

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Low-Dimensional Discrete Reaction Networks for Chemical Space Representation and Reaction Pathway Prediction

Permalink

<https://escholarship.org/uc/item/3230461c>

ISBN

9798197827463

Author

Stulajter, Miko Milano

Publication Date

2026-07-02

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-ShareAlike License, available at <https://creativecommons.org/licenses/by-sa/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, IRVINE
&
SAN DIEGO STATE UNIVERSITY

Low-Dimensional Discrete Reaction Networks for Chemical Space
Representation and Reaction Pathway Prediction

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in Computational Science

by

Miko Stulajter

Dissertation Committee:

Professor Filipp Furche, Advisor
Professor Ricardo Carretero, Co-Advisor
Professor Kieron Burke
Assistant Professor Kyle Hasenstab
Associate Professor Asaf Ferber

DEDICATION

To the unwavering love and support of my family, girlfriend, friends, and loved ones.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	vi
LIST OF TABLES	viii
ACKNOWLEDGMENTS	ix
VITA	xi
ABSTRACT OF THE DISSERTATION	xiv
1 Introduction	1
1.1 Background	3
1.2 Key Concepts	6
1.2.1 Potential Energy Surfaces	6
1.2.2 Reaction Networks	6
1.2.3 Metric and Euclidean Spaces	8
1.2.4 Reaction Pathway Feasibility: arc and karc Criteria	9
1.3 Workflows and Optimizations in colibri2	11
1.3.1 Workflows	11
1.3.2 Optimizations and Improvements	13
1.4 Research Goal	14
1.5 Research Overview	14
1.6 Summary and Significance	16
2 Low-Dimensional Regular Lattice Structure	18
2.1 Introduction	20
2.2 Methods	24
2.2.1 Reaction Networks	24
2.2.2 Network Properties	25
2.2.3 Generative Models	27
2.3 Results	28
2.3.1 Reaction Network Properties	28
2.3.2 Comparisons with Generative Models	34
2.4 Discussion	43
2.4.1 Reaction Network Characterization	43

2.4.2	Comparisons to Generative Network Models	46
2.5	Conclusions	49
3	Euclidean Embeddings of Reaction Networks	51
3.1	Introduction	53
3.2	Methods	56
3.2.1	Reaction Network Representation	56
3.2.2	Conventional Embedding Methods and Error Metrics	58
3.2.3	Distance-Preserving Encoder	60
3.3	Results	63
3.3.1	Linear Methods	63
3.3.2	Nonlinear Methods	65
3.3.3	Neural Network-Based Encoder Embedding	68
3.3.4	Optimal Embedding Dimension	70
3.3.5	Structural Origins of Embeddability	72
3.3.6	Pathway Discovery via Embedding	74
3.4	Discussion	80
3.4.1	Structural Origins and Dimensionality of RN Embeddings	80
3.4.2	Embedding Methods: Performance and Scalability	83
3.4.3	Reaction Pathway Discovery	84
3.5	Conclusions	86
4	Utilizing On-the-Fly Energy Calculations to Construct Bounded Reaction Networks	88
4.1	Introduction	90
4.2	Methods	93
4.2.1	Reaction Network Representation	93
4.2.2	Bounded Reaction Network Generation	96
4.3	Results	99
4.3.1	Network Scaling with Bounded Generation	99
4.3.2	Reaction Discovery in Bounded Networks	104
4.4	Discussion	115
4.4.1	Network Reduction and Bounded Generation	115
4.4.2	Radical Transformation Rules Extension	118
4.4.3	Predicting Reaction Pathways	119
4.5	Conclusions	121
5	Conclusion	123
5.1	Future Work	125
5.2	Concluding Remarks	127
	Bibliography	128
	Appendix A Supplementary Material for Chapter 2	147
	Appendix B Supplementary Material for Chapter 3	164

LIST OF FIGURES

	Page
1.1 Reaction network for two oxygen, one carbon, and four hydrogen atoms starting with formaldehyde and water as the initial reactants.	7
2.1 Variation of reaction network properties with the number of atoms.	29
2.2 Histograms of node degrees and square clustering coefficients in the $C_3H_{19}NO_9$ reaction network.	32
2.3 Variation of reaction network properties with the number of nodes.	33
2.4 Comparison of reaction network properties with generative network models.	35
2.5 Fractal dimension against average square clustering coefficient for periodic and non-periodic regular lattices with rewiring and reaction networks.	42
2.6 Numbers of network edges grouped by transformation rule in the C_2H_3NO homologous series for the C_2H_3NO , $C_2H_9NO_4$, and $C_2H_{33}NO_{16}$ RNs.	43
3.1 Architecture of the distance-preserving encoder.	61
3.2 Comparison of embedding quality metrics for linear and nonlinear embedding methods.	67
3.3 Normalized stress convergence during encoder training for the $C_2H_6N_2O_2$ reaction network.	69
3.4 Embedding distortion for the $CH_{19}NO_9$ reaction network under randomization.	73
3.5 Feasible reaction pathway for enol tautomerization.	75
3.6 Feasible reaction pathway for HCN trimerization.	76
3.7 Feasible reaction pathway for esterification.	77
3.8 Feasible reaction pathway for C3 aldol addition.	78
3.9 Feasible reaction pathway for C4 aldol addition.	79
3.10 Optimal embedding dimension and stress as a function of the number of nodes.	82
4.1 Workflow diagram of the bounded reaction network generation pipeline in <code>colibri2</code>	97
4.2 Scaling of nodes and edges with bounded generation in the CH_4O_3 reaction network.	100
4.3 Scaling of nodes and edges with bounded generation for different starting nodes in the CH_4O_3 reaction network.	102
4.4 Predicted reaction pathways in the O_3 and O_4H_2 reaction networks.	105
4.5 Predicted reaction pathways for decay of the Criegee intermediate CH_2OO to formic acid in the CH_2O_2 reaction network.	107

4.6	Predicted reaction pathways for CH_3OOOH formation and decomposition in the CH_4O_3 reaction network.	109
4.7	Predicted reaction pathways in the CH_4O_3 reaction network starting from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$	111
4.8	Predicted reaction pathways for ethylene ozonolysis in the $\text{C}_2\text{H}_4\text{O}_3$ reaction network.	114
4.9	Cumulative edge energy distribution and accessible nodes retained under bounded generation in the CH_4O_3 reaction network.	116

LIST OF TABLES

	Page
1.1 Workers in <code>colibri2</code> for network generation without energy calculations. . .	11
1.2 Workers in <code>colibri2</code> for network generation with energy calculations.	12
2.1 Polar CHNO transformation rules used in this chapter.	25
2.2 Selected properties of comparison reaction networks.	35
2.3 Selected properties of size-matched Erdős–Rényi networks.	36
2.4 Selected properties of size-matched bipartite random networks.	37
2.5 Selected properties of size-matched Watts–Strogatz networks.	38
2.6 Selected properties of non-periodic regular lattices.	39
2.7 Selected properties of periodic regular lattices.	40
2.8 Selected properties of periodic and non-periodic regular lattices with rewiring.	41
3.1 Embedding quality metrics for classical and landmark multidimensional scaling.	64
3.2 Embedding quality metrics for metric multidimensional scaling and the distance-preserving encoder.	66
3.3 Optimal embedding dimensions, fractal dimensions, and potential energy surface dimensions for test reaction networks.	71
4.1 Bounded reaction network statistics at generation depths 3 and 7.	103
4.2 For each reactant–product pair, the smallest bounded O_3 and O_4H_2 reaction networks that retain the lowest cost pathway.	106
4.3 For each reactant–product pair, the smallest bounded CH_4O_3 reaction networks that retain the lowest cost pathway.	110

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my family, girlfriend, friends, and loved ones, who have been pillars of strength throughout my PhD journey.

To my loving parents, Maria and Milan, and my sister, Tina, your unwavering encouragement and belief in me have been the driving force behind every step I have taken. To my girlfriend, Rochelle, your steadfast faith in my abilities and your constant understanding have been my foundation, guiding me through even the most difficult times.

To my friends, thank you for your constant presence, encouragement, and uplifting spirit, which have made even the most challenging days brighter. I am especially grateful for the long board game nights that allowed me to unwind, as well as for the many conversations where you listened, offered ideas, and helped me think through difficult problems.

I would also like to extend my heartfelt thanks to everyone who stood by me during my battle with cancer in January 2024. Your love, support, and encouragement carried me through one of the most difficult periods of my life and reminded me that I was never alone. This work stands as a testament to that strength and support, and I am deeply grateful to each and every one of you.

I would like to sincerely thank my advisor, Professor Filipp Furche, for his guidance, support, and mentorship throughout my PhD. I am also especially grateful to Dmitrij Rappoport, my unofficial official advisor, whose mentorship has profoundly shaped both this work and my perspective on life.

Finally, I would like to thank my committee members: Ricardo Carretero, Kieron Burke, Kyle Hasenstab, and Asaf Ferber, for their guidance, support, and valuable insights throughout my doctoral studies. I feel incredibly fortunate to have benefited from such strong mentorship.

This dissertation and contained works were supported by funding from the National Science Foundation (CHE-2227112 and DUE-1930546). Part of this work utilized the HPC3 high-performance computing cluster operated by the Research Cyber infrastructure Center (RCIC) at the University of California Irvine (UCI).

Chapter 2 of this dissertation is a reprint of the material as it appears in “Reaction Networks Resemble Low-Dimensional Regular Lattices” in the Journal of Chemical Theory and Computation, used with permission from ACS Publications [1]. The coauthor listed in this publication is Dmitrij Rappoport, who directed and supervised research which forms the basis for Chapter 2.

Chapter 3 of this dissertation is material that has been submitted for publication as “Euclidean Embedding of Reaction Networks for Dimensionality Reduction of Chemical Spaces” in the Journal of Chemical Theory and Computation. The coauthors listed in this publication are Liliana Garcia, Kent Ziti, and Dmitrij Rappoport. Dmitrij Rappoport directed

and supervised research which forms the basis for Chapter 3. Liliana Garcia and Kent Ziti, undergraduate researchers, contributed to the initial constant edge weight embedding study and computational method development, respectively.

During the preparation of this dissertation, the author used large language model tools to assist with proofreading, grammar correction, and improving readability. These tools were not used to generate scientific results, perform data analysis, interpret findings, or formulate conclusions. All scientific content, analyses, and interpretations presented in this dissertation were developed, verified, and approved by the author.

VITA

Miko Stulajter

EDUCATION

Doctor of Philosophy in Computational Science	June 2026
Joint: University of California Irvine	<i>Irvine, CA</i>
San Diego State University	<i>San Diego, CA</i>
Masters in Materials Science and Engineering	December 2020
University of Pennsylvania	<i>Philadelphia, PA</i>
B.S. Double Major in Computer Science & Mathematics, Physics	June 2018
DePaul University	<i>Chicago, IL</i>

RESEARCH EXPERIENCE

Graduate Research Assistant	2022 – 2026
University of California, Irvine	<i>Irvine, California</i>
Graduate Research Intern	2022 – 2023
Lawrence Berkeley National Laboratory	<i>Berkeley, CA</i>
Graduate Research Assistant	2021 – 2023
San Diego State University	<i>San Diego, CA</i>
Graduate Research Assistant	2019 – 2020
University of Pennsylvania	<i>Philadelphia, PA</i>
Undergraduate Research Assistant	2016 – 2019
DePaul University	<i>Chicago, IL</i>

PROFESSIONAL EXPERIENCE

Computational Scientist Intern	2021 – 2026
Predictive Science, Inc.	<i>San Diego, California</i>
Perception Systems Engineering Intern	2025
Nvidia	<i>Santa Clara, California</i>
Software Engineer Intern	2023, 2024
Dassault Systèmes	<i>San Diego, California</i>

TEACHING EXPERIENCE

Teaching Assistant

University of California Irvine

Software Design 1

Data Structure Implementation and Analysis

Data Structure Implementation and Analysis

2023 – 2025

Irvine, California

Spring 2025

Spring 2024

Spring 2023

JOURNAL PUBLICATIONS

- **Stulajter, M. M.**, Garcia, L., & Rappoport, D. (2026) “Optimizing Colibri2 via On-the-Fly Energy Calculations: Applications to Atmospheric Radical Chemistry.” *In preparation*.
- **Stulajter, M. M.**, Garcia, L., Ziti, K., & Rappoport, D. (2026) “Euclidean Embedding of Reaction Networks for Dimensionality Reduction of Chemical Spaces.” *Manuscript submitted*.
- Downs C., et al. (2025) “A near-real-time data-assimilative model of the solar corona.” *Science*.
- Caplan, R. M., **Stulajter, M. M.**, et al. (2025). “Open-source Flux Transport (OFT). I. HipFT – High-performance Flux Transport.” *The Astrophysical Journal Supplement Series*.
- González, G. B., et al. (2025) “Local structure of zinc–indium–tin oxide films via grazing-incidence x- ray pair-distribution functions and theoretical methods.” *Journal of Applied Physics*.
- Gupta, A. K., **Stulajter, M. M.**, Shaidu, Y., & de Jong, W. A. (2024). “Equivariant Neural Networks Utilizing Molecular Clusters for Accurate Molecular Crystal Lattice Energy Predictions.” *ACS Omega*.
- **Stulajter, M. M.**, & Rappoport, D. (2024) “Reaction Networks Resemble Low-Dimensional Regular Lattices.” *Journal of Chemical Theory and Computation*.
- Ryan, J., **Stulajter, M. M.**, Okasinski, J. S., Cai, Z., Gonzalez, G. B., & Stock, S. R. (2020). “Carbonated apatite lattice parameter variation across incremental growth lines in teeth.” *Materialia*.

CONFERENCE PUBLICATIONS

- Caplan, R. M., **Stulajter, M. M.**, et al. “Portability of Fortran’s ‘do concurrent’ on GPUs.” *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024.
- Caplan, R. M., **Stulajter, M. M.**, & Linker, J. A. “Acceleration of a production Solar MHD code with Fortran standard parallelism: From OpenACC to ‘do concurrent’?” *2023 IEEE International Parallel and Distributed Processing Symposium Workshops*, 2023.
- **Stulajter, M. M.**, Caplan, R. M., & Linker J. A. “Can Fortran’s ‘do concurrent’ replace directives for accelerated computing?” *SC21-W: International Workshop on Accelerator Programming Using Directives*, 2021.

PRESENTATIONS

- **Stulajter, M. M.**, Rappoport D., & Furche, F. “Colibri2: Distributed HPC Software for Exploring Chemical Space.” *ACSESS 2026*, San Diego, California, 2026.
- **Stulajter, M. M.**, Garcia, L., & Rappoport D. “Colibri2: Distributed HPC Software for Building Chemical Reaction Networks.” *SIAM Conference on Parallel Processing for Scientific Computing*, Berlin, Germany, 2026.

- **Stulajter, M. M.**, Garcia, L., & Rappoport D. “Exploring Chemical Space Through Reaction Networks.” *NetSci 2025*, Maastricht, Netherlands, 2025.
- **Stulajter, M. M.**, Rappoport D., & Furche, F. “Restoring Coordinates: Embedding Reaction Networks into Euclidean Space.” *ACSESS 2025*, San Diego, California, 2025.
- **Stulajter, M. M.**, & Rappoport, D. “Reaction Networks Resemble Low-Dimensional Regular Lattices.” *ACS Fall 2024*, Denver, Colorado, 2024.
- Caplan, R. M., & **Stulajter, M. M.** “SWQU Workshop Day 2 and 3: Open-source Flux Transport and Solar Wind Generator.” *SWQU Users meeting*, Huntsville, Alabama, 2024.
- **Stulajter, M. M.**, Rappoport, D., & Furche, F. “Exploring Chemical Space: Unraveling the Structure of Reaction Networks.” *ACSESS 2024*, San Diego, California, 2024.
- **Stulajter, M. M.**, & Rappoport, D. “Structure and properties of reaction network representations of potential energy surfaces with different rule sets” *ACS Fall 2023*, San Francisco, California, 2023.
- **Stulajter, M. M.**, Rappoport, D., & Furche, F. “Lower Dimensional Network Representation of Chemical Space” *ACSESS 2023*, San Diego, California, 2023.
- **Stulajter, M. M.**, Tang, Y. H., & de Jong, W. A. “Monte Carlo Tree Search for Inverse Molecular Design” *CS Summer Student Poster Session 2022*, Berkeley, California, 2022.
- **Stulajter, M. M.**, Caplan, R. M., & Carretero-González, R. “WENO3+RK3 with Non-Homogenous Velocity on a 2D Non-Uniform Grid” *ACSESS 2022*, San Diego, California, 2022.
- **Stulajter, M. M.**, Caplan, R. M., & Linker, J. A. “From Directives to DO CONCURRENT: A Case Study in Standard Parallelism” *Nvidia GTC 2022*, Online, 2022.
- Caplan, R. M., **Stulajter, M. M.**, & Linker, J. A. “Speeding up a banded-matrix solver by 3x using the updated cuSparse library” *Nvidia GTC 2022*, Online, 2022.
- **Stulajter, M. M.**, Caplan, R. M., & Linker, J. A. “Can Fortran’s ‘do concurrent’ replace directives for accelerated computing?” *Eighth Workshop on Accelerator Programming Using Directives*, St. Louis, Missouri, 2021.

SOFTWARE

Colibri2

<https://bitbucket.org/rappoport/colibri2/src/master/>

Python software for a distributed computational engine for chemical exploration.

ABSTRACT OF THE DISSERTATION

Low-Dimensional Discrete Reaction Networks for Chemical Space Representation and
Reaction Pathway Prediction

By

Miko Stulajter

Doctor of Philosophy in Computational Science

University of California, Irvine, 2026

Professor Filipp Furche, Chair

Reaction mechanisms and chemical reactivity are governed by high-dimensional potential energy surfaces (PESs). Exploring these surfaces is essential for discovering novel molecules and reaction pathways, yet exhaustive PES exploration becomes computationally prohibitive with each additional atom in a reaction system. Stoichiometry-preserving reaction networks (RNs) provide a tractable discrete representation of chemical space that enables systematic PES exploration. In this representation, nodes correspond to molecular collections of fixed stoichiometry and edges represent stoichiometry-preserving transformations. This dissertation investigates the structural properties of RNs and develops methods for more efficient pathway searching and targeted PES exploration within the open-source `colibri2` software package.

Comparative analysis of RNs against generative network models reveals that RNs closely resemble low-dimensional regular lattices with a small degree of random edge rewiring. Exploiting this underlying regularity, embedding methods are developed to construct low-distortion Euclidean embeddings of RNs. Combined with chemically informed edge-weighting, these embeddings preserve meaningful pathway rankings while enabling reaction pathway prediction without exhaustive graph traversal. RNs thus provide a low-dimensional discrete

representation of PESs.

To handle large chemical systems, this dissertation introduces bounded RN generation through energy threshold checks during node expansion and step limits relative to a starting node. These constraints restrict network growth to chemically relevant regions of the PES for specific reaction studies. Applications to atmospheric chemistry, enabled by the addition of support for radical reaction rules, demonstrate that these bounded RNs efficiently identify feasible pathways while suppressing irrelevant or infeasible ones.

This dissertation establishes stoichiometry-preserving RNs as a scalable framework for PES exploration, supporting both broad chemical space exploration and targeted searches focused on specific regions of the PES within `colibri2`. By exploiting the lattice-like structure of RNs, this dissertation shows that accurate pathway prediction and targeted PES exploration are achievable without exhaustive graph traversal, enabling scalable computational investigation of complex chemical systems.

Chapter 1

Introduction

Understanding and predicting how reactions occur is a fundamental part of chemistry. One of the ways to do this is by looking at chemical potential energy surfaces (PESs), which contain all possible molecular configurations for a given set of atoms and all possible reaction pathways connecting them. Reaction pathways are critical for explaining chemical reactions, discovering new molecules, and gaining insights into complex processes across chemistry subdisciplines, from organic chemistry and combustion to atmospheric chemistry and origin-of-life research. They also help in identifying transition states, which are necessary for understanding reaction kinetics and dynamics. However, PESs are complex, continuous, and high-dimensional landscapes that grow in complexity with each additional atom, making them computationally expensive to explore. For large or complex systems, the sheer size of the corresponding PES and the possible pathways it contains leads to a combinatorial explosion that makes direct, exhaustive exploration intractable.

While various tools have been developed to address the challenge of exploring PESs, there is a tradeoff faced by every method of either being too exhaustive or too narrow, and striking a balance between the two is difficult. Targeted approaches can be fast and focused but

are limited to specific chemical systems or reaction types, and the biases introduced by a restricted set of rules can cause them to overlook less obvious reaction pathways. Exhaustive approaches avoid this bias but generate large numbers of possible pathways, with a majority being either irrelevant to the reaction being studied or chemically infeasible, requiring significant effort to identify which are relevant and chemically feasible. The `colibri2` software package, an open-source code for generating stoichiometry-preserving reaction networks (RNs), takes a systematic enumeration approach, contending with the second issue of being too exhaustive. Instead of having to explore the full continuous PES directly, `colibri2` generates a lower-dimensional, discrete RN representation of the PES, enabling a complete search of chemical space in a discrete and tractable manner. In these networks, each node represents a collection of molecules with fixed atomic composition, with nodes differing only in molecular connectivity, and each edge corresponds to a stoichiometry-preserving transformation defined by a predefined ruleset of bond-breaking and bond-forming rules. When no limits are imposed, `colibri2` generates a finite, complete RN for a given stoichiometry and ruleset containing all chemically possible nodes and transformations between them, including both feasible and infeasible reaction pathways. When generation is instead started from a set of reactants and coupled with energy-informed guided expansion and a maximum step count, a smaller RN containing only the most probable reaction pathways and molecules can be produced. This dual capability of either generating a complete representative map or a targeted region of a PES within a single unified framework is a unique strength of this approach.

This dissertation addresses the challenge of finding feasible reaction pathways while avoiding the generation of too many irrelevant or infeasible ones, by leveraging the underlying structure of RNs and utilizing chemically informed edge weights. The structure of these RNs is found to resemble low-dimensional regular lattices, and this regularity is exploited by embedding RNs into Euclidean space to enable efficient reaction pathway searches. On-the-fly energy calculations during network generation further help by reducing the number of chem-

ically infeasible pathways and molecules generated during node expansion, thereby limiting RN size. The bounded RN approach is demonstrated on several benchmark reactions in atmospheric chemistry. Throughout this dissertation work, the `colibri2` software package was also optimized and improved to scale to larger molecular systems, with key changes reviewed in this chapter. Together, these contributions establish RNs as a structurally regular and geometrically tractable representation of PESs, providing a generalizable framework for unbiased reaction pathway prediction across chemistry subdisciplines. This chapter reviews the fundamental challenges of PES exploration and current approaches, introduces key concepts, describes the `colibri2` software and its optimizations, and outlines the research aims of this dissertation.

1.1 Background

Exploring PESs to identify reaction pathways and transition states is a longstanding and fundamental challenge in computational chemistry. For all but the simplest of systems, exhaustive exploration of the PES is computationally intractable, and targeted strategies that balance completeness against computational cost are required. Chemical reaction networks offer one such framework for representing and exploring chemical reactivity [2–4]. Applications span many areas of chemistry from chemical process modeling [3, 5], reaction mechanism analysis [6, 7], and synthetic path prediction [8] to metabolism [9–12], combustion [13, 14], protein folding [15, 16], atmospheric chemistry [17], and the prebiotic origins of biomolecules [18]. The difficulty is compounded by the fact that not all feasible reaction mechanisms are known. Several recent studies in atmospheric chemistry illustrate this point, including the identification of an additional hydroperoxyl radical formation pathway in formaldehyde via photophysical oxidation [19], the discovery of a previously unrecognized methane oxidation pathway involving water vapor absorption in ultraviolet light that re-

solved long-standing discrepancies in atmospheric methane lifetime predictions [20], and the continued emergence of new reaction pathways involving Criegee intermediates [21]. These discoveries underscore that systematic and unbiased PES exploration remains an open problem.

A wide range of computational methods have been developed to address PES exploration. Cheminformatics-based methods apply graph transformation rules to enumerate reaction networks. Some encode elementary bond-breaking and bond-forming transformations to systematically generate all possible transformations for a given stoichiometry [22–24], while others apply mechanistic reaction templates encoding known multi-step reaction patterns [13, 25–30]. Quantum chemistry-based exploration methods locate transition states and intermediates directly on the PES, including heuristics-guided approaches [7, 31, 32], the Global Reaction Route Mapping method [33, 34], growing string methods [6, 35], and artificial force-based approaches [36–38]. Graph-driven reaction path sampling methods [39–41] and accelerated molecular dynamics approaches, including metadynamics [42, 43], manual steering [44], and *ab initio* nanoreactors [45–48], can discover unexpected pathways by nudging the system across energy barriers. Integrated frameworks such as Chemoton [32] and YARP [49–53] combine reaction network exploration with quantum chemical calculations at scale. Machine learning approaches [54–59] have also been applied with growing success, leveraging data-driven models to predict reaction outcomes, explore chemical space, and prioritize pathways without exhaustive quantum chemical calculations.

Despite this diversity of approaches, two fundamental limitations persist. The first is narrowness. Many methods are tailored to particular chemical systems or reaction classes and exhibit biases that cause them to prioritize well-characterized regions of the PES, potentially overlooking novel or unexpected pathways [60]. The second is scalability, where exhaustive enumeration methods generate combinatorially large networks that are expensive to construct, store, and analyze. Machine learning methods partially address scalability but

require large, diverse training datasets and struggle to generalize beyond known reaction classes [54, 60]. Striking a balance between completeness and computational tractability remains an open challenge.

Stoichiometry-preserving RNs [22, 23] offer a discrete and relatively compact representation of the PES that partially addresses the tradeoff between completeness and computational tractability. These networks have been shown to form low-dimensional metric spaces with dimensionality roughly ten times smaller than the underlying PES [23]. Identifying feasible pathways within a RN is naturally formulated as a shortest-path problem on a weighted graph, and chemically informed edge weights make shortest-path lengths a meaningful measure of reaction feasibility [52, 61, 62]. One possible edge weight is the karc criterion [23, 24], an energy-based measure of pathway feasibility that penalizes both high-energy intermediates and unnecessarily long pathways (see Section 1.2.4). However, the metric space structure of RNs lacks a coordinate representation, which limits the applicability of geometric search algorithms. Euclidean embeddings of RNs provide a coordinate representation that can then be used by such methods. A further limitation of exhaustive RN generation is that networks grow exponentially with system size, producing large numbers of nodes irrelevant to the reaction of interest. On-the-fly energy calculations during network construction enable guided node expansion that can avoid generating less relevant nodes and edges, focusing exploration on the most chemically relevant regions. A full RN of millions of nodes can be reduced to a smaller RN on the order of hundreds to tens of thousands of nodes while retaining the most probable reaction mechanisms.

1.2 Key Concepts

1.2.1 Potential Energy Surfaces

A PES is a multidimensional landscape that represents how the energy of a molecular system changes with the positions of its atoms. For a nonlinear molecule (with atoms that are not all collinear) with N_a atoms, the PES scales as $\mathbb{R}^{3N_a-6}$, accounting for three degrees of freedom per atom minus three for translation and three for rotation. For a linear molecule (with all atoms lying along a single axis), it scales as $\mathbb{R}^{3N_a-5}$, as only two rotational degrees of freedom are removed. In both cases, the dimensionality and number of critical points grow rapidly with each additional atom, making the PES a vast and complex hypersurface that becomes computationally prohibitive to explore exhaustively for all but the smallest systems. Minima on the PES correspond to stable configurations such as reactants, intermediates, and products, while first-order saddle points called transition states represent the lowest energy barrier connecting two minima. Transition states define the activation energy required for a reaction to occur and are necessary for understanding reaction kinetics and dynamics. They are particularly challenging to locate because they correspond to narrow, high-energy saddle points on the PES and therefore require precise and computationally intensive search methods. In principle, a complete PES contains all possible reaction pathways and transition states for a given set of atoms, but in practice only a small fraction of this surface is accessible to current computational methods.

1.2.2 Reaction Networks

In the context of PES exploration, nodes in a RN correspond to chemical states and edges represent transitions between them [22, 23]. The specific type of RN considered throughout this dissertation is the transition-type RN, constructively defined by specifying an initial node

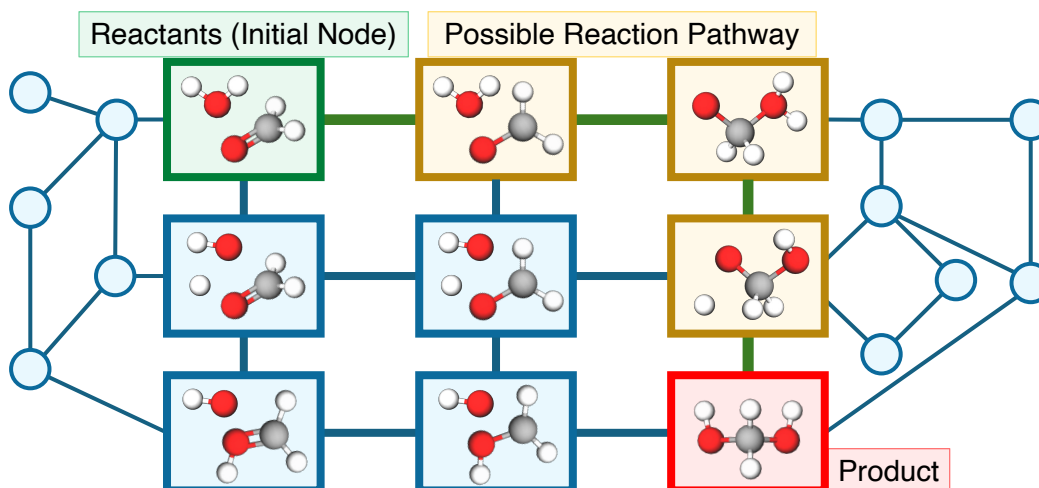


Figure 1.1: Reaction network for two oxygen, one carbon, and four hydrogen atoms, starting with formaldehyde (CH_2O) and water (H_2O) as the initial reactants. The highlighted path (orange nodes) traces the transformation from the reactants (green node) to the product, methanediol ($\text{CH}_2(\text{OH})_2$) (red node). Molecular diagrams along the path depict the specific bonds broken and formed at each step.

and a set of stoichiometry-preserving, reversible transformation rules. Repeated application of these rules terminates after a finite number of steps by combinatorial arguments, producing a finite, connected, undirected graph that defines a well-defined metric space [63]. Chemical distance can be measured as the shortest path length between nodes, corresponding to the minimal number of bond-breaking and bond-forming steps [64, 65]. Unless stated otherwise, the term “reaction network” in this dissertation refers exclusively to transition-type RNs.

RNs consist of molecules grouped into nodes, with each node maintaining constant total stoichiometry. Edges represent stoichiometry-preserving reactive transformations between nodes described using electron-pair transfers, commonly known as “arrow pushing” mechanisms [66, 67]. This formalism enables a broad range of reactions to be expressed in RNs. Generation begins with an initial node, with transformation rules applied iteratively to the initial node and its successors until no further unique nodes can be generated, exhaustively mapping the reachable chemical space. Since the process is exhaustive, only the total count of each atom type in the starting node matters if the transformation ruleset is complete

enough to enumerate all relevant transformations. As an example, a RN consisting of two oxygen, one carbon, and four hydrogen atoms with a reversible and complete transformation ruleset following normal polarity rules contains 20 nodes, as shown in Figure 1.1. All RNs in this dissertation were generated using the open-source `colibri2` software package [68].

1.2.3 Metric and Euclidean Spaces

A metric space is a mathematical structure in which the distance between any two points is defined by a distance function. This function satisfies four properties: non-negativity, identity of indiscernibles, symmetry, and the triangle inequality [69, 70]. Non-negativity means distances are always zero or positive. Identity of indiscernibles means that a distance of zero implies that two points are identical, and conversely, identical points have a distance of zero. Symmetry means the distance from point A to point B equals the distance from B to A . The triangle inequality states that the direct distance between two points is never greater than the sum of distances through an intermediate point. Metric spaces provide a general framework for reasoning about distance and proximity, but they lack a coordinate representation, meaning elements can only be described in terms of their pairwise distances and cannot be explicitly located in space.

A Euclidean space is a special case of a metric space that includes coordinates, allowing elements to be precisely located. This coordinate structure supports geometric operations such as interpolation, vector arithmetic, and nearest-neighbor search. The additional structure of a coordinate representation makes Euclidean spaces significantly more amenable to efficient search and sampling algorithms than general metric spaces.

RNs equipped with a shortest-path distance metric, where the distance between two nodes is the minimum number of bond-breaking and bond-forming steps connecting them, form discrete, low-dimensional metric spaces [23]. While this unweighted metric captures the

topological structure of the RN, it assigns equal weight to all edges and therefore does not account for the energetic barriers separating intermediates. The arc and karc criteria (Section 1.2.4) instead assign each edge a chemically informed weight, defining a weighted shortest-path metric that more closely reflects the underlying PES. Both the unweighted and weighted (karc) metrics are used in this dissertation. Chapter 2 characterizes RN topology using the unweighted metric, Chapter 3 evaluates Euclidean embeddings under both the unweighted and karc-weighted metrics, and Chapter 4 uses the karc-weighted metric to rank candidate pathways during bounded network generation. Regardless of the choice of metric, the absence of coordinates limits the applicability of geometric algorithms. Embedding RNs into low-dimensional Euclidean spaces restores a coordinate representation, at the cost of a small, controllable distortion of pairwise distances, and opens the door to efficient geometric methods for pathway search and sampling.

1.2.4 Reaction Pathway Feasibility: arc and karc Criteria

RNs systematically enumerate all possible reaction pathways between a reactant and a product, making it necessary to assess pathway feasibility. Not all enumerated pathways are chemically meaningful, as some may involve high-energy intermediates, require an excessive number of steps, or violate thermodynamic constraints. The arc and karc kinetic criteria, grounded in thermodynamics and kinetics, define per-edge weights that replace the unweighted edge weight introduced above, providing a principled basis for ranking and filtering pathways by their cumulative weighted cost [22, 24]. Both criteria are designed to satisfy several important properties. These include non-negativity, ensuring the feasibility measure is always non-negative, and additivity, so that the total cost of a pathway equals the sum of its individual steps. The criteria are also consistent with Hammond’s postulate [71], reflecting activation energy through the highest-energy point along the path. In addition, the criteria can include a length penalty that increases the cost with the number of steps to dis-

courage unnecessarily long pathways. They also satisfy detailed balance by relating forward and backward feasibilities through the energy difference between the initial and final states.

Kinetic Criterion: arc The arc criterion $W_{a,K \rightarrow L}$ evaluates pathway feasibility by penalizing longer pathways through an additive constant α at each step:

$$W_{a,K \rightarrow L} = \sum_{i=0}^{N-1} \left[(E_{K_{i+1}} - E_{K_i})^2 + \alpha^2 \right]^{1/2}, \quad (1.1)$$

where E_{K_i} is the energy of intermediate K_i and $\alpha \geq 0$ eV is an adjustable parameter. When $\alpha = 0$ eV, the criterion reduces to the climb criterion, which corresponds to the energy of the highest point along the reaction pathway relative to the initial state and does not penalize longer pathways. For $\alpha > 0$ eV, the penalty increases approximately linearly with the number of steps, penalizing longer pathways. A value of $\alpha = 1$ eV was found to be optimal in most cases [24].

Modified Kinetic arc Criterion: karc The arc criterion penalizes all steps equally, including energy-neutral ones, potentially overestimating pathway cost. The karc criterion addresses this by scaling the penalty according to the magnitude of each energy change:

$$W_{k,K \rightarrow L} = \sum_{i=0}^{N-1} \left[(E_{K_{i+1}} - E_{K_i})^2 + \alpha^2 \left(\frac{1 - \exp(-\kappa (E_{K_{i+1}} - E_{K_i})^2)}{1 + \exp(-\kappa (E_{K_{i+1}} - E_{K_i})^2)} \right) \right]^{1/2}, \quad (1.2)$$

where $\alpha \geq 0$ eV and $\kappa \geq 0$ eV⁻² are adjustable parameters. Energy-neutral steps are effectively excluded from the penalty, while steps with significant energy changes are penalized. Values of $\alpha = 1$ eV and $\kappa = 5$ eV⁻² were found to be optimal in most cases [24].

1.3 Workflows and Optimizations in colibri2

The `colibri2` software package [68] generates stoichiometry-preserving reaction networks using a distributed architecture built on Redis-based queues and caches [72] for inter-process communication and a PostgreSQL database [73] for persistent storage of network data such as nodes and edges. Two workflows are supported depending on whether on-the-fly energy calculations are desired.

1.3.1 Workflows

Network Generation Without Energy Calculations This workflow consists of three main stages, summarized in Table 1.1. First, the `initialize3` worker performs one-time setup of the PostgreSQL database and Redis queues and caches, and inserts the starting node into the *nodes* queue, *nodes* cache, and *nodes* database table. Second, multiple `network3` workers build the network in parallel using a TOML-formatted [74] ruleset of stoichiometry-preserving transformation rules written in SMARTS notation [75]. Each worker dequeues nodes from the Redis *nodes* queue, applies the transformation rules to generate new nodes and edges, and writes the results to the PostgreSQL database. It simultaneously adds each node to the *nodes* cache and enqueues any newly discovered nodes not yet in the *nodes* cache into the *nodes* queue. This process continues until the *nodes* queue is empty and no further unique nodes can be generated. Third, upon completion, one or more `export3` workers read the network from the database and write the nodes and edges to a GraphML file [76].

Table 1.1: Workers in `colibri2` for network generation without energy calculations and their roles.

Worker	Role
<code>initialize3</code>	One-time setup; inserts and enqueues the starting node
<code>network3</code>	Generates nodes and edges and stores them
<code>export3</code>	Writes network nodes and edges to a GraphML file

Network Generation With Energy Calculations This workflow follows the same three-stage structure as the energy-free workflow, but introduces additional workers that evaluate node energies concurrently with network generation, as summarized in Table 1.2.

`initialize5` performs the same one-time setup as before. Once complete, the `pipeline5` orchestrator launches `proton5` to precompute and cache the proton (H^+) reference energy as the free energy difference between water and hydroxide. Upon completion, `pipeline5` spawns four workers concurrently. `network5` generates nodes and edges, expanding a node only when (i) its step count from the starting node is below a set maximum and (ii) the energy difference between the child and parent node is below a set threshold. This is in contrast to the energy-free workflow, which applies no such criteria. `compute5` performs geometry optimization and free energy calculations for newly generated molecules. Application of the reaction rules can generate structures with unphysical electronic configurations, including partially broken bonds and artificial charge-separated states. Transformations that alter bond orders, formal charges, or radical localizations without changing molecular connectivity are termed *zero-energy transformations*. `equivalence5` groups molecules related by such transformations into equivalence classes, assigning each the minimum energy among members with maximal bond order and minimal absolute formal charge, thereby limiting calculations for unphysical configurations. `reaper5` monitors for stalled or failed tasks and restarts them.

Table 1.2: Workers in `colibri2` for network generation with energy calculations and their roles.

Worker	Role
<code>initialize5</code>	One-time setup; inserts and enqueues the starting node
<code>pipeline5</code>	Runs <code>proton5</code> , <code>network5</code> , <code>compute5</code> , <code>equivalence5</code> , and <code>reaper5</code>
<code>proton5</code>	Precomputes and caches the proton (H^+) reference energy
<code>network5</code>	Generates nodes and edges; expands based on criteria
<code>equivalence5</code>	Groups molecules by equivalence class; assigns energies
<code>compute5</code>	Geometry optimization and free energy calculation of molecules
<code>reaper5</code>	Monitors for stalled or failed tasks and restarts them
<code>export5</code>	Writes network nodes and edges to a GraphML file

Additional `network5`, `compute5`, and `equivalence5` workers can be launched to increase throughput. These workers run concurrently until the *nodes* queue is empty, at which point one or more `export5` workers write the final network to a GraphML file [76], with node energies included as node attributes.

1.3.2 Optimizations and Improvements

Over the course of this dissertation work, several optimizations and fixes were implemented to improve the scalability, reliability, and accuracy of `colibri2`.

Several scalability and performance improvements were made to both network generation and export. Network generation was found to be embarrassingly parallel as the architecture relies on queues to coordinate work. Each worker dequeues a task from its task queue and, once complete, writes its results to the appropriate queues, caches, and database tables before dequeuing the next task. Since both PostgreSQL and Redis support hundreds of concurrent connections, hundreds of `network` workers can be launched in parallel with minimal coordination overhead. This enables generation of networks on the order of billions of nodes and edges. This required data storage optimizations to deal with the larger amount of data. The PostgreSQL schema was restructured to use integer primary keys and relational node and edge tables, reducing redundant data storage. Metadata storage was also made configurable, allowing unneeded fields to be excluded and reducing database, queue, and cache sizes. In order to write a larger network to a file, export was parallelized by assigning multiple `export` workers, each stream-writing an assigned batch of the network to an individual GraphML file. The result is a set of smaller files that partition the full network, reducing I/O time and making the output easier to manage and transfer.

Several quirks of the RDKit library [77] were identified and worked around, including inconsistent molecule canonicalization, the inability to apply SMARTS patterns to ring molecules

without explicit ring-opening rules, and the addition of extra hydrogen atoms when applying SMARTS rules in violation of stoichiometry preservation. These were resolved through explicit ring-opening rules and more robust molecule validation. The use of explicit H0 annotations in SMARTS transformation rules, together with extended SMARTS notation specifying radical counts, prevented unintended hydrogen atoms from being added.

1.4 Research Goal

This dissertation is driven by three overarching research questions aimed at making the computational exploration of potential energy surfaces more tractable and improving reaction pathway prediction. First, do stoichiometry-preserving reaction networks possess an underlying structure that can be characterized and leveraged? Second, can reaction networks with chemically informed edge weights be faithfully embedded into low-dimensional Euclidean space, and can such embeddings be used for efficient reaction pathway prediction? Third, can on-the-fly energy calculations during network construction limit network size to the most chemically relevant pathways, and do the resulting bounded networks recover known and novel reaction mechanisms? In addressing these questions, this work develops scalable computational methods that exploit the underlying structure of reaction networks, providing a generalizable framework for unbiased and efficient exploration of potential energy surfaces across chemistry subdisciplines.

1.5 Research Overview

This dissertation focuses on developing and applying computational methods to understand the structure of RNs and leverage that structure for efficient reaction pathway prediction. Each contribution is motivated by gaps in current understanding, with the goal of establish-

ing a foundation for future work and providing scalable tools for exploring chemical potential energy surfaces. The three chapters each represent an independent research project, but the findings and methods of each chapter build directly on those of the previous one, together advancing toward a practical framework for reaction pathway prediction.

Chapter 2 addresses the fundamental question of whether stoichiometry-preserving RNs possess a known, regular structure that can be leveraged. While prior work established that these networks form low-dimensional metric spaces, it remained unclear whether the structure matched a well-characterized class of networks whose properties are understood analytically. Chapter 2 addresses this question through a systematic comparative analysis of RNs containing carbon-, hydrogen-, nitrogen-, and oxygen-containing (CHNO) molecules against a panel of size-matched generative network models. The results show that these networks resemble low-dimensional regular lattices with a small degree of random edge rewiring, independent of system size, stoichiometry, and rule set. This finding is significant beyond structural characterization. Regular lattices are well understood, and knowing that RNs resemble them allows network property trends, such as dimensional scaling, to be known without needing to generate additional networks. It also suggests that RNs are natural candidates for embedding into Euclidean space, where search and sampling algorithms that leverage coordinate representations can be effectively applied.

Chapter 3 builds on the structural findings of Chapter 2 by investigating whether RNs can be faithfully embedded into low-dimensional Euclidean space with low distortion. Chapter 2 established that RNs resemble regular lattices when constant edge weights are used, but it remained an open question whether this embeddability would hold when chemically informed edge weights are introduced. Chapter 3 addresses both questions by evaluating a range of linear, nonlinear, and neural-network-based embedding methods on stoichiometry-preserving RNs with both constant and karc edge weights. A distance-preserving encoder combining landmark multidimensional scaling initialization with neural network refinement

is developed, achieving embedding quality comparable to metric multidimensional scaling at significantly lower computational cost. The results show that RNs with both constant and chemically informed weights can be faithfully embedded, and pathway recovery experiments demonstrate that feasible reaction mechanisms can be identified directly from the embedded space.

Chapter 4 extends the framework developed in Chapters 2 and 3 to practical applications by introducing on-the-fly energy calculations during network construction to restrict network growth to chemically relevant pathways. Rather than generating a complete network and filtering it afterward, node energies are evaluated during expansion, and nodes exceeding either an energy threshold or a maximum step count from the starting reactants are excluded from further expansion. This approach focuses network generation on the most chemically relevant region of the PES. This produces bounded RNs that retain the most probable reaction mechanisms at significantly reduced computational cost. This gives `colibri2` two distinct network generation modes: an exhaustive mode that generates a broad, unbiased map of all chemically possible transformations, and a targeted mode that generates a region of chemical space around a set of starting reactants bounded by energy and step count criteria. The bounded RN method is demonstrated on benchmark reactions in atmospheric chemistry, demonstrating network size reduction and the ability to predict feasible reaction pathways.

Chapter 5 provides a summary and outlook for future research in this area.

1.6 Summary and Significance

In summary, this dissertation advances the understanding and exploitation of reaction network structure for efficient reaction pathway prediction using `colibri2`. This dissertation

characterizes the lattice-like structure of stoichiometry-preserving reaction networks, develops scalable Euclidean embedding methods, and implements energy-informed guided node expansion. Together, these advances demonstrate how reaction networks can be used to identify feasible reaction pathways in complex chemical spaces. The resulting `colibri2` framework supports two complementary approaches: exhaustive generation of a complete chemical space and targeted generation of bounded networks focused on specific regions. Both provide a tractable means of exploring chemical space without bias toward specific mechanisms, helping ensure that novel or unexpected pathways are not overlooked. The exhaustive approach maps all possible transformations for a given stoichiometry, whereas the targeted approach focuses on the most probable pathways. Together, these contributions establish reaction networks as a structurally regular and geometrically tractable representation of the PES. RNs provide a scalable way for unbiased discovery of reaction mechanisms across diverse areas of chemistry, ranging from atmospheric chemistry to origin-of-life research.

Chapter 2

Low-Dimensional Regular Lattice Structure

The motivation for this chapter comes from a fundamental challenge in computational chemistry. The high-dimensional nature of potential energy surfaces (PESs) that govern reactivity makes the exploration, manipulation, and design of complex chemical reactions exceptionally difficult. The `colibri2` software addresses this by modeling the PES as a discrete reaction network, removing its continuous nature and lowering the effective dimensionality. These reaction networks were known to form a low-dimensional metric space for carbon-, hydrogen-, and oxygen-containing (CHO) reaction networks, but it was unclear whether this structure extended to other stoichiometries or whether additional regularity existed that could be exploited. This chapter answers that question and finds that reaction networks resemble low-dimensional regular lattices with minimal distortion, independent of system size, stoichiometry, and rule set. This structural regularity directly motivates the embedding of reaction networks into Euclidean space in Chapter 3, since regular lattices can be naturally represented in Euclidean space. This opens the possibility of searching and sampling reaction networks more efficiently without the need for full graph traversal. The work pre-

sented in this chapter has been published as “*Reaction Networks Resemble Low-Dimensional Regular Lattices*” [1] in the *Journal of Chemical Theory and Computation*. The reaction networks analyzed here were generated using the `colibri2` software package, freely available at <https://bitbucket.org/rappoport/colibri2/src/master/>.

The computational exploration, manipulation, and design of complex chemical reactions face fundamental challenges related to the high-dimensional nature of PESs that govern reactivity. Accurately modeling complex reactions is crucial for understanding the chemical processes involved in, for example, organocatalysis, autocatalytic cycles, and one-pot molecular assembly. Our prior research demonstrated that discretizing PESs using heuristics based on bond breaking and bond formation produces a reaction network representation with a low-dimensional structure (metric space). We now find that these stoichiometry-preserving reaction networks possess additional, though approximate, structure and resemble low-dimensional regular lattices with a small amount of random edge rewiring. The heuristics-based discretization thus generates a nonlinear dimensionality reduction by a factor of ten with an *a posteriori* error measure (probability of random rewiring). The structure becomes evident through a comparative analysis of CHNO reaction networks of varying stoichiometries against a panel of size-matched generative network models, taking into account their local, metric, and global properties. The generative models include random networks (Erdős–Rényi and bipartite random networks), regular lattices (periodic and non-periodic), and network models with a tunable level of “randomness” (Watts–Strogatz graphs and regular lattices with random rewiring). The CHNO networks are simultaneously closely matched in all these properties by 3–4-dimensional regular lattices with 10% or less of edges randomly rewired. The effective dimensionality reduction is found to be independent of the system size, stoichiometry, and rule set, suggesting that search and sampling algorithms for PESs of complex chemical reactions can be effectively leveraged.

2.1 Introduction

Reaction networks (RNs) offer a discrete and relatively compact model of chemical reactivity in metabolism [9–12], combustion [13, 14], protein folding [15, 16], and other types of complex chemical dynamics [6, 7, 31]. Network models with different definitions of their nodes and edges are used depending on the application. In kinetic modeling, it is convenient to assign network nodes to the individual compounds and connect all reactants and products of reactions by network edges, resulting in highly heterogeneous networks [9, 78]. However, chemical dynamics and, in particular, PES exploration, is more naturally represented by a RN, in which network nodes correspond to the states of the chemical system and network edges are transitions between the system states [22, 23]. As we showed previously, these stoichiometry-preserving transition networks, when equipped with shortest-path (chemical) distance as the distance metric, form discrete and low-dimensional metric spaces [23]. In comparison to the underlying PESs, metric-space models have about ten times smaller dimensionality, as measured by their fractal dimension [79–81].

Even though lower dimensionality boosts search algorithm performance [82, 83], metric spaces tend to be more challenging data structures than Euclidean spaces like $\mathbb{R}^d$. The main drawback of metric spaces compared to Euclidean spaces is that they do not support a coordinate representation and cannot effectively locate their elements, having only a concept of a distance between pairs of elements. The low dimensionality of RNs compared to the underlying PESs thus comes with a caveat: the loss of the coordinate representation. However, in this work, we show that it is possible to lift the caveat if we allow for a certain (small) amount of error in the form of random edge rewiring. Specifically, we find that RNs can be closely matched by low-dimensional regular lattices with a small fraction of randomly reassigned (rewired) edges. Therefore, the coordinate representation can be restored by embedding RNs in Euclidean spaces $\mathbb{R}^d$, generating a nonlinear dimensionality reduction between the high-dimensional PES $\mathbb{R}^{3N_a-6}$ (where N_a is the number of atoms) and the low-

dimensional embedding space $\mathbb{R}^d$ with controllable error. The dimensionality reduction is about tenfold and holds for a range of RN sizes, stoichiometries, and rule sets. The RN embeddings can be leveraged by efficient search and sampling algorithms, taking advantage of their low-dimensional coordinate representation.

