Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

An open-source framework for fast-yet-accurate calculation of quantum mechanical features†

Eike Caldeweyher,*[a] Christoph Bauer,[a] and Ali Soltani Tehrani[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

We present the open-source framework *kallisto* that enables the efficient and robust calculation of quantum mechanical features for atoms and molecules. For a benchmark set of 49 experimental molecular polarizabilities, the predictive power of the presented method competes against second-order perturbation theory in a converged atomic-orbital basis set at a fraction of its computational costs. Robustness tests within a diverse validation set of more than 80,000 molecules show that the calculation of isotropic molecular polarizabilities has a low failure-rate of only 0.3%. We present furthermore a generally applicable van der Waals radius model that is rooted on atomic static polarizabilites. Efficiency tests show that such radii can even be calculated for small- to medium-size proteins where the largest system (SARS-CoV-2 spike protein) has 42,539 atoms. Following the work of Domingo-Alemenara *et al.* [Domingo-Alemenara *et al., Nat. Comm.*, 2019, **10**, 5811], we present computational predictions for retention times for different chromatographic methods and describe how physicochemical features improve the predictive power of machine-learning models that otherwise only rely on two-dimensional features like molecular fingerprints. Additionally, we developed an internal benchmark set of experimental super-critical fluid chromatography retention times. For those methods, improvements of up to 17% are obtained when combining molecular fingerprints with physicochemical descriptors. Shapley additive explanation values show furthermore that the physical nature of the applied features can be retained within the final machine-learning models. We generally recommend the *kallisto* framework as a robust, low-cost, and physically motivated featurizer for upcoming state-of-the-art machine-learning studies.

1 Introduction

The efficient and accurate calculation of molecular physicochemical properties is highly desirable because these descriptors provide insight and explanation for experimental and simulation results. Quantum chemistry, including *ab initio* wave function theory (WFT), and density functional theory (DFT), has been the *via regia* in this field.[1] However, *ab initio* and DFT methods often form a bottleneck due to their computational demands. One avenue to expedite these efforts has been the development of modern semi-empirical quantum chemistry methods.[2–4] Additionally, machine learning (ML) methods have recently emerged that accelerate predictions in the chemical sciences.[5–7] These bypass the time-consuming solution of the *ab initio* equations by training on pre-generated data, for example the computation of absolute energies using neural network potentials.[8–10] Applications that make use of these ML-based potentials include *e.g.*, the prediction of reactivity,[11] and molecular ground- and excited-state properties.[12–14] Combined quantum chemistry and ML workflows have shown the great potential of these data-supported computational chemistry approaches within regioselectivity determination,[15,16] activation energies for organic reactions,[17] as well as in tuning molecular excitation energies.[18] However, in those reports the generation of the quantum chemical features is the computational bottleneck, often involving DFT geometry optimizations. The on-the-fly generation of quantum chemistry-like features using ML methods can alleviate these issues, as shown for example in regioselectivity prediction.[19] Moreover, the prediction of partial charges by random forest models[20] has been of use for molecular dynamics simulations. For molecular polarizabilities, empirical dipole-based models have long been available in the literature[21,22] and used to predict, *e.g.*, retention times (RT) in chromatographic applications.[23]

We have developed the open-source computational framework *kallisto*,[24] which enables the efficient calculation of multiple physicochemical descriptors like atomic partial charges or polarizabilities. In the next section, we introduce the theoretical foundations for calculating such atomic and molecular features. Starting at atomic coordination numbers, we show how they are used to build more complex atomic descriptors like, *e.g.*,

[a]*Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Sweden. E-mail: eike.caldeweyher@astrazeneca.com*

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

environment-dependent atomic partial charges or an efficient yet highly accurate model for atomic and molecular dynamic polarizabilities. Atomic static polarizabilities are furthermore used to build a charge- and environment-dependent van der Waals radius model for all elements up to Radon. An example from pharmaceutical science is highlighting how physicochemical descriptors can be incorporated in a quantitative structure-activity relationship (QSAR) to accurately predict RTs for liquid chromatography (LC) and super-critical fluid chromatography (SFC) applications.

2 Methodologies

We introduce in the following the concepts of coordination numbers, electronegativity equilibration atomic partial charges, as well as how dynamic atomic and molecular polarizabilities are created within the presented method.

2.1 Coordination numbers

Coordination numbers (*CN*) represent the atomic hybridization inside a molecular environment that agrees well with chemical intuition.[25] *CNs* are calculated in a pairwise sum that incorporates atomic covalent radii (R^{cov}) as defined by Pyykkö.[26] Recently, the *CN* has been extended by information about the atomic electronegativity for each atom pair[27] as shown in its definition below

$$CN_i = \sum_i \sum_{j \neq i} \frac{\delta_{AB}^{EN}}{2} \left(1 + \mathrm{erf}\left(-k_0 \left(\frac{R_{AB} - R_{AB}^{cov}}{R_{AB}^{cov}}\right)\right)\right)$$
$$\delta_{AB}^{EN} = \frac{k_1 \exp\left(\mathrm{abs}\left(EN_A - EN_B\right) + k_2\right)^2}{k_3} \tag{1}$$

Herein, *EN* is the Pauling electronegativity,[28] R_{AB} is the internuclear distance of the atom pair *AB*, and R_{AB}^{cov} is the sum of both covalent atomic radii – R_A^{cov} and R_B^{cov}. The parameters were obtained by fitting *CN*-values against GFN2-xtb[4] Wiberg bond orders[29] of diatomic molecules having different *EN*-values (final parameters: $k_0 = 7.5$, $k_1 = 4.1$, $k_2 = 19.09$, and $k_3 = 254.56$).

Fig. 1 shows *CN*-values for the bisphosphine palladium complex $Pd(Cy)_3$, whose structure has been extracted from Ref. 30. The *CN*-values of a selection of atoms (bold) inside the complex show that, e.g., sp^3-hybridized carbon atoms have *CN*-values that corresponds to four covalent binding partners. Hydrogen atoms have one covalent binding partner, while Palladium has two and Phosphorous four, respectively. By introducing a computationally efficient measure for atomic hybridization states we are ready to develop more advanced descriptors. Hence, in the next section we describe the development of an atomic partial charge and an atom-in-molecule polarizability model that both incorporate *CN*-values to capture environment effects of each atom.

2.2 Partial charges and dynamic polarizabilities

Electronegativity equilibration partial charges are determined by minimising the isotropic electrostatic (IES) energy expression

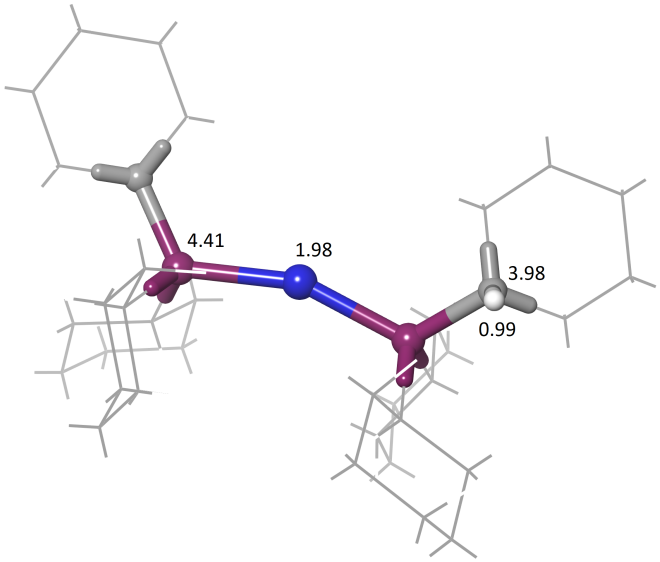


Fig. 1 Palladium catalyst structure of $Pd(Cy)_3$ for which atomic coordination numbers are shown.

with respect to variations in atomic partial charges q

$$E_{IES} = \sum_{i=1}^{N} \left(\chi_i q_i + \frac{1}{2}\left(J_{ii} + \frac{2\gamma_{ii}}{\sqrt{\pi}}\right) q_i^2\right) + \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} q_i q_j \frac{\mathrm{erf}(\gamma_{ij} R_{ij})}{R_{ij}}$$
$$\text{with} \quad \chi_i = EN_i - \kappa_i \sqrt{CN_i}. \tag{2}$$

The first part of the equation describes the on-side interaction for atom i. Environment effects are captured by a first-order term χ_i that incorporates the atomic electronegativity and the κ-scaled coordination number. The second part of the equation describes the pairwise interactions between atom i and all j particles as obtained for interacting charge densities.[31] We apply Lagrangian optimisation to obtain atomic partial charges under the constraint of preserving the total charge q_{tot} of the system

$$\mathscr{L} = E_{IES} + \lambda\left(\sum_{k=1}^{N} q_k - q_{tot}\right)$$
$$\text{with} \quad \frac{\partial \mathscr{L}}{\partial q} = \mathbf{0} \quad \text{and} \quad \frac{\partial \mathscr{L}}{\partial \lambda} = \sum_{k=1}^{N} q_k - q_{total} = 0, \tag{3}$$

which leads to a set of (N + 1) linear equations that can be rewritten in matrix form as

$$\begin{pmatrix} \mathscr{A} & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{q} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathscr{X} \\ q_{tot} \end{pmatrix}. \tag{4}$$

$\mathscr{A}$ incorporates an atomic radius dependent term $\gamma_{ij} = \left(a_i^2 + a_j^2\right)^{-1/2}$ with the atomic radii parameter $a_{i/j}$, as well as the

chemical hardness J_{ii}. Elements of $\mathscr{A}$ are defined as follows

$$\mathscr{A}_{ij} = \begin{cases} J_{ii} + \dfrac{2\gamma_{ii}}{\sqrt{\pi}}, & \text{if } i = j \\ \dfrac{\mathrm{erf}(\gamma_{ij} R_{ij})}{R_{ij}}, & \text{if } i \neq j. \end{cases} \quad (5)$$

There exist overall five parameters per element in this charge model that have been parametrized to match hybrid DFT derived Hirshfeld charges at the PBE0[32]/def2-TZVP[33] level of theory.[27] The parameters per element are: atomic hardness J_{ii}, atomic electronegativity EN_i, *CN*-scaling parameter κ_i, covalent atomic radius R_i^{cov}, and atomic radii a_i used to calculate γ-values as described above.

