KEMP PROTEINS

Protein Problem-Solvers:
QUALITY EXPRESSED

Leveraging Bioinformatics and Machine Learning in Protein Development Workflows – the Expression System is Key

Carter A. Mitchell, CSO
AEIC
RTP - NC
18OCT2023

Reimagining **Bioservices**

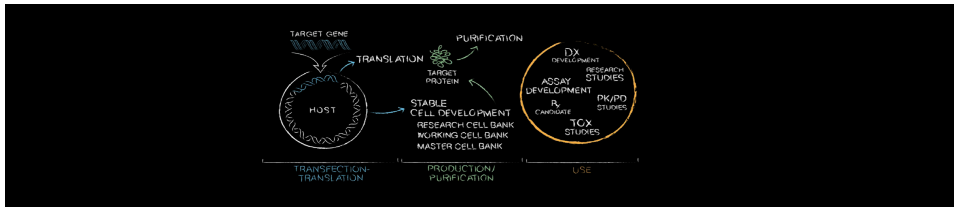
who are KEMP PROTEINS

Kemp Proteins is a US based Bioservices company focused on expressing, purifying and characterizing proteins for use in life sciences

our AIM:

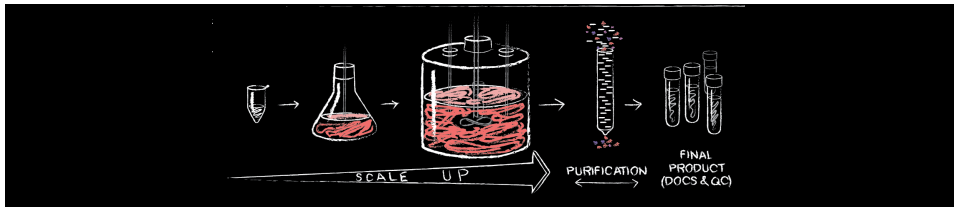
To create value for our clients through a coordinated continuum of unsurpassed quality, communication, efficiency, and satisfaction from design concept through process development to final manufacturing

Range of Complementary Services



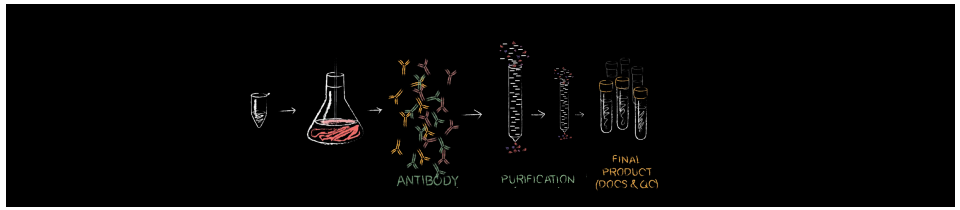
Protein Expression & Purification:

- Gene-to-Protein Services
- Flexible Expression Systems
- Expression Scale mL – 250 L



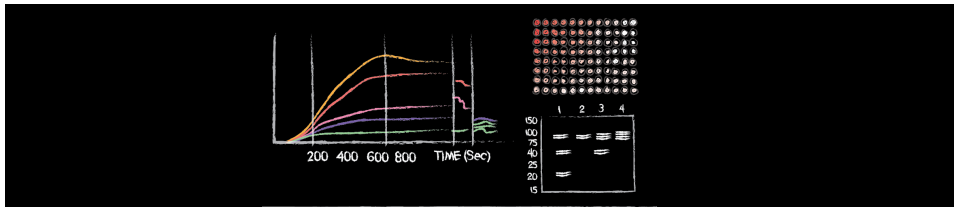
Production:

- USP & DSP Development Services
- Biomanufacturing



Antibodies, Nanobodies, and BiS:

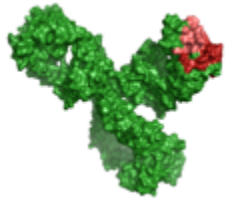
- Custom Ab and anti-ID Development
- mAb and Multi-specific Production
- Hybridoma-to-Recombinant Conversion



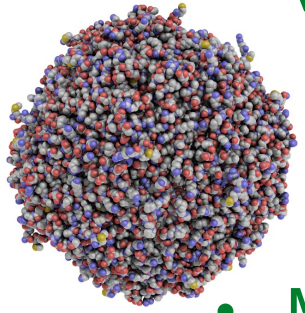
Analytical & Assay:

- Biophysical Characterization
- Analytical Chromatography
- Stability Studies
- Assay Development

Proteins Expressed at Kemp in Various Platforms

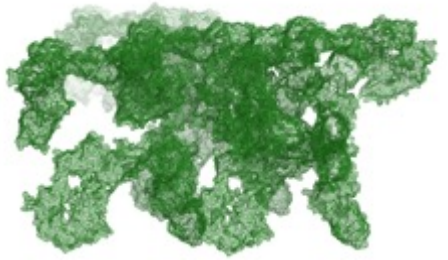


- Antibody Class of Molecule
  - Full-Length - AB Fragments
  - Multi-specific Antibodies
  - Nanobodies – VHH & scFv
  - IgM

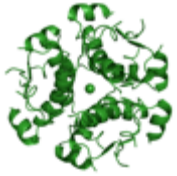


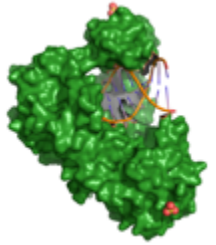
- Virus Related Proteins and Particles
  - Viral Glycoproteins
  - Virus-Like Particles
  - Nanoparticles
- Multi-protein Complexes
  - E3 ligases with targets

- Toxins
  - Ricin, Botox, CtXB



- Membrane Proteins
  - GPCRs
  - Extracellular Domains
  - QTY TM mutants
- Immunologic Proteins / Hormones
  - Hormones
  - Interleukins
    - IL27, IL23, IL13, IL15

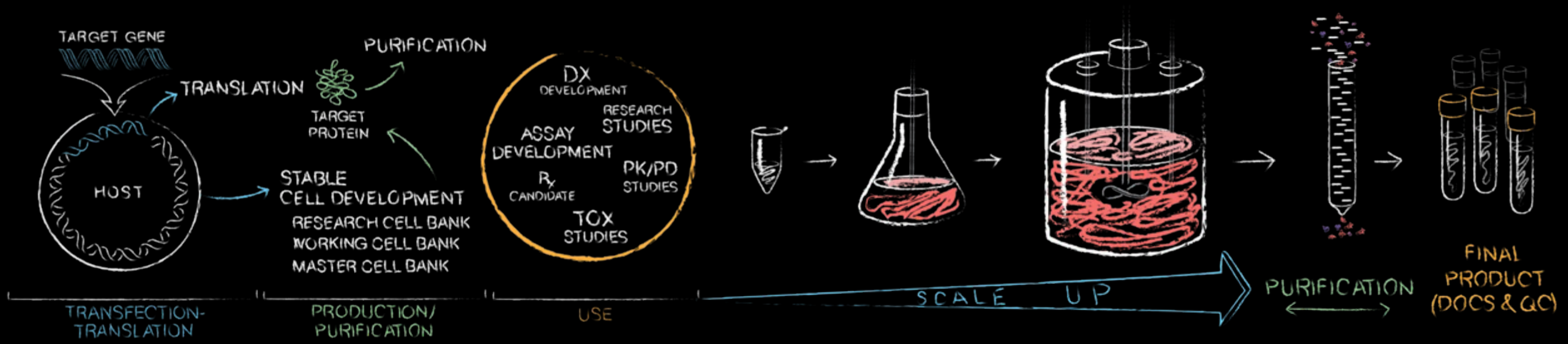




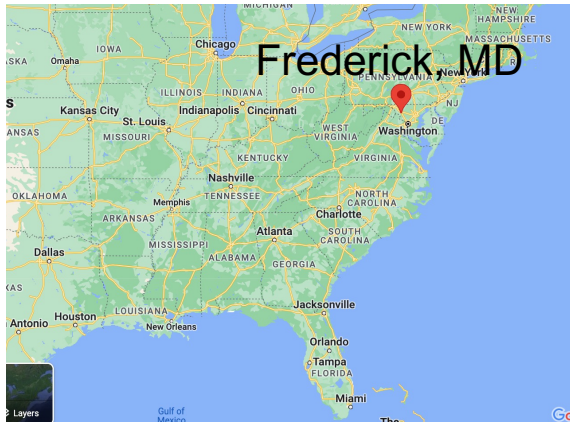
- Enzymes
  - DNA pol
  - Methyltransferases
- Cytoskeletal Elements
  - Dystrophin SILAC and unlabeled[1]

[1]Soderstrom, et. al. 2022: https://doi.org/10.1208/s12248-022-00776-0

KEMP PROTEINS



Kemp Proteins provides a full-continuum of gene-to-protein production services from discovery (mg) to pre-GMP (G)



The Protein Person's Protein People

KEMP PROTEINS

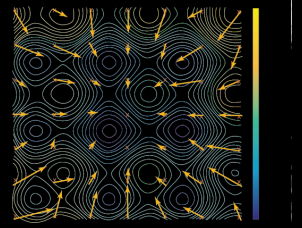
YOUR GMP "READY" PROTEIN

COLLABORATIVE DISCOVERY
WHAT IS THE PROTEIN?
WHAT INFO IS AVAILABLE?
WHAT IS THE USE?
WHAT QUANTITY IS REQUIRED?

DOE EXPRESSIONS TO IDENTIFY OPTIMUM HOST SYSTEM AND UPSTREAM CRITICAL PARAMETERS

VIRUS SPIKE
IDEAS
BISPECIFICS
ANTIBODIES
ENZYMES
VLP
OPTIMIZATION
TRANSIENT EXPRESSION
SCREENING
PURIFICATION
TEST EXPRESSION
DNA
VECTORS
PLASMIDS
SELECT FROM 1000'S OF PROTEINS

"IN HOUSE" SEQUENCE OPTIMIZATION



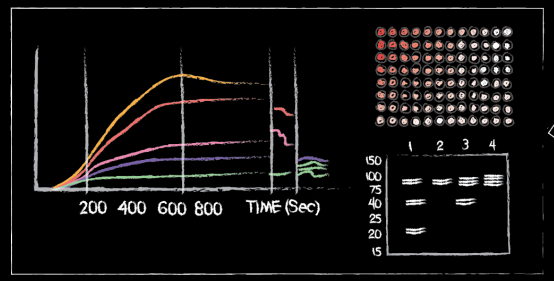
DOWNSTREAM PROCESS OPTIMIZATION

CLARIFICATION
ULTRA/DIA FILTRATION
VIRAL INACTIVATION?
BIND/ELUTE
VIRAL INACTIVATION?
POLISH
VIRAL FILTRATION?
ULTRA/DIA FILTRATE FORMULATE

