

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

1. Modular design

- 1.1 Modularity in engineering
- 1.2 Modularity in nature

2. Modular cells

- 2.1 Conceptual formulation
- 2.2 Mathematical formulation

3. Solution algorithms

- 3.1 What defines a solution?
- 3.2 Two complementary solvers: MOEA and MILP
- 3.3 Measuring MOEA performance

4. Application example

- 4.1 Input: 20 diverse products
- 4.2 Results: Highly compatible chassis

Outline

1. Modular design

- 1.1 Modularity in engineering
- 1.2 Modularity in nature

2. Modular cells

- 2.1 Conceptual formulation
- 2.2 Mathematical formulation

3. Solution algorithms

- 3.1 What defines a solution?
- 3.2 Two complementary solvers: MOEA and MILP
- 3.3 Measuring MOEA performance

4. Application example

- 4.1 Input: 20 diverse products
- 4.2 Results: Highly compatible chassis

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

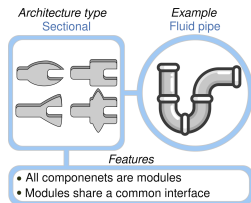
Modular design concepts

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:



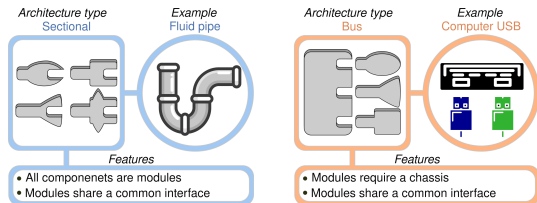
Modular design concepts

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:



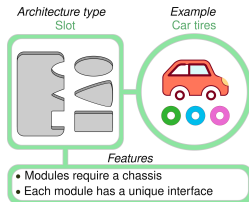
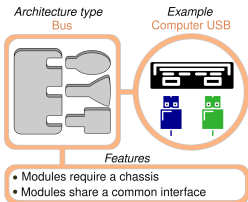
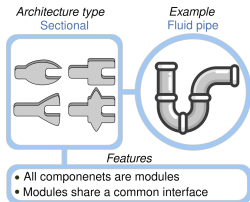
Modular design concepts

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.

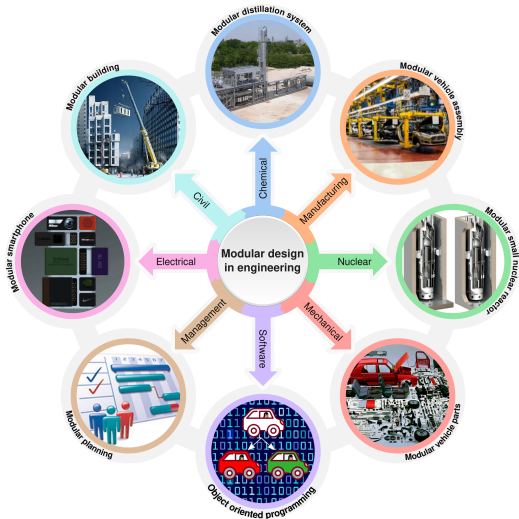
Module:

- ▶ Replaceable
- ▶ Changes system functionality

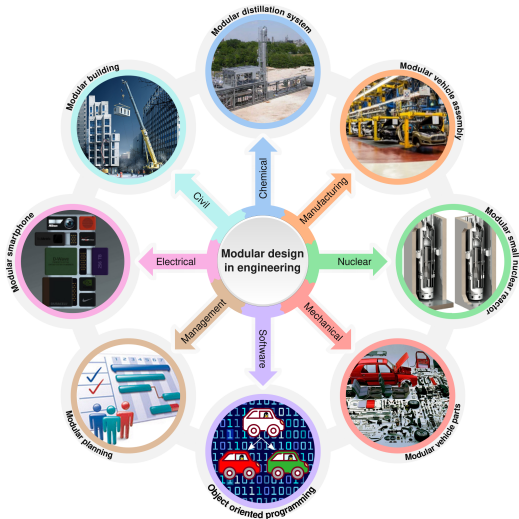
Types of modular architecture:



