Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

DAT121
Introduction to data science

5 Multidimensionality

5.4 Pareto front visualization
5.5 MCO in modelling (& related research)

Glossary terms

Proposed glossary[1] terms:

- How do we best define them? Is the definition controversial?
- What is the best translation into Norwegian bokmål/nynorsk?
- Are there more key concepts that would require an agreed definition?

rationality

Pareto optimality

optimization parameter

optimization objective

agent

[1] https://home.bawue.de/~horsch/teaching/dat121/glossary-en.html

Norwegian University of Life Sciences

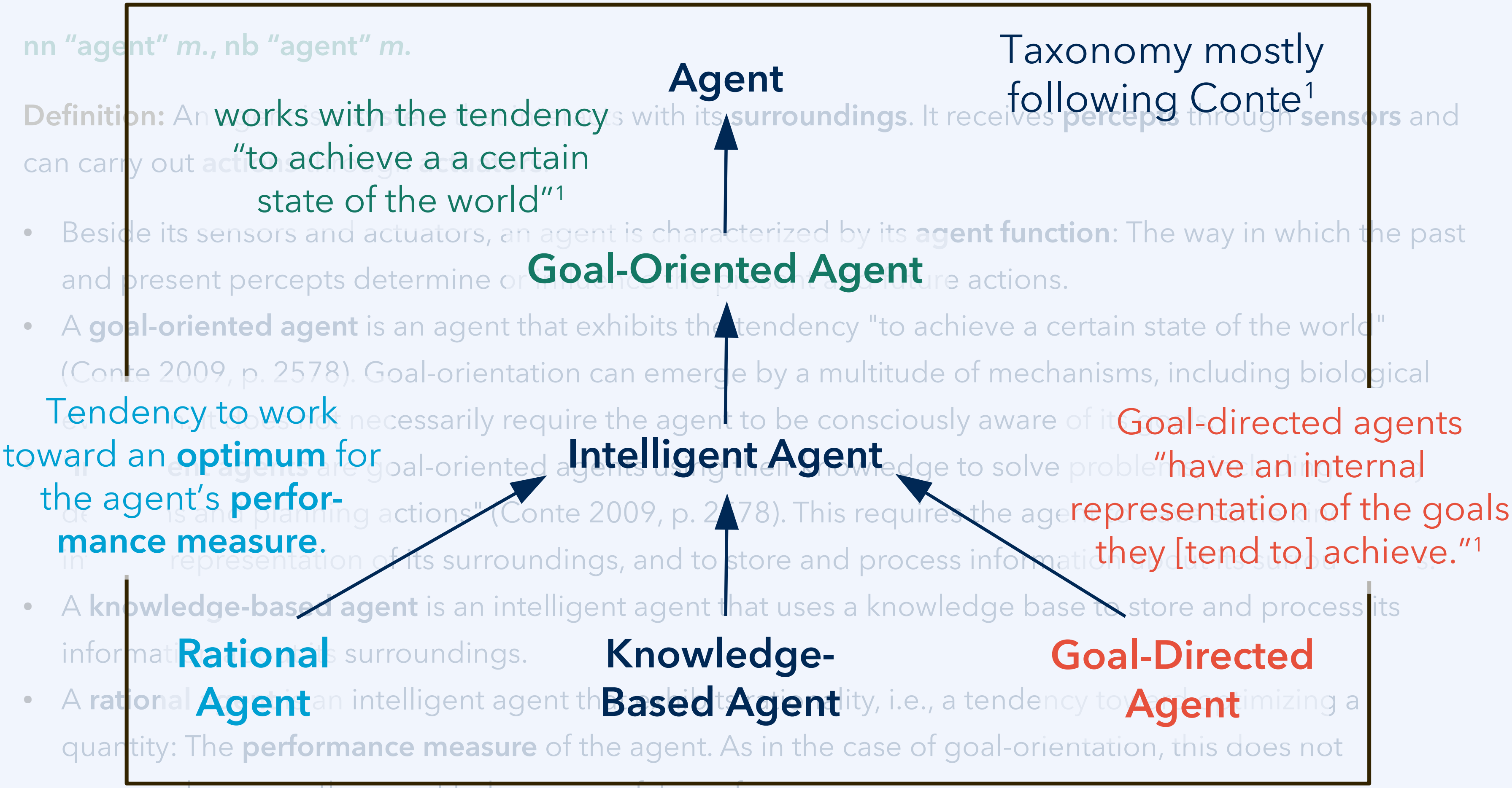
Glossary terms: "Agent"

nn "agent" *m.*, nb "agent" *m.*

Definition: An agent is a **system** that interacts with its **surroundings**. It receives **percepts** through **sensors** and can carry out **actions** through **actuators**.

- Beside its sensors and actuators, an agent is characterized by its **agent function**: The way in which the past and present percepts determine or influence the present and future actions.
- A **goal-oriented agent** is an agent that exhibits the tendency "to achieve a certain state of the world" (Conte 2009, p. 2578). Goal-orientation can emerge by a multitude of mechanisms, including biological evolution. It does not necessarily require the agent to be consciously aware of its goals.
- "**Intelligent agents** are goal-oriented agents using their knowledge to solve problems, including taking decisions and planning actions" (Conte 2009, p. 2578). This requires the agent to have some kind of internal representation of its surroundings, and to store and process information about its surroundings.
- A **knowledge-based agent** is an intelligent agent that uses a knowledge base to store and process its information about its surroundings.
- A **rational agent** is an intelligent agent that exhibits rationality, i.e., a tendency toward optimizing a quantity: The **performance measure** of the agent. As in the case of goal-orientation, this does not necessarily require the agent to be aware of its performance measure.
- "**Goal-directed agents** are intelligent agents that have an internal representation of the goals they [tend to] achieve" (Conte 2009, p. 2578).

Glossary terms: "Agent"



[1]R. Conte, "Rational, goal-oriented agents," in R. A. Meyers (ed.), *Encyclopedia of Complexity and Systems Science*, Springer, **2009**.

Glossary terms: "Rationality"

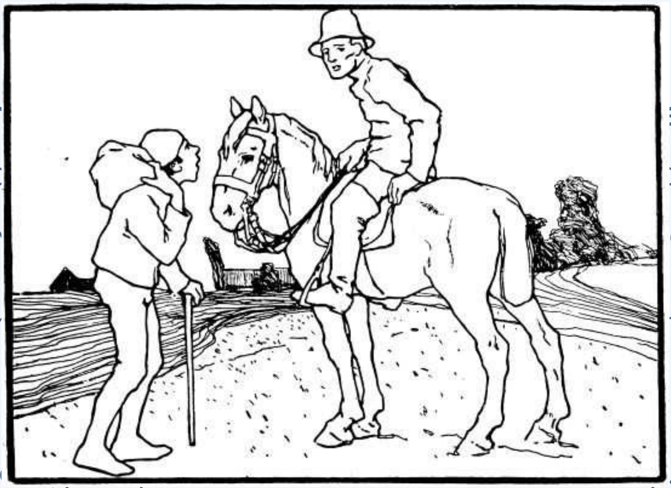
nn "rasjonalitet" *m.*, nb "rasjonalitet" *m.*

Definition: Tendency toward **minimizing a cost function** or **maximizing a performance measure**. In particular, rational preferences, or decisions and choices made by a rational agent, must satisfy the following constraints (Russell & Norvig 2021, *Artificial Intelligence: A Modern Approach*, *p.* 520):

- **Transitivity:** If the agent prefers A over B, and B over C, then the agent also prefers A over C whenever given the choice.
- **Monotonicity:** Assume that the agent prefers A over B. The lotteries (i.e., probability distributions) X and Y both have A and B as their only possible outcomes, where the probability of A is greater in case of the lottery X than in case of the lottery Y. Then the agent prefers X over Y.
- **Continuity:** If the agent prefers A over B, and B over C, then here is exactly one lottery X with A and C as its only possible outcomes such that the agent is indifferent between B and X, i.e., the agent neither prefers B over X nor does the agent prefer X over B. For any other lottery Y with the two possible outcomes A and C, the agent prefers Y over B if the chance of A is greater in case of Y than in case of X; obversely, the agent prefers B over Y if the chance of A is smaller in case of Y than in case of X.

