# Bioinformatics at Novozymes

**Martin S. Borchert**

**Integration Director R&D OrganoBalance**

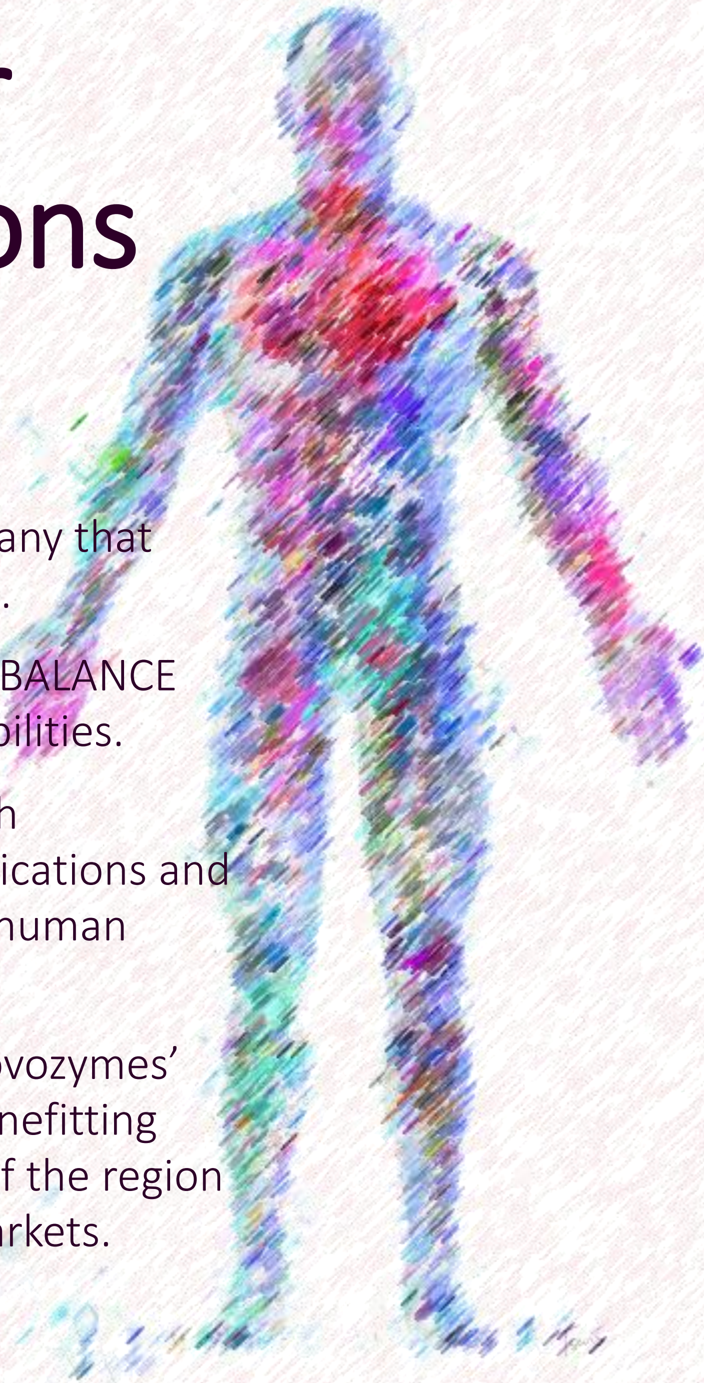
**Head of Bioinformatics  & Microbe Technology**

PRAXISforum DECHEMA, 8 Sept 2016

# Strengthening our position within microbial solutions

# A basis for a wider range of applications

ORGANO BALANCE



- ORGANOBALANCE is a German-based company that researches and develops microbial solutions.
- With 29 accomplished employees, ORGANOBALANCE further enhances our world-class R&D capabilities.
- ORGANOBALANCE applies its strong research capabilities across a number of exciting applications and industries, including food, feed, animal and human health/probiotics and biochemicals.
- ORGANOBALANCE will operate as part of Novozymes' global R&D organization out of Germany, benefitting from the strong biotechnology capabilities of the region and strong ties to German academia and markets.

Find more information at **http://organobalance.de/**

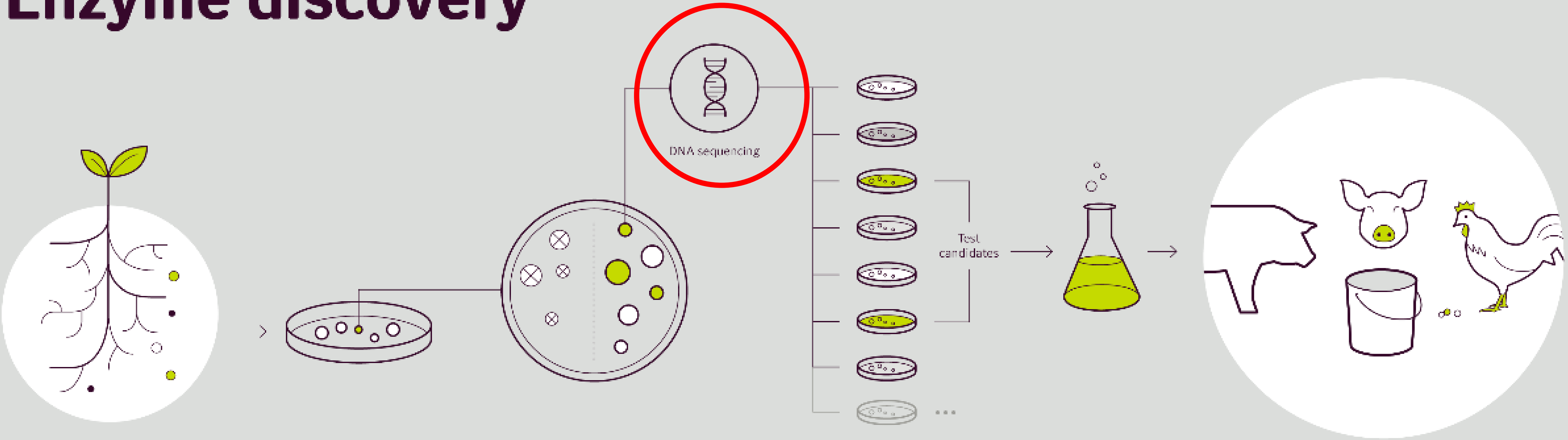
"ORGANOBALANCE will boost our capacity to develop new, sustainable solutions across industries and provide us with additional commercial opportunities"

Sebastian Søderberg, VP of Business Incubation & Acquisitions

novozymes

novozymes

# Novozymes Animal Health & Nutrition: Enzyme discovery



1 **Collect**
Samples collected from targeted sites all around the world by microbiologists

2 **Grow**
From these samples, thousands of microorganisms are grown in special media and under special conditions

3 **Identify & Archive**
Pure colonies of the isolated micro-organisms are DNA-sequenced, identified, characterized and classified in our cell banks

4 **Screening**
Novel assays are developed to screen the identified enzymes for their potential benefits

5 **Testing**
Enzymes are then fermented, formulated and tested

6 **Measure**
In field testing, the potential of the enzymes for increasing animal health and nutrition is measured

novozymes

# Public DNA sequence databases

Mbp
yesterday
today
tomorrow

**Our research makes use of public knowledge**

- Mining for relevant enzyme diversity
- Integration of genome meta-data
  - ecological niche, pH, temperature etc.
- Infer taxonomy of metagenomic contigs
- Comparative genomics: Understand genomic context of enzymes

*"The challenge of Biology is no longer to collect sequence data.*

*Challenge 1:*

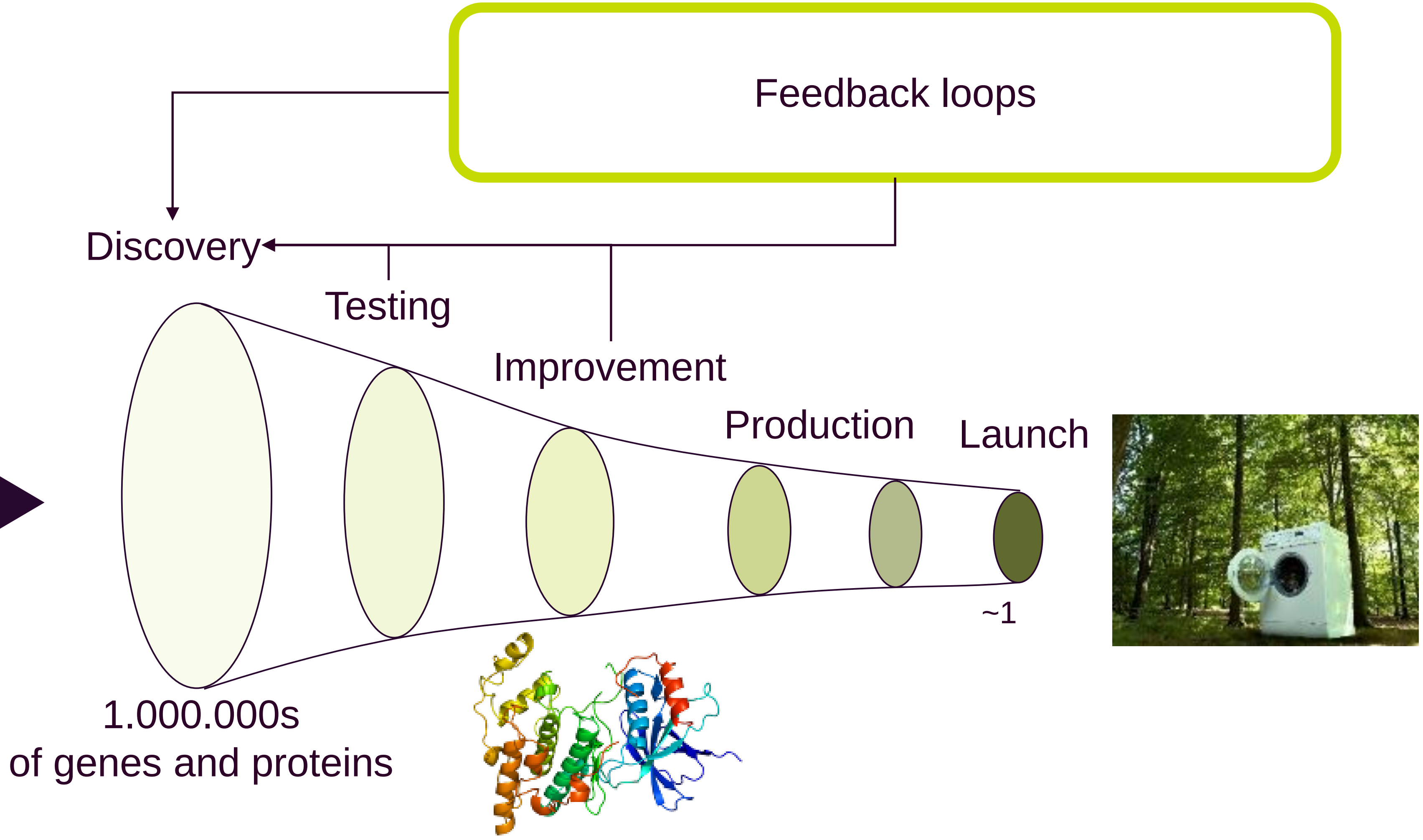
*is to determine **functional data efficiently.***