Driving forces and potential tradeoffs of modular design



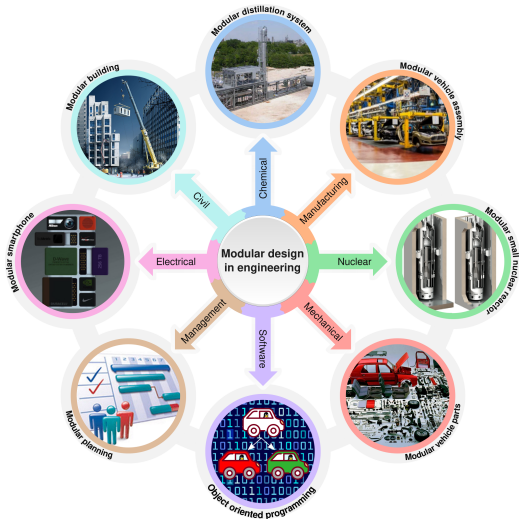
Driving forces and potential tradeoffs of modular design



Driving forces for modularization:

- ▶ Innovation
- ▶ Efficiency
- ▶ Customizability
- ▶ Predictability

Driving forces and potential tradeoffs of modular design



Driving forces for modularization:

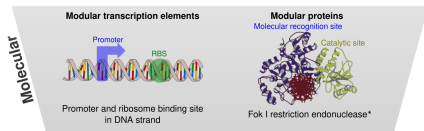
- ▶ Innovation
- ▶ Efficiency
- ▶ Customizability
- ▶ Predictability

Potential drawbacks:

- ▶ Novelty cost
- ▶ Need for special-ization
- ▶ Transportation constraints

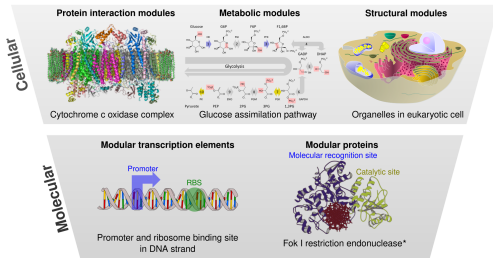
Modularity is an organizational principle across all scales of biology

Modularity levels

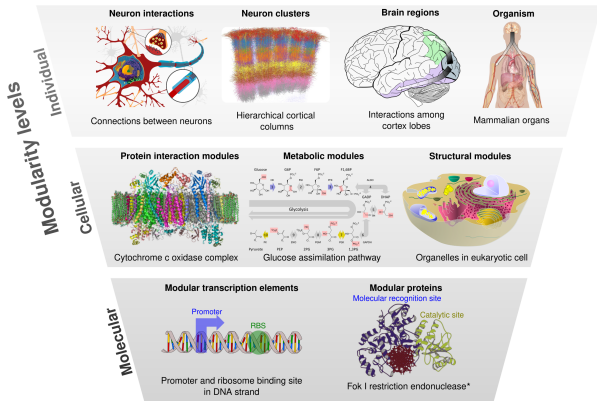


Modularity is an organizational principle across all scales of biology

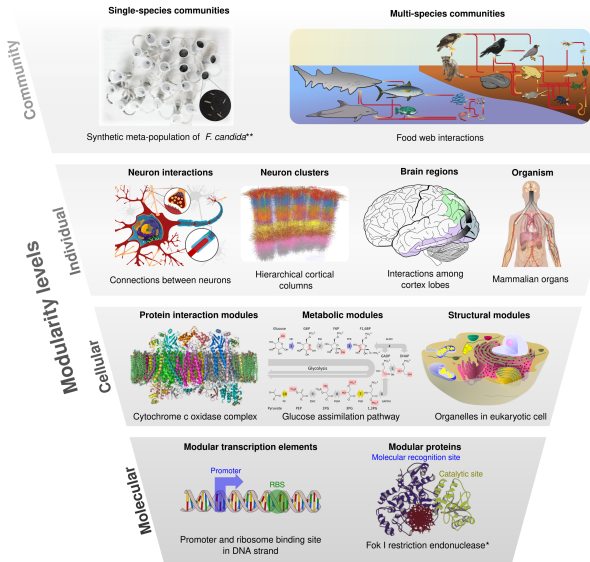
Modularity levels



Modularity is an organizational principle across all scales of biology



Modularity is an organizational principle across all scales of biology



Outline

1. Modular design

- 1.1 Modularity in engineering
- 1.2 Modularity in nature

2. Modular cells

- 2.1 Conceptual formulation
- 2.2 Mathematical formulation

3. Solution algorithms

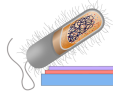
- 3.1 What defines a solution?
- 3.2 Two complementary solvers: MOEA and MILP
- 3.3 Measuring MOEA performance

