Automating Science using Robot Scientists

Ross D. King,
University of Manchester, ATI, ross.king@manchester.ac.uk

The State-of-the-Art - AI

Technology Drivers

- Improved computer hardware:
  - faster processors, more processors, GPUs, …
- Improved data availability:
  - computers recording almost everything, deep data, …
- Improved computer software:
  - new machine learning methods, deep mining, …

Artificial Intelligence (AI)

- There have been multiple AI hype cycles, but this time it seems different.
- AI, especially machine learning, is now the hottest technology on the planet. Speed of Advance has surprised me.
- Machine Learning is the core technology of Google, Facebook, Amazon, …Tencent, Alibaba, Baidu, …

Computer Chess - 1997

Abstract world: 64 squares, 32 pieces.

Computers play chess better than the best humans, and computers can now make strikingly beautiful moves.

Deep Blue v Kasparov

Famous game 2 – Machine had no fear

Jeopardy - 2011

- Watson is a computer system capable of answering questions posed in natural language.
- Watson easily beat Brad Rutter and Ken Jennings the best human players of Jeopardy.

Google Driverless Car - 2016

- The system combines maps with sensors to automate driving.
- Has driven long distances with no human intervention.
- Laws changed in a number of states, countries.

Computer Go - 2016

AlphaGo v Lee Sedol

March 2016, won 4 : 1.

Combination of brute-force search and machine learning (deep learning + reinforcement learning) to evaluate board positions.

Scientific Discovery

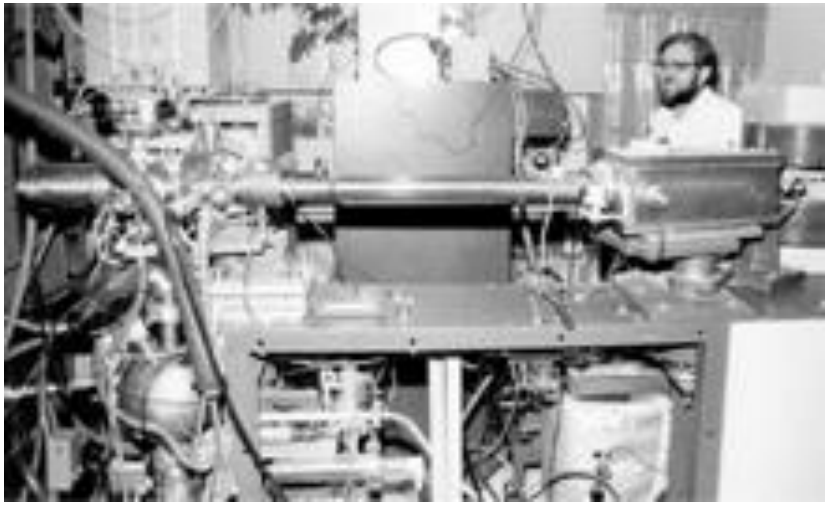
AI Systems have Superhuman Scientific Reasoning Powers

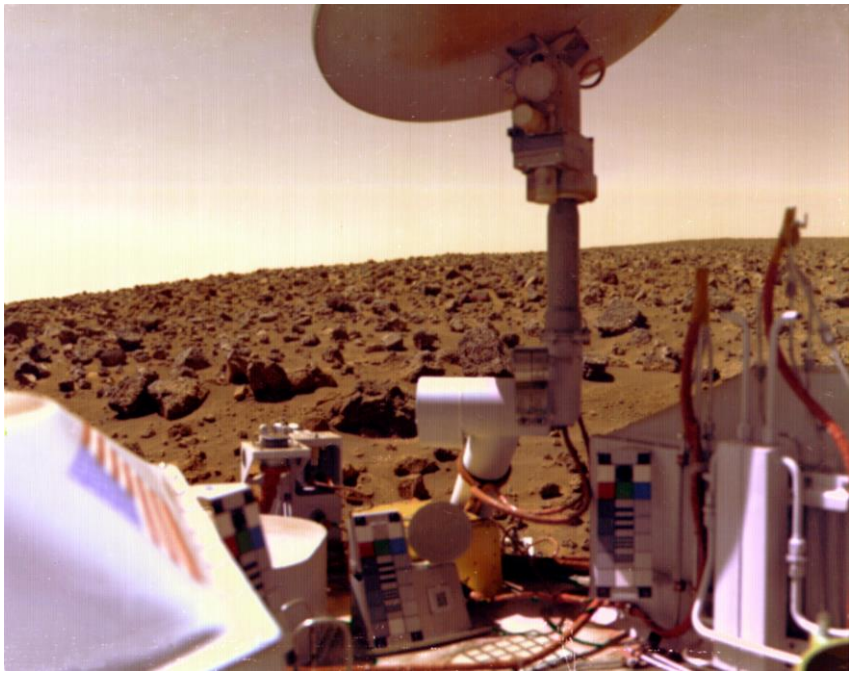
- Flawlessly remember vast numbers of facts
- Execute flawless logical reasoning
- Execute near optimal probabilistic reasoning,
- Learn more rationally than humans
- Learn from vast amounts of data
- Extract information from millions of scientific papers.
- Etc.

Scientific Discovery

- Scientific problems are abstract, but involve the real-world.
- Scientific problems are restricted in scope – no need to know about “Cabbages and Kings”.
- Nature is honest – no malicious agents.
- Nature is a worthy object of our study.
- The generation of scientific knowledge is a public good.

Meta-Dendral

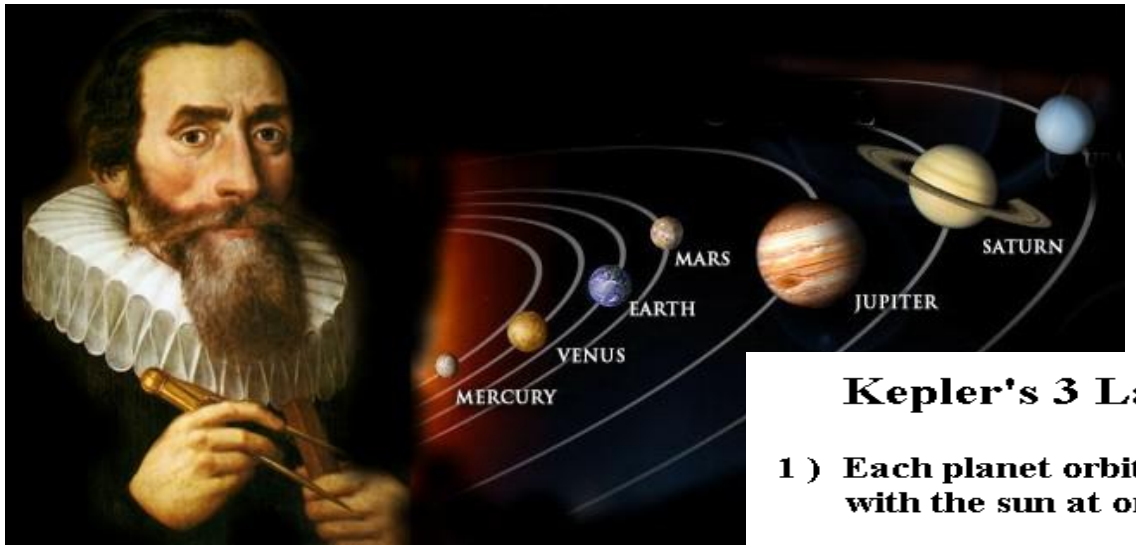




Analysis of mass-spectrometry data.

Joshua Lederburg, Ed. Feigenbaum, Bruce Buchanan, Karl Djerassi, *et al*. 1960-70s.

Bacon



Kepler's 3 Laws of Planetary Motion

1) **Each planet orbits the sun in an elliptical path with the sun at one focus**

2) **The radius vector (from sun to planet) sweeps out equal areas in equal time intervals**

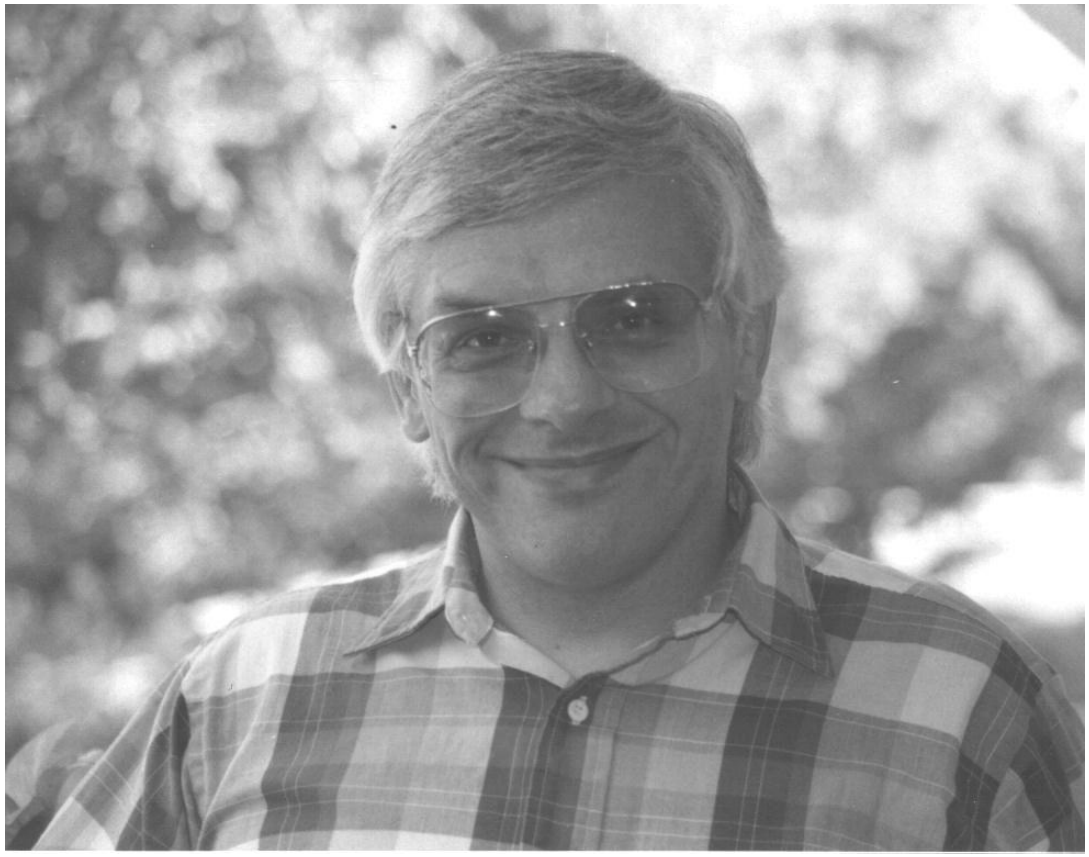
3) **The square of the period is proportional to the cube of the semi-major axis of the orbit**

i.e. $T^2 = k\ a^3$ **for some constant k**

Figure 11.1

Rediscovering physics and chemistry. Langley, Bradshaw, Simon (1979).

Into the Lab



Jan Zytkow (1944-2001)

- Automated discovery in a chemistry laboratory.

Zytkow, et al. (1990)

The State-of-the-Art Laboratory Automation

Lab Automation

- Laboratory Robotics technology is steadily advancing.
- Almost everything that a human can do in the lab can now be automated.
- Robots can work longer, faster, more accurately than humans.
- Problems: high capital-costs, high running-costs, poor integration, poor software.

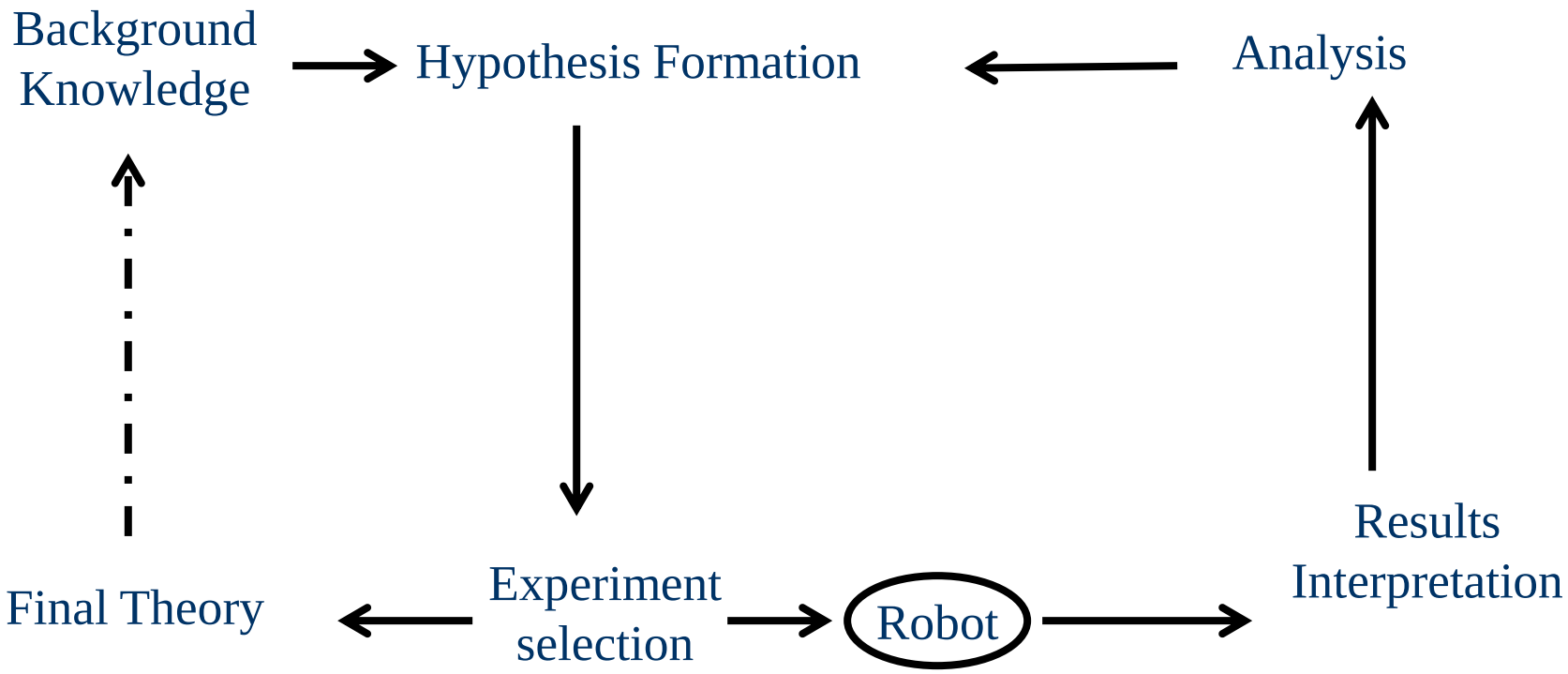
Cloud Laboratory Automation

- Idea:
  - Laboratory automation is expensive and difficult to use.
  - Therefore put automation in the cloud.
  - Customers use an API to design the experiments.
  - Benefits of scale.
- Two main companies: Emerald Cloud Services, Transcriptics.
- I worry about control of the API and/or Balkanisation of APIs

Robot Scientists

The Concept of a Robot Scientist

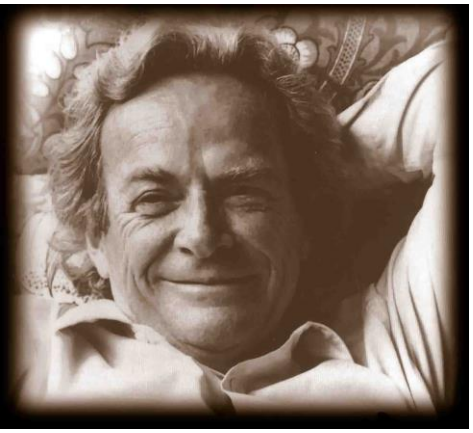
Computer systems capable of originating their own experiments, physically executing them, interpreting the results, and then repeating the cycle.

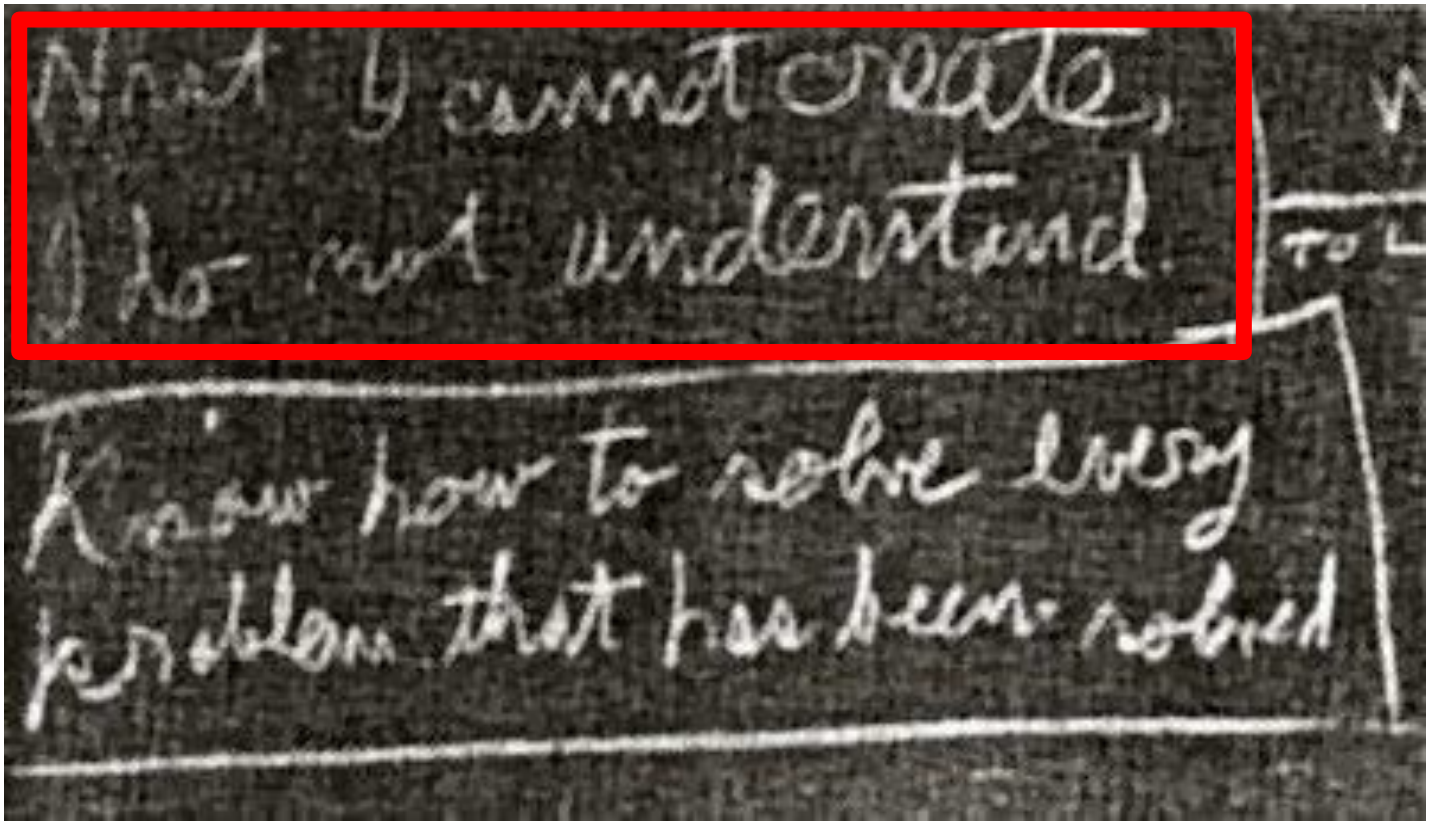


Motivation: Philosophical

- What is Science?
- The question whether it is possible to automate scientific discovery seems to me central to understanding science.
- There is a strong philosophical position which holds that we do not fully understand a phenomenon unless we can make a machine which reproduces it.

Richard Feynman's Blackboard





"What I cannot create, I do not understand"

Motivation: Technological

- Robot Scientists have the potential to increase the productivity of science. They can work cheaper, faster, more accurately, and longer than humans. They can also be easily multiplied.
  - *Enabling the high-throughput testing of hypotheses*.

- Robot Scientists have the potential to improve the quality of science.
  - *by enabling the description of experiments in greater detail and semantic clarity.*

The Complexity of Biological Systems

- Even simple “model” biological systems like that of *E. coli* and yeast are incredibly complicated.
- Thousands of genes, proteins, small-molecules, interacting together in complicated spatial temporal ways.
- Ockham's razor doesn’t work - system evolved.

- Not enough PhDs in the world to disentangle these systems.
- Need help - Robot Scientists.

Robot Scientist Timeline

- 1999-2004 Initial Robot Scientist Project
  - Limited Hardware: Collaboration with Douglas Kell (Aber Biology), Steve Oliver (Manchester), Stephen Muggleton (Imperial)

  King *et al*. (2004) *Nature*, 427, 247-252

- 2004-2011 Adam – Yeast Functional Genomics
  - Sophisticated Laboratory Automation: Collaboration with Steve Oliver (Cambridge).

  King *et al*. (2009) *Science*, 324, 85-89

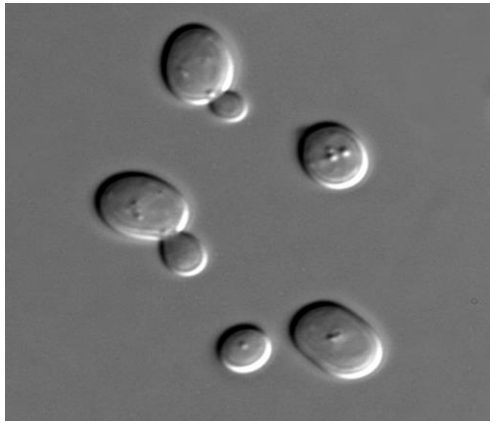
- 2008-2015 Eve – Drug Design for Tropical Diseases
  - Sophisticated Laboratory Automation: Collaboration with Steve Oliver (Cambridge)

  Williams *et al*. (2015) *Royal Society Interface*, DOI 10.1098/rsif.2014.1289

- 2015-2019 Eve – Human cells - Cancer, Yeast - Aging
  - DARPA, CHIST-ERA, EPSRC.

