protera

Protera’s AI technologies in focus

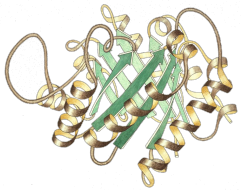
July 2025

Most protein AI hype is about design

protera

Hype: *De novo* design
& functional optimisation

Great idea!



Promising AI-developed candidate

Lab-scale functional sample

Industrial scale-up

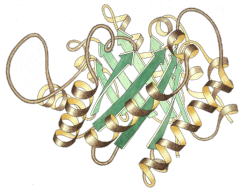
World-changing products

Most protein solutions fail in production

protera

These **roadblocks** come up over and over again

Great idea!



Promising AI-developed candidate

Lab-scale functional sample

Industrial scale-up

World-changing products

Poor expression

Difficulty scaling up.
Poor stability/solubility

Regulation.
IP.

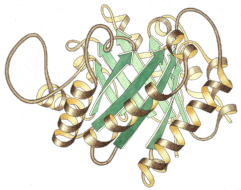
Our AI enforces producibility & compliance

protera

✅ IP-free
✅ Non-allergenic
✅ Non-toxic

✅ Expressible
✅ Soluble
✅ Functional
✅ Stable

Great idea!



High confidence expressible candidate

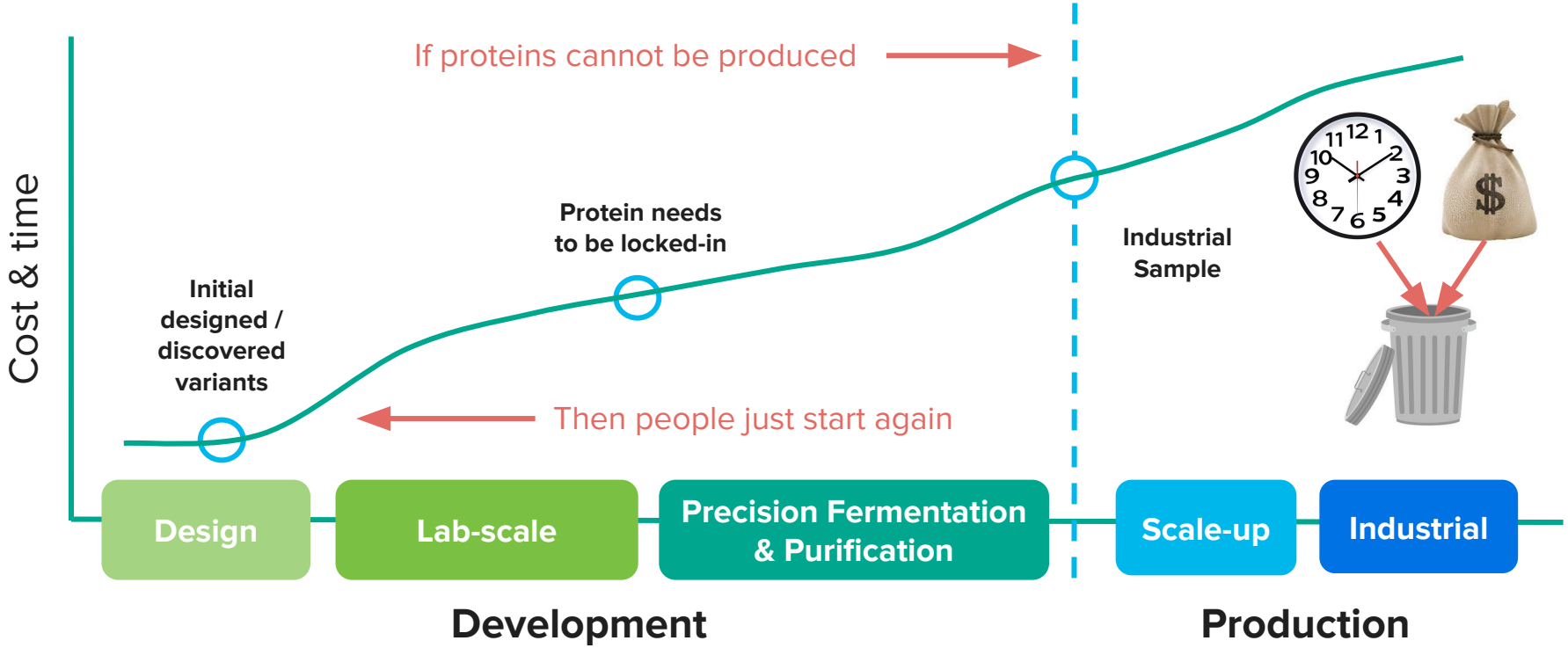
Lab-scale functional sample

Industrial scale-up

World-changing products

Constant lab feedback

Our competitors waste time and money



Our capabilities span three crucial areas

protera

Protein Discovery

Semantic search
Find novel IP-free proteins

Curated GRAS data sets
Patent-free, safe proteins

Safety filtering
Predict allergenicity & toxicity

Function Prediction

Zero-shot models
Data-free protein optimisation

Active supervised learning
Maximise protein function

Protein-peptide interactions
Predict binding affinity

Design for Production

DNA sequence design
AI expression optimisation

Solubility analysis
Avoid aggregation

Stability optimisation
For function & shelf-life

protera

Highlighted technologies:

1 / Semantic protein search

2 / Protein-peptide interactions

3 / Active learning

4 / Predicting expression

protera

Highlighted technologies:

1 / Semantic protein search

2 / Protein-peptide interactions

3 / Active learning

4 / Predicting expression

Problem

Sequence/structure-based search methods miss distant analogs, and often return IP-restricted proteins

Our Solution

Our semantic search can find IP-free proteins with similar function but completely different sequence/structure.

Sequence search doesn't always work

The cat sat on the mat

The cat sat on the flat mat

~~The cat spat on the mat~~

~~They cat-sat Matt~~

~~The cat sat on Matt~~

Sequence-based search **does not always find results with the same meaning**.

