Published in final edited form as:
Pharm Front. 2019 ; 1(1): . doi:10.20900/pf20190005.

Generation of a Small Library of Natural Products Designed to Cover Chemical Space Inexpensively

Steve O'Hagan[1,2], **Douglas B. Kell**[1,2,†,*]

[1]School of Chemistry, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK

[2]Manchester Institute of Biotechnology, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK

Abstract

Natural products space includes at least 200,000 compounds and the structures of most of these compounds are available in digital format. Previous analyses showed (i) that although they were capable of taking up synthetic pharmaceutical drugs, such exogenous molecules were likely the chief 'natural' substrates in the evolution of the transporters used to gain cellular entry by pharmaceutical drugs, and (ii) that a relatively simple but rapid clustering algorithm could produce clusters from which individual elements might serve to form a representative library covering natural products space. This exploited the fact that the larger clusters were likely to be formed early in evolution (and hence to have been accompanied by suitable transporters), so that very small clusters, including singletons, could be ignored. In the latter work, we assumed that the molecule chosen might be that in the middle of the cluster. However, this ignored two other criteria, namely the commercial availability and the financial cost of the individual elements of these clusters. We here develop a small representative library in which we to seek to satisfy the somewhat competing criteria of coverage ('representativeness'), availability and cost. It is intended that the library chosen might serve as a testbed of molecules that may or may not be substrates for known or orphan drug transporters. A supplementary spreadsheet provides details, and their availability via a particular supplier.

*Correspondence: Douglas Kell, dbk@liv.ac.uk; Tel.: +44-151-795-7772.
†Present address: Department of Biochemistry, Institute of Integrative Biology, Biosciences Building, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK

Data Availability

- The dataset generated from (or analyzed in) the study can be found in supplementary materials.
- All data generated from the study are available in the manuscript or supplementary files.

Author Contributions

SOH and DBK conceived the study and discussed it as it unfolded. KNIME workflows were written by SOH. The manuscript was written mainly by DBK. Both authors read and approved the manuscript.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Europe PMC Funders Author Manuscripts

Keywords

drug transporters; cheminformatics; endogenites; metabolomics; clustering

Introduction

It is by now evident (e.g., [1–29]) that pharmaceutical drugs exploit endogenous transporters that normally transport biological metabolites (whether they are endogenous, or are represented by exogenous natural products). The possibly surprising quantitative consequence of these and other studies is that diffusion of such drugs through the phospholipid bilayer portions of undamaged biological membranes is in fact negligible [1,3,5–7,10,11,13,30].

The principle of molecular similarity (e.g., [31]) implies that small molecules with similar structures will bind to the same kinds of proteins and exhibit similar kinds of activity. We [2,16,32–37] and others (e.g., [38–43]) have thus sought to assess the extent to which marketed drugs are similar in structural terms to endogenous human metabolites (that we sometimes refer to as "endogenites"). The criterion of being marketed was used because this implied that the drugs were efficacious and (since almost all were to be taken orally and/or required to interact with intracellular targets) capable of being transported across at least one biological membrane. It turned out [36] that when standard encodings were employed, and a Tanimoto similarity exceeding ~0.8 was used as a criterion of "similarity", all drugs could be seen to be similar to either endogenites (~15%) or (more frequently) to natural products (commonly of plant and microbial origin), but that for similarities below this the various encodings often gave completely different rank orders.

This latter finding, the importance of natural products in the natural selection of transporters, raises a more ecological kind of thinking [44–46], in which it becomes obvious that the ability to take up natural products (such as cocaine [47], ergothioneine [48,49], and many others) is indeed likely to improve the fitness of an organism with a protein transporter capable of transporting them.

As with the products of many other genes uncovered by the systematic genomic sequencing programmes (e.g., [50]), many transporters remain "orphans" [12], with no known substrates. Clearly one strategy to "de-orphanise" them would be to try all kinds of substrates in parallel and use the methods of 'untargeted metabolomics' to assess their uptake differentially in cells expressing different amounts of the transporter of interest (e.g., [48]). Another method is to try many drugs serially, but this would be prohibitively expensive for large libraries. Consequently one strategy (e.g., [37,51]) that we have chosen is to develop a small and 'representative' library that might reasonably cover natural products space efficiently and inexpensively, and that might then be used to assess which of its members were substrates for particular transporters. Having established the greatest activities, those small molecule structures could then be used as "seeds" for the acquisition and analysis of other molecules with which to establish a suitable QSAR. Armed with that, and the concentrations of the transporters themselves, one would then have the information

Europe PMC Funders Author Manuscripts

necessary to permit the calculation of the activity of that transporter for any drugs in different cells.

The only "missing piece" in the generation of this kind of library hinged on the commercial availability and cost of the molecules themselves. As with other programs of this type (e.g., [52–56]), the desire is for a library that is both diverse yet accessible. In collaboration with a commercial partner, we have now developed a library that is at once small, suitably comprehensive, and with a price that is accessible to most reasonably funded laboratories. It is this that we here describe.

Materials and Methods

As in our related projects (e.g., [15,32–35]), we developed and ran our cheminformatics routines in the KNIME environment [57,58], including on occasion two nodes available from the Molport website at https://www.molport.com/shop/knime-nodes. We made considerable use of the RDKit package [59], especially the most recent "patterned" fingerprint encoding. Other software used is referenced in the Results section.

Results

Our previous work [37] separated the large UNPD (Universal Natural Products Database [60] http://pkuxxj.pku.edu.cn/UNPD/)(and the commercial Dictionary of Natural Products (DNP) library) of natural products into appropriate clusters, ranked by cluster size. To create a small and suitably priced library that might nonetheless give good coverage of it, we used the following general algorithm (given as pseudocode):

- Rank each cluster according to its size
- Filter out duplicate molecules
- Pick a subset of each cluster proportional to the square root of the cluster size and such that the total number of subset compounds selected over all clusters is equal to some maximum library size (we initially chose 1920)
- Pick the molecules within the MolPort database closest to each of the cluster subset members
- Continue for any cluster subset containing more than five molecules, stop when no further cluster subsets pass these criteria

Figure 1A shows a PCA plot of 504 molecules that met these criteria. Clearly some molecules are very expensive and fail our criterion of affordability. We also show five representative structures, indicating a variation in complexity over the first PC. Exact matches between Molport molecules and those in the databases are also more common towards the left-hand side of the first PC.

