Parametrization of two-center Lennard-Jones plus pointquadrupole force field models by multicriteria optimization

K. Stöbener[a,*], P. Klein[a], M. Horsch[b], K. Küfer[a], H. Hasse[b]

[a] *Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany*
[b] *Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern, Germany*

Abstract

To ensure the quantitative precision and reliability of molecular simulations, force field models of molecular fluids need to be adjusted to e.g. experimental data. An optimal agreement for different properties is often not achieved by a single model parametrization. Applying multicriteria optimization, based on the evaluation and analysis of the Pareto set solves this problem. The Pareto set contains all optimal compromises between multiple conflicting objectives. Its computation and suitable visualization enables the end user to freely choose a model parametrization, tailored to his particular application scenario.

We apply multicriteria optimization to the two-center Lennard-Jones plus point quadrupole model class (2CLJQ), which has four adjustable parameters. The Pareto set is determined and analyzed for ten real fluids: Ethane, ethylene, acetylene, fluorine, chlorine, bromine, perfluoroethylene, perchloroethylene , nitrogen, and oxygen. Thereby, two multicriteria optimization scenarios are considered, based on two criteria (saturated liquid density and vapor pressure) and three criteria (saturated liquid density, vapor pressure, and surface tension), respectively. It is shown that literature models for these fluids can be further improved in these criteria. We visualize our results by self-organizing patch plots, which facilitate the representation of the entire Pareto set and its corre-

*Corresponding author, email: katrin.stoebener@itwm.fhg.de, tel.: ++49-631-31600-4520

sponding model parametrizations. We developed this plot technique to present Pareto optimal solutions and present it for the first time.

Keywords: optimization, force field models, united-atom, Lennard-Jones, multicriteria, visualization

---

1. Introduction

Molecular simulations of fluids contribute to solving problems in physics, biology, chemistry, and engineering. They rely on the choice of suitable molecular force fields. Adjusting the force field parameters, aiming at an accurate representation of different thermophysical properties is a multicriteria optimization problem. Different properties normally cannot simultaneously be optimized without a trade-off, i.e. the objectives are conflicting and a single optimal solution cannot be determined. The goal of multicriteria optimization based on Pareto is to identify the Pareto set, which represents best possible compromises between conflicting objectives. A solution is defined to be Pareto optimal if a further improvement in one objective can only be achieved at the expense of at least one other objective. The Pareto set is a subset of all feasible solutions in the objective space. To each Pareto optimal point in the objective space corresponds one parametrization in the parameter space, representing a Pareto optimal model. Hence, identifying the Pareto set does not yield one model, but a set of optimal models from which a user can choose the one best fitting a particular application scenario.

It is attractive to use multicriteria optimization for the parametrization of molecular models. In a preceding work, the Pareto optimal Lennard-Jones models for argon and methane were identified by brute force evaluation of 200×200 parameter combinations [1]. The same procedure was used by Werth et al. [2] to identify Pareto optimal parameters for representing carbon dioxide with a two-center Lennard-Jones plus pointquadrupole force field model. The grid in the parameter space employed by Werth et al. [2] was $60 \times 60 \times 60 \times 60$. This brute force enumeration is only feasible for scenarios in which the evaluation of the objective functions is not expensive and the number of parameters is not high. This is rarely the case for optimizations of molecular models. A more efficient strategy to approximate the Pareto set is necessary. Apart from [1] and [2], we are only aware of one other work in which multicriteria optimization was used in the context of developing force field models. Mosthagim et

al. [3] developed Pareto sets by Particle Swarm Optimization with a focus on intramolecular potentials fitted to ab initio data.

In the present work, we determine the Pareto set by a combination of the sandwiching and hyperboxing algorithm [4]. The algorithm aims at efficiently approximating the Pareto set for a predefined approximation quality.

Multicriteria optimization is applied to the parametrization of the two-center Lennard-Jones plus point quadrupole (2CLJQ) potential model in two optimization scenarios. The 2CLJQ potential model has four parameters, which are adjusted to optimize first two and then three different objective functions. However, for both optimization tasks, the Pareto optimal parameter sets are evaluated regarding four functions. They describe the quantitative agreement between the simulation data and the experimental data for the saturated liquid density, the vapor pressure, the surface tension, and the critical temperature. For simplicity, we refer to the saturated liquid density only as liquid density in the following.

In the first optimization scenario the conflicting objective functions for the liquid density and the vapor pressure are used (*two-criteria scenario*). Previous works show, that model parameters adjusted to the liquid density and the vapor pressure are suitable to predict e.g. caloric or further thermodynamic properties [5, 6]. In the second scenario, the surface tension is included as an example for a third objective function (*three-criteria scenario*). For both scenarios the Pareto set is determined for ten different fluids: ethane (C_2H_6), ethylene (C_2H_4), acetylene (C_2H_2), fluorine (F_2), chlorine (Cl_2), bromine (Br_2), nitrogen (N_2), oxygen (O_2), perfluoroethylene (C_2F_4), and perchloroethylene (C_2Cl_4). The obtained Pareto optimal solutions are also compared to the performance of available molecular models from the literature. The discussed literature models were specifically developed for the use in molecular simulations. Molecular models obtained by theoretical studies as e.g. perturbation theory [7, 8] were not taken into account.

In the main text, first the results for the two-criteria scenario are discussed for all fluids. The results for the three-criteria scenario are then discussed for

acetylene and used for a comparison with those of the two-criteria scenario as a representative for all studied fluids. Finally, a brief summary of a comparison between the two- and three-criteria scenario for all fluids is given. The Pareto optimal solutions for both optimization tasks for all studied fluids can be found in the Supplementary Material.

We also introduce a novel technique for simultaneously visualizing the Pareto set in the objective space, the corresponding Pareto optimal model parameter space as well as additional functions, which were not included in the multicriteria optimization as an objective function. It is based on self-organizing maps (SOM), which are also referred to as Kohonen maps in the literature. With a SOM an interpolation of a high-dimensional data set can be displayed in low-dimensional maps [9, 10]. Hunger and Huttner [11] used SOM to gain insight into the dependence of their single-criteria optimization on parameters specifying a force field description for tripod metal templates. SOMs have been used before in a few studies in the literature to represent results from multicriteria optimizations [12, 13]. The drawback of SOMs is, that they do not only represent the given input data set, but an interpolation, also resulting in an approximation of the data set. Thus we enhanced the SOM and used it as a projection method for a given data set, in our case the Pareto set. Furthermore, by combining them with Voronoi diagrams [14], we achieved that only the exact Pareto optimal solutions are displayed. We refer to the representation as *self-organizing patch plot* (SOPP). It is a visualization specifically adapted to display any Pareto set.

By displaying the Pareto sets in the novel SOPP a comprehensive overview over the optimal solutions is gained. It enables assessing how good simultaneous representations of different properties of pure fluids by the 2CLJQ model can be without having to refer to only one single solution for each studied fluid. As many pure fluids are studied and similar results are obtained, the present statements can probably be generalized: They allow an assessment of how good the 2CLJQ model can be.