The additional (approximate) structure of RNs is a feature of the discretization of reaction paths into sequences of elementary transformations. In many instances, the individual bond breaks and bond formations can be composed in different orders while summing up to the same total reaction [23, 84]. The permutations of the elementary transformations create (hyper)cubes within a RN, which correspond to fragments of regular lattices. The dimensionality of the hypercubes is related to the number of permuting RN edges and, thus, to the number of bonds involved in the reaction, which is typically much smaller than the total number of molecular degrees of freedom. The permutation symmetry is not exact since some edges are prohibited by valence rules and are further restricted if thermodynamic and kinetic feasibility are taken into account, for example, using edge weights [23]. However, the deviation from permutation symmetry can be modeled by random edge rewiring, which moreover provides a quantitative measure of the deviation from the symmetry through the rewiring probability p_r . In this chapter, we explore the structures of RNs by comparing them to a panel of size-matched generated network models ranging from regular lattices to random networks. The generative models are useful surrogates for the real networks as they are cheap to construct, scale to arbitrary size, and can be easily tuned using adjustable parameters, for example, dimensionality and degree of “randomness”.

RNs of the transition network type can be defined constructively by the choice of the initial node and the transformation rule set. Given a set of stoichiometry-preserving and reversible rules, repeated application of the rules to the initial node and subsequent nodes must terminate after a finite number of steps by combinatorial arguments. RNs of transition network type are thus finite, connected, undirected graphs that, when combined with an appropriate

graph distance metric, define metric spaces [63]. Chemical distance is defined as the shortest-path distance on the networks of bond breaks and bond formations [64, 65]. The finite size and metric structure distinguish transition networks from other types of RNs, which expand indefinitely unless the construction is stopped. In the following, we will only consider RNs of the transition type and will drop the additional qualification.

We construct and analyze RNs of different stoichiometries containing carbon, hydrogen, nitrogen, and oxygen (CHNO) atoms using polar reactivity rules. Subsequently, we compare these RNs to a panel of generative network models ranging from regular to random. For regular structures, we examine lattices with non-periodic (LNP) and periodic (LP) boundary conditions, both based on a single repeating pattern. Our analysis of random networks includes Erdős–Rényi (ER) [85] and bipartite random (BR), which are formed by creating network edges randomly with a predefined probability. Additionally, Watts–Strogatz (WS) networks [86] and modified regular lattices with rewiring (LNP-R and LP-R for non-periodic and periodic boundary conditions, respectively) contain a tunable degree of randomness via a rewiring probability parameter.

The networks are evaluated based on their structural characteristics, which are categorized into local, metric, and global properties. Local properties take into account only the node adjacencies; examples of local properties are node degree and clustering coefficient. Metric properties involve measures such as average and maximum distances between nodes. Finally, global properties consider aspects like connectedness and network dimension [87–89]. Different network properties are relevant in different circumstances, indicating no simple universal approach for network comparison exists. For example, scale-free (SF) networks are defined by their degree distribution [90, 91], while small-world networks, a network class that includes WS networks, are characterized by their diameter (or average shortest path length) and clustering coefficient [87–89, 92, 93]. Specialized search algorithms have been developed for SF networks, utilizing their heterogeneous structure [94, 95]. In contrast, for

other types of networks, the primary predictor of the search time is the network dimensionality [82, 83, 96, 97], typically defined by the box-covering dimension [79–81, 98], as specialized search algorithms are not usually available.

Exploration of PESs is currently a very active area of research. In addition to rules-based methods by us [22–24] and others, [7, 31, 39–41, 56] accelerated molecular dynamics methods are important tools for exploring PES, for example, metadynamics [42, 43, 99], artificial force-based, [36–38] manual steering [44], and *ab initio* nanoreactor [45–48] approaches. Moreover, reaction template-based methods [49–52] and machine learning approaches [54, 57–59] have been applied with success. Quantum chemistry-based exploration generates stoichiometry-preserving networks, in which the nodes correspond to individual molecular conformations [34, 36–38]. Each node of the RNs considered here can be understood as a collection of all conformations for a given constitution (bonding pattern) and thus a coarse-grained representation of the quantum chemistry-based networks. These and other methods can benefit from a better understanding of the underlying data structure.

This chapter is organized as follows. The details of constructing the discrete RN representations and the generative network models, along with the analysis of the networks, are presented in Section 2.2. Section 2.3 contains the analysis of the local, metric, and global network properties of RNs and the comparison between RNs and size-matched generative network models. We discuss the results in Section 2.4 and give our conclusions in Section 2.5.

2.2 Methods

2.2.1 Reaction Networks

The procedure for generating RNs is described elsewhere [22, 23]; thus, we only give a brief overview. The RNs are collections of molecules subject to the constraint that the total stoichiometry (total number of atoms) within a node across all molecules in that node must be fixed. The systems explored here were initialized from a combination of formaldehyde (CH_2O), hydrogen cyanide (HCN), and water (H_2O) molecules (CHNO networks). These molecules were chosen for their simplicity and the diversity of atoms they contain, within the constraint of only using CHNO atoms. The initial choice of starting molecules is not critical as only the number of each atom type is important, because the network generation process is exhaustive. CHNO networks with up to 4 carbon atoms were generated. For each combination of HCN and CH_2O molecules, up to 15 H_2O molecules were added, creating a homologous series of RNs. The pure CHO networks (which are initialized with CH_2O and H_2O molecules) have an average carbon oxidation state of 0, while in the CHNO networks, the average carbon oxidation number is between 0 and +2. We use Hill notation [100] to denote the RN stoichiometries. The edges of the RNs are stoichiometry-preserving reactive transformations. The transformation rules are chosen to be reversible and complete with respect to normal polarity rules and are presented in Table 2.1. The decomposition into polar transformation covers more than just polar reactions; in fact, all reactions that can be formally written as a combination of electron pair transfers (“arrow pushing”) [66, 67] are representable in this way [24].

All RNs were built using the open-source `colibri2` software package [68]. The Python code was parallelized to decrease the network construction times and to enable the generation and analysis of larger networks. With the parallelization and optimization performed, RNs composed of up to 57 atoms could be built. The construction of the largest RN ($\text{C}_3\text{N}_1\text{O}_{10}\text{H}_{21}$) re-

Table 2.1: Polar CHNO transformation rules used in this chapter.

Bond Dissociation	Bond Association	Bond Breaking	Bond Formation
$\text{H} - \text{C} \rightarrow \text{H}^+ \dots \text{C}^-$	$\text{H}^+ \dots \text{C}^- \rightarrow \text{H} - \text{C}$	$\text{C} = \text{C} \rightarrow \text{C}^+ - \text{C}^-$	$\text{C}^+ - \text{C}^- \rightarrow \text{C} = \text{C}$
$\text{C} - \text{C} \rightarrow \text{C}^+ \dots \text{C}^-$	$\text{C}^+ \dots \text{C}^- \rightarrow \text{C} - \text{C}$	$\text{C} \equiv \text{C} \rightarrow \text{C}^+ = \text{C}^-$	$\text{C}^+ = \text{C}^- \rightarrow \text{C} \equiv \text{C}$
$\text{H} - \text{O} \rightarrow \text{H}^+ \dots \text{O}^-$	$\text{H}^+ \dots \text{O}^- \rightarrow \text{H} - \text{O}$	$\text{C} = \text{O} \rightarrow \text{C}^+ - \text{O}^-$	$\text{C}^+ - \text{O}^- \rightarrow \text{C} = \text{O}$
$\text{H} - \text{O}^+ \rightarrow \text{H}^+ \dots \text{O}$	$\text{H}^+ \dots \text{O} \rightarrow \text{H} - \text{O}^+$	$\text{C} \equiv \text{O} \rightarrow \text{C}^+ = \text{O}^-$	$\text{C}^+ = \text{O}^- \rightarrow \text{C} \equiv \text{O}$
$\text{C} - \text{O} \rightarrow \text{C}^+ \dots \text{O}^-$	$\text{C}^+ \dots \text{O}^- \rightarrow \text{C} - \text{O}$	$\text{C} = \text{O}^+ \rightarrow \text{C}^+ - \text{O}$	$\text{C}^+ - \text{O} \rightarrow \text{C} = \text{O}^+$
$\text{C} - \text{O}^+ \rightarrow \text{C}^+ \dots \text{O}$	$\text{C}^+ \dots \text{O} \rightarrow \text{C} - \text{O}^+$	$\text{C} \equiv \text{O}^+ \rightarrow \text{C}^+ = \text{O}$	$\text{C}^+ = \text{O} \rightarrow \text{C} \equiv \text{O}^+$
$\text{H} - \text{N} \rightarrow \text{H}^+ \dots \text{N}^-$	$\text{H}^+ \dots \text{N}^- \rightarrow \text{H} - \text{N}$	$\text{C} = \text{N} \rightarrow \text{C}^+ - \text{N}^-$	$\text{C}^+ - \text{N}^- \rightarrow \text{C} = \text{N}$
$\text{H} - \text{N}^+ \rightarrow \text{H}^+ \dots \text{N}$	$\text{H}^+ \dots \text{N} \rightarrow \text{H} - \text{N}^+$	$\text{C} \equiv \text{N} \rightarrow \text{C}^+ = \text{N}^-$	$\text{C}^+ = \text{N}^- \rightarrow \text{C} \equiv \text{N}$
$\text{C} - \text{N} \rightarrow \text{C}^+ \dots \text{N}^-$	$\text{C}^+ \dots \text{N}^- \rightarrow \text{C} - \text{N}$	$\text{C} = \text{N}^+ \rightarrow \text{C}^+ - \text{N}$	$\text{C}^+ - \text{N} \rightarrow \text{C} = \text{N}^+$
$\text{C} - \text{N}^+ \rightarrow \text{C}^+ \dots \text{N}$	$\text{C}^+ \dots \text{N} \rightarrow \text{C} - \text{N}^+$	$\text{C} \equiv \text{N}^+ \rightarrow \text{C}^+ = \text{N}$	$\text{C}^+ = \text{N} \rightarrow \text{C} \equiv \text{N}^+$

quired three days on 48 CPU cores. The reaction rules were encoded as SMARTS strings [75] and applied to the molecular structures using the RDKit library [77]. The RN construction terminated when no new nodes could be generated. The networks were stored as compressed GraphML files [76]. The NetworkKit library [101] was used for all RN manipulations and analyses.

2.2.2 Network Properties

We analyzed the constructed CHNO networks with respect to their basic properties (number of nodes n , number of edges e , and density ρ), local properties (average node degree $\bar{k}$ and average square clustering coefficient $\overline{C_4}$), metric properties (diameter D and average shortest path length $\bar{l}$), and global properties (fractal dimension d_f and average growth exponent $\bar{\xi}$). The analysis treated the RNs as undirected simple graphs. To establish notation, we briefly describe the network properties and their computation below.

The network density $\rho = 2e/(n(n-1))$ is defined as the proportion of edges e in the network relative to the complete graph with n nodes. The degree k_u of the node u describes the number of edges incident to u , with the average node degree $\bar{k}$ defined as the mean over all

network nodes. The clustering in the neighborhood of the node u is typically described by its (triangle) clustering coefficient C_u [87]. However, as RNs are bipartite by construction because the number of bonds in a node is either even or odd when the multiplicity of bonds is included, [23] the conventional clustering coefficient is always zero and thus not a useful measure. Instead, we use the square clustering coefficient $C_{4,u}$, which is defined for the node u as the proportion of the pairs of its neighbors v, w that share a neighboring node different from u [102],

$$C_{4,u} = \frac{\sum_{\substack{v < w \\ v, w=1, \dots, k_u}} q_{uvw}}{\sum_{\substack{v < w \\ v, w=1, \dots, k_u}} [(k_v - \eta_{uvw})(k_w - \eta_{uvw}) + q_{uvw}]} . \quad (2.1)$$

Here, q_{uvw} is the number of squares which involve u, v , and w , and $\eta_{uvw} = 1 + q_{uvw} + \theta_{vw}$ with $\theta_{vw} = 1$ if v and w are adjacent and 0 otherwise [103]. The mean of $C_{4,u}$ over all network nodes yields the average square clustering coefficient $\overline{C_4}$.

For an undirected graph with the shortest path distance between nodes $u, v \in 1, \dots, n$ denoted as l_{uv} , the diameter D and the average shortest path length $\bar{l}$ are defined by the maximum and the mean over all pairs of nodes, respectively, $D = \max_{u < v} l_{uv}$ and $\bar{l} = 2/(n(n-1)) \sum_{u < v} l_{uv}$. Small-world networks are characterized by a logarithmic dependence of the metric properties on the network size, $\bar{l} \sim \ln n$ [87–89, 92, 93].

The calculation of the fractal dimension, d_f , was performed using the compact box burning (CBB) algorithm [80, 81, 104, 105], which approximates the box-covering dimension [79, 81, 98, 106]. In preliminary studies, the CBB algorithm proved to be more efficient than the alternative maximum excluded mass burning and greedy clustering algorithms [79]. As the CBB algorithm is stochastic, we determined the standard deviation of d_f values obtained in independent CBB calculations to be less than 1% in all cases. Therefore, only one calculated CBB fractal dimension is reported per RN. The CBB algorithm was parallelized over box sizes, which allowed the calculation of the CBB fractal dimension of the $\text{C}_2\text{H}_{20}\text{N}_2\text{O}_9$ network

with $n = 1,226,147$ nodes in seven days on 40 CPU cores.

The average growth factor $\bar{\xi}$ describes the increase of the number of nodes n_u^i in the cluster or radius i centered at the node u , that is, $n_u^i = |\{v | l_{uv} \leq i\}|$. The logistic fit of n_u^i versus i yields the exponent γ_u . The average growth factor is then defined as $\bar{\xi} = \exp(\bar{\gamma})$, where the mean is taken over all network nodes. The calculation of metric and global properties requires the matrix of shortest path lengths between all pairs of nodes, which was determined using parallel breadth-first search [101].

2.2.3 Generative Models

Size-matched networks were constructed using a variety of generative models, ranging from random to regular network topologies. Erdős–Rényi (ER) networks [85] and bipartite random (BR) networks [107] are obtained by a random assignment of edges with a fixed probability p_c for a given number of nodes n . BR networks differ from ER networks in that they contain random edges between two disjoint sets of nodes with the fraction of nodes in the first node set being b_s . Both ER and BR networks may contain disconnected components. Regular lattices with non-periodic boundary conditions (LNP) are constructed as d -dimensional Cartesian products of the one-dimensional path graph with itself. Connecting the nodes of the outer faces of the LNP networks with their opposite faces generates regular lattices with periodic boundary conditions (LP), which are free of edge effects.

In addition to fully random (ER and BR) and fully regular (LNP and LP) networks, we considered several classes of networks with tunable “randomness”, which is achieved by random rewiring of a regular network. Watts–Strogatz (WS) networks [86] start from a ring graph with n nodes, in which each node is additionally joined by edges with $K/2$ preceding and $K/2$ following nodes in the ring order. Shortcuts are then created by replacing some edges uv with probability p_r by uw , where the node w is chosen with a uniform probability

from the remaining $(n-2)$ nodes. Modified regular lattices (LNP-R and LP-R) generalize the WS construction to higher dimensions by applying random rewiring with probability p_r to the LNP and LP networks, respectively. Random rewiring in the LNP-R and LP-R networks break the bipartite nature of 2D and 3D lattices. As a result of the random rewiring, WS, LNP-R, and LP-R networks may be disconnected.

Since the networks with randomness (ER, BR, WS, LNP-R, and LP-R networks) have variability in network structures, an average of five independent trials is reported for each generative model. In the presence of disconnected components, the weighted average of the network properties by the number of nodes in each component is given. All generative models were constructed using the NetworkX library [108] software package and custom code (available on GitHub [109]). All networks were saved as compressed GraphML files [76], and the NetworKit library [101] was used for all analyses.

2.3 Results

2.3.1 Reaction Network Properties

In this chapter, we consider CHNO networks, which are constructed from combinations of CH_2O , HCN , and H_2O molecules using polar reactivity rules (Table 2.1). Each RN is uniquely identified by its stoichiometry, and the RNs differing only by the number of H_2O molecules form a homologous series. The changes in key RN properties as functions of the number of atoms, N_a , are shown in Figure 2.1, with homologous series indicated by lines. Plots of all properties can be found in Appendix A, Section A.1. The smallest RNs (CH_2O and CHN) have path structures with $n = 4$ nodes and $e = 3$ edges; the largest RN has the $\text{C}_3\text{H}_{21}\text{NO}_{10}$ stoichiometry and consists of $n = 15,244,289$ nodes and $e = 101,328,842$ edges. In total, we investigate 114 RNs belonging to 11 homologous series.

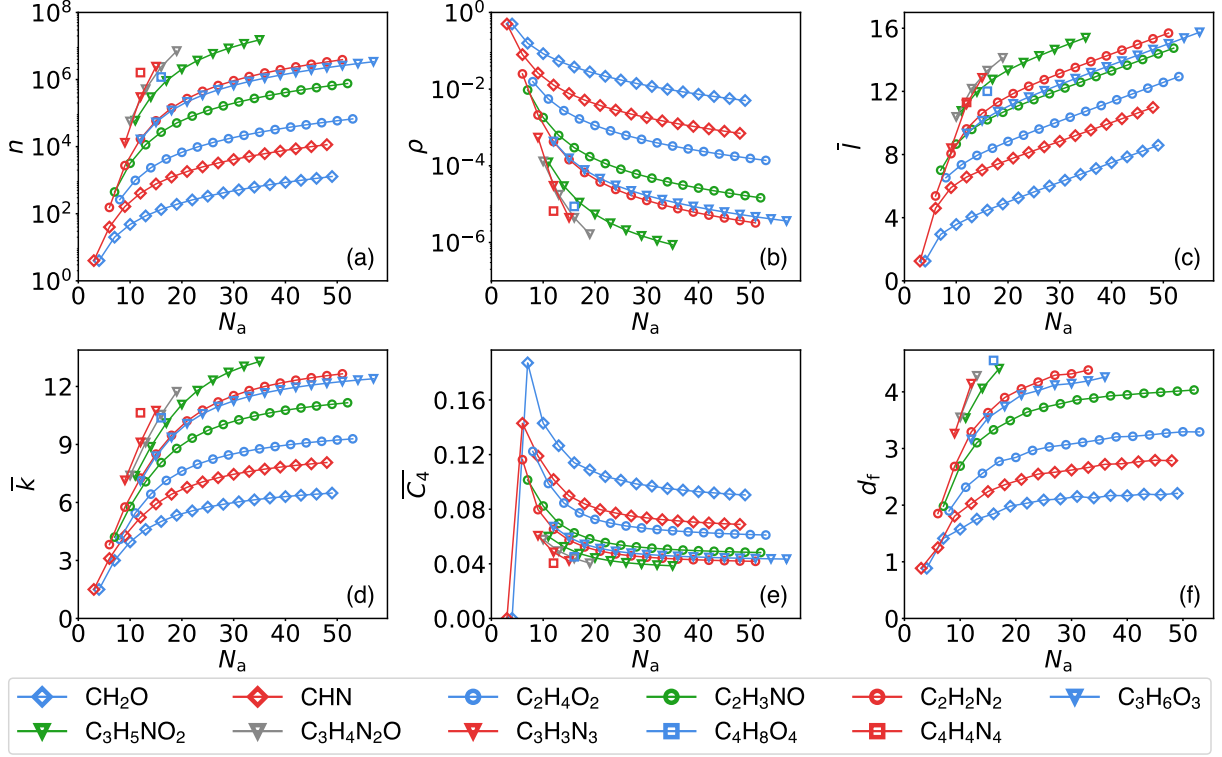


Figure 2.1: Variation of RN properties with the number of atoms, N_a : (a) number of nodes, n , (b) density, ρ , (c) average shortest path length, $\bar{l}$, (d) average node degree, $\bar{k}$, (e) average square clustering coefficient, $\bar{C}_4$, (f) fractal dimension, d_f . Homologous series are indicated by lines and denoted by their base RNs.

The network size, n , increases steeply with the number of carbon atoms; for the same number of carbon atoms, the RNs with higher HCN content are larger than those with higher CH_2O content for the same number of H_2O molecules, see Figure 2.1(a). Within each homologous series, n initially increases rapidly as H_2O molecules are added to the RN stoichiometry but eventually reaches a crossover point, after which the increase in n is much slower and levels off as seen through the asymptotic behavior of the growth. Below the crossover point, the network nodes contain unsaturated molecules; the addition of H_2O molecules introduces additional reactions, which contribute to the network growth. However, above the crossover point, any further increase in network size is due to bond breaking and formation in water, with no new reactivity added. The RNs are relatively sparse ($\rho < 10^{-2}$ in all except the smallest RNs) and become even sparser as they increase in size, see Figure 2.1(b). The

asymptotic behavior observed in n is also seen in ρ as H_2O molecules are added. The reason for increasing sparsity with larger RNs is that the ratio of new nodes to new edges created gets larger with increasing numbers of atoms.

Insights into the distances within a RN are given by the metric properties, D and $\bar{l}$. Figure 2.1(c) shows that even for large networks, $\bar{l}$ remains relatively small. The growth of $\bar{l}$ shows a polynomial nature with N_{a} as H_2O molecules are added, meaning, on average, it takes only a few steps to get from one node to another. For example, even in the largest network we considered, $\text{C}_3\text{H}_{21}\text{NO}_{10}$, the average shortest path length is only $\bar{l} = 15.41$. The smaller (water-poor) members of each homologous series follow quite accurately the logarithmic relationship between the number of nodes, n , and $\bar{l}$, which is characteristic of small-world networks. However, after the crossover to water-rich RNs, $\bar{l}$ increases faster than logarithmically with n . For the same number of H_2O molecules and carbon atoms, the average distances increase with higher HCN content. The network diameter, D , shows qualitatively the same behavior as $\bar{l}$.

The local properties of average node degree, $\bar{k}$, and average square clustering coefficient, $\overline{C}_4$, provide insight into the connectivity among nearby nodes. As Figure 2.1(d) shows, $\bar{k}$ increases rapidly with the number of atoms in water-poor RNs before leveling off in water-rich members of each homologous series as seen through the asymptotic growth. The small $\bar{k}$ means that overall, most nodes have relatively few connections. RNs rich in HCN have higher average node degrees, reflecting a larger diversity of reactivity patterns compared to RNs with higher CH_2O content. The values of $\overline{C}_4$ follow the opposite trend, as Figure 2.1(e) shows. With the exception of the two RNs with linear structures (CHN and CH_2O), which have no square clusters, $\overline{C}_4$ values decrease with increasing network size. Within a homologous series, $\overline{C}_4$ initially decreases rapidly when H_2O molecules are added but appears to approach a constant value for water-rich RNs as seen through the asymptotic growth. Less squares of nodes means fewer nodes are interconnected locally with each other. Similarly, the increase

in the average node degree, $\bar{k}$, with higher HCN content corresponds to a decrease in the average square clustering coefficient, $\overline{C_4}$.

We considered two global RN properties, the average growth factor, $\bar{\xi}$, and the fractal dimension, d_f , which offer an understanding of the broader structures within the networks. The average growth factor, $\bar{\xi}$, which indicates how many nodes can be reached within a given number of steps, shows only small differences between different RN stoichiometries and sizes. All RNs considered in this chapter have an average growth factor close to 2. Moreover, $\bar{\xi}$ decreases slightly with the number of atoms, consistent with lower network connectedness in larger RNs observed in smaller average square clustering coefficients, $\overline{C_4}$. A notable exception is the homologous series derived from CH_2O . We note that the growth factor does not seem to be correlated with the HCN content, unlike all other network properties. The origin of this convergence is not clear.

We have investigated in detail the fractal dimensions, d_f , of CHO RNs approximated by the CBB and two other approximate methods, maximum excluded mass burning (MEMB) and greedy coloring, in previous work [23]. These results and our preliminary studies of CHNO networks showed that the CBB method was sufficiently reliable and computationally efficient for our purposes. Moreover, we estimated the statistical errors of the CBB method due to its stochastic algorithm using the standard deviations over five independent calculations in $\text{C}_2\text{H}_8\text{N}_2\text{O}$, $\text{C}_3\text{H}_8\text{O}_4$, $\text{C}_3\text{H}_5\text{NO}_2$, and $\text{C}_3\text{H}_4\text{N}_2\text{O}$ RNs, which yielded 3.63 ± 0.01 , 3.54 ± 0.01 , 3.540 ± 0.004 , and 3.55 ± 0.02 , respectively. Since the statistical errors were on the order of 1%, only one CBB estimation of the d_f value was performed for all other RNs. For all networks, for which the calculation of d_f was feasible, d_f was found to be less than 5, as Figure 2.1(f) shows. The value of d_f stays below 5 because of the logarithmic growth seen. In the base RNs of each homologous series (containing zero H_2O molecules), the fractal dimension is $d_f \approx 1$ for one-carbon RNs, $d_f \approx 2$ for two-carbon RNs, $d_f \approx 3$ for three-carbon RNs, and $d_f = 4.55$ for $\text{C}_4\text{H}_8\text{O}_4$, the only four-carbon RN, for which the calculation of d_f was

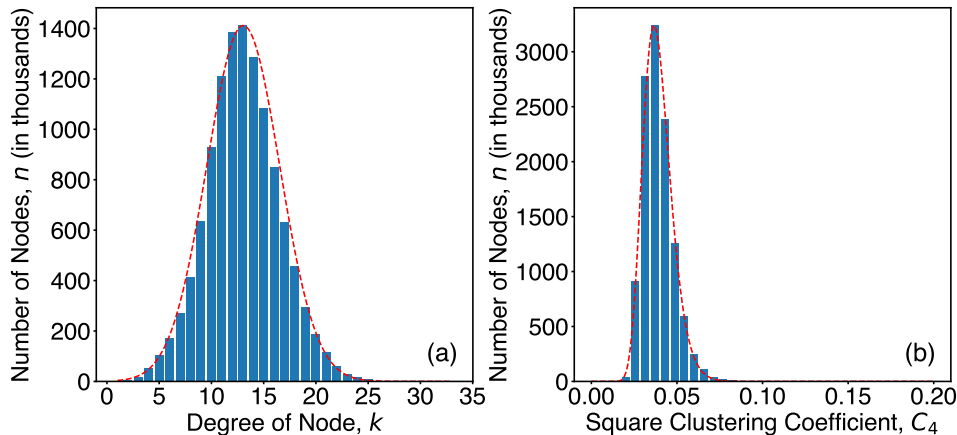


Figure 2.2: Histograms of (a) node degrees, k (bin width $w_k = 1$), and (b) square clustering coefficients, C_4 (bin width $w_C = 5.7 \times 10^{-3}$) in the $C_3H_{19}NO_9$ network. A normal distribution is fitted for k , and a lognormal distribution is fitted for C_4 . The fitted distributions are shown by dashed lines.

feasible. However, the values of d_f of RNs with the same number of carbon atoms diverge as H_2O molecules are added, and the networks with higher HCN content show a faster increase in the value of d_f compared to the RNs with higher CH_2O content. The initial rapid increase levels off with the crossover from water-poor to water-rich RNs.

In addition to the averages of node degrees, $\bar{k}$, and square clustering coefficients, $\overline{C_4}$, we considered their distributions across a network in order to evaluate the heterogeneity of the RNs with respect to local properties. The shape of the degree distribution depends on the network size and composition, with the smaller water-poor networks having a tail-heavy (positively skewed) distribution and the larger, in particular, water-rich networks showing a normal distribution. Figure 2.2(a) shows the distribution of node degrees, k , in the water-rich $C_3H_{19}NO_9$ network with a $\bar{k} = 13.04$ and standard deviation $\Delta k = 3.49$. The water-poor networks show some amount of heterogeneity, but the addition of H_2O molecules makes them more homogeneous, decreasing the relative spread of the node degrees compared to the mean. The Jarque–Bera tests [110] of node degree histograms confirm that the deviation from normality decreases with increasing H_2O content and that water-rich RNs are accepted as normal at $p = 0.05$ significance level. In previous work, we noted

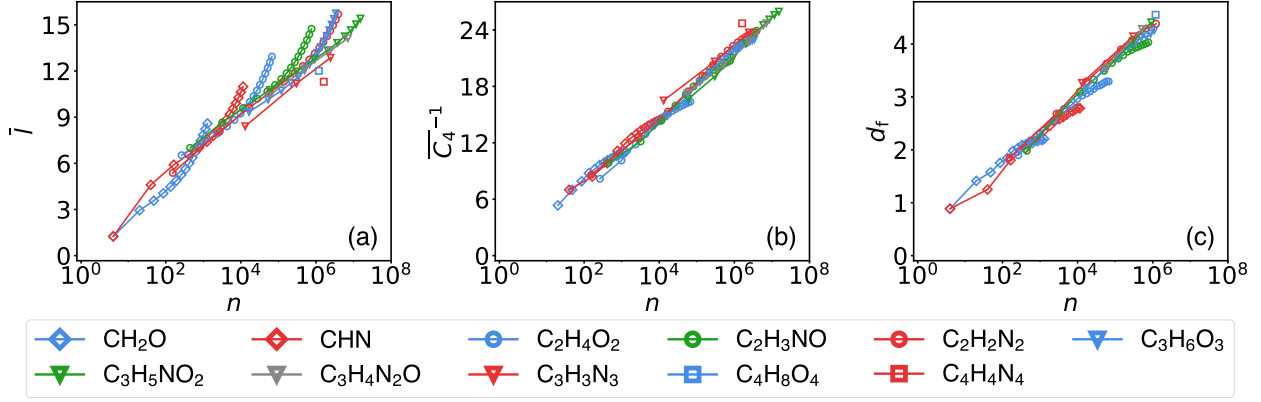


Figure 2.3: Variation of network properties with the number of nodes, n : (a) average shortest path length, $\bar{l}$, (b) inverse average square clustering coefficient, $\overline{C_4}^{-1}$ (excluding CHN and CH_2O networks), (c) fractal dimension, d_f . Homologous series are indicated by lines.

that the node degree distributions in CHO RNs could be described by heavy-tail lognormal distributions [23]. However, this finding seems to hold only for the smaller, water-poor members of each homologous series. The node degree distributions for the entire $\text{C}_3\text{H}_5\text{NO}_2$ homologous series are given in Appendix A, Section A.2.

The distribution of square clustering coefficients, C_4 , is right-skewed, as shown in Figure 2.2(b) for the $\text{C}_3\text{H}_{19}\text{NO}_9$ network. It fits a lognormal distribution, with $\ln(\overline{C_4}) = -3.26$ and $\Delta \ln(C_4) = 0.20$. Nonetheless, the distribution of C_4 values is relatively narrow. The C_4 distributions for the entire $\text{C}_3\text{H}_5\text{NO}_2$ homologous series are compiled in Appendix A, Section A.3.

Structural commonalities of all RNs, irrespective of their composition, can be made visible by considering network properties as a function of the network size, n , across all homologous series. Figure 2.3 shows the dependence of the three selected network properties: average shortest path length, $\bar{l}$ (metric property), average square clustering coefficient, $\overline{C_4}$ (local property), and fractal dimension d_f (global property) as a function of n . The value of $\bar{l}$ varies approximately logarithmically with the network size, $\bar{l} \sim \ln n$, a behavior characteristic of small-world networks, [87–89, 92, 93] see Figure 2.3(a). The linear fit of $\bar{l}$ against $\ln n$ yields

the correlation coefficient $R^2 = 0.973$. However, water-rich RNs have higher values of $\bar{l}$ compared to the logarithmic relationship, likely due to their more regular structure, which deviates from the small-world behavior. The $\overline{C_4}$ values are close to inversely proportional to $\ln n$, as shown in Figure 2.3(b). The correlation coefficient of $1/\overline{C_4}$ against $\ln n$ is $R^2 = 0.995$. The logarithmic scaling of the value of d_f with the network size was already observed in CHO networks [23]. As Figure 2.3(c) shows, the universal relationship also holds for CHN and mixed CHNO networks, despite the addition of rule sets of carbon–nitrogen reactivity (Table 2.1). The linear fit of d_f against $\ln n$ has $R^2 = 0.992$. The water-rich RNs deviate from the scaling relationships for the average square clustering coefficients, $\overline{C_4}$, and d_f yielding $\overline{C_4}$ slightly above and d_f below the general trend.

2.3.2 Comparisons with Generative Models

We selected ten CHNO networks of varying sizes and compositions for comparisons with size-matched generative models. The comparison RNs are grouped into three size classes: small ($\text{C}_2\text{H}_5\text{NO}_2$, $\text{CH}_{19}\text{NO}_9$, and $\text{C}_2\text{H}_{10}\text{O}_5$) with $n \approx 4000$ nodes; medium ($\text{C}_3\text{H}_6\text{O}_3$, $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$, and $\text{C}_2\text{H}_{18}\text{O}_9$) with $n \approx 17,000$; and large ($\text{C}_3\text{H}_8\text{O}_4$, $\text{C}_3\text{H}_4\text{N}_2\text{O}$, $\text{C}_2\text{H}_8\text{N}_2\text{O}_3$, and $\text{C}_3\text{H}_5\text{NO}_2$) with $n \approx 56,000$. These RNs are compared to generative models consisting of Erdős–Rényi networks (ER), bipartite random (BR) networks, Watts–Strogatz small-world networks (WS), regular lattices with non-periodic (LNP) and periodic (LP) boundary conditions as well as regular lattice networks with rewiring (LNP-R and LP-R).

As discussed in Section 2.3.1, the properties of RNs are strongly correlated with the number of network nodes, n . For brevity, we thus show only one RN for each size class in the following: $\text{CH}_{19}\text{NO}_9$ for small comparison RNs, $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ for medium comparison RNs, and $\text{C}_3\text{H}_4\text{N}_2\text{O}$ for large comparison RNs. Their properties are summarized in Table 2.2. The properties of all comparison RNs are listed in Appendix A, Section A.4.

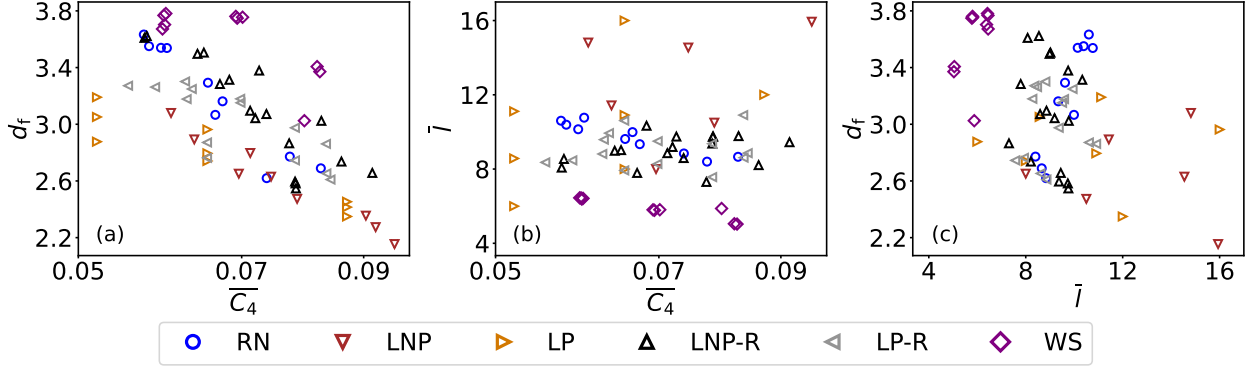


Figure 2.4: Comparisons between network properties of RNs (circles) and generative models: Watts–Strogatz networks (WS, diamonds), regular lattices with $d = 3\text{--}5$, non-periodic boundary conditions (LNP, down triangles) and periodic boundary conditions (LP right-facing triangles), LNP networks with rewiring (LNP-R, up triangles), and LP networks with rewiring (LP-R, left-facing triangles): (a) fractal dimensions, d_f , against average square clustering coefficient, $\overline{C}_4$, (b) average shortest path length, $\bar{l}$, against $\overline{C}_4$, and (c) d_f against $\bar{l}$. See text for more details.

A comprehensive comparison of RNs and generative network models should consider their local, metric, and global properties. In Figure 2.4 we present pairwise plots of average square clustering coefficients, $\overline{C}_4$, average shortest path lengths, $\bar{l}$, and fractal dimensions, d_f , in RNs and size-matched generative network models. The plots exclude the random networks (ER and BR) to improve the readability as these networks differ substantially from RNs with respect to all three properties. The full pairwise plots can be found in Appendix A, Section A.5. In the following, we briefly discuss the properties of the size-matched generative models in comparison to the RNs. The complete data set of generative model networks is available in Appendix A, Section A.6.

Table 2.2: Selected network properties of $\text{CH}_{19}\text{NO}_9$ (small), $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ (medium), and $\text{C}_3\text{H}_4\text{N}_2\text{O}$ (large) comparison RNs: number of nodes, n ; number of edges, e ; average node degree, $\bar{k}$; average square clustering coefficient, $\overline{C}_4$; average shortest path length, $\bar{l}$; fractal dimension, d_f .

RN	n	e	$\bar{k}$	$\overline{C}_4$	$\bar{l}$	d_f
$\text{CH}_{19}\text{NO}_9$	4131	15,404	7.46	7.37×10^{-2}	8.84	2.62
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	17,014	61,762	7.26	6.53×10^{-2}	9.62	3.29
$\text{C}_3\text{H}_4\text{N}_2\text{O}$	56,485	208,955	7.40	5.78×10^{-2}	10.39	3.55

Erdős–Rényi Networks The ER networks are constructed by choosing the number of nodes, n , and creating edges randomly between pairs of nodes with a fixed probability p_c . The properties of the size-matched ER networks are shown in Table 2.3. These models behave very differently from their RN counterparts. We note that the average shortest path length, $\bar{l}$, of ER networks is considerably smaller, being nearly half that of the corresponding RN. For example, the $C_3H_4N_2O$ network has $\bar{l} = 10.39$, while the size-matched ER network has only $\bar{l} = 5.69$ on average. The average square clustering coefficients, $\overline{C_4}$, of ER networks are two to three orders of magnitude smaller than those in RNs, and the fractal dimensions, d_f , are higher by 0.6 to 1.1 units than those of the corresponding RNs. Properties of the other six size-matched ER networks and details of the construction are given in Appendix A, Section A.6.1.

Table 2.3: Selected properties of size-matched Erdős–Rényi (ER) networks. See Table 2.2 for notation.

ER	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
ER($CH_{19}NO_9$)	4131	15,395.2	7.45	8.50×10^{-4}	4.37	3.71
ER($C_2H_6N_2O_2$)	17,014	61,701.4	7.25	1.94×10^{-4}	5.14	3.93
ER($C_3H_4N_2O$)	56,485	209,125.4	7.40	6.60×10^{-5}	5.69	4.24

Bipartite Random Networks The BR networks reflect the bipartite structure of RNs [23] while introducing randomness similar to ER networks. The construction of BR networks uses n nodes, which are split into two partitions according to the fraction b_s . Subsequently, e edges are randomly assigned between the partitions. As the RNs are all very close to parity ($b_s(CH_{19}NO_9) = 0.502$, $b_s(C_2H_6N_2O_2) = 0.528$, and $b_s(C_3H_5NO_2) = 0.505$), we chose $b_s = 0.5$ in all constructed BR networks. Table 2.4 shows the properties of the size-matched BR networks, which are similar to those of the ER networks and very unlike the corresponding RNs. The average square clustering coefficients, $\overline{C_4}$, of BR networks are about one to two orders of magnitude larger than those of ER networks but still much lower than those of the corresponding RNs. Similar to the ER networks, the fractal dimensions, d_f , of BR networks

are higher by 0.6 to 1.0 units compared to the corresponding RNs, while the average shortest path lengths, $\bar{l}$, are nearly half of those in RNs. Properties of the other six size-matched BR networks are given in Appendix A, Section A.6.2.

Table 2.4: Selected properties of size-matched bipartite random (BR) networks with partition fraction $p_s = 0.5$. See Table 2.2 for notation.

BR	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
BR(CH ₁₉ NO ₉)	4131	15,404	7.46	1.75×10^{-3}	4.58	3.57
BR(C ₂ H ₆ N ₂ O ₂)	17,014	61,762	7.26	4.24×10^{-4}	5.35	3.98
BR(C ₃ H ₄ N ₂ O)	56,485	208,955	7.40	1.26×10^{-4}	5.91	4.29

Watts–Strogatz Networks The WS networks allow the tuning of the randomness in a network by the p_r parameter, which corresponds to the probability of random edge rewiring. For moderate values of p_r , WS networks exhibit the small-world behavior, that is, $\bar{l} \sim \ln n$, as well as significant local ordering, expressed by high values of clustering coefficients [87–89, 92, 93]. Table 2.5 shows the properties of WS networks, which balance matching the average square clustering coefficients, $\overline{C_4}$, fractal dimension, d_f , and the number of edges, e , of the corresponding RNs.

The variation of the values of the network properties as a function of p_r shows that the best match between the $\overline{C_4}$ value and the fractal dimension, d_f , of the WS networks and their respective RN counterparts is obtained with a moderate amount of rewiring, $p_r = 0.25$ – 0.3 . However, this level of randomness creates much smaller average shortest path lengths, $\bar{l}$, relative to the corresponding RN. The value of d_f of WS networks is lower than that of random networks but 0.2 to 0.7 units higher than those of the corresponding RNs, and the value of d_f shows a much slower increase with the network size. On the other hand, the closest match of the average shortest path length, $\bar{l}$, is obtained for a significantly reduced amount of rewiring. For example, the C₃H₄N₂O network with $p_r = 0.05$ results in $\bar{l} = 10.69$, an overestimation of the average square clustering coefficient, $\overline{C_4} = 2.5 \times 10^{-1}$,

and an underestimation of the fractal dimension, $d_f = 3.16$. Properties for all size-matched WS networks, with matches based on $\overline{C_4}$, d_f , and balanced between the two are given in Appendix A, Section A.6.3.

Table 2.5: Selected properties of size-matched Watts–Strogatz (WS) networks with $K = 8$ nearest neighbors and rewiring probability p_r . See Table 2.2 for notation.

WS	p_r	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
WS(CH ₁₉ NO ₉)	0.25	4131	16,524	8.00	8.22×10^{-2}	5.03	3.37
WS(C ₂ H ₆ N ₂ O ₂)	0.3	17,014	68,056	8.00	5.98×10^{-2}	5.70	3.66
WS(C ₃ H ₄ N ₂ O)	0.3	56,485	225,940	8.00	5.99×10^{-2}	6.42	3.78

Non-Periodic Regular Lattices The LNP networks are constructed for each given dimensionality, d , by choosing the side length L to match the number of nodes, n , to the corresponding RN. The best match for the number of edges, e , is found if the dimensionalities of the LNP and RN networks are similar, that is, $d = 3, 4$. Since the dimensionality d of LNP networks is integer, only approximate matching of both n and e is possible. The calculations of the fractal dimensions, d_f , of LNP networks help identify the systematic errors of the CBB approximation relative to the underlying dimensionality d . Table 2.6 shows the properties of a one-dimensional LNP network with $n = 5000$, a two-dimensional LNP network with $n = 500$, and size-matched regular lattices with $d = 3, 4$.

The regular lattices show a high degree of clustering, as indicated by high average square clustering coefficients, $\overline{C_4}$, which are similar to that of RNs and WS networks, see Figure 2.4. For example, $\overline{C_4} = 7.12 \times 10^{-2}$, for the LNP network with $d = 4$ and $L = 16$. But LNP networks lack shortcuts typical of WS networks and, as a result, have long average shortest path lengths, $\bar{l}$, above those of RNs and far larger than random and WS networks. The CBB method underestimates the dimension of the underlying space; for example, for the one-dimensional chain of $L = 5000$, the CBB method gives $d_f = 0.86$, and for the two-dimensional $L \times L$ lattice with $L = 250$, the computed fractal dimension is $d_f = 1.73$. The

larger lattice with $L = 500$ has $d_f = 1.75$, which shows that d_f increases towards $d = 2$ but the convergence is slow. We also find convergence toward the underlying dimensionality for $d = 3$, where the size-matched lattice with $L = 15$ has $d_f = 2.15$, while for $L = 50$, we find $d_f = 2.40$. In LNP networks with $d = 4$, we obtain $d_f = 2.41$ for the size-matched lattice with $L = 7$ and $d_f = 2.88$ for $L = 20$. Although the theoretical dimensionality of the networks is underestimated by the CBB method, comparisons between networks of similar size and approximate dimensionality can be expected to be accurate.

Table 2.6: Selected properties of non-periodic regular lattices (LNP) with dimensionality d and side length L . See Table 2.2 for notation.

Lattice	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
$d = 1, L = 5000$	5000	4999	2.00	0	1666.67	0.86
$d = 2, L = 500$	250,000	499,000	3.99	1.25×10^{-1}	333.33	1.75
$d = 3, L = 15$	3375	9450	5.60	9.66×10^{-2}	14.93	2.15
$d = 3, L = 26$	17,576	50,700	5.77	9.22×10^{-2}	25.96	2.27
$d = 3, L = 39$	59,319	173,394	5.85	9.04×10^{-2}	38.97	2.36
$d = 4, L = 7$	2401	8232	6.86	8.08×10^{-2}	9.14	2.41
$d = 4, L = 12$	20,736	76,032	7.33	7.35×10^{-2}	15.89	2.68
$d = 4, L = 16$	65,536	245,760	7.50	7.12×10^{-2}	21.25	2.80

Periodic Regular Lattices The LP networks better approximate an infinite lattice as they include periodic boundary conditions, removing edge effects. Table 2.7 shows the network properties for a one-dimensional LP network (ring) with $n = 5000$ nodes, a two-dimensional LP network with $n = 500$, and size-matched LP networks with $d = 3, 4$. For the one-dimensional ring, the CBB method yields $d_f = 0.93$, which indicates that the error of the CBB method is less than 10%. However, the relative errors seem to grow as the dimensionality increases and the side length of the size-matched LP lattices decreases. In comparison to the corresponding RNs, the fractal dimensions of LP networks with $d = 3, 4$ are too low by 0.1–0.5 units. Periodic boundary conditions slightly reduce average shortest path lengths, $\bar{l}$, of regular networks; however, they remain larger than those of the corresponding RNs, see Figure 2.4. The average square clustering coefficients, $\overline{C_4}$, are somewhat reduced compared

to the LP networks; for example, $\overline{C_4} = 6.52 \times 10^{-2}$, in the LP network with $d = 4$ and $L = 16$. Moreover, the $\overline{C_4}$ values are independent of the side length, L .

Table 2.7: Selected properties of periodic regular lattices (LP) with dimensionality d and side length L . See Table 2.2 for notation.

Lattice	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
$d = 1, L = 5000$	5000	5000	2	0	1250.00	0.93
$d = 2, L = 500$	250,000	500,000	4	1.25×10^{-1}	250.00	1.81
$d = 3, L = 15$	3375	10,125	6	8.70×10^{-2}	11.20	2.29
$d = 3, L = 26$	17,576	52,728	6	8.70×10^{-2}	19.50	2.42
$d = 3, L = 39$	59,319	177,957	6	8.70×10^{-2}	29.23	2.45
$d = 4, L = 7$	2401	9604	8	6.52×10^{-2}	6.86	2.63
$d = 4, L = 12$	20,736	82,944	8	6.52×10^{-2}	12.00	2.87
$d = 4, L = 16$	65,536	262,144	8	6.52×10^{-2}	16.00	2.96

Regular Lattices with Rewiring As Figure 2.4 illustrates, the RNs are located in the same range of average square clustering coefficients, $\overline{C_4}$, as WS networks and regular lattices, and at intermediate values of fractal dimension, d_f , and average shortest path length, $\bar{l}$, between these two generative network models. In analogy to the edge randomization in WS networks, we thus modify LNP and LP networks by randomly rewiring edges with probability p_r . We denote the resulting networks with tunable “randomness” as LNP-R and LP-R networks, respectively.

Table 2.8 shows the properties of the size-matched LNP-R and LP-R networks for $d = 3, 4$. As in WS networks, a small amount of rewiring increases the value of d_f and decreases the value of $\bar{l}$ compared to the parent regular lattices, see Figure 2.4. A modest degree of rewiring suffices to match the fractal dimension d_f of the LNP-R and LP-R networks to that of the corresponding RNs. The LNP-R network with $d = 3$, $L = 39$, and $p_r = 0.1$ has $d_f = 3.62$, while the corresponding LP-R network has $d_f = 3.27$. The average square clustering coefficient is also well matched, $\overline{C_4} = 5.54 \times 10^{-2}$. Smaller RNs need smaller

values of p_r to match the value of d_f of the corresponding RNs. The LNP-R networks with $d = 3$ and $L = 15$ matches the corresponding RN at $p_r = 0.05$, and the best matching LP-R network with $d = 3$ and $L = 15$ has $p_r = 0.025$. With dimensionality $d = 4$, an even smaller amount of rewiring is sufficient; the LNP-R network with $L = 16$ and $p_r = 0.05$ achieves $\overline{C_4} = 5.73 \times 10^{-2}$ and $d_f = 3.61$.

The average shortest path lengths, $\bar{l}$, in these LNP-R and LP-R networks underestimate those of the corresponding RNs somewhat. Smaller amounts of rewiring produce a better match for the $\bar{l}$ value of the RNs. For example, the LNP-R network with $L = 16$ and $p_r = 0.05$ has $\bar{l} = 8.08$, compared to $\bar{l} = 10.39$ in the corresponding RN, see also Table 2.8. At the same time, choosing $p_r = 0.01$ gives $\bar{l} = 10.33$ but overestimates the values of $\overline{C_4}$ and d_f , as shown in Figure 2.4. An intermediate value of $p_r = 0.025$ offers the best trade-off between matching the metric, local, and global properties accurately. Generally, LNP-R networks match the corresponding RNs better than LP-R networks.

Figure 2.5 illustrates the effect of random rewiring on the average square clustering coeffi-

Table 2.8: Selected properties of periodic regular lattices with rewiring (LP-R) and non-periodic regular lattices with rewiring (LNP-R). d denotes the lattice dimensionality, L is the side length, and p_r is the rewiring probability. See Table 2.2 for notation.

