Chapter 3
How to Mix Molecules with Mathematics

Bas van't Hof[1] Jaap Molenaar [2†] Lennart Ros [1] Martijn Zaal [3]

Abstract:
In this paper we develop two methods to calculate thermodynamic properties of mixtures. Starting point are the basic assumptions that also form the basis for the COSMO-RS model. In this approach, the individual molecules are represented by their geometrical shape with an electrical charge density on their surfaces. Next, the surface is split up into surface segments each with its own charge. In COSMO-RS a strong reduction is introduced by treating the segments as if they are completely independent. In the present study we take into account that the coupling between two patches is essentially dependent on the charge distribution on neighboring segments and on the local geometrical structure of the surface. Two approaches are followed. The first one points out how the model equations, which comprise the optimization of the entropy and conservation of internal energy, can efficiently be solved in general, thus also if the dependency between segments and the local geometry is included in the expression for the coupling energy between segments. In the second method the configuration with maximal entropy and prescribed energy is sought via simulation. Successive molecular configurations of the mixture are simulated and updated via a genetic algorithm to optimize the entropy. The second method is more time consuming but very general.

KEYWORDS: *Mixture properties, Entropy, Optimization, COSMO-RS*

[1] Vortech, Delft, The Netherlands
[2] Wageningen University, Wageningen, The Netherlands
[3] Free University, Amsterdam, The Netherlands
[†] corresponding author: `jaap.molenaar@wur.nl`

3.1 Introduction

Thermodynamic properties of a mixture, such as the miscibility of the components and partial vapor pressures, could in principle be calculated by accounting for all the interactions between the constituting molecules. In practice, however, a rigorous approach along these lines is only tractable for a highly restricted number of molecules. In view of the huge number of molecules in a fluid, one has to rely on methods from statistical physics, in which averaging procedures are applied over possible configurations. Even then one has to introduce severe assumptions in order to make calculations for realistic mixtures possible.

In 1995, a promising idea to solve this longstanding problem was worked out by Andreas Klamt [1, 2, 3]. His approach is referred to as COSMO-RS (COnductor like Screening MOdel for Realistic Solvents) and has proven to be quite powerful in some cases. One of the strong points is that the computation times are very modest. The method has its limitations, since it is based on rules that completely ignore the geometry of the molecules. The aim of the present project is to reconsider the problem of mixing anew preferably including the geometrical effects.

We decided to maintain a basic principle of COSMO-RS, namely to represent a molecule via a rigid shell with an electric charge distribution. This will be explained in §3.2. This approach assures that long-range interactions and screening effects are taken into account, but in an averaged manner, and will not lead to unacceptably long computing times.

We followed two lines of research. One line, presented in §3.3 can be looked upon as a natural extension of COSMO-RS with now the geometrical features of the molecules taken into account. In this approach, the optimization the entropy of the mixture under the condition of conserved energy is appropriately done via a fast numerical scheme.

In the second research line, presented in §3.4, the configuration with maximal entropy and prescribed energy is sought via simulation. A molecular configuration is represented in the computer by specifying the positions and orientations of a big number of molecules. An initial configuration is randomly chosen and gradually updated via a genetic optimization algorithm to optimize the entropy.

In §3.5 the results and recommendations are summarized.

3.2 The COSMO-RS model

Basic ingredients

For a clear understanding of the present project it is necessary to first explain the essential ingredients of the COSMO-RS model. Lots of details can also be found in [6, 7].

The first step in this model is taking into account long range interactions and screening effects in an averaged way. To that end the molecule is thought to be embedded in a cavity located in a perfect conductor, that is a material with an infinitely large dielectric constant. Since the molecule will in general have a charge distribution and therefore possess an electric field, it will polarize the embedding medium. That will result in an electric field that can be thought to stem from a charge distribution on the surface of the molecular cavity. In the method the molecule is replaced by the surface of the cavity together with the induced electrical charge distribution. In Figure 3.1 a sketch of such a surface and its charge distribution is given for a water molecule. Such a charge distribution is the result of a quantum mechanical calculation and is throughout this project assumed to be given for each type of molecule in the mixture.

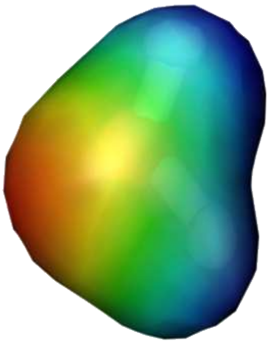

Figure 3.1: *The surface of the cavity of a water molecule with its charge distribution.*

The next step is to divide the surface up into small segments, each with a fixed amount of charge. This segmental charge is obtained by integrating the local charge distribution over the segment. So each molecule is now represented by a number of charged segments on the surface of its cavity. To keep this approach realistic, the size of these segments should be large enough to make the concept of individual pairing of segments meaningful. In practice the segment area is chosen in the range 3–25 (Angstrom)^2.