*Challenge 2:*

*is to **analyze data** to associate complex phenotypes to cheap sequence data."*

*Prof. Bernard Henrissat, CNRS, Head of the Carbohydrate-active enzymes database (CAZy). Novozymes Symposium Oct 2016*

# SHORTEN THE IDEA-TO-PRODUCT TIME



Feedback loops

Discovery

Testing

Improvement

Production

Launch

~1

1.000.000s
of genes and proteins

Biosourcing

Enzyme discovery

Enzyme testing

# THREE DISCOVERY APPROACHES TO NATURAL DIVERSITY

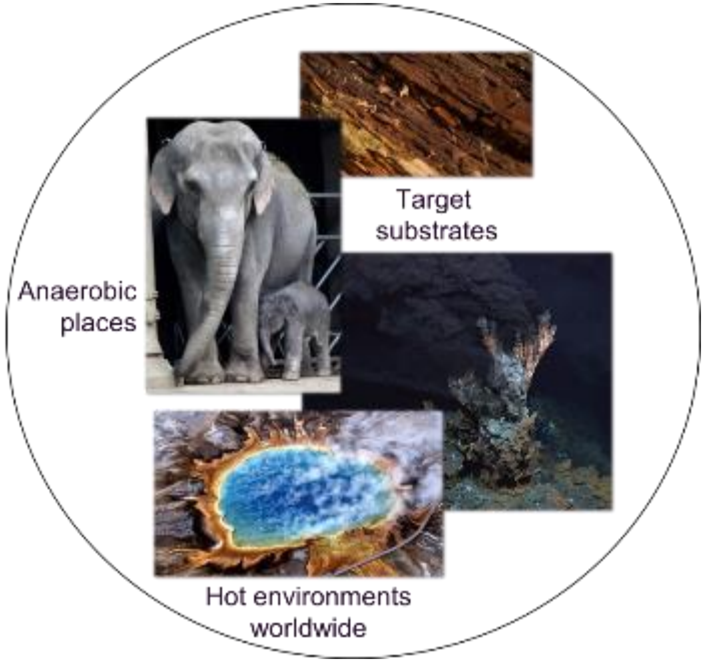
Our Culture Collection contains thousands of strains isolated for their specific characteristics

CLASSIC MICROBIOLOGY

Culturable microorganisms

**Targeted – but slow**

Less than 1% of Nature's microorganisms can be cultured as individual cells

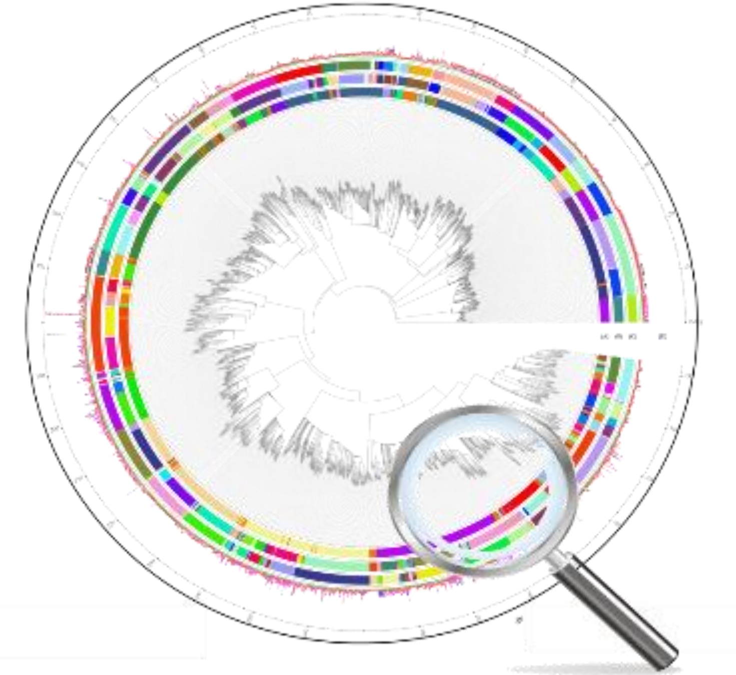


METAGENOMICS

Microbial communities / non-culturable organisms

**Exploratory – but low hit rate**

*In silico* we can search over billions of genes in public and own gene pools/databases



