Large Scale active-learning-guided exploration for in vitro protein production optimization

Olivier Borkowski, Mathilde Koch, Agnès Zettor, Amir Pandi, Angelo Cardoso Batista, Paul Soudier & Jean-Loup Faulon

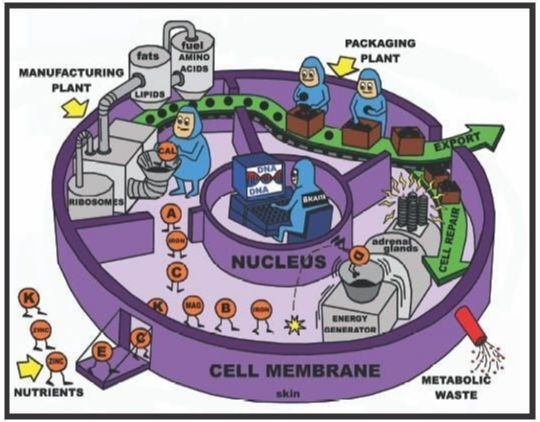
Present: Zeyuan Zuo, Tianqin Li

Outline

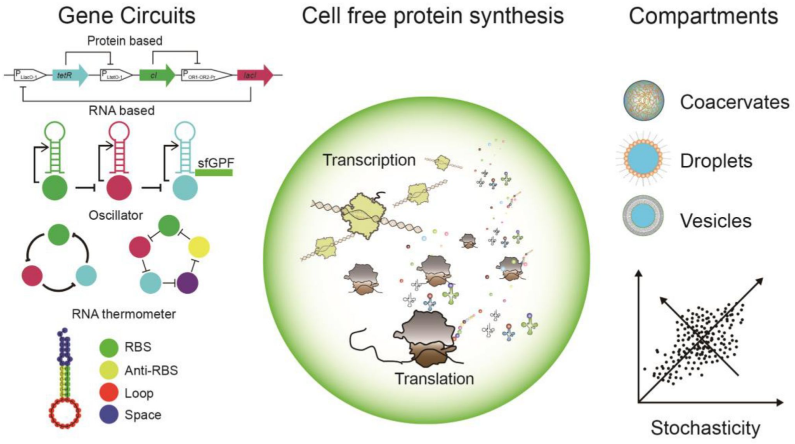
- Scientific context
- Active learning strategies
- Experiments and results

Cell free protein synthesis

- **Cell-free protein synthesis**, also known as ***in vitro* protein synthesis** or **CFPS**, is the production of protein using biological machinery in a cell-free system, that is, without the use of living cells.



Cell based system



Cell free system

Advantages of cell free system

- Fast gene expression kinetics
- Low reaction volumes
- High-throughput measurements
- Simplified gene characterization via decoupling protein production from host physiology
- Disseminate among laboratories, portable and standard

Challenges for cell free system

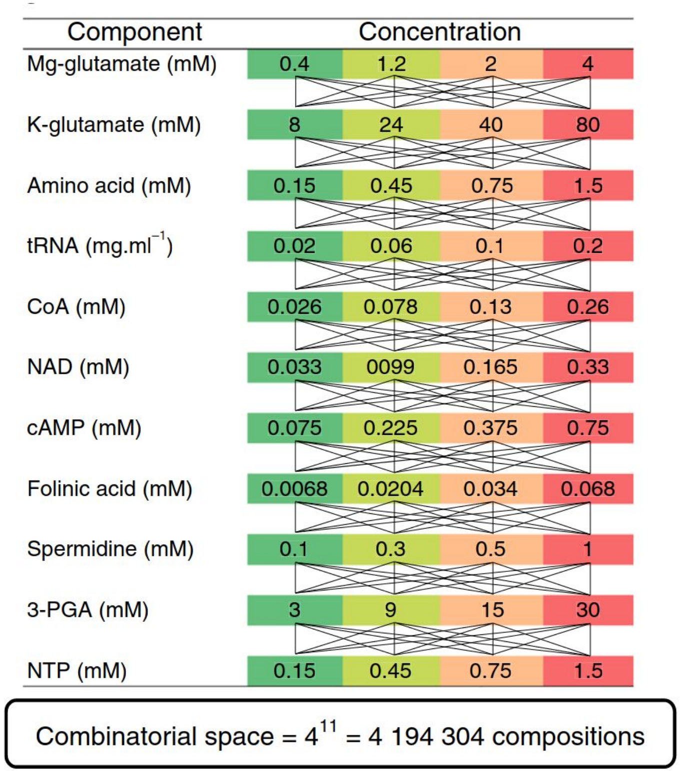
- Ribosomes, native polymerases and cofactors concentrations are hard to control.
- Different concentrations of the cell free system could results in dramatic difference in production efficiency.
- E.g. Caschera et al optimized the concentration so that the protein production efficiency can be increased by 10-fold.
- However, such space is huge and impossible to try all of them.