The methods presented have a much wider scope. They can be used for

parametrizing molecular force fields in general and well beyond for parametrizing any thermodynamic model.

2. Multicriteria Optimization

A multicriteria optimization problem is characterized by multiple objective functions $f_i(x)$ which have to be minimized simultaneously:

$$\min f(x) = (f_1(x), ..., f_r(x)) \in \mathbb{R}^r \tag{1}$$

They span the objective space $\mathbb{R}^r$ and depend on the decision vector $x \in \mathbb{R}^q$ where $\mathbb{R}^q$ is the design space. The solution to such a problem is a set of best compromises: For any improvement in a single objective $f_i(x)$, a decline in at least one other objective $f_k(x)$, $i \neq k$ has to be accepted. The set of all best compromises is called the Pareto set. In the context of force field parametrization the design space is spanned by the parameters describing the model. Thus we refer to it as parameter space in the subsequent text. Mapped to each point in the parameter space, hence, to each model, is one point in the objective space.

For more information on multicriteria optimization see e.g. [15, 16, 17].

2.1. Sandwiching and hyperboxing algorithm

The Pareto set is a subset of all feasible points in the objective space and needs to be approximated by a suitable numerical strategy. In the present work the sandwiching and hyperboxing algorithm is used for determining suitable approximations of the Pareto sets. This algorithm is basically taken from [4].

By scalarizations of the objective functions, single criterion optimization problems are obtained. The solutions for these scalarizations belong to the Pareto set. Thus by solving several single criterion optimizations a point-wise approximation of the Pareto set is obtained. The sandwiching and hyperboxing algorithm is used to subsequently identify suitable scalarizations to efficiently approximate the Pareto set. The algorithm focuses on exploring the Pareto set in regions, where the curvature of the Pareto set is highest and starts by identifying the extreme compromises, i.e. at first the minimum of each objective function is located. The sandwiching algorithm then alternately finds inner and outer approximations for the Pareto set, thereby assuming the Pareto set is convex. The sandwiching algorithm is applied, until a specified approximation quality is

reached. Then, if regions of non-convex behavior are identified, the non-convex regions are approximated with the hyperboxing algorithm. The sandwiching algorithm uses the weighted sum scalarization (see e.g. [17]), whereas the hyperboxing algorithm uses a scalarization proposed by Pascoletti and Serafini [18]. A brief sketch of the concept of the sandwiching and hyperboxing algorithm in a chemical engineering context can be found in Bortz et al. [19]. To solve the single criterion optimization tasks with which the Pareto set is approximated, we employ the Quasi-Newton solver NLPQLP of Schittkowski [20].

During the multicriteria optimization numerical problems may occur. Depending on the choice and settings of the single criterion solver, the sandwiching and hyperboxing algorithm may identify solutions as Pareto optimal, which are not. Thus we additionally check the Pareto sets for Pareto optimality by comparing the solutions to each other and sort out solutions, which are not Pareto optimal.

For further numerical details see Appendix A.

3. Molecular model

The two-center Lennard-Jones plus pointquadrupole (2CLJQ) potential consists of two identical Lennard-Jones sites, which are a distance L apart and which are described by energy and size parameters ϵ and σ, respectively. Furthermore, there is a pointquadrupole with the quadrupole moment Q placed at the geometrical center of the molecule. Hence, the 2CLJQ model has four parameters: ϵ, σ, L, and Q. The potential function u_{2CLJQ} is composed of a sum of two parts and depends on the distance $\mathbf{r}_{ij}$ between the centers of mass of two molecules i and j and their orientations $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$:

$$u_{2CLJQ}\left(\mathbf{r}_{ij}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_j\right) = u_{2CLJ}\left(\mathbf{r}_{ij}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_j\right) + u_Q\left(\mathbf{r}_{ij}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_j\right) \tag{2}$$

The first part u_{2CLJ} describes the Lennard-Jones potential for the four Lennard-Jones sites

$$u_{2CLJ}\left(\mathbf{r}_{ij}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_j\right) = -4\epsilon \sum_{a=1}^{2} \sum_{b=1}^{2} \left[\left(\frac{\sigma}{r_{ab}} \right)^6 - \left(\frac{\sigma}{r_{ab}} \right)^{12} \right] \tag{3}$$

where r_{ab} is the site-site distance between two Lennard-Jones sites.

The second part u_Q describes the interaction of the pointquadrupoles

$$u_Q\left(\mathbf{r}_{ij}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_j\right) = \frac{3}{4} \frac{Q^2}{|\mathbf{r}_{ij}|^5} f(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \tag{4}$$

where the function $f(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)$ depends only on the orientations of the molecules i and j. Detailed information about the 2CLJQ potential can e.g. be found in Gray and Gubbins [21].

4. Data basis and optimization task

4.1. Data basis

The experimental and simulation data sets are taken from the literature. For representing the experimental data, temperature dependent correlations from the DIPPR data bank [22] are used. The following uncertainties are reported: For the liquid density the uncertainties are < 1 % for N_2, O_2, F_2, Cl_2, C_2H_6, and C_2H_4 and < 3 % for Br_2, C_2H_2, C_2F_4, and C_2Cl_4. For the vapor pressure the uncertainties are < 1 % for N_2, O_2, Cl_2, C_2H_6, C_2H_4, and C_2H_2 and < 3 % for F_2, Br_2, C_2F_4, and C_2Cl_4. For the surface tension the uncertainties are < 3 % for N_2, Cl_2, Br_2, C_2H_6, C_2H_4, C_2H_2, and C_2Cl_4, < 1 % for O_2 and F_2 and < 5 % for C_2F_4. The values for the critical temperatures are also taken from the DIPPR data bank [22]. The uncertainties are < 3 % for N_2, O_2, F_2, Cl_2, and Br_2, < 0.2 % for C_2H_6, C_2H_4, C_2H_2, and C_2Cl_4 and < 1 % for C_2F_4.

The simulation data for the liquid density and the vapor pressure are taken from Stoll et al. [23] and for the surface tension from Werth et al. [24]. These authors report correlations in reduced units dependent on the reduced model parameters $Q^* = Q/\sqrt{\epsilon\sigma^5}$ and $L^* = L/\sigma$, and the reduced temperature $T^* = kT/\epsilon$, where k is Boltzmann's constant. The empirical correlations for the reduced liquid density $\rho^* = \rho\sigma^3$, the reduced vapor pressure $p^* = p^S\sigma^3/\epsilon$, and the reduced surface tension $\gamma^* = \gamma\sigma^2/\epsilon$ are based on simulations of 30 individual fluids and valid in a range of $0 \leq Q^* \leq 2$ and $0 \leq L^* \leq 0.8$ along the vapor-liquid coexistence curve for temperatures about $0.55\,T_C^* < T^* < 0.95\,T_C^*$. The statistical uncertainties reported for the liquid density are below 0.1 % for low temperatures and up to 3 % near the critical temperature. For the vapor pressure they are up to 20 % for low temperatures and about 2 % for higher temperatures. The correlations for both properties are reported to represent the simulation data within their uncertainties in most cases. For the surface tension the relative mean deviation between the correlation and the simulation results is 1.9 %. Also, the correlation is reported to agree with the simulation data within the statistical uncertainties.