Semantic search gives better results

The cat sat
on the mat

Semantic embedding

The feline lounged
on the carpet

The mat had a
cat on it

Dave the kitty
sprawled on his rug

Mačka je sjedila
na prostirci

The cat sat on
the flat mat

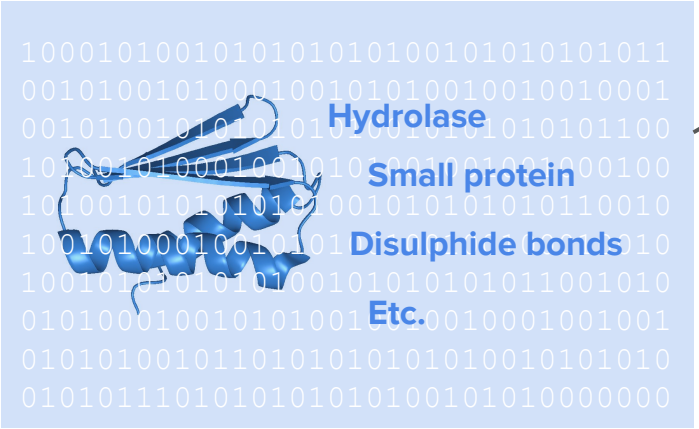
猫坐在 垫
子上

Semantic search can find sentences with the same meaning but **0% sequence similarity**

Semantic search gives better results

MVLSPADKTNVKAAWGKVGA

Semantic embedding



MVIAPSDKTEVLAGWGHVGA

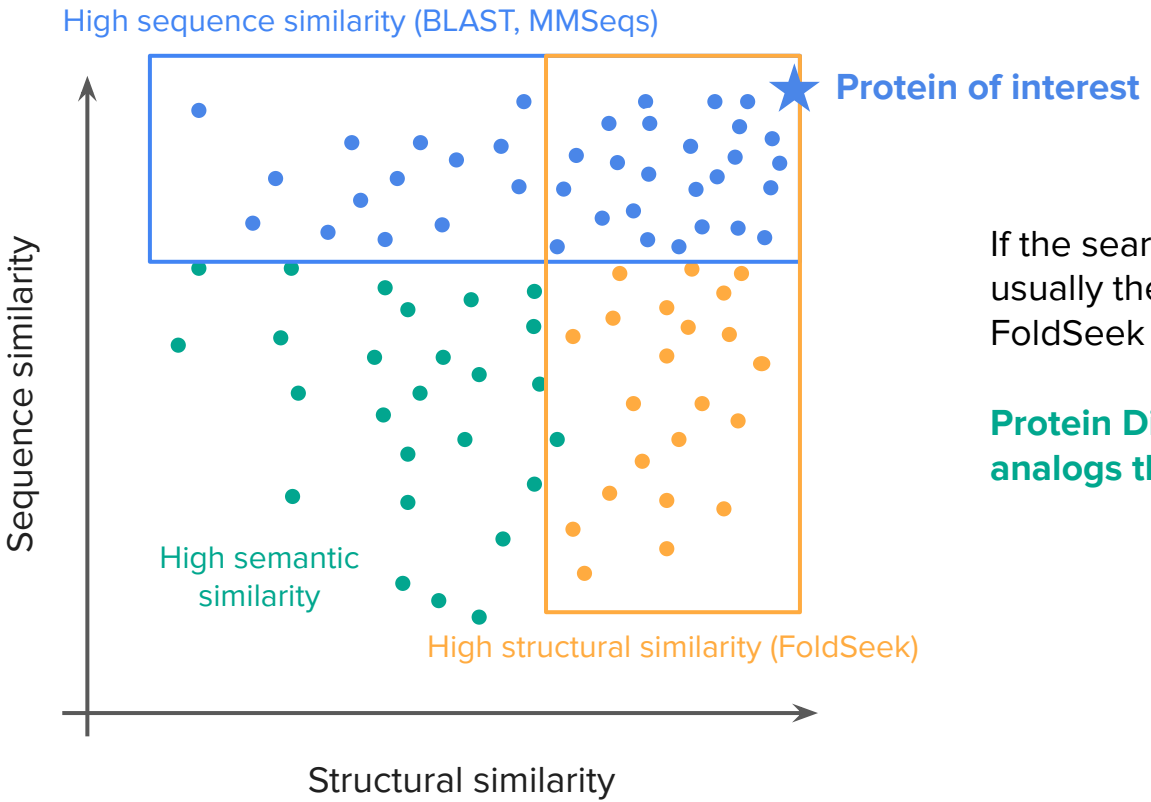
MVLAPADKTNVKAGWGKVGA

MQISTSELHQILGGTAKIAG

KLPOQRTAVNMLGGTAKIAA

Semantic search can find sentences with the same meaning but **0% sequence similarity**

Protera Discovery finds distant analogs

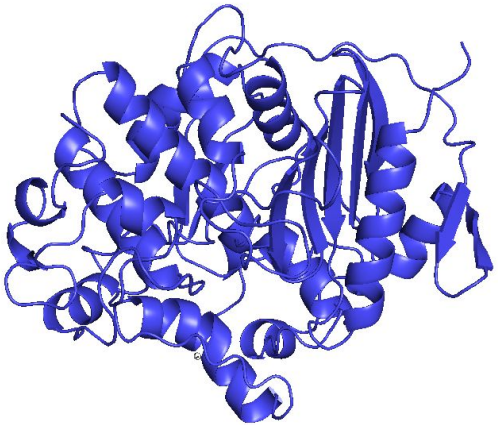


If the search protein is IP-restricted, usually the results found by BLAST and FoldSeek will be too.

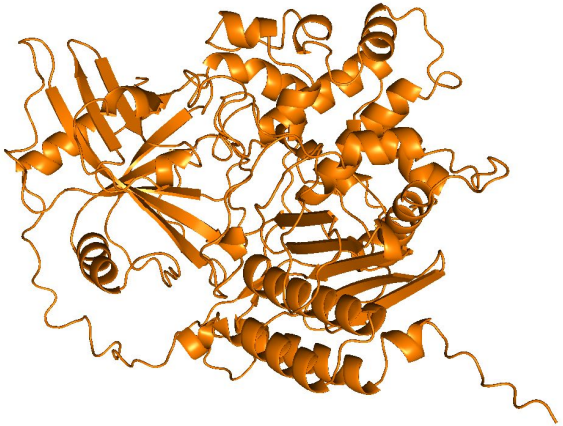
Protein Discovery finds the IP-free analogs that BLAST and FoldSeek miss.

