Evaluating and interpreting uncertainty in QSAR models for toxicology

Charles Gong

Goodman Group / Yusuf Hamied Department of Chemistry

Sulfanilamide tragedy

U. S. Races Death to Save 700 From Elixir

Recovery of Pint Bottles Sold to Patients Goal as Deaths From Poison Reach 36

By Associated Press

CHICAGO, ILL., October 24.—A nation-wide race with death, its object recovery of more than 700 bottles, mostly pints, of a new liquid medicine, named Elixir of Sulfanilamide, which has already caused 36 verified deaths, was described today at the headquarters here of the American Medical Association.

Every agent of the United States Food and Drug Association, said Dr. Morris Fishbein, spokesman of the Medical Association, is scouring the country to recover the bottles. By some time on Monday, said J. O. Clarke of the Food and Drug Administration, it is hoped that all of the "outstanding" shipments will be recovered.

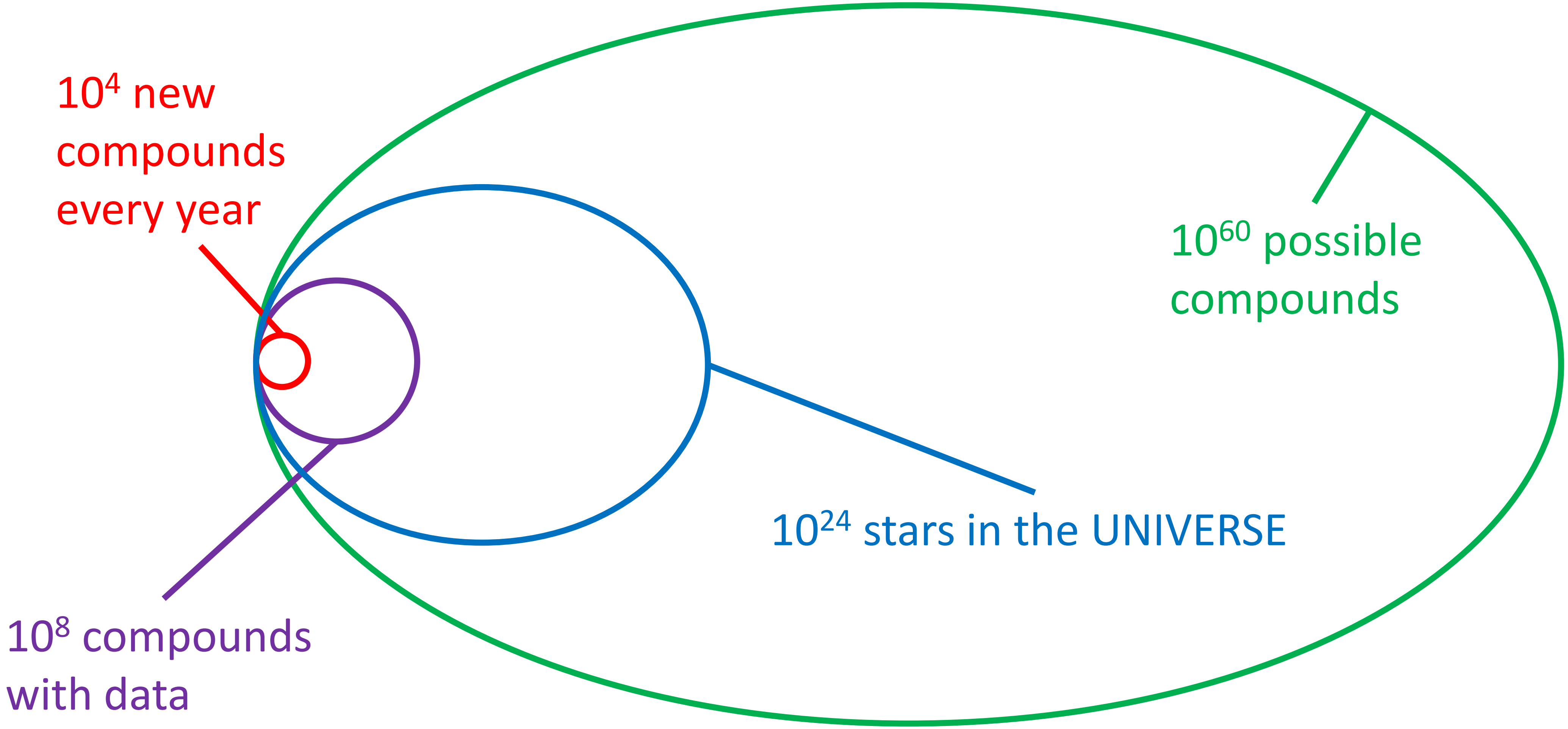
[illegible] lies in the fact that large dosages are customary with sulfanilamide preparations.

The medicine stops the kidneys. At Medical Association headquarters its effects were said to be like those of bichloride of mercury. No antidote is known yet. The "Elixir" is made of sulfanilamide and diethylene glycol, a near relative of auto antifreeze fluid. The diethylene glycol is blamed for the lethal effect, although not itself classed as a poison.

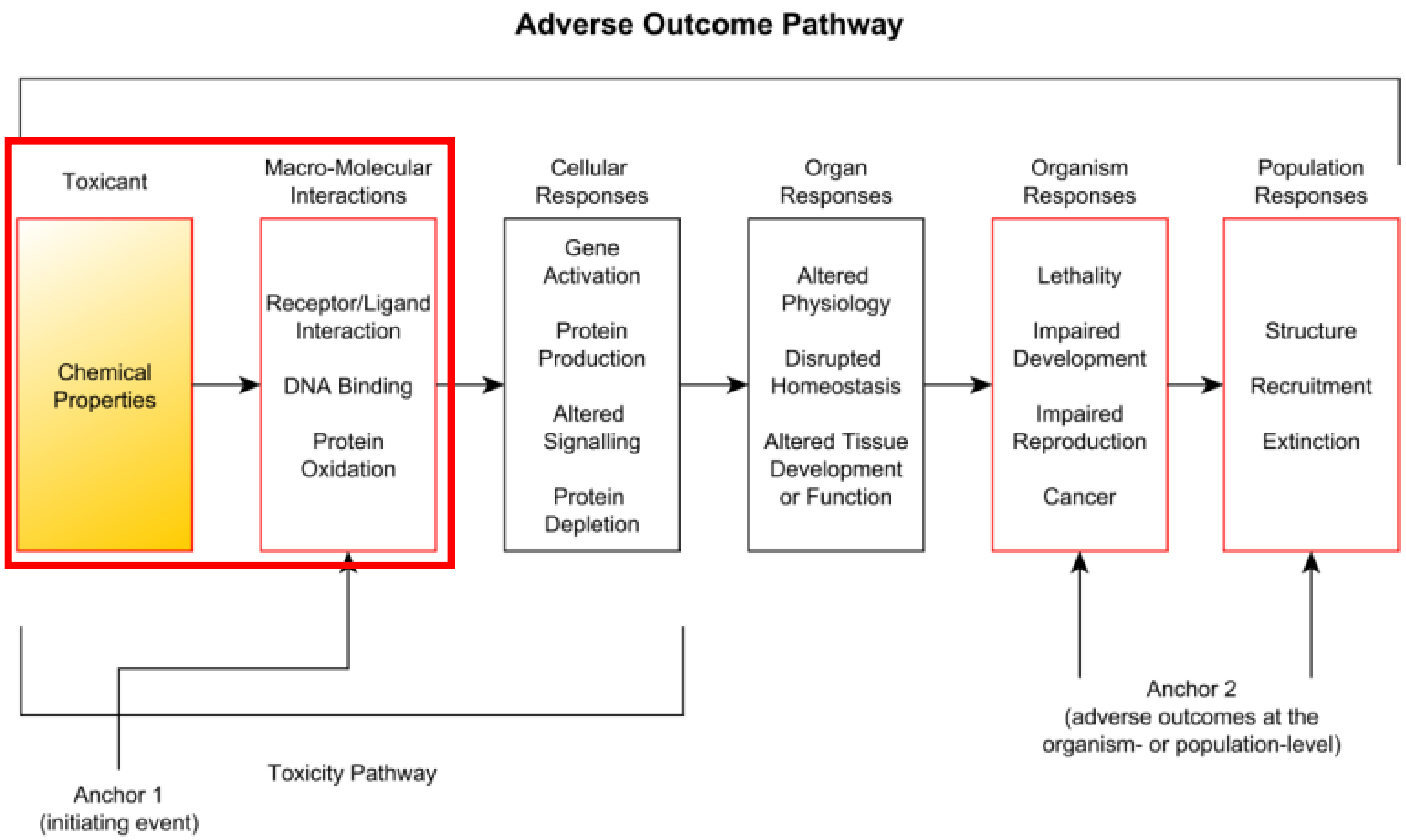
The principal shipments, said Mr. Clarke, went to the South and Midwest. But in addition consignments have been traced to the Northern Peninsula of Michigan, to a distributing house in New York City and to another in San Francisco. An