The reduced critical temperature T_C^* is also taken from Stoll et al. [23] as a correlation dependent on the reduced model parameters Q^* and L^*. The reduced critical temperatures are estimated to be accurate up to the second decimal place by Stoll et al. [23]. The correlation is reported to fit the critical temperature within 0.5 % in most cases.

4.2. Optimization task

To evaluate the quality of a specific 2CLJQ model the correlations of the simulation data in reduced units are transferred to physical units with the corresponding values for the energy parameter ϵ and the length parameter σ. For each model a mean relative deviation function δO for each property O over different temperatures T_j, $j = 1, \ldots, N$ is evaluated. The objective space is spanned by the deviation functions for the chosen thermophysical properties. In this work, the objective function f_i for one property O has the following form:

$$(\delta O)^2 = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{O^{\mathrm{exp}}(T_j) - O^{\mathrm{sim}}(\epsilon, \sigma, L, Q, T_j)}{O^{\mathrm{exp}}(T_j)} \right)^2 \tag{5}$$

Ten equidistant temperatures between 0.55 T_C^{exp} and 0.95 T_C^{exp} are used to evaluate the mean relative deviation, with a maximal temperature T_{max}.

To ensure feasible parameter combinations, the following constraints are set:

$$\begin{aligned} 0 < \quad & L & < \quad & \sigma \\ 0 < \quad & kT_{\mathrm{max}}/T_C^*(Q^*, L^*) & < \quad & \epsilon \\ 0 < \quad & Q & & \end{aligned} \tag{6}$$

Additionally, the objective functions are constrained to ensure staying in a region of interest. This is demonstrated in Fig. 1 which shows the mean relative deviation for the liquid density $\delta\rho'$ versus the vapor pressure δp^S for C_2F_4 obtained for a two-criteria optimization task without constraints in the objective functions (left side) and with constraints in the objective functions (right side). In the unconstrained case, solutions with mean relative deviations in the liquid density of up to 12 % and in the vapor pressure of up to 180 % are found. These solutions represent compromises for which the quality of the

representation of one property is optimal, whereas it is poor for the second. In this work, we are not interested in these compromises and therefore choose to individually constrain our objective functions such that:

$$\begin{aligned} (\delta\rho)^2 &< (0.02)^2 \\ \left(\delta p^S\right)^2 &< (0.1)^2 \\ (\delta\gamma)^2 &< (0.2)^2 \end{aligned} \tag{7}$$

These choices reflect experiences from previous work on the attainable quality of representations with molecular models. In all cases, they are well above the expected limits and the desired quality, respectively.

We first conduct a multicriteria optimization of two objective functions, i.e. the mean relative deviation for the liquid density and the vapor pressure (*two-criteria scenario*). The resulting Pareto optimal models are also evaluated regarding their mean relative deviation in the surface tension and the relative deviation in the critical temperature, calculated as

$$\delta T_\mathrm{C} = \frac{T_\mathrm{C}^\mathrm{exp} - T_\mathrm{C}^\mathrm{sim}}{T_\mathrm{C}^\mathrm{exp}} \tag{8}$$

In a next step, the multicriteria optimization of three objective functions is investigated in which the surface tension is additionally considered together with the two other properties (*three-criteria scenario*). Again, the Pareto optimal models are also evaluated regarding their relative deviation in the critical temperature.

5. Self-organizing patch plot

We propose a new technique, the self-organizing patch plot (SOPP), to display Pareto sets, the corresponding parameters as well as additional functions evaluated for those parameters. The basic idea is, that for each of those quantities a separate diagram is provided. The diagrams each consist of a square area divided in a specific patch structure, which is identical in all diagrams. Each patch in this structure is assigned a data vector representing one specific Pareto optimal point. The vector components consist of the values in the quantities specific for this Pareto point, i.e. the objective function values, its corresponding parameters and the function values of the additional functions. Thus, the number of patches equals the number of Pareto solutions and the number of vector components equals the number of quantities of interest and thus the number of diagrams. In the different diagrams the patches are colored according to their value in the quantity assigned to the diagram, where the value is stored in the vectors assigned to the patches. The patch structure is unique for the Pareto data set and constructed such that the assigned vectors of neighboring patches are similar to each other in all vector components.

For the two- and three-criteria scenario results, eight quantities are considered: the four model parameters σ, ϵ, L and Q, the mean relative deviation in the liquid density, vapor pressure and surface tension, as well as the relative deviation in the critical temperature. Therefore eight diagrams are needed to fully display the results of one optimization scenario for one fluid (cf. Section 6). A SOPP using results of the two-criteria scenario for O_2 is exemplarily shown for the σ values on the right side in Fig. 2. Furthermore Fig. 2 shows the steps of how to create a SOPP. In the subsequent text, the steps consistently are first explained in a general form and then directly related to our specific application, before proceeding with the explanation of the next step of how to generate the SOPP. In total the process is explained in four steps.

Step 1 to generate a SOPP is to prepare a self-organizing map (SOM) for the quantities of interest. The SOM aims at representing the topology of a

high-dimensional discrete data set and arranging its representation in a low dimensional space. The concept to display this representation with a SOM is similar to the one introduced for the SOPPs: a number of diagrams, which is equal to the number of qualities, is prepared. The representation of an input data set is given as an discrete output data set, arranged in a low dimensional output space. To each location in this output space a vector with the values in the quantities of interest is assigned. The output data set has more data points than the input data set and interpolates the input data. Therefore it is not an exact representation, but an approximation of the input data. Further information about SOMs can e.g. be found in [9, 10].

For the two- and three-criteria scenario eight diagrams are necessary to fully display the approximation of the results with a SOM. We prepare our SOMs with a MATLAB implementation of Azzopardi [25], for which the detailed information can be found in the Supplementary Material. The output space is a two-dimensional grid. In Fig. 2 exemplarily the σ values of the representation are shown. Each grid point is colored according to its value in the vector component representing σ.

Step 2. The goal during generating the SOPP is to dispose of the approximative information and instead map the exact input data into the low-dimensional output space. This is achieved by first locating the best matching unit (BMU) for each input data point in the output space. The BMU is simply the vector in the output space that best matches the input vector regarding the Euclidean measure. Once the BMU is identified, the vectors in the output space are substituted by the exact input data vector and all residual output vectors are discarded. The result is a projection of the exact input data vectors into a low-dimensional output space. The advantage of using the SOM instead of conventional linear projection is that the locations of the projected data in the low-dimensional output space is generally non-overlapping, even though some data points might share identical values in some vector components.