The model shows remarkable low deviations compared to higher-level and computationally more demanding density functional methods.[27] Atomic partial charges q are further used to calculate atomic charge-dependent dynamic polarizabilities using an empirical scaling method. For this purpose, a charge-scaling function has been designed previously.[27] This charge scheme requires the formulation of effective nuclear charges that are defined as follows

$$z_i = Z_i + q_i, \quad (6)$$

with Z_i being the nuclear charge of atom i, respectively. The charge dependency is incorporated by multiplying precalculated dynamic reference polarizabilities, termed $\alpha_{j,ref}(i\omega)$, – at the time-dependent PBE38[25]/d-aug-def2-QZVP[34] level of theory – by a Gompertz-like scaling term. The scaling term incorporates the atomic chemical hardness τ_j as steepness regulator and the reference and calculated effective charges $z_{j,ref}/z_j$.

$$\alpha_j(i\omega, z_j) = \alpha_{j,ref}(i\omega) \cdot \exp\left[3\left(1 - \exp\left[\tau_j\left(1 - \frac{z_{j,ref}}{z_j}\right)\right]\right)\right] \quad (7)$$

Every element has a set of $N_{i,ref}$ precalculated reference polarizabilities. We apply a Gaussian nearest-neighbors algorithm to interpolate between reference polarizabilities with respect to *CN*-values. Hence the contribution of every reference value to the final atom-in-molecule polarizability of atom i is given by

$$W_i^{i,ref} = \frac{\sum_{j=1}^{N_s} \exp\left(-j \cdot 6\left(CN_i - CN_{i,ref}\right)^2\right)}{\sum_{i,ref=1}^{N_{i,ref}} \sum_{j=1}^{N_s} \exp\left(-j \cdot 6\left(CN_i - CN_{i,ref}\right)^2\right)}, \quad (8)$$

with the sum of all weights equal to unity. The N_s-limit influences the steepness of the Gaussian and is larger than unity for reference systems having small *CN*-differences (e.g., for references describing carbon inside ethene and benzene). The final charge- and environment dependent atomic polarizability is obtained by

$$\alpha_j(i\omega) = \sum_{j,ref=1}^{N_{j,ref}} \alpha_{j,ref}(i\omega, z_j) \cdot W_j^{j,ref}. \quad (9)$$

Molecular isotropic polarizabilities are calculated by adding all N atomic polarizabilities exploiting the additivity of polarizabilities[35]

$$\alpha_{mol}(i\omega) = \sum_j^N \alpha_j(i\omega). \quad (10)$$

3 Results

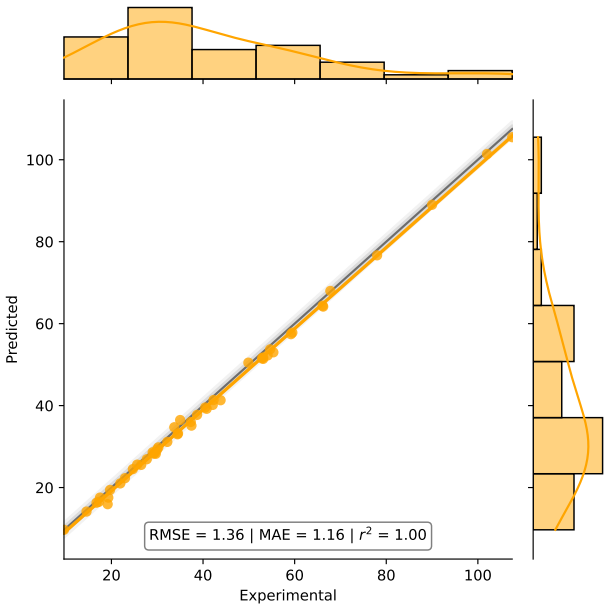
We highlight next how the presented method compares to well-established and computationally more demanding models in terms of predicting experimental molecular polarizabilities. For this purpose, we created a benchmark set of experimental molecular polarizabilities for 49 different organic compounds. We show furthermore how atomic polarizabilities are used to develop a generally applicable model for determining van der Waals (vdW) radii for all elements up to Radon. Last, we show how physicochemical descriptors like the molecular polarizability can be used within a QSAR approach to predict retention times within chromatographic applications in pharmaceutical industry.

3.1 Prediction of experimental molecular polarizabilities

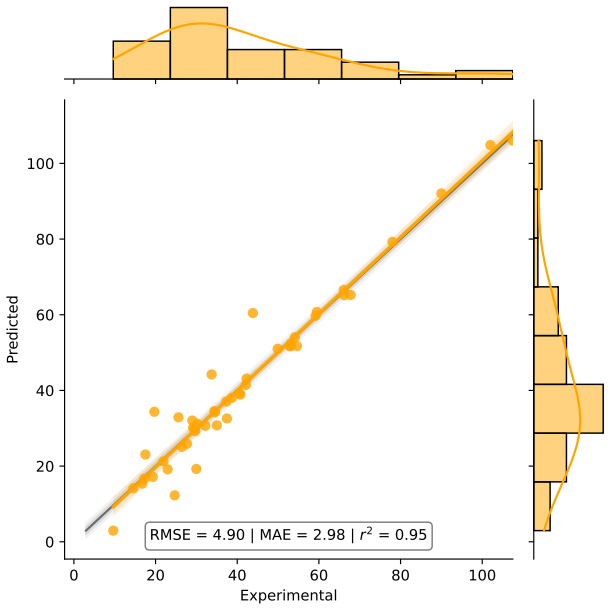
In this section, we compare several methods to predict experimental molecular polarizabilities. Our selection of methods includes an *ab initio* quantum chemical method, a deep-learning framework, a chemoinformatical approach, and lastly the presented method. In order to make a statement about the performance of each method, we created a benchmark set that includes experimentally determined molecular polarizabilities termed MOLPOL135.[36] This benchmark set consists of 135 experimentally obtained static molecular polarizabilities with structures at the CAM-B3LYP[37]-D3(BJ)[25]/def2-TZVP[33] level of theory. Reported experimental molecular polarizabilities are obtained *via* different techniques, *i.e.* dipole oscillator, refractive index, dielectric permittivity, or electron-molecule scattering measurements.

Some of the predictive models are not yet parametrized for the whole periodic table (see below). We therefore extracted a subset of the MOLPOL135 consisting of 49 organic compounds (a list of all compounds is given in the Supplementary Information) that incorporate the following elements: Hydrogen, Carbon, Nitrogen, Oxygen, Sulfur, and Chlorine. Correlation plots for each method (termed as "Predicted") with respect to the experimental molecular polarizabilities (termed as "Experimental") are given in Fig. 2. The root mean squared error (RMSE), the mean absolute error (MAE), the coefficient of determination (r^2), and the computer-timing measure to calculate all systems of consideration once ($\sum t_{\mathrm{CPU}}$) are listed as well.

