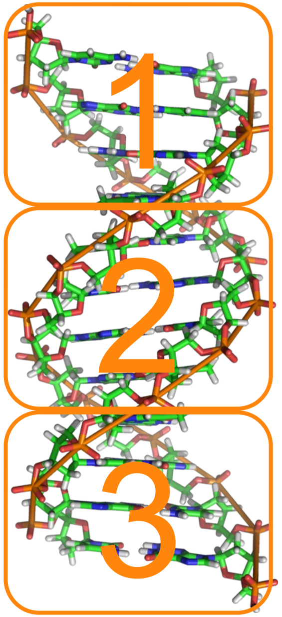
Methylation of cytosine

40 cohort tracked patients (20 developed acute myeloid leukaemia cancer, 20 did not)

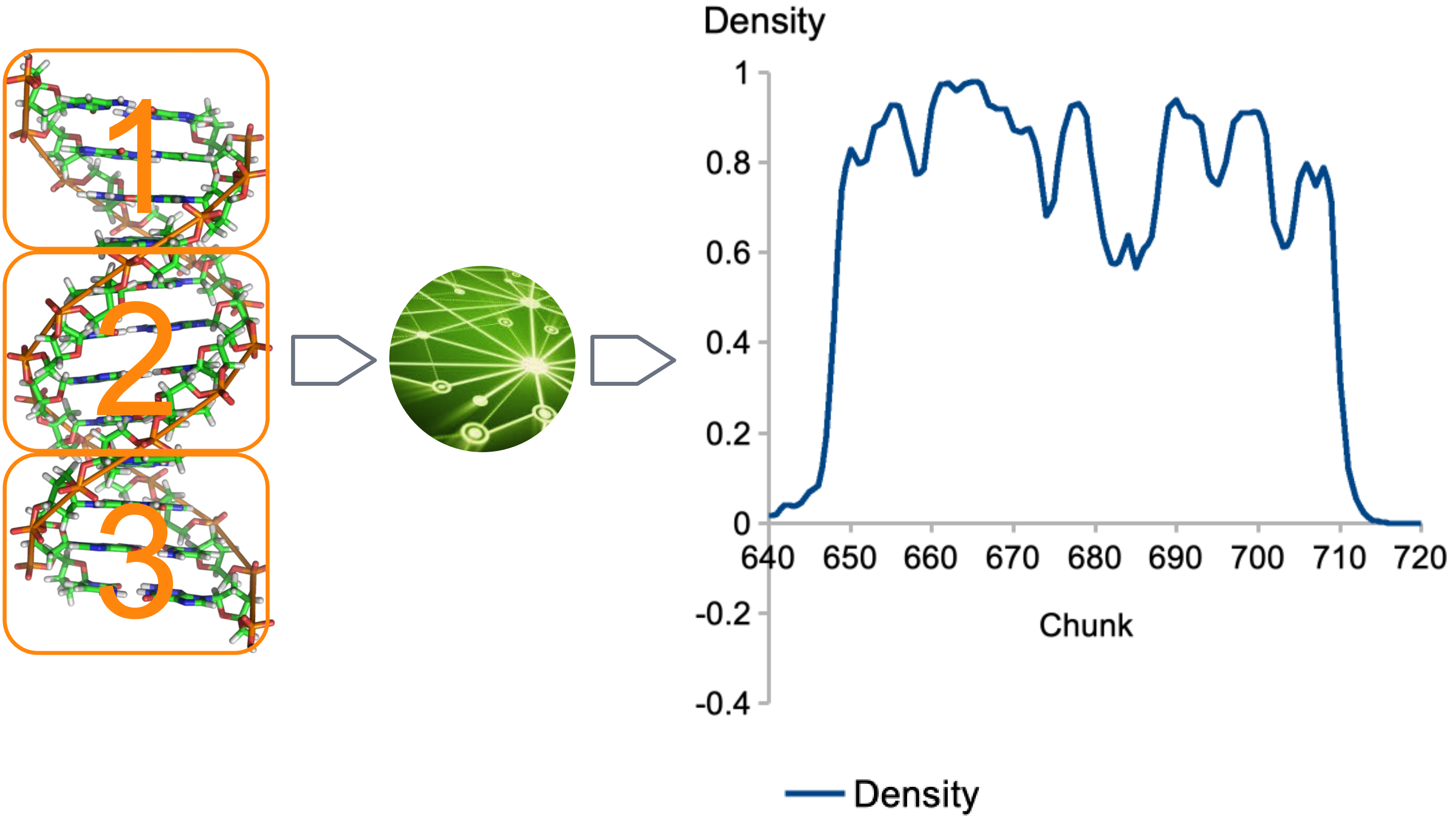


3,054,815,472 base pairs across all chromosomes

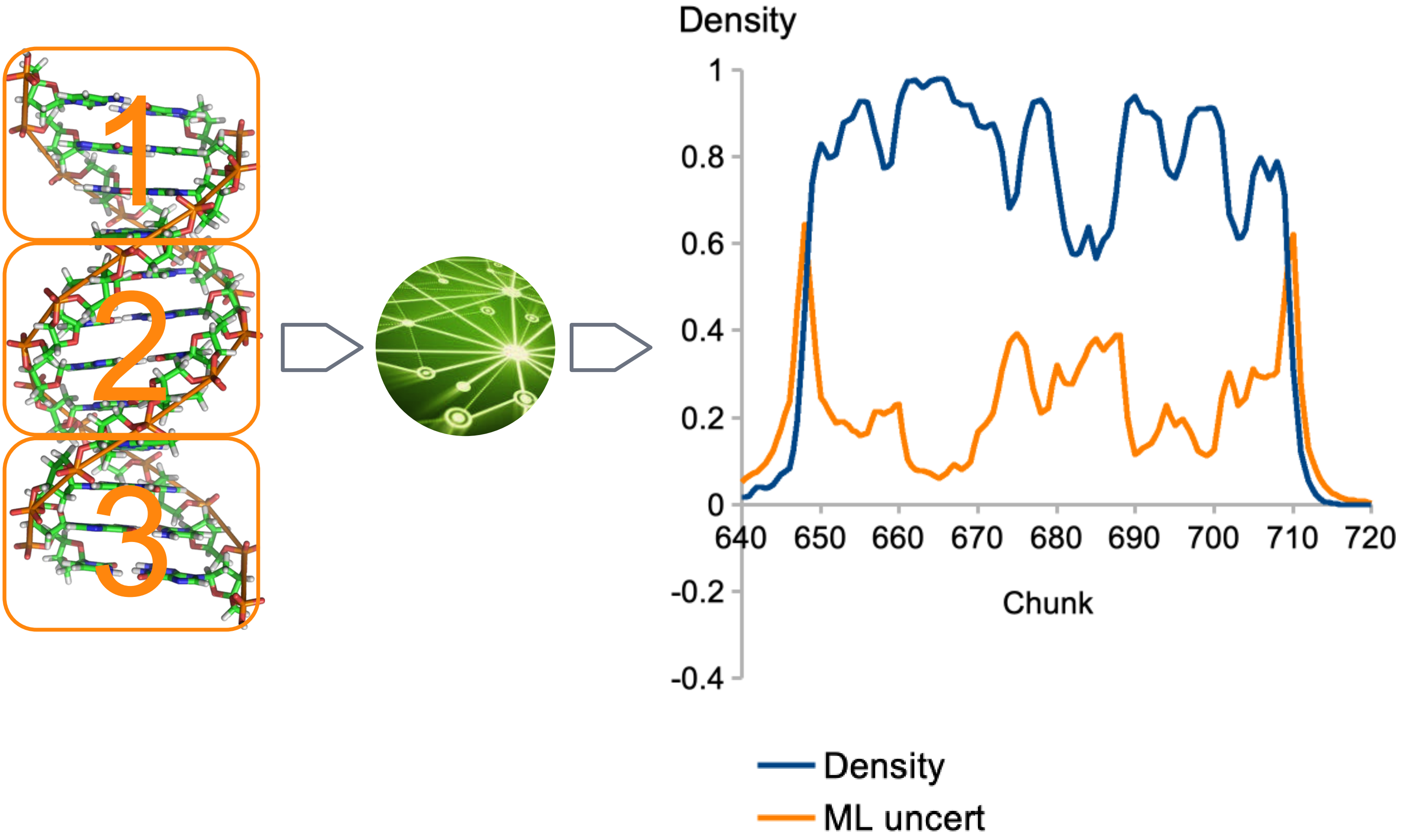
