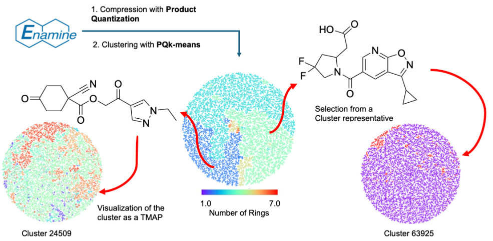
Nested Tree-maps to visualize Billions of Molecules

Alejandro Flores Sepúlveda [a], and Jean-Louis Reymond [a]*

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

e-mail: *jean-louis.reymond@unibe.ch*


1. Compression with **Product Quantization**
2. Clustering with **PQk-means**
Enamine
Selection from a Cluster representative
Visualization of the cluster as a TMAP
1.0
7.0
Number of Rings
Cluster 24509
Cluster 63925

Abstract

Here we present a visualization framework enabling the exploration of billion-sized chemical datasets, exemplified with the REAL database of 9.6 billion make-on-demand molecules. We first organize the dataset in 100,000 clusters by applying Product Quantization (PQ) and PQk-means to the 42-dimensional vectors of MQNs (molecular quantum numbers) describing molecular structures with counts for different atom- and bond-types, polar groups and topological features). We retrieve the molecule closest to cluster center as a representative of each cluster and compute a tree-map (TMAP) displaying these cluster centers organized by MQN-similarity in an approximate nearest neighbor minimum spanning tree. Each cluster center in this primary TMAP is linked to a nested secondary TMAP displaying the corresponding cluster content organized by substructure fingerprint similarity. This nested TMAP approach can be computed on a single workstation and gives direct access to the entire dataset down to single molecular structures in two clicks.

Keywords. clustering, chemical space, product quantization, k-means

Introduction

Advances in cheminformatics and data science are transforming chemical space from a purely conceptual idea into actual and increasingly large molecular datasets, typically billion-sized libraries of screening compounds to support drug discovery.[1–4] Unfortunately, these datasets are difficult to analyze and understand because clustering algorithms as well as visualization tools are typically limited to at most a few million molecules. This limitation applies in particular to dimensionality reduction methods for high-dimensional molecular fingerprint representations,[5] such as Principal Component Analysis (PCA)[6] and the related similarity maps,[7–10] Self-organizing Maps (SOM),[11] generative topographic mapping,[12] t-Distributed Stochastic Neighbor Embedding (t-SNE),[13] Uniform Manifold Approximation and Projection (UMAP),[14] tree-maps (TMAP),[15] and other tools.[7,8,10,16–21]

In our previously reported TMAP visualization tool, we computed an approximate nearest-neighbor graph of a dataset using locality sensitive hashing on vector representations of dataset objects, reduced it to the corresponding minimum spanning tree, and computed a 2D-layout of this minimum spanning tree where each node represents a different object, typically a molecule, and each edge represents an approximate nearest neighbor connection.[15] The TMAP was displayed in an interactive format in a web browser using the application Faerun and Smilesdrawer, in which each molecule can be viewed individually.[22,23] Unfortunately, the TMAP code only runs well for datasets of up to a few million molecules, and the display is only convenient up to ~500,000 molecules because larger datasets tend to be overwhelming.

To adapt TMAP to billion-sized datasets, we now report a “nested TMAP” tool consisting of a primary TMAP displaying similarity relationships between a set of cluster representatives and linked secondary TMAPs displaying corresponding cluster contents. This setup makes it possible to navigate a dataset of billions of molecules interactively from a global overview display in the primary TMAP down to each single molecule within each secondary TMAP.

To form meaningful clusters from the 9.6 billion molecule dataset, we represent the molecules with MQN (Molecular Quantum Numbers), a 42 dimensional count vector describing atom and bond types, polar groups and topological features, and which we had previously found useful to create maps of large molecular datasets by PCA.[27–30] We then reduce the dimensionality of the 42-D MQN vector to 6-D by Product Quantization (PQ) and form clusters using PQk-means, which is a computationally efficient method for very large datasets including molecular datasets,[31,32] and select the molecule closest to cluster center as a representative for each cluster to be featured in the primary TMAP. This primary TMAP and the secondary TMAPs displaying the corresponding cluster contents are displayed in an interactive web-based format supported by the applications Faerun and Smilesdrawer.[22,23] Note that clustering via PQ is computationally efficient and delivers homogeneously sized clusters and computed on a single workstation. This method is therefore more efficient than the recently proposed Bit-BIRCH method requiring high-performance computing and generating clusters of varying size and numbers.[33,34]