The following step is to realize that in a fluid the molecules are nearly space filling. Each molecule is thus in touch with a number of neighboring molecules. The consequence is that most of the time a segment of one molecule is in touch with a segment of another molecule. This contact implies a certain amount of energy, depending on the signs and the values of the segment charges. Segments with opposite charge signs attract each other and segments with equal charge signs will repel each other. The total amount of internal energy U is the sum of all the local contributions.

If the mixture would have vanishing temperature, all positions and orientations of the molecules would be fixed. The system would be *frozen in* and have maximal order. In reality we are interested in mixtures at positive temperature. In such a system the molecules move around and perform so-called Brownian motions and the overall molecular configuration is varying all the time. Macro properties of the system are then calculated by averaging either over time or over all possible microstates with appropriate weighting functions. From statistical mechanics we know that the system most frequently attains those configurations in which the *entropy* is maximal. In fact, the preference for these microstates is so high that we may ignore all the other microstates in the averaging procedure. That's why in the following we will concentrate on the calculation of maximum entropy configurations.

Entropy

Since the number of molecules is in the order of the number of Avogadro (in the order of 10^{26}), it is completely intractable to compute the time evolution of all individual molecules, the so-called *microstate*. Instead, COSMO-RS follows a different approach. To explain this, we first discuss the labeling of segments. For simplicity, let us assume that the mixture consists of two components X and Y: a molecule X has N_X segments and a molecule Y has N_Y segments. Since the molecules of type X are mutually indiscernible and the same holds for type Y, we meet in this system with $N = N_X + N_Y$ essentially different segments. In a particular microstate one could count the frequency that a segment n is coupled to a segment m, and use the frequencies to compute probabilities. However, in the present approach we prefer an alternative scaling based on surface areas involved, which will be explained underneath. We shall denote the scaled frequencies, that do not longer correspond to integers, by $p_{n,m}$. A *macrostate* of the system is now characterized by the values $p_{n,m}, n = 1 \ldots N, m = n \ldots N$. It is clear that one macrostate may be realized by very many different microstates, which in statistical mechanics all together are referred to as an *ensemble*. Shannon proved that the appropriate expression for the entropy S, i.e. of the disorder of such a macrostate, reads as [5]

$$S = -k \sum_{i=1}^{N} \sum_{j=i}^{N} p_{i,j} \log p_{i,j} \tag{3.1}$$

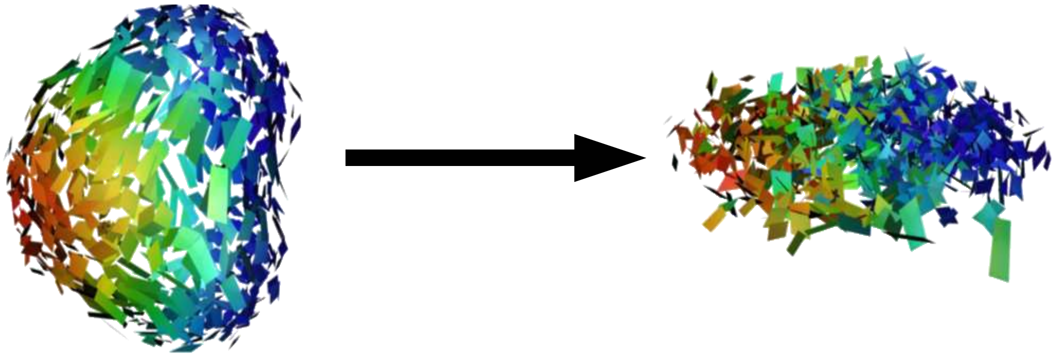

Figure 3.2: *Impression of the surface segments being treated as independent.*

Here, k is the Boltzmann constant ($\sim 1.3806504\,\mathrm{J\,K^{-1}}$).

Model equations and modeling assumptions

In this subsection we state the model equations and discuss the assumptions introduced by COSMO-RS.

In a microstate, two segments are considered to be coupled if they are located next to each other. A highly restrictive assumption of COSMO-RS is that the spatial embedding of a segment between its neighboring segments is completely ignored. In fact, all segments are cut free from their molecules and treated as if they are independent. In this view the mixture consists of a set of segments that move around independently, as illustrated in Figure 3.2.

As a consequence of this approximation, the energy involved in coupling segments n and m is take to be dependent on the charges of these segments only. Denoting the charge of segment n by σ_n, the coupling energy $E_{n,m}$ is assumed to be of the form

$$E_{n,m} = \alpha\left(\sigma_n + \sigma_m\right)^2 \tag{3.2}$$

for some positive coefficient α. Note that segments with equal but opposite charges have zero coupling energy, and segments with equal charges have high coupling energy. Steric hindering and the multipolar nature of the electric field of a molecule are thus not taken into account. Obviously, coupling segment n to segment m is equivalent to coupling segment m to segment n, therefore, both $E_{n,m}$ and $p_{n,m}$ are symmetric: $E_{m,n} = E_{n,m}$ and $p_{m,n} = p_{n,m}$.

