RESEARCH NOTE

ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 1; peer review: 1 approved, 2 approved with reservations]

J. Jesús Naveja[1,2], José L. Medina-Franco[2]

[1]PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico
[2]Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

v1 **First published:** 17 Jul 2017, **6**(Chem Inf Sci):1134
https://doi.org/10.12688/f1000research.12095.1
Latest published: 04 Aug 2017, **6**(Chem Inf Sci):1134
https://doi.org/10.12688/f1000research.12095.2

Abstract
We present a novel approach called ChemMaps for visualizing chemical space based on the similarity matrix of compound datasets generated with molecular fingerprints' similarity. The method uses a 'satellites' approach, where satellites are, in principle, molecules whose similarity to the rest of the molecules in the database provides sufficient information for generating a visualization of the chemical space. Such an approach could help make chemical space visualizations more efficient. We hereby describe a proof-of-principle application of the method to various databases that have different diversity measures. Unsurprisingly, we found the method works better with databases that have low 2D diversity. 3D diversity played a secondary role, although it becomes increasingly relevant as 2D diversity increases. For less diverse datasets, taking as few as 25% satellites seems to be sufficient for a fair depiction of the chemical space. We propose to iteratively increase the satellites number by a factor of 5% relative to the whole database, and stop when the new and the prior chemical space correlate highly. This Research Note warrants the full application of this method for several datasets.

Keywords
chemical space, data visualization, epigenetics, principal components analysis, similarity matrix

This article is included in the Chemical Information Science gateway.

Open Peer Review

Reviewer Status ? ? ?

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 04 Aug 2017	? report	? report	? report
version 1 17 Jul 2017	? report	? report	✓ report

1. **Gerald Maggiora**, University of Arizona, Tucson, USA

2. **Dmitry I. Osolodkin**, Chumakov FSC R&D IBP RAS, Moscow, Russian Federation
Lomonosov Moscow State University, Moscow, Russian Federation

3. **Jean-Louis Reymond**, University of Bern, Bern, Switzerland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: José L. Medina-Franco (jose.medina.franco@gmail.com)

Author roles: Naveja JJ: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Medina-Franco JL**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Consejo Nacional de Tecnología (CONACyT) scholarship 622969 (JJN). Universidad Nacional Autónoma de México (UNAM), Programa de Apoyo a la Investigación y el Posgrado PAIP, grant 5000-9163 (JLMF) and Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica PAPIIT, grant IA204016 (JLMF).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Naveja JJ and Medina-Franco JL. **ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 1; peer review: 1 approved, 2 approved with reservations]** F1000Research 2017, **6**(Chem Inf Sci):1134 https://doi.org/10.12688/f1000research.12095.1

First published: 17 Jul 2017, **6**(Chem Inf Sci):1134 https://doi.org/10.12688/f1000research.12095.1

Introduction

Visual representation of chemical space has multiple implications in drug discovery for virtual screening, library design and comparison of compound collections, among others[1]. Amongst the multiple methods to explore chemical space, principal component analysis (PCA) of pairwise similarity matrices computed with structural fingerprints has been used to analyze compound datasets[2,3]. A drawback of this approach is that it becomes impractical for large libraries due to the large dimension of the similarity matrix[4]. Other approaches use molecular representations different from structural fingerprints, such as physicochemical properties or complexity descriptors, or methods different from PCA, such as multidimensional-scaling and neural networks[5,6].

In representation of the chemical space based on PCA there have been "chemical satellite" approaches, such as ChemGPS, which select satellites molecules that might not be included in the database to visualize, but have extreme features that place them as outliers, with the intention to reach as much of the chemical space as possible[7–9]. Although we concur with the fact that not all compounds in a compound data set should be necessary to generate a meaningful chemical space, there are still obvious limitations of using a fixed set of satellites to which the user is blinded.

We therefore suggest the hybrid approach, ChemMaps, in which a portion of the database to be represented is used as satellite, thereby decreasing the computational effort required to compute the similarity matrix without losing adaptability of the method to any particular database. Since it is expected that more diverse sets would require more satellites, a second goal of this study was to qualitatively explore the relationship between the internal diversity of compound datasets and the fraction of compounds required as satellites, in order to generate a good approximation of the chemical space.

