INPRODAT Innovationszentrum für Prozessdatentechnik e.V.

Technical report no. 2023–A
original version

Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)

Date: 9th January 2023

Authors: Horsch, M. T.; Schembera, B.

Funding information:

- DFG, MaRDI, project no. 460135501
- EC H2020, DOME 4.0, grant agreement no. 953163
- EC H2020, OntoCommons, grant agreement no. 958371

Accessibility:

- doi:10.5281/zenodo.7516532
- https://zenodo.org/communities/inprodat/

Epistemic metadata in molecular modelling: First-stage case-study report (10 cases)

by Martin Thomas Horsch and Björn Schembera

1. Introductory remarks

For this **first-stage report** of our ongoing case study, we went through a series of papers **without consulting these papers' authors** in any way. The aim is to restrict ourselves to analysing explicit statements from the papers under investigation. It will also have happened that we read something into the papers that was not there explicitly, and there may have been the occasional misunderstanding on what the authors intended to claim.

In the subsequent second stage, interviews with authors will be conducted and selected knowledge claims will be explored and documented/annotated in detail. Such discussions may influence the involved researchers: It may lead them to reconsider knowledge claims and other relevant aspects of their work. It will also affect our own point of view, *e.g.*, correcting misconceptions and introducing us to the broader context. Since the common reader's perspective cannot be restored after that, it is wise to write down our first, imperfect but undisturbed understanding at the outset of the case study.

The considered research papers from molecular modelling were selected as follows: First, two leading groups in the discipline at Imperial College London and at TU Berlin were chosen. Second, for each research group, the five papers with a corresponding author from that group, published in 2020, and with the highest citation count on Web of Science by the selection date in January 2022 were determined for inclusion. The limitation to papers with a corresponding author from the respective group is meant to simplify the process for interviews to be conducted as second-stage follow-up work. The selection criterion is a compromise between referring to recent work on the one hand, but being able to foresee what papers will have the greatest impact, so that the present case study (given that it certainly cannot be inclusive of molecular modelling as a whole) is not skewed toward methodologies that have fallen out of use or are of low impact.

Finally, this here is only a step toward sorting out concepts, and you will specially notice some inconsistency in the use of terms related to what papers and claims *are about*, such as the "subject matter" and the "object of research," as we are still working on these.

2. Synopsis: Taxonomy of epistemic metadata

Knowledge claim (KC)

- Conceptual knowledge claim (**CKC**)
- Property claim (**PC**)
    - Model property claim (**MPC**)
        - Abstract model property claim (**AMPC**)
        - Concrete model property claim (**CMPC**); example in 3.1
    - Physical property claim (**PPC**); example in 3.2
- Negative knowledge claim (**NKC**), expressing a gap in our knowledge

Provenance metadata

(beyond the scope of the present report, discussed in detail elsewhere already)

Validity claim (VC)

- Reproducibility claim (**RC**), further categorized as follows, in 8 combinations with a cube-like subsumption hiearchy: **RC**, **ERC**, **PRC**, **TRC**, **EPRC**, **ETRC**, **PTRC**, **EPTRC**.
    - Outcome orthodata-only agreement (ø) or exact agreement (+E)?
        - ERC characterized by absence of outcome paradata (only orthodata)
    - Provenance orthodata specified with which agreement is required (+P)?
        - PRC characterized by presence of provenance orthodata
    - Potentially done by the same team (ø), or only a new/changed team (+T)?
        - TRC characterized by requiring some significant change in the team
- Test-based validity claim (**TVC**), applying to both validation and testing
    - **MTVC**, from metric-quantitative testing (and validation) – how accurate is it?
    - **STVC**, successful positive outcome of qualitative "right/wrong" testing
    - **UTVC**, unsuccessful negative outcome of qualitative "right/wrong" testing
- Proof-based validity/correctness claim from verification (**PVC**)
    - **EPVC**, documenting exact correctness
    - **MPVC**, documenting inexact correctness quantitatively via an accuracy metric

Epistemic grounding

(not very straightforward to categorize, see the examples below)

(Blue: lead example from case study identified)

Remark: Reproducibility claims can be made *without actually doing* a reproducibility study. An actual reproduction attempt *examines* a validity/reproducibility claim, which may or may not have been explicitly stated. A revised validity/reproducibility claim is obtained from the reproduction attempt *as the interpretant* of a semiosis.

