



Cite this: DOI: 10.1039/d6me00094k

Randomness is all you need – studying the effect of molecular conformation on sigma profiles as machine learning features

Fathya Y. M. Salih, ^a Dinis O. Abranches, ^b
Edward J. Maginn ^a and Yamil J. Colón ^{*a}

Sigma profiles (SPs) are powerful molecular descriptors that combine the size-independence and continuity of many machine-learned descriptors with the physical interpretability of structural descriptors like molecular graphs or SMILES. Though SPs have existed for a long time within the context of thermodynamic models like COSMO-RS and COSMO-SAC, there does not seem to be a consensus in the literature on the effect of molecular conformations on SPs or on their performance as inputs to thermodynamic or machine learning (ML) models. This work performs a systematic study on the effect of conformations on SPs as an input to ML models by developing three different conformer generation heuristics and using the resulting SP datasets to train Gaussian process (GP) models as an inexpensive surrogate ML model. All SPs are generated using the open-source tool OpenSPGen following its suggested guidelines for level of theory, tessellation, and post-processing (averaging radius). The study shows that conformations have a small but statistically significant effect on the ML prediction of three out of six considered target properties. Additionally, it shows that the effect of conformations is more significant for smaller molecules, and that the number of conformations included is more important than the heuristic of conformer generation. The final recommendation of this work is to start by generating five random conformers, examine how the ML performance for the target property of interest changes with increasing the number of random conformers, then deciding if the increase in performance from one to three to five conformers justifies generating more conformations and more SPs. A simple computational cost formula depending on molecule size was developed to aide in this decision.

Received 11th June 2026,
Accepted 1st September 2026

DOI: 10.1039/d6me00094k

rsc.li/molecular-engineering

Design, System, Application

Sigma profiles (SPs) are a powerful molecular descriptor for physics-informed machine learning (ML) for molecular design applications because they uniquely combine physical significance and human-interpretability with continuity and size-independence. However, using SPs in ML applications requires guidelines for open-source SP generation at scale. This work develops guidelines for conformer generation and selection to yield the highest-performing SPs for ML applications. A systematic study on the effect of conformations on SPs and their performance as an ML descriptor was performed. The resulting guidelines were implemented as a default option in the open-source SP generation package OpenSPGen allowing consistent treatment of conformations for large-scale SP production.

1 Introduction

Sigma profiles (SPs) are un-normalized histograms of screening charge distribution for a molecule embedded in a conductor-like dielectric medium, *i.e.* a continuum solvent with an infinite dielectric constant.^{1,2} They were developed as part of a thermodynamic framework for evaluating activity coefficients in real solvents called COSMO-RS (COnductor-like Screening MOdel for Real Solvents).^{3,4}

Beyond this conventional use, SPs have drawn interest in machine learning (ML) applications as molecular descriptors (ways of encoding molecules into numerical tensors). In the ML context, SPs offer a unique combination of advantages as molecular descriptors. They combine the physical significance and human-interpretability found in structural descriptors like molecular graphs or fingerprints with the size-independence and continuity found in machine-learned descriptors like autoencoder latent vectors⁵ or chemical foundation models.⁶ Our previous studies have shown that SPs can predict a variety of thermophysical properties when used as inputs to convolutional neural networks (CNNs)⁷ and non-parametric Gaussian process (GP) models with relatively small datasets (~300–1200 labels).⁸ The latter of

^a Department of Chemical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. E-mail: ycolon@nd.edu

^b CICECO – Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Aveiro, 3810-193, Portugal

these works, also, demonstrated the compatibility of SPs with active learning or Bayesian optimization frameworks,⁸ meaning SPs in ML could be the path forward to interpretable and data-scarce molecular design. In the broader literature SPs have been used to predict many properties including thermodynamic properties (e.g. density,^{9,10} solubility,^{11–16} surface tension,^{17,18} Henry's constants,^{19,20} etc.), dynamic/transport properties (e.g. viscosity and ionic conductivity),^{9,10,21} and even higher-level properties further removed from activity coefficients (e.g. toxicity,^{22,23} and odor characteristics).²⁴ These studies span the use of SPs as inputs to both parametric and neural network models^{7,12–14,17,18,21,23,24} as well as probabilistic models.^{8,10,14,20,22} This unique combination of advantages that SPs offer motivated us to improve the accessibility of SPs by developing OpenSPGen – the first end-to-end open-source SP generation tool.²⁵ In that work, in addition to developing the SP generator, a series of parametric studies was conducted to provide recommendations for SP generation settings (e.g. level of theory, segment size, etc.) that ensure the best ML performance for thermophysical property prediction. However, the effect of conformations on SPs and ML performance was not addressed in that work – that is the goal of this paper.

The conformation of a molecule affects the resulting molecular surface and screening charge, and in turn, the SP. This eventually affects the predicted properties, and this has been known and reported for a variety of properties from saturation pressure^{26–28} to aqueous solubility^{29,30} and micellar partition coefficients.^{31–33} For example, Kurtén *et al.*²⁶ used COSMOconf (a software package capable of generating different conformers for the same molecular structure)³⁴ for conformer sampling and COSMO-RS implemented in COSMOtherm³⁵ to predict saturation pressures (P^{sat}) and observed that different stereoisomers of dihydroxy-dihydroperoxide yielded different SPs and, in turn, different predicted P^{sat} values, something that is physically inconsistent.²¹ They additionally noted that the SPs sampled with COSMOconf consistently lead to underestimated P^{sat} .

COSMOconf generates conformations by perturbing the dihedrals of a given starting geometry, generates a SP for each unique conformation, and produces a single molecular SP by weighting the conformer SPs by a Boltzmann factor of their energies in vacuum. This means that, COSMOconf gives a higher weight to conformations that are stable in vacuum. Since then, other works have confirmed the finding that using COSMOtherm with COSMOconf leads to underestimating P^{sat} and aqueous solubility,³⁰ attributing part of the underestimation to the Boltzmann weighting based on gas-phase energies. The problem stems from the fact that conformations with the maximum intramolecular/internal hydrogen bonding (IHB) have very low energies (large Boltzmann weights) in the gas phase while conformations that allow intermolecular/external hydrogen bonds (EHBs) form a more stable system in the condensed

phase. In fact, Hyttinen & Prisle³⁰ found that selecting conformers of citric acid with no IHBs led to better agreement with experimental aqueous solubilities compared to selecting conformers with any other number of IHBs or conformers selected by other Boltzmann-weighting algorithms.³⁰ Subsequently, Besel *et al.*²⁷ considered IHB capability when predicting P^{sat} by generating multiple conformers using COSMOconf and keeping many conformers (e.g. up to 40) with the fewest IHBs.²⁷ The authors concluded that 32 out of 40 conformers were needed to obtain a prediction of P^{sat} because that is when the conformer-averaged P^{sat} prediction converges. The converged value is not the experimental P^{sat} but is approximately five times larger than the prediction obtained from the lowest energy conformer. However, such a large number of conformations may have been necessary only because low energy conformers were selected first and Boltzmann weighting was used – *i.e.* it is likely that many of the 32 included conformations are only stable in the gas phase.

Other examples in literature point to the effect of using low vs. high energy conformations by comparing the use of IHB vs. non-IHB conformers. Cordova *et al.*²⁹ considered 2 conformers of vanillin and ethyl vanillin, each with and without IHBs. Using only one conformer in both systems (vanillin and ethyl vanillin) yielded poorer agreement with experimental solubility than weighting the conformer SPs by a Boltzmann factor (similar to the COSMOconf approach) or by an experimentally parametrized weighting scheme the authors developed, implying that relying on the most stable conformer in a particular solvent/phase is not helpful and that considering multiple conformers is necessary. Wang *et al.*²⁸ reached a similar conclusion when studying the difference in P^{sat} predictions when using conformers obtained from the liquid vs. gas phase of a Gibbs ensemble Monte Carlo simulation. All conformers in a given simulation snapshot were considered, and SPs were combined by taking an arithmetic average of the molecular conformations in the snapshot, assuming the conformations in the snapshot already obey a Boltzmann distribution. On the one hand, they showed that using conformers obtained from only one phase yielded large errors for a molecule capable of IHB (e.g. 2-methoxy-ethanol); relative errors were around –230% and 140% for the liquid-phase and gas-phase predictions, respectively. Meanwhile, taking the average $\log P^{\text{sat}}$ prediction obtained from each phase showed better agreement with experimental data (78% relative error). On the other hand, the conformer selection scheme had a much smaller effect on P^{sat} predictions in a molecule incapable of IHB (e.g. 1-hexanol). Similar trends were observed in the prediction of vapor–liquid equilibria (VLE) curves for mixtures where some of the components are not capable of IHB (e.g. 2-methoxy-ethanol + 2-butanone, and 1-hexanol + cyclohexane).²⁸

In addition to the energy of the conformation (or its IHB), solvent accessible surface area (SASA) appeared as an influencing factor in other studies. Silva *et al.*³⁶ found that using any single conformer obtained from a molecular

dynamics (MD) simulation of poly-ethylene glycol (PEG) in water yielded better agreement with experimental activity coefficients at infinite dilution compared to an all-trans, fully extended conformer optimized in vacuum (*i.e.* maximum SASA conformer). However, they observed that MD conformers with higher surface areas and radii of gyration yielded better predictions than other MD-sampled conformations, even though they occurred less frequently in the MD trajectory. Another work that used MD conformations is Buggert *et al.*,³¹ wherein COSMOtherm was used to predict micelle–water partition coefficients (K_i^{mw}) of different hydrophobic solutes in the presence of water and a surfactant (Triton X100). The solute/surfactant MD conformers considered were extracted from two MD simulations – one of a solute/surfactant molecule in water and one of a solute/surfactant molecule in octanol as a model apolar solvent. The SPs obtained from MD conformers were combined by calculating a Boltzmann weighted average SP. They reported that predictions of $\log K_i^{\text{mw}}$ obtained from MD-conformers were better than those obtained from a fully extended (linear) conformation by 2.4% on average. They also found that the best agreement to experimental $\log K_i^{\text{mw}}$ came from combining solute conformations from both water and octanol simulations, while the worst agreement depended on the system. They found that the poorest predictions for small (molecules with 5 carbons or less), flexible, apolar solutes (*e.g.* butylacetate, acetonitrile) and for inflexible molecules of any size (*e.g.* chlorobenzene, naphthalene) came from linear conformations of the solute optimized in vacuum. Meanwhile, the poorest predictions for large (more than 5 carbons), flexible, apolar solutes (*e.g.* *n*-alkanes, *n*-alkan-1-ols and *n*-ket-2-ones) came from MD conformations of the solute in water. The conformations in this latter case were likely collapsed conformations of minimal SASA, echoing the observations made by Silva *et al.*³⁶ for PEG systems.

More recently, Gomes *et al.*³⁷ examined the effect of conformer selection scheme on the solubility prediction of 942 drug-like molecules and found that the difference between schemes was small, *i.e.* prediction errors for each dataset/scheme were within the 95% confidence interval of each other. They also found that using complex conformer generation methods that implement biased energy sampling to overcome energy barriers like GOAT³⁸ (global optimization algorithm) were only favorable for highly flexible molecules (>13 rotatable bonds). However, based on the previously presented literature, these conclusions are likely affected by the choice to use only one conformation from each sampling method and only selecting the lowest energy conformer.