Active Learning to search the concentration space

- **Design parameters:**

  11 components, each has 4 possible concentration, totally **4,194,304** compositions.

- Therefore, optimally decide which trail to try is critical based on the data already known at each step.



Component	Concentration			
Mg-glutamate (mM)	0.4	1.2	2	4
K-glutamate (mM)	8	24	40	80
Amino acid (mM)	0.15	0.45	0.75	1.5
tRNA (mg.ml^{-1})	0.02	0.06	0.1	0.2
CoA (mM)	0.026	0.078	0.13	0.26
NAD (mM)	0.033	0099	0.165	0.33
cAMP (mM)	0.075	0.225	0.375	0.75
Folinic acid (mM)	0.0068	0.0204	0.034	0.068
Spermidine (mM)	0.1	0.3	0.5	1
3-PGA (mM)	3	9	15	30
NTP (mM)	0.15	0.45	0.75	1.5

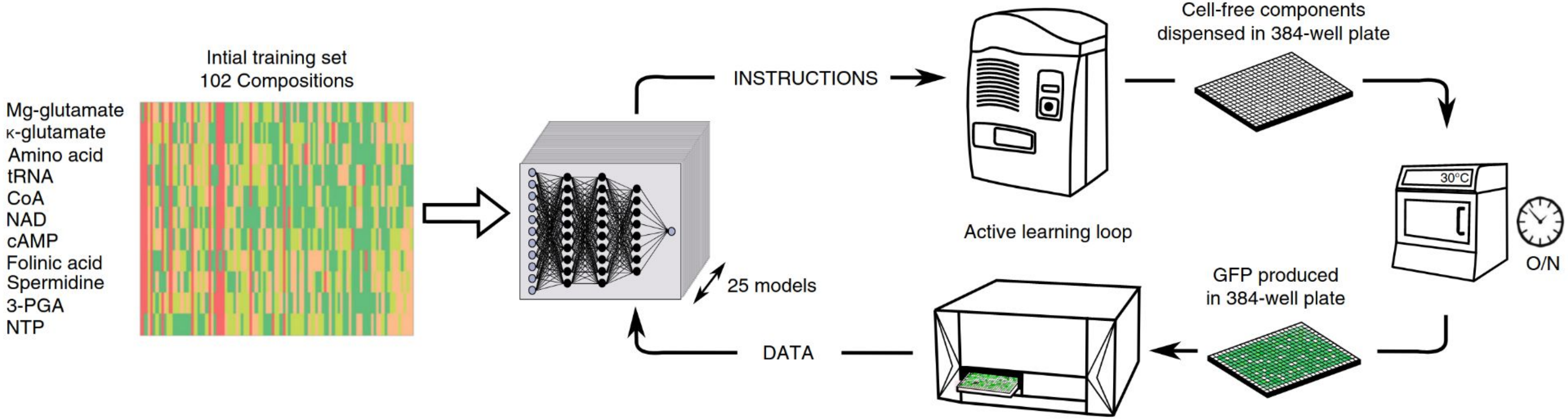
Combinatorial space = 4^{11} = 4 194 304 compositions

Active Learning algorithm

- Objective: Finding the **best combination** out of totally **4,194,304** combination space such that it has the highest yield while using smallest number of experiments possible.

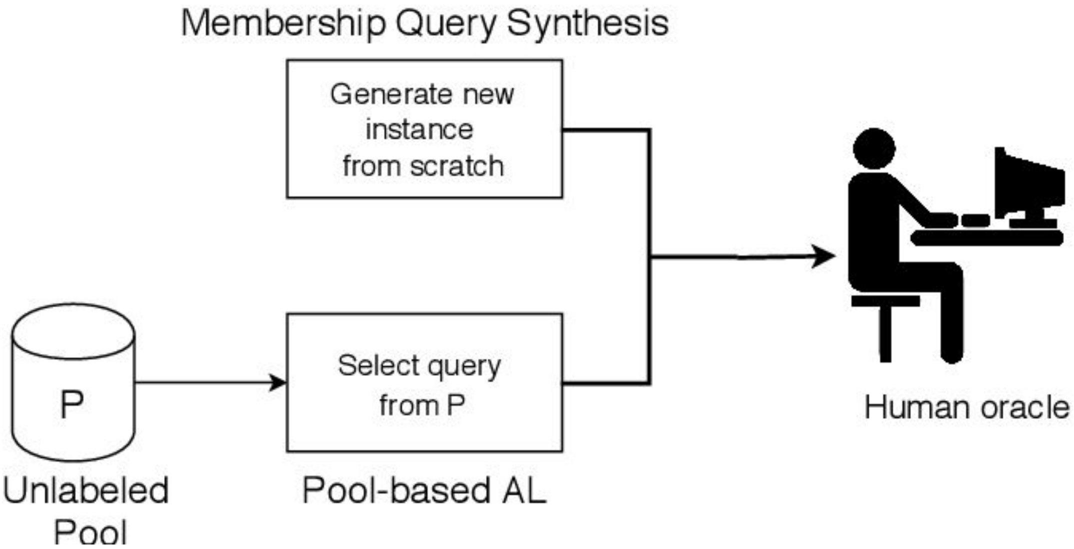
Active Learning algorithm

- Base learner: MLP neural network
- Each MLP predict the yield volume based on the specific combination
- MLP: $f_\theta : R^{11} \rightarrow R^1$