3. Analysis of the individual papers

3.1. Bowskill *et al.*, doi:10.1039/C9ME00089E

Topic/subject matter elements:

1. Organic Rankine cycle (ORC)
2. Working fluid (selection) – only pure fluids – depending on boundary conditions
3. Methodology for working fluid selection
4. SAFT-γ Mie EOS as a group-contribution based model

Research question:

- "Given a set of typical boundary conditions and a typical performance measure for an ORC, how can we proceed to select a working fluid based on a group-contribution based EOS? Can this be demonstrated for the SAFT-γ Mie EOS?"

Objects of research:

- (3) Methodology for working fluid selection
    - Optimization algorithm developed by adjusting "a standard outer-approximation algorithm framework"
- (4) SAFT-γ Mie EOS
    - Design space includes "58 960 candidate structures" (out of which first-stage brute-force exploration is applied to 3175); all can be modelled by the EOS

Epistemic grounding:

- SAFT-γ Mie model taken from established approach (previous work of the group). "In addition, a GC correlation developed by Hukkerikar *et al.* is used to estimate the melting point of each candidate working fluid."
- Other theoretical assumptions: Basic phenomenological thermodynamics (S2.2).
- Three test case scenarios assumed to be representative of real world problems
- From successful application to the test cases (only theoretically) to the conclusion that the methodology is viable

Results and knowledge claims:

- Novel methodology established, status would now be something like "a method that can be employed and has been tested (theory-level only) for realistic cases"
- Lists of best fluids & performance measure evaluation for the three scenarios
    - No uncertainties given – even minor differences used for ranking; **PPC example:** In case study 1, the best fluid is propane with 1.662 MW net power

    output, the second best is propene with 1.663 MW – we are not informed whether that plausibly proves that propane would be better than propene.
  - Experimental validation or other theoretical methods might reuse these results.
- “Computational performance” data (no. of function calls, not CPU times)

3.2. Chatwell and Vrabec, doi:10.1063/1.5142364

Topic/subject matter elements:

1. Considered property: Bulk viscosity (also known as volume viscosity)
2. Considered substances: Noble gases (liquid phase)
3. Newly developed/parameterized model: EOS for the volume viscosity
4. Auxiliary model: LJ fluid
   - The LJ potential is used to supplement experimental data; it is also of interest as such (i.e., we would also like to know the volume viscosity of the LJ fluid)

Research questions:

- “What is the volume viscosity of the liquid phase of noble gases (and/or the LJ fluid) as a function of thermodynamic boundary conditions?”
- “How can we correlate the volume viscosity of noble gases (and LJ) by an EOS?”
  - Observe that the research question as such presupposes that for the present purpose, the noble gases (and, possibly, the LJ model) are a single object of research and any deviations from corresponding states theory are negligible.

Objects of research:

- The main object of research is (1) the bulk viscosity (volume viscosity) as such.
  - Why? a) Sections I and II make it clear that this work deals with that property as its object of research, and b) It is comparably rare to investigate this property.
  - Note how in most other cases we would include the considered thermodynamic properties in a descriptor of what the study is about, but we would not attribute the status of “object of research” to the property as such.
- Secondary objects of research may include (2) the noble gases and (4) the LJ fluid.

Epistemic grounding:

- Experimental data from the literature accepted as basis of research
  - However, “while all available sound attenuation measurements for neon, argon, krypton and xenon were evaluated, data subject to critical attenuation were identified and discarded”

- From experimental viscosity for noble gases to a LJ representation "by a canonical transformation" (reduction by σ, ε, m) with parameters from Rutkai et al. (2017); but density, pressure, temperature are reduced with respect to the critical point.
- EMD simulation of LJ: Solver parameter $r_c = 5.5\sigma$ justified as being "sufficient" with three literature references; similar grounding of other methodological choices
- Novel part of methodology (relaxation model) based on pre-existing theoretical knowledge (according to which "each mode decays exponentially over time following a Kohlrausch-Williams-Watts function")
- Viability of the method made plausible by comparing results with and without using the relaxation model; results look more precise/reliable if the model is used.
- Functional form of the EOS (correlation) made plausible by theoretical arguments
- Parameterization of the correlation/EOS from fit to experimental+simulation data.