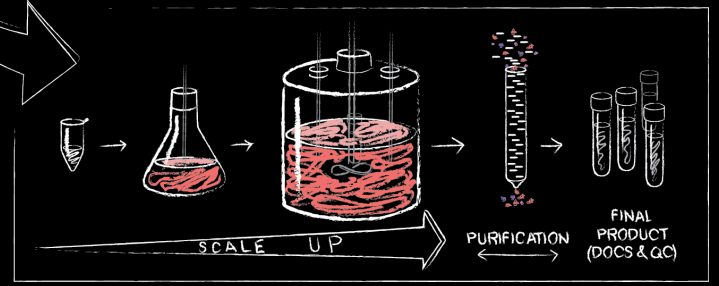
DOWNSTREAM PROCESS DEVELOPMENT IS AN ITERATIVE PROCESS

FINAL PROTEIN
USED IN:
PROCESS DISCOVERY ASSAY DEVELOPMENT
PHASE 1 PHASE 2 PHASE 3
CLINICAL DEVELOPMENT FOR FUTURE

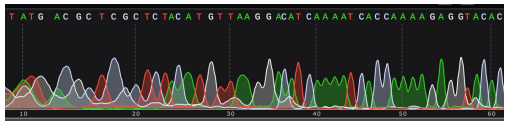
SUPPORT DOSSIER FROM KEMP PROTEINS

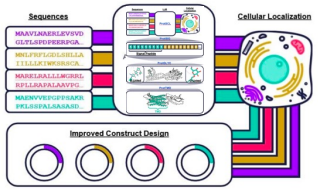


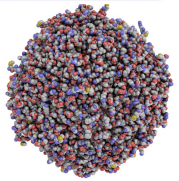
KEMP PROTEINS
QUALITY EXPRESSED

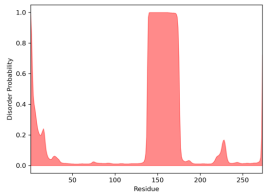


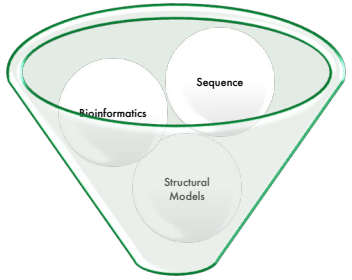
Bioinformatic Informed Process Design











Process Modeling

Sequence

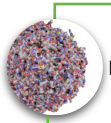
Bioinformatics

Structural Modeling

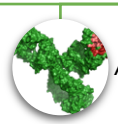
Disorder Prediction

Feasibility Score

Engineering Mode

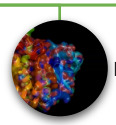


Nanoparticles

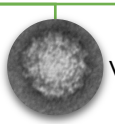


Antibody

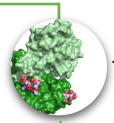
Ectodomains



Fusions

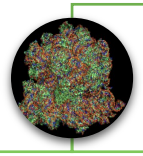


VLP

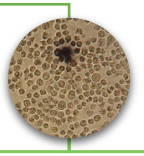


Toxins

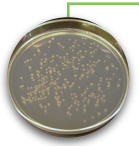
Expression Mode



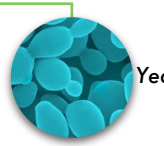
Cell-Free



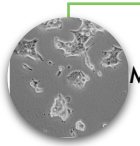
Expression Host



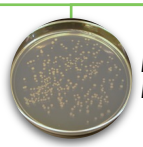
E. coli



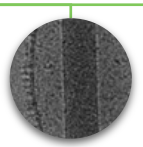
Yeast



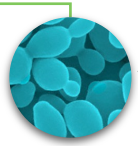
Mammalian



E. coli / Bacillus



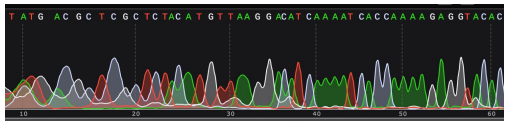
Insect

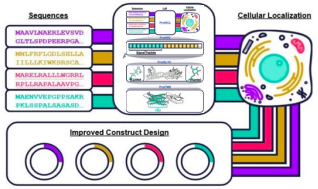


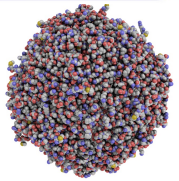
Yeast

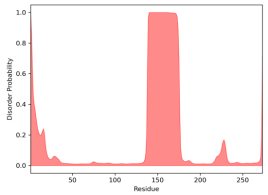
KEMP PROTEINS

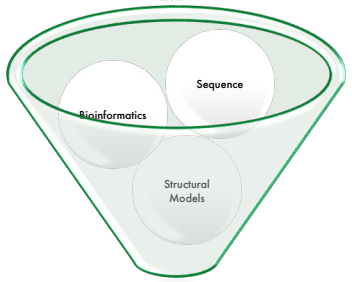
Bioinformatic Informed Process Design











Process Modeling

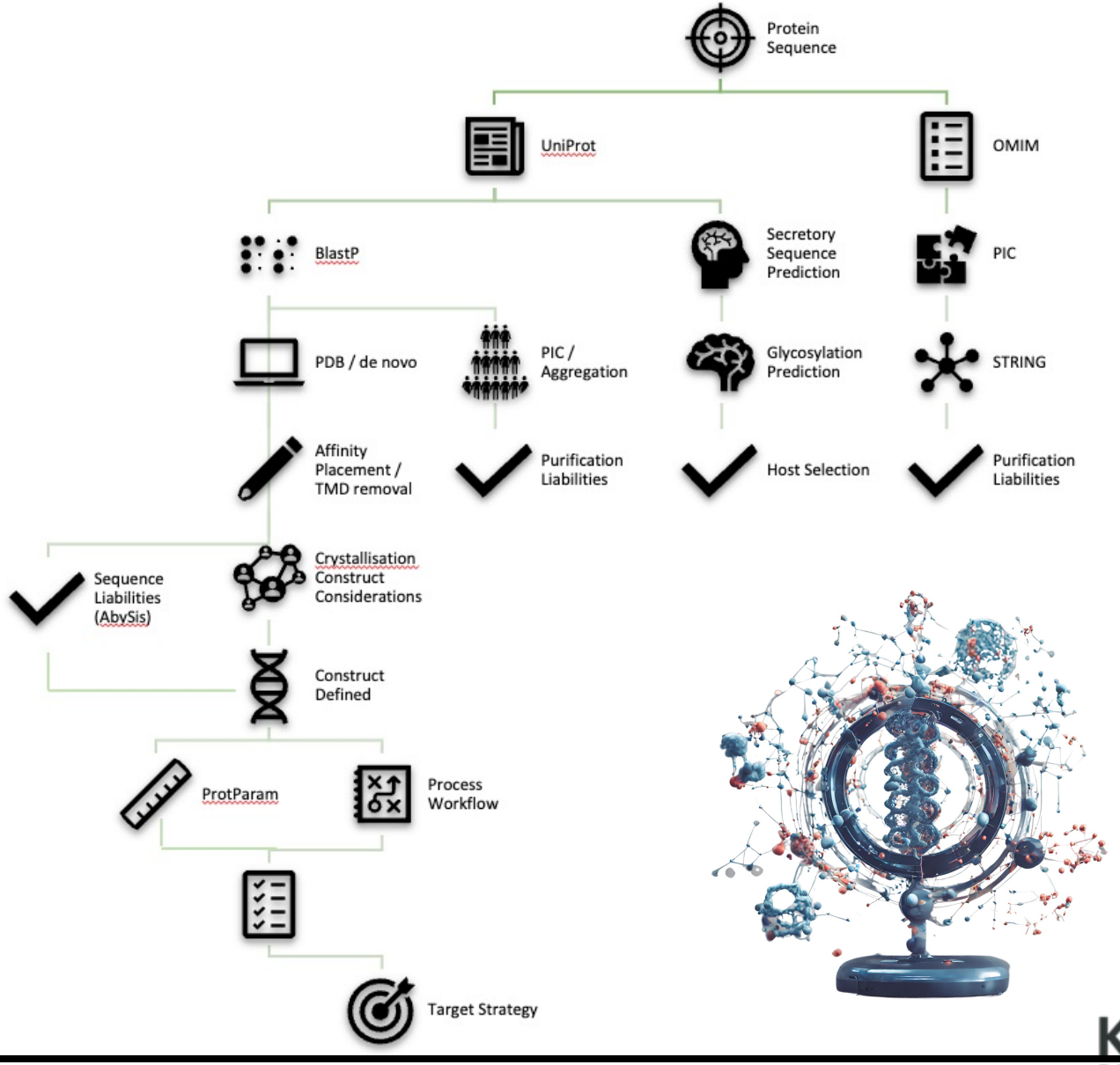
Sequence

Bioinformatics

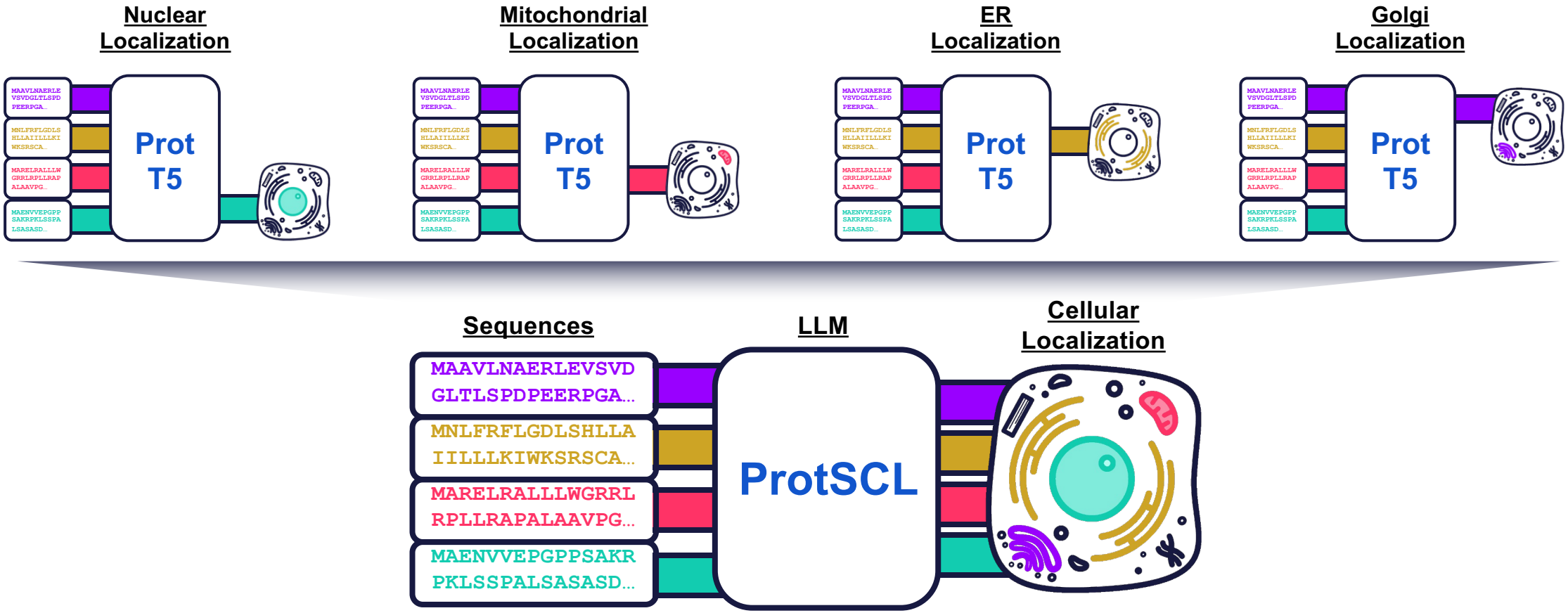
Structural Modeling

Disorder Prediction

Feasibility Score



LLM Implementation for Predicting Localization



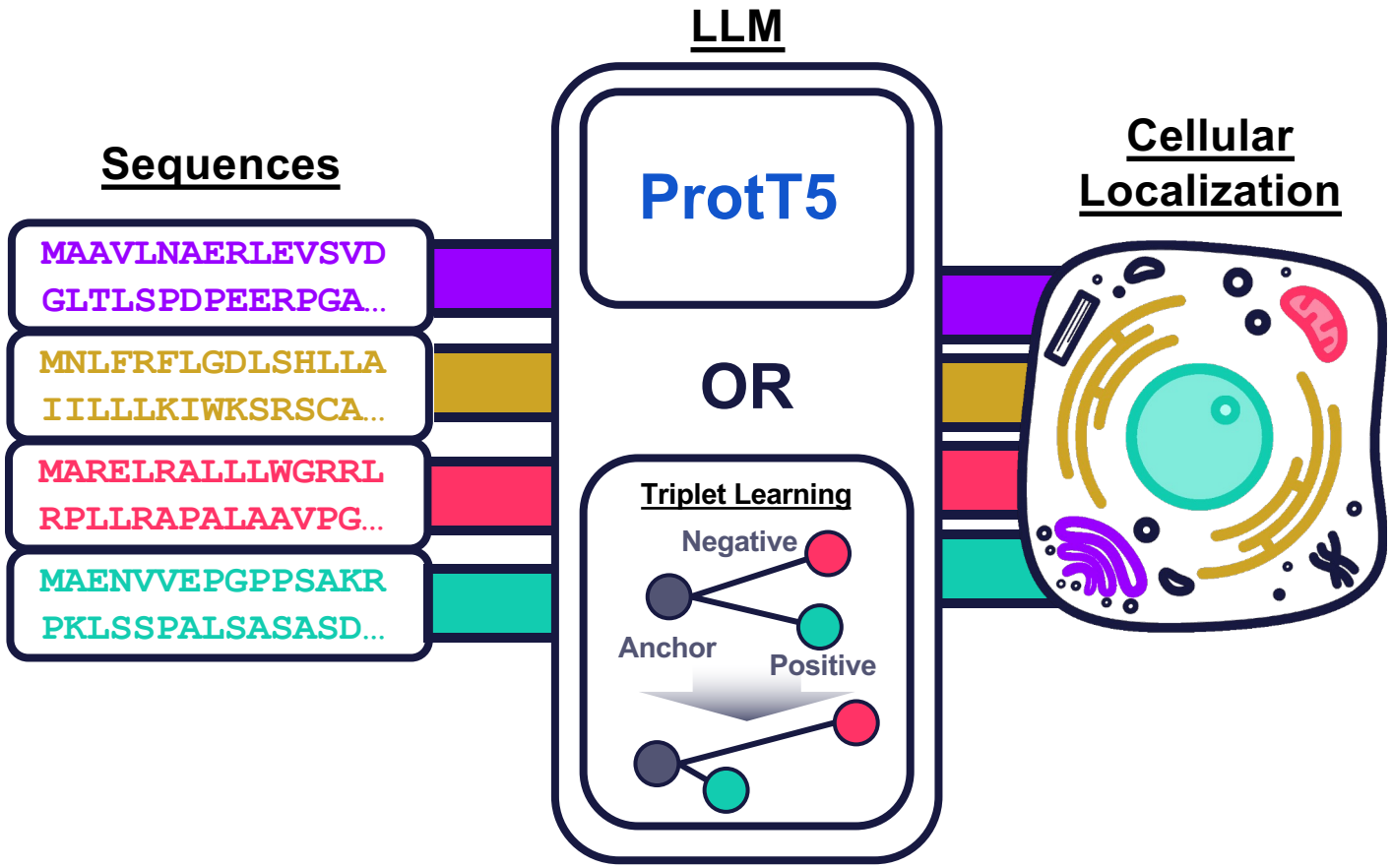
ProtT5 ML models were trained on the entire SWISSPROT DB that enables:

- simultaneous, multi-label prediction
- no need to train models for every cellular localization (potentially less characterized systems too)
- Insight into the Global Proteome as opposed to the hyper-focused, biased systems commonly used

SENTAURI

KEMP PROTEINS

ProtSCL: Subcellular Localization



