A Fragment Library of Natural Products and Compounds Databases for Drug Discovery

Ana Luisa Chávez-Hernández
National Autonomous University of Mexico

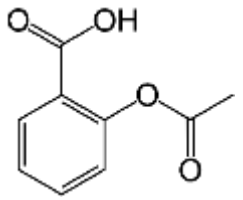
WORKSHOP ON SECONDARY METABOLITE DISCOVERY
Computational Applications in Secondary Metabolite Discovery
9 March 2021

Background

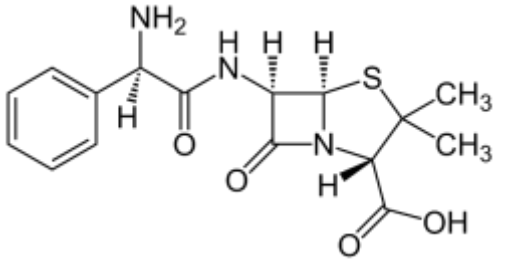
Natural products have been relevant since the beginning of the pharmaceutical era.

1562 Approved drugs for clinical use (1981-2015).

- 4% Natural products
- 21 % Derived from natural products



Acetylsalicylic acid (Aspirin)



Ampicillin

In addition, natural products fragments can be generated with retrosynthetic rules and these substructures can serve as building-blocks in the Pseudo-NP synthesis or *de novo* design.

Pseudo-NP synthesis combine NP-fragments in a way that has not yet been observed in nature.

Vidadala, S. R.; Golz, C.; Strohmann, C. et. al. *Angew. Chem., Int. Ed.*, **2015**, *54*, 651–655.
Gao, D.; Zhou, T.; Da, L.-T. et. Al. *Org. Lett.*, **2019**, *21*, 8577–8581.

Background

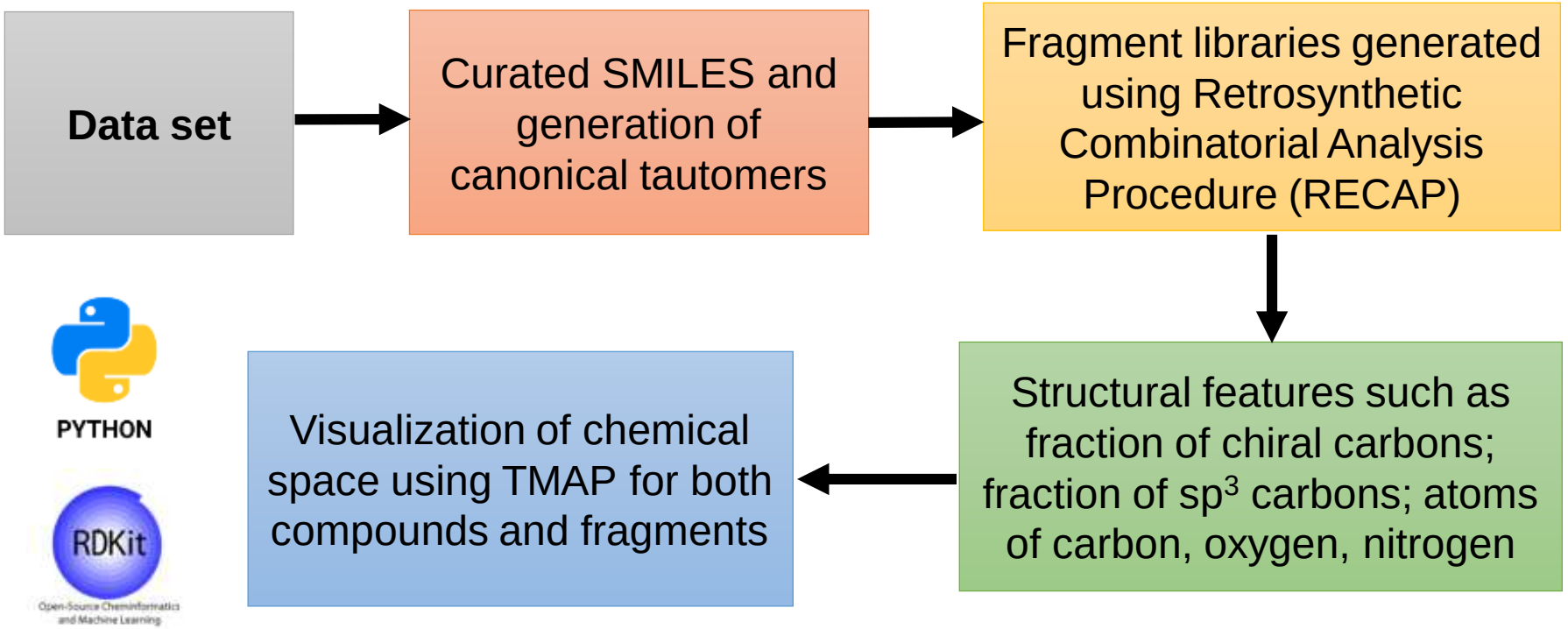
- Natural products (NP) have unique functional groups.
- Few amounts from NP are obtained during extraction and purification procedures.
- To maximize the use of NP, we propose to generate fragments libraries from NP that can be used such as building blocks for the synthesis of the so-called “pseudo-NP.”

