Using Machine Learning to Predict Recombinant Protein Expression

Abel Sousa[1,5], Stephanie K. Ashenden[1,5], Avid Afzal[1,5], Sergio Martinez Cuesta[1,5], Miguel Gancedo Rodrigo[2], William Lee[3], Sandeep K. Talapatra[2,4], Dilrini De Silva[1], Rick Davies[2], Yinhai Wang[1], Ian Barrett[1], Aurelie Bornot[1,6], Lovisa Holmberg Schiavone[3,6]

[1]Data Sciences and Quantitative Biology, Discovery Sciences, AstraZeneca, Cambridge, UK, [2]Discovery Biology Protein Sciences, Discovery Sciences, Cambridge, UK, [3]Discovery Biology Protein Sciences, Discovery Sciences, Gothenburg, Sweden, [4]Present address: Protein Science Group, Department of Protein and Cellular Sciences, GSK, Stevenage, UK, [5]First co-authors, [6]Corresponding authors: lovisa.holmberg.schiavone@astrazeneca.com

Abstract

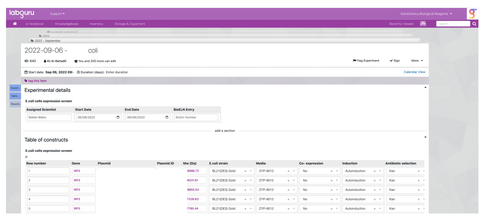
The production of recombinant proteins is critical across several drug discovery stages. The process is costly and lengthy with a minimum of 6 weeks per construct to expression screen campaign, involving multiple steps. We are developing a machine learning platform that uses the primary sequence of proteins represented as physicochemical properties and structural features to support protein scientists by facilitating the design of protein constructs and highlighting sequences expressing at different yield classes. The model is coupled to an *in-silico* screening procedure that systematically designs and assesses thousands of constructs in a high-throughput manner. This method is currently being deployed in drug discovery projects and leads to the design of constructs expressing at a higher yield compared to those designed using human knowledge only. Here we share our initial results and plans to improve and develop the techniques through integrative teamwork and additional resources. We plan to do this by: (1) considering yield values instead of classes aided by GelClick, an automated gel image analysis tool, (2) incorporating deep learning features for sequence representation, and (3) leveraging external public domain datasets. Limited data to train the model is a key blocker so we are putting together a proof-of-concept and a pre-competitive consortium with academic and pharmaceutical industry partners to share data and models in collaboration with EMBL-EBI.

1 – The evolution of our predictive model

2017-2019

Exploring data for model building and implementation of **expression screen module** in AZ's lab management software.

Build of an initial model in collaboration with an external partner. The model was put into production but gained little traction at the time.



2020

Improvements to the model based on lab data with **155** protein targets and more than **1100** constructs.

Discovered that small sequence changes in constructs can result in large changes in production yield.

2021

Implementation of **GelClick** – an automated gel image analysis tool.

Model piloted on **6 projects** predicting and validating designed constructs.

Initial exploration of deep learning embeddings by a **postdoc** (6 months).

2022

Annotation of more data with the use of **GelClick.**

MSc student analyses **public data** to be incorporated in the model and expands deep learning embeddings work.

Initiating recruitment of a **postdoc with EMBL-EBI** and consolidation of a precompetitive **consortium**.

2 – Our current model

Input sequences → Calculate features → Predictive model → Output yield category

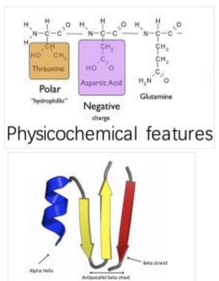
Currently:

AZ's lab data software

Future:

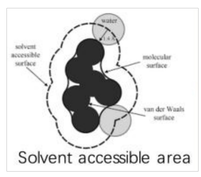
PSI Protein Structure Initiative

SGC



Physicochemical features

Secondary structures



Solvent accessible area

Disorder and flexibility[1]

Prediction of yield category in an expression host using a Random Forest model trained on lab data

- 1139 constructs, 155 proteins
- Assessment using 5-fold cross validation

Yield category	Precision (proportion of correct prediction)	Recall (proportion of true cases identified)
0 (0-1 mg/L)	0.48	0.63
1 (1-10 mg/L)	0.42	0.45
2 (10-20 mg/L)	0.40	0.04
3 (>20 mg/L)	0.37	0.19

Metric	Global performance	Random performance
Accuracy	0.42	0.28

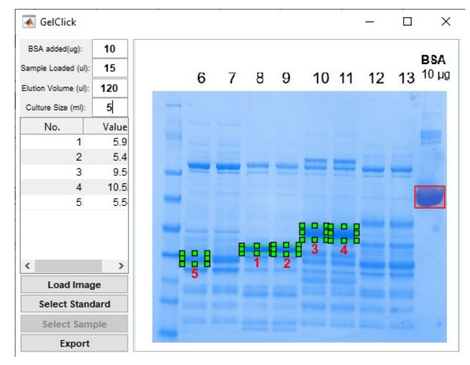
Accuracy better than random, OK at predicting lowly expressing constructs, there is room for improvement

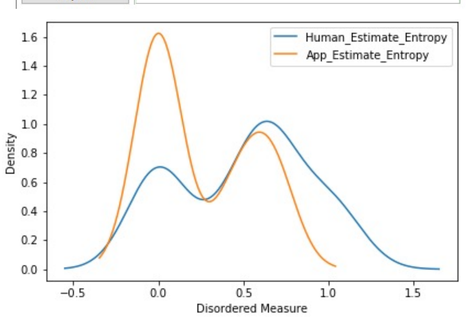
3 – GelClick

A more precise yield estimation can help the predictive model to better learn the relationship between input sequence and yield. GelClick is a tool that accurately estimates the yield of recombinant proteins after expression screens and aims to select the best construct for production.

