CE4 – 06A
Advanced Process Optimisation

Lecture 3

Objectives for today

- Introduction to Computer-Aided Molecular Design Problems (CAMD)
- Understand how to formulate CAMD problems
  - Types of CAMD problems
  - Models for CAMD
  - Molecular feasibility & molecular complexity

Computer-Aided Molecular Design (CAMD)

"Given a desired range for a set of properties and performance criteria,

design the compound that performs best,

while possessing properties within the acceptable range"

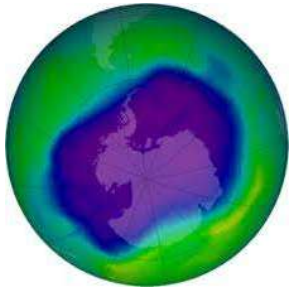
Vaidyanathan and El-Halwagi

EXAMPLES OF CAMD PROBLEMS

New classes of refrigerants

- The 1987 Montreal protocol has led to a phasing out of CFCs
- Chlorine concentration has been decreasing since 1998
- In 2016[a], it was reported that the hole in the Antartic ozone layer is starting to close, a process which should be completed by 2050

- This reversal has been brought about by the discovery of alternative refrigerants: hydrochlorofluorocarbons (HCFCs) and hydrofluorocarbons (HFCs)…
- … but HFCs have a high global warming potential need to be phased out too (Kigali amendment).



NASA - http://www.nasa.gov/vision/earth/lookingatearth/ozone_record.html

[a]Solomon et al, *Science,* 2016, 353:269-274, DOI: 10.1126/science.aae0061

Designing a new polymer

Maranas (1996)

We are seeking a replacement polymer for an existing product. The key properties of the polymer are

- Density D_S with target $D_S^0 = 1.50$ g cm^{-3}
- Water absorption W with target $W^0 = 0.005$ g H_2O / g polymer
- Glass transition temperature T_g with target $T_g^0 = 383$ K

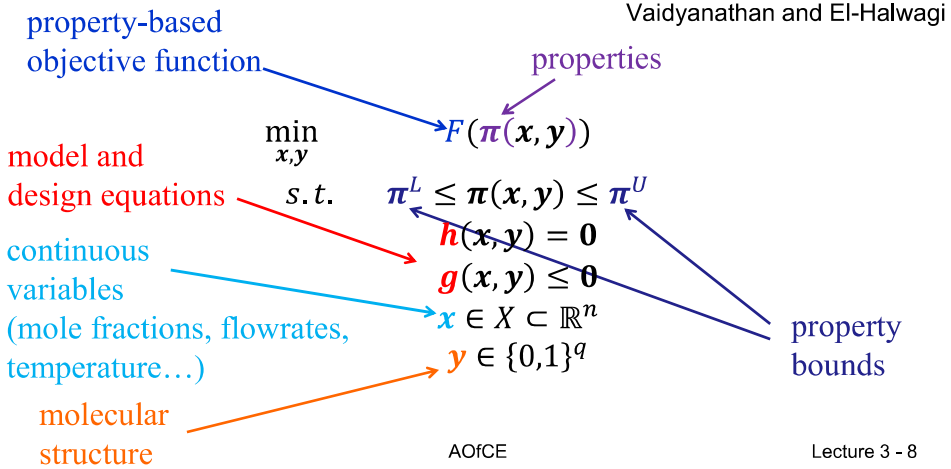
$$\min \left(\frac{D_S - D_S^0}{D_S^0}\right)^2 + \left(\frac{W - W^0}{W^0}\right)^2 + \left(\frac{T_g - T_g^0}{T_g^0}\right)^2$$

- What else do we need?
- How might we find the best molecules?

Computer-Aided Molecular Design (CAMD)

"Given a desired range for a set of properties and performance criteria, design the compound that performs best, while possessing properties within the acceptable range"

Vaidyanathan and El-Halwagi



The design space

- In CAMD, a key step is to define the design space:
  - Specify a set of atom groups:
    - e.g., three atom types: CH_3, CH_2, OH
    - can build all linear alkanes and primary alkanols from these
    - the set of groups is $G = \{CH_3, CH_2, OH\}$
  - Variables represent the number of groups of each type:
    - n_{CH3}, n_{CH2}, n_{OH}
    - can be treated as integer
    - or can be treated as continuous with associated binary variables
      - E.g., $n_{CH3} = y_{1,CH3} + 2y_{2,CH3} + 4y_{3,CH3} + 8y_{4,CH3}$ …
  - In addition we have a set of continuous variables (e.g., composition, temperature, etc)