In Fig. 2, the BMUs of the Pareto data of the two-criteria scenario are marked as black crosses in the SOM on the left side. Note, that the location of

the BMU is not only dependent on the σ value, but on the similarity in all eight components between the vector assigned to the grid point and the Pareto point mapped to the BMU. To each location of the black crosses one exact Pareto point with the full information in all eight quantities is mapped. The residual information in the SOM is dispensable.

Step 3. The output space now only consists of the data vectors assigned to the locations of the BMUs and has to be divided into patches, such that in each patch one of the data vectors is included. This is done with a Voronoi tessellation [14], which partitions a region in patches based on a set of so-called seeds. The seeds are points in the considered region. All other points in the region are each assigned to one seed, based on their distances to the seeds: The set of all points that are closer to one seed than to any other seed form the patch corresponding to this seed. The locations of the BMUs are taken as seeds to generate a Voronoi diagram and the data vectors of the BMUs are mapped to the corresponding patches.

In Fig. 2 in the center, the patches generated with the BMUs as seeds are shown. The Voronoi diagrams were computed with the algorithm implemented in MATLAB [26].

Step 4. Finally, the patches in the individual diagrams are colored in the value of the corresponding quality.

As mentioned above, for the results of the two-criteria scenario for O_2 this is exemplarily shown for the σ values in Fig. 2 on the right side.

Further details can be found in the Supplementary Material.

6. Results and discussion

First an overview of the results for the two-criteria scenario for all fluids is given. Then the results for C_2H_2 are presented and discussed in more detail for both the two- and three-criteria scenario. Finally, the results for both optimization scenarios are briefly discussed for all studied fluids.

The full Pareto sets for all studied fluids are available in the Supplementary Material.

6.1. Two-criteria scenario for all studied fluids

In Fig. 3 the Pareto set for each studied fluid for the two-criteria scenario is shown in the objective space (black line). The mean relative deviation of the vapor pressure is plotted over the mean relative deviation of the liquid density. All Pareto sets show similar behavior and can roughly be divided in three sections: Two sections represent extreme compromises. In the first, the mean relative deviation of the liquid density is good, but that of the vapor pressure is poor. In the second, the situation is reversed: The representation of the vapor pressure is good and that of the liquid density is poor. The third section connects the two others and is known as the "Pareto knee". In general, solutions in the region of the Pareto knee are most attractive, as in the other two sections only a minor deterioration in one objective has to be accepted to considerably enhance the other. This is also true for our results, as the slope of the Pareto set is either very large or almost zero in the other two sections.

Note, that the three sections are not necessarily detected equally well (e.g. for Br_2 one section is missing). In some cases this is due to numerical problems, which are discussed in greater detail in the Appendix.

Fig. 3 also shows that with the 2CLJQ model, mean relative deviations in the liquid density of about $0.1 - 0.3$ % are generally achievable. The respective value for the vapor pressure is about $0.4 - 0.7$ %. These numbers only hold if the single property is optimized. The compromises upon simultaneously optimizing both properties are somewhat worse: The solutions in the region of the Pareto knee typically have deviations of about 0.5 % in the liquid density and 2 % in

the vapor pressure. As an example, for each fluid a Pareto optimal solution in the region of the Pareto knee is marked in Fig. 3 (•). The corresponding numerical data is given in Table 1.

Fig. 3 also shows objective function values for some literature models of the fluids studied in the present work, for which numerical data is given in Table 2. The deviations indicated for the liquid density, the vapor pressure, the surface tension and the critical temperature for the literature models are obtained by evaluating Eqs. 5 and 8 for the parameters of these models under the same conditions as in the multicriteria optimization scenarios from this work (cf. Section 4.2). Not all models from Table 2 are depicted in Fig. 3, as some are out of the considered range and too far from the Pareto set.

When comparing these literature models to the results from the present work, it should be considered, that they were not necessarily optimized for the simultaneous representation of the liquid density and the vapor pressure, even though these properties are obviously important and are generally included in fits. The comparison shows that only in some cases the literature models lie close to the Pareto set and only rarely in the preferred region of the Pareto knee. This indicates that multicriteria optimization opens room for improvements in the parametrization of molecular models as compared to conventional optimization strategies based on a single objective function, even if that function includes different properties. A more detailed study of the literature models is presented in the Supplementary Material.

6.2. Two- and three-criteria scenario for C_2H_2

In Figs. 4 and 5 the results for C_2H_2 for the two- and three-criteria scenarios are shown in self-organizing patch plots. They relate the values of the deviations of the four different properties liquid density, vapor pressure, surface tension and critical temperature to the values of the parameters for the Pareto optimal models.

The results for the liquid density and the vapor pressure of the two-criteria scenario were already discussed above. The Pareto optimal models of the two-

criteria scenario (Fig. 4), optimized for the simultaneous representation in the liquid density and the vapor pressure have mean relative deviations in the surface tension of $20 - 35$ % and relative deviations in the critical temperature of up to almost ± 2 %. Including the surface tension in the optimization leads to improvements (cf. Fig. 5). Under the given constraints in the three-criteria scenario the mean relative deviation of the surface tension can be lowered to approximately 15 %. While the relative deviation in the critical temperature is consistently good, the improvement in the surface tension is only at the expense of accepting significantly increased deviations in the liquid density and the vapor pressure of about 2 % and 10 %, respectively. It should be noted that values for the mean relative deviation of the surface tension well below 5 % can be reached, but only if very large deviations in the liquid density and vapor pressure are accepted. However in this work, these Pareto optimal solutions are excluded by the imposed constraints on the objective functions (cf. Eqs. 7). Under these constraints the 2CLJQ model can simultaneously represent the liquid density and vapor pressure of C_2H_2 well, but this does not hold for the simultaneous representation of these two properties together with the surface tension.

The variations of the Pareto optimal model parameters for C_2H_2 are discussed in the region of the Pareto knee. They are about ± 0.5 % in σ, below 0.1 % in ϵ, ± 3.5 % in L, and ± 2 % in Q for the two-criteria scenario and about ± 2 % in σ, ± 12 % in ϵ, ± 15 % in L, and ± 3.5 % in Q for the three-criteria scenario. A significantly larger parameter variation is necessary to consider Pareto optimal solutions for three objective functions. For C_2H_2 the SOPPs reveal an inverse correlation between σ and L and between ϵ and Q.

The results for the quadrupole moment Q and the elongation L for C_2H_2 can be compared to experimental results from the literature. For the quadrupole moment of C_2H_2 experiments lead to values of $Q = (6.03 - 7.61)$ DÅ [27, 28, 29]. The quadrupole moments gained for C_2H_2 from our optimizations are about $Q = 5$ DÅ and thus somewhat lower. The C $\equiv$ C bond length is 1.203 Å [30], which compares reasonably to our value of about $L = 1.3$ Å. This shows that the 2CLJQ parameters determined with the present method are

effective parameters and should not be misinterpreted as physical quantities of real molecules. Nevertheless it is comforting to see that an adjustment of the parameters of that simple model to thermodynamic data leads to numbers of these parameters which are quite close to results of direct measurements of the related physical properties.

