



UNIVERSITY OF
CAMBRIDGE

Department of Physics

Scientific intuition inspired by machine learning generated hypotheses

P. Friedrich, M. Krenn, I. Tamblyn, A. Aspuru-Guzik

arXiv:2010.14236v2

Paolo Mognini

Cavendish Laboratory

Journal Club on Quantum Physics and Machine Learning

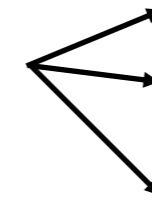
12th January 2021

Organisation

<http://ultracold.org>



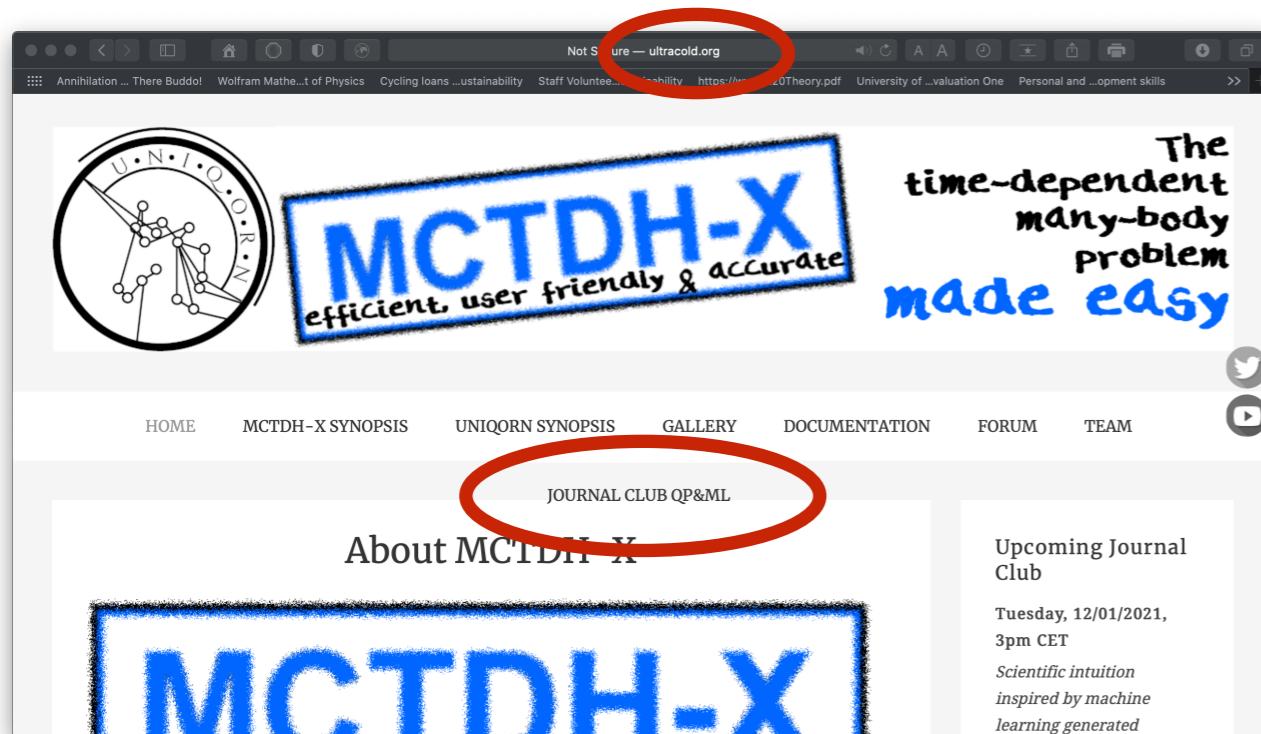
Journal Club QP&ML



Organisation

Schedule

Contact: JC@ultracold.org



Schedule

Participation

The Journal Club takes place on zoom every second Tuesday at 3pm CET. Please send an e-mail to the [mailing list](#) if you would like to receive the Zoom link before every session, or if you want to present or suggest a paper. Contact: JC@ultracold.org

Paper suggestions

This is a list of interesting papers that could be suitable for the Journal Club:

- 1 [Statistical Physics of Unsupervised Learning with Prior Knowledge in Neural Networks](#), Phys. Rev. Lett. 124, 248302 (2020).
- 2 [Extrapolating Quantum Observables with Machine Learning: Inferring Multiple Phase Transitions from](#) [Description of Single Phase](#), Phys. Rev. Lett. 124, 055702 (2020).

- Every second Tuesday, at 3pm CET on Zoom.
- Zoom link and password will be sent out by e-mail.
- Please get in touch if you want to present something: slots available from 16th of February.
- Contribution of own work is welcomed!

Scientific intuition inspired by machine learning generated hypotheses

Pascal Friederich,^{1,2,3,4,*} Mario Krenn,^{1,2,5} Isaac Tamblyn,^{6,5} and Alán Aspuru-Guzik^{1,2,5,7,†}

¹*Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Canada.*

²*Department of Computer Science, University of Toronto, Canada.*

³*Institute of Theoretical Informatics, Karlsruhe Institute of Technology,
Am Fasanengarten 5, 76131 Karlsruhe, Germany.*

⁴*Institute of Nanotechnology, Karlsruhe Institute of Technology,
Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany.*

⁵*Vector Institute for Artificial Intelligence, Toronto, Canada.*

⁶*National Research Council of Canada, Ottawa, Canada.*

⁷*Canadian Institute for Advanced Research (CIFAR) Lebovic Fellow, Toronto, Canada*

(Dated: December 15, 2020)

Abstract Machine learning with application to questions in the physical sciences has become a widely used tool, successfully applied to classification, regression and optimization tasks in many areas. Research focus mostly lies in improving the accuracy of the machine learning models in numerical predictions, while scientific understanding is still almost exclusively generated by human researchers analysing numerical results and drawing conclusions. In this work, we shift the focus on the insights and the knowledge obtained by the machine learning models themselves. In particular, we study how it can be extracted and used to inspire human scientists to increase their intuitions and understanding of natural systems. We apply gradient boosting in decision trees to extract human interpretable insights from big data sets from chemistry and physics. In chemistry, we not only rediscover widely know rules of thumb but also find new interesting motifs that tell us how to control solubility and energy levels of organic molecules. At the same time, in quantum physics, we gain new understanding on experiments for quantum entanglement. The ability to go beyond numerics and to enter the realm of scientific insight and hypothesis generation opens the door to use machine learning to accelerate the discovery of conceptual understanding in some of the most challenging domains of science.

I. INTRODUCTION

Machine learning (ML) recently became a widely used tool with many applications in the physical sciences [1], ranging from chemistry (for example, prediction of quantum chemistry properties [2], solving Schrödinger's equation [3], predicting reactions [4], materials discovery

intelligence in physical sciences aimed to directly answer scientific questions, *e.g.* determine the location of protein encodings in the genome [14]. Further attempts to employ machine learning models to obtain insight and help scientists to develop theories were focused on rediscovering solutions to already solved problems, *e.g.* to rediscover the coordinate transformation in

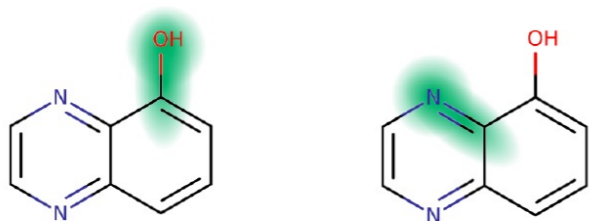
The paper in a nutshell

Gradient Boosting Regression on graph-based structures
to identify features that increase/decrease target properties

physical intuition

Organic chemistry

Which chemical groups in a compound lead to changes in certain properties?



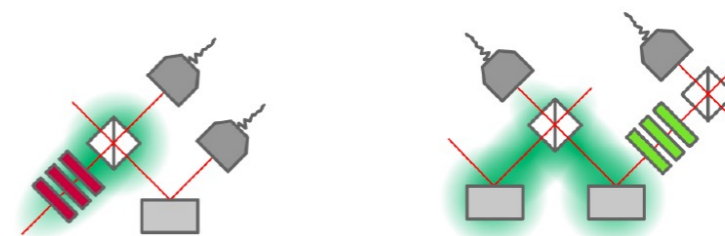
1. better/worse solubility in water vs. octanol.
2. higher/lower HOMO¹.
3. other changes in relevant properties of application-specific data sets, eg larger/smaller HOMO-LUMO² gap...



design of pharmaceutical drugs,
organic solar cells, OLEDs³...