F1 Score	Triplet	ProtT5
Cytoplasm:	0.73	**0.76**
Nucleus:	0.76	**0.80**
Extracellular:	**0.81**	0.79
Cell Membrane:	0.57	**0.70**
ER:	0.33	**0.52**
Golgi:	0.24	**0.40**
Mitochondria:	0.38	**0.63**

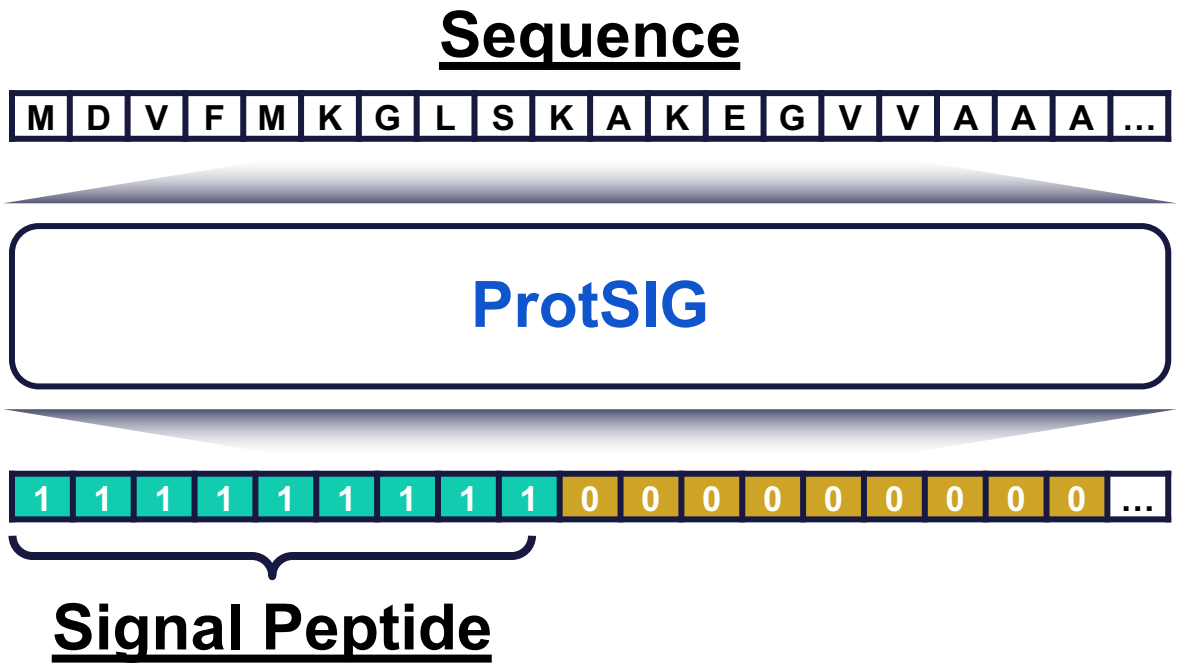
- The ProtT5 model outperforms our Triplet Learning model, which was already state-of-the-art.
- *ProtSCL identified **246 compartments** in mammalian systems (4 compartments from others – purple box).*
- *Understanding the protein maturation process is critical for identifying modes to generate the right protein.*

SENTAURI

[1] Giannakoulias, et. al. *bioRxiv.* **2023**

KEMP PROTEINS

ProtSIG: Signal Peptides Prediction



Accuracy: 99%

	AA not in a SS	AA in a SS
Precision:	**0.99**	**0.98**
Recall:	**0.99**	**0.98**
F1 Score:	**0.99**	**0.98**
Support:	**1,020,657**	**76,729**

- Signal Sequences are often synonymous Secretory Tags – but these are actually targeting sequences.
- SignalP 6.0[1] was trained on only ~25,000 sequences and shows a Matthew's correlation coefficient of 0.87.
- **ProtSIG Matthew's correlation coefficient was >98%.**

SENTAURI

[1] Teufel, et. al. *Nat. Biotech.* **2022**

KEMP PROTEINS

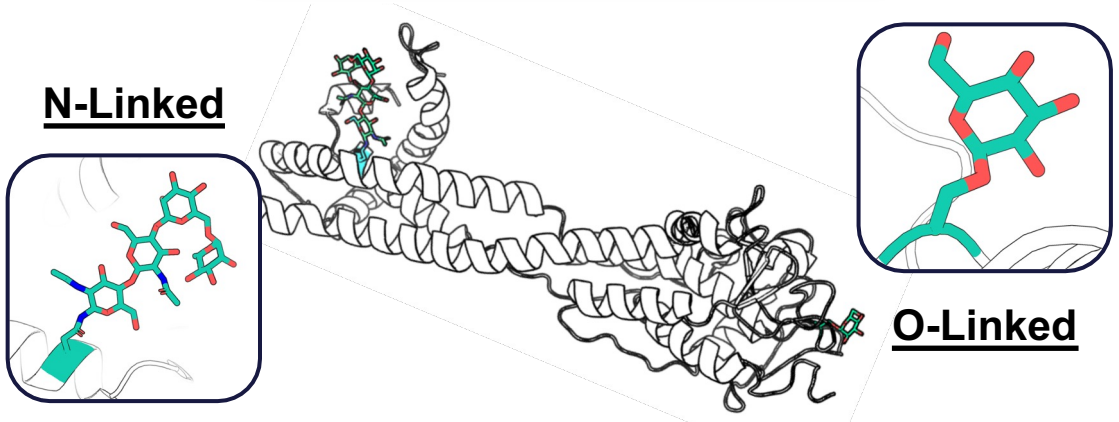
ProtGlyc: N- and O-Linked Glycosylation Prediction

Sequence

M	D	V	F	M	K	G	L	S	K	A	K	E	G	V	V	A	A	A	...

ProtGLYC

N-Linked



O-Linked

Accuracy: 99%

	aglyc. AA	glyc. AA
Precision:	**0.99**	**0.92**
Recall:	**0.99**	**0.94**
F1 Score:	**0.99**	**0.93**
Support:	**1,531,846**	**21,258**

- GPP[1], which outperforms both NetNGlyc and NetOGlyc, has an accuracy of ~92% using a smaller dataset.
- *ProtGLYC has a higher accuracy and was able to detect glycosylations in less characterized microbial and plant species.*
- *The model is being retrained to boost arginyl-glycosyltransferases for better insight*

SENTAURI

[1] Hamby, et. al. *BMC Bioinformatics.* **2008**

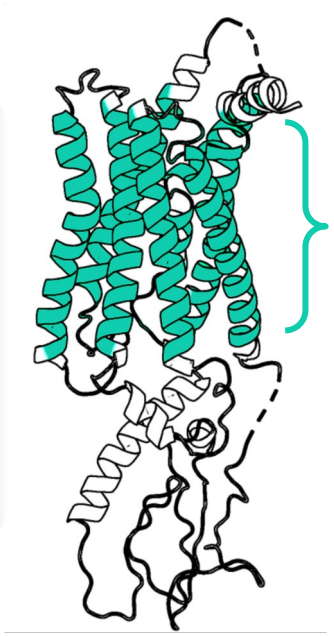
KEMP PROTEINS

ProtTMD: Transmembrane Domain Prediction

Sequence

RPQGATVSLWETVQK
WREYRRQCQRSLTED
PPPATDLFCNRTFDE
YACWPDGEPGSFVNV
SCPWYLPWASSVPQG
HVYRFCTAEGLWLQK
DNSSLPWRDLSECEE
SKRGERSSPEEQLLF
LYYIYTVGYALSFS…

ProtTMD



Accuracy: 97%

	aTMD AA	TMD AA
Precision:	**0.97**	**0.95**
Recall:	**0.98**	**0.95**
F1 Score:	**0.98**	**0.95**
Support:	**2,191,364**	**1,017,832**