Lattice	p_r	n	e	$\bar{k}$	$\overline{C_4}$	$\bar{l}$	d_f
LNP $d = 3, L = 15$	0.05	3375	9450	5.60	7.72×10^{-2}	7.11	2.79
LP $d = 3, L = 15$	0.025	3375	10,125	6.00	7.82×10^{-2}	7.33	2.75
LNP $d = 3, L = 26$	0.075	17,576	50,700	5.77	6.61×10^{-2}	8.00	3.27
LP $d = 3, L = 26$	0.05	17,576	52,728	6.00	6.98×10^{-2}	8.28	3.18
LNP $d = 3, L = 39$	0.1	59,319	173,394	5.85	5.76×10^{-2}	8.54	3.62
LP $d = 3, L = 39$	0.1	59,319	177,957	6.00	5.54×10^{-2}	8.36	3.27
LNP $d = 4, L = 7$	0.0075	2401	8232	6.86	7.81×10^{-2}	7.17	2.68
LP $d = 4, L = 7$	0.0001	2401	9604	8.00	6.52×10^{-2}	6.81	2.64
LNP $d = 4, L = 12$	0.025	20,736	76,032	7.33	6.60×10^{-2}	8.06	3.29
LP $d = 4, L = 12$	0.0001	20,736	82,944	8.00	6.52×10^{-2}	11.52	2.95
LNP $d = 4, L = 16$	0.05	65,536	245,760	7.50	5.73×10^{-2}	8.08	3.61
LP $d = 4, L = 16$	0.025	65,536	262,144	8.00	5.86×10^{-2}	8.47	3.26

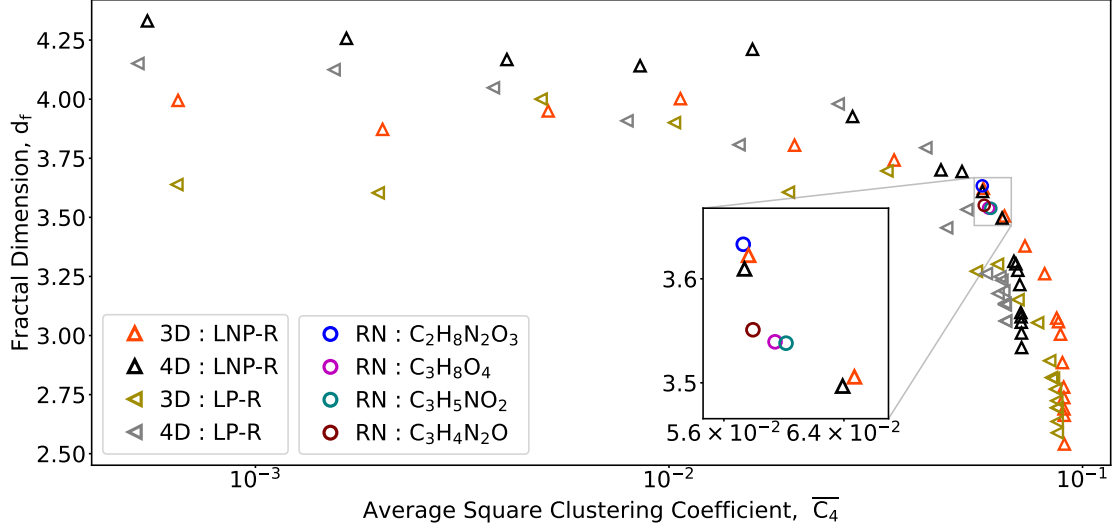


Figure 2.5: Plot of fractal dimension, d_f , against average square clustering coefficient, $\overline{C_4}$, for LNP-R networks with $d = 3$ (black triangles) and $d = 4$ (red triangles) and LP-R networks with $d = 3$ (gray triangles) and $d = 4$ (yellow triangles) in comparison to RNs (circles). The LNP-R and LP-R networks have $L = 39$ and $L = 16$ for $d = 3$ and $d = 4$, respectively.

cient, $\overline{C_4}$, and fractal dimension, d_f , over a larger range of p_r values. Similar to WS networks, two ranges of behavior are observed. For small rewiring probabilities, p_r , the value of d_f decreases by several units due to the occurrence of shortcuts, with only small changes in the average square clustering coefficient, $\overline{C_4}$. The networks are thus still relatively structured but gain the small-world property. For larger amounts of rewiring, the networks become increasingly random, and the values of $\overline{C_4}$ have a much stronger dependence on the rewiring probability, p_r , while the fractal dimension levels off. This behavior is observed for both LNP-R and LP-R networks with $d = 3$ and $d = 4$. Interestingly, the networks that most closely approximate RNs in Figure 2.5 are located close to the transition between the two ranges. The significance of this finding is not yet clear.

2.4 Discussion

2.4.1 Reaction Network Characterization

As Figure 2.1 shows, the properties of the RNs within each homologous series follow similar profiles as H_2O molecules are added to the base RN, with a characteristic crossover point between water-poor (smaller) and water-rich (larger) RNs. For a given homologous series, the number of nodes, n , of water-poor networks scales polynomially with the number of atoms, N_a , in line with our previous observations for CHO networks [23]. The size scaling of water-rich RNs is also polynomial in N_a , however, with a reduced slope. The crossover happens approximately between the third and sixth members of each homologous series and is reflected in local properties, such as average node degree, $\bar{k}$, and average square clustering coefficient, $\overline{C_4}$, metric properties, such as average shortest path length, $\bar{l}$, and diameter, D , and fractal dimension, d_f . The plots of all RN properties against N_a and an illustration of the crossover point for selected networks and properties are given in Appendix A, Section A.1 and A.7, respectively.

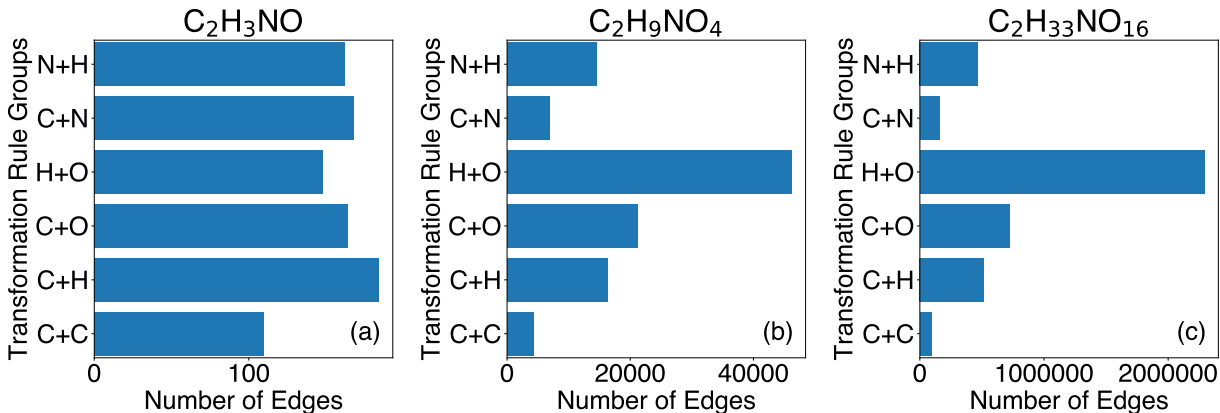


Figure 2.6: Numbers of network edges grouped by transformation rule in RNs of the $\text{C}_2\text{H}_3\text{NO}$ homologous series: (a) Base RN ($\text{C}_2\text{H}_3\text{NO}$), (b) RN containing three H_2O molecules ($\text{C}_2\text{H}_9\text{NO}_4$), and (c) RN containing 15 H_2O molecules ($\text{C}_2\text{H}_{33}\text{NO}_{16}$).

The water-poor RNs contain a wide distribution of transformation rules, including C–C, C–O and/or C–N, and E–H (E = C, O, N) rules and their multiple-bond counterparts (see Table 2.1). For example, Figure 2.6 shows the numbers of edges grouped by transformation rule in three RNs of the $\text{C}_2\text{H}_3\text{NO}$ homologous series. As Figure 2.6(a) shows, the base RN $\text{C}_2\text{H}_3\text{NO}$ has 936 edges, of which 11.8% correspond to C–C, 19.7% to C–H, 17.9% to C–N, 17.5% to C–O, 17.3% to N–H, and 15.8% to O–H transformation rules. In contrast, the transformations in water-rich RNs are dominated by the breaking and formation of O–H bonds in the solvent, masking the differences between RNs of different stoichiometries. In the water-rich $\text{C}_2\text{H}_{33}\text{NO}_{16}$ RN, which results from adding 15 H_2O molecules to the above network, over 53.9% of its 4.25×10^6 edges, as shown in Figure 2.6(c), are due to O–H bond breaks and bond formations. For the RNs above the crossover point, no new chemical transformations are added; instead, each network node increases by an H_2O unit. As a result, the RN with $N_w + 1$ H_2O molecules (where N_w is above the crossover point) is the product (Kronecker) graph of the preceding RN, containing N_w H_2O , and the 2-node graph describing water dissociation [23]. We thus expect RNs to become more regular with increasing dilution. Distributions of edges ungrouped by transformation rule are given in Appendix A, Section A.8.

The increase in RN size as a result of adding CH_2O and/or HCN monomers is exponential, as Figure 2.1(a) shows. In comparison with CHO networks [23], CHN networks show a faster increase in network size, while mixed CHNO networks fall in between. CHN networks also grow faster in the water-poor regime with respect to their average shortest path length, $\bar{l}$, and average node degree, $\bar{k}$, compared to CHO networks. In water-rich RNs, the value of $\bar{l}$ and the diameter, D , increase by a constant for each additional H_2O molecule. The increase of the average node degrees, $\bar{k}$, of the CHN and CHO networks also become similar in the water-rich regime. As for the network sizes, the mixed CHNO networks are intermediate in their increase between CHN and CHO networks.

We observe universal relationships for metric, local, and global network properties in a wide variety of CHNO networks shown in Figure 2.3. For the CHO networks, we have previously shown that the universal relationships for the fractal dimension, d_f , are also independent of the formal carbon oxidation state [23]. These findings indicate that the underlying structural principles of RNs are independent of the specific transformation rules and are the result of the octet rule and the use of polar transformations (electron-pair shifts), see Table 2.1. Moreover, algorithms based on structural features of RNs, including embedding, sampling, and search, are likely to be broadly applicable. Investigations of elements that form hypervalent compounds, for example, phosphorous and sulfur, and of radical reactions will help shed light on this hypothesis. As this chapter excludes some chemistry, if the rule set is expanded to include radical reactions, the bipartite nature of the RNs will be lost. Radical reaction rules are implemented in Chapter 4, though structural properties of the resulting RNs remain an avenue for future work. The small differences observed between CHO and CHNO networks imply that incorporating additional types of atoms and a large set of rules is unlikely to significantly alter the structural properties of the RNs.

As reported previously, the fractal dimension, d_f , of RNs is considerably smaller than the dimensionality of the underlying PES, $d_{\text{PES}} = 3N - 6$ [23]. The computed d_f values range from 0.89 to 4.55 for the RNs, for which these calculations were feasible, see Figure 2.1(f). The fractal dimension $d_f = 4.55$ was obtained for the $\text{C}_4\text{H}_8\text{O}_4$ RN, whose corresponding PES has dimensionality $d_{\text{PES}} = 42$. The dimensionality reduction is nearly tenfold. Despite the expanded rule set, we obtain similar reductions for CHN and mixed CHNO networks. For the $\text{C}_3\text{H}_3\text{N}_3$ and $\text{C}_3\text{H}_9\text{NO}_4$ RNs, the fractal dimensions $d_f = 3.55$ and $d_f = 4.41$ were obtained, with the corresponding PES dimensionality being $d_{\text{PES}} = 21$ and $d_{\text{PES}} = 45$, respectively, showing the nearly tenfold dimensionality reduction is irrespective of the stoichiometry for a range of RNs built with polar transformation rules as heuristics.

2.4.2 Comparisons to Generative Network Models

Generative network models are of interest because they can represent a multitude of network properties with just a few parameters. The dimensionalities of many such models have been studied in the literature. A particularly well-studied class of networks are scale-free (SF) networks, characterized by their highly heterogeneous structure, featuring a few highly connected nodes (hubs) alongside many nodes with low connectivity [90, 111]. In SF networks, the dimension is influenced by the scaling exponent, λ , which is a key parameter in the power-law distribution that describes the probability $P(k)$ that a node has k connections, following the relationship $P(k) \sim k^{-\lambda}$. When $\lambda > 4$, the fractal dimension, d_f , in ER and SF networks, stabilizes at 6, indicating a departure from scale-free behavior due to a rapid decline in degree distribution. Typical SF networks studied have a λ between 2 and 3. As λ approaches 3, d_f tends toward infinity. However, this holds only in the limit of infinitely large networks. For finite networks, an empirical d_f can be estimated, but it would, in effect, only show the finite-size effects of SF networks, which is why they are not included in this chapter for comparison [112].

The fractal nature of ER and SF networks at the boundaries (length scales above the average shortest path length, $\bar{l}$) has been explored and found to have a fractal dimension of $d_f \approx 2$. The nodes at the edges have a low degree, causing the low dimensionality [113]. For ER and other random graph models, analytical expressions for $\bar{l}$ and the diameter, D , have been found [114, 115].

Dimensionality calculations of regular lattices using approximate box-covering methods have been reported in the literature [116–118]. Depending on the method to calculate the dimensionality of regular lattices, the dimensionality is either found to be exact or approach the dimensionality of the underlying space. The spectral dimensions of linear, square, and triangular lattices are found to be 1.0, 2.0, and 3.0 dimensional, respectively [116]. Using a

random sequential box-covering algorithm, the fractal dimension, $d_f \approx 2.0$ was obtained for a 500×500 regular lattice [117]. The 50×50 LNP network was computed to have $d_f = 1.649$ instead of 2 using an approximate method because of the finite size [118], which is comparable to $d_f = 1.68$ obtained in this chapter for a 100×100 LNP network.

The effective dimension of WS-like networks varies as a function of the length scale on which it is measured. A transition from a linear to logarithmic scaling of $\bar{l}$ is observed as the typical length of shortcuts changes, similar to the transition seen in Figure 2.5 for LNP-R and LP-R networks [119]. A transition in $\bar{l}$ from linear scaling to logarithmic scaling was also found on increasing edge rewiring probability, p_r [120]. Higher values of p_r were shown to reduce the characteristic length scale in WS networks, which increases d_f [121]. In the random limit ($p_r = 1$), d_f increases with the number of nodes, n [122]. Similarly, in this chapter, the fractal dimension, d_f , of WS, LNP-R, and LP-R networks increases with larger p_r as shown for LNP-R and LP-R networks in Figure 2.5. The initial increase of d_f with p_r is quite steep but becomes shallower at $p_r \geq 0.2$, as shown in Figure 2.5. The increase of the fractal dimension, d_f , due to random rewiring in regular lattices was previously reported [123, 124].

Figure 2.4 shows the comparison of the average shortest path length, $\bar{l}$ (metric property), the average square clustering coefficient, $\overline{C_4}$ (local property), and the fractal dimension, d_f (global property), between three size classes of RNs: small ($C_2H_5NO_2$, $CH_{19}NO_9$, and $C_2H_{10}O_5$), medium ($C_3H_6O_3$, $C_2H_6N_2O_2$, and $C_2H_{18}O_9$), and large ($C_3H_8O_4$, $C_3H_4N_2O$, $C_2H_8N_2O_3$, and $C_3H_5NO_2$) and size-matched generative network models excluding the random networks (ER and BR). We note that most of these networks belong to the water-poor regime. The water-poor RNs provide a more stringent test of structural similarity to generative models since the networks become more regular with increasing dilution, as discussed above.

The existing body of work has demonstrated that randomness increases fractal dimension, d_f , reduces average shortest path length, $\bar{l}$, and suppresses local clustering [86, 113, 119–127]. Consequently, the size-matched random ER and BR networks have values of d_f higher by 0.6

to 1.1 units and values of $\overline{C_4}$ smaller by one to three orders of magnitude in comparison to the RNs. BR networks show higher local clustering but are still much less ordered than RNs. Although both random networks and RNs show the same scaling of $\bar{l}$ with the network size, that is, belong to small-world networks, the random networks are “smaller” with respect to their metric properties by a factor of two. On the other end of the spectrum, size-matched regular lattice networks with $d = 3, 4$ are similar to the corresponding RNs with respect to their average square clustering coefficients but have longer average shortest path lengths and lower CBB fractal dimensions.

The WS networks demonstrate the effect of randomness on network properties within the same generative network model. The dependence of the clustering coefficient and the diameter on the rewiring probability, p_r , was already described by Watts and Strogatz in their seminal paper [86]. Additionally, random rewiring increases the fractal dimension, d_f , bringing it closer to the values observed in ER and BR networks. However, the WS model cannot simultaneously satisfy the matching conditions for metric, global, and local network properties.

The LNP-R and LP-R networks apply the random rewiring to low-dimensional regular lattices, as shown in Figure 2.5. These models with $d = 3, 4$ require a smaller amount of rewiring than WS networks to match the average square clustering coefficient, $\overline{C_4}$, and the fractal dimension, d_f , of the corresponding RNs with good accuracy. Therefore, it is possible to simultaneously match metric, local, and global properties of RNs by regular lattices with rewiring. The necessary amount of rewiring in LNP-R and LP-R models is small, $p_r \leq 0.1$, indicating that RNs are more regular than random. Thus, RNs are not only low-dimensional *metric spaces*, but moreover, they resemble low-dimensional *Euclidean spaces*, represented by the regular lattices.

The similarities between RNs and LNP-R networks with respect to the local and metric properties suggest the presence of inherent symmetries within the RNs. The symmetry arises from

the fact that the order of bond formations or breaking often does not matter, which creates regular patterns in the RNs, which have been observed in previous work [23, 84]. The regular patterns found in the RNs are also reflected in the relatively narrow distributions of the node degrees, k , and the square clustering coefficients, C_4 , shown in Figure 2.2. Most nodes are thus embedded in similar local neighborhoods, as is the case in regular lattices. The above analysis was performed on water-poor RNs and applies even more to the water-rich RNs, which exhibit a more regular structure. Finally, we note that the RNs studied in this chapter did not take into consideration thermodynamic or kinetic feasibility of the reaction paths and constitute the superset of the experimentally observed reactions [24]. The dimensionality of the LNP-R and LP-R networks is thus an upper bound to the subnetworks of the RN that represent feasible reactions. To gain further insights into RNs, methods such as disconnection graphs [128], which visualize energy landscapes, and persistent homology [129], which captures the global topological features of the network, could be employed.

2.5 Conclusions

In this chapter, we investigated local, metric, and global properties of RNs and showed that they share common structural characteristics irrespective of their size, composition, and rule set and can be modeled by low-dimensional non-periodic regular lattices with a small amount of random rewiring. Thus, in addition to being low-dimensional metric spaces [23], RNs are also close to low-dimensional Euclidean spaces. The coordinate representation of RNs can be constructed by embedding into $\mathbb{R}^d$ with $d = 3, 4$ for the networks considered in this chapter, generating a tenfold dimensionality reduction with a small error. The reduced-dimensionality representations are useful for speeding up sampling and search of RNs. Network embedding is widely used in similarity search, classification, clustering, and visualization tasks [127, 130–132].

RNs have the potential to be effectively utilized for identifying transition states. Transition state searches would first locate a potentially feasible reaction pathway within a RN, and then interpolation between neighboring nodes would be employed to estimate the location of the corresponding transition state. RNs offer a practical starting point for initiating the transition state search. At the same time, the rewiring probability p_r in the LNP-R and LP-R networks offers an *a posteriori* measure of the average deviation from Euclidean structure and is related to the error of the dimensionality reduction between the parent PES and the low-dimensional Euclidean space.

The limitation of the current approach is that it requires explicit RN construction, which is limited by the computational cost and storage needs. However, the understanding of the dimension and amount of rewiring as a function of the RN composition, oxidation state, and rule set developed in this chapter provides the target dimensionality and error measures for the network embedding. These parameters can then be applied to embed subnetworks constructed around a compound or reaction of interest. On-the-fly calculations of thermodynamic and kinetic feasibility [24] further restrict the size of a subnetwork.

Currently, the node representation ignores stereoisomers and conformers, because the three-dimensional molecular structures are not taken into consideration. This could be partially remedied by using low-dimensional representations of conformer equilibria, which have been recently developed [133]. Finally, the expressiveness of RNs is limited by their reliance on simple graph-based rules such as bond breaks and bond formations. This representation, like other graph-based molecular models, excludes reactions of transition-metal compounds and intermolecular complexes, which cannot be expressed by composing discrete rules. Explicit dimensionality reduction methods might be of interest for treating these chemical systems.

Chapter 3

Euclidean Embeddings of Reaction Networks

The motivation for this chapter comes from the desire to embed reaction networks into Euclidean space. In Chapter 2, stoichiometry-preserving reaction networks were shown to resemble low-dimensional regular lattices with minimal edge rewiring, independent of system size, stoichiometry, and rule set. However, those networks used constant edge weights, and it was unclear whether this regular structure would persist when chemically relevant weights are introduced, such as those derived from the karc criterion, or whether the networks could still be embedded into low dimensions. This chapter answers that question and shows that the low-dimensional regular structure is preserved with karc edge weights and can still be leveraged. Having Euclidean coordinates with distances tied to reaction feasibility makes identifying new feasible reactions faster than searching a generic metric space, opening the possibility of efficient search and sampling without the need for exhaustive potential energy surface (PES) studies. We find that nonlinear embedding methods perform best, and develop a distance-preserving encoder combining landmark multidimensional scaling (LMDS) initialization with neural network refinement, as classical methods failed to scale

or produce embeddings that preserved distances sufficiently well. The work presented in this chapter has been submitted for publication as “*Euclidean Embedding of Reaction Networks for Dimensionality Reduction of Chemical Spaces*” in the Journal of Chemical Theory and Computation. Co-authors Liliana Garcia and Kent Ziti were undergraduate researchers who contributed to the initial constant edge weight embedding study and the LMDS implementation, respectively.

Exploring complex chemical reactions is challenging due to the high dimensionality of PESs. Stoichiometry-preserving RNs, constructed using bond-breaking and bond-forming rules, provide a discrete representation of chemical space with an underlying low-dimensional metric structure. Moreover, these RNs resemble regular lattices with a small degree of random edge rewiring. In this chapter, we show that these RNs coupled with weighted-path metrics can be embedded into low-dimensional Euclidean spaces, producing a significant and chemically informed dimensionality reduction of the PES. Across CHNO RNs spanning diverse stoichiometries and sizes, we compare linear embedding methods, including classical multidimensional scaling (cMDS) and LMDS, with nonlinear approaches such as metric MDS (mMDS), Isomap, t-distributed stochastic neighbor embedding (t-SNE), and kernel principal component analysis (KPCA). Incorporating kinetic heuristics as edge weights does not significantly increase embedding error or the optimal embedding dimensionality relative to constant edge weights. Although mMDS provides the best distance preservation, it scales poorly with network size. To address this limitation, we introduce a distance-preserving encoder that combines LMDS initialization with neural network refinement and achieves mMDS-level accuracy at a computational cost comparable to LMDS. The resulting embeddings exhibit high fidelity (Kendall $\tau > 0.84$, normalized stress $\sigma < 0.09$, and $R^2 > 0.95$) and reduce the dimensionality of the reactive chemical space by a factor of 3–10 relative to the underlying PES. The optimal embedding dimension scales as $\mathcal{O}(\log n)$ with the number of nodes n . Randomization experiments indicate that the observed near-Euclidean structure is specific to RNs and is related to their chemically constrained topologies and symmetric

edge weights, in contrast to general sparse graphs. Pathway discovery experiments further demonstrate that shortest-path searches in the embedding space recover chemically feasible pathways consistent with those identified in the original RNs. These results establish low-dimensional Euclidean embeddings of RNs as a compact and searchable representation of chemical reaction spaces.

3.1 Introduction

Chemical reaction networks are discrete algebraic structures that are amenable to the tools of graph theory and combinatorics and provide a versatile framework for representing and exploring chemical reactivity [2–4, 7, 134–136]. Applications of reaction network models range from chemical process modeling [3, 5, 136], reaction mechanism analysis [6, 7], and synthetic path prediction [8] to metabolism [9, 135, 137, 138], atmospheric chemistry [17], and prebiotic evolution [18]. There exist multiple types of reaction networks differing in their definitions of nodes and edges. In stoichiometry-preserving reaction networks, nodes describe collections of molecules with fixed atomic composition and edges represent transformations that conserve stoichiometry, usually bond breaks and bond formations [22–24]. Thus, stoichiometry-preserving reaction networks can be viewed as discretized representations of the underlying PES. In contrast, reaction network models most commonly used in kinetic modeling use molecular species as nodes and reactions as edges connecting reactants to products. Reaction networks can be constructed using a variety of cheminformatics and quantum chemistry approaches. These include rule-based methods [25–29], heuristics-aided quantum chemistry [22, 24], hybrid approaches combining reaction templates with quantum chemical rate estimates [13, 30], and PES exploration techniques including the Global Reaction Route Mapping method [33], growing string methods [6, 35], and molecular dynamics-based approaches such as the *ab initio* nanoreactor [45]. Additional approaches

include deep learning methods combined with quantum chemical calculations [55] and integrated frameworks such as Chemoton [7, 31, 32] and YARP [49, 53].

Identifying thermodynamically and kinetically feasible pathways from reactants to products is a central task in reaction network analysis. With appropriate heuristic edge weighting, standard shortest-path searches [139] can find feasible pathways efficiently. We have previously proposed the *karc* kinetic heuristic [23, 24], which incorporates the energies of the path intermediates based on Hammond’s postulate [71]. *karc* uses the energy of the highest-energy path intermediate as a surrogate for the activation barrier of the rate-limiting step, in combination with a length penalty [24]. Such chemically informed weights make the shortest-path lengths a meaningful measure of reaction feasibility. Weighted shortest-path searches have been applied to identify kinetically relevant pathways in large reaction networks [52, 61, 62]. Alternative approaches such as integer linear programming [140, 141] have also been proposed.

Despite the enormous popularity of chemical reaction networks, relatively little is still known about their underlying structure. Early analyses of biochemical networks showed highly heterogeneous structures with few highly connected hubs and many peripheral nodes [9, 138]. However, these findings were specific to the selection of species and reactions in the reaction network model. In contrast, stoichiometry-preserving reaction networks are finite by construction and have a much simpler, low-dimensional structure [23]. Moreover, as we showed in our previous chapter, key properties of these networks agree with those of low-dimensional regular lattices subject to some random edge rewiring [1]. The simple structure of stoichiometry-preserving reaction networks can be viewed as a regular “scaffold” inside PESs, on which all other non-reactive degrees of freedom are superimposed. To our knowledge, this underlying lattice-like structure had not been previously recognized. These findings can inform the development of structure-aware methods for PES exploration, sampling, and optimization that exploit the intrinsic lattice structure. By restricting the search to the

lattice rather than the full configuration space, such methods could mitigate the effect of the curse of dimensionality.

The analysis of (weighted) undirected graphs can be significantly simplified by embedding them into low-dimensional Euclidean spaces [142, 143]. These embeddings are typically approximate, and the quality of the embedding is determined by the deviation of the pairwise distances in the embedding space from the shortest-path distances on the graph. An undirected graph only encodes pairwise distances and lacks intrinsic coordinates; graph embedding restores the coordinate system, often as a method of dimensionality reduction. The regular lattice-like structure of stoichiometry-preserving reaction networks leads us to hypothesize that they should be embeddable into low-dimensional Euclidean spaces with small distortion. The resulting graph embeddings provide a compact representation of the reaction network and, by extension, of the underlying PES, enabling efficient interpolation and path finding.

In this chapter, we investigate embedding accuracy across a range of linear and nonlinear methods applied to reaction networks. As a linear baseline, we consider classical multidimensional scaling (cMDS) [144–146] and its scalable variant, landmark MDS (LMDS) [147]. We apply nonlinear approaches including metric multidimensional scaling (mMDS) [148], Isomap [149], t-distributed stochastic neighbor embedding (t-SNE) [150], and kernel principal component analysis (KPCA) [151]. For each method, we analyze the embedding accuracy as a function of the embedding dimension and edge-weighting scheme. In addition, we develop a scalable neural network-based distance-preserving encoder and evaluate its performance. The embedding quality is assessed using both rank-based and distance-based metrics [152, 153] across ten CHNO reaction networks spanning a wide range of sizes. To probe the origin of embeddability, we perform randomization experiments on network topology and edge weights. Finally, we evaluate whether low-dimensional embeddings can correctly recover feasible reaction pathways identified by shortest-path analysis. This is demonstrated through

pathway discovery experiments on five benchmark reactions, including enol tautomerization, HCN trimerization, esterification, and two formose-type aldol reactions.

This chapter is organized as follows. Section 3.2 describes the reaction network representation, edge-weighting schemes, embedding methods, and error metrics. Section 3.3 presents embedding results for linear, nonlinear, and encoder-based methods as well as randomization and pathway discovery experiments. Section 3.4 contains the discussion and Section 3.5 summarizes the main conclusions.

3.2 Methods

3.2.1 Reaction Network Representation

The RNs considered in this chapter are of the stoichiometry-preserving type [22, 23]. In these networks, nodes represent collections of molecules with fixed atomic composition and edges correspond to stoichiometry-preserving transformations from a predefined rule set. We restricted ourselves to networks consisting of carbon-, hydrogen-, nitrogen-, and oxygen-containing molecules (CHNO networks). We used reaction rules that describe heterolytic bond breaking and bond formation as well as fictitious polarization and depolarization transformations [24]. The latter help represent the reactivity of multiple bonds in a composable manner. With these definitions, any reaction expressible as a sequence of electron-pair transfer (“arrow-pushing”) steps [66, 67] can be represented within a reaction network. The transformation rules in this chapter were chosen to be complete with respect to normal polarity principles and to be reversible. The full set of reaction rules is provided in Appendix B, Section B.1.1.

Edges can be assigned weights to reflect their thermodynamic and kinetic feasibility. In the

simplest case, all edges are assigned a constant weight (distance) of one (W_c), treating all transformations as equally feasible. To incorporate more chemical information, we apply karc kinetic feasibility heuristics [22, 24] (W_k) as edge weights. The karc heuristic is motivated by Hammond’s postulate [71] and uses the energy of the highest-energy intermediate as a surrogate for the activation barrier of the rate-limiting step, together with a length penalty. For the network edge between nodes K_i and K_j , having energies E_{K_i} and E_{K_j} , the W_k edge weight (distance) is

$$d_{ij}^k = \left[(E_{K_j} - E_{K_i})^2 + \alpha^2 \left(\frac{1 - \exp \left(-\kappa (E_{K_j} - E_{K_i})^2 \right)}{1 + \exp \left(-\kappa (E_{K_j} - E_{K_i})^2 \right)} \right) \right]^{1/2}, \quad (3.1)$$

where the length penalty α and the short-edge rescaling parameter κ are chosen to correctly classify feasible and infeasible reaction pathways in a set of test reactions. In this chapter, we used the values $\alpha = 1$ eV and $\kappa = 5$ eV⁻² found previously [24].

Node energies were computed as sums of molecular energies of all constituent molecules. Molecular energies were calculated at the PM7/COSMO level of theory using MOPAC [154, 155], with the dielectric constant of water ($\epsilon = 78.4$). Intermediate structures arising from fictitious polarization and depolarization rules were handled by grouping molecules into equivalence classes of zero-energy transformations and assigning each class a representative energy. Further details of equivalence class assignment, conformer generation, and energy calculations are given in Appendix B, Section B.1.2.

Embedding methods were evaluated on a set of ten test RNs (referred to as comparison networks in the previous chapter [1]) and five chemically motivated benchmark RNs. The test RNs spanned sizes from approximately $n \approx 4000$ to $n \approx 56,000$ nodes and were grouped by scale into small ($C_2H_5NO_2$, $CH_{19}NO_9$, and $C_2H_{10}O_5$), medium ($C_3H_6O_3$, $C_2H_6N_2O_2$, and $C_2H_{18}O_9$), and large ($C_3H_8O_4$, $C_3H_4N_2O$, $C_2H_8N_2O_3$, and $C_3H_5NO_2$) RNs. For each test RN, embeddings were generated for RNs with W_c and W_k edge weights, while benchmark RNs

were tested with W_k edge weights only. The benchmark RNs included enol tautomerization (C_2H_4O), HCN trimerization ($C_3H_3N_3$), esterification ($C_3H_8O_3$), and two reactions related to the formose reaction [156]: C_3 aldol addition to glyceraldehyde ($C_3H_6O_3$) and C_4 aldol addition to tetrose ($C_4H_8O_4$). Network sizes and edge weight statistics for all RNs are summarized in Appendix B, Section B.1.3.

RNs were constructed using the `colibri2` software package [68] as described in the previous chapter [1]. Reaction rules were encoded as SMARTS patterns [75] and applied to molecular structures using the RDKit library [77]. Network generation was seeded with an initial molecular collection and proceeded iteratively until no new nodes could be produced. Because the network generation procedure is exhaustive, any starting molecular collection of a given stoichiometry produces the same reaction network. The resulting networks were converted to undirected graphs and stored as compressed GraphML files [76]. All subsequent network analysis and manipulation used the NetworkKit [101] and NetworkX [157] libraries (codes used beyond `colibri2` are available on GitHub [158]).

3.2.2 Conventional Embedding Methods and Error Metrics

Several conventional linear and nonlinear dimensionality reduction methods were evaluated for RN embeddings. The linear methods included cMDS [144–146] and LMDS [147]. The nonlinear methods included mMDS [148], Isomap [149], t-Distributed Stochastic Neighbor Embedding (t-SNE) [150] and direct kernel principal component analysis (KPCA) [151] using a kernel constructed from the distance matrix. Brief descriptions of each method are provided in Appendix B, Section B.2.

For each RN, $G = (V, E)$ with $n = |V|$ nodes and $e = |E|$ edges, shortest-path distances between all node pairs were computed using the all-pairs shortest-path implementation in NetworkKit [101], which applies Dijkstra’s algorithm [139] from each node. Distances were

computed for both W_c and W_k edge weights.

Each method constructs a low-dimensional embedding $Z \in \mathbb{R}^{n \times d}$ by optimizing a distinct objective; the rows $Z_i \in \mathbb{R}^d$ are the embedded coordinates of the nodes. cMDS recovers coordinates by eigendecomposition of the Gram matrix $\mathbf{B}$ or inner products. $\mathbf{B}$ is obtained from the distance matrix by double-centering [144, 145],

$$\mathbf{B} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(2)} \mathbf{H}, \quad \mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top, \quad (3.2)$$

where $\mathbf{H}$ is the centering matrix and $\mathbf{D}^{(2)}$ is the matrix of squared pairwise distances with entries $(\mathbf{D}^{(2)})_{ij} = d_{ij}^2$. The leading eigenvectors (ordered by decreasing eigenvalue) give the embedding coordinates. LMDS applies the same procedure to the landmark subset of $n_l = p_l n$ nodes, where p_l is the landmark fraction, and positions the remaining nodes by distance-based triangulation [147]. mMDS avoids the explicit construction and eigendecomposition of $\mathbf{B}$ by iteratively minimizing the normalized stress σ [148] with

$$\sigma = \sqrt{\frac{\sum_{i < j, d_{ij} > 0} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i < j, d_{ij} > 0} d_{ij}^2}}, \quad (3.3)$$

where d_{ij} are the original shortest-path distances and $\hat{d}_{ij} = \|Z_i - Z_j\|$ are the corresponding Euclidean distances in the embedding space. Isomap is the curved-manifold analog of cMDS and performs the eigendecomposition of the geodesic distance matrix estimated on a k -nearest-neighbor graph [149]. The remaining nonlinear methods are often used in high-dimensional data analysis. t-SNE optimizes a probabilistic objective that favors local neighborhood structure over global distances [150]. In contrast, KPCA performs principal component analysis on $\mathbf{B}$ after transformation by a kernel function [151].

For all methods, embedding quality was evaluated by comparing original pairwise distances

with Euclidean distances in the embedding space using rank-based and distance-based error metrics. Rank preservation was quantified using Kendall rank correlation coefficient (τ) [152]. Global distance preservation was assessed using the coefficient of determination (R^2) and normalized stress (σ) [148, 153]. Additional metrics are described in Appendix B, Section B.3.

The quality of the embeddings depends significantly on the embedding dimension d . Low embedding dimensions provide more compact representations and allow for more efficient path finding, but if the dimension is too low, important structural information may be lost. To determine the optimal embedding dimension d_e , normalized stress σ was evaluated as a function of the dimension d . Two distinct classes of behavior were observed depending on whether the embedding method was linear or nonlinear. For linear methods, σ decreased with increasing dimensionality up to a minimum and then increased as additional components were added. This non-monotonic behavior has been previously reported for classical MDS [159]. The optimal embedding dimension for linear methods was thus computed as the value of d that minimized σ . For nonlinear methods, including mMDS, Isomap, t-SNE, and KPCA, stress generally decreased monotonically with increasing d and approached an asymptotic plateau. To identify d_e in this regime, $\log_2 \sigma$ was computed as a function of d . The numerical first derivative of this curve was then fitted to a sigmoidal function, and d_e was determined as the position of the inflection point in the fit, rounded to the nearest integer. d_e values from this procedure are accurate to ± 1 –2 dimensions.

3.2.3 Distance-Preserving Encoder

Neural network-based formulations of MDS have been previously proposed as flexible alternatives to classical eigendecomposition-based approaches [160–163]. In these methods, the vector of initial coordinates is passed through a feed-forward encoder to obtain a nonlinear transformation of the coordinate space, using the distance-preserving objective itself as the

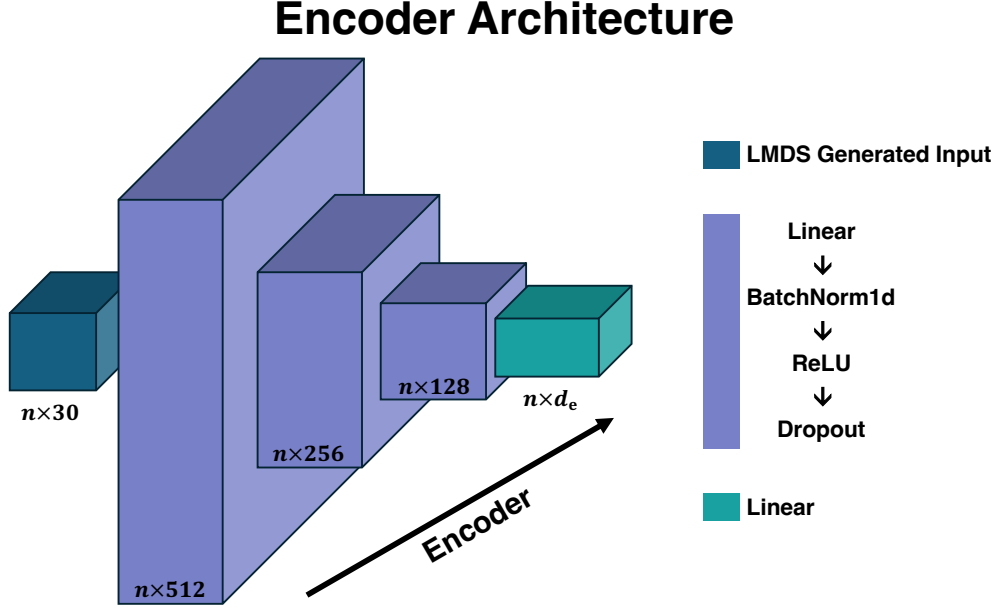


Figure 3.1: Architecture of the distance-preserving encoder. Landmark multidimensional scaling (LMDS) pre-embedded coordinates ($n \times 30$) are passed through a fully connected feed-forward network with hidden-layer dimensions 512, 256, and 128 to an output layer of dimension d . Each hidden layer comprises a linear transformation followed by BatchNorm1d, ReLU, and dropout. The final output layer is a linear transformation.

loss function. The approach used here is related to such neural distance-preserving embedding methods but differs in that the neural network operates as a refinement step from a conventional pre-embedding rather than learning the embedding from scratch.

The encoder generates a nonlinear mapping f_θ of the input coordinate matrix $X \in \mathbb{R}^{n \times d_s}$ with starting dimension d_s to the optimized embedding $Z = f_\theta(X) \in \mathbb{R}^{n \times d}$ of dimension d . The training procedure uses stochastic optimization to minimize the normalized stress σ (Equation 3.3), evaluated on the sampled training subset of node pairs. In contrast, conventional nonlinear dimensionality reduction methods such as mMDS minimize stress by deterministic majorization [164].

The encoder used a fully connected feed-forward network with three hidden layers, $d_s \rightarrow h_1 \rightarrow h_2 \rightarrow h_3 \rightarrow d$, with hidden dimensions $h_1 = 512$, $h_2 = 256$, and $h_3 = 128$. Each hidden layer consisted of a linear transformation followed by batch normalization [165], ReLU acti-

vation [166], and dropout [167] with rate 0.1. The output layer was a linear transformation with no activation function. A schematic representation of the encoder architecture is shown in Figure 3.1.

Two independent random subsets of nodes were drawn. The first, the landmark set (p_l fraction), was used in the LMDS pre-embedding; the second, the encoder set (p_n fraction), defined the node pairs used to train and evaluate the encoder. All pairs of nodes from the encoder set were partitioned into training, validation, and test subsets in a 70/15/15 ratio. In this chapter, we used $p_l = 0.3$ and $p_n = 0.3$. For the largest benchmark RN ($C_4H_8O_4$, 17,841,351 node pairs), $p_l = 0.005$ and $p_n = 0.005$ were chosen. The two subsets were drawn independently.

Distances were computed using the single-pair shortest path (SPSP) implementation in NetworkKit [101], which applies Dijkstra’s algorithm [139] from each unique source node in the sampled set. This approach avoids explicit construction of the full distance matrix and scales to large RNs.

To reduce memory cost, mini-batch training was used [168]. For each batch of node pairs, the forward pass was performed only on nodes present in that batch, followed by backpropagation and an optimizer update. Optimization was carried out by AdamW [169] with cosine annealing learning-rate scheduling [170], gradient clipping [171] with maximum norm 1.0, and early stopping based on validation stress.

The encoder was initialized by vectors with starting dimension $d_s = 30$. While LMDS-based and random initialization were both investigated, LMDS pre-embedding produced faster convergence and lower final σ , and was used for all reported embeddings. Hidden-layer weights were initialized using Kaiming normal initialization [172] (`nonlinearity='relu'`, `mode='fan_out'`), while the output layer weights were sampled from a Gaussian distribution with mean zero and standard deviation 0.01. The biases for hidden and output layers were

initialized to zero. The encoder was implemented in PyTorch [173], and the architecture was selected by comparing several candidate architectures.

3.3 Results

The results are organized as follows. Subsections 3.3.1, 3.3.2, and 3.3.3 evaluate embedding quality across ten test RNs for linear methods, conventional nonlinear methods, and the distance-preserving encoder, respectively. Subsection 3.3.4 presents calculations of the optimal embedding dimension for linear, nonlinear and encoder-based methods. Subsection 3.3.5 probes the structural origins of embeddability through randomization experiments on test RNs. Finally, subsection 3.3.6 reports pathway discovery experiments on five benchmark RNs.

3.3.1 Linear Methods

cMDS [144–146] and its scalable variant, LMDS [147], were evaluated as a linear embedding baseline. In LMDS, a randomly chosen landmark subset of $m = p_1 n$ nodes is embedded using cMDS, and the remaining nodes are positioned via distance-based triangulation using the Nyström method [174–176]. Embeddings were generated for the test RNs with both constant (W_c) and karc (W_k) edge weights. Embedding quality was evaluated with respect to rank correlation (evaluated by Kendall τ) and global distance preservation (using coefficient of determination R^2 and normalized stress σ as metrics). Table 3.1 reports the results for representative small ($\text{CH}_{19}\text{NO}_9$), medium ($\text{C}_2\text{H}_6\text{N}_2\text{O}_2$), and large ($\text{C}_3\text{H}_4\text{N}_2\text{O}$) RNs with embedding dimensions $d = 3$ and $d = 10$. The results for all test RNs with cMDS and LMDS are given in Appendix B, Sections B.4 and B.5, respectively.

Increasing the embedding dimension improved both rank correlation and global distance

Table 3.1: Embedding quality metrics (Kendall τ , coefficient of determination R^2 , and normalized stress σ) for selected test reaction networks (RNs) with constant (W_c) and karc (W_k) edge weights using classical multidimensional scaling (cMDS) and landmark MDS (LMDS, $p_1 = 0.3$) with $d = 3$ and $d = 10$ embedding dimensions.

		CH ₁₉ NO ₉		C ₂ H ₆ N ₂ O ₂		C ₃ H ₄ N ₂ O		
		<i>d</i> = 3	<i>d</i> = 10	<i>d</i> = 3	<i>d</i> = 10	<i>d</i> = 3	<i>d</i> = 10	
<i>W_c</i>	cMDS	<i>τ</i>	0.714	0.867	0.530	0.758	0.530	0.699
		<i>R</i> ²	0.625	0.721	−0.670	0.743	−1.505	0.739
		<i>σ</i>	0.203	0.175	0.348	0.136	0.375	0.121
	LMDS	<i>τ</i>	0.714	0.865	0.529	0.757	0.529	0.699
		<i>R</i> ²	0.626	0.721	−0.683	0.743	−1.507	0.739
		<i>σ</i>	0.203	0.175	0.349	0.137	0.375	0.121
<i>W_k</i>	cMDS	<i>τ</i>	0.794	0.867	0.526	0.729	0.549	0.713
		<i>R</i> ²	0.854	0.873	−0.235	0.803	0.056	0.772
		<i>σ</i>	0.157	0.146	0.427	0.171	0.424	0.208
	LMDS	<i>τ</i>	0.792	0.866	0.523	0.723	0.546	0.712
		<i>R</i> ²	0.852	0.872	−0.257	0.796	0.053	0.766
		<i>σ</i>	0.158	0.147	0.431	0.174	0.424	0.211

preservation metrics for both methods. For the CH₁₉NO₉ RN with W_c edge weights, Kendall τ improved from 0.714 at $d = 3$ to 0.867 at $d = 10$ with cMDS, while R^2 increased from 0.625 to 0.721 and normalized stress σ decreased from $\sigma = 0.203$ to 0.175. The LMDS results track those of cMDS very closely at $d = 3$ and $d = 10$. In medium-sized RNs, low-dimensional embeddings were highly distorted with both methods. For the C₂H₆N₂O₂ RN at $d = 3$ with W_c edge weights, we obtained $R^2 = -0.670$ with cMDS and $R^2 = -0.683$ with LMDS, indicating severe mismatch between original and embedded distances. Increasing the embedding dimension to $d = 10$ resolved most of the observed distortion, yielding $R^2 = 0.743$ with both methods. Similar trends were observed for the larger C₃H₄N₂O RN. We note that the landmark approximation introduced negligible error for all test RNs, remaining at least one order of magnitude smaller than the error due to dimensional truncation.

Embeddings constructed with W_k edge weights showed very similar trends to those con-

structed with W_c edge weights. For the $\text{CH}_{19}\text{NO}_9$ RN, embeddings with W_k edge weights performed better than embeddings with W_c edge weights at both $d = 3$ and $d = 10$ across all metrics. For cMDS at $d = 3$, R^2 increased from 0.625 to 0.854, while stress decreased from 0.203 to 0.157. For medium and large RNs with W_k edge weights, low-dimensional linear embeddings were highly distorted, similar to RNs with W_c edge weights. For the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN at $d = 3$, $R^2 = -0.235$ with W_k edge weights for cMDS, indicating the same severe mismatch between original and embedded distances as seen for W_c edge weights. Increasing the dimension to $d = 10$ resolved most of this distortion, yielding $R^2 = 0.803$ with W_k edge weights with cMDS. The LMDS results were again essentially identical to those for cMDS, showing negligible errors compared to dimensional truncation error. Overall, we find that edge weighting had only a moderate effect on embedding quality metrics.

The computational cost of cMDS has three contributions: all-pairs shortest-path computation ($\mathcal{O}(n^2 \log n)$ time and $\mathcal{O}(n^2)$ space), double-centering of the distance matrix ($\mathcal{O}(n^2)$ time and space), and eigendecomposition of the resulting $n \times n$ kernel matrix ($\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space). The eigendecomposition dominates in time, while peak memory reflects the simultaneous storage of multiple $n \times n$ matrices. For the $\text{C}_2\text{H}_8\text{N}_2\text{O}_3$ RN ($n = 58,791$), the kernel matrix alone requires approximately 27 GB at float64 precision. Peak memory demands thus make naive cMDS impractical for large networks. LMDS reduces the computational cost to $\mathcal{O}(n_1^3) + \mathcal{O}(n_1 n^2)$ in time, where the second term dominates for $m \ll n$, and $\mathcal{O}(n_1 n)$ in space. As a result, the peak memory is decreased to approximately 5 GB for the $\text{C}_2\text{H}_8\text{N}_2\text{O}_3$ RN at $p_1 = 0.3$, a more than five-fold reduction with negligible loss in embedding quality.

3.3.2 Nonlinear Methods

To determine whether nonlinear embeddings improve distance preservation beyond the linear methods of cMDS and LMDS, we evaluated RN embeddings with metric multidimensional scaling (mMDS) [148], Isomap [149], t-distributed stochastic neighbor embedding (t-SNE) [150], and kernel principal component analysis (KPCA) [151]. Unlike cMDS and LMDS, these methods rely on nonlinear transformations or nonlinear geometric constraints. As edge weighting was shown to have only a moderate effect on embedding quality, only RNs with W_k edge weights were considered with nonlinear embedding methods.

We first consider mMDS results for representative small ($\text{CH}_{19}\text{NO}_9$), medium ($\text{C}_2\text{H}_6\text{N}_2\text{O}_2$), and large ($\text{C}_3\text{H}_4\text{N}_2\text{O}$) RNs, which are shown in Table 3.2. Unlike cMDS and LMDS, which exhibited severe distortion at low dimension for medium and large RNs, mMDS produced embeddings with good rank-based and global distance preservation metrics even at $d = 3$. Rank correlation was consistently higher than for the linear methods: at $d = 10$, mMDS reached Kendall $\tau = 0.85 - 0.9$ for all three representative RNs, exceeding cMDS by 0.04 for $\text{CH}_{19}\text{NO}_9$ and by 0.12 and 0.14 for $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ and $\text{C}_3\text{H}_4\text{N}_2\text{O}$, respectively. This gap widened with network size, as cMDS rank correlation degraded on the larger RNs while mMDS held

Table 3.2: Embedding quality metrics for metric multidimensional scaling (mMDS) and the distance-preserving encoder on selected test networks with karc (W_k) edge weights with $d = 3$ and $d = 10$ embedding dimensions. See Table 3.1 for notation.

