Network Analysis Reveals Spatial Clustering and Annotation of Complex Chemical Spaces: Application to Astrochemistry

Alexander Ruf* and Grégoire Danger

ABSTRACT: How are molecules linked to each other in complex systems? In a proof-of-concept study, we have developed the method mol2net (https://zenodo.org/record/7025094) to generate and analyze the molecular network of complex astrochemical data (from high-resolution Orbitrap MS analysis of H₂O:CH₃OH:NH₃ interstellar ice analogs) in a data-driven and unsupervised manner, without any prior knowledge about chemical reactions. The molecular network is clustered according to the initial NH₃ content and unlocked HCN, NH₃, and H₂O as spatially resolved key transformations. In comparison with the PubChem database, four subsets were annotated: (i) saturated C-backbone molecules without N, (ii) saturated N-backbone molecules, (iii) unsaturated C-backbone molecules without N, and (iv) unsaturated N-backbone molecules. These findings were validated with previous results (e.g., identifying the two major graph components as previously described N-poor and N-rich molecular groups) but with additional information about subclustering, key transformations, and molecular structures, and thus, the structural characterization of large complex organic molecules in interstellar ice analogs has been significantly refined.

INTRODUCTION

The complexity of chemical systems is based on species’ interaction patterns. This leads to emergent collective systems’ properties such self-organization, autocatalysis, or global reactivity, properties that are hidden on the level of individual species.

Astrochemical samples are here of specific relevance, as their outstanding molecular diversities exceed those of other chemical samples (e.g., biochemical or environmental data). The evolution of organic material, from molecular clouds toward planetary systems, can be simulated by laboratory experiments. The benefit of such experiments is two-fold: (i) controlled experimental parameter space and (ii) in-depth chemical analysis (e.g., high-resolving, high accurate and sensitive mass spectrometry). The resulting high-dimensional data sets enable to study the complete diversity of astrochemical organic molecules from a system’s chemistry perspective.

The comprehensive and unbiased analysis of complex chemical systems is a huge challenge and implies constant development of analytical methodology with innovative solutions for their data analyses. Network analysis represents an excellent tool to analyze complex systems as complete as possible, in general or in (astro)chemical context. This approach covers a wide range of chemical applications, including astrochemistry, meteorites, prebiotic chemistry, origin(s) of Life, astrobiology, systems chemistry, autacatalysis, environmental science, or biochemistry. Networks allow one to visualize and further quantify complex interactions as they unveil systemic patterns instead of addressing individual reactions only. They thus shed light into the organization and interconnections of chemical systems as a whole. Even though reaction network analysis has powerfully demonstrated the characterization of chemical systems with partially unknown reaction rates, current reaction network methods have two key drawbacks because of their supervised architecture: (i) individually addressed reactions (side reactions are ignored) and (ii) reaction rate dependence on physical environment (e.g., gas-phase vs solid-state-mediated reactions). In other words, current reaction network analysis depends on prior knowledge about the individual reactions.

In this proof-of-concept study, we developed the method mol2net that has analyzed the interactions of many tens of thousands of different chemical compounds. We applied data-driven graph-based network analysis to astrochemical data as...
an example for complex chemical systems. A previously well-described data set of interstellar ice analogs from high-resolution Orbitrap MS data (without fragmentation information, no MS data required) was used that allowed for validation of this novel method. This data set represents 15 different H2O:CH3OH:NH3 ice samples (3:1:0.2, 3:1:1, 3:1:5, and 10:1:1) that are grouped into three categories: (i) NH3 content, (ii) H2O content, and (iii) additional physical processes (over-irradiation) of the formed residue. The molecular network was generated by mass difference matching of 3270 CHNO molecular formulas (between 199 and 365 amu, atomic mass unit, nodes in the network) with a set of in silico transformations (edges in the network). We want to highlight that no prior knowledge about chemical reactions (e.g., reaction rates) was required. In contrast to the previous studies, we have here not only characterized the compounds independently but we also put emphasis on their interconnectedness by discussing (i) the clustering of the molecular network, (ii) the key transformations, (iii) the subnetworks and PubChem database unlock structural information, and (iv) the annotation of complex chemical spaces.

### RESULTS AND DISCUSSION

**Clustering of the Molecular Network.** Figure 1 illustrates the molecular network of interstellar ice analogs, as whole and zoom-in pictures. Using a transformation set of \{H2, O, CO, NH3\} (named as “minimal set”), 95% of all 3270 molecular formulas could be linked within two graph components (components 0 and 1, with 2357 nodes/5608 edges and 752 nodes/1431 edges, respectively, Supporting Figure 1). The two graph components represent two previously reported distinct N-poor and N-rich molecular groups (Supporting Figure 4). Transformations were chosen in a data-driven and unsupervised manner as informed from prior distance matrix analysis. Our goal was to keep the set of transformations as minimal as possible to maximize interpretation of individual transformations. We have tested six other transformations \{CHN, CH3N, H2O, CO2, CH2O, CHNO\} and found that their step-by-step addition is invariant with respect to the network topology and addition sequence, as tested by (i) network clustering, (ii) shape of degree distributions, and (iii) component frequency (Supporting Figure 3). Thus, the four transformations from the “minimal set” explain the complete connectivity of the analyzed greater than 3000 compounds in the interstellar ice analogs. These transformations represent potential reactions in interstellar ices, based on previous laboratory experiments (H2, O, NH3).