Example: Beta lactamase

Industrial **β-lactamase** enzyme from bacteria used as query:

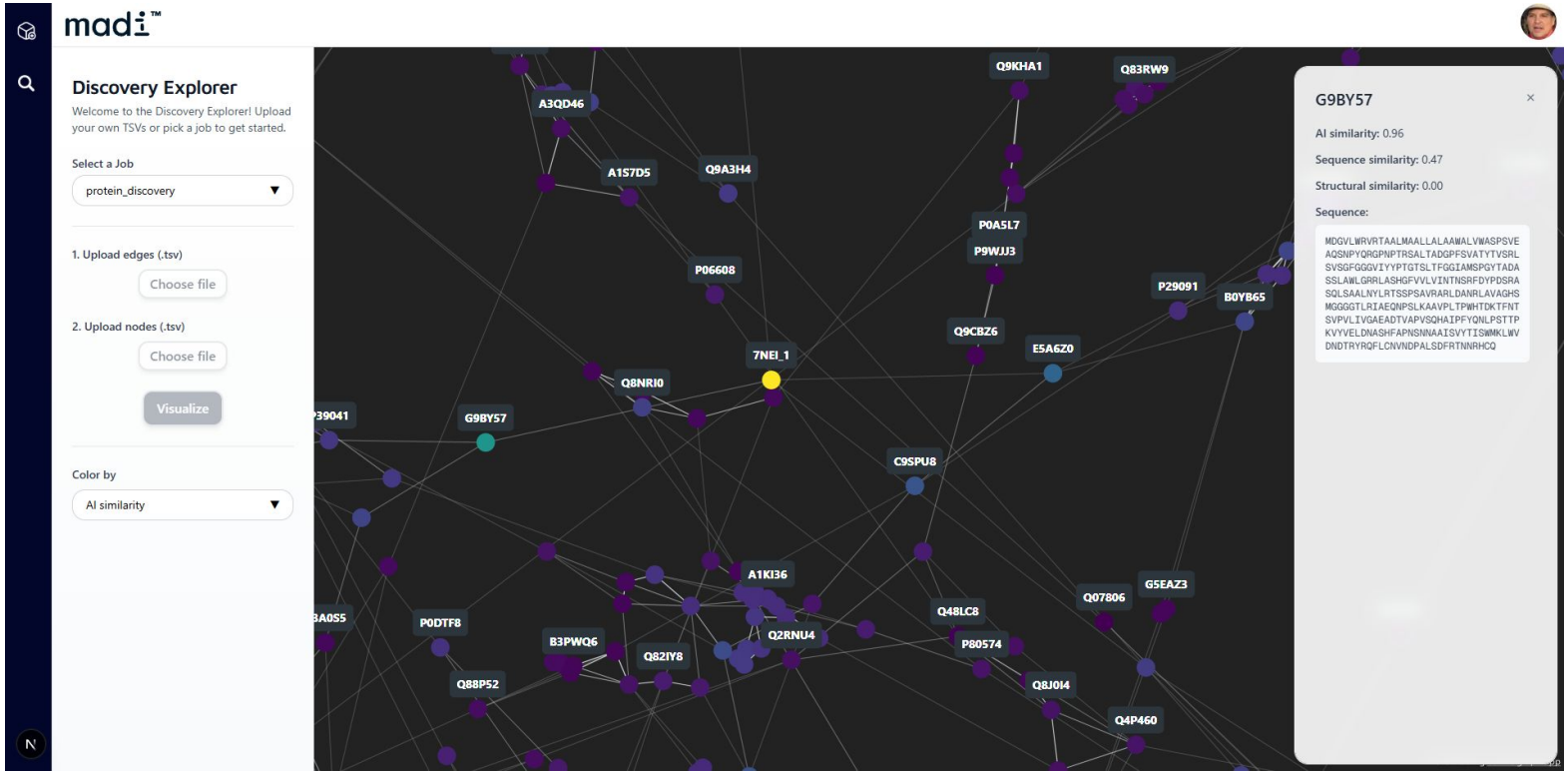


Returned experimentally confirmed **B-lactamase** enzymes from fungi



E. g. 94.6% AI similarity
No sequence or structure similarity

Interactively explore protein space



protera

Highlighted technologies:

1 / Semantic protein search

2 / Protein-peptide interactions

3 / Active learning

4 / Predicting expression

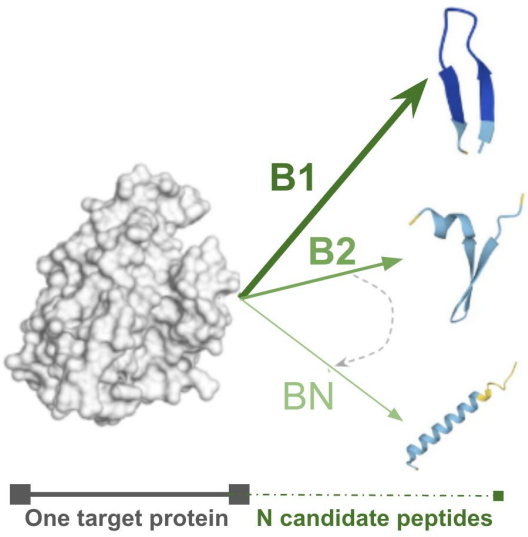
Problem

Understanding peptide-protein interactions (PePIs) is vitally important. Quantitative data to understand their strengths is lacking.

Our Solution

Our unsupervised ranking method can rank peptide-protein interactions without needing labelled data.

Peptide-protein interaction strength matters



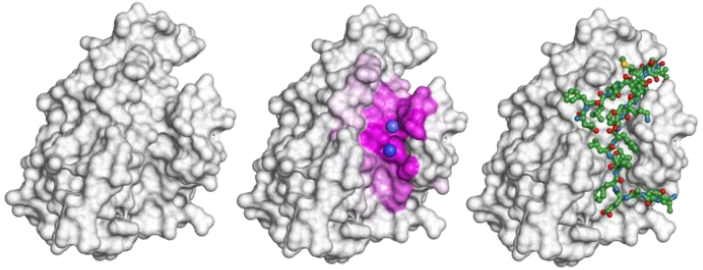
Binding affinities can be ranked

B1 > **B2** > ··· > BN

Identifying the **best protein-peptide matches** is key for bioengineering and therapeutic applications

Unsupervised AIs overcome data scarcity

1 protein-peptide measurement costs **$1000s** and takes **weeks**.