4. Application example

- 4.1 Input: 20 diverse products
- 4.2 Results: Highly compatible chassis

Principles of Modular Cell (ModCell) design

A. Chassis



C. Modules

Ethanol production



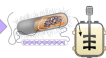
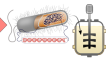
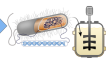
Adipic acid production



Ethyl butyrate production



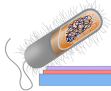
D. Production strains



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

Principles of Modular Cell (ModCell) design

A. Chassis



C. Modules

Ethanol production



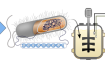
Adipic acid production



Ethyl butyrate production



D. Production strains



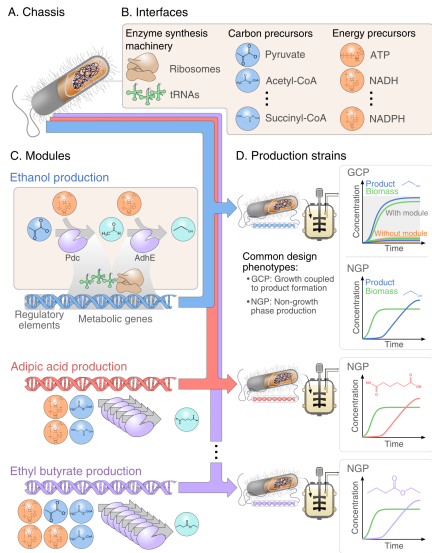
Common design phenotypes:

- GCP: Growth coupled to product formation
- NGP: Non-growth phase production



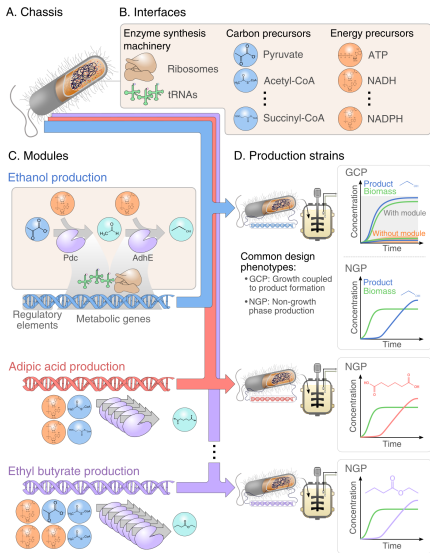
- ▶ The chassis can be combined with various modules in a plug-and-play fashion to obtain *production strains*
- ▶ Each *production strain* displays a desirable phenotype

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

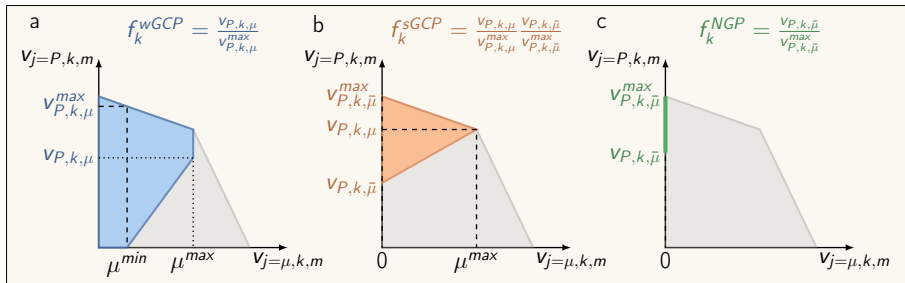
$$\max_{y_j, z_{jk}} (f_1, f_2, \dots, f_{|\mathcal{K}|})^T \quad \text{s.t.}$$

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

Mathematical formulation of ModCell

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

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



Mathematical formulation of ModCell

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

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

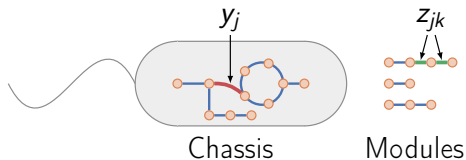
$$\sum_{j \in \mathcal{C}} z_{jk} \leq \beta_k$$

$$z_{jk} \leq (1 - y_j)$$

for all $k \in \mathcal{K}$

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

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

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

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

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

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

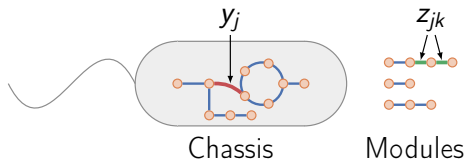
$$\sum_{j \in \mathcal{J}_k} s_{ijk} v_{jk} = 0 \quad \text{for all } i \in \mathcal{I}_k$$