The molecular network is clustered according to the NH3 content (arrow “NH3 content in ice”, NH3 poor (blue), NH3 medium (green), H2O rich (gray), and NH3 rich (red) Figure 1). The trend from NH3 poor to NH3 rich is intuitive but the H2O-rich cluster in between highlights the importance of H2O-mediated organic chemistry inside ices. The over-irradiation cluster (orange, additional irradiation of the resulting ice residue at 300 K) is aligned orthogonal to the “NH3 content gradient”, accompanied by higher DBE (double bond equivalent) and lower #O (decarboxylation/formation of unsaturated, aromatic structures). The same cluster gradient has been observed for component 1 as well and furthermore substructured in each component. This...
Figure 2. Key transformations. Cluster-specific transitions illustrate potential key transformations: CHN, NH₃ and H₂O (with directions) along the NH₃ content gradient in the initial ice (NH₃-poor → NH₃-medium cluster, NH₃-medium → H₂O-rich cluster and H₂O-rich → NH₃-rich cluster).
parameters (e.g., \#H, \#C, or DBE). Each of the four groups for the selected five key components (minimal set - H$_2$, minimal set - O, minimal set - CO, minimal set - NH$_3$) consist of 24–253 molecular formulas and show distinct chemical properties (Figure 3A, Supporting Figure 8). The “minimal set - H$_2$” and “minimal set - NH$_3$” subnetworks specifically show components that are distinct in \#N (e.g., N$_2$ or N$_{10}$ subsets), whereas “minimal set - O” and “minimal set - CO” subnetworks have distinct \#C components (e.g., C$_5$ or C$_{16}$ subsets). Furthermore, the components also cluster with respect to the original ice samples (e.g., NH$_3$-enriched component “minimal set - O”). All four splitted “minimal set” graphs commonly show distinct \#O and in DBE subsets.

The comparison with the PubChem database has revealed further insights into molecular structures and compound classes. We have applied the same workflow for network generation and analysis for the PubChem data as done for those of the ice (transformation set of {H$_2$, O, CO, NH$_3$}, “minimal set” of transformations) for the PubChem network (details about the PubChem data are given in the Supporting Information), unraveling their compound class characteristics.

**Annotation of Complex Chemical Spaces.** All combined results were mapped onto the original molecular network of interstellar ice analogs (Figure 4). This reveals spatial clustering and annotation of the network. The two major graph components were thus organized into different clusters that show distinct elemental characteristics (e.g., C or N counts, inferred from splitting the “minimal set” of transformation) and were assigned with compound class likelihoods (inferred from PubChem database). We further show that a gradient in N- vs C-backbone molecules in each component highlights the competitive two different types of chemistry, namely, the carbon- and nitrogen-driven ones (in relation to N/C ratios). Four subsets were identified: (i) saturated C-backbone molecules without N (e.g., open-chain compounds; component 0, left), (ii) saturated N-backbone molecules (e.g., cyclic secondary amines; component 0, right), (iii) unsaturated C-backbone molecules without N (e.g., polycyclic aromatic hydrocarbons, PAHs; component 1, bottom), and (iv) unsaturated N-backbone molecules (e.g., aromatic N-heterocycles; component 1, top). In short, component 0 represents aliphatic molecules and component 1 unsaturated (potentially aromatic) molecules, with both C- and N-backbones.

Component 0 represents saturated C-backbone molecules and saturated N-backbone ones. From left to right, nitrogen-based chemistry dominates over carbon chemistry, that is supported by N/C ratios. On the total left, a C-backbone cluster with C$_{12}$ unsaturated molecules was identified (DBE = 8).
8, N/C = 0), whereas on the right, a N-backbone cluster represents N₂, saturated molecules (DBE = 3, N/C = 2.5). Component 0 molecules might be specifically enriched in the number of amides (15%), esters (11%), and acids (9%).

Component 1 shows unsaturated C- and N-backbone molecules with low #H and #O. N-Based molecules are enriched from bottom to top. On the bottom, a C-backbone cluster with N₉ unsaturated molecules was identified (DBE = 8, N/C = 0.5), whereas on top, a N-backbone cluster shows N₁₃ unsaturated molecules (DBE = 8, N/C = 3). Component 1 molecules might be specifically enriched in the number of ketones (42%), amines (17%), and nitriles (5%), as well as in CH₃ (e.g., polycyclic aromatic hydrocarbons, PAHs) and N-based aromatic compounds (e.g., N-heterocycles) where the latter were almost absent in component 0.