Though the summarized studies present some interesting observations, none of them are applied at a broad enough scale or in a systematic enough way to offer suggestions on conformer generation or selection methods for SP generation. This is particularly true in the case of SPs as ML inputs. In this work, we aim to develop an automated conformer generation workflow that ensures the best agreement with

experimental properties when SPs are used as the input to an ML model. We consider six target properties (molar mass, boiling point, density, refractive index, aqueous solubility and vapor pressure) and use GP models as a surrogate ML model. We evaluate three different approaches: 1) a publicly available conformer dataset (Mullins VT-2005)¹ in which the authors built conformers by hand without an explicitly mentioned conformer generation heuristic; 2) a heuristic that maximizes SASA and minimizes IHB (labeled “Stretched”); and 3) randomly generated conformers relaxed using a the classical forcefield MMFF94, as implemented in RDKit.^{39,40} The best heuristic/workflow is eventually integrated within the open-source SP generation package OpenSPGen.²⁵

2 Methods

This section describes the methods used in this work. We first describe the tool and settings used to generate SPs, then the conformer generation methods and heuristics evaluated. We then detail the ML study, including the molecule dataset, the target properties, and the ML model.

2.1 SP generation

SPs are most commonly obtained using the commercial software tools TURBOMOLE^{41,42} and DMol3.⁴³ Other commonly-used quantum chemistry packages like GAMESS,^{44,45} Gaussian,⁴⁶ and ORCA⁴⁷ have a COSMO implementation, and there have been efforts to convert their COSMO simulation outputs to SPs and/or parametrize them for conventional SP usage (*i.e.* to predict activity coefficients).^{48,49} However, the licenses for those packages range from “free site license” (GAMESS), to “free for academic use” (ORCA) to fully commercial (Gaussian). This means that none of them are truly open source. We recently released an end-to-end open-source tool for generating SPs called OpenSPGen,²⁵ which is a Python wrapper around NWChem⁵⁰ that allows SPs to be generated from only a molecular descriptor (*e.g.* CAS number, SMILES string, *etc.*) or an atomic coordinate file. The work also developed guidelines for generating SPs for ML applications (*e.g.* level of theory, averaging radius, *etc.*) given known starting molecular geometries. PySCF⁵¹ is another open-source quantum chemistry package with an implementation of COSMO solvation, but the COSMO settings for that tool are an unoptimized degree of freedom. Therefore, all SPs presented in this work were generated using OpenSPGen²⁵ according to the scheme shown in Fig. 1.

NWChem version 7.2.3 (ref. 52) was used with the settings recommended in the original OpenSPGen paper.²⁵ Specifically, the level of theory was HF/def2-SVP⁵³ with the York-Karplus⁵⁴ COSMO model. The starting shape for each atom was an octahedron, refined to a spherical cavity in three tessellation passes using a Delaunay-like meshing algorithm. The resulting SP had 501 bins resulting from the selected charge density range $[-0.25, 0.25] \text{ e } \text{\AA}^{-2}$. The selected charge range was wide enough to accommodate the use of

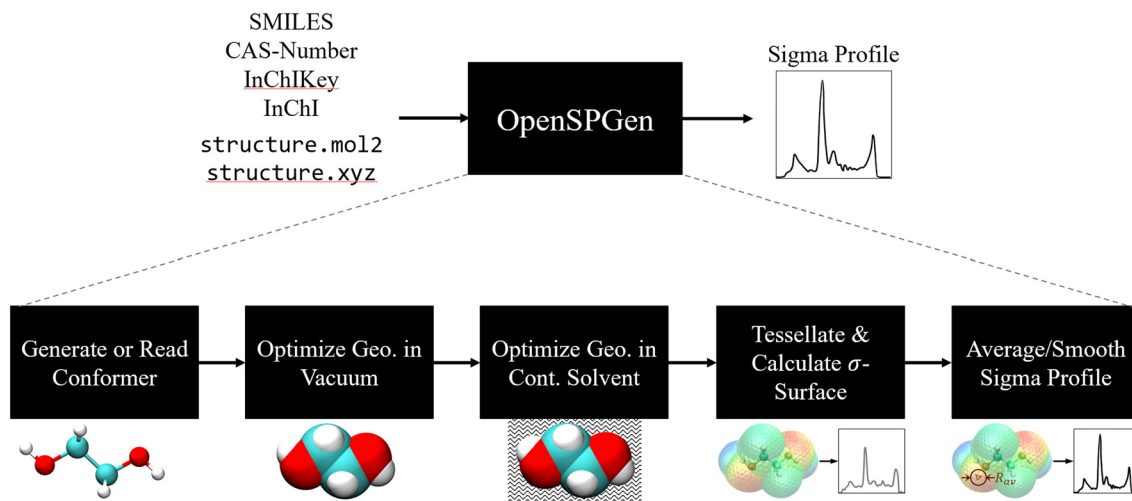


Fig. 1 Outline of SP generation process (top) showing the user-facing structure, where the user only needs to provide a molecular identifier. The hidden SP generation steps (bottom) which start with generating or reading a conformer from a provided file, then optimizing that geometry in NWChem – first in vacuum then in the continuum solvent at the desired level of theory. Continuum optimization requires constructing a solvent-accessible surface, tessellating it and calculating the screening charge or σ -surface. The σ -surface is then binned and averaged to obtain the SP.

unaveraged SPs. Any bins at the edges of the SP dataset that are always zero for a given dataset were removed before model fitting. The number of all-zero bins depends on the width of the SPs in the conformer dataset. The feature dimensions for each considered conformer dataset are summarized in Table S1 of the SI.

2.2 Conformer generation/selection methods

Three conformer selection schemes/heuristics were considered:

1. Mullins (literature): the conformers published in the VT-2005 SP dataset.⁵⁵ These were built by hand by the authors of the original work and no specific heuristic for conformer selection was mentioned.

2. Stretched: a heuristic meant to maximize the amount of detail in a SP, which is explained further below. It targets maximizing the solvent accessible surface area (SASA) and minimizing or breaking IHBs.

3. Random: random conformers were generated using the distance geometry algorithm implemented in RDKit and minimized using MMFF94 (ref. 56) as implemented in RDKit.⁴⁰ Five conformers were generated using a fixed set of random seeds and different methods for combining the random conformer SPs were considered.

RDKit version 2022.03.5 was used for all conformer generation in this work.

2.2.1 Mullins conformers. The VT-2005 SP Database was published by Mullins *et al.*⁵⁵ in 2006 and includes optimized geometries, sigma surfaces, complete DMol3 output files, as well as SPs for 1432 compounds including solvents and small solute molecules. Only one conformation is considered per molecule in the VT-2005 database. The structures obtained from VT-2005 were passed through OpenSPGen for this work, where the geometries were optimized in vacuum and then in

the COSMO solvation medium to generate SPs. These optimizations do not affect the conformation of the molecule (*i.e.* no changes in dihedrals were observed), but they are necessary because optimized atomic positions change slightly between quantum chemistry packages.

2.2.2 Stretched conformers. Stretched conformers maximize SASA and minimize IHB. The molecule was stretched using a biased MMFF94 (ref. 39 and 40) forcefield created by adding a repulsive term implemented as a distance constraint. More details on the development of this forcefield and its parametrization are available in the SI.

2.2.3 Random conformers. To study how much conformer selection heuristics (*e.g.* IHB, SASA, symmetry) matter, an additional dataset of random conformers was created. For each molecule, five random low energy conformers were generated using the distance geometry algorithm implemented in RDKit⁵⁶ and minimized using MMFF94. The random seeds for conformer generation were fixed to integers in the range 42–47 to ensure reproducibility and consistency of the resulting SPs. A SP calculation was carried out for each conformer, then the conformer SPs were combined into a single vector using two different averaging schemes. The first is an arithmetic averaging scheme (labeled AA), and the second is a Boltzmann-weighted averaging scheme (labeled BA), where every conformer SP is weighted according to a Boltzmann factor of the conformer's energy in the COSMO solvation medium at an arbitrarily selected temperature of 300 K. The inclusion of SP standard deviation was also considered in this dataset, which produced two variants of each random conformer SP dataset: one with the standard deviation included (labeled BA or AA depending on the averaging scheme), and one without standard deviation (labeled BAO or AAO for Boltzmann Average Only or Arithmetic Average Only). The standard deviation in SPs is included in the input by simply concatenating it with the

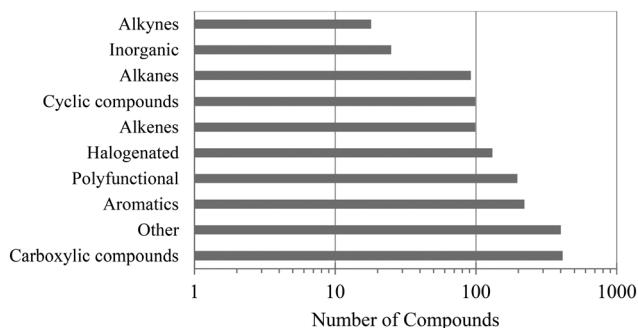


Fig. 2 Histogram of molecule types in the VT-2005 dataset ordered by prevalence. Modified from Table 4 in the original work by Mullins *et al.* (2006).⁵⁵

average SP producing a vector of dimensions (1002×1), since the SP itself is of dimensions (501×1).

2.3 ML study

2.3.1 Molecule dataset. The dataset of molecules used in this study came from the VT-2005 database⁵⁵ (described in more

detail in sections 2.2.1 and 3). The dataset consists of 1432 molecules spanning a wide variety of organic functional groups (hydrocarbons, alcohols, amines, aromatics, halogenated compounds, *etc.*), as well as some inorganic compounds (*e.g.* elemental gases, inorganic acids, *etc.*). Fig. 2 shows a breakdown of the molecules comprising the dataset by chemical family.

Additionally, select molecular features were calculated for the molecules in the set and their distributions are shown in Fig. 3. Panels (a)–(d) of Fig. 3 show that the average molecule in the set has a molecular weight (MW) of 133, with 21 atoms, 9 heavy atoms and 2 heteroatoms. Meaning, the average molecule in the set contains 7 carbons, 2 heteroatoms and 1–2 unsaturated bonds. Fig. 3(e) and (f) show that these heteroatoms are electronegative atoms that can serve as hydrogen bond donors or acceptors. Fig. 3(g) and (h) show that most molecules in the set have no rings, while Fig. 3(i) shows that the average number of rotatable bonds per molecule is approximately 4. Finally, Fig. 3(j) shows the distribution of Tanimoto similarity based on Morgan fingerprints with a radius of two. The average pair similarity is only 13%,

