Solving the Modular Cell Biocatalyst Design Problem with Multi-objective Evolutionary Algorithms

Sergio Garcia and Cong T. Trinh

Department of Chemical and Biomolecular Engineering, The University of Tennessee, Knoxville, TN

Presented at Metabolic Pathway Analysis 2019
Riga, Latvia

Outline

Outline

Modular design concepts

A *module* is an essential and self-contained functional unit relative to the product of which it is part. The module has, relative to a system definition, standardized interfaces and interactions that allow composition of products by combination.

Module:

- Replaceable
- Changes system functionality

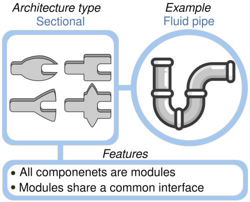
Modular design concepts

A *module* is an essential and self-contained functional unit relative to the product of which it is part. The module has, relative to a system definition, standardized interfaces and interactions that allow composition of products by combination.

Module:

- Replaceable
- Changes system functionality

Types of modular architecture:



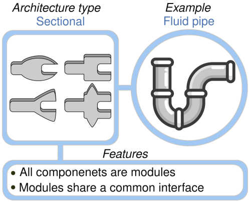
Modular design concepts

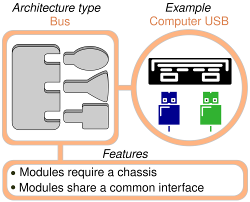
A *module* is an essential and self-contained functional unit relative to the product of which it is part. The module has, relative to a system definition, standardized interfaces and interactions that allow composition of products by combination.

Module:

- Replaceable
- Changes system functionality

Types of modular architecture:





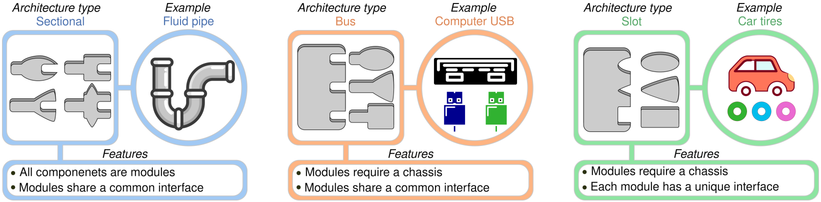
Modular design concepts

A *module* is an essential and self-contained functional unit relative to the product of which it is part. The module has, relative to a system definition, standardized interfaces and interactions that allow composition of products by combination.

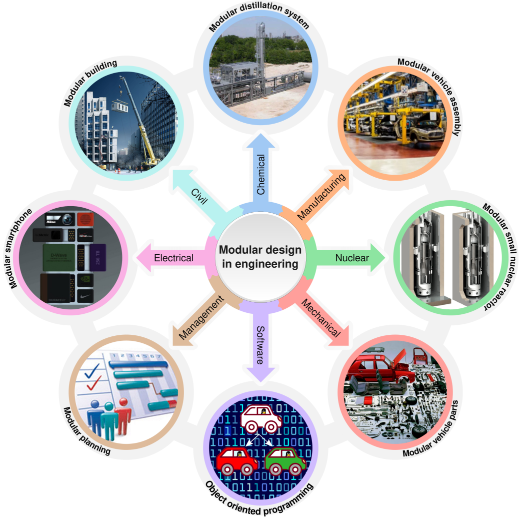
Module:

- Replaceable
- Changes system functionality

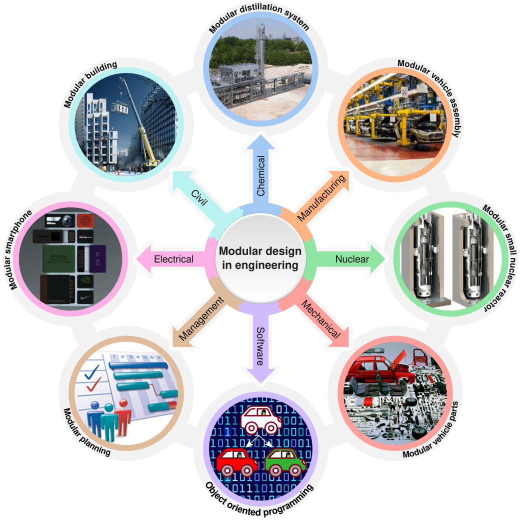
Types of modular architecture:



Driving forces and potential tradeoffs of modular design



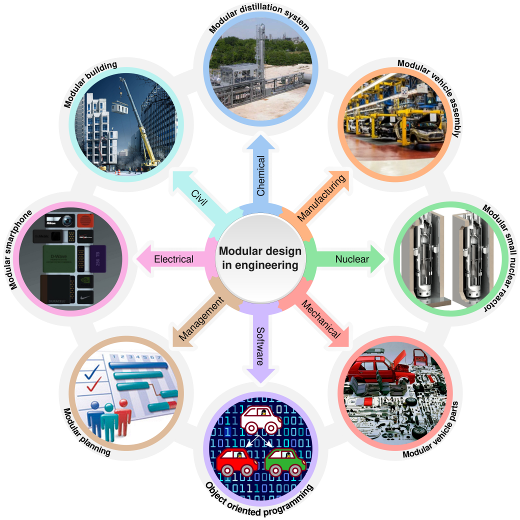
Driving forces and potential tradeoffs of modular design