Results and knowledge claims:

- Dimensionless experimental volume visosities are assigned an uncertainty following linear propagation as given in Eq. (10); these uncertainties turn out to be rather significant (see Fig. 6).
- All experimental and simulation based values, including uncertainties, are given in tables in the supplementary information. While simulation uncertainties are provided (both in the tables and figures), it is not made clear how they were defined and computed.
- The resulting correlation/EOS parameters are given as such, without a confidence interval, and the accuracy expected from the correlation is also not quantified.

3.3. Fingerhut *et al.*, doi:10.1080/00268976.2019.1643046

Topic/subject matter elements:

1. Thermodynamic factor matrix Γ for multicomponent mixtures
2. Kirkwood-Buff integrals

- Both of these elements can also legitimately be considered the *object of research*

Research question:

A) How can the Kirkwood-Buff integral (KBI) approach be applied to thermodynamic correction factors for diffusion coefficients?
   - Or: By what method(s) can we compute the thermodynamic correction factors for diffusion coefficients in multicomponent mixtures?
   - Or: Is there a new method, using Kirkwood-Buff integrals, by which we can …?

B) Can such a new method be successfully demonstrated?

Results, knowledge claims, epistemic grounding:

- Expressions based on Kirkwood-Buff integrals for the thermodynamic correction factor matrix Γ for diffusion coefficients; mathematical results, hence exact
  - Grounding: Mathematical deduction, part of which is laid out in the Appendix
    - Basis: Previous work by Ben Naim on KBI expression for $d\mu_i/dN_j$
  - The authors point out that Γ cannot be accessed directly by experiment
- The method is validated for a sample quaternary mixture by comparison with chemical potentials from Widom test-particle insertion

3.4. Guevara Carrión *et al.*, doi:10.1021/acs.jpcb.0c01625d

Topical elements and/or objects of research:

1. Green-Kubo EMD methodology
2. Fick diffusion coefficient matrix for a multicomponent mixture
3. Water + methanol + ethanol + isopropanol mixtures (and ternary subsystems)
4. Darken correction, relation between Fick and M-S or Onsager coefficients
5. Finite size effects in molecular simulation (with periodic boundary condition)

What are the main research outcomes?

Q (research question or subject matter): What is the diffusion coefficient matrix in the liquid phase, for the molecular models as mentioned above applied to the ternary and quaternary mixtures as mentioned above, for a given composition, pressure, and temperature? (For p = 100 kPa, T = 298.15 K, and $\mathbf{x}$ on a grid as shown in Fig. 2.)
O (object of research):

- The knowledge claims most directly refer to the model: TIP4P/2005 for water, methanol and ethanol from Schnabel *et al.*, isopropanol from Muñoz Muñoz *et al.*, unlike interaction from the unmodified Lorentz-Berthelot rule.

More indirectly the knowledge claims must be taken to also refer to the modelled systems, the actual water-alcohol mixtures. Nobody would be interested in these models and their behaviour as such if they did not represent actual fluids. But there the uncertainty will include an additional contribution, from the model inaccuracy, that is not taken into account for the uncertainties shown here.
K (knowledge claim): Results e.g. given in Tab. 1, also in the Supplementary Information. Uncertainty given in parentheses; “uncertainties of the predicted values were estimated with a block averaging method” (SI), referencing Allen-Tildesley; but it is not stated

whether these uncertainties are single, double, or triple standard deviations. It is also not stated whether there is a contribution to uncertainty from the system size correction.
G (grounding): Code and method are well established, with the exception of the KB part of the method for which some validation is provided (see below). Model accuracy only asserted qualitatively: It "has been shown that all molecular models are suitable".

Q: For the same as above, thermodynamic factor matrix for the Darken correction. (Intermediate result, but potentially also of interest as such.)
O: Same as above.
K: Results are reported mainly in the Supplementary Information, graphically only, about the uncertainty it is just stated that it is within symbol size.
G: Method, see below. Model accuracy, see above.