Dataset	Original compounds	Processed compounds	Generated fragments
COlleCtion of Open NatUral producTs (COCONUT)	432,706	382,248	52,630
Food Database (FooDB)	23,883	21,319	3186
Dark Chemical Matter (DCM)*	139,352	139,326	14,001
Chemical Abstract Service (CAS) set focused on COVID-19	48,876	44,692	8432
Inhibitors of the main protease of SARS-CoV-2 (3CLP)	280	256	108

*Compounds that not showed activity although that have been thoroughly tested (DCM)

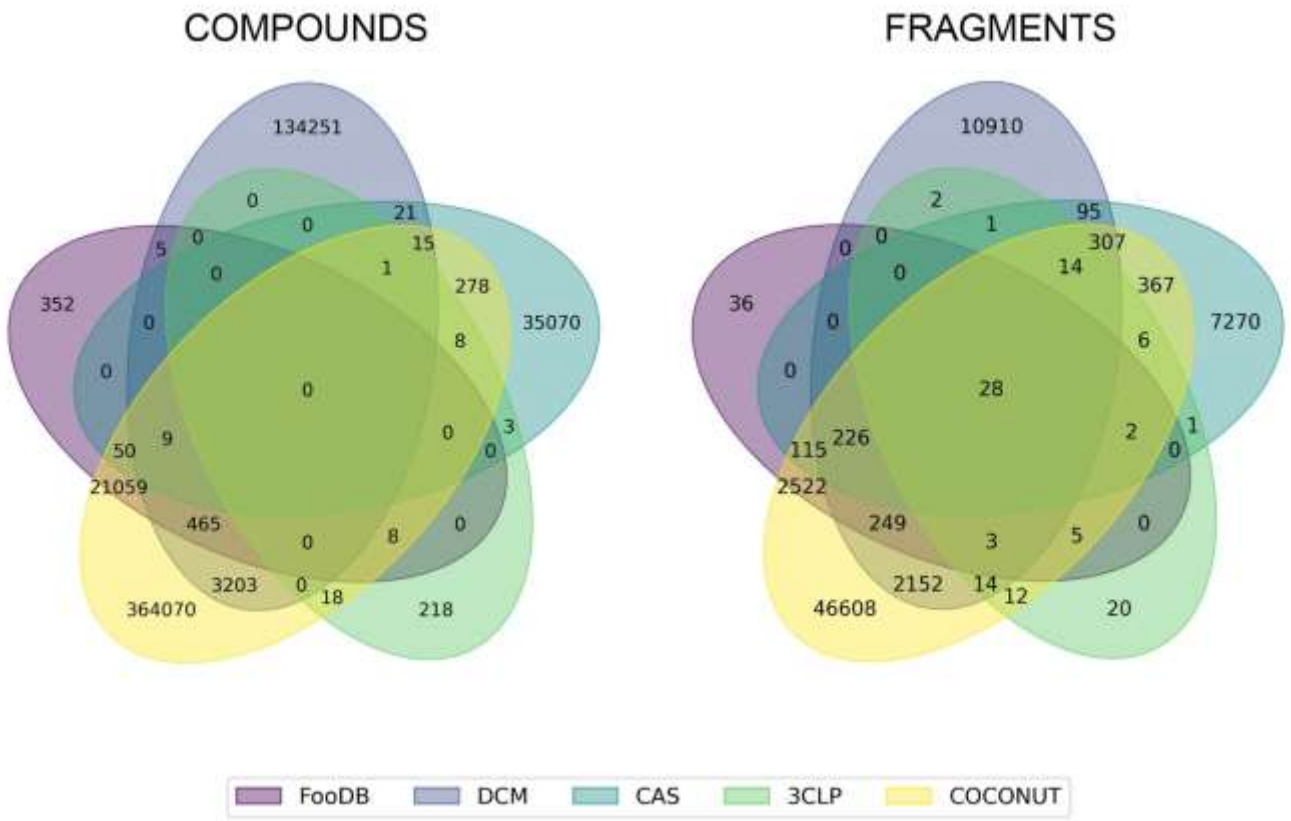
Sorokina, M.; Steinbeck, C. *J Cheminform.* **2020,** *12*, 20.
Chen, Y.; Kirchmair, J. *Mol. Inf.* **2020**, *39*, 2000171.
Christoforow, A.; Wilke, J.; Binici, A. et al. *Angew. Chemie Int. Ed.* **2019**, *58*, 14715–14723.

