Armed with the digital availability of two natural products libraries, amounting to some 195 885 molecular entities, we ask the question of how we can best sample from them to maximize their “representativeness” in smaller and more usable libraries of 96, 384, 1152, and 1920 molecules. The term “representativeness” is intended to include diversity, but for numerical reasons (and the likelihood of being able to perform a QSAR) it is necessary to focus on areas of chemical space that are more highly populated. Encoding chemical structures as fingerprints using the RDKit “patterned” algorithm, we first assess the granularity of the natural products space using a simple clustering algorithm, showing that there are major regions of “denseness” but also a great many very sparsely populated areas. We then apply a “hybrid” hierarchical K-means clustering algorithm to the data to produce more statistically robust clusters from which representative and appropriate numbers of samples may be chosen. There is necessarily again a trade-off between cluster size and cluster number, but within these constraints, libraries containing 384 or 1152 molecules can be found that come from clusters that represent some 18 and 30% of the whole chemical space, with cluster sizes of, respectively, 50 and 27 or above, just about sufficient to perform a QSAR. By using the online availability of molecules via the Molport system (www.molport.com), we are also able to construct (and, for the first time, provide the contents of) a small virtual library of available molecules that provided effective coverage of the chemical space described. Consistent with this, the average molecular similarities of the contents of the libraries developed is considerably smaller than is that of the original libraries. The suggested libraries may have use in molecular or phenotypic screening, including for determining possible transporter substrates.
fingerprints) increased the chance of (that molecule being) another hit by 30-fold, and used that principle to minimize the size of their initial screening deck accordingly. Other workers have adopted a similar strategy for comparable purposes (e.g., [31,52,59–62]).

Our purposes here are precisely to assess natural products on the basis of their chemical similarity to (or diversity from) each other, with a view to building a “small” library (actually several, of different sizes) of molecules that adequately covers the bulk of the known natural products “chemical space.” Since we confine ourselves purely to the structures themselves, and ignore for the purposes of library generation any (and mainly unknown) bioactivities of such molecules, this endeavour requires an unsupervised analysis, involving suitable clustering algorithms, and appropriate analyses of the clusters so formed. The idea then is obviously that to build a small but diverse library one might produce such clusters and then take a small number of molecules from each cluster for an initial screen. Those clusters producing “hits” can then be explored in more depth. In our case, “hits” are to be substrates for particular solute carriers (SLCs [63–70]) where we seek to understand what may be the “natural” substrates with at least some decent activity [71] so that we can then use cheminformatics to select other molecularly similar candidates to build up a QSAR for the transporter in question. A similar strategy was used by Gründemann et al. to discover the main activity of SLC22A4, previously OCTN1, now known as the “ergothioneine transporter,” [72,73] Other assays of interest involve phenotypic screening (e.g., [74,75]), where a “small” but diverse [76] chemical library is the standard starting point. A particular feature of the present paper is that, following our extensive introductory reference to previous methods, and in contrast to previous analyses, we actually provide the suggested contents of two libraries.

Many long-established clustering algorithms exist (e.g., [77–85]), but most of them either require an indication of the number of clusters desired, behave badly for “connected” rather than “compact” clusters, [86] or scale poorly (e.g., O(n3) for Ward’s [87] complete linkage hierarchical clustering method). Many clustering algorithms are also far from robust, in the sense that while the algorithms will (always tend to) assign an individual to a cluster, this is not always the “correct” one. [88]

Classifying groupings of (i.e., partitioning) natural products or the members of any other defined chemical space can be done in a variety of ways (e.g., [89–102]). One variant (e.g., [103]) joins all molecules into a network based on their similarity being above a particular threshold, leading to a network that contains most molecules (with a few orphans living in small or unique groups). That is not suitable for our purposes, as we wish our clusters to be separate from each other, and a priori we are not necessarily interested in the closeness of cluster seeds or midpoints to other clusters.

Although we do not use it here, we also recognize the power of the closely related dimensionality reduction techniques, such as t-SNE. [104–106]

Most clustering or dimensionality reduction algorithms that have been applied to natural products data are rather of the older type. However, with the increasing interest in “data analytics,” many novel and powerful algorithms have come to the fore (e.g., [107–110]), and a number of these are potentially more suited to the specific task of interest here. In particular, we find that a specific form of hierarchical clustering [107] provides a very convenient and computationally tractable approach. We set the scene with a simple clustering method to provide an overview.