Q: What is a good methodology for obtaining Fick diffusion coefficients in multicomponent mixtures by EMD simulation?
O: The object of research is the Fick diffusion coefficient matrix as such.
K: The interesting parts of the methodology are, first, the explicit inclusion of a finite-size correction, where it is specifically novel that this correction is applied to the Onsager coefficients, and second, obtaining the Darken correction from the KB integrals.
G: The KB part of the method is validated against "the Wilson excess Gibbs energy model fitted to binary simulation results," showing good agreement. However, it is not clear what should make us accept the finite-size methodology as correct. It yields a correction of 6% in the Onsager coefficients whereas the "correction following Yeh and Hummer would have led to corrections of around 15%." It is based on a linear regression in $N^{-1/3}$ which looks like an ad hoc fit.

3.5. Haslam *et al.*, doi:10.1021/acs.jced.0c00746

Topical elements:

1. Considered model (or class of models): SAFT-γ Mie EOS
2. Many fluids and fluid mixtures are considered (new group parameters)
3. A variety of properties including VLE data (but also others, see below)
4. Accuracy of the SAFT-γ Mie EOS with the extended set of parameters

Research questions:

- "What would be good group parameters for like and unlike interactions in the SAFT-γ Mie EOS, for overall 58 groups to be considered?" (thereof, about a quarter taken over from previous work)

- "What accuracy can be reached in this way for the most commonly required thermodynamic properties?"

Objects of research:

- Element (1) from above, the SAFT-γ Mie EOS, is the object of research

Epistemic grounding:

- Incremental work: Methodology and pre-existing parameter set taken from previous work of the group
- From optimal agreement with experimental data to optimal model parameters "estimated by optimizing the description of target experimental data" – the exact optimization methodology and critieria are not stated explicitly; the reader is expected to accept without further proof that these are the optimal parameters
- "Illustration of performance" of the extended parameter set, yielding a documented good level of accuracy (agreement with experimental VLE data) – or in case of ion models, agreement with experimental osmotic coefficients and activity coefficients – can this be used as a level of confidence in the model?
    - The deviations are given as "absolute average deviations" ("AAD"), in percent – are the authors using the wrong term? Looks like an average relative deviation.
    - Additionally, LLE data are shown (Fig. 3). Agreement there is rather poor.
    - There is no distinction between validation and test sets or similar; it is probable, or at least not explicitly denied, that the values from the "illustration of performance" are the same to which the model was also adjusted

Results and knowledge claims:

- New group contribution parameters ingested into the SAFT-γ Mie group database
    - "The resulting group parameter table presented here now contains 58 groups, increasing the size of the previous table by a factor of four"
    - This includes unlike interaction parameters, where it was possible to parameterize them; in other cases, a mixing rule is employed
- Deviations ("AAD") for the extended system of model parameters
    - Vapour pressures, enthalpies of vaporization, and isobaric heat capacities are reproduced within 5% in most cases, liquid densities within 2%.
    - It is not established whether this is also the accuracy that is to be expected for predictions (no distinction between validation/test, cross-validation, or similar)

3.6. Lee *et al.*, doi:10.1016/j.compchemeng.2020.106802

Subject matter, research question, and object of research:

- Topic elements:
  - Multiple-Objective Optimization for molecular and process design
- Research Question
  - systematic comparison of the performance of five mixed-integer non-linear programming (MINLP) MOO algorithms on the selection of computer-aided molecular design (CAMD) and computer-aided molecular and process design (CAMPD) problems
- Object of Research
  - CAMD
  - CAMPD
  - MINLP MOO algorithms are WS, SD, NSGA-II

3.7. Malviya and Vrabec, doi:10.1021/acs.jced.9b00571

Subject matter, research question, and object of research:

- Topical elements:

  1. Thermodynamic system consisting of:

     a) The considered solvent, H_2O + MeOH + EtOH at various compositions (including pure and binary subsystems)

     b) Solute N_2, or O_2, or Ar, all at infinite dilution

     At various temperatures

  2. Employed molecular models for the thermodynamic system:

     a) Models for the pure fluids, from previous work

     b) Mixture models based on binary (modified LB) interaction parameters – partly adjusted as an outcome of this work, see Tab. 1

  3. Thermodynamic property: Henry's law coefficients

- Research question – primary (A), secondary (B):

  A) "What is the Henry's law coefficient of N_2/O_2/Ar in the considered solvents as a function of composition and temperature?"