IN SILICO SCREENING

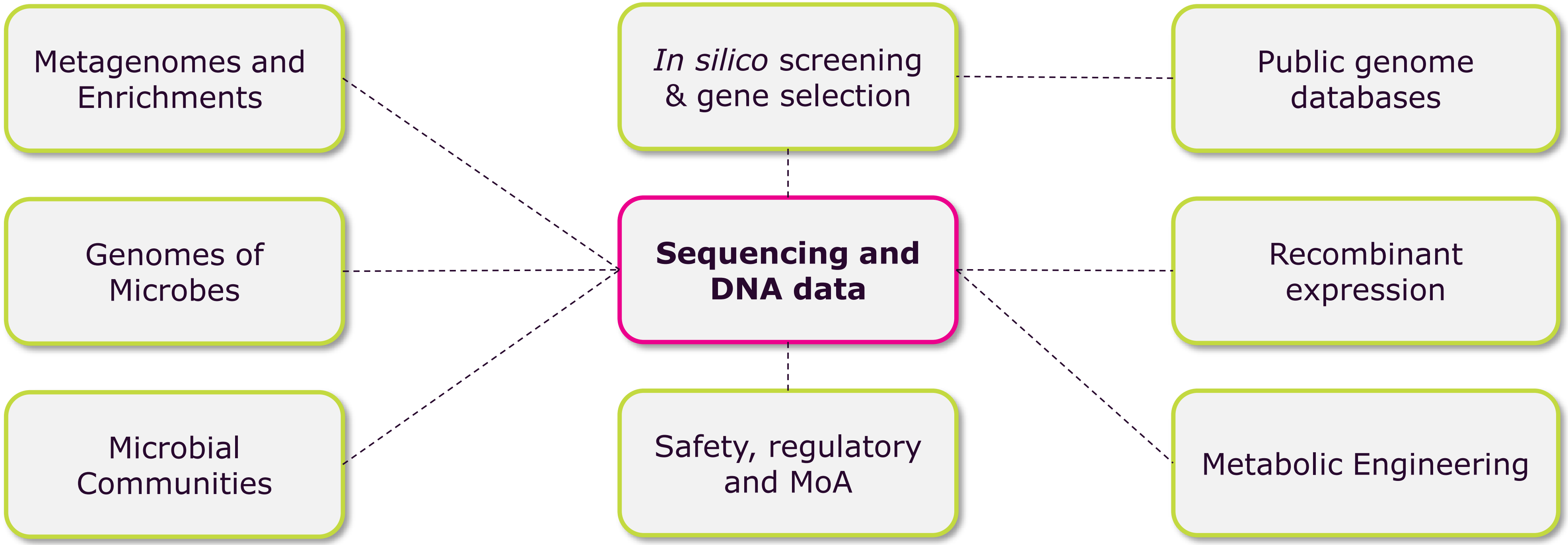
Internal and public genome databases

**Fast – but functionally and performance difficult to predict**

# Bioinformatics

# But Bioinformatics in White Biotech is much more…

novozymes
Rethink Tomorrow



Metagenomes and Enrichments

Genomes of Microbes

Microbial Communities

*In silico* screening & gene selection

**Sequencing and DNA data**

Safety, regulatory and MoA

Public genome databases

Recombinant expression

Metabolic Engineering

Household Care

Food & Beverages

Bioenergy

Animal Health & Nutrition

Agriculture

Biopharma

Textile

Pulp & Paper

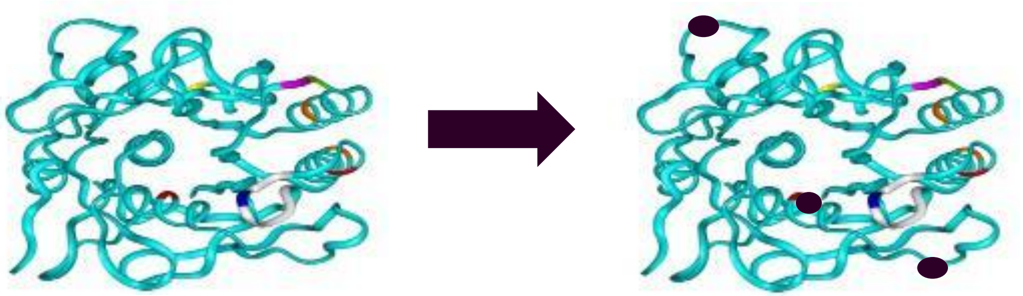
Leather

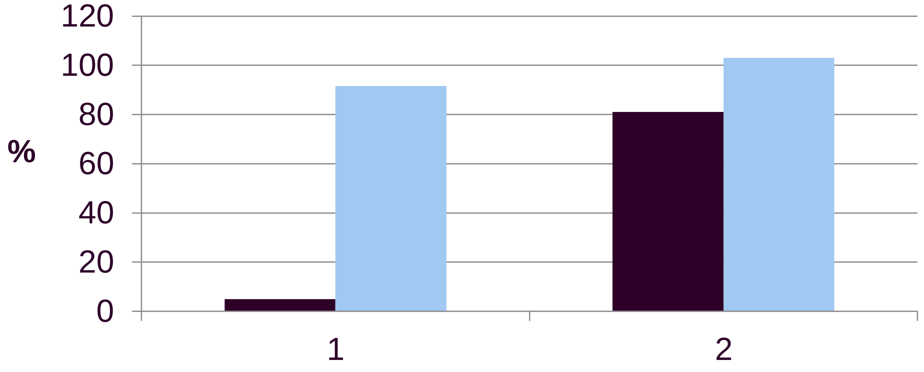
Wastewater Solutions

# Grasping and Enabling Diversity

novozymes

# PE: HIGH NUMBER OR VARIANTS TO IMPROVE NATURE

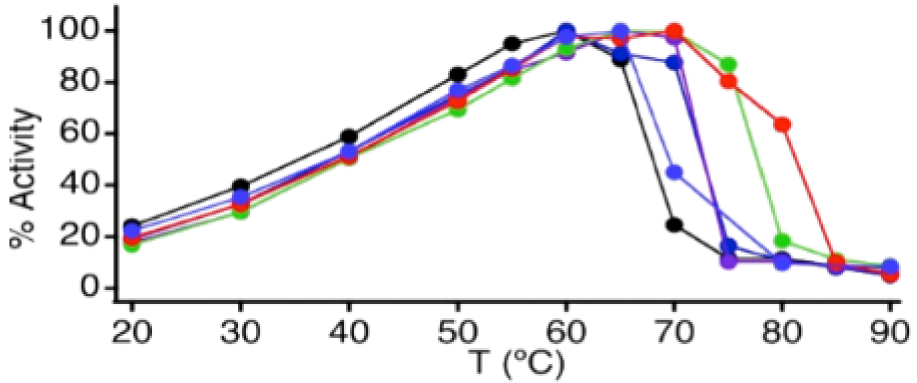




Effect of pepsin/low pH on stability of *P. lycii* phytase (black) and a variant (blue)

```
SWISSPROT:C7Z624  DNSGYCLKDRKQ-KCECFAGFTGSKCDKYTCV
SWISSPROT:B6HRY4  QENGFVDGDGSL---ECFTGFTGTDCTQFTCPN

SWISSPROT:C7Z624  LLIEPTYETESRLGDGDDPAIWISPESPEKSRVV
SWISSPROT:B6HRY4  VGVEPKYETDANGGDGDDPAIWISPVSADQS
```



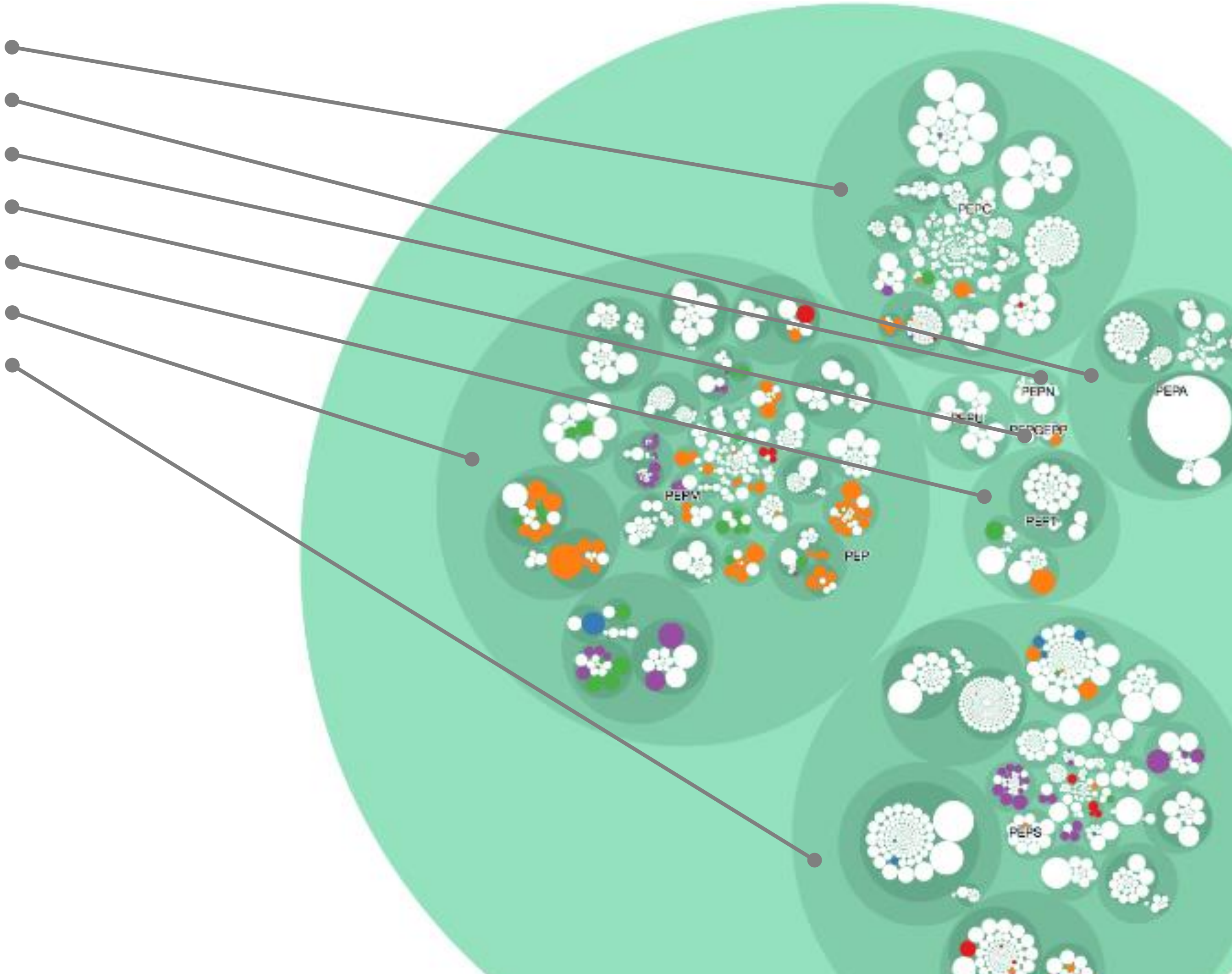
Effect of temperature on stability of *C. braakii* phytase (black) and some variants

# Overviewing [large] enzyme families: proteases

***1.9M annotated peptidases***
***within 7 major clans***

- Secreted peptidases
- Prokaryotic / fungal origin
- Exo-peptidase

Cysteine (C)
Aspartic (A)
Asparagine (N)
Glutamic (G)
Threonine (T)
Metallo (M)
Serine (S)

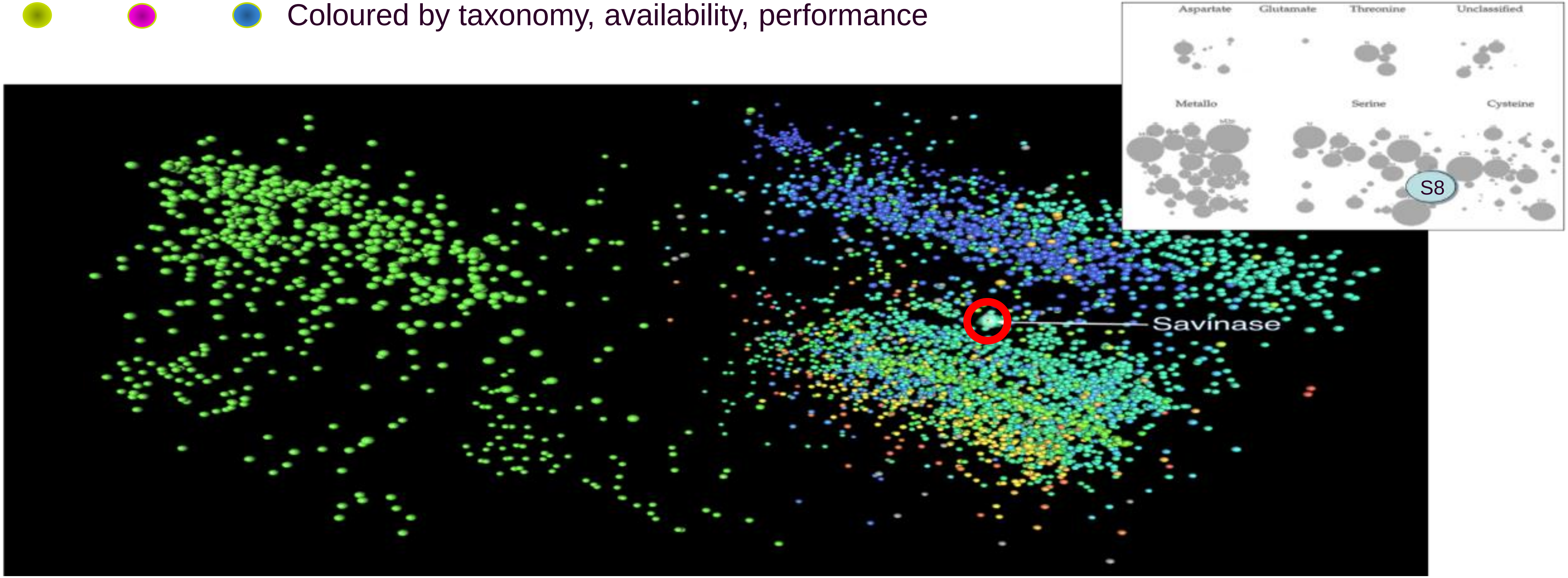


Interactive circle packing (d3.js)