Methods

Table 1 summarizes the six compound data sets considered in this study. Note that small median similarity values imply higher diversity. The datasets were selected from a large scale study of profiling epigenetic datasets (unpublished study, Naveja JJ and Medina-Franco JL) with relevance in epigenetic-drug discovery. We also included DrugBank as a control diverse dataset[10]. Briefly, we selected focused libraries of inhibitors of DNMT1 (a DNA-methyltransferase; library diverse 2D and 3D), L3MBTL3 (a histone methylation reader; diverse 3D and less diverse 2D), SMARCA2 (a chromatin remodeller; diverse 2D, less diverse 3D), and CREBBP (a histone acetyltransferase; less diverse both 2D and 3D). Datasets were selected based on their different internal diversity (as measured with Tanimoto index/MACCS keys for 2D measurements and Tanimoto combo/OMEGA-ROCS for 3D; see Figure S1 in Supplementary File 1). Data sets in this work have approximately the same number of compounds except for HDAC1 and DrugBank, which were selected to benchmark the method in larger databases (Table 2). We evaluated 2D diversity using the median of Tanimoto/MACCS similarity measures in KNIME version 3.3.2, and 3D diversity using the median of Combo Score from the ROCS, version 3.2.2 and OMEGA, version 2.5.1, OpenEye software[11–14].

To assess the hypothesis of this work we performed two main approaches A): *Backwards approach*: start with computing the full similarity matrix of each data set and remove compounds systematically; and B) *Forward approach:* start adding compounds to the similarity matrix until finding the reduced number of

Table 1. Compound data sets used in the study.

Dataset	Description	Size	2D similarity[a]	2D similarity[b]	3D similarity[c]
DNMT1 inhibitors	DNA-methyltransferase	244	0.44	0.12	0.16
SMARCA2 inhibitors	Chromatin remodeller	220	0.51	0.15	0.23
CREBBP inhibitors	Histone acetyltransferase	178	0.67	0.22	0.16
L3MBTL3 inhibitors	Histone methylation reader	115	0.77	0.41	0.03
HDAC1 inhibitors	Histone acetyltransferase	3,257	0.49	0.16	0.12
DrugBank	Approved drugs	1,900	0.35	NC	NC

[a]Median of Tanimoto/MACCS similarity; [b]Median of Tanimoto/ECFP4 similarity; [c]Median of OMEGA-ROCS similarity; NC: not calculated

Table 2. Benchmark with larger databases.

Database	Gold standard timing (s)	Satellites timing (s)	Correlation
DrugBank	162	147	0.92
HDAC1	406	287	0.99

required compounds (called 'satellites') to reach a visualization of the chemical space that is very similar to computing the full similarity matrix. The second approach would be the usual and realistic approach from a user standpoint. Each method is further detailed in the next two subsections.

Backwards approach

The following steps were implemented in an automated workflow in KNIME, version 3.3.2[15]:

1. For each compound in the dataset with *N* compounds, generate the *N* X *N* similarity matrix using Tanimoto/extended connectivity fingerprints radius 4 (ECFP4) generated with CDK KNIME nodes.

2. Perform PCA of the similarity matrix generated in step 1 and selected the first 2 or 3 principal components (PCs).

3. Compute all pair-wise Euclidean distances based on the scores of the 2 or 3 PCs generated in step 2. The set of distances are later used as reference or '*gold standard*'.

4. Repeat steps 1 to 3 with one compound as satellite, generating an *N X 1* similarity matrix. The first compound was selected randomly. In this case, for example, it is only possible to calculate one PC, but as the number of satellites increases, we can again compute 2 or 3 PCs.

5. Calculate the correlation among the pairwise distances generated in step 2 obtained using the whole matrix (e.g., *gold standard*) and those obtained in step 4.

6. Iterate over steps 4 and 5 increasing the number of satellites one by one until *N - 1* satellites are reached. To select the second, third, etc. compounds, two approaches were followed: select compounds at random and select compounds with the largest diversity to the previously selected (i.e., Max-Min approach).

7. Estimate the proportion of satellite compounds required to preserve a 'high' (of at least 0.9) correlation.

8. The prior steps were repeated five times for each dataset in order to capture the stability of the method.

Forward approach

The former approach is useful only for validation purposes of the methodology as a proof-of-principle. However, the obvious objective of a satellite-approach is to avoid the calculation of the complete similarity matrix e.g., step 1 in backwards approach. To this end, we developed a satellite-adding or forward approach, in contrast with the formerly introduced backwards approach. We started with 25% of the database as satellites and for each iteration we added 5% until the correlation of the pairwise Euclidean distances remains high (at least 0.9).