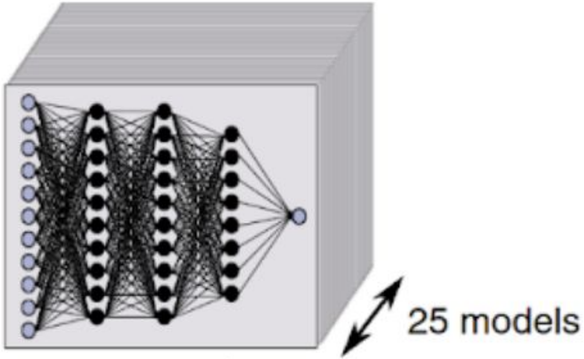
Active Learning algorithm

- Data access model: Membership query synthesis
- Generate combination from scratch (from **4,194,304** combinations)


Membership Query Synthesis
Generate new instance from scratch
P
Unlabeled Pool
Select query from P
Pool-based AL
Human oracle

Active Learning algorithm

- Query selection strategy:
    - Sequential Model-based Optimization
- Learn an ensemble of NN, each model will learn to yield accurate prediction given a input combination.
- The standard deviation of the ensemble represents the uncertainty of the data



Active Learning algorithm

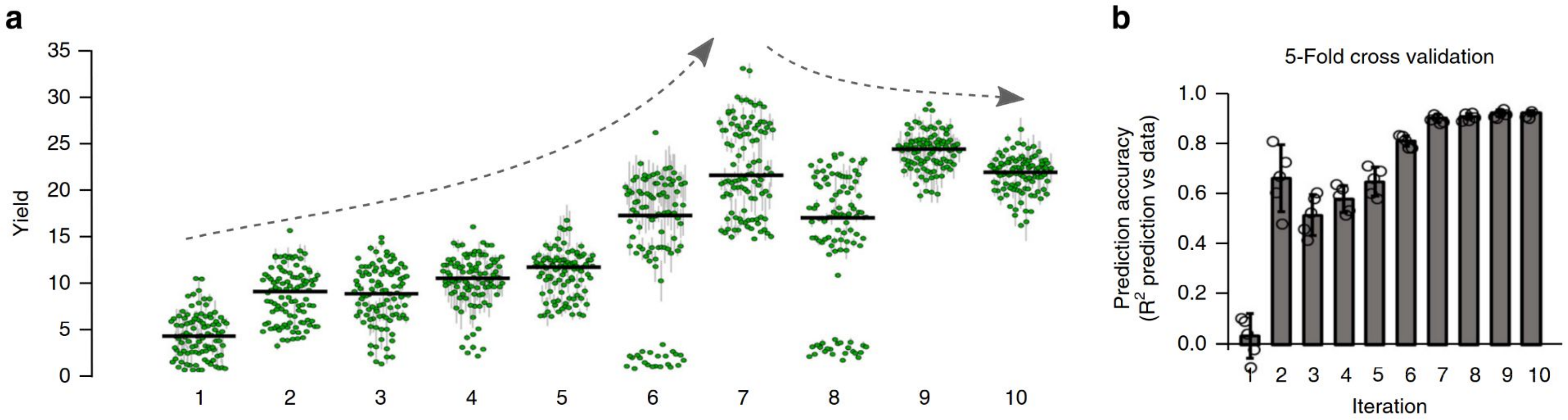
- How to balance between exploitation and exploration?
    - Upper Confidence Bound:
        - Selecting points with the highest UCB score
        - UCB score = exploitation*yield_pred_mean+exploration*yield_pred_std
        - exploitation = 1, exploration = √2
    - Intuition:
        - Want the model to select the combination whose yield can be very high based on what we already know but also want to consider points that the model is uncertain about.