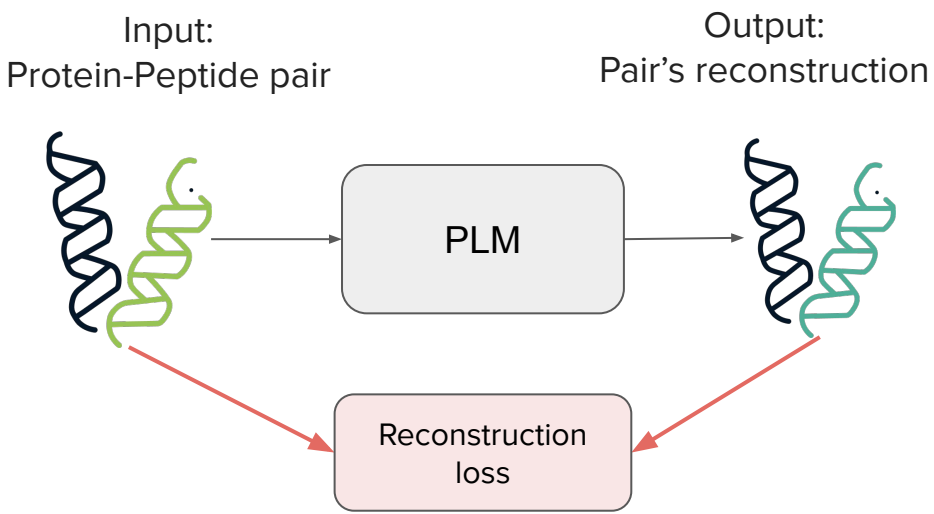


"Supervised" AIs need large amounts of **labelled data**.

We would need thousands of measurements to train a supervised AI.

Unsupervised methods recognise universal patterns and require **0 measurements**.

PLMs understand peptide interactions



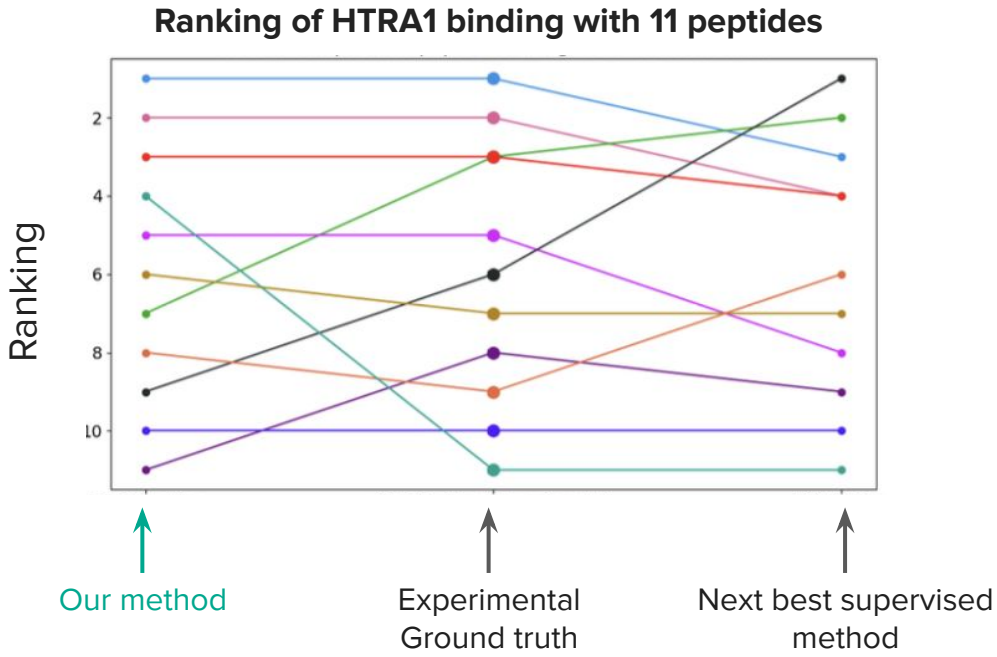
If PLMs (Protein Language Models) understand **intra-sequence interactions**, why not **inter-sequence**?

Our **PLM** ranks **protein-peptide interactions** in an **unsupervised** manner using reconstruction loss

NEURAL INFORMATION PROCESSING SYSTEMS

Our AI can quickly rank binding affinities

We validated the efficacy of our method across **5 different case** studies crucial in cancer, Alzheimer's and HIV research.



Case study: HTRA1 is crucial in **neurodegenerative diseases like Alzheimer's**

Our method can screen **thousands of candidates** in minutes.

protera

Highlighted technologies:

1 / Semantic protein search

2 / Protein-peptide interactions

3 / Active learning

4 / Predicting expression

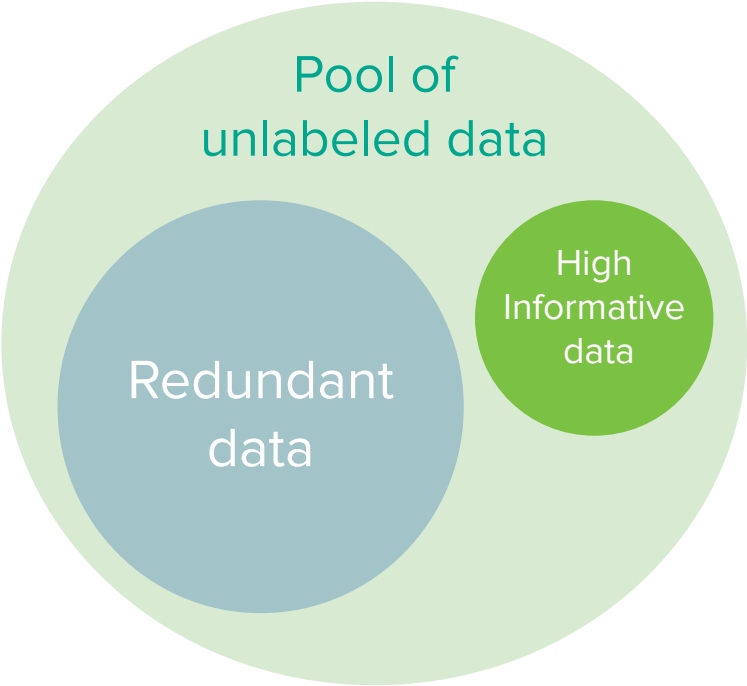
Problem

AI models need real-world data in order to understand proteins. Data is expensive.

