Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Leveraging genetic algorithms to maximise the predictive capabilities of the SOAP descriptor SUPPLEMENTARY INFORMATION

Trent Barnard,[1] Steven Tseng,[1] James Darby,[2] Albert P. Bartók,[3] Anders Broo,[4] and Gabriele C. Sosso[1, a]

[1]*Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom*

[2]*Department of Engineering, University of Cambridge, Cambridge, Trumpington St CB2 1PZ, United Kingdom*

[3]*Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom*

[4]*Data Science and Modelling, Pharmaceutical Sciences, RD, AstraZeneca Gothenburg, Pepparedsleden 1, Mölndal SE-431 83, Sweden*

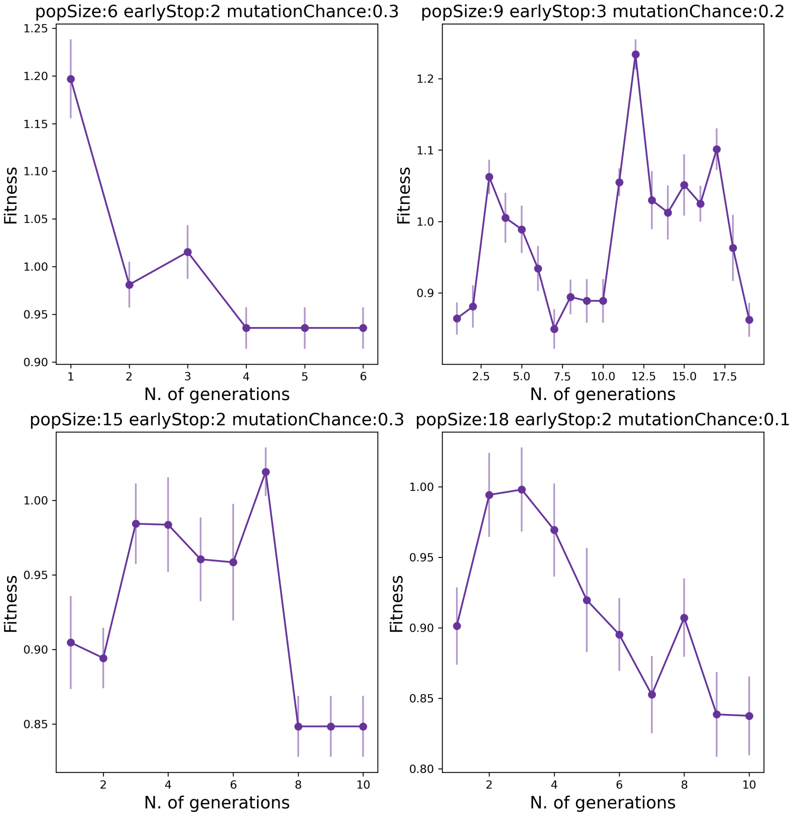
(Dated: 24 October 2022)

We provide supplementary information about:

1. The effect of various SOAP_GAS parameters
2. The multiprocessing scaling of SOAP_GAS using `concurrent futures`
3. The application of the SOAP_GAS algorithm to an additional dataset
4. The effect of different sized datasets on the improvement in score between vanilla and optimised SOAPs
5. A visualisation of the improvement in predictive quality when using an optimised SOAP compared to a vanilla SOAP
6. The correlation between the SOAP hyperparameters for various optimised SOAPs
7. The frequency by which each atomic species appears in the QM7b dataset
8. The difference in optimisation speed and magnitude between SOAP_GAS and a random grid search
9. The performance of the SOAP_GAS algorithm when used with an unseen validation set
10. The statistical significance of the improvement in score achieved with the SOAP_GAS algorithm

[a] Electronic mail: G.Sosso@warwick.ac.uk

I. THE EFFECTS OF VARIOUS SOAP_GAS PARAMETERS



popSize:6 earlyStop:2 mutationChance:0.3

popSize:9 earlyStop:3 mutationChance:0.2

popSize:15 earlyStop:2 mutationChance:0.3

popSize:18 earlyStop:2 mutationChance:0.1

FIG. S1. The impact of various GA parameters on the performance of SOAP_GAS. For each generation, the best performing individual is plotted with error bars denoting the error in the 5-fold cross validation.