Split into chunks



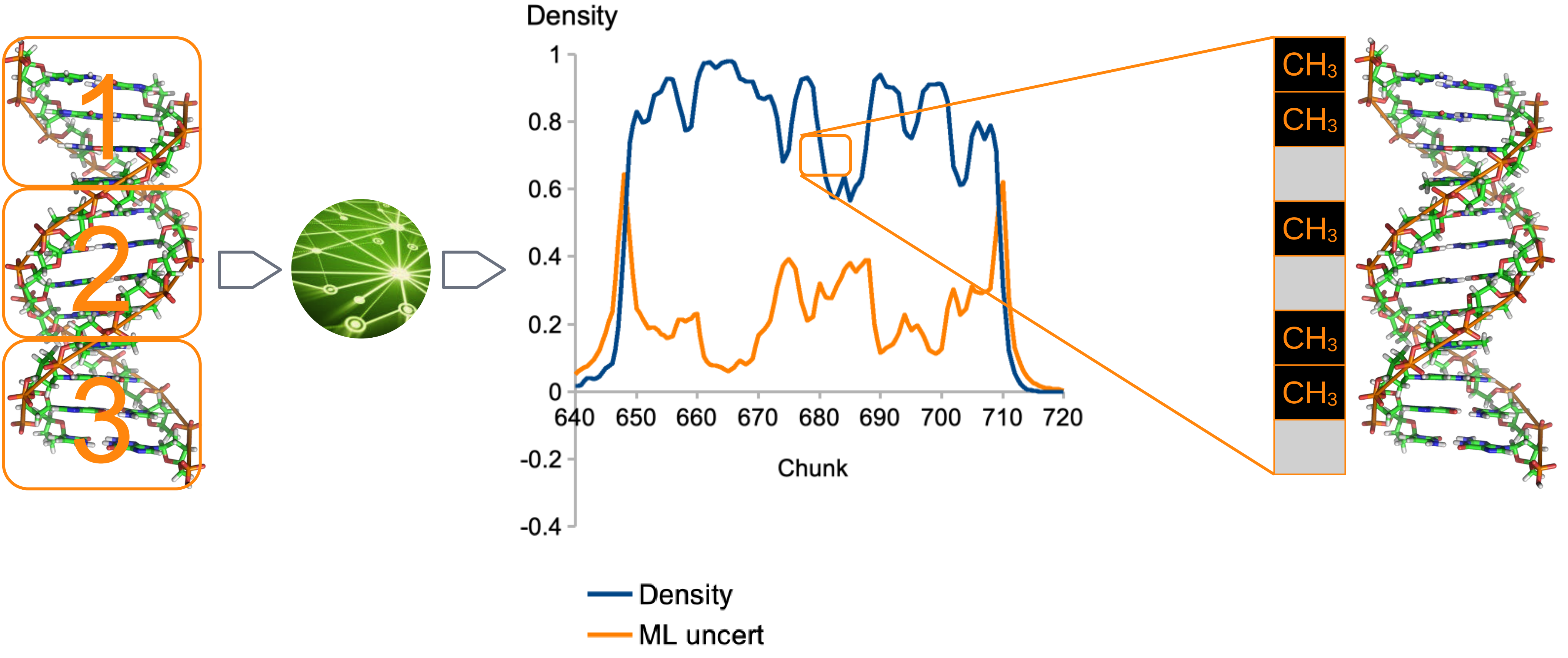
Train a model for methylation density of the chunks