6.3. Two- and three-criteria scenario for all studied fluids

The results for the other substances studied here are similar as for C_2H_2. For all fluids, the two-criteria scenario yields good models for the liquid density and at the same time for the vapor pressure (cf. Fig. 3). The mean relative deviations for the surface tension obtained with these models are about $10 - 30$ %. By including the surface tension in the optimization, better results for the surface tension ($6 - 20$ %) are obtained but the possible improvements in the representation of the surface tension are only moderate while the loss in quality for the liquid density and vapor pressure is grave. Although this finding is interesting in itself, it is more important, that this can be shown and justified by using the comprehensive knowledge gained in the multicriteria optimization.

The inverse correlation between σ and L found for C_2H_2 can be generalized for all studied fluids: large values in σ lead to small values in L and vice versa. Although this is not unexpected, as both parameters relate to the size of the molecule, analyzing the Pareto sets shows this behavior explicitly: In Fig. 6 the Pareto sets of the two- and three-criteria scenarios for all studied fluids are plotted in a σ-L-diagram. Correlating the data linearly yields $\sigma_0 = L + (2.68 \pm 0.4)\sigma$, with a constant value σ_0 specific for the individual fluid. The correlation indicates, that the repulsive interactions could within certain limits be either described by a two-center Lennard-Jones model with elongation L or by a Lennard-Jones model with size σ_0. This knowledge could also be used to speed up parameter fitting for the 2CLJQ model.

The inverse correlation between ϵ and Q found for C_2H_2 cannot be generalized and is only found for some studied fluids (cf. Supplementary Material). Even though both parameters describe the attractive interactions, the fact that

no general trend is observed shows that it is reasonable to model the different types of attractive interactions (dispersion, polarity) separately.

7. Conclusions

In the present work, multicriteria optimization is applied to determine parameters of the 2CLJQ model for two and three objectives based on various thermophysical data (liquid density, vapor pressure and surface tension). A combination of sandwiching and hyperboxing is used for approximating the Pareto sets for ten real fluids.

The work shows that multicriteria optimization of molecular force field models is a promising approach. It gives an overview of what a model can yield in representing the various properties of a given fluid. If many fluids are studied, like in the present work, an overview is obtained, which quality of representation of the experimental data can be expected from the model in more general terms. All this is not possible with single criterion optimization which always yields just one result.

A novel method is introduced for presenting the Pareto set in the objective space and at the same time its picture in the design space: The self-organizing patch plot. It is a useful tool for exploring the Pareto optimal solutions and for choosing one solution which is most suitable to study a chosen application. Good choices often lie in the region of the so-called Pareto knee.

Furthermore, our results show, that with the 2CLJQ model the liquid density and the vapor pressure of many fluids can be well described. It is however not possible to obtain a good fit for the liquid density, the vapor pressure, and the surface tension simultaneously. The studied properties are only examples. Other properties can be included in the multicriteria optimization straightforwardly. A comparison with models from the literature shows that multicriteria optimization opens space for improvements.

The broad studies of the present work were only possible, because correlations for both simulation and experimental data were available. The available algorithms for determining Pareto sets are efficient and could probably be improved further. Pareto sets could therefore be determined on large computing facilities in the near future also if direct molecular simulations are needed.

The methods presented here can be applied and are attractive for optimizing parameters of any thermodynamic model.

Acknowledgements

The authors thank Kai Langenbach and Stephan Werth for fruitful discussions. We gratefully acknowledge financial support by DFG under the Reinhart Koselleck program (grant HA 1993/15-1). The present work was carried out under the auspices of the Boltzmann-Zuse Society for Computational Molecular Engineering (BZS).

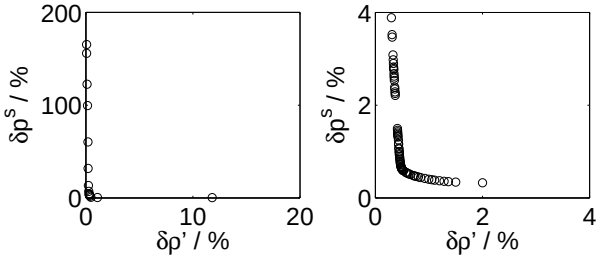

Figure 1: Results of the multicriteria optimization of C_2F_4 in the objective space spanned by the mean relative deviation of the liquid density $\delta\rho'$ and the vapor pressure δp^S. Left side: Unconstrained optimization. Right side: Constrained optimization.

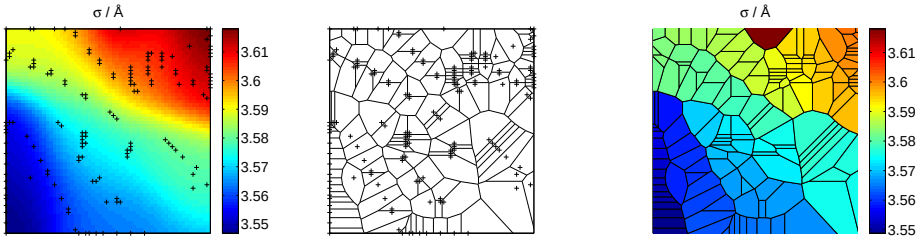

Figure 2: Steps to create a self-organizing patch plot for the two-criteria scenario for C_2H_2. Left: Best matching units BMUs of Pareto optimal solutions in the self-organizing map colored for the parameter values in σ. Middle: Voronoi patching with BMU as input data. Right: Patches colored according to their values in σ.