The objective function

- $F(\pi(\boldsymbol{x}, \boldsymbol{y}))$ is a function of the properties
- It can be focused on property matching
  - Given some property targets, π^0, find a compound that most closely matches these targets:

$$F\left(\pi(\boldsymbol{x}, \boldsymbol{y})\right) = \sum_{i=1}^{P} \left(\frac{\pi_i - \pi_i^0}{\pi_i^0} \right)^2$$

- It can be focused on property optimisation
  - For instance, minimise heat capacity of refrigerant

$$F\left(\pi(\boldsymbol{x}, \boldsymbol{y})\right) = C_{pl}$$

- It can be much broader…

On the utility of CAMD

- CAMD should not be used to identify "the best molecule"
  - CAMD can help to focus resources for more detailed investigation
  - It is a **synthesis** tool

A key question

How do we link performance to the molecular structure of the possible compounds?

$$\boldsymbol{\pi}^L \leq \boldsymbol{\pi}(\boldsymbol{x}, \boldsymbol{y}) \leq \boldsymbol{\pi}^U$$

LINKING MOLECULAR STRUCTURE TO PROPERTIES

What properties are we looking for?

- Consider the following properties:
  - Molecular weight
  - Saturated vapour pressure
  - Molar volume
  - Solubility
- What additional information would you need to be able to find these properties in a database?
- Can you classify the properties based on the information needed?

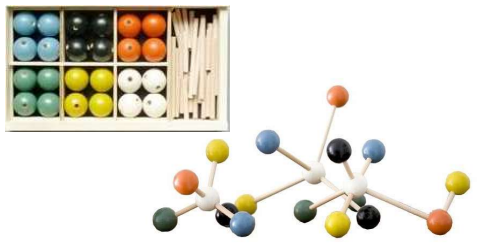
Two approaches to link molecules & properties

1. Use a database of compounds
2. Use a structure-property relations

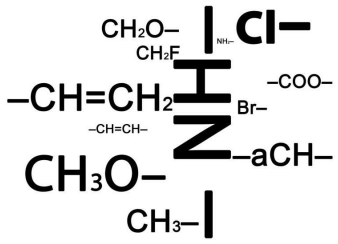
What are the benefits and drawbacks of using a database of properties?

2. Use a structure-property relation

- Start from building blocks: atom groups



Images: http://www.notanothershop.com/ferm.html



- How many possibilities are there?

A truly combinatorial problem

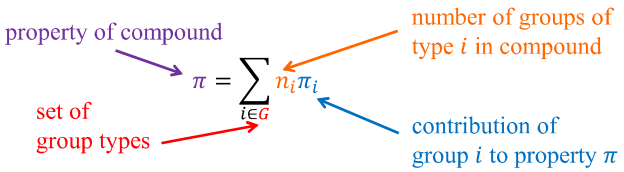
- Consider designing a molecule with between K_{min} and K_{max} groups, chosen from N different atom groups
- There are $\sum_{k=K_{mi}}^{K_{max}} \frac{(N+K-1)!}{K!(N-1)!}$ combinations of the groups *assuming the permutations of groups are indistinguishable*

K_{max}	N	distinct designs
5	5	251
5	10	3,002
5	20	53,129
5	30	324,631
5	40	1,221,758
10	5	3,002
10	10	184,755
10	20	30,045,014
10	30	847,660,527
10	40	10,272,278,169
20	5	53,129
20	10	30,045,014
20	20	137,846,528,819
20	30	47,129,212,243,959
20	40	4,191,844,505,805,494

Table 6 from Maranas (1996)

Group contribution (GC) methods

- Basic principle:
  - Functional/atom groups contribute in a additive and transferable way to properties



  - This is true for molar mass M, e.g., $M_{propanol}$:
    - Propanol is 1xCH_3, 2xCH_2, 1xOH

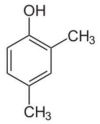
$$M_{propanol} = M_{CH3} + 2 \times M_{CH2} + M_{OH}$$
$$M_{propanol} = 15 + 28 + 17 = 60 \text{ g mol}^{-1}$$