The normalization of the $p_{n,m}$ is chosen to follow from considering the relative area that is involved in such a coupling. For this normalization we take

$$\forall n : \sum_{j=1}^{N} p_{n,j} + p_{n,n} = [X_n]\gamma_n, \tag{3.3}$$

where $[X_n]$ is the molar fraction of the molecule type segment n belongs to, and γ_n is the surface area of segment n. The extra term $p_{n,n}$ stems from the fact that coupling of segment n with itself requires *two* segments n.

Given these normalizations, the internal energy of the mixture U is easily expressed in terms of the frequencies $p_{n,m}$ and the energies $E_{n,m}$:

$$\sum_i \sum_{j \geq i} p_{i,j} E_{i,j} = U. \tag{3.4}$$

The COSMO-RS model formally involves the optimization of the entropy as a function of the variables $p_{n,m}, n = 1 \ldots N, m = n \ldots N$ under the condition that the $p_{n,m}$ are normalized and that the internal energy equals some prescribed value U. In formulae, the required macrostate will be the solution of the following constrained optimization problem:

$$\begin{cases} \text{max} & S(\{p_{i,j}\}) = -k \sum_{i=1}^{N} \sum_{j \geq i}^{N} p_{i,j} \log p_{i,j} \\ \text{under the conditions that} & \forall n : \sum_j p_{n,j} + p_{n,n} = [X_n]\gamma_n \\ \text{and the condition} & \sum_i \sum_{j \geq i} p_{i,j} E_{i,j} = U \end{cases} \tag{3.5}$$

Formally, only $p_{n,m}$ with $m \geq n$ are part of the problem. If in the following $m < n$, $p_{n,m}$ is considered to be shorthand notation for $p_{m,n}$. Although this might seem artificial at first, it makes formulas involving sums easier to read and understand.

The value of U is determined by the external conditions of the system. In practice, one often fixes the temperature T of the mixture. As discussed later on, the value of U is then an outcome, rather than an input of the system. The roles of U and T are in fact interchangeable in the procedure.

3.3 Extended COSMO-RS model

The assumption of independency of segments allows for an explicit solution of this problem along combinatorial lines using the notion of partition function. For this derivation, see Appendix I in [4]. This reduction is a great advantage from a computational point of view. However, this assumption forms a weak point, since it makes the model quite unrealistic, e.g., to deal with irregularly shaped molecules that give rise to steric hindering. In the present approach we want to get rid of this assumption. The consequence is that we have to face the original optimization problem (3.5). It also implies that (3.2) is no longer applicable. The energy involved in coupling two segments should be made to depend on the neighboring segments, too. In the next subsection this point will be touched. For the present procedure we propose for solving (3.5) it is only relevant that some (nonnegative) expression for the coupling energy $E_{n,m}$ is available.

A general method to solve the constrained maximization problem (3.5) is to make use of Lagrange multipliers. For that purpose we form the Lagrangian

$$\begin{aligned} L(\{p_{n,m}\},\{\lambda_n\},\mu) = &- k\sum_{i=1}^{N}\sum_{j\geq i}^{N} p_{i,j}\log p_{i,j} + \sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{N} p_{i,j} + p_{i,i} - [X_i]\gamma_i\right) \\ &+ \mu\left(\sum_{i=1}^{N}\sum_{j\geq i} p_{i,j}E_{i,j} - U\right) \\ = &-k\sum_{i=1}^{N}\sum_{j=i}^{N} p_{i,j}\log p_{i,j} + \sum_{i=1}^{N}\sum_{j=i}^{N}(\lambda_i+\lambda_j)p_{i,j} - \sum_{i=1}^{N}\lambda_i[X_i]\gamma_i \\ &+ \mu\left(\sum_{i=1}^{N}\sum_{j\geq i} p_{i,j}E_{i,j} - U\right) \end{aligned} \tag{3.6}$$

This Lagrangian has as variables the frequencies $p_{n,m}$, $n = 1\ldots N$, $m = n\ldots N$ and the Lagrange multipliers λ_n, $i = n\ldots N$ and μ. For the second identity, the convention $p_{m,n} = p_{n,m}$ has been used in order to eliminate any $p_{n,m}$ with $m < n$. All other quantities such as the internal energy U and the coupling energies $E_{m,n}$ act as parameters. The term containing $\lambda_i + \lambda_j$ follows by replacing $p_{i,j}$ with $p_{j,i}$ whenever $i < j$, and rearranging the double sum:

$$\sum_{i=1}^{N}\lambda_i\sum_{j=1}^{i} p_{i,j} = \sum_{j=1}^{N}\sum_{i=j}^{N}\lambda_i p_{i,j} = \sum_{i=1}^{N}\sum_{j=i}^{N}\lambda_j p_{j,i} = \sum_{i=1}^{N}\sum_{j=i}^{N}\lambda_j p_{i,j} \tag{3.7}$$

Note that the Lagrangian does not include the kinetic energy, since in a fluid the molecules motions are quite slow, so that the total energy is completely dominated by the potential (internal) energy.