Methods



Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. *Biomolecules.* **2020**, *10*, 1518

Overlapping structures between COCONUT, FooDB, DCM, CAS, 3CLP



28 overlapping fragments between five data sets.

Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. *Biomolecules.* **2020**, *10*, 1518

Compounds Structural Features

Structural features	COCONUT	FooDB	DCM	CAS	3CLP
Carbon atoms	25.640	26.563	18.059	22.496	25.828
Oxygen atoms	6.167	7.343	3.252	5.773	4.922
Nitrogen atoms	1.445	0.668	2.859	4.157	3.582
Fraction of sp^3 carbons	0.506	0.620	0.342	0.489	0.291
Fraction of chiral carbons	0.154	0.152	0.028	0.145	0.069

Fragments-Structural Features

Structural features	COCONUT	FooDB	DCM	CAS	3CLP
Carbon atoms	18.504	12.991	10.181	9.904	8.926
Oxygen atoms	3.524	3.173	1.748	3.678	1.556
Nitrogen atoms	0.795	0.394	1.475	0.883	0.713
Fraction of sp^3 carbons	0.557	0.615	0.330	0.656	0.298
Fraction of chiral carbons	0.189	0.199	0.054	0.240	0.071

- Natural products (COCONUT, FooDB) had the largest number of carbon and oxygen atoms as well as compounds in the COVID-19 focused set (CAS and 3CLP). This trend was very similar in their derived fragments.

- COCONUT-fragments and compounds in the COVID-19 focused set had the largest number of nitrogen atoms.