  - An approximation for nearly every other property
    - Proximity effects
    - Works best for medium-size molecules
  - Usually, $f(\pi) = \sum_{i \in G} n_i \pi_i$

Examples of GC methods

- For an extensive collection and discussion, see *The properties of gases and liquids* by Poling, Prausnitz and O'Connell, McGraw Hill, 5th ed, 2001 (available online in library)
- Joback (1984, 1987)
  - Melting point in K: $T_m = 122 + \sum_{i=1}^{N} T_{mi} n_i$
  - Boiling point in K: $T_b = 198 + \sum_{i=1}^{N} T_{bi} n_i$
- Constantinou and Gani (1994, 1995)
  - Melting point in K: $\exp(T_m/102.425) = \sum_{i=1}^{N} T_{mi} n_i$
  - Boiling point in K: $\exp(T_b/204.359) = \sum_{i=1}^{N} T_{bi} n_i$
  - See also Marrero and Gani (2001)
- Different atom groups are used in both methods!!

Examples of using GC methods



Predict the melting and boiling points of 2,4 dimethylphenol

Experimental data: $T_{m,exp} = 297.68$ K and $T_{b,exp} = 484.09$ K

Joback

Group i	T_{mi}	T_{bi}
-CH_3	-5.1	23.58
=CH(ds)	8.13	26.73
=C(ds)	37.02	31.01
-(aC)OH	82.83	76.34

Constantinou and Gani

Group i	T_{mi}	T_{bi}
aCH	1.4669	0.9297
aCCH_3	1.8635	1.9669
aCOH	13.7349	4.4014

Examples of using GC methods

Predict the melting and boiling points of 2,**3** dimethylphenol

Experimental data: $T_{m,exp} = 344$ K and $T_{b,exp} = 490$ K

Joback

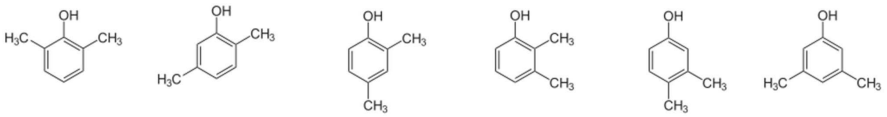
Group i	T_{mi}	T_{bi}
-CH_3	-5.1	23.58
=CH(ds)	8.13	26.73
=C(ds)	37.02	31.01
-(aC)OH	82.83	76.34

Constantinou and Gani

Group i	T_{mi}	T_{bi}
aCH	1.4669	0.9297
aCCH_3	1.8635	1.9669
aCOH	13.7349	4.4014

What about isomers?

- GC methods **cannot** distinguish between **some** isomers



2,6 2,5 2,4 2,3 3,4 3,5

  - Range of T_m is 295 K to 346 K (51 K)
  - Range of T_b is 476 K to 500 K (24 K)

- GC methods can **sometimes** distinguish between isomers

n-pentane

2 CH_3 and 3 CH_2

isopentane

3 CH_3, 1 CH_2 and 1 CH

Higher order GC methods

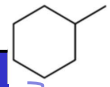
- Address the limitations of GC method
  - correct for proximity effects; allow more isomers to be distinguished

$$f(\pi) = \sum_{i \in G} n_i \pi_i + w_2 \sum_{j \in G_2} n_{2,j} \pi_{2,j} + w_3 \sum_{k \in G_3} n_{3,k} \pi_{3,k}$$

  - w_2 and w_3 are optional weights (0 or 1) that can be used to switch on higher-order contributions

Cycloalkanes

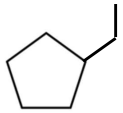
Methylcyclohexane



Group i	Number n_i	
CH_3	1	First order
CH_2	5	
CH	1	
6-ring	1	Second order
5-ring	0	
Alicyclic side chain	1	

	T_m (K)	T_b (K)
Expt.	146.56	374.09
First order	173.54	369.71
Second order	146.46	377.81

Ethylcyclopentane



Group i	Number n_i	
CH_3	1	First order
CH_2	5	
CH	1	
6-ring	0	Second order
5-ring	1	
Alicyclic side chain	1	

	T_m (K)	T_b (K)
Expt.	134.70	376.51
First order	173.54	369.71
Second order	122.14	377.69

Performance of GC methods

- Accuracy of GC methods for boiling point (in K)

Table adapted from Poling et al. (2001).