Driving forces for modularization:

- Innovation
- Efficiency
- Customizability
- Predictability

Driving forces and potential tradeoffs of modular design



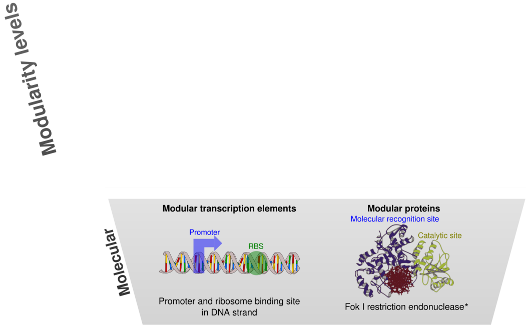
Driving forces for modularization:

- Innovation
- Efficiency
- Customizability
- Predictability

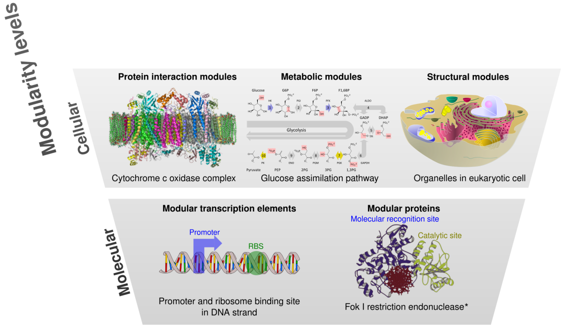
Potential drawbacks:

- Novelty cost
- Need for specialization
- Transportation constraints

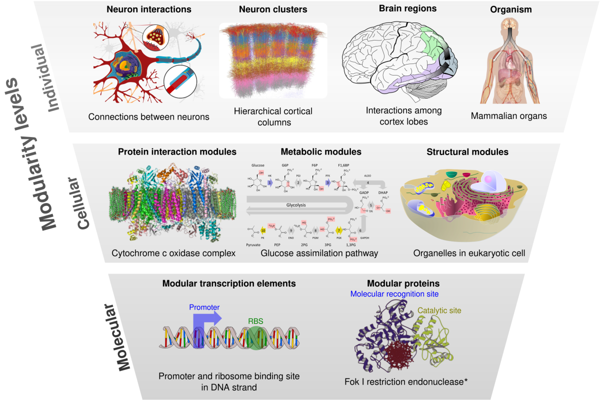
Modularity is an organizational principle across all scales of biology



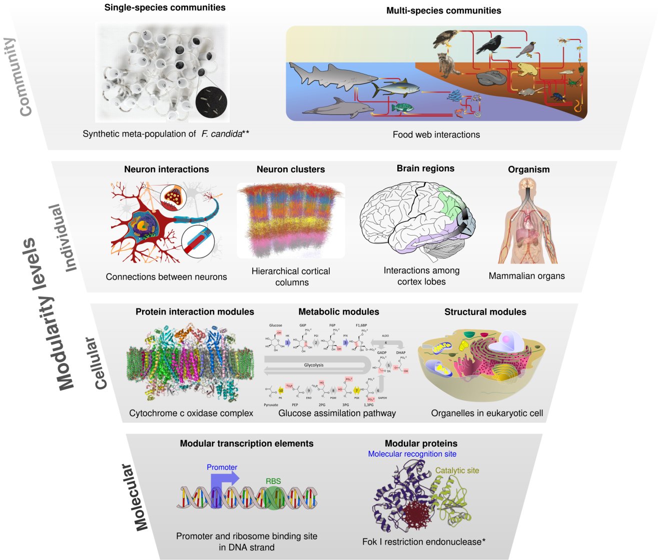
Modularity is an organizational principle across all scales of biology