Methods

Molecular Data Preparation

The Enamine REAL Database (9.6 billion structures) was downloaded as SMILES strings from the Enamine website in March 2025. To remove ordering effects arising from the default heavy-atom-sorted input, the file was first randomly permuted. Molecular Quantum Number (MQN) fingerprints were then computed for every molecule using the RDKit implementation of the 42

MQN indices (RDKit version XXXX) via rdMolDescriptors module without further preprocessing. These were then used as input for PQ encoding.

Product Quantization (PQ)

To reduce the memory footprint of the 42-dimensional MQN vectors sufficiently to allow the clustering of billions of molecules on a single workstation, the vectors were compressed using Product Quantization (PQ). PQ was implemented directly in Python following the formulation in Jégou et al. [31] and the procedure described in Matsui et al. [35]. Each MQN vector $x \in \mathbb{R}^{42}$ was decomposed into $M = 6$ contiguous subvectors of dimension seven, following RDKit index order. For each of the six subspaces, a separated codebook of $L = 256$ centroids was learned by k-means clustering using k-means++ initialization, Euclidean distance (L2) and 20 iterations. Codebook training was performed on a random subsample of 50 million MQN vectors.

After training, all 9.6 billion MQN vectors were reloaded in consecutive shards and encoded into PQ codes. For each molecule, each of the six subvectors was assigned to its nearest centroid by Euclidean distance, yielding a six-integer tuple $(c_1, c_2, \dots, c_6)$ where each $c_i \in \{0, \dots, 255\}$ indexes a codeword in corresponding subspace. PQ reconstruction fidelity was assessed by converting PQ codes back to approximate MQN vectors and computing the Pearson coefficient on a random sample of one million molecules, which yielded $r = 0.99 \pm 0.00$.

Commented [FSA(1]: I don't think Person Coef. is the metrix to be used here...

All computations were performed on a single workstation (AMD Ryzen 7 8700F, 64 GM RAM). PQ codebook training (50 million MQNs) required approximately 70 minutes. Encoding the full dataset into PQ codes in streaming mode required approximately 3 hours, dominated by disk I/O.

Commented [FSA(2]: Not definitive

Clustering Using PQk-means

Clustering of all PQ codes into 100,000 clusters was performed using the original C++ implementation of PQk-means provided by Matsui et al. (https://github.com/DwangoMediaVillage/pqkmeans). PQk-means operates directly on PQ codes and uses the symmetric distance (SD), a lookup-table based approximation of Euclidean distance. Clustering proceeds in two stages, following the recommend PQk-means workflow. First, the model was trained on a uniform random subsample of one billion PQ codes. Second, the remaining PQ codes were assigned to the nearest of the 100,000 learned centroids in streaming batches. Assignment was performed with the same SD metric. For each cluster, the representative molecule was defined as the element whose PQ code had minimal SD distance to the cluster centroid. Training one billion PQ codes required 2 days and 5 hours on the workstation described above. Full-dataset assignment after training required 4 hours.

Commented [FSA(3]: Not definitive

Visualization Using TMAP

A primary TMAP was constructed from the 100,000 representative molecules. For each representative, its original 42-dimensional MQN vector was used as the embedding. An approximate k-nearest neighbor graph was computed and used to create the TMAP. Nodes were colored by structural or physicochemical MQN features (e.g. number of rings, heavy atom count, fraction of aromatic atoms). For each cluster, a secondary TMAP was constructed using the molecules belonging to that cluster. These TMAPs were generated from ECFP4 fingerprints.

Results and Discussion

We exemplify with the 9.6 billion molecules in the REAL database made available by Enamine Ltd., which is a subset of a larger dataset of 76.9 billion products possible by combining a set of building blocks via known reactions, filtering by the drug-likeness criteria Lipinski's Rule of Five and Veber's guidelines (MW: ≤ 500, SlogP ≤ 5, HBA ≤ 10, HBD ≤ 5, Rotatable Bonds ≤ 10, and TPSA ≤ 140).[24,25] This combinatorial dataset is typical of the large libraries currently in use in early phase drug discovery. dataset of make-on-demand molecules.[26] The nested TMAP example realized here is accessible at https://chelombus.gdb.tools.

