Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information for "Accurately Predicting Solvation Free Energy in Aqueous and Organic Solvents beyond 298K by Combining Deep Learning and The 1D Reference Interaction Site Model"

Daniel J. Fowles, Rose G. McHardy, Abdullah Ahmad, David S. Palmer*
* Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, U.K.; E-mail: david.palmer@strath.ac.uk

1 Solvation Free Energy Densities

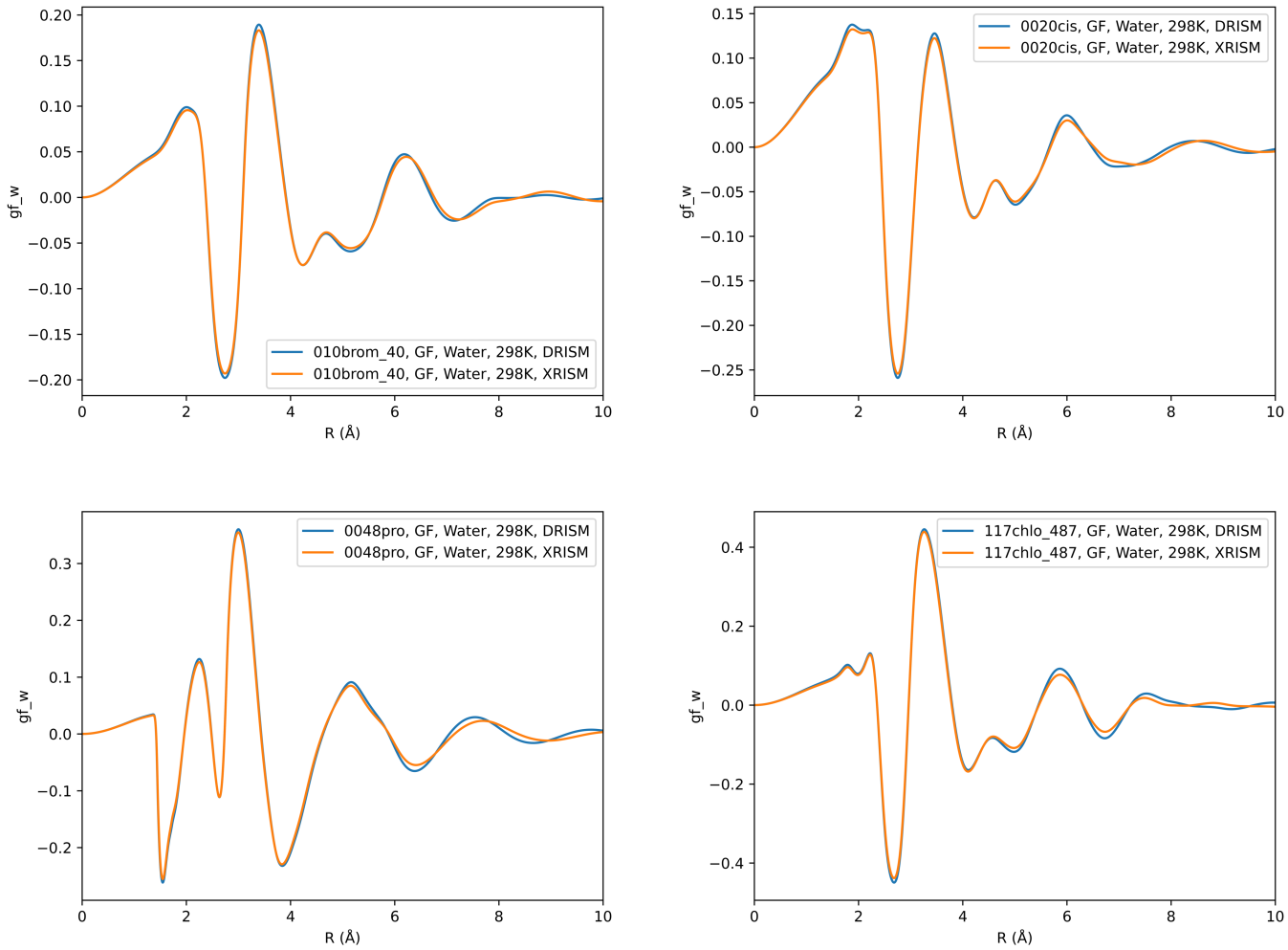

Figure 1 Comparison of the effects of DRISM and XRISM on SFED magnitude and shape. Descriptors were generated using the GF functional and GAFF parameters.