Adam

The Application Domain

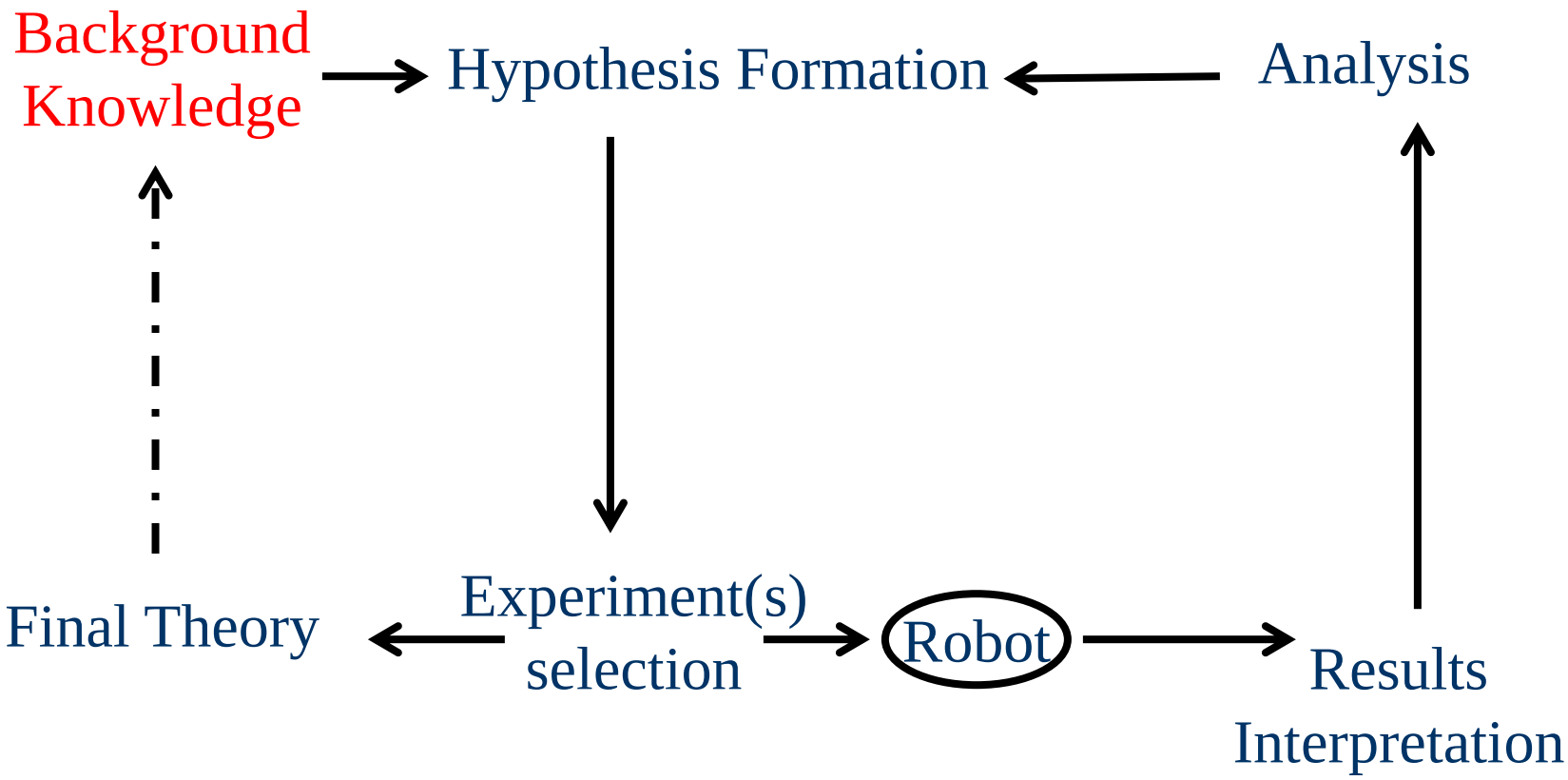


- Functional genomics
- In yeast (*S. cerevisiae*) ~15% of the 6,000 genes still have no known function.
- EUROFAN 2 made all viable single deletant strains.
- Task to determine the “function” of a gene by growth experiments.

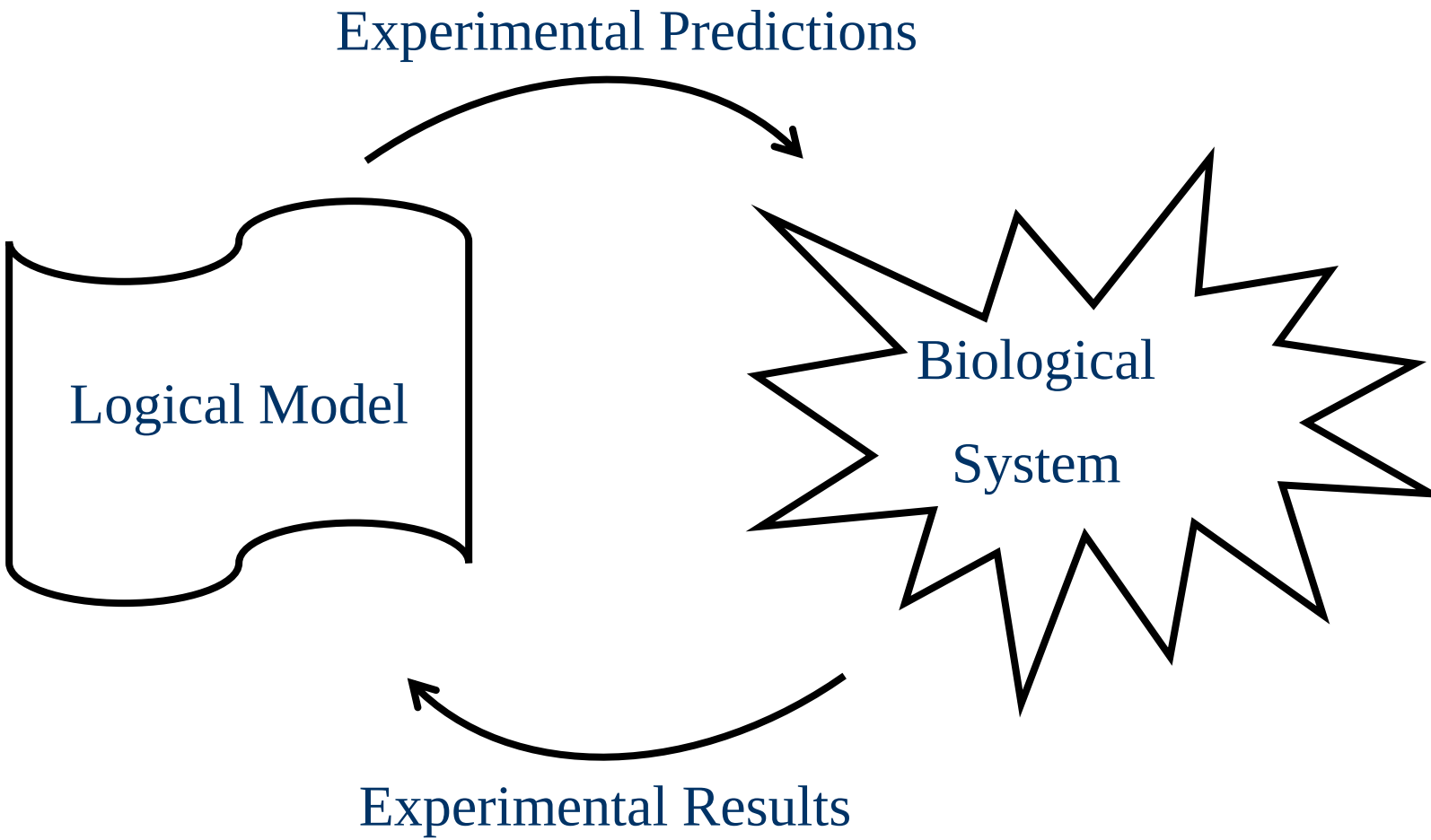
Formalising the Problem

- Use logic programming to represent background knowledge: metabolism modelled as a directed labeled hyper-graph.
- Use abduction to infer new hypotheses:
  - Abductive logic programming.
  - Techniques from Bioinformatics.
- Use active learning to decide efficient experiments: cost of compounds and time.
- Use machine learning to decide meaning of experimental results.

The Experimental Cycle



Model v Real-World

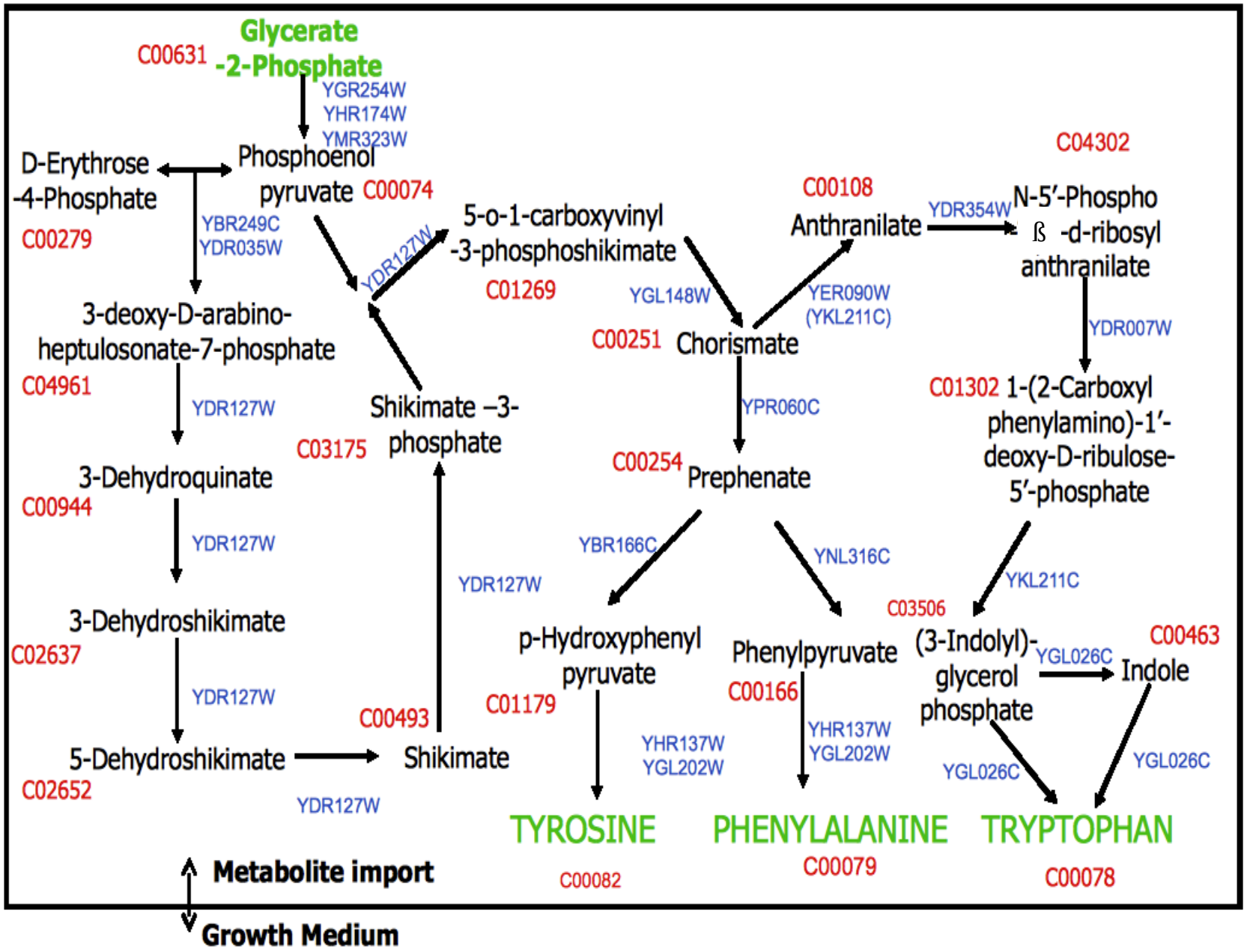


Logical Cell Model

- We have developed a logical formalism for modelling metabolic pathways (encoded in Prolog). This is essentially a directed labeled hyper-graph: with metabolites as nodes and enzymes as arcs.

- If a path can be found from cell inputs (metabolites in the growth medium) to all the cell outputs (essential compounds) then the cell can grow.

Phenyalanine, Tyrosine, and Tryptophan Pathways for *S. cerivisae*



Noise

There were two types of noise in the physical experiments:

- Experiment and measurement noise

(we estimate that ~25% of the experiments are noisy).

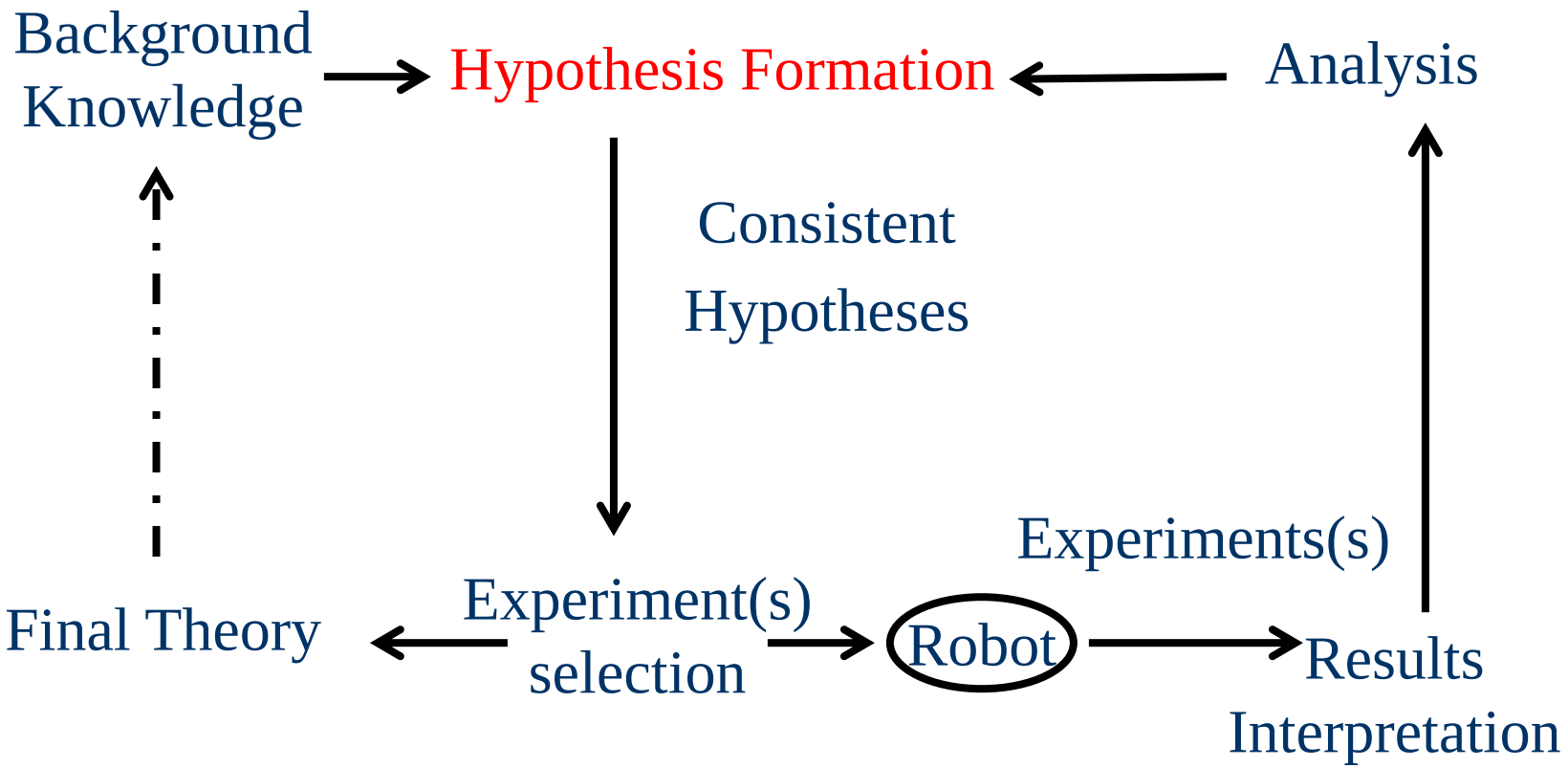
- Noise due to errors in the background knowledge

(the model does not agree with ~2% of experimental results).

Genome Scale Model of Yeast Metabolism

- We have extended our model of aromatic amino acid metabolism to cover most of what is known about yeast metabolism.
- Includes 1,166 ORFs (940 known, 226 inferred)
- Growth if path from growth medium to defined end-points.
- State-of-the-art accuracy in predicting cell viability

The Experimental Cycle



Inferring Hypotheses

- Science is based on the hypothetico-deductive method.
- In the philosophy of science. It has often been argued that only humans can make the "leaps of imagination" necessary to form hypotheses.
- In biology most hypothesis generation is abductive, not inductive.
- Adam used abductive inference to infer missing arcs/labels in its metabolic graph - hypotheses. With these missing nodes Adam could then deductively infer (explain) the observed experimental results.

Types of Logical Inference

Deduction

Rule: All swans are white.
Fact: Daffy is a swan.
∴ Daffy is white.

Abduction

Rule: All swans are white.
Fact: Daffy is white.
∴ Daffy is a swan.

Induction

Fact: Daffy is a swan and white.
Fact: Tweety is a swan and white
∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.
Fact: Daffy is a swan.
∴ Daffy is white.

Abduction

Rule: All swans are white.
Fact: Daffy is white.
∴ Daffy is a swan. Daffy is a duck.

Induction

Fact: Daffy is a swan and white.
Fact: Tweety is a swan and white
∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.
Fact: Daffy is a swan.
∴ Daffy is white.

Abduction

Rule: All swans are white.
Fact: Daffy is white.
∴ Daffy is a swan.

Induction

Fact: Daffy is a swan and white.
Fact: Tweety is a swan and white
∴ All swans are white.

Types of Logical Inference

Deduction

Rule: All swans are white.
Fact: Daffy is a swan.
∴ Daffy is white.

Abduction

Rule: All swans are white.
Fact: Daffy is white.
∴ Daffy is a swan.

Induction

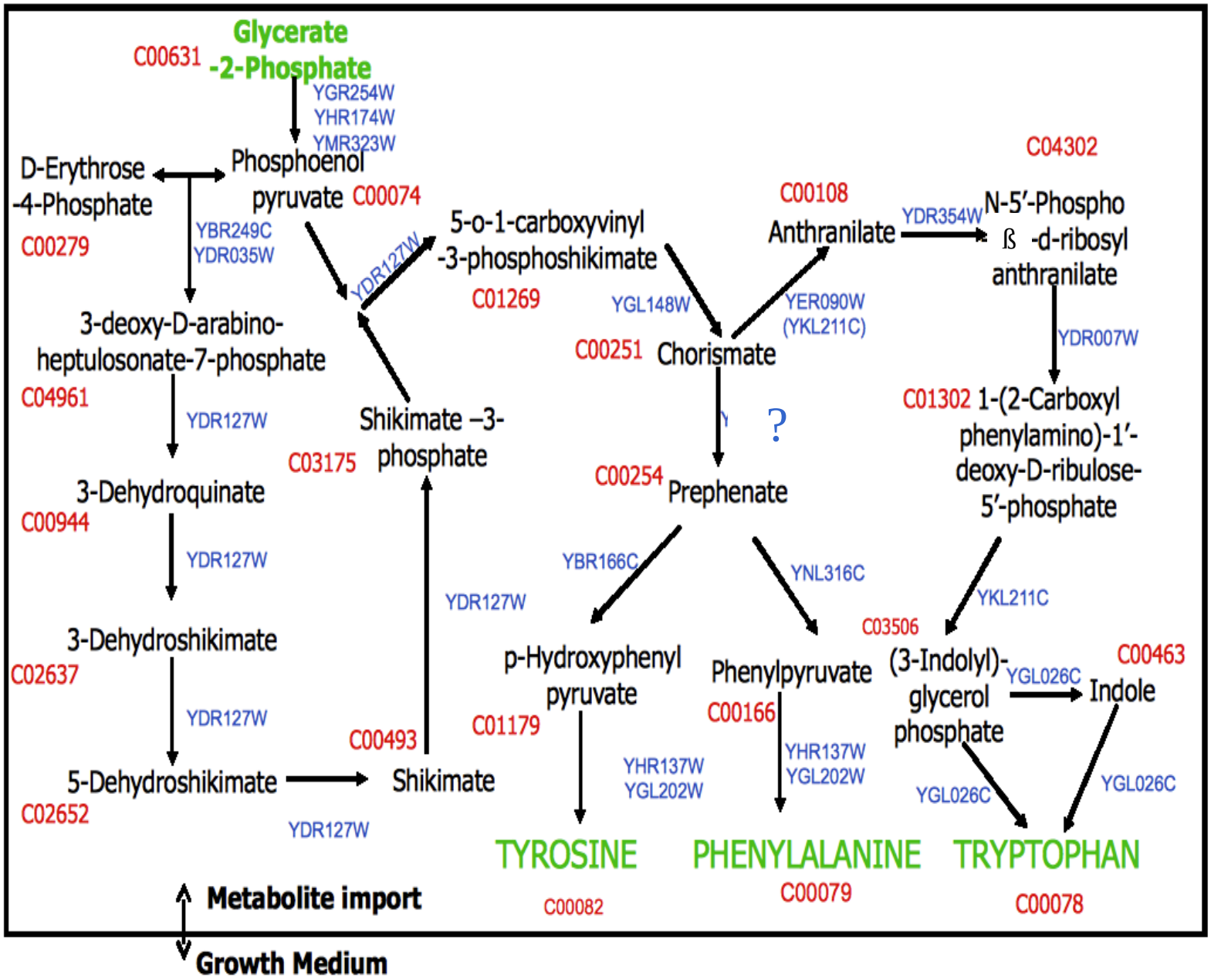
Fact: Daffy is a swan and white.
Fact: Tweety is a swan and white
∴ All swans are white.
Bruce is a black swan.