Method	Joback	Constantinou & Gani	Marrero & Pardillo	Katriztky	Jurs
# substances	353	341	347	175	242
AAE	16.8	13.4	7.5	9.2	5.3
%AAE	5.0	4.0	2.0	2.7	1.5
# %AAE > 10%	42	39	10	6	3
# %AAE < 5%	242	279	318	154	231

- Methods exist for the prediction of many properties: liquid density, heat capacity, heat of vaporization, molar volume, critical properties (pressure, temperature, density, acentric factor), etc.

Beyond pure component properties

- All the GC methods we have seen are for pure component properties
- They are also mostly valid at one temperature
  - E.g., density at 298 K.
  - For heat capacity, temperature dependence is included
- What about mixtures?
  - Assume ideal mixing for some properties
  - But what about phase equilibria?

UNIFAC (Fredenslund et al., 1976)

- Derived from the UNIQUAC activity coefficient model:
  - For a mixture of N_c components, and component i:

$$\ln \gamma_i = \ln \gamma_i^C + \ln \gamma_i^R$$

Residual part →

Combinatorial part ↓

$$\ln \gamma_i^C = (1 - \phi_i + \ln \phi_i) - 5q_i\left(1 - \frac{\phi_i}{\theta_i} + \ln \frac{\phi_i}{\theta_i}\right)$$

$$\phi_i = \frac{r_i}{\sum_{j=1}^{N_c} x_j r_j}; \theta_i = \frac{q_i}{\sum_{j=1}^{N_c} x_j q_j}$$

r_i is the van der Vaals volume of component i

q_i is the van der Vaals area of component i

$$\ln \gamma_i^R = q_i\left(1 - \ln \frac{\sum_{j=1}^{N_c} x_j q_j \tau_{ij}}{\sum_{j=1}^{N_c} x_j q_j} - \sum_{j=1}^{C} \frac{x_j q_j \tau_{ij}}{\sum_{k=1}^{N_c} x_k q_k \tau_{kj}}\right)$$

$$\tau_{ij} = e^{-a_{ij}/RT}$$

a_{ij} is the interaction between components i and j

R is the gas constant

T is temperature

- Concept of a mixture of groups

UNIFAC (Fredenslund et al., 1976)

- Derived from the UNIQUAC activity coefficient model
- Concept of a mixture of groups

$$\ln \gamma_i = \ln \gamma_i^C + \ln \gamma_i^R$$

Residual part →

Combinatorial part ↓

$$\ln \gamma_i^C = \ln \frac{\phi_i}{x_i} + 5q_i \ln \frac{\theta_i}{\phi_i} + l_i - \frac{\phi_i}{x_i}\sum_{i=1}^{Nc} x_i l_i$$

$$l_i = 5(r_i - q_i) - (r_i - 1), \quad i = 1,\ldots, Nc$$

$$\theta_i = q_i x_i \Big/ \sum_{i=1}^{Nc} x_i q_i; \ \phi_i = r_i x_i \Big/ \sum_{i=1}^{Nc} x_i r_i$$

$$r_i = \sum_{k=1}^{Ng} n_{ki} R_k; \ q_i = \sum_{k=1}^{Ng} n_{ki} Q_k, \quad i = 1,\ldots, Nc$$

Group parameters

$$\ln \gamma_i^R = \sum_{k=1}^{Ng} n_{ik}(\ln \Gamma_k - \ln \overline{\Gamma}_k)$$

$$\ln \Gamma_k = Q_k\left[1 - \ln\left(\sum_{m=1}^{Ng} \Theta_m \Psi_{mk}\right) - \sum_{m=1}^{Ng} \frac{\Theta_m \Psi_{km}}{\sum_{p=1}^{Ng} \Theta_p \Psi_{pm}}\right], \quad \forall k \in K$$