Quantum optics

Which combinations of experimental sub-components can increase/decrease the production of high-dimensional, multipartite quantum entanglement?



design of experiments to probe
local realism, or for quantum
communication networks

¹ HOMO = Highest Occupied Molecular Orbital

² LUMO = Lowest Unoccupied Molecular Orbital

³ OLED = Organic Light Emitting Diode

The paper in a nutshell

Gradient Boosting Regression on graph-based structures
to **identify features that increase/decrease target properties**

physical intuition

Organic chemistry

Quantum optics

→ “The carbonyl group increases solubility in water.”
→ “The amine group lifts the HOMO to higher levels.”
→ “Silole rings reduce the HOMO-LUMO gap.”

previously known
more unusual

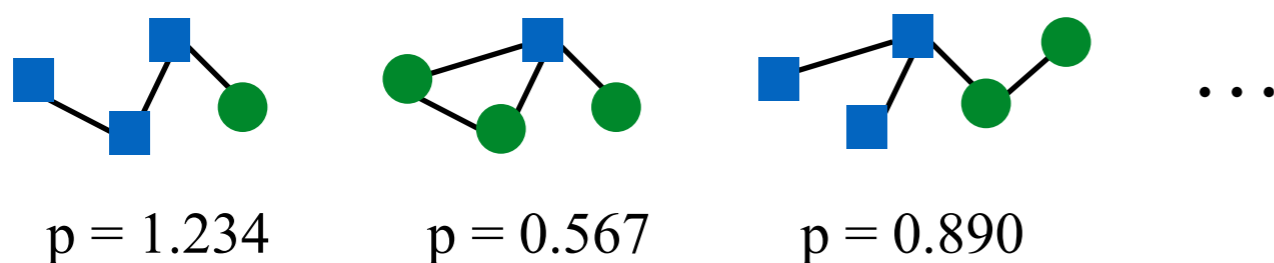
→ “An array of three equal-shift holograms increases multipartite entanglement.”
→ “Two nonlinear crystals connected via a beam splitter decrease multipartite entanglement.”

previously known
previously unknown

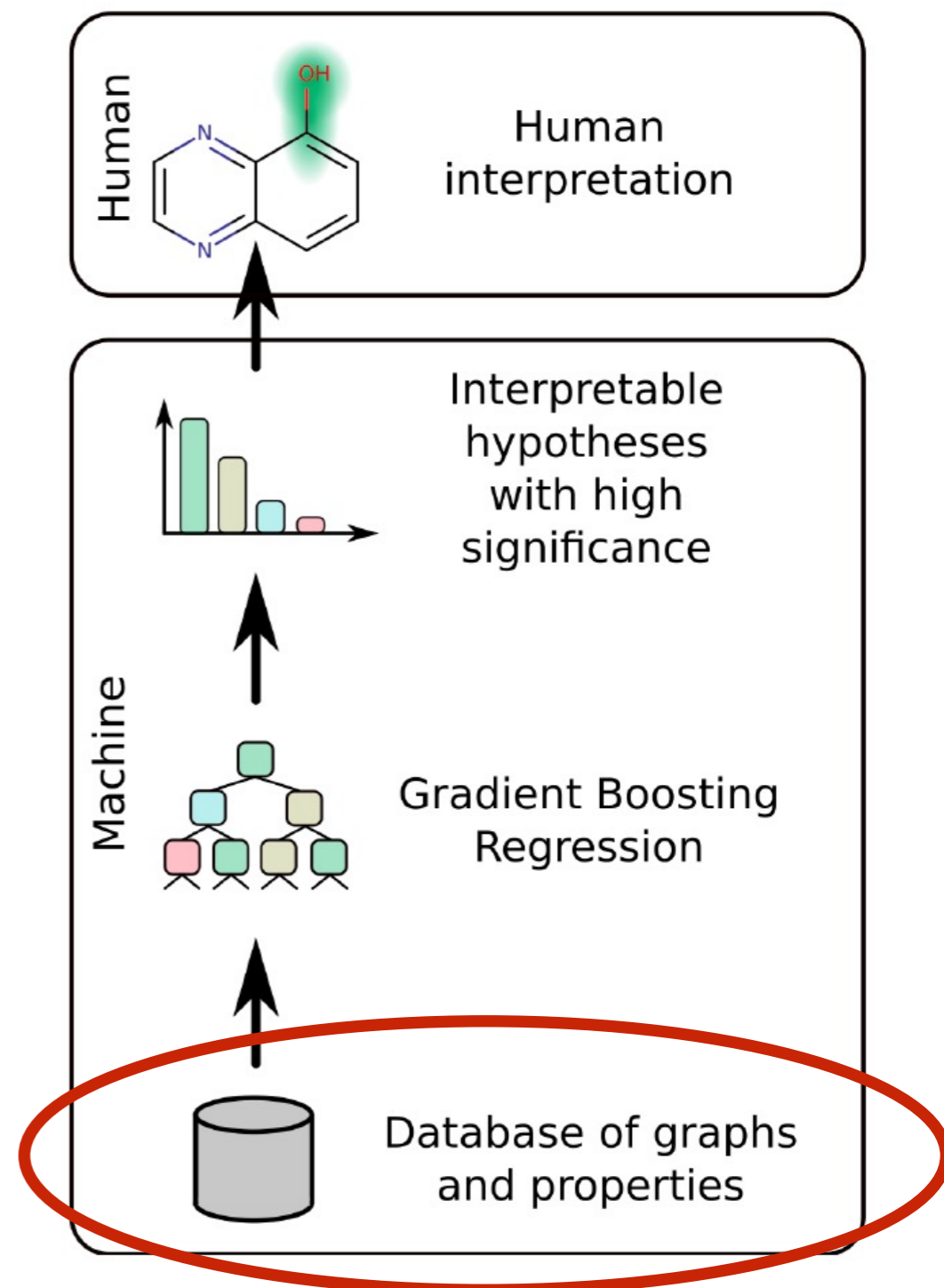
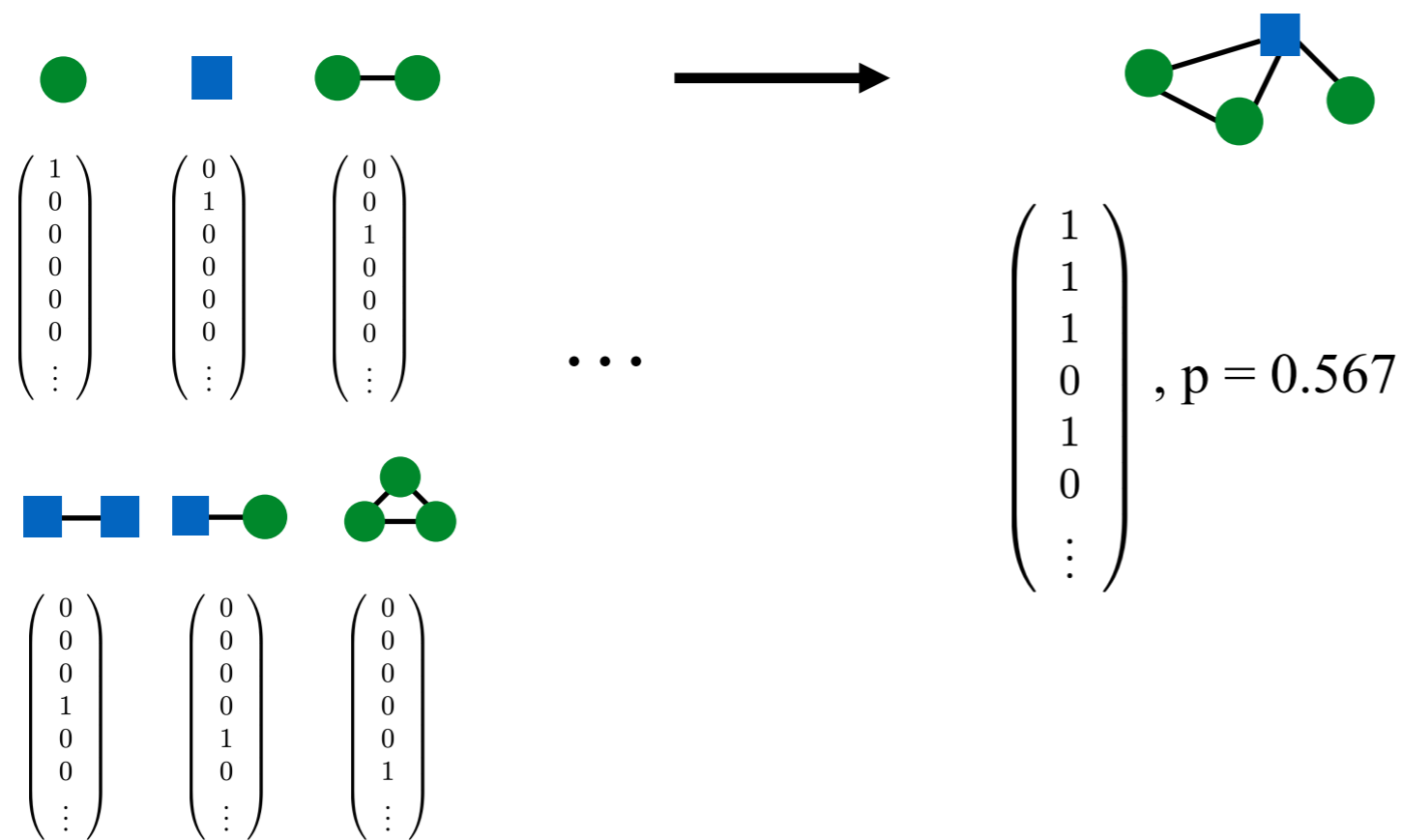
contradicts the intuition in the field !

Workflow for automated hypothesis generation

- Graph = any structure that can be represented as nodes connected by edges.



- All subgraphs are also listed and codified in bit-vectors using fingerprinting techniques.