Orphan Enzymes

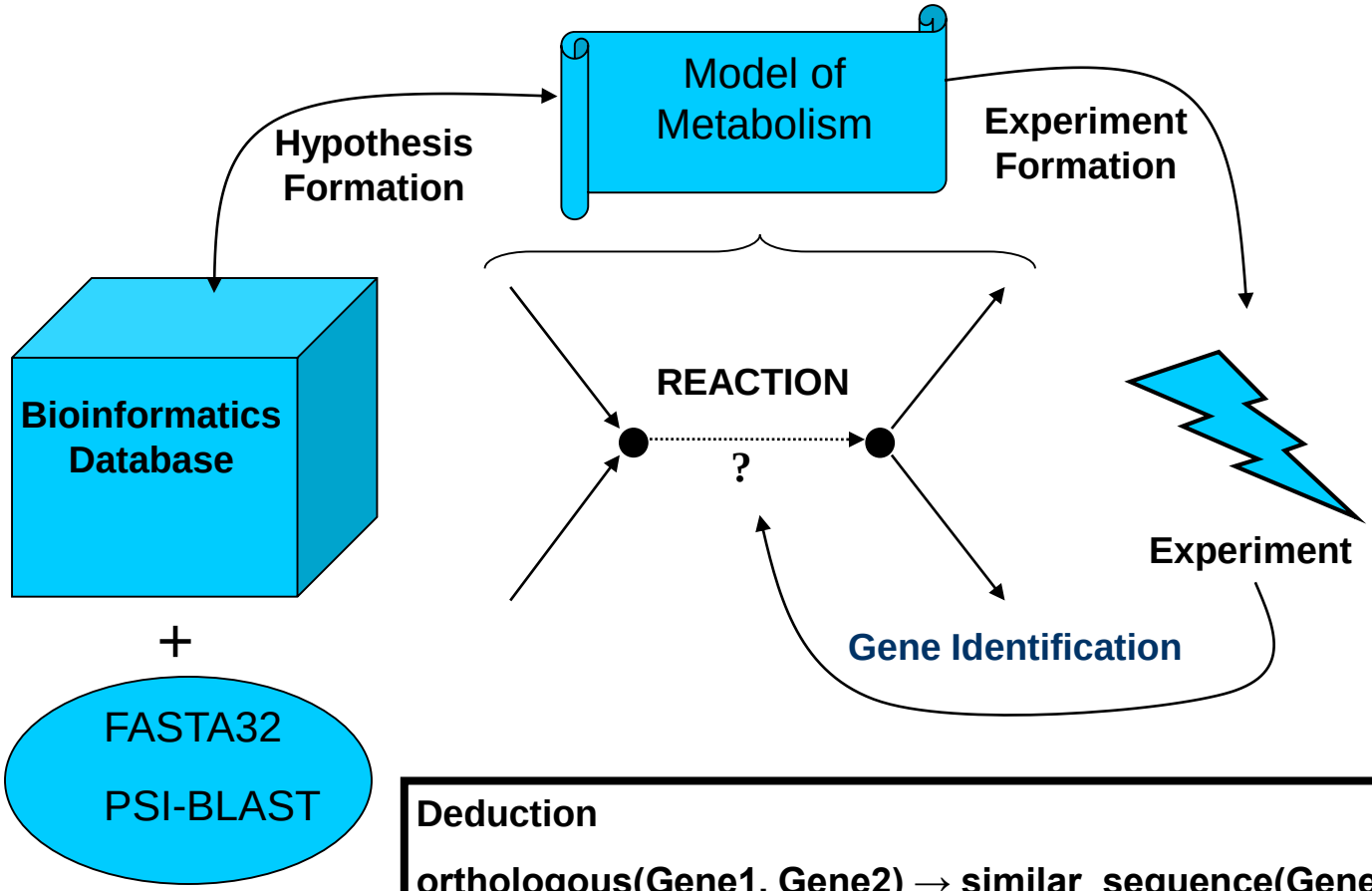
- Our model of yeast metabolism has “orphan enzymes” enzymes which catalyse biochemical reactions known to be in yeast, but which do not have identified parent genes

- We use bioinformatics to abduce genes which encode for these orphan enzymes.

Phenyalanine, Tyrosine, and Tryptophan Pathways for *S. cerivisae*



Automated Model Completion



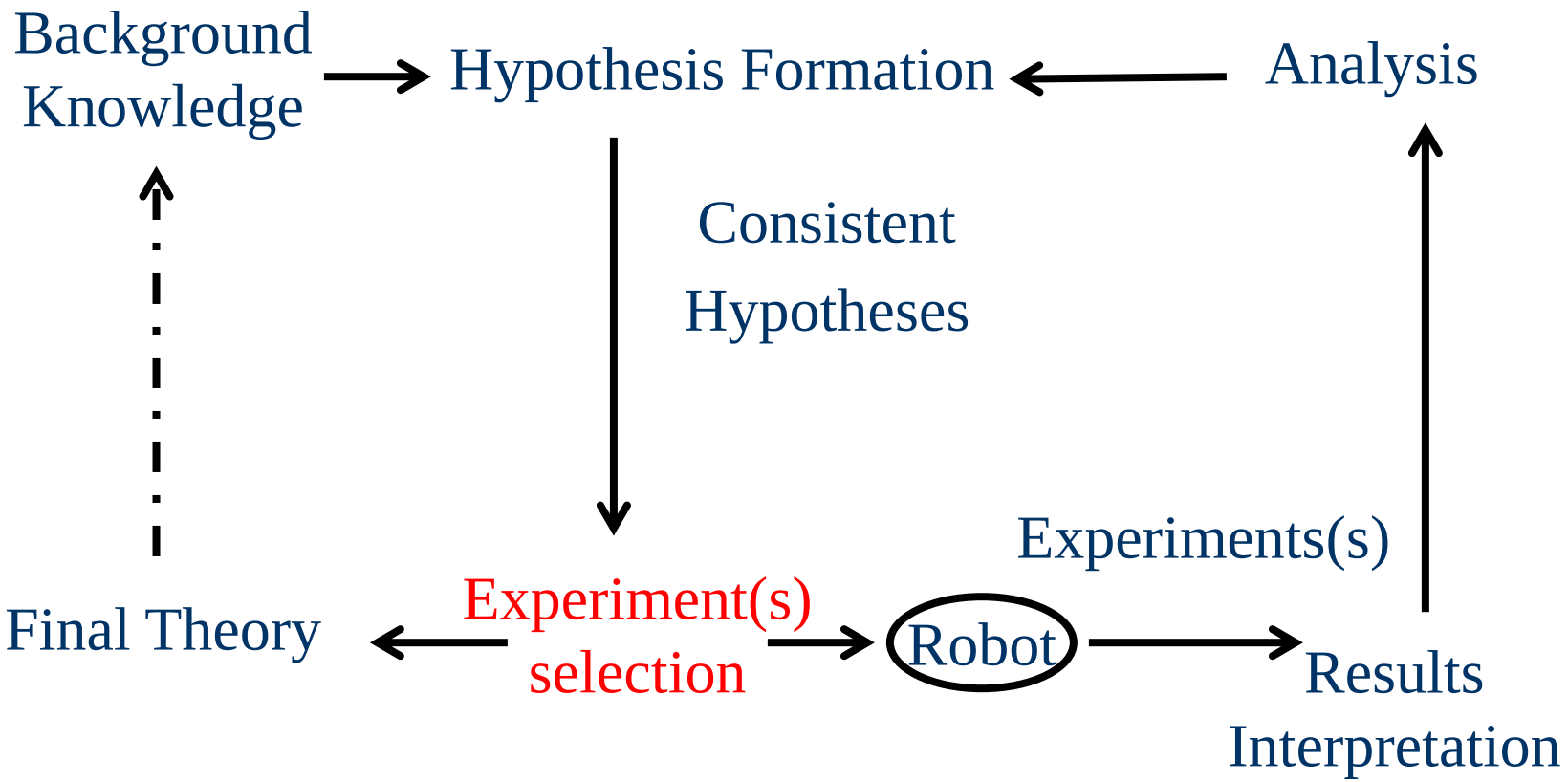
Deduction

orthologous(Gene1, Gene2) → similar_sequence(Gene1, Gene2).

Abduction

similar_sequence(Gene1, Gene2) → orthologous(Gene1, Gene2).

The Experimental Cycle

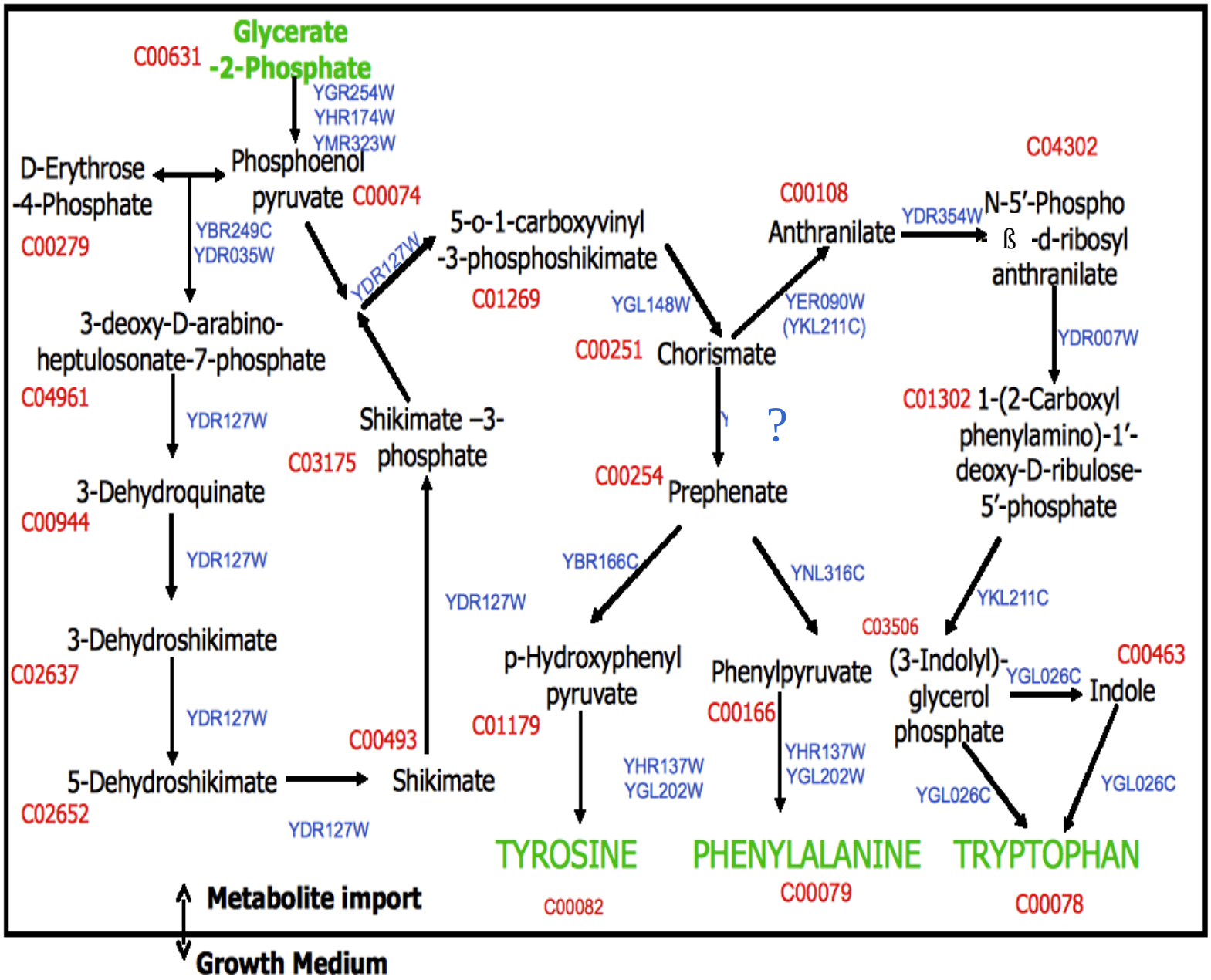


A Discriminating Experiment

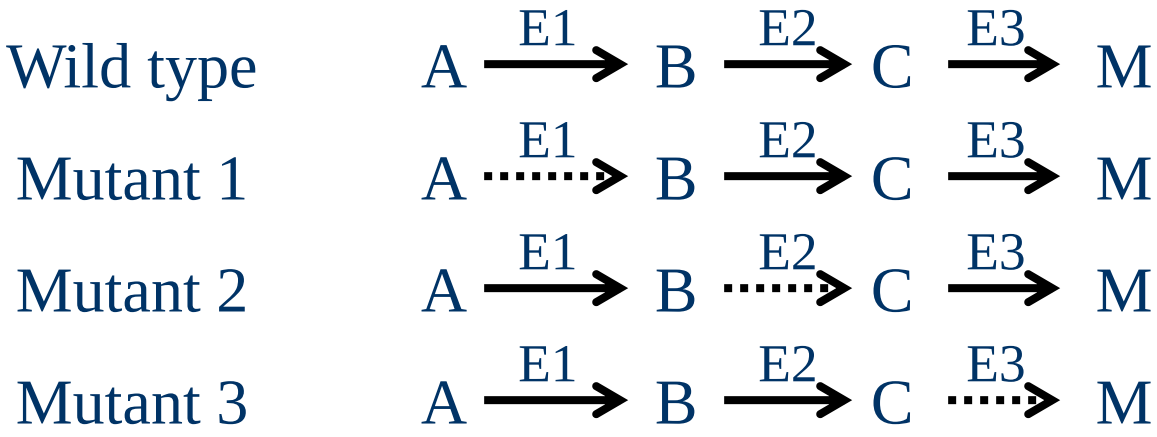
- Hypothesis 1: YDR060C codes for the enzyme the reaction: chorismate $\rightarrow$ prephenate.
- Hypothesis 2: YDR060C codes for the enzyme the reaction: chorismate $\rightarrow$ anthranilate.

- These can be distinguished by growing the knockout YDR060C on prephenate or anthranilate.

Phenyalanine, Tyrosine, and Tryptophan Pathways for *S. cerivisae*



Auxotrophic Experiments



	A	B	C	M
Wild type	+	+	+	+
Mutant 1	-	+	+	+
Mutant 2	-	-	+	+
Mutant 3	-	-	-	+

Experimental Methodology

- Experiments consist of making particular growth media and testing if the mutants can grow (add metabolites to a basic defined medium).
- A mutant is auxotrophic if cannot grow on a defined medium that the wild type can grow on.
- By observing the pattern of chemicals that recover growth the function of the knocked out mutant can be inferred.

Inferring Experiments

Given a set of hypotheses we wish to infer an experiment that will efficiently discriminate between them

Assume:

- Every experiment has an associated cost.
- Each hypothesis has a probability of being correct.

The task:

- To choose a series of experiments which minimise the expected cost of eliminating all but one hypothesis.

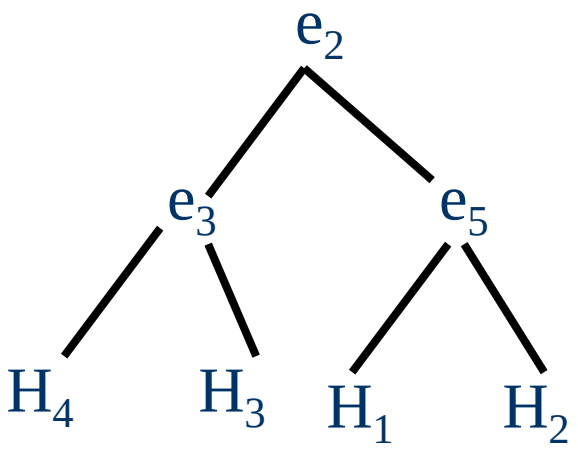
Active Learning

- In the 1972 Fedorov (Theory of optimal experiments) showed that this problem is in general intractable (NP complete).

- However, it can be shown that the problem is the same as finding an optimal decision tree; and it is known that this problem can be solved “nearly” optimally in polynomial time.

How to choose the best experiment

	e_1	e_2	...	e_m
H_1	T	F	...	T
...	F	T	...	F
H_n	F	T	...	T



Choosing the best experiment is equivalent to choosing the best node in a decision tree.

Recurrence Formula

EC(H,T) denote the minimum expected cost of experimentation given the set of candidate hypotheses *H* and the set of candidate trials *T*:

$$EC(\emptyset, T) = 0$$

$$EC(\{h\}, T) = 0$$

$$EC(H, T) = \min_{t \in T} \left[C_t + p(t)(mean_{t' \in (T - t)} C_{t'}) J_{H[t]} + (1 - p(t)) mean_{t' \in (T - t)} C_{t'}) J_{H[\bar{t}]} \right]$$

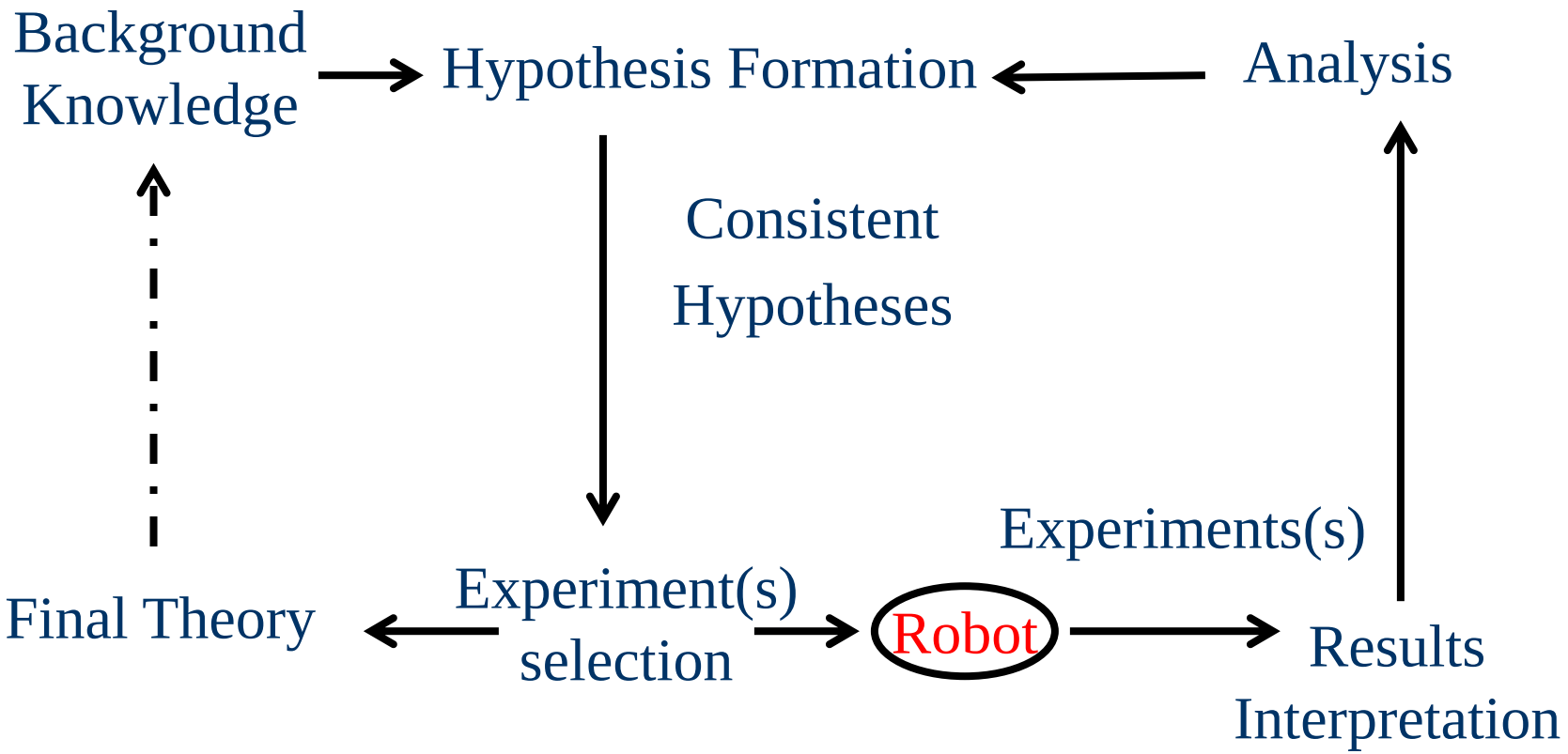
$$J_H = - \sum_{h \in H} p(h) \left[\log_2 (p(h)) \right]$$

C_t is the monetary price of the trial *t*

p(t) is the probability that the outcome of the trial *t* is positive

p(t) can be computed as the sum of the probabilities of the hypotheses (*h*) which are consistent with a positive outcome of *t*

The Experimental Cycle



Adam

- Designed to fully automate yeast growth experiments.
- Had a -20C freezer, 3 incubators, 2 readers, 3 liquid handlers, 3 robotic arms, 2 robot tracks, a centrifuge, a washer, an environmental control system, etc.
- Was capable of initiating ~1,000 new experiments and >200,000 observations per day in a continuous cycle.