The overall workflow consists of five computational stages: MQN calculation from SMILES, PQ codebook training, PQ encoding of all MQNs into PQ codes, clustering of PQ codes by PQk-means to obtain the cluster assignments and representatives, and visualization by a primary TMAP over cluster representatives with nested secondary TMAPs for cluster contents.

Because PQk-means clustering is performed in PQ-code space using the symmetric distance (SD), it is important that SD preserves the similarity relationships of the original MQN vectors. The relationship between Euclidean distances in the original MQN space and SD distances between the corresponding PQ codes is shown in Figure 1, illustrating that SD servers a practical proxy for MQN-based similarity.

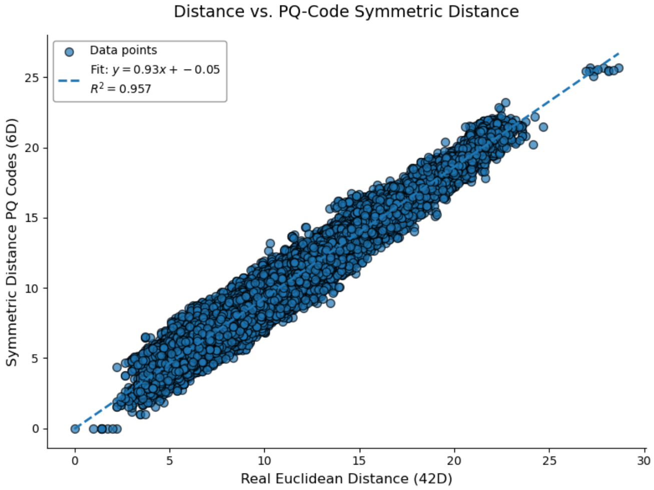


Figure 1. Plot showing the relation between the Real Euclidean distance between a set of MQN vectors (42D) and the SD distance between their corresponding PQ codes (6D) to a reference molecule.

Clustering with $k = 100{,}000$ yields was performed using the PQk-means implementation from Matsui et al.[35] operating directly on PQ codes and the associated symmetric distance approximation to Euclidean distance. Trained centroids were then used to assign the rest of PQ codes to clusters in streaming mode. In practice, not all learned centroids are populated after assignment, and the distribution of populated cluster sizes is shown in Figure 2. Of the 100,000 centroids, 92,434 clusters were populated after assignment. For each populated cluster, the representative molecule was defined as the molecule nearest to the cluster center in PQ-code space and used as the node displayed in the primary TMAP. The primary TMAP was generated from the cluster representatives. Its computation was fast compared to clustering and required on the order of one minute. Secondary TMAPs were generated for cluster contents using ECFP4 fingerprints. Generating a single secondary TMAP required approximately one minute or less. Precomputing secondary TMAPs for all clusters therefore represents a substantial computational effort on a single workstation. In practice, secondary TMAPs can be computed on demand for clusters of interest or generated in parallel on a computing cluster, depending on whether the goal is interactive web deployment or offline analysis.