  B) "What would be good models for the considered solvent+solute mixtures?"

- Objects of research corresponding to research questions A and B:

  A) The considered solvents (1a) and the considered solutes (1b)

  B) The solvent-solute mixture models (2b) and mixtures (1, or 1a + 1b)

Epistemic grounding:

- Validity of the molecular models of pure fluids: From previous work “because they yield a very good agreement with experimental pure solvent properties”; similarly, binary interaction parameters taken from literature/previous work where available. In the literature/previous work, these binary interaction parameters had been adjusted to Henry coefficients.
- As simplification or part of the approach, the binary interaction parameter “values for the solvent-solvent interactions were set to unity, assuming fully predictive liquid mixtures” – no justification, as this is a basic assumption or presupposition
- From experimental data to model parameters, in the five binary solvent+solute subsystems where binary interaction parameters were needed; each parameter is adjusted to a single experimental data point (Henry or other VLE data)
- For pure solvents, agreement of simulation & experiment carries little value (“Due to the fact that the binary interaction parameter was adjusted to experimental H_i data, the data sets from simulation and experiment agree quite well.”)
- For binary solvent mixtures, as far as experimental data are available, models might be validated by demonstrating their predictive capacity; here, the general tendency is to confirm validity of models when acceptable agreement is reached.
- However, the authors *cast doubt on the experimental data* by Tokunaga where the agreement is poor – this sort of arbitration between simulation and experiment in case of a confluct is moderated by independent critical judgment (the experimental data look unusual); combined with deviant simulation results, the validity of the experimental data comes under attack
- Thermodynamic property predictions from molecular modelling and simulation: Henry’s law coefficients obtained
- From properties predicted by simulation to phenomenological correlations over the entire relevant thermodynamic parameter space

Results and knowledge claims:

- Validity of the method corroborated (explicit formulation even: “shown”, not just corroborated): “It was shown that molecular simulation is a reliable method for investigating the Henry’s law constant of gases dissolved in liquid solvents.”
- Model parameters: New binary interaction parameters for five binary mixtures
- Thermodynamic data from simulation: Henry’s law coefficients for certain systems under certain conditions
- Higher-order model/correlation: Empirical correlation expression for the Henry’s law coefficients

3.8. Šarić *et al.*, doi:10.1063/1.5144991

Subject matter, research question, and object of research:

- According to the title, it is "a force field assessment," hence, *mainly about* force fields and their accuracy. The opener of the abstract states that "the concentration dependence of the dielectric constant and the density of 11 aqueous alkali halide solutions (LiCl, NaCl, KCl, RbCl, CsCl, LiI, NaI, KI, CsI, KF, and CsF) is investigated" which makes it *mainly about* the dielectric constant and density of these solutions.
- More clearly: "No study of the dielectric constant covering a wide range of alkali halide salts in aqueous solutions, together with a comparison of several different ion force fields, has been presented to date. The main objective of this paper is thus to provide such a comprehensive comparison. The performance of eight non-polarizable ion force fields combined with the TIP4P/ε water model was assessed with respect to the prediction of the dielectric constant and density of 11 aqueous alkali halide solutions (LiCl, NaCl, KCl, RbCl, CsCl, LiI, NaI, KI, CsI, KF, and CsF)."
- To correctly describe the main topic, we need to combine four elements:
  1. The considered ion models in joint use with the TIP4P/ε water model;
     - note how the paper also reports results using SPC/E water but based on its explicit assertions cannot be taken to be *about* that model;
  2. The considered mixtures, with thermodynamic boundary conditions;
  3. The considered properties (dielectric constant and density);
  4. Predictive accuracy as the main aspect under which this is discussed.
  - It is only this specific combination of these four elements that the paper is about.
  - Observe how the topic or subject matter can here be described by an addressed research question: "What is the predictive accuracy of the molecular models … for the properties … of the substance/mixtures … under the thermodynamic boundary conditions …?" It also has the form of a competency question for ontologizing objects of research.
  - What are the objects of research, *i.e.*, the main referents within the cognitive process? What is "the thing that is investigated"? Intuitively we might say: 1. the models and 2. the mixtures, but *not* 3. the properties and 4. the accuracy. Why
    - Out of the four elements of the topic descriptor, it looks as if we were interested in (3) and (4) only contextually (not in "density" as such but only in "density of X"), whereas we are interested in (1) and (2) not only contextually, but also as such.
    - It is also because we would tend to include (2) the mixtures in the role of a referent only, (1) the ion models both as a referent and a representamen,

but (3) the properties and (4) the accuracy only in the role of a representamen.

