Less may be more: an informed reflection on molecular descriptors for drug design and discovery SUPPLEMENTARY INFORMATION

Trent Barnard, Harry Hagan, Steven Tseng,* and Gabriele C. Sosso†
Department of Chemistry and Centre for Scientific Computing, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

We provide supplementary material about:

- The prediction of selected physical properties for representative structures across the data sets we have considered
- A classification model aimed at showing the reliability of our framework for a biologically relevant target
- An analysis of the variance and correlation of the descriptors we have taken into account

PREDICTIONS FOR REPRESENTATIVE STRUCTURES

We report in Table S1 the predictions of lipophilicity, human hepatocytes intrinsic clearance and glass transition temperature (see main text for further details) for representative molecular structures, using the full set of descriptors described in the main text. In these specific cases, the improvement obtained by using cliques and/or H-wACSFs, especially upon feature selection and optimisation, is particularly clear.

THE TOX21 DATASET

In order to include predictions with respect to a key biological target we have built a classification model, based on random forests, aimed at assessing the toxicity of a number of drugs. In particular, we have considered the very well-known Tox21 dataset, which contains 12 different targets related to the toxicity of a given drug. This dataset is particularly challenging as it is very sparse: only 676 (dataset T_t) out of the 3079 (dataset T_{nt}) data points feature some indication of toxicity with respect to at least one of the 12 targets. Further details about the dataset and said target properties can be found in e.g. Ref. 2.

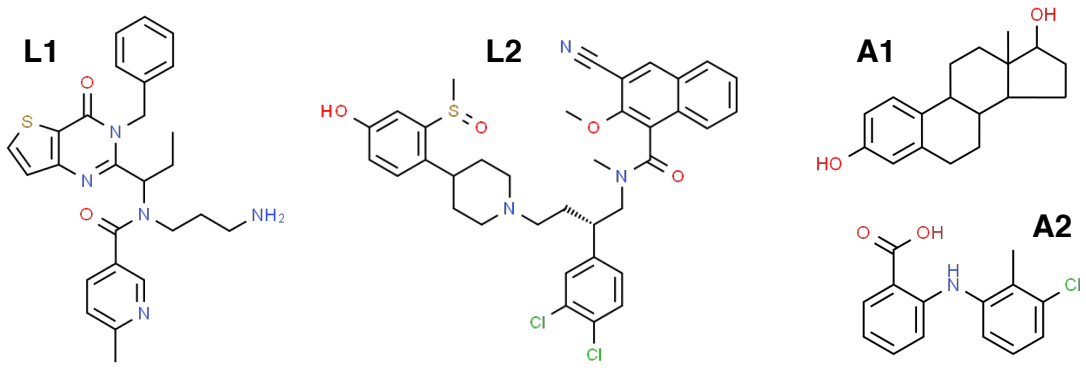
Random forests (RFs) are a machine learning technique that utilises the concept of bagging or bootstrap aggregation to obtain the best prediction from an ensemble of decision trees (DTs). DTs have the advantages of mirroring human decision-making and easy interpretability. However, they suffer from high variance as two decision trees trained on random partitions of data could have very different results [3]. Now notice with a set of n independent observations $Z_1, \cdots, Z_n$ each having variance σ^2, the mean of the observations $\overline{Z}$ has variance σ^2/n, i.e., averaging lowers variance. This method works for regression such that if B different training sets were available to calculate the sets $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_B$ of observations, the average given by

$$\overline{\boldsymbol{Y}} = \frac{1}{B}\sum_{b=1}^{B} Y_b \tag{1}$$

would provide a model with lower variance. Similarly, for a classification model we would choose the mode of the observations for each target instead of taking the mean. Still, as it is uncommon to have multiple training sets available, we may then bootstrap the data we do have through taking repeated samples from it to produce B training sets. This method is called bagging. The RFs algorithm takes bagging one step further by decorrelating the ensemble of decision trees. Consider if there is a strong predictor in the data set. Then the ensemble of bagged trees will use this strong predictor in the top split, producing bagged trees that are similar with predictions that are highly correlated. Instead of this, the RF algorithm at each split takes a random sample of m predictors from the full set of p predictors, usually with $m \approx p$, as split candidates - never allowing a split to consider the majority of predictors available [3]. Once the best predictor is chosen from the sample of m, the process is repeated for the remaining splits. It follows that on average $(p-m)/p$ splits will not consider the strong predictor, giving all other predictors a chance and making the prediction results less correlated [3]. Thus, if m were chosen to be equal to p in building an RF, it would be the same as bagging.