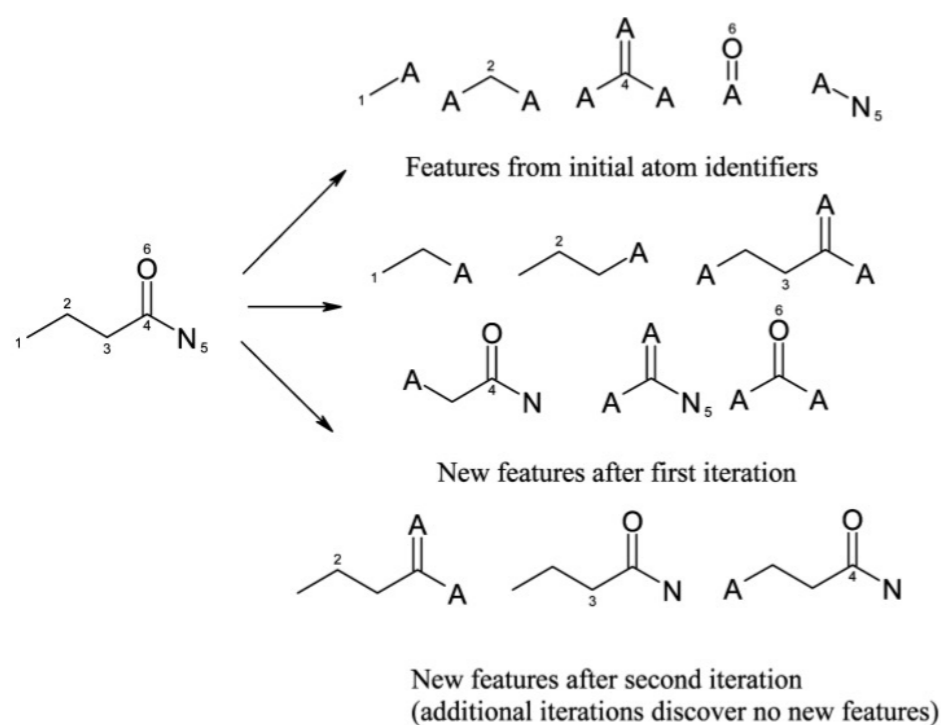
Workflow for automated hypothesis generation

Extended-Connectivity Fingerprints

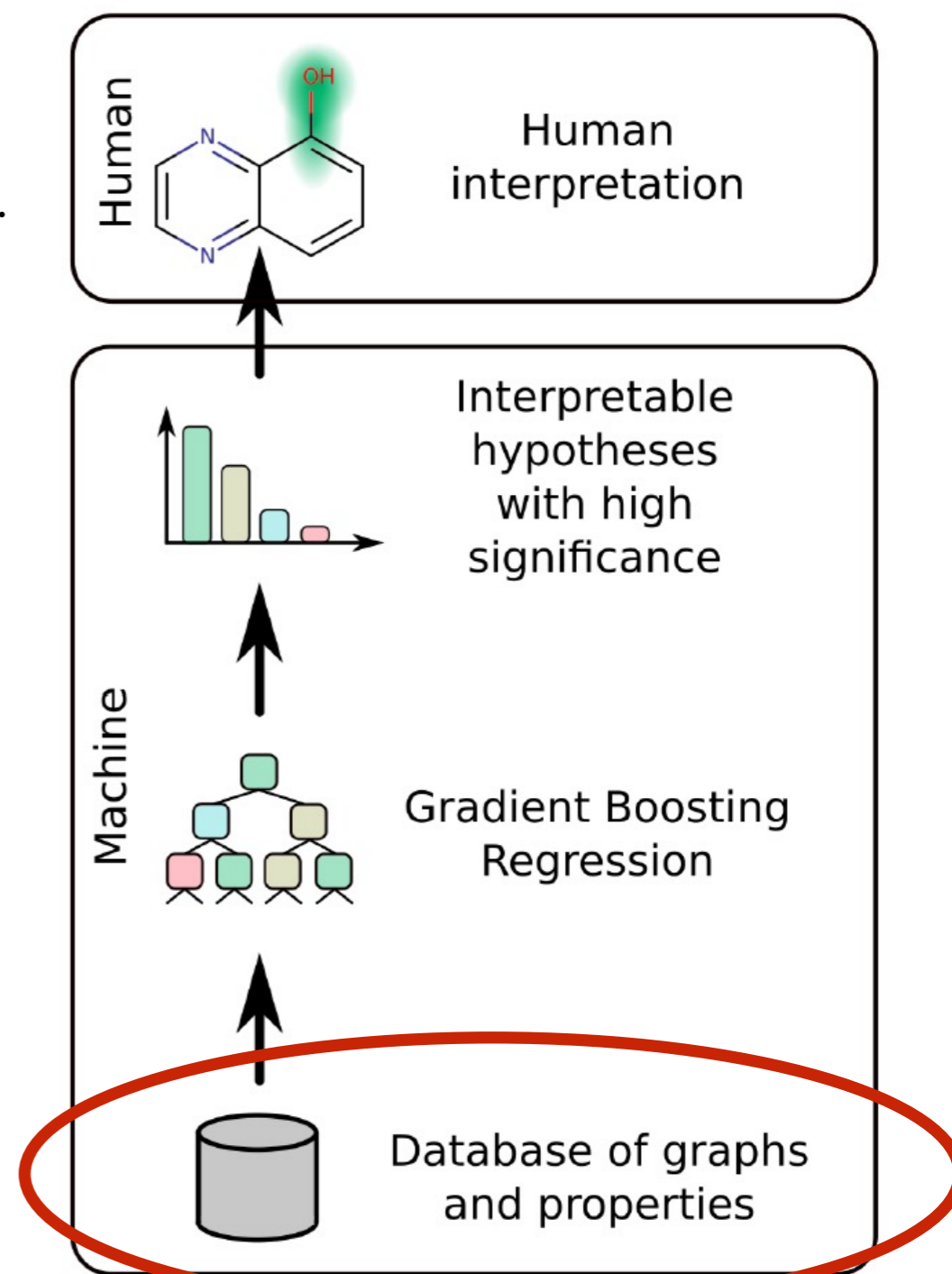
→ Also used to fingerprint optical components!

→ J. Chem. Inf. Model. **50**, 742–754 (2010)

- Numbering of all non-H sites following conventions.
- Iteratively mapping out structures including neighbours.
- Hashing the resulting identifiers into a 32-bit integer.



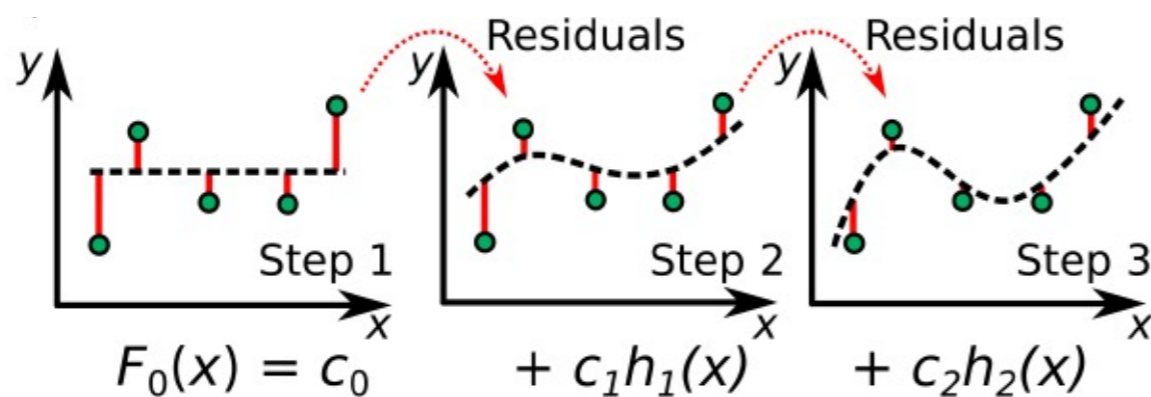
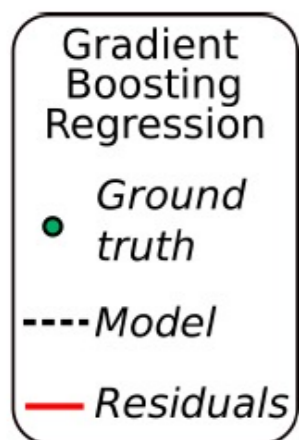
> <ECFP_0>	> <ECFP_2>	> <ECFP_4>	> <ECFP_6>
734603939	734603939	734603939	734603939
1559650422	1559650422	1559650422	1559650422
-1100000244	-1100000244	-1100000244	-1100000244
1572579716	1572579716	1572579716	1572579716
-1074141656	-1074141656	-1074141656	-1074141656
	863188371	863188371	863188371
	-1793471910	-1793471910	-1793471910
	-1789102870	-1789102870	-1789102870
	-1708545601	-1708545601	-1708545601
	-932108170	-932108170	-932108170
	2099970318	2099970318	2099970318
		-87618679	-87618679
		1112638790	1112638790
		-627599602	-627599602