Our Solution

Our active learning pipeline reduces costs by telling us which experiments will help our models learn best.

Not all data points are equally informative



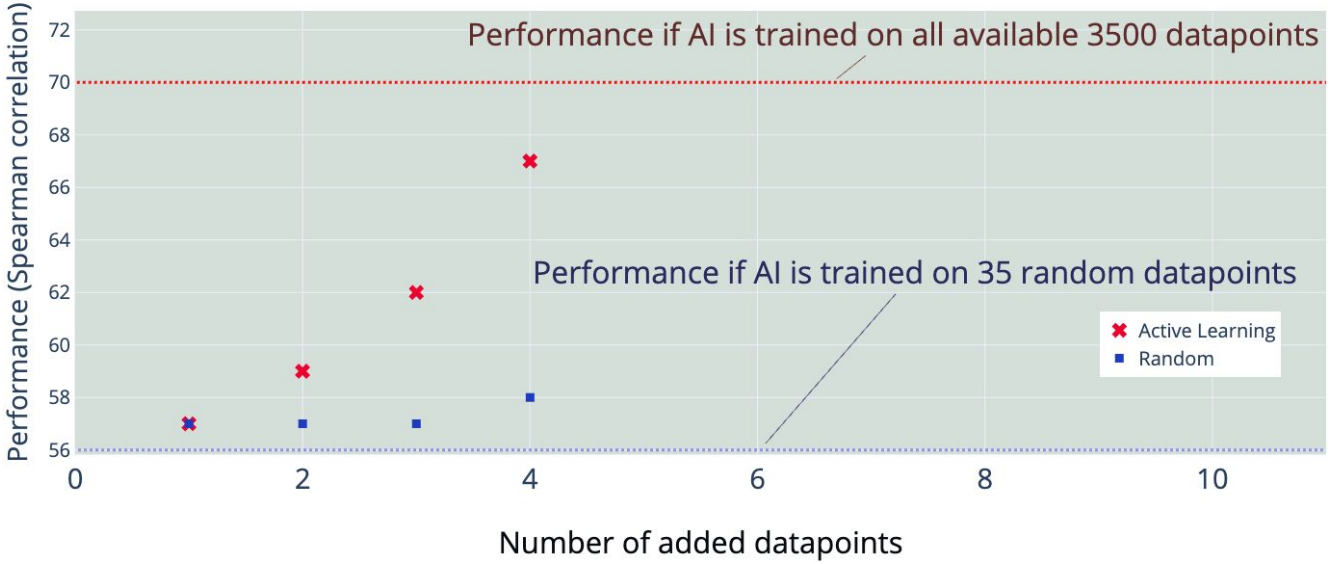
Enriching our training database must involve a **careful selection process**

Active Learning allows our AI to **determine the data it wants to learn from**

Our AI knows best what it needs

Our AI detects the key data points that boost performance, and requests labelling

Performance of our AI when selectively adding highly informative data



10X cost reduction

Boosts project feasibility and success

protera

Highlighted technologies:

1 / Semantic protein search

2 / Protein-peptide interactions

3 / Active learning

4 / Predicting expression

Problem

Standard methods to maximise protein expression are not good enough.

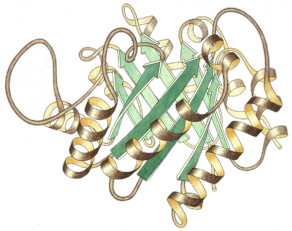
Our Solution

Our proven, best-in-class expression predictor offers precise control over protein expression levels.

DNA sequence choice affects expression

protera

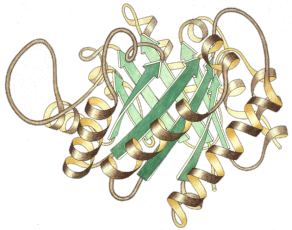
Unoptimised sequence



Same protein

Low yield

Well chosen DNA sequence



Up to 100x higher yield

The choice of DNA sequence can **profoundly affect protein production efficiency**.

Codon “optimisation” tools are blunt

Fixed protein sequence: Met-Ser-Thr-Pro-Gly-Leu-Lys...

Essentially infinite possible RNA encodings:

```
ATG TCC ACC CCT GGT CTC AAA...
ATG TCT ACA CCG GGA CTG AAG...
ATG AGC ACT CCC GGG CTT AAG...
ATG AGC ACT CCT GGG CTT AAG...
ATG TCA ACC CCA GGC CTC AAG...
... ... ... ... ... ... ...
```

Tools like GenScript’s **GenSmart** deliver a single “best” sequence.

- ✗ no context
- ✗ can’t compare existing sequences
- ✗ no fine control
- ✗ only a single option each time

Our AI quantifies expression

Fixed protein sequence: Met-Ser-Thr-Pro-Gly-Leu-Lys...

Essentially infinite possible RNA encodings:

ATG	TCC	ACC	CCT	GGT	CTC	AAA...	6.2
ATG	TCT	ACA	CCG	GGA	CTG	AAG...	34.2
ATG	AGC	ACT	CCC	GGG	CTT	AAG...	92.3
ATG	AGC	ACT	CCT	GGG	CTT	AAG...	65.2
ATG	TCA	ACC	CCA	GGC	CTC	AAG...	0.9
...	...	...	...	...	...	...	

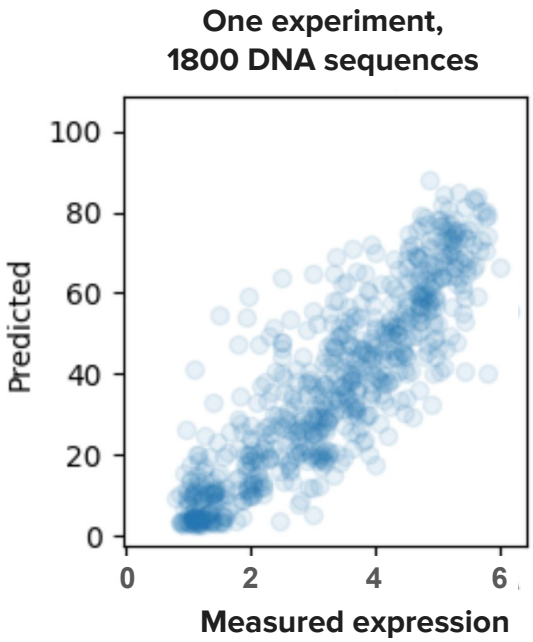
NEURAL INFORMATION PROCESSING SYSTEMS

✅ contextual ✅ can be used to compare known sequences

✅ tune down expression to keep cells happy ✅ can try multiple options *in vitro*

Our AI model performs extremely well

validated on **120k sequences** across **72 experiments**.