Modularity is an organizational principle across all scales of biology

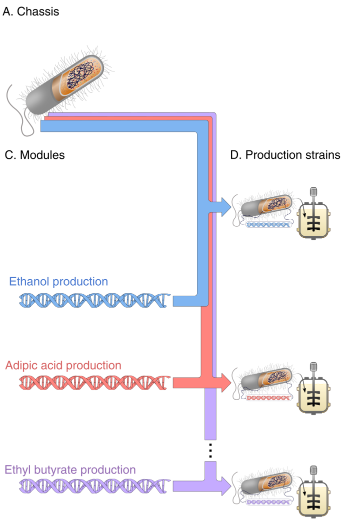


Modularity is an organizational principle across all scales of biology



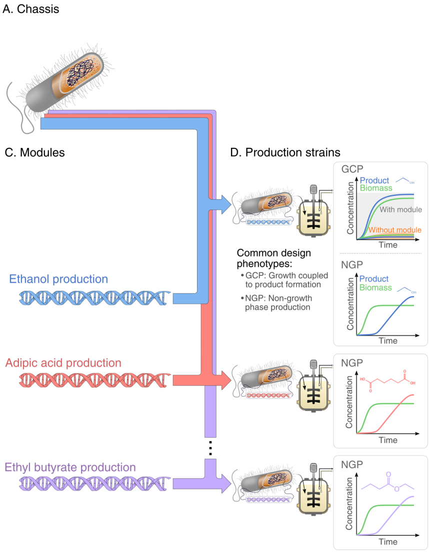
Outline

Principles of Modular Cell (ModCell) design



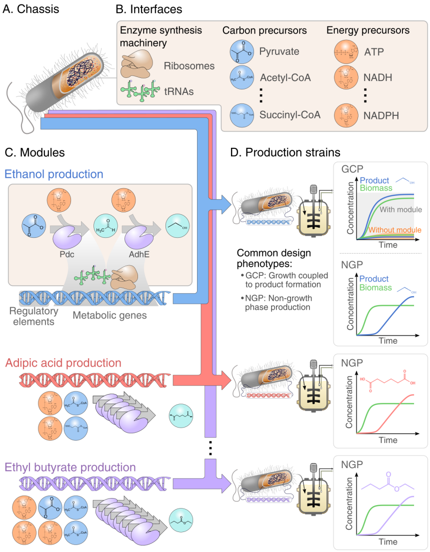
- The chassis can be combined with various modules in a plug-and-play fashion to obtain *production strains*

Principles of Modular Cell (ModCell) design



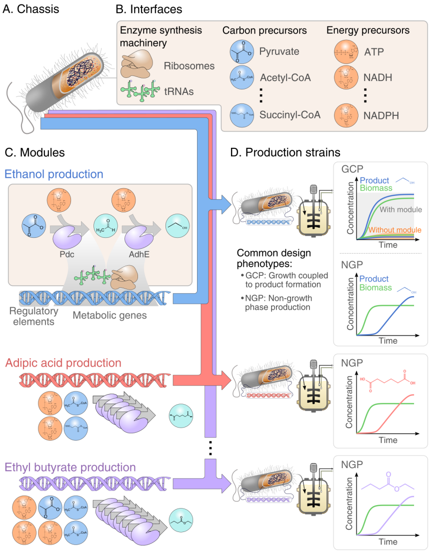
- The chassis can be combined with various modules in a plug-and-play fashion to obtain *production strains*
- Each *production strain* displays a desirable phenotype

Principles of Modular Cell (ModCell) design



- The chassis can be combined with various modules in a plug-and-play fashion to obtain *production strains*
- Each *production strain* displays a desirable phenotype
- Chassis, modules, and interfaces have to be designed in accordance to this desirable functions

Principles of Modular Cell (ModCell) design



- The chassis can be combined with various modules in a plug-and-play fashion to obtain *production strains*
- Each *production strain* displays a desirable phenotype
- Chassis, modules, and interfaces have to be designed in accordance to this desirable functions
- ModCell brings the same advantages of modularity in conventional engineering to metabolic engineering: efficiency and robustness

Mathematical formulation of ModCell

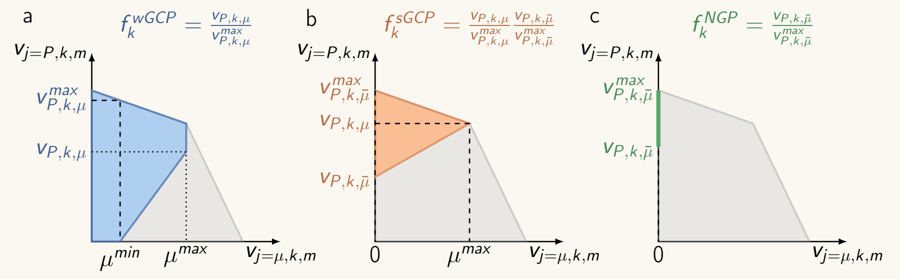
$$\max_{y_j, z_{jk}} \quad (f_{\mathbf{1}}, f_{\mathbf{2}}, \ldots, f_{|\mathcal{K}|})^T \quad \textbf{s.t.}$$

- Multi-objective optimization. Design objective f_k is the target phenotype of production strain k.

Mathematical formulation of ModCell

$$\max_{y_j, z_{jk}} \quad (f_1, f_2, \ldots, f_{|\mathcal{K}|})^T \quad \textbf{s.t.}$$

- Multi-objective optimization. Design objective f_k is the target phenotype of production strain k.