# WE WANT TO NAVIGATE THE SEQUENCE SPACE!

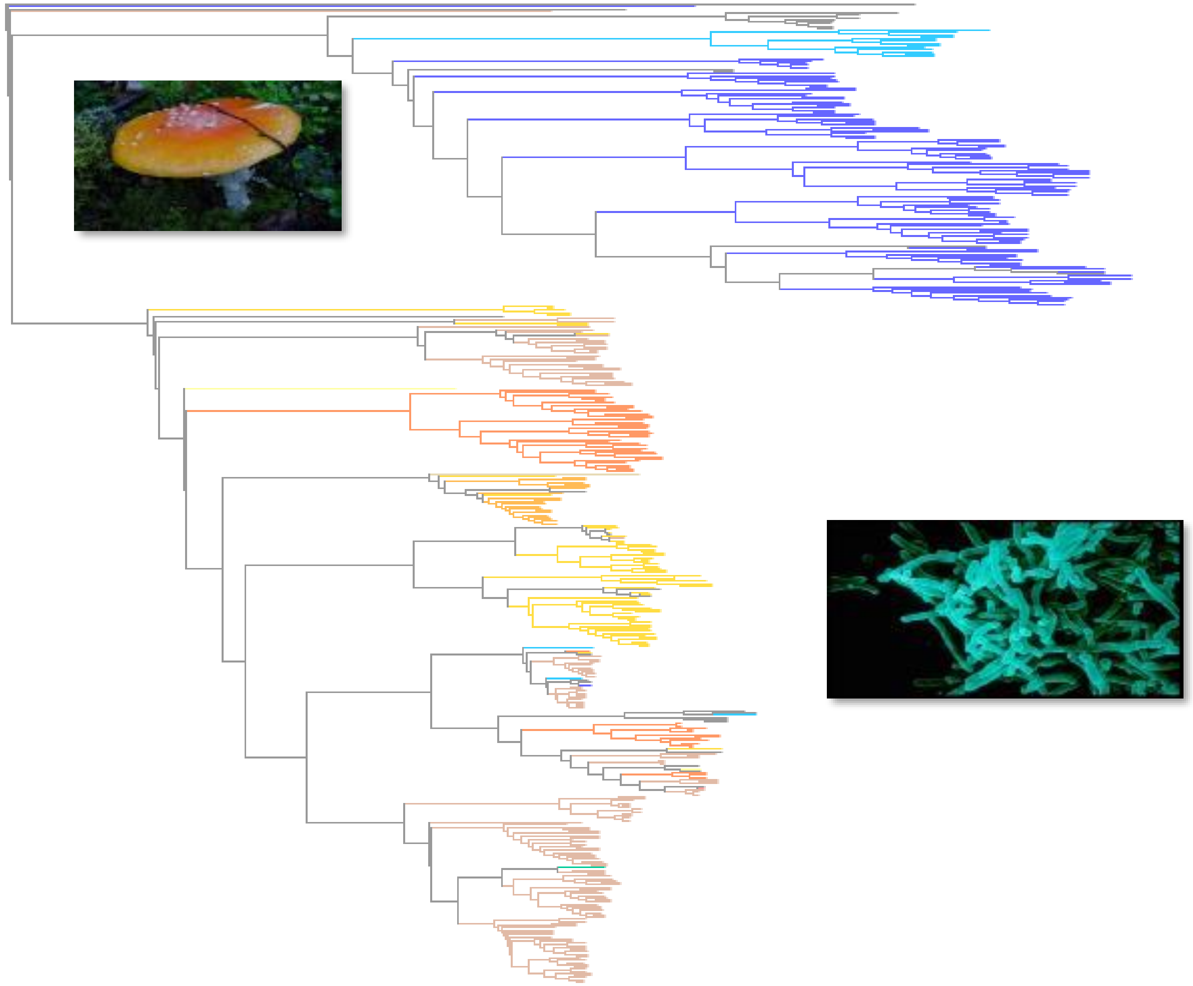
● ↔ ● Calculated distance from protein sequence similarity

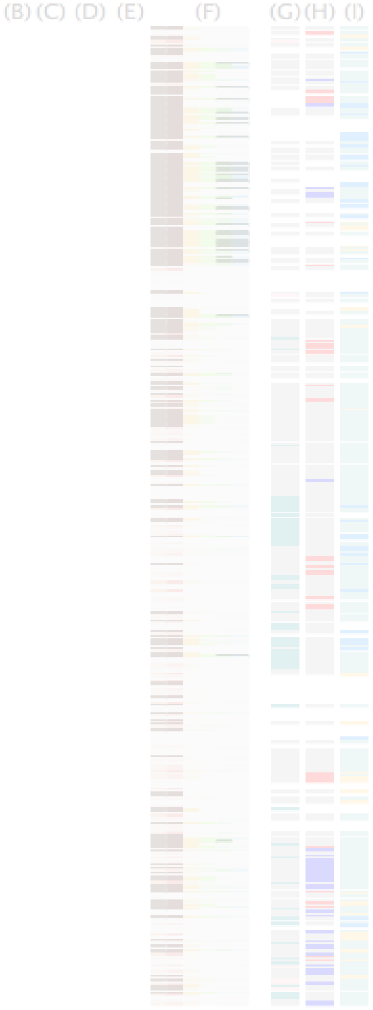
● ● ● Coloured by taxonomy, availability, performance


Protease families

Multi-dimentional Principal Analysis Plot of the DNA sequence diversity of proteases belonging to **the S8 family** (one of the Serine protease families)

# MAPPING GENOME DIVERSITY





Map meta data to whole genome or single enzyme

- Average Nucleotide Identity
- Enzymatic profile – eg. CAZYmes and/or internal domains
- Diversity of house-keeping genes or genes with verified association with performance

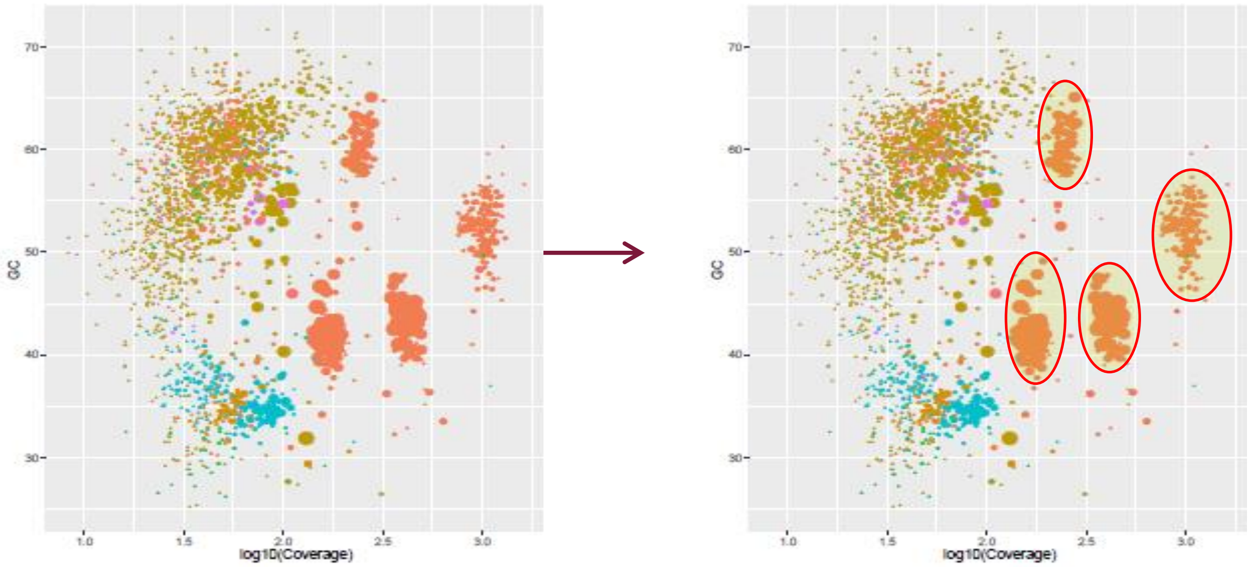
Internal interactive tree visualization software

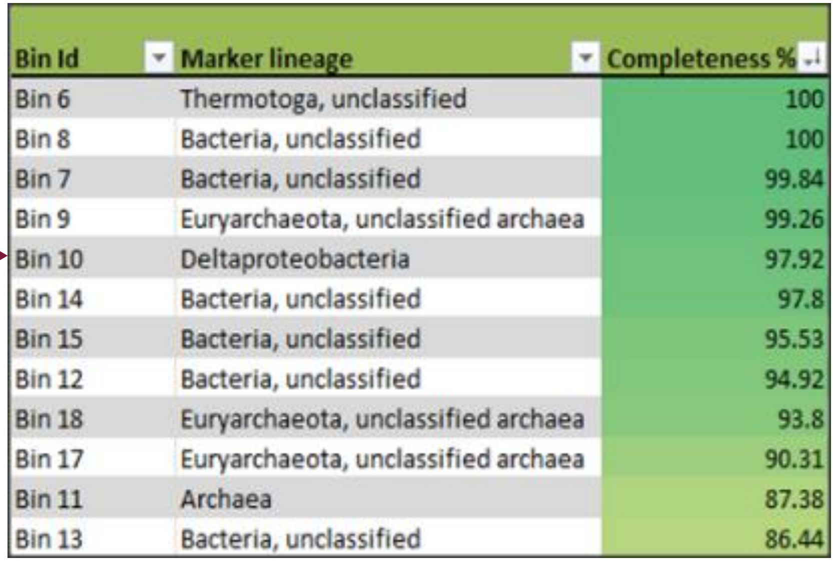
# *In silico* isolation of genomes from metagenomes