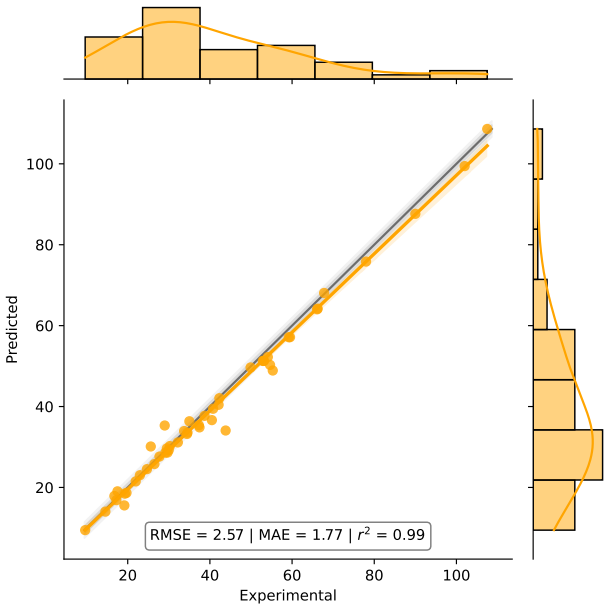
We extracted second-order Møller-Plesset perturbation theory (MP2) molecular polarizabilities obtained in a large atomic-orbital basis set (def2-QZVPD[34]) from Ref. 38 (abbreviated as MP2/QZ in the following). MP2/QZ has the highest $\sum t_{\mathrm{CPU}}$-value, which is expected since we need to solve the *non*-relativistic Schrödinger equation and perform perturbation theory of second order on-top where a molecular-orbital (MO) transformation is necessary that formally scales $\mathscr{O}(N^5)$ with respect to the applied basis set functions N. The high computer cost comes along with the lowest RMSE (1.36 Bohr^3) and MAE (1.16 Bohr^3) values across all tested methods for the prediction of experimental molecular polarizabilities. The coefficient of determination



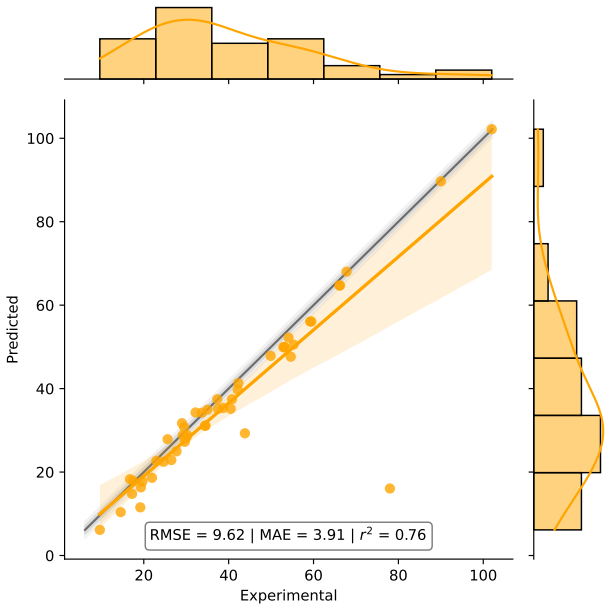
(a) MP2/def2-QZVPD ($\sum t_{\mathrm{CPU}} \propto$ days)



(b) *AlphaML* ($\sum t_{\mathrm{CPU}} \propto$ minutes)



(c) *kallisto* ($\sum t_{\mathrm{CPU}} \propto$ seconds)



(d) *Meanpol* ($\sum t_{\mathrm{CPU}} \propto$ milliseconds)

Fig. 2 Correlation plots of experimental molecular polarizabilities for 49 organic molecules against MP2/def2-QZVPD[34], *AlphaML*, *kallisto* , and *Meanpol* listed with respect to decreasing computer time t_{CPU}. Given are furthermore the root mean squared error (RMSE), the mean absolute error (MAE), and the coefficient of determination (r^2). Two outliers – CS_2 and O_3 – were removed from the *AlphaML* set (listed in the Supplementary Information).

shows a perfect correlation to the experimental references. However, due to the high computational costs arising from diagonalizing the Hamiltonian and performing the MO transformation, this method is not routinely applicable on large scales, *e.g.*, screening hundreds to thousands of compounds. Hence we can only recommended MP2/QZ as a theoretical reference method, but not generally for the fast prediction of molecular polarizabilities.

Another promising candidate of being an efficient and accurate model for the prediction of molecular polarizabilities is the recently published *AlphaML* deep learning model as developed by Ceriotti and co-workers.[39] References for this method have been obtained at the linear-response coupled-cluster method including single and double excitations (CCSD) in a d-aug-cc-pVDZ basis set. A symmetry-adapted Gaussian process regression scheme has been created and tested on a set of 52 larger molecular systems to validate the accuracy for predicting molecular polarizabilities.[39] We calculate *AlphaML* molecular static polarizability as the trace of the obtained polarizability tensor $\mathscr{P}$ – as introduced by Ceriotti and co-workers in a recent work.[40]

$$\alpha_{mol}^{AlphaML} = \frac{1}{3}\,\mathrm{Tr}\,\mathscr{P} \tag{11}$$

Computer times for this method are orders of magnitudes below the MP2/QZ method reaching an overall scope of minutes in computer time. This, however, comes with the drawback of increasing RMSE (4.90 Bohr^3) and MAE values (2.98 Bohr^3). Unfortunately, *AlphaML* suffers from robustness issues in terms of predicting physically correct molecular polarizabilities. Within our set, two molecular systems obtained unphysical molecular polarizabilities: First, the molecular polarizability of ozone (O_3) has been overestimated several orders of magnitudes. Second, the molecular polarizability of carbon disulfide (CS_2) was predicted to be negative. Even though the performance of the *AlphaML* method is good in terms of computational speed, this could not compensate the missing availability of elements across the periodic system and the mentioned robustness issues. These limitations hinder the general application of *AlphaML* for the fast, robust, and accurate prediction of molecular polarizabilities.

The next predictive method has been developed within the field of chemoinformatics.[41] This method is the conceptionally simplest one among all of our tested approaches. Obtaining the molecular polarizability is straightforward since we only need to add up averaged atomic polarizabilities $\overline{\alpha}$ to obtain this molecular property. We therefore term this method as *Meanpol* in the following.

$$\alpha_{mol}^{Meanpol} = \sum_i \overline{\alpha}_i \tag{12}$$

In their work, Bosque and Sales published $\overline{\alpha}$-values for 10 different elements (Carbon, Hydrogen, Oxygen, Nitrogen, Sulfur, Phosphorous, Fluorine, Chlorine, Bromine, and Iodine). Since only the chemical composition is necessary, this method is by far the most efficient one, calculating all molecular polarizabilities in only milliseconds of computer time. Nevertheless, this gain in efficiency also comes with the price of the overall highest RMSE (9.62 Bohr^3) and MAE (3.91 Bohr^3) values across all tested methods. Nevertheless, it is remarkable that such a simple method is able to predict experimental polarizabilities with a reasonable accuracy. Though this method is not accurate enough to be generally applicable in the sense of offering a fast yet reliable predictive model, however, it can be used for the fast guess of molecular polarizabilities when averaged atomic references are available for the system of interest.

Last, we discuss results for the *kallisto* method whose theoretical framework has already been introduced in section 2. This method uses the same strategy as the *Meanpol* approach, however, instead of applying averaged atomic polarizabilities, references are interpolated with respect to a Gaussian nearest neighbour approach using atomic *CNs*. This enables the explicit description of environment effects and the implicit modulation of many-body dispersion effects that generally lower atomic polarizabilities. Compared to MP2/QZ, and *AlphaML*, the computer time is lowered achieving a scope of seconds to calculate all molecular polarizabilities within our test set. Note that the computer time can even be lowered below one second when additionally a pre-compiled shared library that handles intense linear algebra calculations in the back-end is included.[27] However, for the sake of applicability and user-friendliness we decided to stick to an implementation that does not incorporate pre-compiled sources and is as such easier to distribute and easier to install by the user.

With the second lowest RMSE (2.57 Bohr^3) and MAE (1.77 Bohr^3) the *kallisto* method hits the "sweet-spot" between accuracy and efficiency. Also the coefficient of determination ($r^2 = 0.99$) shows almost the same quality of correlation as the orders of magnitudes computationally more demanding MP2/QZ method. Furthermore, this method is parameterized for all elements up to Radon ($Z = 86$), which enables a far wider scope of applicability compared to *AlphaML* and *Meanpol*. We tested the robustness for obtaining molecular polarizabilities on a molecular set having more than 80,000 organic compounds.[42] Only 0.3% of the cases were not successful, which shows generally a high robustness for this method. Overall, this enables the efficient, robust, and accurate prediction of molecular polarizabilities for large parts of the periodic table reaching from small to large system sizes. We therefore generally recommend the *kallisto* method for the calculation of accurate molecular polarizabilities. In the next section, we use atomic polarizabilities as obtained by *kallisto* to develop a quantum mechanical van der Waals radii model applicable for all elements up to Radon.