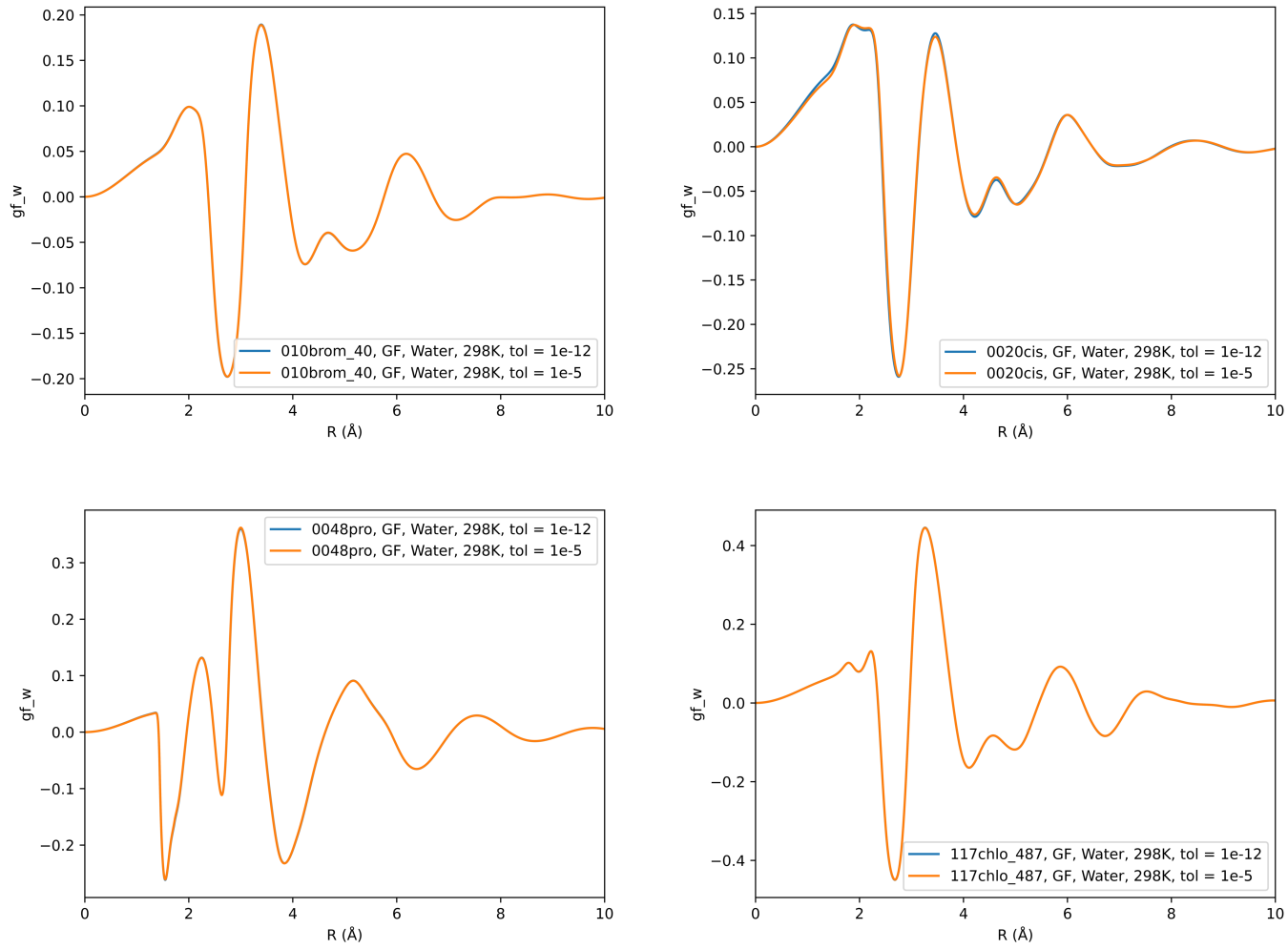

Figure 2 Comparison of the effects of convergence tolerance on SFED magnitude and shape. Descriptors were generated using the GF functional and GAFF parameters.

2 Convolutional Neural Network

Single Optimisation Method					GAFF			
Solvent	Temp.	Temp. Descr.	Functional	Datapoints	R^2	RMSE	Bias	SDEP
Multi-solvent	273-373K	No	KH	3629	0.93(0.04)	0.98(0.26)	0.10(0.16)	0.97(0.25)
			HNC		0.92(0.04)	1.07(0.25)	0.11(0.14)	1.05(0.25)
			GF		0.94(0.04)	0.95(0.24)	0.09(0.15)	0.94(0.24)
Separate Optimisation Methods					GAFF			
Solvent	Temp.	Temp. Descr.	Functional	Datapoints	R^2	RMSE	Bias	SDEP
Multi-solvent	273-373K	No	KH	3629	0.93(0.03)	1.01(0.21)	0.10(0.14)	1.00(0.20)
			HNC		0.93(0.05)	1.00(0.29)	0.09(0.14)	0.98(0.29)
			GF		0.94(0.04)	0.97(0.28)	0.04(0.16)	0.96(0.27)

Table 1 Solvation free energy predictions taken from CNN models using KH, HNC and GF calculated descriptors with GAFF parameters. Two separate optimisation methods are shown. The single method performed a conformational search using a weighted rotor search with the GAFF forcefield on all solute geometries. The lowest energy conformer for each solute was used in 1D-RISM calculations to generate SFED datasets. The separate optimisation method used a different procedure for the MSD and Chamberlin *et al.* datasets. The full breakdown of the different methods can be found in Section 3.1 of the manuscript.