In comparison to previous studies by Fresneau et al. (based on average values for atomic parameters) and Gautier et al. (based on ternary diagrams of atomic-ratios), the here presented network analysis sets emphasis species' interactions, and both confirm but also refine current knowledge. The previously reported two distinct molecular groups were validated, but the network analysis unlocked variation (Supporting Figure 4). Whereas Gautier et al. reported that NH₃-rich ices “lead to a group of unsaturated molecules in the final residue, while H₂O rich ices lead to saturated ones”, but networks show that there are both saturated and unsaturated molecules for both NH₃- and H₂O-rich ices.

**Conclusions.** In a proof-of-concept study, we have analyzed the network of complex astrochemical data (from high-resolution Orbitrap MS³ analysis of H₂O:CH₃OH:NH₃ interstellar ice analogs). The molecular network is clustered according to the initial NH₃ content and unlocked HCN, NH₃, and H₂O as spatially resolved key transformations. In comparison with the PubChem database, four subsets were structurally annotated. This method was validated with previous results by Fresneau et al. and Gautier et al. Current knowledge was confirmed but also refined (e.g., characterization of two distinct molecular groups).

The implications of the presented network analysis are not specific to astrochemistry but manifold for complex chemical systems, including (i) the characterization of nondetected species in any sample, (ii) the comprehensive elucidation of chemical reactivity (e.g., radical vs thermal chemistry), or (iii) studying of emergence behavior. For astrochemistry, these findings provide the very first insights about transformation patterns and structures of large complex organic molecules in interstellar ice analogs (>200 amu, those of prebiotic interest). To date, information about large molecules’ chemistries is hidden to traditional chemical models as detailed information on their individual reaction rates and mechanisms is missing. Even though the here newly described graph-based network analysis does not directly address mechanistic insights, it acts as a screening tool to unveil all possible transformation pathways that can now be further investigated by upcoming
targeted experimental and/or computational studies. Furthermore, the identified transformation patterns might be generalized that would allow the characterization of non-detected species as well. This has strong implications for astrochemistry to probe the chemistries of large complex organic molecules that would otherwise be hidden to observations.

■ DATA SET AND METHODS

Data Set of Interstellar Ice Analogs. This data set represents 15 different H$_2$O:CH$_3$OH:NH$_3$ ice samples: (3:1:0.2, blue, 3 replicates; 3:1:1, green, 3 replicates; 1:1:5, red, 2 replicates; 10:1:1, gray, replicates/over-irradiation, orange, 3 replicates). Ice deposition/irradiation time was 48 h except for the “3:1:1 16 h” ice with a deposition/irradiation time of 16 h. All details about the experiments and their high resolution Orbitrap mass spectrometric analysis have been described by Fresneau et al. and Gautier et al. Data will be made available on request.

Method: Molecular Network Generation and Analysis. The algorithmic workflow of the method mol2net is organized as follows (Figure 5): (1) input, (2) identify transformations, and (3) network analysis (using NetworkX).

1. INPUT
   - List of detected molecules
   - List of transformations

2. IDENTIFY TRANSFORMATIONS
   - Check if $n_i = n_j = g_i$ with $(n_i, n_j) \in \text{NODES}, g_i \in \text{GUESS}$
   - Transformations (combinatorial initial guess reactions)

3. NETWORK ANALYSIS (using NetworkX)
   - Filter graph components
   - Degree distributions
   - Network statistics
   - Key transformation patterns
   - Network layout

Figure 5. Schematic algorithmic workflow of the method mol2net: (1) input, (2) identify transformations, and (3) network analysis (using NetworkX).

PubChem Database. The PubChem database was downloaded via the Python package pubchempy (https://pubchempy.readthedocs.io/en/latest/). We have filtered the downloaded data for the chemical elements (CHNO) and mass range (199 and 365 amu) as present in the ice data set. Two subsets with 10,000 compounds each were randomly sampled (1–10,000 and 10,001–20,000; details about CID identifiers are given in the Supporting Information). Compound classes have been defined by data mining their IUPAC name endings as following: CH_aromatics = classes.str.endswith(‘benzene’), CH_aromatics = classes.str.endswith(‘naphthalene’), CH_aromatics = classes.str.endswith(‘pyrene’), Acids = classes.str.endswith(‘acid’), Alcohols = classes.str.endswith(‘ol’), Ketones = classes.str.endswith(‘one’), Esters = classes.str.endswith(‘oate’), Esters = classes.str.endswith(‘acetate’), Amides = classes.str.endswith(‘amide’), Amides = classes.str.endswith(‘carbamate’), Nitriles = classes.str.endswith(‘nitrile’), Amines = classes.str.endswith(‘amine’), N cyclic aliphatics = classes.str.endswith(‘piperidine’), N aromatics = classes.str.endswith(‘imidazole’), N aromatics = classes.str.endswith(‘quinoline’), N aromatics = classes.str.endswith(‘triazine’), N aromatics = classes.str.endswith(‘indole’), N aromatics = classes.str.endswith(‘oxazole’). Based on qualitative element map comparison, detailed analyses were performed with the PubChem subset 1–10,000 only.