2. Experimental Section

We have access to two main sources of natural products. First and most important is the UNPD database (Ref.[97] and see also Ref. [33]). In addition, we had a license to the commercial Dictionary of Natural Products, that allowed us to make public up to 2% of its contents in any individual paper. The DNP as provided contained 286 310 molecules, meaning that we can (if appropriate) disclose up to 5726 such molecules in this paper. There was and is, unsurprisingly, considerable overlap between these two databases. [72] As in our related projects (e.g., [51,76,111–113]), cheminformatics routines were created and performed using KNIME. [114] We made considerable use therein of the RDKit
package,[49] especially the recent “patterned” fingerprint encoding. For our general overview, we chose fingerprint methods over descriptor-based methods as these retain better the essence of the structural similarities that we are seeking here.

3. Results

3.1. Simple Clustering With Random Seeds

Our first interest was to get a feel for the “graininess” of the different clusters (this may be imagined in much the same way as we consider the “clustering” of stars into galaxies [or not]). Thus, the kind of question we had in mind was “are most molecules all within more-or-less large clusters of similarity, that might be seen as being merged into a continuum, or were many such molecules effectively singletons in their own area of natural products space, and not really like any other kinds of molecule? Among the simplest algorithms for assessing this kind of question is one that runs as follows:

1. Select one compound at random as a cluster center, using the desired encoding (here, we used the RDKit “patterned” encoding).
2. Calculate the Tanimoto similarity (TS) of all remaining unclustered compounds versus the selected center.
3. Add all compounds with TS ≥ a stated threshold to this (newest) cluster and to the clustered list.
4. Select a compound with the lowest TS to the initial center as the new center.
5. REPEAT 2–4 until unclustered count = 0.

We used our union[21] of the cleaned DNP and UNPD databases (195 885 molecules) in which 123 443 were from UNDP, 40 837 from DNP and 31 605 were from both. We looked at a number of thresholds of TS. Those above 0.8 tended to have far too many singletons, while those below 0.8 would necessarily include in the same cluster molecules that were not especially similar. Thus for the threshold that we here publish, we used a TS of 0.8.

The results (Figure 1A and B for an individual run), which were broadly similar when averaged over several runs with random starting molecules (Figure 1C), show the expected type of distribution, with one very large cluster followed by increasing numbers of smaller clusters. Each run took about a day on a modern 8-core PC. The number of singletons, doubletons, etc are given in Table 1, up to clusters of 10.

Thus, only some 16 000 molecules (but 5053 clusters) can be discarded on the grounds that they are especially unusual or rare; this inevitably leaves a huge number of molecules to select from for a “small” library (we shall pick 96, 384, 1152, and 1920 as possible “small” library sizes), and nothing changes the fact that some clusters must be discarded if 195 885 molecules are to be reduced to something like one molecule per cluster we might still miss some “hits.” In addition, a better strategy would include more molecules from the very largest clusters, weighted in some way. Other criteria might include availability, structural complexity, and price.

To assess the broad “quality” of the clusters, we have chosen to illustrate the first cluster with 5 members (arbitrarily numbered as cluster 3083) and show it in Figure 2A; it is clear that this is a very sensible cluster, with the overwhelming majority of its members containing a fused indole-pyridine ring (i.e., β-carboline or norharmane) motif, a widespread grouping[115] of psychoactives[116] that can also serve as a scaffold for synthetics (e.g.,[117]). The trivial names of the same molecules are given in Figure 2B. With any analysis of this type, it does assume that the structures are correct, and, for instance, at least that for Hylamine (left-hand-most structure, marked green) is in doubt.[118] However, we do not rehearse these in further detail, save to mention that the largest cluster contains a great many sterols, molecules that were among the most abundant both in terms of approved drugs and of endogenous metabolites.[51]

Another way to show the benefits of the clustering is to compare the Tanimoto similarities of a subset of the starting library with that of one of the libraries after clustering. Thus Figure 2C shows a random subsample of 1152 molecules, where some of the larger clusters emerge naturally, while Figure 2D shows that this is much reduced when the chosen library is assessed in the same way (and note that 19, 8, and 7 molecules are purposely taken from the first three clusters, respectively).