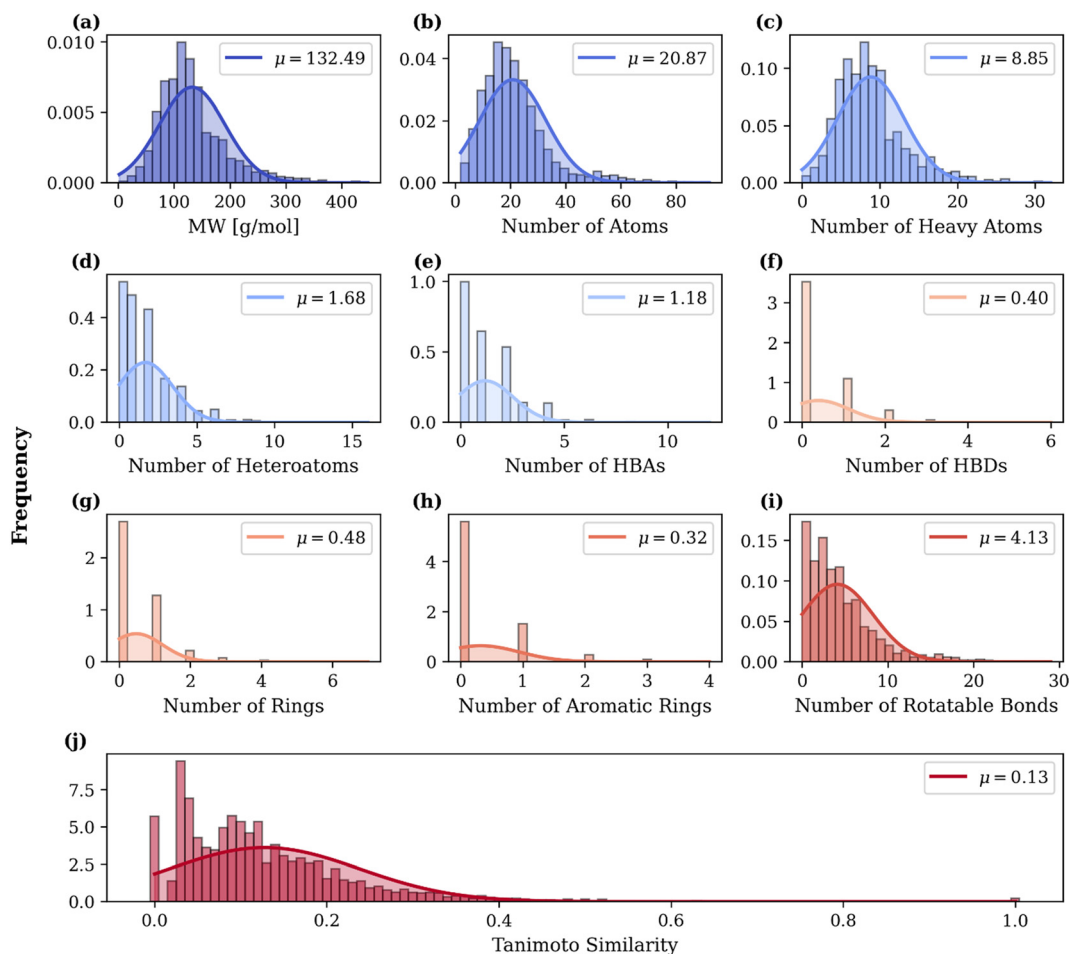


Fig. 3 Molecular feature distributions in the VT-2005 dataset. These include: size-related features in (a) molecular weight, (b) number of atoms, and (c) number of heavy atoms; and features indicating polarity like (d) number of heteroatoms, (e) number of hydrogen bond acceptors (HBAs), and (f) number of hydrogen bond donors (HBDs). The number of rings (g), number of aromatic rings (h), and the number of rotatable bonds (i), are indicators of molecular rigidity/flexibility. And Tanimoto similarity (j) shows the structural diversity of the molecules in the set.

which is very low compared to most commonly used similarity thresholds (0.85, 0.5 and 0.3).

2.3.2 Target properties dataset. The target data considered included molar mass and five thermophysical properties: boiling temperature, vapor pressure at 25 °C, density at 20 °C, refractive index (at wavelength 589 nm and 20 °C), and aqueous solubility at 25 °C. The datasets of target data for the molecules of the Mullins dataset were collected from the CRC Handbook of Chemistry and Physics.⁵⁷ Table 1 shows the size of the dataset for each property. Each property value was discretized into 15 bins, then split into 10 stratified folds. One k -fold was used for testing while the remaining nine were used for training. This provided 10 possible training-testing splits and allowed for cross validation and visualizing the effect of data splitting on ML performance for these small datasets.

2.3.3 Model selection and description. Since the objective of this work is to study the effect of conformer selection on the resulting SP and its performance as an input to ML models, the generated conformer datasets were used as inputs to GP models as an affordable surrogate model. GPs were selected since their training is relatively hands off once certain hyperparameters (*i.e.* the kernel and data normalization method) are selected. These hyperparameters were selected in previous work⁸ and depend on the target property being modeled. Specifically, P^{sat} and S_{aq} were log-scaled then standardized to a normal distribution, while all other properties considered were standardized to a normal distribution without transformation. Additionally, the GPs were all based on a radial basis function (RBF) kernel with trainable variance and length scale. A white noise kernel with trainable variance was added to mitigate over-fitting. This ML study follows the same methodology outlined in our previous paper.²⁵ All GP-related work was implemented using GPflow version 2.9.2,⁵⁸ and the exact environment file used can be found in the repository of the original OpenSPGen work.²⁵

Another conclusion made in our previous work⁸ on using SPs as GP model inputs, was that GP performance did not suffer with the sizes of the available target property datasets. In fact GP performance did not change beyond 250 samples for all target properties (refer to Fig. S5 in the SI of the previous work).⁸

2.3.4 ML model inputs. Here we describe all considered conformer datasets along with equations/descriptions for how SPs are prepared for the final ML model input vector x .

1. Mullins: literature conformers were taken as is. The only treatment applied to the SP dataset was dropping any all-zero bins.

2. Stretched: the stretched conformer obtained from the repulsive forcefield was used. The only treatment applied to the SP dataset was dropping any all-zero bins.

3. Random: a SP was calculated for each random conformer generated. The conformer SPs were then combined using different averaging methods and using a different number of random conformers (N_{conf}). The naming scheme of each method followed the format: “Rand-⟨averaging method⟩- N_{conf} ”. Detailed descriptions of each scheme are listed below:

i) Rand-AAO- N_{conf} : the SPs of N_{conf} random conformers were combined using an arithmetic averaging scheme, and the arithmetic average only (AAO) is used as model input as in eqn (1)

$$x = \overline{\text{SP}}_{\text{Rand-AA-}N_{\text{conf}}} = \frac{\sum_{\text{conf}=1}^{N_{\text{conf}}} \text{SP}_{\text{conf}}}{N_{\text{conf}}} \quad (1)$$

where SP_{conf} is the SP of conformer number ‘conf’ and N_{conf} is the total number of conformers included in the average where $N_{\text{conf}} \in \{1, 3, 5\}$. $\overline{\text{SP}}_{\text{Rand-AA-}N_{\text{conf}}}$ is the arithmetic average (AA) of N_{conf} random conformer SPs.

ii) Rand-BAO- N_{conf} : the SPs of N_{conf} random conformers were combined using a Boltzmann-weighted averaging scheme, and the Boltzmann average only was used as model input, calculated according to eqn (2)

$$x = \overline{\text{SP}}_{\text{Rand-BA-}N_{\text{conf}}} = \frac{\sum_{\text{conf}=1}^{N_{\text{conf}}} e^{-\beta \Delta E_{\text{conf}}} \text{SP}_{\text{conf}}}{\sum_{\text{conf}=1}^{N_{\text{conf}}} e^{-\beta \Delta E_{\text{conf}}}} \quad (2)$$

where ΔE_{conf} is the energy of conformer ‘conf’ relative to the lowest energy conformer out of the considered N_{conf} conformers to ensure numerical stability. β is the Boltzmann factor calculated at a temperature of 300 K and has units of Hartree to match the units of ΔE_{conf} returned by NWChem. $\overline{\text{SP}}_{\text{Rand-BA-}N_{\text{conf}}}$ is the Boltzmann weighted average of N_{conf} SPs. This scheme considered the energies of the conformer in the COSMO solvation medium.

iii) Rand-AA- N_{conf} : the SPs of N_{conf} random conformers were combined using an arithmetic averaging scheme, and both the average and standard deviation in SPs were used as model input. The average and standard deviation were concatenated to create the ML input vector as in eqn (3)

$$x = [\overline{\text{SP}}_{\text{Rand-AA-}N_{\text{conf}}} \sigma_{\text{Rand-AA-}N_{\text{conf}}}] \quad (3)$$

where $\overline{\text{SP}}_{\text{Rand-AA-}N_{\text{conf}}}$ is the arithmetic average SP of N_{conf} random conformers, and $\sigma_{\text{Rand-AA-}N_{\text{conf}}}$ is the standard deviation in those SPs.

iv) Rand-BA- N_{conf} : the SPs of N_{conf} random conformers were combined using a Boltzmann-weighted averaging scheme, and the average and standard deviation were concatenated to create the ML input vector as in eqn (4)

Table 1 List of target properties and the size of the data set for each

Target property	Code	$N_{\text{Property Set}}$	$\frac{N_{\text{Property Set}}}{N_{\text{Mullins SPs}}}$
Molar mass	MM	1432	100.0%
Boiling point	BP	1208	84.4%
Density at 20 °C	D ₂₀	711	49.7%
Refractive index at 20 °C	RI	834	58.3%
Aqueous solubility at 25 °C (g kg ⁻¹)	S_{aq}	327	22.9%
Vapor pressure	P^{sat}	594	41.5%

$$x = [\overline{\text{SP}}_{\text{Rand-BA-}N_{\text{conf}}} \sigma_{\text{Rand-BA-}N_{\text{conf}}}] \quad (4)$$

where $\overline{\text{SP}}_{\text{Rand-BA-}N_{\text{conf}}}$ is the Boltzmann-weighted average SP of N_{conf} random conformers, and $\sigma_{\text{Rand-BA-}N_{\text{conf}}}$ is the Boltzmann-weighted standard deviation in those SPs. Boltzmann factors were calculated as in eqn (2).

3 Conformers

This section explains the rationale of the conformer datasets/heuristics used in this work in more detail.

Mullins conformers were used because the VT-2005 dataset is commonly used for benchmarking and developing conventional COSMO-RS and COSMO-SAC models. The authors mention that conformers were validated by ensuring that the resulting SPs can reproduce experimental activity coefficients and normal boiling points from COSMO-based models, but no specifics were given on how to generate such conformers. The conformers of three molecules from the VT-2005 database were visualized to study the preference of factors that appeared to be important from literature review and the development of the stretched conformer dataset (mainly, SASA and IHB). Fig. 4 shows sample conformations from all considered conformer generation heuristics with Mullins conformers in the left column. After inspection, the heuristic for the Mullins dataset seems to be: 1) maximize SASA, 2) maximize symmetry, and 3) maximize IHB.

The rationale for the “stretched” heuristic came from the observations made in our previous work,²⁵ wherein, increasing the averaging radius had an adverse effect on ML performance. Fig. 5a shows the averaging radius (R_{av}) around a surface segment in the sigma surface of a sample molecule, while Fig. 5b shows how increasing R_{av} affects the SP. Fig. 5b shows that averaging with a larger radius was essentially removing details (or information) from the SP. One can imagine a limit where at a large enough averaging radius, the only information remaining in the SP is the total SASA and the net charge (*i.e.* the SP would be a single peak at the net

charge of the molecule and with a height equal to the surface area of the molecule). Accordingly, increasing R_{av} lead to diminished ML performance as shown in Fig. 5c. These observations lead to the interpretation that a SP with non-zero values distributed over a wider charge density range is more information rich. Hence the goal for this heuristic was to maximize detail in the SP, assuming that would be a surrogate for information.

To illustrate this idea of “maximizing detail”, two molecules are used for illustration. Fig. 6 shows two stable conformers of salicylic acid as an IHB-capable molecule along with their SPs. Because of the internal hydrogen bond in conformer (a), the SP peaks around $\sigma = 0.020 \text{ e } \text{\AA}^{-2}$ and $\sigma = -0.020 \text{ e } \text{\AA}^{-2}$ disappear and are replaced by increased area under the curve in the apolar region (typically defined as the range $\sigma \in [-0.007, 0.007] \text{ e } \text{\AA}^{-2}$).⁴⁸ This is due to the charge screening on the hydrogen-bonded atoms.