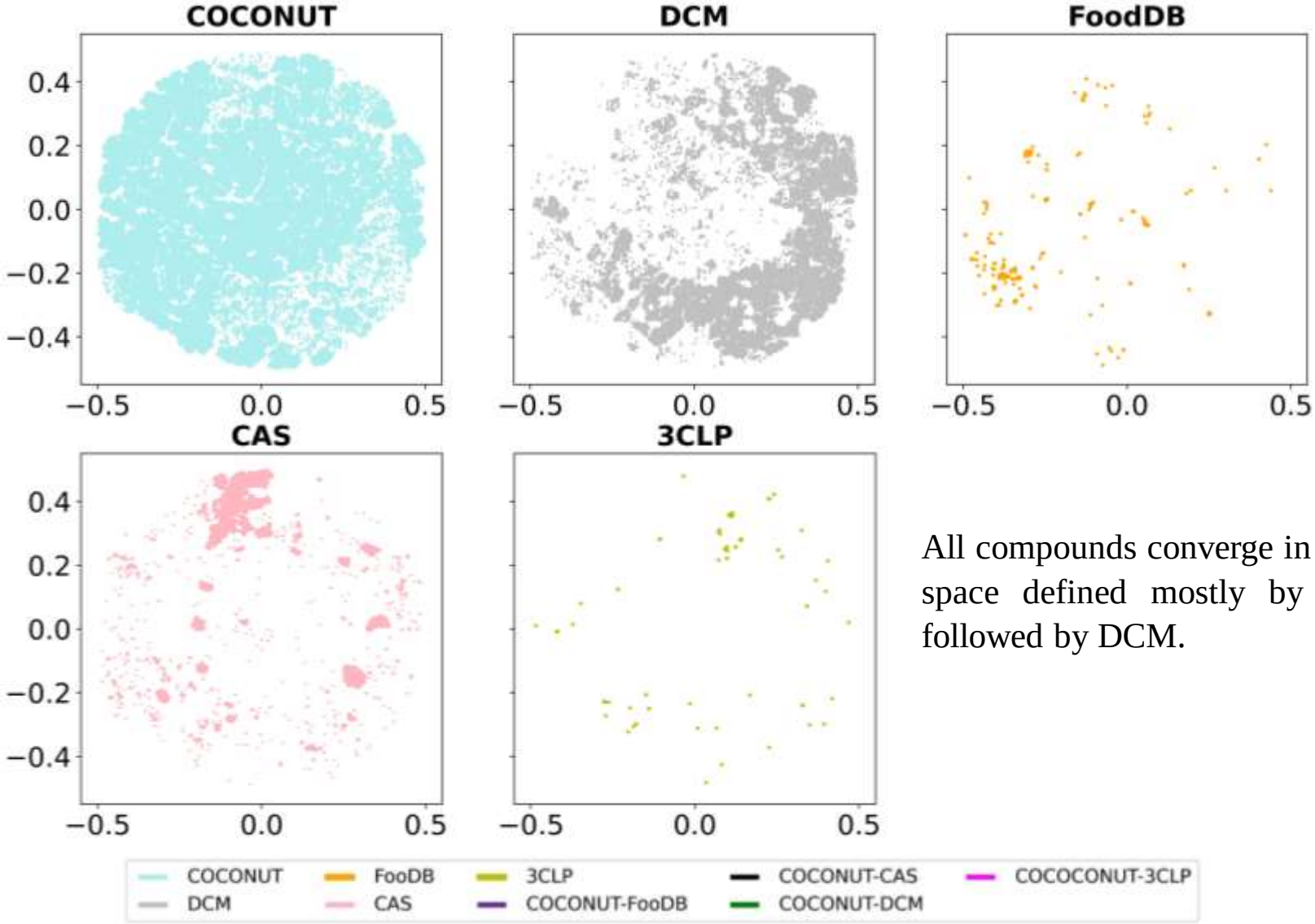
- Natural products (fragments and compounds) had the highest structural complexity (fraction of sp^3 carbons and fraction of chiral carbons). Also, compounds in the COVID-19 focused set.

Natural product fragments could be used as building-blocks for synthesis of bioactivity molecules.

Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. *Biomolecules.* **2020**, *10*, 1518

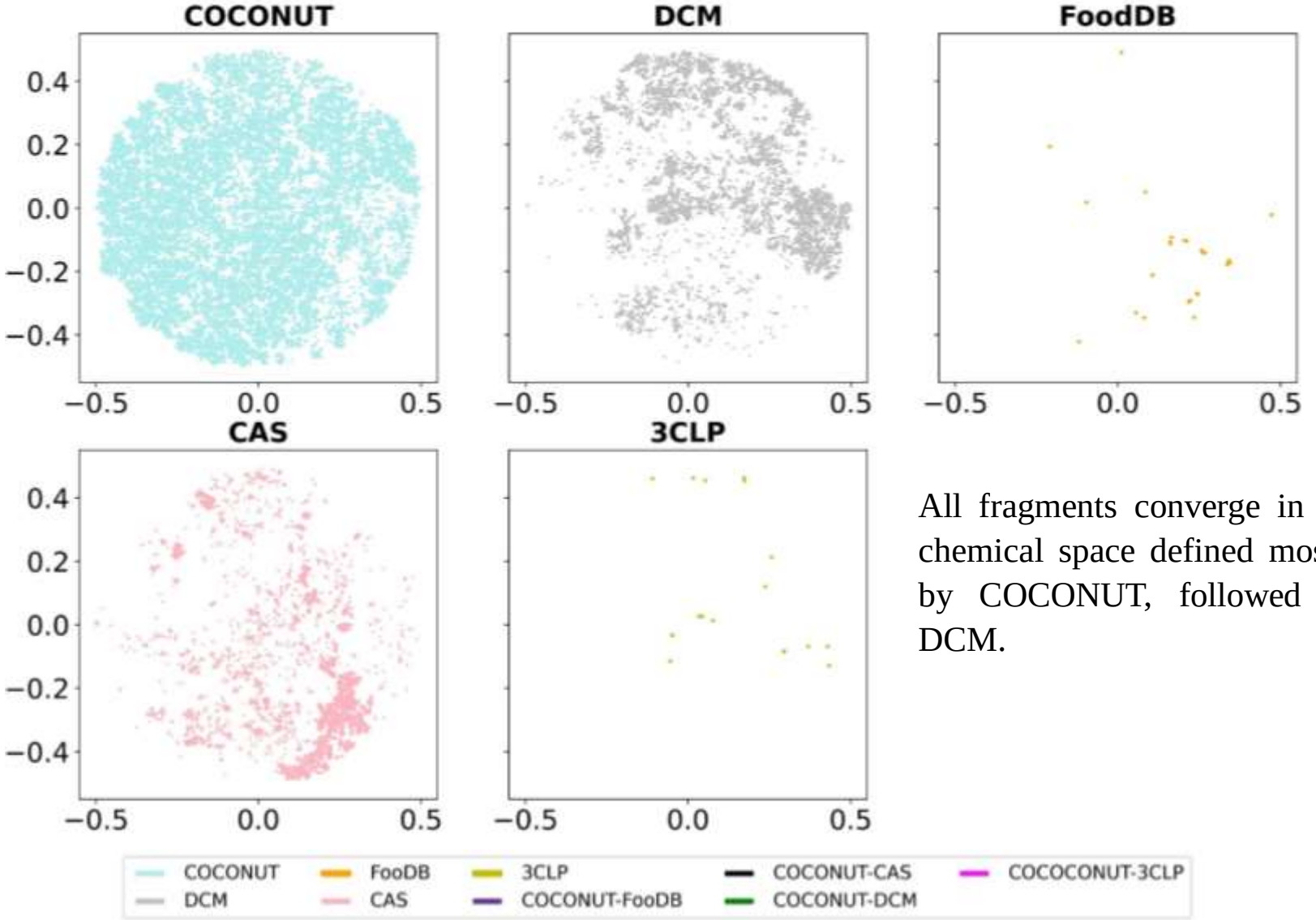
Visualization of the chemical space of the compounds data sets generated with Tree Maps

TMAP is a method to visualize thousands of compounds. Each set of similar compounds is linked together to form branches which are then linked to each other.



All compounds converge in the chemical space defined mostly by COCONUT, followed by DCM.