Here, for each target property, we have constructed a dataset containing 200 molecules: 100 "toxic" molecules (i.e. those which are flagged as toxic with respect to that particular target) taken from the T_t dataset and 100 "non-toxic" molecules taken from the T_{nt} dataset. In some cases the number n_t of toxic molecules available for a given target properties is less than 100: in that case, we have included 200-n_t non-toxic molecules in the dataset instead. This choice represents one way to

* Present address: Department of Engineering Technology, University of Twente, De Horst 2, 7522 LW Enschede, The Netherlands
† G.Sosso@warwick.ac.uk



L1: CCC(N(CCCN)C(=O)c1ccc(C)nc1)C2=Nc3ccsc3C(=O)N2Cc4ccccc4

L2: COc1c(cc2ccccc2c1C(=O)N(C)C[C@@H](CCN3CCC(CC3)c4ccc(O)cc4[S+](C)[O-])c5ccc(Cl)c(Cl)c5)C#N

A1: CC12CCC3C(C1CCC2O)CCC4=C3C=CC(=C4)O

A2: CC1=C(C=CC=C1Cl)NC2=CC=CC=C2C(=O)O

Molecule	STD		Cliques		Cliques [FS]		H-wACSFs		H-wACSFs [GAs]	
	MSE	σ	MSE	σ	MSE	σ	MSE	σ	MSE	σ
L1	0.212	0.506	0.166	0.623	0.066	0.233	0.057	0.218	0.087	0.118
L2	0.163	0.601	0.074	0.348	0.035	0.123	0.062	0.203	0.056	0.044
H1	5.188	2.743	0.080	0.078	0.449	0.514	0.118	0.074	0.033	0.026
H2	3.685	2.201	0.107	0.154	0.015	0.016	0.156	0.123	0.010	0.010
A1	1.651	1.306	0.709	0.407	0.396	0.271	0.284	0.285	0.047	0.058
A2	1.347	1.193	1.162	0.710	0.755	0.466	0.214	0.334	0.133	0.221

TABLE S1: **Figure**: 2D structure of selected molecules within the Lipo data set (L1 and L2) and the Amo data set (A1 and A2). The corresponding SMILES [1] string are also included. **Table**: Mean square error (MSE) and corresponding standard deviation (σ) for selected molecules within the Lipo (L1, L2), Hepa (H1, H2) and Amo (A1, A2) data sets. Results for "standard" RDKit descriptors (STD), molecular cliques (Cliques) and histograms of weighted atom-centred symmetry functions (H-wACSFs) are reported. Cliques [FS] and H-wACSFs [GAs] refer to the results obtained for cliques upon feature selection and H-wACSFs upon optimisation, respectively - see main text for further details.

compare prediction across different target properties: we should note that we experimented with alternative strategies - obtaining results similar to those described below. The same k-fold (with k-5) cross-validation strategy we have used throughout this work (see main text) has been used to quantify the accuracy of our model: with a 80%-20% train-test split, the data reported in Fig. S1 refer to 160*5 = 800 and 40*5 = 200 data points for the training and test predictions, respectively, thus including the 5 different splits.

As we discuss in the main text, the cliques descriptor should be particularly suitable to pinpoint the structural properties responsible for biological targets such as toxicity. Indeed, as illustrated in Fig. S1 via the confusion matrices summarising the results we have obtained when testing our classification model on the 12 Tox21 toxicity targets, the cliques descriptor seems to be performing quite well. The model performs quite differently in terms of accuracy according to different target properties: for instance, predictions concerning the SR-MMP target are quite accurate, while in the case of the SR-ATAD5 target our model fails to single out any of the 11 molecules that display some toxicity in terms of this target. The performance of the model is obviously related to the abundance of training data with respect to the specific target: the SR-MMP and SR-ATAD5 targets contain 142 and 11 toxic molecules, respectively - similar examples of the obvious importance of the data points abundance within the Tox21 dataset can be easily found in Fig. S1.

In addition, we have used the very same framework of feature selection described in the main text to identify specific molecular fragments that might be especially relevant for this dataset. Interestingly, we find that there exists a set of 10 cliques only (see Table S2) which are ranked within the top 25 in terms of their Gini importance (see main text) *for all the 12 toxicity targets* in the Tox21 dataset. In fact, while we are in no position to draw any robust conclusion in terms of structure-function relation at this stage, it is intriguing to note that this selection of cliques contains aromatic rings as well as and S-rich functional groups - a staple of many pharmaceuticals currently available (see e.g. Ref. 4).