GelClick interface: The user selects the reference band (red) and the construct bands (green), enters parameters of the experiment and GelClick analyses the image components to estimate the yield of the constructs

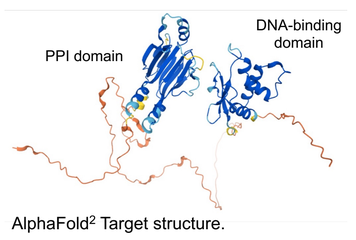
Quantification of the estimation variability for the two methods: The disorder or scattering of estimations is lower when using the application. That can be seen in the orange curve that has a higher density around zero. A lower value of disordered is better since it means higher consistency.





4 – Project case study and summary of results

Target background → Construct design → Predictions → Ranking → Validation

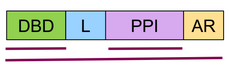


AlphaFold[2] Target structure.

Our target is a transcription factor involved in the development of immune cells, including B/T cells and dendritic cells.

Its deregulation is involved in the development of different types of cancers.

Protein constructs are designed *in silico* by systematically combining multiple protein regions; protein tags; and removing residues from the C/N-terminal domains.



DBD = DNA-binding domain, L = linker, PPI = protein-protein interaction domain, AR = autoregulatory domain

Target regions considered: DBD, IAD and Full-length

Protein tag	Terminus
6His-GS-TEV-Target	N
6His-GS-ZZ-GS-TEV-Target	N
6His-SUMO-Target	N
GST-TEV-Target	N
M	N
GS-AVI	C
None	C

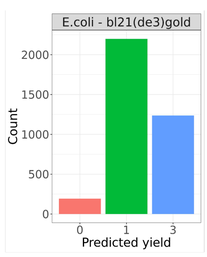
Up to 10 residues removed from the N/C-terminus

>3,500 constructs designed.

Bioinformatics features of the *in silico* constructs are extracted using an in-house pipeline.

The current ML model can provide expression predictions for 8 hosts.

Organism	Host
E. coli	bl21(de3)gold
E. coli	bl21(de3)codonplus
E. coli	bl21(de3)rosettaplyss
E. coli	bl21(de3)star
E. coli	arcticexpress
Insect	sf21
Insect	sf9
Mammal	expi293



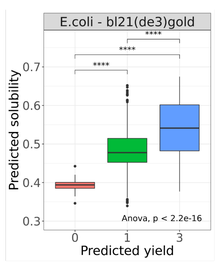
Yield class / Yield (mg/L): 0 / 0-1; 1 / 1-10; 2 / 10-20; 3 / >20

The constructs are ranked using a desirability function that transforms and integrates multiple ranking features into a 0-1 scale, such that high values are desirable[3].

Ranking features:

- Prediction confidence
- Residues removed
- Solubility

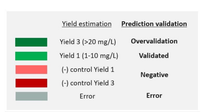
Construct	Score
A	1
B	0.98
C	0.95
…	…

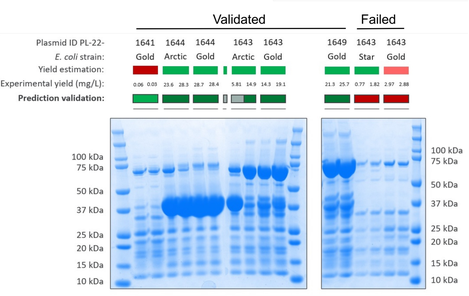


Solubility is predicted using NetSolP[4] model.

The constructs are validated using expression screenings to confirm the predictions.

Target constructs (21)	Number / Percentage
Validated	11 / 52%
Failed	5 / 24%
Inconclusive	5 / 24%

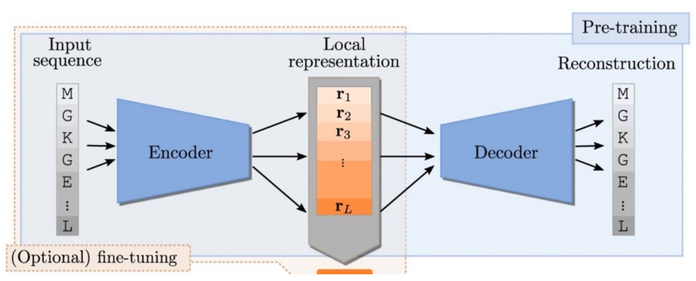




5 – Future work

Model improvements

- Generate more yield data, using GelClick, from future projects and include within the model.
- Upgrade model using deep-learning representations of protein sequences[7].



Postdoc

- Improve end-to-end modelling approach in general e.g. augment internal data with publicly available resources.
- Joint collaboration with the Leach group at EMBL-EBI.

EMBL-EBI

Consortium

- Championed by Lovisa Holmberg Schiavone and Aurelie Bornot since 2021 in collaboration with EMBL-EBI.
- Bring pharma and academic partners together to (a) share pre-competitive protein production data and (b) collaborate on the development of predictive models.
- Some discussions have been had within the EMBL-EBI industry forum
- Currently at proposal stage for proof-of-concept study.

References

1. Liu *et al.*, Molecules, 23(10), 2535 (2018)
2. Jumper *et al.*, Nature, 596, 583–589 (2021)
3. Lazic S. E., PeerJ, 3:e1444 (2015)
4. Thumuluri *et al.*, Bioinformatics, 38(4), 941–946 (2022)
5. Detlefsen *et al.*, Nature Communications, 13, 1914 (2022)

Acknowledgements

We thank the members of the Data Sciences and Quantitative Biology (QuBi) and Discovery Biology departments in AstraZeneca for helpful inputs throughout the duration of the project and while creating this poster.

AstraZeneca
What science can do