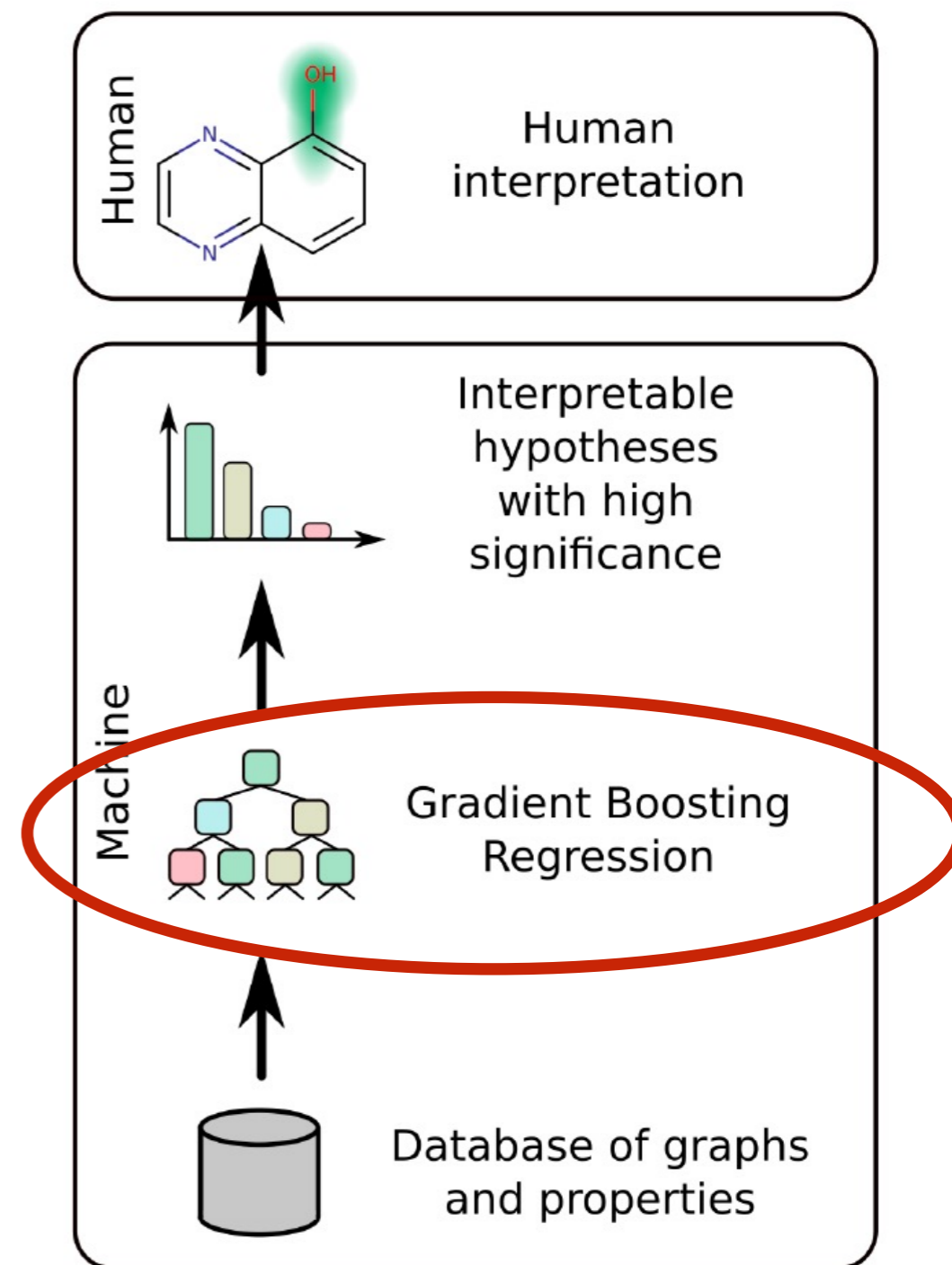
Gradient Boosting Regression

Minimise loss function L (eg least squares) by iteratively minimising the residuals, eg via decision trees.

$$h_m(x) = y - F_m(x) = -\frac{\partial L_{MSE}}{\partial F}$$



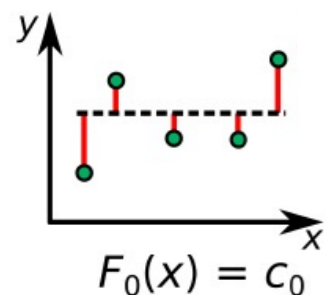
Ref: Cory Maklin, “Gradient Boosting Decision Tree Algorithm explained”, [towardsdatascience.com](https://towardsdatascience.com/gradient-boosting-decision-tree-algorithm-explained/) (May 18, 2019)



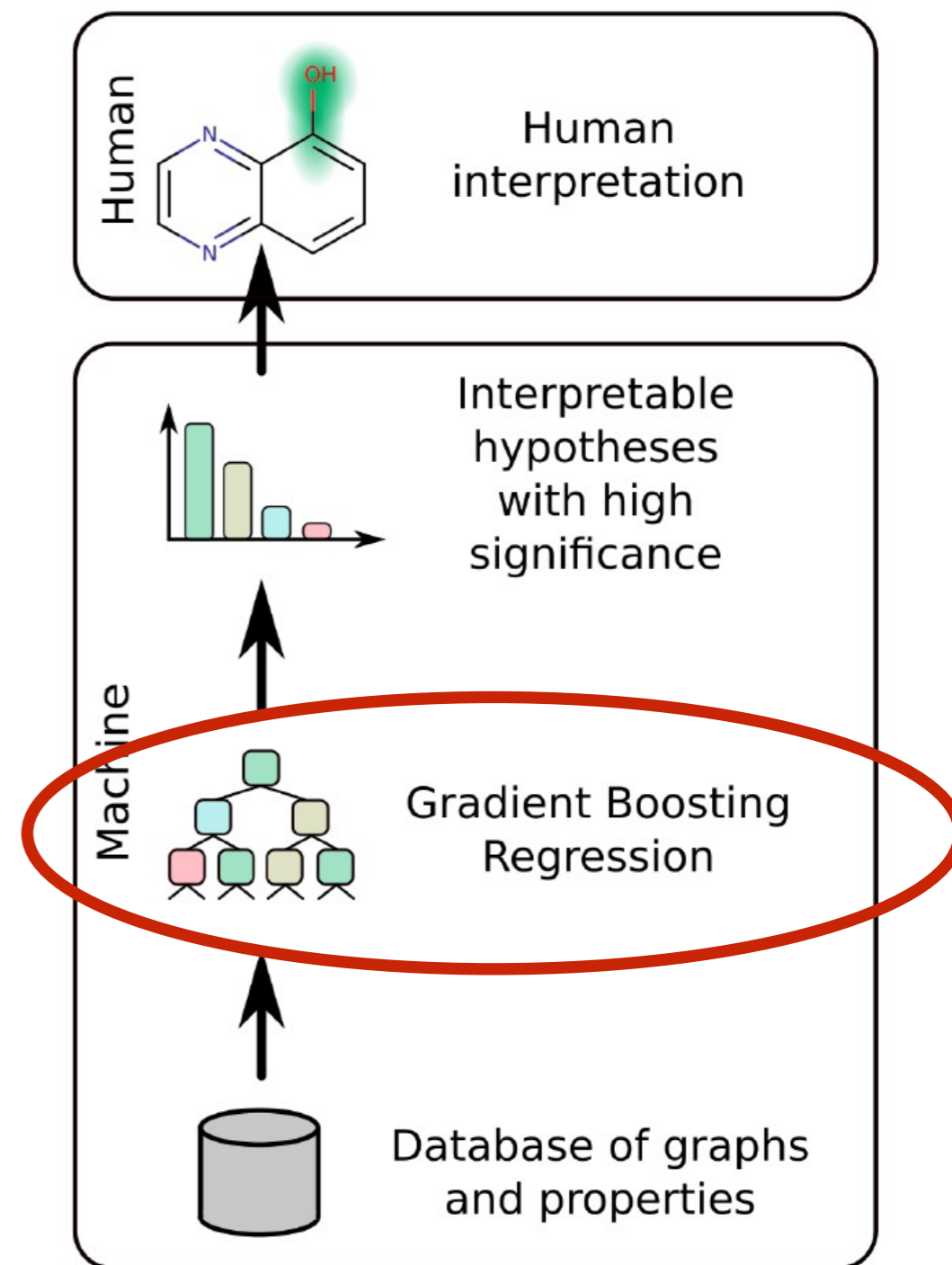
Gradient Boosting Regression

Example: determine house price

Age	Sq. footage	Location	Price
5	1500	5	480
11	2030	12	1090
14	1442	6	350
8	2501	4	1310
12	1300	9	400
10	1789	11	500



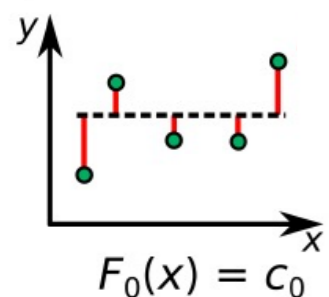
1. Calculate average: $c_0=688$
2. Calculate first residuals.