3.2 van der Waals radii model

The concept of a vdW radius was pioneered by Pauling and Bondi,[28,43] who both defined this radius as half of the distance between two atoms of the same chemical element. At this distance Pauli exchange repulsion and London dispersion attraction forces exactly balance each other. However, the determination of the atomic vdW radius is unambiguous for noble gases only, while the definition breaks apart for other elements. Hence, a robust determination of vdW radii for most elements in the periodic table requires a high amount of experimental structural data.[44] It is therefore of interest to have theoretical models that allow the accurate prediction of vdW radii across large parts of the periodic

1.54 H	Radius in Ångström																1.34 He
2.20 Li	2.19 Be	Element										2.05 B	1.90 C	1.79 N	1.71 O	1.63 F	1.56 Ne
2.25 Na	2.40 Mg											2.39 Al	2.32 Si	2.23 P	2.14 S	2.06 Cl	1.97 Ar
2.34 K	2.70 Ca	2.63 Sc	2.57 Ti	2.52 V	2.33 Cr	2.42 Mn	2.37 Fe	2.33 Co	2.29 Ni	2.17 Cu	2.22 Zn	2.33 Ga	2.34 Ge	2.31 As	2.24 Se	2.19 Br	2.12 Kr
2.40 Rb	2.79 Sr	2.74 Y	2.69 Zr	2.51 Nb	2.44 Mo	2.52 Tc	2.37 Ru	2.33 Rh	2.15 Pd	2.25 Ag	2.38 Cd	2.46 In	2.48 Sn	2.46 Sb	2.42 Te	2.38 I	2.32 Xe
2.49 Cs	2.93 Ba		2.64 Hf	2.58 Ta	2.53 W	2.49 Re	2.44 Os	2.40 Ir	2.30 Pt	2.26 Au	2.29 Hg	2.42 Tl	2.49 Pb	2.50 Bi	2.50 Po	2.47 At	2.43 Rn

Fig. 3 Van der Waals radii derived from static polarizabilities for all elements up to Radon. Shown are atomic values for $CN = 0$ and $q = 0$ using the default vdW parametrization. All values have been calculated using the *kallisto* command-line interface.

table.

The classical relationship between the atomic polarizability and the vdW radius defines an atom as a positive charge q surrounded by a uniform electron density within a hard sphere having radius r. Once an external electrical field ε is acting on the point charge, it undergoes a displacement d with respect to the center of the sphere. By including the induced dipole moment ($qd = \alpha\varepsilon$) we obtain

$$q\varepsilon - \frac{q^2 d}{r_{vdW}^3} = 0 \Leftrightarrow q\varepsilon - \frac{q\alpha\varepsilon}{r_{vdW}^3} = 0 \Leftrightarrow r_{vdW} = \alpha^{1/3}. \quad (13)$$

Another definition has recently been popularized by Fedorov *et al.*[45] showing the quantum-mechanical relation between the two quantities based on the Tang–Toennies model.[46,47] Their model consists purely of a London dispersion and a Pauli exchange repulsion part and, furthermore, a dipole approximation is employed to the Coulomb potential. The application of quantum Drude oscillators[48–50] leads to simplified expressions for both the attractive London dispersion (Disp) and the repulsive Pauli repulsion (X). Force equilibration finally reveals the important relation (see Ref. 45 for the exact derivation)

$$F_{\mathrm{X}} + F_{\mathrm{Disp}} = 0 \rightarrow r_{vdW} = \theta_a \alpha^{1/7}. \quad (14)$$

By fitting this scaling law to reference data of noble gases, the authors obtained $\theta_a = 2.54$ as their central result. Since the present model should be applicable to all elements up to Radon, an additional element-wise parameter θ_b is introduced, which is fitted to reproduce theoretically obtained atomic vdW radii.

$$r_{vdW} = \theta_b \theta_a \alpha^{1/7} \quad (15)$$

Two parametrizations are available – termed "rahm" (default) and "truhlar" –, which have been obtained by fitting against atomic vdW radii as reported in Ref. 51 and Ref. 52, respectively. For the calculation of vdW radii, we apply the static atomic polarizability. Fig. 3 shows calculated atomic vdW radii in Ångström for all elements up to Radon. Fitted θ_b values are shared in the reference implementation.[24]

Due to many-body effects, the atomic polarizability decreases significantly when additional covalent bonds are formed.[53] Since the proposed vdW scheme is directly dependent on the atomic polarizability, we expect that the radii decrease with increasing coordination numbers. This trend is correctly observed for all tested systems (see Table 1 in the Supplementary Information).

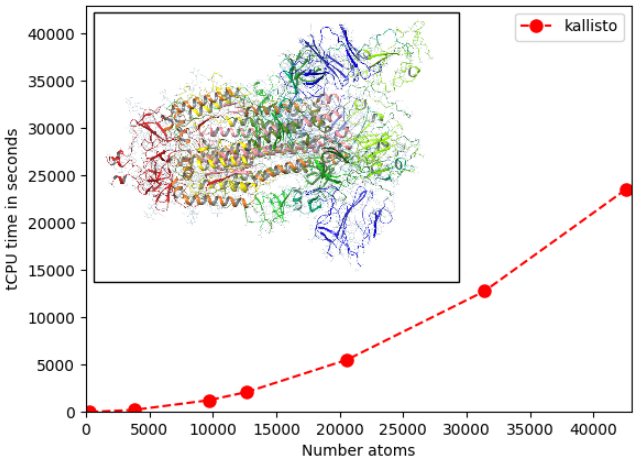


Fig. 4 Computer times as obtained by *kallisto* for different sized proteins. Our selection of proteins is given as follows: PDB code, name (number of atoms). 1L2Y[54], Trp-Cage miniprotein (302); 1EMA[55], Green fluorescent protein (3,784); 1GZX[56], Haemoglobin (9,689); 1CC1[57], Ni-Fe-Se hydrogenase (12,689); 1GPE[58], Glucose oxidase (20,561); 6LZ3[59], Cryptochrome (31,396); 7AD1[60], SARS-CoV-2 spike protein (42,539). Timings obtained on a single Intel(R) Xeon(R) Gold 6140 CPU@2.30GHz processor.

Additionally, charge-scaling effects show a decrease for vdW radii with increasing cationic character. This effect, however, becomes less pronounced for atoms having larger ordinal numbers. Changing the electronic state from a neutral atom to a cation results for the vdW radius of Carbon in a 5% decrease while the Iridium radius decreases only by 2%. This smaller influence is expected due to the application of effective charges in the charge-

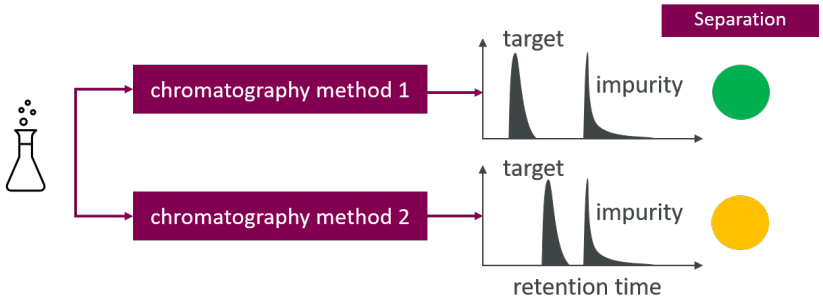


Fig. 5 The impact of chromatography method on the separation of target analytes and impurities. Each method has its discrete set of conditions (stationary phase, mobile phase, including gradients). A larger retention time difference implies better separation quality (green and amber markers).

scaling procedure (equation 7), which reduce the relative charge effect for higher nuclear charges. To the best of the authors knowledge, this enables for the first time the efficient and robust calculation of environment- and charge-dependent vdW radii using quantum mechanical data across large parts of the periodic table.

Apart from small molecules applications, other interesting areas might include larger system sizes. We exemplify the generation of vdW radii for the scope of small- to medium-sized proteins, *i.e.* several hundred to thousands of atoms as shown in Fig. 4. To estimate its feasibility and efficiency, we begin with a designed 20-residue miniprotein[54] consisting of 304 atoms and end with the SARS-CoV-2 spike protein[60] having 42,539 atoms (shown as ribbon inlet in Fig. 4). In our selection, we include proteins with transition metals (Haemoglobin and Ni-Fe-Se hydrogenase) to show the generality with respect to applied atom types.

Note that even for the largest SARS-CoV-2 spike protein no memory issues appeared, which highlights the robustness of the presented method. Possible areas of application for vdW radii are, *e.g.*, the construction of new or the enhancement of well-established force fields,[61,62] the improvement of state-of-the-art implicit solvation models as commonly applied in quantum chemistry,[63] or the calculation of features for protein QSAR approaches.[64] Ongoing work applies the proposed vdW radius model for the construction of energy terms within a low-cost free-energy approach that is intended to share insights about hydrogen-bond strengths.

We discuss in the following the impact of the presented method within one example applications from the pharmaceutical industry. We show how some of the proposed physicochemical molecular descriptors can be incorporated in a machine-learning approach to predict experimental RTs for different chromatographic methods.

3.3 Pharmaceutical industry machine-learning application

In the pharmaceutical industry, LC and SFC methods coupled with mass spectrometry are commonly used to separate the target compound from impurities. Depending on the compound and the analytical method, each compound (target or impurity) elutes at different times from the chromatographic columns, commonly referred to as its retention time (RT). Generally, RTs are highly influenced by method-specific conditions, *e.g.*, stationary phase, mobile phase, and applied gradients. Purification is achieved by only collecting the solution at the appropriate RT where the target compound is expected to appear.