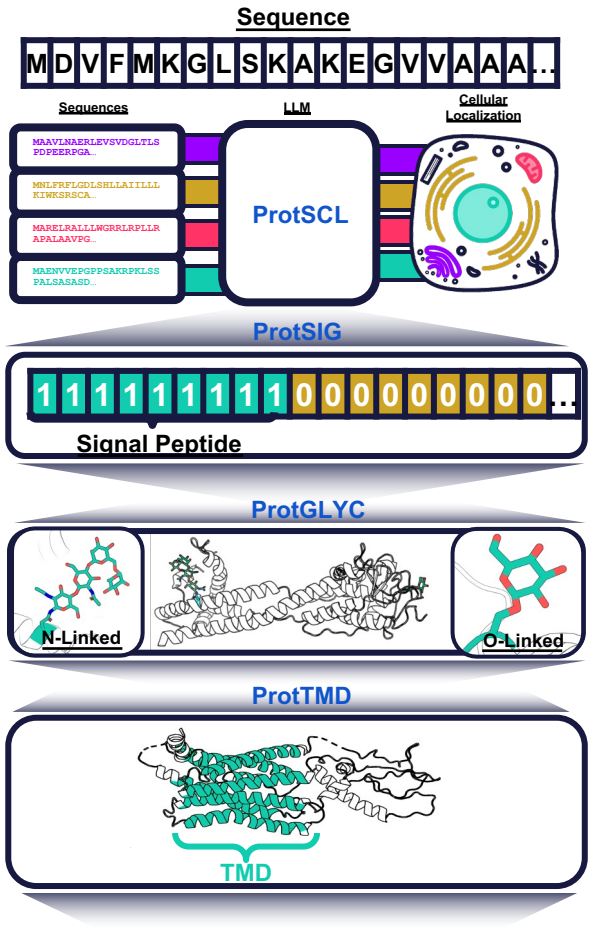
- DeepTMHMM[1] was trained on only 3,574 sequences and shows classification accuracy ~90%.
- **ProtTMD was trained on >3.2 million sequences with a higher accuracy, but more importantly, the model is better at predicting less characterized non-human systems.**

SENTAURI

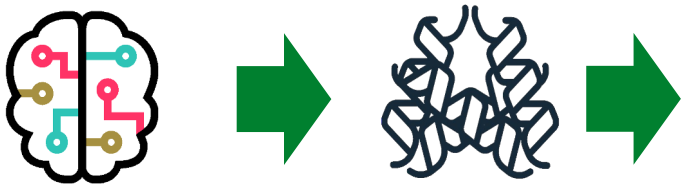
[1] Hallgren, et. al. *bioRxiv* **2022**

KEMP PROTEINS

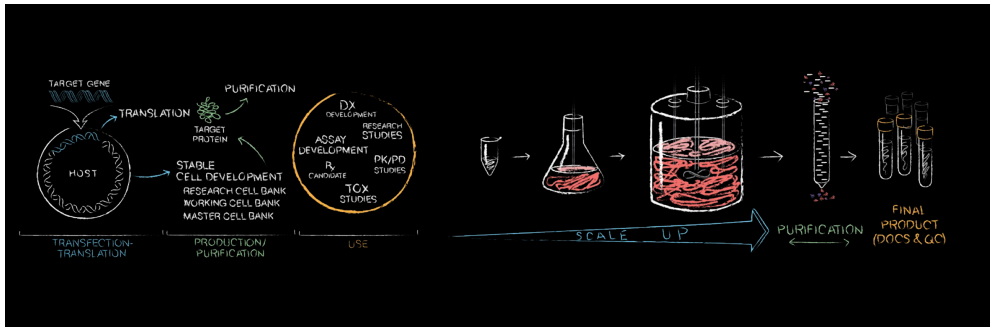
ProtIQ™ Automates Process Modeling



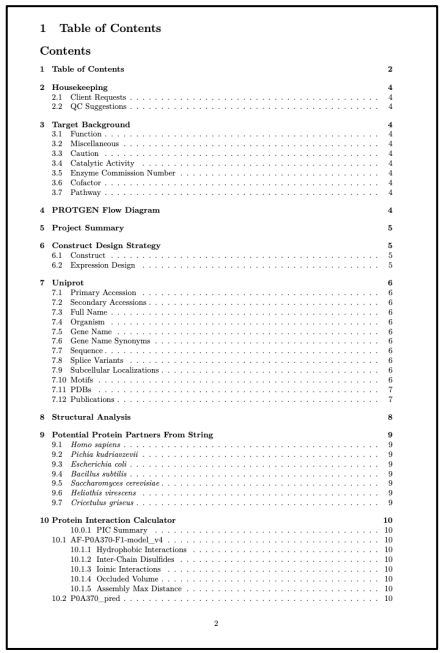
Expert System



ProtIQ™ **ProtDES**



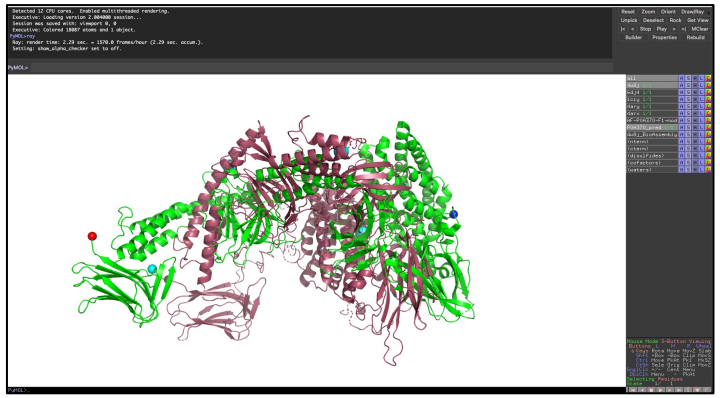
ProtGEN



1 Table of Contents

Contents

2

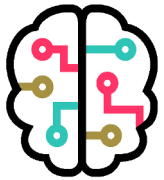


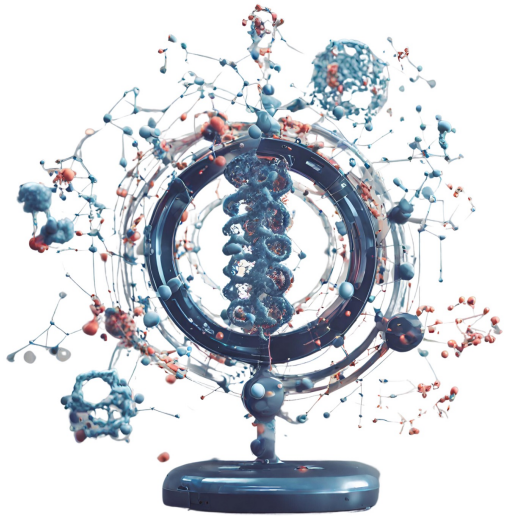
Report and PyMol Session are generated in seconds on a local machine – your IP is protected.

KEMP PROTEINS

Using Bioinformatics and ML to Access the Global Hypothetical Proteome

Expert System





ProtIQ™ **ProtDES** **ProtGEN**

The implementation of multiomics for a specific program can further increases the likelihood of success

- Genomic, transcriptomics, proteomics, and epigenomics datasets
- Development of an organisms specific ML system
- Marginal costs relative to IP
- Provides confidence in less explored systems – we have the capability, why not?