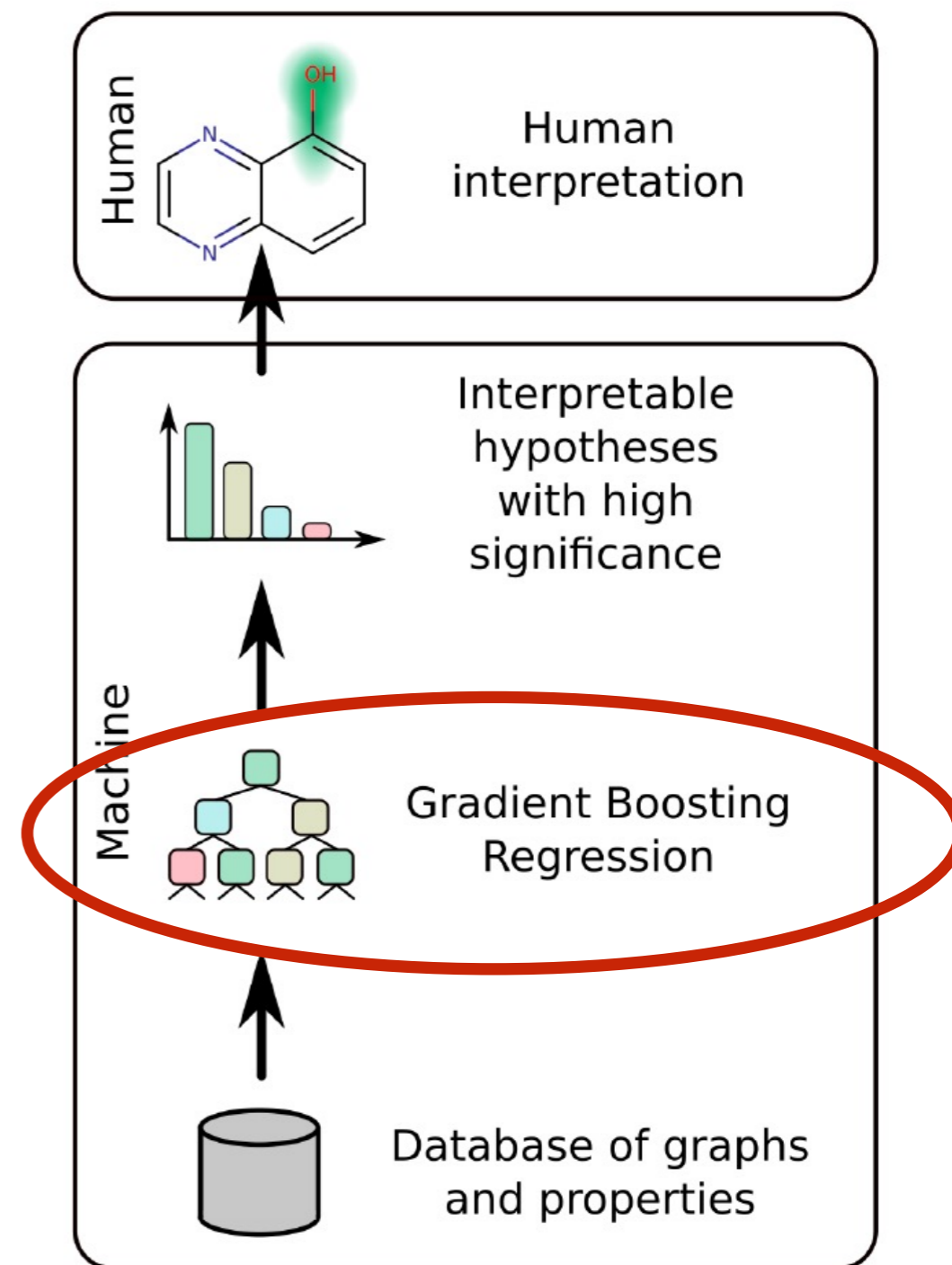
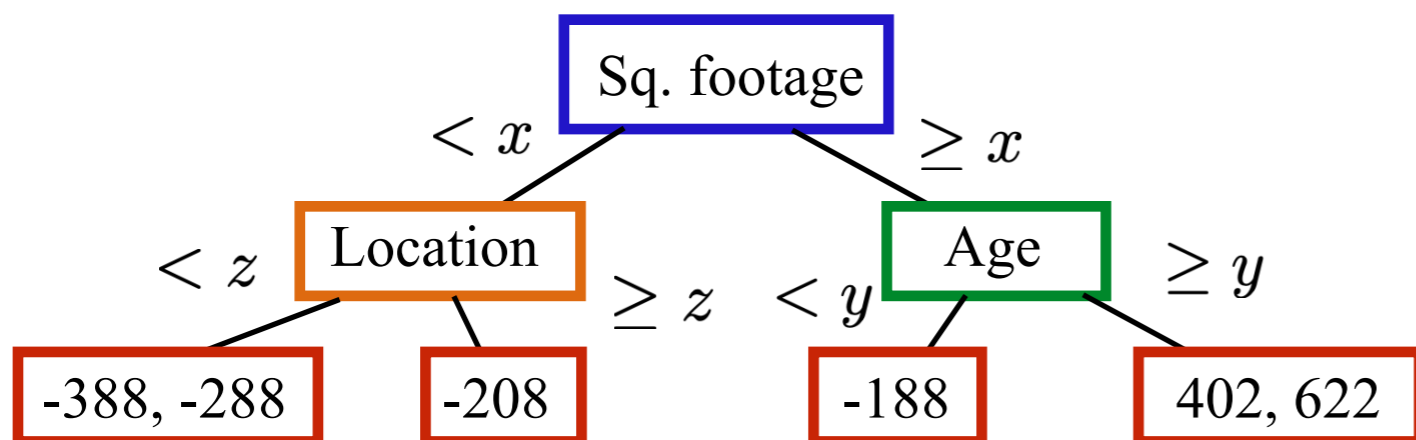
Gradient Boosting Regression

Example: determine house price

Age	Sq. footage	Location	Price	Res. 1
5	1500	5	480	-208
11	2030	12	1090	402
14	1442	6	350	-338
8	2501	4	1310	622
12	1300	9	400	-288
10	1789	11	500	-188



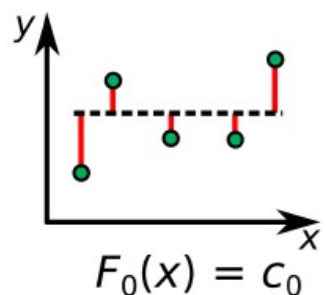
1. Calculate average: $c_0=688$.
2. Calculate first residuals.
3. Construct a decision tree from features.



Gradient Boosting Regression

Example: determine house price

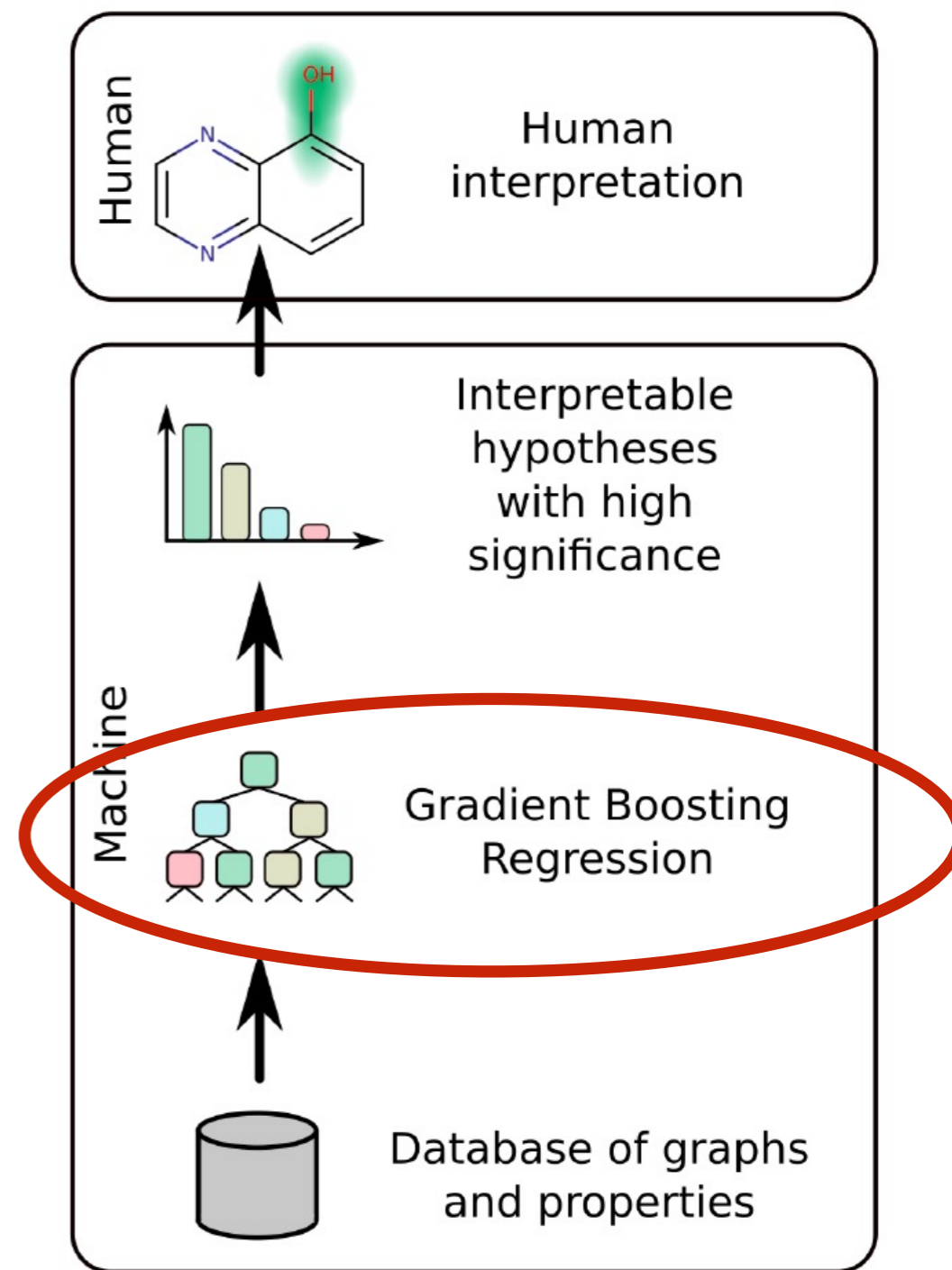
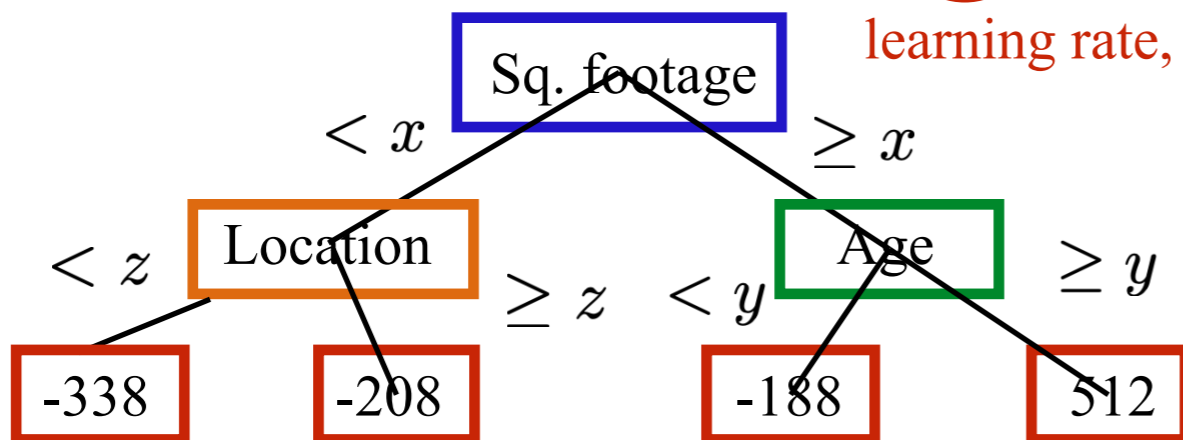
Age	Sq. footage	Location	Price	Res. 1	Pred.
5	1500	5	480	-208	667
11	2030	12	1090	402	739
14	1442	6	350	-338	654
8	2501	4	1310	622	739
12	1300	9	400	-288	654
10	1789	11	500	-188	669