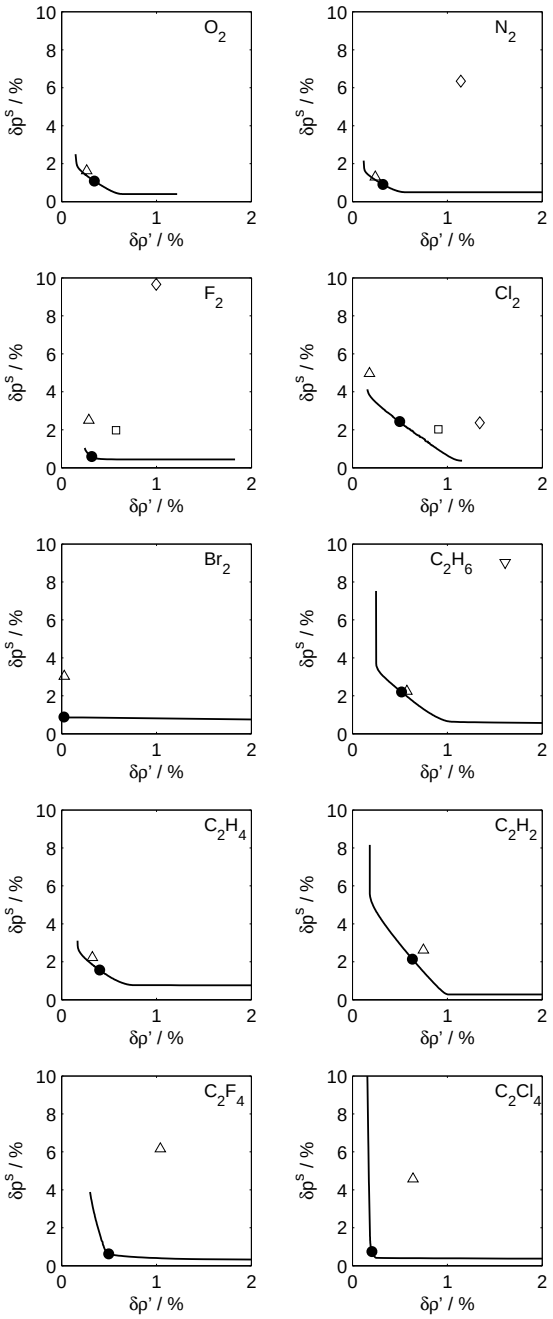

Figure 3: Results for all studied fluids for the two-criteria scenario: the mean relative deviation in the vapor pressure δp^S plotted over the mean relative deviation in the liquid density $\delta\rho'$ (black line). Models proposed in Table 1 (•). Performance of literature models (cf. Table 2): Vrabec et al. [31] (△), Singer et al. [32] (□), and McGuigan et al. [33] (♢)

.

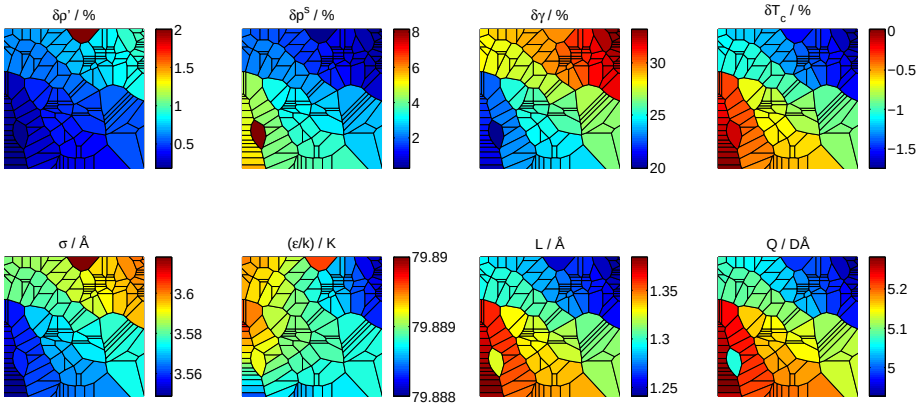

Figure 4: Pareto optimal solutions for the two-criteria scenario for C_2H_2 in a self-organizing patch plot. First row: mean relative deviation for the liquid density $\delta\rho'$, the vapor pressure δp^S, the surface tension $\delta\gamma$, and the relative deviation for the critical temperature δT_C. Second row: two-center Lennard-Jones plus pointquadrupole parameters energy ϵ, size σ, elongation L, and quadrupole moment Q.

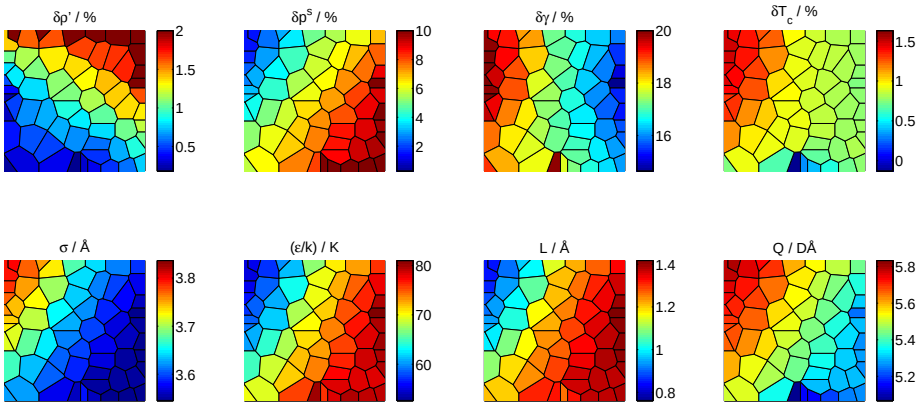

Figure 5: Pareto optimal solutions for the three-criteria scenario for C_2H_2 in a self-organizing patch plot. First row: mean relative deviation for the liquid density $\delta\rho'$, the vapor pressure δp^S, the surface tension $\delta\gamma$, and the relative deviation for the critical temperature δT_C. Second row: two-center Lennard-Jones plus pointquadrupole parameters energy ϵ, size σ, elongation L, and quadrupole moment Q.

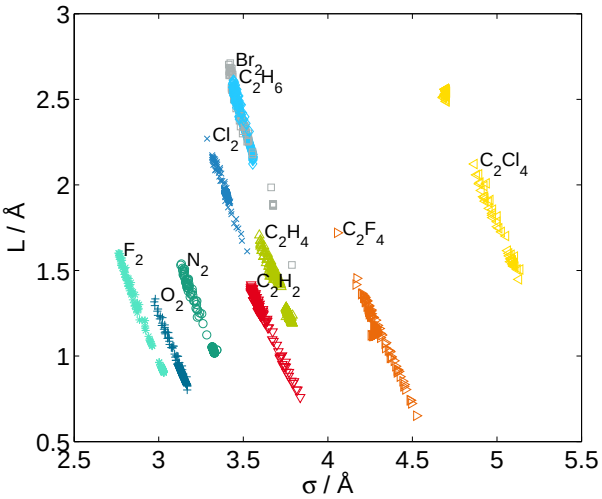

Figure 6: Pareto optimal model parameters σ and L for all studied fluids. The plot contains the Pareto solutions of the two- and three-criteria scenarios and shows linear correlations with similar slopes for all fluids.

Table 1: Selected compromises for the studied fluids for the two-criteria (I) and the three-criteria (II) scenario.