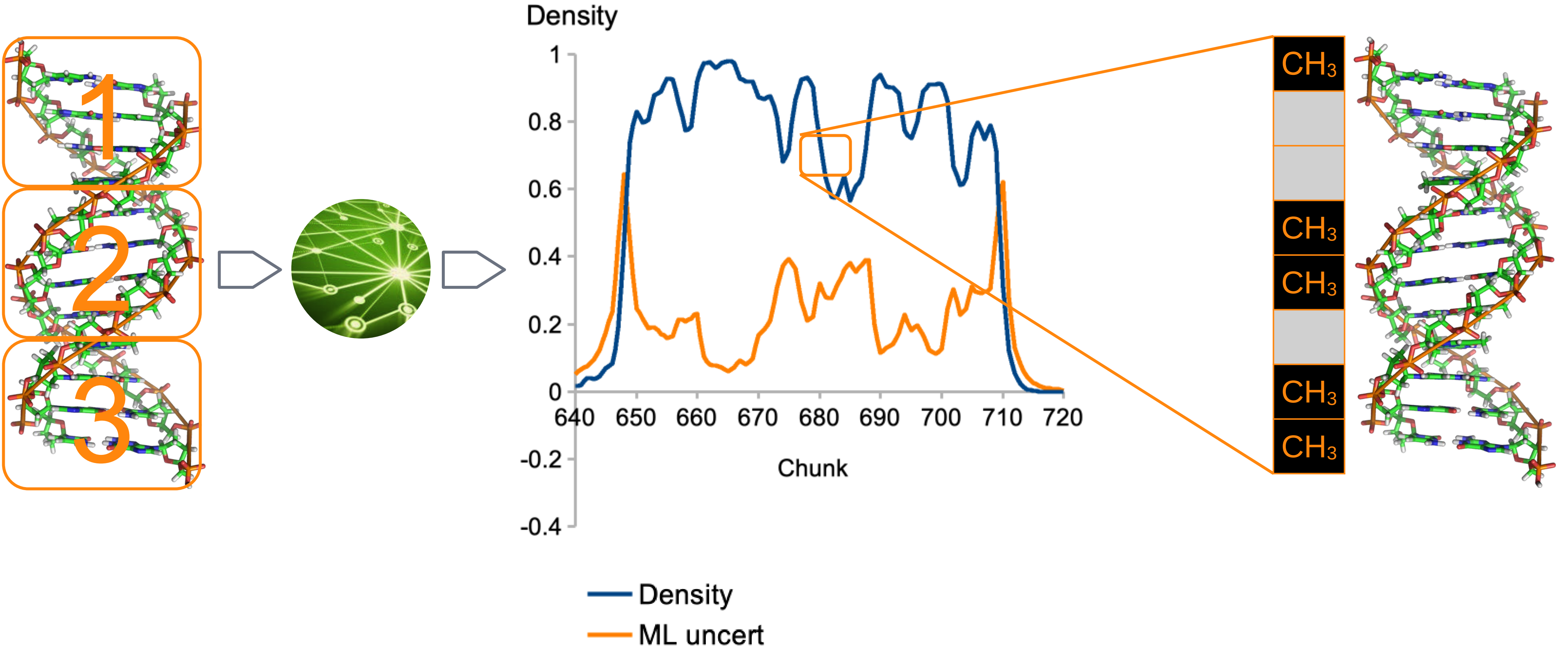
Model for chunk methylation density has uncertainty