1. Calculate average: $c_0=688$.
2. Calculate first residuals.
3. Construct a decision tree from features.
4. Fit the data with the decision tree.

$$F_1(x) = c_0 + c_1 h_1(x)$$

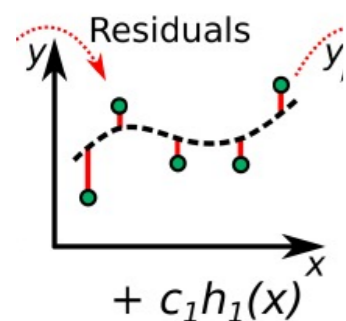
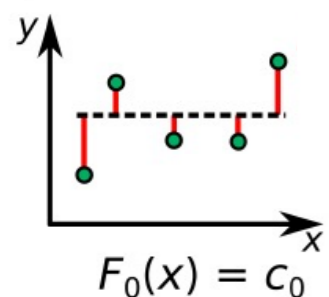
learning rate, $c_1 = 0.1$



Gradient Boosting Regression

Example: determine house price

Age	Sq. footage	Location	Price	Res. 1	Pred.
5	1500	5	480	-208	667
11	2030	12	1090	402	739
14	1442	6	350	-338	654
8	2501	4	1310	622	739
12	1300	9	400	-288	654
10	1789	11	500	-188	669

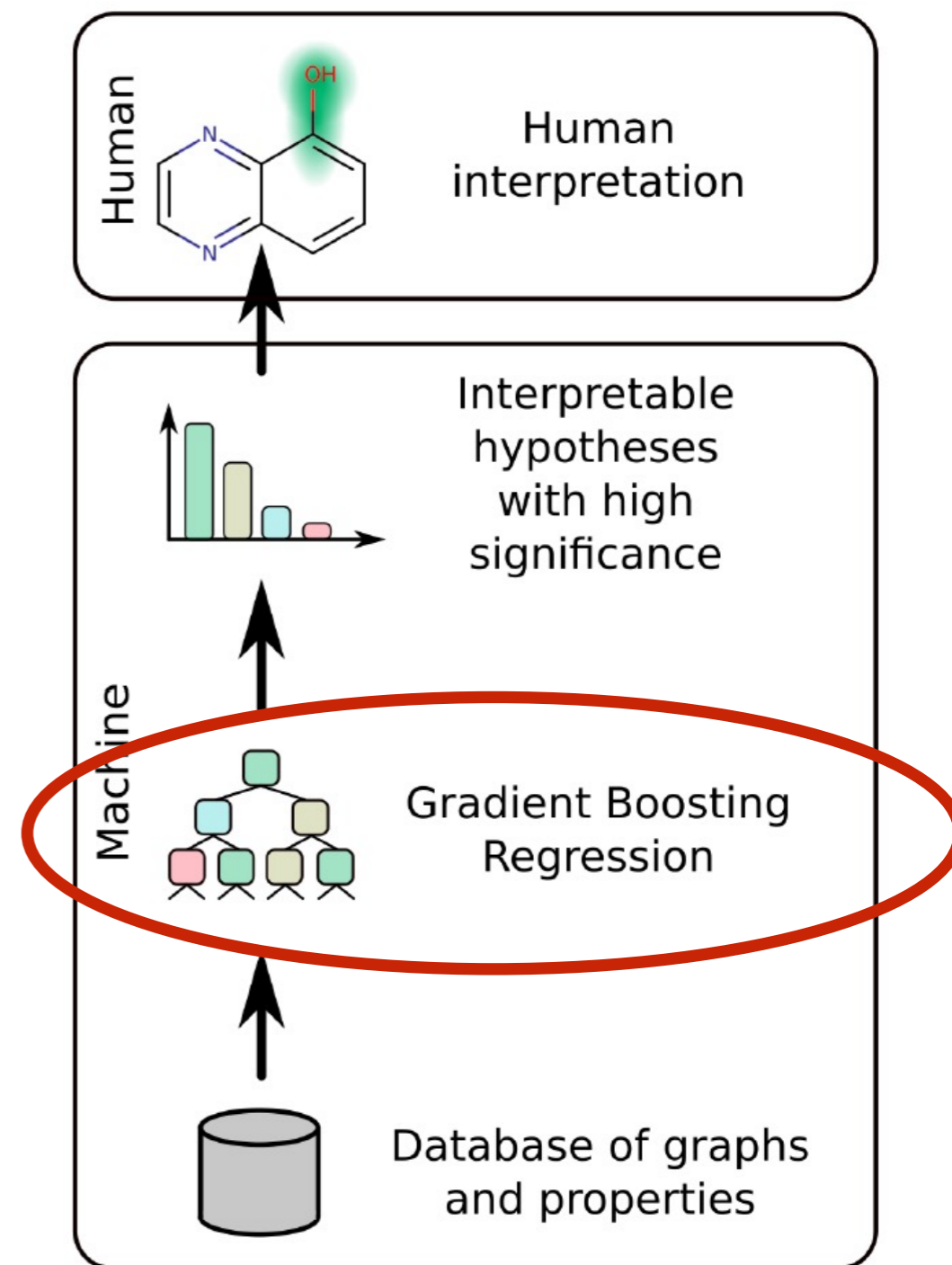


1. Calculate average: $c_0=688$.
2. Calculate first residuals.
3. Construct a decision tree from features.
4. Fit the data with the decision tree.

$$F_1(x) = c_0 + c_1 h_1(x)$$

learning rate, $c_1=0.1$

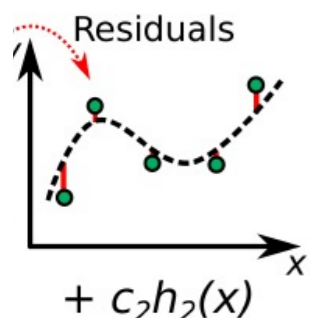
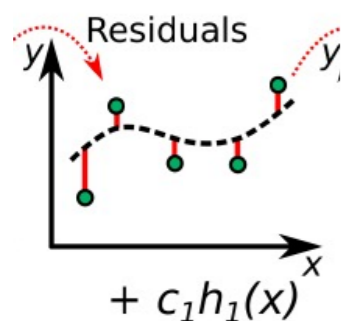
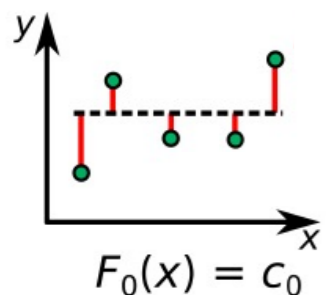
5. Compute new residuals.



Gradient Boosting Regression

Example: determine house price

Age	Sq. footage	Location	Price	Res. 1	Pred.	Res. 2
5	1500	5	480	-208	667	-187
11	2030	12	1090	402	739	351
14	1442	6	350	-338	654	-304
8	2501	4	1310	622	739	571
12	1300	9	400	-288	654	-254
10	1789	11	500	-188	669	-169