Active Learning algorithm

- For one batch in 10 batches:
    - a. For N times (N=100,000):
        - i. Randomly sample a composition in the composition space
        - ii. If a composition was drawn previously (either in a previous experiment or during current selection), neglect it.
        - iii. Predict mean and standard deviation for all 100,000 points using the ensemble of 25 models previously trained.
    - b. Select the best set of compositions, according to the following Upper Confidence Bound (UCB) formula
    - c. Do experiments on these selected points
    - d. repeat

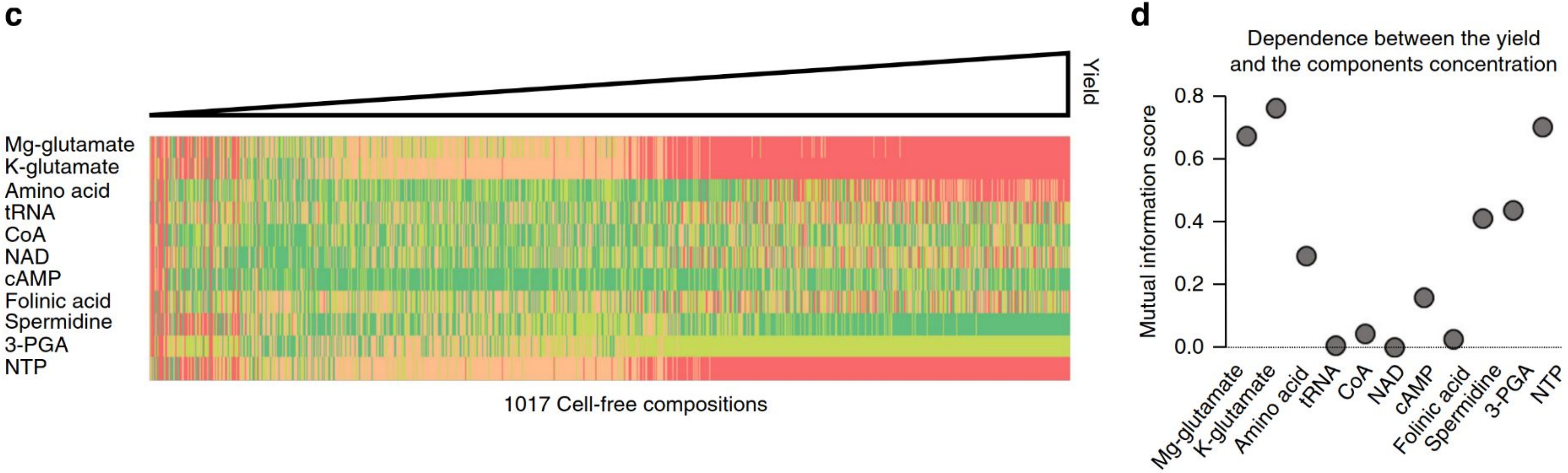
Results

- The data queried become better and better
- The model’s prediction accuracy also grows



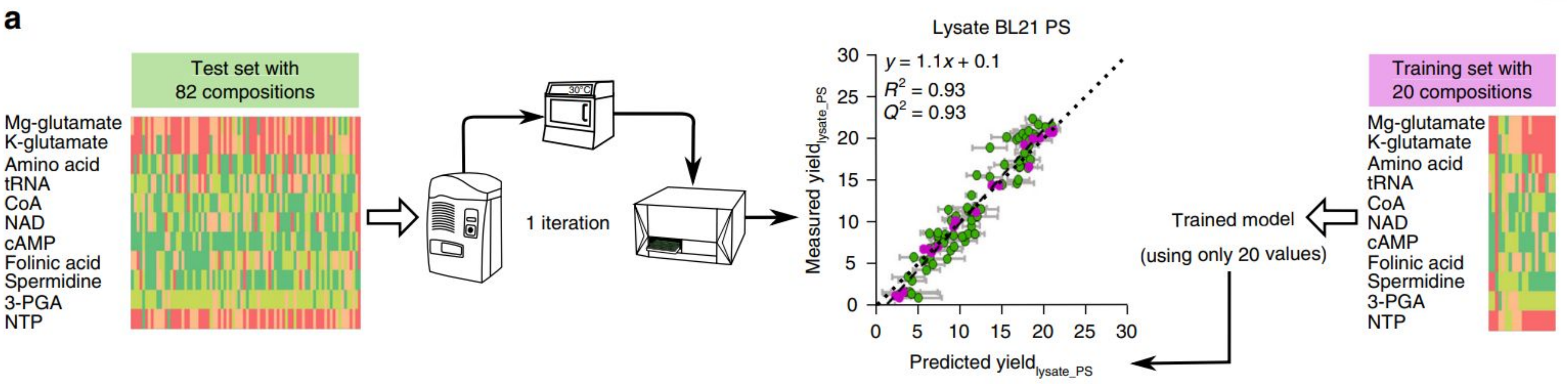
Results

- Calculate Mutual information between each variable and the yield
- Find Mg-glutamate, K-glutamate and NTP has high correlation with the yield value, i.e. more important and sensitive.