Fig. 5 exemplifies this for two different chromatographic methods that obtain different RTs for the target and impurity compound. Here, "chromatographic method 1" is superior in terms of separating the target and impurity compound and hence this setup both simplifies the purification step and ensures higher purity in the resulting solution. Accurate predictions of RTs for the target and impurity compound are therefore of general interest to develop, *e.g.*, an automated recommender system that predicts the best chromatographic method for a given target-impurity pair. The basis for such a recommender system is given by computer-supported algorithms that enable the accurate prediction of RTs for different chromatographic methods and hence enable chemists to make quantitative statements with respect to their chromatographic method choice. Recently, researchers have been using machine-learning methods to accurately predict RTs, for both the target compound and impurities.[42,65–67]

We present in the following how the accuracy of predicted RTs for LC and SFC methods is impacted by including certain molecular information *via* physicochemical descriptors as obtained by the *kallisto* featurizer. Many different molecular descriptors and fingerprints have been proposed in the literature, and comparing *kallisto* descriptors to all available descriptors is out of scope of this work. We follow the recently published approach of Domingo-Almenara *et al.*, who proposed that extended connectivity fingerprints (ECFP) generally outperform physicochemical based molecular descriptors for RT modelling.[42]

We take their recommendation as baseline and enhance it by physicochemical descriptors as calculated by the presented method. Physicochemical descriptors are selected to account qualitatively for basic interactions between chromatographic column material and target or impurity compound. We choose the isotropic electrostatic energy (IES) of a molecule to account for electrostatic interactions and molecular polarizabilities (MolPol) to account for London dispersion (LD) interactions. Since the pair-wise and leading order LD term is directly proportional to the polarizability of a molecule (Casimir–Polder integration),[68] the MolPol should be suitable for this interaction type. In the next section, we present the data acquisition process and featurization

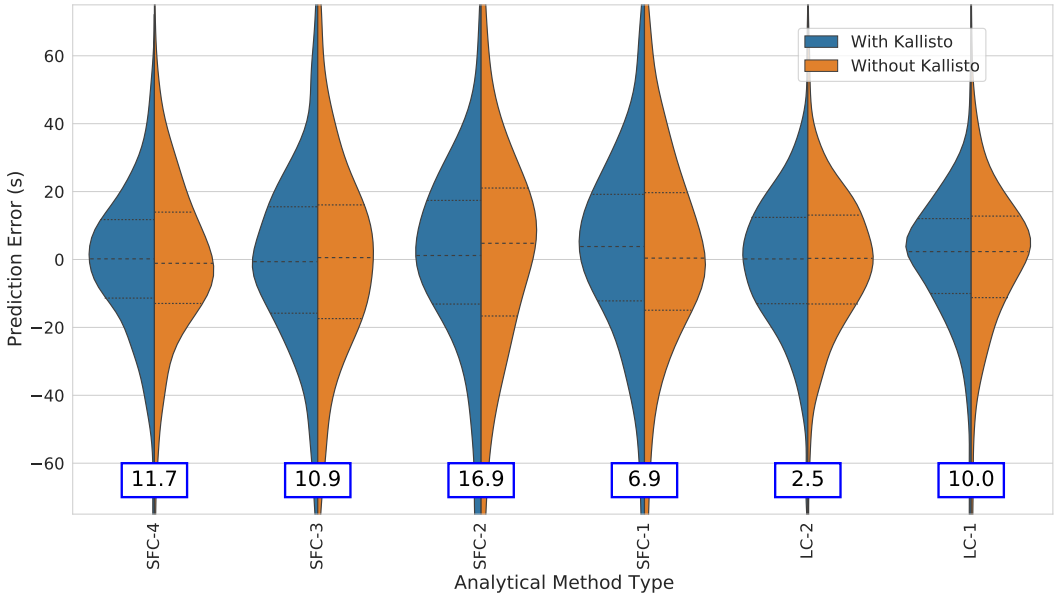


Fig. 6 Retention time prediction MAE for the various methods with and without *kallisto* descriptors. The left side of distributions show errors with *kallisto*, while the right side shows them without. Number in the boxes show the improvement in MAE% using *kallisto* descriptors.

strategy, which is followed by a comparison between modelling performances of the baseline and the enhanced model.

3.3.1 Defining the training and validation set

In this work we use the *METLIN* dataset[42] consisting of 80,308 small molecule chromatograms and an internal dataset of 17,108 small molecules experimented on two different LC mass spectrometry and five different SFC mass spectrometry analytical methods (termed *AZset*). For the *METLIN* dataset, the validation set is derived from a (75/25) train/validation split, while for the *AZset*, the training set is composed of *non*-publicly available compounds, while the validation set are the compounds that are publicly available. This public set is shared in the Supplementary Information.

3.3.2 Predicting experimental retention times

Following the strategy of Domingo-Almenara *et al.*, we extract ECFP4 fingerprints for each compound using RDKIT.[69] After generating three-dimensional coordinates (*xmol* files) *via* OPENBABEL,[70] MolPol and IES values are calculated using the *kallisto* command-line interface.[24] With the fingerprints and *kallisto* descriptors as features, and the experimental RTs as the labels, we train a random forest regressor with 100 decision trees and max depth of 10 using SCIKIT LEARN.[71] Due to the high number of features, we also apply a feature reduction scheme to remove features with variance lower than 0.05, and remove correlated features where the correlation is higher than 0.9. The resulting regressor is used to predict RTs on the validation set.

We reproduce the results of Domingo-Almenara *et al.* on the *METLIN* dataset, where the trained random forest model achieves a RT prediction MAE and RMSE of 42s and 66s, respectively. Using the *kallisto* descriptors yields an MAE of 39s and an RMSE of 62s. This corresponds to improvements in RT predictions of 7.1% and 6.1%.

Furthermore, we create models for the *AZset* methods – namely LC-1, LC-2, SFC-1, SFC-2, SFC-3, SFC-4 – and predict RTs on the validation set for which violin plots of the MAE for all analytical method RTs are shown in Fig. 6.

In this figure, the prediction error using *kallisto* descriptors is shown in blue on the left-hand split of the distributions, while the *non-kallisto* distribution is given in orange on the right-hand split. The improvements in MAE in terms of percentages are shown in the boxes below each distribution. The modelling and RT prediction using *kallisto* show promising results, with improvements ranging from 2.5% to 16.9% with generally SFC methods seeming to see the highest improvement in RT predictions.

Last, we prepared Shapley additive explanations (*SHAP*) to analyze the RT prediction by computing the contribution of each feature.[72,73] An overview of which features are the most important for the trained model is given in Fig. 7. This plot sorts features by the sum of *SHAP* value magnitudes over all samples, and uses *SHAP* values to show the distribution of the impacts each feature has on the model output. The color code represents the feature value, where red is high and blue is low. Overall, the MolPol and IES features are rated as the most and third-most important features within the random forest regressor.

Fig. 7 furthermore shows that high MolPol values increase the predicted RT as expected from increasing LD interactions with increasing molecular size (additivity of polarizabilities). Hence, we reproduce this physical effect within our trained random forest regressor. The IES feature behaves somehow different, where low IES-values increase the predicted RT. Since the IES is an energetic contribution, its sign is generally negative, but positive values

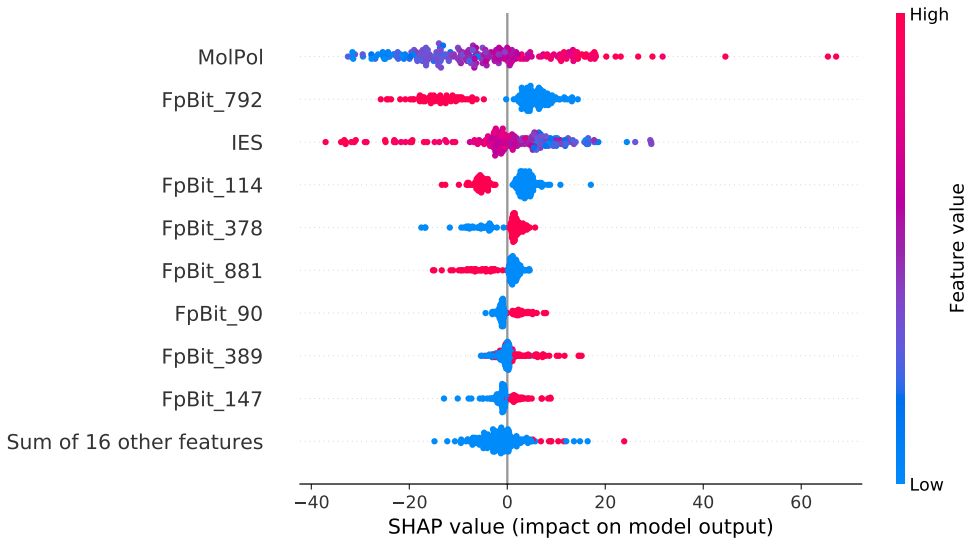


Fig. 7 SHAP importances of the top 25 features for the random forest model trained on the SFC-2 dataset. Molecular Polarizability ranks as the top feature, while IES ranks as third.

can occur for charged molecules. However, we excluded charged molecules in our selection and hence no positive IES-values are present. The molecular size is implicitly captured within this feature, showing a more negative value for larger molecular size.

Compared to the ECFP4 fingerprint (FpBit) features, both MolPol and IES have a larger span of *SHAP* values showing that those physicochemical features might have a broader impact on the model performance. Further analysis in terms of feature importance showed that both MolPol and IES ranked in the top 3 important features for model predictions (see Supplementary Information), which shows their importance and suitability for use in RT modelling.