Mathematical formulation of ModCell

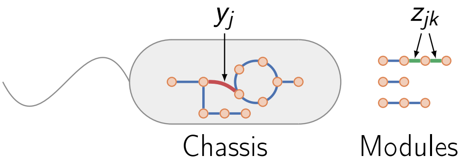
$$\max_{y_j, z_{jk}} \quad (f_1, f_2, \ldots, f_{|\mathcal{K}|})^T \quad \text{s.t.}$$

$$\sum_{j \in \mathcal{C}} (1 - y_j) \leq \alpha$$

$$\sum_{j \in \mathcal{C}} z_{jk} \leq \beta_k \qquad \text{for all } k \in \mathcal{K}$$

$$z_{jk} \leq (1 - y_j) \qquad \text{for all } j \in \mathcal{C},\ k \in \mathcal{K}$$

- Multi-objective optimization. Design objective f_k is the target phenotype of production strain k.
- Simultaneous design of chassis (y_j) and modules (z_{jk})



Mathematical formulation of ModCell

$$\max_{y_j, z_{jk}} \quad (f_1, f_2, \ldots, f_{|\mathcal{K}|})^T \quad \text{s.t.}$$

$$\sum_{j \in \mathcal{C}} (1 - y_j) \leq \alpha$$

$$\sum_{j \in \mathcal{C}} z_{jk} \leq \beta_k \qquad \text{for all } k \in \mathcal{K}$$

$$z_{jk} \leq (1 - y_j) \qquad \text{for all } j \in \mathcal{C},\ k \in \mathcal{K}$$

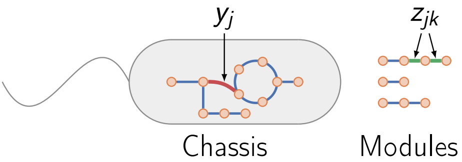
$$f_k \in \text{arg max} \left\{ \frac{1}{f_k^{max}} \sum_{j \in \mathcal{J}_k} c_{jk} v_{jk} \quad \text{s.t.} \right.$$

$$\sum_{j \in \mathcal{J}_k} S_{ijk} v_{jk} = 0 \qquad \text{for all } i \in \mathcal{I}_k$$

$$l_{jk} \leq v_{jk} \leq u_{jk} \qquad \text{for all } j \in \mathcal{J}_k$$

$$l_{jk} d_{jk} \leq v_{jk} \leq u_{jk} d_{jk} \qquad \text{for all } j \in \mathcal{C}$$

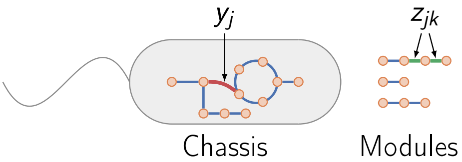
$$\left. \text{where } d_{jk} = y_j \vee z_{jk} \right\} \qquad \text{for all } k \in \mathcal{K}$$



- Multi-objective optimization. Design objective f_k is the target phenotype of production strain k.
- Simultaneous design of chassis (y_j) and modules (z_{jk})
- Flux prediction based, but not limited, in constraint-based models

Mathematical formulation of ModCell

$$
\begin{aligned}
&\max_{y_j, z_{jk}} \quad (f_1, f_2, \ldots, f_{|\mathcal{K}|})^T \quad \textbf{s.t.} \\
&\quad \sum_{j \in \mathcal{C}} (1 - y_j) \le \alpha \\
&\quad \sum_{j \in \mathcal{C}} z_{jk} \le \beta_k && \text{for all } k \in \mathcal{K} \\
&\quad z_{jk} \le (1 - y_j) && \text{for all } j \in \mathcal{C},\ k \in \mathcal{K} \\
&\quad f_k \in \textbf{arg max} \Bigg\{ \frac{1}{f_k^{max}} \sum_{j \in \mathcal{J}_k} c_{jk} v_{jk} \quad \textbf{s.t.} \\
&\qquad \sum_{j \in \mathcal{J}_k} S_{ijk} v_{jk} = 0 && \text{for all } i \in \mathcal{I}_k \\
&\qquad l_{jk} \le v_{jk} \le u_{jk} && \text{for all } j \in \mathcal{J}_k \\
&\qquad l_{jk} d_{jk} \le v_{jk} \le u_{jk} d_{jk} && \text{for all } j \in \mathcal{C} \\
&\qquad \text{where } d_{jk} = y_j \vee z_{jk} \Bigg\} && \text{for all } k \in \mathcal{K}
\end{aligned}
$$



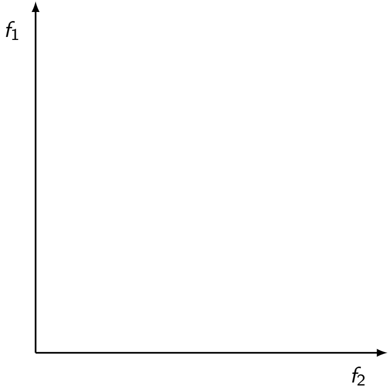
- Multi-objective optimization. Design objective f_k is the target phenotype of production strain k.
- Simultaneous design of chassis (y_j) and modules (z_{jk})
- Flux prediction based, but not limited, in constraint-based models
- New strain design approach to simultaneously consider an arbitrary number of target phenotypes, thus reducing redundant engineering efforts

Outline

What defines a solution?

Multi-objective optimization problem

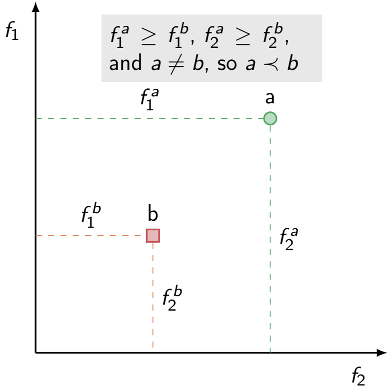
$$\max_{x \in X} \quad F(x) = (f_1(x), f_2(x), \dots)^T$$



What defines a solution?

