Biologically-oriented mud volcano database: muddy_db

Alexei Remizovschi and Rahela Carpa

Department of Molecular Biology and Biotechnology, Faculty of Biology and Geology, Babes-Bolyai University, Cluj-Napoca, Cluj, Romania

ABSTRACT

Mud volcanoes (MVs) are naturally occurring hydrocarbon hotbeds with continuous methane discharge, contributing to global warming. They host microbial communities adapted to hydrocarbon oxidation. Given their research value, MVs still represent a niche topic in microbiology and are neglected by hydrocarbon-oriented research. All the data regarding MVs is sporadic and decentralized. To mitigate this problem, we built a custom Natural Language Processing pipeline (muddy_mine), and collected all the available MV data from open-access articles. Based on this data, we built the muddy_db database. The muddy_db represents the first biologically oriented database rendered as a user-friendly web app. This database includes all the relevant MV data, ranging from microbial taxonomy to hydrocarbon occurrence and geology. The muddy_mine and muddy_db tools are licensed under the GPLv3. muddy_db R Shiny web app: https://muddy-db.shinyapps.io/muddy_db/ muddy_db R package: https://github.com/TracyRage/muddy_db muddy_mine Conda package: https://github.com/TracyRage/muddy_mine.

Subjects Bioinformatics, Microbiology, Soil Science, Environmental Contamination and Remediation

Keywords Mud volcano, PAH, Database, Data mining, Hydrocarbon

Submitted 27 April 2021
Accepted 19 October 2021
Published 9 November 2021

Corresponding author
Rahela Carpa,
rahela.carpa@ubbcluj.ro

Academic editor
James Procter

Additional Information and Declarations can be found on page 6

DOI **10.7717/peerj.12463**

Copyright
2021 Remizovschi and Carpa

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

INTRODUCTION

Mud volcanoes (MVs) represent hydrocarbon discharging landforms (*Mazzini & Etiope, 2017*). They are distributed worldwide in both marine and terrestrial environments (*Milkov, 2000*). The most distinctive feature of MVs is recurrent methane emission. Due to methane emissions, MVs contribute extensively to global warming (*Etiope, Feyzullayev & Baciu, 2009*).

MV genesis is mainly caused by a naturally mediated process - kerogen maturation (*Vandenbroucke & Largeau, 2007*). Therefore, the surrounding area of MVs can provide valuable data regarding both aerobic and anaerobic hydrocarbon microbial oxidation (*Cheng et al., 2012*).

Over the years, MV research has mainly focused on anaerobic oxidation of methane (AOM) and the implicit interaction between sulfate-reducing bacteria and methane oxidizing archaea (ANME) (*Bose et al., 2013*; *Cui et al., 2014*). In addition to AOM research, MVs were also investigated in the context of hydrocarbon research. A myriad of MV studies discussed the thermogenic and biogenic origin of evolved methane (*Etiope, Feyzullayev & Baciu, 2009*; *Sano et al., 2017*).

Despite these studies, the biological aspect of MVs is still a niche and unexplored topic. The biological data regarding MVs are sporadic and mostly biased towards AOM. Even worse, the already available data is not centralized. An MV-dedicated database would facilitate this, however, making it easier for researchers to conduct comparative studies.

Meanwhile, mainstream biomedical fields have extensively employed natural language processing (NLP) techniques to mine meaningful data from research articles (*Wang et al., 2020*). Simultaneously, the number of databases related to biomedical fields is considerable (*Luo et al., 2016*). Niche environmental science fields have not entirely caught up. This lack of tools limits the possibility to mine environmental-oriented articles and build field-specific databases, delaying the publication of the meta-analyses or any comparative studies.

Fortunately, democratic NLP models and tools have been published over the last years. Some of them can be easily used by environmental scientists with limited computer science (CS) experience, for example, the spaCy library, ScispaCy models, and S2ORC database (*Honnibal & Johnson, 2015*; *Neumann et al., 2019*; *Lo et al., 2020*).

These advancements in NLP provide opportunities for consolidating and promoting niche environmental topics such as MV microbiology.

In this paper, we used them to build the first biologically oriented mud volcano database, muddy_db. This niche database consolidates all relevant biological and environmental data, that will be of great use for researchers specializing in bacterial hydrocarbon oxidation or MV microbiology. Our pipeline can serve as a methodological blueprint for other research communities interested in employing NLP to build their specialized databases.