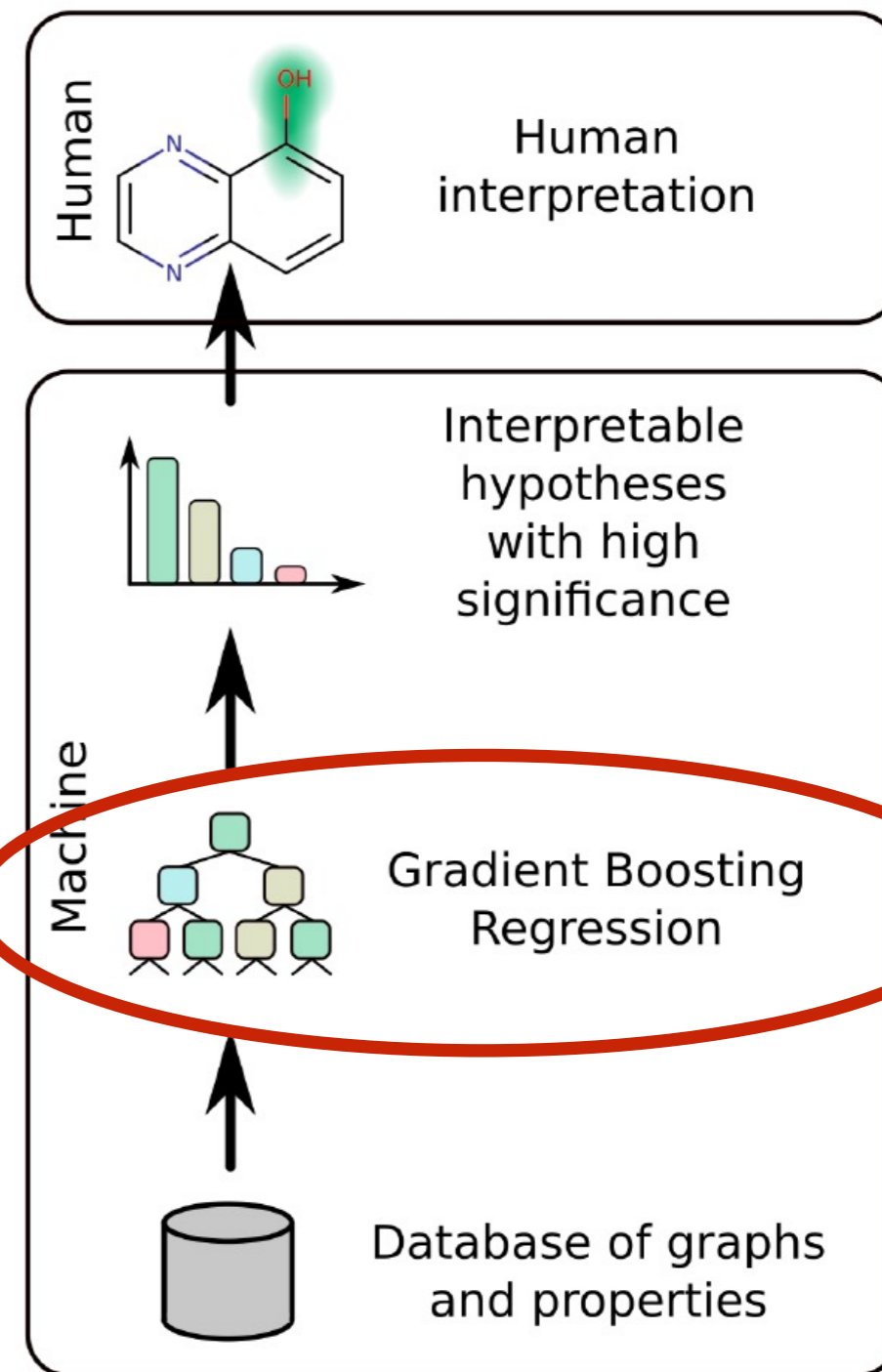
1. Calculate average: $c_0=688$.
2. Calculate first residuals.
3. Construct a decision tree from features.
4. Fit the data with the decision tree.

$$F_1(x) = c_0 + c_1 h_1(x)$$

learning rate, $c_1 = 0.1$

5. Compute new residuals.
6. Repeat 3-5 until convergence.

$$F_n(x) = c_0 + c_1 h_1(x) + c_2 h_2(x) + \dots + c_n h_n(x)$$



Gradient Boosting Regression for hypothesis generation

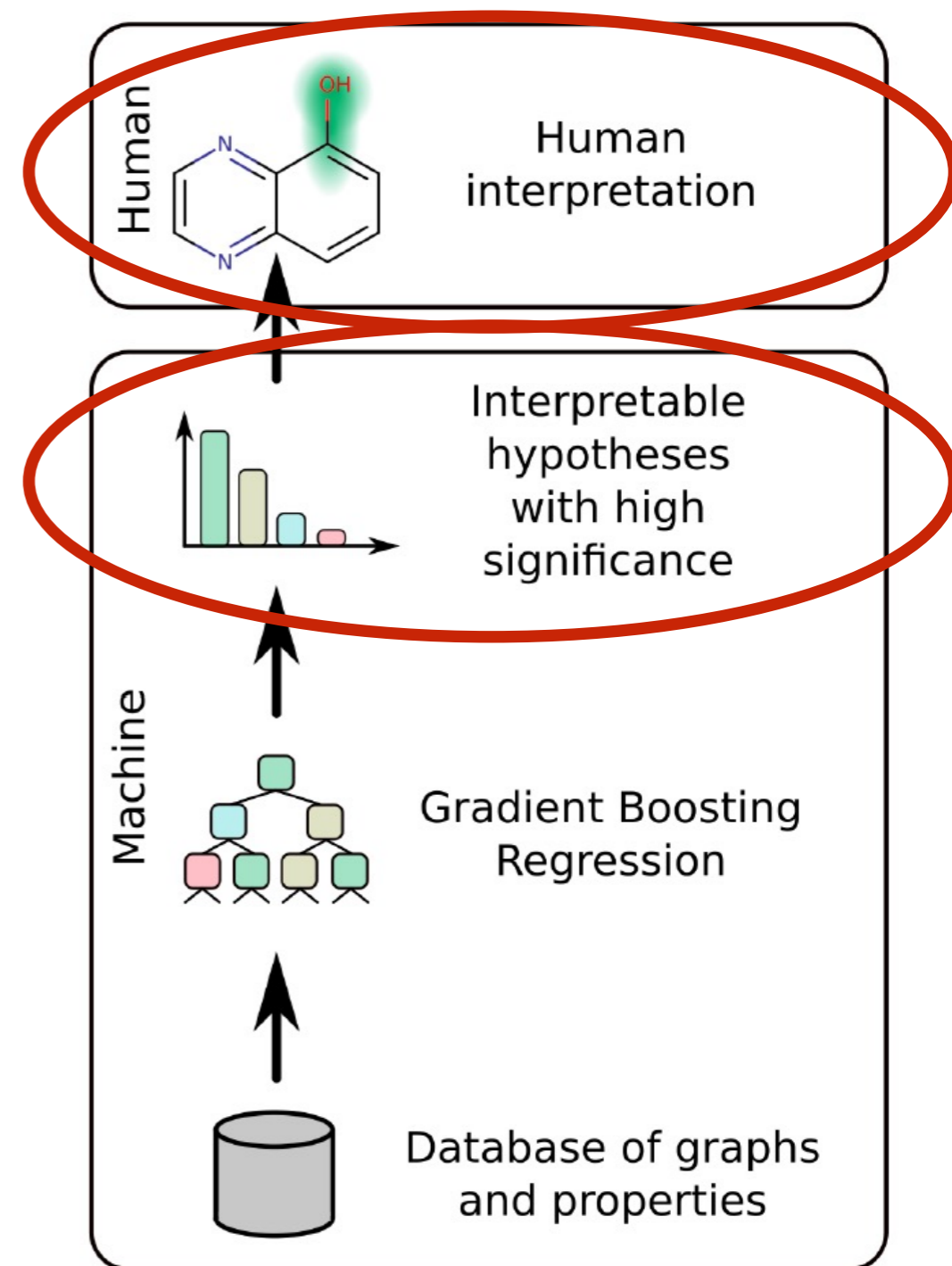
“A binary feature vector describing presence/absence of automatically generated subgraphs is used to train a tree ensemble method *eg* Gradient Boosting.”

subgraph 1	subgraph 2	subgraph 3	...	property
Yes	No	Yes		1.234
Yes	No	No		0.456
No	Yes	No		0.789
No	No	Yes		1.011
Yes	No	Yes		1.213
...

→ Quantification of feature importance

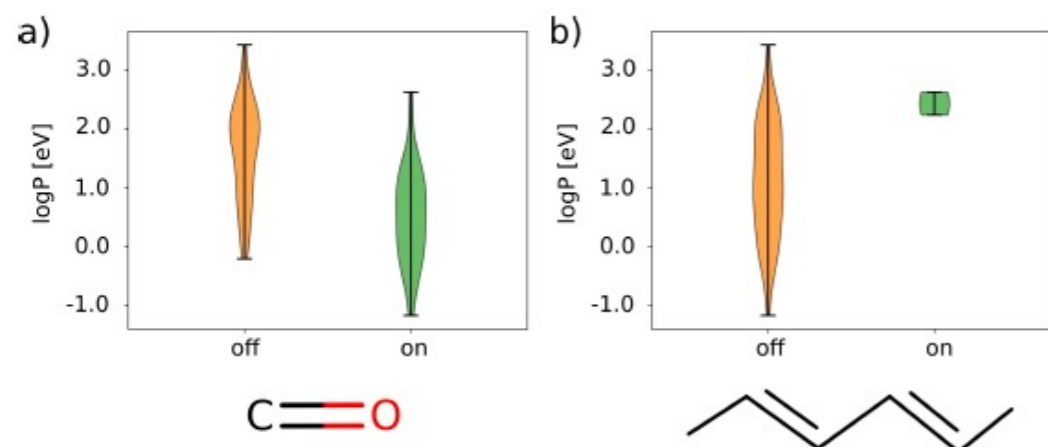
“Feature i leads to an increase/decrease of target property of strength s ”

→ Aggregate features to formulate hypothesis, *eg*
“The carbonyl group increases solubility in water.”