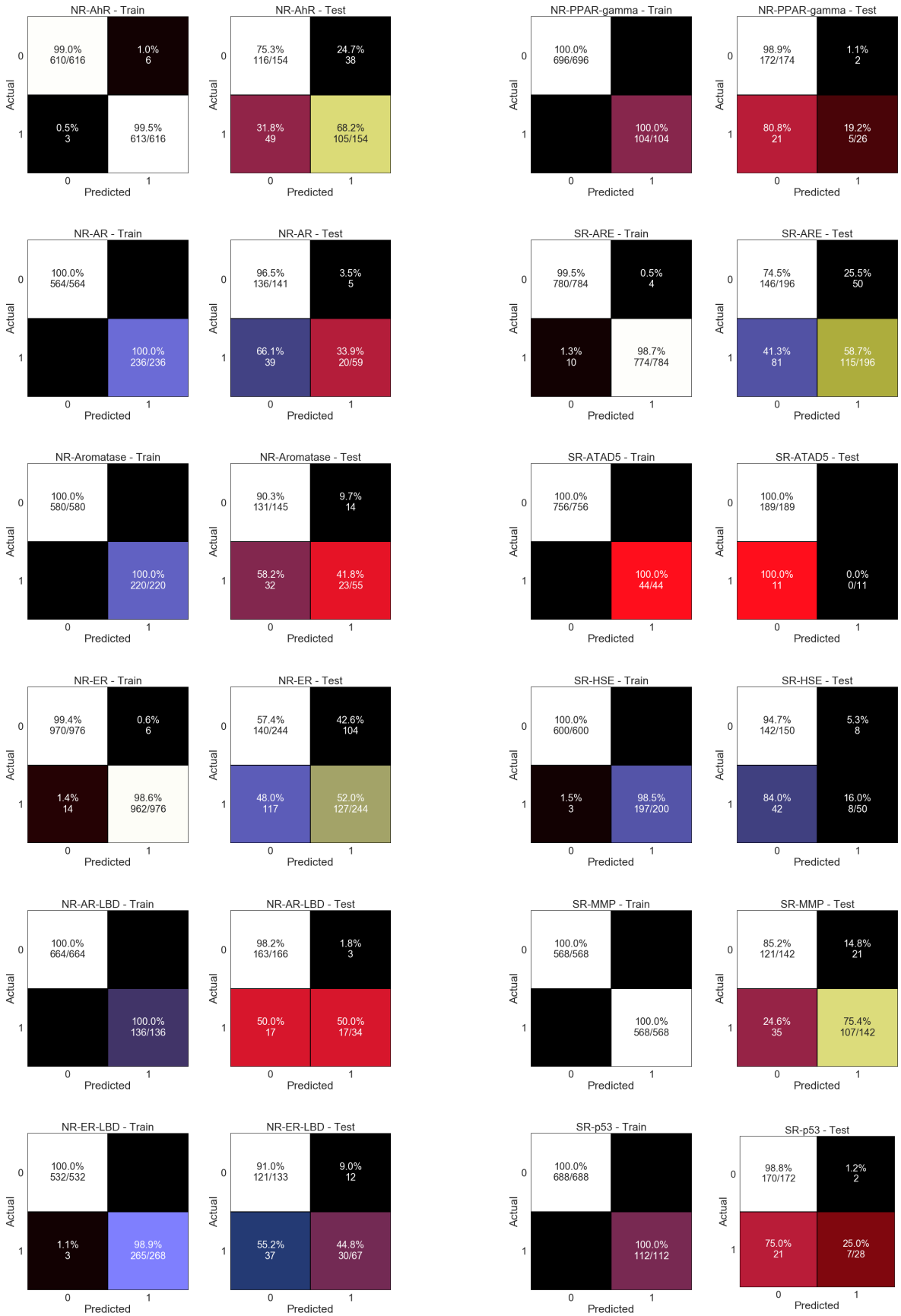


FIG. S1: Confusion matrices summarising the performance of our Random Forest classification algorithm with respect to the (12) target properties of the Tox21 dataset [2]. The molecular cliques descriptor (see main text for further details) alone has been used. Labels of 0 and 1 correspond to non-toxic and toxic activity, respectively.