Figure 1B shows the same data when they were subject to a price ceiling of $100 per molecule (regardless of quantity). We then added two more criteria.

- Filter out any molecules with SLogP > 5.0

- Keep only molecules that pass chosen price and availability criteria (usually this was at least 25 mg for less than $100)

The final filtered list of 167 library compounds, taken either at or near 167 unique cluster centres out of the total of 7363 clusters represents 2.27% of clusters. Taking cluster membership into account, these 167 clusters represent approx. 8200 compounds out of a total of 195,000 compounds (~4.2%). Whilst these figures seem small, they give no clear indication of how well the library covers natural product chemical space because most clusters are in fact tiny.

For purposes of visualization, we extracted a random subset of 5000 molecules from the UNPD dataset studied previously [37]. We used the full set of RDKit's numerical scalar descriptors, except that correlated descriptors were filtered out with a correlation threshold of 0.98, and z-score normalized (descriptors as available in KNIME were used, see https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors).

Figure 2A shows a Principal Components Analysis of those 5000 molecules (dots) together with the 117 molecules chosen (triangles) with filter criteria on SLogP, price and availability. It is clear that apart from the more sparsely populated part of chemical space to the right we do indeed have good coverage of the whole natural products (and natural-product-like) space. A more principled way of performing and visualizing dimensionality reduction is represented by the now-well-known variant of Stochastic Neighbor Embedding known as t-SNE [61]. In contrast to PCA, t-SNE is a nonlinear algorithm that does not admit projection of new data. To get round this, we first calculated the t-SNE coordinates in the normal way; we used Python Scikit-learn TSNE with default parameters and pre-computed distance matrix whose elements were (1.0—RDKit Pattern Fingerprint Tanimoto similarity) with t-SNE parameters: n_components = 2, perplexity = 30.0, early_exaggeration = 12.0, learning_rate = 200.0, n_iter = 1000, n_iter_without_progress = 300, min_grad_norm = 1e–07, metric = "precomputed", init = "random", method = "barnes_hut", angle = 0.5). We trained a random forest model [62] using RDKit Pattern fingerprints as the input and the two t-SNE values as the output. We could then project in the new compounds of interest (cluster representatives) by passing them through the trained RF model in the same way. Thus Figure 2B shows a t-SNE plot of the same data, indicating that indeed this library covers the great majority of the chemical space. Those parts least covered (in orange) were not in fact from clusters that had only a very few members (and thus unable to provide sufficient members for a sensible QSAR analysis), but mainly from clusters containing compounds that did not meet our price or availability criteria.

96% of library compounds were exact matches to their target (TS = 1.0), most of the rest were either close isomers, tautomers or alternate charge states. The worst Tanimoto similarity between target and library compound found was 0.858 for the charged and non-charged versions of Chlorin e6.

Because of issues related to the same compound being represented by different tautomeric forms and charge states, *etc.*, we have not been able to find a foolproof procedure to standardize compound representations into a truly 'canonical' form, hence Tanimoto similarities somewhat less than 1.0 can nevertheless correspond to identical molecules.

Europe PMC Funders Author Manuscripts

To assess the extent to which our clustering and subsetting has provided a much more widely separated set of molecules, we again encoded the molecules using the RDKit Pattern fingerprint. Figure 3A shows a heat map [63] of the 5000 subsample molecules as judged by their Tanimoto similarities, with a mode value being around 0.7. Figure 3B shows a similar analysis for the cluster representatives in the Molport library, where it is clear that far fewer have a mutual Tanimoto similarity exceeding 0.8, *i.e.*, we have covered the available space much more sparsely, as intended. Figure 3C shows the heat map for library samples *vs.* the 5000 subsample.

In a similar vein, Figure 4A,B show the similarities to each other of the 5k and cluster representatives when the fingerprint Euclidean distances (rather than Tanimoto similarities) are used. In this case the abscissa represents the square root of the number of different bits and blue represents more similar. Again the extraction of cluster representatives has pulled the average similarities away from each other.

Finally, Figure 5 shows a PCA plot of the 117 molecules that passed our criteria; the amounts are encoded by colour and the cost by size, while the shape encodes whether their partial charge at neutral pH ≥ 0.5 and thus whether they are likely to be observed in positive ionization mode in a mass spectrometer. Again, notwithstanding some outliers to the right, there is a reasonable coverage of the available chemical space. The set of molecules is given in the supplementary spreadsheet. In practice, molecules go in and out of availability, and at the time of finalizing this manuscript only 116 of the 117 were in fact available. Consequently, we have not extended our analyses beyond this.

For those with larger budgets, we have also listed other representative quantities and guide prices in different tabs in the attached Supplementary Excel sheet. Both guide prices were optimized considering the total cost of compounds and shipping combinations.

Discussion

Our aim in the present work, as part of a programme aimed at deorphanising (*i.e.*, finding the substrates for) membrane transporters, was to build on the recognition that many evolved and were selected to take up (or to efflux, or both) exogenous natural products (e.g., [36,48]). Although natural products space is occupied by far fewer known molecules (e.g., [60,64–68]) than either those possible [69] or the set of ~230 million mainly synthetic molecules collated e.g., at ZINC [70](http://zinc.docking.org/), it is still very large. Purchasing every possible molecule is prohibitively expensive, even for the subset of known (~200,000) natural products, and even if it were not many are either commercially unavailable or just singletons unsuited to our purposes (which aims to build a QSAR based on an initial hit followed by possible candidates that bear at least some chemical similarity to it). This is a simple extrapolation of the principle of molecular similarity, and the finding that molecules close in structure to a molecule with a certain activity are substantially enriched for that activity. In the landscape metaphor (e.g., [71,72]), this is equivalent to the assumption that a "starting" hit should at least be in the foothills of a more or less isolated mountain range that one would wish to explore (noting that in phenotypic sceening the objective function may involve or even favour polypharmacology).