Fluid	Scenario	σ/ Å	(ϵ/k) K	L / Å	Q / DÅ	$\delta\rho'/\%$	$\delta p^S/\%$	$\delta\gamma/\%$	$\delta T_c/\%$
O_2	I	3.1339	41.7	0.9053	0.81	0.35	1.08	16.78	-0.63
	II	3.089	45.05	1.0578	0	2	1.85	9.23	0.02
N_2	I	3.3232	34.88	1.035	1.34	0.32	0.91	17.41	-0.53
	II	3.2621	38.21	1.1895	0	0.71	0.85	14.05	-0.54
F_2	I	3.0278	38.76	0.9088	1.35	0.32	0.59	18.08	-0.83
	II	2.8582	48.94	1.2943	0	0.45	8.92	6.07	0.05
Cl_2	I	3.4111	160.84	1.9203	3.6	0.5	2.44	11.22	-0.9
	II	3.3894	169.67	2.0056	0	2	3.8	6.99	-1.06
Br_2	I	3.6759	204.62	1.8879	6.33	0.02	0.88	10.35	-0.25
	II	3.4166	280.39	2.6404	0	0.57	5.04	6.07	-0.5
C_2H_6	I	3.4818	138.56	2.4161	1.45	0.52	2.2	22.44	-1.58
	II	3.4623	143.11	2.5577	0	2	4.9	13.72	-0.82
C_2H_4	I	3.7639	76.95	1.2531	4.22	0.4	1.56	26.34	-0.7
	II	3.7126	81.99	1.4421	4.4	2	3.71	15.83	0.3
C_2H_2	I	3.5819	79.89	1.2976	5.08	0.63	2.14	27.97	-1.08
	II	3.7504	59.71	0.9659	5.73	1.99	3.15	18.6	1.27
C_2F_4	I	4.2875	57.35	1.1295	8.28	0.5	0.62	28.08	1.98
	II	4.2228	69.78	1.3216	7.82	1.19	1.27	15.12	-1.02
C_2Cl_4	I	4.6997	211.1	2.5535	15.67	0.21	0.74	29.41	-1.87
	II	5.0852	132.45	1.5922	18.36	1.3	2	14.95	0.92

Table 2: Performance of literature models for objective and evaluation functions (Eqs. 5 and 8) considered in this work .

Fluid	Ref.	σ/ Å	(ϵ/k) K	L / Å	Q / DÅ	$\delta\rho'/\%$	$\delta p^S/\%$	$\delta\gamma/\%$	$\delta T_c/\%$
O_2	[31]	3.1062	43.183	0.9699	0.8081	0.26	1.63	15.63	-0.5
N_2	[33]	3.308	36.67	1.0883	–	1.14	6.35	26.26	-1.7
	[34]	3.314	35.3	1.101	1.48	2.62	6.64	7.67	0.91
	[34]	3.31	37.3	1.09	–	2.36	16.57	45.46	-3.4
	[35]	3.31	36	1.0897	–	2.14	6.13	7.09	0.19
	[31]	3.3211	34.897	1.0464	1.4397	0.24	1.3	16.68	-0.39
F_2	[32]	2.825	52.8	1.4266	–	0.57	1.97	13.38	-0.89
	[33]	2.832	53.47	1.4302	–	1	9.66	26.03	-2.17
	[31]	2.8258	52.147	1.4129	0.892	0.29	2.5	14.38	-0.73
Cl_2	[32]	3.353	173.5	2.0386	–	0.91	2.02	10.1	-1.5
	[33]	3.332	178.3	2.0992	–	1.34	2.37	12.82	-1.86
	[31]	3.4016	160.86	1.9819	4.2356	0.18	4.97	12.11	-0.36
Br_2	[32]	3.538	257.2	2.2289	–	3.53	21.52	44.71	-4.93
	[31]	3.5546	236.76	2.1777	4.8954	0.03	3.03	9.31	-0.4
C_2H_6	[36]	3.64	112.2	1.8382	–	1.49	14.54	16.9	-1.2
	[37]	3.825	100.6	1.54	–	2.94	8.1	38.63	-3.73
	[38]	3.52	137.5	2.45	–	4.47	5.27	9.34	-0.55
	[39]	3.75	98	1.54	–	0.22	27.79	6.84	0.06
	[40]	3.93	114	1.54	–	6.68	69.72	206.93	-19.31
	[41]	3.775	104.1	1.53	–	5.7	22.83	77.33	-6.94
	[42]	3.506	137.5	2.349	–	1.61	9.02	36.19	-3.05
	[31]	3.4896	136.99	2.3762	0.8277	0.57	2.24	22.04	-1.6
C_2H_4	[36]	3.63	91.4	1.48	–	1.05	12.29	23.05	-1.2
	[31]	3.7607	76.95	1.2695	4.331	0.32	2.23	25.98	-0.61
C_2H_2	[31]	3.5742	79.89	1.2998	5.073	0.75	2.62	27.92	-1.02
C_2F_4	[31]	3.8611	106.32	2.2394	7.0332	1.04	6.16	17.13	-1.01
C_2Cl_4	[31]	4.6758	211.11	2.652	16.143	0.64	4.57	22.58	-1.01

Appendix A. Numerical details

Appendix A.1. Approximation quality sandwiching and hyperboxing

The sandwiching and hyperboxing algorithm measures the quality of the approximated Pareto set by comparing an inner and outer approximation. The difference between those approximations can be calculated by different types of quality measures. In this work the chosen quality measure is *outerNadir*, which gives a relative deviation between the inner and outer approximation related to the Nadir point. The Nadir point is a fixed reference point used by the sandwiching and hyperboxing algorithm to find the Pareto set. A predefined approximation quality can be set as a stopping criterion: If so, the sandwiching and hyperboxing algorithm stops, when the approximation of the Pareto set reaches this quality. In this work the approximation quality was set to 10^{-5} for the two-criteria scenario and to 10^{-2} for the three-criteria scenario.

Appendix A.2. Solver settings NLPQLP

The NLPQLP solver is used to solve the single criterion optimization tasks identified by the sandwiching and hyperboxing algorithm. In this work, the termination quality for the NLPQLP algorithm is set to 10^{-11} for both optimization scenarios. For all studied fluids the model parameters from Vrabec et al. [31] are taken as starting values for the single criterion optimizations.

Appendix A.3. Numerical effort

The numerical effort can be characterized by the number of points needed to approximate the Pareto set with the specified quality and by the number of function calls of the NLPQLP solver to find one point of the Pareto set. For the two-criteria scenario on average over all studied fluids 142 points are needed to approximate the Pareto set. For the three-criteria scenario less Pareto points are calculated: On average only 57 points are needed, because a lower termination quality was set for the three-criteria scenario. The number of function calls by the NLPQLP on average over all studied fluids and Pareto points are 7.6 for the two-criteria scenario and 41.2 for the three-criteria scenario. The

significantly higher number of function calls for the three-criteria scenario indicates the difficulties of simultaneously optimizing the three considered objective functions under the given constraints.

Appendix A.4. Numerical problems

As already mentioned in the main text, the approximation of the Pareto set constructed by the sandwiching and hyperboxing algorithm may include solutions which are not Pareto optimal. Two reasons can be given for this behavior. Firstly, the choice of the termination criterion of the NLPQLP can be responsible. If it is fulfilled, the single criterion optimization terminates, but might not yet be in the true minimum for the current scalarization. To prevent this, the termination criterion can be lowered further. Then increased number of iterations have to be accepted. The second reason for not Pareto optimal solutions can be, that the objective function resulting from the scalarization for the single criterion optimization task is not necessarily convex, as assumed by the NLPQLP solver. Instead of finding the global minimum, the algorithm may get stuck in a local minimum. This can lead to large differences between the approximated and the true Pareto set. To avoid local minima e.g. the NLPQLP solver can be restarted with different starting values or different solver settings. Also, a different solver can be used, e.g. one with a global instead of a local optimization strategy. A further option is to check the parameter space in a region of interest by a brute force enumeration e.g. used in [1].