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

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

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

- ▶ 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

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

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

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

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

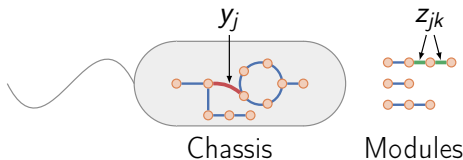
$$f_k \in \arg \max \left\{ \frac{1}{f_k^{\max}} \sum_{j \in \mathcal{J}_k} c_{jk} v_{jk} \right\} \quad \text{s.t.}$$

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

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

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

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



- ▶ 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
- ▶ New strain design approach to simultaneously consider an arbitrary number of target phenotypes, thus reducing redundant engineering efforts

Outline

1. Modular design

- 1.1 Modularity in engineering
- 1.2 Modularity in nature

2. Modular cells

- 2.1 Conceptual formulation
- 2.2 Mathematical formulation

3. Solution algorithms

- 3.1 What defines a solution?
- 3.2 Two complementary solvers: MOEA and MILP
- 3.3 Measuring MOEA performance

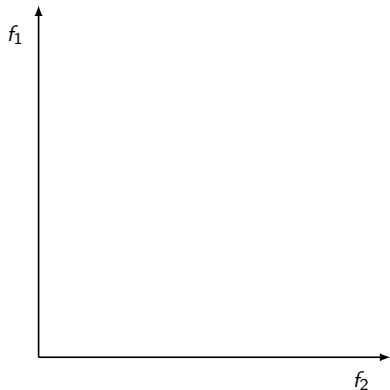
4. Application example

- 4.1 Input: 20 diverse products
- 4.2 Results: Highly compatible chassis

What defines a solution?

Multi-objective optimization problem

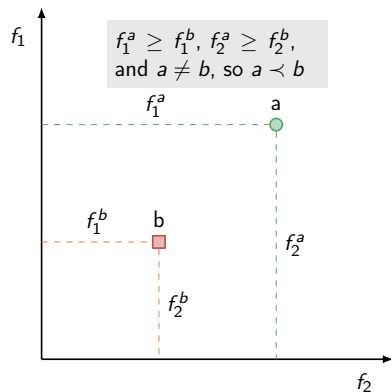
$$\max_{x \in X} F(x) = (f_1(x), f_2(x), \dots)^T$$



What defines a solution?

Multi-objective optimization problem

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



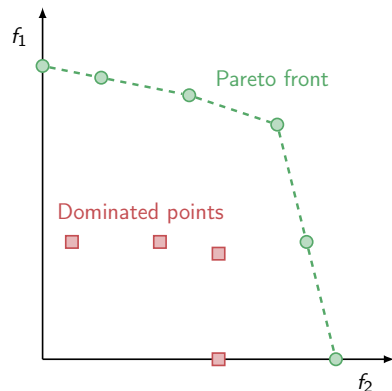
Definition of domination

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

What defines a solution?

Multi-objective optimization problem

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



Definition of domination

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

Pareto set

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

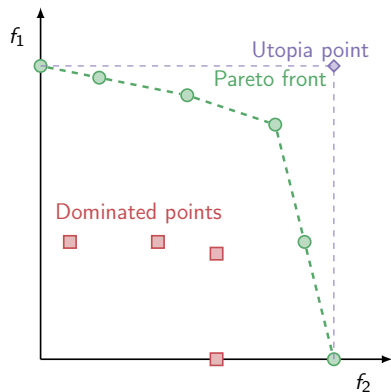
Pareto front

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

What defines a solution?