Feature selection - Cliques *Tox21 dataset*		
Smiles	Gini (mean)	Gini (σ)
CC	0.112	0.007
C1=CC=CC=C1	0.094	0.012
CO	0.085	0.007
C	0.062	0.004
CN	0.058	0.005
C=O	0.053	0.004
CCl	0.035	0.006
C=C	0.029	0.005
CS	0.013	0.002
N	0.013	0.002

TABLE S2: Feature selection for the cliques descriptor in the case of the Tox21 dataset. The full cliques vocabulary contains in this case 310 cliques. We report the average Gini importance index (calculated as described in the main text) and its corresponding standard deviation σ for the ten "most important" cliques: these cliques are within the top 25 cliques with the highest Gini index across all the 12 target properties considered.

DESCRIPTORS: VARIANCE AND CORRELATION

When pre-processing descriptors for machine learning, it is common practice to remove those features with zero or near-zero variance with respect to the dataset under consideration. The reason being that the values of these features are constant or *almost* constant (in the case of near-zero variance) across the entire dataset: as such, one might think that they must be non-informative and thus they can/should be safely removed. In Fig. S2 we report the variance for each descriptors we have used with respect to the three data sets we have considered. A few interesting trends emerge:

- In the situation where they perform the best (the Lipo dataset), the STD descriptors are characterised by the highest amount of near-zero variance features compared to the other two data sets. This evidence seems to suggest that the presence of near-zero features does not necessarily imply the need for their removal.

- Feature selection consistently reduces the amount of near-zero variance features within the cliques descriptor. This is expected, as the cliques descriptor is very sparse - thus with lot of potential to include "rare" features (i.e. cliques corresponding to extremely rare molecular fragments across a given dataset) that might not be selected as very relevant in term of the Gini importance index discussed in the main text. However, in many cases these rare features are exactly what makes a certain molecular species unique in terms of its functional property. Again, we feel that the outright removal of near-zero feature might not be the best way forward in the attempt to increase the predictive capabilities of machine learning models in the context of drug design.

- Optimised sets of H-wACSFs show a very similar amount of near-zero features compared to the non-optimised ones. As discussed in the main text, optimised H-wACSFs perform on average significantly better than non-optimised sets, which once more is suggestive of the fact that near-zero variance features might very well represent a meaningful section of our descriptors.

	MSE		PCC	
	Train	Test	Train	Test
Opt. H-wACSFs	0.124 ±0.019	0.838 ± 0.084	0.936 ± 0.009	0.497 ± 0.134
Opt. H-wACSFs, no 0-var	0.203±0.018	0.975 ± 0.158	0.892 ± 0.015	0.484 ± 0.171

TABLE S3: Mean square error (MSE) and Pearson correlation coefficient (PCC) for both the training and the test set, including the corresponding uncertainties as calculated according to the cross-validation strategy discussed in the main text. *Opt. H-wACSFs* and *Opt. H-wACSFs, no 0-var* refer to the optimised (see main text) set of H-wACSFs for the Amo dataset and the same set where near-zero variance features have been removed.

To further support this claim we have removed near-zero variance descriptors from the H-wACSFs set optimised for the Amo dataset: the results are summarised in Table S3, and they clear show that in this case said removal is actually detrimental in terms of the accuracy of our model.

On a similar note, the correlation between different features also deserves to be discussed. To this end, we report in figure Fig. S3 the correlation matrices for each descriptor we have considered in this work - with respect to different data sets. The values reported in the colour map refer to the Pearson correlation coefficient, calculated pairwise for each feature within a given descriptor: it is evident that the degree of correlation is strongly dependent on the specific nature of the descriptor. In the case of the STD set, we observe blocks of highly correlated features which - interestingly - are consistently found for the three data sets we have considered: this is an indication of the fact that many STD features are intrinsically highly correlated irrespectively of the molecular structures we have considered. This is expected, as features such as "the number of n-membered rings" and "the number of aromatic rings" contain a great extent of redundant information. Conversely, features in the cliques descriptor are very much uncorrelated throughout the whole set of them, even after feature selection - where we only have a handful of molecular fragments involved. This is somehow surprising in that some cliques are definitely redundant: for instance, a clique defined as a 6-membered ring of carbon atoms contains six carbon-carbon bonds, which are also considered as

cliques. Finally, the H-wACSFs features are consistently highly correlated - which is expected, given that many of them do contain redundant information about e.g. the local atomic environment of adjacent atoms within a given molecular structure. Interestingly, the optimised H-wACSFs sets are even more correlated, on average, than the non-optimised ones, and yet the usage of optimised sets did lead to a substantial improvement of the predictive power of our models (see main text). Thus, we conclude that "off-the-shelf" strategies such as the removal of either near-zero variance or highly correlated features should be used with special care.