In this work, all Pareto sets were checked by comparison. Not Pareto optimal solutions were excluded. In the two-criteria scenario, for the fluids C_2H_2 and C_2Cl_4 no erroneous solutions were found. The average ratio of non-Pareto optimal solutions is 8.6 %. In the three-criteria scenario erroneous solutions were only found for Cl_2, Br_2, C_2F_4, and C_2Cl_4. The average ratio of non-Pareto optimal solutions is 1 %.

References

[1] K. Stöbener, P. Klein, S. Reiser, M. Horsch, K.-H. Küfer, H. Hasse, Fluid Phase Equilib. 373 (2014) 100–108.

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

[3] S. Mostaghim, M. Hoffmann, P. H. König, T. Frauenheim, J. Teich, Molecular force field parametrization using multi-objective evolutionary algorithms, in: Congress on Evolutionary Computation (CEC), volume 1, IEEE Conference Publications, 2004, pp. 212–219.

[4] J. I. Hernández, Multi-objective Optimization in Mixed Integer Problems: with application to the Beam Selection Optimization Problem in IMRT, *International Series in Operations Research and Management Science*, Mensch und Buch, Berlin, 2012.

[5] J. Fischer, B. Saager, M. Bohn, H. Oelschläger, J. M. Haile, Mol. Phys. 62 (1987) 1175–1185.

[6] B. Saager, J. Fischer, Fluid Phase Equilib. 66 (1991) 103 – 111.

[7] J. Fischer, R. Lustig, H. Breitenfelder-Manske, W. Lemming, Mol. Phys. 52 (1984) 485–497.

[8] M. Bohn, R. Lustig, J. Fischer, Fluid Phase Equilib. 25 (1986) 251 – 262.

[9] T. Kohonen, Biological Cybernetics 43 (1982) 59–69.

[10] T. Kohonen, Self-organizing Maps, Springer Series in Information Sciences, Springer, Heidelberg, 2001.

[11] J. Hunger, G. Huttner, J. Comput. Chem. 20 (1999) 455–471.

[12] S. Obayashi, D. Sasaki, Visualization and data mining of Pareto solutions using self-organizing map, in: Evolutionary multi-criterion optimization, Springer, Heidelberg, 2003, pp. 796–809.

[13] K. Chiba, Y. Makino, T. Takatoya, Design-Informatics Approach for Intimate Configuration of Silent Supersonic Technology Demonstrator, in: 47th AIAA Aerospace Sciences Meeting, American Institute of Aeronautics and Astronautics, 2009.

[14] F. Aurenhammer, R. Klein, D. Lee, Voronoi diagrams and Delaunay triangulations, World Scientific, Singapore, 2013.

[15] K. M. Miettinen, Nonlinear Multiobjective Optimization, volume 12 of *International Series in Operations Research and Management Science*, Kluwer Academic, Dordrecht, 1999.

[16] C. Hillermeier, Nonlinear Multiobjective Optimization, volume 135 of *International Series of Numerical Mathematics*, Birkhäuser, Basel, 2001.

[17] M. Ehrgott, Multicriteria Optimization, volume 491 of *Lecture Notes in Economics and Mathematical Systems*, Springer, Heidelberg, 2005.

[18] A. Pascoletti, P. Serafini, J. Optim. Theory Appl. 42 (1984) 499–524.

[19] M. Bortz, J. Burger, N. Aspiron, S. Blagov, R. Böttcher, U. Nowak, A. Scheithauer, R. Welke, K.-H. Küfer, H. Hasse, Comput. Chem. Eng. 60 (2014) 354–363.

[20] K. Schittkowski, NLPQLP: A Fortran implementation of a sequential quadratic programming algorithm with distributed and non-monotone line search, Version 4.2, http://www.klaus-schittkowski.de, 2014.

[21] C. G. Gray, K. E. Gubbins, Theory of Molecular Fluids. Volume 1: Fundamentals, Clarendon Press, Oxford, 1984.

[22] Design Institute for Physical Properties Sponsored by AIChE, DIPPR Project 801 - Full Version (2012).

[23] J. Stoll, J. Vrabec, H. Hasse, J. Fischer, Fluid Phase Equilib. 179 (2001) 339–362.

[24] S. Werth, M. Horsch, H. Hasse, Fluid Phase Equilib. 392 (2015) 12 – 18.

[25] G. Azzopardi, Self-organizing map - simple demonstration, `http://www.mathworks.com/matlabcentral/fileexchange`, 2013. File ID #39930.

[26] MATLAB, R2014a, The MathWorks Inc., Massachusetts, 2014.

[27] M. H. Coonan, G. L. D. Ritchie, Chem. Phys. Lett. 202 (1993) 237 – 241.

[28] A. J. Russell, M. A. Spackman, Mol. Phys. 88 (1996) 1109–1136.

[29] H. Kling, H. Geschka, W. Hüttner, Chem. Phys. Lett. 96 (1983) 631 – 635.

[30] K. P. E. Vollhardt, N. E. Schore, Organic Chemistry: Structure and Function, Freeman, New York, 2011.

[31] J. Vrabec, J. Stoll, H. Hasse, J. Phys. Chem. B 105 (2001) 12126–12133.

[32] K. Singer, A. Taylor, J. Singer, Mol. Phys. 33 (1977) 1757–1795.

[33] D. B. McGuigan, M. Lupkowski, D. M. Paquet, P. A. Monson, Mol. Phys. 67 (1989) 33–52.

[34] P. S. Y. Cheung, J. G. Powles, Mol. Phys. 32 (1976) 1383–1405.

[35] J. L. Rivera, J. Alejandre, S. K. Nath, J. J. de Pablo, Mol. Phys. 98 (2000) 43–55.

[36] T. Spyriouni, I. G. Economou, D. N. Theodorou, Phys. Rev. Lett. 80 (1998) 4466–4469.

[37] S. K. Nath, F. A. Escobedo, J. J. de Pablo, J. Chem. Phys. 108 (1998) 9905–9911.

[38] D. Fincham, N. Quirke, D. J. Tildesley, J. Chem. Phys. 84 (1986) 4535–4546.

[39] M. G. Martin, J. I. Siepmann, J. Phys. Chem. B 102 (1998) 2569–2577.

[40] J. I. Siepmann, S. Karaborni, B. Smit, Nature 365 (1993) 330–332.

[41] W. L. Jorgensen, J. D. Madura, C. J. Swenson, J. Am. Chem. Soc. 106 (1984) 6638–6646.

[42] M. Wojcik, K. Gubbins, J. Powles, Mol. Phys. 45 (1982) 1209–1225.