Standard theory tells us that the solution of (3.5) is also the solution of the set of equations obtained by setting the derivatives of the Lagrangian with respect to each of its variables equal to zero. So, (3.5) is equivalent to solving the system

$$\begin{cases} -k(\log p_{n,m}+1)+(\lambda_n+\lambda_m)+\mu E_{n,m}=0, & \forall n, \forall m \geq n \\ \sum_j p_{n,j}+p_{n,n}=[X_n]\gamma_n, & \forall n \\ \sum_i \sum_{j\geq i} p_{i,j}E_{i,j}=U. \end{cases} \tag{3.8}$$

The term $(\lambda_n+\lambda_m)$ follows from the second equality in (3.6).

A result from thermodynamics states that the Lagrange multiplier μ is related to the absolute temperature via

$$\mu=-\frac{1}{T}.$$

Since the temperature of the mixture can be controlled, μ will from now on be considered as a parameter. This implies that we only need to solve the equations in the first two lines of (3.8) for the variables $p_{n,m}$, $n=1\ldots N$, $m=n\ldots N$ and λ_n, $n=1\ldots N$. The equation in the third line will be used afterwards to calculate the internal energy U.

Solving the first equation in (3.8) for $p_{n,m}$ and substituting in the second one, we obtain the following set of equations:

$$\begin{cases} p_{n,m}=e^{-1+\frac{\lambda_n+\lambda_m+\mu E_{n,m}}{k}} & \forall n, \forall m \geq n \\ \sum_j e^{-1+\frac{\lambda_n+\lambda_j+\mu E_{n,j}}{k}}+e^{-1+\frac{2\lambda_n+\mu E_{n,n}}{k}}=[X_n]\gamma_n & \forall n \end{cases} \tag{3.9}$$

To rewrite these equations in a more tractable form we introduce the vector

$$\Lambda_n := e^{\lambda_n/k}, n=1\ldots N$$

and the matrix

$$F_{n,m} := e^{\mu E_{n,m}/k}+\delta_{n,m}e^{\mu E_{n,n}/k}$$

with the Kronecker delta as is usual defined as $\delta_{n,m}=1$ if $n=m$ and $\delta_{n,m}=0$ if $n\neq m$. The last equation of (3.9) can then be written as

$$\forall n : \Lambda_n \sum_j F_{n,j}\Lambda_j = e[X_n]\gamma_n =: \alpha_n \tag{3.10}$$

The right hand sides and the matrix $F_{n,m}$ are known. So, we arrive upon a set of N quadratic equations for the unknowns $\Lambda_n, n=1\ldots N$. This system is not simple to solve explicitly, but it has a pretty nice form for numerical evaluation. The Jacobian matrix of the set of equations (3.10) is easy to obtain explicitly. So, we resort to a numerical, and thus iterative approach and need therefore an initial guess for the Λ_n. To that end we observe that the exponentials in $F_{n,m}$ are expected to be close to one, since the coupling energies $E_{n,m}$ are small. Setting $F_{n,m} = 1$ for all $n \neq m$ and $F_{n,n} = 2$ for all n, we obtain the approximating equation

$$\Lambda_n^2 + \Lambda_n \sum_j \Lambda_j = \alpha_n.$$

Neglecting the first term Λ_n^2 since it is expected to be small compared to the sum in the second term, we find as initial guess

$$\Lambda_n^0 := \frac{\alpha_n}{\sqrt{\sum_j \alpha_j}}.$$

Once the Λ_n are known, the values of the variables $p_{n,m}$ follow from

$$p_{n,m} = e^{-1+\frac{\lambda_n+\lambda_m+\mu E_{n,m}}{k}} = \Lambda_n \Lambda_m e^{-1+\frac{\mu E_{n,m}}{k}} \tag{3.11}$$

Example

To solve Λ_n from (3.10), we choose as iterative scheme the Newton-Raphson method. As a toy model we consider a fluid with only one molecule type with $N = 4$ segments of equal size. Furthermore, we use $\gamma_n = 1$ for all n. Taking for the $E_{n,m}$ matrix

$$E_{n,m} = \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix},$$

representing charges of equal size, but opposite sign, we found for the $p_{n,m}$ matrix

$$p_{n,m} = \begin{pmatrix} 0.0945 & 0.3583 & 0.0945 & 0.3583 \\ 0.3583 & 0.0945 & 0.3583 & 0.0945 \\ 0.0945 & 0.3583 & 0.0945 & 0.3583 \\ 0.3583 & 0.0945 & 0.3583 & 0.0945 \end{pmatrix}$$

at $T = 300$ K. This clearly shows that segments with opposite charges tend to attract each other, whereas segments with charges of equal signs repel each other. As expected, the lower the temperature, the stronger the influence of the energy. The convergence appeared to be very fast, thanks to the system being quadratic.