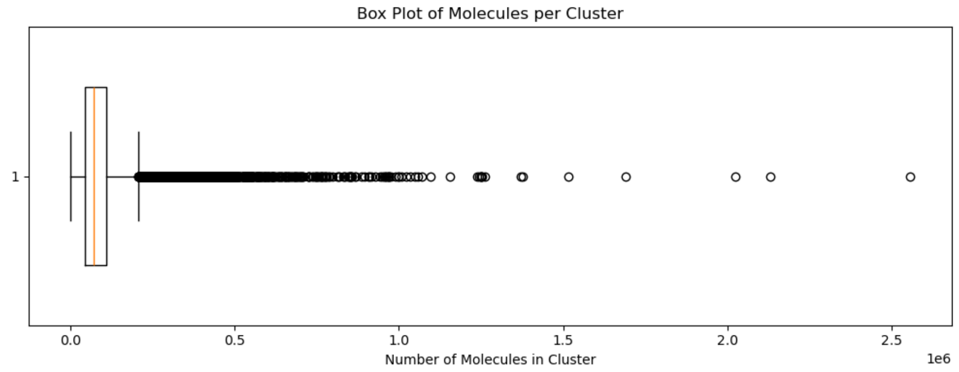


Figure 2. Box plot of the number molecules per cluster

To further evaluate the quality of our clusters we also measured the dispersion statistics. Table 1 summarize the degree of internal consistency across our clusters using nine molecular descriptors that were chosen to capture a spectrum of chemical properties. For example, Molecular Weight (MW) and Heavy Atom Count (HAC) were selected to assess cluster homogeneity in terms of molecular size. Similarly, Polarity was evaluated through descriptors like Total Polar Surface Area (TPSA), Hydrogen Bond Acceptors (HBA), Hydrogen Bond Donors (HBD). It should also be noted that some of these descriptors (MW, TPSA, HBA and HBD) have previously been employed effectively to compare property space coverage between virtual libraries and reference databases[36,37].

Commented [FSA(4]: cites 36,37 probably need to be re-indexed

For each descriptor three complementary measures were computed for every cluster: Range, Inter-Quartile ($IQR = 3 - 1$) and Coefficient of Variation ($CV = \frac{\sigma}{\mu}$). Molecular Weight and HAC showed low median CV (~5% and 3.5%, respectively), indicating that clusters are highly size-homogeneous. TPSA and fraction cSP3 showed somewhat larger variability, meaning some clusters contain both very aromatic and highly aliphatic molecules; however, these descriptors still indicate that polar surface area is largely conserved intracluster for most clusters. For ring count, rotatable bonds, HBD (IQR = 0) and HBA (IQR=1) where the median Inter-Quantile is low, the longer upper tails in CV (e.g. aromatic atom $CV\ \approx 6$) are probably artefacts of small means which is why IQR and Range are probably more reliable here. Only 5% of clusters span less than 20 aromatic atoms or less than 8 rotatable bonds, again pinpointing obvious "mis-clustering" candidates.

Commented [FSA(5]: This is my interpretation of the table. It could be obviously wrong or better explained in other ways. We can definitively discuss this more in detail or present it in a different way.

Descriptor	CV 5th %; Median; 95th %	IQR 5th %; Median; 95th %	Range 5th %; Median; 95th %
Molecular Weight	0.03; 0.05; 0.1	14; 24; 44	112; 159; 254
Heavy Atom Count	0.02; 0.03; 0.06	0; 1; 2	4; 7; 11
Number of Rings	0.02; 0.04; 0.06	0; 1; 2	4; 7; 11
Rotatable Bonds	0.06; 0.12; 0.25	0; 1; 2	2; 5; 8
Hydrogen Bond Donors	0.08; 0.32; 7.92	0; 0; 1	1; 3; 4
Hydrogen Bond Acceptors	0.12; 0.16; 0.28	0; 1; 2	4; 6; 7
Topological Polar Surface Area	0.08; 0.11; 0.22	5.1; 12.6; 18	50; 69.5; 90.1
Fraction of sp3 carbons	0.02; 0.09; 0.31	0; 0; 1	1; 2; 3
Aromatic Atom Count	0.17; 0.33; 5.8	0; 4; 6	6; 12; 20

Table 1. Median, 5th and 95th percentiles for the Covariance (CV), Interquartile Range (IQR) and Range values calculated from the clusters values for different molecular descriptors.

In addition, within-cluster similarity was compared to between-cluster similarity by computing distances from random molecule pairs sampled within and across clusters. Table 2 reports mean within-cluster versus between-cluster distances using Euclidean and Manhattan distances on MQN fingerprints and SD on PQ codes, showing that molecules assigned to the same cluster are, on average, substantially closer than molecules drawn from different clusters.

Metric	Within Clusters (Mean ± SD)	Between Clusters (Mean ± SD)
Euclidean Distance	4.27 ± 1.63	11.61 ± 3.06
Manhattan Distance	16.79 ± 8.15	50.14 ± 14.49
Symmetric Distance	4.68 ± 2.63	12.5 ± 3.54

Table 2. Euclidean, Cosine and Manhattan Distance and Tanimoto Similarity on ECFP Fingerprints from 10 randomly selected molecules from 100 randomly selected clusters.