One can imagine that an ML model fed these two SPs might come up with different predictions for something like water solubility. With conformer (b), the model might be able to recognize that the peak at $\sigma = 0.020 \text{ e } \text{\AA}^{-2}$ indicates the presence of a HB donating group and the peak at $\sigma = -0.020 \text{ e } \text{\AA}^{-2}$ indicates the presence of a HB accepting group. The ML model would in turn be able to predict that salicylic acid is highly soluble in water and capable of IHB. With the SP of conformer (a), however, that information is not present due to IHB. This leads to the first rule of this heuristic: to maintain detail or information in the polar regions of the SP, internal hydrogen bonding should be minimized.

As a second example, Fig. 7 shows two stable conformers of a long *n*-alkane (triacontane, $n\text{-C}_{30}$). Conformer (a) is a randomly generated conformer of $n\text{-C}_{30}$ and exhibits a good amount of chain bending/folding. Conformer (b) on the other hand is a fully stretched conformer of $n\text{-C}_{30}$. Because the folded areas of conformer (a) are not accessible to a solvent molecule, the height of the apolar SP peak is noticeably lower than it is for conformer (b). This leads to the second rule of

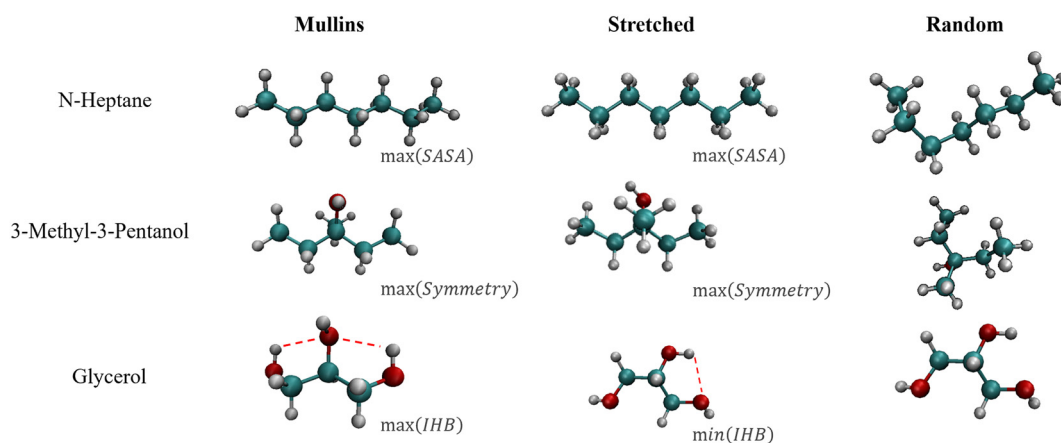


Fig. 4 Conformer samples from each of the considered conformer datasets for three different molecules, showing the preference of maximizing vs. minimizing IHB and SASA in each dataset. Red dotted lines are added to clarify the presence of an IHB.

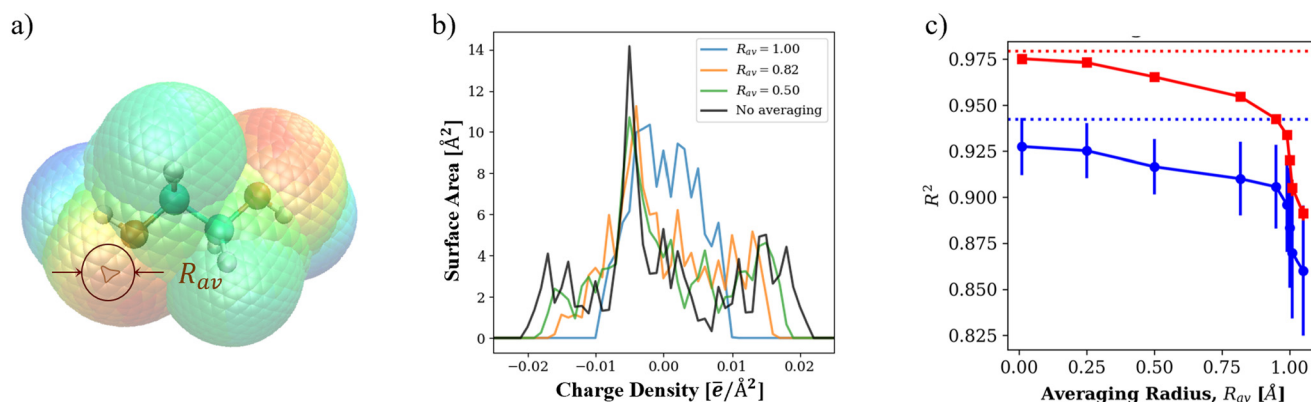


Fig. 5 Effect of averaging radius on the ML performance and SPs. a) shows a visual of averaging radius around a specific surface segment. b) shows the effect of averaging radius R_{av} on the SPs of glycerol, where larger R_{av} lead to SPs with a tighter charge density range and a higher apolar surface area. And c) shows the effect of averaging radius on ML performance for predictions of boiling point, where the red squares indicate training set performance, the blue circles indicate testing set performance, the dashed line indicates performance without any SP smoothing/averaging and error bars indicate the variance in performance with cross validation using 10 folds. Figures in panels (b) and (c) were modified from Salih (2025).²⁵

this heuristic: to ensure the SP contains the greatest amount of information in the apolar region, SASA should be maximized.

The center column of Fig. 4 shows sample conformers generated using the “stretched” method. The figure shows that the minimum IHB conformer was difficult to obtain for small symmetric molecules (e.g. glycerol) without manual manipulation. This issue of not obtaining the minimum number of IHBs was not observed in other small non-symmetric molecules with multiple IHB sites and such examples are shown in the SI (e.g. malic acid in Fig. S21(q)).

Finally, as a baseline heuristic-free method, random conformers were also considered. As can be seen in the right

column of Fig. 4, these conformers showed no preference in terms of maximizing or minimizing SASA and IHB.

4 Results & discussion

4.1 Effect of conformers on property prediction

Fig. 8 shows the effect of different conformer generation strategies on ML performance in terms of the mean absolute error (MAE) on the training and testing sets (performance results based on R^2 and mean absolute percent error (MAPE) are available in the SI Fig. S1 and S2, respectively). Three of the random conformer averaging approaches mentioned in section 2.3.4 (the Boltzmann averaging approaches, Rand-BA and Rand-BAO, and the arithmetic averaging with SP standard deviation Rand-AA) were omitted from Fig. 8 for the sake of clarity. All these omitted approaches display worse or

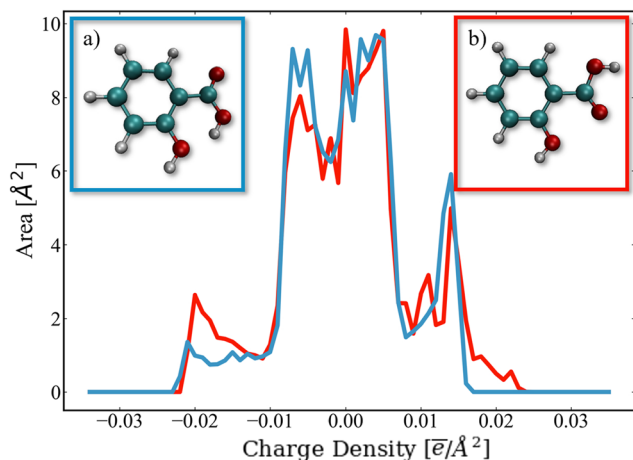


Fig. 6 Effect of conformers on the sigma profile of a molecule – sample shown for 2 stable conformers of salicylic acid. The blue SP belongs to conformer (a) in the blue box (exhibits IHB) and the red SP belongs to conformer (b) in the red box (no IHB). Dashed line in conformer (a) indicates a hydrogen bond and was only added for illustration purposes. Two SP peaks in conformer (a) are lost at $\sigma = -0.019, 0.020$ (polar regions of the SP) due to charge screening by the IHB.

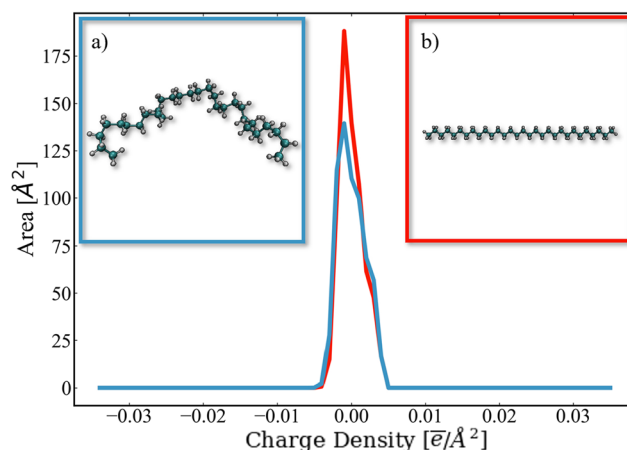


Fig. 7 Effect of conformers on the sigma profile of a molecule – sample shown for two stable conformers of $n\text{-C}_{30}$. The blue SP belongs to conformer (a) in the blue box (exhibits chain folding) and the red SP belongs to conformer (b) in the red box (no chain folding). Peak height in apolar regions of SP is reduced due to chain folding.

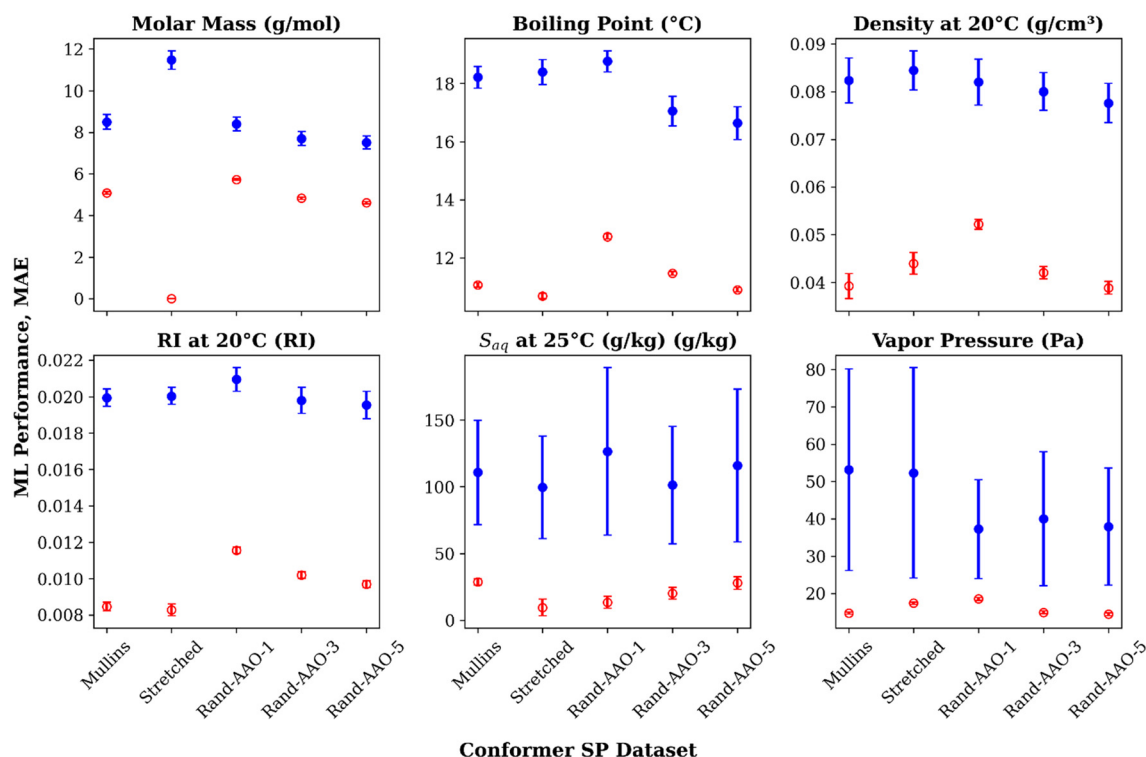


Fig. 8 Effect of conformers on GP performance for different target properties. Solid blue dots indicate performance on the testing set, while hollow red dots indicate performance on the training set. Error bars are the standard error in MAE values for 10 stratified data splits (cross-validation folds).

the same performance as the AAO approach. Including SP standard deviation always negatively affected performance, while the choice of SP averaging scheme had no effect on performance (see SI Fig. S3–S6).