Dataset 1. This file contains five compound datasets used in this work in SDF format

http://dx.doi.org/10.5256/f1000research.12095.d168322

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. The HDAC1 dataset is available from ChEMBL, version 23 at https://www.ebi.ac.uk/chembl/.

Results

Backwards approach

In this pilot study, we assessed a few variables to tune up the method, such as the number of PCs used (2 or 3) and the selection of satellites at random or by diversity. We found that selection at random is more stable, above all in less diverse datasets (Figure 1 and Figure 2; Figure S2 and Figure S3). Likewise, selecting 2 PCs the performance is slightly better and more stable (compare Figure 1 and Figure 2 against Figure S2 and Figure S3).

Therefore, from this point onwards we will focus on the results of the at random satellites selection and using 2 PCs (Figure 2). From the four datasets, we conclude that for datasets with lower 2D diversity (CREBBP and L3MBTL3, see Table 1), around 25% of satellite compounds are enough to obtain a high correlation (≥ 0.9) with the gold standard (e.g., PCA on the whole matrix). Whereas for 2D-diverse datasets i.e., DNMT1 and SMARCA2, up to 75% of the compounds could be needed to ensure a high correlation. Nonetheless, even for these datasets, using 25% of the compounds as satellites the correlation with the gold standard is already between 0.6 and 0.8; using 50% of the compounds as satellites the correlation is between 0.7 and 0.9. Hence, the higher the diversity of a dataset (especially 2D), the higher the number of satellites required.

Forward approach

Evidently, a useful method for reducing computing time and disk usage space should not use the PCA on the whole similarity matrix to determine an adequate number of satellites for each dataset. With that in mind, we decided to design a method that starts with a given percentage of the database as satellites, and then keeps adding a proportion of them until the correlation between the former and the updated data is of at least 0.9. In Figure 3 we depict this approach on the same databases in Table 1 for step sizes of 5% and starting from zero. Similarly as what we saw in the backwards method, around 5 steps (25% of the database) are usually necessary to reach a stable, high correlation between steps. Figure S4 shows that for step sizes of 10% there is no further improvement. Therefore we suggest that the method should, for default, start with 25% of compounds as satellites and then keep adding 5% until a correlation between steps of at least 0.9 is reached.

Application

In this pilot study we applied the method to visualize the chemical space of two larger datasets (HDAC1 and DrugBank with 3,257 and 1,900 compounds, respectively, Table 1). As shown in Table 2, a significant reduction in time performance was achieved as compared to the gold standard, and the correlation between the gold standard and the satellites approach was in both cases higher than 0.9. Figure 4 depicts the chemical spaces generated in both instances. Although the orientation of the map changed for HDAC1, the shape and distances remain quite similar, which is the main objective. This preliminary work supports the hypothesis that a reduced number of compounds is sufficient to generate a visual representation of the chemical space (based on PCA of the similarity matrix) that is quite similar to the chemical space of the PCA of the full similarity matrix.

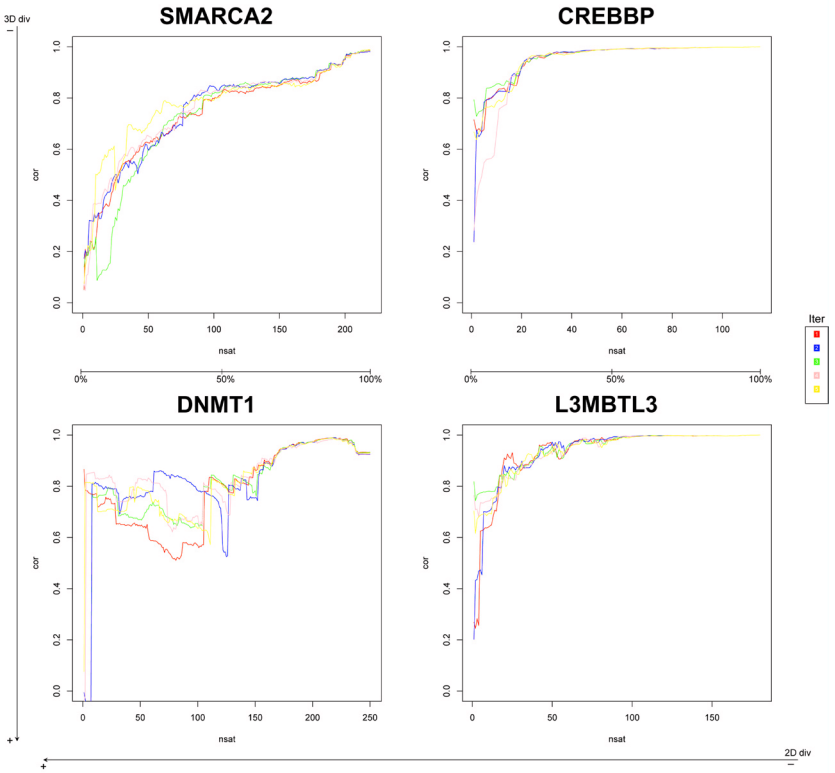

Figure 1. Backwards analysis with 2PCs picking satellites by diversity. The correlation with the results from the whole matrix was calculated with increasing numbers of satellites. Each colored line represents one of the five random sets.