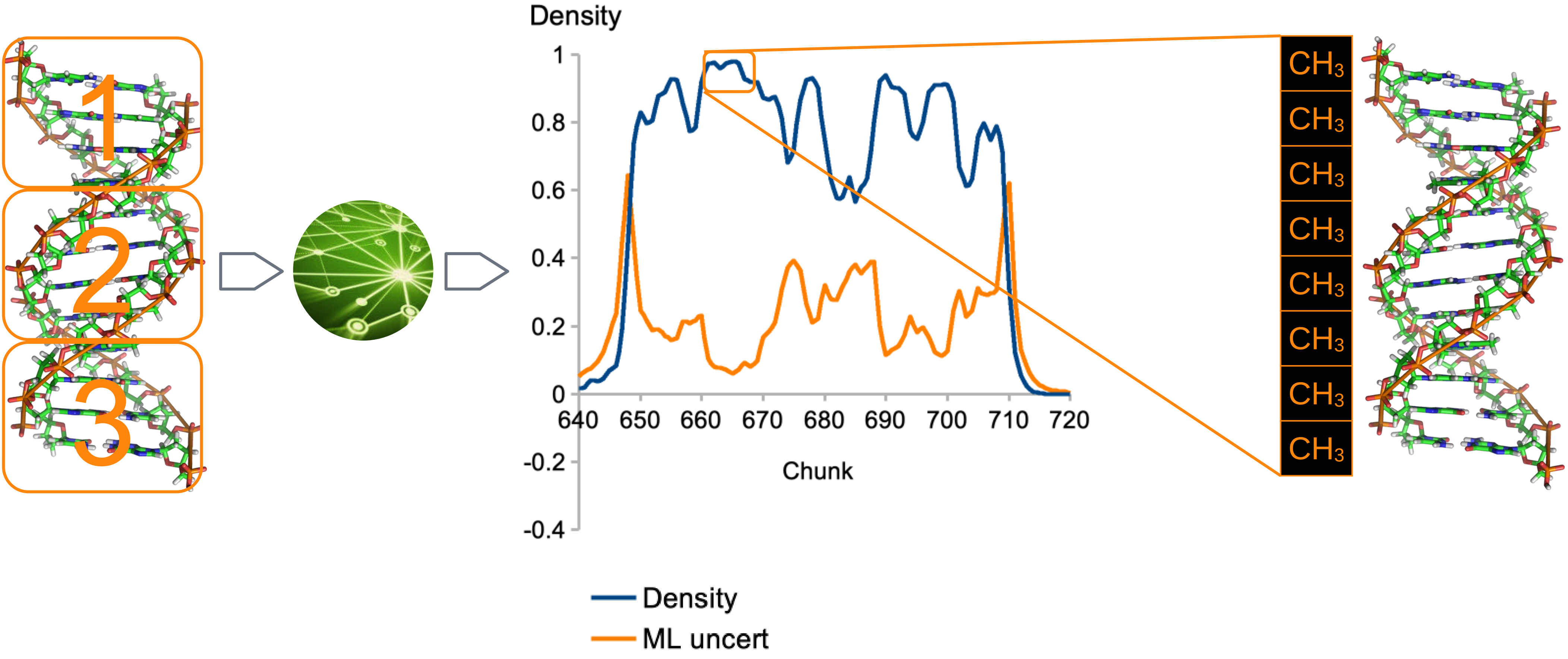
Uncertainty is natural owing to permutations of methylation sites