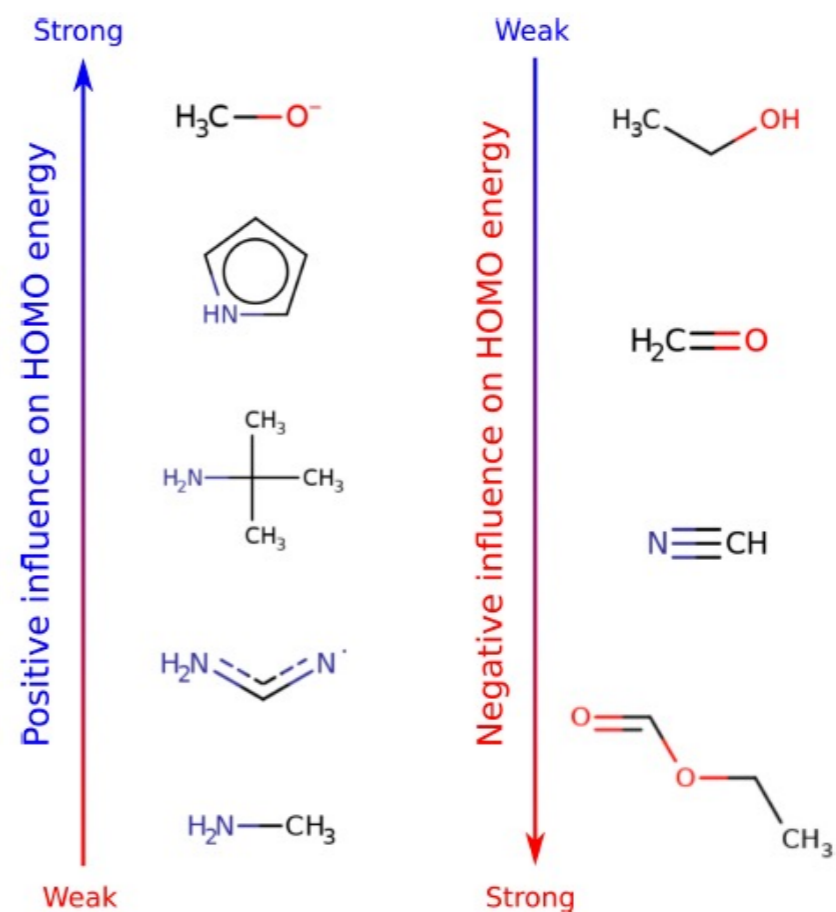


Results 1: chemistry

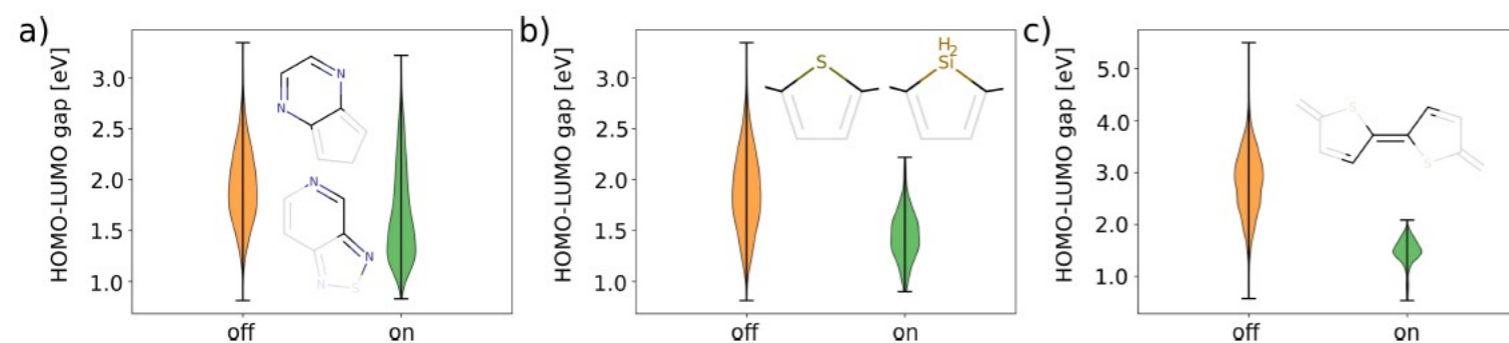
Influence on solubility in polar vs. nonpolar molecules



Influence on HOMO energy



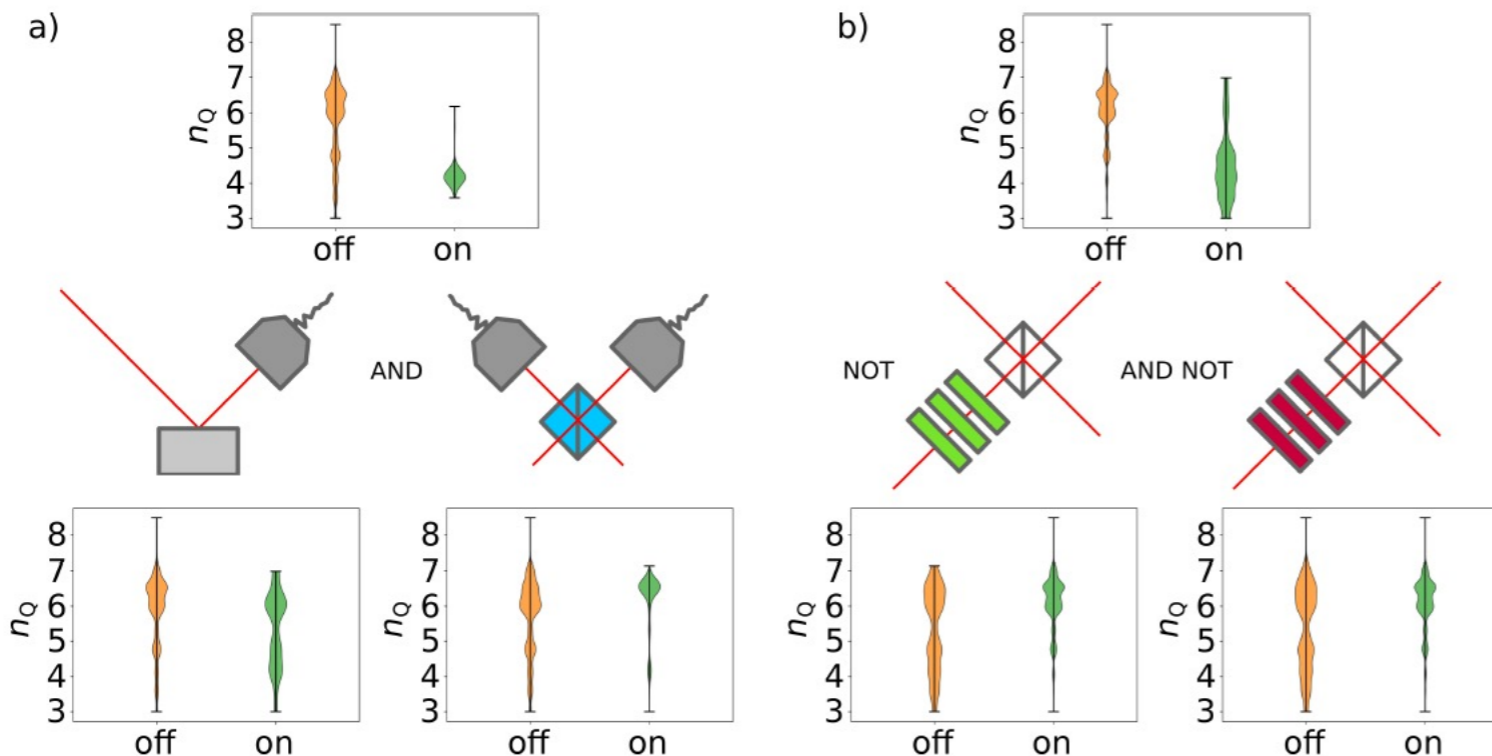
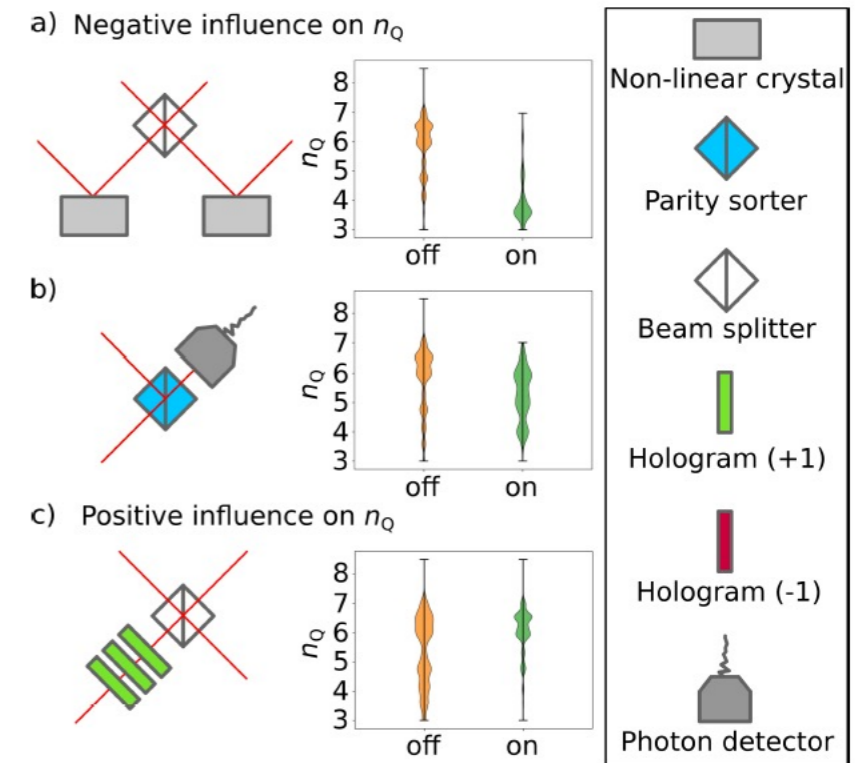
Influence on HOMO-LUMO gap



Results 2: quantum optics

- Various experimental setups to engineer high-dimensional multipartite entanglement.
- 3 photons + 1 photon as a trigger
- Measure of entanglement = size of involved Hilbert space⁴:

$$n_Q = \log_2(d_1 d_2 d_3) \quad d_i = \text{rank}(\text{Tr}_i[\rho])$$



⁴ M. Huber and J. I. de Vicente, Structure of multidimensional entanglement in multipartite systems. Phys. Rev. Lett. **110**, 030501 (2013).

Discussion

Extend same approach to other areas?

- Identify features (symmetries?) that increase localization of edge modes in topological materials?
- Identify compounds or elements that stabilise certain phases of matter, *eg* superconductivity?
- Identify laser regimes in ultracold systems that maximise observable readout?

Issues:

- Need for graph-based database.
- Size of the database?

Meta-analysis:

- Hyperparameter optimisation to determine which features increase performance for other graph-based machine learning algorithms, *eg* neural networks.