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.87	0.66	0.03	0.59	0.90	0.64	0.03	0.59
Carbon Tet.	298K	No	109	0.88	0.40	0.00	0.34	0.88	0.42	-0.00	0.34
Water	298K	No	521	0.98	0.59	0.00	0.58	0.98	0.52	0.01	0.51
	273-373K	No	3053	0.95	0.51	-0.03	0.49	0.94	0.57	-0.02	0.55
	273-373K	Yes	3053	0.97	0.44	-0.01	0.42	0.39	0.61	-0.01	0.59
Multi-solvent	298K	No	709	0.96	0.70	0.02	0.67	0.97	0.63	-0.01	0.61
	273-373K	No	3241	0.93	0.57	-0.03	0.55	0.94	0.59	-0.03	0.58
	273-373K	Yes	3241	0.95	0.55	-0.02	0.54	0.92	0.61	-0.02	0.60
	273-373K	No	3629	0.53	0.97	-0.04	0.94	0.78	0.98	-0.05	0.96
	273-373K	Yes	3629	0.94	0.75	0.05	0.73	0.46	0.98	-0.04	0.96
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.89	0.67	0.04	0.60	0.89	0.66	-0.01	0.60
Carbon Tet.	298K	No	109	0.92	0.38	-0.01	0.32	0.84	0.39	0.02	0.31
Water	298K	No	521	0.98	0.58	-0.03	0.56	0.98	0.52	0.00	0.51
	273-373K	No	3053	0.90	0.52	-0.01	0.51	0.90	0.58	-0.04	0.57
	273-373K	Yes	3053	0.52	0.62	-0.03	0.61	0.89	0.56	-0.00	0.55
Multi-solvent	298K	No	709	0.96	0.71	0.03	0.69	0.97	0.64	0.04	0.62
	273-373K	No	3241	0.95	0.54	-0.02	0.53	0.94	0.57	-0.03	0.56
	273-373K	Yes	3241	0.68	0.61	-0.01	0.60	0.94	0.57	-0.03	0.55
	273-373K	No	3629	0.87	0.81	-0.09	0.78	0.89	0.87	-0.07	0.85
	273-373K	Yes	3629	0.90	0.77	0.04	0.75	0.91	0.84	-0.05	0.82
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.90	0.63	0.04	0.58	0.87	0.71	0.01	0.65
Carbon Tet.	298K	No	109	0.84	0.44	0.10	0.36	0.88	0.44	-0.01	0.37
Water	298K	No	521	0.98	0.59	-0.07	0.58	0.98	0.54	-0.01	0.53
	273-373K	No	3053	0.95	0.51	-0.03	0.49	0.95	0.55	-0.02	0.53
	273-373K	Yes	3053	0.96	0.47	-0.01	0.45	0.96	0.46	-0.00	0.45
Multi-solvent	298K	No	709	0.96	0.70	0.00	0.68	0.97	0.62	-0.02	0.60
	273-373K	No	3241	0.79	0.64	-0.05	0.63	0.93	0.60	-0.04	0.58
	273-373K	Yes	3241	0.77	0.57	-0.01	0.56	0.96	0.53	-0.02	0.51
	273-373K	No	3629	0.83	0.82	-0.10	0.79	0.92	0.81	-0.03	0.79
	273-373K	Yes	3629	0.94	0.72	-0.08	0.70	0.93	0.80	-0.05	0.78

Table 2 Solvation free energy predictions taken from CNN models using KH, HNC and GF calculated descriptors, OPLS and GAFF parameters and temperature descriptors. Results shown are the average five-fold cross validation test set statistics within each resample, averaged over the number of resamples run for each model. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and aqueous solvent was modelled with MSPC/E.