A standard activity in cheminformatics is thus the production of reduced chemical libraries that cover the chemical space of interest [51], and that should still contain molecules that are (i) commercially available, and (ii) reasonably cheap. Cost provides a particularly clear filter [73]. Obviously this latter is a function of a laboratory's budget, so we focused on the smallest library of this type that one might purchase in reasonable quantities for a somewhat arbitrary $5000 or so.

Plate-based screens are well known to be rather prone to edge effects [74,75], so while one might have suggested that we specify a number of molecules that might have a multiple of 90 wells or so (to allow for controls), we do not feel bound by this as numbers such as 117 allow arraying in a manner that easily avoids them.

Conclusions

Conventional cheminformatics based on a prior cluster analysis of natural products space has allowed us to provide a set of small and relatively inexpensive libraries that may be useful in drug discovery and other assays (such as those seeking the substrates of orphan transporters).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Jose Garcia-Tenorio (Molport) for very helpful engagement with us in the development of these collections.

Funding

We thank the UK BBSRC (grant BB/P009042/1) for financial support.

References

1. Dobson PD, Kell DB. Carrier-mediated cellular uptake of pharmaceutical drugs: an exception or the rule? Nat Rev Drug Disc. 2008; 7:205–20.
2. Dobson PD, Patel Y, Kell DB. "Metabolite-likeness" as a criterion in the design and selection of pharmaceutical drug libraries. Drug Disc Today. 2009; 14:31–40.
3. Dobson P, Lanthaler K, Oliver SG, Kell DB. Implications of the dominant role of cellular transporters in drug uptake. Curr Top Med Chem. 2009; 9:163–84. [PubMed: 19200003]
4. Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KL, Chu X, et al. Membrane transporters in drug development. Nat Rev Drug Discov. 2010; 9(3):215–36. [PubMed: 20190787]
5. Kell DB, Dobson PD, Oliver SG. Pharmaceutical drug transport: the issues and the implications that it is essentially carrier-mediated only. Drug Disc Today. 2011; 16(15/16):704–14.
6. Kell DB, Dobson PD, Bilsland E, Oliver SG. The promiscuous binding of pharmaceutical drugs and their transporter-mediated uptake into cells: what we (need to) know and how we can do so. Drug Disc Today. 2013; 18(5/6):218–39.
7. Kell DB. Finding novel pharmaceuticals in the systems biology era using multiple effective drug targets, phenotypic screening, and knowledge of transporters: where drug discovery went wrong and how to fix it. FEBS J. 2013; 280:5957–80. [PubMed: 23552054]

Europe PMC Funders Author Manuscripts

8. Sugiyama, Y, Steffansen, B, editors. Transporters in Drug Development: Discovery, Optimization, Clinical Study and Regulation. New York (US): Springer; 2013.
9. Kell DB, Goodacre R. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. Drug Disc Today. 2014; 19(2):171–82.
10. Kell DB, Oliver SG. How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion. Front Pharmacol. 2014; 5:231. [PubMed: 25400580]
11. Winter GE, Radic B, Mayor-Ruiz C, Blomen VA, Trefzer C, Kandasamy RK, et al. The solute carrier SLC35F2 enables YM155-mediated DNA damage toxicity. Nat Chem Biol. 2014; 10:768–73. [PubMed: 25064833]
12. César-Razquin A, Snijder B, Frappier-Brinton T, Isserlin R, Gyimesi G, Bai X, et al. A call for systematic research on solute carriers. Cell. 2015; 162(3):478–87. [PubMed: 26232220]
13. Kell DB. What would be the observable consequences if phospholipid bilayer diffusion of drugs into cells is negligible? Trends Pharmacol Sci. 2015; 36(1):15–21. [PubMed: 25458537]
14. Mendes P, Oliver SG, Kell DB. Fitting transporter activities to cellular drug concentrations and fluxes: why the bumblebee can fly. Trends Pharmacol Sci. 2015; 36:710–23. [PubMed: 26538313]
15. O'Hagan S, Kell DB. The apparent permeabilities of Caco-2 cells to marketed drugs: magnitude, and independence from both biophysical properties and endogenite similarities. PeerJ. 2015; 3:e1405. [PubMed: 26618081]
16. Kell DB. Implications of endogenous roles of transporters for drug discovery: hitchhiking and metabolite-likeness. Nat Rev Drug Disc. 2016; 15(2):143–4.
17. Kell DB. How drugs pass through biological cell membranes–a paradigm shift in our understanding? Beilstein Magaz. 2016; 2(5)doi: 10.3762/bmag.5
18. Mooij MG, Nies AT, Knibbe CAJ, Schaeffeler E, Tibboel D, Schwab M, et al. Development of Human Membrane Transporters: Drug Disposition and Pharmacogenetics. Clin Pharmacokinet. 2016; 55(5):507–24. [PubMed: 26410689]
19. Govindarajan R, Sparreboom A. Drug Transporters: Advances and Opportunities. Clin Pharmacol Ther. 2016; 100(5):398–403. [PubMed: 27718234]
20. Grixti J, O’Hagan S, Day PJ, Kell DB. Enhancing drug efficacy and therapeutic index through cheminformatics-based selection of small molecule binary weapons that improve transporter-mediated targeting: a cytotoxicity system based on gemcitabine. Front Pharmacol. 2017; 8:155. [PubMed: 28396636]
21. Yee SW, Brackman DJ, Ennis EA, Sugiyama Y, Kamdem LK, Blanchard R, et al. Influence of Transporter Polymorphisms on Drug Disposition and Response: A Perspective From the International Transporter Consortium. Clin Pharmacol Ther. 2018; doi: 10.1002/cpt.1098
22. Tournier N, Stieger B, Langer O. Imaging techniques to study drug transporter function *in vivo*. Pharmacol Ther. 2018; 189:104–22. [PubMed: 29684469]
23. Dickens D, Rädisch S, Chiduza GN, Giannoudis A, Cross MJ, Malik H, et al. Cellular uptake of the atypical antipsychotic clozapine is a carrier-mediated process. Mol Pharm. 2018; 15(8):3557–72. [PubMed: 29944835]
24. Wang Y, Moussian B, Schaeffeler E, Schwab M, Nies AT. The fruit fly *Drosophila melanogaster* as an innovative preclinical ADME model for solute carrier membrane transporters, with consequences for pharmacology and drug therapy. Drug Discov Today. 2018:1746–60. [PubMed: 29890226]
25. Kermani AA, Macdonald CB, Gundepudi R, Stockbridge RB. Guanidinium export is the primal function of SMR family transporters. Proc Natl Acad Sci U S A. 2018; 115(12):3060–5. [PubMed: 29507227]
26. César-Razquin A, Girardi E, Yang M, Brehme M, Sáez-Rodríguez J, Superti-Furga G. *In silico* prioritization of transporter-drug relationships from drug sensitivity screens. bioRxiv [Preprint]. 2018; doi: 10.1101/381335
27. Kell DB, Wright Muelas M, O’Hagan S, Day PJ. The role of drug transporters in phenotypic screening. Drug Target Rev. 2018; 4:16–9.
28. Yang NJ, Hinner MJ. Getting across the cell membrane: an overview for small molecules, peptides, and proteins. Meth Mol Biol. 2015; 1266:29–53.