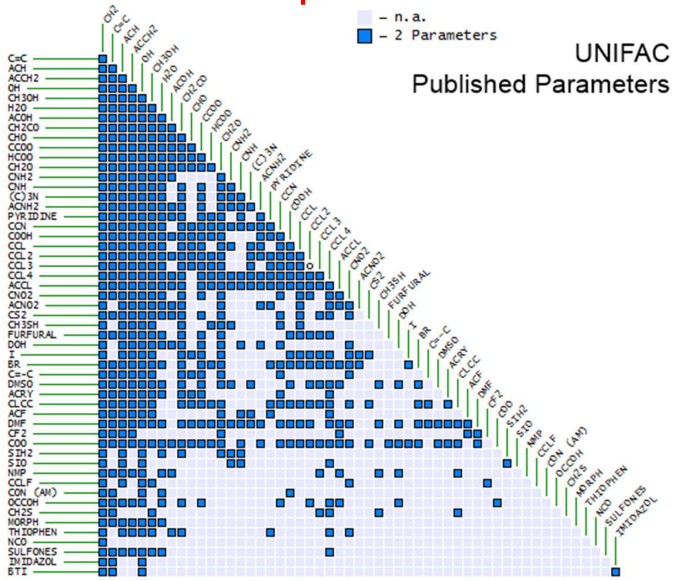
$$\ln \overline{\Gamma}_k = Q_k\left[1 - \ln\left(\sum_{m=1}^{Ng} \overline{\Theta}_m \Psi_{mk}\right) - \sum_{m=1}^{Ng} \frac{\overline{\Theta}_m \Psi_{km}}{\sum_{p=1}^{Ng} \overline{\Theta}_p \Psi_{pm}}\right], \quad \forall k \in K$$

$$\Theta_m = Q_m X_m \Big/ \sum_{k=1}^{Ng} Q_k X_k; \ X_m = \sum_{i=1}^{Nc} n_{mi} x_i \Big/ \sum_{j=1}^{Nc}\sum_{k=1}^{Ng} n_{kj} x_j, \forall m \in K$$

$$\overline{\Theta}_m = Q_m \overline{X}_{m,IB} \Big/ \sum_{k=1}^{Ng} Q_k \overline{X}_{k,IB}; \ \overline{X}_{m,IB} = n_{m,IB} \Big/ \sum_{k=1}^{Ng} n_{k,IB}, \quad \forall m \in K$$

$$\Psi_{mn} = \exp\left(-a_{mn}/RT\right), \quad \forall m,n \in K$$

UNIFAC parameter table



http://www.ddbst.com/published-parameters-unifac.html#InteractionParametersMatrix

Other approaches

- Group contribution equations of state, e.g.:
  - VTPR (Ahlers et al., 2004)
  - SAFT family of equations of state (including SAFT-γ Mie, developed at Imperial; Papaioannou et al. 2014).
- Methods based on connectivity indices or connectivity matrices
- Quantitative structure-property relationships
- Machine learning models
- Atomistic/quantum models

$$\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = \mathbf{0}$$
$$\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) \leq \mathbf{0}$$

MOLECULAR FEASIBILITY

Molecular feasibility constraints – basic idea

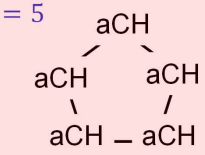
- Not all combinations of groups are meaningful
  - Need to add some chemical knowledge in our mathematical formulation!
- Write down a set of constraints that exclude some combinations

Examples of forbidden combinations

- Different types of groups exist:
  - aliphatic groups
    - Saturated chain: CH_3-, -CH_2-, -OH, -NH<
    - Unsaturated chain: -CH=CH-
  - aromatic groups
    - -aCH-, -aC-
- Among the $\sum_{k=K_{min}}^{K_{max}} \frac{(N+K-1)!}{K!(N-1)!}$ combinations, we have:

unphysical

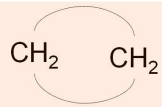
$n_{aCH} = 5$



$CH_3 - CH_2 - CH_2 -$

$n_{CH3} = 1, n_{CH2} = 2$

$n_{cCH2} = 2$



undesirable

Making sure only chemically-feasible combinations are generated

- Strategy
  1. Define group types and valencies v_i
  2. Define a molecule type (acyclic, monocyclic, etc)
  3. Add constraints to prevent certain combinations

1. Define a type and valency for each group

- Valency = number of free bonds

Group	Type	Valency i
CH_3	general	1
CH_2	general	2
CH	general	3
C	general	4
OH	general	1

Group	Type	Valency i
CH=CH	unsaturated	2
aCH	aromatic	2
aC	aromatic	3
$aCCH_3$	aromatic	2
$aCCH_2$	aromatic	3