Multi-objective optimization problem

$$\max_{x \in X} \quad F(x) = (f_1(x), f_2(x), \ldots)^T$$



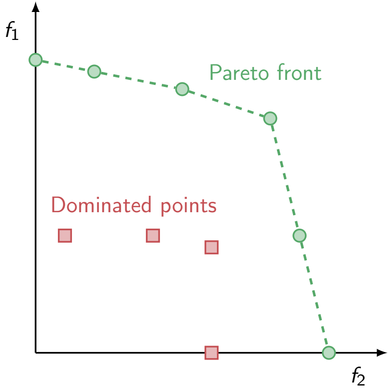
Definition of domination

A vector a dominates another vector b (denoted $a \prec b$) iff $a_i \geq b_i \; \forall i \in \{1, 2, \ldots, K\}$ and $a_i \neq b_i$ for at least one i.

What defines a solution?

Multi-objective optimization problem

$$\max_{x \in X} \quad F(x) = (f_1(x), f_2(x), \ldots)^T$$



Definition of domination

A vector a dominates another vector b (denoted $a \prec b$) iff $a_i \geq b_i \; \forall i \in \{1, 2, \ldots, K\}$ and $a_i \neq b_i$ for at least one i.

Pareto set

$$PS := \{x \in X : \nexists x' \in X, F(x') \prec F(x)\}$$

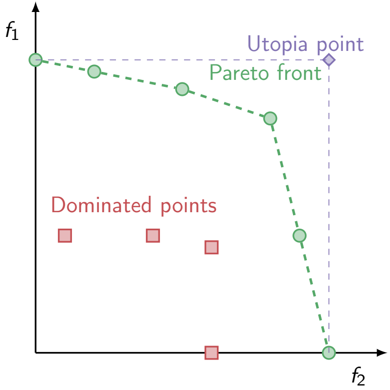
Pareto front

$$PF := \{F(x) : x \in PS\}$$

What defines a solution?

Multi-objective optimization problem

$$\max_{x \in X} \quad F(x) = (f_1(x), f_2(x), \ldots)^T$$



Definition of domination

A vector a dominates another vector b (denoted $a \prec b$) iff $a_i \geq b_i \; \forall i \in \{1, 2, \ldots, K\}$ and $a_i \neq b_i$ for at least one i.

Pareto set

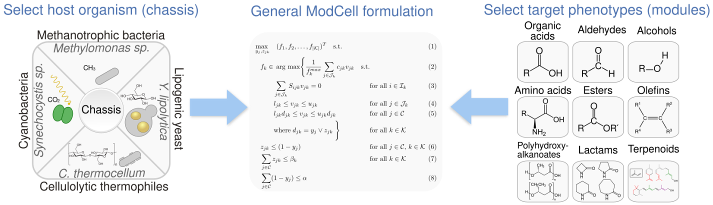
$$PS := \{x \in X : \nexists x' \in X, F(x') \prec F(x)\}$$