In Figure 3.3 it is illustrated that some couplings are geometrically impossible. In a second example we illustrate how to deal with such a situation. In the example we consider again the fluid in the example above, but now we assume that segments one and

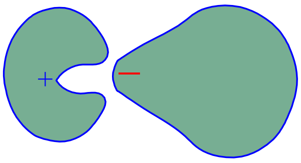

Figure 3.3: *Sketch of a situation in which a coupling is geometrically impossible, although the involved charges would favor it.*

two cannot touch each other. two. This can be taken into account by a very high entry in the energy matrix, say $E_{1,2} = 20$:

$$E_{n,m} = \begin{pmatrix} 4 & 20 & 4 & 0 \\ 20 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix},$$

The coupling frequencies now become

$$p_{n,m} = \begin{pmatrix} 0.2073 & 0.0010 & 0.1219 & 0.4625 \\ 0.0010 & 0.2073 & 0.4625 & 0.1219 \\ 0.1219 & 0.4625 & 0.0717 & 0.2721 \\ 0.4625 & 0.1219 & 0.2721 & 0.0717 \end{pmatrix}$$

As expected, the coupling frequency between segments one and two dropped to almost zero. Note that also the other entries have changed. The highest frequency is now found between one and four, as was to be expected, since this is energetically speaking the most favorable coupling.

Choice of coupling energies

Using the above model, the macrostate with the highest entropy can be easily calculated, provided that the coupling energies $E_{n,m}$ are given. It remains to specify them such that the geometrical effects are accounted for. In the present project we developed some ideas, which are worth to be worked out. out.

- Include neighboring effects. If two segments couple, also the neighbors come close together. It depends on the charges on the neighboring segments and their distances what the effect will be on the energy. A possibility to take this into account is to choose

$$E_{n,m} = \alpha \left(\sigma_n + \sigma_m\right)^2 + \beta \sum_{i_n, j_m} d_{i,j} \left(\sigma_i + \sigma_j\right)^2,$$

where i_n runs over all neighbors of segment n and j_m runs over all neighbors of segment m and $d_{i,j}$ is some appropriate distance function. The factor β has to be finetuned in order to get the correct balance between the two terms. In this way we introduce a penalty if a coupling involves neighbors that repel each other. So the second term acts as a penalty function. Including higher order neighbor effects might also be an option.

- An alternative would be to include the local curvatures into $E_{n,m}$, for instance a term proportional to
$$(H_n + H_m)^2 ,$$
where H_n is the (average) mean curvature of the molecule surface around the position of segment n. The advantage of this criterion is that it is much less subjective than defining penalties for individual couplings.

- Forbidden couplings. If illustrated in the example above, if some coupling is physically infeasible due to the shape of molecules, it can be forbidden simply by assigning to it a very high energy cost. It is to be expected that this will somewhat reduce the quality of the initial guess discussed above, which means that the numerical method will need more time to find the solution.

3.4 Entropy optimization via simulation

In this section we follow an approach that is considerably different from the one presented in the preceding section. The aim is the same: to find a configuration with maximum entropy and prescribed energy. The idea is to do perform this via simulation. We focus on a part of the fluid, a so-called parcel, with a tractable number of molecules. The rest of the fluid is represented by periodic boundary conditions, as explained below. The molecular configuration in this fluid parcel is represented in the computer by specifying the positions and orientations of all molecules in it. An initial configuration is randomly chosen and gradually updated via a genetic optimization algorithm to optimize the entropy, meanwhile keeping the energy at or close to the prescribed value. This approach has the complication that randomly placed molecules will in general overlap. So, this leads to an extra optimization goal: minimization of the overlap.

The present approach has the following features:

- As we already did above in the (extended) COSMO-RS model, we ignore the kinetic energy. So, our search space is the set of static configurations in the fluid parcel.

- The surface of the molecule is approximated by segments, each with its own charge. The geometry of the surface is taken into account, so the segments are connected.
- The state of a molecule consists of is 6 parameters per molecule: 3 coordinates for the location and 3 angles for the orientation. From these the position of each segment directly follows.
- In the coupling energy between segments we incorporate the geometry, in the way discussed in §3.3.

Periodic boundary conditions

In the simulation approach we calculate the properties in a small fluid parcel. this is based on the assumption that on average the parcel represents the fluid as a whole quite well. To avoid boundary effects, periodic boundary conditions are applied. This results in a periodic domain, as illustrated in Figure 3.4. Now, we deal with an infinitely large domain, but represented with only a finite amount of information because of the repeating patterns.