The same effect is visualized as pairwise-distance heatmaps in Figure 7, where the within-cluster blocks along the diagonal display systematically lower distances than off-diagonal blocks. Because the primary aim of this work is to show that very large collections of molecules can be clustered and visualized on commodity hardware, we deliberately adopted a minimalist set-up: we decided to use the standard MQN fingerprint, which part of the reason why larger count values like HAC, Number of Carbons or Cyclic Single Bonds (all descriptors in MQN) dominate the clustering process and thus a lower covariance is observed in the clusters for these features. A normalized or weighted version of MQN could point towards a different result in the clustering process. Also, PQk-means was applied with k fixed a-priory to yield a map-friendly granularity (100.000 clusters). Despite the absence of hyper-parameter optimization or MQN weighting, the clusters exhibit good intra-group homogeneity across the nine physicochemical descriptors (median $CV \leq 0.20$ for all, and ≤ 0.06 for size-related properties) and a lower within-cluster distance across several distances metrics.

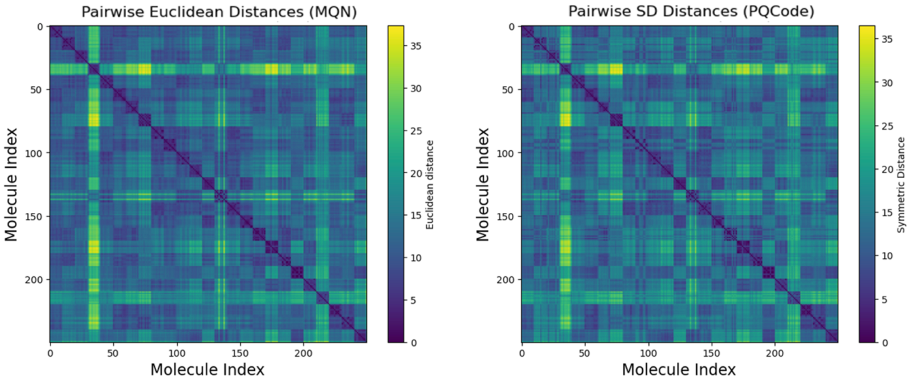


Figure 3. Heatmap for the pairwise Euclidean Distances between MQN fingerprints and Symmetric Distance (approximated Euclidean) for the PQ Codes for 10 randomly selected molecules from 25 randomly selected clusters. The diagonal squares correspond to within-cluster distances whereas the rest of squares are between-cluster distances.

For visualization, the primary TMAP was constructed from the representative molecule of each populated cluster, selected as the molecule nearest the cluster centroid in PQ-code space. Coloring the primary TMAP by representative properties reveals coherent subregions associated with MQN-derived structural trends. This is illustrated in Figure 4, where different areas of the map correspond to gradients in features such as ring count, aromaticity, and size-related descriptors.