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Carbon Tet.	298K	No	79	0.93(0.06)	0.44(0.15)	0.06(0.13)	0.42(0.14)	0.91(0.17)	0.45(0.21)	0.03(0.25)	0.40(0.13)
Chloroform	298K	No	109	0.92(0.05)	0.74(0.21)	0.00(0.19)	0.72(0.20)	0.90(0.08)	0.76(0.30)	-0.01(0.15)	0.75(0.30)
Water	298K	No	521	0.95(0.03)	0.91(0.31)	0.04(0.22)	0.89(0.30)	0.95(0.05)	0.89(0.38)	0.01(0.27)	0.85(0.36)
	273-373K	No	3053	0.93(0.05)	0.66(0.24)	-0.01(0.11)	0.65(0.24)	0.91(0.08)	0.75(0.27)	0.02(0.13)	0.74(0.27)
	273-373K	Yes	3053	0.95(0.05)	0.55(0.22)	0.01(0.09)	0.54(0.22)	0.93(0.07)	0.65(0.26)	0.01(0.12)	0.64(0.26)
Multi-solvent	298K	No	709	0.95(0.03)	0.83(0.22)	-0.01(0.15)	0.82(0.22)	0.95(0.05)	0.82(0.32)	-0.02(0.18)	0.80(0.31)
	273-373K	No	3241	0.88(0.10)	0.94(0.26)	0.08(0.23)	0.92(0.25)	0.89(0.06)	0.91(0.20)	0.03(0.24)	0.88(0.18)
	273-373K	Yes	3241	0.88(0.10)	0.97(0.28)	0.06(0.20)	0.95(0.27)	0.87(0.06)	1.01(0.23)	0.03(0.27)	0.98(0.21)
	273-373K	No	3629	0.93(0.03)	1.01(0.21)	0.10(0.14)	1.00(0.20)	0.91(0.06)	1.12(0.36)	0.00(0.24)	1.10(0.35)
	273-373K	Yes	3629	0.93(0.03)	1.03(0.21)	0.21(0.12)	1.00(0.21)	0.92(0.05)	1.06(0.31)	-0.02(0.15)	1.05(0.31)
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Carbon Tet.	298K	No	79	0.93(0.06)	0.45(0.11)	0.03(0.14)	0.43(0.11)	0.91(0.12)	0.47(0.22)	-0.01(0.17)	0.44(0.21)
Chloroform	298K	No	109	0.89(0.08)	0.78(0.27)	-0.00(0.16)	0.76(0.27)	0.90(0.08)	0.74(0.27)	0.03(0.17)	0.72(0.27)
Water	298K	No	521	0.94(0.06)	0.94(0.39)	-0.14(0.25)	0.91(0.36)	0.95(0.06)	0.88(0.39)	-0.01(0.24)	0.85(0.38)
	273-373K	No	3053	0.94(0.04)	0.64(0.21)	-0.01(0.09)	0.63(0.21)	0.90(0.11)	0.75(0.33)	0.03(0.11)	0.74(0.33)
	273-373K	Yes	3053	0.94(0.05)	0.57(0.23)	0.02(0.10)	0.56(0.23)	0.91(0.09)	0.71(0.33)	0.03(0.13)	0.70(0.32)
Multi-solvent	298K	No	709	0.95(0.03)	0.84(0.22)	0.06(0.15)	0.83(0.22)	0.95(0.03)	0.80(0.26)	0.02(0.14)	0.79(0.26)
	273-373K	No	3241	0.89(0.04)	0.93(0.17)	0.08(0.18)	0.90(0.17)	0.77(0.62)	1.15(0.86)	0.05(0.29)	1.11(0.86)
	273-373K	Yes	3241	0.88(0.05)	0.98(0.18)	0.11(0.21)	0.95(0.17)	0.86(0.08)	1.04(0.25)	0.07(0.24)	1.01(0.24)
	273-373K	No	3629	0.93(0.05)	1.00(0.29)	0.09(0.14)	0.98(0.29)	0.93(0.04)	0.98(0.37)	0.00(0.14)	0.97(0.27)
	273-373K	Yes	3629	0.93(0.04)	0.97(0.25)	0.12(0.11)	0.95(0.24)	0.92(0.05)	1.08(0.31)	0.03(0.17)	1.07(0.31)
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Carbon Tet.	298K	No	79	0.91(0.09)	0.47(0.20)	0.05(0.20)	0.44(0.17)	0.90(0.06)	0.51(0.17)	-0.01(0.22)	0.47(0.14)
Chloroform	298K	No	109	0.89(0.13)	0.80(0.36)	0.01(0.21)	0.77(0.35)	0.91(0.07)	0.73(0.21)	0.00(0.16)	0.72(0.21)
Water	298K	No	521	0.94(0.04)	0.96(0.29)	-0.06(0.32)	0.91(0.27)	0.95(0.05)	0.88(0.35)	0.02(0.21)	0.86(0.35)
	273-373K	No	3053	0.93(0.08)	0.65(0.26)	0.02(0.14)	0.64(0.25)	0.92(0.06)	0.68(0.23)	0.03(0.09)	0.67(0.23)
	273-373K	Yes	3053	0.94(0.05)	0.62(0.21)	0.01(0.09)	0.61(0.21)	0.94(0.05)	0.59(0.20)	0.01(0.07)	0.58(0.20)
Multi-solvent	298K	No	709	0.95(0.03)	0.87(0.22)	0.05(0.18)	0.85(0.22)	0.95(0.05)	0.77(0.31)	0.01(0.14)	0.76(0.30)
	273-373K	No	3241	0.90(0.05)	0.91(0.21)	0.03(0.20)	0.89(0.20)	0.36(3.42)	1.27(1.75)	0.07(0.39)	1.24(1.73)
	273-373K	Yes	3241	0.88(0.06)	0.97(0.22)	0.10(0.23)	0.93(0.21)	0.74(1.02)	1.10(1.10)	-0.02(0.26)	1.07(1.09)
	273-373K	No	3629	0.94(0.04)	0.97(0.28)	0.04(0.16)	0.96(0.27)	0.94(0.04)	0.93(0.27)	0.05(0.21)	0.91(0.25)
	273-373K	Yes	3629	0.94(0.04)	0.95(0.24)	0.04(0.12)	0.95(0.23)	0.93(0.05)	1.04(0.28)	0.03(0.20)	1.03(0.27)

Table 3 Solvation free energy predictions for all solvents using convolutional neural network models trained on KH, HNC and GF calculated descriptors. Predictions are separated by temperature, inclusion of a temperature descriptor, solute forcefield parameters and number of datapoints. Results for each model are taken from test set predictions. Units are in kcal/mol. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and water solvent was modelled with MSPC/E. The standard deviation (SD) for each model statistic over 50 resamples is also shown.