Plan of Adam

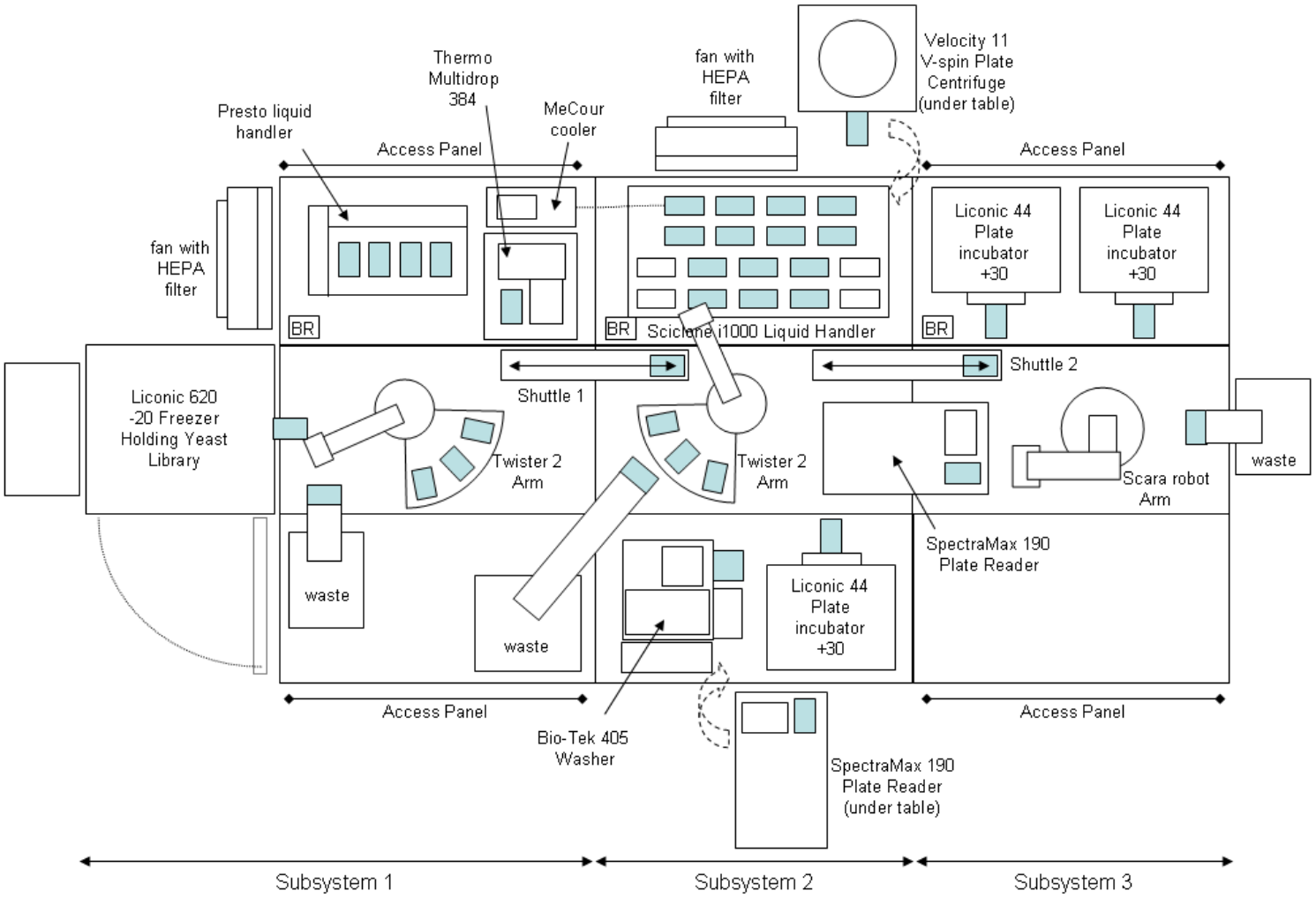
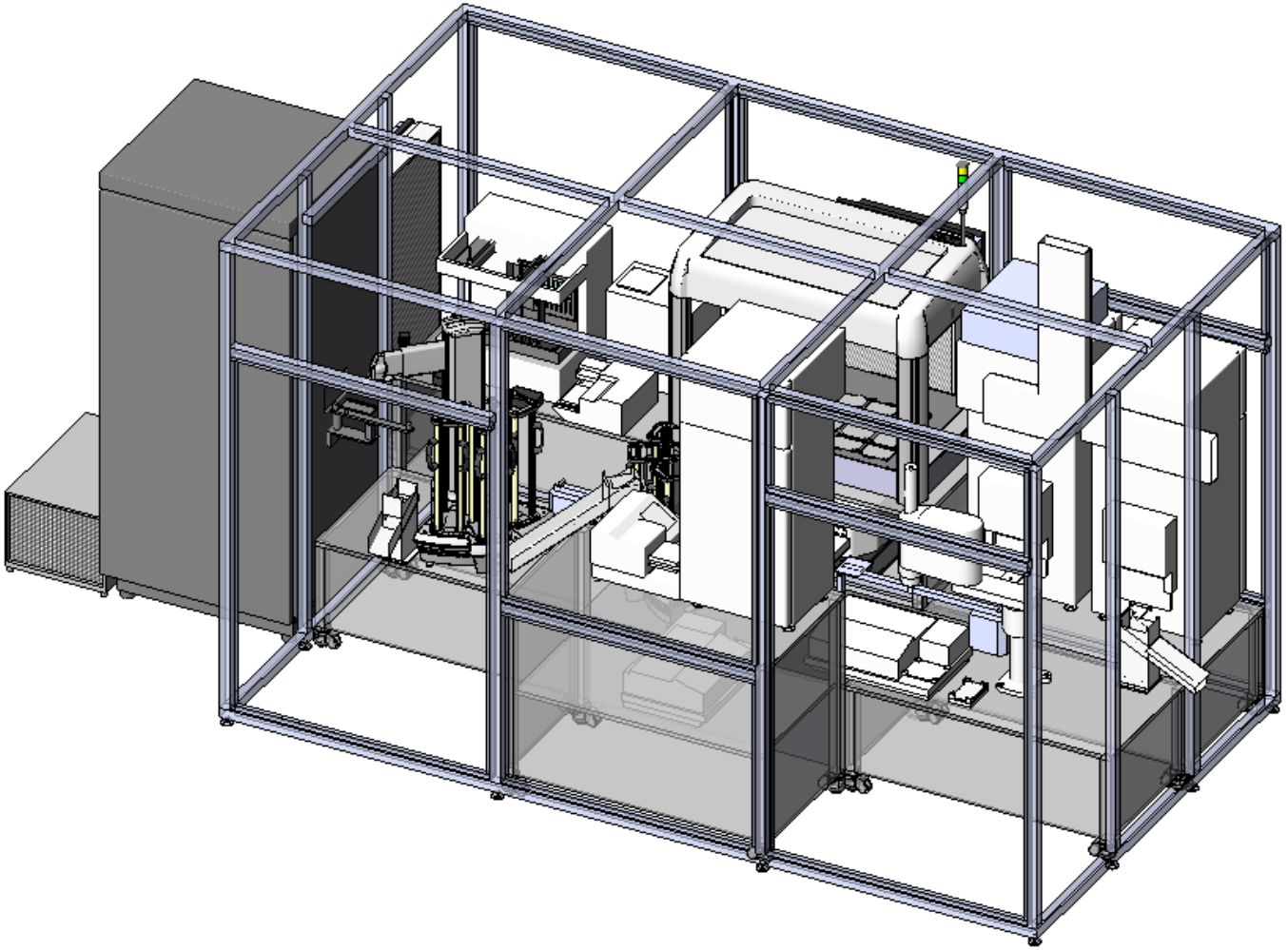
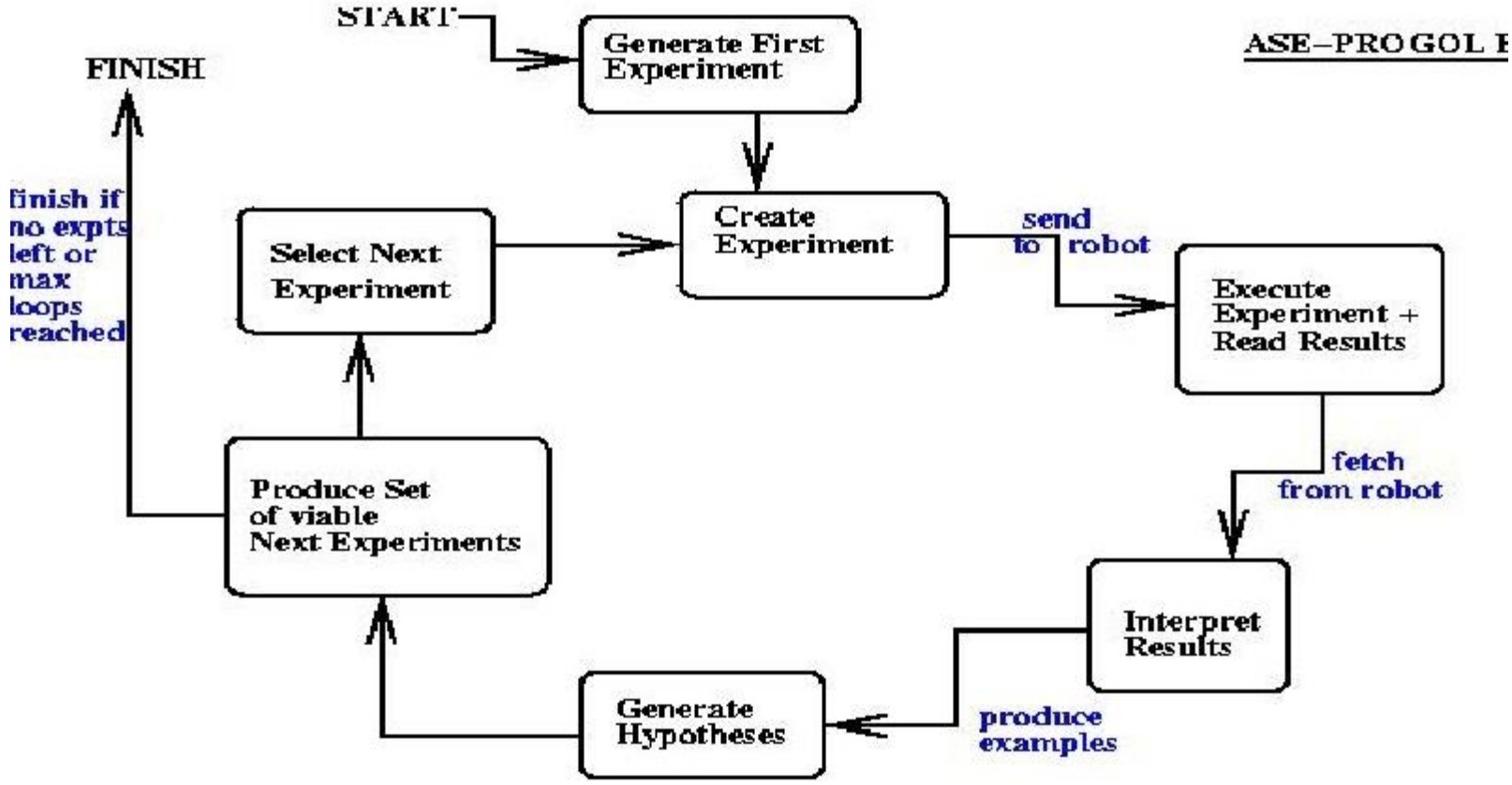

Velocity 11
V-spin Plate
Centrifuge
(under table)
Thermo
Multidrop
384
fan with
HEPA
filter
Presto liquid
handler
MeCour
cooler
Access Panel
Access Panel
fan with
HEPA
filter
Liconic 44
Plate
incubator
+30
Liconic 44
Plate
incubator
+30
BR
BR
Sciclone i1000 Liquid Handler
BR
Shuttle 2
Shuttle 1
Liconic 620
-20 Freezer
Holding Yeast
Library
Twister 2
Arm
Twister 2
Arm
Scara robot
Arm
waste
SpectraMax 190
Plate Reader
waste
waste
Liconic 44
Plate
incubator
+30
Access Panel
Access Panel
Bio-Tek 405
Washer
SpectraMax 190
Plate Reader
(under table)
Subsystem 1
Subsystem 2
Subsystem 3

Diagram of Adam

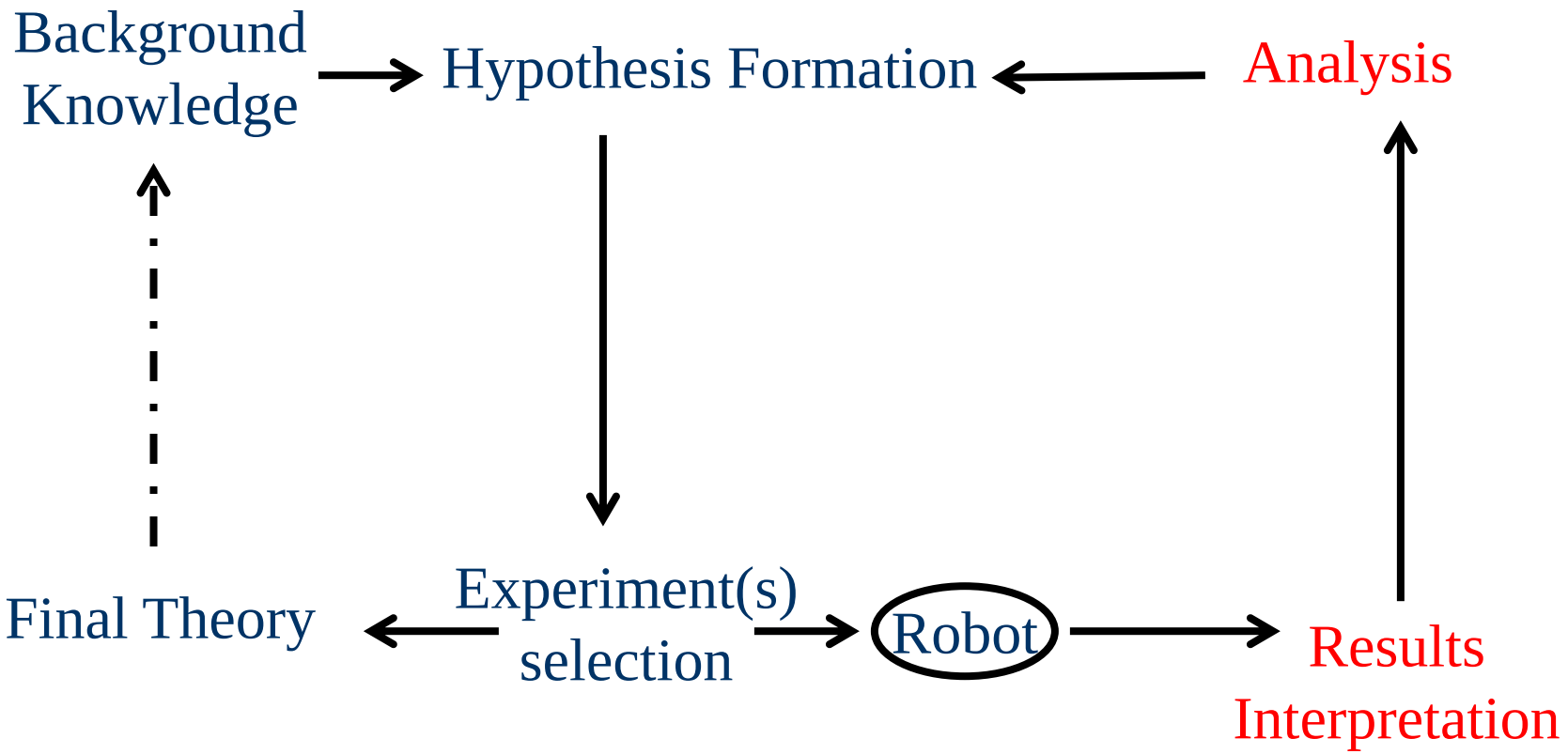


LIMS Setup

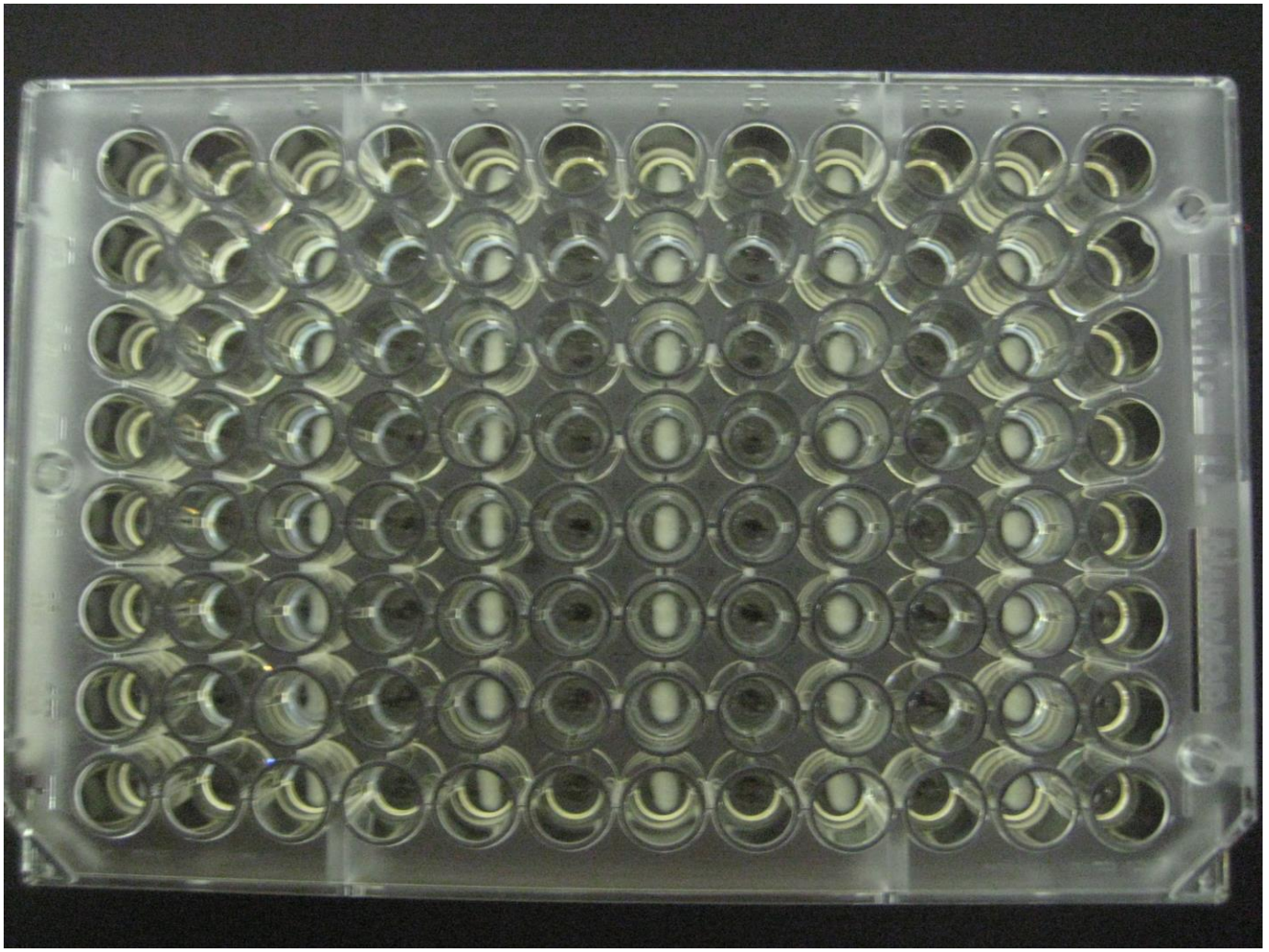

START
Generate First Experiment
ASE–PROGOL
FINISH
finish if no expts left or max loops reached
Select Next Experiment
Create Experiment
send to robot
Execute Experiment + Read Results
fetch from robot
Produce Set of viable Next Experiments
Interpret Results
Generate Hypotheses
produce examples