As shown in Fig S1 the performance of SOAP_GAS can vary greatly based on the parameters of the algorithm. In theory a low earlyStop should take longer to converge, a high mutationChance and larger popSize should explore more of the parameter space. However, due to the stochastic nature of the algorithm this is not always the case. This stochastic nature is also the reason why the score is not monotonically decreasing, the individuals mutate and change throughout the various generations meaning that it is possible for a score to get worse from one generation to the next. This is why we select the overall lowest score throughout the entire algorithm as out optimised SOAP.

II. MULTIPROCESSING SCALING OF SOAP_GAS

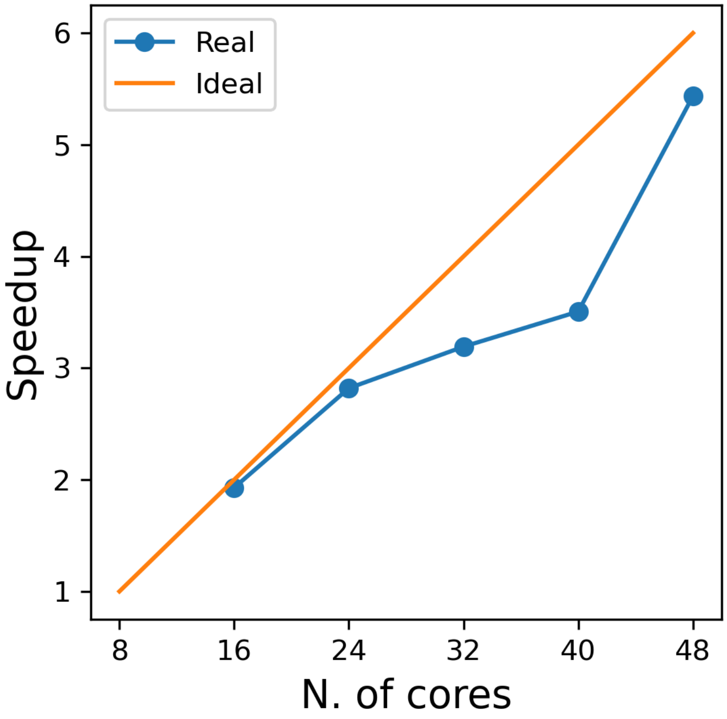


FIG. S2. Scaling test for SOAP_GAS using `concurrent.futures` with a popSize of 48

Clearly, by using `concurrent.futures` we are able to achieve close to ideal multiprocessing speedup. There is a bit of a drop in performance between 24 and 48 cores and this is an artifact of the popSize parameter that we used. So we can conclude that when the N. of cores is a factor of the popSize we achieve near ideal multiprocessing speedup.

III. ADDITIONAL DATASET

The SOAP_GAS framework has also been applied to an additional dataset, namely the QM7b dataset[1,2]. This (relatively low-noise, particularly if compared to our original Solubility dataset) dataset contains 7,211 molecules and 13 target properties, information about the number of atoms and molecular composition is given in table S2. We have chosen the polarizability as our target, as this property has a substantial impact on the biochemical functions of these molecules. The SOAP descriptors have been generated in exactly the same fashion as for the original (solubility) dataset. No atomic species were excluded as either neighbours or centres. The performance of the SOAP_GAS framework has been evaluated in exactly the same fashion as for the original (solubility) dataset. The results are summarised in Table S1 (Full). Similarly to what we have previously obtained for the Solubility dataset, the SOAP_GAS consistently improves on the results obtained by means of “vanilla” (i.e., un-optimised) SOAPs. Notably, the improvement is much more substantial for the QM7b dataset then it is for the Solubility dataset. This is not entirely unexpected: whilst the polarizability of a given molecular specie is entirely determined by its structure, the relationship between structure and solubility depends on other factors (such as the inter-molecular interactions within the solid phase) that have not been considered in this work.

IV. DATASETS OF VARIOUS SIZES

In order to test how the SOAP_GAS algorithm performs using datasets of different sizes, we have considered subsets of the QM7b dataset containing 500, 1000, and 3500 molecules (randomly selected from the full dataset)

The results are reported in Table S1. Again, the SOAP_GAS framework consistently improves on the results of the vanilla SOAPs. Interestingly, the improvement (relative to the vanilla SOAPs) is very similar notwithstanding the dimension of the dataset. The fact that our “Score” metrics increases (i.e. worsens) slightly as a function of the number of datapoints reflects the slightly worse performance in terms of the training set (which is weighted in our definition of the overall Score, see main text) – not the test set. Figure S3 offers a rather clear visual indication of the significant (86%) performance improvement achieved by the SOAP_GAS when applied to the whole QM7b dataset.