3 Partial Least Squares

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.64	1.39	-0.06	1.33	0.52	1.55	-0.04	1.48
Carbon Tet.	298K	No	109	0.81	0.67	-0.01	0.63	0.74	0.76	-0.01	0.72
Water	298K	No	521	0.77	1.94	-0.04	1.93	0.74	2.10	-0.00	2.08
	273-373K	No	3053	0.55	1.73	-0.05	1.68	0.67	1.50	-0.00	1.47
	273-373K	Yes	3053	0.61	1.62	-0.04	1.58	0.70	1.41	-0.02	1.37
Multi-solvent	298K	No	709	0.64	2.31	-0.01	2.29	0.67	2.21	-0.01	2.20
	273-373K	No	3241	0.45	1.95	0.16	1.86	0.55	1.80	0.17	1.73
	273-373K	Yes	3241	0.52	1.85	0.14	1.78	0.55	1.74	0.16	1.70
	273-373K	No	3629	0.48	2.29	0.28	2.20	0.53	2.17	0.42	2.05
	273-373K	Yes	3629	0.52	2.21	0.21	2.15	0.55	2.13	0.37	2.03
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.63	1.38	-0.05	1.32	0.59	1.47	-0.02	1.40
Carbon Tet.	298K	No	109	0.80	0.69	-0.02	0.65	0.78	0.70	0.01	0.66
Water	298K	No	521	0.76	1.96	-0.04	1.94	0.73	2.12	-0.01	2.10
	273-373K	No	3053	0.54	1.74	-0.06	1.70	0.67	1.51	-0.02	1.47
	273-373K	Yes	3053	0.60	1.65	-0.06	1.61	0.70	1.40	-0.03	1.37
Multi-solvent	298K	No	709	0.63	2.32	-0.02	2.30	0.67	2.22	-0.01	2.21
	273-373K	No	3241	0.44	2.01	0.16	1.92	0.50	1.86	0.18	1.79
	273-373K	Yes	3241	0.48	1.90	0.16	1.82	0.54	1.77	0.16	1.70
	273-373K	No	3629	0.46	2.33	0.30	2.23	0.52	2.21	0.42	2.08
	273-373K	Yes	3629	0.52	2.21	0.24	2.14	0.53	2.18	0.39	2.07
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.67	1.31	-0.05	1.26	0.66	1.30	-0.05	1.24
Carbon Tet.	298K	No	109	0.81	0.66	-0.00	0.62	0.82	0.63	0.01	0.60
Water	298K	No	521	0.77	1.96	-0.05	1.94	0.73	2.14	-0.02	2.12
	273-373K	No	3053	0.59	1.65	-0.04	1.60	0.68	1.46	-0.02	1.43
	273-373K	Yes	3053	0.65	1.56	-0.03	1.52	0.73	1.34	-0.01	1.31
Multi-solvent	298K	No	709	0.67	2.18	-0.02	2.17	0.65	2.28	-0.03	2.26
	273-373K	No	3241	0.50	1.85	0.16	1.77	0.59	1.70	0.15	1.64
	273-373K	Yes	3241	0.55	1.77	0.13	1.71	0.60	1.63	0.14	1.58
	273-373K	No	3629	0.53	2.18	0.19	2.12	0.56	2.11	0.36	2.01
	273-373K	Yes	3629	0.54	2.14	0.17	2.09	0.58	2.07	0.34	1.98

Table 4 Solvation free energy predictions taken from PLS models using KH, HNC and GF calculated descriptors, OPLS and GAFF parameters and temperature descriptors. Results shown are the average five-fold cross validation test set statistics within each resample, averaged over the number of resamples run for each model. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and aqueous solvent was modelled with MSPC/E.