Multi-objective optimization problem

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



Definition of domination

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

Pareto set

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

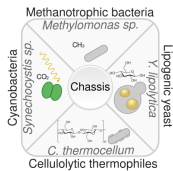
Pareto front

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

► Utopia point

Two complementary solvers: MOEA and MILP

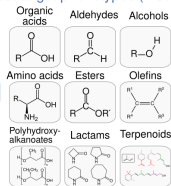
Select host organism (chassis)



General ModCell formulation

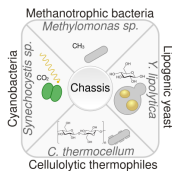
$$\begin{aligned} \max_{\{v_{ja}\}} & \{f_1, f_2, \dots, f_{|K|}\}^T \quad \text{s.t.} & (1) \\ f_j & \in \arg \max \left\{ \frac{1}{|I_j|} \sum_{i \in I_j} v_{ja} \mid v_{ja} \right. & (2) \\ & \left. \sum_{j \in I_a} v_{ja} = 0 \right\} & \text{for all } i \in I_a & (3) \\ v_{ja} & \leq v_{ja} \leq w_{ja} & \text{for all } j \in J_a & (4) \\ v_{ja} d_{ja} & \leq v_{ja} \leq w_{ja} d_{ja} & \text{for all } j \in C & (5) \\ & \left. \text{where } d_{ja} = \delta_j \vee \lambda_{ja} \right\} & \text{for all } k \in K & \\ v_{ja} & \leq (1 - \eta_j) & \text{for all } j \in C, k \in K & (6) \\ \sum_{j \in C} v_{ja} & \leq \beta_k & \text{for all } k \in K & (7) \\ \sum_{j \in C} (1 - \eta_j) & \leq \alpha & (8) \end{aligned}$$

Select target phenotypes (modules)



Two complementary solvers: MOEA and MILP

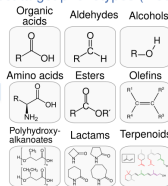
Select host organism (chassis)



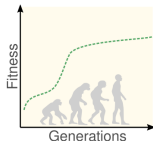
General ModCell formulation

$$\begin{aligned}
 & \max_{v_{jka}} \{f_1, f_2, \dots, f_{|K|}\}^T \quad \text{s.t.} & (1) \\
 & f_k \in \text{arg max} \left\{ \frac{1}{|I_k|} \sum_{j \in I_k} v_{jka} \mid \text{s.t.} \right. & (2) \\
 & \quad \left. \sum_{j \in I_k} v_{jka} = 0 \right. & \text{for all } i \in I_k & (3) \\
 & \quad v_{jka} \leq v_{ja} \leq v_{ka} & \text{for all } j \in J_k & (4) \\
 & \quad v_{jka} d_{jka} \leq v_{ja} \leq v_{ka} d_{jka} & \text{for all } j \in C & (5) \\
 & \quad \left. \text{where } d_{jka} = \theta_j \vee \lambda_{jka} \right\} & \text{for all } k \in K & (6) \\
 & v_{jka} \leq (1 - \eta_j) & \text{for all } j \in C, k \in K & (7) \\
 & \sum_{j \in C} v_{jka} \leq \beta_k & \text{for all } k \in K & (8) \\
 & \sum_{j \in C} (1 - \eta_j) \leq \alpha & &
 \end{aligned}$$

Select target phenotypes (modules)



MOEA Solver

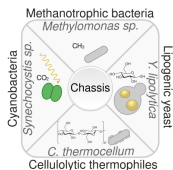


- Sample many unbiased solutions
- Determine important candidates for genetic manipulation

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

Two complementary solvers: MOEA and MILP

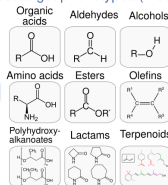
Select host organism (chassis)



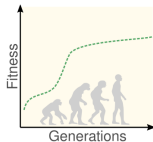
General ModCell formulation

$$\begin{aligned}
 & \max_{\forall i \in \mathcal{I}_i} \{f_1, f_2, \dots, f_{|\mathcal{I}_i|}\} \quad \text{s.t.} & (1) \\
 & f_i \in \arg \max \left\{ \frac{1}{|\mathcal{I}_i|} \sum_{j \in \mathcal{I}_i} v_{j,i} \quad \text{s.t.} \right. & (2) \\
 & \quad \sum_{j \in \mathcal{I}_i} s_{j,i} v_{j,i} = 0 & (3) \\
 & \quad v_{j,i} \leq v_{j,k} \leq w_{j,k} & (4) \\
 & \quad v_{j,i} w_{j,i} \leq v_{j,k} \leq w_{j,i} w_{j,k} & (5) \\
 & \quad \text{where } w_{j,i} = w_j \forall j, i & \\
 & v_{j,i} \leq (1 - \eta_j) & (6) \\
 & \sum_{j \in \mathcal{I}_i} v_{j,i} \leq \alpha_i & (7) \\
 & \sum_{j \in \mathcal{I}_i} (1 - \eta_j) \leq \alpha & (8)
 \end{aligned}$$