## 99 % of all new public diversity are from metagenomes[1]

- **Including 85%-99% “non-cults”/“hard to cultivate” bacteria and archaea cannot easily be isolated in a lab**

## Pan metagenome analysis or read binning

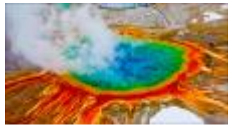




| Bin Id | Marker lineage | Completeness % |
|---|---|---|
| Bin 6 | Thermotoga, unclassified | 100 |
| Bin 8 | Bacteria, unclassified | 100 |
| Bin 7 | Bacteria, unclassified | 99.84 |
| Bin 9 | Euryarchaeota, unclassified archaea | 99.26 |
| Bin 10 | Deltaproteobacteria | 97.92 |
| Bin 14 | Bacteria, unclassified | 97.8 |
| Bin 15 | Bacteria, unclassified | 95.53 |
| Bin 12 | Bacteria, unclassified | 94.92 |
| Bin 18 | Euryarchaeota, unclassified archaea | 93.8 |
| Bin 17 | Euryarchaeota, unclassified archaea | 90.31 |
| Bin 11 | Archaea | 87.38 |
| Bin 13 | Bacteria, unclassified | 86.44 |

**GC-DNA coverage plot**

- Sequencing coverage of species in metagenomes is *not* random
- Neither is GC content and kmer profiles



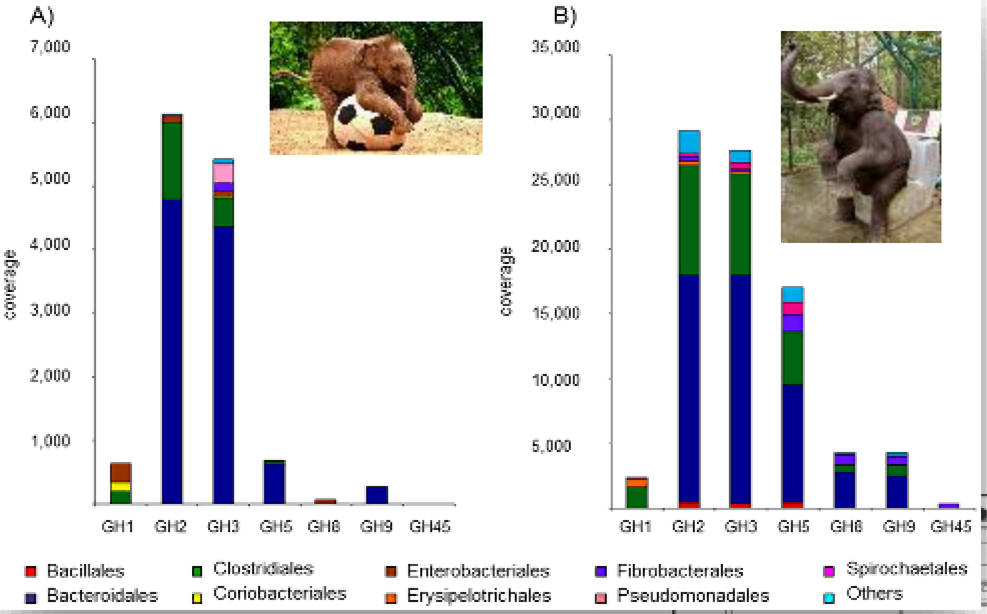
INMARE

EU H2020 Program

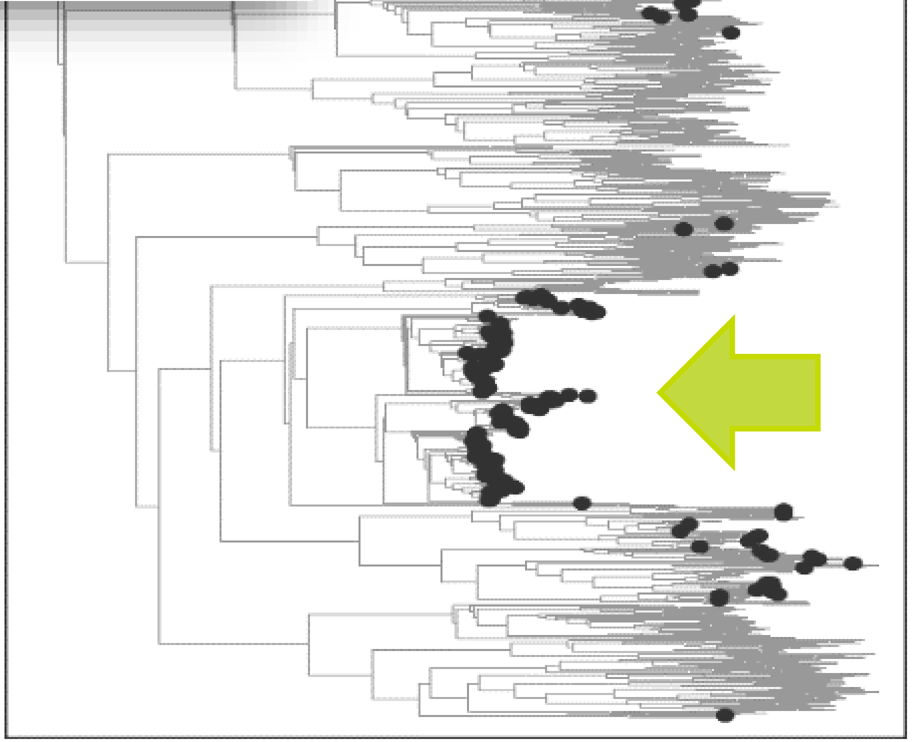
novozymes

[1] Number of genes. JGI Annual Report 2015 link, ~10bn/genes/year

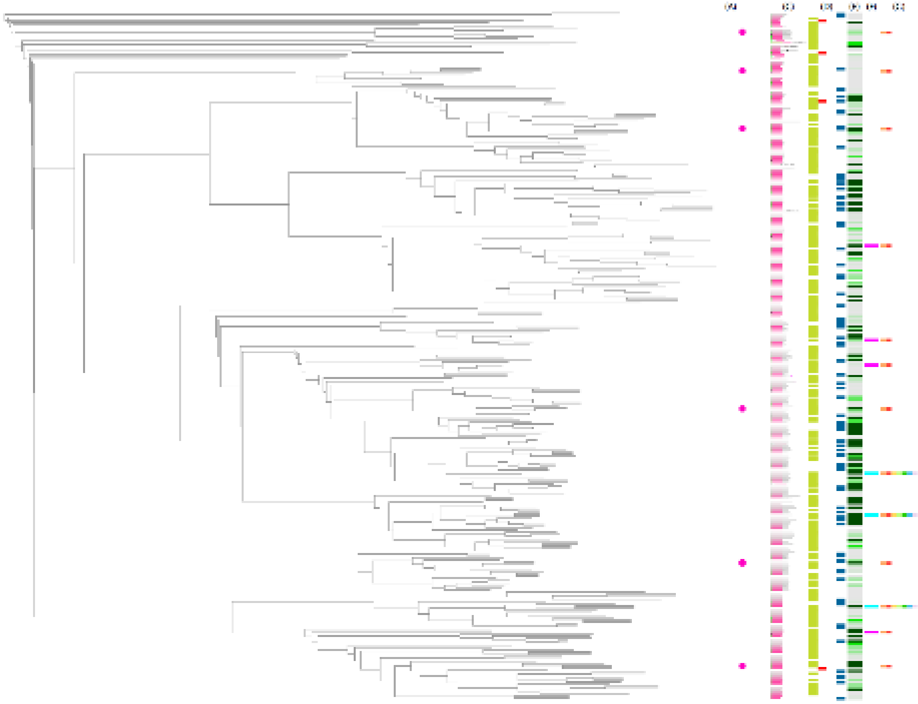
# Elephant dung metagenome provides novel enzyme diversity



Relative abundance of GH enzymes in feces:
(A) 3 months old and
(B) 6 years old elephant



Novel diversity such as GH5_1



Integration of many data: *in silico* screening results, performance & characterization

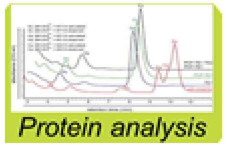
# Connecting DNA and Function

novozymes

# Finding the unknowns by “Secretomics”

Induction of microbes

Mass spectrometry with / without induction



Genome sequencing

Clone and express enzymes

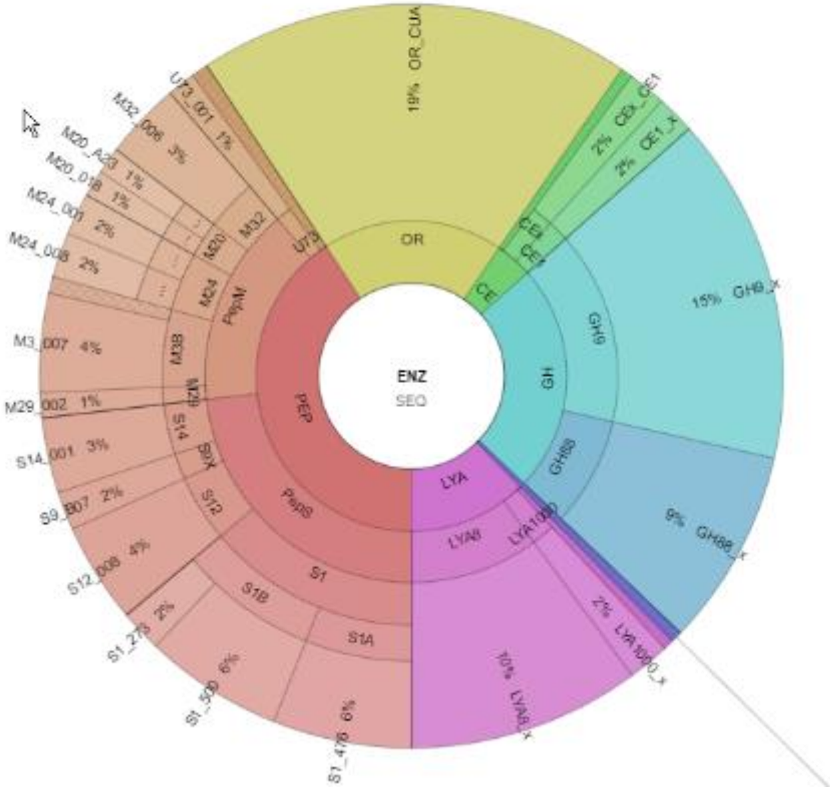
Confirm activity

Performance Testing

# Example of Secretomics

Comparison of induced versus non-induced sample to speed up the enzyme discovery process