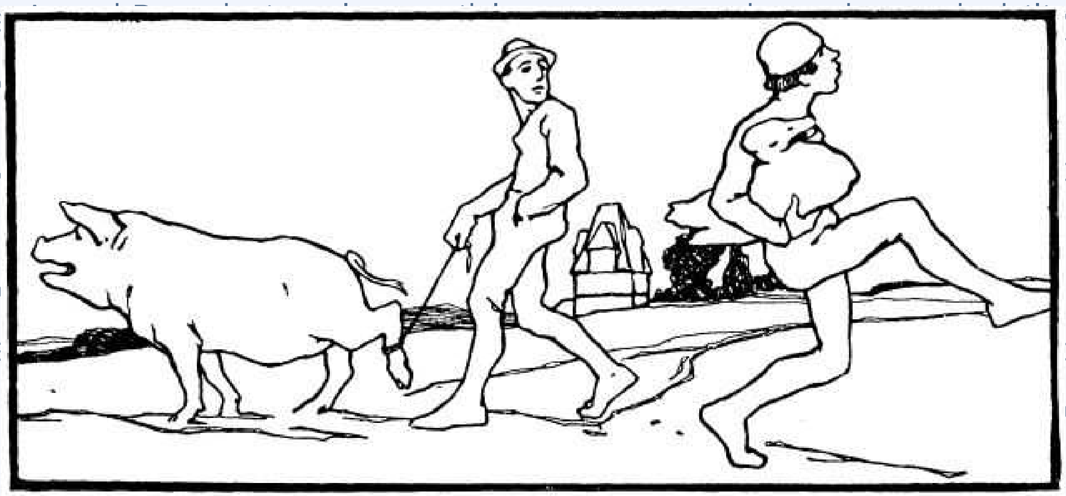
Glossary terms: “Rationality”

nn “rasjonalitet” *m.*, nb “rasjona



Definition: Tendency toward **mi** **a performance measure**. In particular, rational preferences, c onal agent, must satisfy the following constraints (Russell & N *dern Approach*, *p.* 520):

- **Transitivity:** If the agent prefer t also prefers A over C whenever given the choice.
- **Monotonicity:** Assume that th .e., probability distributions) X and Y both have y of A is greater in case of the lottery
- **Continuity:** If the ottery X with A and C as its only possi X, i.e., the agent neither prefers B Y with the two possible outcome er in case of Y than in case of X; er in case of Y than in case of X.



Glossary terms: “Optimization parameter/objective”

nn “optimaliseringsparameter” *m.*,
nb “optimaliseringsparameter” *m.*

Definition: An **optimization parameter** is a quantity over which the decision maker has direct control; a parameter value (or parameterization) is selected in order to obtain the best possible outcome for the optimization objective(s).

- In **multivariate optimization**, there are *multiple optimization parameters*; accordingly, the parameter space is multidimensional.
- If an optimization problem with multiple parameters is formulated adequately, it should be possible to **vary all optimization parameters independently**. If that is not the case and one of the parameters can be expressed as a function of the others, the problem needs to be reformulated, eliminating redundant parameter(s).

nn “optimaliseringsmål” *n.*,
nb “optimaliseringsmål” *n.*

Definition: An **optimization objective** is a quantity that is used to formulate preferences for the outcome of a decision making scenario. In case of a maximization objective, greater values are preferred, and in case of a minimization objective, smaller values are preferred.

- An optimization objective can also be called an optimization criterion or a **key performance indicator** (KPI). If it is a minimization objective, it can also be called **cost**, and if it is a maximization objective, it can also be called **utility**.
- In **multicriteria optimization** (MCO), multiple conflicting optimization objectives are used simultaneously. In this case, there is a multidimensional objective space; the dimension of the objective space is given by the number of optimization objectives.

Glossary terms: “Optimization parameter/objective”

nn “optimaliseringsparameter” *m.*,
nb “optimaliseringsparameter” *m.*

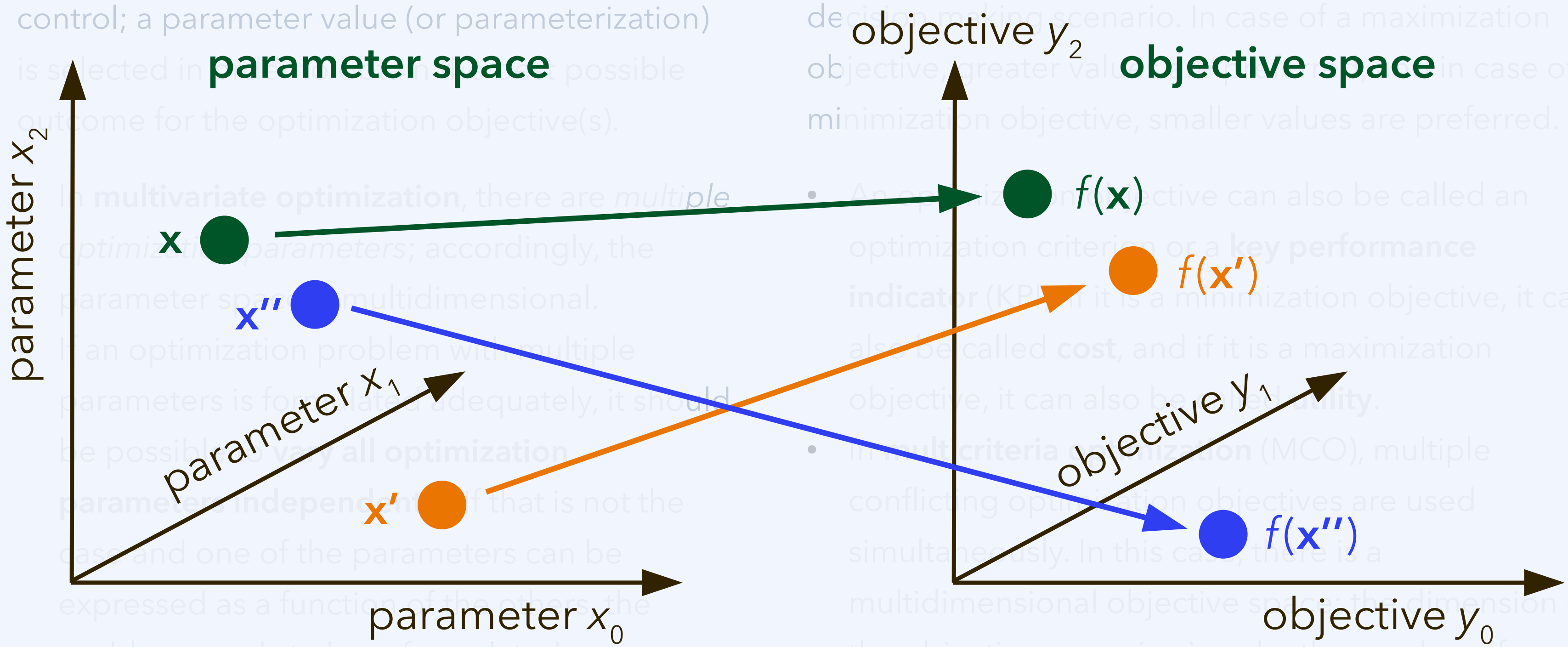
Definition: An **optimization parameter** is a quantity over which the decision maker has direct control; a parameter value (or parameterization) is selected in order to obtain the best possible outcome for the optimization objective(s).

- In **multivariate optimization**, there are *multiple optimization parameters*; accordingly, the parameter space is multidimensional.
- If an optimization problem with multiple parameters is formulated adequately, it should be possible to **vary all optimization parameters independently**. If that is not the case and one of the parameters can be expressed as a function of the others, the problem needs to be reformulated, eliminating redundant parameter(s).

nn “optimaliseringsmål” *n.*,
nb “optimaliseringsmål” *n.*

Definition: An **optimization objective** is a quantity that is used to formulate preferences for the outcome of a decision-making scenario. In case of a maximization objective, greater values are preferred; in case of a minimization objective, smaller values are preferred.

- An optimization objective can also be called an optimization criterion or a **key performance indicator** (KPI). If it is a minimization objective, it can also be called **cost**, and if it is a maximization objective, it can also be called **utility**.
- In **multicriteria optimization** (MCO), multiple conflicting optimization objectives are used simultaneously. In this case, there is a multidimensional objective space; the dimension of the objective space is given by the number of optimization objectives.



Glossary terms: “Pareto optimality”

nn “Pareto-optimalitet” *m.*, nb “Pareto-optimalitet” *m.*

Definition: Within the framework of *multicriteria optimization* (MCO), a *point in objective space* is *Pareto optimal if it is accessible* and *no other accessible point in objective space dominates it*.