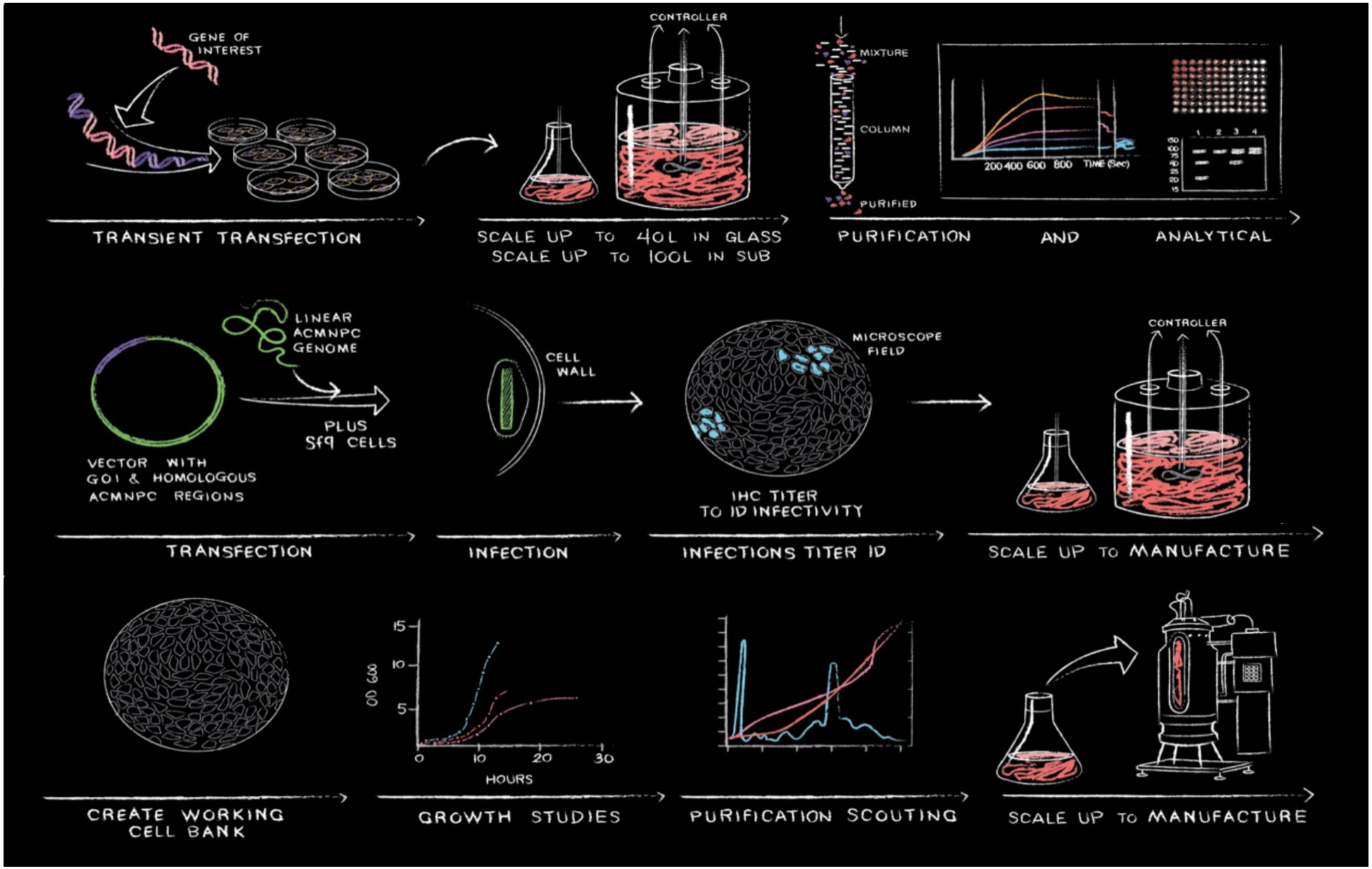
KEMP PROTEINS

Flexibility is Critical - all expression systems assessed

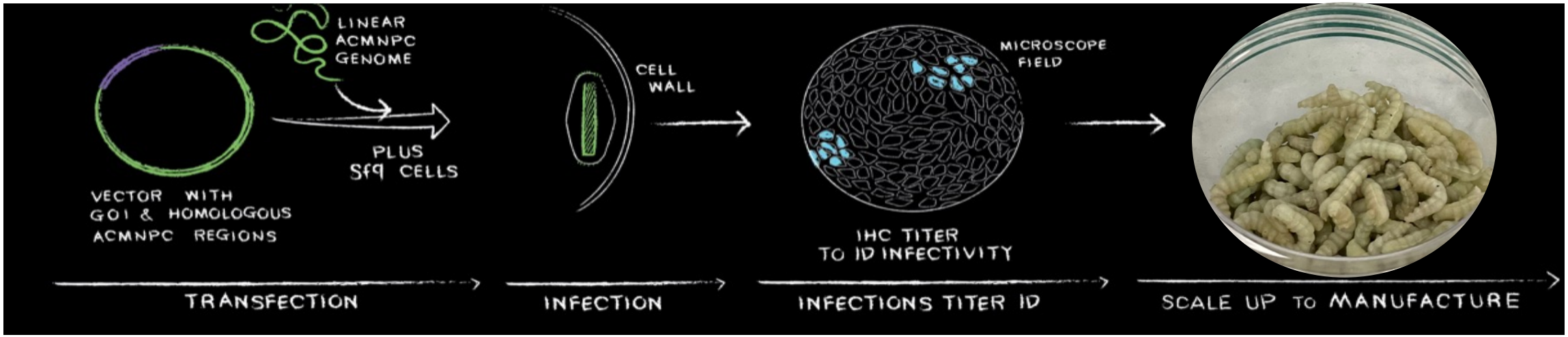
Mammalian

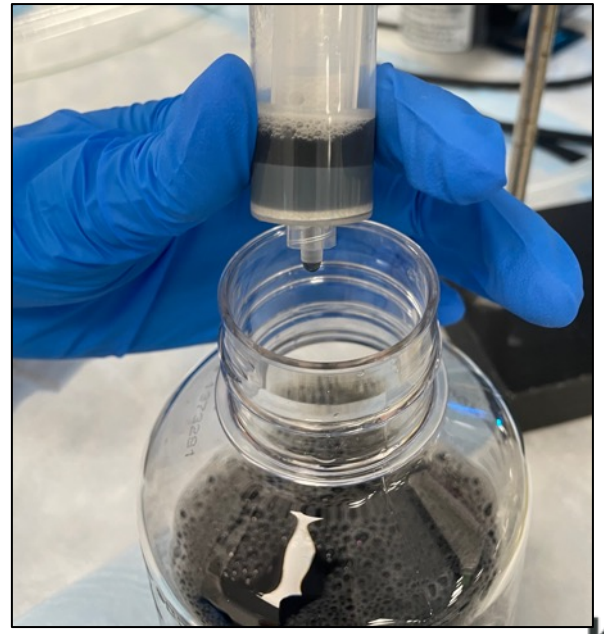
Insect

Microbial



Trichoplusia ni Larvae for a Difficult Target

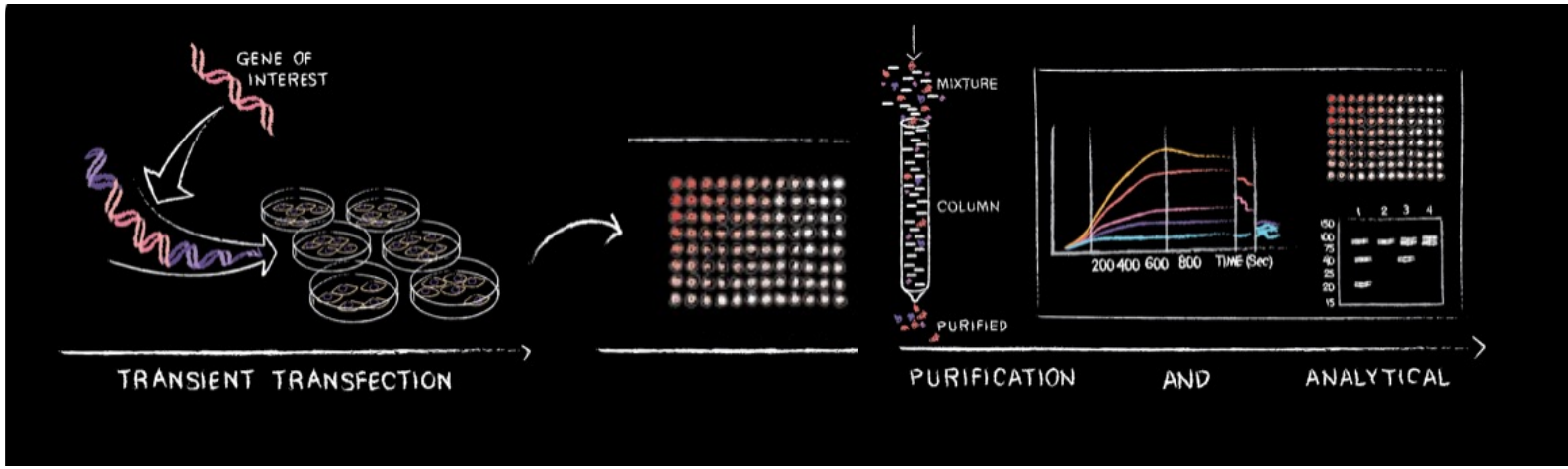




KEMP PROTEINS

Rapid Small Scale Expression & Purification

DNA Synthesis → Plasmid → Expression → Purification → QC → QA



ML Derived Sequence Assessment

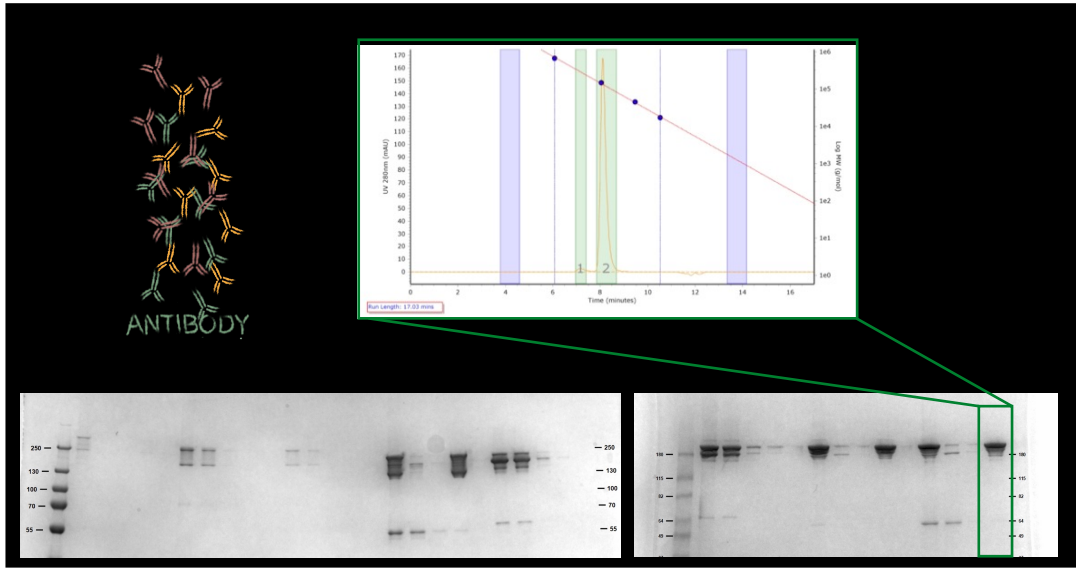
- 24–96 purifications from 0.5 mL - >1L
- Flexible expression & purification modalities
- Quick TAT ~15-20 days to get 100's protein variants and associated data*

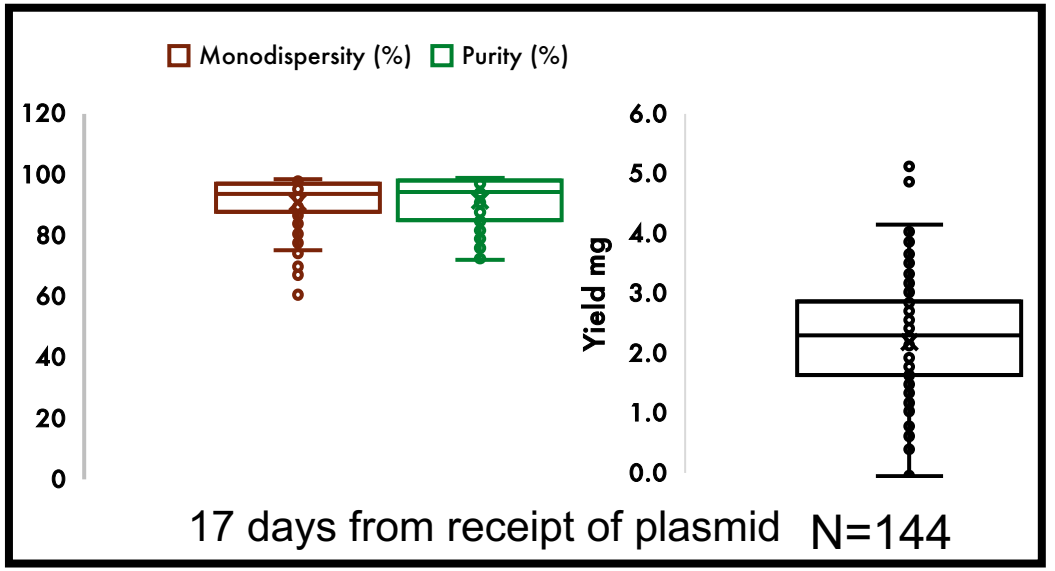
Final Deliverables:

- Go / No-Go on Constructs
- Customizable QC packages to feed into ML models
- ng to >100 mg of purified protein buffer exchanged

CASE STUDY: 144 Multispecifics

DNA Synthesis → Plasmid → Expression → Purification → QC → QA





ML Derived Sequence Assessment

- 144 expressions in CHO system
  - 4 combinations resulted in ultra low titer
- GORE ProA Membrane
  - Scalable Affinity Capture
- 17 day TAT DNA Provided

Developability and Lead Identification:

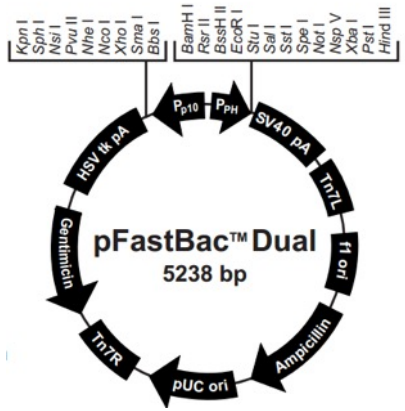
- Identified:
  - HC point mutations that destroyed affinity capture.
  - Light chain sequences that decreased titer
- Results used in the next round of seq