Select target phenotypes (modules)

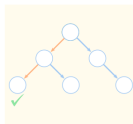


MOEA Solver



- Sample many unbiased solutions
- Determine important candidates for genetic manipulation

MILP Solver

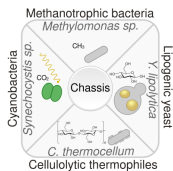


- Identify optimal design

- ▶ 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

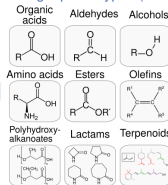
Select host organism (chassis)



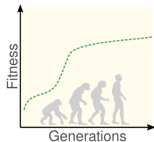
General ModCell formulation

$$\begin{aligned} \max_{\mathbf{v}, \mathbf{a}} & \{f_1, f_2, \dots, f_n\}^T \quad \text{s.t.} & (1) \\ f_i & \in \arg \max \left\{ \frac{1}{V_{i,j}} \sum_{j \in \mathcal{I}_i} v_{j,i} v_{j,i} \quad \text{s.t.} \right. & (2) \\ & \sum_{j \in \mathcal{I}_i} s_{j,i} v_{j,i} = 0 & (3) \\ & v_{j,i} \leq v_{j,i} \leq w_{j,i} & (4) \\ & v_{j,i} a_{j,i} \leq v_{j,i} \leq w_{j,i} a_{j,i} & (5) \\ & \text{where } a_{j,i} = 0, \forall j, i & \\ & v_{j,i} \leq (1 - \beta_j) & (6) \\ & \sum_{j \in \mathcal{I}_i} v_{j,i} \leq \alpha_i & (7) \\ & \sum_{j \in \mathcal{I}_i} (1 - \beta_j) \leq \alpha & (8) \end{aligned}$$

Select target phenotypes (modules)

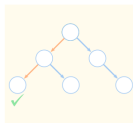


MOEA Solver



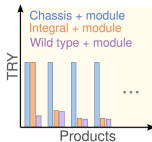
- Sample many unbiased solutions
- Determine important candidates for genetic manipulation

MILP Solver



- Identify optimal design

Experimental validation



- ✓ Outperform conventional specialized strains (integral) without tradeoffs
- ✓ Faster strain design cycle

- ▶ 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)

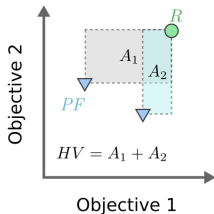
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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 assess 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):

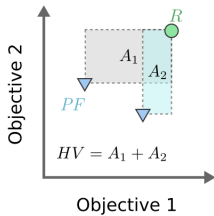
a.



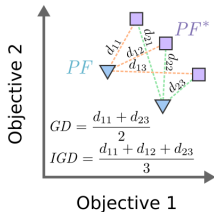
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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.



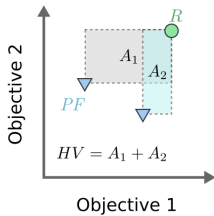
b.



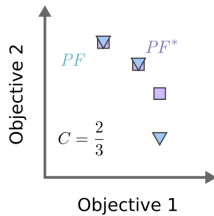
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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):

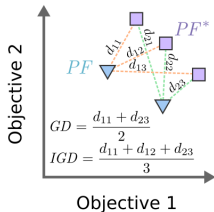
a.



c.



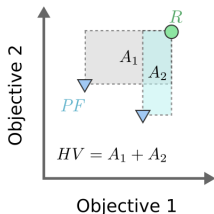
b.



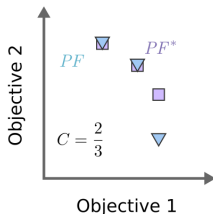
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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):

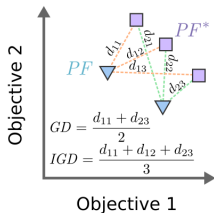
a.



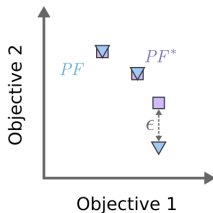
c.



b.