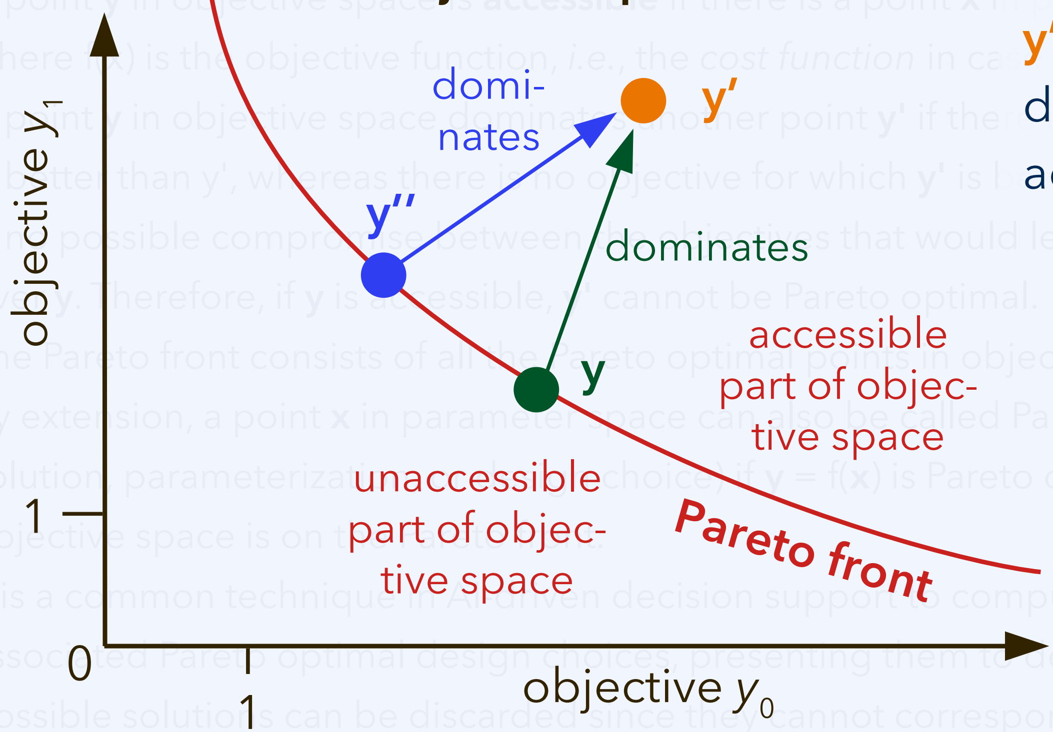
- A point **y** in objective space is **accessible** if there is a point **x** in parameter space such that f(**x**) = **y**, where f(x) is the objective function, *i.e.*, the *cost function* in case of minimization objectives.
- A point **y** in objective space dominates another point **y'** if there is at least one objective for which **y** is better than y', whereas there is no objective for which **y'** is better than **y**. If that is the case, there is no possible compromise between the objectives that would lead a rational agent to prefer y' over **y**. Therefore, if **y** is accessible, **y'** cannot be Pareto optimal.
- The Pareto front consists of all the Pareto optimal points in objective space.
- By extension, a point **x** in parameter space can also be called Pareto optimal (*e.g.*, a Pareto optimal solution, parameterization, or design choice) if **y** = f(**x**) is Pareto optimal, *i.e.*, if the point **y** in objective space is on the Pareto front.
- It is a common technique in AI-driven decision support to compute the Pareto front and the associated Pareto optimal design choices, presenting them to decision makers. All the other possible solutions can be discarded since they cannot correspond to a rational compromise between the objectives.

Glossary terms: “Pareto optimality”

nn “Pareto-optimalitet” *m.*, nb “Pareto-optimalitet” *m.*

Definition: Within the framework of *multicriteria optimization* (MCO), a *point in objective space* is *Pareto optimal if it is accessible* and *no other accessible point in objective space dominates it.*

- A point **y** in objective space is **accessible** if there is a point **x** in parameter space such that f(**x**) = **y**, where f(**x**) is the objective function, *i.e.*, the *cost function* in case of minimization objectives.
- A point **y** in objective space dominates another point **y**' if there is at least one objective for which **y** is better than y', whereas there is no objective for which **y**' is better than **y**. In the case, there is no possible compromise between the objectives that would lead a rational agent to prefer y' over **y**. Therefore, if **y** is accessible, **y**' cannot be Pareto optimal.
- The Pareto front consists of all the Pareto optimal points in objective space.
- By extension, a point **x** in parameter space can also be called Pareto optimal (*e.g.*, a Pareto optimal solution, parameterization, or design choice) if **y** = f(**x**) is Pareto optimal, *i.e.*, if the point **y** in objective space is on the Pareto front.
- It is a common technique in AI-driven decision support to compute the Pareto front and the associated Pareto optimal design choices, presenting them to decision makers. All the other possible solutions can be discarded since they cannot correspond to a rational compromise between the objectives.



y and **y′′** are rational compromises between the two objectives.

y′ is not, because it is dominated by other accessible points.

Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

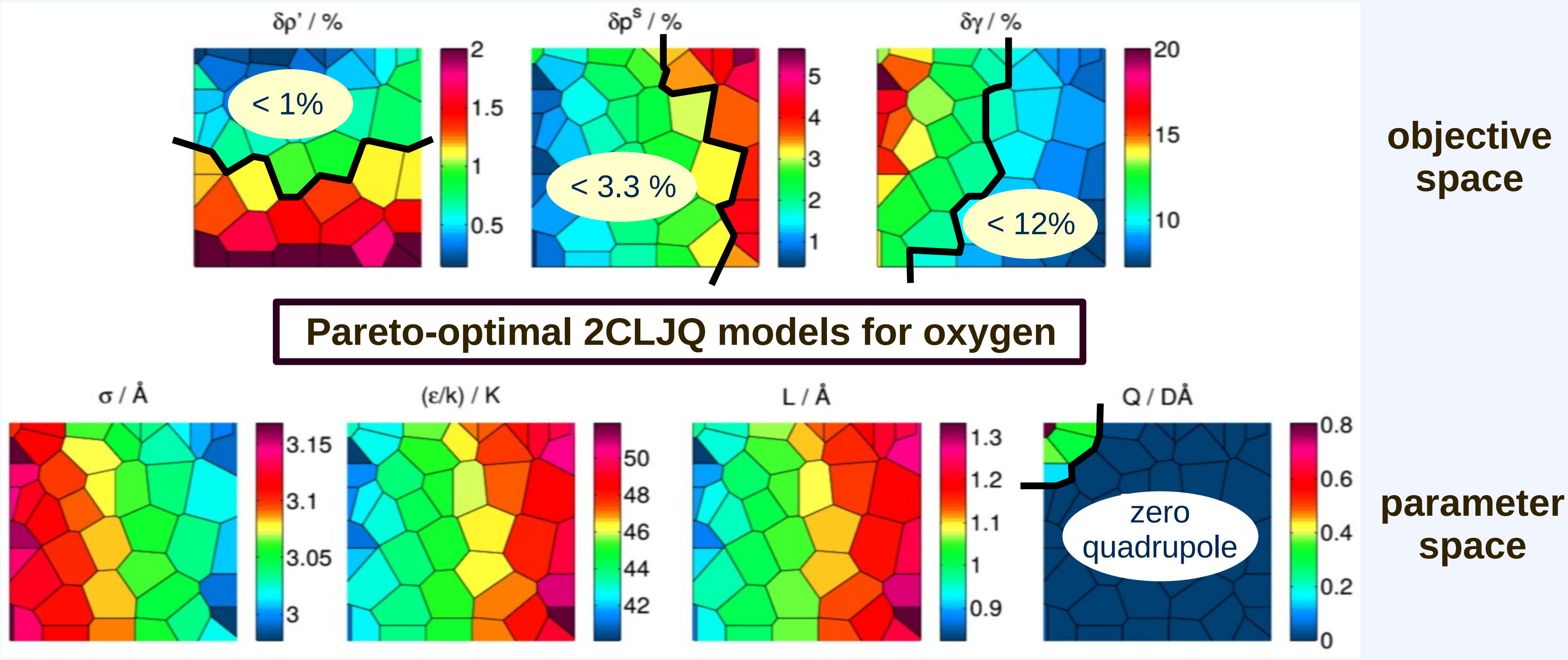
Digitalisering på Ås

5 Multidimensionality

5.4 Pareto front visualization

Example (molecular model parameterization)

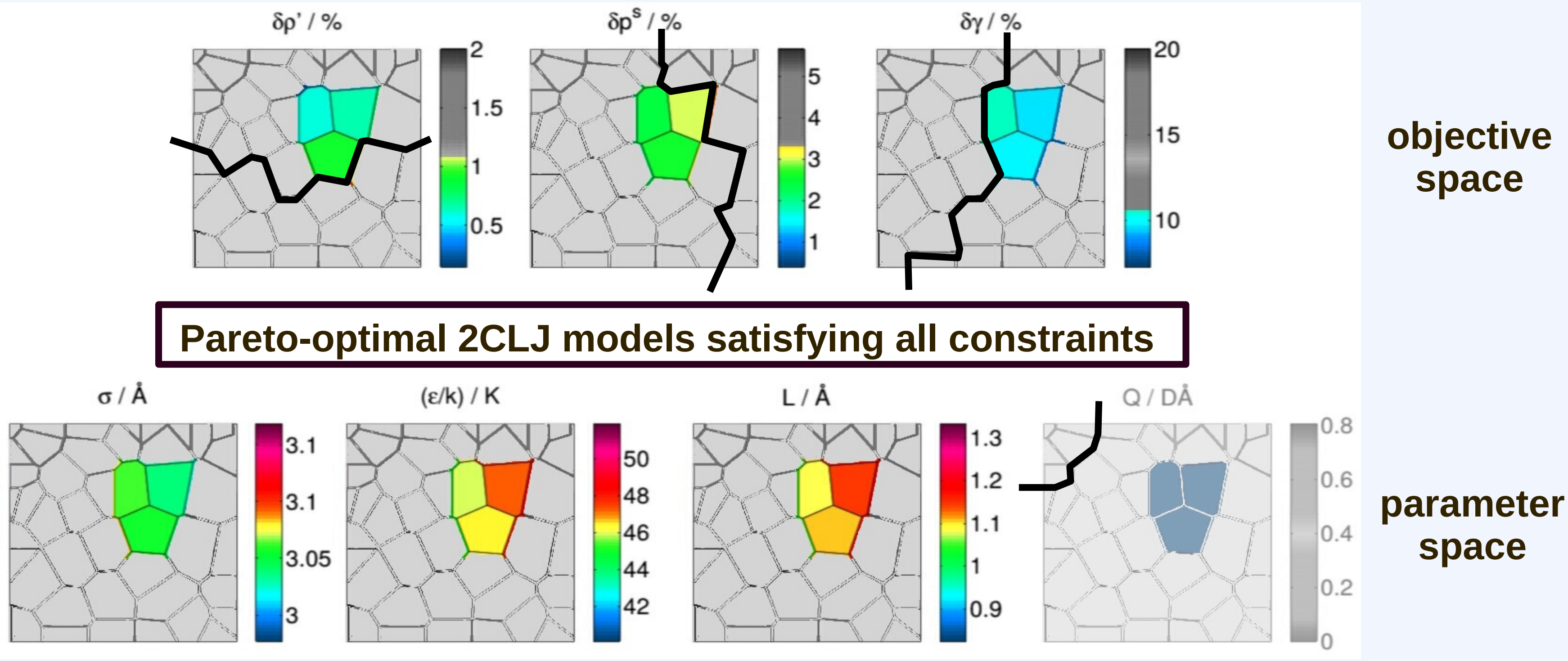
Self-organized patch plots[1] visualizing the Pareto front and the Pareto-optimal models:



K. Stöbener, P. Klein, M. Horsch, K. Küfer, H. Hasse, *Fluid Phase Equilib.* 411, 33 – 42, **2016**.

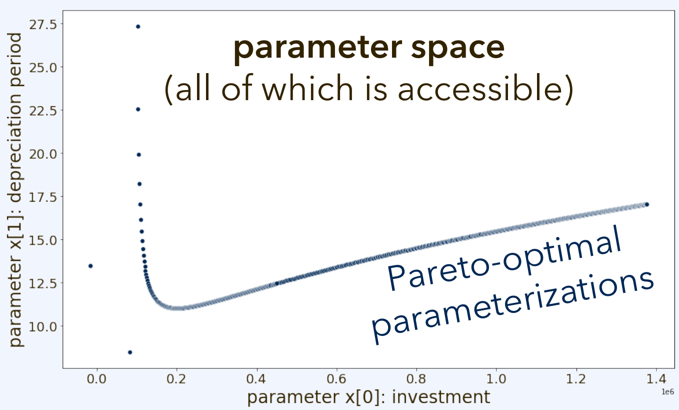
Example (molecular model parameterization)

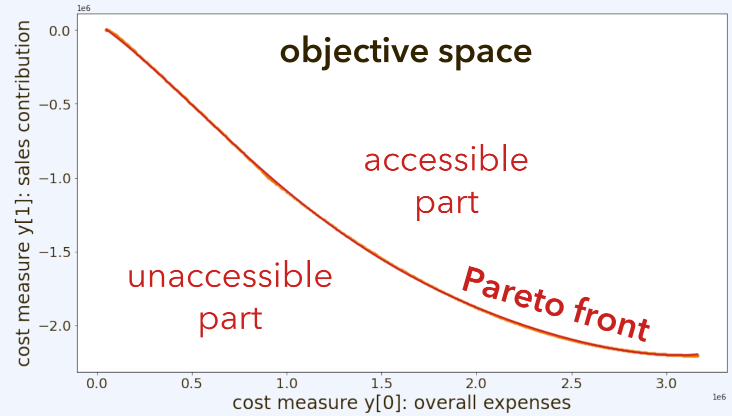
Self-organized patch plots[1] visualizing the Pareto front and the Pareto-optimal models:



K. Stöbener, P. Klein, M. Horsch, K. Küfer, H. Hasse, *Fluid Phase Equilib.* 411, 33 – 42, **2016**.

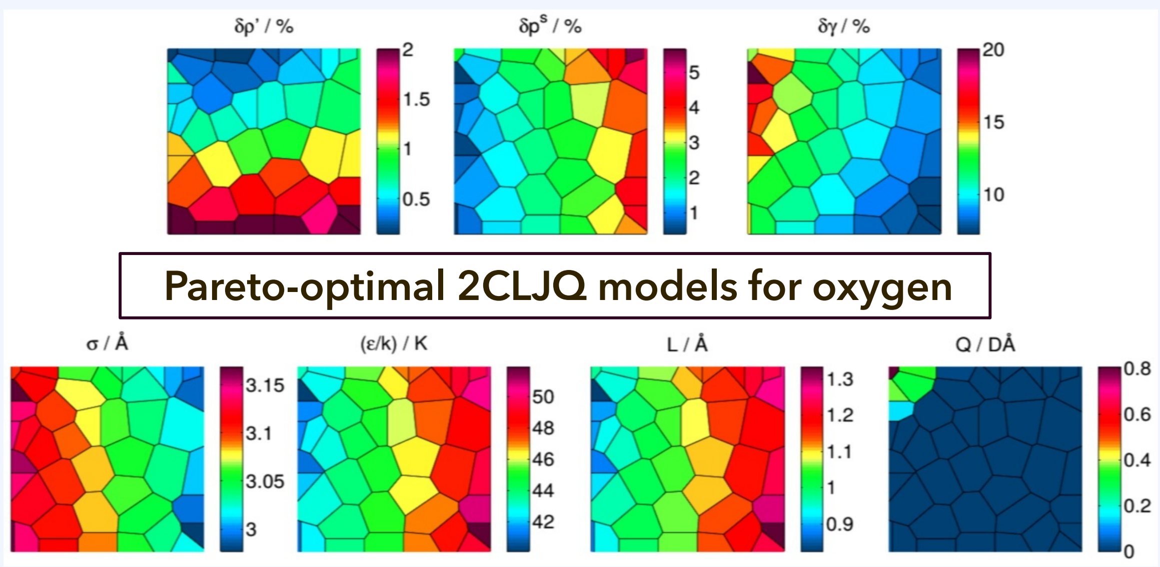
Two methods for visualizing an MCO problem





objective space

parameter space



K. Stöbener, P. Klein, M. Horsch, K. Küfer, H. Hasse, *Fluid Phase Equilib.* 411, 33 – 42, **2016**.

Example notebook for Pareto front visualization

```
def cost_function(x, debug_output):
    if x[0] < 0 or x[1] < 1 or x[0] < salary*x[1]:
        return [math.inf, math.inf]
    expenses = x[0]
    acquired_equipment = (x[0] - salary*x[1]) / unit_cost
    upgraded_units = min(num_units, acquired_equipment, x[1]/fte_per_unit)

    y = [expenses, num_units - upgraded_units]
    return y
```

- In cell [1], replace the body of cost_function(**x**, debug_output).
- The constant coefficients need to be included.
- It is advisable to implement a **penalty for values outside the specified parameter space**, since scipy.optimize will not be aware of constraints.

Example notebook for Pareto front visualization

```
def random_parameters():
    max_expenses = num_units * (unit_cost + salary*fte_per_unit)
    expenses = random.uniform(0, max_expenses)
    total_labour_cost = random.uniform(0, expenses)
    return [expenses, total_labour_cost/salary]

objective_scale = [180000, 600]
sigma = 2
```

- In cell [1], replace the body of cost_function(**x**, debug_output).
- In cell [2], edit random_parameters() such that it returns a random point in parameter space, and **objective_scale** such that objective_scale[i] is of the order of variations expected in the outcome for objective y[i]. Increase/decrease sigma if you want weights to vary more/less.
- In cells [4] and [6], adjust local and global optimizer settings.

Example notebook for Pareto front visualization

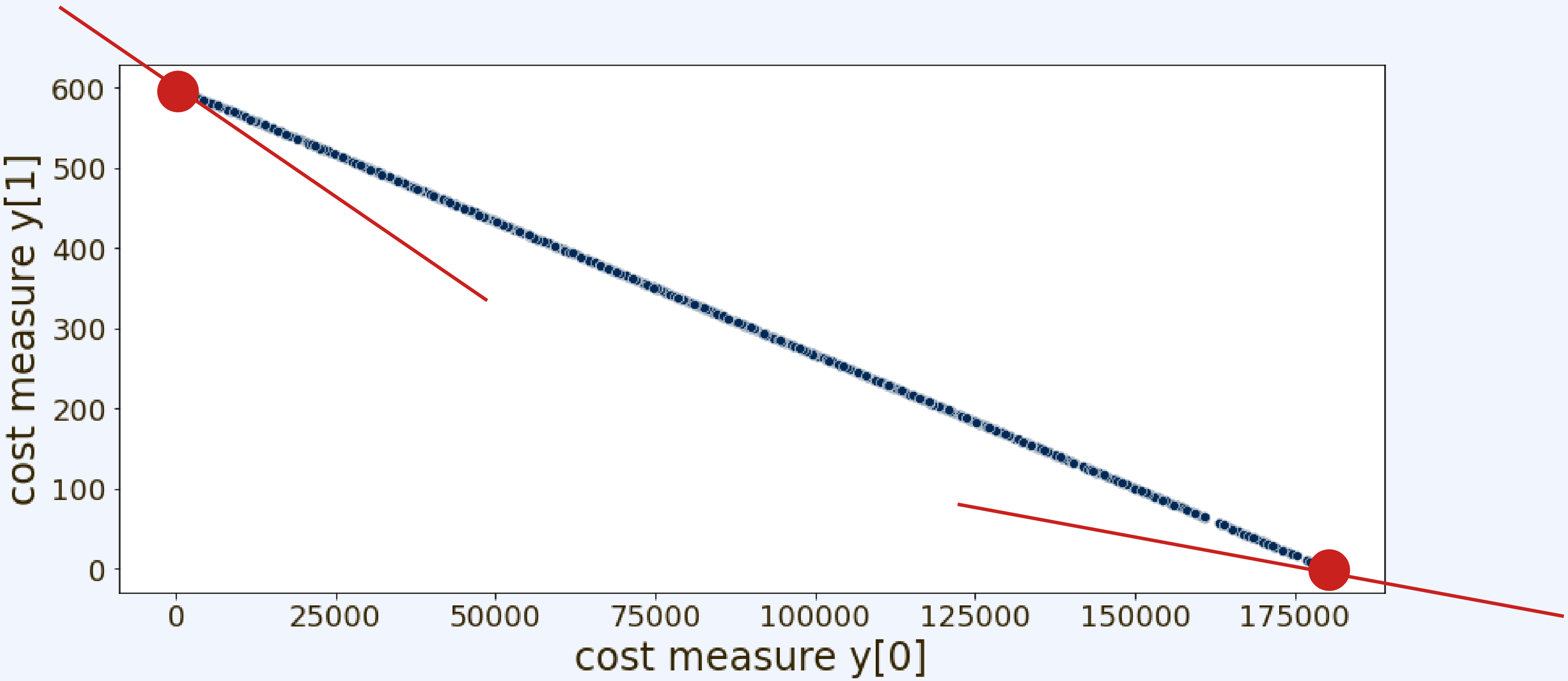
In **cell [6]**, adjust:

- number of parameters *m* and number of objectives *n*.
- number of points to be determined by linear combinations and by hyperboxing, respectively; their sum should be a square number.
  - linear combinations only work for a convex Pareto front: It can happen that this part needs to be removed; in this case, the lists **objective_space_lower** and **objective_space_upper** need to be initialized appropriately.
- local and global optimizer settings.

In **cell [8]**, select the axes to be shown for the 2D projection (*e.g.*, 0 and 1).

```
sbn.scatterplot(x=pareto_optimal_compromises[0], \
                y=pareto_optimal_compromises[1], color="#002855")
```

Example notebook for Pareto front visualization



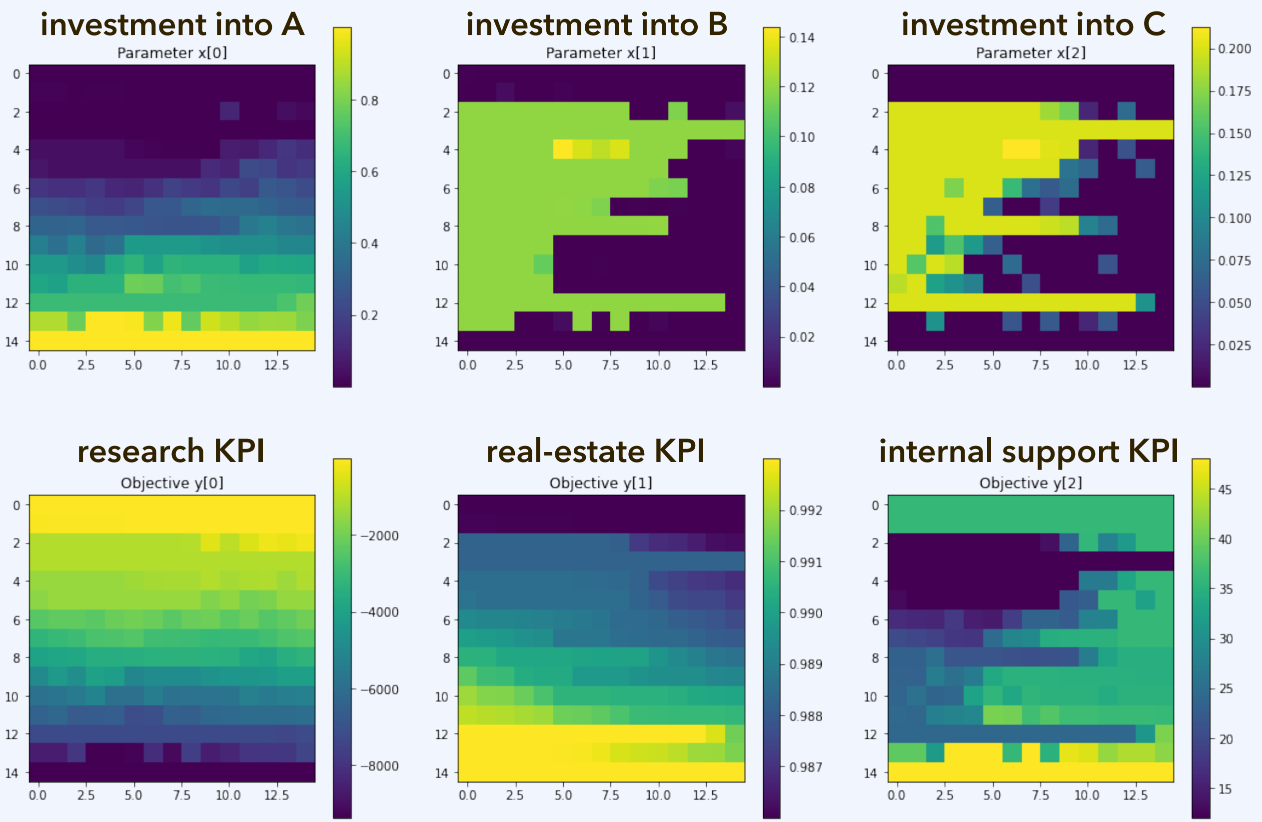
In **cell [8]**, select the axes to be shown for the 2D projection (*e.g.*, 0 and 1).

```
sbn.scatterplot(x=pareto_optimal_compromises[0], \
                y=pareto_optimal_compromises[1], color="#002855")
```

Example notebook for Pareto front visualization

In **cell [10]**, set square_size to the square root of the number of determined Pareto optimal solutions. Pass indices of the criteria for ordering (*e.g.*, 0 and 1):

idx_order = arrange_indices(square_size, n, pareto_optimal_compromises, 0, 1)



Noregs miljø- og biovitskaplege universitet

Institutt for datavitskap

Digitalisering på Ås

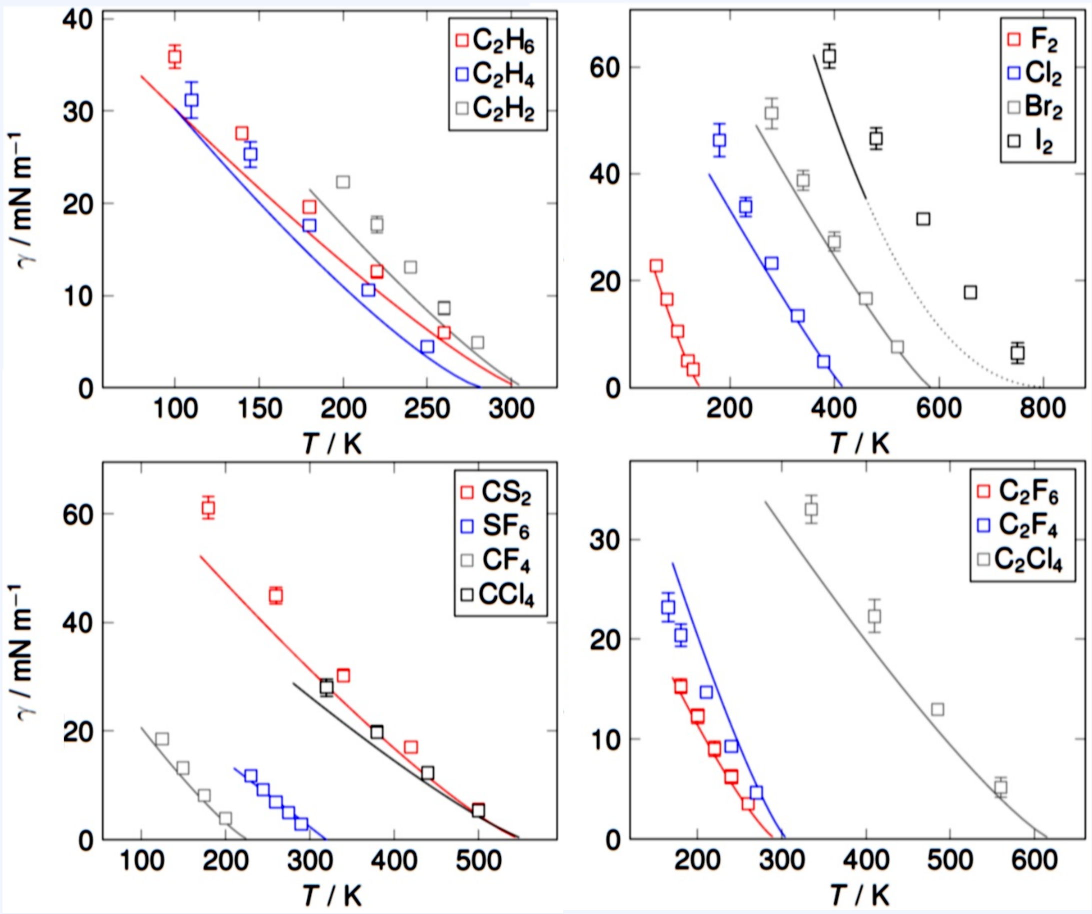
5 Multidimensionality

5.4 Pareto front visualization

5.5 MCO in modelling

Background of the model optimization problem

2CLJQ: Two LJ centers + quadrupole[1]