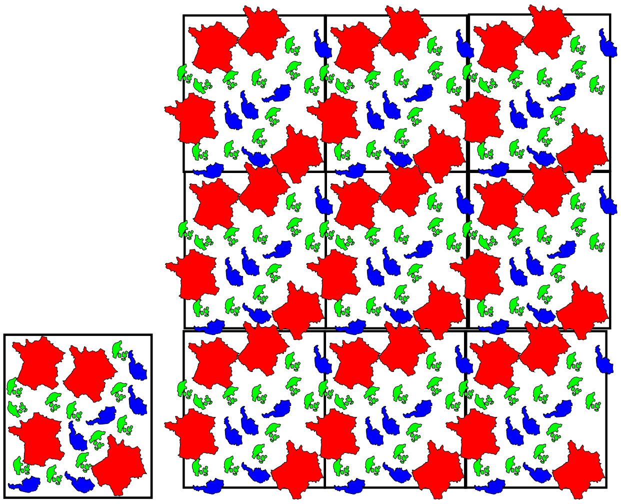

Figure 3.4: *Left: A small fluid parcel. Right: A periodic domain. A periodic domain has no boundaries.*

Optimization procedure

Let us consider n molecules (maybe of different species) in the fluid parcel. We use the following noattions:

state The state $\boldsymbol{x} \in \mathbb{R}^{6n}$ of the configuration consists of the locations and orientations of all n molecules;

Energy function The energy function $E : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the binding energy for the given state;

Entropy function The entropy function $S : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the entropy for the given state;

Overlap function The function $V : \mathbb{R}^{6n} \to \mathbb{R}_+$ returns the amount of space occupied by two or more molecules at the same time.

For a given *target energy* E_t we have to solve the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & S(\boldsymbol{x}) \\ \text{under the restrictions that} \quad & V(\boldsymbol{x}) = 0, \\ \text{and} \quad & (E(\boldsymbol{x}) - E_t)^2 = 0. \end{aligned} \tag{3.12}$$

3.4.1 Technical details

The optimization problem (3.12) has many local optima. By the way, it is good to realize that it also has many global optima. For example, if we have an optimal solution and we shift the whole solution a little bit (and/or rotate it) we again have an optimal solution. In general, it is typical for many-particles systems that one and the same macro state may correspond to a huge amount of micro states, all having the same entropy and energy. In the present approach we need to find only one of the global optima. Since the system has so many degrees of freedom, optimization may lead to unacceptably long computation times. The success of the method will therefore heavily depend on how efficiently the functions E, S and V and their gradients are evaluated. In this section we discuss several related technical details.

Efficient evaluation of overlap V

Each molecule may be described as a set of tetrahedra. The overlap in a configuration can therefore be determined by comparing every one of these tetrahedra to every other tetrahedron, calculating the volume they share and adding all these overlap volumes. Such a process is quadratic in the number of tetrahedra in the configuration and would become prohibitive very quickly when many molecules are to be modelled, or when detailed shapes are to be used to model them.

The calculation of the overlap can be sped up considerably by keeping track of the *circumscribed spheres* of the molecules, as illustrated in Figure 3.5. This is very simple to do, because the circumscribed sphere of the molecules does not change when the molecule is rotated and because its radius only depends on the molecule species. If the circumscribed spheres do not intersect, the molecules do not intersect and their tetrahedrons need not be compared. In this way, every molecule is only seriously compared to the molecules near it. A similar speed-up may be achieved by comparing the circumscribed spheres of the individual tetrahedra before calculating their overlap.

A further reduction in the calculation can be achieved using a *grid*, as illustrated in Figure 3.6. The domain is split up into grid cells. For every grid cell, a list is made of all molecules in or near it (i.e. whose center of gravity is in the shaded area). Molecules near a grid cell boundary may be in more than one list.

In this case the calculation of the overlap consists of the following steps:

```
for all molecules do
  place it in a list of all grid cells in or near which it is located
end for

for all molecules $M_1$ do
  for all molecule $M_2$ in or near the grid cell where molecule $M_1$ is located do
    compare circumscribed spheres:
    if spheres do not intersect then
      Overlap $V$ is zero.
    else
      compare all tetrahedra of $M_1$ to all tetrahedra of $M_2$:
      if there is no intersection then
        Overlap $V$ is zero.
      else
        a detailed calculation is needed
      end if
    end if
  end for
end for
```

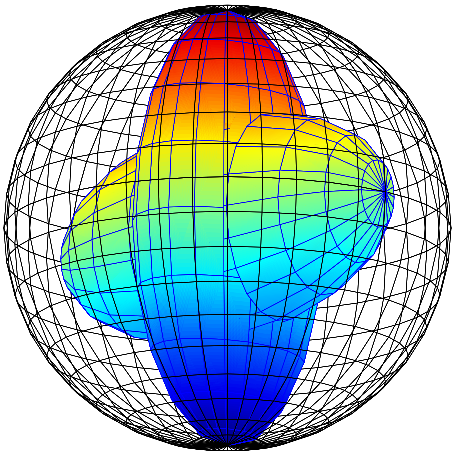

Figure 3.5: *A molecule and its circumscribed sphere: molecules do not overlap if their circumscribed spheres do not do.*

Efficient evaluation of coupling frequencies

For the evaluation of the energy and the entropy, it is necessary to determine for every segment of the molecule shell to which segment(s) it is 'coupled'. A simple way to determine these couplings is by the overlap calculation of slightly enlarged molecules. This idea is illustrated in Figure 3.7. The molecules $M1$ and $M2$ (dark colors) do not overlap. The enlarged molecules (lighter colors), however, have some overlap. Segment A1, or rather the tetrahedron that it is a face of, overlaps with B2 and a little bit with A2. Hence, we say that A1 is coupled mostly to B2 and a bit to A2 and we let both couplings contribute to the entropy, but in a weighted fashion.