METHODS

To collect all the available data regarding the biological aspects of MVs, we exclusively relied on open-access articles. This minimized potential issues arising from both copyright concerns and the lack of standardized XML encoding of open-access literature.

We then built a custom pipeline muddy_mine to extract all the biologically oriented tokens, including taxonomy-, chemicals-, geology-, and MV-specific terms (Fig. 1). Tables generated by muddy_mine are presented *via* the muddy_db framework (Fig. 2).

Data collection

We used the S2ORC (20200705v1) database to collect open-access articles. S2ORC represents a centralized database that includes 12.7 million articles with a fully preserved paper structure. S2ORC is quite comprehensive and includes niche environment science articles (*Lo et al., 2020*). Given these facts, we extracted all the available MV-related titles ($N = 118$ total, $N = 115$ deduplicated) from the S2ORC.

Token extraction and muddy_mine pipeline

Having MV articles, we proceeded with token extraction (*i.e.,* the extraction of the terms of interest) using the muddy_mine pipeline.

Taxonomy extraction represented a difficult challenge due to the fact that we intended to collect as many tokens as possible. To overcome this problem, we used the spaCy library

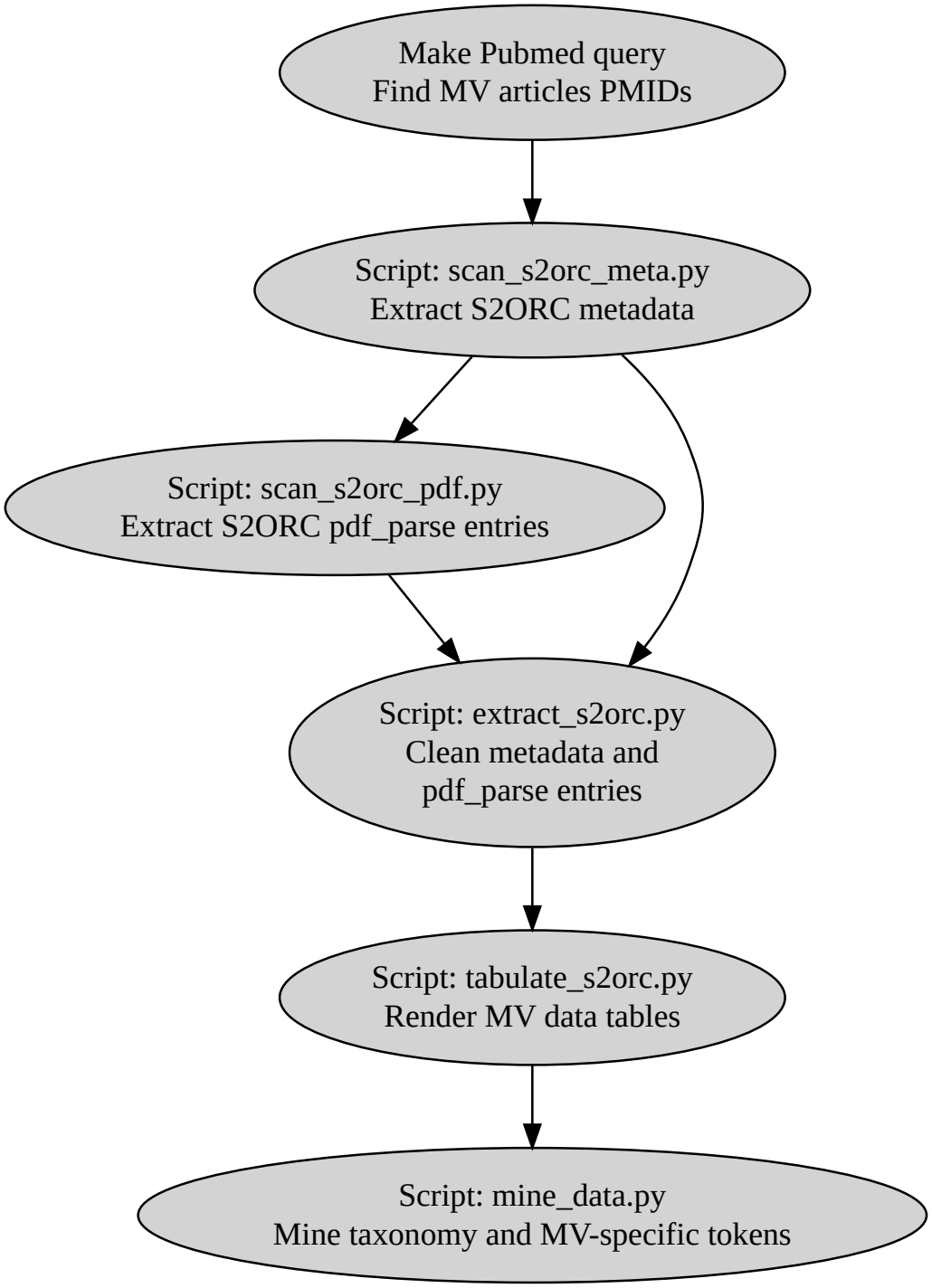

Figure 1 **muddy_mine-pipeline used to build the muddy_db database.** MV, mud volcano; PMIDs, Pubmed IDs.

Full-size DOI: 10.7717/peerj.12463/fig-1

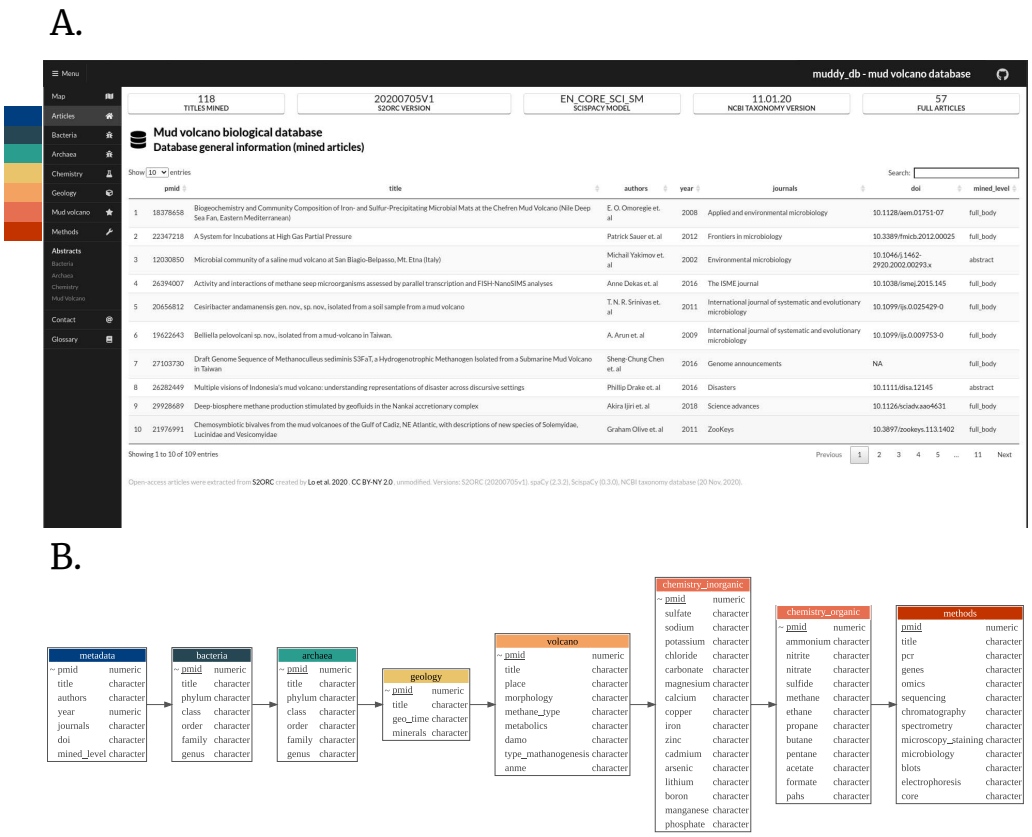

Figure 2 **(A) muddy_db general appearance. (B) muddy_db schema.** For each muddy_db tab, there is a reciprocal table in the database (colorwise). Pubmed ID (PMID), database primary key.
Full-size DOI: 10.7717/peerj.12463/fig-2

(2.3.2), third-party ScispaCy NLP models (0.3.0), and the most recent NCBI Taxonomy database (20 November, 2020) (*Honnibal & Johnson, 2015*; *Neumann et al., 2019*; *Schoch et al., 2020*). First, we extracted all the taxon tokens using en_core_sci_sm ScispaCy model (2.2.5). Second, we checked those tokens against a local NCBI Taxonomy database. Third, we counted the extracted tokens. The more often a token occurs, the counting number, the more likely the token was explicitly discussed in the article.