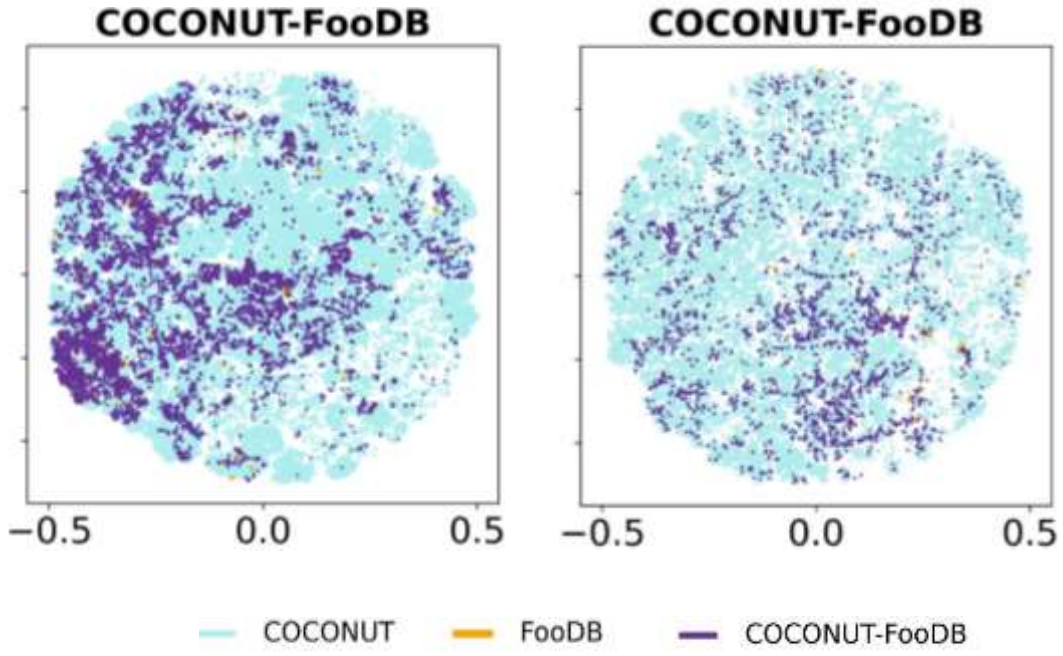
Visualization of the chemical space of the compounds data sets generated with Tree Maps



All fragments converge in the chemical space defined mostly by COCONUT, followed by DCM.

Visualization of the chemical space with Tree Maps

Compounds **Fragments**

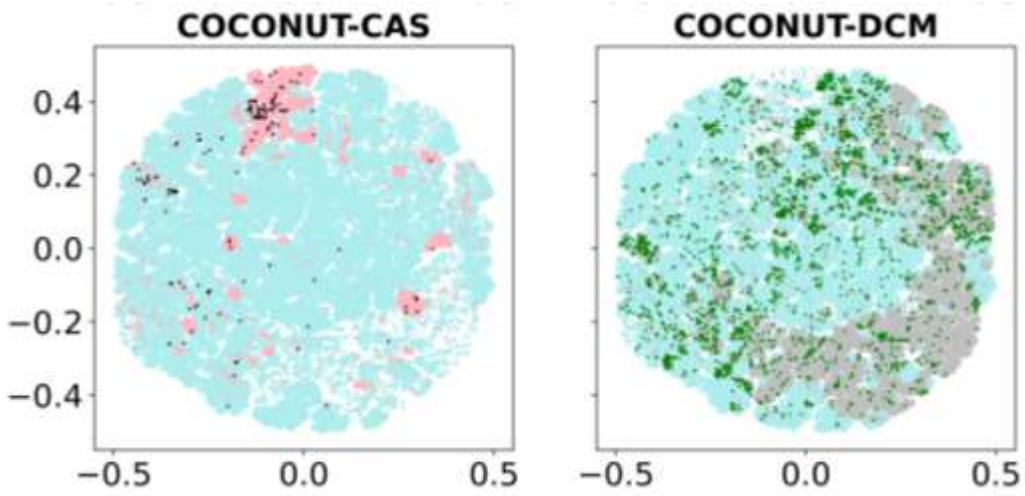


There are more overlapping compounds between COCONUT and FooDB, compared to their fragments.