Data and Code Availability. The Python code of mol2net used in this study is available on Zenodo and GitHub. Data will be made available on request.

■ ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c01271.
Several authors and institutions contributed to the research described in the document. The authors and their affiliations include Alexander Ruf, Thomas Javelle, and Fabrice Duvernay. The work was supported by various grants and fellowships, including the Excellence Strategy EXC 2000 and the Centre National d’Etudes Spatiales. The authors acknowledge contributions from other researchers and institutions, and express their gratitude for the support received.

The references cited in the document cover a range of topics, from astrochemistry to molecular analysis, and include notable works by authors such as Adam S., Fahrenbach, A.; Jacobson, S. M.; Kacar, B.; and Zubarev, D. Y. The specific references are too numerous to list here, but they provide a comprehensive overview of the research and its context.

The document also includes a section on ACKNOWLEDGMENTS, where the authors thank various individuals and organizations for their support and contributions. This includes the Centre National d’Etudes Spatiales, the Department of Chemistry and Pharmacy, Ludwig-Maximilians-University, and the Deutsche Forschungsgemeinschaft.

In conclusion, the document is a detailed piece of research that highlights the interdisciplinary nature of the scientific community and the collaborative effort required to advance knowledge in complex fields such as astrochemistry and molecular analysis.
Chiavassa, T.; Le Sergeant d’Hendecourt, L. Mon. Not. R. Astron. Soc. 2017, 464, 114–120.
(40) Ruf, A.; Bouquet, A.; Schmitt-Kopplin, P.; Boduch, P.; Mousis, O.; Danger, G. Astronomy & Astrophysics 2021, 655, A74.
(41) Van Krevelen, D. Fuel 1950, 29, 269–84.
(42) Rimola, A.; Skouteris, D.; Balucani, N.; Ceccarelli, C.; Enrique-Romero, J.; Taquet, V.; Ugliengo, P. ACS Earth and Space Chemistry 2018, 2, 720–734.
(43) Gerakines, P.; Moore, M. H.; Hudson, R. L. Astron. Astrophys. 2000, 357, 793–800.
(44) Zheng, W.; Kaiser, R. I. Chem. Phys. Lett. 2007, 450, 55–60.
(45) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. Nucleic Acids Res. 2016, 44, D1202–D1213.
(46) Wakelam, V.; Smith, I. W. M.; Herbst, E.; Troe, J.; Geppert, W.; Linnartz, H.; Oberg, K.; Roueff, E.; Agundez, M.; Pernot, P.; Cuppen, H. M.; Loison, J. C.; Talbi, D. Space science reviews 2010, 156, 13–72.
(47) Wakelam, V.; Herbst, E.; Loison, J.-C.; Smith, I. W. M.; Chandrasekaran, V.; Pavone, B.; Adams, N. G.; Bacchus-Montabonel, M.-C.; Bergeat, A.; Beroff, K.; Bierbaum, V. M.; Chabot, M.; Dalgaro, A.; van Dishoeck, E. F.; Faure, A.; Geppert, W. D.; Gerlich, D.; Galli, D.; Hebrard, E.; Hersant, F.; Hickson, K. M.; Honvault, P.; Klippenstein, S. J.; Le Picard, S.; Nyman, G.; Pernot, P.; Schlemmer, S.; Selsis, F.; Sims, I. R.; Talbi, D.; Tennison, J.; Troe, J.; Wester, R.; Wiesenfeld, L. Astrophysical Journal Supplement Series 2012, 199, 21.
(48) McElroy, D.; Walsh, C.; Markwick, A. J.; Cordiner, M. A.; Smith, K.; Millar, T. J. Astronomy & Astrophysics 2013, 550, A36.
(49) Cuppen, H. M.; Walsh, C.; Lamberts, T.; Semenov, D.; Garrod, R. T.; Penteado, E. M.; Ioppolo, S. Space Science Reviews 2017, 212, 1–58.
(50) Van Rossum, G.; Drake, F. L., Jr. Python Tutorial; Centrum voor Wiskunde en Informatica Amsterdam, 1995; Vol. 620.
(51) Hagberg, A.; Swart, P.; S Chult, D. Exploring Network Structure, Dynamics, and Function Using NetworkX; U.S. Department of Energy, 2008.