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.71(0.12)	1.35(0.29)	0.04(0.32)	1.31(0.27)	0.63(0.17)	1.55(0.34)	-0.09(0.36)	1.50(0.34)
Carbon Tet.	298K	No	109	0.78(0.11)	0.76(0.19)	0.01(0.21)	0.73(0.18)	0.74(0.11)	0.85(0.16)	-0.05(0.21)	0.82(0.15)
Water	298K	No	521	0.77(0.06)	1.96(0.27)	-0.04(0.25)	1.94(0.26)	0.73(0.05)	2.17(0.19)	-0.04(0.19)	2.17(0.19)
	273-373K	No	3053	0.62(0.08)	1.59(0.14)	0.08(0.21)	1.58(0.14)	0.66(0.08)	1.51(0.15)	0.07(0.19)	1.50(0.15)
	273-373K	Yes	3053	0.65(0.07)	1.55(0.15)	0.10(0.18)	1.54(0.15)	0.69(0.06)	1.43(0.16)	0.06(0.22)	1.41(0.16)
Multi-solvent	298K	No	709	0.64(0.08)	2.35(0.26)	-0.03(0.23)	2.34(0.26)	0.65(0.07)	2.27(0.25)	-0.07(0.25)	2.26(0.25)
	273-373K	No	3241	-1.27(0.54)	4.33(0.30)	3.16(0.45)	2.94(0.22)	-0.90(0.53)	3.91(0.33)	2.84(0.43)	2.66(0.19)
	273-373K	Yes	3241	-1.25(0.61)	4.15(0.27)	3.01(0.42)	2.84(0.18)	-0.96(0.53)	4.04(0.31)	2.88(0.48)	2.80(0.18)
	273-373K	No	3629	0.27(0.10)	3.31(0.21)	1.20(0.25)	3.08(0.28)	0.30(0.06)	3.27(0.16)	1.55(0.17)	2.87(0.17)
	273-373K	Yes	3629	0.31(0.11)	3.23(0.23)	0.99(0.21)	3.07(0.27)	0.29(0.07)	3.27(0.17)	1.49(0.21)	2.91(0.15)
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.67(0.13)	1.42(0.25)	0.00(0.39)	1.37(0.25)	0.62(0.15)	1.60(0.33)	-0.06(0.36)	1.56(0.32)
Carbon Tet.	298K	No	109	0.78(0.12)	0.76(0.17)	-0.03(0.21)	0.74(0.16)	0.77(0.10)	0.81(0.16)	0.05(0.22)	0.78(0.16)
Water	298K	No	521	0.78(0.06)	1.96(0.26)	-0.07(0.20)	1.95(0.26)	0.74(0.06)	2.12(0.20)	-0.00(0.25)	2.11(0.20)
	273-373K	No	3053	0.61(0.08)	1.60(0.16)	0.11(0.19)	1.59(0.17)	0.64(0.07)	1.59(0.16)	0.05(0.19)	1.55(0.16)
	273-373K	Yes	3053	0.66(0.06)	1.53(0.13)	0.11(0.22)	1.52(0.13)	0.68(0.09)	1.47(0.19)	-0.01(0.21)	1.45(0.18)
Multi-solvent	298K	No	709	0.64(0.08)	2.33(0.23)	-0.04(0.20)	2.32(0.23)	0.67(0.06)	2.22(0.22)	-0.03(0.22)	2.21(0.21)
	273-373K	No	3241	-1.38(0.72)	4.52(0.29)	3.32(0.46)	3.05(0.25)	-1.32(0.61)	4.19(0.30)	3.05(0.45)	2.84(0.24)
	273-373K	Yes	3241	-1.17(0.54)	4.22(0.30)	3.03(0.42)	2.92(0.22)	-1.36(0.59)	4.36(0.29)	3.21(0.43)	2.93(0.20)
	273-373K	No	3629	0.28(0.12)	3.37(0.25)	1.31(0.25)	3.09(0.30)	0.28(0.07)	3.30(0.16)	1.60(0.20)	2.88(0.14)
	273-373K	Yes	3629	0.25(0.11)	3.36(0.22)	1.02(0.17)	3.19(0.25)	0.28(0.07)	3.31(0.17)	1.46(0.23)	2.96(0.17)
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.70(0.14)	1.40(0.36)	-0.03(0.33)	1.36(0.35)	0.69(0.16)	1.44(0.30)	0.00(0.34)	1.40(0.29)
Carbon Tet.	298K	No	109	0.76(0.14)	0.82(0.24)	-0.01(0.24)	0.79(0.24)	0.81(0.09)	0.72(0.17)	-0.05(0.22)	0.68(0.16)
Water	298K	No	521	0.78(0.05)	1.92(0.24)	-0.02(0.17)	1.91(0.24)	0.75(0.05)	2.07(0.19)	0.02(0.22)	2.06(0.19)
	273-373K	No	3053	0.64(0.08)	1.59(0.16)	0.05(0.18)	1.58(0.16)	0.66(0.09)	1.49(0.18)	0.01(0.19)	1.47(0.17)
	273-373K	Yes	3053	0.65(0.07)	1.49(0.14)	0.01(0.20)	1.48(0.14)	0.70(0.06)	1.42(0.17)	0.03(0.17)	1.41(0.17)
Multi-solvent	298K	No	709	0.67(0.08)	2.22(0.25)	0.05(0.23)	2.21(0.25)	0.68(0.05)	2.21(0.19)	-0.03(0.19)	2.20(0.19)
	273-373K	No	3241	-1.21(0.55)	4.26(0.28)	3.11(0.41)	2.89(0.22)	-0.51(0.38)	3.55(0.25)	2.43(0.34)	2.57(0.19)
	273-373K	Yes	3241	-1.05(0.46)	4.17(0.31)	3.03(0.42)	2.84(0.19)	-0.69(0.48)	3.77(0.30)	2.66(0.39)	2.65(0.17)
	273-373K	No	3629	0.34(0.12)	3.18(0.30)	1.01(0.29)	2.99(0.36)	0.32(0.06)	3.25(0.16)	1.42(0.17)	2.91(0.17)
	273-373K	Yes	3629	0.36(0.12)	3.14(0.26)	0.84(0.22)	3.02(0.27)	0.29(0.08)	3.30(0.18)	1.33(0.27)	3.00(0.19)

Table 5 Solvation free energy predictions for all solvents using partial least squares models trained on KH, HNC and GF calculated descriptors. Predictions are separated by temperature, inclusion of a temperature descriptor, solute forcefield parameters and number of datapoints. Results for each model are taken from test set predictions. Units are in kcal/mol. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and water solvent was modelled with MSPC/E. The standard deviation (SD) for each model statistic over 50 resamples is also shown.