			$\text{CH}_{19}\text{NO}_9$		$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$		$\text{C}_3\text{H}_4\text{N}_2\text{O}$	
			$d = 3$	$d = 10$	$d = 3$	$d = 10$	$d = 3$	$d = 10$
W_k	mMDS	τ	0.835	0.902	0.699	0.852	0.746	0.851
		R^2	0.924	0.973	0.736	0.955	0.832	0.959
		σ	0.113	0.068	0.198	0.082	0.179	0.089
	Encoder	τ	0.832	0.894	0.696	0.850	0.750	0.855
		R^2	0.920	0.968	0.737	0.950	0.839	0.960
		σ	0.116	0.073	0.197	0.086	0.175	0.087

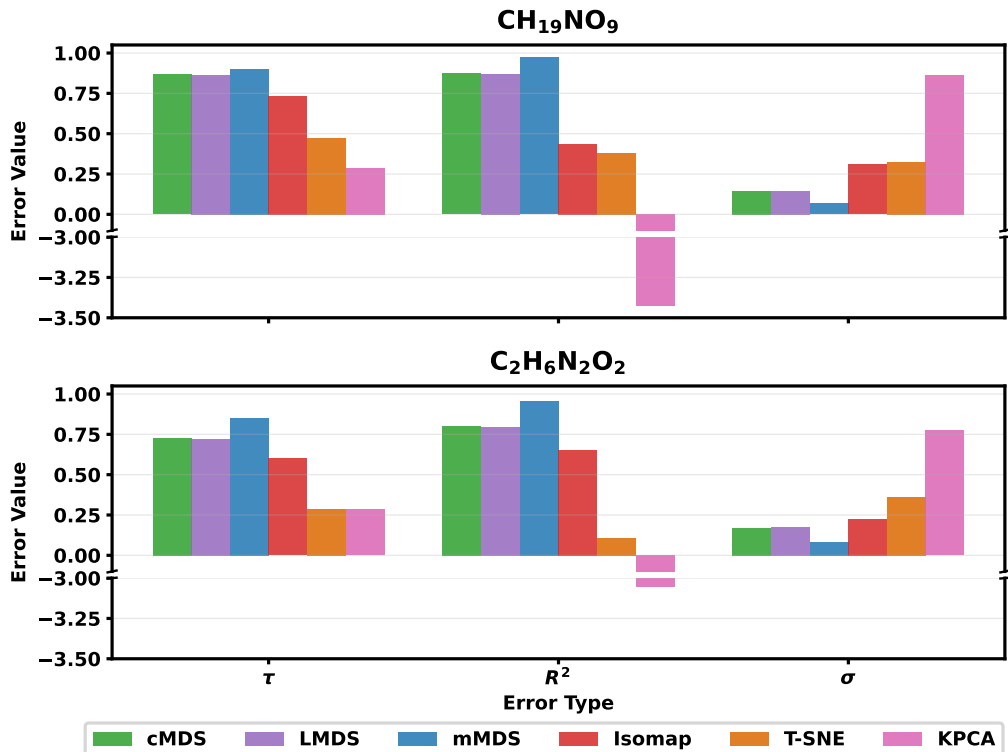


Figure 3.2: Comparison of embedding quality metrics for linear and conventional nonlinear embedding methods (cMDS, LMDS, mMDS, Isomap, t-SNE, and KPCA) for the $\text{CH}_{19}\text{NO}_9$ and $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ reaction networks (RNs) at $d = 10$ embedding dimensions. Kendall τ , coefficient of determination R^2 , and normalized stress σ are reported. The results for the best-performing hyperparameter set (based on R^2) are given for each method. See text for hyperparameter details.

nearly constant. Global distance preservation showed the same pattern. For the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN at $d = 3$, $R^2 = 0.736$ and $\sigma = 0.198$, compared to negative R^2 values and $\sigma > 0.4$ for cMDS and LMDS. For the same network at $d = 10$, we obtained $R^2 = 0.955$ and $\sigma = 0.082$. Similar trends held for the other test RNs. Results for all test RNs with mMDS are given in Appendix B, Section B.6.1.

The comparison of linear and nonlinear methods (cMDS, LMDS, mMDS, Isomap, t-SNE, and KPCA) is presented in Figure 3.2, which summarizes Kendall τ , coefficient of determination R^2 , and normalized stress σ for the $\text{CH}_{19}\text{NO}_9$ and $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RNs at $d = 10$. Given the substantial global distortion observed for Isomap, t-SNE, and KPCA on small and medium RNs, these methods were not tested on larger RNs. Among the nonlinear approaches, only

mMDS consistently improved both global distance preservation and rank correlation relative to cMDS and LMDS. Isomap embeddings with $N_n = 5$ nearest neighbors for the $\text{CH}_{19}\text{NO}_9$ RN and $N_n = 30$ for the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN yielded moderate rank preservation but poor global distance fidelity. For the $\text{CH}_{19}\text{NO}_9$ RN at $d = 10$, we obtained $\tau = 0.733$ but $R^2 = 0.435$ and $\sigma = 0.308$. For t-SNE, the best-performing hyperparameter choices were learning rate $L = 200$, maximum iterations $M = 1000$, perplexity $P = 150$, and early exaggeration factor $T = 24$ for the $\text{CH}_{19}\text{NO}_9$ RN and $L = 200$, $M = 5000$, $P = 75$, and $T = 12$ for the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN. However, even with the optimal hyperparameter choices, R^2 did not exceed 0.38. These results showed that t-SNE is unsuitable for applications requiring faithful preservation of global shortest-path distances. For KPCA [151], several kernels were evaluated by applying each to the Gram matrix $\mathbf{B}$ from Equation 3.2. No kernel produced faithful global distance preservation. The sigmoid kernel ($c = 4$ and $\gamma = 0.001$) gave the highest $R^2 = -3.425$ ($\tau = 0.286$ for $\text{CH}_{19}\text{NO}_9$ at $d = 10$); a linear kernel improved rank preservation ($\tau = 0.776$ for $\text{CH}_{19}\text{NO}_9$) but left R^2 numerically unstable ($R^2 \approx -6 \times 10^5$). Complete results for all small and medium test RNs with Isomap, t-SNE, and KPCA are given in Appendix B, Sections B.7, B.8, and B.9, respectively.

Although mMDS provided the best distance preservation among the nonlinear methods tested, its computational cost is substantially higher than both cMDS and LMDS. Like cMDS, mMDS requires the full $n \times n$ distance matrix. While mMDS avoids the $\mathcal{O}(n^3)$ eigendecomposition of cMDS, the iterative SMACOF optimization [164] used by scikit-learn [177] for mMDS requires $\mathcal{O}(n^2)$, for an effective $\mathcal{O}(n^2 N_i)$ over N_i iterations.

In practice, cMDS is often faster on large datasets because eigendecomposition benefits from highly optimized and parallelized linear algebra routines, whereas the iterative SMACOF algorithm is memory-bandwidth bound and may require hundreds of iterations to converge.

3.3.3 Neural Network-Based Encoder Embedding

The distance-preserving encoder was evaluated as a scalable alternative to mMDS. A feed-forward network refines a pre-embedding into a nonlinear embedding by minimizing the normalized stress σ (Equation 3.3) directly. Two initializations were compared: random initialization and LMDS pre-embedding. Figure 3.3 shows the normalized stress σ as a function of training epoch for the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN with W_k edge weights, with only the training subset shown as the convergence for training, validation, and test subsets was indistinguishable. LMDS initialization converged faster and to lower final stress, and was used for all reported embeddings.

Embedding quality for the encoder is summarized alongside mMDS in Table 3.2 for representative small ($\text{CH}_{19}\text{NO}_9$), medium ($\text{C}_2\text{H}_6\text{N}_2\text{O}_2$), and large ($\text{C}_3\text{H}_4\text{N}_2\text{O}$) RNs at $d = 3$ and $d = 10$ with W_k edge weights. The encoder achieved embedding quality comparable to mMDS across all test RNs. At $d = 10$, R^2 ranged from 0.950–0.968 and normalized

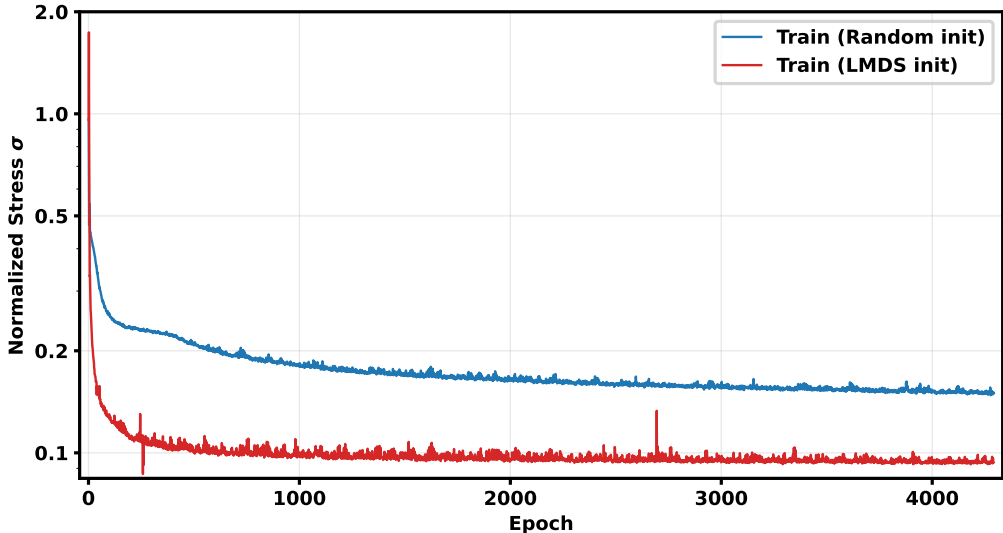


Figure 3.3: Normalized stress σ convergence for the training subsets during encoder training for the $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$ RN (encoder fraction $p_n = 0.3$) with random initialization and LMDS pre-embedding. W_k edge weights, starting dimension $d_s = 30$, and output dimension $d = 10$ were used.

stress from $\sigma = 0.073\text{--}0.087$, with $\tau > 0.84$ in all cases, compared to $R^2 = 0.955\text{--}0.973$ and $\sigma = 0.068\text{--}0.089$ for mMDS. At $d = 10$, the encoder and mMDS performed nearly identically across the test RNs: the encoder reached $\tau = 0.84 - 0.90$, $R^2 = 0.92 - 0.97$, and $\sigma = 0.06 - 0.11$, compared with $\tau = 0.84 - 0.90$, $R^2 = 0.94 - 0.97$, and $\sigma = 0.06 - 0.09$ for mMDS. The two methods differed by less than 0.01 in σ and R^2 on most networks, with neither consistently outperforming the other. The encoder was additionally evaluated with W_c edge weights, for which the nonlinear methods were not run; here it outperformed the linear methods ($\text{C}_2\text{H}_6\text{N}_2\text{O}_2$, $d = 10$: $\tau = 0.803$, $R^2 = 0.877$, $\sigma = 0.095$, compared with $\tau = 0.758$, $R^2 = 0.743$, $\sigma = 0.137$ for both cMDS and LMDS). Like mMDS, the encoder avoided the degradation in metrics observed for the linear methods at low embedding dimensions. At $d = 3$, cMDS and LMDS gave severely distorted embeddings for the $\text{C}_3\text{H}_4\text{N}_2\text{O}$ RN ($R^2 = -1.505$ and $R^2 = -1.507$, respectively), whereas the encoder reached $R^2 = 0.839$, which the linear projection can only represent at substantially higher dimension. Encoder results for all test RNs with W_k and W_c edge weights are given in Appendix B, Section B.10.1.

The encoder introduces additional cost over LMDS due to neural network training ($\mathcal{O}(n_n^2 d_s t)$ time over t epochs, where $n_n = p_n n$ is the size of the encoder subset). It remains of the same overall computational order since the LMDS initialization stage dominates for large networks. Unlike mMDS, the encoder avoids materializing the full $n \times n$ distance matrix. Each mini-batch forward pass operates only on nodes present in that batch, with memory dominated by the node feature matrix ($\mathcal{O}(n d_s)$) and the per-batch pairwise distance submatrix ($\mathcal{O}(n_b^2)$, where n_b is the batch size). The encoder thus achieved embedding quality comparable to mMDS at a computational cost closer to LMDS, making it practical for large RNs.

3.3.4 Optimal Embedding Dimension

The optimal embedding dimension d_e for cMDS was determined from the minimum of stress curves, while the corresponding values for mMDS and the distance-preserving encoder were computed using the derivative-based inflection-point criterion described in Section 3.2.2. Table 3.3 shows a comparison of the optimal embedding dimension values for W_c and W_k edge weights. For reference, the dimension of the underlying potential energy surface d_{PES} and the fractal dimensions d_f of the RNs from compact box-burning (CBB) calculations [1, 23, 81] are also included. The CBB algorithm covers the network with boxes of increasing size and extracts d_f from the scaling of the box count. The procedure is the same for W_c and W_k networks: the box-size parameter ℓ_B takes integer values from 1 to the network diameter, and two nodes share a box if their shortest-path distance does not exceed ℓ_B . For RNs with W_k edge weights, weighted shortest-path distances and the weighted diameter were used.

Table 3.3: Number of network nodes n , potential energy surface (PES) dimension d_{PES} , optimal embedding dimension d_e for cMDS, mMDS, and the distance-preserving encoder, and fractal dimension d_f for test RNs with W_c and W_k edge weights. Optimal embedding dimensions were obtained from a sigmoidal fit to the stress-dimension curve and are reported as integers, and mMDS was evaluated only with W_k edge weights.

Network	n	d_{PES}	W_c			W_k			
			d_e	d_e	d_f	d_e	d_e	d_e	d_f
			cMDS	Encoder		cMDS	mMDS	Encoder	
$\text{C}_2\text{H}_5\text{NO}_2$	3201	24	7	7	2.69	13	8	6	2.22
$\text{CH}_{19}\text{NO}_9$	4131	84	5	6	2.65	6	4	5	2.17
$\text{C}_2\text{H}_{10}\text{O}_5$	4280	45	7	15*	2.75	12	7	6	2.28
$\text{C}_3\text{H}_6\text{O}_3$	16,553	30	9	9	3.17	22	8	9	2.61
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	17,006	30	8	10	3.27	18	10	9	2.68
$\text{C}_2\text{H}_{18}\text{O}_9$	17,252	81	7	8	3.09	10	7	7	2.46
$\text{C}_3\text{H}_8\text{O}_4$	53,493	39	10	10	3.56	26	10	9	2.95
$\text{C}_3\text{H}_4\text{N}_2\text{O}$	56,464	24	11	11	3.56	24	10	8	2.71
$\text{C}_2\text{H}_8\text{N}_2\text{O}_3$	58,791	39	9	11	3.63	22	10	9	2.95
$\text{C}_3\text{H}_5\text{NO}_2$	59,742	27	10	10	3.55	29	9	9	2.82

* Sigmoidal fit outlier.

For cMDS with W_c edge weights, d_e ranged from 5 to 11 across the test RNs, whereas the minimal stress was $\sigma \approx 0.115$ – 0.126 (Appendix B, Section B.4.2). With W_k edge weights d_e was roughly twice as large, ranging from 6 to 29, while the minimal stress remained comparable ($\sigma \approx 0.108$ – 0.131). The principal effect of using W_k edge weights was therefore not larger irreducible error but higher optimal embedding dimension. The fractal dimension d_f showed no comparable sensitivity and remained in a similar range under both schemes, $d_f = 2.65$ – 3.63 for W_c and 2.17 – 2.95 for W_k .

For mMDS with W_k edge weights, the optimal embedding dimension ranged from 4 to 10 across the test RNs, substantially below the values for cMDS. The encoder required comparable dimensions, with d_e from 5 to 9. For the $C_3H_6O_3$ RN, for example, $d_e = 22$ for cMDS, compared with 8 for mMDS and 9 for the encoder. At their optimal dimension, both nonlinear methods reached lower residual stress than cMDS: $\sigma \approx 0.079$ – 0.099 for mMDS and 0.089 – 0.107 for the encoder, compared with 0.108 – 0.131 for cMDS. Residual stress showed no systematic increase with network size for either method, indicating that embedding distortion at d_e remains bounded as RNs grow. This dimensional efficiency arises because mMDS and the encoder minimize stress directly, rather than recovering coordinates by eigendecomposition as in cMDS. Direct minimization represents the non-uniform W_k metric in fewer dimensions.

d_e exceeds the fractal dimension d_f by a factor of about 3 across the test RNs (e.g. $C_2H_8N_2O_3$: $d_f = 2.95$ compared to $d_e \approx 9$ using the encoder). The difference reflects the distinct aspects of RN dimensionality that these definitions address. d_f is determined by network scaling under coarsening, while d_e depends on the structure of the shortest-path metric. Both quantities are much smaller than the dimensionality of the underlying PES $d_{PES} = 3N_a - 6$, where N_a is the number of atoms. In the test RNs, d_{PES} ranges from 24 to 84. Embedding the discrete RN therefore reduces the dimensionality of the reactive subspace relative to the full PES by a factor of 3–4 for most RNs, increasing to 8–17 for the hydrogen-rich $C_2H_{10}O_5$,

C₂H₁₈O₉, and CH₁₉NO₉ networks, which correspond to high-dimensional PESs but contain relatively few nodes.

3.3.5 Structural Origins of Embeddability

To determine whether the near-Euclidean structure of RNs is generic to sparse graphs or specific to chemical networks, randomization experiments were performed on the test RNs. In each case, the distance-preserving encoder was used to embed the networks at $d = 3$. Four conditions were compared: the original RN with W_k edge weights, topology-rewired with symmetric W_k edge weights preserved, topology preserved with edge weights independently resampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$ (truncated to positive values), and topology preserved with edge-weight symmetry broken by assigning forward and reverse weights independently. All randomization experiments were repeated across three independent runs. Results for the CH₁₉NO₉ RN are shown in Figure 3.4, and complete results for all test RNs are given in Appendix B, Section B.11.

The original CH₁₉NO₉ RN with W_k edge weights achieved $\tau = 0.832$, $R^2 = 0.920$, and $\sigma = 0.116$ at $d = 3$ (Figure 3.4(a)). Randomly rewiring all edges while preserving the symmetric edge-weight distribution reduced embedding quality to $R^2 = -20 \pm 23$ and $\sigma = 3.4 \pm 1.9$, while $\tau = 0.952 \pm 0.002$ remained high (Figure 3.4(b)). The high τ obscures the loss of structural information in this case. The small-world collapse of shortest-path distances under random rewiring [85, 86] compresses all pairwise distances into a narrow range, which preserves their rank order while destroying the global distance structure. With topology preserved but edge weights replaced by values independently resampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$, embedding quality was lower than for the original RN, with $\tau = 0.721 \pm 0.005$, $R^2 = 0.662 \pm 0.019$, and $\sigma = 0.171 \pm 0.004$ (Figure 3.4(c)). The spread of embedded versus network shortest-path distances was substantially wider than in the original RN. This indicates that chemically derived

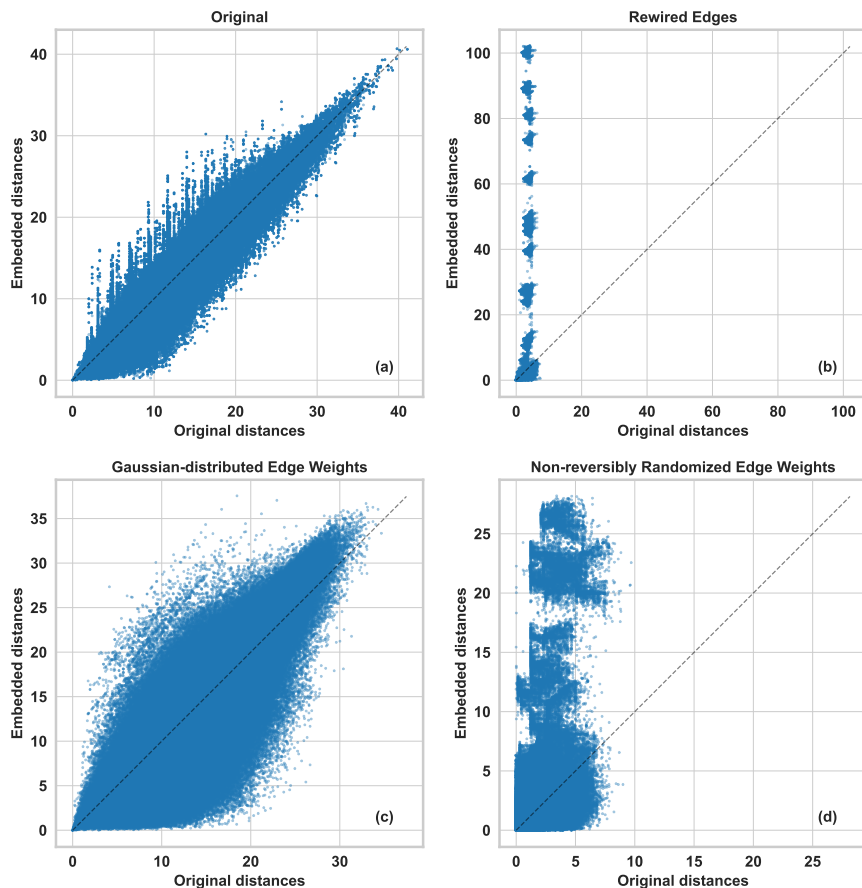


Figure 3.4: Embedding distortion for the $\text{CH}_{19}\text{NO}_9$ reaction network under four randomization conditions: (a) original W_k edge weights; (b) topology rewired, symmetric W_k edge weights preserved; (c) symmetric edge weights resampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$; (d) W_k edge weights randomized without symmetry preservation. Embedding-space distances from distance-preserving encoder embeddings with starting dimension $d_s = 30$, output dimension $d = 3$, $p_l = 0.3$, and $p_n = 0.3$ are plotted against network shortest-path distances.

W_k edge weights encode geometric structure beyond what a generic weight distribution of comparable scale provides. Breaking edge-weight symmetry by assigning forward and reverse weights independently drastically degraded rank preservation, with τ falling to 0.419 ± 0.020 (Figure 3.4(d)). Asymmetric weights violate the detailed-balance condition implicit in the reversible reaction rule set [24]. This makes the shortest-path metric asymmetric and rules out low-distortion Euclidean embedding [178]. Together, these experiments show that the embeddability of RNs in Euclidean spaces is driven by the chemically constrained topology and edge-weight symmetry.

3.3.6 Pathway Discovery via Embedding

Encoder embeddings of RNs using W_k edge weights were evaluated for their ability to support reaction pathway discovery. Shortest-path searches in the embedding Euclidean space were compared against chemically feasible pathways identified in the original RNs. We considered five benchmark reactions including tautomerization of vinyl alcohol (C_2H_4O RN), HCN trimerization ($C_3H_3N_3$ RN), esterification of acetic acid with methanol ($C_3H_8O_3$ RN), aldol addition of acetaldehyde and formaldehyde ($C_3H_6O_3$ RN), and glycolaldehyde dimerization by aldol addition ($C_4H_8O_4$ RN). Embeddings used the distance-preserving encoder with $d_s = 30$ starting dimension, $d = 10$ output dimension, $p_l = 0.3$ landmark fraction and $p_n = 0.3$ encoder fraction ($p_l = 0.005$ and $p_n = 0.005$ for $C_4H_8O_4$ RN).

For a path π in the RN, the network cost $C(\pi)$ is the sum of edge weights along the path and the corresponding embedding cost $\hat{C}(\pi)$ is the sum of Euclidean distances between consecutive embedded nodes,

$$C(\pi) = \sum_{(i,j) \in \pi} d_{ij}^k, \quad \hat{C}(\pi) = \sum_{(i,j) \in \pi} \hat{d}_{ij}, \quad (3.4)$$

where d_{ij}^k is the weight of the network edge between nodes K_i and K_j (Equation 3.1) and $\hat{d}_{ij}$ is the Euclidean distance between the images of K_i and K_j under the embedding. In most cases, multiple reaction pathways shared the same total path cost, differing only in the order of zero-weight steps such as polarization and depolarization. We represented each group of cost-degenerate paths by a single pathway, with multiplicity m giving the group size.

Enol tautomerization (ET). The tautomerization of vinyl alcohol to acetaldehyde is a prototypical enol-keto tautomerization. The direct gas-phase process is described as a symmetry-constrained 1,3-H shift [179, 180], whereas tautomerization in water proceeds by acid- or base-catalyzed proton transfer [181–183]. In the C_2H_4O RN, one feasible pathway

(ET-0) was identified with a length of 4 steps and $m = 5$ (Figure 3.5). The key steps of the ET-0 mechanism consist of the deprotonation of the alcohol group and reprotonation at the α -carbon atom. The embedding recovers this pathway as its lowest-cost path, with embedding cost $\hat{C} = 4.35$ compared to $C = 3.75$ in the original RN.

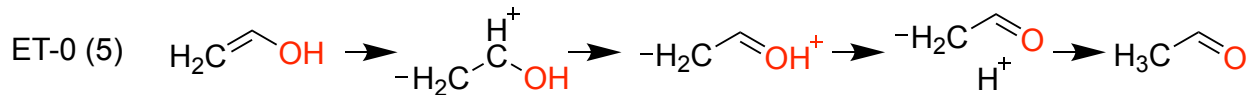


Figure 3.5: Feasible reaction pathway for enol tautomerization (ET). Path multiplicity is given in parentheses.

HCN trimerization (HT). The trimerization of HCN to form *s*-triazine is formally a cycloaddition of three nitrile groups to form the 1,3,5-triazine ring [184–186]. In the $C_3H_3N_3$ RN, one feasible pathway (HT-0) was identified with a length of 6 steps and $m = 9$ (Figure 3.6). The mechanism consists of three sequential C–N bond-forming steps. The first two assemble an open-chain trimeric C/N framework, and the final C–N bond formation results in ring closure and aromatization to *s*-triazine. The embedding recovers this pathway as its lowest-cost path, with $\hat{C} = 6.07$ compared to $C = 5.65$.

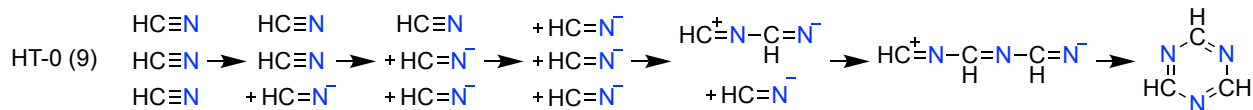


Figure 3.6: Feasible reaction pathway for HCN trimerization (HT). Path multiplicity is given in parentheses.

Esterification (ES). The esterification of acetic acid with methanol to form methyl acetate can proceed by one of several theoretically known reaction mechanisms [187, 188]. Acid-catalyzed esterification typically follows the $A_{AC}2$ mechanism [187]. In this mechanism, the carbonyl oxygen is first protonated, activating the acyl group toward nucleophilic attack by methanol and forming the tetrahedral intermediate. The latter undergoes proton transfer and elimination of water to yield methyl acetate. Alternative $A_{AC}1$ and $A_{AL}1$

pathways are possible under harsh conditions or for specific substrates [187].

In the $C_3H_8O_3$ RN, six feasible A_{AC2} -like pathways were identified (ES-0 through ES-5), all with a length of 6 steps and multiplicities of 1, 2, 2, 3, 4, and 3, respectively (Figure 3.7). All six pathways share the same key steps: C–O bond formation, O–H deprotonation, O–H protonation, and C–O bond cleavage, connected by polarization and depolarization steps. They differ in the ordering of these transformations. The lowest-cost reference pathway ES-0 involves a nucleophilic attack by a neutral methanol molecule, followed by elimination of a water molecule. The remaining pathways include elimination of hydroxide anion (ES-1), attack by the methoxide anion (ES-2 and ES-3, differing in the ordering of deprotonation relative to C–O bond formation), or both (ES-4 and ES-5, differing in the ordering of methoxide formation and hydroxide elimination). The embedding recovers four of the six

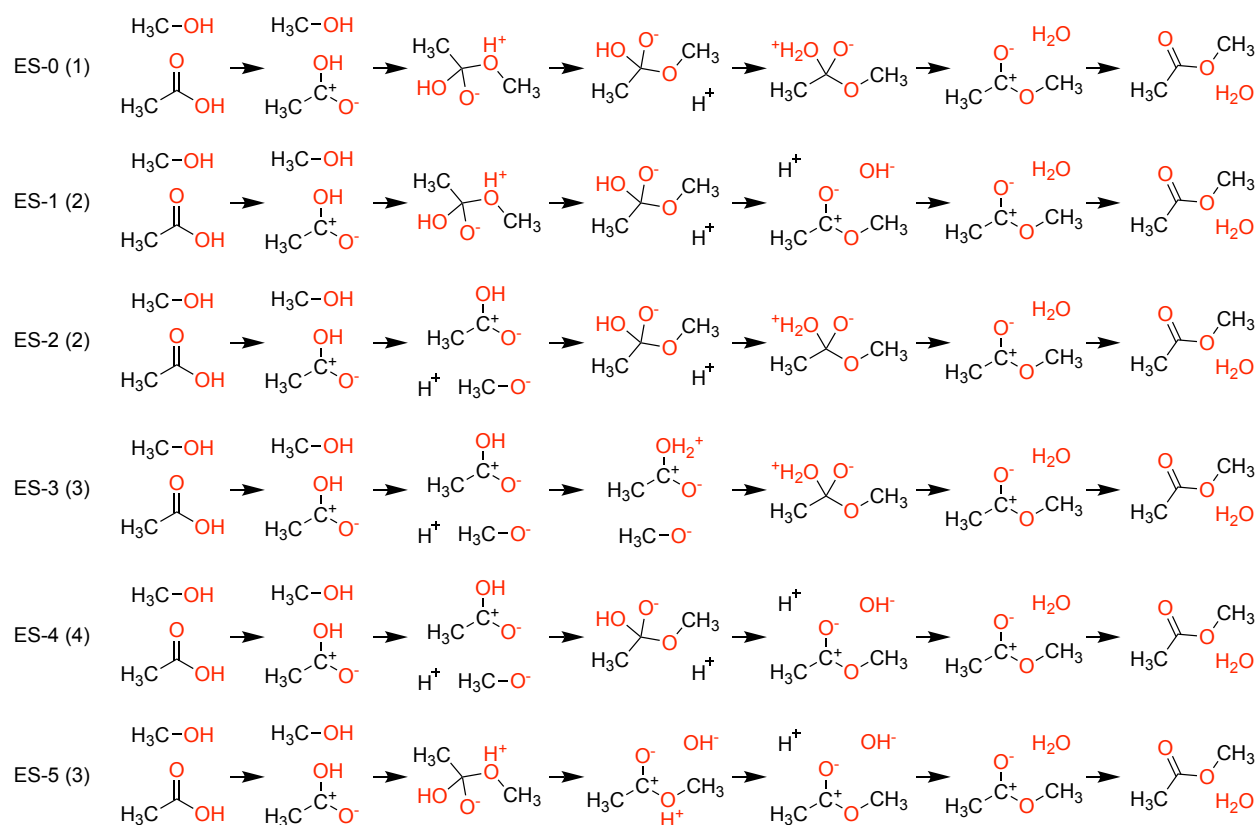


Figure 3.7: Feasible reaction pathway for esterification (ES). Path multiplicity is given in parentheses.

reference pathways among its six lowest-cost paths, with ES-0 ranked first ($\hat{C} = 10.53$, $C = 4.21$). The four recovered pathways preserve their relative ordering in the embedding space. The two missed pathways rank seventh and tenth in the embedding space ordering despite ranking second and fifth in the original RN, reflecting moderate reordering among near-degenerate mechanisms.

C₃ aldol addition (A3). The aldol addition of glycolaldehyde and formaldehyde to glyceraldehyde (C₃ aldol addition) is a central step of the formose reaction [156, 189, 190]. The base-catalyzed aldol mechanism proceeds via deprotonation of the α -carbon of the donor carbonyl compound to form an enolate. The C–C bond formation between the enolate intermediate and the acceptor carbonyl group and the protonation of the resulting oxyanion intermediate generate the aldol product [188, 191, 192]. Under acid catalysis, the enol tautomer rather than the enolate acts as the nucleophile [188, 192]. Two feasible pathways (A3-0 and A3-1) were identified in the C₃H₆O₃ RN with 4 steps and multiplicities of 2 and 3, respectively (Figure 3.8). Both pathways share the key steps of C–H deprotonation of glycolaldehyde, C–C bond formation, and O–H protonation, connected by polarization and depolarization steps. However, they differ in the relative ordering of C–C bond formation and protonation. The lowest-cost reference pathway A3-0 ($\hat{C} = 6.52$, $C = 4.57$) does C–C bond formation before final protonation, whereas in A3-1 ($\hat{C} = 6.18$, $C = 4.90$) the protonation of the formaldehyde oxygen happens before C–C bond formation. The embedding recovers

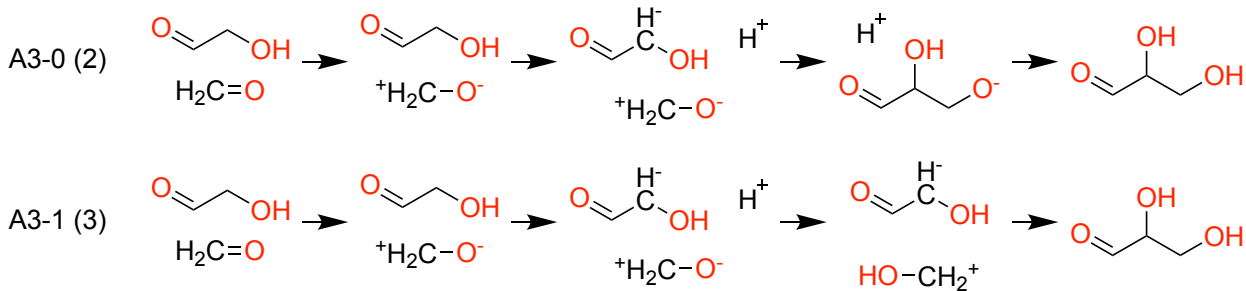


Figure 3.8: Feasible reaction pathway for C₃ aldol addition (A3). Path multiplicities are given in parentheses.

both pathways as its two lowest-cost paths, but in reverse order relative to the original RN.

C₄ aldol addition (A4). The glycolaldehyde dimerization to form a tetrose sugar (C₄ aldol addition) follows the same aldol mechanism as the C₃ aldol addition described above. However, its C₄H₈O₄ RN is richer due to the presence of additional degrees of freedom. The reverse of this reaction concludes the autocatalytic Breslow cycle of the formose reaction [156, 189, 190]. Two feasible length-4 pathways (A4-0 and A4-1) with multiplicities 2 and 3 were identified in the C₄H₈O₄ RN (Figure 3.9). They follow the same step pattern as A3-0 and A3-1: A4-0 ($\hat{C} = 4.45$, $C = 4.55$) performs C-C bond formation before protonation, while A4-1 ($\hat{C} = 5.59$, $C = 4.92$) reverses the order (here the protonation is of the acceptor glycolaldehyde’s carbonyl oxygen). By W_k cost, the length-4 pathways rank only fifth and twenty-third, behind longer pathways of length 6 and 8. The embedding ranks them as its top two. Two of the longer length-8 paths are provided in Appendix B, Section B.12.

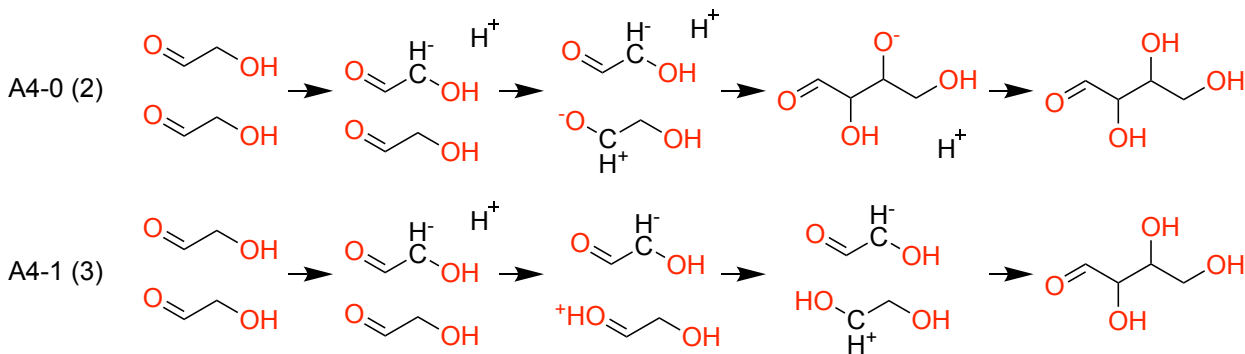


Figure 3.9: Feasible reaction pathway for C₄ aldol addition (A4). Path multiplicities are given in parentheses.

Overall, the encoder embeddings preserve pathway structure sufficiently well to recover either the exact lowest-cost pathway or a substantial fraction of chemically feasible pathways among the lowest-cost candidates. The changes in ordering between the original network and its embedding are primarily observed in networks containing many competing mechanisms of comparable cost. For the C₄ aldol, cumulative Euclidean distance in the embedding penalizes

longer pathways more strongly than the W_k edge weights do, which incidentally compensates for a limitation of the karc heuristic in balancing path length against intermediate energy demand.

3.4 Discussion

3.4.1 Structural Origins and Dimensionality of RN Embeddings

Stoichiometry-preserving reaction networks (RNs) admit low-dimensional Euclidean embeddings with small distortion. This is a consequence of their chemically constrained lattice-like topology and symmetric edge weights [1]. Regular lattices embed isometrically into Euclidean space by construction [178], and the lattice-like topology of RNs supports similar low-distortion embeddings. This contrasts with generic sparse graphs such as Erdős–Rényi random graphs [85] or Watts–Strogatz small-world networks [86], which do not generally admit such embeddings [178, 193]. Randomization experiments confirm that both ingredients are required: sparse topology alone is insufficient, and breaking edge-weight symmetry yields an asymmetric shortest-path metric that rules out low-distortion Euclidean embedding [178]. For W_c edge weights (treating all transformations as equally feasible), low-distortion Euclidean embedding was expected based on prior rewiring experiments characterizing RN topology as lattice-like [1]. The W_k case, which overlays a continuous, chemically derived weight distribution onto the topology and introduces non-uniform edge lengths, had no analogous prior characterization.

The optimal embedding dimension d_e and its relationship to the fractal dimension d_f and PES dimension d_{PES} are summarized in Table 3.3 across the test RNs for the cMDS, mMDS, and distance-preserving encoder methods. For cMDS with W_c edge weights, d_e ranged from 5–11. With W_k edge weights, d_e roughly doubled (ranging from 6–29), while the minimal

stress remained comparable across the two weightings. The non-uniform edge lengths of W_k increase the dimensionality required for faithful embedding without increasing the irreducible error. As RNs resemble regular lattices with a small amount of edge rewiring [1], the non-uniform edge lengths of W_k introduce additional distortions from regular lattice geometry. This makes linear methods require additional dimensions to represent the resulting shortest-path metric faithfully. This difference in d_e stands in contrast to the fractal dimension d_f , computed using the compact box-burning (CBB) algorithm [23, 81], which showed no comparable sensitivity to edge weighting. This confirms that the non-uniform edge lengths increase the dimensionality required to represent the shortest-path metric faithfully without changing the self-similar scaling of the network topology.

Nonlinear methods recover substantially lower d_e for W_k edge weights than cMDS. mMDS reaches $d_e \approx 4$ –11, compared to 6–29 for cMDS, and the encoder achieves comparable values. Both reach lower minimal stress at their respective d_e than cMDS, with no systematic increase with network size. This reduction arises because nonlinear methods directly minimize the stress objective instead of relying on eigendecomposition, allowing them to represent RNs with non-uniform edge weights more efficiently in fewer dimensions.

The empirical ratio $d_e/d_f \approx 3$ in the test RNs is comparable to results for smooth manifolds [194] and dynamical attractors [195, 196], where faithful embedding requires on the order of twice the intrinsic dimension. Although d_e represents the optimal dimension, useful embeddings can be obtained below d_e ; for applications where approximate distance preservation is sufficient, dimensions closer to d_f may be adequate. Moreover, the scaling of d_e with network size is consistent with Bourgain’s theorem [178, 193], which states that n -point metric spaces admit embeddings with $\mathcal{O}(\log n)$ dimension at bounded distortion. Figure 3.10 shows d_e and the residual stress σ as a function of network size n for the test RNs ($n \approx$

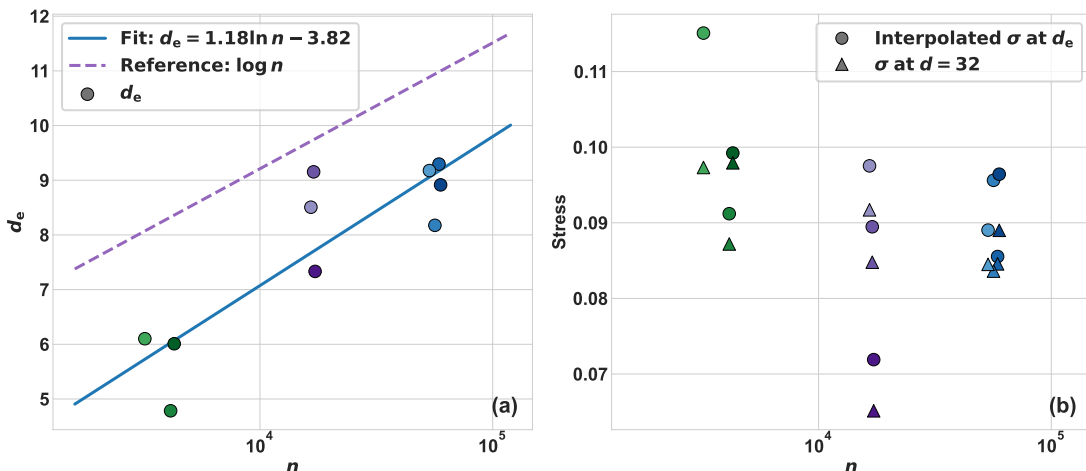


Figure 3.10: (a) Optimal embedding dimension d_e as a function of the number of nodes n for the test RNs, with linear fit. (b) Stress σ versus n for embeddings at d_e (circles) and $d = 32$ (triangles).

3200–58,800). A linear fit of d_e against $\ln n$ gives

$$d_e = (1.183 \pm 0.237) \ln n - 3.824 \pm 2.334. \quad (3.5)$$

The fit captures the slow logarithmic growth expected for lattice-like graphs over the two orders of magnitude in n available here. With only three size clusters, the data do not sharply discriminate this $\mathcal{O}(\log n)$ form from the $\mathcal{O}(\log^2 n)$ general worst-case behavior of Bourgain’s theorem. The stress at d_e is nearly constant in n and comparable to the stress at the much higher reference dimension $d = 32$, indicating that d_e captures essentially all the achievable distance preservation, with little gained by embedding at substantially higher dimension.

The low dimensionality of these embeddings stands in contrast to the dimensionality of the underlying PESs. The continuous PES of a molecular collection of N_a atoms requires $3N_a - 6$ dimensions, ranging from 24–84 across the test RNs. Stoichiometry-preserving RNs retain only the connectivity between molecular collections, discarding conformational degrees of freedom. The combinatorial constraints imposed by the reaction rules produce the lattice-

like topology that underlies their near-Euclidean character [1, 23]. For the $\text{C}_2\text{H}_8\text{N}_2\text{O}_3$ RN ($n = 58,791$ nodes, $N_a = 15$), for example, the underlying PES has $d_{\text{PES}} = 39$ continuous degrees of freedom per molecular collection, yet the discrete network is faithfully embedded in only $d_e = 9$ Euclidean dimensions. RNs thus serve as a low-dimensional scaffold of the reactive degrees of freedom, with conformational degrees of freedom representing a local expansion around each network node. This is complementary to approaches that reduce dimensionality by identifying collective variables from continuous PES data [197].

3.4.2 Embedding Methods: Performance and Scalability

Among MDS-based methods, cMDS provides a useful linear baseline but its $\mathcal{O}(n^3)$ eigendecomposition and $\mathcal{O}(n^2)$ memory requirement make it impractical for large networks. LMDS resolves the scalability issue at negligible loss in embedding quality; the dominant source of error is dimensional truncation, not the landmark approximation. Iterative Krylov subspace algorithms [198–200] are an alternative cost-reduction strategy that compute only the d leading eigenvectors of $\mathbf{B}$. However, both cMDS and LMDS share a fundamental limitation. The linear projection of both methods cannot efficiently represent the non-uniform shortest-path metric introduced by W_k edge weights in low dimensions. The non-monotonic behavior of stress with increasing dimension, previously reported for cMDS [159], further means that the optimal embedding dimension must be identified carefully. It cannot be determined simply by increasing dimensionality until stress plateaus.

mMDS overcomes the linear projection limitation by directly minimizing the normalized stress objective via iterative SMACOF optimization [148, 164]. It achieves substantially better distance preservation than cMDS and LMDS at low dimensions and requires fewer dimensions to reach comparable stress. Its iterative optimization incurs $\mathcal{O}(n^2 N_i)$ cost and requires the full $n \times n$ distance matrix, making it impractical for large networks. More

memory-efficient implementations could extend its practical range, but the quadratic memory requirement for the full distance matrix remains a fundamental bottleneck regardless of the linear algebra backend.

The distance-preserving encoder achieves mMDS-level embedding quality at a computational cost closer to LMDS by operating on a sampled subset of node pairs instead of the full $n \times n$ distance matrix. The encoder produces a faithful embedding at the default p_n . The regular lattice-like structure of RNs (Section 3.4.1) ensures that a random subsample of distances is representative of the full RN. The $\text{C}_4\text{H}_8\text{O}_4$ benchmark RN, trained with only $p_n = 0.005$ due to its size ($n = 1,194,872$), still achieved good embedding quality, confirming that the sampling fraction can be reduced further for large networks. For faster convergence and lower distortion, LMDS initialization is critical. By placing the encoder in a geometrically meaningful region of parameter space, neural network refinement recovers local metric structure efficiently instead of learning global geometry from scratch.

Isomap, t-SNE, and KPCA consistently fail to preserve global shortest-path distances regardless of hyperparameter configuration. These methods optimize objectives that prioritize local structure or cluster separation [201], and have no mechanism for bounding global distance distortion. The probabilistic objective of t-SNE and the kernel-based projection of KPCA cannot bound global distance distortion. Isomap’s sensitivity to neighborhood graph construction introduces systematic errors in long-range distances. These methods are therefore unsuitable for distance-preserving embedding of RNs.

3.4.3 Reaction Pathway Discovery

The pathway discovery experiments across five benchmark networks demonstrate that encoder embeddings preserve RN structure sufficiently well to recover chemically feasible pathways through shortest-path searches in the embedding space. For reactions with a sin-

gle clearly favorable mechanism, such as enol tautomerization and HCN trimerization, the lowest-cost reference pathway is ranked first in the embedding-space search. For reactions with multiple competing mechanisms of comparable cost, most reference pathways are recovered among the embedding’s lowest-cost paths, though the relative ordering may differ from that of the original RN. For esterification, four of the six reference pathways are recovered. The two missed pathways rank seventh and tenth in the embedded ordering, despite being ranked second and fifth in the original RN. For the C_3 and C_4 aldol additions, both reference pathways are recovered, though for the C_3 aldol addition the two pathways are ranked in reverse order relative to the original RN.

When multiple competing reaction pathways are predicted, examining a larger subset of candidate pathways is advisable to avoid overlooking feasible mechanisms. Even over a large candidate set, the search remains substantially more tractable than exhaustive PES exploration. RNs provide a discrete representation of PESs, and when embedded into Euclidean space they become easier to directly search. Imposing an upper bound on embedded pathway cost serves as a useful first-pass filter for narrowing the candidate set to more likely pathways. The narrowed set is then more manageable for higher accuracy quantum chemical characterization.

The pathway discovery results suggest that the accuracy of embedding-based pathway discovery is coupled to the quality of the edge-weighting scheme. Embeddings better identify feasible pathways when there is a clear cost gap between feasible and infeasible mechanisms, and when the original shortest-path metric ranks them correctly. In the C_4 aldol reaction, longer pathways have lower W_k costs than the mechanistically preferred length-4 pathways in the original network, yet the embedding ranks the length-4 pathways first and second. The embedding’s cumulative Euclidean cost penalizes longer pathways more strongly than the W_k edge weights do, incidentally compensating for karc’s imbalance between path length and intermediate energy. More broadly, when some feasible pathways are missed, as in esterifica-

tion, the karc costs do not clearly separate feasible from infeasible pathways. Co-developing chemically informed edge-weighting schemes with Euclidean embedding methods is therefore a natural direction for improving pathway discovery. Weights that better separate feasible from infeasible pathways would produce a more faithful metric and, in turn, more reliable embedding-space searches.

3.5 Conclusions

In this chapter, we showed that stoichiometry-preserving RNs can be embedded into low-dimensional Euclidean space with small distortion under both constant (W_c) and karc (W_k) edge weights, consistent with their previously established lattice-like structure [1]. The near-Euclidean character of RNs arises specifically from their chemically constrained topology and symmetric edge weights, as established by randomization experiments. Rewiring the topology severely degrades embeddability; generic symmetric weight distributions of comparable scale largely preserve it; breaking edge-weight symmetry collapses rank preservation. The optimal embedding dimension scales as $\mathcal{O}(\log n)$, consistent with Bourgain’s theorem [178, 193]. This corresponds to a 3–4 $\times$ reduction relative to the dimensionality of the underlying PES for most RNs, increasing to 8–17 $\times$ for hydrogen-rich networks. The optimal embedding dimension d_e is approximately three times the fractal dimension d_f , consistent with analogous results for smooth manifolds [194] and dynamical attractors [195, 196].