The effect of conformers on ML performance is property dependent and small but statistically significant for three out of the six considered target properties. Specifically, the effect of conformers was not statistically significant (*i.e.* differences in MAE_{test} are within uncertainty limits) for density, aqueous solubility and vapor pressure, despite S_{aq} and P^{sat} being the most well-studied properties in the conformer literature for COSMO-RS. Meanwhile, differences between conformer dataset testing MAEs are small but greater than the standard error bars for boiling point, molar mass, and refractive index. Specifically, the maximum difference in MAE between conformer datasets is 2 °C for boiling point prediction, 4 g mol⁻¹ for molar mass, and 0.002 [unitless] for refractive index. For density, the difference between conformer datasets is smaller than the error bars, though we observe and discuss interesting trends later. An analysis of the statistical significance of the claim that one dataset outperforms another in terms of MAE_{test} is available in the SI (SI section 2.4 and Fig. S7). The *p*-values in Fig. S7 are calculated for each pair of reference and alternative conformer datasets and show that there are statistically significant pairwise improvements (or bolded *p*-value cells) in molar mass, boiling point, density and RI. However, for RI there was only statistically significant improvement in MAE when switching

from Rand-AAO-1 to Rand-AAO-3 and from Rand-AAO-1 to Rand-AAO-5. Meanwhile for density, the only statistically significant pair wherein MAE improvement was statistically significant was where Rand-AAO-3 was the reference and Rand-AAO-5 was the alternate conformer dataset. The switch from Rand-AAO-1 to Rand-AAO-5 for density was not statistically significant because of the larger variance in MAE_{test} in the Rand-AAO-1 conformer dataset.

The stretched conformer dataset (maximum SASA, minimum IHB) is the poorest performer (in terms of MAE_{test}) for molar mass predictions, while the Rand-AAO-1 (1 random conformer) dataset has the worst refractive index and boiling point predictions. These two datasets share the poorest performance for the predictions of density when considering the standard error bars. Interestingly, the Mullins dataset (maximum SASA and maximum IHB) has the same performance as the stretched dataset for molar mass, refractive index, boiling point, and density (all considered target properties except S_{aq} and P^{sat}).

One final observation is made on the random conformer datasets: increasing the number of conformers improved performance for all properties (except S_{aq} and P^{sat}). This is clear for molar mass, boiling point, and refractive index from both training and testing MAE. For density, however, the trend of improved performance with increased number of conformers was only statistically significant in the training MAE and appreciable – but not exceeding error bars – in testing MAE. This study only considers up to five random

conformers, but the performance does not plateau at five conformers for all target properties. It only plateaus at five conformers for molar mass and refractive index.

This trend of improved ML performance with the number of random conformers implies that the number of conformers used is more important than the selection heuristic. This agrees with observations from previous publications summarized in the introduction of this work. For instance, Wang²⁸ reported that using molecular simulation conformers obtained from both the gas and liquid phase gave better P^{sat} predictions than using conformations from only one phase. Similarly, Buggert *et al.*³¹ reported obtaining the best agreement with experimental micelle–water partition coefficients (K_i^{mw}) when using conformers obtained from both a polar solvent MD simulation and an apolar solvent MD simulation. It is likely that randomly sampling conformers serves as a surrogate to running molecular simulations in the phases relevant to the target property at hand (*e.g.* pure liquid and gas for BP and P^{sat} , solute in liquid water for S_{aq} , and pure liquid for RI). For example, randomly sampling conformers to predict glycerol boiling point leads to conformers with no IHBs (more likely in a pure dense phase) and conformers with one or more IHBs (more likely in the gas phase). This is demonstrated in Fig. 9 where the five conformers of glycerol obtained using the random heuristic are shown in panels a–e along with their energies and relative energies in panels f and g, respectively. Fig. 9 shows the heuristic produced three conformers with one IHB and two conformers with no IHBs. Meaning, the random heuristic can produce a sampling of glycerol conformers that would be likely in a dense polar phase (less IHB) and in a gas phase or dense apolar phase (more IHB). Fig. 9, panels f and g also show that the

difference in conformer energies is small ($<8 \text{ kJ mol}^{-1}$) even in the COSMO solvation medium. This explains why the Boltzmann averaging schemes had the same ML performance as arithmetic averaging schemes (see SI Fig. S5 and S6).

4.2 Correlating molecular properties with ML performance

Considering that the observed differences between conformer datasets are minimal through the metric of MAE, we analyze the per-molecule performance. The study aims to correlate per-molecule performance with per-molecule properties to see, for instance, if stretched conformations perform better for molecules that can form IHBs. This section considers how per-molecule performance varies with molecular properties, to attempt to understand if conformers have a more significant effect on larger molecules. Four molecular features that were hypothesized to be significant were extracted for all molecules in the set. These molecular features are summarized below, along with the rationale for including them:

1. Size: larger molecules have more possible conformers, and thus conformations of larger molecules are expected to have a greater impact on the SP and ML performance. The expectation is that per-sample MAE increases with molecular size.
2. Polarity: the properties of small polar molecules could be dominated by their intermolecular interactions, and conformations would have less impact on ML performance than for other molecules. The expectation is that MAE shows no correlation with polarity metrics, except the spread of errors could possibly increase with increased polarity.
3. Flexibility: a more flexible molecule has more available conformations and should be impacted by the conformer dataset.

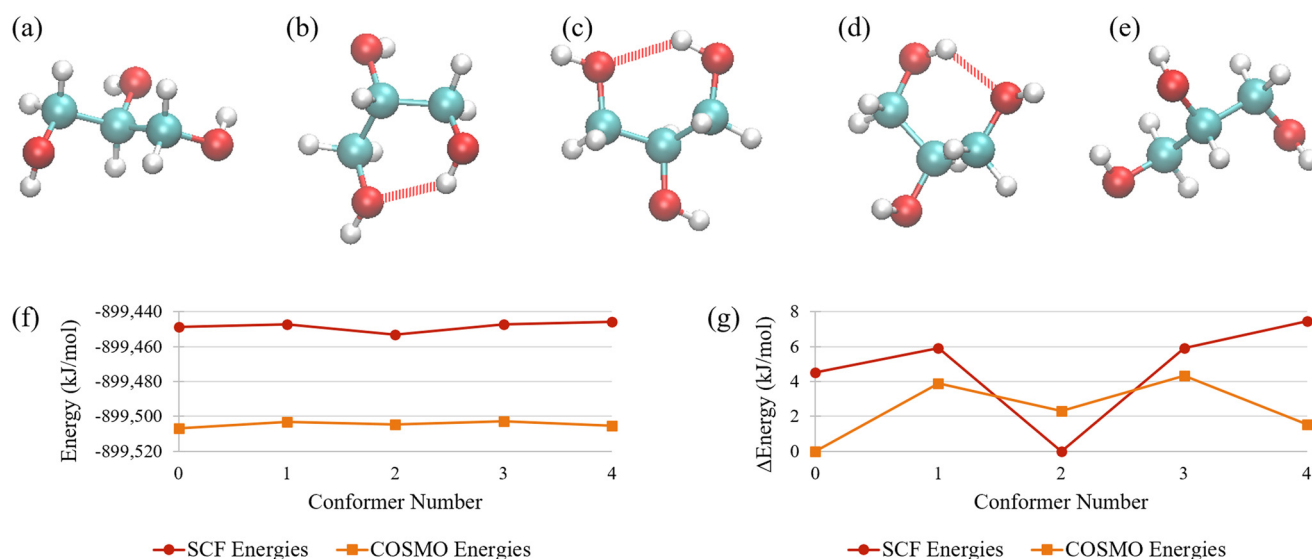


Fig. 9 Conformations of glycerol obtained by the random generation protocol along with their energies. Panels (a) through (e) show conformations 0 through 4, respectively, with hydrogen bonds indicated by a dashed red line. Plots (f) and (g) show the energies of these conformers in vacuum (SCF) and in the COSMO solvation medium, wherein (f) shows the raw energies and (g) shows energies relative to the lowest energy conformation in the ensemble.

Table 2 Molecular features used to study correlations between ML performance of different conformation datasets

General molecular feature	Quantifiable molecular feature ^a
Size	Molecular weight Total number of atoms Solvent accessible surface area
Polarity	Number of HB donors Number of HB acceptors Ratio of polar to apolar SASA
Flexibility	Number of rotatable bonds Number of rings (implies rigidity)
Internal HB capability	Capable or not Number of possible IHBs

^a All features are calculated on a per-molecule basis.

4. Internal HB capability: the property predictions for a molecule capable of IHB should be more sensitive to conformations. This hypothesis is based on prior literature findings highlighted.^{26,28,30}

Table 2 lists the different parameters used to quantify these molecular features.

All features are self-explanatory, except for two: the ratio of polar to apolar SASA and the number of possible IHBs. For the ratio of polar to apolar SASA, polar and apolar regions of the SP are identified by specifying an absolute value of the charge density of $|\sigma| = 0.007 \text{ e } \text{\AA}^{-2}$, beyond which the surface charges are classified as polar and below which the surface charges are taken as apolar. This cutoff was selected based on the HB regime limits chosen for COSMO-SAC implementations.⁴⁸ Polar and apolar SASAs are then calculated by summing up the SP bins in those regions. Eqn (5) and (6) show the formulas used for calculating polar and apolar SASA, respectively.

$$\text{SASA}_{\text{polar}} = \sum_{\sigma=-0.007}^{\sigma=-0.250} \text{SP} + \sum_{\sigma=0.007}^{\sigma=0.250} \text{SP} \quad (5)$$

$$\text{SASA}_{\text{apolar}} = \sum_{\sigma=-0.007}^{\sigma=0.007} \text{SP} \quad (6)$$

As for the number of possible IHBs, an IHB is deemed “possible” if the HBD hydrogen is at least three intramolecular bonds away from an HBA, since that is the shortest distance that would still allow the formation of a HB. Carboxylic acid groups are excluded to avoid counting the oxygen of the carbonyl group and the acidic hydrogen from the same carboxylic group as a possible IHB. However, no further consideration is made for molecule rigidity or steric effects. If the HBD–HBA pairs found fit these criteria, it is assumed that there is a conformation where the pair in question could form a HB. This is a reasonable assumption for the type of small molecules considered in this work.

Finally, since the performance of a particular GP model on a given molecule depends on whether that molecule is part of the training set, the per-molecule performance is evaluated as an average of the performances of all 10 models obtained from the different data splits.