29. Jindal S, Yang L, Day PJ, Kell DB. Involvement of multiple influx and efflux transporters in the accumulation of cationic fluorescent dyes by *Escherichia coli*. BMC Microbiol. 2019; doi: 10.1101/603688
30. Kell DB. The transporter-mediated cellular uptake of pharmaceutical drugs is based on their metabolite-likeness and not on their bulk biophysical properties: Towards a systems pharmacology. Perspect Sci. 2015; 6:66–83.
31. Brindefalk B, Dessailly BH, Yeats C, Orengo C, Werner F, Poole AM. Evolutionary history of the TBP-domain superfamily. Nucleic Acids Res. 2013; 41(5):2832–45. [PubMed: 23376926]
32. O'Hagan S, Swainston N, Handl J, Kell DB. A 'rule of 0.5' for the metabolite-likeness of approved pharmaceutical drugs. Metabolomics. 2015; 11(2):323–39. [PubMed: 25750602]
33. O'Hagan S, Kell DB. Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. Front Pharmacol. 2015; 6:105. [PubMed: 26029108]
34. O'Hagan S, Kell DB. MetMaxStruct: a Tversky-similarity-based strategy for analysing the (sub)structural similarities of drugs and endogenous metabolites. Front Pharmacol. 2016; 7:266. [PubMed: 27597830]
35. O'Hagan S, Kell DB. Analysis of drug-endogenous human metabolite similarities in terms of their maximum common substructures. J Cheminform. 2017; 9:18. [PubMed: 28316656]
36. O'Hagan S, Kell DB. Consensus rank orderings of molecular fingerprints illustrate the 'most genuine' similarities between marketed drugs and small endogenous human metabolites, but highlight exogenous natural products as the most important 'natural' drug transporter substrates. ADMET DMPK. 2017; 5(2):85–125.
37. O'Hagan S, Kell DB. Analysing and navigating natural products space for generating small, diverse, but representative chemical libraries. Biotechnol J. 2018; 13(1)
38. Karakoc E, Sahinalp SC, Cherkasov A. Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds. J Chem Inf Model. 2006; 46(5):2167–82. [PubMed: 16995747]
39. Gupta S, Aires-de-Sousa J. Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness. Mol Divers. 2007; 11(1):23–36. [PubMed: 17447158]
40. Khanna V, Ranganathan S. Physicochemical property space distribution among human metabolites, drugs and toxins. BMC Bioinformatics. 2009; 10(Suppl 15):S10.
41. Peironcely JE, Reijmers T, Coulier L, Bender A, Hankemeier T. Understanding and classifying metabolite space and metabolite-likeness. PLoS One. 2011; 6(12):e28966. [PubMed: 22194963]
42. Hamdalla MA, Mandoiu II, Hill DW, Rajasekaran S, Grant DF. BioSM: Metabolomics Tool for Identifying Endogenous Mammalian Biochemical Structures in Chemical Structure Space. J Chem Inf Model. 2013; 53(3):601–12. [PubMed: 23330685]
43. Nigam SK. What do drug transporters really do? Nat Rev Drug Discov. 2015 Jan; 14(1):29–44. [PubMed: 25475361]
44. Kell DB, Kaprelyants AS, Grafen A. On pheromones, social behaviour and the functions of secondary metabolism in bacteria. Trends Ecol Evol. 1995; 10:126–9. [PubMed: 21236981]
45. Kell DB, Swainston N, Pir P, Oliver SG. Membrane transporter engineering in industrial biotechnology and whole-cell biocatalysis. Trends Biotechnol. 2015; 33:237–46. [PubMed: 25746161]
46. Kell DB. Control of metabolite efflux in microbial cell factories: current advances and future prospects. OSF Preprints. 2018; doi: 10.31219/osf.io/xg9jh
47. Chapy H, Smirnova M, Andre P, Schlatter J, Chiadmi F, Couraud PO, et al. Carrier-mediated cocaine transport at the blood-brain barrier as a putative mechanism in addiction liability. Int J Neuropsychopharmacol. 2014; 18(1):1–10.
48. Gründemann D, Harlfinger S, Golz S, Geerts A, Lazar A, Berkels R, et al. Discovery of the ergothioneine transporter. Proc Natl Acad Sci U S A. 2005; 102(14):5256–61. [PubMed: 15795384]
49. Kerley RN, McCarthy C, Kell DB, Kenny LC. The potential therapeutic effects of ergothioneine in pre-eclampsia. Free Radic Biol Med. 2018; 117:145–57. [PubMed: 29284116]