A one-step method for lysate-specific optimization

- Batchsize: 102
- Train: 20 informative compositions
- Test: 82 remaining compositions

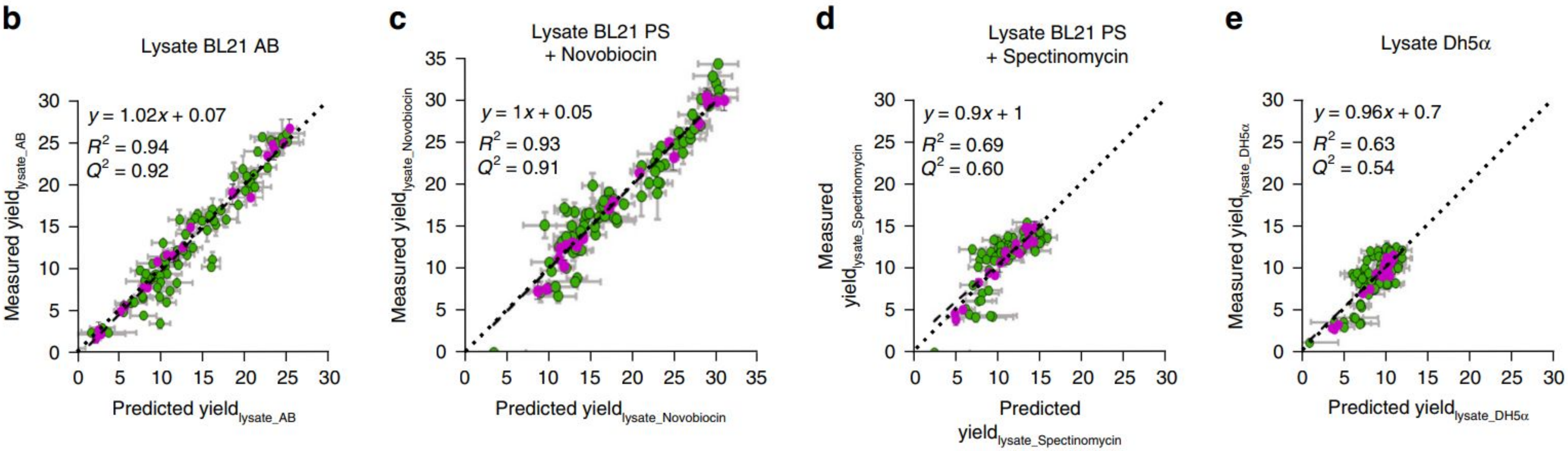


A one-step method for lysate-specific optimization

- Informative compositions
- For 1000 iteration:
- Randomly sample 20 combinations from the dataset(102).
- Train models on those points using the strategy before.
- Predict on the other points(82).
- Obtain the average score on all lysate.
- Keep those combinations if this average is better.

A one-step method for lysate-specific optimization

- Regression model with different lysates and antibiotics



Conclusion

- Active Learning to search in huge parameter space
- It’s an application of sequential model based optimization
- Balance between model exploitation and exploration
- Good example of performance membership query synthesis

- Code available at: ***https://github.com/brsynth/active_learning_cell_free***

Large Scale active-learning-guided exploration for in vitro protein production optimization

Olivier Borkowski, Mathilde Koch, Agnès Zettor, Amir Pandi, Angelo Cardoso Batista, Paul Soudier & Jean-Loup Faulon

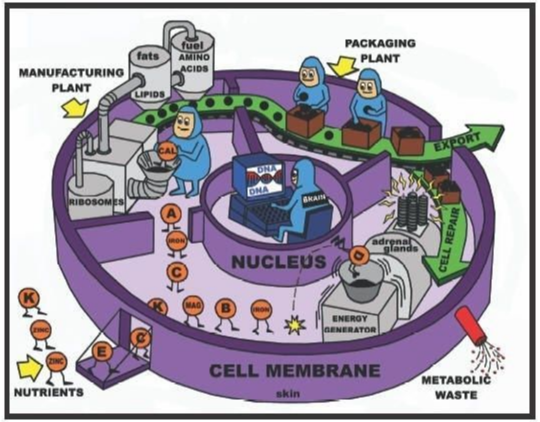
Present: Zeyuan Zuo, Tianqin Li

Outline

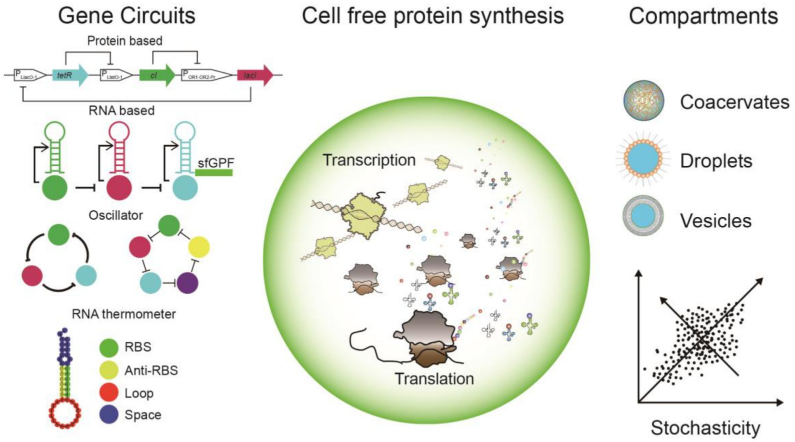
- Background
- Maximize mean yield
- Maximize yield regression accuracy
- Conclusion

Background: Cell free protein synthesis

- **Cell-free protein synthesis**, also known as ***in vitro*** **protein synthesis** or **CFPS**, is the production of protein using biological machinery in a cell-free system, that is, without the use of living cells.