Fig. 10 shows the correlations between molecular size (in terms of SASA) and per-molecule performance in terms of absolute relative error (ARE) for two selected target properties – boiling point and S_{aq} . The same figure including all target properties is available in the SI as Fig. S8, however, P^{sat} follows the same trend as S_{aq} and all other properties follow the same trend as BP. Unexpectedly, Fig. 10 shows that the model performs better for larger molecules for BP, but not for S_{aq} where there was no trend between ARE and SASA. This was consistent with other extracted molecular features related to size (see Fig. S8–S10 in the SI). Also, there is no clear difference in the distribution of errors between different conformer datasets. The only difference is in outlier molecules (see the points highlighted in red in Fig. 10). This is supported by previous observations made by Mullins¹ on

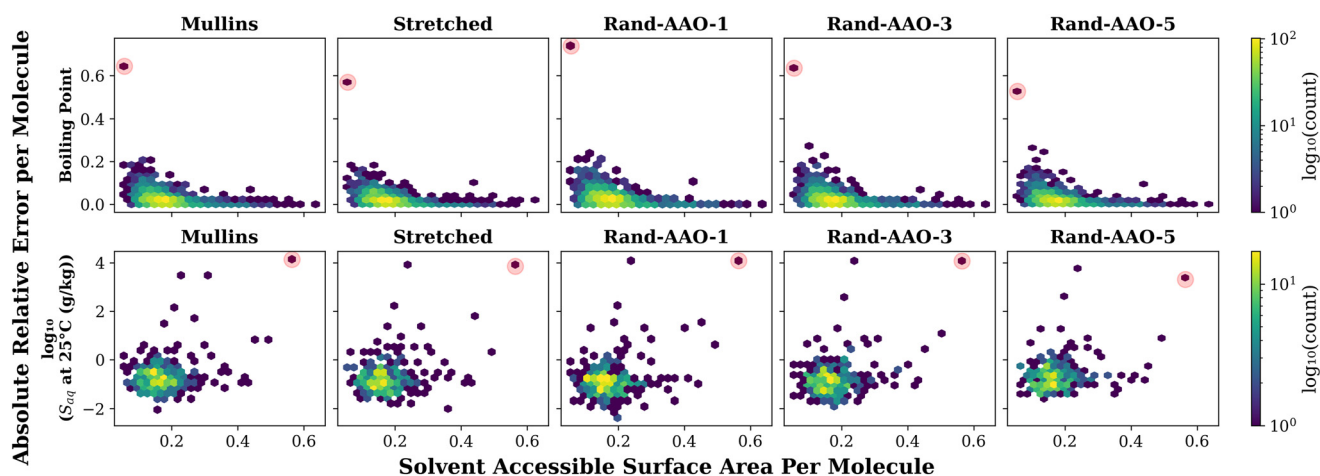


Fig. 10 Correlations/trends between SASA as a measure of molecular size and per-molecule ARE for select target properties. The diagrams are hexbins (2d histograms), where the color indicates the occurrence frequency of a certain ARE–SASA bin, the discretization is 25 bins in each direction, and the lowest visible occurrence frequency is one. Each row shows the AREs in the prediction of a certain material property, while the columns are different conformer datasets. Points highlighted in red are the outliers for selected properties.

the effect of molecular conformations on the solubility prediction of COSMO-SAC models. They observed that solubility predictions for a large molecule (cholesterol) were less sensitive to conformational changes than those of small molecules (aspirin and lidocaine). It was, however, unclear in that study whether cholesterol was a representative sample of a large molecule, as it contains a fused-ring cluster which makes it rigid. The results presented here show this conclusion is generalizable. It is possible that the effect of conformations does not increase with size because the SPs of larger molecules experience degeneracy, where changing the position of a few atoms in a large apolar molecule will still result in apolar surface segments of approximately the same size leading to little change to the final SP.

Quantities related to other molecular features (*e.g.* polarity, flexibility, *etc.*) show either the same trend as size, because features like the number of HB donors also indicates molecular size, or no correlation with per-sample error (see Fig. S11–S16 in the SI).

The last interesting trend to note concerns the ratio of polar:apolar SASA. Fig. 11 presents the correlations between per-molecule ARE and molecule polarity in terms of polar:apolar SASA. Fig. 11 only shows the correlations for BP and S_{aq} for clarity, but the same plot for all six target properties can be found in Fig. S17. Fig. 11 (and Fig. S17) shows the spread of ARE increases with increasing polarity, following the original expectation mentioned. Additionally, plotting molecules by their ratio of polar:apolar SASA reveals two clusters in the dataset of molecules considered separated at $\log_{10}(\text{SASA}_{\text{polar}}:\text{SASA}_{\text{apolar}}) \approx -2$ or $\text{SASA}_{\text{polar}}:\text{SASA}_{\text{apolar}} \approx 0.01$. Below this threshold molecules are large and apolar and above the threshold molecules are small and polar. The expected trend of the spread of errors increasing with polarity is observed above the threshold (for small polar molecules). However, below the threshold, *i.e.* for large apolar molecules, there is no

correlation between per-molecule performance and polarity. The difference between conformer datasets is still only visible in the outliers. However, Fig. S17 shows that depending on the target property, the effect of adding more random conformers on outlier AREs changes. For instance, the outlier ARE values for molar mass, boiling point, and refractive index decrease as the number of random conformers increases. However, the outlier ARE values for density increase as the number of random conformers increases in the polar region ($\text{SASA}_{\text{polar}}:\text{SASA}_{\text{apolar}} > 0.01$) and decrease with increased random conformers in the apolar region. This explains why the test MAE for density in Fig. 8 decreased with increased random conformers, but not by an amount greater than the standard error bars.

Finally, since the highlighted points in Fig. 10 showing the difference between how different conformer datasets perform are only visible for outliers, section 4.3 examines the outlier molecule highlighted in Fig. 10 for boiling point along with its SP in different conformer datasets.

4.3 Studying outliers of ML performance

The selected outlier highlighted in Fig. 10 is discussed in this section. The outlier highlighted for boiling point is carbon monoxide (index 387), which has only one unique conformation. Looking at Fig. 12(top) the SPs of $\text{C}\equiv\text{O}$ obtained from each conformer dataset are extremely similar. Fig. 12(bottom), which shows the differences in the SPs obtained from each conformer dataset with the Mullins dataset SP as the reference, shows that the SPs of $\text{C}\equiv\text{O}$ do not vary within a considerable range. SP variations at this scale are essentially noise and do not indicate different conformations. They might emerge from different initializations of the surface mesh, slight differences in nuclei positions or floating-point error propagations. Even in the case of small differences in nuclei positions, the C–O

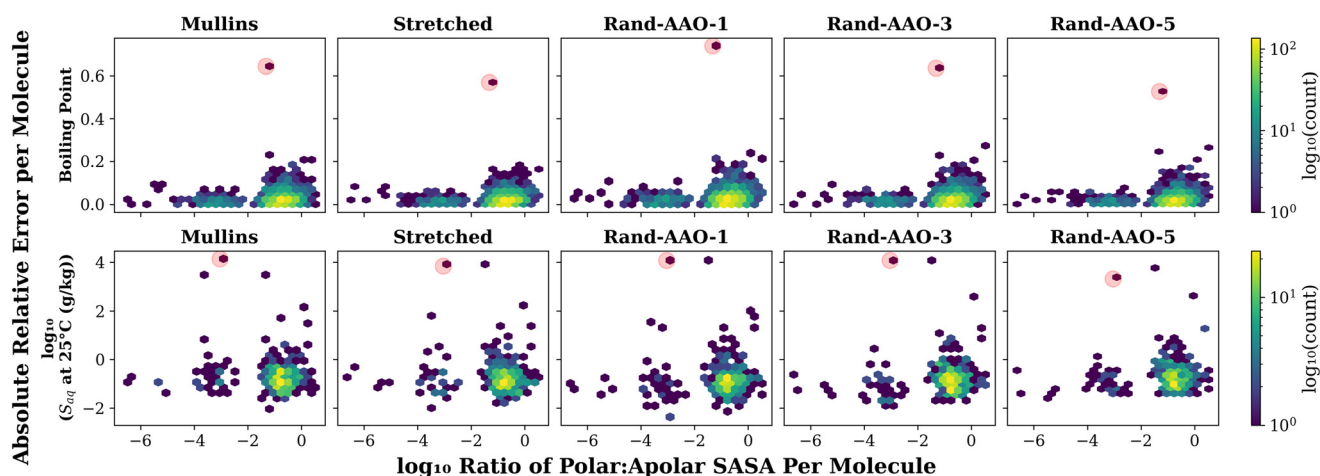


Fig. 11 Correlations/trends between ratio of polar:apolar SASA as a measure of polarity and per-molecule ARE. The diagrams are hexbins (2d histograms), where the color indicates the occurrence frequency of a certain ARE-molecular feature bin, the discretization is 25 bins in each direction, and the lowest visible occurrence frequency is one. Each row shows the AREs in the prediction of a particular material property, while the columns are different conformer datasets. Points highlighted in red are the outliers for selected properties.

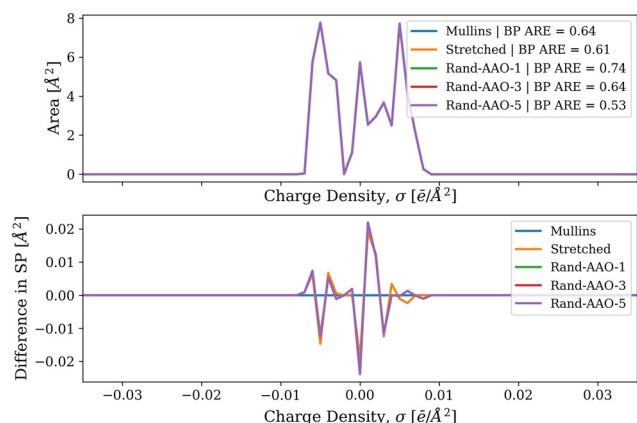


Fig. 12 SPs of a select outlier molecule (carbon monoxide – index 387) from boiling point predictions. Top figure shows the SPs themselves, while the bottom shows the difference between a given SP and the Mullins SP as reference. Legend on top figure reports the absolute relative error (ARE) in boiling point predictions corresponding to each SP.

bond length in carbon monoxide is $1.128 \text{ Å} \pm 0.004 \text{ Å}$ according to NIST,⁵⁹ and, normally, if the weighted local similarity between atom positions in two mapped conformers is less than 0.5 Å the conformers are not considered unique.²⁷ Additionally, our previous studies on reproducibility have shown that per-bin perturbations in the SP on the scale of 10^{-5} ē Å^{-2} are not “visible” to a pretrained GP and the same material property value should be predicted by each model.²⁵ Thus, these differences in prediction are likely coming from the fact that $\text{C}\equiv\text{O}$ falls at the edge of the distribution for boiling point. Therefore, predicting the properties of $\text{C}\equiv\text{O}$ requires the GP to extrapolate beyond the range of target values it was trained on, leading to large relative errors. This conclusion is supported by plotting the distribution of ARE against the target properties themselves (Fig. S18). We observe that all the outliers appear near the edges of the target property distribution or in data-sparse regions. However, the predicted properties, even for outliers like $\text{C}\equiv\text{O}$, do not appear to be random.