	Score			
	Vanilla		GA	
500	0.492		0.036	
1000	0.405		0.037	
3500	0.326		0.046	
Full	0.37		0.052	
	MSE			
	Vanilla		GA	
	Test	Train	Test	Train
500	1.182 ± 0.321	0.391 ± 0.038	0.352 ± 0.071	0.107 ± 0.007
1000	0.992 ± 0.163	0.453 ± 0.030	0.325 ± 0.035	0.136 ± 0.007
3500	0.758 ± 0.072	0.510 ± 0.012	0.295 ± 0.033	0.200 ± 0.003
Full	0.735 ± 0.021	0.588 ± 0.004	0.295 ± 0.007	0.229 ± 0.005
	PCC			
	Vanilla		GA	
	Test	Train	Test	Train
500	0.652 ± 0.072	0.922 ± 0.008	0.914 ± 0.017	0.977 ± 0.002
1000	0.689 ± 0.016	0.893 ± 0.008	0.912 ± 0.01	0.968 ± 0.002
3500	0.761 ± 0.023	0.859 ± 0.003	0.918 ± 0.004	0.947 ± 0.002
Full	0.771 ± 0.005	0.828 ± 0.002	0.917 ± 0.002	0.938 ± 0.002

TABLE S1. Performance of the SOAP_GAS framework applied to the QM7b dataset(target property: polarizability). 500, 1000, 3500 refers to the results obtained by considering a subset of 500, 1000, and 3500 molecules within the whole dataset (Full)

V. SCATTER PLOT OF THE OPTIMISATION OF THE QM7B DATASET

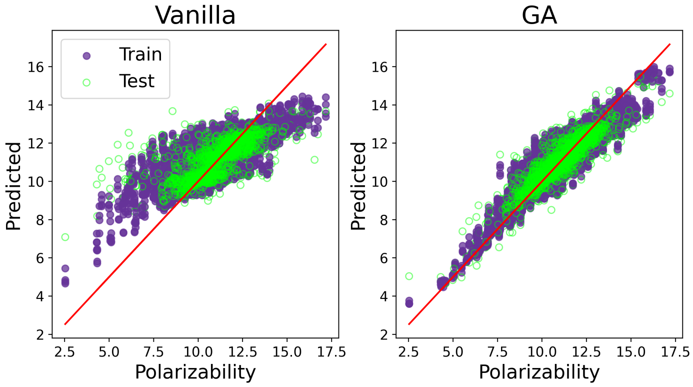


FIG. S3. Performance of the SOAP_GAS framework applied to the QM7b dataset (target property: polarizability, SOAP:all-all). The left panel illustrates the results obtained via "vanilla" SOAPs, whilst the right panel illustrates the results obtained via the SOAP_GAS framework

VI. CORRELATION PLOT FOR OPTIMISED SOAP VALUES

It is also instructive to investigate any potential correlation between the SOAPs parameters (upon optimisation via SOAP_GAS), in the same way we did for the Solubility dataset – where we did not find any evidence of any specific correlation. The result is reported in Figure S4 . Similarly to what we have observed in the case of the Solubility dataset, strong correlations between the

SOAPs parameters l_{max}, cutoff and atom_sigma and the accuracy of the SOAPs (quantified by means of the Score metrics) are absent. Increasing the number of radial basis functions n_{max}, however, leads to consistent, if small gains in terms of performance. The same trend is visible, albeit to a lesser extent, in FIG.3 in the main text for the Solubility dataset.