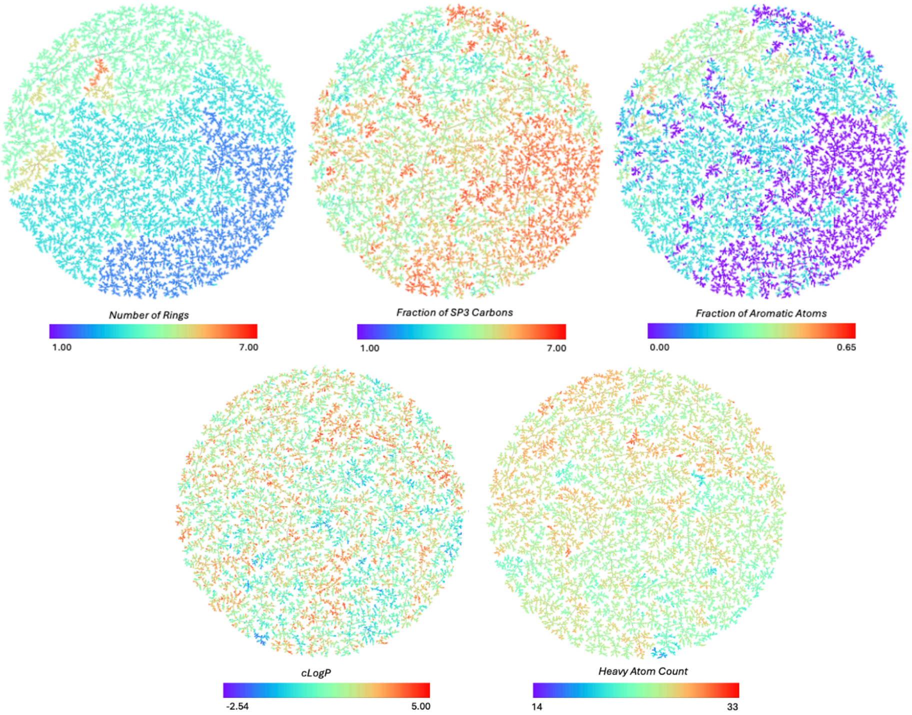


Figure 4. Primary TMAP (each node is a cluster) colored by each cluster representative Number of Rings, Fraction of SP3 Carbons, Fraction of Aromatic Atoms, cLogP and Heavy Atom Count.

Commented [FSA(6]: I think this looks better than the current TMAP color/node scheme we have on the website where it's harder to distinguish the subareas. (https://chelombus.gdb.tools/maps/enamine_representative_TMAP.html)

At least for the figures we could consider changing to as they are in Figure 4 (same TMAP as the website but different color/node). It's up to you. I don't think its a good idea to not have a consistent layout for the paper and the website beacuse it can lead the reviewers to confussion.

Each node of the primary TMAP links to a secondary TMAP representing the corresponding cluster contents. Secondary TMAPs are typically homogeneous in the MQN features used for clustering, while still resolving substructure-based organization when constructed using substructure fingerprints. Representative examples are shown in Figure 5. Inspection of individual clusters further illustrates that clusters correspond to chemically interpretable motif families, as exemplified by Figure 6 and Figure 7.

Commented [FSA(7]: We can also include a Figure showing individual cluster TMAPs.

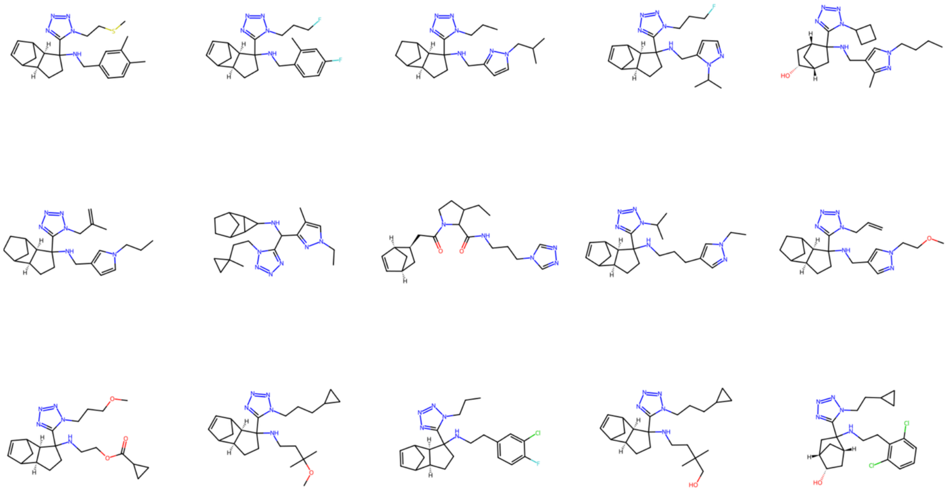


Figure 11. 12 randomly selected molecules for the cluster 15345.