Europe PMC Funders Author Manuscripts

50. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature. 2018; 557(7706):503–9. [PubMed: 29769716]

51. Harper G, Pickett SD, Green DVS. Design of a compound screening collection for use in high throughput screening. Comb Chem High Throughput Screen. 2004; 7(1):63–71. [PubMed: 14965262]

52. Grabowski K, Baringhaus KH, Schneider G. Scaffold diversity of natural products: inspiration for combinatorial library design. Nat Prod Rep. 2008; 25(5):892–904. [PubMed: 18820757]

53. Cragg GM, Newman DJ. Natural products: A continuing source of novel drug leads. Biochim Biophys Acta. 2013; 1830(6):3670–95. [PubMed: 23428572]

54. Laraia L, Waldmann H. Natural product inspired compound collections: evolutionary principle, chemical synthesis, phenotypic screening, and target identification. Drug Discov Today Technol. 2017; 23:75–82. [PubMed: 28647090]

55. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov. 2017; 16(8):531–43. [PubMed: 28685762]

56. Laraia L, Robke L, Waldmann H. Bioactive Compound Collections: From Design to Target Identification. Chem. 2018; 4(4):705–30.

57. Mazanetz MP, Marmon RJ, Reisser CBT, Morao I. Drug discovery applications for KNIME: an open source data mining platform. Curr Top Med Chem. 2012; 12(18):1965–79. [PubMed: 23110532]

58. O'Hagan S, Kell DB. The KNIME workflow environment and its applications in Genetic Programming and machine learning. Genetic Progr Evol Mach. 2015; 16:387–91.

59. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform. 2013; 5(1):26. [PubMed: 23721588]

60. Gu JY, Gui YS, Chen LR, Yuan G, Lu HZ, Xu XJ. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. PLoS One. 2013; 8(4):e62839. [PubMed: 23638153]

61. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Machine Learning Res. 2008; 9:2579–605.

62. Breiman L. Random forests. Machine Learning. 2001; 45(1):5–32.

63. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95(25):14863–8. [PubMed: 9843981]

64. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, et al. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci U S A. 2005; 102(48):17272–7. [PubMed: 16301544]

65. Rosen J, Gottfries J, Muresan S, Backlund A, Oprea TI. Novel chemical space exploration via natural products. J Med Chem. 2009; 52(7):1953–62. [PubMed: 19265440]

66. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015; 14(2):111–29. [PubMed: 25614221]

67. Skinnider MA, Magarvey NA. Statistical reanalysis of natural products reveals increasing chemical diversity. Proc Natl Acad Sci U S A. 2017; 114(31):E6271–2. [PubMed: 28710332]

68. Chen Y, Garcia de Lomana M, Friedrich NO, Kirchmair J. Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. J Chem Inf Model. 2018; 58(8):1518–32. [PubMed: 30010333]

69. Reymond JL. The Chemical Space Project. Acc Chem Res. 2015; 48(3):722–30. [PubMed: 25687211]

70. Sterling T, Irwin JJ. ZINC 15-Ligand Discovery for Everyone. J Chem Inf Model. 2015; 55:2324–37. [PubMed: 26479676]

71. Currin A, Swainston N, Day PJ, Kell DB. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. Chem Soc Rev. 2015; 44(5):1172–239. [PubMed: 25503938]

72. Kell DB, Lurie-Luke E. The virtue of innovation: innovation through the lenses of biological evolution. J R Soc Interface. 2015; 12(2)

73. Clark RD, Kar J, Akella L, Soltanshahi F. OptDesign: extending optimizable k-dissimilarity selection to combinatorial library design. J Chem Inf Comput Sci. 2003; 43(3):829–36. [PubMed: 12767140]

74. Lundholt BK, Scudder KM, Pagliaro L. A simple technique for reducing edge effect in cell-based assays. J Biomol Screen. 2003; 8(5):566–70. [PubMed: 14567784]

75. Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R. Statistical practice in high-throughput screening data analysis. Nat Biotechnol. 2006; 24(2):167–75. [PubMed: 16465162]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