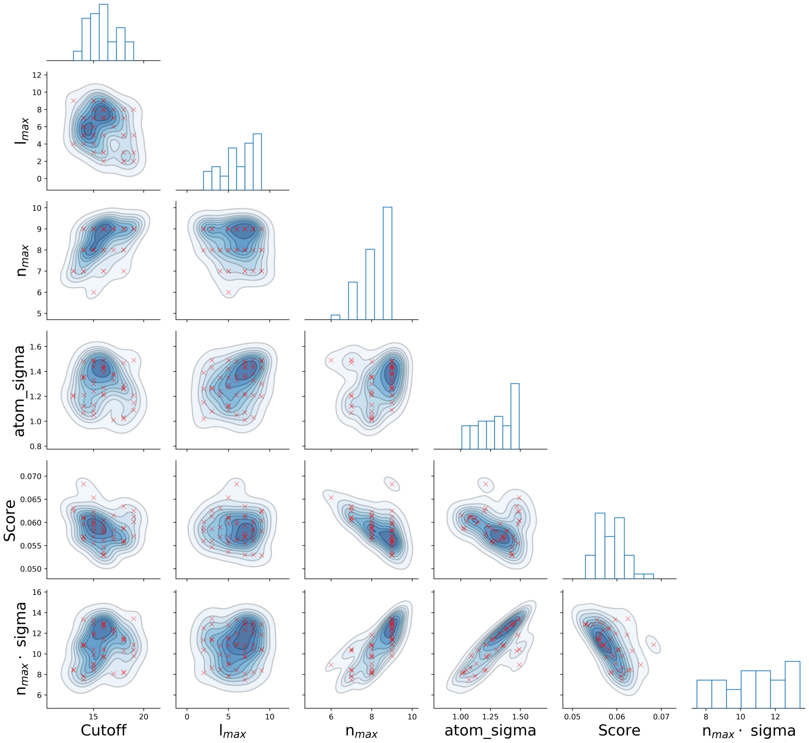


FIG. S4. Correlation between the parameters of the SOAP vectors, optimised via the SOAP_GAS algorithm, for the QM7b dataset(target property: polarizability, SOAP: all-all). The correlation wuth the accuracy of the SOAPs (represented by the "Score" metrics) is also reported.

VII. ATOMIC SPECIES COUNTS

H **61966** (7150)	C **35724** (7150)	O **5965** (4582)
N **6637** (4707)	S **306** (306)	

TABLE S2. Frequency by which each atomic species appears in the dataset. The overall occurrences are reported in bold text, whilst the number of molecules containing a given atomic species are reported in parenthesis.

VIII. COMPARISON TO RANDOM GRID SEARCH

It may also be of interest to the reader how the SOAP_GAS algorithm compares to a randomised grid search approach for the QM7b dataset. This aspect is illustrated in Figure S5, where we report for the QM7b dataset a comparison between the two approaches. It is evident that the SOAP_GAS framework consistently outperforms the randomised grid search approach. In fact, in line with the results reported in Table S1, the performance improvement is more pronounced for the QM7b than for the Solubility dataset.

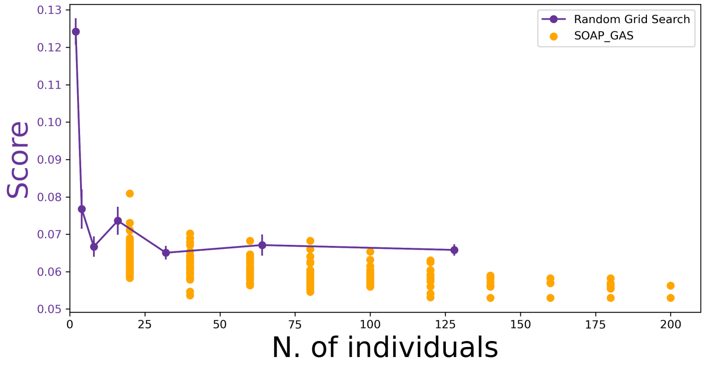


FIG. S5. A comparison between the number of individuals created and the associated scores between the SOAP_GAS algorithm and a random grid search.

IX. VALIDATION SET

In order to evaluate the true performance of the SOAP_GAS algorithm and compare it fairly to vanilla SOAPs, we have included results using a validation set.

In the case of SOAP_GAS, we have built (random selection) a validation set containing 10% of the datapoints. We then applied the SOAP_GAS framework on the remaining 90% of the original dataset, applying a 5-fold cross-validation. Subsequently, we selected the resulting best set of SOAP parameters and used those to train a model (using the remaining 90% of the original dataset) and using the validation set as our test set. In the case of the "vanilla" SOAPs, we have used the same validation set as in the case of SOAP_GAS. We trained a model on the remaining 90% of the original dataset (again, same datapoints as for the SOAP_GAS), using the validation set as our test set. The results are summarized in Table S3 and demonstrate that indeed the performance increase observed for the SOAP_GAS framework if compared to vanilla SOAPs is genuine, and not an artefact due to the lack of a validation dataset. In fact, the relative improvement of the performance is slightly better when considering a validation set as opposed to splitting the whole dataset in just a training and test set – this is likely because when using a validation set the proportion of training data is slightly higher. As such, we are now confident that utilizing a validation set does not have an impact on the robustness of the performance increase reported for the SOAP_GAS framework.