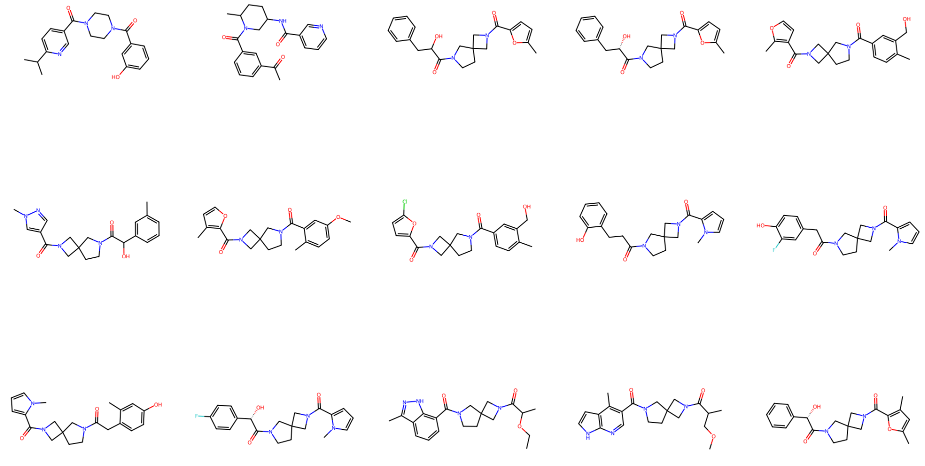


Figure 12. 12 randomly selected molecules for the cluster 1184

This nested representation provides a global-to-local navigation scheme: the primary map summarizes relationships between cluster representatives, and secondary maps enable direct inspection of the underlying molecular content without attempting to render the full dataset simultaneously.

Commented [FSA(8]: Probably remove. Repeating too much...

Conclusion

We report a nested TMAP framework enabling interactive exploration of billion-sized chemical datasets on commodity hardware. The method combines an interpretable, low-dimensional count fingerprint (MQN) with product quantization and PQk-means clustering to organize an ultra large chemical space into a set of cluster representatives. Visualization is achieved by a primary TMAP describing relationships between cluster representatives, where each node links to a nested secondary TMAP displaying the corresponding cluster content. The resulting representation provides a global overview of chemical diversity while preserving access to individual molecular structures, enabling navigation of a 9.6 billion scale space in a manner that is both computationally feasible and chemically interpretable.

Data availability

All code written for PQ implementation, TMAP generation, and preprocessing is available along with trained models and tutorials for using the tools can be found at https://github.com/afloresep/chelombus-package. The datasets used to train the models can be downloaded from https://enamine.net/compound-collections/real-compounds/real-database.

Author Contributions

AFS designed and realized the project and wrote the paper. JLR designed and supervised the project. Both authors read and approved the final manuscript.

Acknowledgments

This work was supported financially by the Swiss National Science Foundation (grant no. 200020_178998).

References

(1) Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discovery Today* **2019**, *24* (5), 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.

(2) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62* (9), 2021–2034. https://doi.org/10.1021/acs.jcim.2c00224.

(3) Neumann, A.; Marrison, L.; Klein, R. Relevance of the Trillion-Sized Chemical Space “eXplore” as a Source for Drug Discovery. *ACS Med. Chem. Lett.* **2023**, *14* (4), 466–472. https://doi.org/10.1021/acsmedchemlett.3c00021.

(4) Reymond, J.-L. Chemical Space as a Unifying Theme for Chemistry. *J. Cheminform.* **2025**, *17* (1), 6. https://doi.org/10.1186/s13321-025-00954-0.

(5) Koutsoukas, A.; Paricharak, S.; Galloway, W. R.; Spring, D. R.; Ijzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2014**, *54* (1), 230–242. https://doi.org/10.1021/ci400469u.

(6) Dunteman, G. H. *Principal Components Analysis*; SAGE, 1989.

(7) Oprea, T. I.; Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **2001**, *3* (2), 157–166. https://doi.org/10.1021/cc0000388.

(8) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chemical Biology & Drug Design* **2007**, *70* (5), 393–412. https://doi.org/10.1111/j.1747-0285.2007.00579.x.

(9) Awale, M.; Reymond, J. L. Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2015**, *55* (8), 1509–1516. https://doi.org/10.1021/acs.jcim.5b00182.

(10) Awale, M.; Probst, D.; Reymond, J.-L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2017**, *57* (4), 643–649. https://doi.org/10.1021/acs.jcim.6b00690.

(11) Kohonen, T. The Self-Organizing Map. *Proceedings of the IEEE* **1990**, *78* (9), 1464–1480. https://doi.org/10.1109/5.58325.

(12) Gaspar, H. A.; Baskin, I. I.; Marcou, G.; Horvath, D.; Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **2015**, *55* (1), 84–94. https://doi.org/10.1021/ci500575y.

(13) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579–2605.

(14) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **2018**, *3* (29), 861. https://doi.org/10.21105/joss.00861.

(15) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.