4 Conclusions

We have presented the theory and main features of the open-source framework *kallisto*, which enables the efficient and robust calculation of atomic and molecular features designed for machine-learning applications. Atom-in-molecule dipole polarizabilities are obtained by a Gaussian nearest-neighbor algorithm (equation 8) that interpolates charge-scaled polarizabilties such that charge and environment effects are represented. The partial charges used within the charge-scaling step are obtained by a classical electronegativity equilibration model. For a benchmark set of 49 experimental molecular polarizabilities, the presented method obtains similar accuracies as our *ab initio* method within a nearly converged atomic-orbital basis set (MP2/def2-QZVPD), but at only a fraction of its computational costs. Robustness test for the creation of molecular polarizabilies on a test set including more than 80,000 molecules are conducted showing a remarkably low failure rate of 0.3%.

Furthermore, a model is presented that exploits the connection between the quantum mechanically derived atomic polarizability and the van der Waals (vdW) radius of an atom as initially proposed by Fedorov *et al.*. The effect of many-body perturbations is retained showing that vdW radii decrease for increasing crowdness of an atom. Furthermore, charge effects show that vdW radii shrink for atoms that have a higher cationic nature and *vice versa*. Efficiency tests exhibit that this vdW model is amenable for small- to medium-sized proteins including also transition-metal atoms, where the largest system tested has 42,539 atoms. To the best of the authors knowledge this enables the low-cost calculation of quantum mechanically derived vdW radii for the first time.

An example from pharmaceutical industry highlights that physicochemical descriptors improve the accuracy of machine-learning models that otherwise only include two-dimensional features like molecular fingerprints. The molecular polarizability and the isotropic electrostatic energy are used within a quantitative structure activity relationship procedure to predict retention times for different chromatographic methods including liquid chromatography (LC) as well as super-critical fluid chromatography (SFC). On the established *METLIN* dataset improvements of up to 7% are archived for LC methods. An internally created benchmark set for experimentally determined SFC retention times exhibits an even higher improvement showing an accuracy gain of up to 17% while including the two physicochemical features within a random-forest model. Furthermore, to the best of the authors knowledge this is the first time that predictions for SFC methods are published within the literature.

An online documentation that covers bits of the underlying theory has been created including a selection of copy-paste recipes for a quick-start into production.[74] The framework is listed in the python package index and hence easily installable *via* the *pip* command-line interface.* The program code of the presented framework is available and maintained at GitHub[75] and published under the Apache 2.0 license.[24]

* To install the *kallisto* framework into your virtual environment: *pip install kallisto*

Author Contributions

Eike Caldeweyher: Conceptualization; supervision; methodology; data curation; data analysis; writing-original draft. **Christoph Bauer:** Data curation; data analysis; writing-original draft. **Ali Soltani Tehrani:** Machine-learning modelling; data curation; data analysis; writing-original draft.

Conflicts of interest

There are no conflicts of interest.

Acknowledgements

The authors would like the thank Hanna Leek and Kristina Öhéln from the *Separation Sciences Lab* at AstraZeneca for providing support and guidance, LC/SFC data, and help with interpreting the results. We would also like to thank Anders Bertilson for help in extracting some relevant metadata from the experiments and Johan Ulander for driving the data acquisition process. EC is a fellow of the AstraZeneca post doc programme.

Notes and references

1 W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory*, Wiley-VCH Verlag GmbH, Weinheim, Germany, 2nd edn, 2001.

2 A. S. Christensen, T. Kubař, Q. Cui and M. Elstner, *Chem. Rev.*, 2016, **116**, 5301–5337.

3 P. O. Dral, X. Wu and W. Thiel, *J. Chem. Theory Comput.*, 2019, **15**, 1743–1760.

4 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, *WIREs Comput. Mol. Sci.*, 2021, **11**, e1493.

5 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.

6 A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.*, 2019, **59**, 2545–2559.

7 D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, *Mol. Syst. Des. Eng.*, 2019, **4**, 828–849.

8 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

9 J. Behler, *Angew. Chem. Int. Ed.*, 2017, **56**, 12828–12840.

10 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

11 K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P.-O. Norrby, *Nat. Rev. Chem.*, 2021, **5**, 240–255.

12 W. P. Walters and R. Barzilay, *Acc. Chem. Res.*, 2021, **54**, 263–270.

13 J. Westermayr and P. Marquetand, *Chem. Rev.*, 2021, **121**, 9873–9926.

14 P. O. Dral and M. Barbatti, *Nat. Rev. Chem.*, 2021, **5**, 388–405.

15 A. Tomberg, M. J. Johansson and P. O. Norrby, *J. Org. Chem.*, 2019, **84**, 4695–4703.

16 N. Ree, A. H. Goller and J. H. Jensen, *J Cheminform*, 2021, **13**, 10.

17 K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.

18 M. Sumita, X. Yang, S. Ishihara, R. Tamura and K. Tsuda, *ACS Cent. Sci.*, 2018, **4**, 1126–1133.

19 Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 2198–2208.

20 P. Bleiziffer, K. Schaller and S. Riniker, *J. Chem. Inf. Model*, 2018, **58**, 579–590.

21 K. J. Miller and J. Savchik, *J. Am. Chem. Sco.*, 1979, **101**, 7206–7213.

22 L. Jensen, P.-O. Åstrand, A. Osted, J. Kongsted and K. V. Mikkelsen, *J. Chem. Phys.*, 2002, **116**, 4001–4010.

23 R. Rohrbaugh and P. Jurs, *Anal Chem*, 1988, **60**, 2249–2253.

24 E. Caldeweyher, *J. Open Source Softw.*, 2021, **6**, 3050.

25 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

26 P. Pyykkö and M. Atsumi, *Chem. Eur. J.*, 2009, **15**, 186–197.

27 E. Caldeweyher, S. Ehlert, A. Hansen, H. Neugebauer, S. Spicher, C. Bannwarth and S. Grimme, *J. Chem. Phys.*, 2019, **150**, 154122.

28 L. Pauling *et al.*, *The Nature of the Chemical Bond*, Cornell university press Ithaca, NY, 1960, vol. 260.

29 K. B. Wiberg, *Tetrahedron*, 1968, **24**, 1083–1096.

30 M. Bursch, E. Caldeweyher, A. Hansen, H. Neugebauer, S. Ehlert and S. Grimme, *Acc. Chem. Res.*, 2018, **52**, 258–266.

31 S. A. Ghasemi, A. Hofstetter, S. Saha and S. Goedecker, *Phys. Rev. B*, 2015, **92**, 045131.

32 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.

33 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

34 F. Weigend, F. Furche and R. Ahlrichs, *J. Chem. Phys.*, 2003, **119**, 12753–12762.

35 K. J. Miller, *J. Am. Chem. Soc.*, 1990, **112**, 8533–8542.

36 E. Caldeweyher, *MolPol135: A benchmark set for static molecular polarizabilities*, `https://github.com/f3rmion/molpol135`, Accessed: 2021-05-12.

37 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51–57.

38 A. J. Thakkar and T. Wu, *J. Chem. Phys.*, 2015, **143**, 144302.

39 D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio and M. Ceriotti, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 3401–3406.

40 Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti and R. A. DiStasio, *Sci. data*, 2019, **6**, 1–10.

41 R. Bosque and J. Sales, *J. Chem. Inform. Comput. Sci.*, 2002, **42**, 1154–1163.

42 X. Domingo-Almenara, C. Guijas, E. Billings, R. Montenegro-Burke, W. Uritboonthai1, A. Aisporna1, E. Chen, P. Benton and G. Siuzdak, *Nat Commun*, 2019, **10**, 5811.