Among the embedding methods evaluated, only multidimensional scaling (MDS)-based approaches consistently preserve global shortest-path distances. Isomap, t-distributed stochastic neighbor embedding, and kernel principal component analysis are unsuitable for this application. Among MDS-based methods, classical MDS and landmark MDS are limited by linear projection. Metric MDS achieves better distance preservation but is impractical for large networks. The distance-preserving encoder achieves mMDS-level accuracy at a com-

putational cost comparable to LMDS by operating on a sampled subset of node pairs. It also scales to networks where mMDS is intractable. Pathway discovery experiments demonstrate that embedding-space shortest-path searches recover chemically feasible pathways. The principal limitation lies in the edge-weighting scheme instead of the embedding itself. The low-dimensional Euclidean representation is useful for accelerating sampling and search of RNs, and the embedding space enables direct interpolation between network nodes as a practical starting point for transition-state searches.

The current approach requires explicit RN construction, which is limited by computational cost and storage. However, the d_e scaling characterized here provides target dimensionality and error measures for subnetwork embeddings around a specific compound or reaction. Co-developing chemically informed edge-weighting schemes with embedding methods is a natural direction for further improvement. Weights that better separate feasible from infeasible pathways would produce a more faithful metric and more reliable embedding-space searches. More accurate energetic models beyond PM7/COSMO would improve both network distances and pathway rankings, albeit at increased computational cost.

The RN representation carries inherent limitations, including rule-based construction and the exclusion of stereoisomers and conformational degrees of freedom, as discussed in previous work [23, 24]. The partial ordering structure of stoichiometry-preserving RNs [23] suggests that embeddings trained on one RN may transfer to related networks of different stoichiometries. This provides a promising direction for constructing transferable embeddings across chemical space.

Chapter 4

Utilizing On-the-Fly Energy Calculations to Construct Bounded Reaction Networks

The motivation for this chapter comes from the goal of applying reaction network generation to larger, more chemically interesting systems. In earlier chapters, stoichiometry-preserving reaction networks were generated exhaustively for carbon-, hydrogen-, nitrogen-, and oxygen-containing (CHNO) systems of up to 57 atoms under the 40 normal polarity transformation rules of Chapter 2. However, extending this approach to systems with more atoms or a broader set of transformation rules is fundamentally limited by the combinatorial explosion. This chapter addresses that obstacle in two ways: by expanding the rule set to include radical reactions, enabling the network to capture open-shell chemistry, and by introducing on-the-fly energy calculations to bound network growth to chemically relevant regions of the potential energy surface (PES). The work presented in this chapter is in preparation for publication. Co-author Liliana Garcia, an undergraduate researcher, contributed to the analysis of reaction networks for feasible pathways and reactions.

Radical intermediates are ubiquitous in reactive chemistry, yet their incorporation into reaction network generation requires both an expanded rule set and a strategy for controlling the resulting growth in network size. Stoichiometry-preserving RNs provide a discrete representation of reactive chemical space that enables systematic exploration of chemically accessible regions of the PES. In this chapter, we extend the `colibri2` software to support radical chemistry through reaction rules that explicitly encode homolytic bond cleavage, radical recombination, and unpaired-electron tracking using extended SMARTS notation. To mitigate the resulting combinatorial growth in RN size arising from the expanded reaction rule set, we introduce a bounded-generation workflow. The workflow incorporates on-the-fly energy calculations during node generation to restrict network expansion according to two user-defined parameters, a maximum generation depth from the starting node (G_{max}) and a maximum permitted energy difference between connected nodes (ΔE_{max}). The rapid growth in network size with generation depth allows even modest bounds to produce dramatic reductions in network size, often by orders of magnitude, while preserving chemically relevant pathways. Full exhaustive RNs can contain hundreds of thousands to millions of nodes, whereas bounded networks targeting specific regions of chemical space may contain only hundreds to tens of thousands of nodes while still retaining kinetically feasible reaction pathways. Energy bounds preferentially exclude high-energy intermediates and transformations, thereby focusing exploration on low-energy regions of chemical space that are more likely to be chemically relevant. Bounded RNs are applied to benchmark reactions in atmospheric radical chemistry, where $\Delta E_{\text{max}} \approx 3\text{--}4$ eV provides an effective balance between network reduction and retention of chemically relevant pathways. The kinetic heuristic karc path-cost ordering correctly identifies the dominant channel in each benchmark system, consistent with established atmospheric mechanisms. These results demonstrate that bounded RN generation provides a computationally efficient strategy for exploring chemically relevant reactive pathways while avoiding exhaustive enumeration of the full chemical space.

4.1 Introduction

Exploration of chemical reaction space becomes increasingly difficult with larger and more complex systems because the number of accessible intermediates and reaction pathways increases combinatorially with system size [23, 32]. Computational methods for reaction discovery therefore face a fundamental tradeoff between breadth of chemical coverage and computational tractability. To address this challenge, a wide range of computational exploration strategies have been developed, including rule-based and template-driven reaction generation [25–29], heuristic and graph-based reaction-network approaches [22–24, 39], quantum-chemical transition-state and reaction-path searches [6, 7, 31, 33, 35, 202], molecular dynamics-based sampling methods such as the *ab initio* nanoreactor [45], hybrid cheminformatics–quantum chemistry frameworks [13, 30], and machine-learning-assisted reaction exploration methods [49, 53, 55, 203]. These approaches differ in how chemical states and transformations are represented, sampled, and prioritized during exploration. Despite these advances, exhaustive exploration rapidly becomes computationally prohibitive even for relatively small systems, making the identification of chemically relevant subspaces a central problem in reaction discovery.

Stoichiometry-preserving RNs [22–24] provide a framework for representing chemical reactivity in a discrete and exhaustive manner. In this representation, nodes correspond to collections of molecules with fixed atomic composition, while edges represent stoichiometry-preserving bond transformations. RNs can capture large regions of chemically accessible reaction space and recover competing mechanistic pathways within a single unified representation [22, 23]. This is achieved through exhaustive generation for a given stoichiometry and reaction-rule set, generating all nodes and edges permitted by the rules. The resulting network growth follows a pattern analogous to logistic population growth [7, 204, 205]. Early generations produce relatively few new nodes, after which a combinatorial expansion occurs as each node generates many descendants. Growth eventually slows and saturates once the

finite stoichiometric chemical space has been exhausted. This behavior follows from the lattice-like structure of RNs [1], which produces relatively regular local branching patterns and approximately exponential growth throughout the early and intermediate stages of network generation. Exhaustive generation also produces large numbers of radicals, ions, and high-energy intermediates that may be irrelevant to a target reaction. Consequently, network size increases rapidly with stoichiometric complexity, limiting exhaustive RN construction to relatively small systems [1, 23].

A natural strategy for controlling this growth is to restrict network generation to chemically relevant regions of reaction space. Existing computational exploration techniques employ a variety of prioritization strategies, including reactive-site selection [7, 31], transition-state and barrier-based filtering [6, 35], adaptive reaction-network exploration [44, 206], and molecular dynamics-based sampling of dynamically accessible regions of the potential energy surface [45, 207]. Prior work on stoichiometry-preserving RNs introduced chemically motivated edge-weighting schemes for identifying kinetically favorable paths within exhaustively generated networks [24], but did not restrict network growth during generation. As a result, exhaustive enumeration remains computationally expensive for larger stoichiometries. Although energetic pruning and kinetic steering strategies are well established in automated reaction discovery [32, 33, 44], their incorporation directly into stoichiometry-preserving RN generation has, to our knowledge, not been systematically explored. Such constraints could retain mechanistically relevant pathways while avoiding exhaustive enumeration of distant high-energy regions of reaction space.

Stoichiometry-preserving RNs have primarily been applied to closed-shell systems and have not previously been extended to radical reaction chemistry. Radical processes are central to many areas of chemical reactivity, including atmospheric oxidation, [17, 208] combustion and pyrolysis [13, 25], polymerization [209], electrochemistry [210], radiation chemistry [211], and biological redox chemistry [212]. Compared to closed-shell systems, radical chemistry of-

ten permits substantially larger numbers of accessible intermediates and competing reaction pathways because homolytic bond cleavage and recombination can rapidly generate highly reactive species with diverse subsequent reactivity [17, 208, 213]. Radical mechanisms also frequently involve short-lived intermediates, chain-propagation and chain-branching processes, and spin-dependent reactivity [208, 214–216]. Multiple competing unimolecular and bimolecular reaction channels further complicate systematic exploration [17, 208, 213, 214]. Within stoichiometry-preserving RNs, representing these processes requires extending reaction-rule sets beyond conventional electron-pair-transfer transformations to include homolytic bond cleavage, radical recombination, and explicit tracking of unpaired electrons while maintaining chemically meaningful network connectivity.

In this chapter, we extend the `colibri2` software [68] to support radical reaction chemistry through reaction rules that explicitly represent homolytic bond cleavage, radical recombination, and unpaired-electron tracking using extended SMARTS notation [77, 217]. Building on this extension, we introduce a bounded-generation workflow that incorporates on-the-fly semiempirical molecular energy calculations [154, 155]. Network expansion is constrained by two user-defined bounds: the maximum generation distance from the initial starting node, $G_{\max}$, and the maximum allowed molecular energy difference between connected nodes, $\Delta E_{\max}$. Together, these criteria restrict network growth to low-energy regions near the starting node while retaining pathways that are chemically relevant to a target reaction or reaction set. We apply this workflow to stoichiometry-preserving RNs relevant to atmospheric radical chemistry and conduct pathway discovery experiments for known atmospheric reactions.

This chapter is organized as follows. Section 4.2 describes the RN representation, the extension to radical transformations, and the bounded-generation workflow. Section 4.3 presents network scaling as a function of $G_{\max}$ and $\Delta E_{\max}$ and demonstrates pathway discovery for known reactions in atmospheric chemistry. Section 4.4 interprets the findings in terms of RN size reduction, the impact of radical reaction rules, and pathway discovery. Finally, the

conclusions are given in Section 4.5.

4.2 Methods

4.2.1 Reaction Network Representation

Stoichiometry-preserving RNs [22, 23] were constructed for carbon, hydrogen, and oxygen (CHO) stoichiometries relevant to atmospheric chemistry. In RNs, nodes represent collections of molecules with identical atomic composition that differ only in molecular connectivity. Edges correspond to elementary bond transformations that conserve stoichiometry. Networks were generated with the open-source `colibri2` software [68], which iteratively applies a predefined transformation rule set until no new nodes can be generated. For a fixed stoichiometry and rule set, this procedure is exhaustive, such that all molecules and reactive intermediates permitted by the rules appear as nodes in a RN. Consequently, the final RN is independent of the initial molecular collection.

The reaction rules describe heterolytic and homolytic bond breaking and formation (in heterolytic cleavage both bonding electrons go to one atom, producing ions, whereas in homolytic cleavage one electron goes to each atom, producing radicals), as well as fictitious polarization and depolarization transformations used to treat multiple-bond reactivity compositionally [24]. All transformations were ensured to be reversible, with each bond-breaking transformation paired to a corresponding bond-forming rule. The rules are encoded in SMARTS format [75], which can represent reactions expressible as electron-pair-transfer (“arrow-pushing”) steps [66, 67]. Radical transformations were included using the extended SMARTS notation implemented in `RDKit` [77, 217]. Unpaired electrons are encoded through trailing `|n:i,...|` annotations, where `n` specifies the number of unpaired electrons and `i` indexes the atom within the SMARTS pattern. Within `RDKit`, these annotations act as

pattern-matching and product-construction constraints by defining atom-centered unpaired-electron counts for molecular representation. The assigned radical counts do not necessarily correspond to physically realizable radical multiplicities. Rather, `RDKit` assigns formal atom-centered unpaired-electron counts based on valence and bond-counting rules, which are treated here as radical counts.

The rule set used in this chapter contains 43,526 SMARTS-encoded transformations, comprising 17,795 unimolecular transformations, 5855 bimolecular bond-forming transformations, 4969 cyclization transformations, and 14,907 ring-opening transformations. The rules include all charge and radical combinations for which the total absolute formal charge on either side does not exceed two, each side contains at most four radical electrons, and each elementary transformation changes the total radical-electron count and total formal charge by no more than two. As the radical notation is formal rather than physical, multiple radical electrons may be assigned to the same atom. However, transformations involving hydrogen atoms were restricted such that the unpaired electrons remained distributed across the two atoms participating in the transformation. Generated molecules in the RNs were restricted to a maximum net formal charge of two, a total absolute formal charge of eight, four radical electrons, and four rings per molecule. Further details on the allowed radical counts for specific atom types and additional rule-set details are provided in Appendix C, Section C.1.

Application of the reaction rules can generate structures with unphysical electronic configurations, including partially broken bonds and artificial charge-separated states. These structures were grouped into molecular equivalence classes (Q_M) differing by zero-energy transformations, defined as transformations that alter bond orders, formal charges, or radical localizations without changing molecular connectivity. A subset of 11,940 unimolecular rules from the complete reaction-rule set defines the equivalence zero-energy transformations. These include zero-energy transformations involving C–C, C–O, and O–O bonded pairs. Each equivalence class was assigned the minimum energy among members with maximal

bond order and minimal absolute formal charge. Molecular energies were calculated at the PM7/COSMO level of theory using MOPAC [154, 155], with the dielectric constant of water ($\epsilon = 78.4$). Each node represents a multiset of molecule formulas with integer stoichiometric coefficients c_m , and its energy is $E_K = \sum_m c_m \mu_m$, where μ_m is the representative energy of the equivalence class to which formula m belongs. Node energies were thus computed as sums of these representative molecular energies. Further details of the equivalence rules and of conformer generation and energy evaluation are provided in Appendix C, Section C.2 and C.3, respectively. Edge weights were assigned to reflect thermodynamic and kinetic feasibility. In this chapter, the kinetic heuristic (karc) criterion W_k was employed [22, 24], which approximates the rate-limiting activation barrier through Hammond’s postulate [71]. For a transformation from node K_i to node K_{i+1} with energies E_{K_i} and $E_{K_{i+1}}$, respectively, the edge weight is

$$w_{K_i \rightarrow K_{i+1}} = \left[(E_{K_{i+1}} - E_{K_i})^2 + \alpha^2 \left(\frac{1 - e^{-\kappa(E_{K_{i+1}} - E_{K_i})^2}}{1 + e^{-\kappa(E_{K_{i+1}} - E_{K_i})^2}} \right) \right]^{1/2}, \quad (4.1)$$

where $\alpha = 1$ eV and $\kappa = 5$ eV⁻², as determined previously for CHO RNs [24]. The karc weight $w_{K_i \rightarrow K_{i+1}}$ is a single-edge quantity. Pathway ranking uses the cumulative path cost $C = \sum_i w_{K_i \rightarrow K_{i+1}}$, the sum of karc weights along the full sequence of elementary steps connecting reactant to product. Lower C corresponds to a more kinetically favorable overall reaction sequence.

In this chapter, seven RNs composed of carbon, hydrogen, and oxygen atoms were studied across atmospherically relevant stoichiometries using bounded generation. These include the pure-oxygen systems O₃ and O₄H₂; the single-carbon systems CH₂O₂, CH₄O₃, and CH₆O₄; and the two-carbon systems C₂H₄O₃ and C₂H₆O₄, encompassing stoichiometries containing 3–12 atoms. Exhaustive generation RNs were created for four of these RNs, O₃, O₄H₂, CH₂O₂, and CH₄O₃. The O₃ RN targets ozone decomposition chemistry, and O₄H₂ extends this to hydrated ozone reactions. The CH₂O₂ RN covers Criegee intermediate chemistry.

The CH_4O_3 RN targets methylperoxy radical oxidation and methyl trioxide decomposition, and CH_6O_4 extends this to the hydrated system. The $\text{C}_2\text{H}_4\text{O}_3$ RN covers ethylene ozonolysis, and $\text{C}_2\text{H}_6\text{O}_4$ is examined at selected bounded-generation levels as a size comparison to assess the effect of water on bounded RN size. Within these RNs, nodes composed exclusively of neutral molecules define the neutral-node subset (N_N), representing the physically accessible portion of the network. Collapsing these nodes by molecular equivalence class yields the reduced neutral-node set Q_N .

All generated RNs were converted to undirected graphs and stored as compressed GraphML files [76]. Network manipulation and analysis were performed with the NetworKit [101] and NetworkX [157] libraries. All custom code developed beyond `colibri2` is available on GitHub [218].

4.2.2 Bounded Reaction Network Generation

Exhaustive RN generation quickly becomes computationally prohibitive for large stoichiometries and reaction-rule sets because iterative rule application causes a combinatorial explosion in the number of generated nodes and edges. Rather than exhaustively enumerating the full chemical space defined by a stoichiometry and rule set, we introduce a bounded-generation workflow in `colibri2` that restricts expansion to a localized region of chemical space. This enables exploration of regions of the underlying PES relevant to a target reaction or reaction set while avoiding the combinatorial growth associated with exhaustive enumeration. The workflow combines on-the-fly molecular energy calculations with two expansion criteria: a maximum generation depth from the starting node (G_{max}) and a maximum allowed molecular energy difference between connected nodes (ΔE_{max}). Together, these criteria enable expansion to be restricted to low-energy regions near the starting node N_s , substantially reducing network size relative to exhaustive enumeration. Consequently, the generated RN

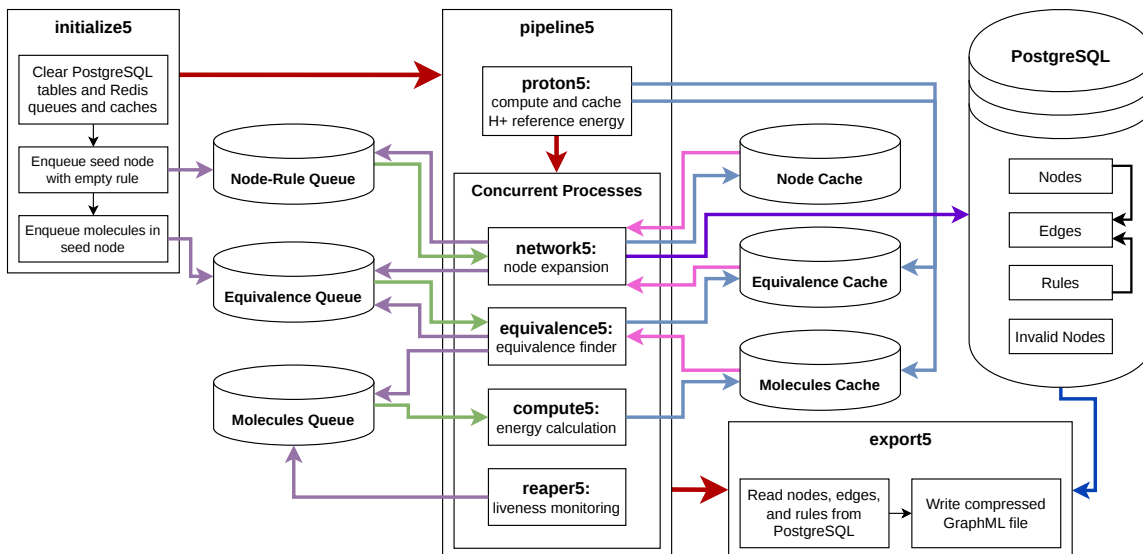


Figure 4.1: Workflow diagram of the bounded RN-generation pipeline implemented in `colibri2`, consisting of three sequential stages: `initialize5`, `pipeline5`, and `export5`. Within `pipeline5`, `proton5` runs to completion before four concurrent workers are launched: `network5`, `equivalence5`, `compute5`, and `reaper5`. Workers communicate through Redis queues and caches, while persistent data are stored in PostgreSQL. Details of each worker and the data flow between components are described in the text. Arrow colors indicate communication type: red, sequential stage execution; light purple, enqueue; green, dequeue; pink, cache read; blue, cache write; dark purple, database write; dark blue, database read; and black, table relationships within PostgreSQL.

depends on the initial starting node, as only nodes reachable within the specified bounds are generated. Additional implementation details are provided in Appendix C, Section C.4.

The workflow consists of three sequential stages (Figure 4.1). The `initialize5` stage prepares the PostgreSQL [73] database together with Redis [72] queues and caches and clears any existing data. The starting node at generation $g = 0$, paired with an empty rule, is seeded into the node-rule queue, while its constituent molecules are queued for equivalence processing. The `pipeline5` stage performs network generation and energy calculations. Within this stage, the `proton5` subprocess first computes and caches the H^+ reference energy from the energy difference between water and hydroxide. Four concurrent workers are then launched: `network5` for node expansion, `equivalence5` for equivalence-class processing, `compute5` for molecular energy evaluation, and `reaper5` for failure monitoring and retry management.

Additional **network5**, **equivalence5**, and **compute5** workers may be launched in parallel to accelerate generation. Once generation is complete, the **export5** stage reads the PostgreSQL database and writes the final RN as a compressed GraphML file.

Persistent data are stored in PostgreSQL, including nodes, invalid nodes, transformation rules, and optional metadata. Inter-worker communication is handled through Redis, which maintains queues for pending tasks and caches for completed calculations. These components decouple the workers and allow multiple instances of **network5**, **equivalence5**, and **compute5** to operate concurrently without direct coordination.

Most computational time is spent in three workers that form a closed demand-driven loop in which node expansion drives equivalence screening, which in turn drives molecular energy calculations. **network5** dequeues a node–rule pair and checks whether molecular energies are available for all constituent molecules. If energies are missing, the node is requeued and the molecules with missing energies are requeued in the equivalence queue. Once all molecular energies are available, the energy difference between the parent and child node is evaluated against ΔE_{max} . If the criterion is satisfied, the node and its connecting edge are stored in the database. Otherwise, the edge is discarded and the child node is not queued for further expansion, although it may still be stored or later expanded if reached through another valid edge. If the current generation g has not yet reached G_{max} , the node is expanded by applying the full reaction-rule set to generate new node–rule tasks, and each child node is assigned generation $g + 1$. **equivalence5** processes formula SMILES from the equivalence queue by applying the equivalence zero-energy transformations to enumerate all members of each equivalence class. An equivalence class groups all molecular representations that share identical atomic connectivity but differ in bond order, formal charge, or radical localization, analogous to resonance structures of the same molecule. The maximally bonded, lowest-charge representatives are identified, and their absolute molecular energies are retrieved from the molecular-energy cache when available. Representatives without cached energies are

added to the molecule queue for energy calculation. The minimum successful representative energy is then assigned as the class energy and cached. `compute5` processes molecules from the molecule queue, performs molecular energy calculations, and records energies together with success, failure, or timeout status in Redis. An active-task registry prevents duplicate energy calculations. Nodes containing molecules without successful energy calculations are assigned null energies.

The two expansion criteria, $G_{\max}$ and $\Delta E_{\max}$, may be applied independently or in combination. $G_{\max}$ controls the topological radius explored from the starting node, and $\Delta E_{\max}$ constrains thermodynamic accessibility along individual reaction steps. In the present chapter, both parameters were varied independently and in combination to examine their effect on network size and the discovery of chemically relevant pathways. The $\Delta E_{\max}$ bound is best understood as a broad exploratory cutoff rather than a strict kinetic threshold. Thermally accessible organic reaction barriers typically fall in the range of $\sim 0.1\text{--}2$ eV [219, 220], and semiempirical methods carry intrinsic energy uncertainties of comparable magnitude [154]. Accordingly, a cutoff of $\Delta E_{\max} = 3\text{--}4$ eV provides a reasonable and permissive threshold for network expansion. Unless otherwise noted, a value of $\Delta E_{\max} = 3$ eV is used throughout this chapter, with $\Delta E_{\max} = 4$ eV employed in select cases.

4.3 Results

4.3.1 Network Scaling with Bounded Generation

To characterize how maximum generation depth $G_{\max}$ and allowed molecular energy difference between connected nodes $\Delta E_{\max}$ jointly control RN size and composition, we applied bounded generation to a set of CHO stoichiometries spanning a range of molecular complexity: O_3 , O_4H_2 , CH_2O_2 , CH_4O_3 , CH_6O_4 , $\text{C}_2\text{H}_4\text{O}_3$, and $\text{C}_2\text{H}_6\text{O}_4$. For each system we

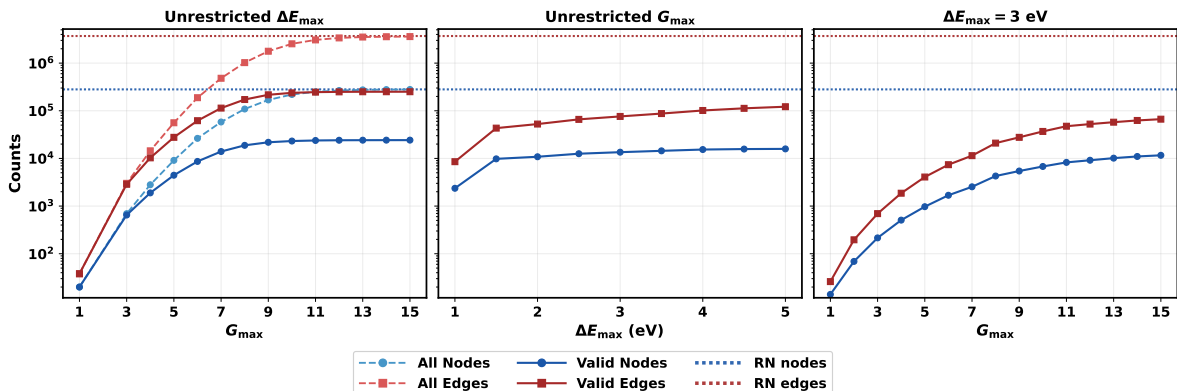


Figure 4.2: Scaling of nodes (n) and edges (e) with bounded generation in the CH₄O₃ RN ($N_s = \text{C000}$) as functions of (a) $G_{\max}$ with unrestricted $\Delta E_{\max}$, showing all generated and energy-validated nodes separately, (b) $\Delta E_{\max}$ with unrestricted $G_{\max}$, and (c) $G_{\max}$ at fixed $\Delta E_{\max} = 3$ eV. Dotted lines mark the exhaustive RN numbers.

tracked nodes (n), edges (e), the number of unique molecules (M), molecule equivalence classes (Q_M), neutral nodes (N_N), and neutral equivalence classes (Q_N) as functions of the two bounding parameters.

Figure 4.2 characterizes the scaling behavior of bounded generation for the CH₄O₃ RN starting from $N_s = \text{C000}$. The exhaustive CH₄O₃ RN contains $N = 281,005$ and $E = 3,702,400$, representing 58,821 unique molecules grouped into 55,776 equivalence classes by zero-energy transformations. However, many of these species are chemically inaccessible due to unphysical charge states or radical multiplicities. When energy validation is applied, only 3047 molecules are recoverable, reducing the network to $N = 23,527$ and $E = 239,332$. Network sizes and molecule count before and after energy validation are given for several exhaustive RNs in Appendix C, Section C.5.

Figure 4.2(a) shows the dependence of RN size on generation depth $G_{\max}$ with unrestricted $\Delta E_{\max}$. Both the total number of n and e generated for a specific $G_{\max}$ and energy-validated counts are shown. The separation between the two curves quantifies the fraction of nodes rejected during energy validation at each generation depth. Both sets of curves exhibit logistic-like growth characteristic of stoichiometry-preserving RNs because of their lattice-

like structure [1]. After an initial period of approximately exponential growth, the curves approach the exhaustive RN limits marked by the horizontal dotted lines at large $G_{\max}$. In the worst case, bounded generation recovers the exhaustive RN once $G_{\max}$ reaches the network diameter, which is the longest of the shortest paths in the network.

Unlike the strong generation-dependent growth observed with increasing $G_{\max}$, RN size varies more gradually with $\Delta E_{\max}$. This behavior is illustrated in Figure 4.2(b), which shows the dependence of RN size on $\Delta E_{\max}$ for unrestricted $G_{\max}$. For the CH_4O_3 RN with $N_s = \text{C000}$, increasing $\Delta E_{\max}$ from 2 to 3 eV increases n from 10,840 to 13,518 ($\approx 25\%$), while a further increase from 3 to 4 eV produces a smaller increase ($13,518 \rightarrow 15,360$, $\approx 14\%$). Even at $\Delta E_{\max} \approx 4$ eV, the network remains substantially smaller than the exhaustive RN limit of 23,527 nodes, which is reached only at $\Delta E_{\max} \approx 23$ eV.

Combining $G_{\max}$ and $\Delta E_{\max}$ provides much finer control over RN size (Figure 4.2(c)). For $N_s = \text{C000}$ with $\Delta E_{\max} = 3$ eV, the network exhibits growth behavior similar to that in Figure 4.2(a), but with substantially reduced node and edge counts. In this regime, $\Delta E_{\max}$ primarily determines the maximum accessible RN size, while $G_{\max}$ controls the rate at which that limit is approached. As RN growth is approximately sigmoidal with respect to $G_{\max}$, even modest reductions in generation depth produce substantial decreases in network size.

Figure 4.3 shows the dependence of bounded RN growth on the choice of starting node N_s . At fixed $\Delta E_{\max} = 3$ eV and $G_{\max} = 7$, $N_s = \text{C0[0].[OH]}$ yields $N = 2208$ and $E = 10,002$, compared with $N = 2538$ and $E = 11,428$ for $N_s = \text{C000}$ ($\approx 15\%$ fewer nodes). The difference is more pronounced at lower $G_{\max}$ (e.g. $N = 215$ vs $N = 117$ at $G_{\max} = 3$), reflecting the distinct local topologies around each starting node. As $G_{\max}$ increases, both networks expand toward the same accessible region of chemical space and the curves converge. The influence of N_s is therefore strongest for shallow, targeted explorations of local RN structure.

Table 4.1 summarizes bounded RN statistics at $\Delta E_{\max} = 3$ eV and $G_{\max} = 3$ and 7 across

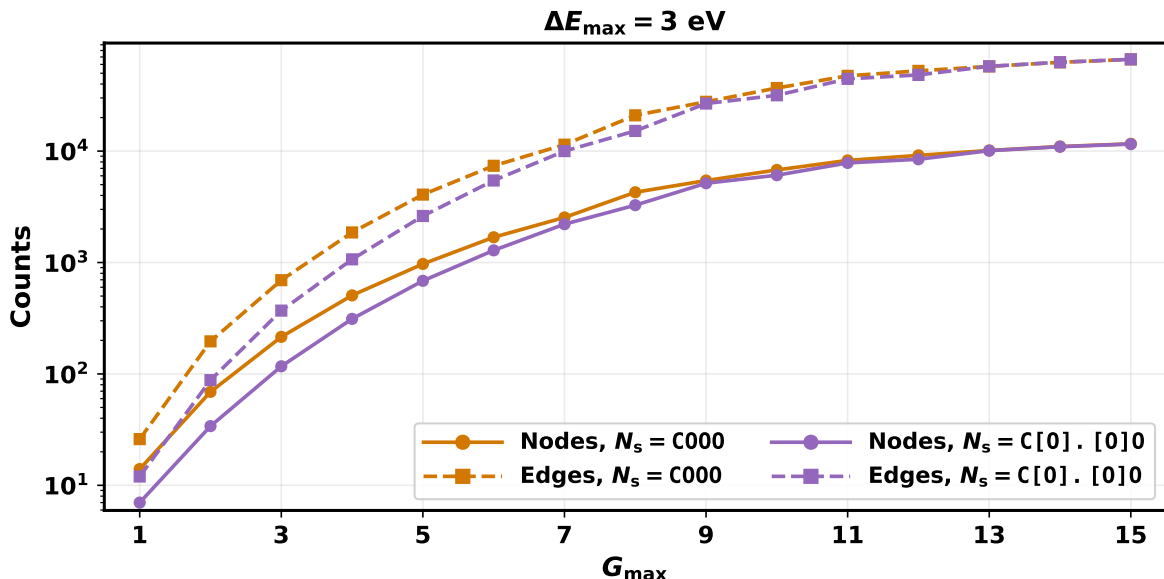


Figure 4.3: Scaling of n and e with bounded generation for the CH_4O_3 RN with generation depth $G_{\max}$ at fixed $\Delta E_{\max} = 3$ eV for the starting nodes $N_s = \text{CO}[0] \cdot [\text{OH}]$ and $N_s = \text{C000}$.

all stoichiometries studied. Network size increases strongly with stoichiometric complexity. Smaller systems, such as O_3 and CH_2O_2 , remain compact even at $G_{\max} = 7$, reaching at most $N = 214$ and $E = 790$. For O_3 , the node count changes only slightly, from $N = 9$ at $G_{\max} = 3$ to $N = 11$ at $G_{\max} = 7$, consistent with the limited chemical space available for this stoichiometry. In contrast, the larger CH_4O_3 , CH_6O_4 , $\text{C}_2\text{H}_4\text{O}_3$, and $\text{C}_2\text{H}_6\text{O}_4$ systems exhibit substantially more rapid growth with generation depth, reaching up to $N = 15,960$ and $E = 70,122$ for $\text{C}_2\text{H}_6\text{O}_4$ at $G_{\max} = 7$. Additional data for other combinations of $\Delta E_{\max}$ and $G_{\max}$ are provided in Appendix C, Section C.6.

Adding a single water molecule can substantially expand accessible chemical space. Adding water to the $\text{C}_2\text{H}_4\text{O}_3$ RN to give the $\text{C}_2\text{H}_6\text{O}_4$ RN nearly doubles n at $G_{\max} = 3$ ($N = 54$ to $N = 94$) and increases n by a factor of approximately 2.5 at $G_{\max} = 7$ ($N = 6258$ to $N = 15,960$). An analogous trend is observed when a water molecule is added to the CH_4O_3 RN to give the CH_6O_4 RN. These comparisons show that increasing stoichiometric complexity substantially enlarges the accessible reaction space by enabling additional stoichiometry-conserving transformations. However, this growth eventually becomes bounded once ad-

Table 4.1: Network statistics at allowed molecular energy difference between connected nodes of $\Delta E_{\text{max}} = 3$ eV and maximum generation depths $G_{\text{max}} = 3$ and 7. Notation: N_s : starting node (SMILES); n : nodes; e : edges; M : unique molecules; Q_M : molecule equivalence classes; N_N : neutral nodes; Q_N : neutral nodes collapsed by equivalence.

Stoichiometry	N_s	G_{max}	n	e	M	Q_M	N_N	Q_N
$\text{C}_2\text{H}_4\text{O}_3$	<chem>C=C.O=[O+] [O-]</chem>	3	54	170	44	36	31	20
		7	6258	24,270	3644	2286	2078	1311
$\text{C}_2\text{H}_6\text{O}_4$	<chem>C=C.O.O.O=[O+] [O-]</chem>	3	94	330	58	48	36	24
		7	15,960	70,122	5437	3786	3413	2357
CH_2O_2	<chem>C=[O+] [O-]</chem>	3	31	84	34	22	13	10
		7	214	790	178	81	103	43
CH_4O_3	<chem>C000</chem>	3	215	694	137	129	82	78
		7	2538	11,428	953	614	697	416
	<chem>CO[O] . [OH]</chem>	3	117	370	95	91	53	50
		7	2208	10,002	849	559	644	390
CH_6O_4	<chem>C000.O</chem>	3	375	1354	163	155	100	96
		7	9080	50,184	1636	1182	1434	1043
	<chem>CO[O] .O. [OH]</chem>	3	165	528	105	101	64	61
		7	6788	36,344	1233	905	1185	869
O_3	<chem>O=[O+] [O-]</chem>	3	9	18	11	7	7	4
		7	11	30	14	10	8	5
O_4H_2	<chem>O.O.O=[O+] [O-]</chem>	3	22	60	22	16	10	7
		7	228	1062	91	80	70	54

ditional atoms no longer introduce qualitatively new chemistry, as observed previously for larger stoichiometries [1].

Equivalence grouping ($M \rightarrow Q_M$) substantially reduces the number of energy calculations required: intermediate partially broken-bond states and related equivalent representations are merged into a single class, so only one representative energy need be computed per class rather than one per SMILES. This grouping does not affect the number of nodes or edges in the network. At $G_{\text{max}} = 7$ and $\Delta E_{\text{max}} = 3$ eV, the reduction ranges from $\approx 12\%$ (O_4H_2) to $\approx 55\%$ (CH_2O_2), with most CHO stoichiometries near 30–40%. For example, CH_2O_2 decreases from $M = 178$ to $Q_M = 81$ ($\approx 55\%$ reduction), whereas CH_4O_3 ($N_s = \text{C000}$)

decreases from $M = 953$ to $Q_M = 614$ ($\approx 36\%$ reduction). The reduction is generally smaller at $G_{\max} = 3$, where the accessible chemical space is still limited.

Neutral nodes account for approximately 25–55% of nodes for most CHO stoichiometries at $G_{\max} = 3$ (20–80% overall, with O_3 notably higher). The neutral fraction N_N/N generally decreases at larger $G_{\max}$ (6 of 7 stoichiometries; CH_2O_2 is an exception), reflecting the increasing accessibility of new radical and ionic intermediates. After collapsing by equivalence, Q_N/Q_M exceeds N_N/N for most networks. For example, in the CH_4O_3 RN with $N_s = C000$ at $G_{\max} = 7$, neutral nodes account for only $697/2538 \approx 27\%$ of all nodes, whereas neutral equivalence classes account for $416/614 \approx 68\%$ of molecule classes. This difference suggests that non-neutral species contribute disproportionately to combinatorial node growth, because many distinct node configurations can be formed from comparatively few unique molecules after collapsing into molecular equivalence classes.

4.3.2 Reaction Discovery in Bounded Networks

To evaluate whether bounded generation preserves chemically relevant reactivity, we examine a series of benchmark reactions spanning radical association, unimolecular decomposition, peroxide rearrangement, and Criegee-intermediate chemistry. Literature mechanisms are compared against the lowest-cost pathways recovered within the bounded RNs.

As the RN formalism represents reactions through individual bond-breaking and bond-forming transformations, concerted mechanisms are decomposed into sequences of elementary bond operations. Some intermediates generated along these paths may therefore contain formal charge separations or radical localizations that reflect the graph-based decomposition of the mechanism rather than physically stable species. The cumulative path cost, C , defined as the sum of edge weights along a reaction pathway, provides a heuristic measure of overall kinetic accessibility, with lower values corresponding to more kinetically favorable reaction

sequences.

Reactions in the O₃ and O₄H₂ Reaction Networks

The gas-phase thermal decomposition of ozone to molecular oxygen and ground-state atomic oxygen, $\text{O}_3 \longrightarrow \text{O}_2 + \text{O}$, is a key step in stratospheric chemistry and forms part of the Chapman cycle, $\text{O}_3 + \text{O} \longrightarrow 2 \text{O}_2$ [221]. The unimolecular decomposition step of ozone is studied in the O_3 RN as a test case for pathway discovery. In the O_4H_2 RN, the bimolecular reaction of ozone with water, $\text{O}_3 + \text{H}_2\text{O} \longrightarrow \text{HOOOOH}$, is explored as a route to hydrogen tetraoxide (HOOOOH). This metastable higher polyoxide has been characterized computationally as a minimum on the singlet potential energy surface and proposed as an intermediate in the HO_2 self-reaction and ozone–water oxidation chemistry [222, 223]. The gas-phase reaction of O_3 with H_2O has been studied computationally, with the dominant thermodynamic channel yielding H_2O_2 and O_2 [224]. The study of bounded O_4H_2 RN targets both the dominant channel yielding H_2O_2 and O_2 and the higher-cost pathway leading to HOOOOH.

The bounded-generation workflow recovers all three reaction pathways, shown in Figure 4.4 and summarized in Table 4.2. The unimolecular ozone decomposition (O3-1) is recovered with $C = 2.81$ as a two-step homolytic sequence. It proceeds via a diradical intermediate, $[O] [O^+] [O^-]$, in which the terminal O=O π -bond is broken homolytically before the re-

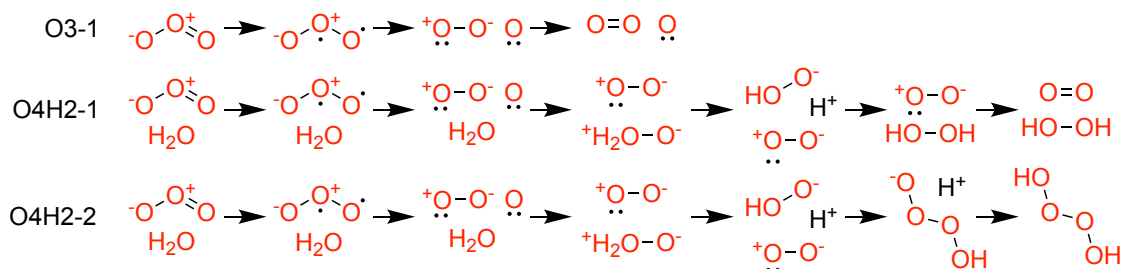


Figure 4.4: Predicted reaction pathways for ozone decomposition to molecular oxygen and atomic oxygen in the O₃ RN (O3-1), and ozone hydration to H₂O₂ and O₂ (O4H2-1) and to hydrogen tetraoxide HOOOOH (O4H2-2) in the O₄H₂ RN.

Table 4.2: For each reactant–product pair, the table reports the smallest bounded RN generated from the reactant that retains the lowest cost pathway to the product. The table reports the applied generation depth ($G_{\max}$), the maximum allowed molecular energy difference between connected nodes ($\Delta E_{\max}$), n , e , Q_N , and the cost C of the lowest karc-weighted path.

RN	Reactant	Product	$G_{\max}$	$\Delta E_{\max}$	n	e	Q_N	C
O ₃	0=[O+] [O-]	0=0. [0]	3	3	9	18	4	2.81
O ₄ H ₂	0.0=[O+] [O-]	0=0.00	4	3	47	150	15	6.36
O ₄ H ₂	0.0=[O+] [O-]	0000	6	3	170	734	42	6.76

maining O–O bond cleaves to release ground-state atomic oxygen and O₂. The bounded O₃ network uses $N_s = 0=[O^+][O^-]$ as the starting node, generated with $G_{\max} = 3$ and $\Delta E_{\max} = 3$ eV, and contains only $N = 9$ nodes and $E = 18$ edges, with four equivalence-class-reduced neutral nodes: $0=0. [0]$, $0=[O^+][O^-]$, $[O^+]0[O^-]$, and $[0]0[0]$.

Bounded O₄H₂ RNs with $N_s = 0.0=[O^+][O^-]$ as the starting node recover both bimolecular channels. Both channels share an initial sequence in which the charge-separated O₃ species undergoes the same two-step homolytic O–O cleavage identified in O3–1, breaking the terminal O=O π -bond and then the remaining O–O bond to give an oxygen-atom di-radical adjacent to water. A proton is then transferred from water to this oxygen atom, generating a zwitterionic intermediate that loses a further proton to give a common species consisting of a bare proton, a charge-separated O₂, and a charge-separated HO₂-type fragment. From this common intermediate the two channels diverge. The dominant channel (O4H2–1) yielding H₂O₂ + O₂ is recovered with $C = 6.36$ in a bounded RN with $G_{\max} = 4$ and $\Delta E_{\max} = 3$ eV, containing $N = 47$ nodes and $E = 150$ edges. Here the free proton recombines with the HO₂-type fragment to give H₂O₂, with the charge-separated O₂ fragment unchanged as a spectator. A final charge recombination within the O₂ fragment then gives neutral O₂, yielding H₂O₂ + O₂. The HOOOOH-forming channel (O4H2–2) is recovered with $C = 6.76$ in a bounded RN with $G_{\max} = 6$ and $\Delta E_{\max} = 3$ eV, containing $N = 170$ nodes and $E = 734$ edges. Here, instead, the HO₂-type fragment bonds directly to the charge-

separated O_2 to form a neutral HOOO-type species paired with a separate charge-separated O_2 , and sequential O–O bond formation and proton rearrangement between these fragments build up the O–O–O–O backbone, ultimately yielding HOOOOH. The karc cost ordering, O4H2–1 ($C = 6.36$) < O4H2–2 ($C = 6.76$), correctly identifies the $\text{H}_2\text{O}_2 + \text{O}_2$ channel as the lower-cost pathway, consistent with it being the dominant thermodynamic channel [224].

Reactions in the CH_2O_2 Reaction Network

The simplest Criegee intermediate, formaldehyde oxide (CH_2OO), undergoes unimolecular decay to formic acid (HCOOH). This well-established pathway is initiated by 1,3-ring closure to dioxirane, followed by O–O bond homolysis to a methylene bisoxy biradical and subsequent rearrangement to formic acid [225, 226]. A bounded RN with $N_s = \text{C}=[\text{O}^+][\text{O}^-]$, $G_{\max} = 6$, and $\Delta E_{\max} = 4$ eV recovers this mechanism. The network contains $N = 227$ nodes and $E = 824$ edges, with $Q_N = 42$ equivalence-class-reduced neutral nodes. The pathway (CH2O2–1) has $C = 6.16$, recovering the literature mechanism’s three stages across six steps, as shown in Figure 4.5. The first two steps recover the 1,3-ring closure. Starting from the charge-separated carbonyl oxide $\text{C}=[\text{O}^+][\text{O}^-]$, charge recombination gives the Criegee intermediate resonance form $[\text{CH}_2^+]\text{O}[\text{O}^-]$, which closes to dioxirane (C1001). The third step recovers the O–O homolysis, opening the ring to give the methylene bisoxy biradical $[\text{O}]\dot{\text{C}}[\text{O}]$. The remaining three steps recover the rearrangement to formic acid, with the literature’s 1,2-H shift represented in the RN via a radical-H intermediate. A hydrogen atom is first lost from carbon to give $[\text{H}] \cdot [\text{O}][\dot{\text{C}}\text{H}][\text{O}]$, a C–O π -bond forms to give $[\text{H}] \cdot [\text{O}]\text{C}=\text{O}$, and recombination of the hydrogen atom yields formic acid ($\text{O}=\text{C}\text{O}$). Recovery of this mechanistic

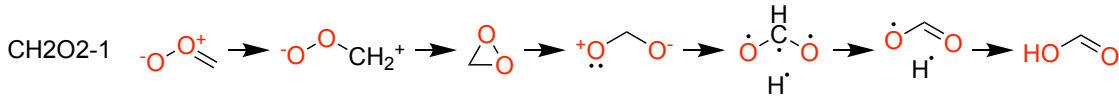


Figure 4.5: Predicted reaction pathways for decay of the Criegee intermediate CH_2OO to formic acid (CH2O2–1) in the CH_2O_2 RN.

sequence confirms that the bounded-generation workflow correctly identifies the established dioxirane-mediated pathway as an accessible route to formic acid [225, 226].

Reactions in the CH₄O₃ and CH₆O₄ Reaction Networks

The reaction of a methylperoxy radical (CH₃O₂•) with a hydroxyl radical (OH•) proceeds on the singlet potential energy surface via barrierless O–O bond formation to give the methyl trioxide adduct CH₃OOOH (TRIOX, C000) [227, 228]. At ambient pressure TRIOX is short-lived and dissociates primarily by O–O homolysis adjacent to carbon, giving CH₃O• + HO₂• (channel 1a), the dominant atmospheric channel [227, 229–231]. Two further decomposition channels are accessible. A retro-peroxidation delivers CH₃OH + O₂ via heterolytic O–O cleavage and 1,3-proton transfer, assisted by water [227, 228, 232], and a β -elimination yields CH₂O + H₂O₂, both minor channels under atmospheric conditions [228, 233]. We first examine the step-by-step formation and decomposition of methyl trioxide within the CH₄O₃ RN.

Formation of CH₃OOOH from CH₃O₂• + OH• is recovered as a single O–O bond-forming step with $C = 1.34$ (Figure 4.6 CH4O3–1), consistent with the literature mechanism of barrierless O–O association which requires only a single bond change [227]. The bounded RN uses $N_s = \text{C0}[0] \cdot [\text{OH}]$, $G_{\text{max}} = 1$, and $\Delta E_{\text{max}} = 3$ eV and contains only $N = 7$ nodes and $E = 12$ edges. Likewise, O–O homolysis of CH₃OOOH to CH₃O• + HO₂• is recovered as a single bond-breaking step with $C = 0.99$ (Figure 4.6 CH4O3–2). The mechanism is found in a bounded RN with $N_s = \text{C000}$, $G_{\text{max}} = 1$, and $\Delta E_{\text{max}} = 3$ eV.

The β -elimination from CH₃OOOH to CH₂O + H₂O₂ is recovered at $C = 2.68$ via a four-step ionic path (Figure 4.6 CH4O3–3). The path proceeds through heterolytic C–H cleavage to give the charge-separated trioxide [CH₂–]OOO and a free proton. Heterolytic O–O cleavage then gives the charge-separated formaldehyde-type fragment [CH₂–][O⁺] and a ter-

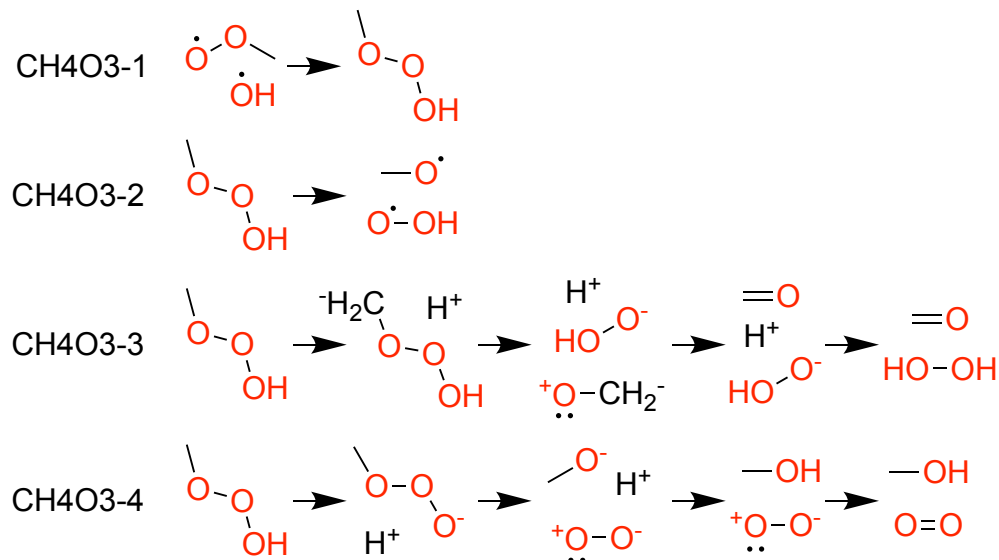


Figure 4.6: Predicted reaction pathways in the CH_4O_3 RN for CH_3OOOH formation from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$ (CH4O3-1), and for CH_3OOOH decomposition to $\text{CH}_3\text{O}^\bullet + \text{HO}_2^\bullet$ (CH4O3-2), $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$ (CH4O3-3), and $\text{CH}_3\text{OH} + \text{O}_2$ (CH4O3-4).

minimal HO_2^- -type fragment, with the free proton unchanged. Charge recombination within $[\text{CH}_2^-][\text{O}^+]$ then gives neutral $\text{C}=\text{O}$. A final proton transfer to the HO_2^- -type fragment yields $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$. This ionic sequence represents the bond-level decomposition of the β -elimination motif of C–H and O–O cleavage. The path is recovered in a bounded RN with $N_s = \text{C000}$, $G_{\max} = 4$, and $\Delta E_{\max} = 3$ eV containing $N = 507$ and $E = 1862$.