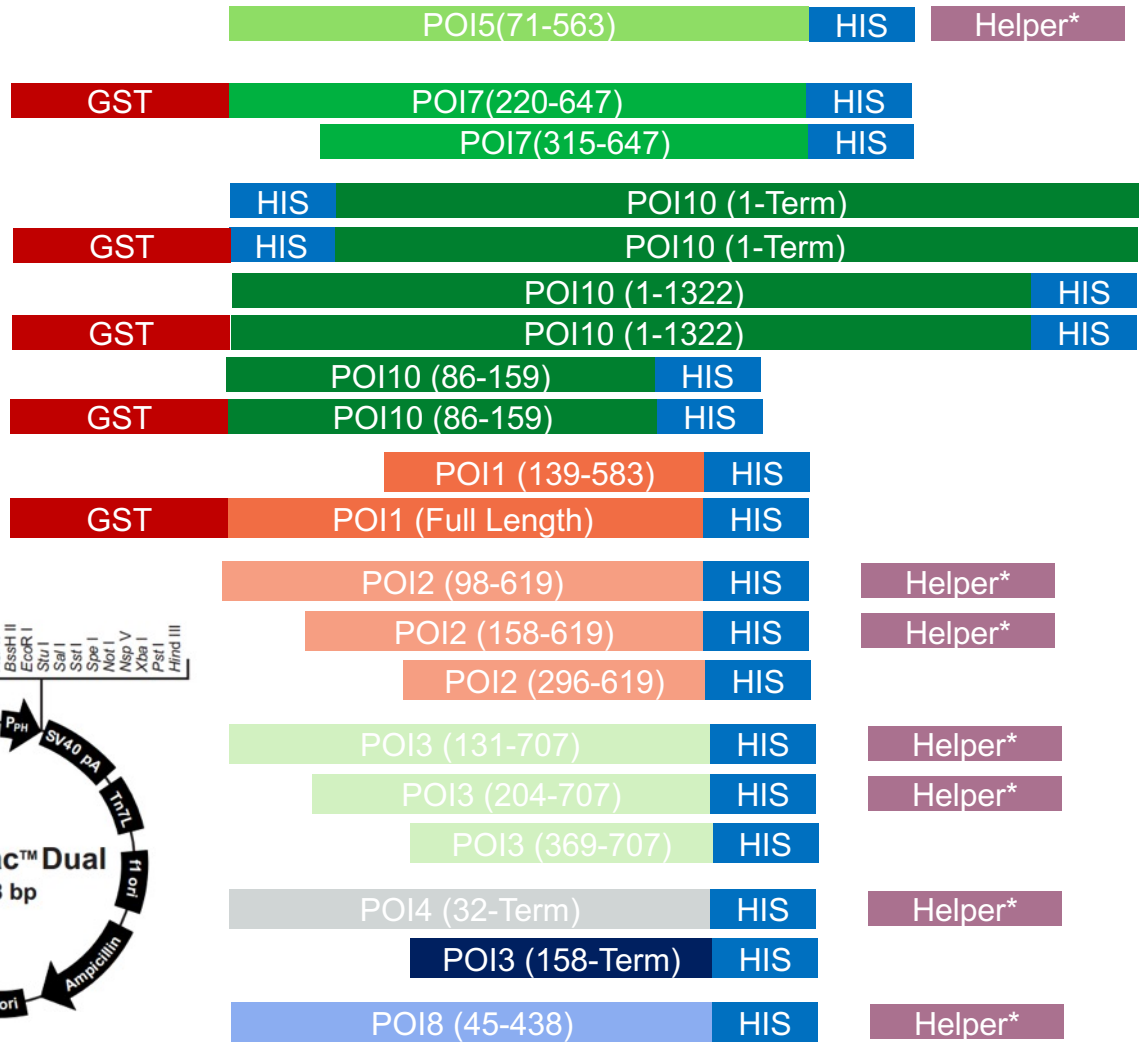
Case Study: Expression and Purification of Enzymes

Insert	Expected MW
POI5(71-563)-His6; Helper*	56.2 kDa + 19.1 kDa
GST-3c-POI7(220-647)-His6; Helper*	76.1 kDa + 19.1 kDa
POI7(315-647)-His6	38.2 kDa
His6-POI10	254.8 kDa
His6-GST-3c-POI10	281.6 kDa
POI10(1-1322)-His6	150.9 kDa
GST-3c-POI10(1-1322)-His6	177.6 kDa
POI10(86-159)-His6	9.1 kDa
GST-3c-bcLig10(86-159)-His6	35.8 kDa

Insert	Expected MW
POI1(139-583)-His6	49.6 kDa
POI2(98-619)-His6; Helper*	58.5 kDa + 19.1 kDa
POI2(158-619)-His6; Helper*	51.7 kDa + 19.1 kDa
POI8(45-438)-His6; Helper*	45.7 kDa + 19.1 kDa

Insert	Expected MW
GST-3c-POI1-HIS6; Helper*	98.6 kDa + 19.1 kDa
POI2(296-619)-His6	36.4 kDa
POI3(131-707)-His6; Helper*	65.9 kDa + 19.1 kDa
POI3(204-707)-His6; Helper*	57.6 kDa+ 19.1 kDa
POI3(369-707)-His6	38.5 kDa
POI4(His32-C-term)-His6; Helper*	65.2 kDa+ 19.1 kDa
POI4(158-C-term)-His6	31.6 kDa





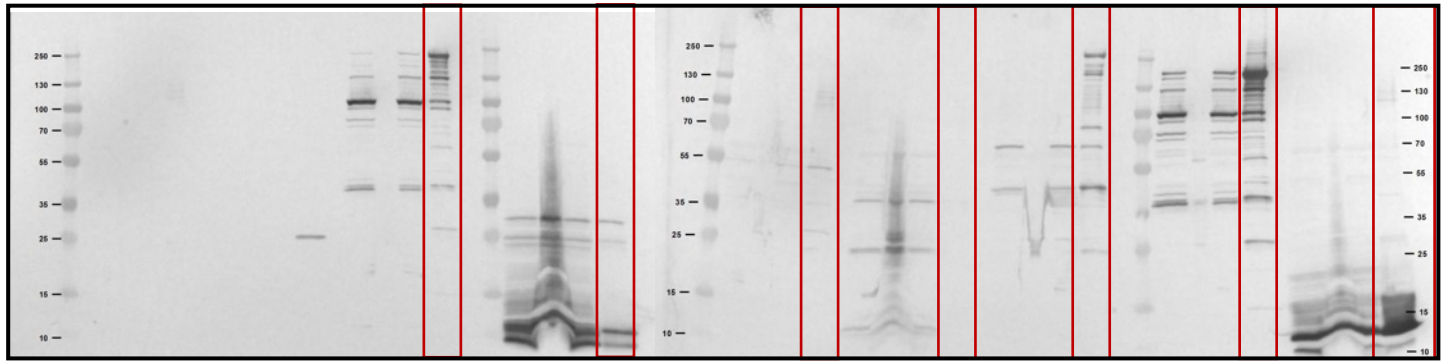
***Helper was included in bicistronic expression system pFastBac™Dual**

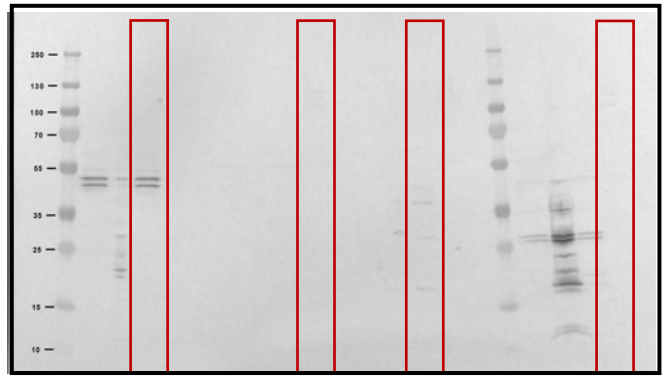
Enzymes/Fragments – Constructs Matter

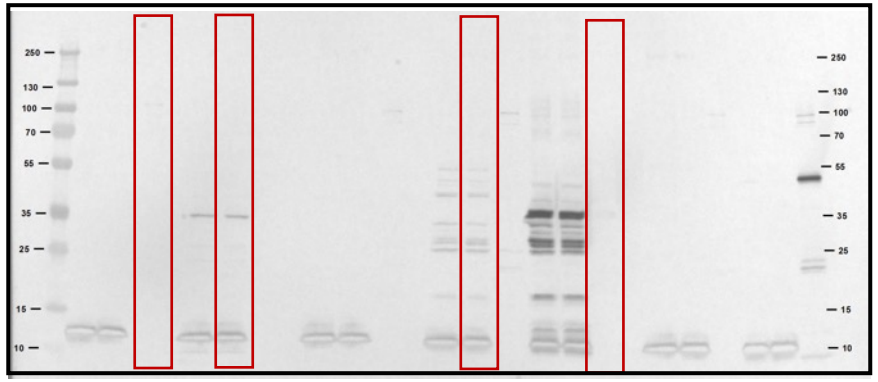
Insert	POI in Eluate	Expected MW
POI5(71-563)-His6; Helper*	Yes	56.2 kDa + 19.1 kDa
GST-3c-POI7(220-647)-His6; Helper*	Maybe	76.1 kDa + 19.1 kDa
POI7(315-647)-His6	Low	38.2 kDa
His6-POI10	Yes	254.8 kDa
His6-GST-3c-POI10	NO	281.6 kDa
POI10(1-1322)-His6	Yes	150.9 kDa
GST-3c-POI10(1-1322)-His6	Yes	177.6 kDa
POI10(86-159)-His6	Yes	9.1 kDa
GST-3c-POI10(86-159)-His6	Yes	35.8 kDa

Insert	POI in Eluate	Expected MW
POI1(139-583)-His6	In FT	49.6 kDa
POI2(98-619)-His6; Helper*	Low	58.5 kDa + 19.1 kDa
POI2(158-619)-His6; Helper*	Yes Degrade?	51.7 kDa + 19.1 kDa
POI8(45-438)-His6; Helper*	Low SOL	45.7 kDa + 19.1 kDa

Insert	POI in Eluate	Expected MW
GST-3c-POI1-HIS6; Helper*	Low Yield	98.6 kDa + 19.1 kDa
POI2(296-619)-His6	FT	36.4 kDa
POI3(131-707)-His6; Helper*	ND	65.9 kDa + 19.1 kDa
POI3(204-707)-His6; Helper*	FT	57.6 kDa+ 19.1 kDa
POI3(369-707)-His6	Inef Cap	38.5 kDa
POI4(His32-C-term)-His6; Helper*	NO	65.2 kDa+ 19.1 kDa
POI4(158-C-term)-His6	YES	31.6 kDa