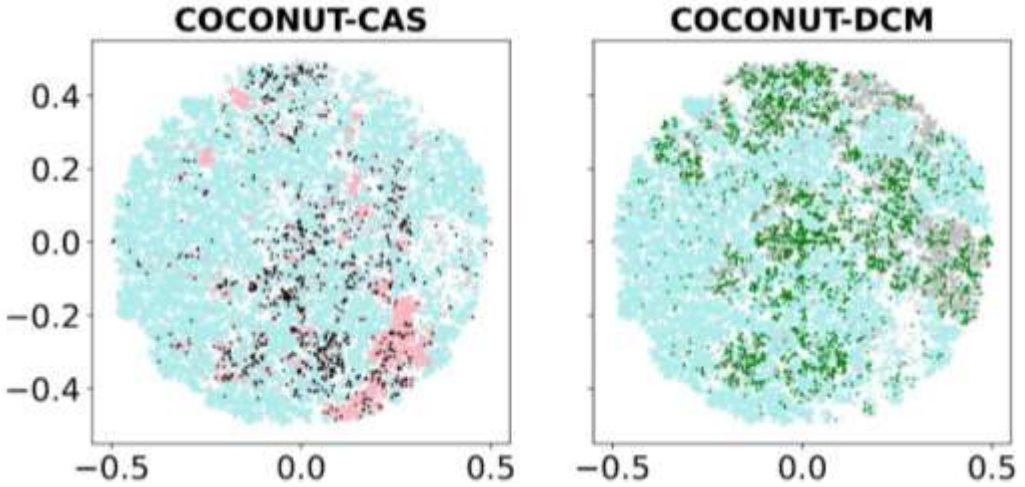
Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. *Biomolecules.* **2020**, *10*, 1518

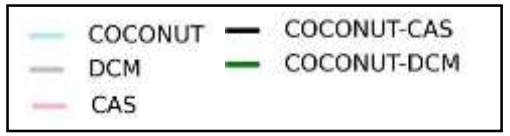
Visualization of the chemical space of the compounds data sets generated with Tree Maps