For the retro-peroxidation of CH_3OOOH to $\text{CH}_3\text{OH} + \text{O}_2$, the lowest-cost path of $C = 4.39$ (Figure 4.6 CH4O3-4) proceeds via a four-step ionic sequence. The terminal O–H bond cleaves heterolytically first to give $[\text{H}^+]$ and CH_3OOO^- . Heterolytic O–O cleavage then gives CH_3O^- alongside a charge-separated O_2 . The proton recombines at the methoxide oxygen to give CH_3OH , still paired with the charge-separated O_2 . Finally, charge recombination within the O_2 gives neutral O_2 . This ionic, heterolytic sequence is consistent with the water-assisted 1,3-proton-transfer mechanism proposed for the heterolytic fragmentation of CH_3OOOH [232]. This path is recovered with $N_s = \text{C000}$, $G_{\max} = 4$, and $\Delta E_{\max} = 3$ eV.

The path cost ordering $\text{CH}_3\text{O}^\bullet + \text{HO}_2^\bullet$ ($C = 0.99$) $<$ $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$ ($C = 2.68$) $<$ $\text{CH}_3\text{OH} +$

Table 4.3: For each reactant–product pair, the table reports the smallest bounded CH_4O_3 RN generated from the reactant that retains the lowest cost pathway to the product.

Reactant	Product	$G_{\max}$	$\Delta E_{\max}$	n	e	Q_N	C
$\text{CO}[\text{O}] \cdot [\text{OH}]$	C000	1	3	7	12	5	1.34
C000	$\text{C}[\text{O}] \cdot [\text{O}]\text{O}$	1	3	14	26	9	0.99
C000	$\text{C}=\text{O}.\text{O0}$	4	3	507	1862	170	2.68
C000	$\text{CO}.\text{O}=\text{O}$	4	3	507	1862	170	4.39
$\text{CO}[\text{O}] \cdot [\text{OH}]$	$\text{C}[\text{O}] \cdot [\text{O}]\text{O}$	2	3	34	88	17	2.33
$\text{CO}[\text{O}] \cdot [\text{OH}]$	$\text{C}=[\text{O}+] [\text{O}-] .\text{O}$	7	3	2208	10,002	390	5.94
$\text{CO}[\text{O}] \cdot [\text{OH}]$	$\text{CO}.\text{O}=\text{O}$	5	3	685	2616	219	5.73
$\text{CO}[\text{O}] \cdot [\text{OH}]$	$\text{CO}.\text{O}=\text{O}$	3	4	194	652	67	6.55
$\text{CO}[\text{O}] \cdot [\text{OH}]$	$\text{C}=\text{O}.\text{O0}$	5	3	685	2616	219	4.02

O_2 ($C = 4.39$) ranks O–O homolysis as most accessible, with β -elimination and retro-peroxidation as comparably accessible higher-cost channels. This is consistent with the dominance of the homolysis channel under atmospheric conditions [228]. The network parameters for all bounded RNs discussed in this section are summarized in Table 4.3.

We now consider paths from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$ directly to each product, without manually staging through the intermediate CH_3OOOH . This tests whether the paths found recover the TRIOX intermediate and reproduces the same pathways as identified above.

The lowest-cost path from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$ to $\text{CH}_3\text{O}^\bullet + \text{HO}_2^\bullet$ was recovered with $C = 2.33$ in two steps routing through CH_3OOOH (Figure 4.7 CH4O3–5), consistent with channel 1a proceeding via the TRIOX adduct on the singlet surface [227]. This path is retained in a bounded RN with $N_s = \text{CO}[\text{O}] \cdot [\text{OH}]$, $G_{\max} = 2$, and $\Delta E_{\max} = 3$ eV containing $N = 34$ and $E = 88$.

The path from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$ to $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ (channel 1b) is recovered with $C = 5.94$ in eight steps routing through CH_3OOOH (Figure 4.7 CH4O3–6). The path proceeds by C–O heterolysis of TRIOX to give a free proton and a charge-separated trioxide fragment. O–O heterolysis and proton transfer to the terminal oxygen then give a charge-separated intermediate bearing an OH_2^+ group. Loss of water gives a charge-separated diradical-type

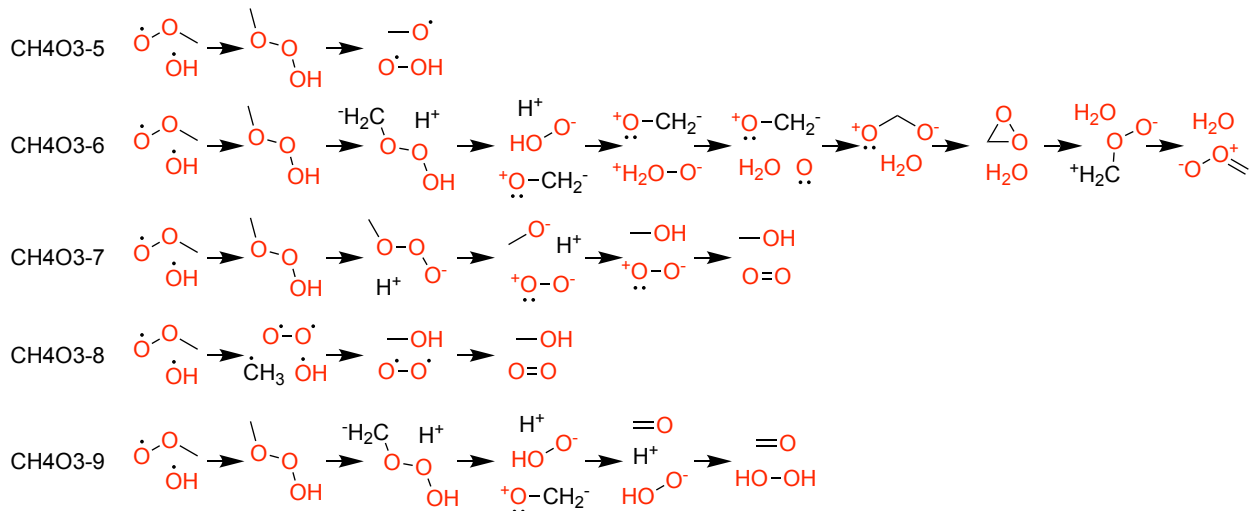


Figure 4.7: Predicted reaction pathways in the CH_4O_3 RN starting from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$: O–O homolysis to $\text{CH}_3\text{O}^\bullet + \text{HO}_2^\bullet$ via CH_3OOOH (CH4O3–5), formation of $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ via the dioxirane manifold (CH4O3–6), retro-peroxidation to $\text{CH}_3\text{OH} + \text{O}_2$ via CH_3OOOH (CH4O3–7) and bypassing CH_3OOOH via direct radical fragmentation (CH4O3–8), and β -elimination to $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$ via CH_3OOOH (CH4O3–9).

species, which rearranges and undergoes ring closure to neutral dioxirane. Ring opening gives the charge-separated carbonyl oxide, and final charge recombination yields $\text{CH}_2\text{OO} + \text{H}_2\text{O}$, recovering channel 1b through the dioxirane manifold, consistent with the known interconversion of methyldioxirane-type rings and carbonyl oxide [225]. This path is retained in a bounded RN with $N_s = \text{CO}[0] \cdot [\text{OH}]$, $G_{\max} = 7$, and $\Delta E_{\max} = 3$ eV.

For $\text{CH}_3\text{OH} + \text{O}_2$, two paths are recovered. The lowest-cost path, with $C = 5.73$ and five steps (Figure 4.7 CH4O3–7), routes through CH_3OOOH . It follows the same ionic retro-peroxidation sequence identified in CH4O3–4, with the additional initial O–O bond-formation step from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$. This path is recovered with $N_s = \text{CO}[0] \cdot [\text{OH}]$, $G_{\max} = 5$, and $\Delta E_{\max} = 3$ eV. A second path with a higher cost of $C = 6.55$ but only three steps (Figure 4.7 CH4O3–8) bypasses CH_3OOOH entirely via direct radical fragmentation. Homolytic C–O cleavage of the methylperoxy radical gives $\text{CH}_3^\bullet$, $\text{OH}^\bullet$, and an O_2 -type diradical as three separate fragments. Recombination of $\text{CH}_3^\bullet$ and $\text{OH}^\bullet$ then gives CH_3OH , while the diradical recombines to give O_2 . This path is recovered with $G_{\max} = 3$ and $\Delta E_{\max} = 4$ eV.

and carries a substantially higher cost than the TRIOX-mediated route, consistent with its higher energy threshold. A direct bimolecular channel yielding $\text{CH}_3\text{OH} + \text{O}_2$ without TRIOX stabilization has been proposed but remains poorly characterized experimentally [231, 233].

The path from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$ to $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$ was found with $C = 4.02$ and five steps routing through CH_3OOOH (Figure 4.7 CH4O3–9). It follows the same ionic β -elimination sequence identified in CH4O3–3, with the additional initial O–O bond-formation step from $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet$. It is recovered in a bounded RN with $N_s = \text{CO}[\text{O}] \cdot [\text{OH}]$, $G_{\max} = 5$, and $\Delta E_{\max} = 3$ eV containing $N = 685$ and $E = 2616$.

With the exception of CH4O3–8, all lowest-cost paths route through CH_3OOOH , confirming that the RNs recover TRIOX as the central intermediate. The path cost ordering across all direct bimolecular channels, $\text{CH}_3\text{O}^\bullet + \text{HO}_2^\bullet$ ($C = 2.33$) $<$ $\text{CH}_2\text{O} + \text{H}_2\text{O}_2$ ($C = 4.02$) $<$ $\text{CH}_3\text{OH} + \text{O}_2$ ($C = 5.73$) $<$ $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ ($C = 5.94$), places O–O homolysis as the most accessible channel and $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ as least accessible. This is in qualitative agreement with the atmospheric branching hierarchy in which channel 1a dominates [231, 233]. The minimal networks range from $N = 7$ nodes for TRIOX formation to $N = 2208$ for the $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ channel, as reported in Table 4.3.

The CH_6O_4 RN adds water to the CH_4O_3 , allowing us to test whether the pathways identified in the anhydrous system are preserved with hydrated conditions. It also provides a means of assessing how hydration affects bounded-RN size. Water is known to strongly influence Criegee-intermediate chemistry through hydrogen-bonded complexes, water-assisted proton transfer, and water-dimer reaction channels, which can substantially enhance the reactivity of CH_2OO relative to the monomer-water pathway [234–237]. In the CH_6O_4 RN, however, the inclusion of water does not alter the lowest-cost paths or their costs relative to the anhydrous CH_4O_3 system. The same mechanistic sequences are recovered with the same karc cost ordering, and water acts as a spectator rather than an active participant in the lowest-cost channels. The primary effect of the additional water molecule is a substantial

increase in bounded-RN size. The β -elimination network grows from $N = 507$ to $N = 1899$ nodes and from $E = 1862$ to $E = 8850$ edges, and the $\text{CH}_2\text{OO} + \text{H}_2\text{O}$ channel grows from $N = 2208$ to $N = 6788$ nodes and from $E = 10002$ to $E = 36344$ edges. This reflects the combinatorial expansion of the chemical space accessible when an additional water molecule is present.

Reactions in the $\text{C}_2\text{H}_4\text{O}_3$ Reaction Network

Ethylene ozonolysis provides a compact two-carbon benchmark for testing bounded generation in a larger CHO stoichiometry. The reaction proceeds through concerted [3+2] cycloaddition of ozone across the C=C bond to form a primary ozonide (1,2,3-trioxolane, **C1C0001**), followed by fragmentation of the cyclic peroxide to formaldehyde and the simplest Criegee intermediate, CH_2OO [17, 238, 239]. This chemistry is a prototypical route for alkene oxidation in the troposphere and has served as a benchmark for experimental and computational studies of Criegee-intermediate formation [17, 239]. We examine whether bounded generation recovers the 1,2,3-trioxolane intermediate and the corresponding fragmentation pathway within a bounded $\text{C}_2\text{H}_4\text{O}_3$ RN.

The bounded-generation workflow recovers both the 1,2,3-trioxolane formation and fragmentation pathways, illustrated in Figure 4.8. The lowest-cost path from $\text{C}_2\text{H}_4 + \text{O}_3$ to 1,2,3-trioxolane **C1C0001** is recovered with $C = 3.23$ in six steps ($\text{C}_2\text{H}_4\text{O}_3\text{-1}$), proceeding through sequential bond-forming steps that first activate the ozone fragment homolytically before ring closure. The path passes through a four-membered zwitterionic ring intermediate $[\text{O}^-][\text{O}^+]\text{1CC01}$, a 1,2-dioxetane-type species carrying formal charges on the oxygen atoms. As in the CH_4O_3 and O_4H_2 networks, this is an artifact of the single-bond-change decomposition rather than a chemically stable species. It represents the bond-level decomposition of the concerted [3+2] cycloaddition into sequential steps [17]. A second, higher-cost path to the same product is recovered with $C = 4.17$ ($\text{C}_2\text{H}_4\text{O}_3\text{-2}$) via an alternative zwitterionic se-

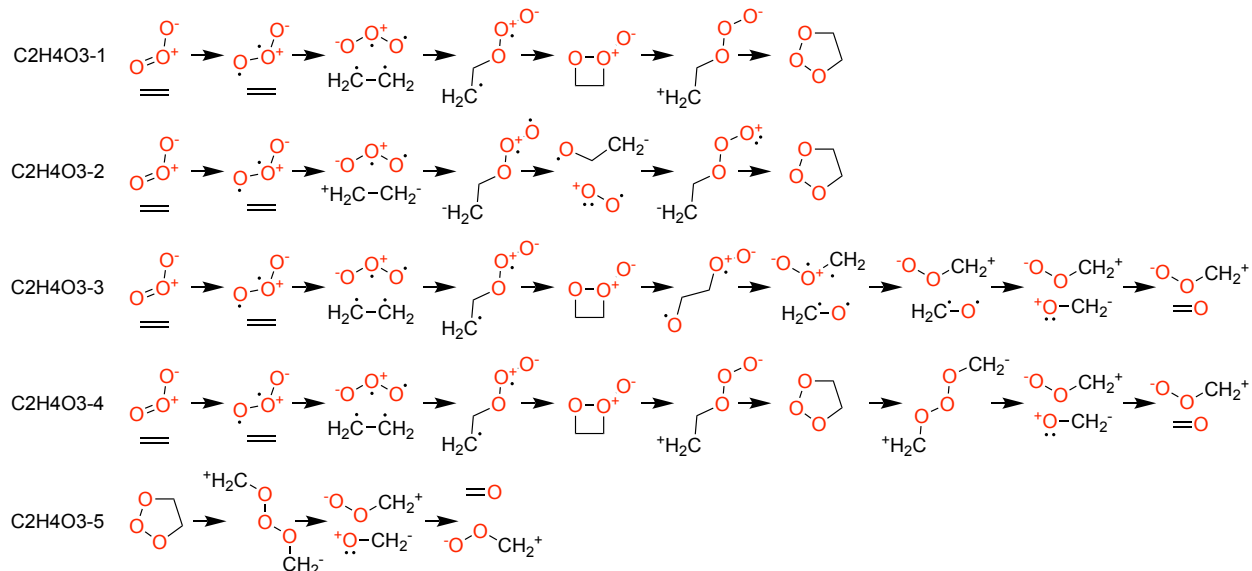


Figure 4.8: Predicted reaction pathways for ethylene ozonolysis in the $\text{C}_2\text{H}_4\text{O}_3$ RN: formation of the primary ozonide **C1C0001** from $\text{C}_2\text{H}_4 + \text{O}_3$ (C2H4O3-1 and C2H4O3-2), fragmentation of the primary ozonide to $\text{CH}_2\text{O} + \text{CH}_2\text{OO}$ (C2H4O3-5), a direct path from $\text{C}_2\text{H}_4 + \text{O}_3$ to $\text{CH}_2\text{O} + \text{CH}_2\text{OO}$ bypassing the trioxolane (C2H4O3-3), and an alternative route through the trioxolane (C2H4O3-4).

quence. Both paths are retained in a bounded RN with $N_s = \text{C}=\text{C}.0=[\text{O}^+][\text{O}^-]$, $G_{\max} = 7$, and $\Delta E_{\max} = 3$ eV containing $N = 6258$ nodes and $E = 24270$ edges, and $Q_N = 1311$.

Fragmentation of 1,2,3-trioxolane to $\text{CH}_2\text{O} + \text{CH}_2\text{OO}$ is recovered with $C = 0.86$ in three steps starting from **C1C0001** (C2H4O3-5). The path proceeds via $[\text{CH}_2^+]\text{O}[\text{CH}_2^-] \rightarrow [\text{CH}_2^+]\text{O}[\text{O}^-] \cdot [\text{CH}_2^-][\text{O}^+] \rightarrow \text{C}=\text{O} \cdot [\text{CH}_2^+]\text{O}[\text{O}^-]$. This represents the bond-level decomposition of the concerted cycloreversion in which simultaneous O–O and C–C bond cleavage of the trioxolane ring delivers formaldehyde and the Criegee intermediate, consistent with the literature mechanism [238, 239]. The direct path from $\text{C}_2\text{H}_4 + \text{O}_3$ to $\text{CH}_2\text{O} + \text{CH}_2\text{OO}$ without passing through an explicit 1,2,3-trioxolane intermediate is recovered with $C = 3.61$ in nine steps (C2H4O3-3), and an alternative route through the trioxolane with $C = 4.09$ in nine steps (C2H4O3-4). The latter path proceeds through the same six steps identified for C2H4O3-1 to form **C1C0001**, followed by the same three-step fragmentation sequence identified for C2H4O3-5, so that its cost is approximately the sum of the two constituent

paths ($3.23 + 0.86 = 4.09$), confirming that once the trioxolane is formed, fragmentation is highly accessible and that the network correctly identifies 1,2,3-trioxolane as the central intermediate in the reaction.

4.4 Discussion

4.4.1 Network Reduction and Bounded Generation

Bounded generation explores only the region of chemical space reachable from a chosen starting node N_s within specified $G_{\max}$ and $\Delta E_{\max}$ bounds. Unlike the exhaustive RN, which is uniquely determined by the stoichiometry and rule set, bounded RNs depend explicitly on the starting node and therefore trade global completeness for substantial reductions in network size. This enables targeted exploration of chemically relevant regions of reactive chemical space without exhaustive enumeration of the full stoichiometric chemical space.

The use of reversible transformation rules enables mechanism searches to proceed from either reactants or products. Each bond-breaking transformation is paired with a corresponding bond-forming rule. Consequently, any pathway reachable from the product is also reachable from the reactant, allowing bounded generation to recover mechanisms in either direction from any starting node. However, reversibility alone does not guarantee pathway recovery. Every intermediate and every edge along a pathway must remain within the specified $G_{\max}$ and $\Delta E_{\max}$ bounds. If either constraint excludes a required intermediate, the pathway will not be recovered even if it is formally accessible under the reaction rules. The bounds must therefore be chosen with some knowledge of the expected pathway depth and energy range, or set sufficiently loose to avoid excluding relevant intermediates.

The maximum generation depth $G_{\max}$ is the dominant parameter controlling RN size because

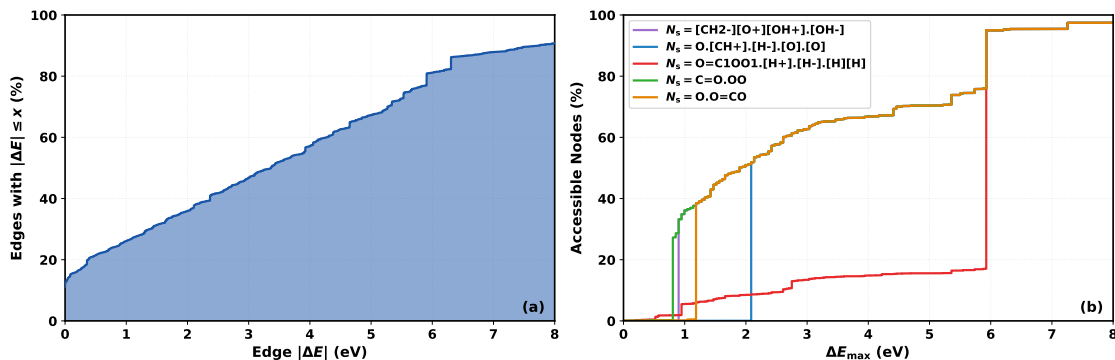


Figure 4.9: Cumulative percentage of (a) edges with $|\Delta E| \leq x$ for the exhaustive CH_4O_3 RN and (b) accessible nodes retained as a function of the bounded-generation energy threshold ΔE_{max} for different N_s .

RNs exhibit approximately exponential growth over early and intermediate generations. This behavior follows from their lattice-like topology, where successive generation shells expand through relatively regular local branching patterns [1]. Even modest reductions in G_{max} can therefore produce orders-of-magnitude decreases in network size before network size saturation is reached.

By comparison, ΔE_{max} provides more gradual and secondary control over RN size. Figure 4.9(a) shows that approximately 42% of exhaustive CH_4O_3 edges lie within 3 eV and 54% within 4 eV. The energy bound therefore acts primarily as a permissive thermodynamic filter against high-energy transformations rather than the primary determinant of network size. The fraction of nodes accessible at a given ΔE_{max} nevertheless depends strongly on the location of the starting node within the underlying energy landscape [240]. As shown in Figure 4.9(b), low-energy chemically feasible starting configurations retain approximately 63–67% of accessible nodes in the CH_4O_3 RN at $\Delta E_{\text{max}} = 3$ –4 eV, whereas high-energy starting configurations retain only ≈ 13 –15% under the same bounds. This dependence is strongest for shallow targeted explorations and diminishes as G_{max} or ΔE_{max} increases, consistent with the convergence behavior observed in Figures 4.3 and 4.9(b). Although only a modest fraction of edges satisfy low ΔE_{max} thresholds, these edges still permit access to large regions of the RN. Recovery of all edges in the CH_4O_3 RN requires $\Delta E_{\text{max}} \approx 23$ eV.

Figures showing $\Delta E_{\max}$ scaling for the O_3 , O_4H_2 , and CH_2O_2 exhaustive RNs are provided in Appendix C, Section C.7.

An important limitation of $\Delta E_{\max}$ as a bounding criterion is that it filters thermodynamic energy differences between connected nodes rather than transition-state barriers. Consequently, low- ΔE transformations are not necessarily kinetically accessible, and high barriers may connect energetically similar nodes [55, 219]. $\Delta E_{\max}$ should be interpreted as a broad thermodynamic filter that excludes highly unstable or unlikely transformations and energetically remote regions of chemical space rather than as a direct criterion for kinetic feasibility.

A substantial fraction of generated molecules cannot be assigned semiempirical energies. For the exhaustive CH_4O_3 RN, only ≈ 3047 of 58,821 uniquely generated molecules yielded successful PM7/COSMO calculations ($\approx 5\%$), visible as the separation between the generated and energy-validated node curves in Figure 4.2(a). This large loss likely reflects the presence of electronically unusual charge-separated and radical configurations generated by the rule set that fail to converge in MOPAC [154, 241]. Equivalence grouping partially mitigates this representational problem by merging structures connected through zero-energy transformations that alter bond orders, formal charges, or radical localizations without changing molecular connectivity. This treatment reduces the number of chemically redundant graph states requiring energetic evaluation. The reduction from M molecules to Q_{M} equivalence classes ranges from $\approx 12\%$ for O_4H_2 to $\approx 55\%$ for CH_2O_2 at $G_{\max} = 7$ and $\Delta E_{\max} = 3$ eV, with most RNs seeing a reduction of 30–40%.

These observations suggest the following practical strategy for bounded RN generation. First, select a starting node corresponding to a physically realizable species relevant to the target chemistry. Next, increase $G_{\max}$ incrementally until the target pathway is recovered, or choose sufficiently large bounds. Finally, apply $\Delta E_{\max} \approx 3\text{--}4$ eV as a permissive filter against high-energy transformations. For the stoichiometries examined here, this strategy reliably recovers chemically relevant pathways while maintaining computationally tractable network sizes.

4.4.2 Radical Transformation Rules Extension

Several intermediates appearing along recovered reaction pathways carry formal ionic or radical configurations that do not correspond to physically stable species. Examples include the 1,2-dioxetane-type zwitterion $[O^-][O^+]1CCO1$ in the $C_2H_4O_3$ network, the methylene bisoxy biradical $[O^+]C[O^-]$ in the CH_2O_2 network, and proton-transfer zwitterions in the O_4H_2 network. These species arise because the RN formalism represents elementary transformations as sequential single-bond-breaking or bond-forming events [24, 66]. Concerted mechanisms involving simultaneous bond rearrangements are therefore decomposed into multiple sequential steps, producing pathways that pass through formal graph intermediates. Such intermediates should be interpreted as bookkeeping constructs within the graph representation rather than as predictions of experimentally observable species.

Radical chemistry introduces an additional layer of formalism. Within `RDKit` [77], atom-centered unpaired-electron counts are assigned using valence and bond-counting rules rather than explicit electronic spin states. These formal radical counts function primarily as pattern-matching and product-construction constraints during reaction-rule application and do not necessarily correspond to physically realizable radical multiplicities. In particular, multiple formal radical electrons may be assigned to the same atom, producing configurations with no direct physical analogue. Transformations involving hydrogen atoms were restricted to distribute unpaired electrons across the two participating atoms, partially reducing such unphysical assignments. Nevertheless, the formal nature of the radical representation remains an inherent limitation of the SMARTS-based rule encoding.

These formal graph representations are directly connected to the failed energy calculations discussed in Section 4.4.1. Charge-separated and high-radical-count configurations generated by the rule set frequently fail to converge in PM7/COSMO calculations [154]. This is partly because PM7 was parameterized primarily for closed-shell organic species and ex-

hibits known limitations for highly charged and open-shell systems [154, 241]. More robust energetic models could reduce this failure rate, although it remains unclear which of these formal intermediates correspond to chemically relevant configurations and which simply reflect unimportant graph representations. Discarding failed molecules, as done here, is therefore conservative but excludes chemistry that formally passes through these states. The appropriate treatment ultimately depends on whether the goal is exhaustive enumeration of all rule-permissible pathways or restriction to more physically interpretable ones.

A related practical issue arises from implicit hydrogen handling within `RDKit`. During reaction-rule application, hydrogens may be added or removed automatically to satisfy valence constraints in ways that differ from the intended transformation. Consequently, some edges are generated in only one direction because the reverse transformation fails to reproduce the expected implicit hydrogen count in the product SMILES. As a result, certain transformations appear in bounded RNs only through indirect routes, and complete bidirectional recovery of all edges cannot be guaranteed.

Despite these limitations, the radical extension correctly captures the key radical processes across all benchmark systems: homolytic O–O cleavage in O_3 decomposition and CH_4O_3 channel 1a, diradical intermediates in CH_2OO unimolecular decay, and barrierless O–O radical recombination in $\text{CH}_3\text{O}_2^\bullet + \text{OH}^\bullet \rightarrow \text{CH}_3\text{OOOH}$. The extended SMARTS notation [217] provides sufficient expressiveness to encode these transformations, and the benchmark results confirm that the formal radical representation does not prevent recovery of physically meaningful mechanistic pathways.

4.4.3 Predicting Reaction Pathways

The `karc` edge cost provides a thermodynamic ordering of competing pathways that is broadly consistent with established atmospheric mechanisms across the benchmark systems studied

here. In O_3 decomposition and $\text{C}_2\text{H}_4\text{O}_3$ ozonolysis, the lowest-cost paths correspond to the dominant atmospheric channels. In CH_2OO unimolecular decay, the lowest-cost pathway proceeds through dioxirane to form formic acid, consistent with the accepted mechanism and observed product distribution. The relative separation between competing channels is also informative. Trioxolane fragmentation in the $\text{C}_2\text{H}_4\text{O}_3$ RN has a cost of $C = 0.86$, substantially lower than trioxolane formation at $C = 3.23$. This supports the primary ozonide intermediate as an accessible entry point and suggests trioxolane fragmentation is a comparatively less favorable step. In contrast, the CH_4O_3 β -elimination ($C = 2.68$) and retro-peroxidation ($C = 4.39$) channels have higher costs, consistent with both appearing as minor atmospheric pathways without a strong preference for either mechanism.

Bounded generation occupies a distinct position among reaction-network exploration methods. Stochastic approaches such as ab initio molecular dynamics nanoreactors and Monte Carlo graph-traversal methods [45, 242, 243] can efficiently sample high-dimensional chemical space, but they do not guarantee coverage of all accessible pathways within a given run. Transition-state search and barrier-based filtering methods [35, 207, 244] provide rigorous activation barriers and direct kinetic information, but require many quantum-chemical calculations and scale poorly to large networks. Adaptive and context-driven expansion schemes [44, 206, 245, 246] focus exploration iteratively using learned or user-defined scoring functions. These approaches can be highly efficient for targeted problems, although they introduce dependence on training data or scoring design. Machine-learning-guided exploration methods similarly require prior data for the systems of interest [53, 55]. By contrast, bounded generation is deterministic and exhaustive within the specified G_{max} and ΔE_{max} bounds. Every node and edge satisfying the bounds is enumerated without any stochastic element. The tradeoff is that bounded generation replaces activation barriers with thermodynamic proxies and remains sensitive to the choice of starting node and parameter values. This makes the method well suited for exploratory mechanism discovery in systems where prior data are sparse and completeness within a defined region of chemical space is more

valuable than maximal computational efficiency.

4.5 Conclusions

Bounded stoichiometry-preserving reaction networks (RNs) provide a computationally efficient approach for exploring chemically relevant regions of reactive chemical space without exhaustive enumeration of the full stoichiometric RN [22, 23]. Across all benchmark systems studied here, bounded generation recovered established atmospheric reaction pathways while reducing RN size by orders of magnitude relative to exhaustive generation. The maximum generation depth $G_{\max}$ is the dominant parameter controlling network size, reflecting the lattice-like structure of RNs [1], where successive generations expand through relatively regular local branching patterns and produce approximately exponential growth in node count. In contrast, $\Delta E_{\max}$ acts primarily as a permissive thermodynamic filter against high-energy transformations rather than as a direct kinetic criterion [55, 219]. Values of $\Delta E_{\max} \approx 3\text{--}4$ eV retained chemically relevant pathways across the systems examined here while substantially reducing network size. The karc edge-cost ordering was also broadly consistent with established atmospheric mechanisms.

Several limitations remain. Although bounded generation substantially reduces RN size, it does not eliminate combinatorial growth, and large rule sets remain computationally expensive. Pathway recovery is also sensitive to the reaction rules, starting-node choice, and parameter selection. Incomplete or overly permissive rules can produce missing or unphysical transformations. Sequential bond-breaking and bond-forming rules further generate formal charge-separated and radical graph intermediates that are difficult to evaluate reliably with semiempirical methods [154, 241]. The karc edge-cost scheme also carries known limitations as a predictor of reaction feasibility [24, 247]. Additional constraints arise from radical representations in SMILES and RDKit, including implicit hydrogen handling that can produce

unexpected product SMILES and asymmetric transformations. Consequently, some edges may only be recovered indirectly during bounded generation. The RN representation itself also excludes stereoisomerism and conformational degrees of freedom [23, 24]. Bounded RNs should therefore be interpreted as discrete graph representations of accessible reactive chemical space rather than as complete kinetic models.

Several future directions are promising. Coupling bounded generation with Euclidean embedding of reaction networks [247] could enable low-dimensional sampling and visualization of bounded chemical space. Exploration efficiency for large stoichiometries could also be improved through stochastic strategies such as Monte Carlo sampling of reaction space [45, 242]. Instead of constructing a single large RN, mechanism discovery could also proceed through multiple smaller bounded RNs centered on specific intermediates or product channels. Systematic enumeration of neutral-node subsets within bounded RNs could further identify which neutral molecules are accessible within specified energetic and depth constraints. Finally, replacing or augmenting on-the-fly semiempirical energies with more accurate energetic models would improve the reliability of energy-based pruning, particularly for radical, open-shell, and charge-separated species.

Chapter 5

Conclusion

This dissertation has analyzed stoichiometry-preserving reaction networks (RNs) and presented methods for both searching and constructing them more efficiently through Euclidean embeddings and energetically bounded network generation. RNs provide a tractable representation of chemical potential energy surfaces (PESs) and address a central challenge in computational chemistry of identifying chemically feasible reaction pathways while avoiding the combinatorial explosion associated with direct exhaustive PES exploration. By combining network analysis, Euclidean embeddings, and energetically informed network generation, this dissertation demonstrates that large and complex regions of chemical space can be explored in a systematic and computationally efficient manner.

More broadly, this work shows that RNs generated with `colibri2` are not arbitrary graphs, but instead possess meaningful structural and geometric properties that can be exploited computationally. Leveraging the understanding developed throughout this dissertation and the methods introduced here, `colibri2` provides a framework for balancing broad, relatively unbiased exploration of chemical space with targeted exploration guided by energetic constraints. As a result, these methods provide a generalizable foundation for reaction discovery

and pathway prediction across diverse areas of chemistry.

Three primary contributions were highlighted in this dissertation:

- **Chapter 2** established that stoichiometry-preserving RNs exhibit structural properties resembling low-dimensional regular lattices with a small amount of random edge rewiring. These findings demonstrate that RNs possess an underlying structure that can be leveraged to develop more efficient methods for RN construction, sampling, and search.
- **Chapter 3** demonstrated that RNs can be embedded into low-dimensional Euclidean spaces with relatively low distortion. This embeddability extends to weighted RNs, where edge weights encode chemical feasibility. Encoder-based embeddings were shown to preserve chemically meaningful pathways while maintaining a relatively low computational cost. By incorporating chemically informed edge-weighting schemes, embedded-space searches were able to identify feasible reaction pathways efficiently while reducing the need for exhaustive graph traversal. These embeddings effectively reduce the dimensionality of the underlying PES while enabling more efficient search and sampling techniques.
- **Chapter 4** introduced the generation of bounded RNs guided by on-the-fly energy calculations and step limits relative to a starting node. These constraints reduced the generation of chemically infeasible and irrelevant pathways and enabled the construction of smaller, chemically meaningful RNs. This chapter also incorporated radical reaction rules and applied the workflow to atmospheric chemistry systems, demonstrating that bounded generation substantially reduces network size while retaining chemically relevant reaction pathways.

5.1 Future Work

Several avenues of future work naturally follow from the results of this dissertation.

For the work in Chapter 2, the observed lattice-like structure of stoichiometry-preserving reaction networks (RNs) suggests opportunities to further leverage their underlying regular structure. Since these networks can be modeled as low-dimensional regular lattices with a small amount of random edge rewiring, future work may exploit this structure to accelerate RN construction, sampling, and search. Another possible direction involves generating smaller subnetworks, each corresponding to an individual species or localized region of chemical space, which could later be assembled to construct larger RNs. Machine learning approaches that leverage the known structural properties of these networks may help predict how such subnetworks connect, enabling assembly of larger RNs that retain the geometric and topological characteristics of fully generated networks.

The work in Chapter 3 demonstrated that Euclidean embeddings of weighted stoichiometry-preserving RNs can preserve chemically meaningful pathways while reducing the need for exhaustive graph traversal. These results open several additional research directions. One important area is the co-development of chemically informed edge-weighting schemes together with embedding methods. Improved weighting functions that better distinguish chemically feasible from infeasible transformations would produce more faithful embeddings and more reliable embedded-space searches. Additionally, the partially ordered structure of stoichiometry-preserving RNs suggests that embeddings trained on one RN may transfer to related networks with different stoichiometries. This raises the possibility of constructing transferable embeddings across chemical space, reducing the need to train new embeddings for every new RN.

Several extensions are possible for the bounded RNs developed in Chapter 4. One promising direction is to couple bounded RNs with Euclidean embeddings to enable efficient pathway

searches directly within compact, energy-bounded networks. Another is to incorporate a Monte Carlo-like expansion criterion, where only selected edges based on an energy-based probability function are expanded. This would further reduce the size of a bounded RN while retaining chemically relevant pathways. Additional work could involve analyzing all molecules within a bounded RN to identify species of interest and determine chemically feasible reaction pathways connecting them. Generating bounded RNs under different chemical environments, such as varying solvents or including additional reacting species, may also provide insight into how reaction mechanisms change and whether alternative pathways become accessible in different molecular environments.

More generally, future work should continue expanding the application of `colibri2` and stoichiometry-preserving RNs to broader classes of chemical systems. Additional studies in combustion, atmospheric chemistry, catalysis, and prebiotic chemistry may uncover previously unknown reaction pathways and mechanisms while further evaluating the scalability and transferability of RNs. Further integration of reaction-network generation, network embeddings, and automated pathway analysis may enable increasingly autonomous exploration of chemical reactivity and prediction of novel molecules and reaction pathways. A complementary direction is the integration of large language models (LLMs) into the `colibri2` workflow. As a postprocessing step, an LLM could evaluate pathways extracted from exhaustive or bounded RNs and assess their chemical feasibility, helping distinguish mechanistically reasonable pathways from those that only satisfy energetic or distance-based criteria but are unlikely to be chemically realistic. LLMs could also analyze generated RNs to identify chemically feasible and potentially novel pathways. As RNs may contain millions or billions of nodes and an enormous number of possible pathways, manual exploration is often impractical. An LLM could automate this process by highlighting promising pathways for further computational or experimental study. More broadly, an LLM could directly interact with a RN generating hypotheses, proposing additional bounded RN generations or pathway searches, and analyzing the resulting data.

5.2 Concluding Remarks

The long-term goal of this research is to enable the discovery of new chemical reactions, mechanisms, and molecules through scalable exploration of chemical space. Achieving this goal requires methods that can systematically explore large regions of PESs while remaining computationally tractable and chemically meaningful. This dissertation demonstrated that stoichiometry-preserving reaction networks implemented in `colibri2` provide a framework for achieving this balance. The work presented throughout this dissertation, including the structural characterization of reaction networks, Euclidean embedding approaches for pathway discovery, and energetically informed bounded reaction network generation, represents progress toward more efficient and scalable exploration of PESs. Importantly, stoichiometry-preserving reaction networks are capable of broad and relatively unbiased exploration of chemical reaction space without relying on prior knowledge of specific reaction mechanisms or pathways. Instead, reaction networks are generated through reversible rules composed of elementary bond-breaking and bond-forming transformations while still allowing energetic constraints to focus exploration on chemically relevant regions of the PES when desired. These contributions not only address key limitations in current reaction network methodologies, but also provide an adaptable framework with `colibri2` for future studies in reaction discovery, pathway prediction, and exploration of chemical reactivity across diverse chemistry subdisciplines.

Bibliography

- [1] Miko M. Stulajter and Dmitrij Rappoport. Reaction networks resemble low-dimensional regular lattices. *J. Chem. Theory Comput.*, 20(18):8035–8048, 2024.
- [2] Oleg N Temkin, Andrew V Zeigarnik, and Danail Bonchev. *Chemical Reaction Networks: A Graph-Theoretical Approach*. CRC Press, Boca Raton, 1996.
- [3] Martin Feinberg. *Foundations of Chemical Reaction Network Theory*. Springer, Cham, 2019.
- [4] Mingjian Wen, Evan Walter Clark Spotte-Smith, Samuel M Blau, Matthew J McDermott, Aditi S Krishnapriyan, and Kristin A Persson. Chemical reaction networks and opportunities for machine learning. *Nature Comput. Sci.*, pages 1–13, 2023.
- [5] R. Vinu and Linda J. Broadbelt. Unraveling reaction pathways and specifying reaction kinetics for complex systems. *Annu. Rev. Chem. Biomol. Eng.*, 3:29–54, 2012.
- [6] Amanda L. Dewyer, Alonso J. Argüelles, and Paul M. Zimmerman. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.*, 8(2):e1354, 2018.
- [7] Jan P Unsleber and Markus Reiher. The exploration of chemical reaction networks. *Annu. Rev. Phys. Chem.*, 71(1):121–142, 2020.
- [8] Bartosz A. Grzybowski, Kyle J. M. Bishop, Bartłomiej Kowalczyk, and Christopher E. Wilmer. The ‘wired’ universe of organic chemistry. *Nat. Chem.*, 1(1):31–36, 2009.
- [9] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [10] Vassily Hatzimanikatis, Chunhui Li, Justin A Ionita, Christopher S Henry, Matthew D Jankowski, and Linda J Broadbelt. Exploring the diversity of complex metabolic networks. *Bioinformatics*, 21(8):1603–1609, 2005.
- [11] Natalie C. Duarte, Scott A. Becker, Neema Jamshidi, Ines Thiele, Monica L. Mo, Thuy D. Vo, Rohith Srivas, and Bernhard Ø. Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Nat. Acad. Sci.*, 104(6):1777–1782, 2007.

- [12] Raul Montañez, Miguel Angel Medina, Ricard V. Solé, and Carlos Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *BioEssays*, 32(3):246–256, 2010.
- [13] Connie W Gao, Joshua W Allen, William H Green, and Richard H West. Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.*, 203:212–225, 2016.
- [14] Sergio Vernuccio and Linda J. Broadbelt. Discerning complex reaction networks using automated generators. *AIChE J.*, 65(8):e16663, 2019.
- [15] John D. Chodera, Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope. Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126(15):155101, 2007.
- [16] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.
- [17] Luc Vereecken and Joseph S. Francisco. Theoretical studies of atmospheric reaction mechanisms in the troposphere. *Chem. Soc. Rev.*, 41(19):6259–6293, 2012.
- [18] Sara Imari Walker and Cole Mathis. Network theory in prebiotic evolution. In César Menor-Salvá'n, editor, *Prebiotic Chemistry and Chemical Evolution of Nucleic Acids*, pages 263–291. Springer, Cham, 2018.
- [19] Blair A Welsh, Maggie E Corrigan, Emmanuel Assaf, Klaas Nauta, Paolo Sebastianelli, Meredith JT Jordan, Christa Fittschen, and Scott H Kable. Photophysical oxidation of hcho produces ho2 radicals. *Nature Chemistry*, 15(10):1350–1357, 2023.
- [20] Michael J Prather and Lei Zhu. Resetting tropospheric oh and ch4 lifetime with ultraviolet h2o absorption. *Science*, 385(6705):201–204, 2024.
- [21] Cangtao Yin and Gábor Czakó. Revealing new pathways for the reaction of criegee intermediate ch2oo with so2. *Communications Chemistry*, 7(1):157, 2024.
- [22] Dmitriy Rappoport, Cooper J. Galvin, Dmitry Yu. Zubarev, and Alán Aspuru-Guzik. Complex chemical reaction networks from heuristics-aided quantum chemistry. *J. Chem. Theory Comput.*, 10(3):897–907, 2014.
- [23] Dmitriy Rappoport. Reaction networks and the metric structure of chemical space(s). *J. Phys. Chem. A*, 123(13):2610–2620, 2019.
- [24] Dmitriy Rappoport and Alán Aspuru-Guzik. Predicting feasible organic reaction pathways using heuristically aided quantum chemistry. *J. Chem. Theory Comput.*, 15(7):4099–4112, 2019.
- [25] Linda J. Broadbelt, Scott M. Stark, and Michael T. Klein. Computer generated pyrolysis modeling: On-the-fly generation of species, reactions, and rates. *Ind. Eng. Chem. Res.*, 33(4):790–799, 1994.

- [26] Stacey D Finley, Linda J Broadbelt, and Vassily Hatzimanikatis. Computational framework for predictive biodegradation. *Biotechnol. Bioeng.*, 104(6):1086–1097, 2009.
- [27] Noushin Hadadi, Jasmin Hafner, Adrian Shajkofci, Aikaterini Zisaki, and Vassily Hatzimanikatis. ATLAS of biochemistry: A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS Synth. Biol.*, 5(10):1155–1166, 2016.
- [28] Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski. Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem. Int. Ed.*, 55(20):5904–5937, 05 2016.
- [29] Jakob L Andersen, Rolf Fagerberg, Christoph Flamm, Walter Fontana, Jurai Kolčák, Christophe V F P Laurent, Daniel Merkle, and Nikolai Nøjgaard. Graph transformation for enzymatic mechanisms. *Bioinformatics*, 37(Suppl_1):i392–i400, 2021.
- [30] Mengjie Liu, Alon Grinberg Dana, Matthew S. Johnson, Mark J. Goldman, Agnes Jocher, A. Mark Payne, Colin A. Grambow, Kehang Han, Nathan W. Yee, Emily J. Mazeau, Katrin Blondal, Richard H. West, C. Franklin Goldsmith, and William H. Green. Reaction mechanism generator v3.0: Advances in automatic mechanism generation. *J. Chem. Inf. Model.*, 61(6):2686–2696, 2021.
- [31] Gregor N. Simm, Alain C. Vaucher, and Markus Reiher. Exploration of reaction pathways and chemical transformation networks. *J. Phys. Chem. A*, 123(2):385–399, 2019.
- [32] Jan P. Unsleber, Stephanie A. Grimm, and Markus Reiher. Chemoton 2.0: Autonomous exploration of chemical reaction networks. *J. Chem. Theory Comput.*, 18(9):5393–5409, 2022.
- [33] Satoshi Maeda, Koichi Ohno, and Keiji Morokuma. Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (grrm) strategy using the addf and afir methods. *Phys. Chem. Chem. Phys.*, 15:3683–3701, 2013.
- [34] Koichi Ohno and Satoshi Maeda. A scaled hypersphere search method for the topography of reaction pathways on the potential energy surface. *Chemical Physics Letters*, 384:277–282, 2004.
- [35] Paul M. Zimmerman. Automated discovery of chemically reasonable elementary reaction steps. *Journal of Computational Chemistry*, 34(16):1385–1392, 2013.
- [36] W. M. C. Sameera, Satoshi Maeda, and Keiji Morokuma. Computational catalysis using the artificial force induced reaction method. *Acc. Chem. Res.*, 49(4):763–773, 2016.
- [37] Satoshi Maeda, Yu Harabuchi, and Kenichiro Saita. Global reaction route mapping strategy: A tool for finding new chemistry in computers. In Hisashi Yamamoto

- and Takashi Kato, editors, *Molecular Technology: Life Innovation*, volume 1, pages 173–199. Wiley-VCH, Weinheim, 2018.
- [38] Satoshi Maeda and Yu Harabuchi. Exploring paths of chemical transformations in molecular and periodic systems: An approach utilizing force. *WIREs Comput. Mol. Sci.*, 11(6):e1538, 2021.
 - [39] Scott Habershon. Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling. *J. Chem. Theory Comput.*, 12(4):1786–1798, 2016.
 - [40] Idil Ismail, Holly B. V. A. Stuttford-Fowler, Curtis Ochan Ashok, Christopher Robertson, and Scott Habershon. Automatic proposal of multistep reaction mechanisms using a graph-driven search. *J. Phys. Chem. A*, 123(15):3407–3417, 2019.
 - [41] Christopher Robertson and Scott Habershon. Fast screening of homogeneous catalysis mechanisms using graph-driven searches and approximate quantum chemistry. *Catal. Sci. Technol.*, 9(22):6357–6369, 2019.
 - [42] Alessandro Laio and Francesco L Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Progr. Phys.*, 71(12):126601, 2008.
 - [43] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *WIREs Comput. Mol. Sci.*, 1(5):826–843, 2011.
 - [44] Miguel Steiner and Markus Reiher. A human-machine interface for automatic exploration of chemical reaction networks. *Nat. Commun.*, 15:3680, 2024.
 - [45] Lee-Ping Wang, Alexey Titov, Robert McGibbon, Fang Liu, Vijay S. Pande, and Todd J. Martínez. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.*, 6(12):1044–1048, 2014.
 - [46] Lee-Ping Wang, Robert T. McGibbon, Vijay S. Pande, and Todd J. Martinez. Automated discovery and refinement of reactive molecular dynamics pathways. *J. Chem. Theory Comput.*, 12(2):638–649, 2016.
 - [47] Jason Ford, Stefan Seritan, Xiaolei Zhu, Michael N. Sakano, Md Mahbub Islam, Alejandro Strachan, and Todd J. Martínez. Nitromethane decomposition via automated reaction discovery and an ab initio corrected kinetic model. *J. Phys. Chem. A*, 125(7):1447–1460, 2021.
 - [48] Rui Xu, Jan Meisner, Alexander M. Chang, Keiran C. Thompson, and Todd J. Martínez. First principles reaction discovery: from the schrodinger equation to experimental prediction for methane pyrolysis. *Chem. Sci.*, 14(27):7447–7464, 2023.
 - [49] Qiyuan Zhao and Brett M Savoie. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. *Nat. Comput. Sci.*, 1(7):479–490, 2021.

- [50] Qiyuan Zhao, Yinan Xu, Jeffrey Greeley, and Brett M. Savoie. Deep reaction network exploration at a heterogeneous catalytic interface. *Nat. Commun.*, 13(1):4860, 2022.
- [51] Qiyuan Zhao and Brett M. Savoie. Algorithmic explorations of unimolecular and bimolecular reaction spaces. *Angew. Chem. Int. Ed.*, 61(46):e202210693, 2022.
- [52] Qiyuan Zhao and Brett M. Savoie. Deep reaction network exploration of glucose pyrolysis. *Proc. Nat. Acad. Sci.*, 120(34):e2305884120, 2023.
- [53] Michael Woulfe and Brett M. Savoie. Chemical reaction networks from scratch with reaction prediction and kinetics-guided exploration. *J. Chem. Theory Comput.*, 21(3):1276–1291, 2025.
- [54] Philippe Schwaller, Alain C. Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. Machine intelligence for chemical reaction space. *WIREs Comput. Mol. Sci.*, 12(5):e1604, 2022.
- [55] Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Deep learning of activation energies. *J. Phys. Chem. Lett.*, 11(8):2992–2997, 2020.
- [56] Jan H. Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chem. Sci.*, 10(12):3567–3572, 2019.
- [57] Ruud van Deursen and Jean-Louis Reymond. Chemical space travel. *ChemMedChem*, 2(5):636–640, 2007.
- [58] Kentaro Kawai, Naoya Nagata, and Yoshimasa Takahashi. De novo design of drug-like molecules by a fragment-based molecular evolutionary approach. *J. Chem. Inf. Model.*, 54(1):49–56, 2014.
- [59] David Hoksza, Petr Škoda, Milan Voršilák, and Daniel Svozil. Molpher: a software framework for systematic chemical space exploration. *J. Cheminf.*, 6(1):7, 2014.
- [60] Markus Meuwly. Machine learning for chemical reactions. *Chemical Reviews*, 121(16):10218–10239, 2021.
- [61] Samuel M Blau, Hetal D Patel, Evan Walter Clark Spotte-Smith, Xiaowei Xie, Shyam Dwaraknath, and Kristin A Persson. A chemically consistent graph architecture for massive reaction networks applied to solid-electrolyte interphase formation. *Chem. Sci.*, 12(13):4931–4939, 2021.
- [62] Paul L. Türtcher and Markus Reiher. Pathfinder—navigating and analyzing chemical reaction networks with an efficient graph-based approach. *Journal of Chemical Information and Modeling*, 63(1):147–160, 01 2023.
- [63] Michel Marie Deza and Elena Deza. Distances in graph theory. In *Encyclopedia of Distances*, pages 249–278. Springer, Berlin Heidelberg, 2013.
- [64] James Dugundji and Ivar Ugi. An algebraic model of constitutional chemistry as a basis for chemical computer programs. *Top. Curr. Chem.*, 39:19–64, 1973.