Pareto front

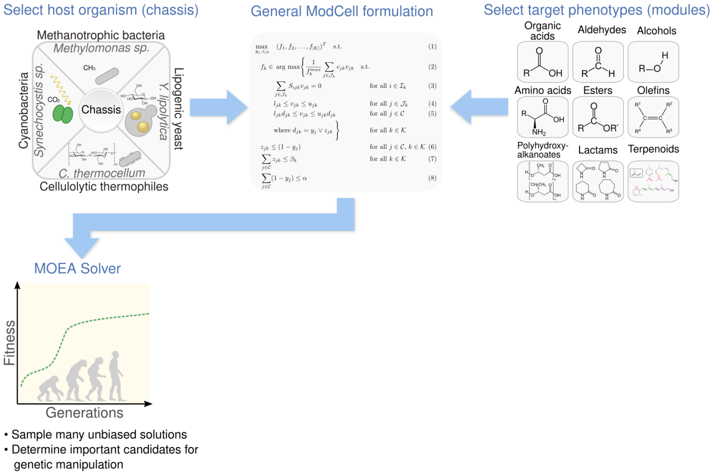
$$PF := \{F(x) : x \in PS\}$$

- Utopia point

Two complementary solvers: MOEA and MILP

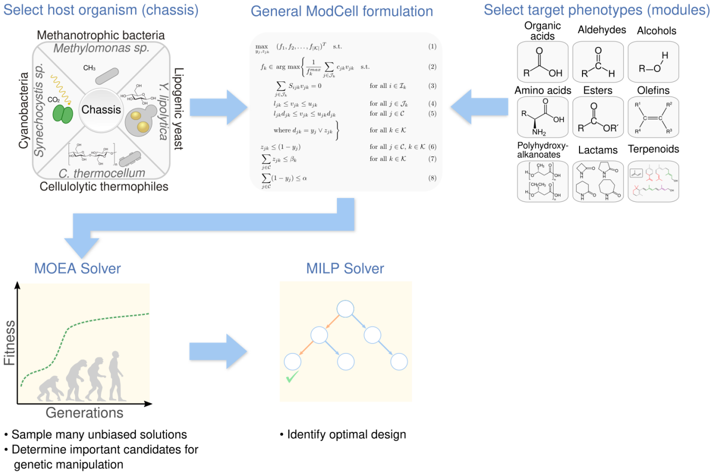


Two complementary solvers: MOEA and MILP



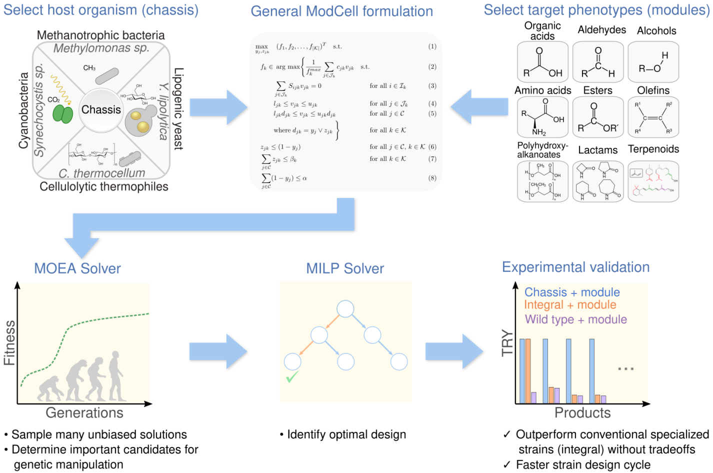
- MOEA is a highly flexible heuristic optimization method but cannot guarantee optimality.

Two complementary solvers: MOEA and MILP



- MOEA is a highly flexible heuristic optimization method but cannot guarantee optimality.
- MILP is more restricted than MOEA in formulation and harder to solve but can ensure optimality (See Poster P6)

Two complementary solvers: MOEA and MILP



- MOEA is a highly flexible heuristic optimization method but cannot guarantee optimality.
- MILP is more restricted than MOEA in formulation and harder to solve but can ensure optimality (See Poster P6)

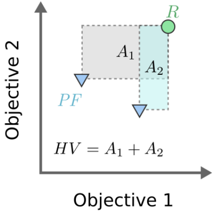
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):

Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):

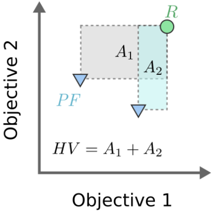
a.



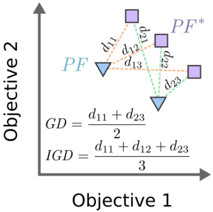
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):

a.

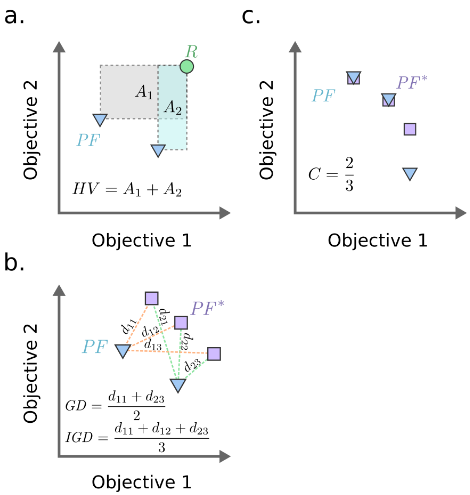


b.



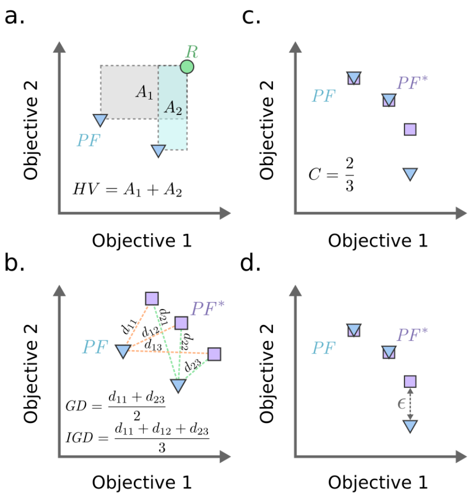
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):



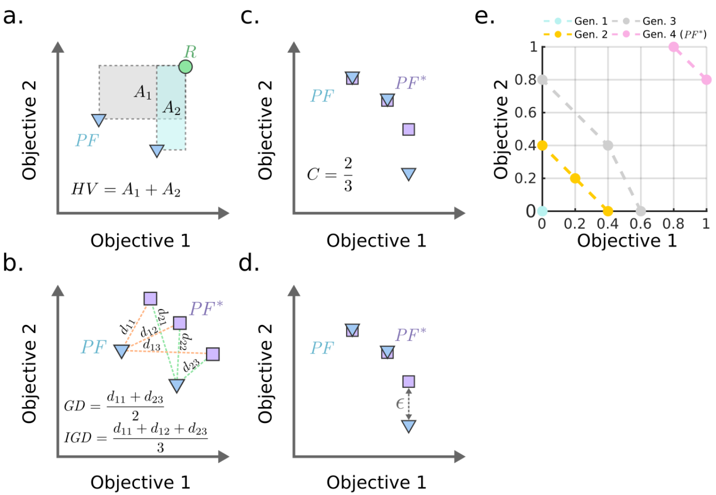
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):