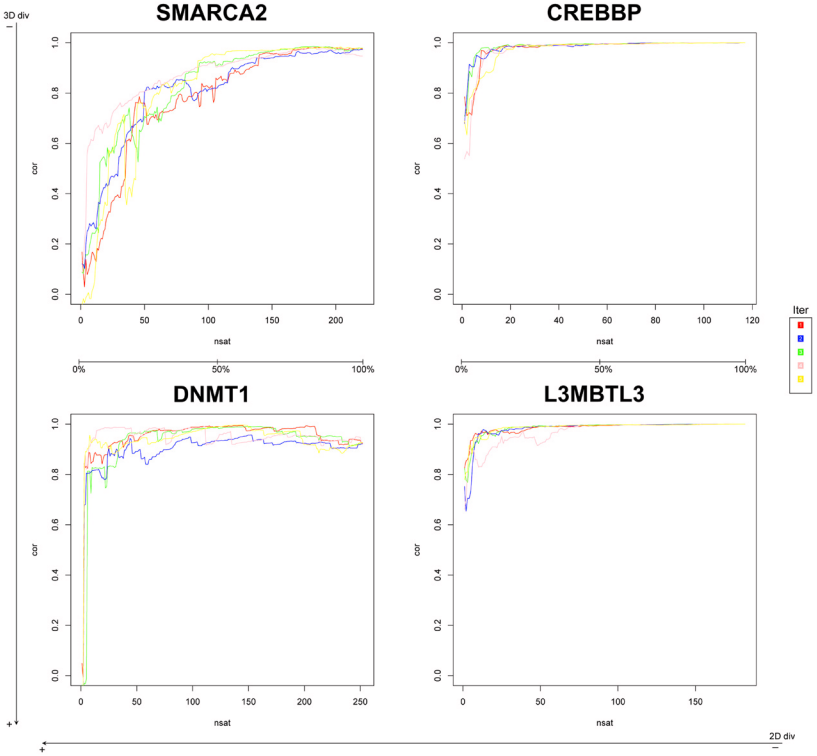

Figure 2. Backwards analysis with 2PCs picking satellites at random. The correlation with the results from the whole matrix was calculated with increasing numbers of satellites. Each colored line represents one of the five random sets.

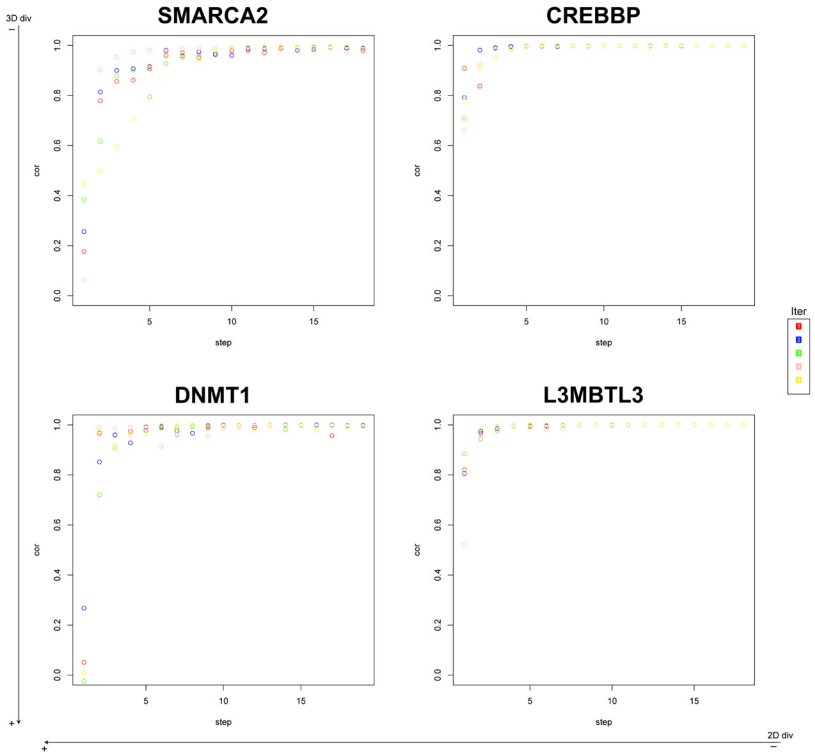

Figure 3. Forward analysis with 2PCs picking satellites at random step sizes of 5%.