4 Random Forest

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.65	1.38	0.00	1.32	0.58	1.54	0.00	1.48
Carbon Tet.	298K	No	109	0.67	0.88	0.03	0.83	0.77	0.73	0.03	0.69
Water	298K	No	521	0.85	1.50	0.02	1.48	0.80	1.88	-0.01	1.86
	273-373K	No	3053	0.70	1.45	-0.03	1.42	0.60	1.67	-0.05	1.63
	273-373K	Yes	3053	0.70	1.44	-0.04	1.41	0.61	1.63	-0.06	1.60
Multi-solvent	298K	No	709	0.83	1.61	0.00	1.60	0.76	1.89	-0.03	1.88
	273-373K	No	3241	0.69	1.51	-0.05	1.48	0.64	1.63	-0.08	1.60
	273-373K	Yes	3241	0.69	1.50	-0.07	1.47	0.64	1.62	-0.09	1.60
	273-373K	No	3629	0.76	1.52	-0.02	1.49	0.67	1.80	-0.11	1.75
	273-373K	Yes	3629	0.77	1.51	-0.01	1.48	0.68	1.79	-0.11	1.75
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.67	1.37	-0.00	1.32	0.56	1.55	-0.02	1.47
Carbon Tet.	298K	No	109	0.69	0.85	0.02	0.80	0.77	0.75	0.02	0.70
Water	298K	No	521	0.86	1.50	0.00	1.49	0.79	1.88	-0.01	1.86
	273-373K	No	3053	0.69	1.46	-0.02	1.43	0.60	1.68	-0.04	1.64
	273-373K	Yes	3053	0.71	1.42	-0.03	1.39	0.62	1.64	-0.04	1.61
Multi-solvent	298K	No	709	0.83	1.62	0.01	1.61	0.76	1.88	-0.02	1.87
	273-373K	No	3241	0.68	1.53	-0.07	1.50	0.63	1.66	-0.11	1.63
	273-373K	Yes	3241	0.69	1.50	-0.06	1.48	0.64	1.62	-0.09	1.60
	273-373K	No	3629	0.76	1.52	-0.03	1.48	0.68	1.78	-0.11	1.74
	273-373K	Yes	3629	0.77	1.50	-0.03	1.47	0.68	1.77	-0.13	1.72
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.57	1.55	-0.02	1.48	0.50	1.70	-0.03	1.63
Carbon Tet.	298K	No	109	0.50	1.09	0.01	1.03	0.70	0.85	0.02	0.79
Water	298K	No	521	0.87	1.50	0.01	1.49	0.80	1.83	-0.02	1.82
	273-373K	No	3053	0.71	1.42	-0.04	1.39	0.64	1.59	-0.06	1.56
	273-373K	Yes	3053	0.70	1.44	-0.04	1.41	0.63	1.63	-0.04	1.58
Multi-solvent	298K	No	709	0.82	1.64	0.01	1.63	0.77	1.86	-0.03	1.85
	273-373K	No	3241	0.69	1.51	-0.08	1.47	0.66	1.60	-0.08	1.57
	273-373K	Yes	3241	0.69	1.50	-0.05	1.46	0.66	1.57	-0.08	1.54
	273-373K	No	3629	0.76	1.53	-0.03	1.50	0.69	1.76	-0.11	1.71
	273-373K	Yes	3629	0.77	1.50	-0.03	1.47	0.70	1.73	-0.10	1.69

Table 6 Solvation free energy predictions taken from RF models using KH, HNC and GF calculated descriptors, OPLS and GAFF parameters and temperature descriptors. Results shown are the average five-fold cross validation test set statistics within each resample, averaged over the number of resamples run for each model. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and aqueous solvent was modelled with MSPC/E.