43 A. v. Bondi, *J. Phys. Chem.*, 1964, **68**, 441–451.

44 S. S. Batsanov, *Inorg. Mater.*, 2001, **37**, 871–885.

45 D. V. Fedorov, M. Sadhukhan, M. Stöhr and A. Tkatchenko, *Phys. Rev. Lett.*, 2018, **121**, 183401.

46 K. Tang, J. Toennies and C. Yiu, *Phys. Rev. Lett.*, 1995, **74**, 1546.
47 K. Tang, J. Toennies and C. Yiu, *Int. Rev. Phys. Chem.*, 1998, **17**, 363–406.
48 F. Wang and K. Jordan, *J. Chem. Phys.*, 2001, **114**, 10717–10724.
49 T. Sommerfeld and K. D. Jordan, *J. Phys. Chem. A*, 2005, **109**, 11531–11538.
50 A. P. Jones, J. Crain, V. P. Sokhan, T. W. Whitfield and G. J. Martyna, *Phys. Rev. B*, 2013, **87**, 144103.
51 M. Rahm, R. Hoffmann and N. Ashcroft, *Chem. Eur. J.*, 2017, **23**, 4017–4017.
52 M. Mantina, A. C. Chamberlin, R. Valero, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, 2009, **113**, 5806–5812.
53 J. F. Dobson, *Int. J. Quantum Chem.*, 2014, **114**, 1157–1161.
54 J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, *Nat. Struct. Biol.*, 2002, **9**, 425–430.
55 M. Ormö, A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien and S. J. Remington, *Science*, 1996, **273**, 1392–1395.
56 M. Paoli, R. Liddington, J. Tame, A. Wilkinson and G. Dodson, *J. Mol. Biol*, 1996, **256**, 775–792.
57 E. Garcin, X. Vernede, E. Hatchikian, A. Volbeda, M. Frey and J. Fontecilla-Camps, *Structure*, 1999, **7**, 557–566.
58 G. Wohlfahrt, S. Witt, J. Hendle, D. Schomburg, H. M. Kalisz and H.-J. Hecht, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1999, **55**, 969–977.
59 K. Shao, X. Zhang, X. Li, Y. Hao, X. Huang, M. Ma, M. Zhang, F. Yu, H. Liu and P. Zhang, *Nat. Struct Mol. Biol.*, 2020, **27**, 480–488.
60 J. Juraszek, L. Rutten, S. Blokland, P. Bouchier, R. Voorzaat, T. Ritschel, M. J. Bakkers, L. L. Renault and J. P. Langedijk, *Nat. Commun.*, 2021, **12**, 1–8.
61 W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.
62 W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.
63 A. Klamt, *WIREs Comput. Mol. Sci.*, 2011, **1**, 699–709.
64 L. Sun, H. Yang, J. Li, T. Wang, W. Li, G. Liu and Y. Tang, *ChemMedChem*, 2018, **13**, 572–581.
65 P. R. Haddad, M. Taraji and R. Szücs, *Anal. Chem.*, 2021, **93**, 228–256.
66 M. Witting and S. Böcker, *J. Sep. Sci.*, 2020, **43**, 1746–1754.
67 Q. Yang, H. Ji, H. Lu and Z. Zhang, *Anal. Chem.*, 2021, **93**, 2200–2206.
68 S. Grimme, A. Hansen, J. G. Brandenburg and C. Bannwarth, *Chem. Rev.*, 2016, **116**, 5105–5154.
69 G. A. Landrum, *RDKit: Open-Source Cheminformatics Software; version XX*, `https://github.com/rdkit/rdkit`, Accessed: 2021-08-27.
70 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminformatics*, 2011, **3**, 1–14.
71 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn Res.*, 2011, **12**, 2825–2830.
72 S. M. Lundberg and S.-I. Lee, *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.
73 S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, *Nat. Mach. Intell.*, 2020, **2**, 2522–5839.
74 E. Caldeweyher, *User Guide to Atomic Featurizer Kallisto*, `https://app.gitbook.com/@ehjc/s/kallisto/`, Accessed: 2021-05-12.
75 E. Caldeweyher and P. Pracht, *The kallisto atomic featurizer*, `https://github.com/AstraZeneca/kallisto`, Accessed: 2021-09-23.
76 *AlphaML: machine learning of molecular polarizabilities*, `https://tools.materialscloud.org/alphaml/input_structure/`, Accessed: 2021-05-12.
77 *RCSB Protein Data Bank*, `https://www.rcsb.org/pages/policies#References`, Accessed: 2021-06-28.

5 Supplementary Information

5.1 Extended statistical measures

As statistical measure for a set $\{x_1,\ldots,x_n\}$ of data points with references $\{r_1,\ldots,r_n\}$ we use

- Root mean squared error: RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i-x_i)^2}$
- Mean: $m = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mean absolute error: MAE $= \frac{1}{n}\sum_{i=1}^{n}\mathrm{abs}\,(r_i-x_i)$
- Coefficient of determination: $r^2 = 1-\frac{SS_{res}}{SS_{tot}}$

  Residual of squares: $SS_{res} = \sum_{i=1}^{n}(r_i-f_i)^2 = \sum_{i=1}^{n} e_i^2$

  Total sum of squares: $SS_{tot} = \sum_{i=1}^{n}(x_i-m)^2$

5.2 Charge- and environment effects for van der Waals radii

The table below shows the environment- and charge dependency of van der Waals (vdW) radii calculated for different atom types (Carbon, Nitrogen, Oxygen, Phosphorous, and Iridium). Overall, three different CN-values ($CN = 0,1,2$) and three different atomic partial charges q ($q = 0.0, 0.5, 1.0$) have been chosen within the determination of vdW radii. All radii have been calculated using the *kallisto* command-line interface with its default parameterization ("rahm").

5.3 Molecular polarizabilities

This benchmark set is a subset of the *MOLPOL135* benchmark set[36], whose experimental molecular polarizabilities have been determined by either dipole oscillator, refractive index, dielectric

Table 2 Molecular polarizabilities given in Bohr^3 obtained by experiment, MP2, *kallisto*, *AlphaML*, and *Meanpol* and statistical measures - root mean squared error (RMSE), mean absoulte error (MAE), and the coefficient of determination (r^2).

Name	Formula	Exp.	MP2	*kallisto*	*AlphaML*	*Meanpol*
1-3-butadiene	C_4H_6	54.64	53.73	50.29	51.72	47.64
1-butene	C_4H_8	52.88	51.56	51.27	51.69	49.94
2-methyl-1-propene	C_4H_8	53.13	51.40	51.32	51.72	49.94
acetaldehyde	C_2H_4O	30.25	29.81	30.23	31.15	28.82
acetone	C_3H_6O	42.30	41.37	42.06	43.13	41.30
adamantane	$C_{10}H_{16}$	107.50	105.50	108.62	106.03	120.26
benzene	C_6H_6	67.79	67.99	68.08	65.22	68.02
C_2H_2	C_2H_2	22.96	22.29	23.02	19.15	22.67
C_2H_4	C_2H_4	27.72	26.91	27.63	25.89	24.97
C_2H_6	C_2H_6	29.69	28.23	28.62	29.22	27.26
CH_3Cl	CH_3Cl	29.98	29.29	29.22	19.24	28.21
CH_3CN	C2H3N	29.52	28.48	29.65	29.68	30.77
CH_3NH_2	CH_5N	26.50	25.53	25.78	25.08	22.88
CH_3OH	CH_4O	21.94	21.01	21.49	21.26	18.63
CH_3SH	CH_4S	35.00	36.48	36.36	30.76	34.96
CH_4	CH_4	17.24	16.50	16.86	16.71	14.78
CO_2	CO_2	17.50	17.55	19.04	23.06	17.88
CS_2	CS_2	55.30	56.53	48.90	*(−98.55)*	50.55
cyclopropane	C_3H_6	37.32	36.00	35.38	37.08	37.45
dimethylamine	C_2H_7N	38.70	37.74	37.71	38.05	35.36
dimethylether	C_2H_6O	34.54	33.20	33.56	34.22	31.11
E-2-butene	C_4H_8	53.13	51.75	51.28	52.28	49.94
ethanol	C_2H_6O	34.43	33.00	33.22	34.55	31.11
ethoxyethane	$C_4H_{10}O$	59.50	57.80	57.15	60.73	56.08
H_2CO	CH_2O	19.32	17.54	18.46	17.14	16.33
H_2O	H_2O	9.64	9.69	9.44	2.95	6.14
H_2S	H_2S	24.68	24.50	24.52	12.28	22.47
HCN	CHN	16.75	16.30	17.89	15.37	18.29
methyl-propyl-ether	$C_4H_{10}O$	59.20	57.43	57.13	59.74	56.08
N_2O	N_2O	19.70	19.42	18.66	34.36	17.75
N_2O_4	N_2O_4	43.83	41.31	34.07	60.46	29.29
n-butane	C_4H_{10}	54.10	52.24	52.23	54.02	52.23
NCCN	C_2N_2	32.20	31.14	31.13	30.65	34.28
neopentane	C_5H_{12}	66.23	64.16	64.20	65.16	64.72
NH_3	H_3N	14.56	14.14	14.02	14.15	10.39
n-heptane	C_7H_{16}	90.00	89.01	87.64	92.06	89.69
n-hexane	C_6H_{14}	78.00	76.67	75.84	79.28	16.06
n-octane	C_8H_{18}	102.00	101.40	99.44	104.85	102.17
n-pentane	C_5H_{12}	66.10	64.39	64.04	66.57	64.72
O_3	O_3	19.18	15.94	15.55	*(1572.84)*	11.54
OCS	COS	33.72	34.72	33.94	44.22	34.21
oxirane	C_2H_4O	29.19	28.28	28.60	30.15	28.82
propadiene	C_3H_4	40.48	39.54	36.64	39.24	35.16
propane	C_3H_8	42.12	40.17	40.43	41.46	39.75
propene	C_3H_6	40.79	39.22	39.46	38.90	37.45
propyne	C_3H_4	37.47	35.11	34.84	32.58	35.16
SO_2	O_2S	25.61	25.59	30.13	32.89	27.87
SO_3	O_3S	29.00	28.60	35.31	32.05	31.72
trimethylamine	C_3H_9N	49.90	50.48	49.66	51.01	47.85
		RMSE	0.90	2.35	4.95*(223.54)*	9.28
		MAE	1.16	1.76	2.98*(37.70)*	4.09
		r^2	0.99	0.99	0.95	0.80

Table 3 Timings given in seconds for the calculation of small- to medium-sized protein structures. PDB codes are given for each entry. All calculations were performed on a single Intel(R) Xeon(R) Gold 6140 CPU@2.30GHz. 1L2Y[54]: Trp-Cage miniprotein; 1EMA[55]: Green fluorescent protein; 1CC1[57]: Active form of the Ni-Fe-Se hydrogenase; 1GPE[58]: Gluose oxidase; 6LZ3[59]: Cryptochrome in active conformation; 7AD1[60]: Prefusion stabilized SARS-CoV-2 spike protein.