[1] E. Anderson, G. Veith, and D. Weininger, Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021 (1987).

[2] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, Front. Environ. Sci. **3** (2016), 10.3389/fenvs.2015.00080.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Vol. 112 (Springer, 2013).

[4] M. Feng, B. Tang, S. H. Liang, and X. Jiang, Curr Top Med Chem **16**, 1200 (2016).

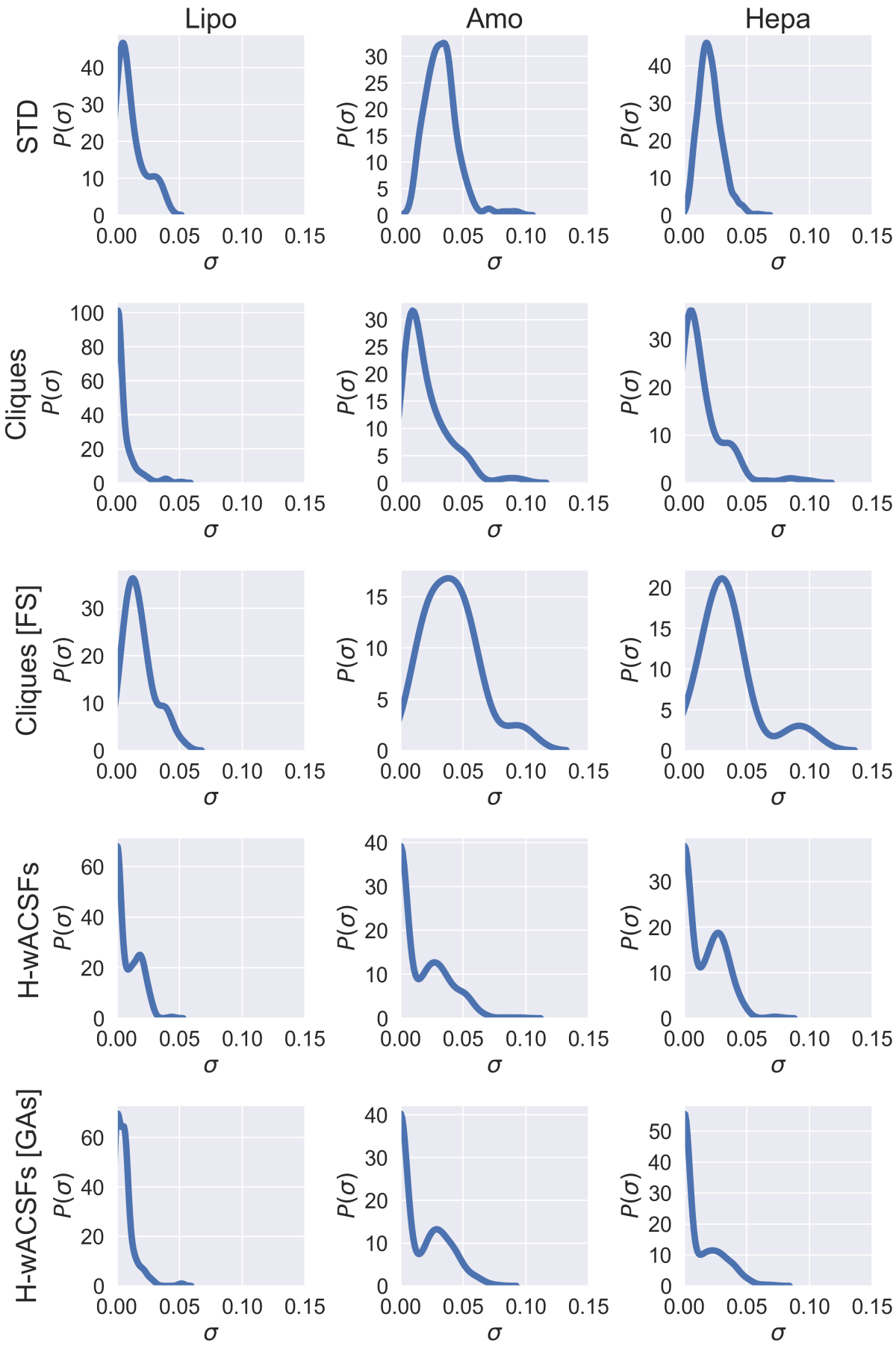


FIG. S2: Probability distribution of the variance of each feature within the different classes of descriptors used - for each dataset we have considered.