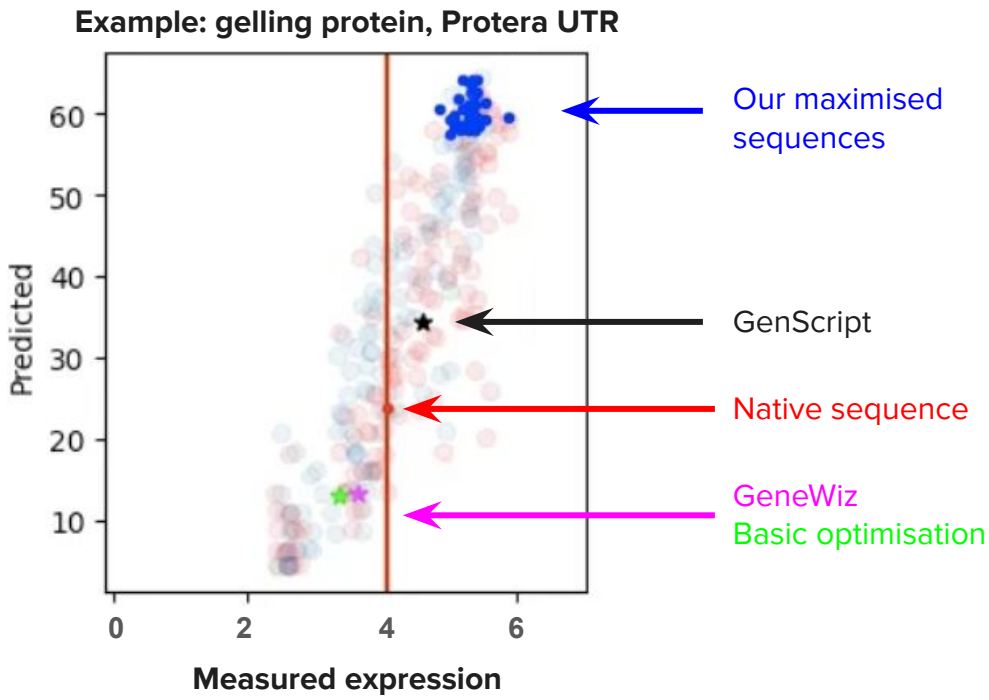


Spearman ρ:

0.88

Largest ever DNA expression study

Our AI outperforms competition



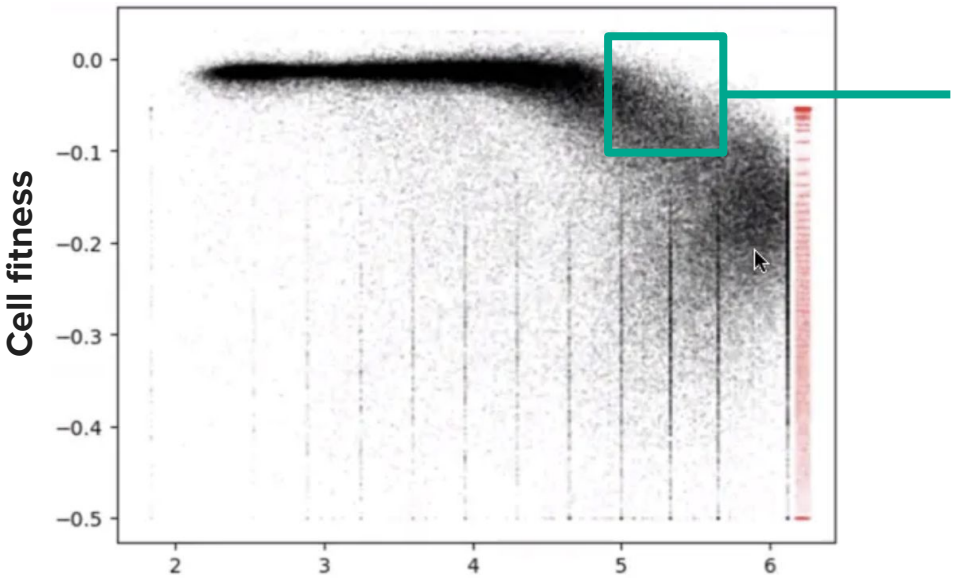
Our AI picked higher-expressing sequences than **GenScript** in

81%

of cases

Our AI helps find the expression sweet spot

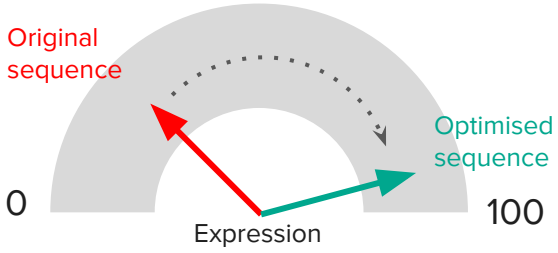
Expression v. cell fitness, all 120k sequences



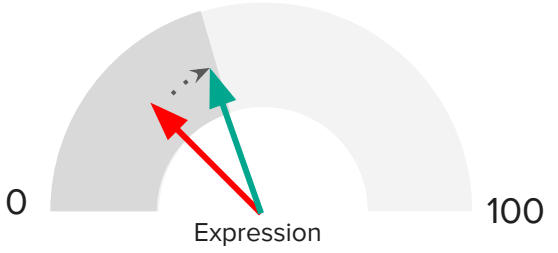
High expression
High fitness

Our AI helps us compare & contextualise

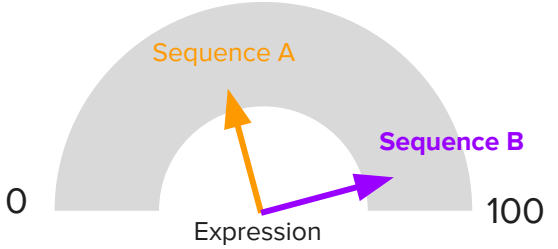
Everyone **assumes** expression optimisation works like this:



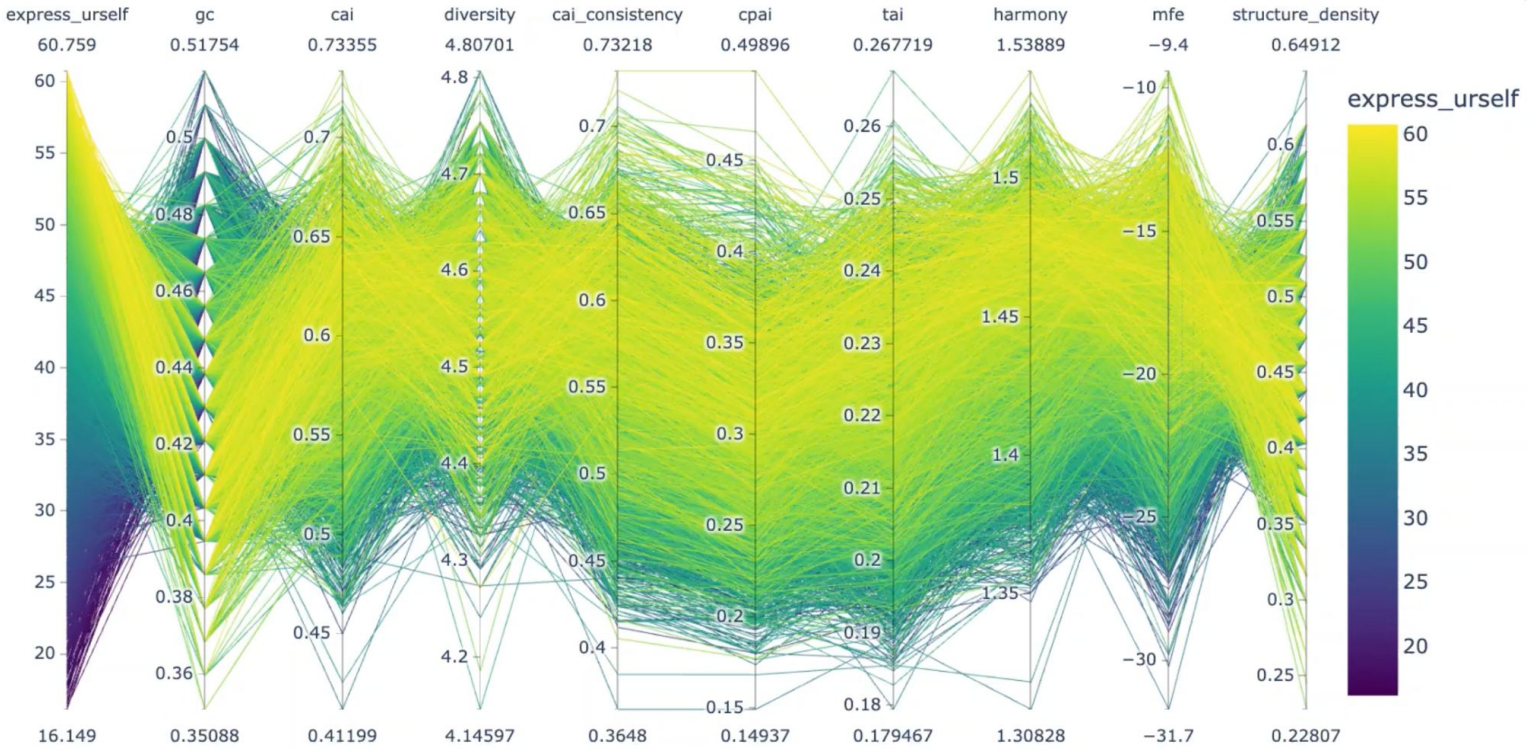
In reality, many proteins don't have **any** high-expressing DNA sequences.
Traditional methods will not tell you this.



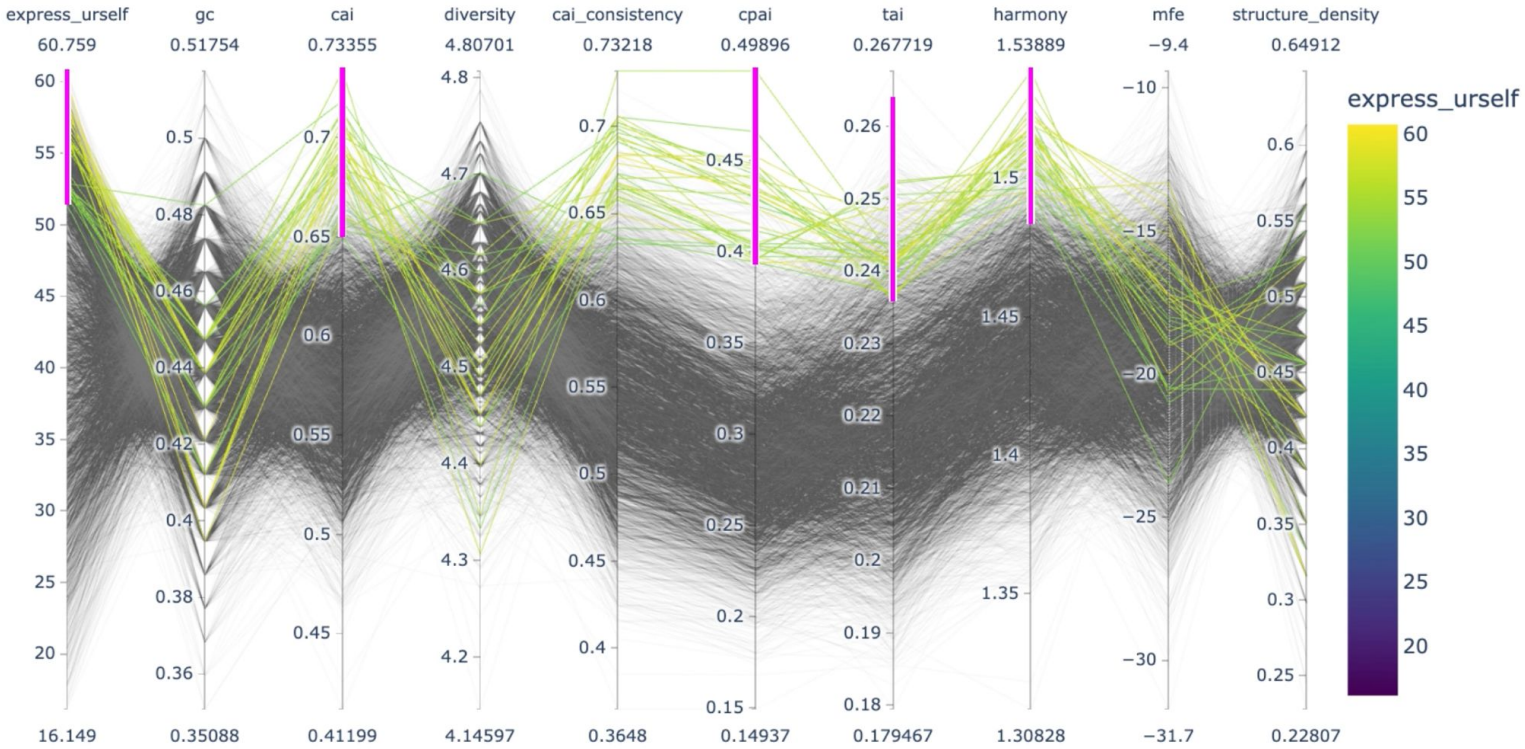
Our method can also compare the **sequences you already have** in stock