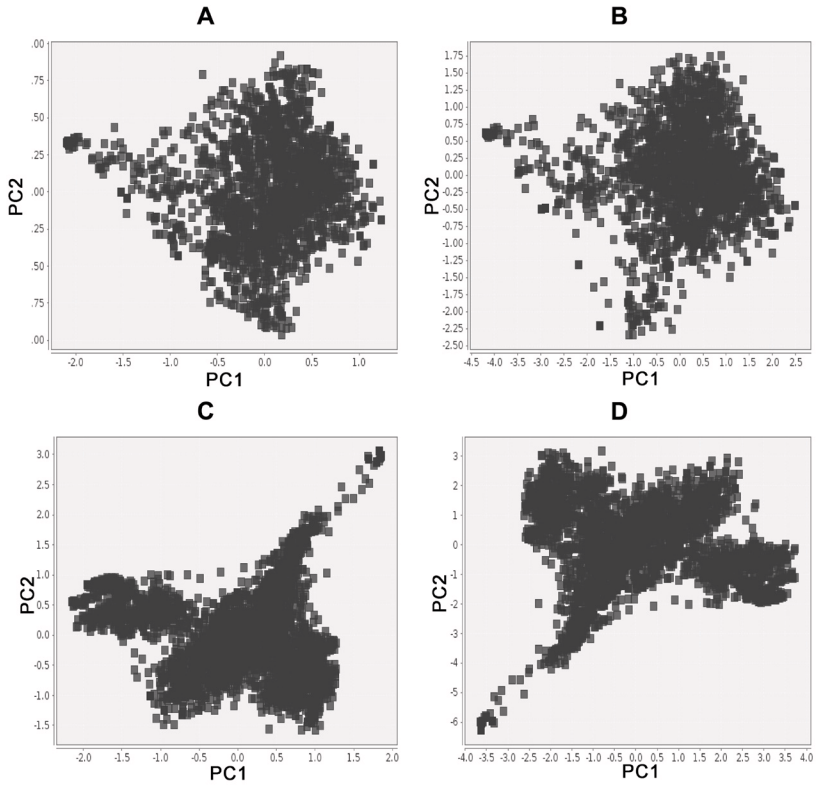

Figure 4. Chemical space of DrugBank using (**A**) the adaptive satellites approach or (**B**) the gold standard. As well as for HDAC1 using (**C**) the adaptive satellites approach or (**D**) the gold standard.

Conclusion and future directions

This proof-of-concept study suggests that using the adaptive satellite compounds ChemMaps is a plausible approach to generate a reliable visual representation of the chemical space based on PCA of similarity matrices. The approach works better for relatively less-diverse datasets, although it seems to remain robust when applied to more diverse datasets. For datasets with small diversity, fewer satellites seem to be enough to produce a representative visual representation of the chemical space. The higher relevance of 2D diversity over 3D in this study could be importantly related to the fact that the chemical space depiction is based on 2D fingerprints. Therefore, the performance of the methods depicting the chemical space based on 3D fingerprints could also be assessed.

A major next step is to conduct a full benchmark study to assess the general applicability of the approach proposed herein, and also in larger databases, in which we anticipate this method would be even more useful. A second step is to propose a metric that determines the number of compounds required as satellites for PCA representation of the chemical space based on similarity matrices.

Data availability

Dataset 1: This file contains five compound datasets used in this work in SDF format. No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. The HDAC1 dataset is available from ChEMBL, version 23 at https://www.ebi.ac.uk/chembl/. doi, 10.5256/f1000research.12095.d168322[16]

Competing interests

No competing interests were disclosed.

Grant information

Consejo Nacional de Tecnología (CONACyT) scholarship 622969 (JJN). Universidad Nacional Autónoma de México (UNAM), *Programa de Apoyo a la Investigación y el Posgrado* PAIP, grant 5000-9163 (JLMF) and *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* PAPIIT, grant IA204016 (JLMF).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Insightful discussions with Dr. Jakyung Yoo (Daewoong Life Science Research Institute) are highly appreciated. The authors thank OpenEye for the academic license granted.