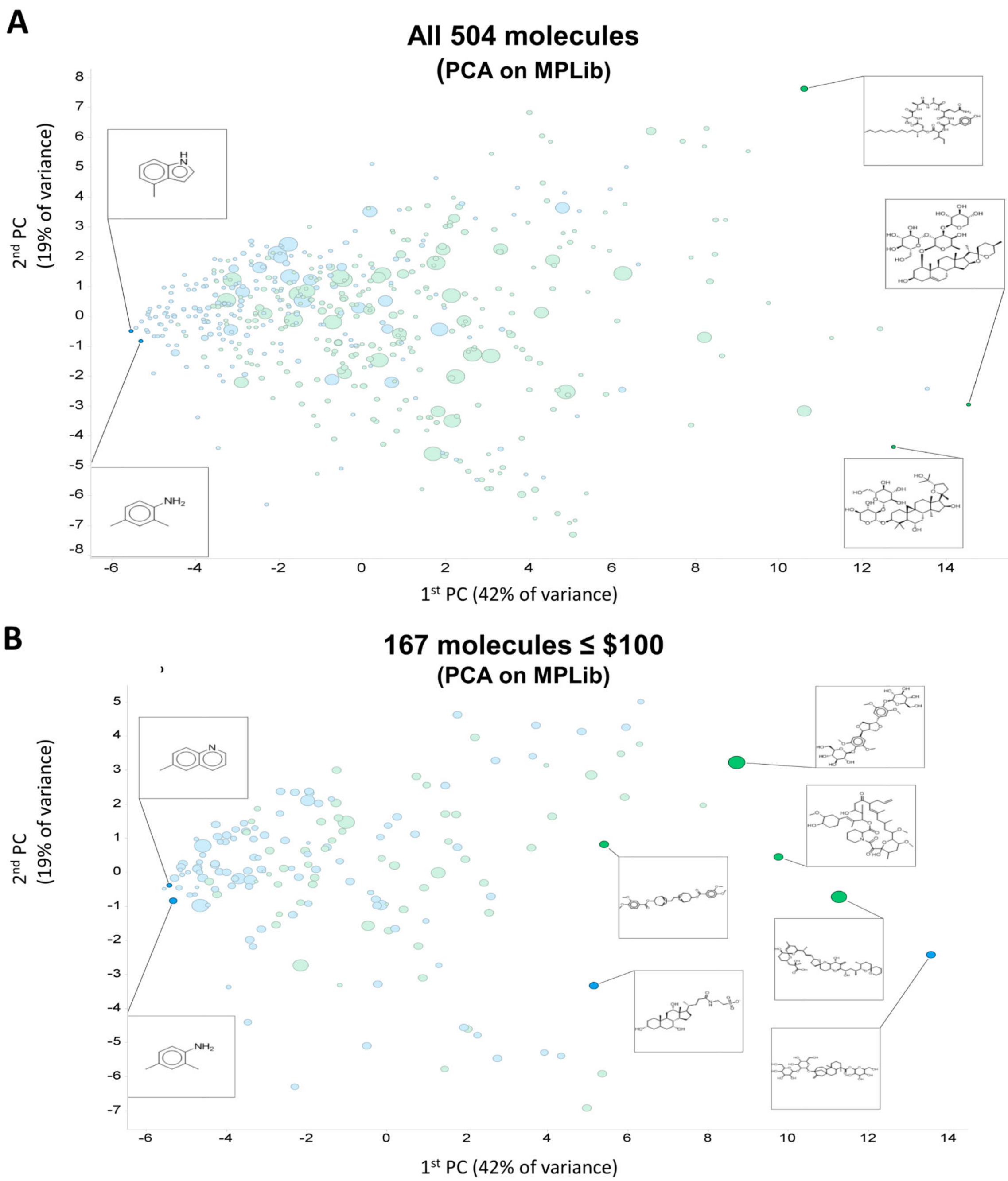

Figure 1.

Initial coverage of natural products space as described in the text. Exact matches to cluster centres (blue) or nearby isomers (green) available in the Molport collection are labelled accordingly. Price is encoded via symbol size from $10 to $5713. (**A**) Full set of 504 molecules. (**B**) Reduced set of 167 molecules costing $100 or less.

Europe PMC Funders Author Manuscripts

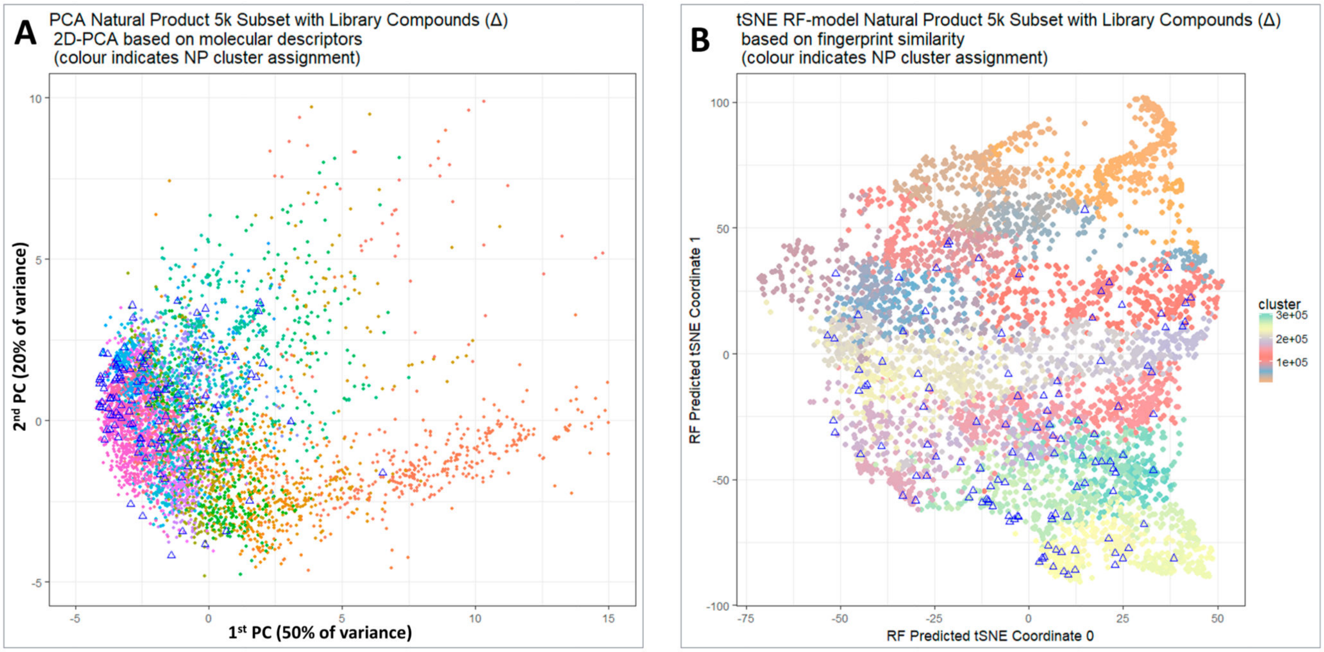

Figure 2.