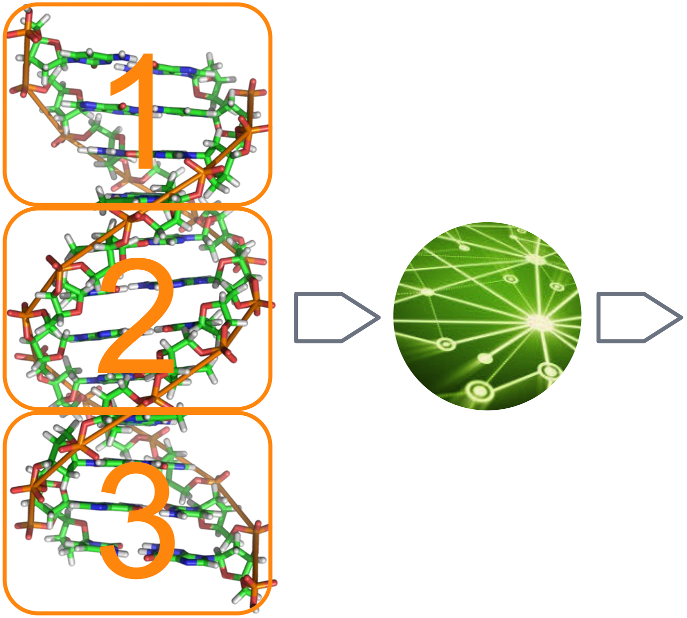
Many permutations are possible

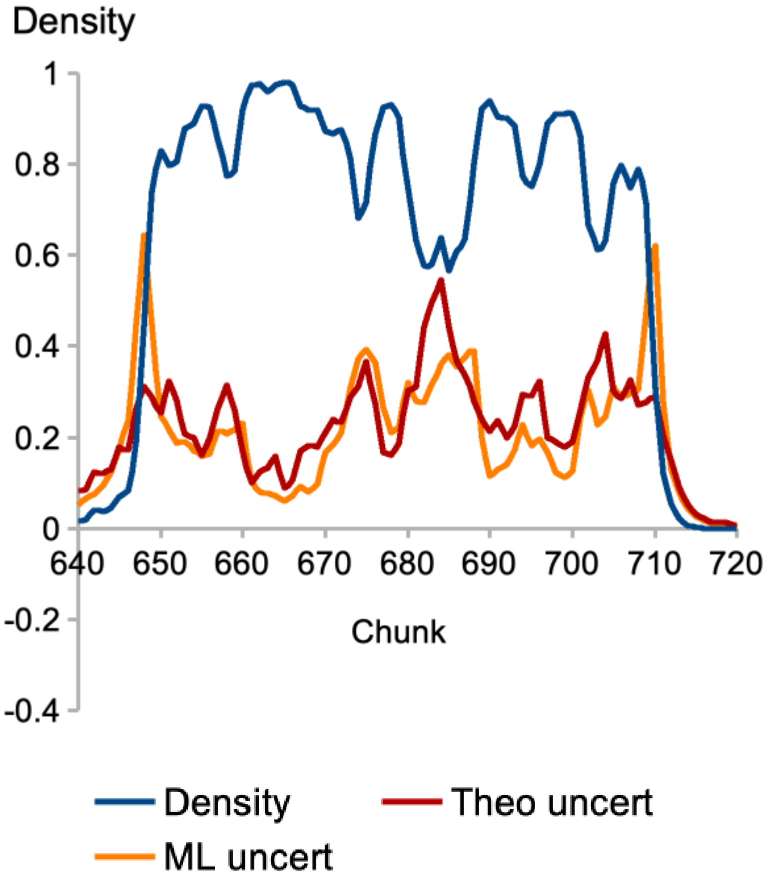


No uncertainty when either fully or not methylated



For each chunk train model for methylation density and uncertainty

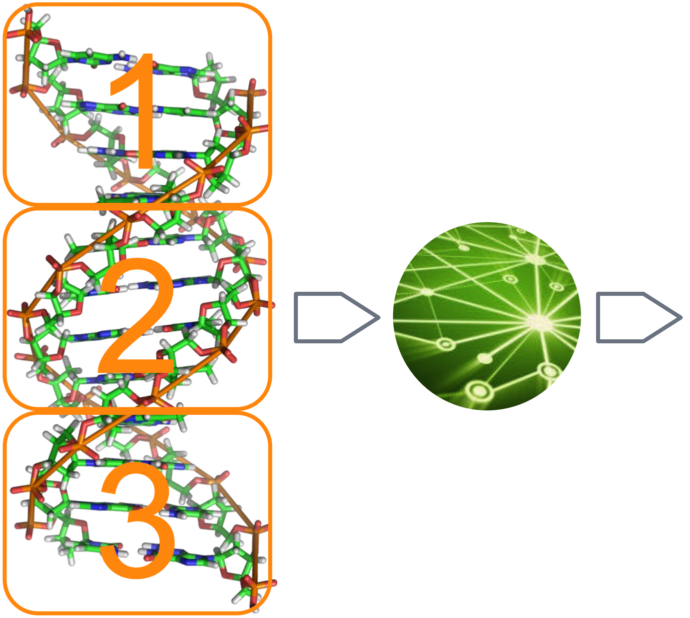