Animal testing

Chemical space

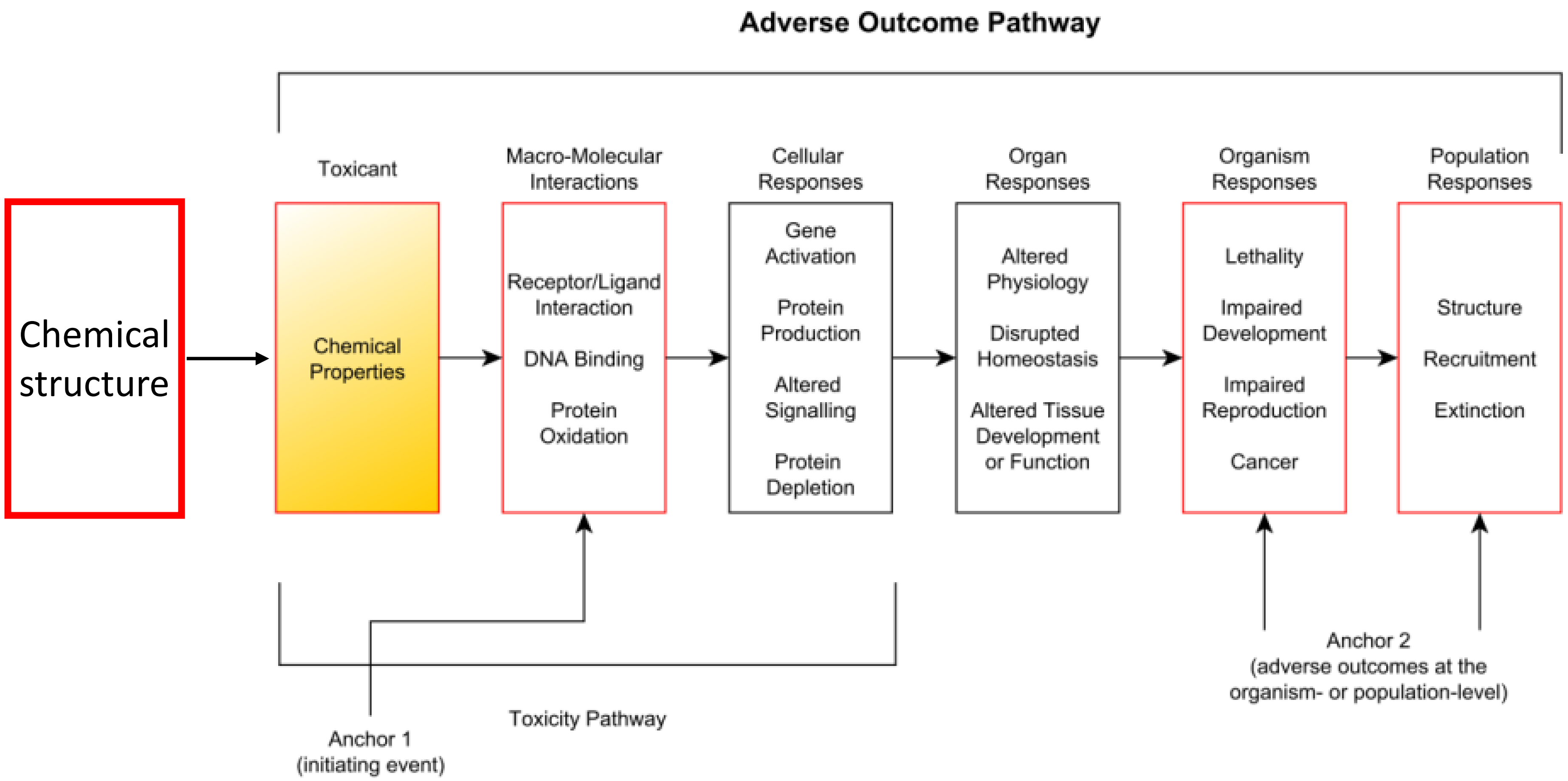


Adverse Outcome Pathway

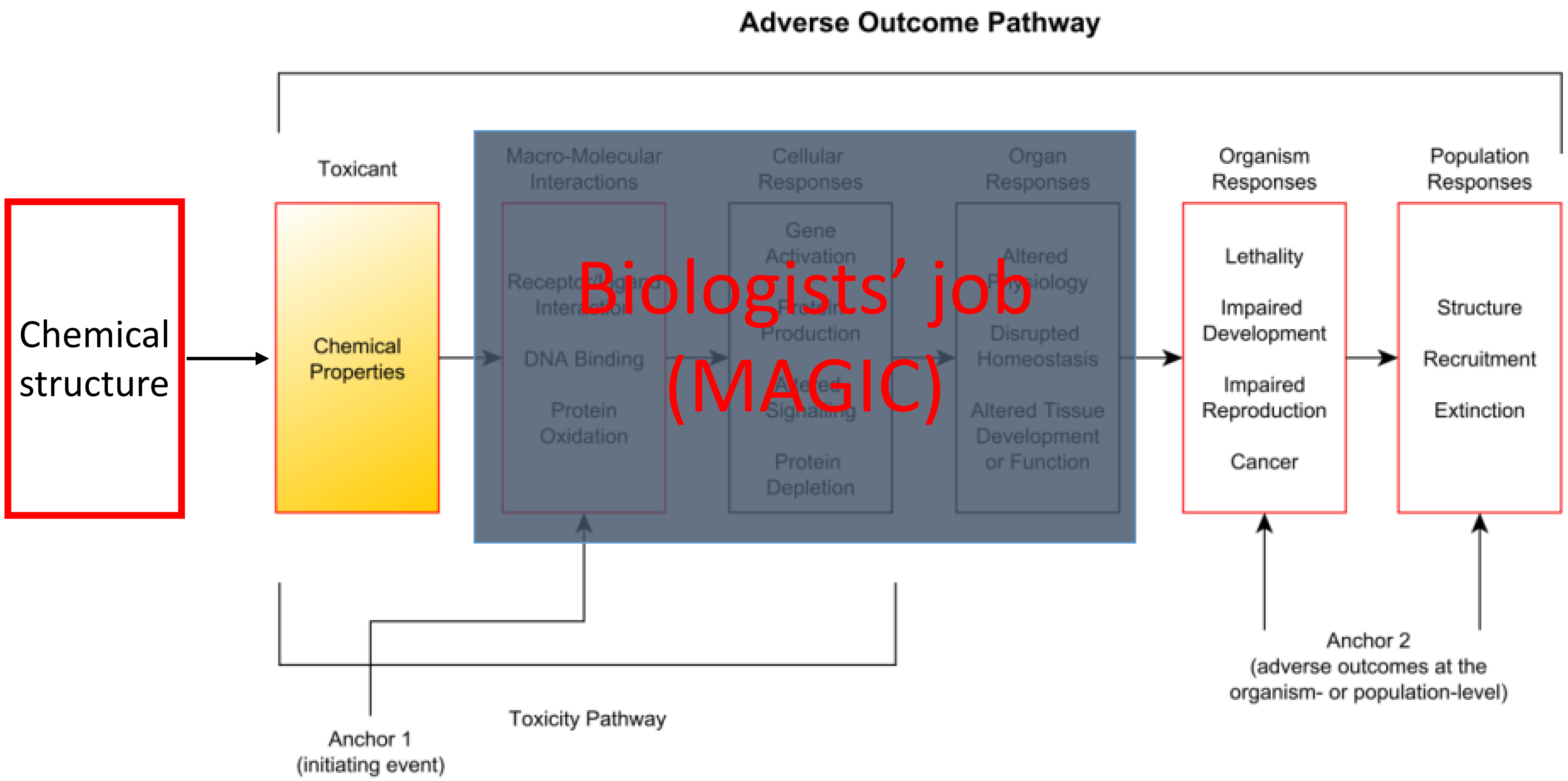


Structure – Activity Relationships

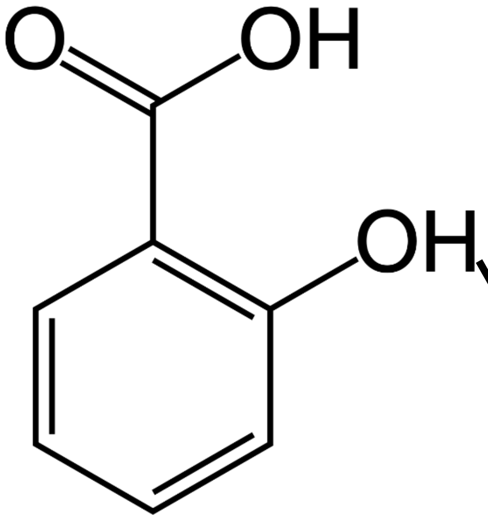
Adverse Outcome Pathway



Adverse Outcome Pathway



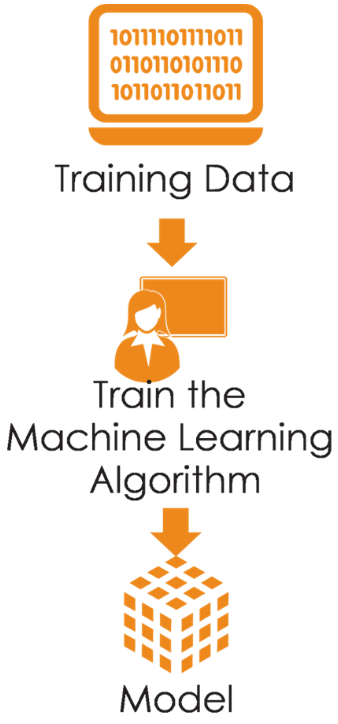
Adverse Outcome Pathway (idealized)