	Score	
Vanilla	0.435	
SOAP_GAS	0.302	
	MSE	
	Train	Validation
Vanilla	1.172	1.227
SOAP_GAS	0.945	1.088
	PCC	
	Train	Validation
Vanilla	0.881	0.872
SOAP_GAS	0.905	0.887

TABLE S3. Performance of the SOAP_GAS framework applied to the Solubility dataset, obtained by utilising a validation dataset (10% of the total number of datapoints)

X. TESTING TO SEE IF THE SOAP_GAS IMPROVEMENTS ARE STATISTICALLY SIGNIFICANT

We have now performed a Z-test (i.e., a statistical test to determine whether two population means are different when the variances are known) on our results in terms of the mean squared error (MSE). The Z-test gives us the probability that the MSE for either the vanilla or the SOAP_GAS optimised SOAPs come from the same distribution. Hence, if the SOAP_GAS optimised SOAP belongs to a different distribution (and that distribution is characterised by a lower MSE), we can safely assume that the performance differences between vanilla and SOAP_GAS results are genuine. Note that this analysis is based on the assumption that the MSE scores from each CV fold are normally distributed. We have verified that this assumption is valid by performing multiple iterations of the 5 fold cross validation using different train/test splits. The results from this is shown in figure S6. A D'Agostino K-squared test[3,4] was performed on this data and returned a p value of 0.103 which indicates that this assumption is statistically valid at the 95% confidence interval. Also note that we have chosen to not use the Z-test to the Pearson correlation coefficient (PCC), as the score would be prone to multicollinearity issues - since MSE and PCC are clearly correlated metrics.

The results are reported in Table S4 for the Solubility dataset. For the All-All (and Double) SOAP as well as the SOAPs centred on C, H, O, N and S, the p-values are < 0.05, which implies the performance differences we observe for SOAP_GAS compared to vanilla SOAP are significant with respect to a 95% confidence interval. The same cannot be said for the SOAP centred on the far less frequent atomic species, namely the halogens. This is somewhat expected, in that our predictions are more accurate when using SOAP centred on the most frequent atomic species, which in turn is reflected in the extent of the accuracy improvement upon applying the SOAP_GAS framework

SOAP	**p-value**
C	1.89e-187
All-All	8.41e-101
H	7.02e-79
O	3.11e-30
Double	1.75e-13
N	3.82e-10
S	0.03
Br	0.31
F	0.31
Cl	0.34
P	0.46
I	0.50

TABLE S4. p-values re: the Z test performed on the MSE of vanilla and SOAP_GAS results. p-values < 0.05 are significant with respect to a 95% confidence interval.

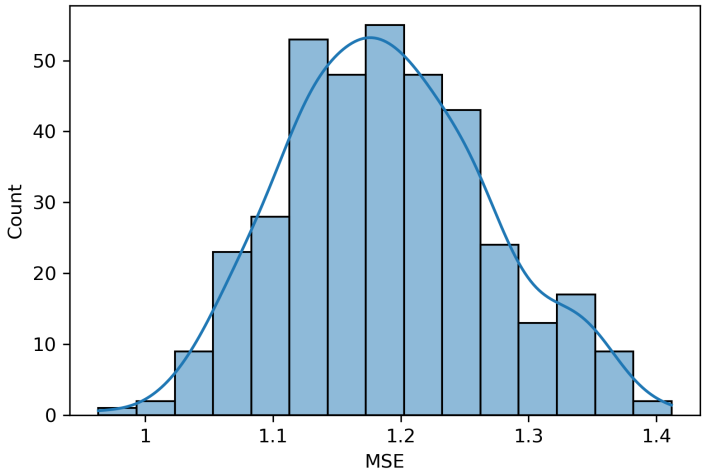


FIG. S6. The distribution of the MSE for the SOAP_GAS optimised All-All SOAP. This data was verified to be normal at the 95% confidence interval using a D'Agostino K-squared test.

[1] L. C. Blum and J.-L. Reymond, "970 million druglike small molecules for virtual screening in the chemical universe database GDB-13," J. Am. Chem. Soc. **131**, 8732 (2009).

[2] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New Journal of Physics **15**, 095003 (2013).

[3] R. B. D'agostino, A. Belanger, and R. B. D'Agostino Jr, "A suggestion for using powerful and informative tests of normality," The American Statistician **44**, 316–321 (1990).

[4] R. B. D'AGOSTINO, "Transformation to normality of the null distribution of g1," Biometrika **57**, 679–681 (1970), https://academic.oup.com/biomet/article-pdf/57/3/679/705358/57-3-679.pdf.