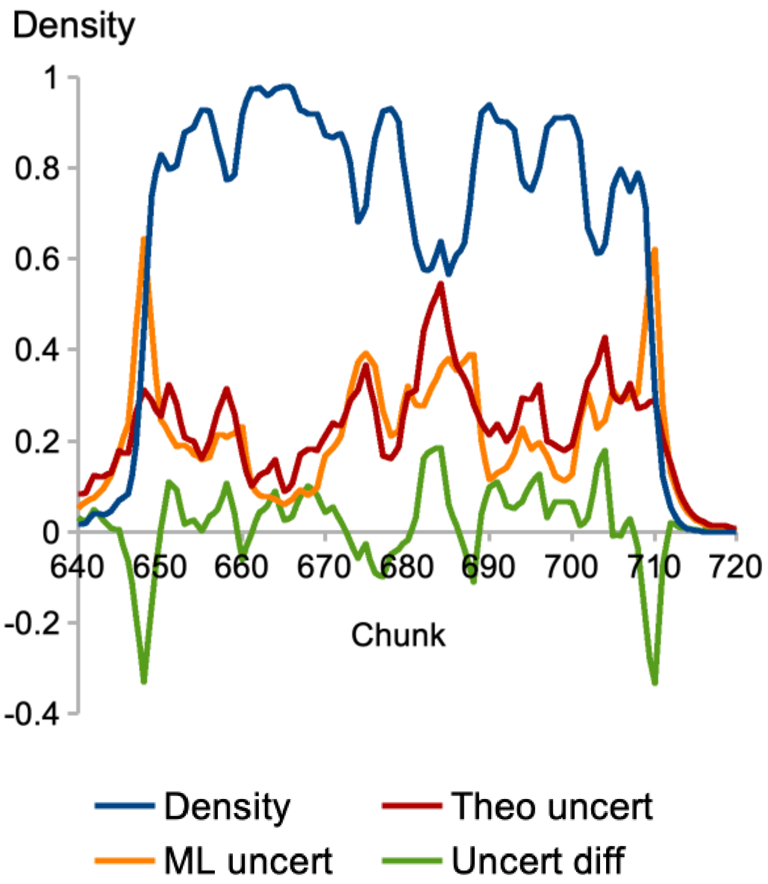


Density follows a binomial distribution so its inherent uncertainty is

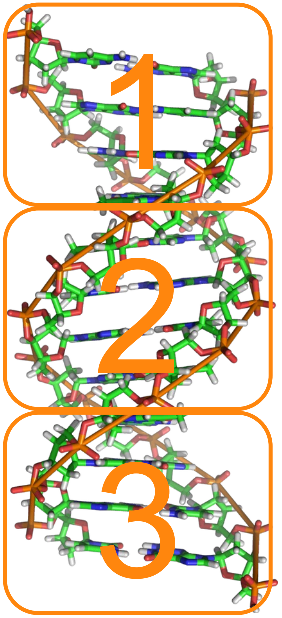
$$\sqrt{\rho(1-\rho)}$$

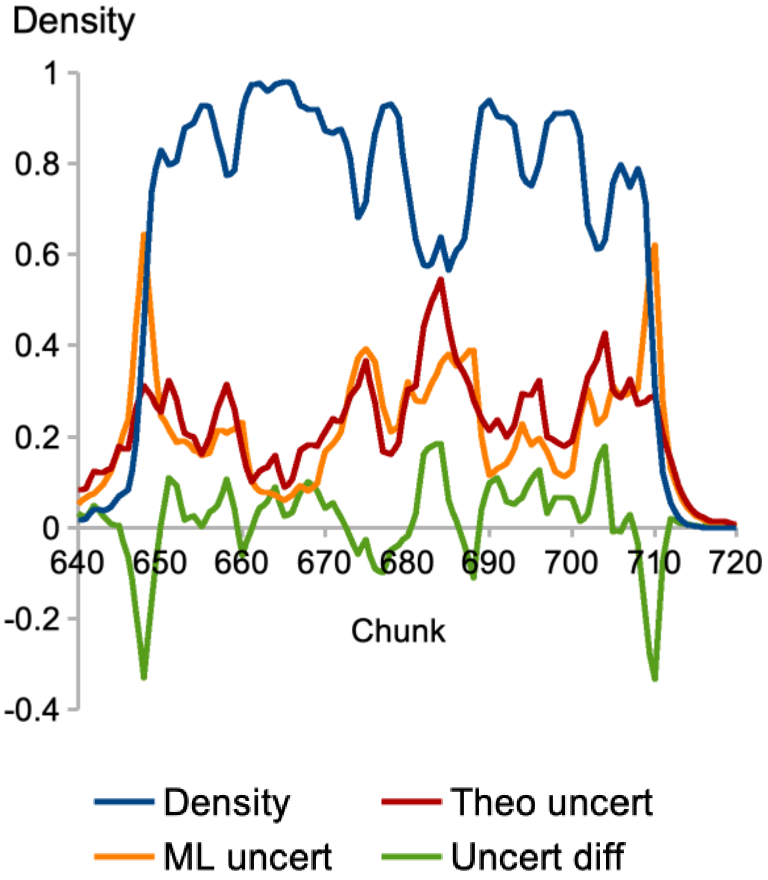
Extract the unexplained chaotic methylation due to rogue biology