(16) Jin, X.; Awale, M.; Zasso, M.; Kostro, D.; Patiny, L.; Reymond, J.-L. PDB-Explorer: A Web-Based Interactive Map of the Protein Data Bank in Shape Space. *BMC Bioinformatics* **2015**, *16* (1), 339. https://doi.org/10.1186/s12859-015-0776-9.

(17) Schwartz, J.; Awale, M.; Reymond, J.-L. SMIfp (SMILES Fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules. *J Chem Inf Model* **2013**, *53* (8), 1979–1989. https://doi.org/10.1021/ci400206h.

(18) Awale, M.; van Deursen, R.; Reymond, J.-L. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013**, *53* (2), 509–518. https://doi.org/10.1021/ci300513m.

(19) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. https://doi.org/10.1021/ci500588j.

(20) Hoksza, D.; Škoda, P.; Voršilák, M.; Svozil, D. Molpher: A Software Framework for Systematic Chemical Space Exploration. *Journal of Cheminformatics* **2014**, *6* (1), 7. https://doi.org/10.1186/1758-2946-6-7.

(21) Takahashi, Y.; Konji, M.; Fujishima, S. MolSpace: A Computer Desktop Tool for Visualization of Massive Molecular Data. *Journal of Molecular Graphics and Modelling* **2003**, *21* (5), 333–339. https://doi.org/10.1016/S1093-3263(02)00180-8.

(22) Probst, D.; Reymond, J.-L. FUn: A Framework for Interactive Visualizations of Large, High-Dimensional Datasets on the Web. *Bioinformatics* **2018**, *34* (8), 1433–1435. https://doi.org/10.1093/bioinformatics/btx760.

(23) Probst, D.; Reymond, J.-L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J. Chem. Inf. Model.* **2018**, *58* (1), 1–7. https://doi.org/10.1021/acs.jcim.7b00425.

(24) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.

(25) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.

(26) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681. https://doi.org/10.1016/j.isci.2020.101681.

(27) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4* (11), 1803–1805. https://doi.org/10.1002/cmdc.200900317.

(28) van Deursen, R.; Blum, L. C.; Reymond, J. L. A Searchable Map of PubChem. *J. Chem. Inf. Model.* **2010**, *50* (11), 1924–1934.

(29) Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and Subsets of the Chemical Universe Database GDB-13 for Virtual Screening. *J. Comput.-Aided Mol. Des.* **2011**, *25* (7), 637–647.

(30) Ruddigkeit, L.; Blum, L. C.; Reymond, J. L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2013**, *53* (1), 56–65. https://doi.org/10.1021/ci300535x.

(31) Jégou, H.; Douze, M.; Schmid, C. Product Quantization for Nearest Neighbor Search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33* (1), 117–128. https://doi.org/10.1109/TPAMI.2010.57.

(32) Kirchoff, K. E.; Wellnitz, J.; Hochuli, J. E.; Maxfield, T.; Popov, K. I.; Gomez, S.; Tropsha, A. Utilizing Low-Dimensional Molecular Embeddings for Rapid Chemical Similarity Search. In *Advances in Information Retrieval*; Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I., Eds.; Springer Nature Switzerland: Cham, 2024; pp 34–49. https://doi.org/10.1007/978-3-031-56060-6_3.

(33) Pérez, K. L.; Jung, V.; Chen, L.; Huddleston, K.; Miranda-Quintana, R. A. BitBIRCH: Efficient Clustering of Large Molecular Libraries. *Digital Discovery* **2025**, *4* (4), 1042–1051. https://doi.org/10.1039/D5DD00030K.

(34) Pickering, I.; Zsigmond, K.; Pérez, K. L.; Lžičař, M.; Miranda-Quintana, R. A. BitBIRCH-Lean: Chemical Space in the Palm of Your Workstation. bioRxiv October 23, 2025, p 2025.10.22.684015. https://doi.org/10.1101/2025.10.22.684015.

(35) Matsui, Y.; Ogaki, K.; Yamasaki, T.; Aizawa, K. PQk-Means: Billion-Scale Clustering for Product-Quantized Codes. arXiv September 12, 2017. https://doi.org/10.48550/arXiv.1709.03708

(36) Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. J. Chem. Inf. Model. 2007, 47 (2), 342–353. https://doi.org/10.1021/ci600423u

(37) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. J. Chem. Inf. Model. 2009, 49 (4), 1010–1024. https://doi.org/10.1021/ci800426u