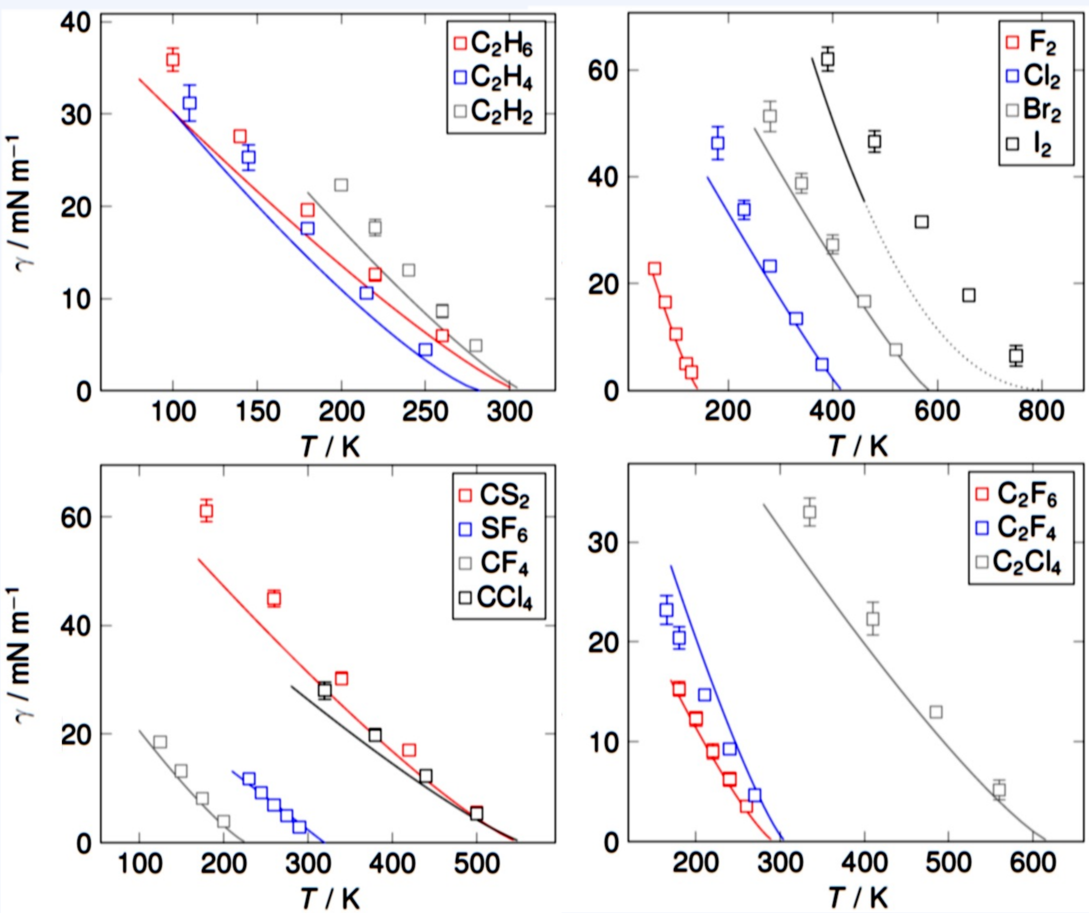
Fit to bulk properties

About 20 % overestimation of the surface tension

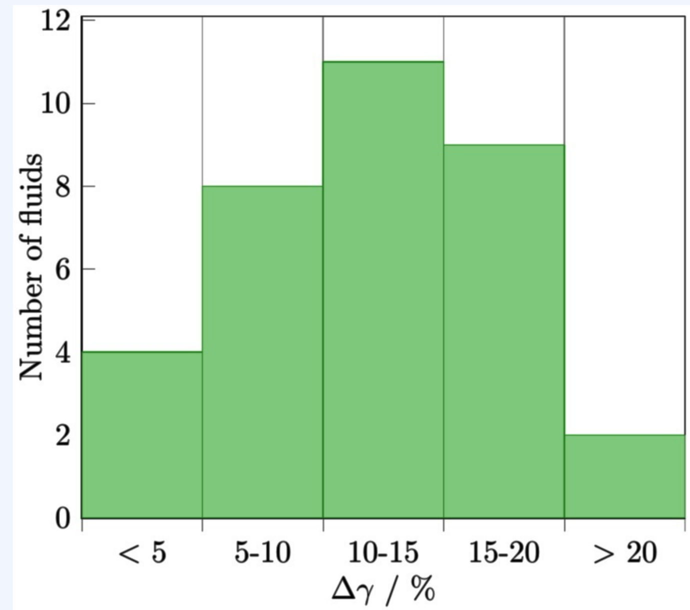
[1]S. Werth, K. Stöbener, P. Klein, K.-H. Küfer, M. Horsch, H. Hasse, *Chem. Eng. Sci.* 121, 110–117, **2015**.

Background of the model optimization problem

2CLJQ: Two LJ centers + quadrupole[1]



2CLJD: Two LJ + dipole[2]



[1]S. Werth, K. Stöbener, P. Klein, K.-H. Küfer, M. Horsch, H. Hasse, *Chem. Eng. Sci.* 121, 110–117, **2015**;

[2]S. Werth, M. Horsch, H. Hasse, *J. Chem. Phys.* 144, 054702, **2016**.

Background of the model optimization problem

Non-polar: 1CLJ

Neon (Ne)
Argon (Ar)
Krypton (Kr)
Xenon (Xe)
Methane (CH_4)

Quadrupolar: 2CLJQ

+20 %

Fluorine (F_2)
Chlorine (Cl_2)
Bromine (Br_2)
Iodine (I_2)
Nitrogen (N_2)
Oxygen (O_2)
Carbon dioxide (CO_2)
Carbon sulfide (CS_2)
Ethane (C_2H_6)
Ethylene (C_2H_4)
Acetylene (C_2H_2)
R116 (C_2F_6)
R1114 (C_2F_4)
R1110 (C_2Cl_4)
Propadiene (CH_2=C=CH_2)
Propyne (CH_3-C≡CH)
Propylene (CH_3-CH=CH_2)
R846 (SF_6)
R14 (CF_4)
R10 (CCl_4)
R113 ($CFCl_2$-CF_2Cl)
R114 (CF_2Cl-CF_2Cl)
R115 (CF_3-CF_2Cl)
R134 (CHF_2-CHF_2)
R150B2 (CH_2Br-CH_2Br)
R114B2 ($CBrF_2$-$CBrF_2$)
R1120 (CHCl=CCl_2)

Dipolar: 2CLJD

+12 %

Carbon monoxide (CO)
R11 ($CFCl_3$)
R12 (CF_2Cl_2)
R13 (CF_3Cl)
R13B1 ($CBrF_3$)
R22 (CHF_2Cl)
R23 (CHF_3)
R41 (CH_3F)
R123 ($CHCl_2$-CF_3)
R124 (CHFCl-CF_3)
R125 (CHF_2-CF_3)
R134a (CH_2F-CF_3)
R141b (CH_3-$CFCl_2$)
R142b (CH_3-CF_2Cl)
R143a (CH_3-CF_3)
R152a (CH_3-CHF_2)
R40 (CH_3Cl)
R40B1 (CH_3Br)
CH_3I
R30B1 (CH_2BrCl)
R20 ($CHCl_3$)
R20B3 ($CHBr_3$)
R21 ($CHFCl_2$)
R32 (CH_2F_2)
R30 (CH_2Cl_2)
R30B2 (CH_2Br_2)
CH_2I_2
R12B2 (CBr_2F_2)
R12B1 ($CBrClF_2$)
R10B1 ($CBrCl_3$)
R161 (CH_2F-CH_3)
R150a ($CHCl_2$-CH_3)
R140 ($CHCl_2$-CH_2Cl)
R140a (CCl_3-CH_3)
R130a (CH_2Cl-CCl_3)
R160B1 (CH_2Br-CH_3)
R150B2 ($CHBr_2$-CH_3)
R131b (CH_2F-CCl_3)
R123B1 (CHClBr-CF_3)
R112a (CCl_3-CF_2Cl)
R1141 (CHF=CH_2)
R1132a (CF_2=CH_2)
R1140 (CHCl=CH_2)
R1122 (CHCl=CF_2)
R1113 (CFCl=CF_2)
R1113B1 (CFBr=CF_2)