We have also analyzed the larger clusters in terms of subclusters using the K-Medoids algorithm, varying K according to the (square root of the) size of the clusters. The square root weighting avoids extracted sub-clusters being dominated by those few clusters that have a large size. K-Medoids provides cluster centers that are exemplars of actual data points rather than the abstract vectors that are returned by the K-Means algorithm, and additionally, K-Medoids as implemented in KNIME, can use fingerprint based Tanimoto similarity directly. While K-Medoids is somewhat slower than K-Means, this is less of an issue when sub-clustering a relatively small cluster rather than clustering the entire library.

3.2. Hierarchical K-Means Clustering Strategy (KMHC)

The ideas behind this were developed, in particular for purposes of chemical clustering, by Böcker et al.[107] based in part on earlier implementations in bioinformatics.[119] They recognized[107] that while hierarchical methods were potentially better from the point of view of interpretability, they had unfeasibly long run times for large libraries. However, while non-hierarchical methods such as K-means[120] are fast (O(N²) in time), they can lead to large and heterogeneous or small and exclusive clusters, depending in particular on the number and nature of predefined clusters and the distance (boundary) conditions. This KMHC strategy seeks the benefits of each by setting K = 2 at each level and splitting each cluster hierarchically. Much as with a heatmap representation common in transcriptomics[121] (and as we have used in structure similarity comparisons[51,75]), the clusters may also be displayed as dendrograms, making interpretation comparatively easy.
K-Means is based on Euclidean distance and the cluster centers returned are thus abstract vectors; we have, however, used the nearest actual data points to the abstract vectors to represent the cluster centers as this is more chemically informative.

In the KMHC method\[107\] there is a variable threshold that may be set to determine the minimum cluster size in terms of Euclidean distance; this inevitably affects, initially inversely (Figure 2 of Ref.\[107\]), the number of singletons and the number of clusters. The former decreases monotonically, while the latter peaks. Based on this (the peak), we first chose a (Euclidean) distance threshold of 5.7.

Table 1. Relation between cluster size and number of molecules for low cluster sizes.

| Cluster size (molecules) | Number of clusters | Number of molecules + size x number | Cumulative number of molecules |
|-------------------------|--------------------|-------------------------------------|--------------------------------|
| 1                       | 1802               | 1802                                | 1802                           |
| 2                       | 1017               | 2034                                | 3836                           |
| 3                       | 595                | 1785                                | 5621                           |
| 4                       | 429                | 1716                                | 7337                           |
| 5                       | 324                | 1620                                | 8957                           |
| 6                       | 252                | 1512                                | 10469                          |
| 7                       | 188                | 1316                                | 11785                          |
| 8                       | 171                | 1368                                | 13153                          |
| 9                       | 147                | 1323                                | 14476                          |
| 10                      | 128                | 1280                                | 15756                          |

Figure 1. Simple clustering of the union of molecules in UNPD and DNP natural products databases. A) Number of clusters vs cluster size. B) Count of molecules in cluster, ranked according to size (largest first), vs cumulative count of molecules. C) Box-and-whisker plot of the variation in the general shape of cluster distribution over 10 runs.

K-Means is based on Euclidean distance and the cluster centers returned are thus abstract vectors; we have, however, used the nearest actual data points to the abstract vectors to represent the cluster centers as this is more chemically informative.

In the KMHC method\[107\] there is a variable threshold that may be set to determine the minimum cluster size in terms of Euclidean distance; this inevitably affects, initially inversely (Figure 2 of Ref.\[107\]), the number of singletons and the number of clusters. The former decreases monotonically, while the latter peaks. Based on this (the peak), we first chose a (Euclidean) distance threshold of 5.7.