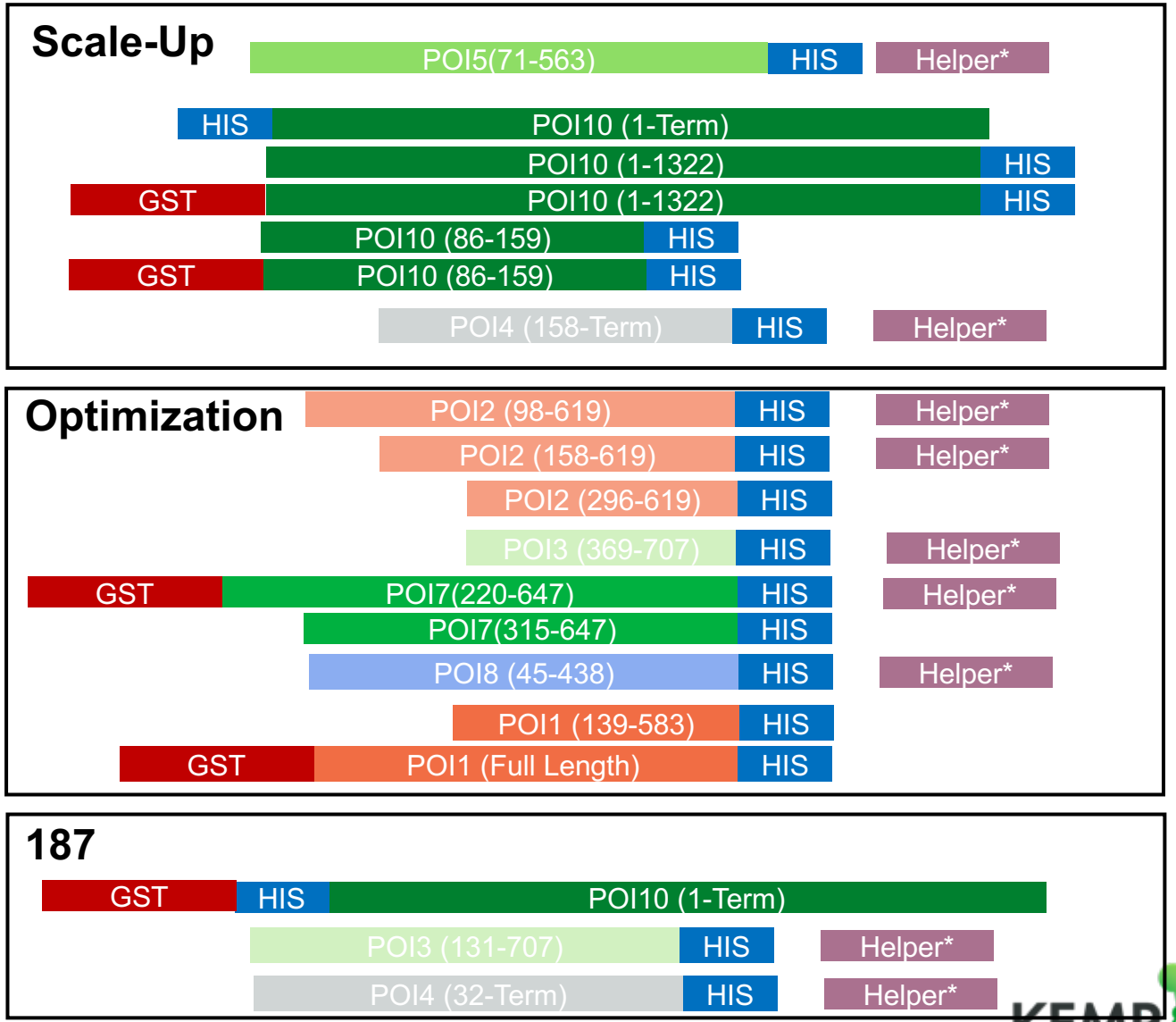


Selection, Optimization, Scale, and …

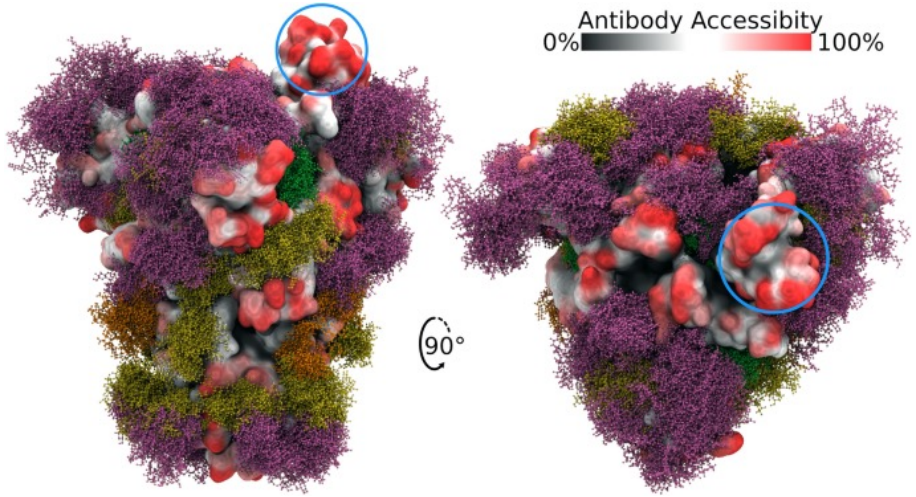
Insert	POI in Eluate	Expected MW
POI5(71-563)-His6; Helper*	Yes	56.2 kDa + 19.1 kDa
GST-3c-POI7(220-647)-His6; Helper*	Maybe	76.1 kDa + 19.1 kDa
POI7(315-647)-His6	Low	38.2 kDa
His6-POI10	Yes	254.8 kDa
His6-GST-3c-POI10	NO	281.6 kDa
POI10(1-1322)-His6	Yes	150.9 kDa
GST-3c-POI10(1-1322)-His6	Yes	177.6 kDa
POI10(86-159)-His6	Yes	9.1 kDa
GST-3c-POI10(86-159)-His6	Yes	35.8 kDa

Insert	POI in Eluate	Expected MW
POI1(139-583)-His6	In FT	49.6 kDa
POI2(98-619)-His6; Helper*	Low	58.5 kDa + 19.1 kDa
POI2(158-619)-His6; Helper*	Yes Degrade?	51.7 kDa + 19.1 kDa
POI8(45-438)-His6; Helper*	Low SOL	45.7 kDa + 19.1 kDa

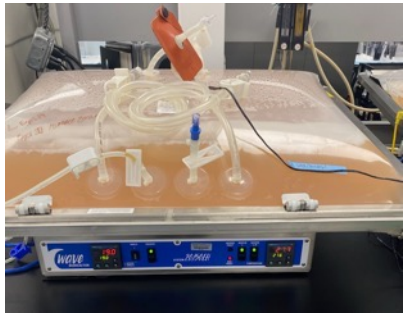
Insert	POI in Eluate	Expected MW
GST-3c-POI1-HIS6; Helper*	Low Yield	98.6 kDa + 19.1 kDa
POI2(296-619)-His6	FT	36.4 kDa
POI3(131-707)-His6; Helper*	ND	65.9 kDa + 19.1 kDa
POI3(204-707)-His6; Helper*	FT	57.6 kDa+ 19.1 kDa
POI3(369-707)-His6	Inef Cap	38.5 kDa
POI4(His32-C-term)-His6; Helper*	NO	65.2 kDa+ 19.1 kDa
POI4(158-C-term)-His6	YES	31.6 kDa

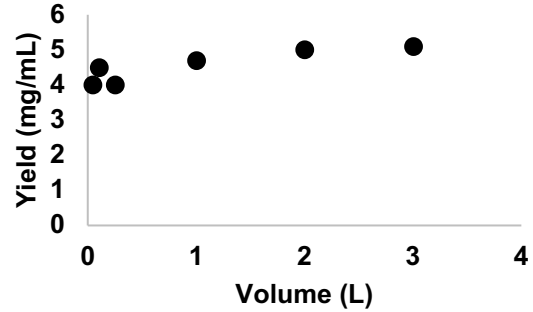


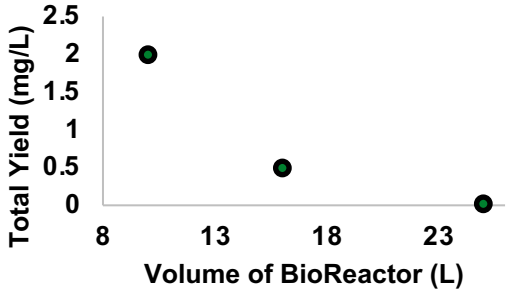
Diagnosing a Broken Process
Scale Up of the Upstream Process

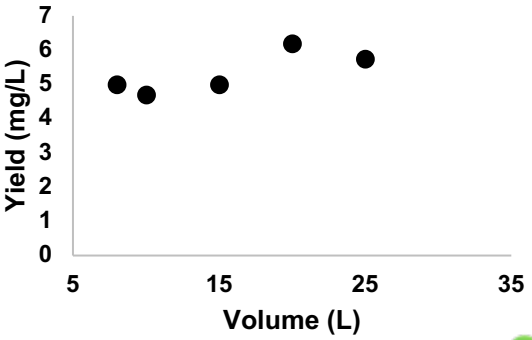


Signal Sequences:
HBM - alright
Gp64 – no
MRJ – fine
Gp67 - fine
Native – why is this decent?



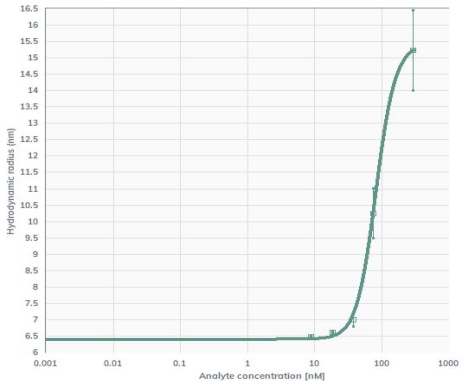




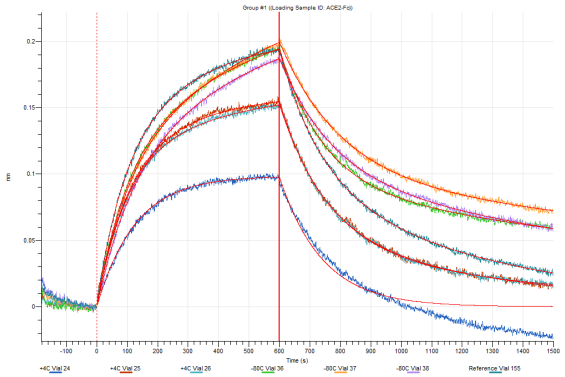


Grant et al. https://www.nature.com/articles/s41598-020-71748-7

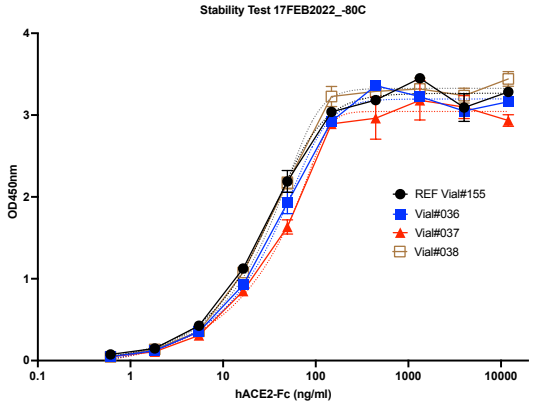
Process Analytics Guide the Development Cycle



