5 Supporting 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").

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

5.3 Molecular polarizabilities

This benchmark set is a subset of the *MOLPOL135* benchmark set[40], whose experimental molecular polarizabilities have been determined by either dipole oscillator, refractive index, dielectric permittivity, or electron-molecule scattering. MP2 molecular polarizabilities have been extracted from Ref. 42. *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).[45] *AlphaML* molecular polarizabilities have been obtained by their webinterface[82] and *kallisto* molecular polarizabilities by its command-line interface.[78]

5.4 Timings for the calculation of van der Waals radii

All structures have been extracted from the protein data bank[83] 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 experimental 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.

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 use feature reduction techniques to remove highly correlated (with correlation higher than 0.9) and low variance features (features with variance lower than 0.05). Then we further choose

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

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

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

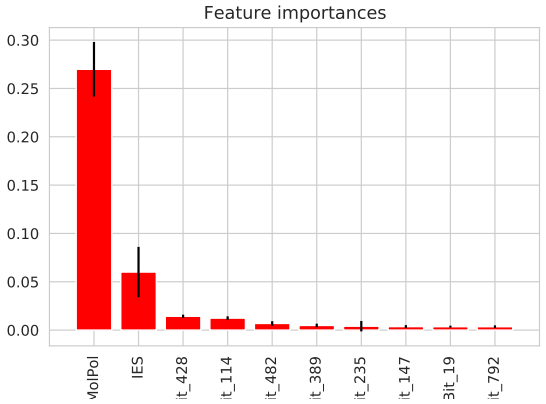
Table 4 Description of number of data points (#) in the dataset (T=training, V=validation, U=unique).

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[59]: Trp-Cage miniprotein; 1EMA[60]: Green fluorescent protein; 1CC1[62]: Active form of the Ni-Fe-Se hydrogenase; 1GPE[63]: Gluose oxidase; 6LZ3[64]: Cryptochrome in active conformation; 7AD1[65]: Prefusion stabilized SARS-CoV-2 spike protein.

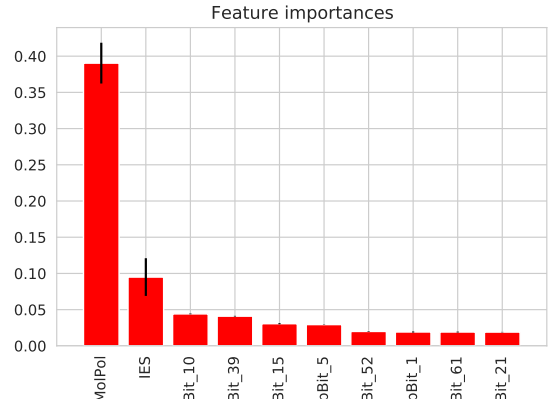
PDB	Number of atoms	t_{CPU} / seconds
1L2Y	302	2.4
1EMA	3784	212.9
1GZX	9686	1238.7
1CC1	12689	2110.9
1GPE	20561	5582.2
6LZ3	31396	12765.6
7AD1	42539	23522.6

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

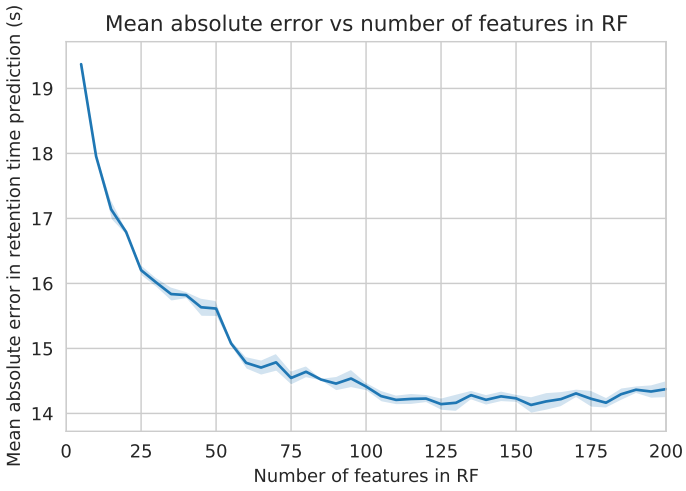


Fig. 8 Mean squared error between model prediction and retention time vs. the number of top features selected for SFC-4 shaded areas shows 2 standard deviations. Other analytical method plots are available in the papers GitHub link.[80]

the top K features and plot vs. prediction error. This is shown in Fig. 8.

This analysis is done for each experiment, and the lowest number of features where the model performance converges to is selected. From this figure for SFC-4, we decide to select the top 65 features for our further analysis as we achieve most of the modelling performance with acceptable number of parameters. A new model is trained with these features and retention time is predicted for the both Metlin and AstraZeneca datasets and analyzed.

We analyze feature importance in two ways, by using the mean decrease in impurity with SCIKIT LEARN,[75] 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. 7 and 6, 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 5 important features for the various experiments. 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.