Literature models by J. Stoll, H. Hasse, J. Vrabec *et al.*, **2001 – 2016**

Multicentric United Atom Models

Isobutane (C_4H_{10})
Cyclohexane (C_6H_{12})
Methanol (CH_3OH)
Ethanol (C_2H_5OH)
Formaldehyde (CH_2=O)
Dimethyl ether (CH_3-O-CH_3)
Acetone (C_3H_6O)
Ammonia (NH_3)
Methylamine (NH_2-CH_3)
Dimethylamine (CH_3-NH-CH_3)
R227ea (CF_3-CHF-CF_3)
Sulfur dioxide (SO_2)
Ethylene oxide (C_2H_4O)
Dimethyl sulfide (CH_3-S-CH_3)
Hydrogen cyanide (HCN)
Acetonitrile (NC_2H_3)
Thiophene (SC_4H_4)
Nitromethane (CH_3NO_2)
Phosgene ($COCl_2$)
Benzene (C_6H_6)
Toluene (C_7H_8)
Chlorobenzene (C_6H_5Cl)
Dichlorobenzene ($C_6H_4Cl_2$)
Cyclohexanol (C_6H_{11}OH)
Cyclohexanone (C_6H_{10}O)
Cyanogen (C_2N_2)
Cyanogen chloride (CClN)
Formic acid (CH_2O_2)
Ethylene glycol ($C_2H_6O_2$)
TIP4P/2012 water (H_2O)
Hydrazine (N_2H_4)
Monomethylhydrazine (CH_6N_2)
Dimethylhydrazine ($C_2H_8N_2$)
Perfluorobutane (C_4F_{10})
Ethyl acetate ($C_4H_8O_2$)
HMDSO ($C_6H_{12}OSi_2$)
D4 ($C_8H_{24}O_4Si_4$)

+22 %

MCO problem specification for 2CLJQ models

a model parameters

(here, a = 4)

- LJ size parameter σ
- LJ energy parameter ε
- Model elongation L
- Quadrupole moment Q

Dimension of Pareto set $d \leq a$.

b optimization criteria

(here, b = 3)

- Saturated liquid density ρ'
- Saturated vapor pressure p^s
- Vapor-liquid surface tension γ

Dimension of Pareto set $d \leq b - 1$.

$d = \min(a, b - 1)$. (here, d = 2)

[1]M. Bortz *et al.*, *Comput. Chem. Eng.* 60, 354, **2014**; [2]Stöbener *et al.*, *Fluid Phase Equilib.* 411, 33, **2016**.

MCO problem specification for 2CLJQ models

Multicriteria optimization problem

Simultaneously minimized objective functions f_ξ with $\xi \in \{\rho', p^s, \gamma\}$ given by

$$f_\xi = \langle \delta \xi^2 \rangle_{0.55 T_c^{exp} < T < 0.95 T_c^{exp}} = \lim_{N \to \infty} \frac{1}{N+1} \sum_{i=0}^{N} \left(1 - \frac{\xi^{sim}(T)}{\xi^{exp}(T)} \right)^2_{T/T_c = 0.55 + 0.4i/N} \qquad \text{(here: } N = 9\text{).}$$

Sandwiching

Alternating construction of inner (reachable) and outer (unreachable) approximations, in regions where the Pareto set is locally convex.

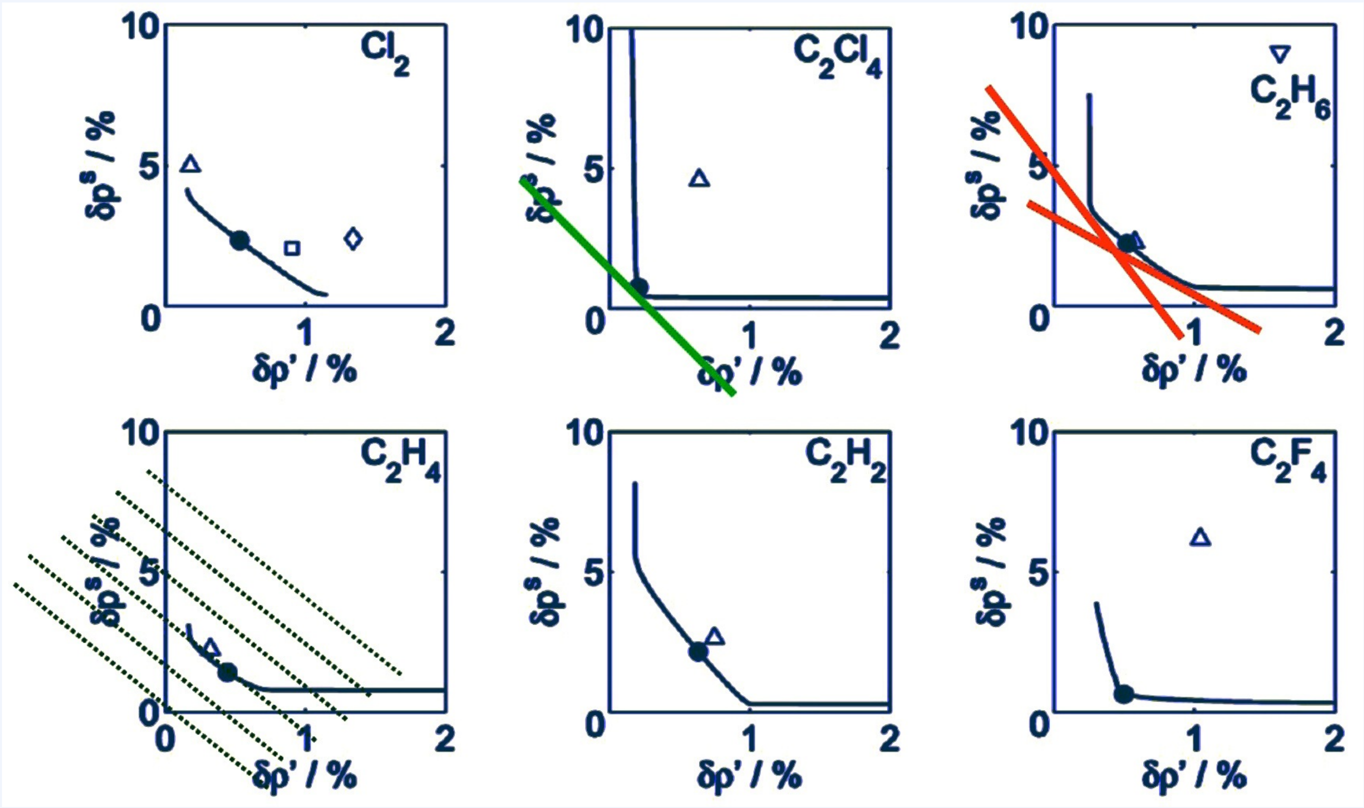
Hyperboxing

In non-convex regions (hyperboxes), Pascoletti-Serafini scalarization is used to formulate an appropriately constrained single-criterion problem.

[1]M. Bortz *et al.*, *Comput. Chem. Eng.* 60, 354, **2014**; [2]Stöbener *et al.*, *Fluid Phase Equilib.* 411, 33, **2016**.

The Pareto knee

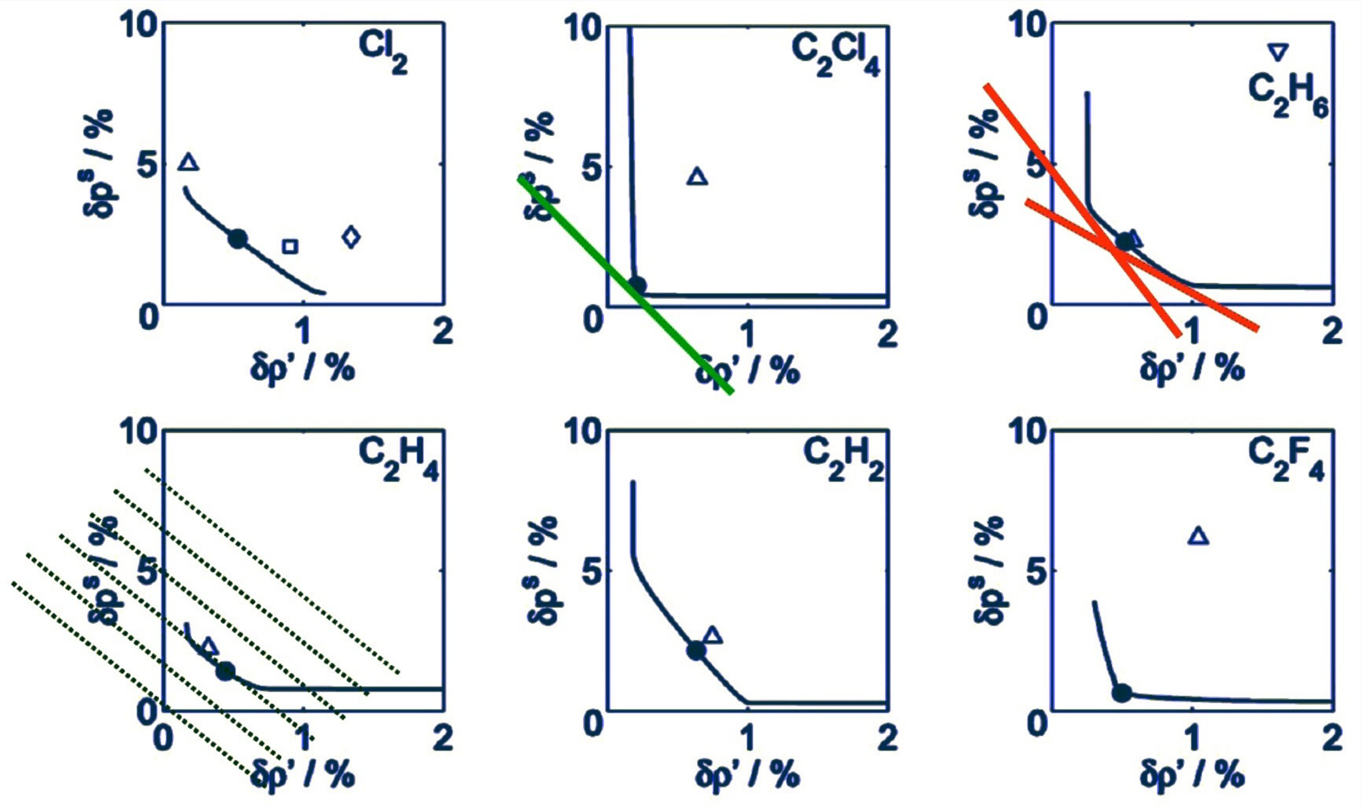
(Example: Two-criteria optimization of molecular models.)