KH				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.66(0.14)	1.45(0.25)	0.02(0.35)	1.41(0.24)	0.63(0.13)	1.56(0.24)	-0.04(0.31)	1.53(0.24)
Carbon Tet.	298K	No	109	0.70(0.13)	0.91(0.17)	0.02(0.25)	0.88(0.16)	0.78(0.09)	0.75(0.14)	-0.02(0.23)	0.72(0.14)
Water	298K	No	521	0.86(0.03)	1.53(0.19)	-0.01(0.17)	1.53(0.18)	0.78(0.04)	1.90(0.20)	-0.00(0.26)	1.89(0.19)
	273-373K	No	3053	0.69(0.07)	1.45(0.20)	0.05(0.21)	1.44(0.20)	0.65(0.05)	1.56(0.15)	0.06(0.23)	1.54(0.15)
	273-373K	Yes	3053	0.70(0.07)	1.39(0.19)	0.04(0.17)	1.37(0.19)	0.65(0.08)	1.57(0.25)	0.03(0.21)	1.56(0.25)
Multi-solvent	298K	No	709	0.82(0.03)	1.65(0.14)	0.01(0.19)	1.64(0.14)	0.76(0.03)	1.94(0.18)	-0.01(0.21)	1.93(0.17)
	273-373K	No	3241	0.23(0.34)	2.50(0.45)	1.02(0.50)	2.25(0.32)	0.47(0.18)	2.12(0.31)	0.49(0.40)	2.03(0.26)
	273-373K	Yes	3241	0.31(0.23)	2.37(0.35)	0.94(0.39)	2.16(0.26)	0.44(0.17)	2.14(0.30)	0.49(0.38)	2.06(0.25)
	273-373K	No	3629	0.75(0.05)	1.93(0.22)	0.37(0.14)	1.89(0.21)	0.66(0.04)	2.30(0.16)	0.49(0.19)	2.24(0.14)
	273-373K	Yes	3629	0.76(0.04)	1.91(0.17)	0.39(0.15)	1.87(0.16)	0.67(0.04)	2.27(0.16)	0.45(0.17)	2.22(0.15)
HNC				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.66(0.10)	1.45(0.18)	-0.00(0.33)	1.41(0.17)	0.64(0.11)	1.53(0.28)	0.01(0.38)	1.49(0.27)
Carbon Tet.	298K	No	109	0.71(0.12)	0.91(0.16)	0.04(0.27)	0.86(0.16)	0.77(0.09)	0.78(0.13)	0.05(0.25)	0.74(0.12)
Water	298K	No	521	0.85(0.03)	1.62(0.14)	-0.04(0.17)	1.61(0.14)	0.80(0.03)	1.88(0.18)	-0.03(0.22)	1.86(0.17)
	273-373K	No	3053	0.71(0.07)	1.40(0.18)	0.03(0.22)	1.38(0.17)	0.67(0.07)	1.51(0.20)	0.03(0.19)	1.50(0.20)
	273-373K	Yes	3053	0.70(0.07)	1.45(0.21)	0.03(0.23)	1.43(0.20)	0.66(0.07)	1.51(0.17)	-0.02(0.22)	1.49(0.17)
Multi-solvent	298K	No	709	0.82(0.03)	1.64(0.17)	0.01(0.20)	1.63(0.17)	0.76(0.03)	1.93(0.16)	-0.03(0.20)	1.92(0.15)
	273-373K	No	3241	0.23(0.30)	2.50(0.34)	1.05(0.41)	2.24(0.25)	0.50(0.18)	2.06(0.33)	0.41(0.43)	1.98(0.28)
	273-373K	Yes	3241	0.27(0.25)	2.45(0.36)	1.04(0.38)	2.19(0.30)	0.45(0.18)	2.10(0.29)	0.44(0.42)	2.02(0.24)
	273-373K	No	3629	0.76(0.04)	1.92(0.19)	0.39(0.16)	1.87(0.17)	0.65(0.04)	2.33(0.19)	0.49(0.20)	2.26(0.17)
	273-373K	Yes	3629	0.76(0.05)	1.91(0.20)	0.40(0.17)	1.86(0.18)	0.66(0.03)	2.25(0.15)	0.47(0.18)	2.20(0.14)
GF				GAFF				OPLS			
Solvent	Temp.	Temp. Descr.	Datapoints	R^2	RMSE	Bias	SDEP	R^2	RMSE	Bias	SDEP
Chloroform	298K	No	79	0.60(0.13)	1.62(0.29)	-0.03(0.40)	1.57(0.29)	0.54(0.12)	1.70(0.21)	-0.06(0.37)	1.66(0.22)
Carbon Tet.	298K	No	109	0.57(0.11)	1.13(0.15)	0.06(0.28)	1.10(0.14)	0.72(0.10)	0.87(0.15)	0.00(0.25)	0.83(0.14)
Water	298K	No	521	0.87(0.02)	1.52(0.15)	-0.00(0.17)	1.51(0.14)	0.80(0.04)	1.88(0.20)	-0.05(0.23)	1.86(0.20)
	273-373K	No	3053	0.70(0.07)	1.41(0.21)	0.07(0.19)	1.40(0.21)	0.67(0.05)	1.48(0.12)	-0.00(0.17)	1.47(0.12)
	273-373K	Yes	3053	0.71(0.06)	1.38(0.18)	0.08(0.17)	1.37(0.18)	0.68(0.08)	1.44(0.17)	-0.01(0.19)	1.43(0.17)
Multi-solvent	298K	No	709	0.82(0.03)	1.65(0.14)	0.02(0.17)	1.64(0.14)	0.77(0.04)	1.87(0.16)	-0.04(0.17)	1.86(0.16)
	273-373K	No	3241	0.29(0.26)	2.36(0.33)	0.95(0.36)	2.14(0.27)	0.43(0.19)	2.19(0.32)	0.61(0.48)	2.06(0.25)
	273-373K	Yes	3241	0.32(0.28)	2.33(0.40)	0.93(0.41)	2.11(0.32)	0.40(0.17)	2.17(0.30)	0.57(0.40)	2.07(0.25)
	273-373K	No	3629	0.76(0.04)	1.92(0.16)	0.37(0.16)	1.88(0.15)	0.65(0.04)	2.30(0.18)	0.47(0.20)	2.24(0.16)
	273-373K	Yes	3629	0.76(0.04)	1.92(0.16)	0.43(0.15)	1.86(0.15)	0.66(0.05)	2.28(0.22)	0.47(0.20)	2.22(0.19)

Table 7 Solvation free energy predictions for all solvents using random forest models trained on KH, HNC and GF calculated descriptors. Predictions are separated by temperature, inclusion of a temperature descriptor, solute forcefield parameters and number of datapoints. Results for each model are taken from test set predictions. Units are in kcal/mol. Carbon Tet. refers to Carbon Tetrachloride. Organic solvents were modelled with GAFF parameters, and water solvent was modelled with MSPC/E. The standard deviation (SD) for each model statistic over 50 resamples is also shown.