Supplementary material

Supplementary File 1: File with four supporting figures. Figure S1: 3D-Consensus Diversity Plot depicting the diversity of the datasets used for the backwards approach; Figure S2: Backwards analysis with 3PCs picking satellites by diversity; Figure S3: Backwards analysis with 3PCs picking satellites at random; Figure S4: Forward analysis with 2PCs picking satellites at random with step sizes of 10%.

Click here to access the data.

References

1. Medina-Franco J, Martinez-Mayorga K, Giulianotti M, *et al.*: **Visualization of the chemical space in drug discovery.** *Curr Comput-Aided Drug Discov.* 2008; **4**(4): 322–333.
Publisher Full Text
2. Reymond JL: **The chemical space project.** *Acc Chem Res.* 2015; **48**(3): 722–730.
PubMed Abstract | Publisher Full Text
3. Naveja JJ, Medina-Franco JL: **Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors.** *RSC Adv.* 2015; **5**(78): 63882–63895.
Publisher Full Text
4. Maggiora GM, Bajorath J: **Chemical space networks: a powerful new paradigm for the description of chemical space.** *J Comput Aided Mol Des.* 2014; **28**(8): 795–802.
PubMed Abstract | Publisher Full Text
5. Medina-Franco JL: **Interrogating novel areas of chemical space for drug discovery using chemoinformatics.** *Drug Dev Res.* 2012; **73**(7): 430–438.
Publisher Full Text
6. Osolodkin DI, Radchenko EV, Orlov AA, *et al.*: **Progress in visual representations of chemical space.** *Expert Opin Drug Discov.* 2015; **10**(9): 959–973.
PubMed Abstract | Publisher Full Text
7. Larsson J, Gottfries J, Muresan S, *et al.*: **ChemGPS-NP: tuned for navigation in biologically relevant chemical space.** *J Nat Prod.* 2007; **70**(5): 789–794.
PubMed Abstract | Publisher Full Text
8. Larsson J, Gottfries J, Bohlin L, *et al.*: **Expanding the ChemGPS chemical space with natural products.** *J Nat Prod.* 2005; **68**(7): 985–991.
PubMed Abstract | Publisher Full Text
9. Rosén J, Lövgren A, Kogej T, *et al.*: **ChemGPS-NP(Web): chemical space navigation online.** *J Comput Aided Mol Des.* 2009; **23**(4): 253–259.
PubMed Abstract | Publisher Full Text
10. Wishart DS, Knox C, Guo AC, *et al.*: **DrugBank: a comprehensive resource for *in silico* drug discovery and exploration.** *Nucleic Acids Res.* 2006; **34**(Database issue): D668–72.
PubMed Abstract | Publisher Full Text | Free Full Text
11. OpenEye Scientific Software, Santa Fe NM: **ROCS 3.2.1.4.** 2017.
Reference Source
12. OpenEye Scientific Software, Santa Fe NM: **OMEGA 2.5.1.4.** 2017.
Reference Source
13. Hawkins PC, Skillman AG, Warren GL, *et al.*: **Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database.** *J Chem Inf Model.* 2010; **50**(4): 572–584.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Hawkins PC, Skillman AG, Nicholls A: **Comparison of shape-matching and docking as virtual screening tools.** *J Med Chem.* 2007; **50**(1): 74–82.
PubMed Abstract | Publisher Full Text
15. Berthold MR, Cebron N, Dill F, *et al.*: **KNIME - the Konstanz information miner.** *SIGKDD Explor Newsl.* 2009; **11**(1): 26.
Publisher Full Text
16. Naveja JJ, Medina-Franco JL: **Dataset 1 in: ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds.** *F1000Research.* 2017.
Data Source

Open Peer Review

Current Peer Review Status: ? ? ✓

Version 1

Reviewer Report 31 July 2017

https://doi.org/10.5256/f1000research.13086.r24277

✓ **Jean-Louis Reymond**
Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

J. Jesús Naveja *et al* present a methodology for representation of chemical space of small sets of compounds. In general, the approach involves selection of satellite compounds from the database, computing the similarities of all compounds in the database to these satellites, and finally projection of the resulting similarity matrix using principal component analysis. J. Jesús Naveja *et al* further report various methods for selecting satellite compounds (backward or forward selection approach; selection at random or selection by diversity check) and show how the number of selected satellite compounds influence the quality of projection.