By iterating this three-step algorithm (each iteration being focused on a specific taxonomic rank), we managed to centralize MV-specific taxonomy on all the possible levels: phylum, class, order, family, and genus.

The same procedure was applied to mine other types of information. We extracted and counted the tokens related to the following categories: chemistry (inorganic ions, hydrocarbons), geology (geological periods, minerals), MV terminology (ANME, methanogenesis type), and experimental methods (PCR types, amplified genes, chromatography). The full list can be found in the muddy_db repository.

The raw output of the muddy_mine pipeline is a set of csv tables with MV data.

Building muddy_db database

After obtaining muddy_mine raw output, we can advance to the next step - building a web based database interface to present the output in a user-friendly manner. To this end, we created a Shiny web app, entitled muddy_db. In order to build it, we used the following R packages: shiny (1.5.0), semantic.dashboard (0.2.0), and golem (0.2.1) (*Filip & Igras, 2021*; *Chang et al., 2021*; *Fay et al., 2021*). This app includes all the output generated by the muddy_mine pipeline. Specifically, it displays tokens and their counts, extracted both from the article bodies ($N = 57$) and abstracts ($N = 115$). Additionally, we added annotated map, which displays the geographical distribution of MVs and their affiliated research metadata. To build it, we used leaflet package (2.0.4.1) (*Cheng, Karambelkar & Xie, 2021*).

System requirements

The muddy_mine pipeline was designed to run on systems with modest memory requirements. We achieved this feature by using Python generators (*Van Rossum & Drake, 2009*). Intel Core i3 (3rd Gen) 3217U / 1.8 GHz processor (Intel, USA) and 4GB RAM system was used to build muddy_db. The mining process took 24 h.

RESULTS

The aim of the muddy_db is to gather all the available MV biologically relevant data and include it in a user-friendly database (Fig. 2). First, we collected all the known taxa associated with MVs. The muddy_db includes data regarding archaeal and bacterial taxonomy on all the possible taxonomy levels. Second, we gathered information regarding metabolic pathways, geology, and chemical substrate availability. Together, these conveniently centralized taxonomic and physicochemical data can facilitate routine MV-related documentation. Additionally, muddy_db contains data regarding experimental methods, applied in the context of MV studies, which can guide specialists to implement appropriate research strategies.

DISCUSSION

MVs are considered to be one of the settings where early life evolved (*Pons et al., 2011*). They sustain a plethora of bacterial metabolic pathways, ranging from methane oxidation and synthesis to sulphate reduction (*Kleindienst et al., 2014*; *Cheng et al., 2012*). These pathways and their affiliated microbial communities could provide valuable data regarding(1) origin of microbial life, (2) the effect of the naturally occurring methane discharging systems on global warming, (3) the contribution of microbial consortia to oil souring (*Gieg, Jack & Foght, 2011*; *Etiope, Feyzullayev & Baciu, 2009*; *Pons et al., 2011*). The accumulation of MVs data could enhance knowledge regarding topics that range from fundamental studies to ecology and engineering.

Given these facts, MVs should be the main focus of hydrocarbon-oriented research and ecology. Unfortunately, data regarding the biological aspects of MVs are scarce. Additionally, the data already gathered are not combined in a dedicated database. The creation of a specialized MV database will help future research efforts, and thus help raise the profile of MV microbiology research in environmental science

Biomedical fields have always represented the cutting-edge subset of natural science, which actively implement CS techniques, and are tightly intertwined with big data (*Luo et al., 2016*). Simultaneously, the implementation of CS methods in niche environmental fields lags. With the muddy_mine NLP pipeline and muddy_db database, we have both advanced the application of CS methods in environmental contexts and addressed a critical lack in the toolbox of MV microbiologists.