The accuracy of a particular model's prediction deployed on $\text{C}\equiv\text{O}$ could be related to the underlying trained model. For instance, the ARE in boiling point for $\text{C}\equiv\text{O}$ drops as the number of random conformers is

increased from Rand-AAO-1 to Rand-AAO-5. It is possible that the error for $\text{C}\equiv\text{O}$ drops because the model trained on five random SPs of molecules with multiple conformers contains more information relevant to boiling point prediction even for a single-conformer molecule. This indicates that the outliers are not driving the average model performance. Instead, the SP dataset the model was trained on determined the prediction accuracy for outliers. Perhaps a more meaningful way of describing per-molecule error distribution would be to look at the parameters of the 2D (or joint) normal distributions shown in Fig. S18 and analyzed in Table 3.

Table 3 shows the size of the confidence ellipse for the joint distribution of per-molecule ARE and target property values (*i.e.* the hexbin plots in Fig. S18). This property is calculated as the determinant of the covariance matrix of the joint Gaussian distribution and is equivalent to the standard deviation in a 1D Gaussian distribution. The values shown in Table 3 support our observations; the size of the confidence ellipse decreases (*i.e.* the quality of per-sample predictions improves) with the addition of more conformers. The extent of improvement is highly property dependent, as observed with average ML performance in section 4.1 and Fig. 8. Contrary to the metrics shown in Fig. 8, however, Table 3 shows that there is an improvement in ML performance for all properties except vapor pressure when more random conformers are added.

5 Conclusions

The effect of conformations on SPs and their performance as ML descriptors was systematically studied by generating multiple conformer datasets using three conformer generation heuristics. The first, “Mullins”, is a literature dataset of manually created conformers where the only heuristic mentioned was to minimize energy, but a look at the selected conformers revealed a preference to maximize SASA and IHB. The second, “Stretched”, was developed to maximize SASA and minimize IHB and is based on the hypothesis that this would increase the amount of detail carried by the SP (*i.e.* expose surfaces hidden by chain folding and expose the maximum number of available HB sites). The final was “Random”, where up to five random

Table 3 Size of confidence ellipse for the correlation between for the ground-truth value of molecular target properties and per-molecule ARE for each combination of target property and conformer dataset

Target property	Conformer datasets				
	Mullins	Stretched	Rand-AAO-1	Rand-AAO-3	Rand-AAO-5
Molar mass	13.94	13.27	14.97	15.24	14.54
Boiling point	10.1	9.22	12.69	10.92	9.45
Density at 20 °C	1.50×10^{-4}	1.77×10^{-4}	2.19×10^{-4}	1.69×10^{-4}	1.63×10^{-4}
RI at 20 °C	1.69×10^{-7}	1.62×10^{-7}	2.95×10^{-7}	2.25×10^{-7}	2.08×10^{-7}
S_{aq} at 25 °C (g/kg) ^a	1.39	1.37	1.62	1.48	1.25
Vapor pressure ^a	0.42	0.41	0.45	0.44	0.45

^a Properties are transformed to log-scale before characterizing the joint gaussian distribution.

conformers were generated and an average of the resulting SPs was used as the ML model input. Both an arithmetic average of the random conformer SPs and a Boltzmann-weighted average were considered. These three conformer generation approaches were then used to generate SPs as inputs to GP models to predict six different target properties – molar mass and five thermophysical properties.

The effect of conformations on GP performance was observed to be small and property dependent. We observed a small effect for the prediction of boiling point, molar mass, and refractive index, but there was no statistically significant effect (greater than standard error bars) on density, vapor pressure, and aqueous solubility predictions. For the properties where conformations had an effect, “Rand-AAO-5” (five random conformers averaged arithmetically) performed the best. We note that the performance of random conformers improved with the addition of more conformers. The dataset “Rand-AAO-5” performed the same or better than Mullins for most properties. This implies that the number of conformers is more important than the conformer generation method and that seeking the lowest energy conformation is not necessary or preferred in the ML case. This conclusion can be extended to the class of methods followed by many works in the literature wherein a large number of conformers is generated only to find and use the lowest energy conformation – as those are the closest reproducible and scalable proxy to the non-scalable hand-tuning approach used to assemble the Mullins dataset.

Finally, a per-sample analysis of performance showed that the conformer dataset had the greatest effect on small molecules. This supports previous literature findings but is likely impacted by the fact that small molecules have property values at the extremes of the target property distribution and require the GP to extrapolate. This was confirmed by an analysis of the confidence ellipses of the 2D distributions of prediction error *vs.* target property. We observe the size of the confidence ellipse (and hence the ML prediction errors) decreased with increasing number of conformers.

The final recommendation of this work, based on the findings summarized above, is to use at least five random conformers when generating SPs for ML applications. Using fixed random seeds is also recommended to ensure consistency. However, we acknowledge that the use of five conformers does not guarantee that ML performance converges. Given this and the fact that conformer effects are dependent on the target property, users are encouraged to examine how ML performance for predicting their target property of interest trends with the number of random conformers up to five. If a plot of the ML performance on the desired target property with one, three and five random conformers plateaus at five random conformers, the user can stop there. However, if the plot shows a steep increase in performance, the user should consider adding five more conformers and reexamining. The cost of additional conformer generation can be estimated using the speed tests

discussed in the SI. Equation S6 shows that wall-clock time in seconds scales exponentially with number of atoms as $\sim 10^{0.06 \times N_{\text{atoms}} + 0.65}$. Meaning, the added cost of using five conformations instead of one is not significant (<10 minutes) for molecules with 25 atoms or less and becomes more significant (>1 hour) for molecules with 40 atoms or more. The option to generate five random conformers has been set as the default in the latest version of the open-source SP generator OpenSPGen (commit hash: 5cc98c2). The repository can be accessed at: <https://github.com/FaSalih/OpenSPGen>.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article, mainly, compiled SP and target property datasets are available in the Github repository: <https://github.com/FaSalih/OpenSPGen> (commit hash: 5cc98c2).

Additionally, OpenSPGen outputs like geometry files, COSMO files, sigma surface files and complete NWChem output files are available on Zenodo through the DOI: <https://doi.org/10.5281/zenodo.20618532>. The Zenodo database, also, includes the scripts used to generate conformers, compile SPs, train GP models, and extract the molecular features described in section 4.2.

Supplementary information (SI) is available. See DOI: <https://doi.org/10.1039/d6me00094k>. The SI includes additional results, details on the developed forcefield for the stretched conformer dataset, and information on the conducted speed tests.

Acknowledgements

This work was supported by the Breakthrough Electrolytes for Energy Storage (BEES) Energy Frontier Research Center (award #DE-SC0019409), funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences program, and by the U.S. Department of Energy through subcontract 630340 from Los Alamos National Laboratory, Materials and Chemical Sciences Division. This research is also based upon work partially supported by the National Science Foundation under award number EEC-2330175 for the Engineering Research Center EARTH. Additionally, partial funding was provided by the project CICECO-Aveiro Institute of Materials, UID/50011/2025 (DOI <https://doi.org/10.54499/UID/50011/2025>) & LA/P/0006/2020 (DOI <https://doi.org/10.54499/LA/P/0006/2020>) & UID/PRR/50011/2025 (DOI <https://doi.org/10.54499/UID/PRR/50011/2025>), financed by Portuguese funds through the FCT/MCTES (PIDDAC).

References

- 1 E. Mullins, Y. A. Liu, A. Ghaderi and S. D. Fast, Sigma profile database for predicting solid solubility in pure and mixed solvent mixtures for organic pharmacological compounds with COSMO-based thermodynamic methods, *Ind. Eng. Chem. Res.*, 2008, **47**, 1707–1725.

- 2 A. Klamt and G. Schuurmann, *COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient*, 1993.
- 3 A. Klamt, V. Jonas, T. Bu and J. C. W. Lohrenz, *Refinement and Parametrization of COSMO-RS*, 1998.
- 4 A. Klamt, F. Eckert and W. Arlt, COSMO-RS: An Alternative to Simulation for Calculating Thermodynamic Properties of Liquid Mixtures, *Annu. Rev. Chem. Biomol. Eng.*, 2010, **1**, 101–122.
- 5 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules, *ACS Cent. Sci.*, 2018, **4**, 268–276.
- 6 E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira and K. Schmidt, An open-source family of large encoder-decoder foundation models for chemistry, *Commun. Chem.*, 2025, **8**, 193.
- 7 D. O. Abranches, Y. Zhang, E. J. Maginn and Y. J. Colón, Sigma profiles in deep learning: towards a universal molecular descriptor, *Chem. Commun.*, 2022, **58**, 5630–5633.
- 8 D. O. Abranches, E. J. Maginn and Y. J. Colón, Stochastic machine learning via sigma profiles to build a digital chemical space, *Proc. Natl. Acad. Sci. U. S. A.*, 2024, **121**, e2404676121.
- 9 T. Lemaoui, N. E. H. Hammoudi, I. M. Alnashef, M. Balsamo, A. Erto, B. Ernst and Y. Benguerba, Quantitative structure properties relationship for deep eutectic solvents using Σ -profile as molecular descriptors, *J. Mol. Liq.*, 2020, **309**, 113165.
- 10 O. Nordness, P. Kelkar, Y. Lyu, M. Baldea, M. A. Stadtherr and J. F. Brennecke, Predicting thermophysical properties of dialkylimidazolium ionic liquids from sigma profiles, *J. Mol. Liq.*, 2021, **334**, 116019.
- 11 G. Járvas, C. Quellet and A. Dallos, COSMO-RS based CFD model for flat surface evaporation of non-ideal liquid mixtures, *Int. J. Heat Mass Transfer*, 2011, **54**, 4630–4635.
- 12 J. Wang, Z. Song, L. Chen, T. Xu, L. Deng and Z. Qi, Prediction of CO₂ solubility in deep eutectic solvents using random forest model based on COSMO-RS-derived descriptors, *Green Chem. Eng.*, 2021, **2**, 431–440.
- 13 A. Niederquell, N. Wyttenbach and M. Kuentz, New prediction methods for solubility parameters based on molecular sigma profiles using pharmaceutical materials, *Int. J. Pharm.*, 2018, **546**, 137–144.
- 14 J.-E. Li, S.-C. Chien and C.-M. Hsieh, Modeling solid solute solubility in supercritical carbon dioxide by machine learning algorithms using molecular sigma profiles, *J. Mol. Liq.*, 2024, **395**, 123884.
- 15 G. Hernández Morales, E. I. Sanchez Medina, A. Jiménez-Gutiérrez and V. M. Zavala, Graph neural networks for CO₂ solubility predictions in Deep Eutectic Solvents, *Comput. Chem. Eng.*, 2024, **187**, 108750.
- 16 P. Zhou, J. Yu, K. L. Sánchez-Rivera, G. W. Huber and R. C. Van Lehn, Large-scale computational polymer solubility predictions and applications to dissolution-based plastic recycling, *Green Chem.*, 2023, **25**, 4402–4414.
- 17 A. Kondor, G. Járvas, J. Kontos and A. Dallos, Temperature dependent surface tension estimation using COSMO-RS sigma moments, *Chem. Eng. Res. Des.*, 2014, **92**, 2867–2872.
- 18 W. Benmouloud, C. Si-Moussa and O. Benkortbi, Machine learning approach for the prediction of surface tension of binary mixtures containing ionic liquids using σ -profile descriptors, *Int. J. Quantum Chem.*, 2023, **123**, e27026.
- 19 C. Mehler, A. Klamt and W. Peukert, Use of COSMO-RS for the prediction of adsorption equilibria, *AIChE J.*, 2002, **48**, 1093–1099.
- 20 Y.-H. Cheng, I.-T. Sung, C.-M. Hsieh and L.-C. Lin, Module-based machine learning models using sigma profiles of organic linkers to predict gaseous adsorption in metal-organic frameworks, *J. Taiwan Inst. Chem. Eng.*, 2024, **165**, 105728.
- 21 Y. Zhao, Y. Huang, X. Zhang and S. Zhang, A quantitative prediction of the viscosity of ionic liquids using Σ -profile molecular descriptors, *Phys. Chem. Chem. Phys.*, 2015, **17**, 3761–3767.
- 22 D. Fan, K. Xue, R. Zhang, W. Zhu, H. Zhang, J. Qi, Z. Zhu, Y. Wang and P. Cui, Application of interpretable machine learning models to improve the prediction performance of ionic liquids toxicity, *Sci. Total Environ.*, 2024, **908**, 168168.
- 23 D. Fan, K. Xue, Y. Liu, W. Zhu, Y. Chen, P. Cui, S. Sun, J. Qi, Z. Zhu and Y. Wang, Modeling the toxicity of ionic liquids based on deep learning method, *Comput. Chem. Eng.*, 2023, **176**, 108293.
- 24 L. Zhang, H. Mao, Y. Zhuang, L. Wang, L. Liu, Y. Dong, J. Du, W. Xie and Z. Yuan, Odor prediction and aroma mixture design using machine learning model and molecular surface charge density profiles, *Chem. Eng. Sci.*, 2021, **245**, 116947.
- 25 F. Y. M. Salih, D. O. Abranches, E. J. Maginn and Y. J. Colón, Open-source generation of sigma profiles: impact of quantum chemistry and solvation treatment on machine learning performance, *Digital Discovery*, 2025, **4**(10), 2711–2723.
- 26 T. Kurtén, N. Hyttinen, E. L. D'Ambro, J. Thornton and N. L. Prisle, Estimating the saturation vapor pressures of isoprene oxidation products C₅H₁₂O₆ and C₅H₁₀O₆ using COSMO-RS, *Atmos. Chem. Phys.*, 2018, **18**, 17589–17600.
- 27 V. Besel, M. Todorović, T. Kurtén, P. Rinke and H. Vehkamäki, Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules, *Sci. Data*, 2023, **10**, 450.
- 28 S. Wang, J. M. Stubbs, J. I. Siepmann and S. I. Sandler, Effects of Conformational Distributions on Sigma Profiles in COSMO Theories, *J. Phys. Chem. A*, 2005, **109**, 11285–11294.
- 29 I. W. Cordova, G. Teixeira, P. J. A. Ribeiro-Claro, D. O. Abranches, S. P. Pinho, O. Ferreira and J. A. P. Coutinho, Using Molecular Conformers in COSMO-RS to Predict Drug Solubility in Mixed Solvents, *Ind. Eng. Chem. Res.*, 2024, **63**, 9565–9575.
- 30 N. Hyttinen and N. L. Prisle, Improving Solubility and Activity Estimates of Multifunctional Atmospheric Organics by Selecting Conformers in COSMOtherm, *J. Phys. Chem. A*, 2020, **124**, 4801–4812.