(**A**) Visualisation of the coverage of natural product(-like) space when molecules are selected from individual clusters. Principal components analysis was performed after normalizing to unit variance using a standard KNIME workflow. 5000 molecules are shown for purposes of visualization, and the 118 molecules closest to cluster centres that fulfilled our other criteria are indicated with triangles. (**B**) A t-SNE plot of the same data as in Figure 2A.

Europe PMC Funders Author Manuscripts

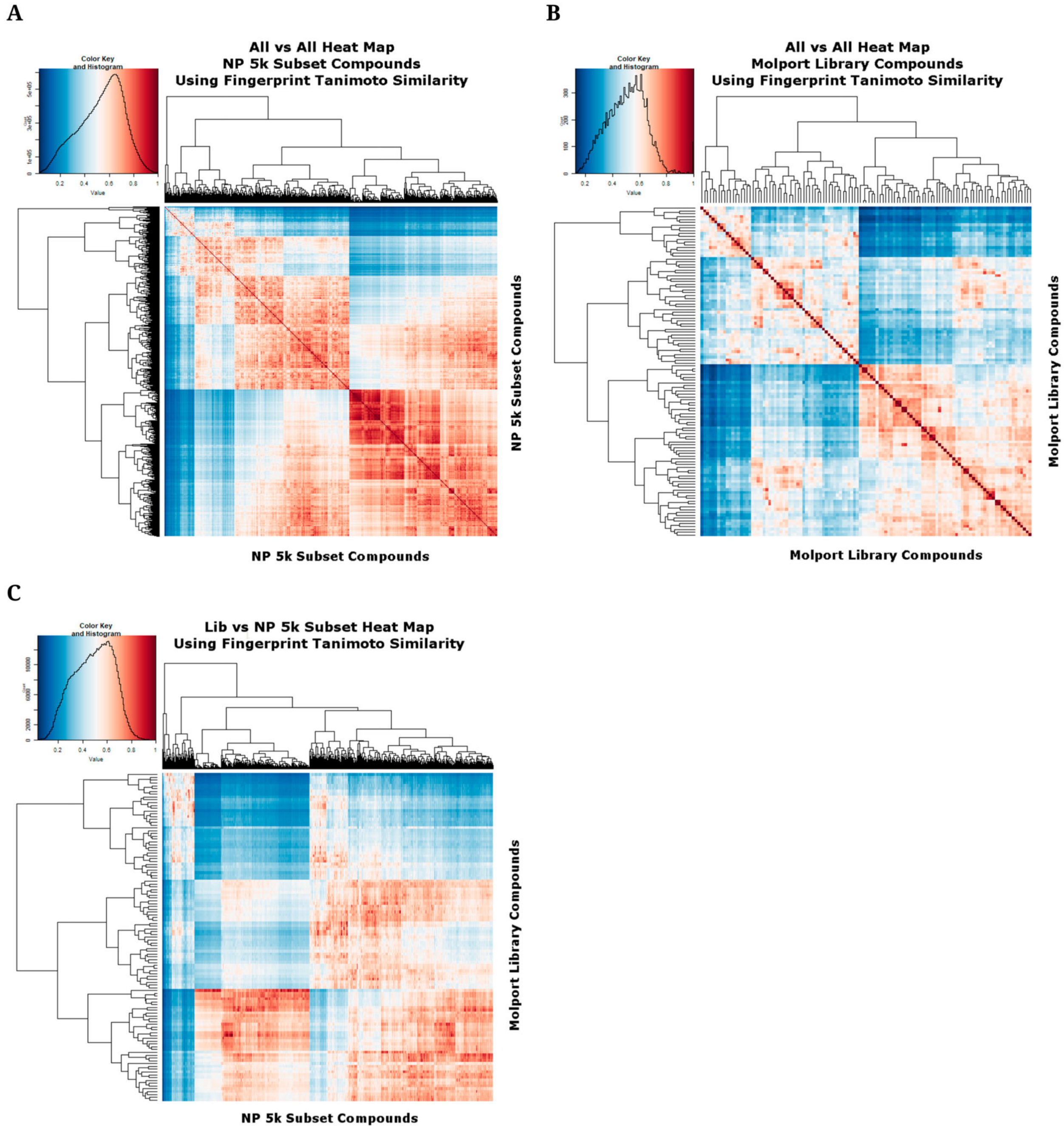

Figure 3.

Heat map analyses of (**A**) The 5000-molecule subset and (**B**) The 117-molecule subset, based on their Tanimoto similarities. The analyses used the same workflows as those described in [32]. (**C**) The 5000-molecule subset versus the 117-molecule subset, based on their Tanimoto similarities. The analyses used the same workflows as those described in [32].