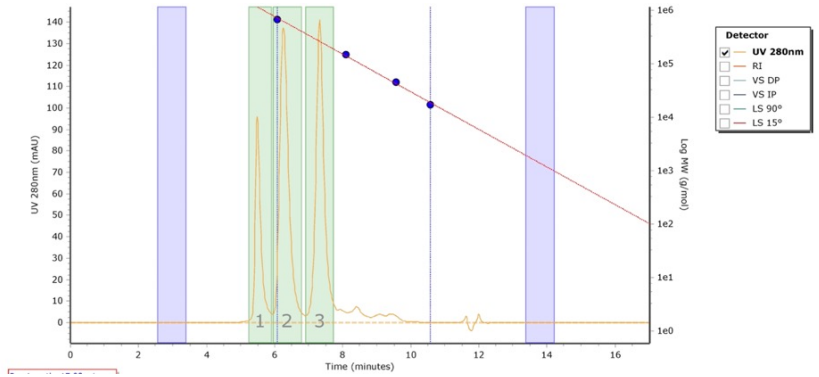
FIDA: Kd = 13.4 nM



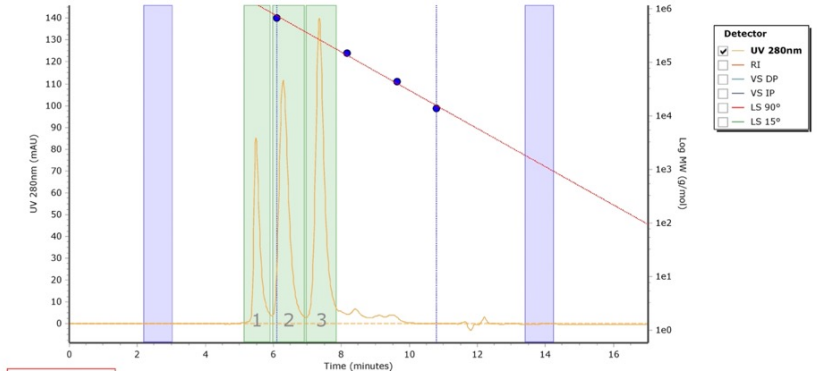
BLI: Kd = 22.125 nM



ELISA: Kd = 32.127 nM



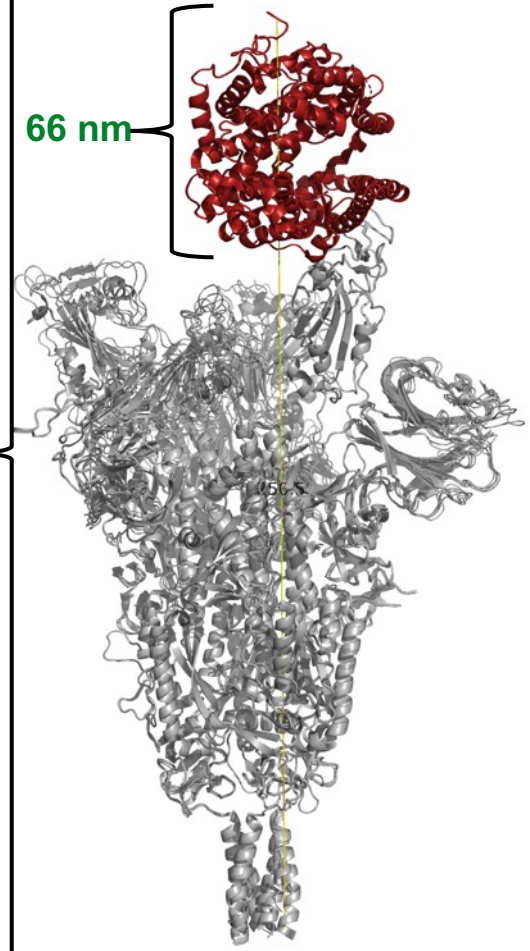
Fresh Freeze



3 Months -80 ºC

66 nm

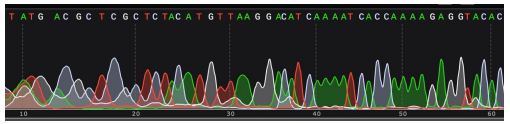
25.5 nm



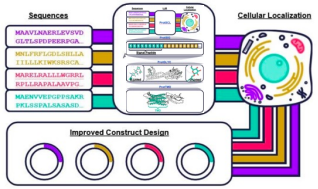
The spike maintained affinity out to 4 months at 4, 27, and -80 ºC

PDBID: 7VXM, 7VXK, &VXF - Wang et al. DOI: 10.1038/s41467-021-27350-0

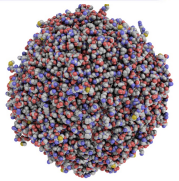
How to Increase Successes in Protein Workflows?



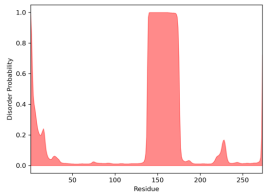
Sequence



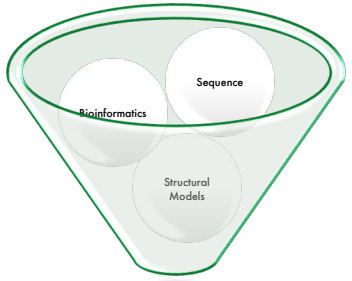
Bioinformatics



Structural Modeling



Disorder Prediction



Feasibility Score

Process Modeling

Understanding the goals of the program is critical

- Protein, Intent, & Scale help define the process

Process Modeling with ML and Bioinformatics

- Using the right tools increases the foundational understanding of the program
- Helps with risk assessment and prediction of liabilities

<u>Flexibility</u> with the expression and purification systems is key to increasing the likelihood of success

- You should not believe that a single expression and purification platform will work for the Global Proteome

Empirical data and reproducibility at a minimum of ~10% scale

- Identifies Critical Process Parameters that ENSURE Critical Quality Attributes are maintained
- Development of the manufacturing batch record

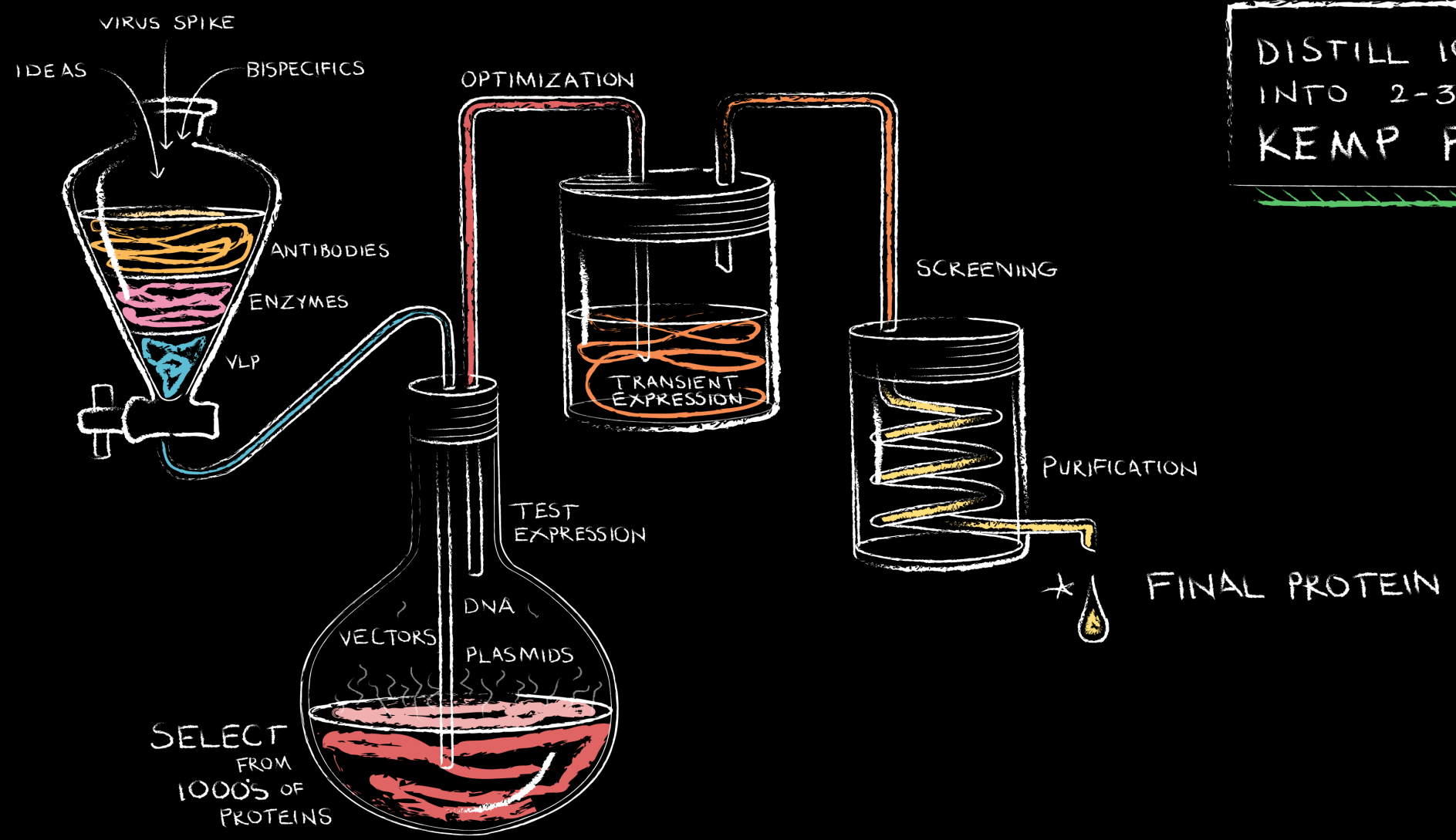
KEMP PROTEINS

Thank you for your time!

AEIC
ADVANCING BIOANALYTICAL TECHNOLOGIES

SENTAURI

BASF
We create chemistry



DISTILL 1000'S OF PROTEINS INTO 2-3 TARGETS USING KEMP PROTEINS HTP

USED IN:

PROCESS DISCOVERY

ASSAY DEVELOPMENT

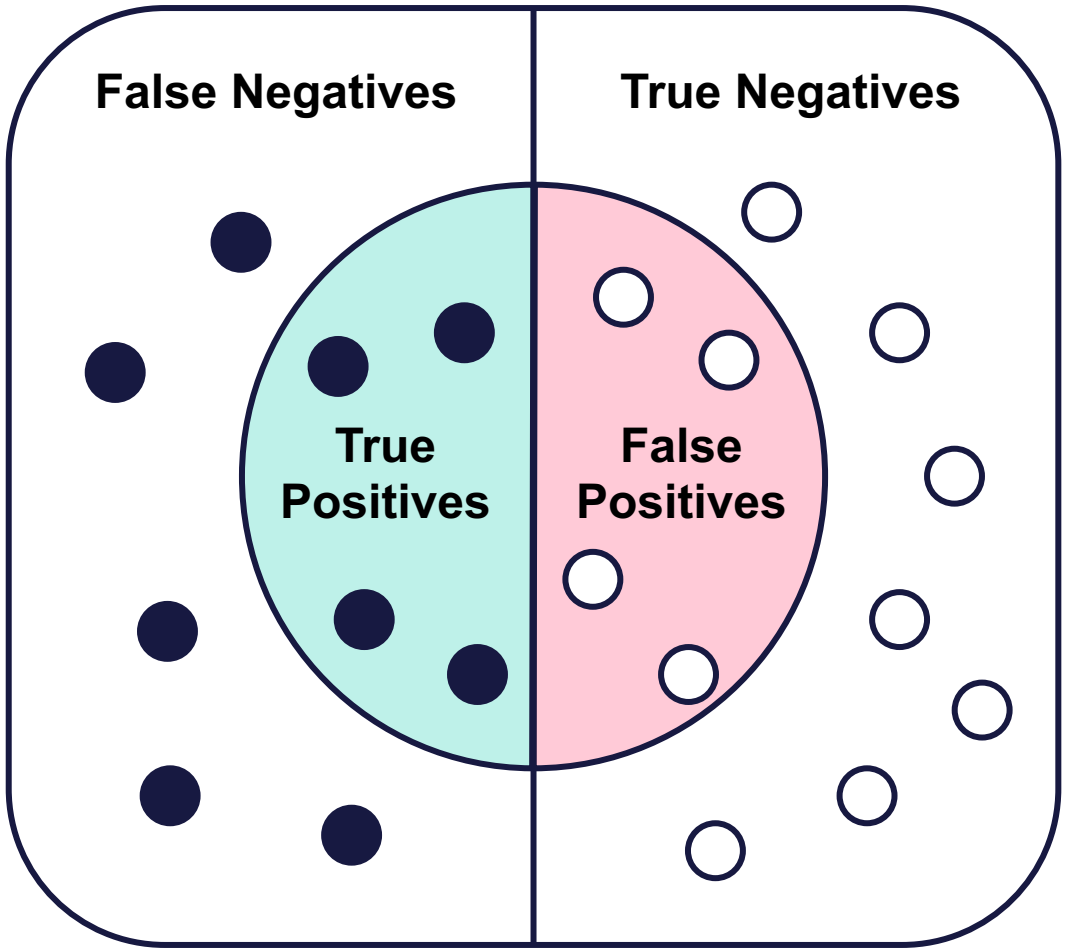
PHASE 1	PHASE 2	PHASE 3
CLINICAL DEVELOPMENT		

SUPPORT DOSSIER FROM KEMP PROTEINS

KEMP PROTEINS
QUALITY EXPRESSED

solutions@kempproteins.com

Refresher on Benchmarking Statistics



$$\textbf{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\textbf{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\textbf{F1 Score} = \frac{2\,(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

$$\textbf{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total}}$$

SENTAURI

KEMP PROTEINS