- 31 M. Buggert, C. Cadena, L. Mokrushina, I. Smirnova, E. J. Maginn and W. Arlt, COSMO-RS Calculations of Partition Coefficients: Different Tools for Conformation Search, *Chem. Eng. Technol.*, 2009, **32**, 977–986.
- 32 M. Buggert, L. Mokrushina, I. Smirnova, R. Schomäcker and W. Arlt, Prediction of Equilibrium Partitioning of Nonpolar Organic Solutes in Water- Surfactant Systems by UNIFAC and COSMO-RS Models, *Chem. Eng. Technol.*, 2006, **29**, 567–573.
- 33 L. Mokrushina, M. Buggert, I. Smirnova, W. Arlt and R. Schomäcker, COSMO-RS and UNIFAC in Prediction of Micelle/Water Partition Coefficients, *Ind. Eng. Chem. Res.*, 2007, **46**, 6501–6509.
- 34 BIOVIA COSMOconf, Dassault Systèmes, 2024, preprint, Dassault Systèmes:2025.
- 35 F. Eckert and A. Klamt, COSMOtherm, COSMOlogic GmbH & Co. KG, preprint, COSMOlogic GmbH & Co. KG:19.
- 36 G. M. C. Silva, D. A. Pantano, S. Loehlé and J. A. P. Coutinho, The Challenges of Using COSMO-RS To Describe Polymer Solution Behavior, *Ind. Eng. Chem. Res.*, 2023, **62**, 20936–20944.
- 37 A. M. M. Gomes, J. F. O. Granjo and P. J. Costa, Conformer Generation Workflows for COSMO-RS Calculations: Are They All the Same?, *J. Chem. Inf. Model.*, 2026, **66**(16), 10004–10018.
- 38 B. de Souza, GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters, *Angew. Chem., Int. Ed.*, 2025, **64**, e202500393.
- 39 T. A. Halgren and Merck molecular force field. V., Extension of MMFF94 using experimental data, additional computational data, and empirical rules, *J. Comput. Chem.*, 1996, **17**, 616–641.
- 40 P. Tosco, N. Stiefl and G. Landrum, Bringing the MMFF force field to the RDKit: implementation and validation, *J. Cheminform.*, 2014, **6**, 37.
- 41 Y. J. Franzke, C. Holzer, J. H. Andersen, T. Begušić, F. Bruder, S. Coriani, F. Della Sala, E. Fabiano, D. A. Fedotov, S. Fürst, S. Gillhuber, R. Grotjahn, M. Kaupp, M. Kehry, M. Krstić, F. Mack, S. Majumdar, B. D. Nguyen, S. M. Parker, F. Pauly, A. Pausch, E. Perlt, G. S. Phun, A. Rajabi, D. Rappoport, B. Samal, T. Schrader, M. Sharma, E. Tapavicza, R. S. Treß, V. Voora, A. Wodyński, J. M. Yu, B. Zerulla, F. Furche, C. Hättig, M. Sierka, D. P. Tew and F. Weigend, TURBOMOLE: Today and Tomorrow, *J. Chem. Theory Comput.*, 2023, **19**, 6859–6890.
- 42 S. G. Balasubramani, G. P. Chen, S. Coriani, M. Diedenhofen, M. S. Frank, Y. J. Franzke, F. Furche, R. Grotjahn, M. E. Harding, C. Hättig, A. Hellweg, B. Helmich-Paris, C. Holzer, U. Huniar, M. Kaupp, A. Marefat Khah, S. Karbalaee Khani, T. Müller, F. Mack, B. D. Nguyen, S. M. Parker, E. Perlt, D. Rappoport, K. Reiter, S. Roy, M. Rückert, G. Schmitz, M. Sierka, E. Tapavicza, D. P. Tew, C. van Wüllen, V. K. Voora, F. Weigend, A. Wodyński and J. M. Yu, TURBOMOLE: Modular program suite for ab initio quantum-chemical and condensed-matter simulations, *J. Chem. Phys.*, 2020, **152**, 184107.
- 43 B. Delley, From molecules to solids with the DMol3 approach, *J. Chem. Phys.*, 2000, **113**, 7756–7764.
- 44 F. Zahariev, P. Xu, B. M. Westheimer, S. Webb, J. Galvez Vallejo, A. Tiwari, V. Sundriyal, M. Sosonkina, J. Shen, G. Schoendorff, M. Schlinsog, T. Sattasathuchana, K. Ruedenberg, L. B. Roskop, A. P. Rendell, D. Poole, P. Piecuch, B. Q. Pham, V. Mironov, J. Mato, S. Leonard, S. S. Leang, J. Ivanic, J. Hayes, T. Harville, K. Gururangan, E. Guidez, I. S. Gerasimov, C. Friedl, K. N. Ferreras, G. Elliott, D. Datta, D. D. A. Cruz, L. Carrington, C. Bertoni, G. M. J. Barca, M. Alkan and M. S. Gordon, The General Atomic and Molecular Electronic Structure System (GAMESS): Novel Methods on Novel Architectures, *J. Chem. Theory Comput.*, 2023, **19**, 7031–7055.
- 45 G. M. J. Barca, C. Bertoni, L. Carrington, D. Datta, N. De Silva, J. E. Deustua, D. G. Fedorov, J. R. Gour, A. O. Gunina, E. Guidez, T. Harville, S. Irle, J. Ivanic, K. Kowalski, S. S. Leang, H. Li, W. Li, J. J. Lutz, I. Magoulas, J. Mato, V. Mironov, H. Nakata, B. Q. Pham, P. Piecuch, D. Poole, S. R. Pruitt, A. P. Rendell, L. B. Roskop, K. Ruedenberg, T. Sattasathuchana, M. W. Schmidt, J. Shen, L. Slipchenko, M. Sosonkina, V. Sundriyal, A. Tiwari, J. L. Galvez Vallejo, B. Westheimer, M. Włoch, P. Xu, F. Zahariev and M. S. Gordon, Recent developments in the general atomic and molecular electronic structure system, *J. Chem. Phys.*, 2020, **152**, 154102.
- 46 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 16 Rev. C.01*, 2016.
- 47 F. Neese, F. Wennmohs, U. Becker and C. Riplinger, The ORCA quantum chemistry program package, *J. Chem. Phys.*, 2020, **152**, 224108.
- 48 I. H. Bell, E. Mickoleit, C. M. Hsieh, S. T. Lin, J. Vrabec, C. Breitkopf and A. Jäger, A Benchmark Open-Source Implementation of COSMO-SAC, *J. Chem. Theory Comput.*, 2020, **16**, 2635–2646.
- 49 T. Gerlach, S. Müller, A. G. de Castilla and I. Smirnova, An open source COSMO-RS implementation and parameterization supporting the efficient implementation of multiple segment descriptors, *Fluid Phase Equilib.*, 2022, **560**, 113472.
- 50 E. Aprà, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y.

- Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauët, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Früchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Götz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jónsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, Á. Vázquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Woliński, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao and R. J. Harrison, NWChem: Past, present, and future, *J. Chem. Phys.*, 2020, **152**, 184102.
- 51 Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S.-N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Yu. Sokolov and G. K.-L. Chan, Recent developments in the PySCF program package, *J. Chem. Phys.*, 2020, **153**, 024109.
- 52 NWChem 7.2.3, 2024, preprint, 7.2.3, <https://github.com/nwchemgit/nwchem/releases/tag/v7.2.3-release>.
- 53 F. Weigend and R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
- 54 D. M. York and M. Karplus, A smooth solvation potential based on the conductor-like screening model, *J. Phys. Chem. A*, 1999, **103**, 11060–11079.
- 55 E. Mullins, R. Oldland, Y. A. Liu, S. Wang, S. I. Sandler, C.-C. Chen, M. Zwolak and K. C. Seavey, Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods, *Ind. Eng. Chem. Res.*, 2006, **45**, 4389–4415.
- 56 J. M. Blaney and J. S. Dixon, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, 1994, pp. 299–335.
- 57 *CRC Handbook of Chemistry and Physics*, ed. J. R. Rumble, CRC Press/Taylor & Francis, Boca Raton, United States, 102nd edn., 2021.
- 58 A. G. de G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrà, Z. Ghahramani and J. Hensman, GPflow: A Gaussian process library using TensorFlow, *J. Mach. Learn. Res.*, 2017, **18**, 1–6.
- 59 F. Lovas, *Diatomc spectral database, NIST standard reference database 114*, National Institute of Standards and Technology, 2002, preprint.