Adam in Action

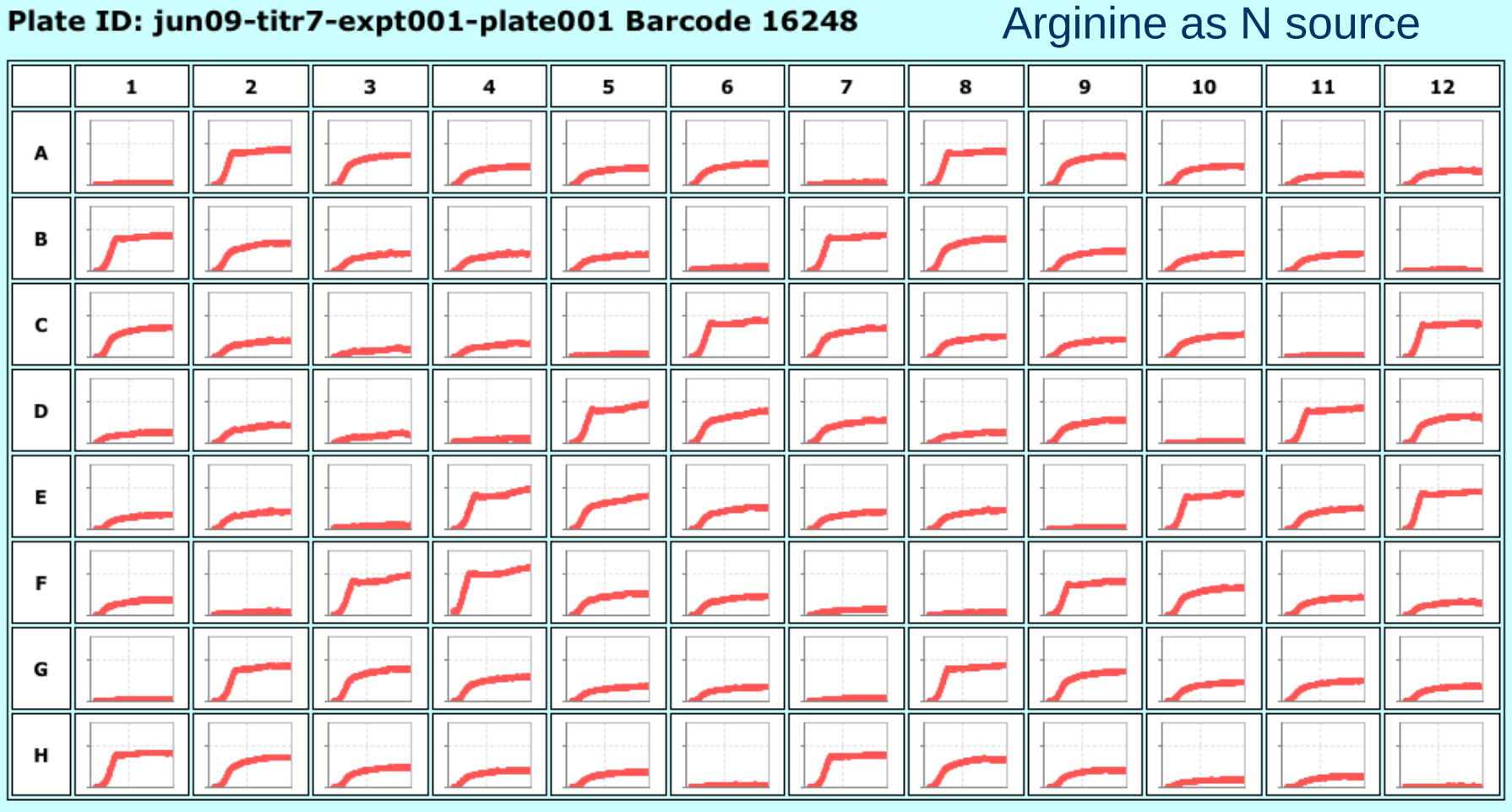
The Experimental Cycle



Growth plates

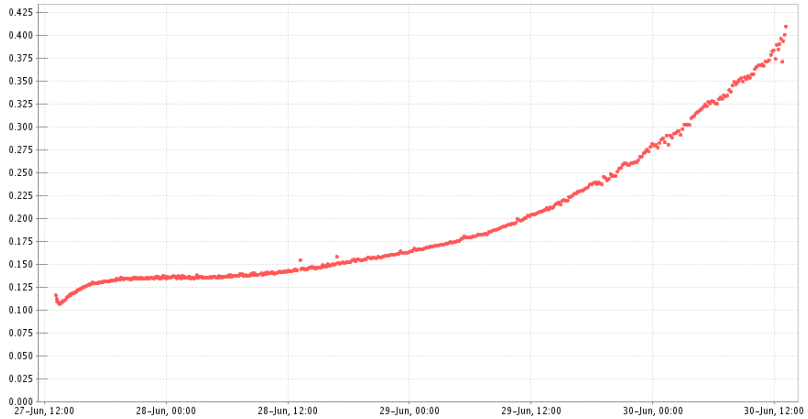


Example Growth Curves



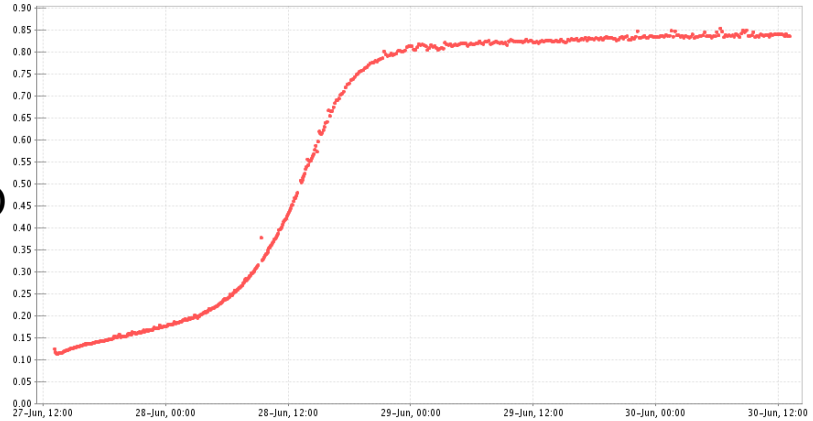
Growth curves

OD



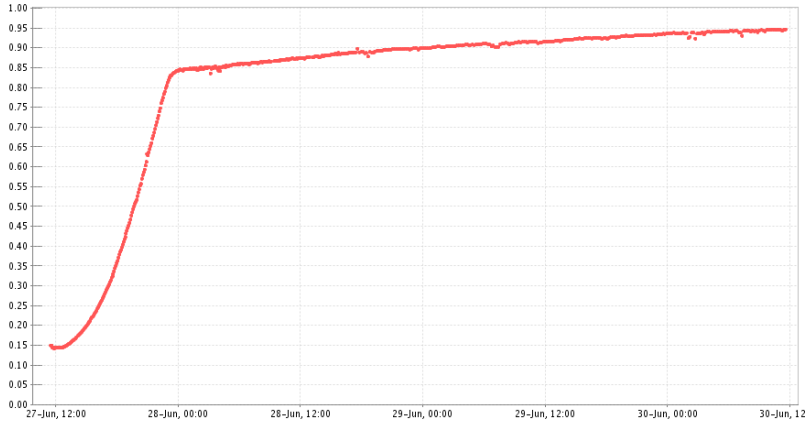
Time

OD



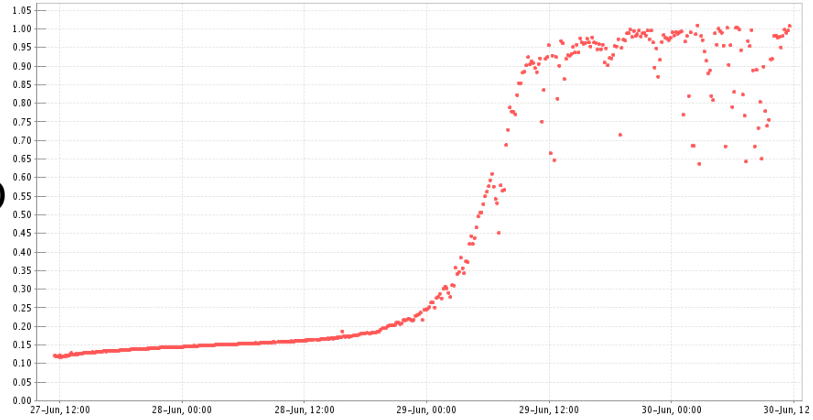
Time

OD



Time

OD



Time

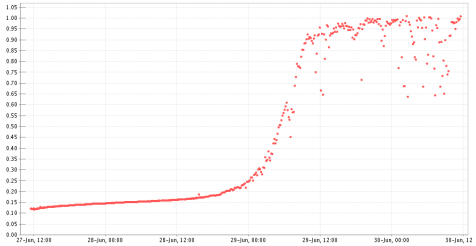
Qualitative to Quantitative

- The functions of most genes in *S. cerevisiae* that when deleted result in auxotrophy (no growth) have already been discovered.
- Most genes of unknown function only affect growth quantitatively.
- They may have slower growth (bradytrophs), faster growth, higher/lower biomass yield, etc.

Test for Growth

- We needed a reliable and automatic method to decide whether growth of a mutant strain had occurred or not.
- Controls - positive (wild type) and negative (no yeast)
- Trained a decision tree to fit predictions of model, attributes aggregates of measurements.
- Simple tree with only a few leafs

Results

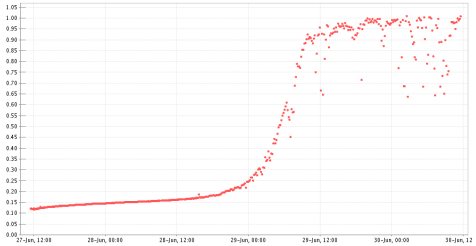


Spline based curve fitting, smoothing

Curve parameters
Lag time: 3.5
Maximum OD: 1.4
Growth rate: 0.44

Random forests,
Statistical tests

Results



Spline based curve fitting, smoothing

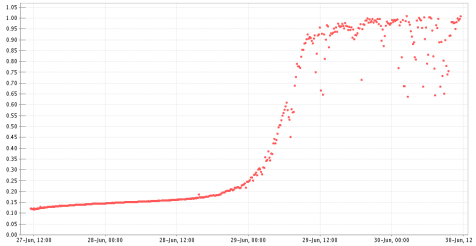
Curve parameters
Lag time: 3.5
Maximum OD: 1.4
Growth rate: 0.44

Random forests,
Statistical tests

Two factor design:

Wildtype	Deletant
Wildtype with metabolite	Deletant with metabolite

Results



Spline based curve fitting, smoothing

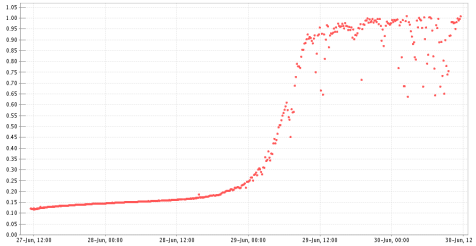
Curve parameters
Lag time: 3.5
Maximum OD: 1.4
Growth rate: 0.44

Random forests, Statistical tests

Two factor design:

Wildtype	Deletant
Wildtype with metabolite	Deletant with metabolite

Results

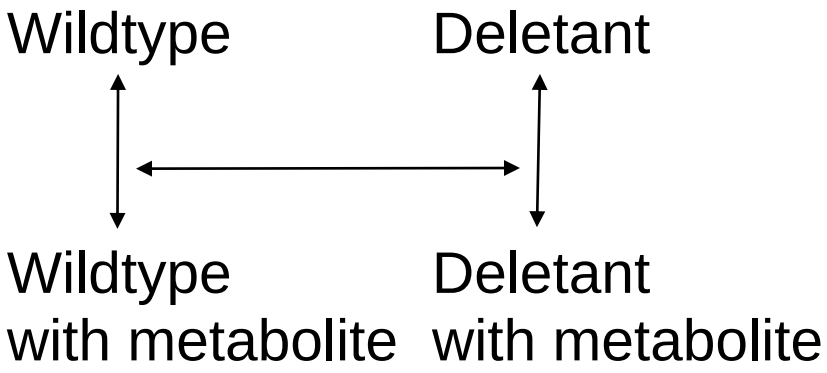


Spline based curve fitting, smoothing

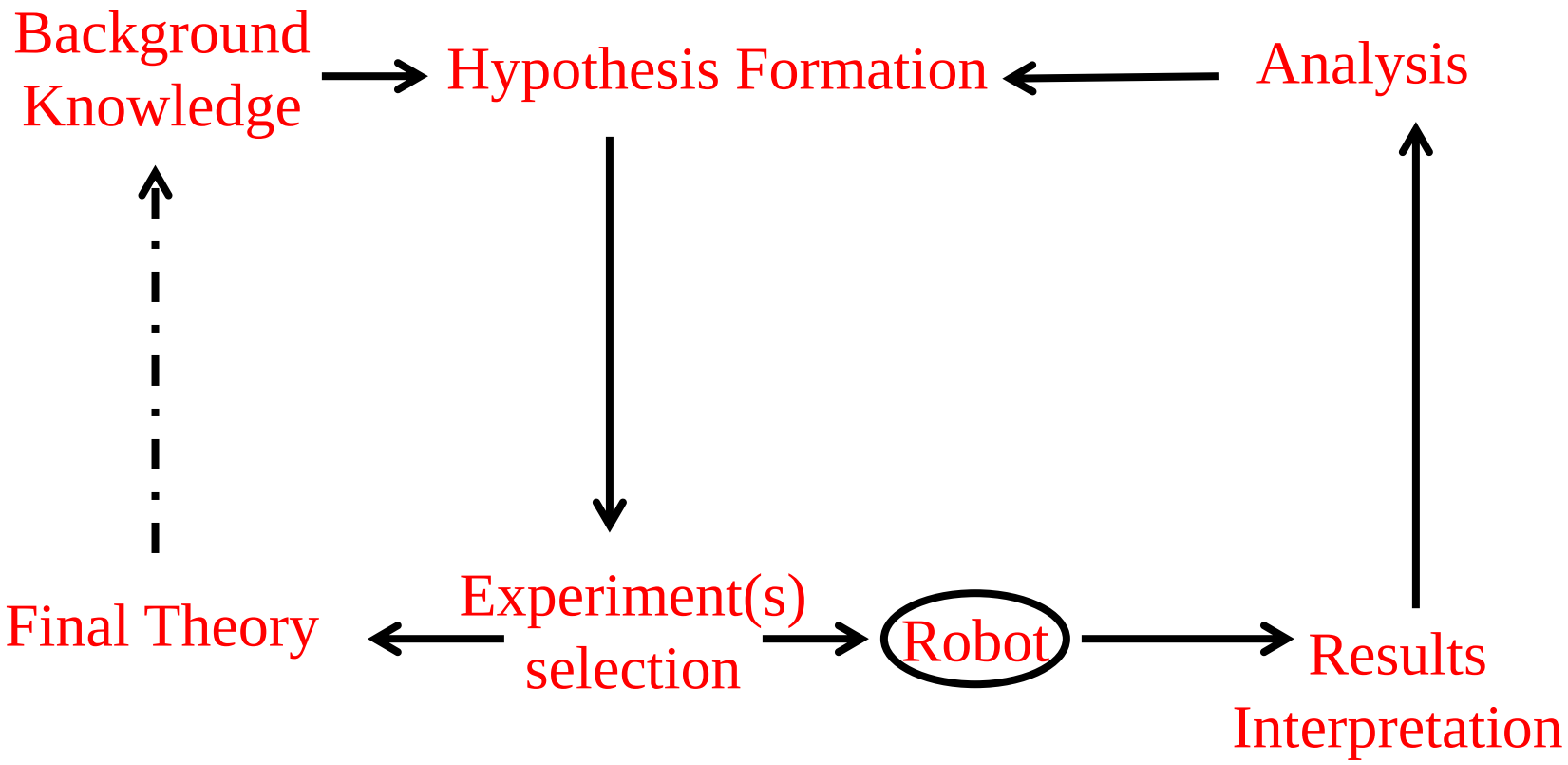
Curve parameters
Lag time: 3.5
Maximum OD: 1.4
Growth rate: 0.44

Random forests,
Statistical tests

Two factor design:



The Experimental Cycle



Closing the Loop

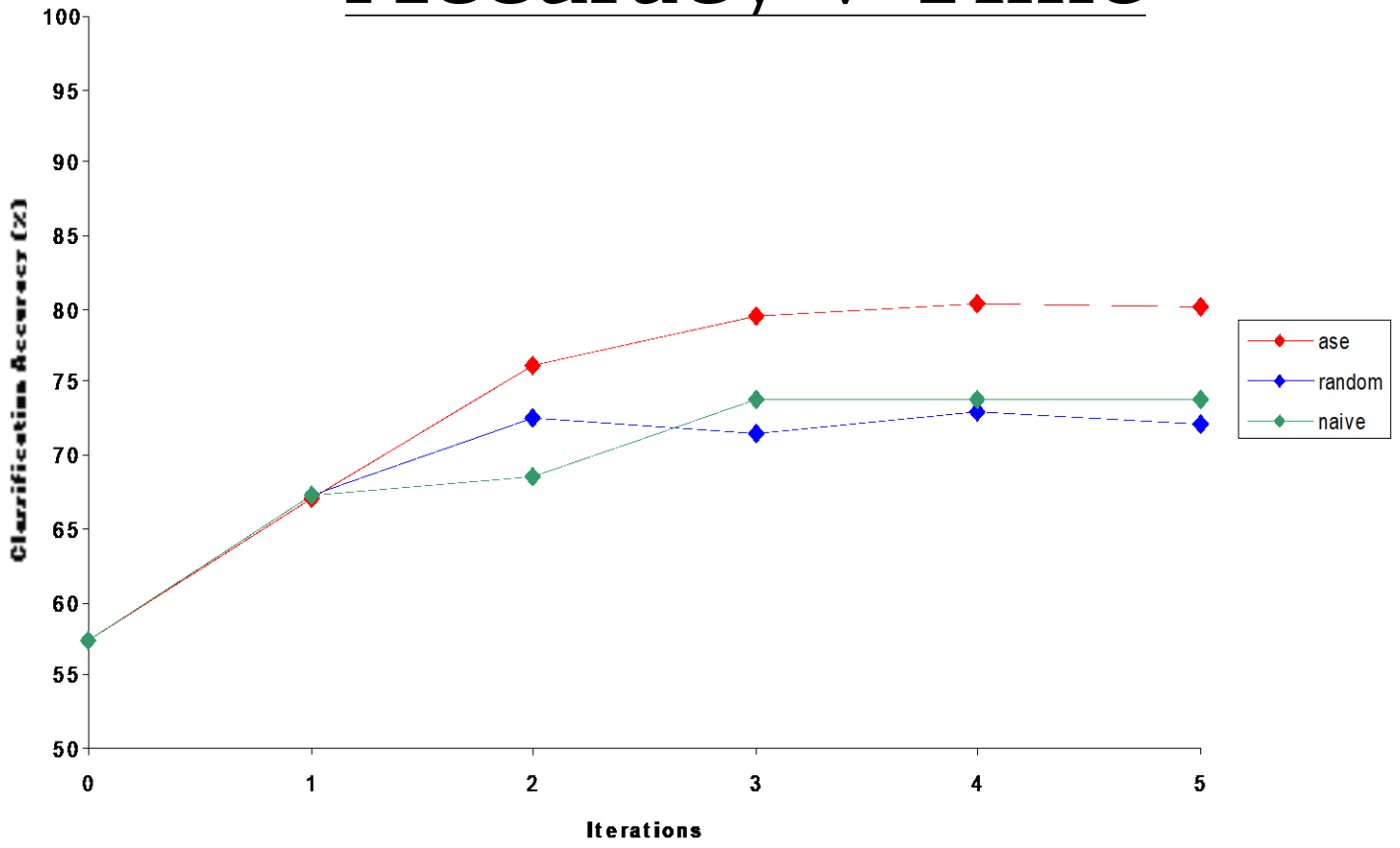
- We physically implemented all aspects of Adam.
- To the best of our knowledge Adam was the first AI system that can both explicitly form hypotheses and experiments, and physically do the experiments.

Time and Money

- “Cost” is a positive function of time & money.
- ASE dominates for both, therefore ASE dominates for any reasonable cost function.

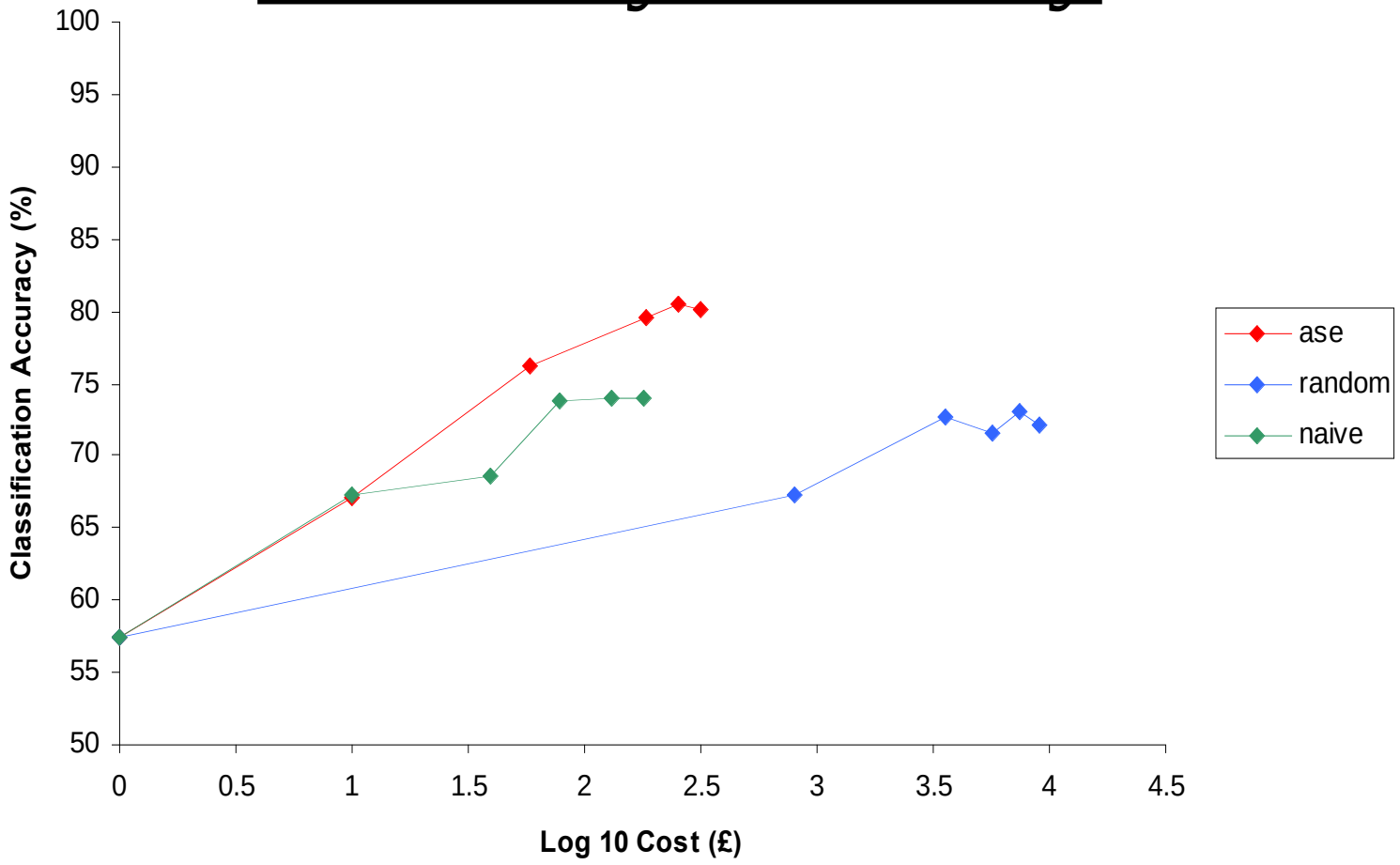
- For example: to achieve an accuracy of ~70%, ASE requires fewer trial iterations, and a hundredth of the price, of Random; and almost half the number of iterations, and a third of the price of Naïve.