Cell based system



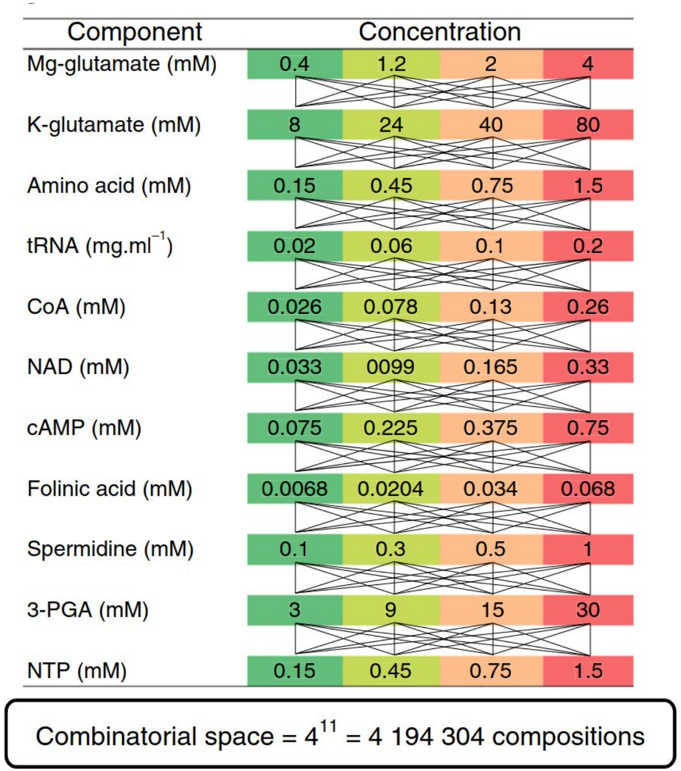
Cell free system

Active Learning to search the concentration space

- **Design parameters:**

  11 components, each has 4 possible concentration, totally **4,194,304** compositions.

- Therefore, optimally decide which trail to try is critical based on the data already known at each step.

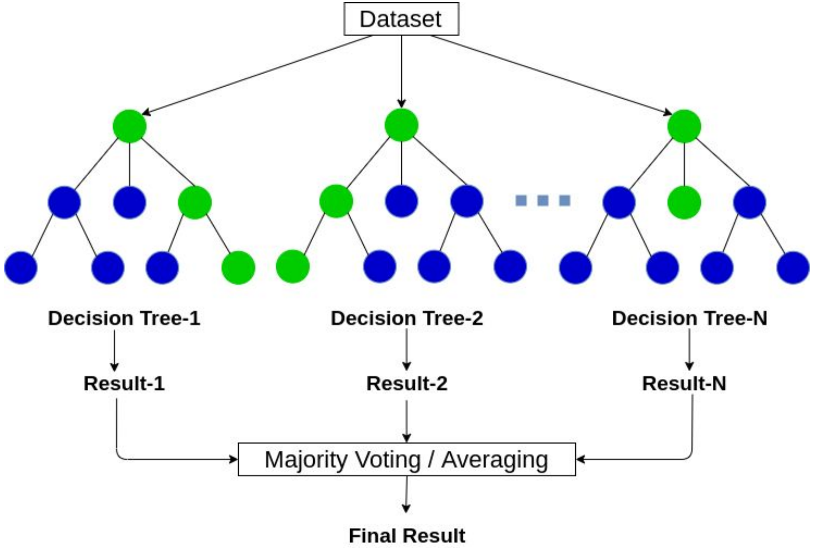


Component	Concentration			
Mg-glutamate (mM)	0.4	1.2	2	4
K-glutamate (mM)	8	24	40	80
Amino acid (mM)	0.15	0.45	0.75	1.5
tRNA (mg.ml^{-1})	0.02	0.06	0.1	0.2
CoA (mM)	0.026	0.078	0.13	0.26
NAD (mM)	0.033	0099	0.165	0.33
cAMP (mM)	0.075	0.225	0.375	0.75
Folinic acid (mM)	0.0068	0.0204	0.034	0.068
Spermidine (mM)	0.1	0.3	0.5	1
3-PGA (mM)	3	9	15	30
NTP (mM)	0.15	0.45	0.75	1.5