The creation of muddy_db tools aims to create a platform, that would provide sufficient data to perform meta-analyses or comparative studies of the MVs. We hope that muddy_db would facilitate the discovery of atypical taxonomic patterns and point out the influence of chemical substrate and geography on MV microbial characteristics. For example, muddy_db offers data regarding MV essential dichotomies such as the predominance of biogenic/thermogenic methane, presence of organic acids/hydrocarbons, and terrestrial/marine localization. All of these conjugated parameters influence the microbial distribution and metabolic patterns in MV sediments (*Bhattarai, Cassarini & Lens, 2019*; *Lazar et al., 2012*; *Wrede et al., 2012*; *Remizovschi et al., 2020*; *Sano et al., 2017*). Simultaneously, muddy_mine represents a reproducible example of a mining technique applied in the context of environmental studies. The muddy_mine pipeline, however, employs general NLP tools, and so could be used by researchers to mine their own data of interest.

Whilst we have shown muddy_mine and muddy_db to be effective, however, there are some caveats. We designed muddy_mine around the S2ORC repository, which provides a standardized encoding of open access literature. Additional parsers would be needed to ingest other article formats, such as TEI-XML and JATS-XML (*J4R, 2021*; *TEI Consortium, 2021*). Our reliance on S2ORC also ensures that the derived resource, muddy_mine, can be publicly shared without infringing author or journal copyright. This may not be the case for article collections obtained from other sources.

CONCLUSIONS

The muddy_db represents the first biologically oriented mud volcano database. It was designed to provide a comprehensive data corpus that can facilitate mud volcano research and shed light on the topic as a whole. The muddy_db contains data ranging from taxonomy to geology and experimental methods. Simultaneously, the muddy_mine NLP pipeline can serve as an example of accessible implementation of NLP techniques in environmental sciences.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Alexei Remizovschi conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Rahela Carpa conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The muddy_db R Shiny web app is available at: https://muddy-db.shinyapps.io/muddy_db/

The muddy_db R package is available at GitHub: https://github.com/TracyRage/muddy_db

The muddy_mine Conda package is available at GitHub: https://github.com/TracyRage/muddy_mine.

REFERENCES

Bhattarai S, Cassarini C, Lens P. 2019. Physiology and distribution of archaeal methanotrophs that couple anaerobic oxidation of methane with sulfate reduction. *Microbiology and Molecular Biology Reviews* **83(3)**:e00074–18 DOI 10.1128/mmbr.00074-18.

Bose A, Rogers D, Adams M, Joye S, Girguis P. 2013. Geomicrobiological linkages between short-chain alkane consumption and sulfate reduction rates in seep sediments. *Frontiers in Microbiology* **4**:386 DOI 10.3389/fmicb.2013.00386.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. 2021. shiny: web application framework for R. R package version 1.6.0. *Available at https://CRAN.R-project.org/package=shiny*.

Cheng J, Karambelkar B, Xie Y. 2021. leaflet: create interactive web maps with the JavaScript 'Leaflet' Library. R package version 2.0.4.1. *Available at https://CRAN.R-project.org/package=leaflet*.

Cheng T, Chang Y, Tang S, Tseng C, Chiang P, Chang K, Sun C, Chen Y, Kuo H, Wang C, Chu P, Song S, Wang P, Lin L. 2012. Metabolic stratification driven by surface and subsurface interactions in a terrestrial mud volcano. *ISME Journal* **6(12)**:2280–2290 DOI 10.1038/ismej.2012.61.

Cui M, Ma A, Qi H, Zhuang X, Zhuang G. 2014. Anaerobic oxidation of methane: an "active" microbial process. *Microbiology Open* **4(1)**:1–11 DOI 10.1002/mbo3.232.

Etiope G, Feyzullayev A, Baciu C. 2009. Terrestrial methane seeps and mud volcanoes: a global perspective of gas origin. *Marine and Petroleum Geology* **26(3)**:333–344 DOI 10.1016/j.marpetgeo.2008.03.001.

Fay C, Guyader V, Rochette S, Girard C. 2021. golem: a framework for robust shiny applications. R package version 0.3.1. *Available at https://CRAN.R-project.org/package=golem*.

Filip S, Igras K. 2021. semantic.dashboard: dashboard with fomantic UI support for shiny. R package version 0.2.0. *Available at https://CRAN.R-project.org/package=semantic.dashboard*.

Gieg L, Jack T, Foght J. 2011. Biological souring and mitigation in oil reservoirs. *Applied Microbiology and Biotechnology* **92(2)**:263–282 DOI 10.1007/s00253-011-3542-6.

Honnibal M, Johnson M. 2015. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. Association for Computational Linguistics, DOI 10.18653/v1/d15-1162.