Identify the critical chunk driving the emergence of cancer

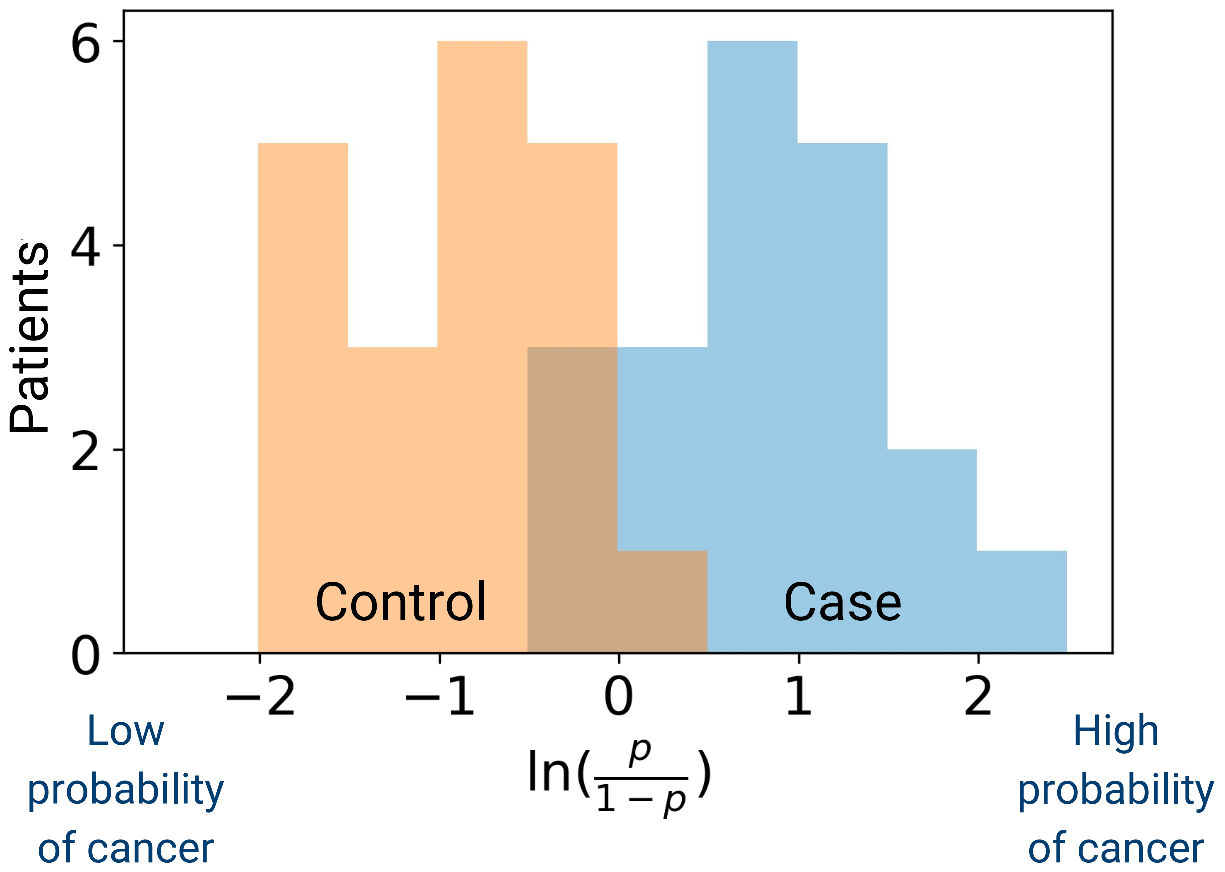




Critical uncertainty threshold for
Chr 3
Pos 31801200

Probability cancerous in one year

90% accuracy on 40 patients available during study



Blind test on five more patients

Patient	Probability cancerous	Later outcome
19317_U017	0.946	Case
19316_U007	0.451	Control
19317_U015	0.966	Case
19317_U018	0.449	Control
19316_U012	0.934	Case

Next steps

Further blind tests with data from US collaborators

Explore science behind identified critical nucleobase

chromosome 3, positions 31800700-31801700

Develop machine learning formalism that can extract information from noise itself

Design and experimentally verify two concrete mixes

Exploit uncertainty to predict emergence of cancer

Generic approach applicable across the sciences