Combinatorial space = 4^{11} = 4 194 304 compositions

Maximize mean yield

Base learner	Random forest regressor
Query strategy	max_UCB
Training samples	510
Batch size	102
Number of batches	5

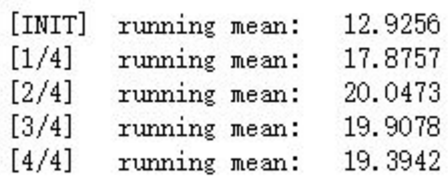


Maximize mean yield

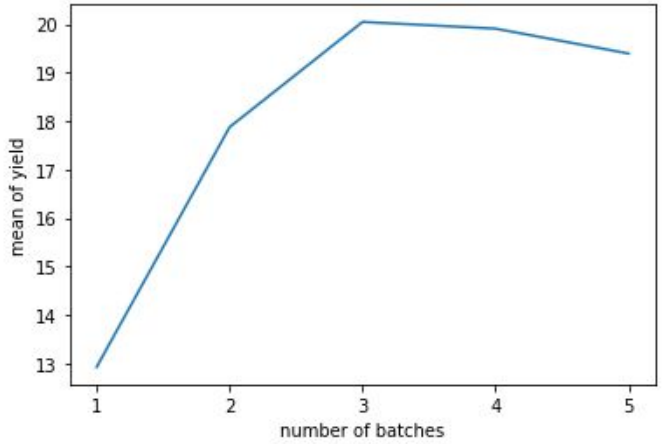
- Randomly selected 102 samples in the pool to train the regression model
- Delete the samples from pool
- For t = 1,2,...,T
    - Use the model to compute UCB for rest of the samples
    - Select the 102 samples with maximum UCB
    - Use the samples to update the regression model
    - Delete the samples from pool
    - Compute mean yield of all selected samples

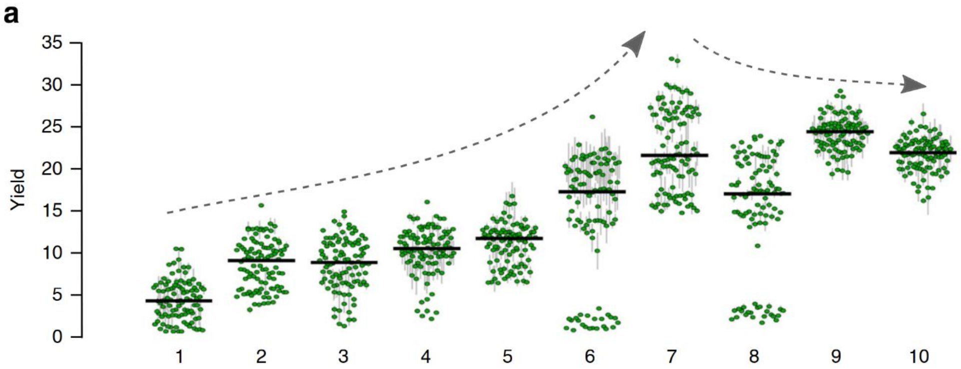
UCB = exploitation * mean_yield + exploration * std_yield
Exploitation = 1
Exploration = sqrt(2)

Maximize mean yield

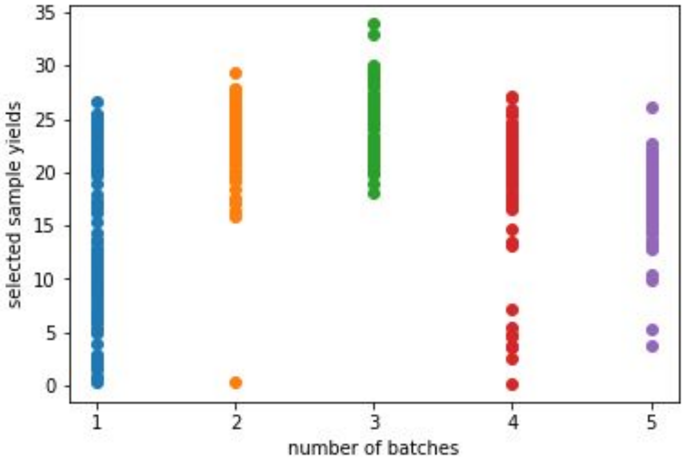


```
[INIT]  running mean:   12.9256
[1/4]   running mean:   17.8757
[2/4]   running mean:   20.0473
[3/4]   running mean:   19.9078
[4/4]   running mean:   19.3942
```