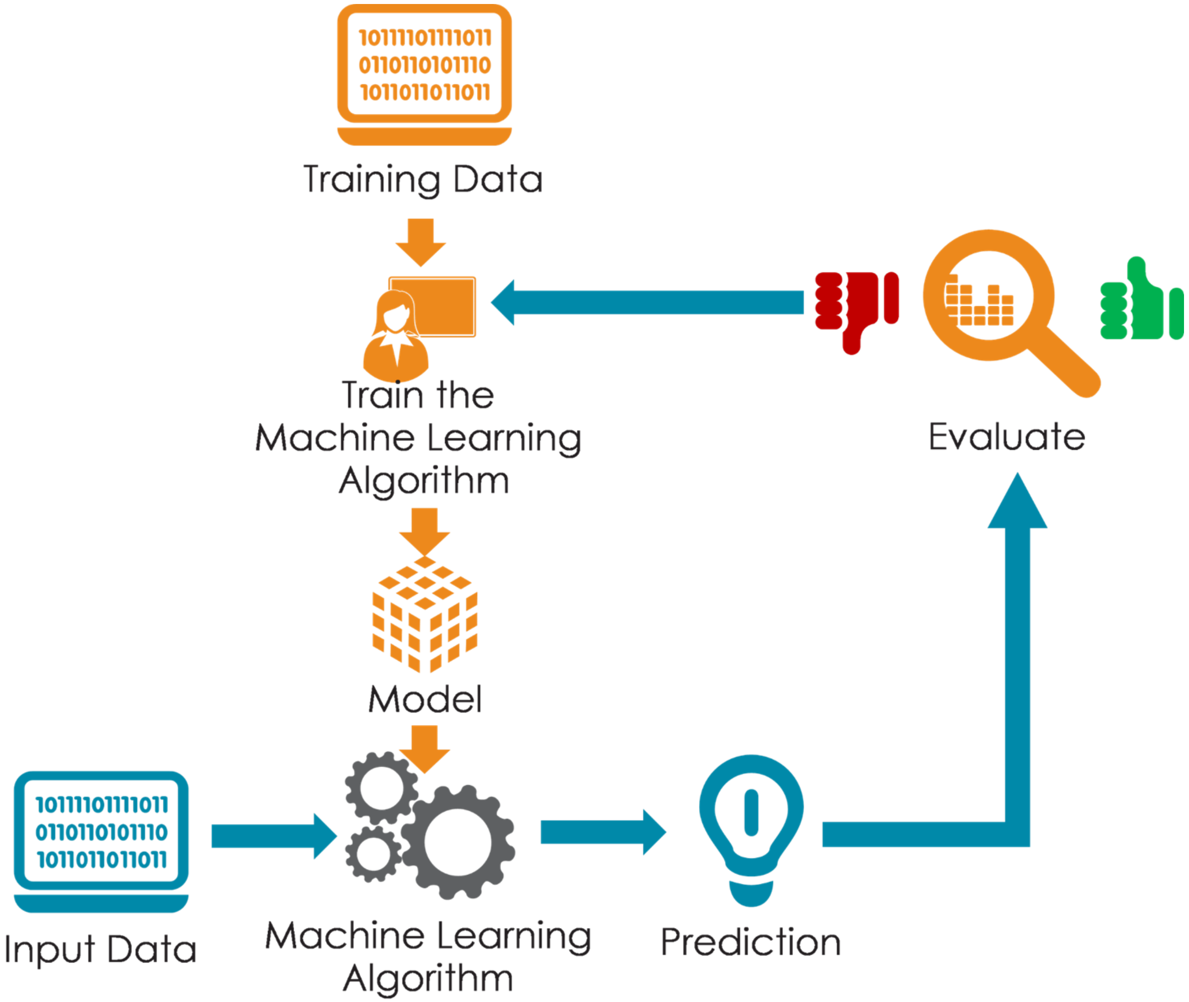
Chemical structure → Toxicological profile