Krona Plot

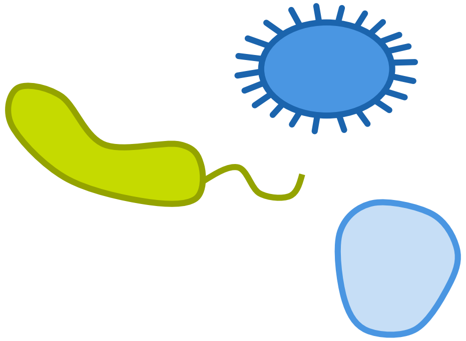


# Applied Microbiomics

# Microbiome studies at Novozymes

Are the **microbial communities** in soil/plant/gut environments shifted upon treatment with Novozymes products (enzymes, bacterial/fungal strains)?

- *What is the **function** of the microbial community?*
- *Why does that change take place and is it a **temporary** change?*
- *How does the change **influence the host**?*
- *How long does the microbe **remain** in the gut/soil – does it progress or recess?*

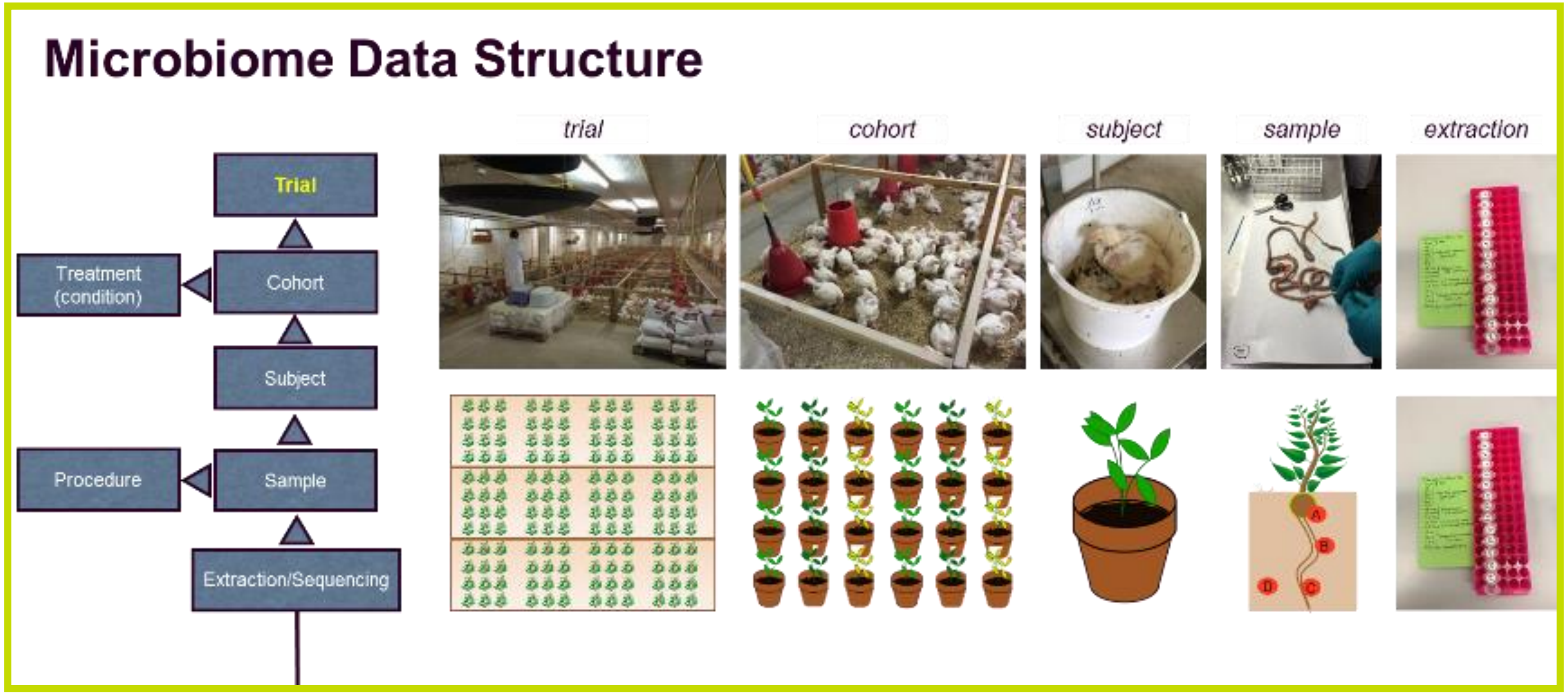


Animal Health & Nutrition

Bioag

Household Care

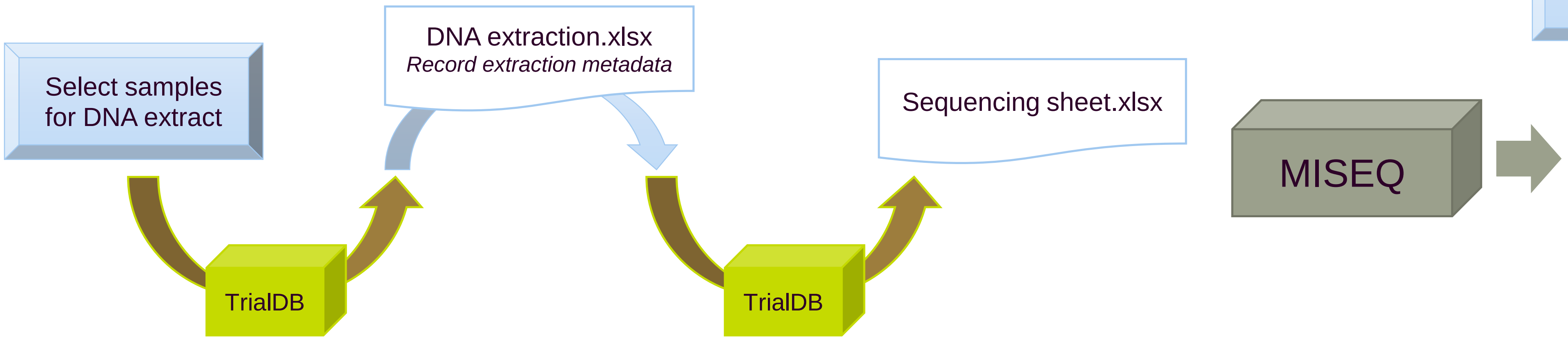
novozymes

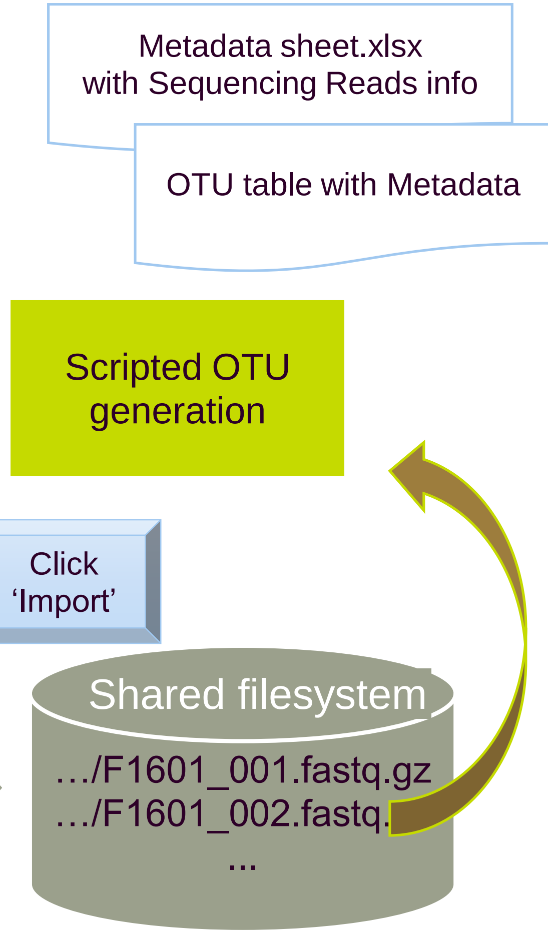


## Microbiome Data Structure

trial
cohort
subject
sample
extraction
Trial
Treatment (condition)
Cohort
Subject
Procedure
Sample
Extraction/Sequencing

# … and then build the necessary tools around it





novozymes

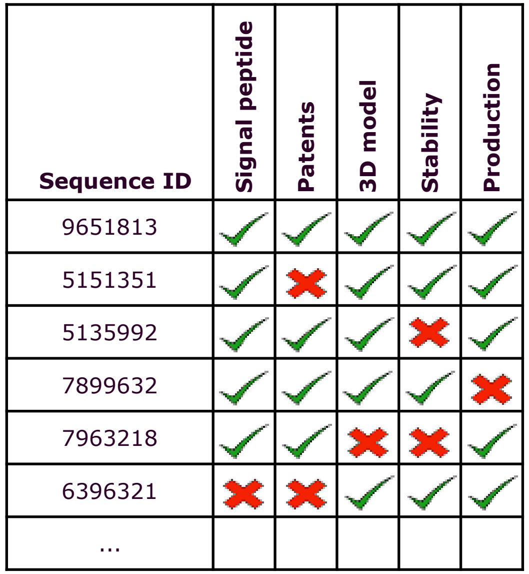
# Machine Learning

novozymes

## This is not machine learning

(Novozymes and Beta Renewables have established a strategic partnership to market cellulosic-ethanol solutions.)