Epistemic grounding:

- “These salts were chosen as they are the only ones out of the 20 alkali halides for which reliable experimental data are available for comparison.” → The available pre-existing experimental data are accepted as a reliable point of departure.
- From experimental data to model accuracy, from quantitative accuracy comparison for multiple models to qualitative assessment of a model (TIP4P/ε as opposed to SPC/E water): “The role of the underlying water model was investigated by comparing dielectric constant predictions for aqueous NaCl solutions at 298 K using different ion force fields together with the TIP4P/ε or the SPC/E water model. As TIP4P/ε yielded better predictions for all investigated ion force fields, all subsequent studies were carried out with that water model.”
- Validity of methodology accepted by (researcher interpretation of) previous work:
  - ”none of the ion force fields studied in the present work was adjusted explicitly for the use with the TIP4P/ε water model” but various previous studies reapplied ion models from one water model to another, and it worked (“good compatiblity”, “similar results”). “Hence, combining different ion model sets with the TIP4P/ε water model appears to be reasonable.”
  - Applicability of some of the models under certain conditions questionable due to formation of clusters (taken here as potentially anticipating phase separation), concluding that “the seemingly good reproduction of both the dielectric constant and molar density with some ion models is, therefore, probably associated with premature crystallization. Hence, it may be only a coincidence of an unphysical behavior of the respective model, which is thus not useful.”

Results and knowledge claims:

- Qualitative preferability of one water model (TIP4P/ε) over another (SPC/E) for a particular purpose – predicting the dielectric behaviour of electrolyte solutions.
  - Subject matter the same as for paper? Yes; the simulations are done for a restricted set of boundary conditions, but the claim is held to be generalizable.
- Quantitative accuracy, “mean relative error (MRE),” for the dielectric constant and the density of aqueous alkali halide solutions using eight ion force fields.
  - This is the main knowledge claim from the paper – recognizable by its being the answer to the main competency/research question.
  - As a quantitative knowledge claim, this includes numerical data.

- Qualitative inaccuracy of some models found due to “premature crystallization”.
  - Subject matter of this knowledge claim: While it is primarily applied to the research question, it is held to be more general (not only density and dielectric constant are not reflected, the inaccuracy goes beyond this).

- **CMPC example:** Tab. III shows the number of contact ion pairs from a variety of force fields, all referring to actual aqueous salt solutions.

3.9. Zheng *et al.*, doi:10.3390/en13112770

Q (research questions or subject matter): How to represent carbonate rock surfaces, and their wettability by hydrocarbons, through molecular/mesoscopic modelling. How does the roughness of carbonate-rock like materials influence the contact angle of oil?
O (object of research): Carbonate rock wettability by oils (properties); methodology for modelling and simulating the contact angle of rough surfaces.
K (knowledge claim): Results shown in Fig. 7 and also Tab. 1; error bars are included in the figure, some of which are surprisingly small.
G (grounding):

- AFM data on an actual rock surface → Fourier transform of the surface → characteristic wavelength Λ = 7.55 nm and amplitude 4 nm → molecular level (σ = 0.33 nm) model surfaces with the same amplitude and a spectrum of wavelengths, including the one from the actual physical sample.
- SAFT-based Mie force field parameterized to *n*-decane, from the molecular EOS, based on previous knowledge that EOS-based model parameterization usually carries over with good accuracy to molecular simulation.
- Fluid-solid interaction “arbitrarily taken in this case to provide a contact angle of the fluid on the flat surface of around 90°, *i.e.*, neutrally wetting conditions”.
- Finite-size effect checked by scaling up to N = 50 000, showing convergence.