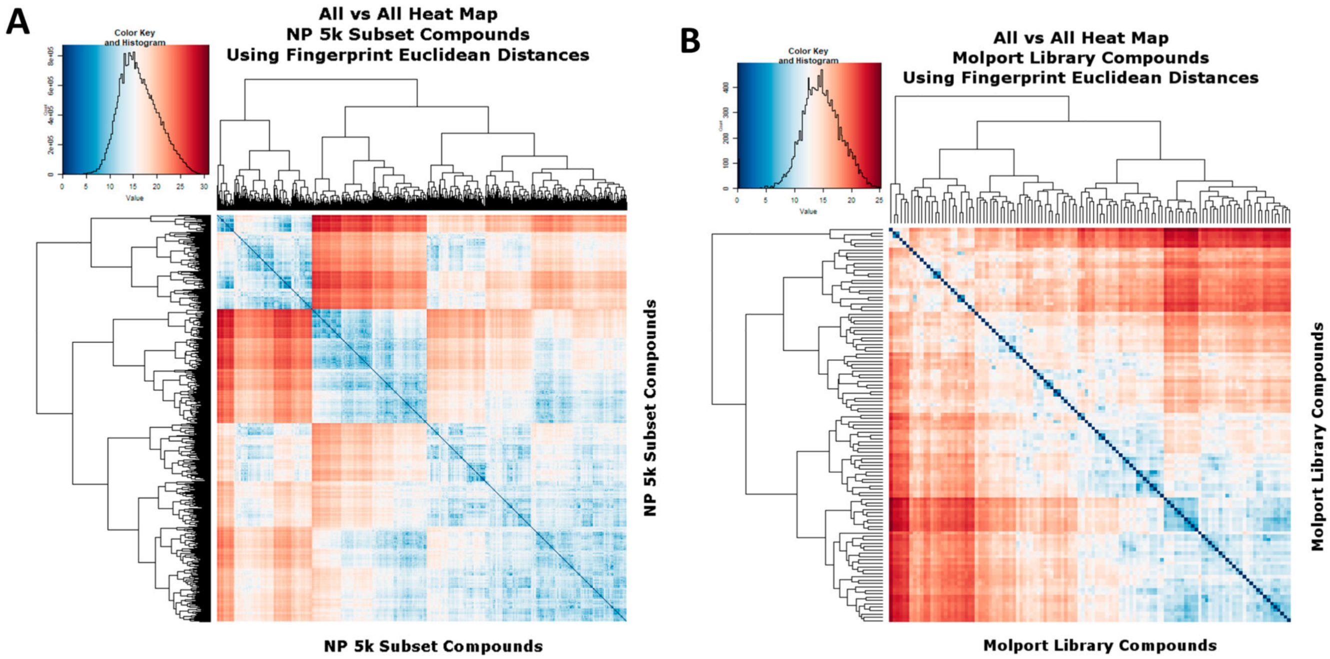

Figure 4.

Heat map analyses of (**A**) The 5000-molecule subset, based on their Euclidean distances, as described in the text. Analyses and displays were otherwise as per Figures 2 and 3. (**B**) The 117-molecule subset, based on their Euclidean distances, as described in the text. Analyses and displays were otherwise as per Figures 2 and 3.

Europe PMC Funders Author Manuscripts

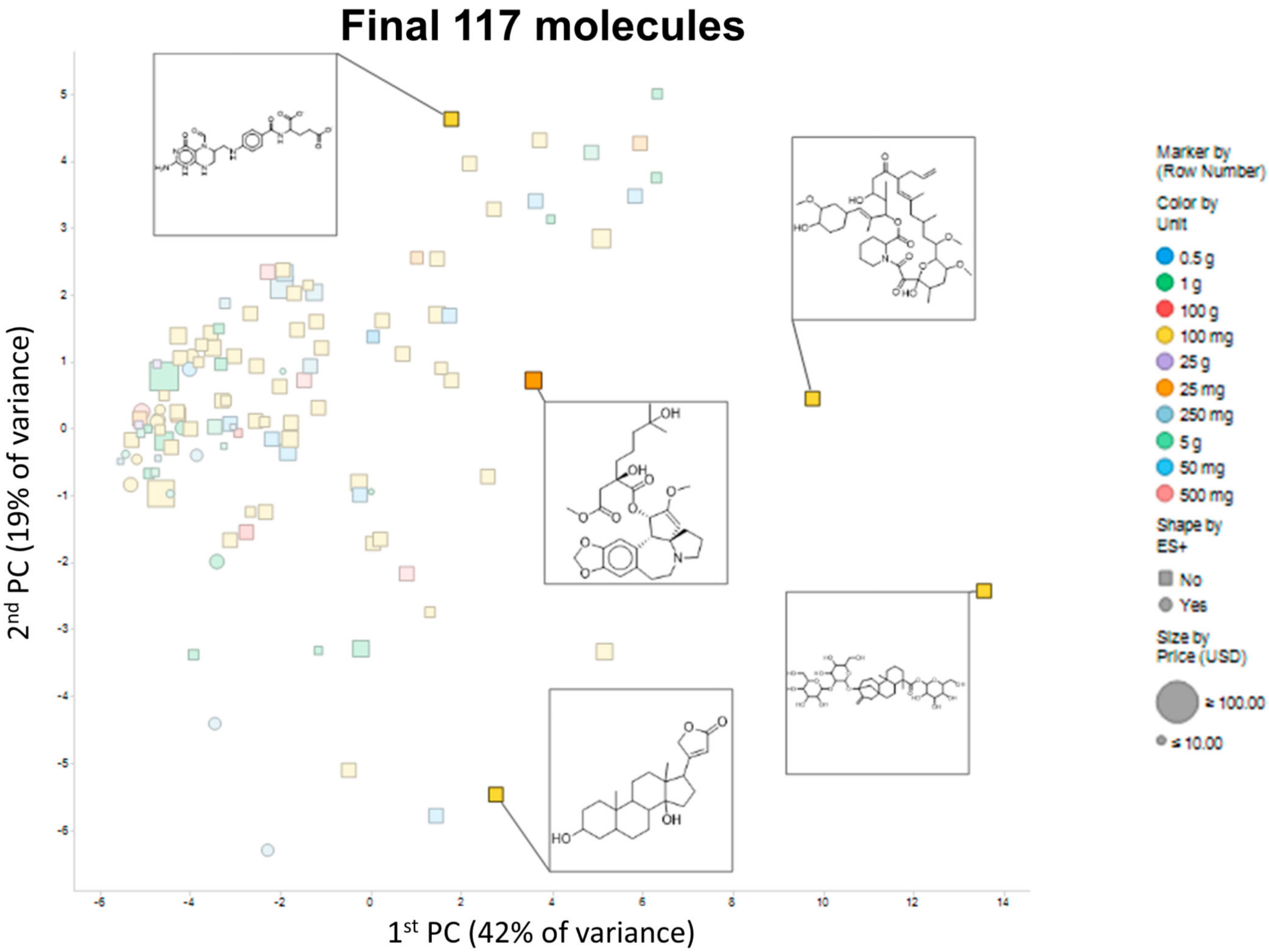

Figure 5.
PCA plot of the 117-molecule subset, showing 5 representative molecules.