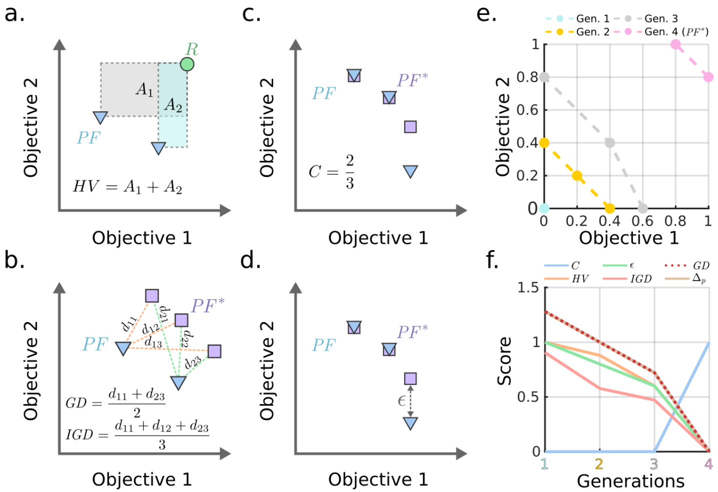
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):

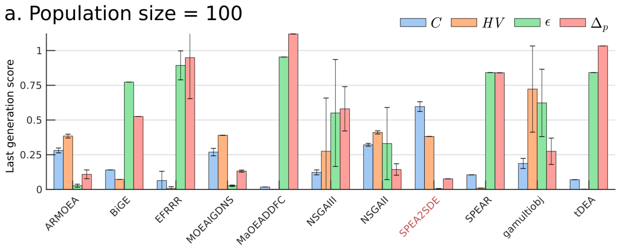


Measuring MOEA performance

MOEAs do not guarantee optimality, how do we asses the performance to choose the best algorithm and parameters? We measure the distance between the best known Pareto front (PF^*) and the current solution (PF):

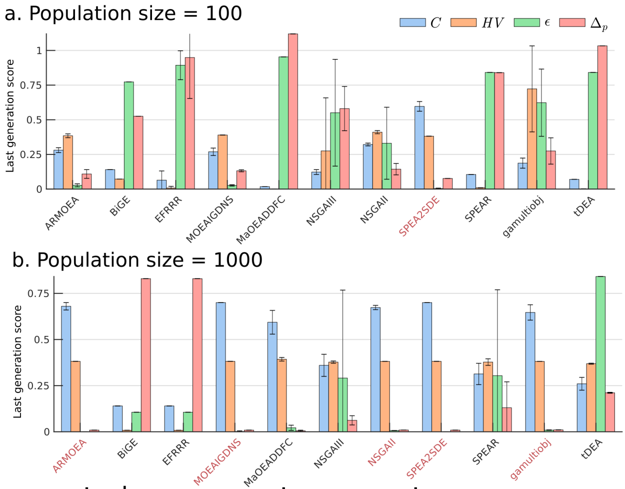


Population size is the most important factor in MOEA



- Coverage is the most consistent metric.
- For small population algorihtm heuristics matter.

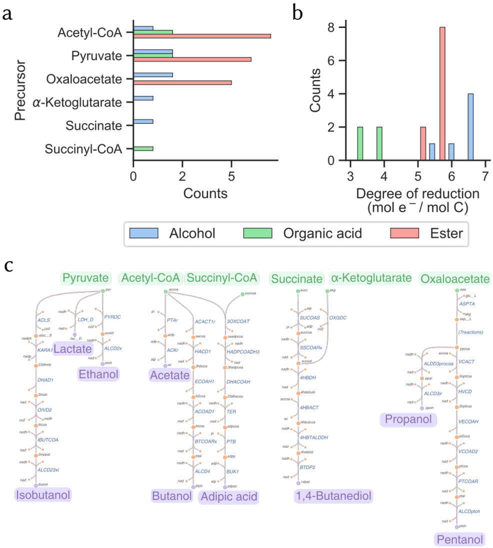
Population size is the most important factor in MOEA



- Coverage is the most consistent metric.
- For small population algorihtm heuristics matter.
- For large population sizes several algorithms attain the best results.