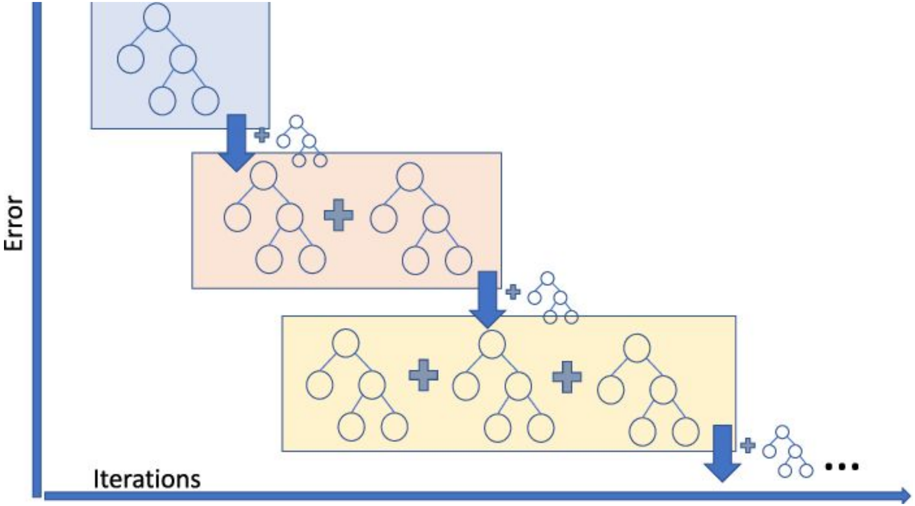
Paper’s results with 1017 budget



Our results with 510 budget

Maximize yield regression accuracy

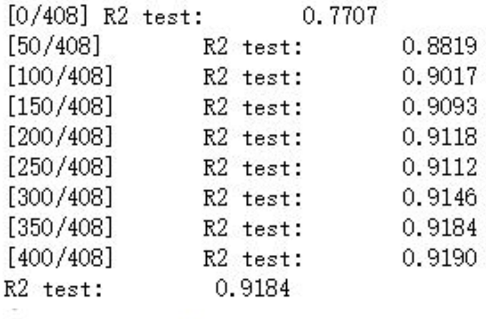
Base learner	Gradient boosting regressor
Committee	5 base learners
Query strategy	max_std_sampling
Training samples	510



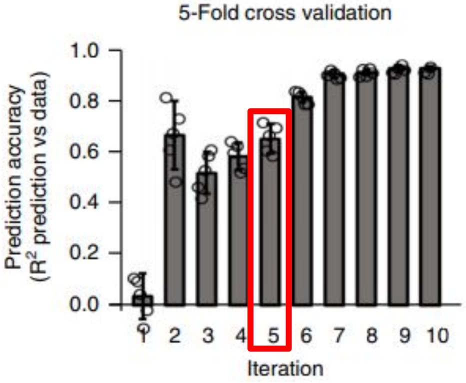
Maximize yield regression accuracy

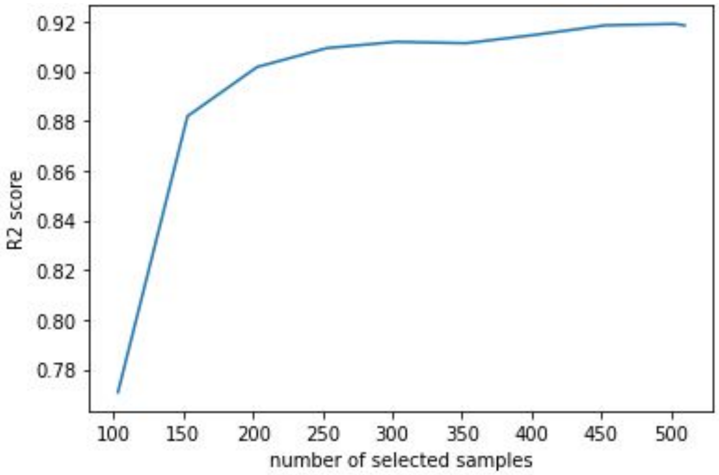
- Randomly selected 102 samples in the pool to train the regression
    - Each regressor in the committee equally assigned with training samples
- Delete the samples from pool
- For t = 1,2,...,T
    - Query the sample with max_std
    - Update the committee regressors with the new sample
    - Delete the sample from pool
    - Compute R2 score for the rest of samples in the pool

Maximize yield regression accuracy



```
[0/408] R2 test:        0.7707
[50/408]        R2 test:        0.8819
[100/408]       R2 test:        0.9017
[150/408]       R2 test:        0.9093
[200/408]       R2 test:        0.9118
[250/408]       R2 test:        0.9112
[300/408]       R2 test:        0.9146
[350/408]       R2 test:        0.9184
[400/408]       R2 test:        0.9190
R2 test:        0.9184
```





Paper’s results: ~0.6 after 5 iterations (510 budget)

Conclusion

- Reach a mean of ~19 for maximizing mean yield
- Reach a R2 score of ~0.9 for maximizing yield regression

- Code available at: ***https://github.com/zeyuanz/750hw4***