Comments:

1. The authors are completely hiding the fact that similarity mapping is quite well-known and absolutely not new, the authors should read and cite Awale et al., J. Chem. Inf. Model., 2015, 55 (8), pp 1509–1516 and the detailed discussion of literature precedents on similarity mapping presented therein.

2. The authors compare their satellites to the satellite compounds used by T. Oprea in his 2001 approach to mapping chemical space. Obviously either they did not read Oprea's paper or they misunderstood it: Oprea's satellites are artificial molecules with extreme properties such as to orient the PCA projection and stretch its dimensions in reproducible directions. However the projection is simply PCA, and does not involve similarity mapping. In similarity mapping the satellites are molecules from within the database to which similarities are calculated.

3. In the abstract, author mentioned that "3D diversity played a secondary role, although it becomes increasingly relevant as 2D diversity increases". However, I didn't found the relevant explanation in main text supporting this statement.

4. Figure 1 and Figure 2: The five random sets in the legend. Its not clear exactly what the author meant by five random sets. As per my understanding the author used the complete set of compounds for each target and what is changing is the random selection of satellites,

which is repeated for five times.

5. In case of forward selection approach: "..With that in mind, we decided to design a method that starts with a given percentage of the database as satellites, and then keeps adding a proportion of them until the correlation between the former and the updated data is of at least 0.9. " The correlation between projections obtained from the current set of satellites and projections obtained from former set of satellites might well be high, but still the correlation to the projection obtained from the complete similarity matrix is low. How one can assure the quality of projection in this case?

6. For all plots axis labels are too small to read.

References

1. Awale M, Reymond JL: Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces.*J Chem Inf Model*. 2015; **55** (8): 1509-16
PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Cheminformatics and drug design

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 02 Aug 2017

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

Dear Dr. Reymond, thank you for your comments to this Research Note.

Regarding the modifications we have done considering your comments:

- In the Introduction we briefly discuss other similarity approaches to visualize the chemical space. We have expanded that discussion there with a reference to the Similarity Mapplet approach.
- We did not intend to imply that Oprea's and Gottfires' ChemGPS approach is based on structural similarity. To clarify this point we rephrased that in the introduction.
- In the corresponding Figures legends we changed "random sets" with "iterations".
- We added to the Supplementary Information a discussion on the correlation of the complete similarity matrix Euclidean distances and using only 2 and 3PCs. However, we would like to highlight that our approach is intended to approximate the best possible chemical space visualization using PCA. This last is given by the first 3PCs at most.
- We augmented the font size in all figures.

Competing Interests: No competing interests were disclosed.

Reviewer Report 28 July 2017

https://doi.org/10.5256/f1000research.13086.r24276

?

Dmitry I. Osolodkin
[1] Institute of Poliomyelitis and Viral Encephalitides, Chumakov FSC R&D IBP RAS, Moscow, Russian Federation
[2] Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation

The paper under consideration presents an elegant approach to efficient mapping of chemical space using principal component analysis. Being technically sound in general, well-written and easily understandable, the paper lacks several technical details without which it is not complete. In particular:

1. The concept of 'chemical satellites' is discussed in a rather concise manner, a bit more details may be added and the seminal paper by Oprea & Gottfries [1] needs to be cited. The approach suggested here is rather different from the Oprea's one, because satellites are defined there as intentional outliers, whereas in the current work they are just extracted

from the mapped dataset. This difference should be stated in a clearer way.

2. Dataset processing routine is not presented. Although the suggested technique would work on totally random datasets (by the way, addition of such a dataset to the list of examples would be beneficial and illustrative), standardization of structures should be performed for consistency and for more informative application of similarity measures. Targeted datasets in the supplement look standardized, but DrugBank contains metal ions, unconnected molecules, and macromolecules, all of which may significantly distort the comparison. For HDAC1 inhibitors the procedure to obtain this dataset from ChEMBL should be provided, because simple target keyword search for 'hdac1' gives 9 different datasets.

3. Diversity of datasets may be additionally illustrated by any of currently available visualization methods. A method that clearly shows compound clustering or diversity of the dataset would be preferred.

4. Visual comparison of figures is not sufficient to make conclusions about preference of random selection over diversity-based (Figures 1, 2, S2, S3). Differences are visible, but their importance and significance are not obvious (maybe just for me), so use of a quantitative measure would be highly appreciated. Random selection shows sometimes lower stability of the backwards analysis (larger difference between the iterations), and this observation could be discussed.