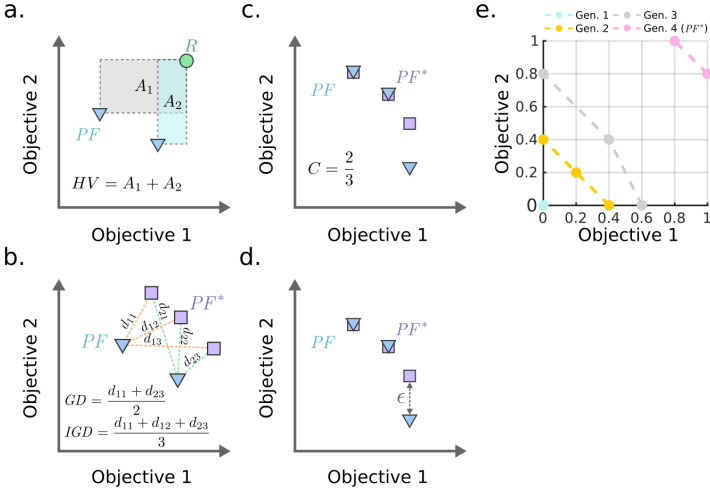


d.



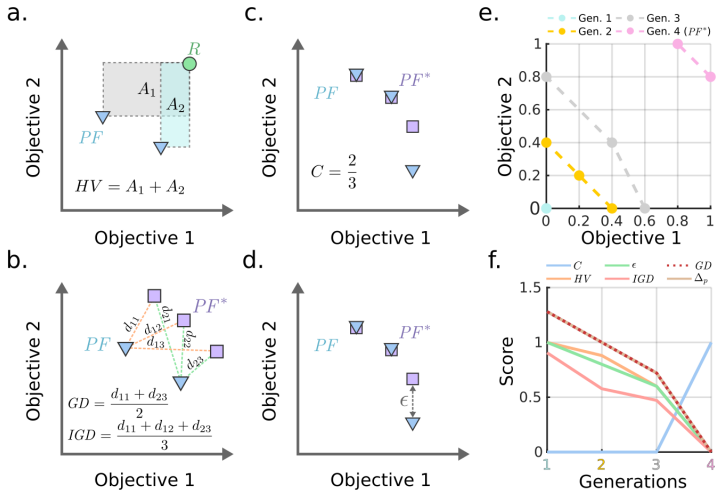
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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):



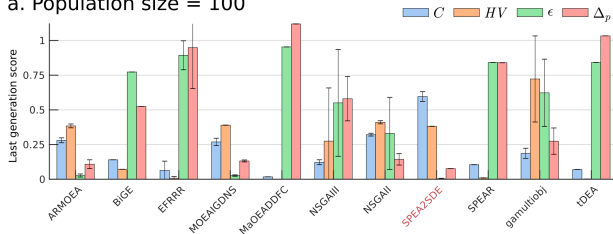
Measuring MOEA performance

MOEAs do not guarantee optimality, how do we assess 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):



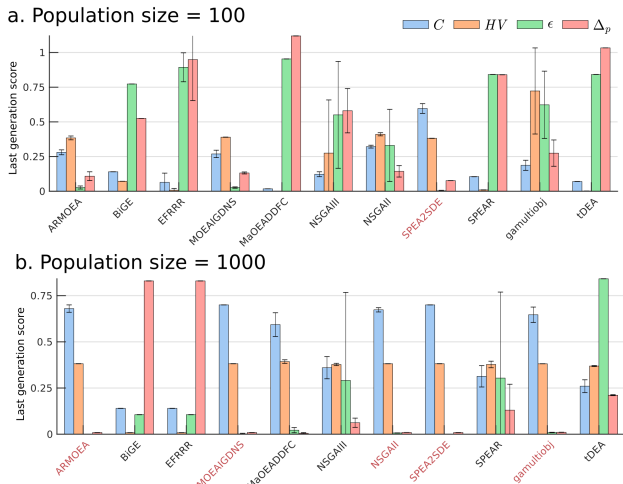
Population size is the most important factor in MOEA

a. Population size = 100



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

Population size is the most important factor in MOEA



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

Outline

1. Modular design

- 1.1 Modularity in engineering
- 1.2 Modularity in nature

2. Modular cells

- 2.1 Conceptual formulation
- 2.2 Mathematical formulation

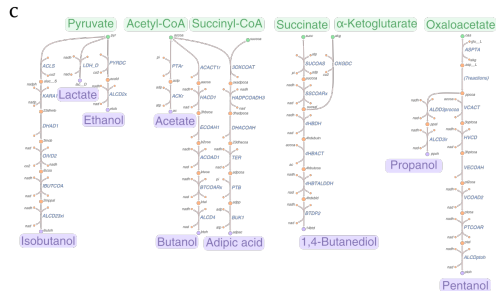
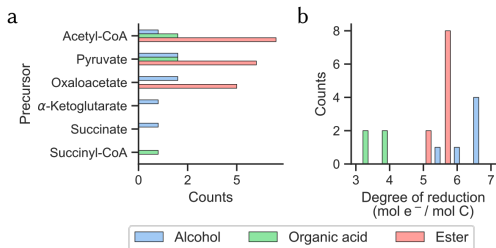
3. Solution algorithms