Smoothing the functions

The overlap-function V and the coupling frequencies (and hence the energy E and entropy S) are continuous and differentiable functions of the state $\boldsymbol{x}$. Their derivatives, however, are not continuous, so the Hessian matrices of the functions V, E and S do not exist. Since many optimization techniques need Hessian matrices, it is useful to smooth these functions. A simple way to do this is to 'soften' the tetrahedra. When doing so, the original overlap V_{ij} between two tetrahedra i and j is modified to V'_{ij} according to

$$V'_{ij} := \frac{V_{ij}^2}{\epsilon \min(V_i, V_j) + V_{ij}}, \tag{3.13}$$

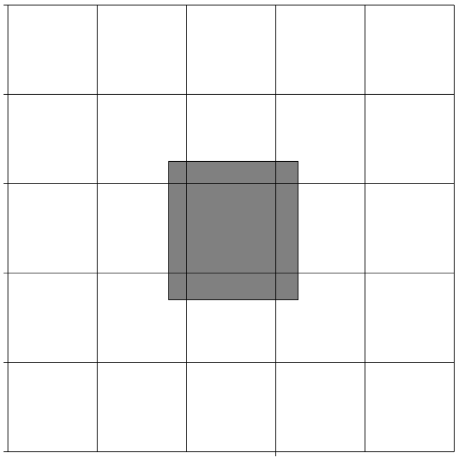

Figure 3.6: *The grid used to speed up the calculation of the overlap.*

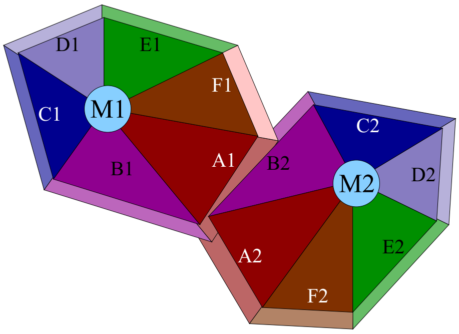

Figure 3.7: *Example for the calculation of the couplings: segment A1 is mostly coupled to segment B2, and also a little bit to A2.*

where V_i and V_j are the volumes of tetrahedra i and j, and ϵ is a 'small' parameter. Larger values for ϵ make 'softer' overlap functions.

3.4.2 Efficient optimization of the configuration

The original optimization problem (3.12) involves a target function and constraints. The constraints can be incorporated in the target function by giving a penalty for constraint violation. The modified optimization method is then

$$\text{maximize} \qquad S(\boldsymbol{x}) - c_V V(\boldsymbol{x}) - c_E (E(\boldsymbol{x}) - E_t)^2. \tag{3.14}$$

with c_V and c_E weighting factors that determine the relative contributions of the two penalty functions. This optimization problem is standard problem and may be solved using steepest descent or variations of Newton's method. In the present context some problems might be expected:

- Local optimization methods are very likely to find local optima which are not global optima.

- Local search techniques may also converge very slowly. This may happen for instance in configurations with regions that are too crowded and regions which are too empty. A lot of molecules have to move in order to even this out. They will moreover have to move in complicated patterns because the target function is not allowed to increase on the way.

To find a global optimum, additional techniques may be needed. When a local optimum is found or when convergence slows down, the solution has to be 'shaken up' in order to move away from a local optimum. Sudden changes which may help are for example

- Some (randomly chosen) molecules may be taken from the most crowded regions and placed in the emptiest regions;

- Some (randomly chosen) molecules are moved and rotated to a random place and orientation in the domain.

3.4.3 Preliminary results

The simulation approach requires a lot of programming. Due to time limitations it was not possible to produce a working molecular simulation model in only a few days. A modest start in 2D was made, which provides us with some understanding of what is involved in the calculations. The evaluation of overlap turned out to be not too complicated. The couplings were evaluated only in a simple way: every segment was considered to couple to the nearest segment of another molecule. Local search was not yet applied. For purpose of demonstration, optimization was studied via a simple random search algorithm. In that approach, a configuration $\boldsymbol{x}$ is chosen entirely randomly, after which the target function (3.14)is evaluated. The first configuration is saved and a new configuration is randomly produced. If this configuration turns out to have a higher value of the target function, then the latter replaces the former. This can be repeated many times. Obviously, this method has very slow convergence. The results of this procedure are shown in Table 3.1 and Figure 3.8. Two types of molecules are mixed: 18 of one type and 7 of another type. The dimensions of the molecules, the domain and charges were not realistically chosen, that's why no units are shown in the results. The coefficients c_V and c_E were set at one and for the target energy we use $E_t = 40$. A thousand configurations were produced, and 8 times a new 'best sofar' configuration was encountered. Table 3.1 shows that in this instance the overlap is indeed minimized, but the entropy and energy are still