2. Define a molecule type: Indicator variables

- Define binary molecule types
  - $y_a = 1$ if molecule is acyclic; 0 otherwise
  - $y_m = 1$ if molecule is monocyclic; 0 otherwise
  - $y_b = 1$ if molecule is bicyclic; 0 otherwise
  - Then $y_a + y_m + y_b = 1$

- Define "continuous" molecule type, m

$$m = \begin{cases} 1 & \text{for an acyclic molecule} \\ 0 & \text{for a monocyclic molecule} \\ -1 & \text{for a bicyclic molecule} \end{cases}$$

- Relate m to binary molecule types:

$$m - (y_a - y_b) = 0$$

2. Define a molecule type: Relate indicator variables to numbers of group

- To illustrate concept, assume all cyclic molecules are aromatic (i.e., no cyclohexane)

- Then:
  - a monocyclic molecule must have 6 aromatic groups
  - a bicyclic molecule must have 10 aromatic groups
  - G_a is the set of aromatic groups

- How can this be captured in an algebraic equation?

3. Constraints to prevent unphysical combinations
Octet rule

- The octet rule: there should be no free bonds in a molecule

  Odele and Macchietto (1993)

$$\left(\sum_{i \in G}(2 - v_i)n_i\right) - 2m = 0$$

(G is the set of all groups)

Applying the octet rule

Are the following combinations feasible?

1. $n_{CH3} = 1$ and $n_{CH2} = 2$
2. $n_{CH3} = 2$ and $n_{CH2} = 2$
3. $n_{CH3} = 1, n_{aCH} = 4, n_{aCCH2} = 1, n_{aCCH3} = 1$

3. Constraints to prevent undesired combinations
Bonding rule

- Two adjacent groups in a molecule cannot be bonded together twice

vs

CH_3-CH=CH-CH_3

Octet rule is satisfied!!!

$n_{CH3} = 2, n_{CH} = 2, m = 0$

$\Rightarrow (2-1) \times 2 + (2-3) \times 2 - 0 = 0$

- The bonding rule:

$$n_j(v_j - 1) + 2 - \sum_{i \in G} n_i \leq 0, \forall j \in G$$

Applying the bonding rule

- Are the following combinations feasible?
  1. $n_{CH2} = 2$
  2. $n_{CH3} = 2$ and $n_{CH} = 2$
  3. $n_{CH3} = 2$ and $n_{CH=CH} = 1$
  4. $n_{CH3} = 2$ and $n_{CH2} = 2$

Additional constraints

- Some constraints can be based on chemical insights
  - For example, compounds with several unsaturated bonds tend be less stable, so limit no. bonds to 1:

$$n_{CH2=CH} + n_{CH=CH} + n_{CH2=C} + n_{CH=C} + n_{C=C} \leq 1$$

- Some constraints aim to reduce the size of the design space
  - For example, an upper bound on the number of groups in the molecule:

$$\sum_{i \in G} n_i \leq 10$$

  - An upper bound on the number of groups of a given type:

$$n_{OH} \leq 3$$

AN EXAMPLE PROBLEM

Designing a new polymer - revisited

Case Study 1 – Maranas (1996)

We are seeking a replacement polymer for an existing product. The key properties of the polymer are:

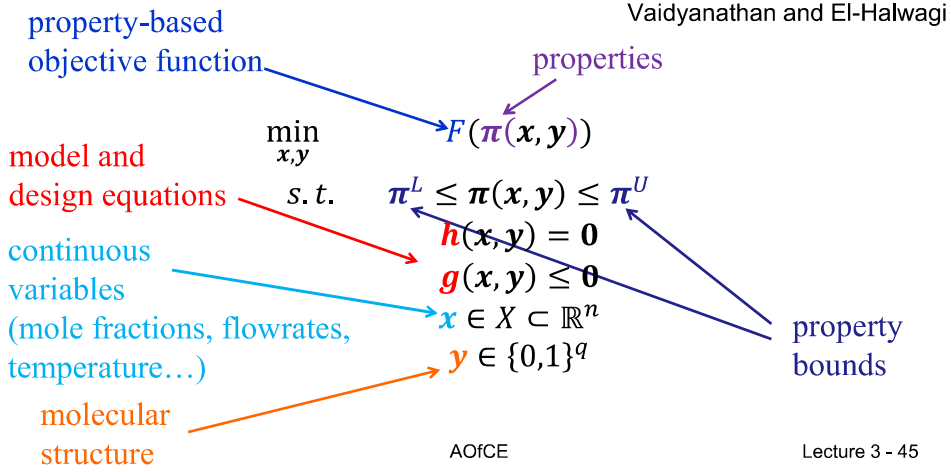
- Density D_S with target $D_S^0 = 1.50$ g cm^{-3}
- Water absorption W with target $W^0 = 0.005$ g H_2O / g polymer
- Glass transition temperature T_g with target $T_g^0 = 383$ K