J4R. 2021. *NISO JATS4R Peer Review Materials v1.0*. National Information Standards Organization DOI 10.3789/niso-rp-39-2021.

Kleindienst S, Herbst F, Stagars M, von Netzer F, von Bergen M, Seifert J, Peplies J, Amann R, Musat F, Lueders T, Knittel K. 2014. Diverse sulfate-reducing bacteria of the Desulfosarcina/Desulfococcus clade are the key alkane degraders at marine seeps. *ISME Journal* **8(10)**:2029–2044 DOI 10.1038/ismej.2014.51.

Lazar C, John Parkes R, Cragg B, L'Haridon S, Toffin L. 2012. Methanogenic activity and diversity in the centre of the Amsterdam Mud Volcano, Eastern Mediterranean Sea. *FEMS Microbiology Ecology* **81(1)**:243–254 DOI 10.1111/j.1574-6941.2012.01375.x.

Lo K, Wang L, Neumann M, Kinney R, Weld D. 2020. In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, DOI 10.18653/v1/2020.acl-main.447.

Luo J, Wu M, Gopukumar D, Zhao Y. 2016. Big data application in biomedical research and health care: a literature review. *Biomedical Informatics Insights* **8**:BII.S31559 DOI 10.4137/bii.s31559.

Mazzini A, Etiope G. 2017. Mud volcanism: an updated review. *Earth-Science Reviews* **168**:81–112 DOI 10.1016/j.earscirev.2017.03.001.

Milkov A. 2000. Worldwide distribution of submarine mud volcanoes and associated gas hydrates. *Mar Geol* **167(1–2)**:29–42 DOI 10.1016/s0025-3227(00)00022-0.

Neumann M, King D, Beltagy I, Ammar W. 2019. In: *Proceedings of the 18th BioNLP workshop and shared task*. Association for Computational Linguistics, DOI 10.18653/v1/w19-5034.

Pons M, Quitte G, Fujii T, Rosing M, Reynard B, Moynier F, Douchet C, Albarede F. 2011. Early Archean serpentine mud volcanoes at Isua, Greenland, as a niche for early life. *Proceedings of the National Academy of Sciences of the United States of America* **108(43)**:17639–17643 DOI 10.1073/pnas.1108061108.

Remizovschi A, Carpa R, Forray F, Chiriac C, Roba C, Beldean-Galea S., Andrei A, Szekeres E, Baricz A., Lupan I, Rudi K, Coman C. 2020. Mud volcanoes and the presence of PAHs. *Scientific Reports* **10(1)**:1253 DOI 10.1038/s41598-020-58282-2.

Sano Y, Kinoshita N, Kagoshima T, Takahata N, Sakata S, Toki T, Kawagucci S, Waseda A, Lan T, Wen H, Chen A, Lee H, Yang T, Zheng G, Tomonaga Y, Roulleau E, Pinti D. 2017. Origin of methane-rich natural gas at the West Pacific convergent plate boundary. *Scientific Reports* **7(1)**:15646 DOI 10.1038/s41598-017-15959-5.

Schoch C, Ciufo S, Domrachev M, Hotton C, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan J, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020** DOI 10.1093/database/baaa062.

TEI Consortium. 2021. TEI P5: guidelines for electronic text encoding and interchange. P5, version 4.3.0. TEI Consortium. *Available at https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSFLT*.

Van Rossum G, Drake FL. 2009. *Python 3 reference manual.* Scotts Valley: CreateSpace.

Vandenbroucke M, Largeau C. 2007. Kerogen origin, evolution and structure. *Organic Geochemistry* **38(5)**:719–833 DOI 10.1016/j.orggeochem.2007.01.001.

Wang J, Deng H, Liu B, Hu A, Liang J, Fan L, Zheng X, Wang T, Lei J. 2020. Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on PubMed. *Journal of Medical Internet Research* **22(1)**:e16816 DOI 10.2196/16816.

Wrede C, Brady S, Rockstroh S, Dreier A, Kokoschka S, Heinzelmann S, Heller C, Reitner J, Taviani M, Daniel R, Hoppert M. 2012. Aerobic and anaerobic methane oxidation in terrestrial mud volcanoes in the Northern Apennines. *Sedimentary Geology* **263–264**:210–219 DOI 10.1016/j.sedgeo.2011.06.004.