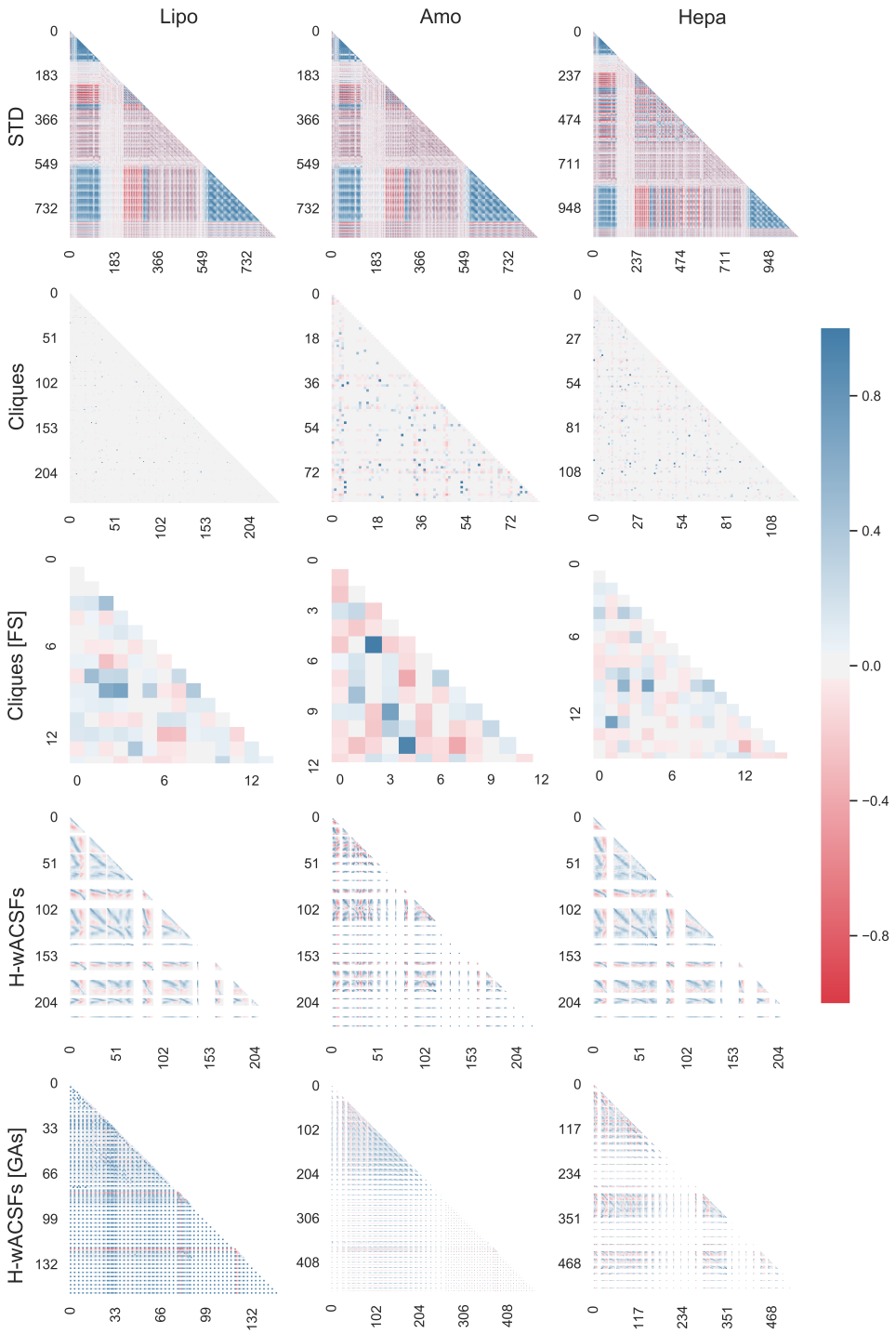


FIG. S3: Correlation matrices for each descriptor we have considered in this work - with respect to different data sets. The values reported in the colour map refer to the Pearson correlation coefficient, calculated pairwise for each feature within a given descriptor. Highly correlated and anti-correlated features are highlighted in blue and red, respectively.