3.10. Zhu and Müller, doi:10.1021/acs.jpcb.0c05806

Topic/subject matter:

- Equations of State (EoS) for fluids are considered are central and require an enormous design effort. In this paper it should be shown how and evaluated if machine learning techniques (specifically ANN and GPR) can be used to speed up and/or improve the design process of EoS
- Lack of Data is a problem: “The current engineering folklore for dealing with this lack of data is to use empirical correlations and/or fitted simplified theories to essentially interpolate and/or extend the results. Historians of thermodynamics

have given a detailed account of the quest of scientists to rationalize the interrelation between the thermophysical properties of fluids." → "need for correlation"

- It is examined if "the ability of both nonlinear ANN and GPR to correlate thermodynamic properties in the way a traditional EoS would."
- Exploring the feasibility of being able to correlate massive amounts of physical property data both rapidly and effectively

Research questions:

- Effectiveness of a machine learned model to replicate the statistical associating fluid theory (SAFT-VR Mie)
- "exploring the question of whether current ML approaches can be employed to substitute for analytical EoS"
- Is the error associated with the correlation acceptable?
- Are ANNs the optimal choice for an ML algorithm for this purpose?

Objects of research:

- EOS, specifically SAFT-VR Mie
- ML:ANN/GPR

Epistemic grounding:

- "start with an expected simplified molecular model and develop corresponding approximations in order to arrive at a closed-form expression. The resulting model is further commonly employed as a correlation tool, with parameters fit to experimental properties"
- "In a typical use of Mie force-field parameters (σ, ε, λ_r, λ_a) to represent a fluid, we look to the SAFT-VR Mie EoS to estimate the parameters that provide the best representation of available macroscopic experimental data. [...] As the equation of state explicitly calculates Helmholtz free energy, macroscopic thermodynamic properties such as the vapor pressure P_v, saturated liquid density ρ_L and equilibrium vapor−liquid conditions can be derived"
- Experimental data is noisy and poorly distributed; large abundance in data points; Since the ML algorithm relies on the processing of large amounts of data, the direct use of experimental data might turn out to produce inconclusive results, as the typical data points will be biased in quantity and quality toward the experimentally "easier" state points;
- By generating a larger database of pseudoexperimental data employing the SAFT EoS, we remove these limitations and biases and gauge the ability of ML models to both recognize and correlate the data

- By introducing prior scientific knowledge of the system behavior into transforming the data, the data-driven model can be trained more effectively to the desired correlations we wish to observe.

Results and knowledge claims:

- Critical properties: The ANNML model was fitted to predict the critical pressure and critical temperature with a statistical high accuracy. GPR can also predict these critical points with a similar performance.
- Vapor pressure: Data transformation is conducted based on the Clausius-Clapeyron equation. For the normal model it suggests a good model fit for the vapor pressure.
- Saturated densities: ANN is also employed to predict saturated densities. Good model performance and some tweaking needed to capture the VLE envelope shape. GPR model fails to capture the VLE shape even when including critical points and employing over 2000 data points in the fitting process.
- Critical density: Many more training iterations are required to a achieve a good accuracy.
- Through the analysis of ML models for different thermodynamic problems, we conclude that both ANN and GPR can be used effectively as surrogates for analytical EoS.
- GPR needs much less data to achieve working accuracy, but ANN fitting proves to be a more flexible, robust and accurate approach.
- Although no attempt is made here to match the experimental data of real compounds, a value of m_s = 20 would roughly correspond to a 60 carbon linear alkane chain (n-hexacontane), and repulsive exponents of $\lambda_r > 20$ are useful to describe highly fluorinated compounds, while a soft potential $\lambda_r = 8$ is essential for modeling water.
- The impressive point here is the relative ease with which ML models can both “develop” and “learn” an equation of state, which in our experience typically requires years of dedicated effort.
- The other side of the coin is that ML approaches, by their nature, are devoid of any physics-based insights (other than the scaling of the features) and as such cannot provide for anything other than a simple and accurate correlation of the data.

Acknowledgment. Co-author B.S. acknowledges funding from the NFDI programme of the Joint Science Conference (GWK) within DFG project no. 460135501 (MaRDI). Co-author M.T.H. acknowledges funding from the EU’s Horizon 2020 research and innovation programme under grant agreements no. 953163 (DOME 4.0), and 958371 (OntoCommons) and activities of Inprodat e.V.