carcinogenicity,
liver toxicity,
reproductive toxicity

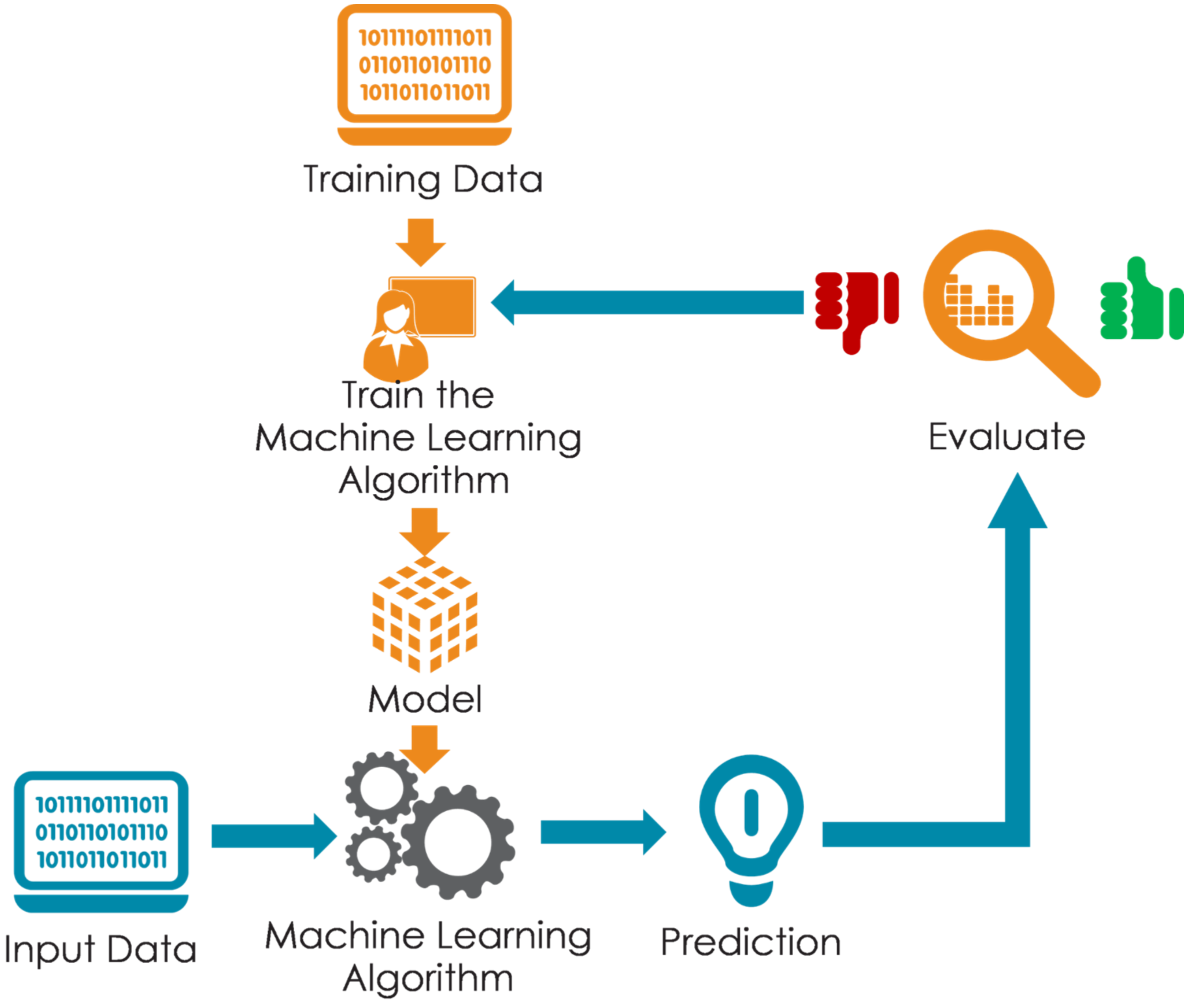
Model building cycle



Model building cycle

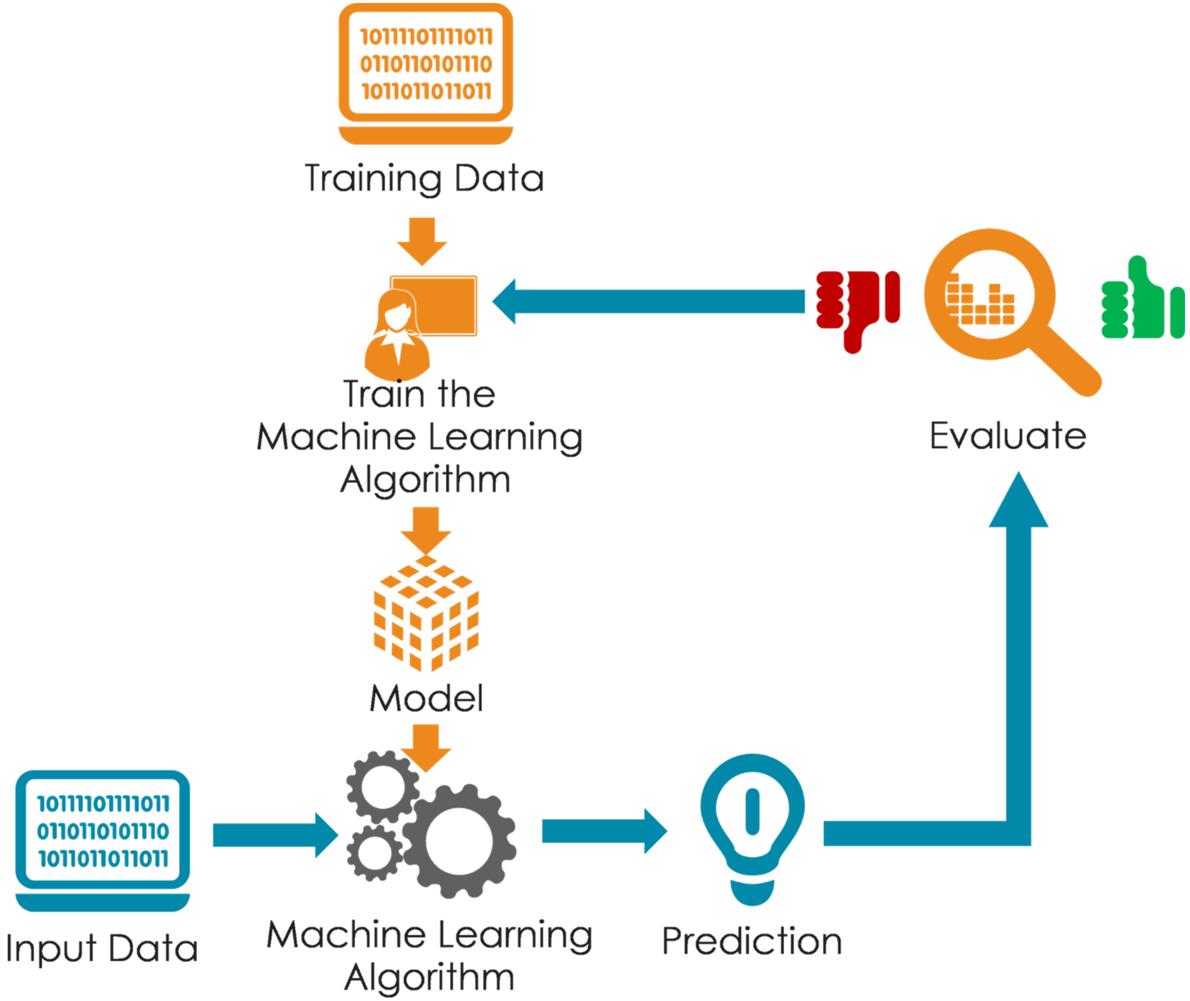


Model building cycle



Validation

Motivation



But how much can you trust the prediction?

pIC50 = 6 ± 0.1

vs

pIC50 = 6 ± 1.0

Contents

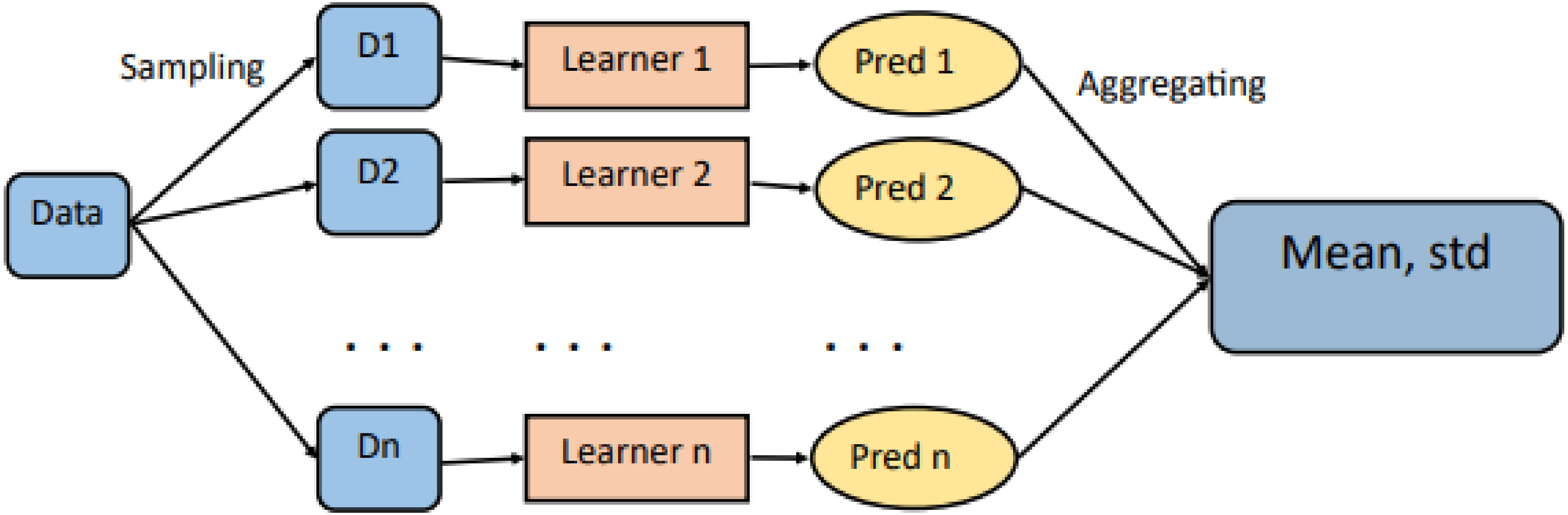
1. How we introduce uncertainty to models

2. How we know our uncertainty is correct

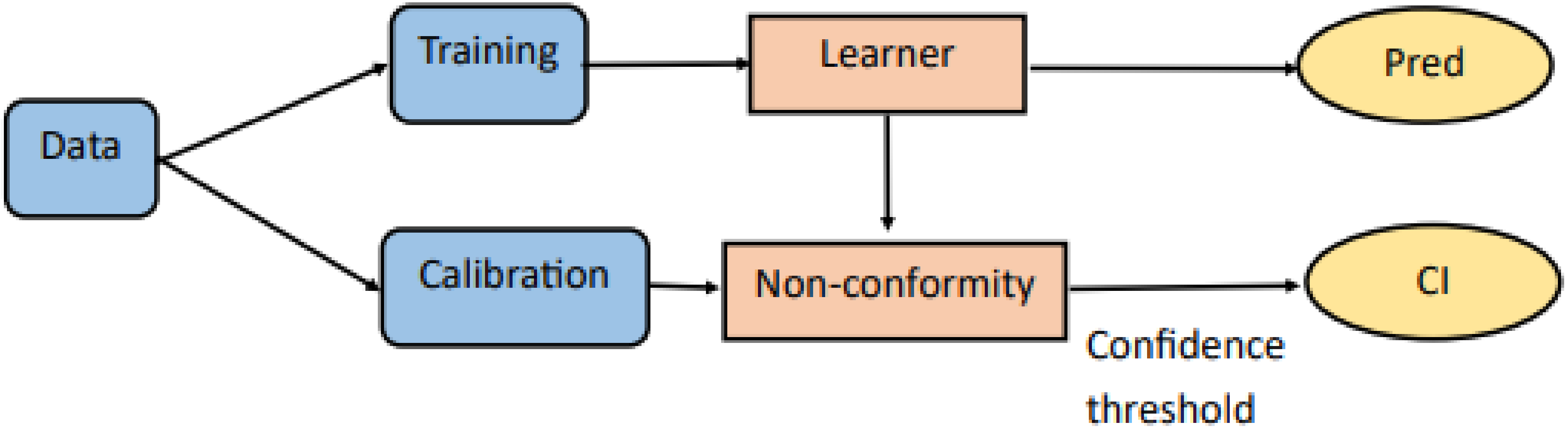
3. How to interpret uncertainty

Contents

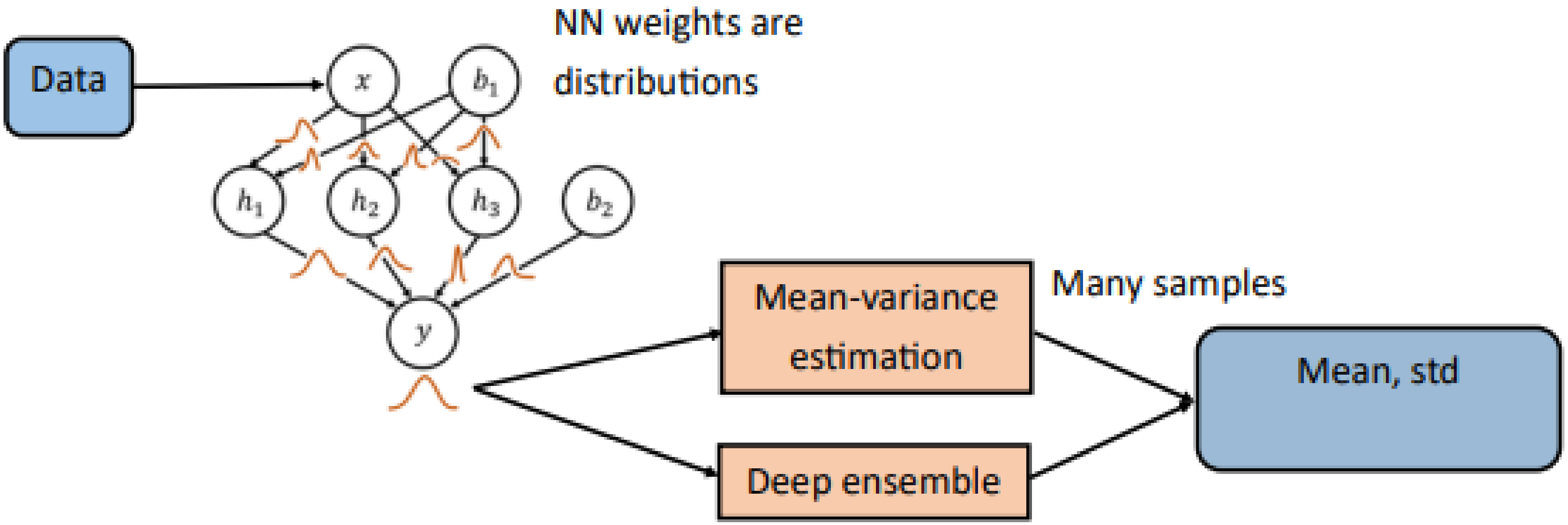
Bootstrapping



Conformal prediction



Bayesian neural network



Contents

1. How we introduce uncertainty to models
2. How we know our uncertainty is correct
3. How to interpret uncertainty

Calibration

We want models to:

Return range of values Y±X, such that out of all predictions of this range the true value lies within this range $\boxed{95\%}$ of the time.
[confidence interval can be changed]

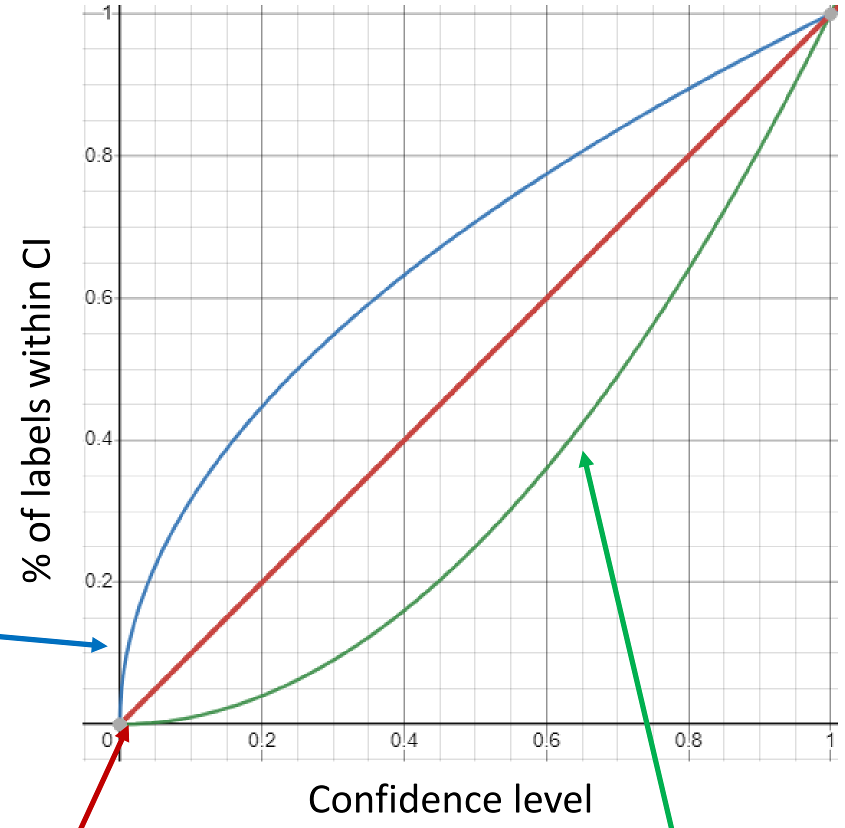
A model that does this for every choice of confidence interval is called "well-calibrated"

Calibration curve

X axis: confidence level
Y axis: % of labels within given CI

Metric: calibration curve R^2 score

Confidence intervals too wide



Well-calibrated

Confidence intervals too narrow

Efficiency

We want models to:

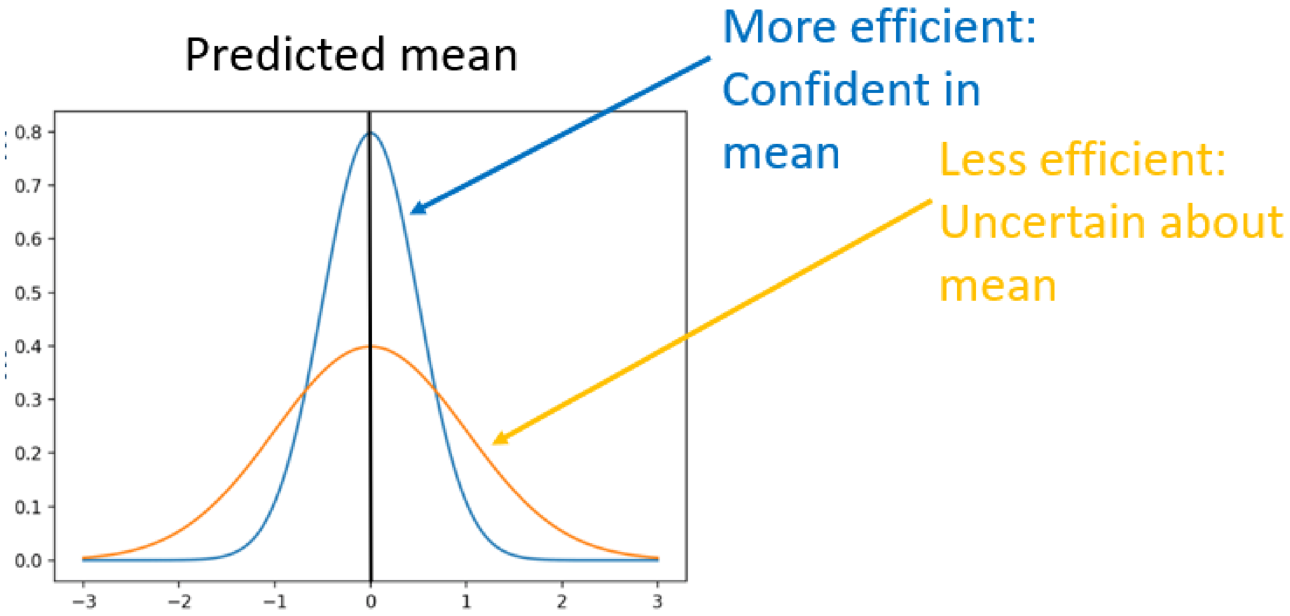
Return range of values Y±X, such that out of all predictions of this range the true value lies within this range 95% of the time.
[confidence interval can be changed]

* And we also want this range to be as small as possible
Mean of the predicted standard deviations is <u>efficiency</u>

Efficiency

And we also want this range to be as small as possible

Mean size of the confidence interval is called <u>efficiency</u>



Distribution for <u>a single prediction</u>

Dispersion

We want models to:

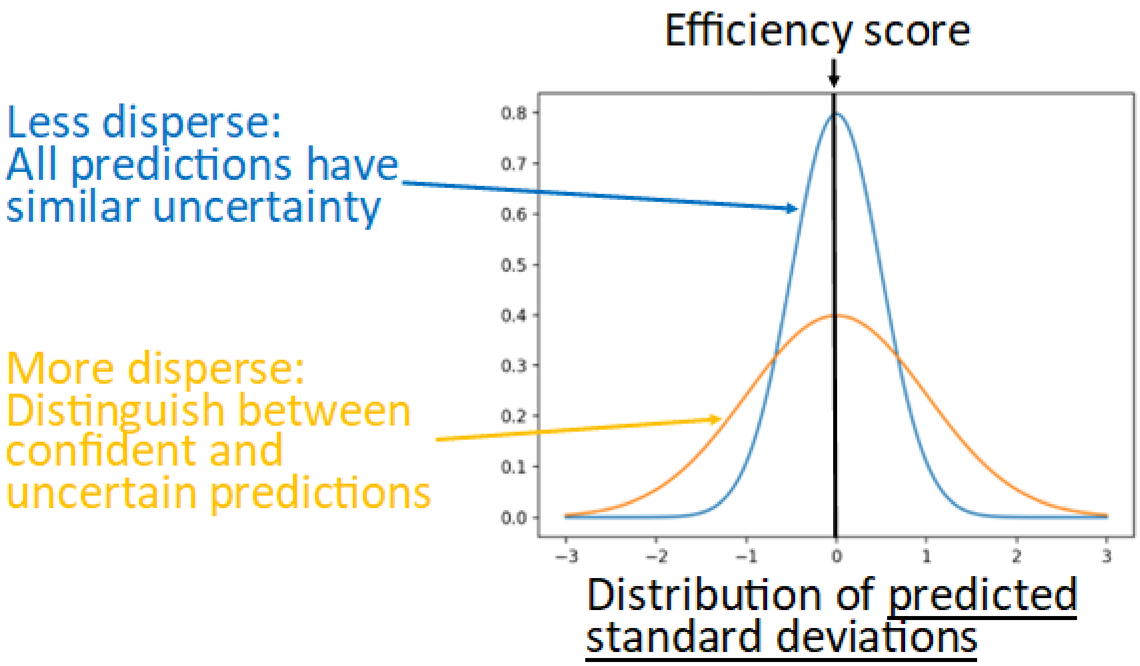
Return range of values Y±X, such that out of all predictions of this range the true value lies within this range 95% of the time.
[confidence interval can be changed]

* And we want each molecule to give a different range
Standard deviation of the ranges is called dispersion