Compounds



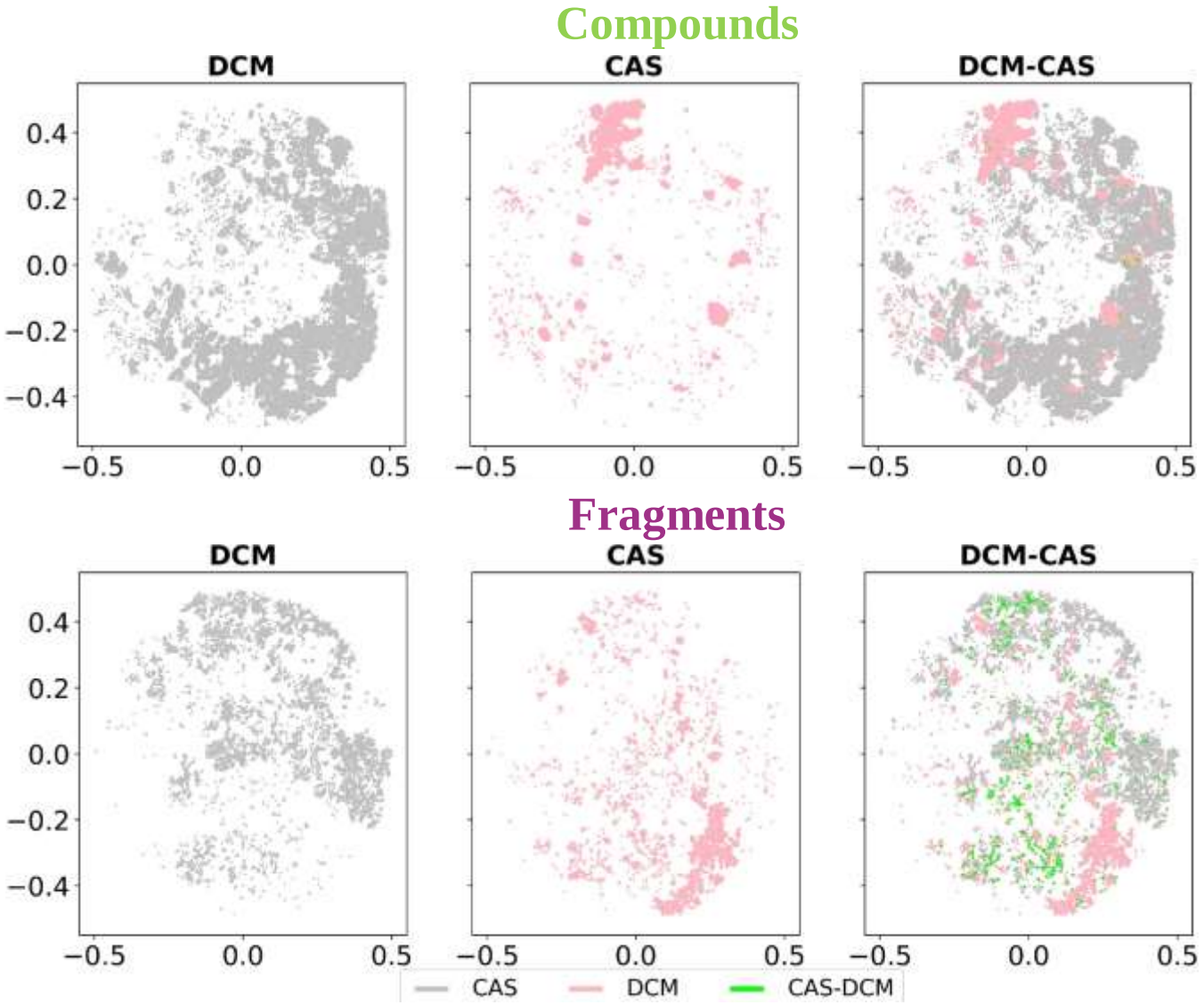
Fragments





There are more fragments in common between COCONUT-CAS and COCONUT-DCM, compared to their compounds.

Visualization of the chemicla space of the compounds data sets generated with Tree Maps



Hardly, compounds between CAS and DCM converge in chemical space, contrary to their fragments.

DCM fragments, too, can be used as building blocks in the *de novo* design of bioactive molecules even though the original compounds have weak biological activity.

Conclusions

- The natural product fragment libraries we generated in this work are public available at: https://doi.org/10.6084/m9.figshare.13064231.v1
- Molecular fragments retained the structural characteristics from original compounds which was consistent with the results of our previous work. Natural products (**COCONUT** and **FooDB**) were structurally more complex as well as compounds in a COVID-19 focused set (**CAS**). Therefore, they would be a an important starting point for the synthesis of new novel and bioactive molecules (*de novo* design).
- The chemical space defined by the structural diversity of each data set is mostly defined by NP (**COCONUT**) followed by **DCM**: for compounds and fragments.
- **DCM**-fragments can be used as building-blocks in *de novo* design of bioactive molecules even though the parent compounds have weak biological activity.

Acknowledgements

- **José L. Medina-Franco, PhD** for accepting me into your research group (DIFACQUIM). PhD Supervisor.
- **Norberto Sánchez-Cruz, PhD** for his help with python and sharing his experience and knowledge in cheminformatics.
- **Maria Sorokina, PhD** for developing COlleCtion of Open NatUral producTs (COCONUT) and making it public-available.

DIFACQUIM UNAM

UNAM POSGRADO Ciencias Químicas

CONACYT Consejo Nacional de Ciencia y Tecnología

Scholarship 847870

miztli

Project LANCAD-UNAM-DGTIC-335