# FINDING THE RIGHT ENZYME – FAST

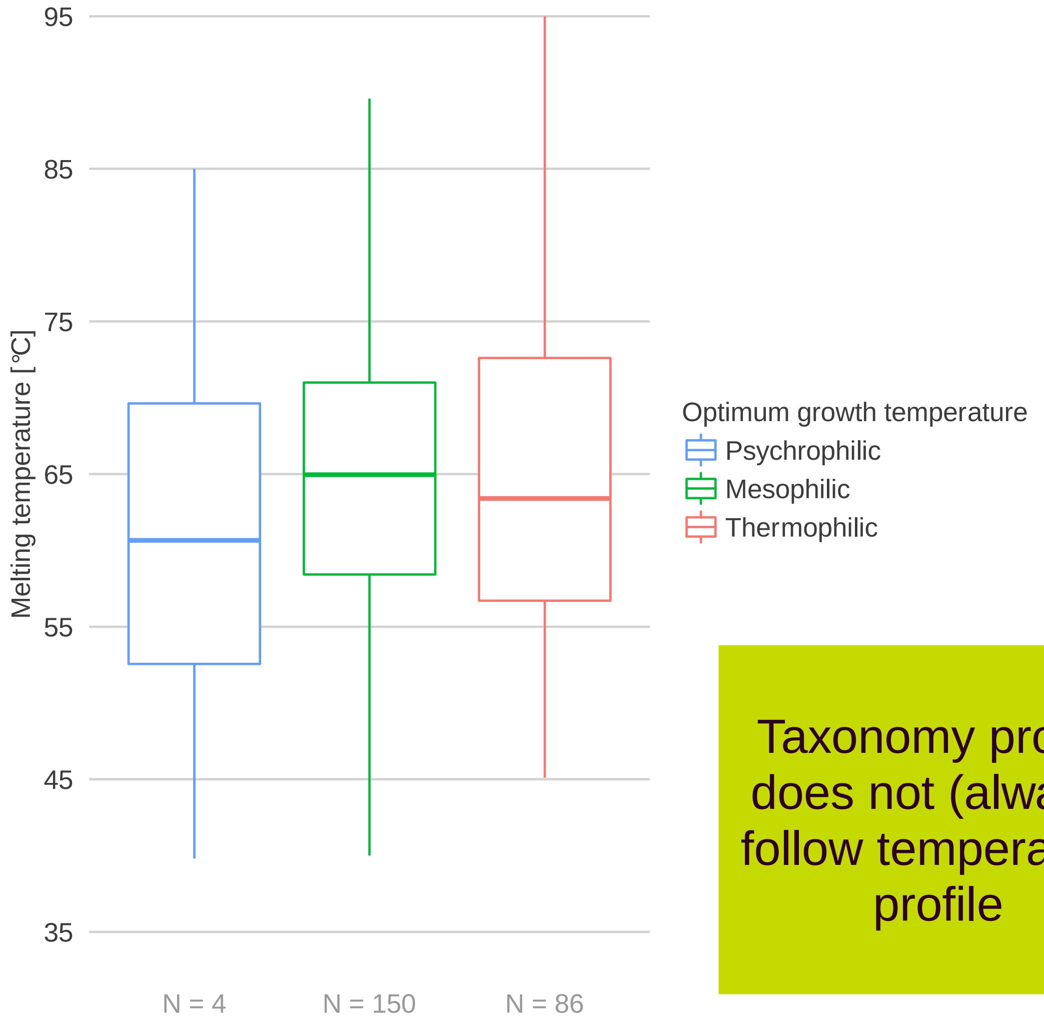


| Sequence ID | Signal peptide | Patents | 3D model | Stability | Production |
|---|---|---|---|---|---|
| 9651813 | ✓ | ✓ | ✓ | ✓ | ✓ |
| 5151351 | ✓ | ✗ | ✓ | ✓ | ✓ |
| 5135992 | ✓ | ✓ | ✓ | ✗ | ✓ |
| 7899632 | ✓ | ✓ | ✓ | ✓ | ✗ |
| 7963218 | ✓ | ✓ | ✗ | ✗ | ✓ |
| 6396321 | ✗ | ✗ | ✓ | ✓ | ✓ |
| … | | | | | |

- Secondary protein structure
- Motifs
- Tertiary structure (determined or calculated)
- AA content
- Surface charge
- Local charge

…

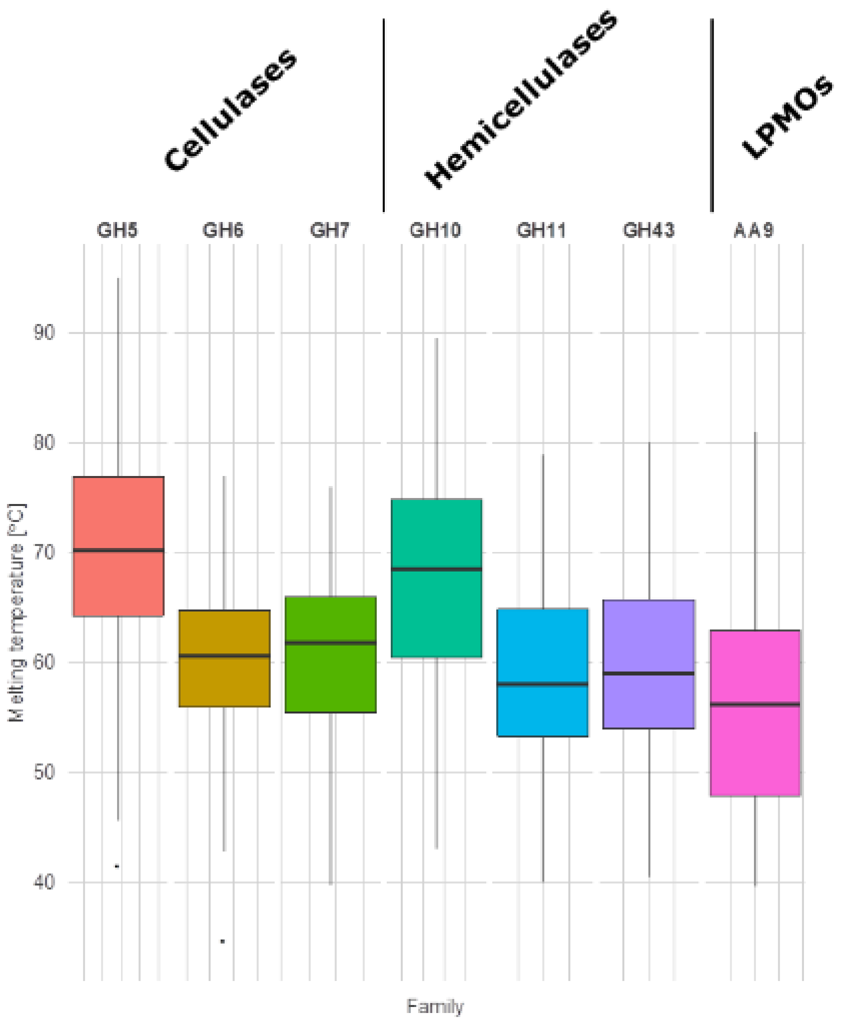
# PREDICTION



Taxonomy profile does not (always) follow temperature profile

Genome meta-data from GOLD

# FACT

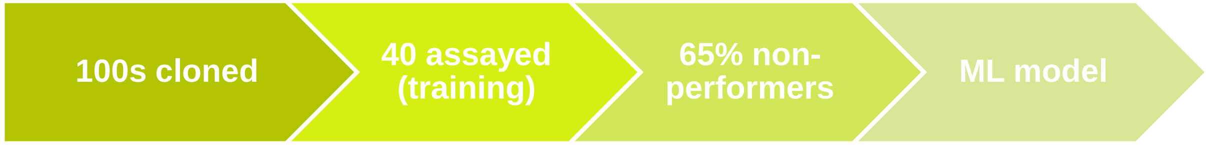


| GH5 | GH6 | GH7 | GH10 | GH11 | GH43 | AA9 | | |
|---|---|---|---|---|---|---|---|---|
| 130 | 99 | 108 | 87 | 61 | 45 | 72 | = | 602 samples |

Expressed at Novozymes

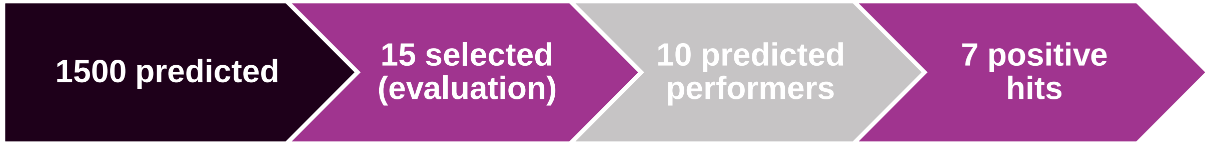
# Machine learning example – Predicting corn fibre solubilisation performance

a)