Figure 3A and B shows the decrease in cluster size vs cluster number in a manner equivalent to Figure 1. Note that cluster numbers are here simply labels. It is clear that clusters are much better behaved, with the largest having only 558 members (albeit now 5320 singletons rather than 1802). Although the suggested sampling regime (according to the square root of the number of cluster members) affects this slightly, the number of molecules included in n clusters as n = 96, 384, 1152, and 1920 is, respectively, 16 286 (8.2% of the total; smallest cluster size of 98),
This would suggest that a library of just 96 molecules is probably too small to be usefully representative, but that three or five 384-well plates can give a very decent coverage of one third or two fifths of natural products space. To achieve 50, 60, or 75% coverage on this basis would require 2913 wells (smallest cluster size 15), 4830 wells (smallest cluster size 11), and 7672 wells (smallest cluster size 7), respectively. These latter cluster sizes, and maybe that of the 1920 library, represent numbers from which it would be impossible to create a sensible QSAR, while the fact that one can sample 18 or 32% of the variance with one or three 384-well plates is very attractive.

Figure 4A and B shows the first cluster of 70 molecules (in decreasing cluster size); in this case it is again obviously rather homogeneous, being based on unsaturated long-chain fatty alcohols, that may exhibit activity as insect pheromones.\textsuperscript{[122,123]}

We note that there is no simple mapping between a Euclidean distance (as used in KMHC) and the Tanimoto similarity (TS), but having a starting molecule allows one to run “one-vs-all” to assess those other molecules that are closest; as a standard similarity search, this is a rapid and efficient procedure. When we did this for 1152 random samples, the closest cluster as judged by TS was the same as that found by the KMHC/Euclidean distance some 82% of the time, consistent with a good albeit not perfect degree of robustness of the clusters.

Finally, to illustrate the effect of threshold choice, we did the same analysis (Figure 5A and B) using a threshold of 8.0. As expected, this results in a large initial cluster (of 2875) and fewer clusters (~2900) overall. As before, we show structures (Figure 5C) and names (Figure 5D) of the members of the first cluster of size 70. Finally, to illustrate the “coverage” of the natural products chemical space by our clusters, we used CDK\textsuperscript{[124,125]} to extract 30 standard molecular descriptors (Table 2), normalized each of them (to the range zero to 1, via [value minus minimum]/[maximum value minus minimum value]), and assessed the principal components of their variance (the first two of which explained 60.91 and 24.00% of the variance). Figure 5E shows the cluster medoids (in red) as well as every molecule (blue) visualized as a plot of X log P vs value of the first principal component. It is clear that our clustering method does permit excellent coverage of the chemical space. Supplementary file S1 NPKMHc_T8.0_limited_DNP_hc_filter_out.xlsx gives a subset of the clustered data that contain both UNPD data and 2000 from DNP and whether or not they were cluster centers.
We then sought to assemble a slightly less “virtual” library by choosing molecules that were (i) as suggested by sampling the clusters appropriately, (ii) actually available. The latter compared the central point with molecules available via MolPort (www.molport.com), whether the exact molecules themselves or ones enjoying a Tanimoto threshold above a certain level (we again used 0.8). Under these circumstances, the same molecule can serve more than one cluster, and when this was done the 1920-member library was effectively reduced to 1240. Some of these were rather unexciting hydrocarbons not considered bioactive, so these have been removed. The overall picture is illustrated in Figure 3F, and — in contrast to previous analyses of this type — is given in Supplementary Information as Molport_KMHC1920ssThr8.0_FoundFiltered.xlsx.

### Figure 3. Hierarchical K-means clustering. The threshold was set at 5.7. A) A number of clusters vs cluster size B. Count of molecules in cluster, ranked according to size (largest first), vs cumulative count of molecules.

### 4. Discussion

There is a considerable body of literature (e.g., [126–133]) devoted to the in silico assessment of chemical library diversity and its maximization (usually without too much worry as to its size).

Our chief purpose here was slightly different, as it was to exploit the availability of digital representations of two large (and partially overlapping) natural products libraries so as to select (in silico) smaller subsets of these molecules that might adequately represent the much larger chemical space occupied by the full set. Since we required generality, rather than seeking to understand bioactivity in any given assay, we focussed on a purely unsupervised, clustering strategy. As recognized by
Everitt\cite{29} there is no “best” clustering, and clustering methods should ultimately be judged by their utility. Certainly one could imagine any number of alternative clusters or subsets of 195 885 molecules (indeed even the number of pairs and triples one can obtain from 195 885 molecules exceed $10^{10}$ and $10^{15}$, respectively). However, we believe that the simple and efficient methods that we have applied have produced, and then sampled from, robust clusters, as judged by standard robustness methods (reviewed in Ref.\cite{88}). We have also provided evidence that the average similarities between the subsets we have chosen are far greater than were those of the original dataset. What we have not done is to assess the “drug-likeness” of our libraries, since it is