$$\min \left(\frac{D_S - D_S^0}{D_S^0}\right)^2 + \left(\frac{W - W^0}{W^0}\right)^2 + \left(\frac{T_g - T_g^0}{T_g^0}\right)^2$$

Computer-Aided Molecular Design (CAMD)

"Given a desired range for a set of properties and performance criteria, design the compound that performs best, while possessing properties within the acceptable range"

Vaidyanathan and El-Halwagi



Identify property models

- Group contribution method of van Krevelen (1990), based on polymer repeat unit, e.g., -($CH_2(CHCl)_2$)-
  - $D_S = \frac{\sum_{i=1}^{N} M_i n_i}{\sum_{i=1}^{N} V_i n_i}$ g cm^{-3}
  - $T_g = \frac{\sum_{i=1}^{N} Y_{gi} n_i}{\sum_{i=1}^{N} M_i n_i}$ K
  - $W = \frac{\sum_{i=1}^{N} 18 H_i n_i}{\sum_{i=1}^{N} M_i n_i}$ g H_2O/g polymer

Parameters:
- N total number of groups
- V_i, Y_{gi}, H_i, M_i contributions of group i
- NB: M_i molar mass of group i

- Design space:

  Use a subset of the groups defined in van Krevelen.

1	2	3	4	5	6	7
-CH_2-	-CO-	-COO-	-O-	-CONH-	-CHOH-	-CHCl-

- Strong relation between GC method and design space

Problem formulation

$$\min \left(\frac{D_S - D_S^0}{D_S^0}\right)^2 + \left(\frac{W - W^0}{W^0}\right)^2 + \left(\frac{T_g - T_g^0}{T_g^0}\right)^2$$

s.t.
$$D_S = \frac{\sum_{i=1}^{N} M_i n_i}{\sum_{i=1}^{N} V_i n_i}$$
$$T_g = \frac{\sum_{i=1}^{N} Y_{gi} n_i}{\sum_{i=1}^{N} M_i n_i}$$
$$W = \frac{\sum_{i=1}^{N} 18\ _i n_i}{\sum_{i=1}^{N} M_i n_i}$$
$$n_i = \sum_{k=0}^{K} 2^k y_{ik}, i = 1, \ldots, N$$
$$0 \le n_i \le n_i^U,\ i = 1, \ldots, N$$

- What is the octet rule in this case? Why is it not needed?
- See Maranas (1996) for a linear formulation of the problem

The ten best solutions

Table 9. Results of Property Matching for Case Study 1 from Maranas (1996)

rank	repeating unit	violation	W	T (K)	D
1	−($CH_2(CHCl)_2$)−	0.0163	0.0049	384.68	1.4889
2	−($CH_2(CHCl)_3$)−	0.0264	0.0051	393.10	1.5351
3	−(($CH_2)_2(CHCl)_3$)−	0.0526	0.0047	376.94	1.4489
4	−(CHCl)−	0.1134	0.0056	412.37	1.6524
5	−(CH_2CHCl)−	0.1169	0.0044	363.20	1.3827
6	−(($CH_2)_3$O$(CHCl)_3$)−	0.1674	0.0058	354.30	1.3977
7	−(($CH_2)_3$O$(CHCl)_2$)−	0.1843	0.0059	336.12	1.3333
8	−(($CH_2)_3(CHCl)_2$)−	0.1974	0.0040	346.04	1.3082
9	−(($CH_2)_3$OCHCl)−	0.2166	0.0061	301.41	1.2255
10	−(($CH_2)_2$O$(CHCl)_3$)−	0.2474	0.0062	366.23	1.4605

0.005 g/g 383 K 1.5 g cm^{-3}

Extension

- How would you write the following constraint:

  "Use at most two types of groups from the set of possible groups"