- [65] Vladimír Kvasnička, Jiří Pospíchal, and Vladimír Baláž. Reaction and chemical distances and reaction graphs. *Theor. Chim. Acta*, 79(1):65–79, 1991.
- [66] Robert B Grossman. *The Art of Writing Reasonable Organic Reaction Mechanisms*. Springer, Cham, 3rd edition, 2019.
- [67] Daniel E Levy. *Arrow-Pushing in Organic Chemistry. An Easy Approach to Understanding Reaction Mechanisms*. Wiley, Hoboken NJ, 2008.
- [68] D Rappoport and M Stulajter. colibri2 is a distributed computational engine for chemical exploration. (accessed 2024-02-26).
- [69] Mícheál Ó Searcóid. *Metric spaces*. Springer, 2007 edition, 2006.
- [70] Irving Kaplansky. *Set theory and metric spaces*, volume 298. American Mathematical Society, 2020.
- [71] George S Hammond. A correlation of reaction rates. *Journal of the American Chemical Society*, 77(2):334–338, 1955.
- [72] Redis Ltd. Redis, 2024. [Software].
- [73] The PostgreSQL Global Development Group. Postgresql, 2024. [Software].
- [74] Tom Preston-Werner. TOML: Tom’s obvious, minimal language, 2013. [Software].
- [75] D. Weininger. Smarts – a language for describing molecular patterns. (accessed 2025-05-21).
- [76] Ulrik Brandes, Markus Eiglsperger, Ivan Herman, Michael Himsolt, and M Scott Marshall. Graphml progress report. structural layer proposal. In P. Mutzel, M. Jünger, and S. Leipert, editors, *Graph Drawing 2001*, pages 501–512, Berlin Heidelberg, 2002. Springer.
- [77] Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, gedec, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, guillaume godin, Axel Pahl, Juuso Lehtivarjo, Francois Berenger, jasondbiggs, and strets123. Rdkit: Open-source cheminformatics software; ver. 2021.09.2. (accessed 2024-02-26).
- [78] Andreas Wagner and David A. Fell. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B. Biol. Sci.*, 268(1478):1803–1810, 2001.
- [79] Chaoming Song, Shlomo Havlin, and Hernan A Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.

- [80] Chaoming Song, Lazaros K Gallos, Shlomo Havlin, and Hernán A Makse. How to calculate the fractal dimension of a complex network: the box covering algorithm. *J. Stat. Mech.*, 2007(03):P03006, 2007.
- [81] Lazaros K. Gallos, Chaoming Song, and Hernán A. Makse. A review of fractality and self-similarity in complex networks. *Physica A*, 386(2):686–691, 2007.
- [82] K L Clarkson. Nearest-neighbor searching and metric space dimensions. In Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk, editors, *Nearest-Neighbor Methods in Learning and Vision*, pages 15–59. MIT Press, Cambridge MA, 2005.
- [83] A Andoni and P Indyk. Nearest neighbors in high-dimensional spaces. In Jacob E. Goodman, Joseph O’Rourke, and Csaba D. Tóth, editors, *Handbook of Discrete and Computational Geometry*, chapter 43, pages 1135–1155. CRC Press, Boca Raton FL, 3rd edition, 2018.
- [84] Dmitrij Rappoport. Statistics and bias-free sampling of reaction mechanisms from reaction network models. *J. Phys. Chem. A*, 127(24):5252–5263, 2023.
- [85] Paul Erdős and Alfréd Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.
- [86] Duncan J Watts and Steven H Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442, 1998.
- [87] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 1st edition, 2010.
- [88] Reuven Cohen and Shlomo Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, Cambridge, 2010.
- [89] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, 2011.
- [90] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [91] Guido Caldarelli. *Scale-Free Networks*. Oxford University Press, Oxford, 2007.
- [92] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In Frances Yao and Eugene Luks, editors, *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163—170. ACM, 2000.
- [93] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Cambridge University Press, Cambridge, 2008.
- [94] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Phys. Rev. E*, 64(4):046135, 2001.

- [95] L A Adamic, R M Lukose, and B A Huberman. Local search in unstructured networks. In Stefan Bornholdt and Heinz Georg Schuster, editors, *Handbook of Graphs and Networks*, chapter 13, pages 295–317. Wiley–VCH, Weinheim, 2003.
- [96] Alessandro P. S. de Moura, Adilson E. Motter, and Celso Grebogi. Searching in small-world networks. *Phys. Rev. E*, 68(3):036106, 2003.
- [97] Tongfeng Weng, Michael Small, Jie Zhang, and Pan Hui. Lévy walk navigation in complex networks: A distinct relation between optimal transport exponent and network dimension. *Sci. Rep.*, 5:17309, 2015.
- [98] Eric Rosenberg. *Fractal Dimensions of Networks*. Springer, Cham, 2020.
- [99] Philipp Pracht and Stefan Grimme. Calculation of absolute molecular entropies and heat capacities made simple. *Chem. Sci.*, 12(19):6551–6568, 2021.
- [100] Edwin A Hill. On a system of indexing chemical literature; adopted by the classification division of the us patent office. *J. Am. Chem. Soc.*, 22(8):478–494, 1900.
- [101] Christian L Staudt, Aleksejs Sazonovs, and Henning Meyerhenke. Networkit: A tool suite for large-scale complex network analysis. *Netw. Sci.*, 4(4):508–530, 2016.
- [102] Pedro G Lind, Marta C Gonzalez, and Hans J Herrmann. Cycles and clustering in bipartite networks. *Phys. Rev. E*, 72(5):056127, 2005.
- [103] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A*, 387(27):6869–6875, 2008.
- [104] Peter Tamas Kovacs, Marcell Nagy, and Roland Molontay. Comparative analysis of box-covering algorithms for fractal networks. *Appl. Netw. Sci.*, 6(1):73, 2021.
- [105] Tao Wen and Kang Hao Cheong. The fractal dimension of complex networks: A review. *Inf. Fusion*, 73:87–102, 2021.
- [106] Jin Seop Kim, Kwang-Il Goh, Byungnam Kahng, and Doochul Kim. Fractality and self-similarity in scale-free networks. *New J. Phys.*, 9(6):177, 2007.
- [107] Vladimir Batagelj and Ulrik Brandes. Efficient generation of large random networks. *Phys. Rev. E*, 71(3):036113, 2005.
- [108] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena CA, 2008. ACM Press.
- [109] M Stulajter. Reaction networks resemble low-dimensional regular lattices : Generation and analysis code. (accessed 2024-05-07).

- [110] Carlos M. Jarque and Anil K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259, 1980.
- [111] Mark EJ Newman. Power laws, pareto distributions and zipf’s law. *Contemporary Physics*, 46(5):323–351, 2005.
- [112] Reuven Cohen and Shlomo Havlin. Fractal dimensions of percolating networks. *Physica A*, 336(1–2):6–13, 2004.
- [113] Jia Shao, Sergey V Buldyrev, Reuven Cohen, Maksim Kitsak, Shlomo Havlin, and H Eugene Stanley. Fractal boundaries of complex networks. *EPL*, 84(4):48004, 2008.
- [114] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97, 2002.
- [115] Agata Fronczak, Piotr Fronczak, and Janusz A. Hołyst. Average path length in random networks. *Phys. Rev. E*, 70(5):056110, 2004.
- [116] Jayanta K Rudra and John J Kozak. Spectral dimension of regular and fractal lattices. *Phys. Lett. A*, 151(8):429–432, 1990.
- [117] JS Kim, K-I Goh, B Kahng, and D Kim. A box-covering algorithm for fractal scaling in scale-free networks. *Chaos*, 17(2), 2007.
- [118] Guo Long and Xu Cai. The fractal dimensions of complex networks. *Chin. Phys. Lett.*, 26(8):088901, 2009.
- [119] Mark EJ Newman and Duncan J Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E*, 60(6):7332–7342, 1999.
- [120] A Lochmann and M Requardt. An analysis of the transition zone between the various scaling regimes in the small-world model. *J. Stat. Phys.*, 122(2):255–278, 2006.
- [121] XS Yang. Fractals in small-world networks with time-delay. *Chaos, Solitons Fractals*, 13(2):215–219, 2002.
- [122] Pablo Pavon-Dominguez and Soledad Moreno-Pulido. Sandbox fixed-mass algorithm for multifractal unweighted complex networks. *Chaos, Solitons Fractals*, 156:111836, 2022.
- [123] O Shanker. Defining dimension of a complex network. *Mod. Phys. Lett. B*, 21(06):321–326, 2007.
- [124] KH Chang, BC Choi, Seong-Min Yoon, and Kyungsik Kim. Multifractal measures on small-world networks. *Fractals*, 14(02):119–123, 2006.
- [125] Benjamin F. Maier. Generalization of the small-world effect on a model approaching the erdős–rényi random graph. *Sci. Rep.*, 9:9268, 2019.

- [126] Eytan Katzav, Ofer Biham, and Alexander K. Hartmann. Metric properties of subcritical erdős-rényi networks. *Phys. Rev. E*, 98(1):012301, 2018.
- [127] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. Network representation learning: A survey. *IEEE Trans. Big Data*, 6(1):3–28, 2020.
- [128] Oren M. Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of Chemical Physics*, 106(4):1495–1517, 1997.
- [129] Burai Murayama, Masato Kobayashi, Masamitsu Aoki, Suguru Ishibashi, Takuya Saito, Takenobu Nakamura, Hiroshi Teramoto, and Tetsuya Taketsugu. Characterizing reaction route map of realistic molecular reactions based on weight rank clique filtration of persistent homology. *Journal of Chemical Theory and Computation*, 19(15):5007–5023, 2023.
- [130] Hai-Cheng Yi, Zhu-Hong You, De-Shuang Huang, and Chee Keong Kwoh. Graph representation learning in bioinformatics: trends, methods and applications. *Brief. Bioinform.*, 23(1):bbab340, 2022.
- [131] Alejandro F Rozenfeld, Reuven Cohen, Daniel Ben-Avraham, and Shlomo Havlin. Scale-free networks on lattices. *Phys. Rev. Lett.*, 89(21):218701, 2002.
- [132] Bianka Kovács and Gergely Palla. Model-independent embedding of directed networks into euclidean and hyperbolic spaces. *Commun. Phys.*, 6(1):28, 2023.
- [133] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24240–24253, Red Hook NY, 2022. Curran Associates.
- [134] Alexandru T Balaban. Reaction graphs. In Danail Bonchev and Ovanes Mekenyan, editors, *Graph Theoretical Approaches to Chemical Reactivity*, pages 137–180. Kluwer, Dordrecht, 1994.
- [135] Bernhard Ø Palsson. *Systems Biology. Constraint-based Reconstruction and Analysis*. Cambridge University Press, Cambridge, 2015.
- [136] Guy B Marin, Gregory S Yablonsky, and Denis Constales. *Kinetics of Chemical Reactions. Decoding Complexity*. Wiley-VCH, Weinheim, 2nd edition, 2019.
- [137] Oliver Ebenhöh and Reinhart Heinrich. Evolutionary optimization of metabolic pathways. theoretical reconstruction of the stoichiometry of ATP and NADH producing systems. *Bull. Math. Biol.*, 63(1):21–55, 2001.
- [138] E Ravasz, A L Somera, D A Mongru, Z N Oltvai, and A L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

- [139] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numer. Math.*, 1(1):269–271, 1959.
- [140] Anne-Susann Abel, Nino Lauber, Jakob Lykke Andersen, Rolf Fagerberg, Daniel Merkle, and Christoph Flamm. Computational approaches in chemical space exploration for carbon fixation pathways. *NPJ Syst. Biol. Appl.*, 12(1):17, 2026.
- [141] Adittya Pal, Rolf Fagerberg, Jakob Lykke Andersen, Peter Dittrich, and Daniel Merkle. Finding pathways in reaction networks guided by energy barriers using integer linear programing. *Mol. Inform.*, 45(3):e70021, 2026.
- [142] HongYun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 30(9):1616–1637, 2018.
- [143] Mengjia Xu. Understanding graph embedding methods and their applications. *SIAM Rev.*, 63(4):825–853, 2021.
- [144] Warren S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.
- [145] J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.
- [146] Ingwer Borg and Patrick JF Groenen. *Modern Multidimensional Scaling*. Springer, New York, 2nd edition, 2005.
- [147] Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004. Technical Report.
- [148] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.
- [149] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- [150] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.
- [151] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In Wolfram Gerstner, Alain Germond, Martin Hasler, and Jean-Daniel Nicoud, editors, *Artificial Neural Networks — ICANN’97*, pages 583–588, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg.
- [152] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
- [153] I. Borg and P. Groenen. Modern multidimensional scaling: Theory and applications. *J. Educ. Meas.*, 40(3):277–280, 2003.

- [154] James J. P. Stewart. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.*, 19(1):1–32, 2013.
- [155] Jonathan E. Moussa and James J. P. Stewart. MOPAC; ver. 23.2.2. <https://doi.org/10.5281/zenodo.17470335> (accessed 2026-05-15).
- [156] Ronald Breslow. On the mechanism of the formose reaction. *Tetrahedron Lett.*, 1(22):22–26, 1959.
- [157] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15, Pasadena, CA USA, 2008.
- [158] Miko M. Stulajter. Euclidean embedding of reaction networks for dimensionality reduction of chemical spaces: Scripts and reaction network files. (accessed 2026-04-29).
- [159] Rishi Sonthalia, Greg Van Buskirk, Benjamin Raichel, and Anna Gilbert. How can classical multidimensional scaling go wrong? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12304–12315. Curran Associates, Inc., 2021.
- [160] Michiel van Wezel, Joost Kok, and Walter Kusters. Two neural network methods for multidimensional scaling. In *ESANN '97 : European symposium on artificial neural networks*, pages 97–102, Brussels, 1997. D Facto.
- [161] Luís Garrido, Sergio Gómez, and Jaume Roca. Improved multidimensional scaling analysis using neural networks with distance-error backpropagation. *Neural Comput.*, 11(3):595–600, 1999.
- [162] Gautam Pai, Ronen Talmon, Alex Bronstein, and Ron Kimmel. Dimal: Deep isometric manifold learning using sparse geodesic sampling. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 819–828, 2019.
- [163] Gautam Pai, Alex Bronstein, Ronen Talmon, and Ron Kimmel. Deep isometric maps. *Image Vision Comput.*, 123:104461, 2022.
- [164] Patrick J. F. Groenen and Jan de Leeuw. The majorization approach to multidimensional scaling for minkowski distances. *J. Classif.*, 10:73–93, 1993.
- [165] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37, ICML'15*, pages 448–456. JMLR.org, 2015.
- [166] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, pages 807–814, Madison, WI, 2010. Omnipress.

- [167] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.
- [168] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. *Neural Networks: Tricks of the Trade*, pages 9–50, 1998.
- [169] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
- [170] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations (ICLR’17)*, 2017.
- [171] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1310–1318, Atlanta, GA, 2013. PMLR.
- [172] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.
- [173] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019.
- [174] Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Proceedings of the 14th International Conference on Neural Information Processing Systems*, pages 661–667, Cambridge, MA, 2000. MIT Press.
- [175] John Platt. Fastmap, metricmap, and landmark mds are all nyström algorithms. In Robert G Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 261–268. PMLR, 2005.
- [176] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the nystrom method. In David van Dyk and Max Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 304–311, Clearwater Beach, FL, 2009. PMLR.
- [177] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

- [178] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.
- [179] R. B. Woodward and R. Hoffmann. Stereochemistry of electrocyclic reactions. *J. Am. Chem. Soc.*, 87:395–397, 1965.
- [180] R. B. Woodward and R. Hoffmann. *The Conservation of Orbital Symmetry*. Academic Press, New York, 1970.
- [181] Brian Capon, Bao-Zhen Guo, Fanny C. Kwok, Asit K. Siddhanta, and Cesare Zucco. Generation of vinyl alcohol in solution and its slow conversion into acetaldehyde. *J. Chem. Soc., Chem. Commun.*, pages 724–725, 1979.
- [182] Brian Capon and Cesare Zucco. Simple enols. 1. Vinyl alcohol. *J. Am. Chem. Soc.*, 104:7567–7568, 1982.
- [183] Y. Chiang, A. J. Kresge, Y. S. Tang, and J. Wirz. The pK_a and rate constants for ketonization of vinyl alcohol in aqueous solution by flash photolysis protonation. *J. Am. Chem. Soc.*, 106:460–462, 1984.
- [184] Ch. Grundmann. Syntheses with *s*-triazine. *Angew. Chem. Int. Ed.*, 2(6):309–323, 1963.
- [185] Edwin M. Smolin and Lorence Rapoport. *s-Triazines and Derivatives*. Interscience Publishers, New York, 1959.
- [186] D. Martin, M. Bauer, and V. A. Pankratov. Cyclotrimerisation of cyano-compounds into 1,3,5-triazines. *Russ. Chem. Rev.*, 43(10):975–990, 1978.
- [187] C. K. Ingold. *Structure and Mechanism in Organic Chemistry*. Cornell University Press, Ithaca, NY, 2nd edition, 1969.
- [188] Michael B. Smith and Jerry March. *March’s Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*. Wiley, Hoboken, NJ, 6 edition, 2007.
- [189] Arthur Omran, Cesar Menor-Salván, Greg Springsteen, and Matthew Pasek. The messy alkaline formose reaction and its link to metabolism. *Life*, 10(8):125, 2020.
- [190] Scot M. Sutton, Sunil Pulletikurti, Huacan Lin, Ramanarayanan Krishnamurthy, and Charles L. Liotta. Abiotic aldol reactions of formaldehyde with ketoses and aldoses—implications for the prebiotic synthesis of sugars by the formose reaction. *Chem*, 2025.
- [191] Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic Chemistry*. Oxford University Press, Oxford, 2nd edition, 2012.
- [192] Francis A. Carey and Richard J. Sundberg. *Advanced Organic Chemistry, Part A: Structure and Mechanisms*. Springer, New York, 5th edition, 2007.
- [193] Jean Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Isr. J. Math.*, 52:46–52, 1985.

- [194] Hassler Whitney. The self-intersections of a smooth n -manifold in $2n$ -space. *Ann. Math.*, 45(2):220–246, 1944.
- [195] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, 1981.
- [196] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. *J. Stat. Phys.*, 65(3–4):579–616, 1991.
- [197] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. U.S.A.*, 108:13023–13029, 2011.
- [198] David S. Watkins. *The Matrix Eigenvalue Problem. GR and Krylov Subspace Methods*. SIAM, Philadelphia, 2007.
- [199] Yousef Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, 2011.
- [200] Dmitrij Rappoport, Samuel Bekoe, Luke Nambi Mohanam, Scott Le, Naje’ George, Ziyue Shen, and Filipp Furche. Libkrylov: A modular open-source software library for extremely large on-the-fly matrix computations. *J. Comput. Chem.*, 44(11):1105–1118, 2023.
- [201] Laurens van der Maaten, Eric Postma, and Jaap van den Herik. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.*, 10:66–71, 2009.
- [202] Emilio Martínez-Núñez, George L. Barnes, David R. Glowacki, Stanisław Kopec, Daniel Peláez, Aurelio Rodríguez, Roberto Rodríguez-Fernández, Robin J. Shannon, James J. P. Stewart, Pablo G. Tahoces, and Robert S. Tranter. AutoMeKin2021: An open-source program for automated reaction discovery. *J. Comput. Chem.*, 42:2036–2048, 2021.
- [203] Weiqi Ji and Sili Deng. Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network. *J. Phys. Chem. A*, 125(4):1082–1092, 2021.
- [204] Pierre-François Verhulst. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique*, 10:113–121, 1838.
- [205] A. Tsoularis and J. Wallace. Analysis of logistic growth models. *Math. Biosci.*, 179(1):21–55, 2002.
- [206] Gregor N. Simm and Markus Reiher. Context-driven exploration of complex chemical reaction networks. *J. Chem. Theory Comput.*, 13:6108–6119, 2017.
- [207] Satoshi Maeda, Tetsuya Taketsugu, Koichi Ohno, and Keiji Morokuma. From roaming atoms to hopping surfaces: Mapping out global reaction routes in photochemistry. *J. Am. Chem. Soc.*, 137:3433–3445, 2015.

- [208] P.D Lightfoot, R.A Cox, J.N Crowley, M Destriau, G.D Hayman, M.E Jenkin, G.K Moortgat, and F Zabel. Organic peroxy radicals: Kinetics, spectroscopy and tropospheric chemistry. *Atmos. Environ.*, 26(10):1805–1961, 1992.
- [209] George Odian. *Principles of Polymerization*. Wiley-Interscience, Hoboken, NJ, 4th edition, 2004.
- [210] Richard G. Compton and Craig E. Banks. *Understanding Voltammetry*. Imperial College Press, London, 3rd edition, 2018.
- [211] J. W. T. Spinks and R. J. Woods. *An Introduction to Radiation Chemistry*. Wiley-Interscience, New York, 3rd edition, 1990.
- [212] Toren Finkel and Nikki J. Holbrook. Oxidants, oxidative stress and the biology of ageing. *Nature*, 408:239–247, 2000.
- [213] L. Vereecken, B. Aumont, I. Barnes, J.W. Bozzelli, M.J. Goldman, W.H. Green, S. Madronich, M.R. Mcgillen, A. Mellouki, J.J. Orlando, B. Picquet-Varraut, A.R. Rickard, W.R. Stockwell, T.J. Wallington, and W.P.L. Carter. Perspective on mechanism development and structure-activity relationships for gas-phase atmospheric chemistry. *Int. J. Chem. Kinet.*, 50:435–469, 2018.
- [214] Roger Atkinson and Janet Arey. Atmospheric degradation of volatile organic compounds. *Chem. Rev.*, 103(12):4605–4638, 2003.
- [215] N. N. Semenov. *Chemical Kinetics and Chain Reactions*. Oxford University Press, Oxford, 1935.
- [216] Jeremy N. Harvey. Spin-forbidden reactions: computational insight into mechanisms and kinetics. *WIREs Comput. Mol. Sci.*, 4:1–14, 2013.
- [217] ChemAxon. Chemaxon extended smiles and smarts (cxsmiles and cxsmarts), 2024. Accessed: 2026-05-08.
- [218] Miko M. Stulajter. Optimizing colibri2 via on-the-fly energy calculations: Applications to atmospheric radical chemistry: Scripts and files. (accessed 2026-05-25).
- [219] Kevin Spiekermann, Lagnajit Pattanaik, and William H. Green. High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions. *Sci. Data*, 9:417, 2022.
- [220] Kevin A. Spiekermann, Lagnajit Pattanaik, and William H. Green. Fast predictions of reaction barrier heights: Toward coupled-cluster accuracy. *J. Phys. Chem. A*, 126:3976–3986, 2022.
- [221] Sydney Chapman. A theory of upper-atmospheric ozone. *Mem. R. Meteorol. Soc.*, 3:103–125, 1930.
- [222] Pablo A. Denis and Fernando R. Ornellas. Theoretical characterization of hydrogen polyoxides: HOOH, HOOOH, HOOOOH, and HOOO. *J. Phys. Chem. A*, 113(2):499–506, 2009.

- [223] Alexander V. Levanov, Dmitry A. Sakharov, Anna V. Dashkova, Ewald E. Antipenko, and Valerii V. Lunin. Structure of hydrogen tetraoxide in gas phase and in aqueous environments: Relationship to the hydroperoxyl radical self-reaction. *Struct. Chem.*, 27:1013–1023, 2016.
- [224] M. Tozihi, M. Vahedpour, and F. Nazari. A theoretical study on the mechanism and thermodynamics of ozone–water gas phase reaction. *J. Iran. Chem. Soc.*, 7:585–596, 2010.
- [225] Dieter Cremer, Elfi Kraka, and Peter G. Szalay. Decomposition modes of dioxirane, methyldioxirane and dimethyldioxirane — a CCSD(T), MR-AQCC and DFT investigation. *Chem. Phys. Lett.*, 292:97–109, 1998.
- [226] Yujie Qian, Thanh Lam Nguyen, Peter R. Franke, John F. Stanton, and Marsha I. Lester. Nonstatistical unimolecular decay of the CH_2OO criegee intermediate in the tunneling regime. *J. Phys. Chem. Lett.*, pages 6222–6229, 2024.
- [227] Chao Yan and Lev N. Krasnoperov. Pressure-dependent kinetics of the reaction between CH_3O_2 and OH : TRIOX formation. *J. Phys. Chem. A*, 123:8349–8357, 2019.
- [228] Feng Zhang and Can Huang. Pressure-dependent kinetics of the reaction between CH_3OO and OH focusing on the product yield of methyltrioxide (CH_3OOOH). *J. Phys. Chem. Lett.*, 10:3598–3603, 2019.
- [229] Adriana Bossolasco, Eszter P. Faragó, Coralie Schoemaeker, and Christa Fittschen. Rate constant of the reaction between CH_3O_2 and OH radicals. *Chem. Phys. Lett.*, 593:7–13, 2014.
- [230] Emmanuel Assaf, Bo Song, Alexandre Tomas, Coralie Schoemaeker, and Christa Fittschen. Rate constant of the reaction between CH_3O_2 radicals and OH radicals revisited. *J. Phys. Chem. A*, 120:8923–8932, 2016.
- [231] Chao Yan, Stefani Kocavska, and Lev N. Krasnoperov. Kinetics of the reaction of CH_3O_2 radicals with OH studied over the 292–526 K temperature range. *J. Phys. Chem. A*, 120(31):6111–6121, 2016.
- [232] Božo Plešničar, Jože Koller, Janez Cerkovnik, Tell Tuttle, and Dieter Cremer. ^{17}O NMR spectroscopic characterization and the mechanism of formation of alkyl hydrotrioxides (ROOOH) and hydrogen trioxide (HOOOH) in the low-temperature ozonation of isopropyl alcohol and isopropyl methyl ether: Water-assisted decomposition. *Chem. Eur. J.*, 6:809–819, 2000.
- [233] Emmanuel Assaf, S. Tanaka, Y. Kajii, S. Berrehane, and C. Fittschen. The reaction of hydroxyl and methylperoxy radicals is not a major source of atmospheric methanol. *Nat. Commun.*, 9:4343, 2018.
- [234] Oliver Welz, John D. Savee, David L. Osborn, Subith S. Vasu, Carl J. Percival, Dudley E. Shallcross, and Craig A. Taatjes. Direct kinetic measurements of criegee intermediate (CH_2OO) formed by reaction of CH_2I with O_2 . *Science*, 335:204–207, 2012.

- [235] Mica C. Smith, Chun-Hung Chang, Wen Chao, Liang-Chun Lin, Kaito Takahashi, Kristie A. Boering, and Jim Jr-Min Lin. Strong negative temperature dependence of the simplest criegee intermediate CH₂OO reaction with water dimer. *J. Phys. Chem. Lett.*, 6:2708–2713, 2015.
- [236] Leonid Sheps, Brandon Rotavera, Arkke J. Eskola, David L. Osborn, Craig A. Taatjes, Kendrew Au, Dudley E. Shallcross, M. Anwar H. Khan, and Carl J. Percival. The reaction of criegee intermediate CH₂OO with water dimer: primary products and atmospheric impact. *Phys. Chem. Chem. Phys.*, 19(33):21970–21979, 2017.
- [237] Eleni Dovrou, Kelvin H. Bates, Jonathan M. Moch, Loretta J. Mickley, Daniel J. Jacob, and Frank N. Keutsch. Catalytic role of formaldehyde in particulate matter formation. *Proc. Natl. Acad. Sci. U.S.A.*, 119(6):e2113265119, 2022.
- [238] Mark Pfeifle, Yong-Tao Ma, Ahren W. Jasper, Lawrence B. Harding, William L. Hase, and Stephen J. Klippenstein. Nascent energy distribution of the criegee intermediate CH₂OO from direct dynamics calculations of primary ozonide dissociation. *J. Chem. Phys.*, 148(17):174306, 2018.
- [239] Mixtli Campos-Pineda, Lei Yang, and Jingsong Zhang. Direct measurement of the criegee intermediate ch₂oo in ozonolysis of ethene. *Nature Communications*, 16:6515, 2025.
- [240] David J. Wales, Mark A. Miller, and Tiffany R. Walsh. Archetypal energy landscapes. *Nature*, 394:758–760, 1998.
- [241] Anders S. Christensen, Tomáš Kubár, Qiang Cui, and Marcus Elstner. Semiempirical quantum mechanical methods for noncovalent interactions for chemical and biochemical applications. *Chem. Rev.*, 116:5301–5337, 2016.
- [242] Xiao-Jie Zhang and Zhi-Pan Liu. Reaction sampling and reactivity prediction using the stochastic surface walking method. *Phys. Chem. Chem. Phys.*, 17:2757–2769, 2015.
- [243] Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555:604–610, 2018.
- [244] Maike Bergeler, Gregor N. Simm, Jonny Proppe, and Markus Reiher. Heuristics-guided exploration of reaction mechanisms. *J. Chem. Theory Comput.*, 11:5712–5722, 2015.
- [245] Daniel Barter, Evan Walter Clark Spotte-Smith, Nikita S. Redkar, Aniruddh Khanwale, Shyam Dwaraknath, Kristin A. Persson, and Samuel M. Blau. Predictive stochastic analysis of massive filter-based electrochemical reaction networks. *Digital Discovery*, 2:123–137, 2023.
- [246] Moritz Bensberg and Markus Reiher. Uncertainty-aware first-principles exploration of chemical reaction networks. *J. Phys. Chem. A*, 128:4532–4547, 2024.

- [247] Miko M. Stulajter, Liliana Garcia, Kent Ziti, and Dmitrij Rappoport. Euclidean embedding of reaction networks for dimensionality reduction of chemical spaces. Manuscript submitted for publication, 2026.
- [248] Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015.
- [249] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. III Goddard, and W. M. Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.*, 114(25):10024–10035, 1992.
- [250] Andreas Klamt and GJGJ Schüürmann. Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2*, pages 799–805, 1993.
- [251] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.*, 15:1652–1671, 2019.
- [252] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements ($Z = 1\text{--}86$). *J. Chem. Theory Comput.*, 13:1989–2009, 2017.

Appendix A

Supplementary Material for Chapter 2

Contents

A.1	Reaction Network Properties as a Function of the Number of Atoms	148
A.2	C ₃ H ₅ NO ₂ Reaction Network Degree Histograms	150
A.3	C ₃ H ₅ NO ₂ Reaction Network Average Square Clustering Coefficient Histograms	152
A.4	Properties of Comparison Reaction Networks	153
A.5	Generative Model Networks and Reaction Network Properties Compared	154
A.5.1	Fractal Dimension as a Function of Average Square Clustering Coefficient	154
A.5.2	Average Path Length as a Function of Average Square Clustering Coefficient	155
A.5.3	Average Path Length as a Function of Fractal Dimension	156
A.6	Properties of Generative Model Networks	158
A.6.1	Erdős-Rényi Networks	158
A.6.2	Bipartite Random Networks	159
A.6.3	Watts Strogatz Networks	160
A.7	Crossover Point in Reaction Networks	162
A.8	Reaction Network Transformation Rules Histograms	163

A.1 Reaction Network Properties as a Function of the Number of Atoms

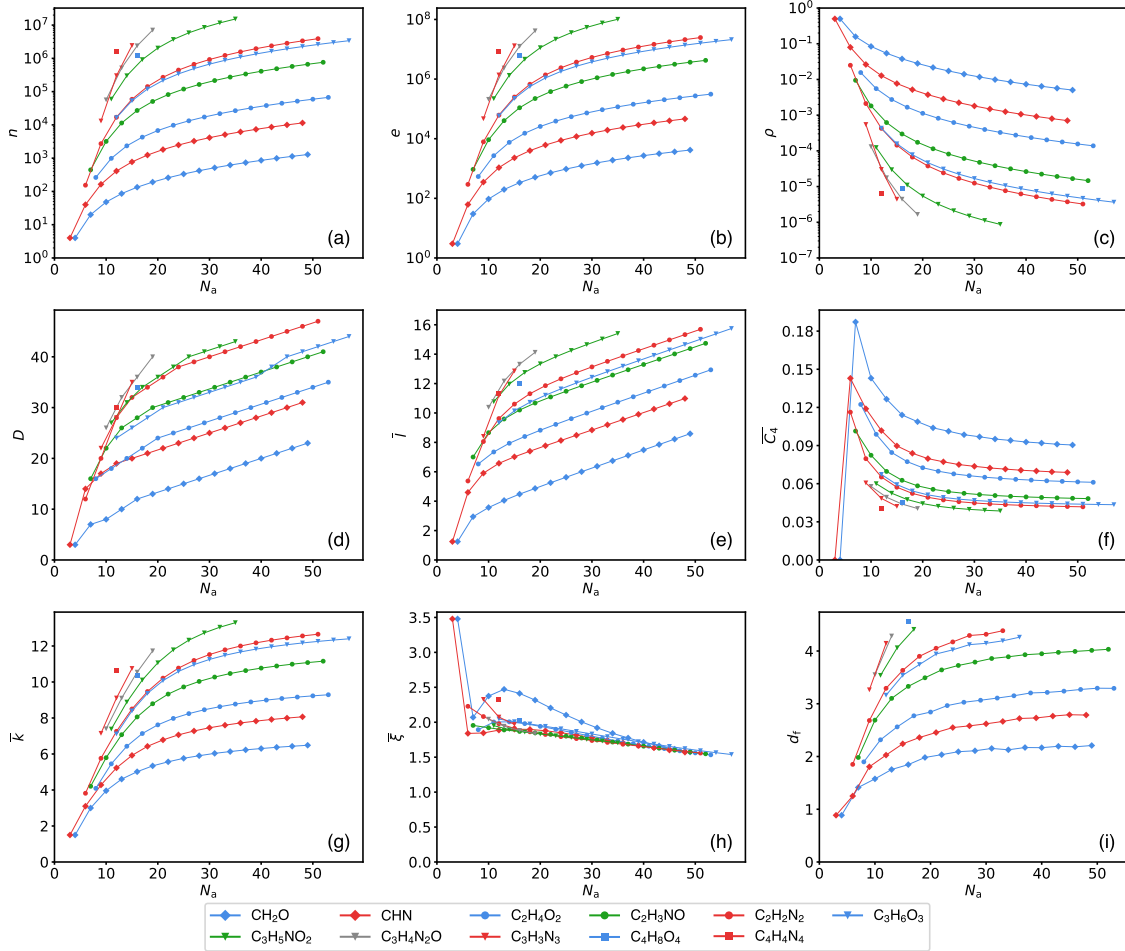


Figure A.1: Variation of RN properties with number of atoms, N_a : (a) number of nodes, n , (b) the number of edges, e , (c) the density, ρ , (d) the diameter, D , (e) the average shortest path length, $\bar{l}$, (f) the average square clustering coefficient, $\overline{C_4}$, (g) the average node degree, $\bar{k}$, (h) the average growth factor, $\bar{\xi}$, and (i) the fractal dimension, d_f . Homologous series are indicated by lines and denoted by their base RNs.

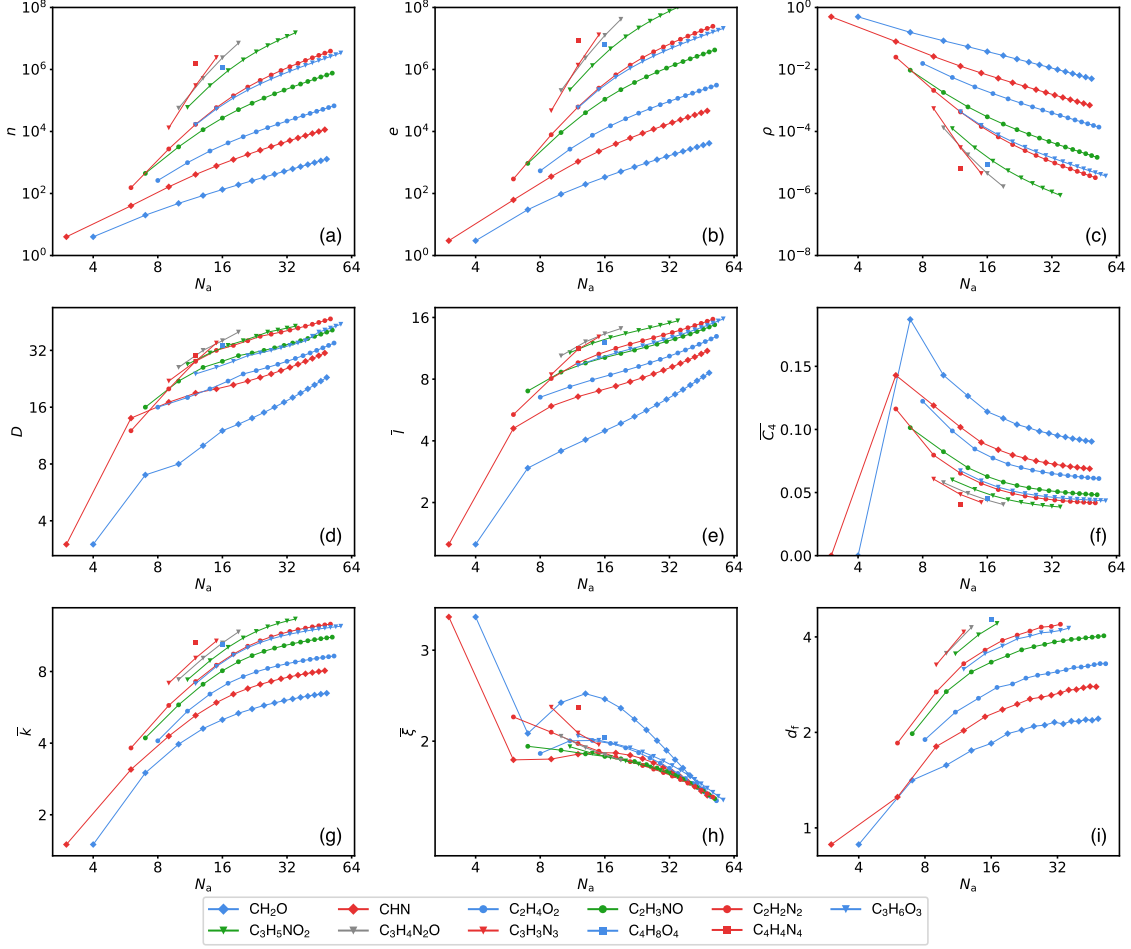


Figure A.2: Variation of RN properties with number of atoms, N_a , on a log-log plot: (a) number of nodes, n , (b) the number of edges, e , (c) the density, ρ , (d) the diameter, D , (e) the average shortest path length, $\bar{l}$, (f) the average square clustering coefficient, $\bar{C}_4$, (g) the average node degree, $\bar{k}$, (h) the average growth factor, $\bar{\xi}$, and (i) the fractal dimension, d_f . Homologous series are indicated by lines and denoted by their base RNs.

A.2 C₃H₅NO₂ Reaction Network Degree Histograms

Histograms of node degrees, k , for reaction networks in the C₃H₅NO₂ homologous series. Normal distributions are fitted for k histograms as visual aids and are shown by red dashed lines. A table at the end of each subsection contains the Jarque-Bera goodness of fit p-value, P_{JB} , and skewness value, SK , for the node degree distributions.

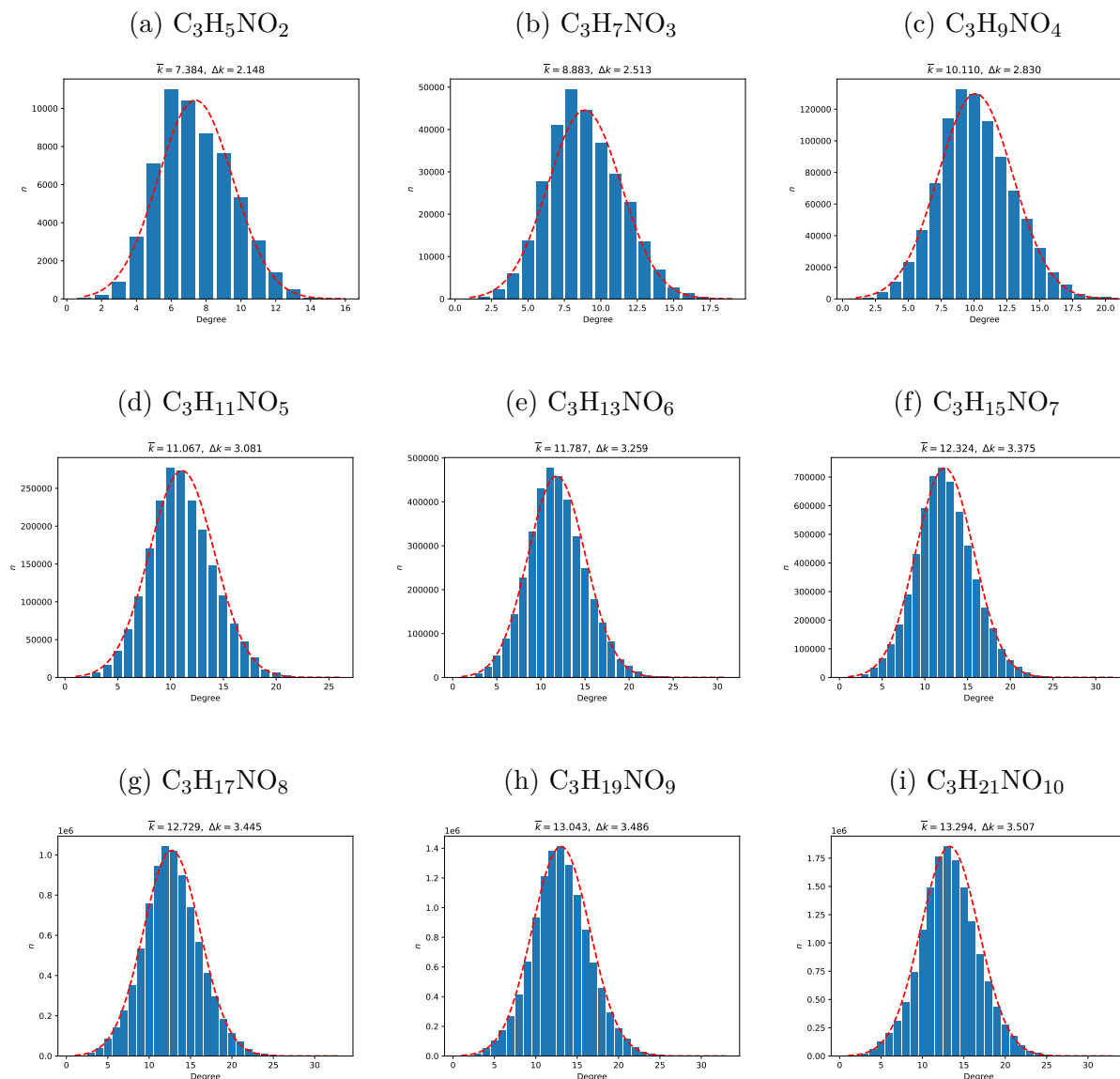


Figure A.3: Histograms of node degree, k , for RNs in the C₃H₅NO₂ homologous series, (a) C₃H₅NO₂, (b) C₃H₇NO₃, (c) C₃H₉NO₄, (d) C₃H₁₁NO₅, (e) C₃H₁₃NO₆, (f) C₃H₁₅NO₇, (g) C₃H₁₇NO₈, (h) C₃H₁₉NO₉, and (i) C₃H₂₁NO₁₀, with fitted normal distributions shown as red dashed lines.

Table A.1: $\text{C}_3\text{H}_5\text{NO}_2$ homologous series Jarque-Bera goodness of fit p-value, P_{JB} , and skewness value, SK .

Network	P_{JB}	SK
$\text{C}_3\text{H}_5\text{NO}_2$	3.42×10^{-200}	0.266
$\text{C}_3\text{H}_7\text{NO}_3$	0.00	0.244
$\text{C}_3\text{H}_9\text{NO}_4$	0.00	0.253
$\text{C}_3\text{H}_{11}\text{NO}_5$	0.00	0.245
$\text{C}_3\text{H}_{13}\text{NO}_6$	0.00	0.224
$\text{C}_3\text{H}_{15}\text{NO}_7$	0.00	0.196
$\text{C}_3\text{H}_{17}\text{NO}_8$	0.00	0.167
$\text{C}_3\text{H}_{19}\text{NO}_9$	0.00	0.141
$\text{C}_3\text{H}_{21}\text{NO}_{10}$	0.00	0.119

A.3 C₃H₅NO₂ Reaction Network Average Square Clustering Coefficient Histograms

Histograms of square clustering coefficients, $\ln(C_4)$, for reaction networks in the C₃H₅NO₂ homologous series. Log-normal distributions are fitted for C_4 histograms as visual aids and are shown by red dashed lines.

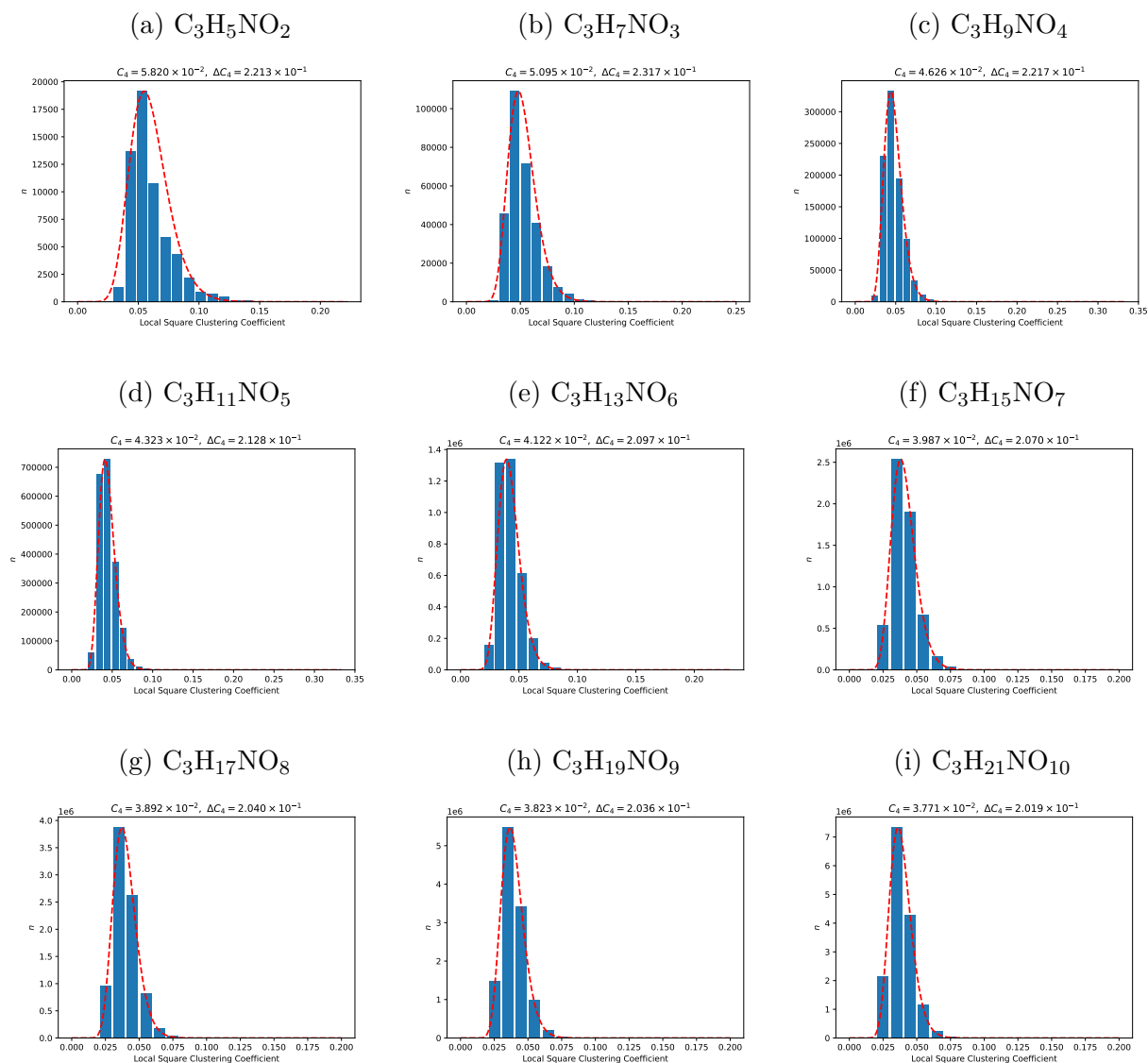


Figure A.4: Histograms of square clustering coefficient, $\ln(C_4)$, for RNs in the C₃H₅NO₂ homologous series, (a) C₃H₅NO₂, (b) C₃H₇NO₃, (c) C₃H₉NO₄, (d) C₃H₁₁NO₅, (e) C₃H₁₃NO₆, (f) C₃H₁₅NO₇, (g) C₃H₁₇NO₈, (h) C₃H₁₉NO₉, and (i) C₃H₂₁NO₁₀, with fitted log-normal distributions shown as red dashed lines.

A.4 Properties of Comparison Reaction Networks

Table A.2: Comparison reaction network properties. Notation: n is number of nodes; e is number of edges; ρ is density; $\bar{k}$ is average degree; $\overline{C_4}$ is average square clustering coefficient; D is diameter; $\bar{l}$ is average shortest path length; d_f is fractal dimension; $\bar{\xi}$ is the growth factor.