5. Some analysis of the technique applicability domain would significantly improve the conclusions of the paper. One parameter that deserves attention is dataset diversity threshold above which the technique becomes unstable or less useful. Will it work good for totally random or intentionally diverse compounds or for datasets with two or three large congeneric series? A slightly more thorough characterization of example datasets would be useful to deal with this question.

References

1. Oprea T, Gottfries J: Chemography: The Art of Navigating in Chemical Space. *Journal of Combinatorial Chemistry*. 2001; **3** (2): 157-166 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Aug 2017

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

Dear Dr Osolodkin, thank you, we highly appreciate your comments to this Research Note.

Regarding the modifications we have done considering your comments:

- We added a citation to the first publication related to ChemGPS by Oprea and Gottfries. In the Introduction, we further, although briefly (given the extension limit of a Research Note), explained the differences among these two approaches.
- We added a Supplementary Information file describing the data curation methodology used. -We also added the HDAC1 dataset to the supplementary files.
- Supplementary Figure 1 should address the visualization of the diversity of the datasets.
- We find quite interesting your observation about quantifying the stability of the iterations, as well as that about determining the applicability domain of the approach (including defining a diversity threshold). Based on this Research Note we are planning an extensive study fully addressing these concerns.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 July 2017

https://doi.org/10.5256/f1000research.13086.r24274

?

Gerald Maggiora

BIO5 Institute, University of Arizona, Tucson, AZ, USA

Graphically representing coordinate-based chemical spaces requires some type of dimensionality reduction. One method involves the use of similarity matrices treated as data matrices that are subsequently subjected to principal component analysis (PCA). The first two or three PCs are then used as a basis to graphically depict the chemical space. Although this approach works reasonably well, the size of chemical spaces that can be treated is somewhat limited, since the PCA transformation requires diagonalizing a matrix whose dimension is equal to the number of molecules in the chemical space of interest. The work of Naveja and Medina-Franco seeks to overcome this limitation by building a lower dimensional representation of chemical space in a stepwise manner using “backwards” or “forward” procedures. While the method has the potential for accomplishing their goals, it does not in my estimation provide a sufficiently rigorous test of the approximations that are the foundation of their approach. For this reason additional work needs to be done before their method can be applied with confidence.

My objection is based on the authors’ use of the first 2 or 3 PCs as the ‘gold standard’ for representing of the entire chemical space, and as a basis for all subsequent comparisons of the approximate chemical spaces. I would at least like to see what percent of the total sample variance is accounted for by these PCs. If it is an insignificant amount, then approximating these PCs by whatever method will not produce a sufficiently accurate model of the chemical space and their model will have to be improved. The true ‘gold standard’ is the original set of column vectors in their data matrix from which the PCs are obtained. This will produce the ‘true’ distance between ‘molecular points’ in the full dimensional chemical space, but because of its very high dimension computing distances in the original chemical space can be a problem. An alternative is to carry out the PCA and choose a larger subset of PCs (say 6 or 8) that do account for most of the sample variance and then use these in the correlation or error analysis.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Physical chemistry, biophysics, computer-aided drug design, chemical informatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 Jul 2017

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

Dear Dr. Maggiora,
We thank you for your feedback on this Application Note. We entirely agree with your comment that if the variance captured by the first 2 or 3 PCs is not high enough, the visual representation of the chemical space will not be meaningful. For the data sets included in this work, we have seen that the variance is high. We also agree that formally speaking the ´true gold standard´ would involve computing the distances for the full matrix. Based on your feedback we are preparing a revised version of this manuscript.

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 1

Reader Comment 27 Jul 2017

José L. Medina-Franco, Universidad Nacional Autónoma de México, Mexico

Dear Dr. Oprea:
Thank you for your comment. In the first version of the manuscript we cited three papers about ChemGPS published between 2005 and 2009 (references 7-9). In a revised version of our manuscript we will include the citation to the first paper you wrote about ChemGPS (*J. Comb. Chem.* **2001,** *3,* 157-166). Thanks also for your suggestion to compare directly ChemMaps with ChemGPS.

Competing Interests: No competing interests were disclosed.

Reader Comment ( ) 21 Jul 2017

Tudor Oprea, Department of Internal Medicine, Translational Informatics Division, University of New Mexico School of Medicine, USA

Dear authors,

You mention ChemGPS, but do not cite the original papers [disclosure: I wrote them].
Given that you expand on the same concept, it would make sense to compare your work with the original ChemGPS set, perhaps the "expanded" one as well (ChemGPS-NP).

Competing Interests: No competing interests were disclosed.