Iteration	Overlap	Energy	Entropy
1	40.7	0.30	4.53
2	38.4	0.36	4.50
4	37.5	0.35	4.44
5	30.2	0.34	4.54
10	27.6	0.39	4.46
31	20.2	0.32	4.52
593	18.3	0.39	4.43
939	17.2	0.41	4.41

Table 3.1: *Results when maximizing the target function (3.14)during a random search approach. The overlap indeed reduces in the course of the time*

varying much. The initial and final (after 8 improvement steps) configurations are shown in Figure 3.8

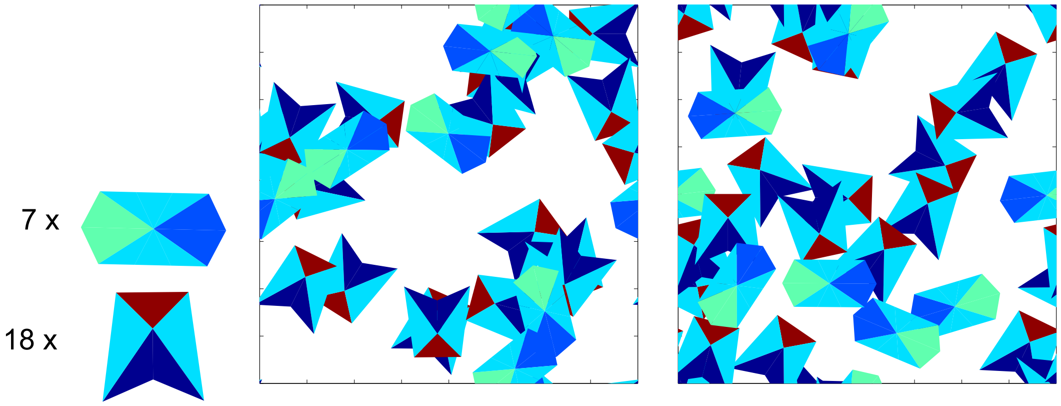

Figure 3.8: *First (left) and final (right) configurations in the a simple random search summarized in Table 3.1.*

3.5 Conclusions and Recommendations

We have shown that the COSMO-RS procedure to calculate the properties of mixtures can be extended to incorporate the geometrical effect of constraints that may drastically influence the chance that two surface segments of the constituting molecules couple. The general problem concerns the optimization of the entropy under the condition that the energy has a prescribed value. To perform this task while accounting for the geometrical effects, we followed two lines.

In the first approach, we show that the optimization problem can be very efficiently solved, by putting it in a form that is appropriate for numerical optimization methods. The geometrical constraints are included via specification of the energy involved in coupling two segments. We discuss suggestions for the effective choice of these coupling energies, such that the effect of the local geometry and the local charge distribution is taken into account.

In the second approach, we tackle the optimization problem via simulation. We focus on a part of the fluid, a so-called parcel, with a tractable number of molecules. The rest of the fluid is represented by periodic boundary conditions. The molecular configuration in this fluid parcel is represented in the computer by specifying the positions and orientations of all molecules in it. The idea is to start from a randomly chosen configuration, that is gradually updated via a genetic optimization algorithm. The object function consists of the entropy together with penalty functions that have to assure that the procedure converges to a configuration with the correct energy and without overlapping molecules. A fairly complete image of the computational aspects was obtained from developing a simple piece of software, that is restricted to 2D.

Our conclusion is that the first approach answers the original specific question quite efficiently, while the second approach is highly general and could also be applied to answer many other questions concerning mixtures.

3.6 Acknowledgements

The authors wish to thank Jaap Louwen and Peter Daudeij from Albemarle Catalysts Company BV, Amsterdam, for bringing the problem to their attention and inspiring and pleasant discussions on this topic.

Bibliography

[1] A. Klamt, *Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena*, J. Phys. Chem. 99 (1995) 2224..

[2] A. Klamt, V. Jonas, T. Brger and J.C. Lohrenz, *Refinement and Parametrization of COSMO-RS*, J. Phys. Chem. A 102 (1998) 5074.

[3] A. Klamt, *COSMO-RS From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, Elsevier, Amsterdam (2005), ISBN 0-444-51994-7.

[4] S.T. Lin and S.I. Sandler, *A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model*, Ind. Eng. Chem. Res. 41 (2002), pp. 899 - 913

[5] E.T. Jaynes, *Information Theory and Statistical Mechanics*, The Physical Review, Vol. 106, No. 4, (1957), pp 620 - 630.

[6] C.C. Pye, T. Ziegler, E. van Lenthe, J.N. Louwen, *An implementation of the conductor-like screening model of solvation within the Amsterdam density functional package. Part II. COSMO for real solvents* accepted for publication in Can. J. Chem. (2009).

[7] See www.scm.com