\textbf{Figure 4.} Analysis of the “homogeneity” of a subcluster from KMHC chosen as the first cluster in diminishing size to have 70 members. A) Structures, B names (mainly trivial). Colors encode the origin of the molecules: Blue DNP only, red UNPD only, green both.
known (e.g., from the original analysis of UNPD[97]) that marketed drugs occupy only a small portion of the natural products space. Nor have we applied diversity metrics of the kind widely used in metagenomics. This is because (i) the latter are only relative, and (ii) we did not want the most diverse library (since this would almost certainly have included obscure singletons that were not easily available); our aim was to find a reasonable library that was in some sense representative of the

Figure 5. Hierarchical K-means clustering. The threshold was this time set at 8.0. A) A number of clusters vs cluster size B. Count of molecules in cluster, ranked according to size (largest first), vs cumulative count of molecules. C) Structures of members of the first cluster with 70 members. D) Names of the molecules in C. E) Illustration of the coverage of chemical space (blue dots) by a small library (red dots) visualized in a 2D plot of XlogP against the first principal component of the variance. F) Analysis of Representative molecules from different clusters and their the availability of the medoids or chemically very similar molecules for purchase (assessed using the MolPort KNIME node).
mainstream of known natural products, and this was to be done (and was done) by selecting from the larger clusters only. A second benefit of selecting from clusters with more representatives is that one would more likely have access to sufficient material for a QSAR without the need for extra chemical synthesis.

Finally, in our selections, we also ignored any thoughts of molecular target promiscuity (see e.g.,[65]) because (i) we are mainly interested in phenotypic screening (ii) we consider that docking procedures and the number of available target structures are insufficiently developed to make this a useful addition at this time. At all events, we do stress the point that, uniquely so far as we are aware, we have actually provided the contents of our chosen library for workers to assemble and use as they see fit.

On the flip side, an inevitable consequence of being selective is that we lose some things in which we might have an interest; thus, the “berberine” cluster contained only seven members, despite berberines having a wide array of bioactivities,[134] and would not feature. Under these circumstances, manual adjustment of the libraries would be required.

5. Conclusions

In conclusion, our analyses suggest that natural products space in different regions is at once both sparse and dense; the former adds very significantly to diversity, while the latter allows us to select very much smaller libraries for initial assays. The fact that larger libraries likely originated earlier in natural evolution also implies[71] that there has been a longer period for transporters to be selected to take them up. Such considerations suggest strongly that small but representative libraries will be a very useful resource for assessing the substrate specificities of solute carriers. We consider that the provision in digital form of the contents of actual library will provide a particularly convenient resource for workers who wish to assemble it in chemical form and use it for various purposes.

Abbreviations

KMHC, K-Means Hierarchical Clustering; QSAR, Quantitative Structure Activity Relationship; TS, Tanimoto similarity.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

We thank the UK BBSRC (grants BB/P009042/1 and BB/M017702/1) for financial support.

Conflict of Interest

The authors declare no financial or commercial conflict of interest.

Keywords

drug transporters, cheminformatics, encodings, endogenites, metabolomics, maximum common substructure
The elements of statistical learning: data mining, inference and prediction, (2nd ed.), Springer-Verlag, Berlin 2009.

Clustering in bioinformatics and drug discovery, CUP, Cambridge 2009.

Algorithms for clustering data, Prentice Hall, Englewood Cliffs, NJ 1988.

The elements of statistical learning: data mining, inference and prediction, (2nd ed.), Springer-Verlag, Berlin 2009.

Numerical taxonomy, Edward Arnold, London 1963.

Mathematics for machine learning, Cambridge University Press, 2019.

An introduction to statistical learning, Springer, 2017.

Probability and statistics for engineers and scientists, 9th edition, Cengage Learning, 2018.

Introduction to information retrieval, CUP, Cambridge 2009.

An introduction to statistical learning, Springer, 2018.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2018.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.

An introduction to statistical learning, Springer, 2017.