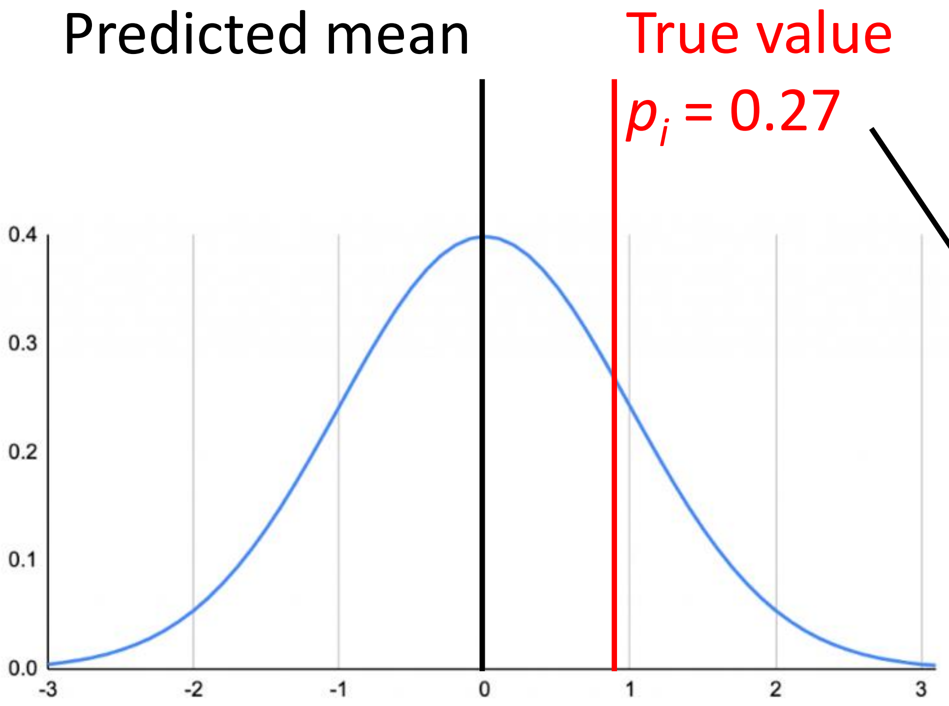
Dispersion

And we want each molecule to give a different range

Standard deviation of the ranges is called dispersion



Geometric mean of probabilities (GMP)



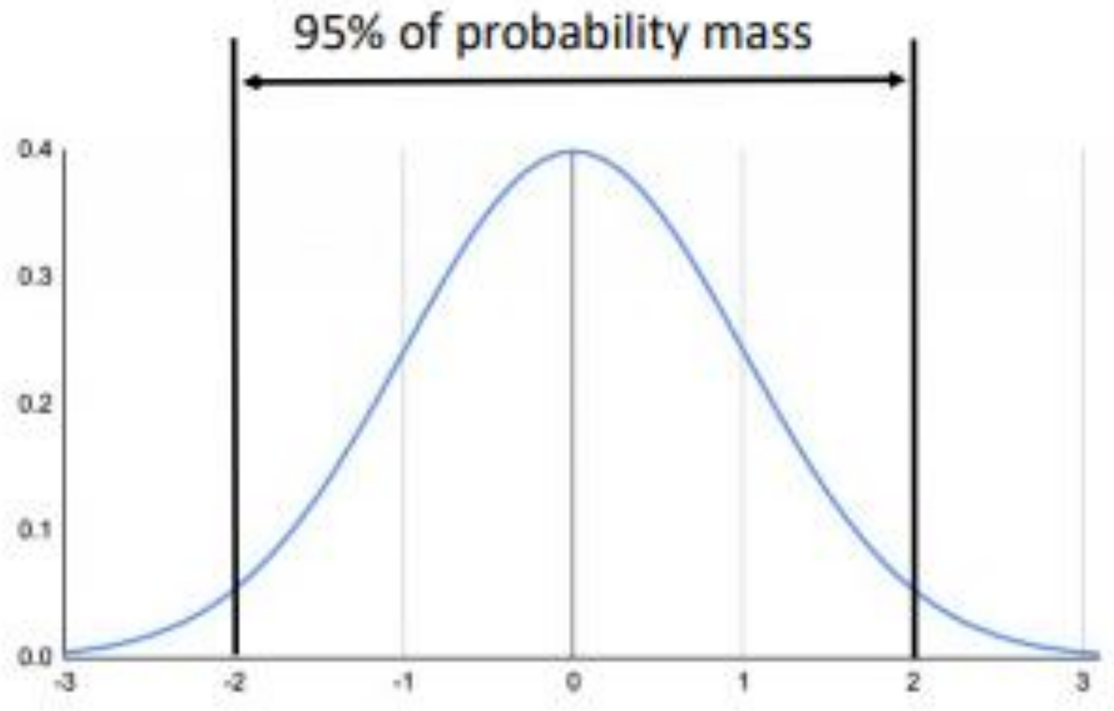
Proper scoring rule analogous to NLL that has units of probability.

When error is low, rewards confidence
When error is high, rewards uncertainty

$$GMP = \left(\prod_{i=1}^{n} p_i\right)^{\frac{1}{n}} = \sqrt[n]{p_1 p_2 \cdots p_n}$$

Contents

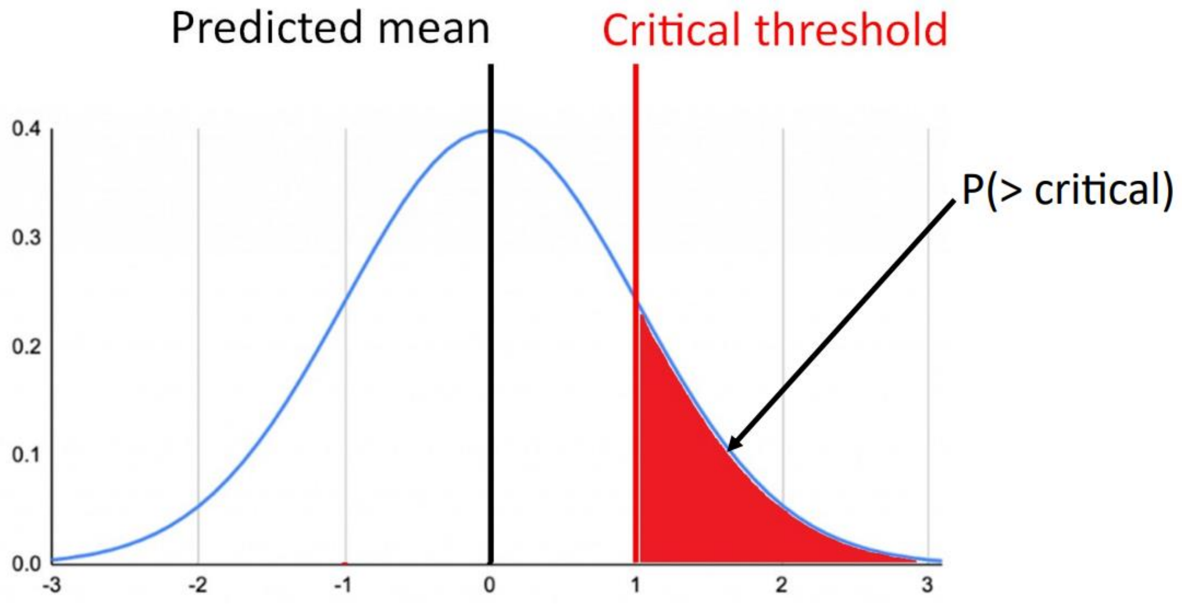
Confidence intervals



Given model is **well-calibrated**:

Can define **confidence interval,** where there is X% chance that the true value lies within this range

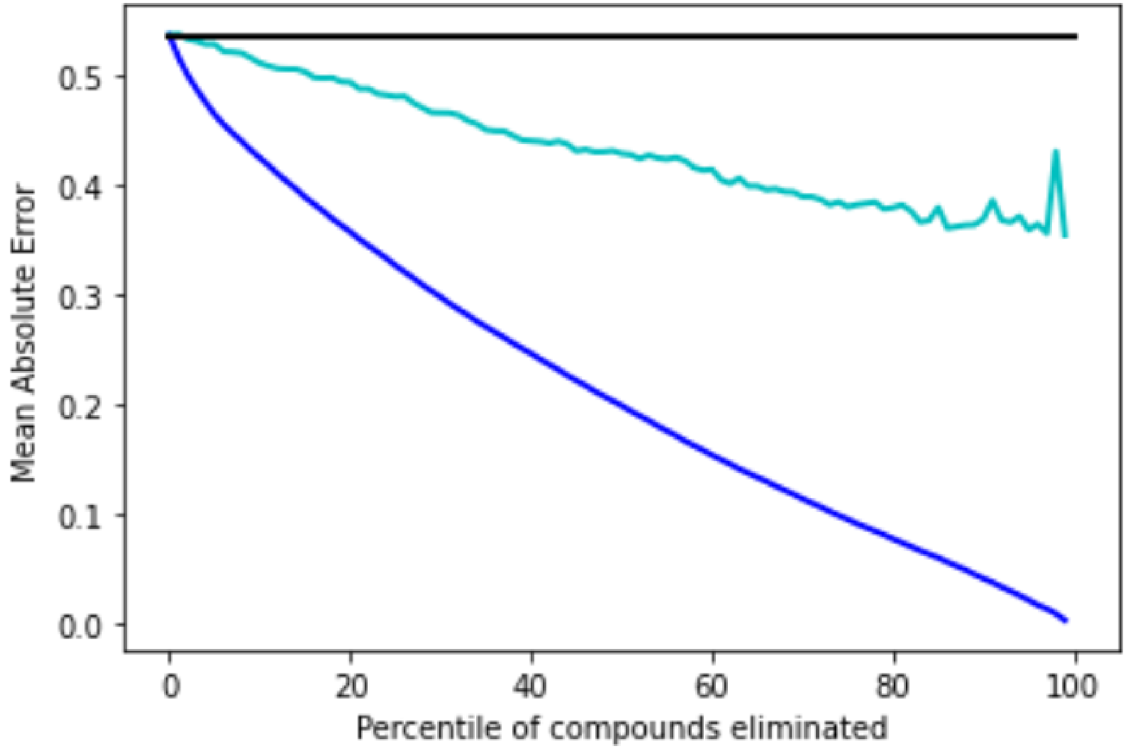
Probability of exceeding threshold



If interested in a critical threshold:

Can find P(> critical) which is more meaningful than simply comparing the mean to the critical threshold

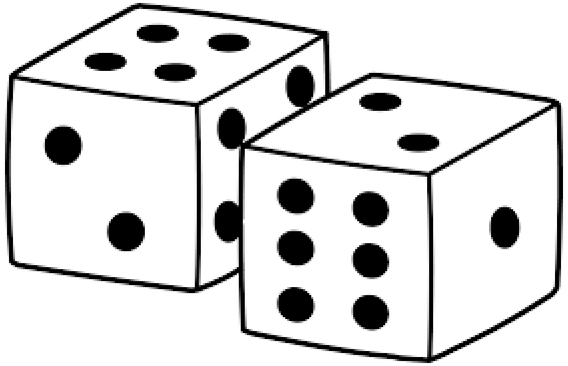
Uncertainty vs Mean Error



Dark blue (oracle):
Remove compounds with the highest **error**

Light blue (uncertainty):
Remove compounds with the highest **uncertainty**

Aleatoric vs Epistemic uncertainty



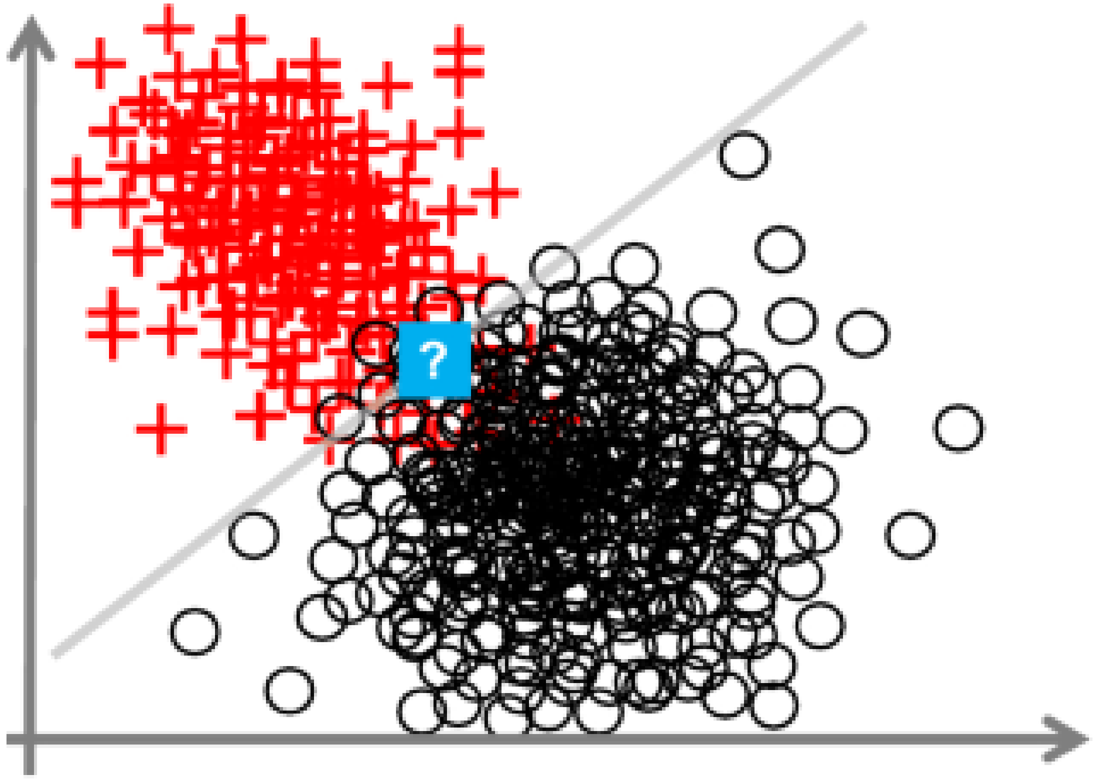
Alea: Dice (Latin)

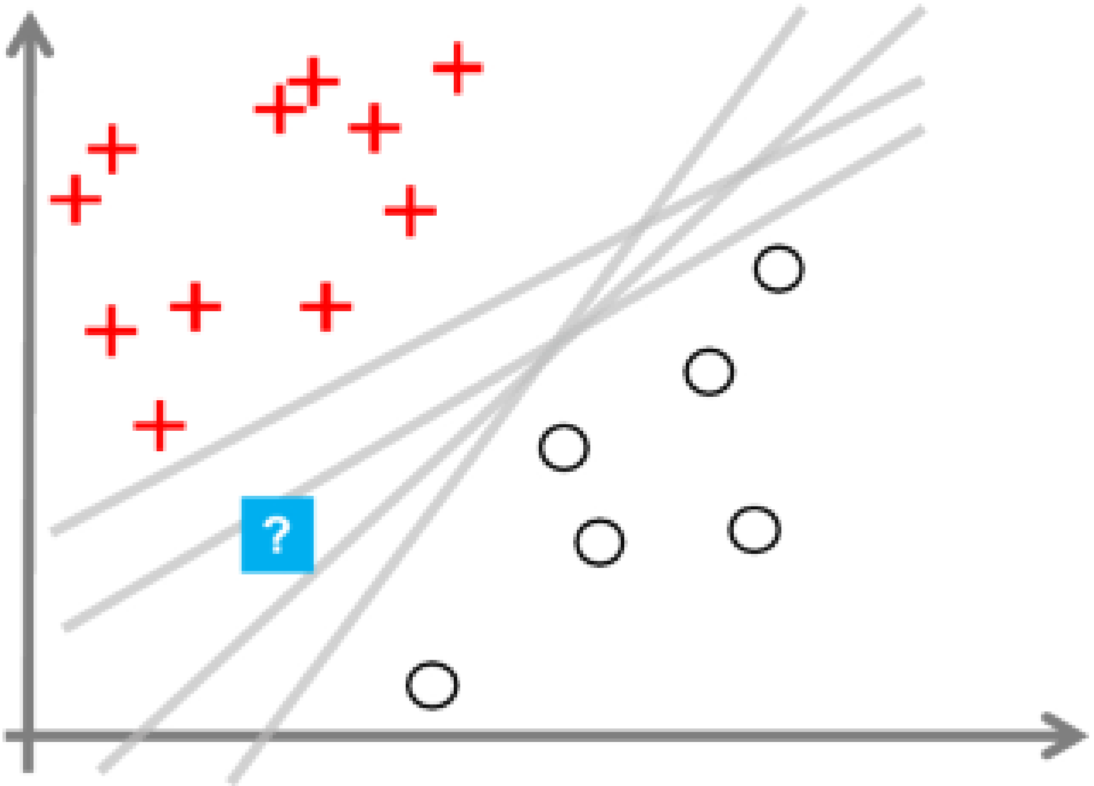
- Aleatoric uncertainty represents randomness inherent in the model
- Additional knowledge cannot eliminate it

Episteme: Knowledge (Greek)

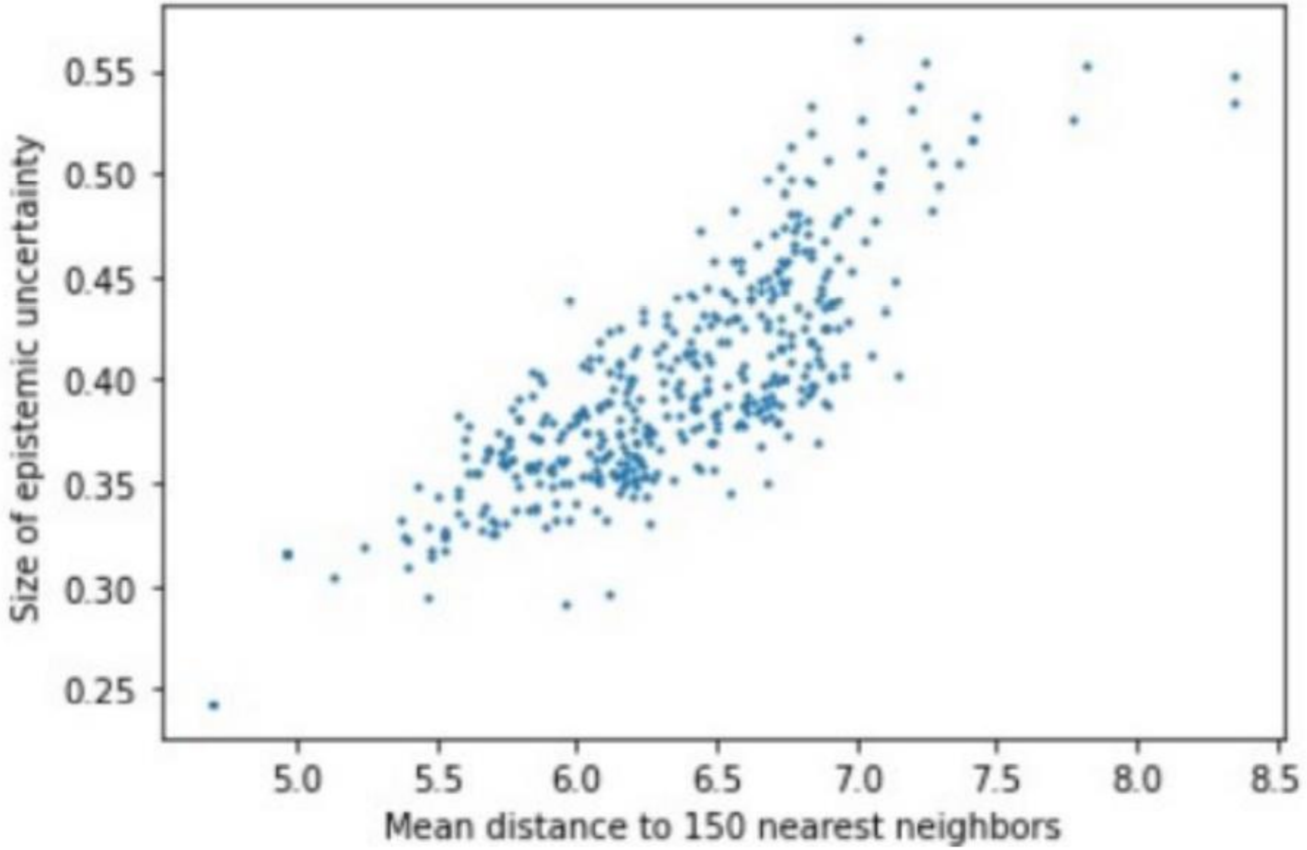
- Epistemic uncertainty represents lack of knowledge about the system
- Can be overcome with additional data and learning