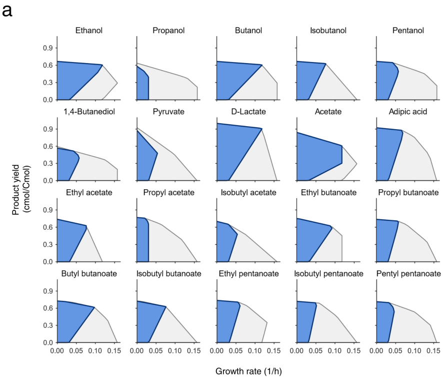
Outline

Input: 20 diverse products



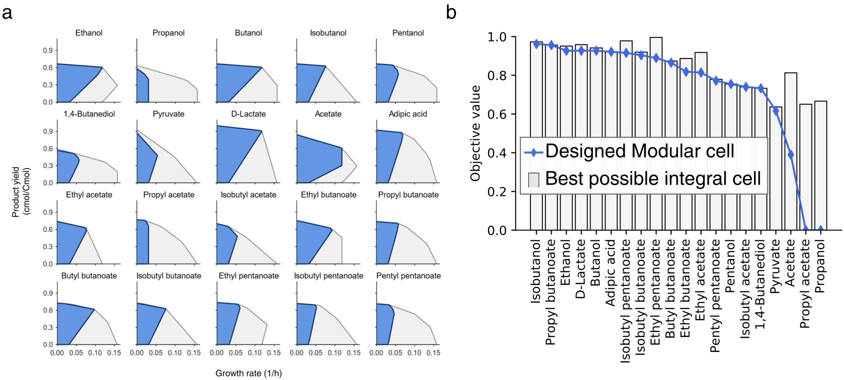
- *E. coli* as a parent to build the chassis.
- 6 alcohols from C2 to C5.
- 4 carboxylic acids from C2 to C6.
- 10 derived esters from C4 to C10.
- Esters are synthesized from an acyl-CoA and alcohol by the AAT enzyme.

Results: Highly compatible chassis



- 4 gene-knockouts (*adhE*, *ldhA*, *ack-pta*, *zwf*) obtain *wGCP* design objective above 60% of maximum for 17 out of 20 products

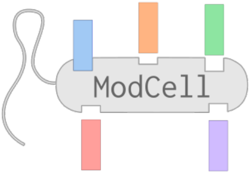
Results: Highly compatible chassis



- 4 gene-knockouts (*adhE, ldhA, ack-pta, zwf*) obtain *wGCP* design objective above 60% of maximum for 17 out of 20 products
- No loss of performance with respect to conventional single-product (integral) strain design

Summary

- Inspired by conventional engineering modularity, develop modular cell design concept to accelerate biocatalyst R&D cycles.



Summary

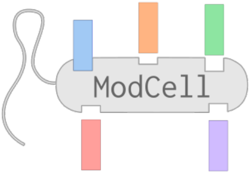
- Inspired by conventional engineering modularity, develop modular cell design concept to accelerate biocatalyst R&D cycles.
- Propose modular cell design as a multi-objective optimization problem, this framework allows to simultaneously design multiple target phenotypes minimizing redundant efforts.

Summary

- Inspired by conventional engineering modularity, develop modular cell design concept to accelerate biocatalyst R&D cycles.
- Propose modular cell design as a multi-objective optimization problem, this framework allows to simultaneously design multiple target phenotypes minimizing redundant efforts.
- Demonstrate MOEA and MILP approaches to solve the optimization problem.

Summary

- Inspired by conventional engineering modularity, develop modular cell design concept to accelerate biocatalyst R&D cycles.
- Propose modular cell design as a multi-objective optimization problem, this framework allows to simultaneously design multiple target phenotypes minimizing redundant efforts.
- Demonstrate MOEA and MILP approaches to solve the optimization problem.
- Design a chassis cell compatible with growth-coupled synthesis of 17 out of 20 products without loss of performance with respect to integral (conventional) strain design.



Acknowledgements

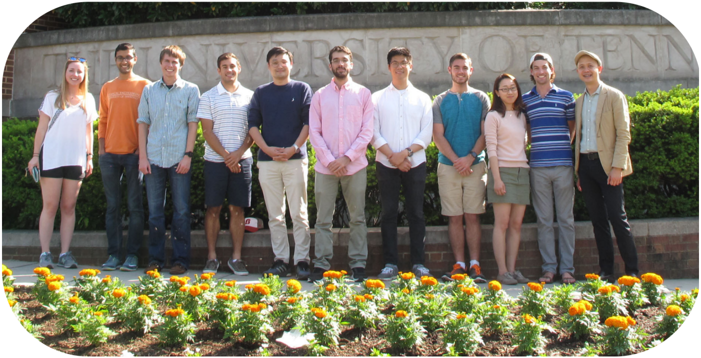
Funding Sources

NSF

THE UNIVERSITY OF TENNESSEE KNOXVILLE

cbi THE CENTER FOR BIOENERGY INNOVATION

Trinh Lab



References

- Garcia, S. & Trinh, C. T. Multiobjective strain design: A framework for modular cell engineering. *Metabolic Engineering* **51** (2019).
- Garcia, S. & Trinh, C. T. Modular design: Implementing proven engineering principles in biotechnology. *Biotechnology Advances* (2019).
- Garcia, S. & Trinh, C. T. Comparison of Multi-Objective Evolutionary Algorithms to Solve the Modular Cell Design Problem for Novel Biocatalysis. *Processes* **7** (2019).

All programs and data analysis scripts are available on Github with detailed documentation to enable reproducibility and further use:
`https://github.com/trinhlab`