Accuracy v Time



At the end of the 5th iteration: ASE 80.1%, Naïve 74.0%, Random 72.2%. ASE was significantly more accurate than either Naïve ($p < 0.05$) or Random ($p < 0.07$) using a paired t-test.

Accuracy v Money



Given a spend of ≤£$10^{2.26}$, ASE 79.5%, Naïve 73.9%, Random 57.4%. ASE was significantly more accurate than either Naïve ($p < 0.05$) or Random ($p < 0.001$).

Human Comparisons

- We were interested to compare the performance of the Robot Scientist with that of humans.
- We adopted the simulator to allow humans to chooses and interpret the results of cycles of experimentation.
- Compared nine graduate computer scientists and biologists.

- No significant difference between the best humans and the Robot

Discovery of Novel Science

Novel Science

- Adam generated and confirmed novel functional-genomics hypotheses concerning the identify of genes encoding enzymes catalysing orphan reactions in the metabolic network of the yeast *S. cerevisiae.*

- Adam's conclusions have been manually verified using bioinformatic and biochemical evidence.

Novel Scientific Knowledge

Orphan Enzyme		Hypothesised Gene	Prob.	Acc.	No.	Existing Annotation	Dry	Wet
1	glucosamine-6-phosphate deaminase (3.5.99.6)	YHR163W (SOL3)	$<10^{-4}$	97	8	'6-phosphogluconolactonase' ida	-	-
2	glutaminase (3.5.1.2)	YIL033C (BCY1)	$<10^{-4}$	92	11	'cAMP-dependent protein kinase inhibitor ' ida	✗ ?	-
3	L-threonine 3-dehydrogenase (1.1.1.103)	YDL168W (SFA1)	$<10^{-4}$	83	6	'alcohol dehydrogenase' ida	-	-
4	purine-nucleoside phosphorylase (2.4.2.1)	YLR209C (PNP1)	$<10^{-4}$	82	11	'purine-nucleoside phosphorylase' ida	✓	-
5	2-aminoadipate transaminase (2.6.1.39)	YGL202W (ARO8)	$<10^{-4}$	80	3	'aromatic-amino-acid transaminase' ida	✓	✓
6	5,10-methenyltetrahydrofolate synthetase (6.3.3.2)	YER183C (FAU1)	$<10^{-4}$	80	4	'5,10 formyltetrahydrofolate cyclo-ligase' ida	✓	-
7	glucosamine-6-phosphate deaminase (3.5.99.6)	YNR034W (SOL1)	$<10^{-4}$	79	2	'possible role in tRNA export'	-	-
8	pyridoxal kinase (2.7.1.35)	YPR121W (THI22)	$<10^{-4}$	78	1	'phosphomethylpyrimidine kinase' iss	-	-
9	mannitol-1-phosphate 5-dehydrogenase (1.1.1.17)	YNR073C	$<10^{-4}$	78	6	'putative mannitol dehydrogenase ' iss	-	-
10	1-acylglycerol-3-phosphate O-acyltransferase (2.3.1.51)	YDL052C (SLC1)	0.0001	80	6	'1-acylglycerol-3-phosphate O-acyltransferase' ida	✓	-
11	glucosamine-6-phosphate deaminase (3.5.99.6)	YGR248W (SOL4)	0.0002	78	2	'6-phosphogluconolactonase' ida	-	-
12	maleylacetoacetate isomerase (5.2.1.2)	YLL060C (GTT2)	0.0003	76	3	'glutathione S-transferase' ida	-	-
13	serine O-acetyltransferase (2.3.1.30)	YJL218W	0.0005	78	2	'unknown function'	-	-
14	L-threonine 3-dehydrogenase (1.1.1.103)	YLR070C (XYL2)	0.0052	75	6	'xylitol dehydrogenase' ida	-	-
15	2-aminoadipate transaminase (2.6.1.39)	YJL060W (BNA3)	0.0084	73	3	'kynurenine aminotransferase' ida	-	✓
16	pyridoxal kinase (2.7.1.35)	YNR027W	0.0259	76	2	'involved in bud-site selection' iss	-	-
17	polyamine oxidase (1.5.3.11)	YMR020W (FMS1)	0.0289	78	4	'polyamine oxidase' ida	✓	-
18	2-aminoadipate transaminase (2.6.1.39)	YER152C	0.0332	74	3	'uncharacterized'	-	✓
19	L-aspartate oxidase (1.4.3.16)	YJL045W	0.1300	72	1	'succinate dehydrogenase isozyme' iss	-	-
20	purine-nucleoside phosphorylase (2.4.2.1)	YLR017W (MEU1)	0.1421	72	6	'methylthioadenosine phosphorylase' ida	✓	-

A 50 Year Old Puzzle

- The enzyme 2-aminoadipate: 2-oxoglutarate aminotransferase was a locally orphan enzyme.
- It is in the lysine biosynthesis pathway which has been studied for 50 years in fungi: target for antibiotics, and on path to penicillin.

- Adam formed three hypotheses for the gene to encode this enzyme: YER152C, YJL060W, and YGL202W (in that order of probability).

- Currently KEGG states that YGL202W is the gene.
- Evidence from 1960's that at least 2 isoenzymes are involved.

Confirmed New Knowledge

- Adam's differential growth experiments were consistent with all three genes encoding 2-oxoglutarate aminotransferase.
- Manual experiments: purified protein + enzyme assays are consistent.
  - YGL202W literature confirmed.
  - YJL060W (was annotated as an arylformamidase, new (08) annotation kynurenine aminotransferase)
  - YER152C (currently not annotated)
- YGL202W & YJL060W double knockout is lethal

Formalising Science

Formalisation of Science

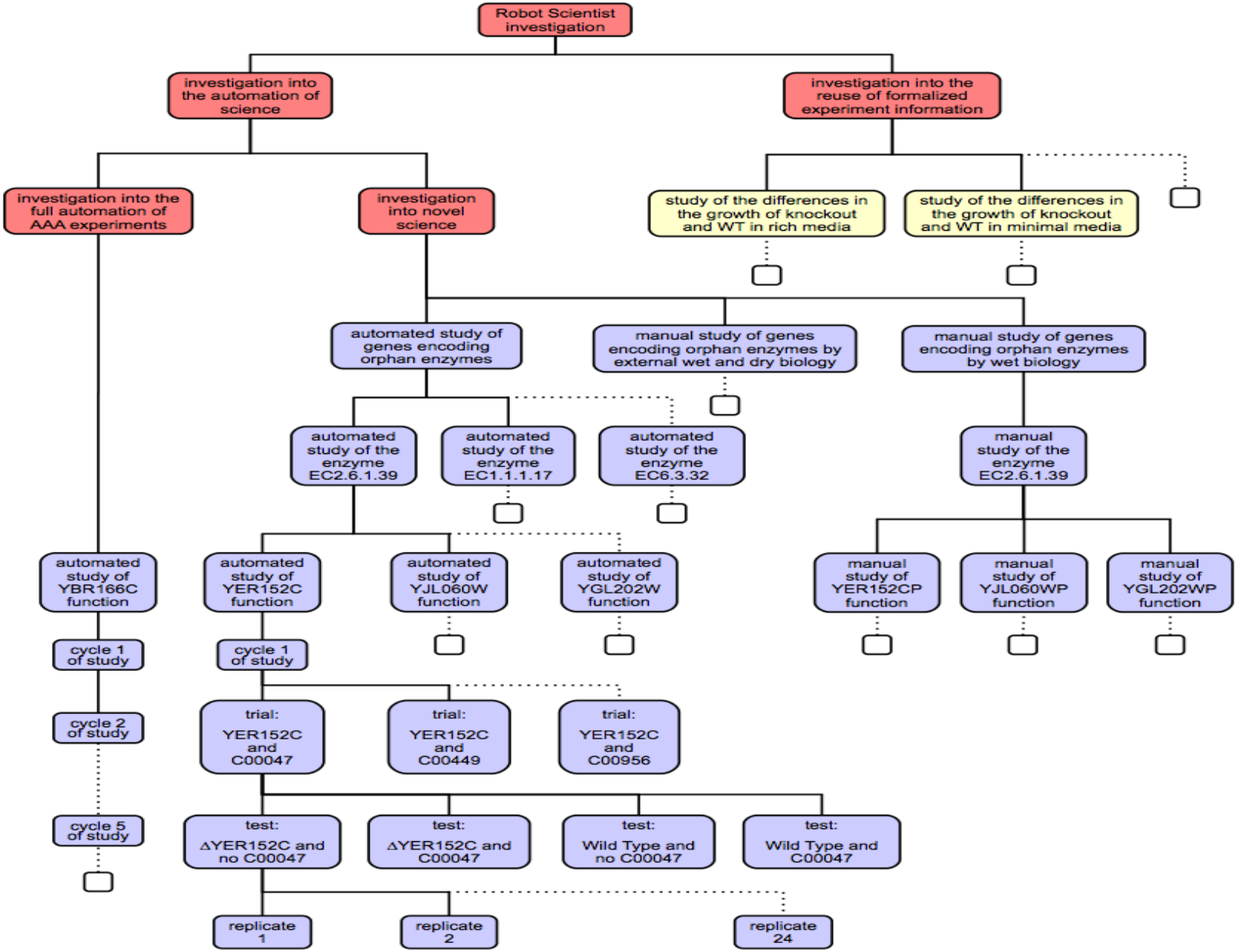
- The goal of science is to increase our knowledge of the natural world through the performance of experiments.
- This knowledge should be expressed in *formal logical languages*.
- Formal languages promote semantic clarity, which in turn supports the free exchange of scientific knowledge and simplifies scientific reasoning.

Robot Scientist & Formalisation

- Robot Scientists provide excellent test-beds for the development of methodologies for formalising science.
- Using them it is possible to completely capture and digitally curate all aspects of the scientific process.
- The ontology LABORS is designed to enable the open access of the Robot Scientist experimental data and metadata to the scientific community.

Adam's Investigations

- This formalisation involves >10,000 different research units in a nested tree-like structure 11 levels deep.
- It logically connects >6.6 million $OD600_{nm}$ measurements to hypotheses, experimental goals, results, etc.
- No previous large-scale experimental work has been so comprehensively described and recorded.


Robot Scientist investigation
investigation into the automation of science
investigation into the reuse of formalized experiment information
investigation into the full automation of AAA experiments
investigation into novel science
study of the differences in the growth of knockout and WT in rich media
study of the differences in the growth of knockout and WT in minimal media
automated study of genes encoding orphan enzymes
manual study of genes encoding orphan enzymes by external wet and dry biology
manual study of genes encoding orphan enzymes by wet biology
automated study of the enzyme EC2.6.1.39
automated study of the enzyme EC1.1.1.17
automated study of the enzyme EC6.3.32
manual study of the enzyme EC2.6.1.39
automated study of YBR166C function
automated study of YER152C function
automated study of YJL060W function
automated study of YGL202W function
manual study of YER152CP function
manual study of YJL060WP function
manual study of YGL202WP function
cycle 1 of study
cycle 2 of study
cycle 5 of study
cycle 1 of study
trial: YER152C and C00047
trial: YER152C and C00449
trial: YER152C and C00956
test: ΔYER152C and no C00047
test: ΔYER152C and C00047
test: Wild Type and no C00047
test: Wild Type and C00047
replicate 1
replicate 2
replicate 24

Levels in the Formalisation

Investigation into the automation of Science

Investigation into the automation of novel science

Investigation into the automated discovery of genes encoding orphan enzymes

Automated study of E.C.2.6.1.39 encoding

Cycle 1 of automated study of YER152C function

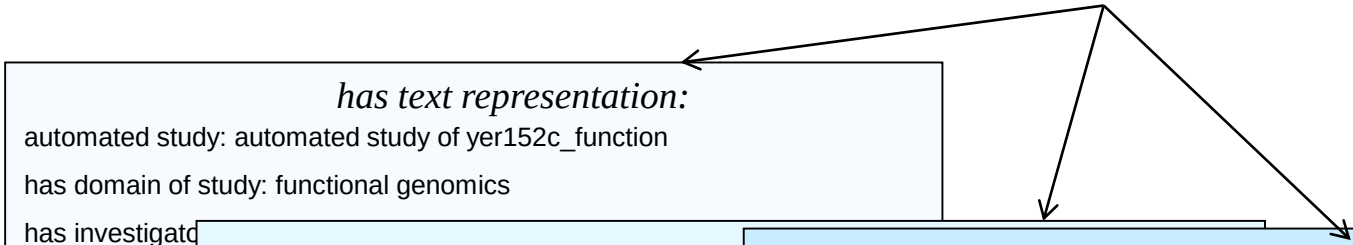
YER152C and Lysine automated trial

Experiment 1 (wild-type no metabolite)

Replicate 1 (well)

Observation 1

automated study of yer152c function



has text representation:

automated study: automated study of yer152c_function

has domain of study: functional genomics

has investigato

has goal: 'To te
with enzyme cl

has organism

has ncbi taxo

has hypothese

has researcl

has negative

has cycle 1 of

has study resu
encodes(yer15

highest

proporti

has study cond

has datalog representation:

a:automated study(X) :- a:automated_study_(
a:hypotheses-set(X) :- a:research_hypothesis
a:cycle_of_study(X) :- a:cycle_1_of_study_(X
a:hypotheses-set(X) :- a:negative_hypothesis
a:domain_of_study(Y) :- a: automated study(
a:investigator(Y) :- a: automated study(X), a:h
a:goal(Y) :- a: automated study(X), a:has_goa
a:organism_of_study (Y) :- a: automated stud
a:hypotheses-set(Y) :- a: automated study(X)
a:cycle_of_study(Y) :- a: automated study(X),
a:study_result(Y) :- a: automated study(X), a:
a:study_conclusion(Y) :- a: automated study(
a:domain_of_study(X) :- a:functional_genomi
a: investigator(X) :- a:adam.
a:goal(X) :- a: to_test_the_hypothesis_that_g
_encodes_an_enzyme_with_enzyme_class_
a:organism_of_study(X) :- a:saccharomyces_
a:study_result(X) :- a:the_strength_of_eviden
a:study_conclusion(X) :- a:hypothesis_1_con

has OWL representation:

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns="http://www.owl-ontologies.com/Ontology1204198571.owl#"
<owl:Class rdf:ID="goal"/>
  <owl:Class rdf:ID="study_result"/>
  <owl:Class rdf:ID="ncbi_taxonomy_ID"/>
  <owl:Class rdf:ID="cycle_of_study"/>
  <owl:Class rdf:ID="negative_hypothesis">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="hypotheses-set"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="domain_of_study"/>
  <owl:Class rdf:ID="organism_of_study"/>
  <owl:Class rdf:ID="cycle_1_of_study_">
    <rdfs:subClassOf rdf:resource="#cycle_of_study"/>
  </owl:Class>
  <owl:Class rdf:ID="automated_study">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom rdf:resource="#goal"/>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="has_goal"/>
        </owl:onProperty>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:someValuesFrom rdf:resource="#organism_of_study"/>
        <owl:onProperty>
          <owl:ObjectProperty rdf:ID="has_organism_of_study"/>
……..
```

Eve

Drug Design

Parasitic Diseases targeted

Malaria

Shistosomaisis

Leishmania

Chagas

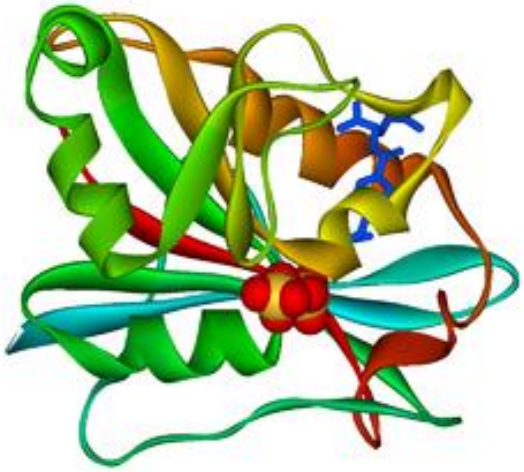
Why Tropical Diseases?

- Millions of people die of these diseases, and hundreds of millions of people suffer infection.
- It is clear how to cure these diseases – kill the parasites.
- They are “neglected”, so avoid competition from the Pharmaceutical industry.

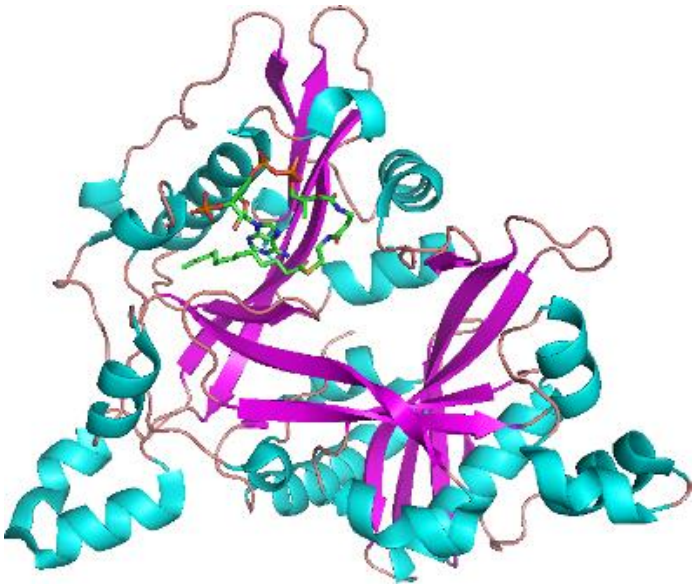
Parasites targeted

- Organism/disease:
  - *Plasmodium falciparum* (malaria),
  - *Plasmodium vivax* (malaria),
  - *Trypanosoma brucei* (sleeping sickness),
  - *Trypanosoma cruzi* (Chagas),
  - *Leishmania major* (leishmania),
  - *Schistosoma mansoni* (shistosomaisis),
  - Staphylococcus aureus
  - …

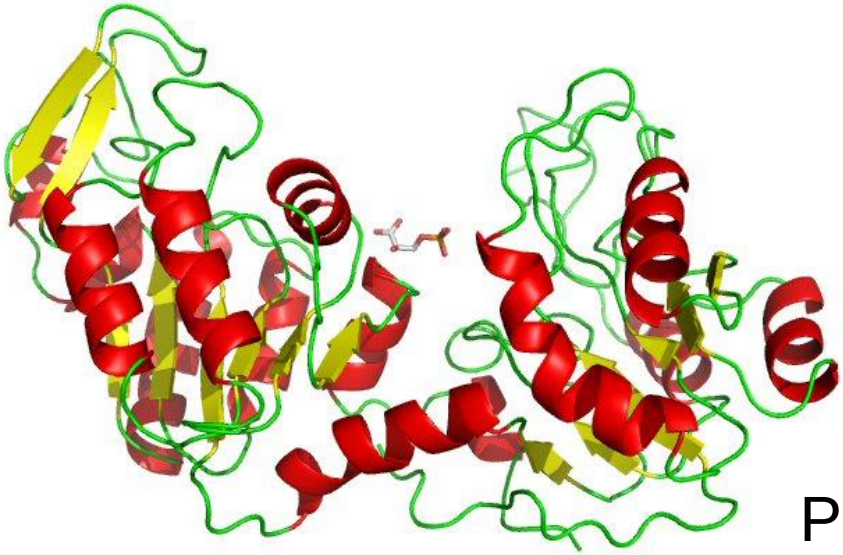
Enzymes Targeted



Dihydrofolate Reductase (DHFR)



N-myristoyl transferase

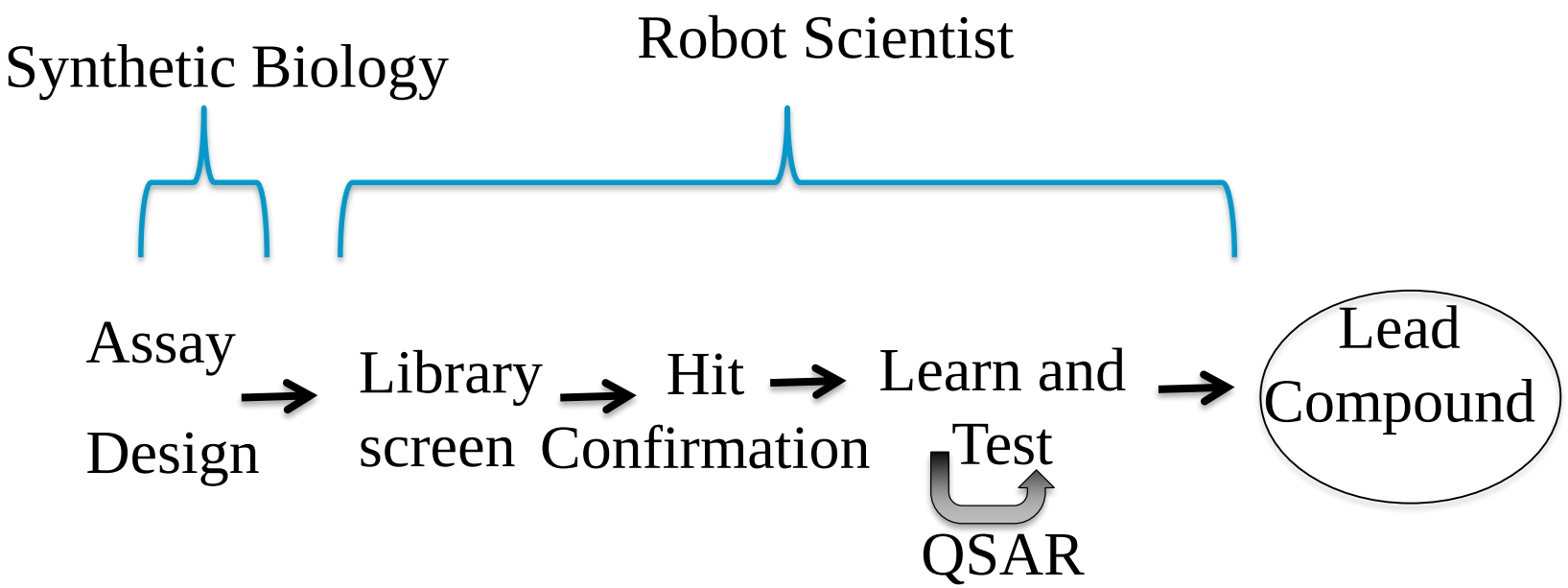


Phosphoglycerate kinase

Formalising the Problem

- Use graphs and standard chemoinformatic methods to represent background knowledge - the use of relations is planned.
- Uses induction (quantitative structure activity relationship – QSAR learning) to infer new hypotheses.
- Use active learning to decide efficient experiments, and econometric model to decide what compounds to test.

Automating Early Drug Development



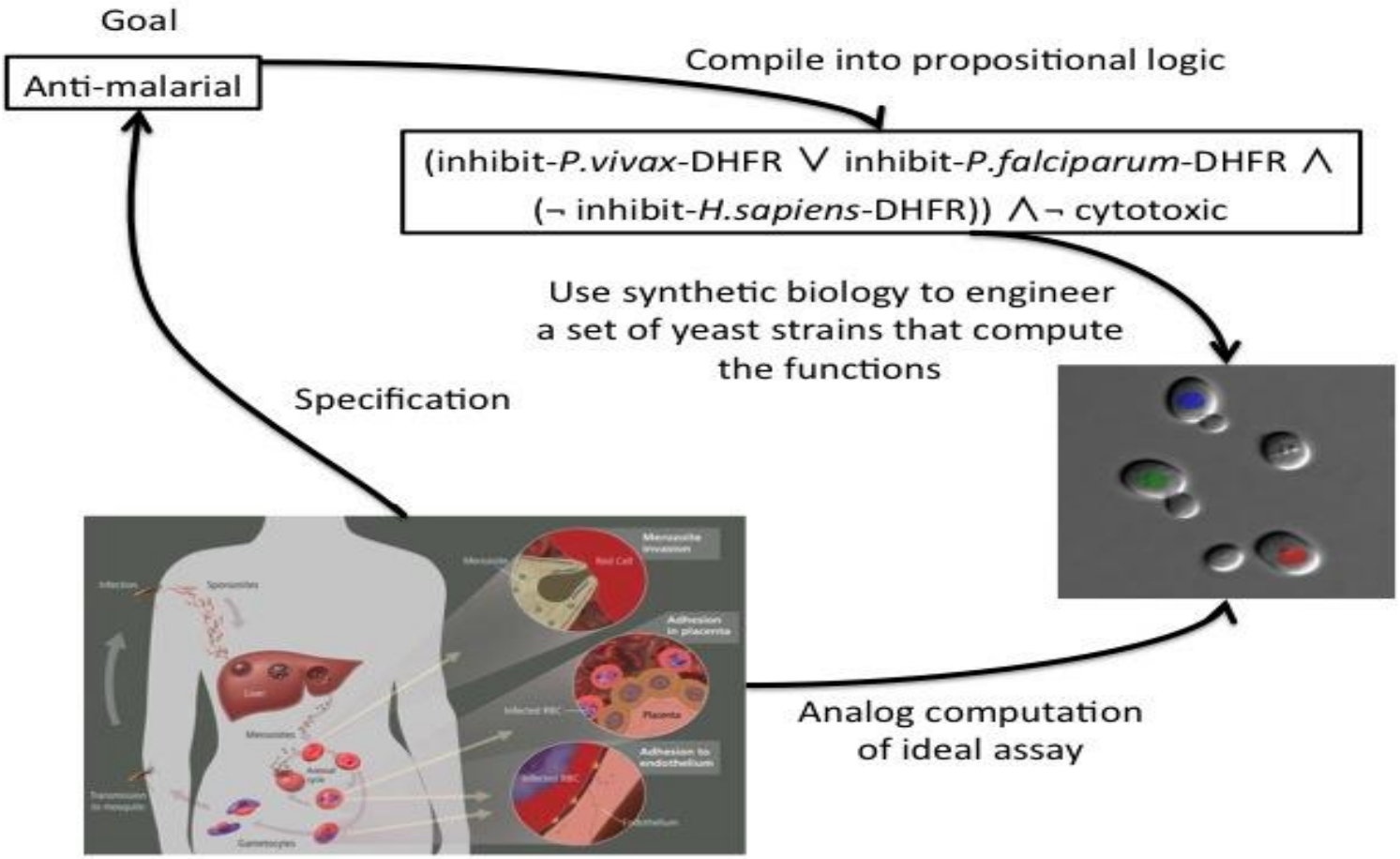
Design

- Novel design features: Combines: screening, hit confirmation, and QSAR learning.
  - Uses synthetic biology based yeast assays.
  - During the standard screening process Eve is able to decide to switch to QSAR mode.
  - Use cycles of active learning to improve QSARs.

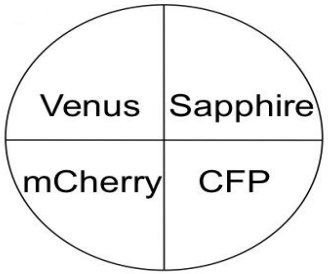
Synthetic Biology based Assays

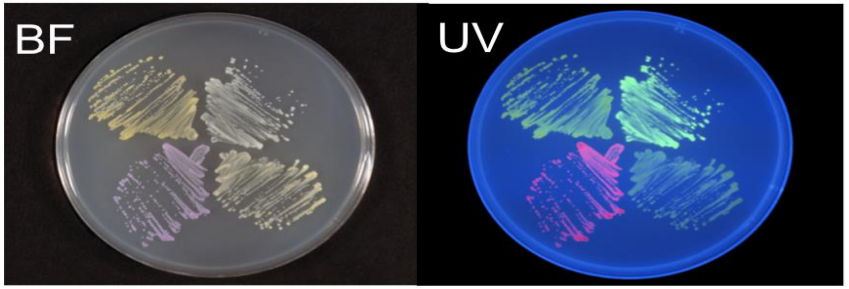
- Our idea is to engineer cells to be analog computers.
- These computers will accurately estimate a biological function that corresponds to the set of desired assay properties.
- The function estimated is the utility of a compound against a disease.
- E.g. ((inhibit *P. vivax* DHFR) ∧ (¬ inhibit *H. sapiens* DHFR) ∧ (¬ cytotoxic)).

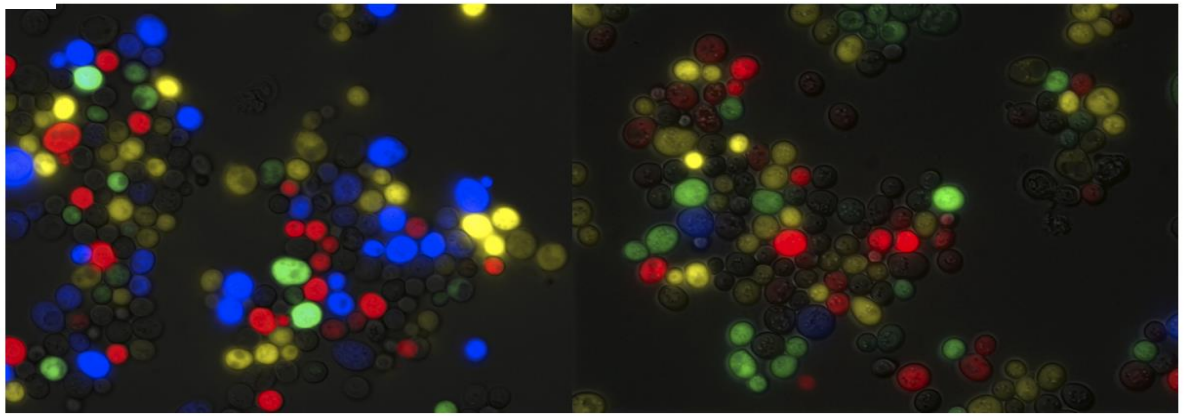
Synthetic Biology Workflow



Yeast Strains







No drug

Pyrimethamine

y^{Hs}DHFR+yEpmCherry

y^{Pf}DHFR+yEpSapphire

y^{PfR}dhfr+yEpVenus

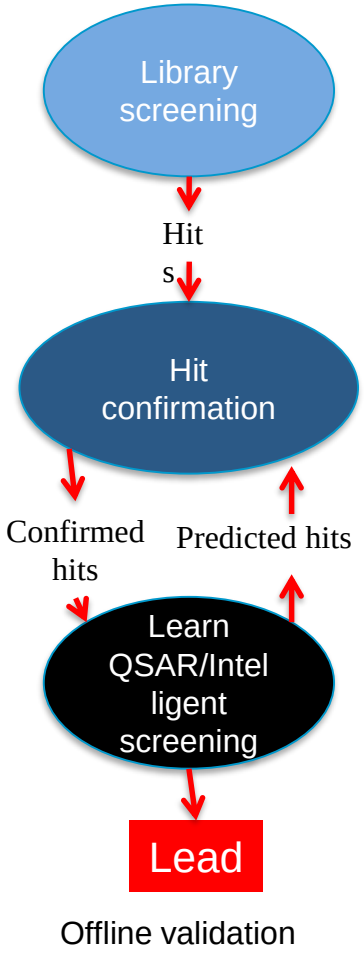
y^{Sm}DHFR+yEpCFP

Learning QSARs

- Almost every form of statistical and machine learning method you can think of has been applied to QSAR learning.
- Leading methods are logistic regression, support vector machines, random forests. …

- Eve currently uses Gaussian process models. Has the advantages of being generative and outputting probabilities – helps active learning.

Eve's Automation of Pipeline



- Standard library screening is brute force:
- Eve uses intelligent screening

- In the standard "pipeline" the 3 processes are not integrated.
- In Eve automated and integrated.

Intelligent v Brute-force Screening

- We wished to compare our AI based screening against the standard brute-force approach: “begin at the beginning and go on till you come to the end: then stop” (Lewis Carroll).

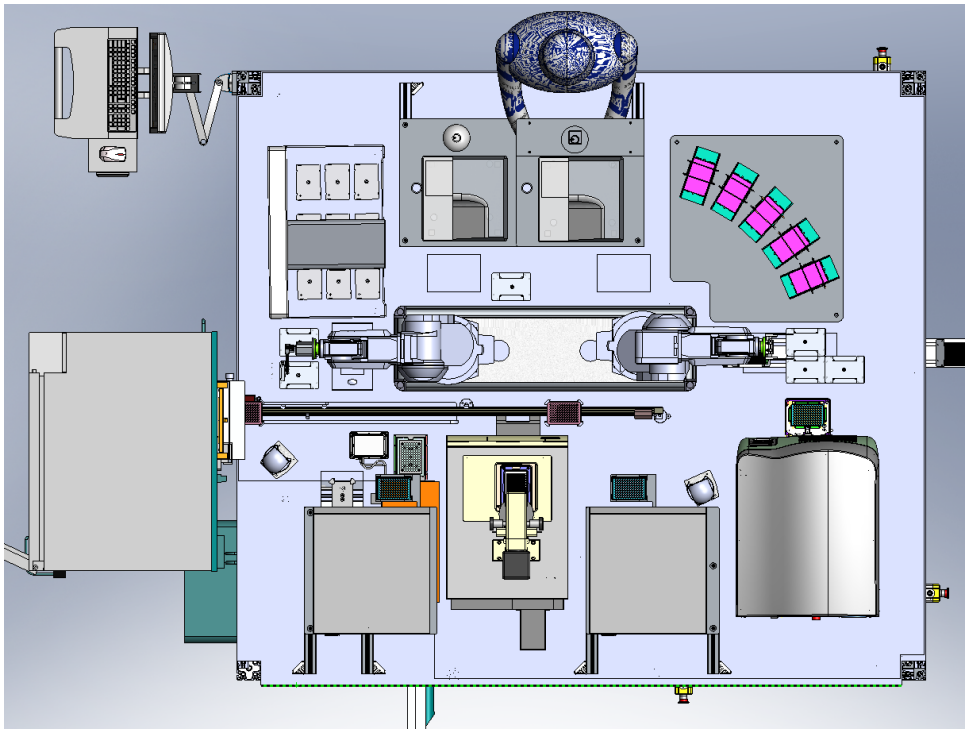
- While simple to automate standard screening is slow and wasteful of resources, since every compound in the library is tested. It is also unintelligent, as it makes no use of what is learnt during screening.

- Use money to decide.

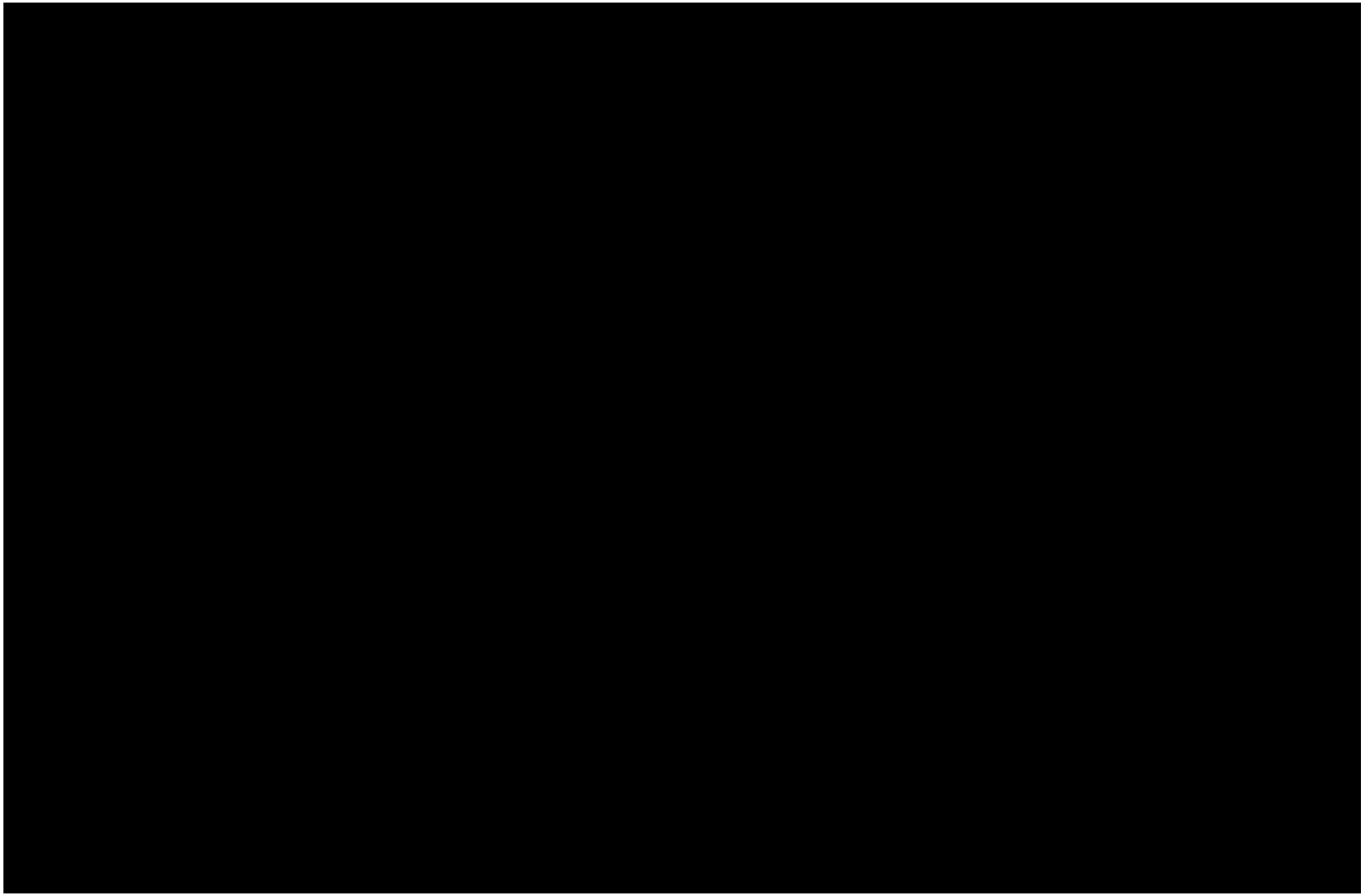
Eve's Hardware

Highlights of Eve's hardware:

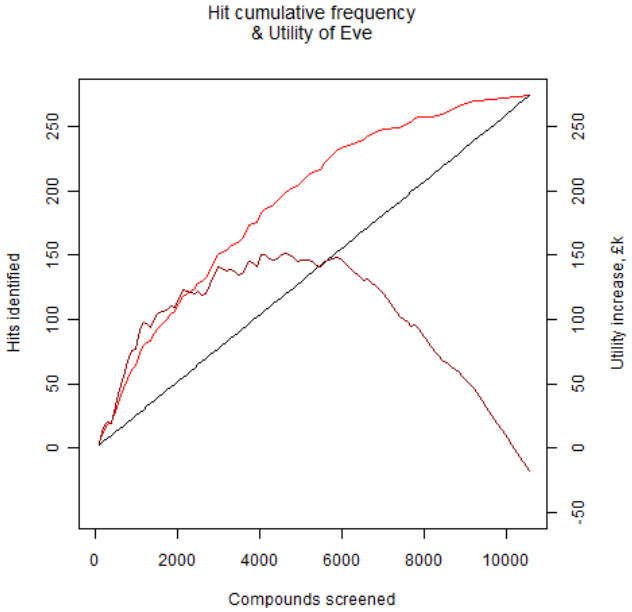
- Acoustic liquid handling
- High throughput 384 well plates
- Two industrial robot arms
- Automated 60x microscope
- Liquid handlers, fluorescence readers, barcode scanners, dry store, incubator, tube decapper ...

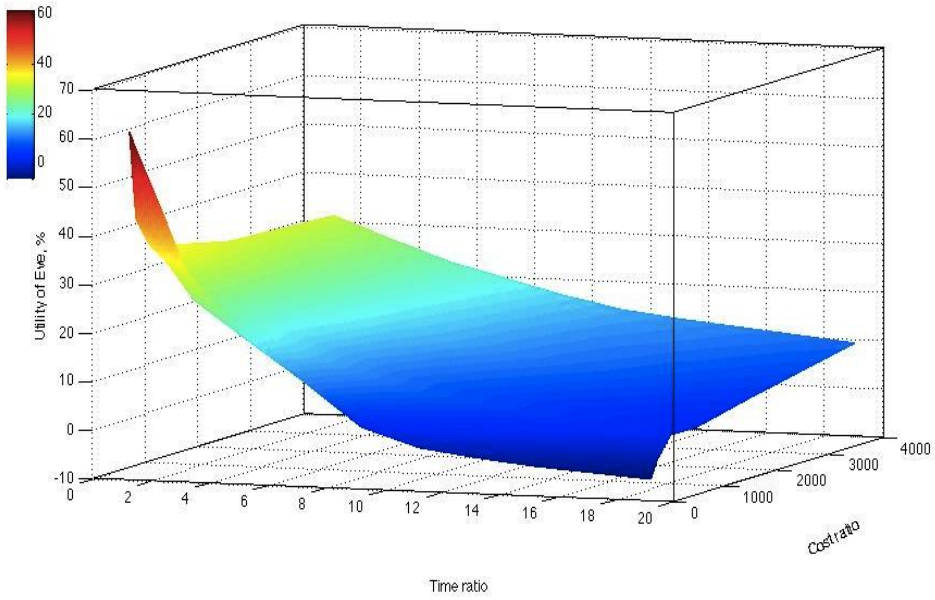


Eve



The Economics of Intelligent Screening



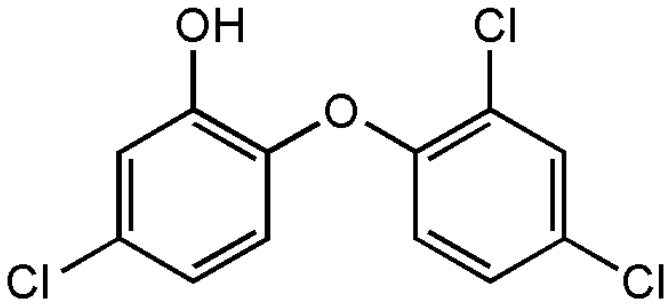


$$\Delta\, Utility\ of\ Eve = \sum_{1}^{Nm}(Tm + Cm) + \sum_{1}^{Nx}(Tc + Cc - Uh) + \sum_{1}^{Ne}(Tm - Tc + Cm - Cc)$$

Nm	-	Number of compounds not assayed by Eve
Tm	-	Cost of the time to screen a compound using the mass screening assay
Cm	-	Cost of the loss of a compound in the mass screening assay