Applicability domain is like epistemic uncertainty



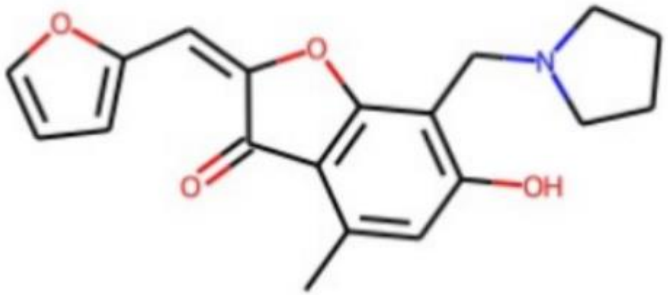


Epistemic uncertainty and neighbour density



Epistemic uncertainty is related to neighbour density and hence applicability domain

Case study

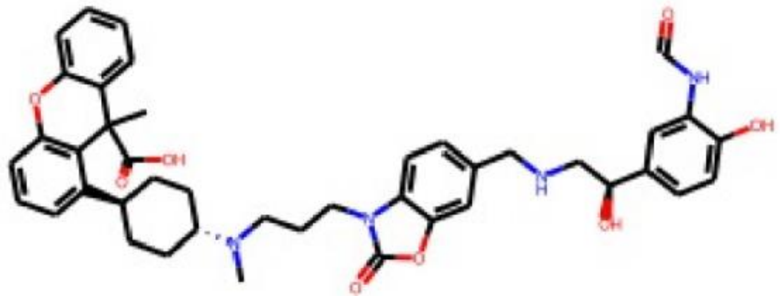


A

SMILES: Cc1cc(O)c(CN2CCCC2)c2c1C(=O)/C(=C\c1ccco1)O2

Predicted mean: **[4.54]**

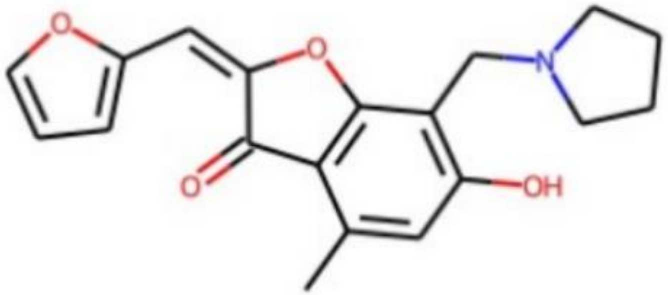
B



SMILES: CN(CCCn1c(=O)oc2cc(CNC[C@H](O)c3ccc(O)c(NC=O)c3)ccc21)

Predicted mean: **[6.28]**

Case study

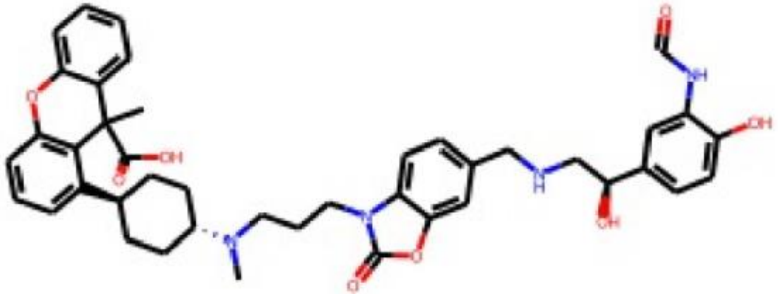


A

SMILES: Cc1cc(O)c(CN2CCCC2)c2c1C(=O)/C(=C\c1ccco1)O2

Predicted mean:	**[4.54]**
95% CI:	**[3.12] — [5.96]**
Epistemic uncertainty:	**[0.244]**
P(> 5):	**[0.255]**

B



SMILES: CN(CCCn1c(=O)oc2cc(CNC[C@H](O)c3ccc(O)c(NC=O)c3)ccc21)

Predicted mean:	**[6.28]**
95% CI:	**[4.53] — [8.04]**
Epistemic uncertainty:	**[0.574]**
P(> 5):	**[0.929]**

Conclusions

1. Uncertainty can allow us to better interpret model predictions

2. Need metrics to determine the quality of uncertainty

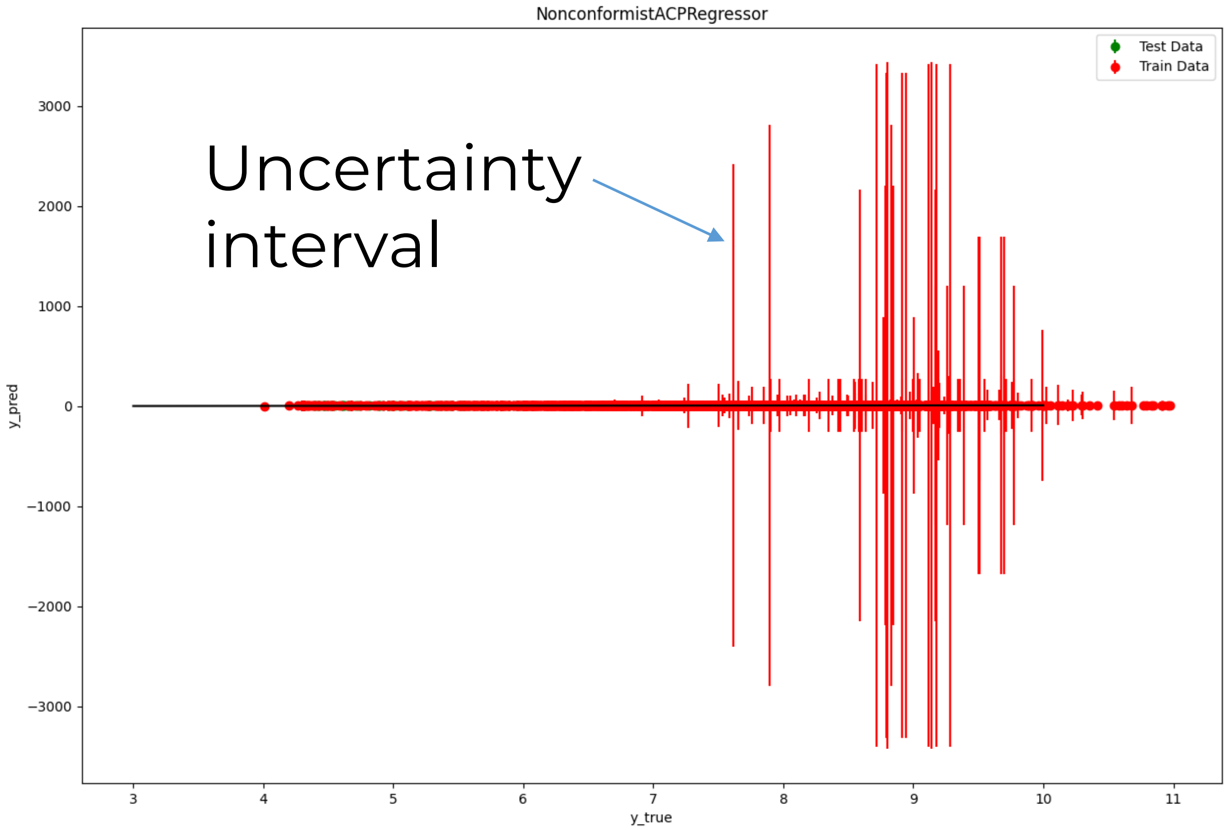
3. Uncertainty is weakly correlated to mean error

4. Epistemic uncertainty is a notion of applicability domain

Acknowledgements

- Prof Jonathan Goodman
- Dr Katie Przybylak
- Dr Tim Allen
- Dr Dawei Tang, Dr Joe Reynolds, Dr Alistair Middleton
- Goodman group

Blooper


NonconformistACPRegressor
Test Data
Train Data
3000
2000
1000
0
−1000
−2000
−3000
y_pred
3
4
5
6
7
8
9
10
11
y_true
Uncertainty interval

Questions and discussions!

(Thank you for listening to me)