PDB	Number of atoms	t_{CPU} / seconds
1L2Y	302	2.4
1EMA	3784	212.9
1GZX	9686	1238.7
1CC1	12689	2110.9
1GPE	20561	5582.2
6LZ3	31396	12765.6
7AD1	42539	23522.6

Table 1 Consequence of environment- and charge effects on the absolute vdw-radius size. We calculate vdw radii for every coordination number at three different atomic partial charges. All values are given in Ångström.

	CN								
	0			1			2		
	q								
	0.00	0.50	1.00	0.00	0.50	1.00	0.00	0.50	1.00
C	1.90	1.85	1.80	1.73	1.69	1.64	1.82	1.77	1.73
N	1.79	1.74	1.69	1.76	1.71	1.66	1.77	1.72	1.67
O	1.71	1.65	1.60	1.70	1.65	1.60	1.70	1.65	1.60
P	2.23	2.20	2.18	2.22	2.20	2.18	2.23	2.19	2.17
Ir	2.40	2.38	2.36	2.30	2.28	2.25	2.28	2.26	2.24

permittivity, or electron-molecule scattering. MP2 molecular polarizabilities have been extracted from Ref. 38. *Meanpol* molecular polarizabilities are obtained by adding up averaged atomic polarizabilities using the chemical formula of the molecule as exemplified below for ethane

$$\alpha_{mol}^{C_2H_6} = 2 \cdot \alpha_C + 6 \cdot \alpha_H. \quad (16)$$

We applied averaged polarizabilities to calculate *Meanpol* molecular polarizabilities (Carbon: 10.19, Hydrogen: 1.15, Oxygen: 3.85, Nitrogen: 6.95, Sulfur: 20.18, and Chlorine: 14.58 all given in Bohr^3).[41] *AlphaML* molecular polarizabilities have been obtained by their webinterface[76] and *kallisto* molecular polarizabilities by its command-line interface.[74]

5.4 Timings for the calculation of van der Waals radii

All structures have been extracted from the protein data bank[77] and in all cases hydrogen atoms were added using the Maestro suite.

5.5 Retention times: Data Acquisition and Experimental Setup

Tentative: For this work, data gathered by the separation sciences laboratory at AstraZeneca used for purifying novel compounds was used. The lab uses different instruments, analytical columns and solvents for purification, where the scientist analyzes mass and UV chromatograms, and decides on the most appropriate experimental setup to use for purification.

The preparative samples, submitted dissolved in dimethyl sulfoxide (DMSO), were diluted 20-200 μL DMSO, and injected on a Waters supercritical fluid chromatography system (UPC2 description) coupled to a Waters 3100 mass detector. A Waters diode array detector (DAD) was used in the range of 200-500 nm. The mass detector was set to detect in the m/z range 100-1200 kDa. The electrospray source conditions were as follows: Capillary voltage 3 kV, cone voltage 30 kV, source temperature 150 C, with a desolvation gas flow of 650 L/h. The stationary phase was a Waters Viridis BEH Column, 130Å, 3.5 μm, 3 mm X 100 mm, 1/pk. A mobile phase, 5-50% gradient, of methanol with 20 mM ammonia in supercritical CO2, with a 4.1 minutes total run time was used. The flow was 2.5 mL/min, the back pressure was set to 1740 psi and the temperature was 40°C. Retention times were based on the peak time of the positive ESI mass trace of the protonated target compound. A summary of the experimen-

tal setups for preparatory step of the different experiments and the number of data points for each of the Liquid Chromatogram Mass Spectrometry (LC/MS) and Superfluid Critical Mass Spectrometry (SFC/MS) used for this analysis is presented in Table 4 and 5. It should be noted that a compound may have more than one or no datapoints for an experiment type. These experiments were done on a total 14627 unique compounds. Of these 3031 are publicly available, and separated from the original dataset to be used as validation.

Analytical Method	# T	# T U	# V	# V U
LC-1	15818	13170	876	869
LC-2	13172	13171	851	833
SFC-1	10939	9395	487	480
SFC-2	6020	5333	227	227
SFC-3	10848	9297	495	487
SFC-4	11170	9571	502	495

Table 4 Description of number of data points (#) in the dataset (T=training, V=validation, U=unique).

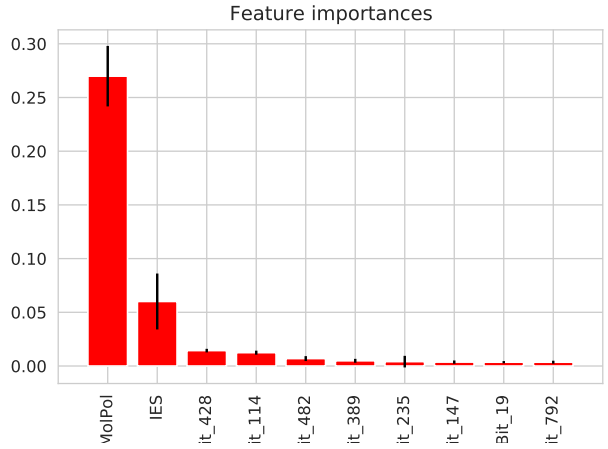
The data was processed into a machine learning-ready format using the ProteoWizard MS Converter for Linux[†], where the *.raw* data files were converted to an *.mzXML* format. The data files were further processed and the SMILES, analytical method used in the experiment and retention time were inserted into a Pandas dataframe to be used for analysis and modelling purposes.

We analyze feature importance in two ways, by using the mean decrease in impurity with SCIKIT LEARN,[71] and by the *SHAP* importance metrics cite that use a game theoretic approach with Shapley values to explain the outputs of the model. The results are shown in Fig. 8 and 7, and can be seen that the *kallisto* descriptors rank highly in terms of feature importance. Further analysis showed that the *kallisto* descriptors consistently ranked in top 3 important features. This result, in conjunction with the other results, show that the *kallisto* features are indeed describing aspects of the compounds that are not properly captured by the fingerprints, indeed they capture 3D features in a meaningful way, and thus are enhancing the modelling performance.

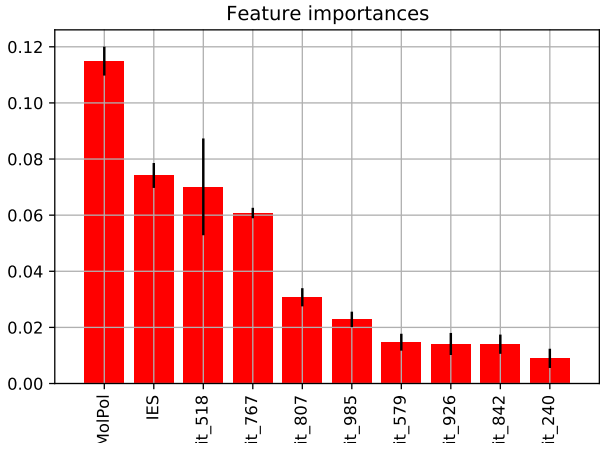
† https://hub.docker.com/r/chambm/pwiz-skyline-i-agree-to-the-vendor-licenses

Analytical Method	Stationary Phase	Mobile Phase
LC-1	Waters Acquity BEH C18 1.7μ 2.1x50mm	Gradient 5-95% ACN, in 0.1M NH4HCO3, pH9
LC-2	Waters Acquity HSS C18 1.8μ 2.1x50mm	Gradient 5-95% ACN, in 0.1M HCO2H, pH3
SFC-1	Waters Acquity BEH 3.5μ 3x100mm	Gradient 5-50% MeOH, in 20mM MeOH/NH3
SFC-2	Waters Acquity BEH 3.5μ 3x100mm	Gradient 5-50% MeOH, in 20mM MeOH/H2O/NH3
SFC-3	Phenomenex Luna Hilic 3.5μ 3x100mm	Gradient 5-50% MeOH, in 20mM MeOH/NH3
SFC-4	Waters Acquity BEH-2EP 3.5μ 3x100mm	Gradient 5-50% MeOH, in 20mM MeOH/NH3

Table 5 Experimental setup for the different analytical methods.



(a) Molecular Polarizability ranks as the top feature, while IES ranks as second.



(b) Molecular Polarizability ranks as the top feature, while IES ranks as second.

Fig. 8 Feature importance for top 10 features of the random forest applied to the (a) AstraZeneca SFC-2 dataset and (b) *METLIN* dataset