40 xylanases were screened, but the majority (65%) did not perform well

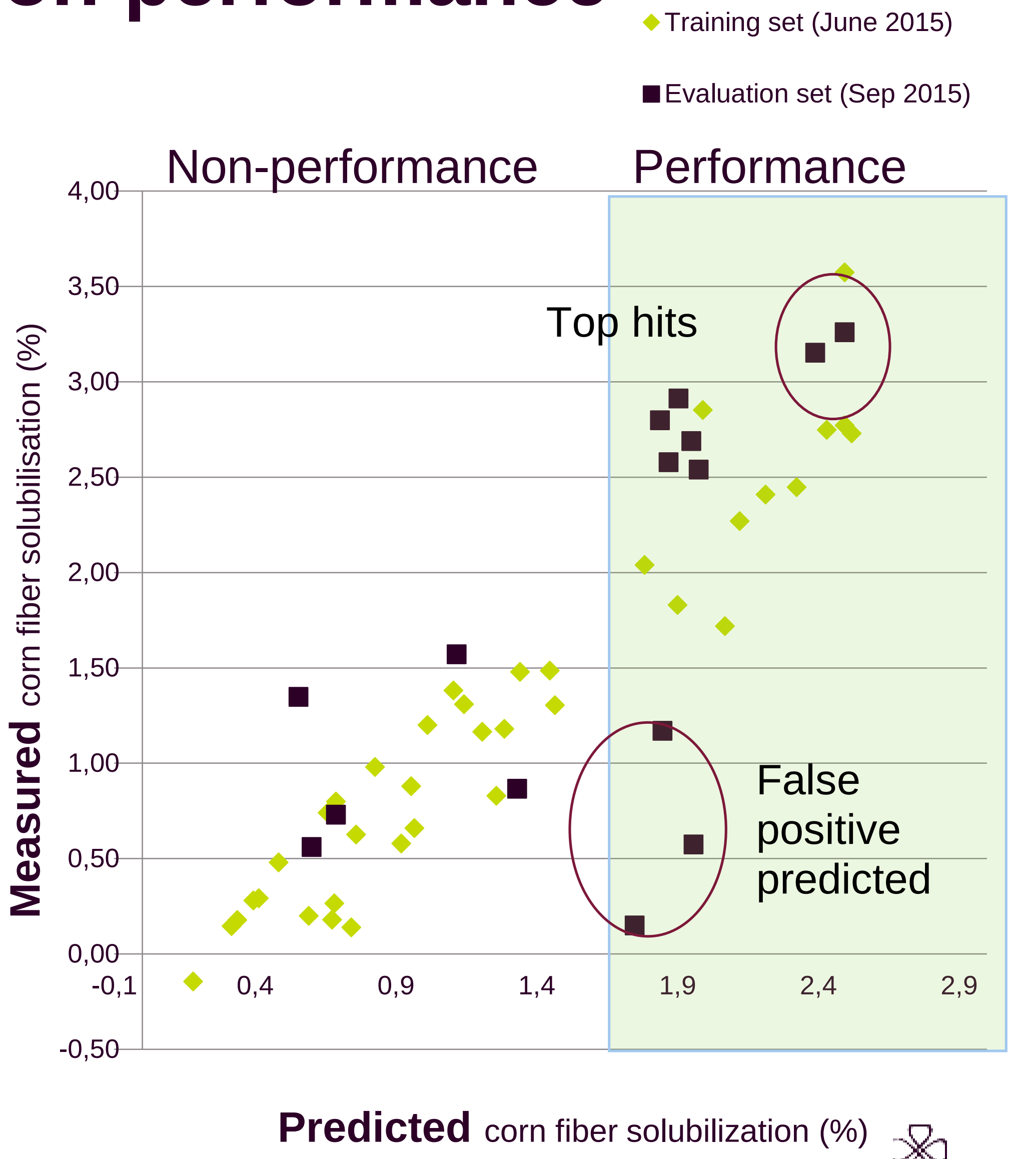
b)



A machine learning model with 100s of examined protein features

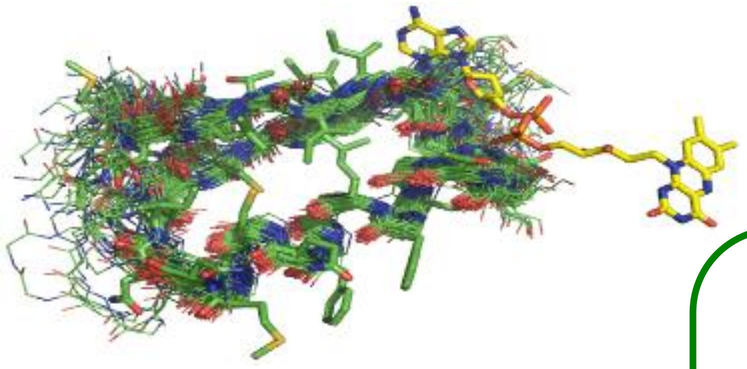
Charge, pI, and hydrophobicity keys to predicting performance

**An increase in "hit rate" from 30% to 70%**



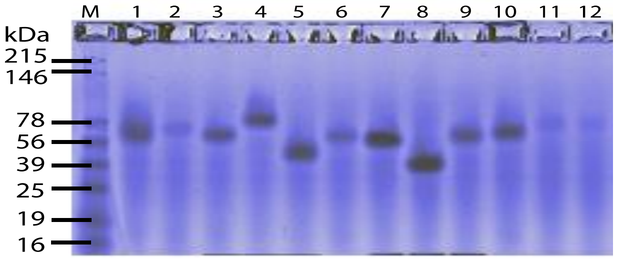
novozymes

# OTHER MACHINE LEARNING ACTIVITIES

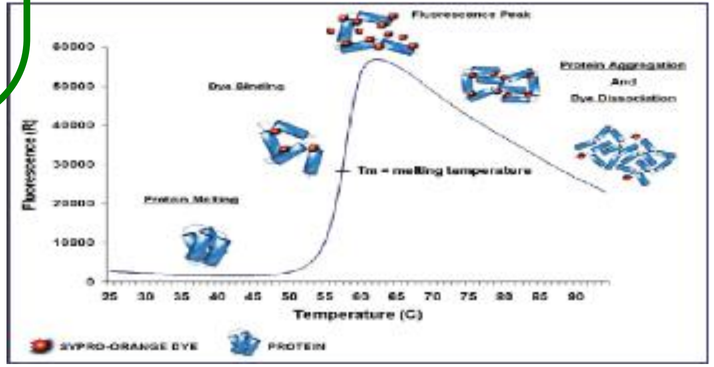


Substrate specificity

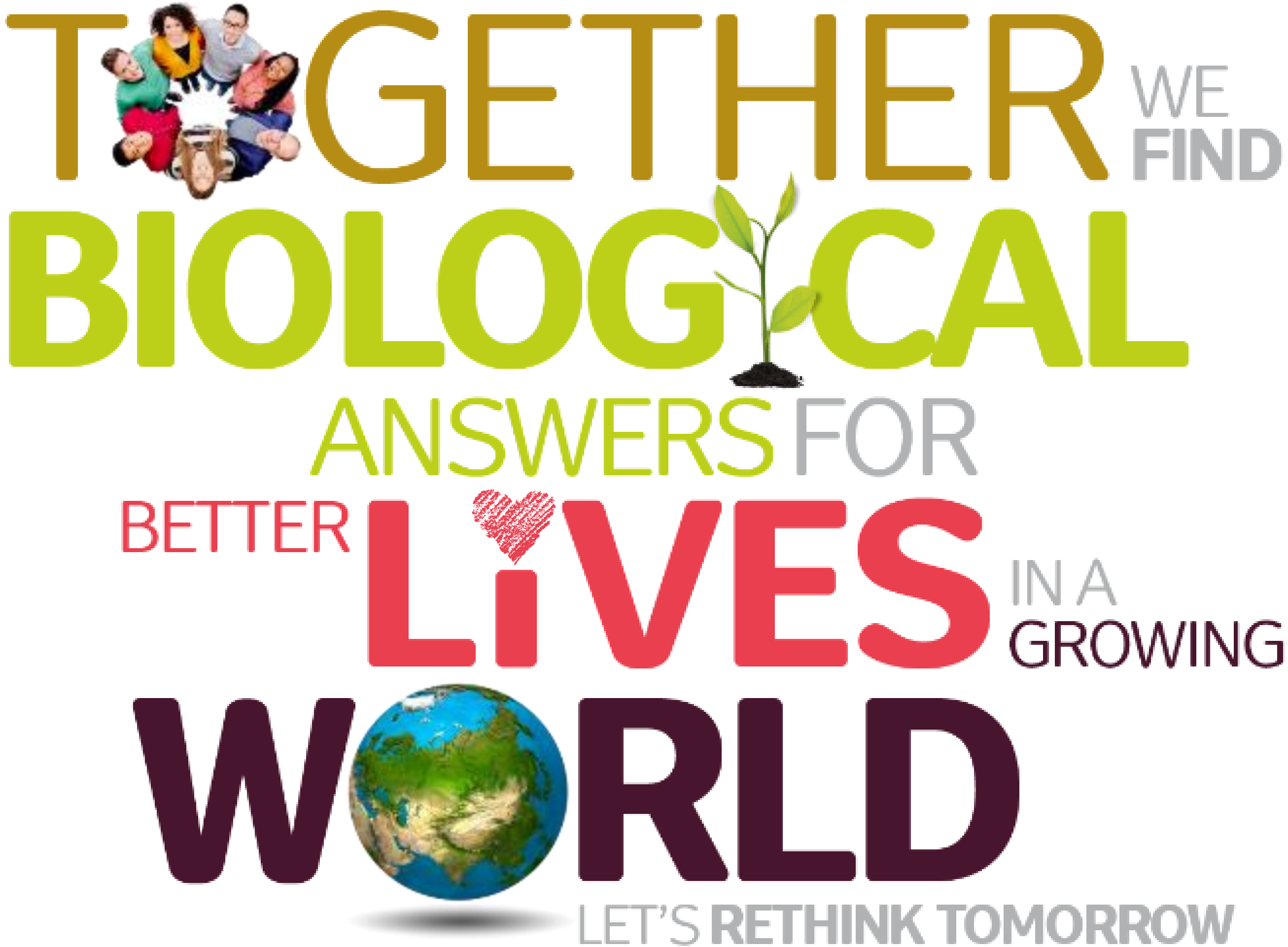
1: YPG, 26°C, 4 days, 7 µL culture supernatant



Expressibility



Stability


TOGETHER WE FIND
BIOLOGICAL
ANSWERS FOR
BETTER LIVES IN A GROWING
WORLD
LET'S RETHINK TOMORROW

novozymes

novozymes®

Rethink Tomorrow