Nx	-	Number of hits missed by Eve
Tc	-	Cost of the time to screen a compound using a cherry-picking (confirmation or intelligent) assay
Cc	-	Cost of the loss of a compound in a cherry-picking assay
Uh	-	Utility of a hit
Ne	-	Number of compounds assayed by Eve

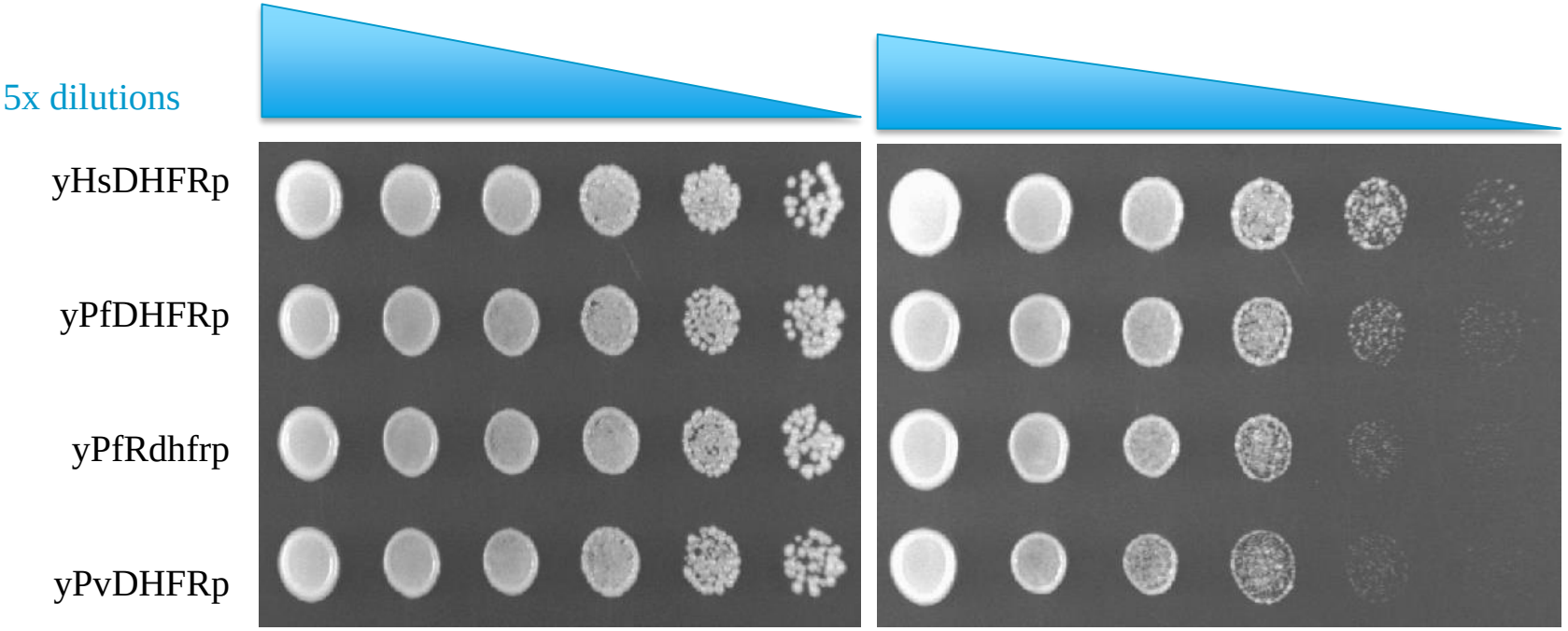
Triclosan

Triclosan Repositioned for Malaria



- Simple compound
- Known to be safe – used in toothpaste.
- Targets both DHFR and FAS-II – well established targets.
- Demonstrated activity using multiple wet experimental techniques.
- Works against wild-type and drug-resistant *Plasmodium falciparum*, and *Plasmodium vivax*.

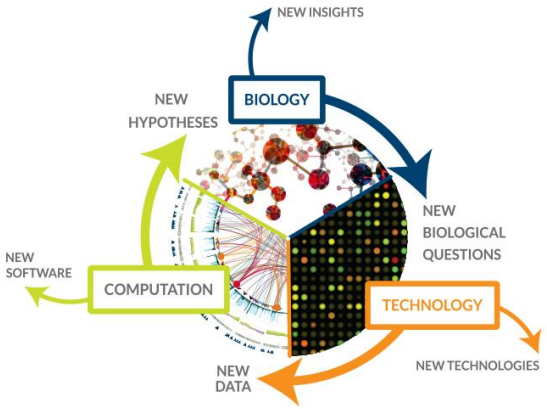
Triclosan specifically inhibits *Plasmodium* DHFRs



Eve

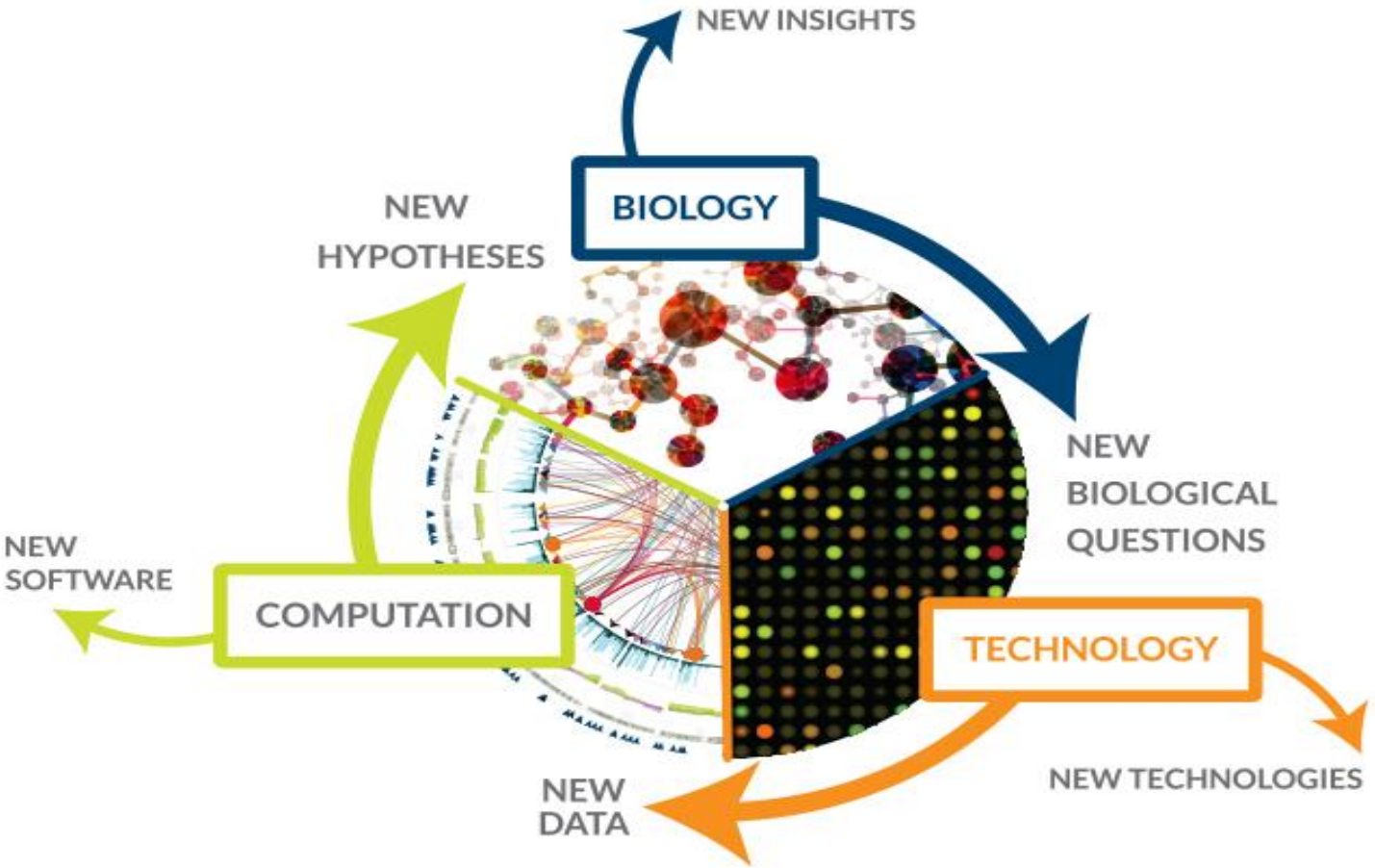
Systems Biology

The Application Domain



- Systems Biology.
- The modelling of Biological Systems.
- Diauxic shift in yeast (*S. cerevisiae*).
- Glucose ➔ Ethanol ➔ CO_2
- A model of Biological Switching.
- Understanding important for cancer – Warburg effect.
- Understanding important for aging.

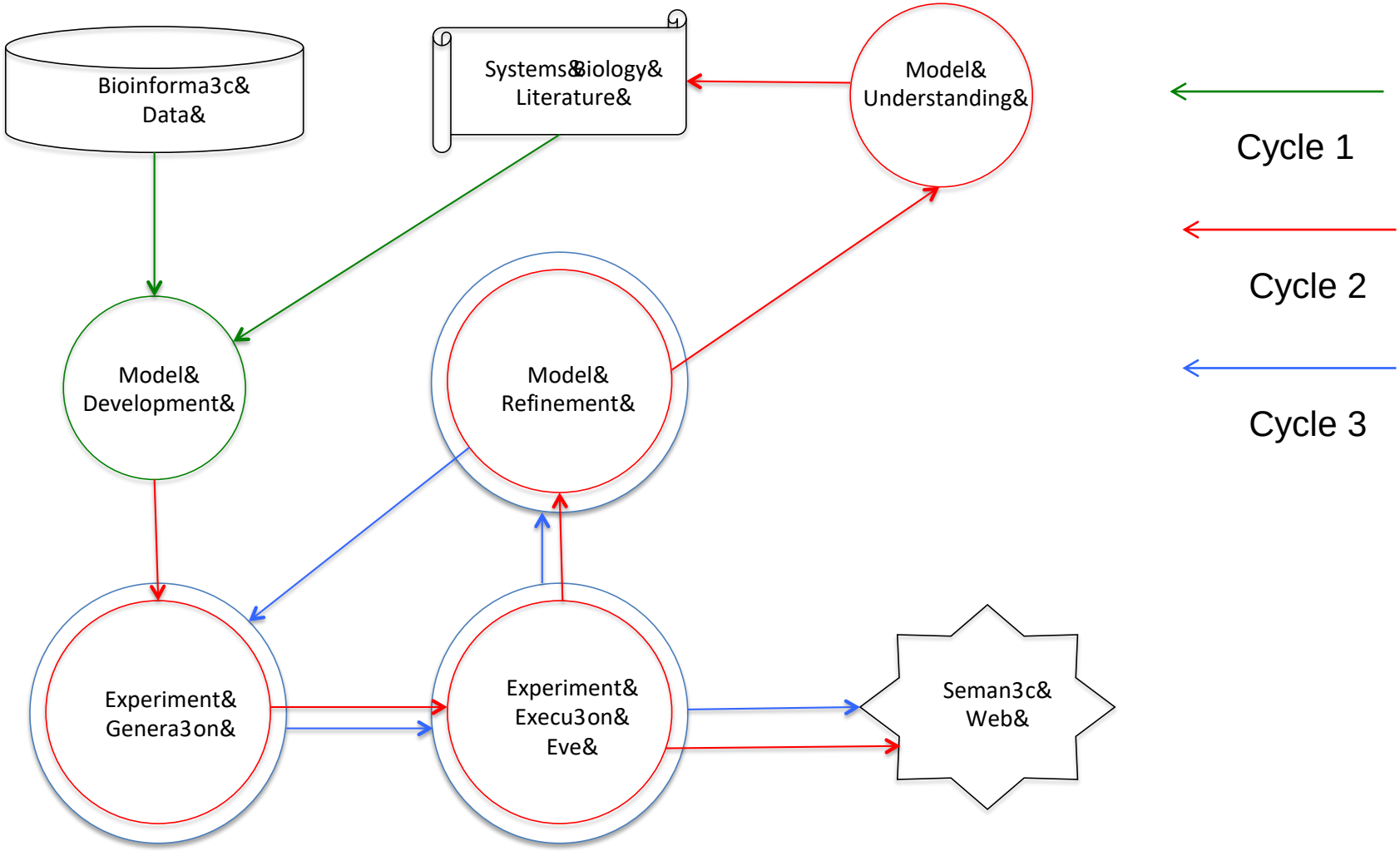
Systems Biology

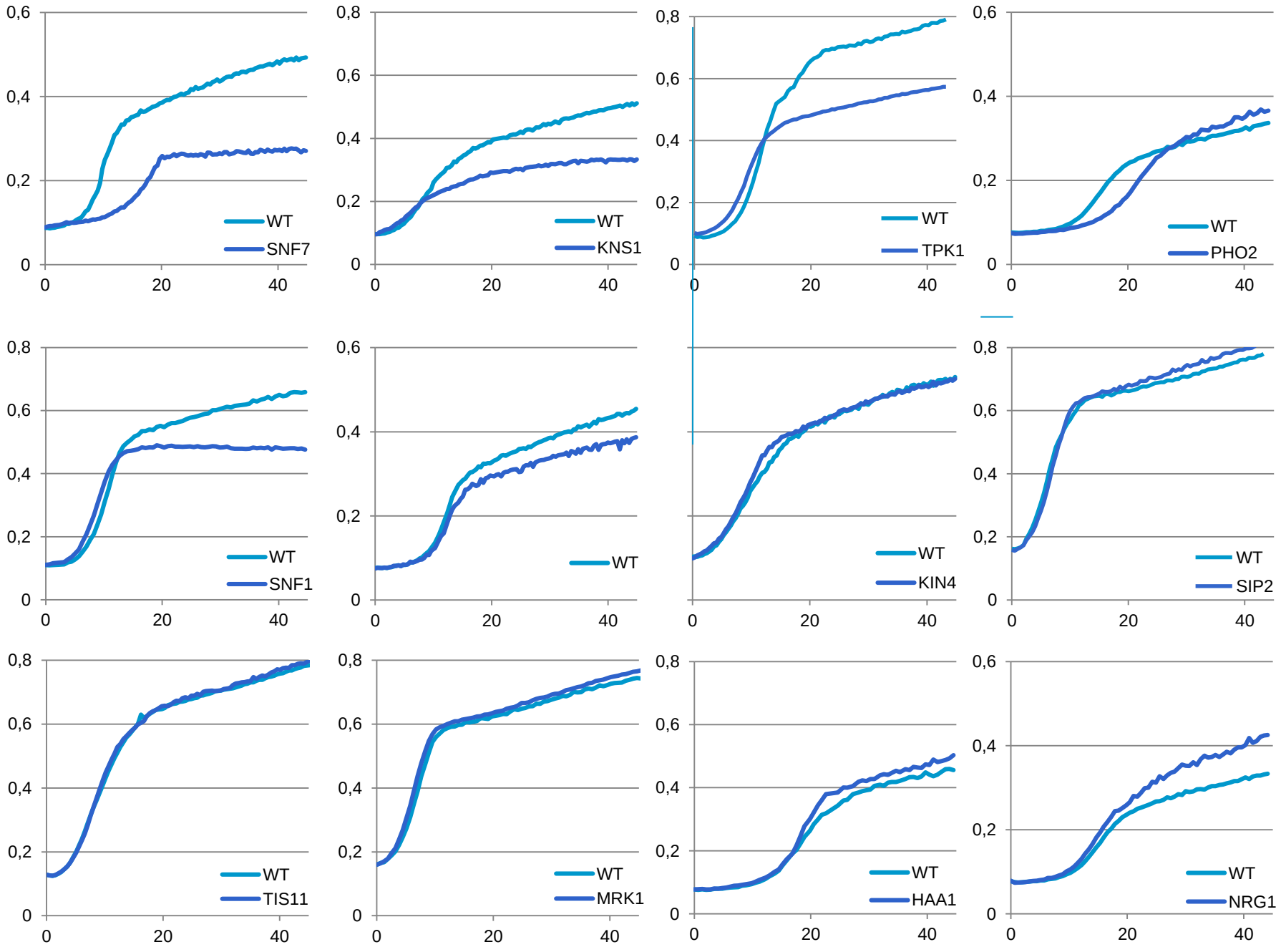


AdaLab Software

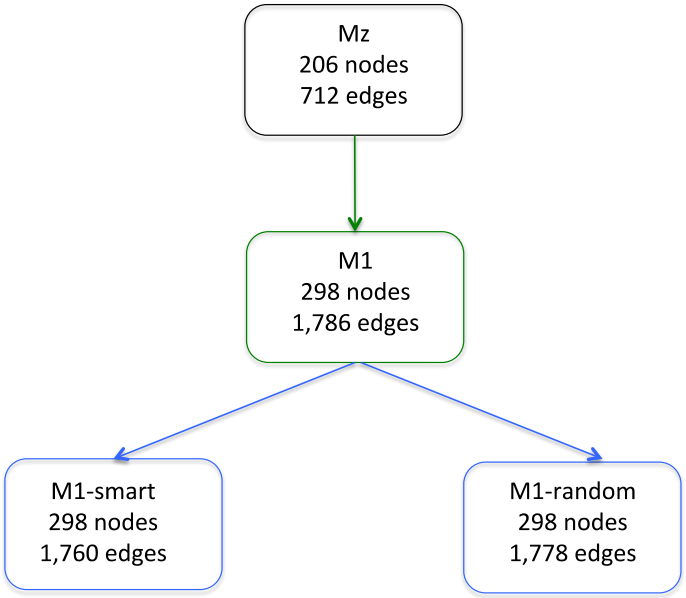
Software Type	Name	Details
AI Tools	CoRegNet	Reconstruction and integrated analysis of co-regulatory networks.
	CoRegMine	Experiment design
	ELSA	Ensemble Learning of Spanning Arborescences.
	AdactiveFB	Active learning based on forward & backward simulation.
	MinerLC*	Graph mining tool.
	Adarev	Experiment design
	Adana	Data analysis
Semantic Web Tools and Ontologies	AdaLab-meta	An ontology for the description of metadata about datasets.
	AdaLab	A domain ontology to represent relevant system biology biological entities.
	UNO	An ontology of uncertainties.
	Eve-CV	Eve experiments control vocabulary
	RDF	RDF Knowledgebase
Bioinformatic Resources	Brauer	Microarray data of the yeast diauxic shift.
	Yeastract	A curated repository of regulatory associations between transcription factors and target genes in yeast.
	YeastKinome	A yeast kinase and phosphatase interactome resource.
Systems Biology Models	iMM904	Model of yeast metabolism: a Flux Balance Analysis (FBA) model.
	Mz	Diauxic shift model derived from the literature.
	M1	Diauxic shift model enhanced using bioinformatic data.
	M1-random	Diauxic shift model enhanced using bioinformatic data and high-throughput experiments.
	M1-smart	Diauxic shift model enhanced using bioinformatic data and hypothesis led experiments.
Systems Biology Simulation	DBN	Simulation of yeast cell signaling: a dynamic two-time slice Bayesian network, with linear Gaussian parameters.
	DFBA	Simulation of yeast metabolism: Dynamic Flux Balance Analysis.
Statistics	Yeast-stats	Yeast growth parameter estimation.
Laboratory Robotics	Overlord	Laboratory automation control.

AdaLab Project: 3 Cycles




WT
SNF7
WT
KNS1
WT
TPK1
WT
PHO2
WT
SNF1
WT
WT
KIN4
WT
SIP2
WT
TIS11
WT
MRK1
WT
HAA1
WT
NRG1

AdaLab Models



M1-smart > M1-random > M1 > Mz

Cycle	Test Strains	Mz error	M1 error	M1-random error	M1-smart error	Wins	Ratio	Significance
1	192	0.17	0.0033	-	-	192	98%	$< 2.2 \times 10^{-16}$
2	281	-	5.533×10^{-3}	3.892×10^{-3}	-	259	30%	$< 2.2 \times 10^{-16}$
2	281	-	5.533×10^{-3}	-	2.158×10^{-3}	269	74%	$< 2.2 \times 10^{-16}$
3	81	-	-	1.757×10^{-3}	7.231×10^{-4}	79	58.4	$< 7.6 \times 10^{-9}$

Future Prospects

Next Generation Robot Scientist

- Domain Systems Biology.
- One of the most challenging tasks in modern science.
- Ability to do ~10,000 individual hypothesis-led experiments in parallel.
- AI able to automatically design, plan, execute experiments, and modify systems biology model.

The Future?

- In Chess/Go there is a continuum of ability from novices up to Grandmasters.
- I argue that this is also true in science, from the simple research of Eve, through what most human scientists can achieve, up to the ability of a Newton or Einstein.
- If you accept this, then just as in Chess/Go, it is likely that advances in computer hardware and software will drive the development of ever smarter Robot Scientists.
- In favour of this argument are the ongoing development of AI and laboratory robotics.

Vision

- The collaboration between Human and Robot Scientists will produce better science than either can alone – human/computer teams still play better chess that either alone.
- Scientific knowledge will be primarily expressed in logic with associated probabilities and published using the Semantic Web.
- The improved productivity of science leads to societal benefits: better food security, better medicines, etc.
- The Physics Nobel Frank Wilczek is on record as saying that in 100 years' time the best physicist will be a machine

Conclusions

- Science is a wonderful application area for AI.
- Automation is becoming increasingly important in scientific research e.g. DNA sequencing, drug design.
- The Robot Scientist concept is the logical next step in scientific automation.
- The Robot Scientist Adam was the first machine to have discovered novel scientific knowledge.
- The Robot Scientist Eve has found new lead compounds for neglected tropical diseases.