- 3.1 What defines a solution?
- 3.2 Two complementary solvers: MOEA and MILP
- 3.3 Measuring MOEA performance

4. Application example

- 4.1 Input: 20 diverse products
- 4.2 Results: Highly compatible chassis

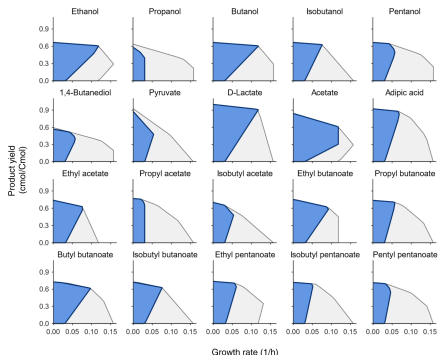
Input: 20 diverse products



- ▶ *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.

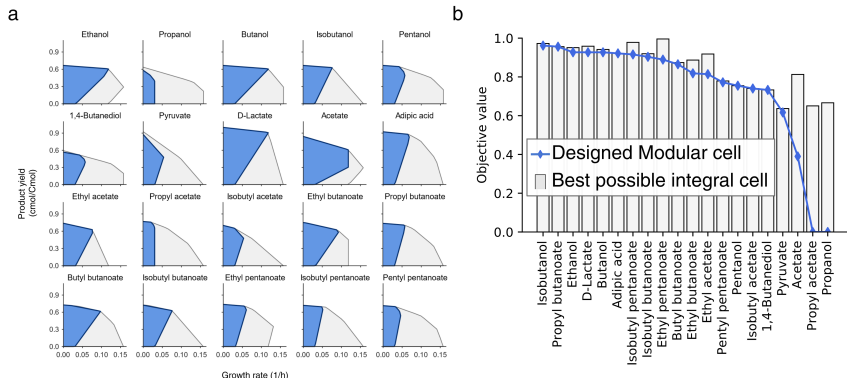
Results: Highly compatible chassis

a



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

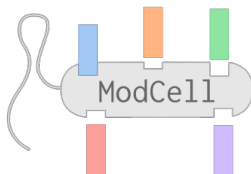
Results: Highly compatible chassis



- ▶ 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

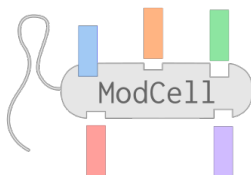
Summary

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



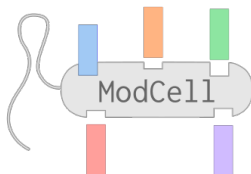
Summary

- ▶ 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.



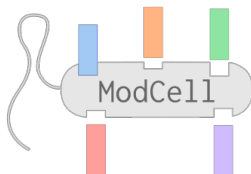
Summary

- ▶ 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

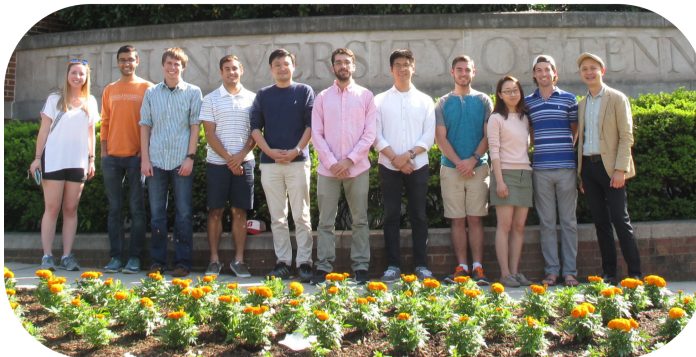
- ▶ Inspired by conventional engineering modularity, develop modular cell design concept to accelerate biocatalyst R&D cycles.
- ▶ Propose modular cell design as a multi-objective optimization problem, this framework allows to simultaneously design multiple target phenotypes minimizing redundant efforts.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






Funding Sources



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>