A **Pareto knee** is a highly curved region on the Pareto front.

The Pareto knee

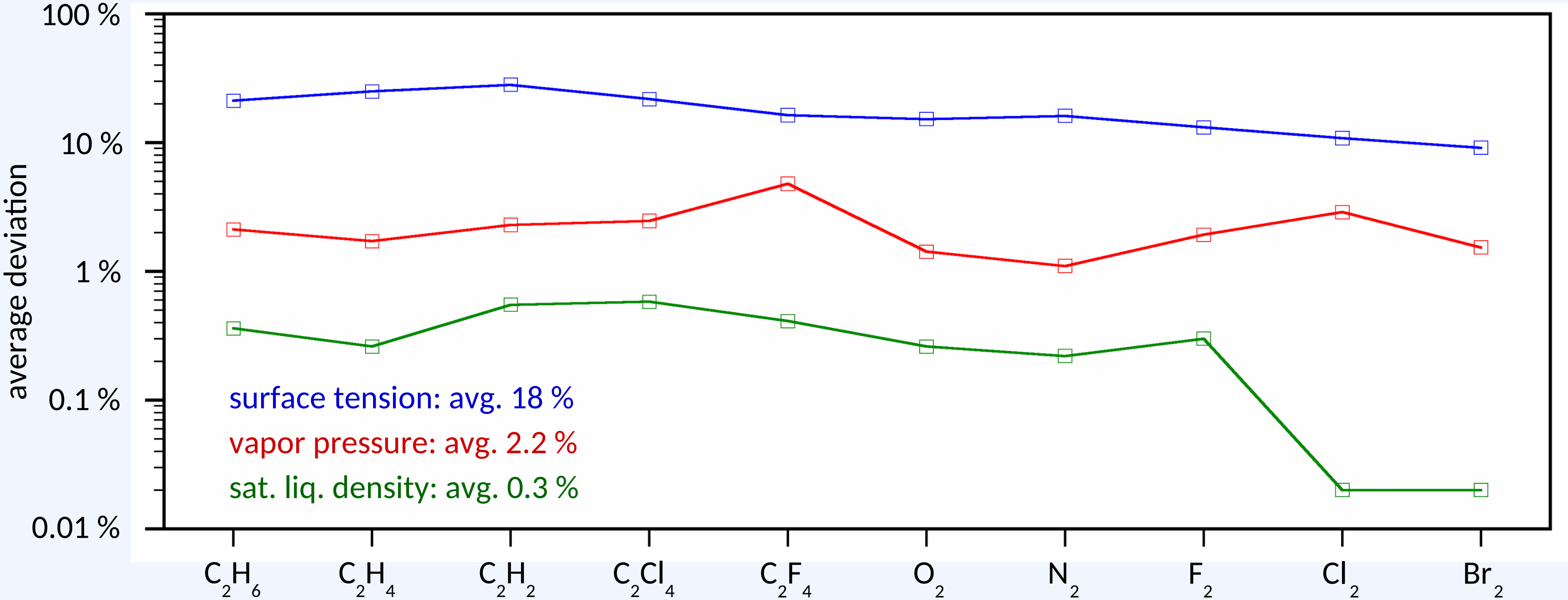
The viability of models close to a **Pareto knee** is comparably resilient even when priorities shift. Example: Two-criteria optimization of molecular models.



A **Pareto knee** is a highly curved region on the Pareto front.

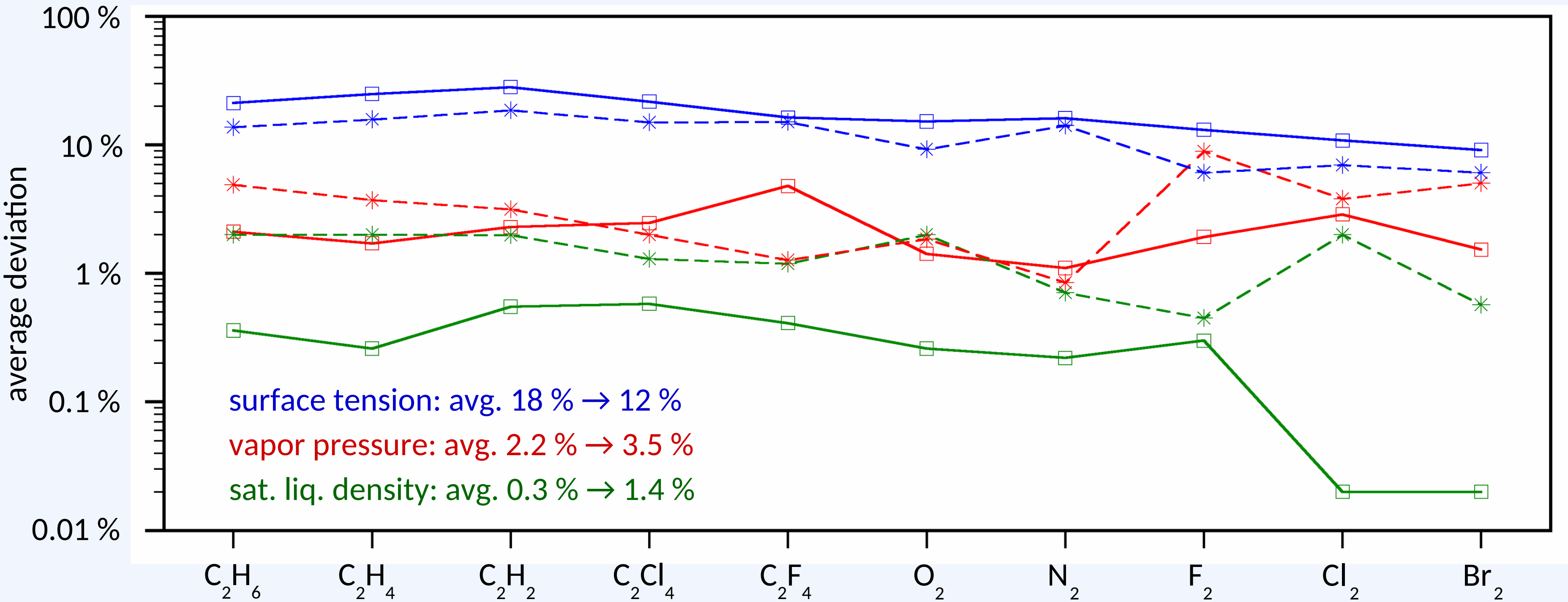
In general, a systematic **exploration of the Pareto front** is needed to find such regions.

Pre-existing models' performance



[1]J. Vrabec, J. Stoll, H. Hasse, *J. Phys. Chem. B* 105(48), 12126–12133, **2001**;
[2]S. Werth, K. Stöbener, P. Klein, K.-H. Küfer, M. Horsch, H. Hasse, *Chem. Eng. Sci.* 121, 110–117, **2015**.

Pre-existing models vs. the Pareto knee



[1]J. Vrabec, J. Stoll, H. Hasse, *J. Phys. Chem. B* 105(48), 12126–12133, **2001**;

[2]S. Werth, K. Stöbener, P. Klein, K.-H. Küfer, M. Horsch, H. Hasse, *Chem. Eng. Sci.* 121, 110–117, **2015**;

[3]K. Stöbener, P. Klein, M. Horsch, K.-H. Küfer, H. Hasse, *Fluid Phase Equilib.* 411, 33–42, **2016**.

Modelling paradigm shift due to MCO

Norwegian University of Life Sciences

The art of molecular modelling

An **expert modelling artist** designs and publishes

- a single optimized model for a particular fluid,
- according to his choice of criteria (often unknown to the public),
- users are passive, they have to live with the artists' decision.

The science of molecular modelling

For well-characterized model classes and **multiple optimization criteria**,

- the dependence of thermodynamic properties on the model parameters is determined and correlated,
- the deviation between model properties and real fluid behaviour is characterized, and the Pareto set is published,
- users can design their own tailored model **with minimal effort**.

Norges miljø- og biovitenskapelige universitet

Institutt for datavitenskap

Digitalisering på Ås

DAT121
Introduction to data science

5 Multidimensionality

5.4 Pareto front visualization

5.5 MCO in modelling (& related research)