Our tool lets you interactively optimise



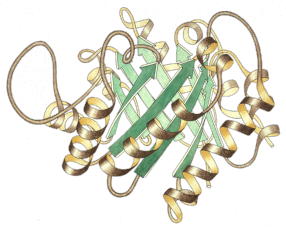
Our tool lets you interactively optimise

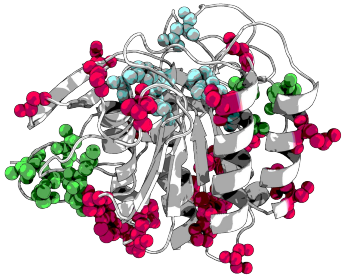


protera

Case studies

BASF — Optimizing a commercial enzyme





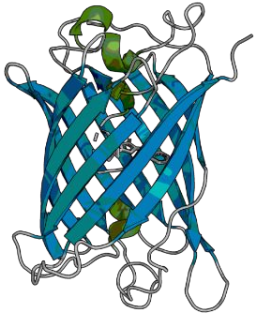
Client optimised an enzyme for **over a decade**

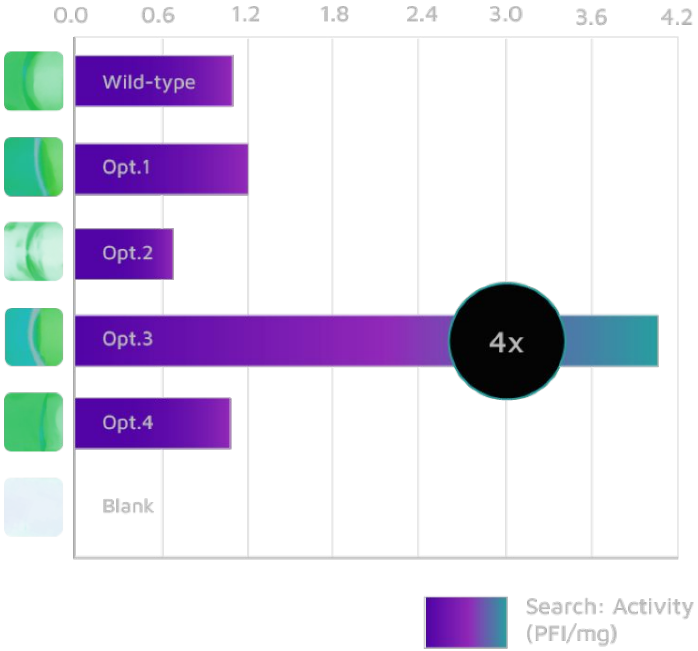
Target:
Protera to optimise application performance

Protera created fine-tuned models and delivered **IP-ready** enzymes with **4x higher performance**

Our optimised GFP – 4x brighter

GFP
Source organism: Aequorea victoria
pI: 5.85
Mw: ~27 kDa
Disulphide bonds: No





Models are trained on **single point mutations to predict high-order mutations** (i.e. more than 2-3 mutations per variant)

protera

Our platform on antimicrobials

Our new focus:

Developing antimicrobials, leveraging our expertise and progress in bakery antifungals

4 years of R&D, lab testing, consumer insights and machine learning give us a **competitive advantage in antimicrobials:**

- Safeguard against **bacteria, fungi, and viruses**
- **€113 billion** market in food, personal care, crop protection, and healthcare

And a **promising candidate for personal care**:

Key microorganisms	Protera Activity
Candida albicans	✅
Escherichia coli	✅
Staphylococcus aureus	✅
Pseudomonas aeruginosa	✅
Aspergillus brasiliensis	=

Our first candidate product ready for testing in application:

- Outperforms existing preservatives in just one AI iteration
- Low allergenicity + high purity
- High performance in neutral pH aqueous solutions
- Production scaled to 40L fermenter, yielding **50g in one month**

Antimicrobials for personal care :
a public health issue

- Preservatives are critical ingredients, yet most are **harmful chemicals**

- Negative consumer perception about **Parabens** ➔ Proven Endocrine Disruptors
  ⇒ will probably be banned in EU in 2029

- **Current replacements** create irritation, allergic reaction, and endocrine disruption and natural ones **are considered hazardous** (CLP classification).

➔ Replacement of harmful chemical preservatives
➔ Low allergenicity product
➔ Non-toxic product
➔ Product compatible with skin needs

Protera product,
antimicrobial plant based protein

Bioactive Natural Compound

specialised protein
with low allergenicity

≥ 90% protein purity

Extended shelf life storage
at room temperature

High efficiency against
Candida albicans and *E. coli*

Conclusion

A quick recap

protera

1 /

Our **semantic protein search** finds IP-free candidates with **completely distinct sequences**.

2 /

Our **protein-peptide ranking** method lets us quantify interactions to find beneficial variants.

3 /

Our **active learning** pipeline reduces costs by telling us which experiments will help our models learn best.

4 /

Our proven, **best-in-class expression predictor** offers innovators precise control over protein expression levels.

protera

Thanks!