Network	n	e	ρ	$\bar{k}$	$\overline{C_4}$	D	$\bar{l}$	d_f	$\bar{\xi}$
<chem>C2H5NO2</chem>	3189	9238	1.82×10^{-3}	5.79	8.24×10^{-2}	22	8.66	2.69	1.92
<chem>CH19NO9</chem>	4131	15,404	1.81×10^{-3}	7.46	7.37×10^{-2}	25	8.84	2.62	1.78
<chem>C2H10O5</chem>	4280	15,274	1.67×10^{-3}	7.14	7.73×10^{-2}	22	8.40	2.77	1.98
<chem>C3H6O3</chem>	16,469	58,581	4.32×10^{-4}	7.11	6.73×10^{-2}	24	9.34	3.16	2.05
<chem>C2H6N2O2</chem>	17,014	61,762	4.27×10^{-4}	7.26	6.53×10^{-2}	28	9.62	3.29	1.98
<chem>C2H18O9</chem>	17,252	72,955	4.90×10^{-4}	8.46	6.63×10^{-2}	27	9.99	3.07	1.81
<chem>C3H8O4</chem>	53,262	223,586	1.58×10^{-4}	8.40	5.93×10^{-2}	26	10.14	3.54	2.01
<chem>C3H4N2O</chem>	56,485	208,955	1.31×10^{-4}	7.40	5.78×10^{-2}	26	10.39	3.55	2.04
<chem>C2H8N2O3</chem>	58,834	250,059	1.44×10^{-4}	8.50	5.72×10^{-2}	32	10.60	3.63	1.91
<chem>C3H5NO2</chem>	59,634	220,172	1.24×10^{-4}	7.38	6.00×10^{-2}	27	10.77	3.54	1.95

A.5 Generative Model Networks and Reaction Network Properties Compared

The plots below compare properties for comparison RNs ($\text{C}_2\text{H}_5\text{NO}_2$, $\text{CH}_{19}\text{NO}_9$, $\text{C}_2\text{H}_{10}\text{O}_5$, $\text{C}_3\text{H}_6\text{O}_3$, $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$, $\text{C}_2\text{H}_{18}\text{O}_9$, $\text{C}_3\text{H}_8\text{O}_4$, $\text{C}_3\text{H}_4\text{N}_2\text{O}$, $\text{C}_2\text{H}_8\text{N}_2\text{O}_3$, and $\text{C}_3\text{H}_5\text{NO}_2$, circles), Erdős–Rényi networks (ER, squares), bipartite random networks (BR, stars), Watts–Strogatz networks (WS, diamonds), 1D to 5D regular lattices with non-periodic boundary conditions (LNP, down triangles) and periodic boundary conditions (LP right-facing triangles), LNP networks with rewiring (LNP-R, up triangles), and LP networks with rewiring (LP-R, left-facing triangles).

A.5.1 Fractal Dimension as a Function of Average Square Clustering Coefficient

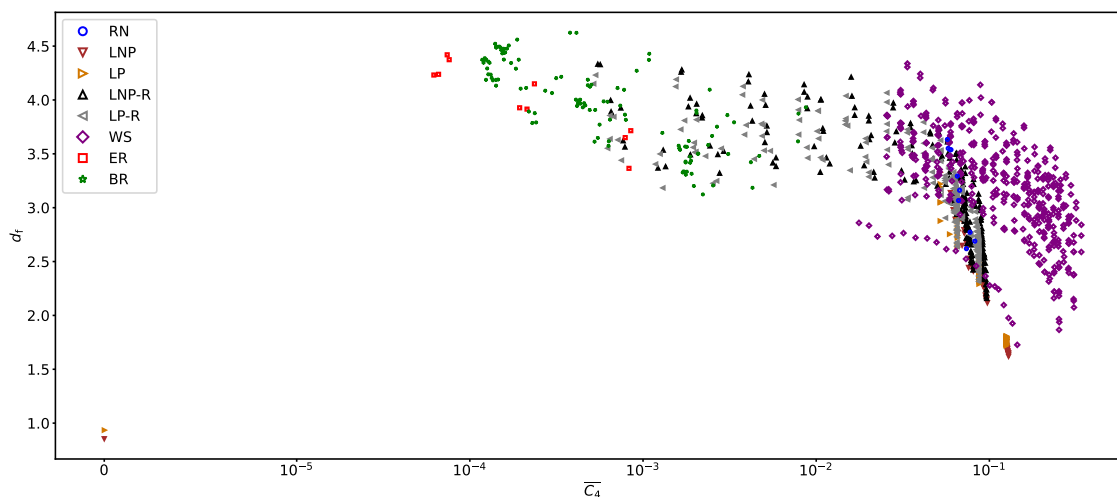


Figure A.5: Plot of fractal dimensions, d_f , against average square clustering coefficient, $\overline{C}_4$.

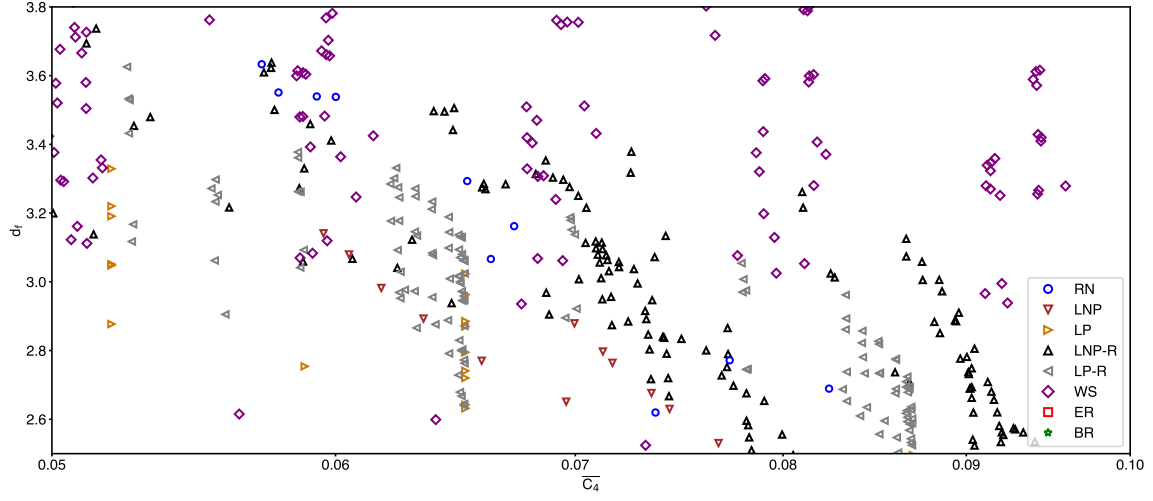


Figure A.6: Zoomed plot of fractal dimensions, d_f , against average square clustering coefficient, $\overline{C_4}$.

A.5.2 Average Path Length as a Function of Average Square Clustering Coefficient

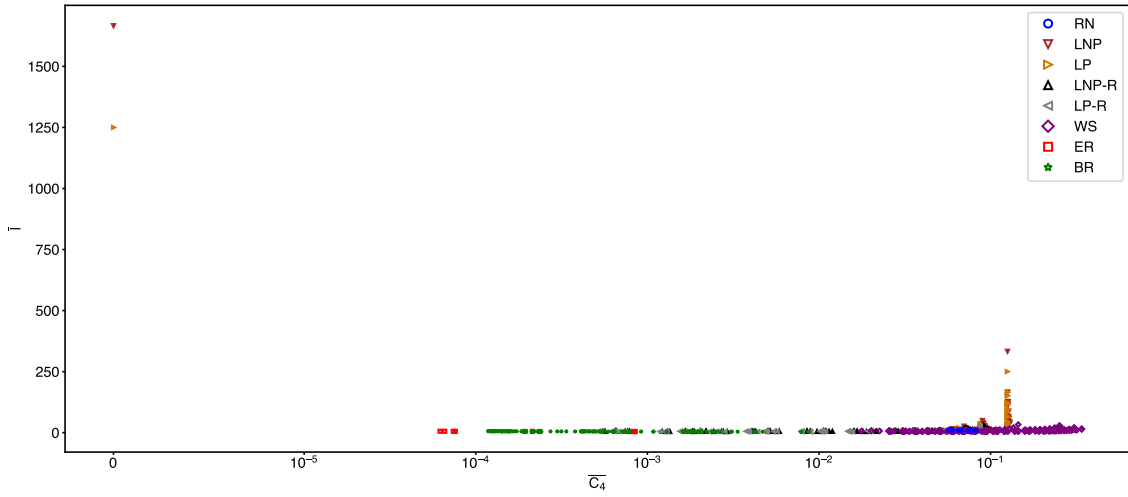


Figure A.7: Plot of average shortest path length, $\bar{l}$, against average square clustering coefficient, $\overline{C_4}$.

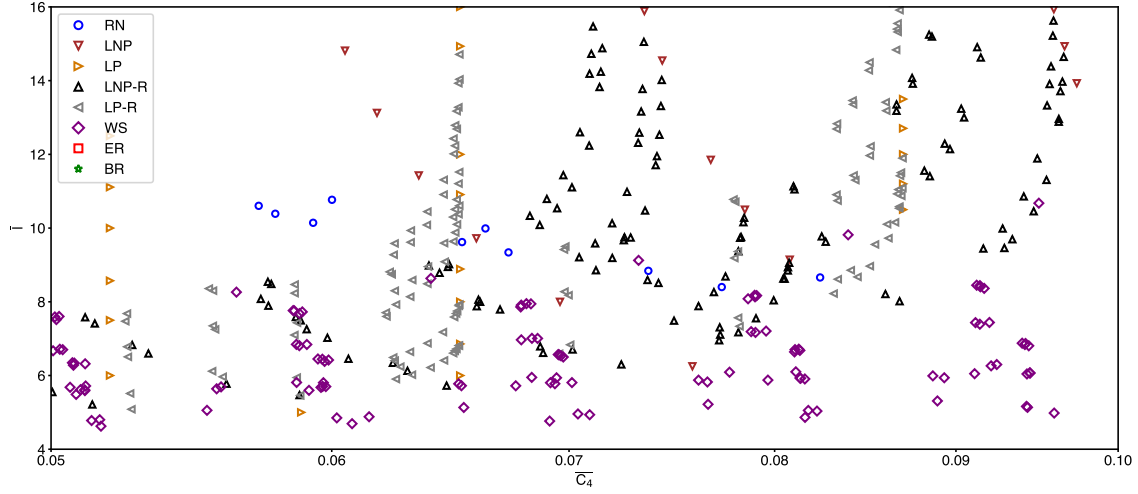


Figure A.8: Zoomed plot of average shortest path length, $\bar{l}$, against average square clustering coefficient, $\overline{C_4}$.

A.5.3 Average Path Length as a Function of Fractal Dimension

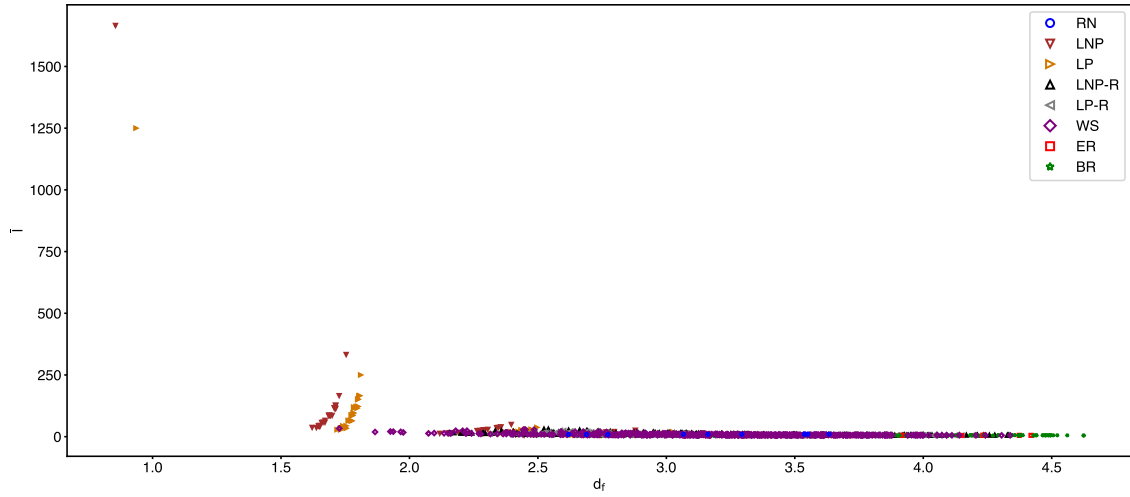


Figure A.9: Plot of average shortest path length, $\bar{l}$, against fractal dimensions, d_f .

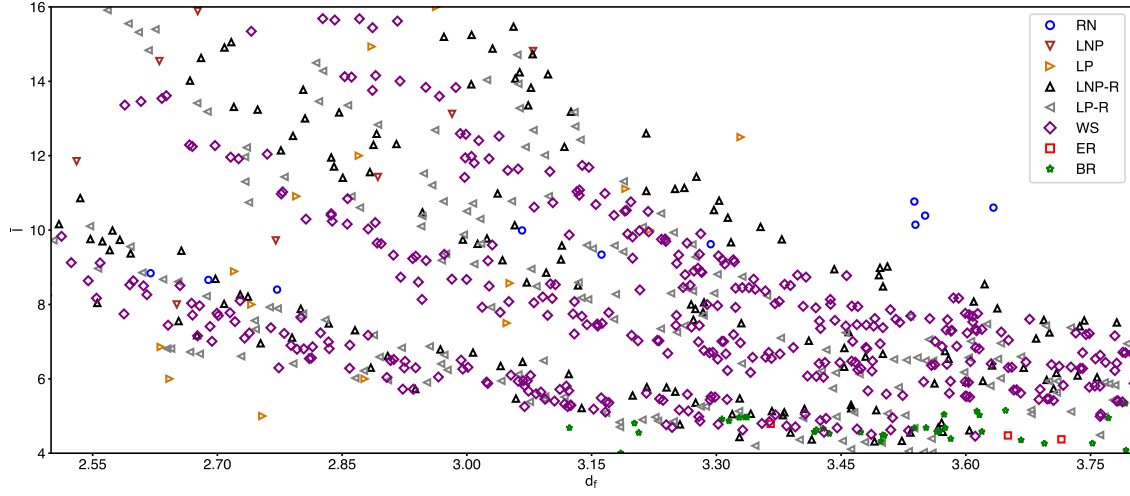


Figure A.10: Zoomed plot of average shortest path length, $\bar{l}$, against fractal dimensions, d_f .

A.6 Properties of Generative Model Networks

A.6.1 Erdős-Rényi Networks

For comparison with RNs, we fixed the model parameters to reproduce the number of nodes and edges of the RNs. The resulting number of edges, e , deviates slightly from that of the RN due to the random generation process. The properties are averaged over five independently generated ER networks, which produces non-integer results for shortest path distances.

Table A.3: Table of sized-matched Erdős-Rényi (ER) networks. See Table A.2 for notation.

ER	n	e	ρ	$\bar{k}$	$\overline{C_4}$	D	$\bar{l}$	d_t	$\bar{\xi}$
C ₂ H ₅ NO ₂	3189	9188.2	1.81×10^{-3}	5.76	8.30×10^{-4}	9.37	4.79	3.37	10.48
CH ₁₉ NO ₉	4131	15,395.2	1.80×10^{-3}	7.45	8.50×10^{-4}	8.00	4.37	3.71	14.92
C ₂ H ₁₀ O ₅	4280	15,231.2	1.66×10^{-3}	7.12	7.89×10^{-4}	8.19	4.48	3.65	14.93
C ₃ H ₆ O ₃	16,469	58,466.0	4.31×10^{-4}	7.10	2.14×10^{-4}	9.19	5.17	3.92	13.10
C ₂ H ₆ N ₂ O ₂	17,014	61,701.4	4.26×10^{-4}	7.25	1.94×10^{-4}	8.99	5.14	3.93	13.49
C ₂ H ₁₈ O ₉	17,252	72,847.2	4.90×10^{-4}	8.45	2.36×10^{-4}	8.40	4.81	4.15	22.11
C ₃ H ₈ O ₄	53,262	223,351.0	1.57×10^{-4}	8.39	7.60×10^{-5}	9.20	5.36	4.37	17.36
C ₃ H ₄ N ₂ O	56,485	209,125.4	1.31×10^{-4}	7.40	6.60×10^{-5}	9.99	5.69	4.24	17.18
C ₂ H ₈ N ₂ O ₃	58,834	250,083.0	1.44×10^{-4}	8.50	7.40×10^{-5}	9.20	5.38	4.42	18.01
C ₃ H ₅ NO ₂	59,634	220,281.6	1.24×10^{-4}	7.39	6.20×10^{-5}	9.99	5.72	4.23	17.03

A.6.2 Bipartite Random Networks

Table A.4: Table of size-matched bipartite random networks to $\text{C}_2\text{H}_5\text{NO}_2$ with partition fraction $p_s = 0.5$. See Table A.2 for other notation.

BR	n	e	ρ	$\bar{k}$	$\overline{C_4}$	D	$\bar{l}$	d_f	$\bar{\xi}$
$\text{C}_2\text{H}_5\text{NO}_2$	3189	9238	1.82×10^{-3}	5.79	1.79×10^{-3}	9.77	4.97	3.33	6.13
$\text{CH}_{19}\text{NO}_9$	4131	15,404	1.81×10^{-3}	7.46	1.75×10^{-3}	8.39	4.58	3.57	7.38
$\text{C}_2\text{H}_{10}\text{O}_5$	4280	15,274	1.67×10^{-3}	7.14	1.61×10^{-3}	8.59	4.68	3.58	7.05
$\text{C}_3\text{H}_6\text{O}_3$	16,469	58,581	4.32×10^{-4}	7.11	4.34×10^{-4}	9.99	5.38	3.97	7.29
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	17,014	61,762	4.27×10^{-4}	7.26	4.24×10^{-4}	9.79	5.35	3.98	7.39
$\text{C}_2\text{H}_{18}\text{O}_9$	17,252	72,955	4.90×10^{-4}	8.46	4.80×10^{-4}	9.00	5.04	4.14	7.63
$\text{C}_3\text{H}_8\text{O}_4$	53,262	223,586	1.58×10^{-4}	8.40	1.54×10^{-4}	10.00	5.58	4.47	8.01
$\text{C}_3\text{H}_4\text{N}_2\text{O}$	56,485	208,955	1.31×10^{-4}	7.40	1.26×10^{-4}	10.79	5.91	4.29	7.18
$\text{C}_2\text{H}_8\text{N}_2\text{O}_3$	58,834	250,059	1.44×10^{-4}	8.50	1.40×10^{-4}	10.00	5.60	4.52	8.03
$\text{C}_3\text{H}_5\text{NO}_2$	59,634	220,172	1.24×10^{-4}	7.38	1.18×10^{-4}	10.99	5.94	4.37	7.17

A.6.3 Watts Strogatz Networks

In the construction of WS networks, the number of nodes, n , was matched to the RNs, and the number of nearest neighbors, K , was chosen to match the number of edges, e , as closely as possible.

Table A.5: Watts-Strogatz networks closest-matching each reaction network (RN) by $\overline{C}_4$, minimising $|\overline{C}_4^{\text{WS}} - \overline{C}_4^{\text{RN}}|$. See Table A.2 for notation.

RN	p_r	n	e	ρ	$\bar{k}$	$\overline{C}_4$	D	$\bar{l}$	d_f	$\bar{\xi}$
<chem>C2H5NO2</chem>	0.225	3189	9567	1.88×10^{-3}	6.00	7.97×10^{-2}	9.80	5.88	3.03	5.63
<chem>CH19NO9</chem>	0.275	4131	16,524	1.94×10^{-3}	8.00	7.09×10^{-2}	7.80	4.94	3.43	9.46
<chem>C2H10O5</chem>	0.25	4280	17,120	1.87×10^{-3}	8.00	8.18×10^{-2}	8.00	5.05	3.41	8.57
<chem>C3H6O3</chem>	0.275	16,469	65,876	4.86×10^{-4}	8.00	6.94×10^{-2}	9.00	5.78	3.75	10.18
<chem>C2H6N2O2</chem>	0.275	17,014	68,056	4.70×10^{-4}	8.00	7.01×10^{-2}	9.00	5.81	3.76	9.95
<chem>C2H18O9</chem>	0.275	17,252	69,008	4.64×10^{-4}	8.00	6.92×10^{-2}	9.00	5.80	3.76	10.05
<chem>C3H8O4</chem>	0.3	53,262	213,048	1.50×10^{-4}	8.00	5.97×10^{-2}	9.20	6.38	3.70	10.21
<chem>C3H4N2O</chem>	0.3	56,485	225,940	1.42×10^{-4}	8.00	5.99×10^{-2}	9.40	6.42	3.78	10.47
<chem>C2H8N2O3</chem>	0.3	58,834	235,336	1.36×10^{-4}	8.00	5.96×10^{-2}	9.40	6.44	3.77	10.64
<chem>C3H5NO2</chem>	0.3	59,634	238,536	1.34×10^{-4}	8.00	5.95×10^{-2}	9.20	6.45	3.67	10.68

Table A.6: Watts-Strogatz networks closest-matching each reaction network (RN) by d_f , minimising $|d_f^{\text{WS}} - d_f^{\text{RN}}|$. See Table A.2 for notation.

RN	p_r	n	e	ρ	$\bar{k}$	$\overline{C_4}$	D	$\bar{l}$	d_f	$\bar{\xi}$
$\text{C}_2\text{H}_5\text{NO}_2$	0.1	3189	9567	1.88×10^{-3}	6.00	1.60×10^{-1}	13.20	7.41	2.69	3.46
$\text{CH}_{19}\text{NO}_9$	0.05	4131	16,524	1.94×10^{-3}	8.00	2.50×10^{-1}	13.80	7.71	2.67	3.26
$\text{C}_2\text{H}_{10}\text{O}_5$	0.075	4280	17,120	1.87×10^{-3}	8.00	2.20×10^{-1}	11.60	6.86	2.81	4.08
$\text{C}_3\text{H}_6\text{O}_3$	0.1	16,469	65,876	4.86×10^{-4}	8.00	1.91×10^{-1}	12.00	7.46	3.14	4.87
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	0.2	17,014	68,056	4.70×10^{-4}	8.00	1.09×10^{-1}	9.20	6.22	3.31	7.53
$\text{C}_2\text{H}_{18}\text{O}_9$	0.075	17,252	69,008	4.64×10^{-4}	8.00	2.20×10^{-1}	14.00	8.20	3.09	4.06
$\text{C}_3\text{H}_8\text{O}_4$	0.175	53,262	213,048	1.50×10^{-4}	8.00	1.25×10^{-1}	11.00	7.23	3.60	6.98
$\text{C}_3\text{H}_4\text{N}_2\text{O}$	0.175	56,485	225,940	1.42×10^{-4}	8.00	1.26×10^{-1}	11.00	7.28	3.58	7.01
$\text{C}_2\text{H}_8\text{N}_2\text{O}_3$	0.175	58,834	235,336	1.36×10^{-4}	8.00	1.26×10^{-1}	11.00	7.31	3.58	7.05
$\text{C}_3\text{H}_5\text{NO}_2$	0.175	59,634	238,536	1.34×10^{-4}	8.00	1.26×10^{-1}	11.00	7.31	3.58	7.06

Table A.7: Watts-Strogatz networks closest-matching each reaction network (RN) by the combined normalised distance $\delta = \sqrt{(\Delta\overline{C_4}/\overline{C_4})^2 + (\Delta d_f/d_f)^2}$. An asterisk (*) indicates that this row coincides with the $\overline{C_4}$ match. See Table A.2 for notation.

RN	p_r	n	e	ρ	$\bar{k}$	$\overline{C_4}$	D	$\bar{l}$	d_f	$\bar{\xi}$
$\text{C}_2\text{H}_5\text{NO}_2^*$	0.225	3189	9567	1.88×10^{-3}	6.00	7.97×10^{-2}	9.80	5.88	3.03	5.63
$\text{CH}_{19}\text{NO}_9$	0.25	4131	16,524	1.94×10^{-3}	8.00	8.22×10^{-2}	8.00	5.03	3.37	8.60
$\text{C}_2\text{H}_{10}\text{O}_5^*$	0.25	4280	17,120	1.87×10^{-3}	8.00	8.18×10^{-2}	8.00	5.05	3.41	8.57
$\text{C}_3\text{H}_6\text{O}_3$	0.3	16,469	65,876	4.86×10^{-4}	8.00	5.96×10^{-2}	8.20	5.68	3.48	11.34
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	0.3	17,014	68,056	4.70×10^{-4}	8.00	5.98×10^{-2}	8.60	5.70	3.66	11.24
$\text{C}_2\text{H}_{18}\text{O}_9$	0.3	17,252	69,008	4.64×10^{-4}	8.00	5.96×10^{-2}	8.60	5.71	3.66	11.25
$\text{C}_3\text{H}_8\text{O}_4^*$	0.3	53,262	213,048	1.50×10^{-4}	8.00	5.97×10^{-2}	9.20	6.38	3.70	10.21
$\text{C}_3\text{H}_4\text{N}_2\text{O}^*$	0.3	56,485	225,940	1.42×10^{-4}	8.00	5.99×10^{-2}	9.40	6.42	3.78	10.47
$\text{C}_2\text{H}_8\text{N}_2\text{O}_3^*$	0.3	58,834	235,336	1.36×10^{-4}	8.00	5.96×10^{-2}	9.40	6.44	3.77	10.64
$\text{C}_3\text{H}_5\text{NO}_2^*$	0.3	59,634	238,536	1.34×10^{-4}	8.00	5.95×10^{-2}	9.20	6.45	3.67	10.68

A.7 Crossover Point in Reaction Networks

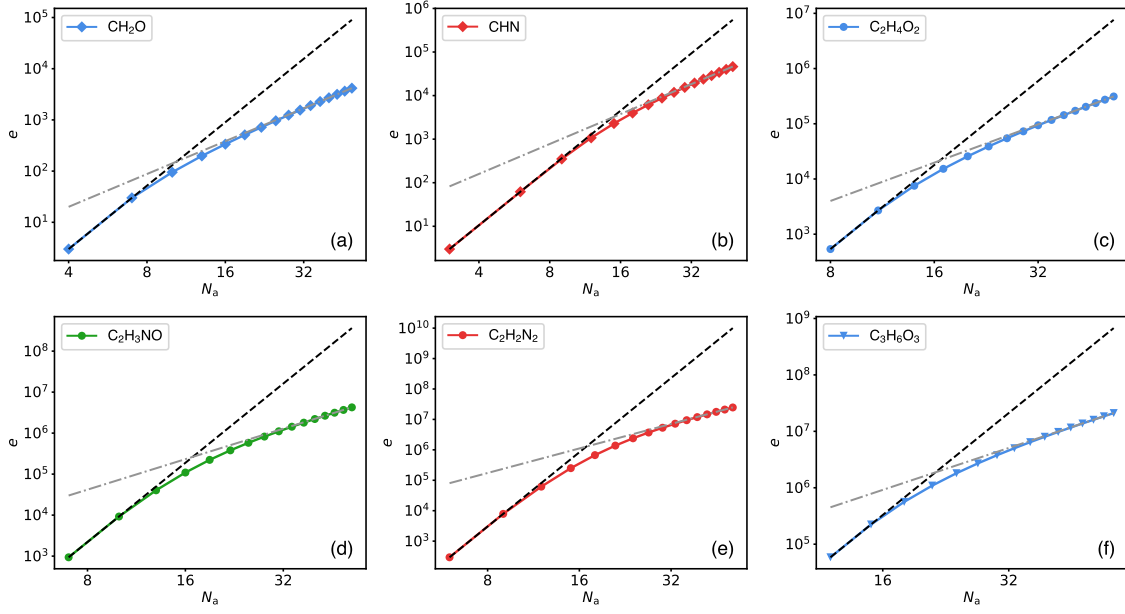


Figure A.11: Scaling of e and N_a on a log-log plot for the RN of (a) CH_2O , (b) CHN , (c) $\text{C}_2\text{H}_4\text{O}_2$, (d) $\text{C}_2\text{H}_3\text{NO}$, (e) $\text{C}_2\text{H}_2\text{N}_2$, and (f) $\text{C}_3\text{H}_6\text{O}_3$. Linear fits are done of the first two points, black line, and the last two points, gray line.

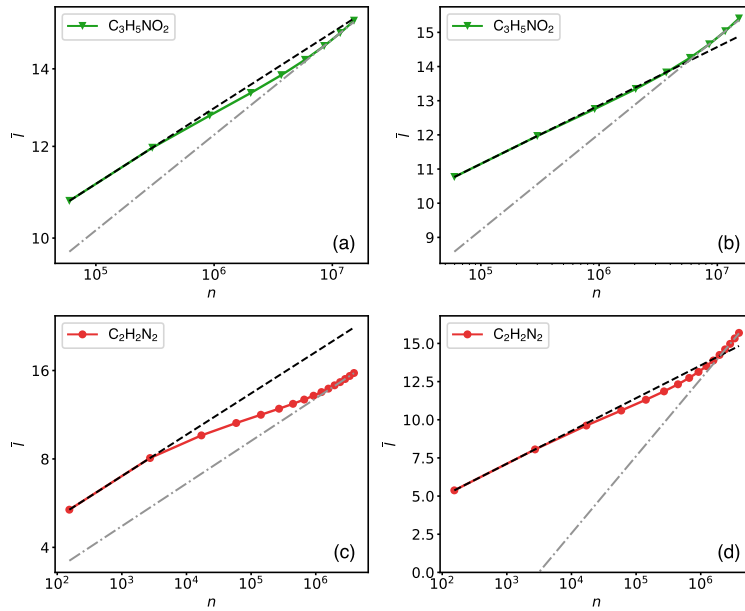


Figure A.12: Scaling of $\bar{l}$ against n for (a-b) $\text{C}_3\text{H}_5\text{NO}_2$ and (c-d) $\text{C}_2\text{H}_2\text{N}_2$ on (a,c) a log-log plot and (b,d) a semilog x-axis plot. Linear fits are done of the first two points, black line, and the last two points, gray line.

A.8 Reaction Network Transformation Rules Histograms

Histograms of rule distributions for select reaction networks in the C_2H_3NO homologous series. The notations for the rules is $N_1XN_2 - T$, where N_1 and N_2 are the atoms of the bond (a lower case p next to an atom denotes that the atom is positively charged, otherwise an atom is neutral), X is the bond order, and T is either F or B denoting either a bond formation or break, respectively.

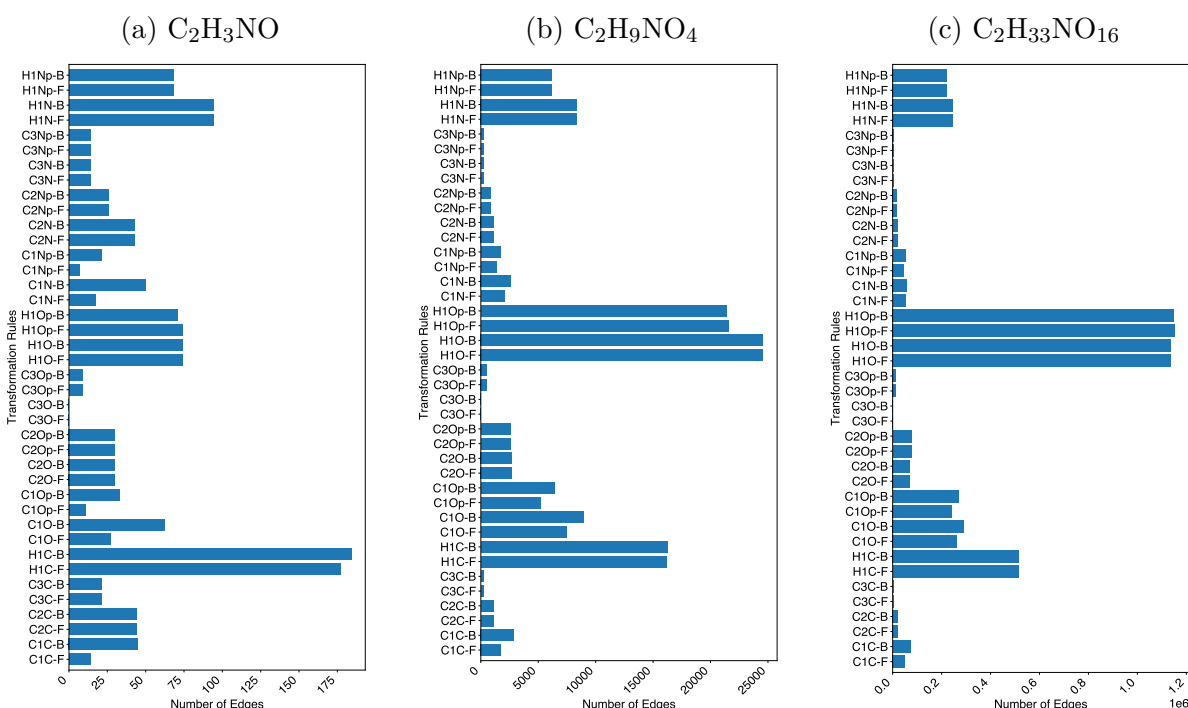


Figure A.13: Histograms of network edges by individual transformation rule for (a) the base C_2H_3NO RN, (b) the $C_2H_9NO_4$ RN containing three H_2O molecules, and (c) the water-rich $C_2H_{33}NO_{16}$ RN containing 15 H_2O molecules. These ungrouped distributions correspond to the grouped transformation-rule distributions shown in Figure 2.6.

Appendix B

Supplementary Material for Chapter 3

Contents

B.1	Reaction Networks	166
B.1.1	Stoichiometry Preserving Rules	166
B.1.2	Energy Calculation Details	167
B.1.3	Reaction Network Statistics	168
B.2	Embedding Methods	169
B.3	Error Metrics	171
B.4	Classical Multidimensional Scaling	173
B.4.1	Embedding Metrics	173
B.4.2	Optimal Embedding Dimension	174
B.5	Landmark Multidimensional Scaling: Embedding Metrics	175
B.6	Metric Multidimensional Scaling	176
B.6.1	Embedding Metrics	176
B.6.2	Optimal Embedding Dimension	178
B.7	Isomap: Embedding Metrics	179
B.8	T-distributed Stochastic Neighbor Embedding: Embedding Metrics	180

B.9	Kernel Principal Component Analysis: Embedding Metrics	182
B.10	Distance-preserving Encoder	184
B.10.1	Embedding Metrics	184
B.10.2	Optimal Embedding Dimension	186
B.11	Randomized tests	187
B.12	Pathway Discovery Costs	195
B.12.1	Enol Tautomerization ($\text{C}_2\text{H}_4\text{O}$)	195
B.12.2	HCN Trimerization ($\text{C}_3\text{H}_3\text{N}_3$)	195
B.12.3	Esterification ($\text{C}_3\text{H}_8\text{O}_3$)	196
B.12.4	C_3 Aldol Addition ($\text{C}_3\text{H}_6\text{O}_3$)	197
B.12.5	C_4 Aldol Addition ($\text{C}_4\text{H}_8\text{O}_4$)	197

B.1 Reaction Networks

B.1.1 Stoichiometry Preserving Rules

The reaction rules used in this work encode heterolytic bond transformations over C, H, N, and O atoms, preserving atomic stoichiometry. Each rule operates on a specific bond type and charge state, covering bond dissociation (homolytic cleavage into a cation–anion pair), association (the reverse recombination), bond breaking (reduction of bond order with charge separation), and bond formation (the reverse). Due to a quirk in RDKit’s SMARTS ring-perception, bonds within ring systems require dedicated ring-opening rules distinct from their acyclic counterparts. Therefore, we included ring-opening rules for C–C, C–O, C–O⁺, C–N, and C–N⁺ bonds in rings of size 3 through 6. The acyclic rules are summarized in Table B.1.

Table B.1: Polar bond transformation rules over carbon, hydrogen, nitrogen, and oxygen used in this work.

Bond Dissociation	Bond Association	Bond Breaking	Bond Formation
$\text{H} - \text{C} \rightarrow \text{H}^+ \dots \text{C}^-$	$\text{H}^+ \dots \text{C}^- \rightarrow \text{H} - \text{C}$	$\text{C} = \text{C} \rightarrow \text{C}^+ - \text{C}^-$	$\text{C}^+ - \text{C}^- \rightarrow \text{C} = \text{C}$
$\text{C} - \text{C} \rightarrow \text{C}^+ \dots \text{C}^-$	$\text{C}^+ \dots \text{C}^- \rightarrow \text{C} - \text{C}$	$\text{C} \equiv \text{C} \rightarrow \text{C}^+ = \text{C}^-$	$\text{C}^+ = \text{C}^- \rightarrow \text{C} \equiv \text{C}$
$\text{H} - \text{O} \rightarrow \text{H}^+ \dots \text{O}^-$	$\text{H}^+ \dots \text{O}^- \rightarrow \text{H} - \text{O}$	$\text{C} = \text{O} \rightarrow \text{C}^+ - \text{O}^-$	$\text{C}^+ - \text{O}^- \rightarrow \text{C} = \text{O}$
$\text{H} - \text{O}^+ \rightarrow \text{H}^+ \dots \text{O}$	$\text{H}^+ \dots \text{O} \rightarrow \text{H} - \text{O}^+$	$\text{C} \equiv \text{O} \rightarrow \text{C}^+ = \text{O}^-$	$\text{C}^+ = \text{O}^- \rightarrow \text{C} \equiv \text{O}$
$\text{C} - \text{O} \rightarrow \text{C}^+ \dots \text{O}^-$	$\text{C}^+ \dots \text{O}^- \rightarrow \text{C} - \text{O}$	$\text{C} = \text{O}^+ \rightarrow \text{C}^+ - \text{O}$	$\text{C}^+ - \text{O} \rightarrow \text{C} = \text{O}^+$
$\text{C} - \text{O}^+ \rightarrow \text{C}^+ \dots \text{O}$	$\text{C}^+ \dots \text{O} \rightarrow \text{C} - \text{O}^+$	$\text{C} \equiv \text{O}^+ \rightarrow \text{C}^+ = \text{O}$	$\text{C}^+ = \text{O} \rightarrow \text{C} \equiv \text{O}^+$
$\text{H} - \text{N} \rightarrow \text{H}^+ \dots \text{N}^-$	$\text{H}^+ \dots \text{N}^- \rightarrow \text{H} - \text{N}$	$\text{C} = \text{N} \rightarrow \text{C}^+ - \text{N}^-$	$\text{C}^+ - \text{N}^- \rightarrow \text{C} = \text{N}$
$\text{H} - \text{N}^+ \rightarrow \text{H}^+ \dots \text{N}$	$\text{H}^+ \dots \text{N} \rightarrow \text{H} - \text{N}^+$	$\text{C} \equiv \text{N} \rightarrow \text{C}^+ = \text{N}^-$	$\text{C}^+ = \text{N}^- \rightarrow \text{C} \equiv \text{N}$
$\text{C} - \text{N} \rightarrow \text{C}^+ \dots \text{N}^-$	$\text{C}^+ \dots \text{N}^- \rightarrow \text{C} - \text{N}$	$\text{C} = \text{N}^+ \rightarrow \text{C}^+ - \text{N}$	$\text{C}^+ - \text{N} \rightarrow \text{C} = \text{N}^+$
$\text{C} - \text{N}^+ \rightarrow \text{C}^+ \dots \text{N}$	$\text{C}^+ \dots \text{N} \rightarrow \text{C} - \text{N}^+$	$\text{C} \equiv \text{N}^+ \rightarrow \text{C}^+ = \text{N}$	$\text{C}^+ = \text{N} \rightarrow \text{C} \equiv \text{N}^+$

B.1.2 Energy Calculation Details

Nodes in the RNs were assigned energies by computing the sum of the molecular energies of all molecules in a given node. The reaction rule set includes fictitious depolarization and polarization transformations that can generate intermediate structures containing partially opened bonds. Because such structures are artifacts of the rule-based representation rather than physically realizable species, their energies cannot be computed directly. Molecular structures were therefore grouped into equivalence classes consisting of molecules that differ only by zero-energy transformations, defined as bond opening and closing operations that modify bond orders and formal charges without changing molecular connectivity. Each equivalence class was assigned the minimum molecular energy among structures with maximal total bond order and minimal sum of absolute formal charges. Node energies were then defined as the sum of these representative energies. Nodes for which any molecular energy calculation failed were removed together with their associated edges.

Molecular energies were computed from SMILES strings as follows. Three-dimensional geometries were generated using RDKit [77], with 1000 conformers produced by the ETDG algorithm [248] (maximum 5000 embedding iterations). All conformers were optimized with the UFF force field [249] (maximum 3000 steps), and the lowest-energy conformer was submitted to MOPAC [155] for geometry optimization and molecular energy calculation. MOPAC calculations used the PM7 semiempirical method with solvent effects modeled by COSMO [250] at $\epsilon = 78.4$ (aqueous conditions), a gradient norm threshold of 0.0, an SCF convergence criterion of 10^{-9} , and a maximum of 1000 optimization cycles. If the selected conformer yielded an unphysically positive molecular energy, up to four additional conformers were reprocessed through MOPAC before the calculation was marked as failed. Further details of the energy calculation implementation in `colibri2` are given in previous work [24].

B.1.3 Reaction Network Statistics

Tables B.2 and B.3 report the size and fractal dimensions of all reaction networks used in this work. Fractal dimensions were computed using the compact box-burning (CBB) method [80, 81, 104, 105]. CBB was applied to reaction networks with both constant edge weights (W_c) and edge weights derived from the karc criterion [22, 24] (W_k), yielding fractal dimensions denoted (const. d_f) and (wt. d_f), respectively. The mean and standard deviation of W_k edge weights are also reported. The fractal dimension of C₄ Aldol Addition (A4) (C₄H₈O₄) could not be computed due to its size. The glyceraldehyde benchmark network is identical to the C₃H₆O₃ test network.

Table B.2: Size and fractal dimensions of the ten test networks with W_c and W_k edge weights.

Formula	Nodes	Edges	const. d_f	wt. d_f	$\bar{w}$	STD
CH ₁₉ NO ₉	4131	30,808	2.652	2.173	1.707	1.342
C ₂ H ₅ NO ₂	3201	18,544	2.687	2.219	1.905	1.817
C ₂ H ₆ N ₂ O ₂	17,006	123,636	3.273	2.682	1.832	1.859
C ₂ H ₈ N ₂ O ₃	58,791	500,140	3.632	2.952	1.842	1.737
C ₂ H ₁₀ O ₅	4280	30,548	2.750	2.277	2.004	1.664
C ₂ H ₁₈ O ₉	17,252	145,910	3.093	2.460	1.893	1.611
C ₃ H ₄ N ₂ O	56,464	418,526	3.557	2.709	1.748	1.993
C ₃ H ₅ NO ₂	59,742	441,790	3.555	2.823	1.910	1.972
C ₃ H ₆ O ₃	16,553	117,864	3.167	2.613	2.061	1.942
C ₃ H ₈ O ₄	53,493	449,232	3.558	2.949	2.073	1.844

Table B.3: Size and fractal dimensions of the five benchmark reaction networks with W_c and W_k edge weights.

Reaction	Formula	Nodes	Edges	const. d_f	wt. d_f	$\bar{w}$	STD
Enol tautomerization (ET)	C ₂ H ₄ O	49	144	1.268	1.096	2.586	1.812
HCN trimerization (HT)	C ₃ H ₃ N ₃	13,033	93,028	3.276	2.400	1.591	2.047
Esterification (ES)	C ₃ H ₈ O ₃	13,143	94,272	3.141	2.591	2.318	1.904
C ₃ Aldol Addition (A3)	C ₃ H ₆ O ₃	Identical to C ₃ H ₆ O ₃ test network					
C ₄ Aldol Addition (A4)	C ₄ H ₈ O ₄	1,194,872	12,418,444	—	—	2.057	2.077

B.2 Embedding Methods

Classical Multidimensional Scaling [144–146] (cMDS), also known as principal coordinates analysis, embeds dissimilarity data by seeking a low-dimensional Euclidean representation in which pairwise distances approximate the original dissimilarities as closely as possible. Let $D = (d_{ij})$ denote the matrix of pairwise distances. cMDS is computed by double-centering the matrix $-\frac{1}{2}(d_{ij}^2)$, giving the Gram matrix, and performing an eigendecomposition, yielding coordinates that minimize the squared error between original and embedded distances under a least-squares criterion.

Landmark Multidimensional Scaling [147] (LMDS) extends cMDS to improve scalability for larger networks. LMDS selects a subset of nodes as landmarks, embeds them using cMDS, and then places the remaining nodes relative to the landmarks using distance-based triangulation. The proportion of nodes used as landmarks can be varied, and landmarks can be selected either randomly or using a selection criterion such as maximally distant points.

Metric Multidimensional Scaling [148] (mMDS) embeds by directly minimizing a stress objective that quantifies discrepancies between original dissimilarities and embedded Euclidean distances. Unlike cMDS, which admits a closed-form solution, mMDS is solved iteratively and does not rely on eigenvalue decomposition. The Scikit-learn [177] library implementation of mMDS was used.

Isomap [149] is a nonlinear embedding method that modifies the input distance matrix prior to embedding. Given a distance matrix, Isomap constructs a neighborhood graph by connecting each node to its k nearest neighbors, with edge weights given by the original distances. Approximate geodesic distances are then computed as shortest-path distances on this graph, and classical MDS is applied to the resulting geodesic distance matrix. The neighborhood size controls the balance between local and global structure preservation, with large neighborhoods yielding behavior closer to classical MDS. The Scikit-learn [177] library

implementation of Isomap was used.

Kernel Principal Component Analysis [151] (KPCA) generates nonlinear embeddings via the kernel trick. KPCA operates by implicitly mapping data into a high-dimensional feature space defined by a kernel function and performing PCA in that space, allowing nonlinear structure to be captured. In this work, KPCA was applied to the Gram matrix. The linear, sigmoid, and polynomial kernels were mainly considered in this work. For the polynomial kernel, the degree and independent term (coef0) were treated as tunable parameters, along with the scaling parameter γ . For the sigmoid kernel, both γ and the independent term (coef0) were treated as tunable parameters. The parameter γ controls the influence of individual data points by scaling pairwise similarities in the feature space. The Scikit-learn [177] library implementation of KPCA was used.

t-Distributed Stochastic Neighbor Embedding [150] (t-SNE) is applied directly to distance matrices to obtain embeddings that emphasize local neighborhood structure. The method converts pairwise distances into probability distributions representing neighborhood affinities and computes a low-dimensional embedding by minimizing the Kullback–Leibler divergence between the original and embedded distributions. A Student’s t-distribution is used in the embedded space to mitigate crowding effects. The parameters that can be tuned are perplexity, early exaggeration, learning rate, and the maximum number of optimization iterations. The Scikit-learn [177] library implementation of t-SNE was used.

B.3 Error Metrics

Embedding quality was assessed using the following metrics, computed between the original shortest-path distances D_{ij}^o and the Euclidean distances in the embedded space D_{ij}^e .

Normalized stress (σ) measures global distance preservation as the normalized root mean squared error between original and embedded distances:

$$\sigma = \sqrt{\frac{\sum_{i < j} (D_{ij}^o - D_{ij}^e)^2}{\sum_{i < j} (D_{ij}^o)^2}} \quad (\text{B.1})$$

Spearman's rank correlation (ρ) quantifies rank preservation between original and embedded distances, independent of their absolute values.

Pearson correlation measures the linear association between original and embedded distances.

Kendall's rank correlation (τ) measures rank preservation as the normalized difference between concordant and discordant pairs among all (i, j) distance pairs:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)} \quad (\text{B.2})$$

where N_c and N_d are the numbers of concordant and discordant pairs respectively.

Coefficient of determination (R^2) is computed with respect to the identity prediction $\hat{D}_{ij}^e = D_{ij}^o$, measuring how well the embedding recovers the original distances without scaling:

$$R^2 = 1 - \frac{\sum_{i < j} (D_{ij}^o - D_{ij}^e)^2}{\sum_{i < j} (D_{ij}^o - \bar{D}^o)^2} \quad (\text{B.3})$$

Mean squared error (MSE) is the mean of the squared pointwise differences between original and embedded distances:

$$\text{MSE} = \frac{1}{N} \sum_{i < j} (D_{ij}^o - D_{ij}^e)^2 \quad (\text{B.4})$$

Mean absolute error (MAE) is the mean of the absolute pointwise differences between original and embedded distances:

$$\text{MAE} = \frac{1}{N} \sum_{i < j} |D_{ij}^o - D_{ij}^e| \quad (\text{B.5})$$

Normalized mean squared error (NMSE) normalizes the squared error by the original distance at each pair, reducing the influence of large distances.

Mean Canberra distance ($\bar{d}_C$) is the mean of the Canberra metric over all pairs, which weights errors by the magnitude of the original distances and is sensitive to small distances.

B.4 Classical Multidimensional Scaling

B.4.1 Embedding Metrics

Table B.4: cMDS embedding metrics at $d = 10$ for all test networks with both W_c and W_k edge weights. All metrics are defined in Section B.3.

Network	Weight	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
CH ₁₉ NO ₉	W_c	0.1747	0.7214	0.9626	0.8674	0.9675	0.0796	1.4458	2.6787	0.0883
	W_k	0.1462	0.8727	0.9730	0.8665	0.9741	0.0500	1.3634	3.1076	0.0680
C ₂ H ₅ NO ₂	W_c	0.1609	0.7028	0.9382	0.8218	0.9449	0.0663	1.2500	2.1297	0.0785
	W_k	0.1274	0.8814	0.9206	0.7681	0.9410	0.0305	0.8622	1.3575	0.0587
C ₂ H ₁₀ O ₅	W_c	0.1546	0.7440	0.9439	0.8322	0.9513	0.0636	1.1537	1.8605	0.0750
	W_k	0.1119	0.9202	0.9515	0.8197	0.9606	0.0248	0.9120	1.5563	0.0532
C ₂ H ₆ N ₂ O ₂	W_c	0.1364	0.7431	0.8970	0.7579	0.9129	0.0377	1.1111	1.8558	0.0620
	W_k	0.1708	0.8028	0.8934	0.7289	0.9274	0.0311	1.0121	1.9382	0.0741
C ₂ H ₁₈ O ₉	W_c	0.1433	0.7592	0.9420	0.8253	0.9493	0.0448	1.2463	2.2418	0.0667
	W_k	0.1093	0.9209	0.9604	0.8358	0.9643	0.0228	1.0703	2.1787	0.0500
C ₃ H ₆ O ₃	W_c	0.1233	0.7851	0.8852	0.7425	0.9016	0.0296	0.9591	1.4274	0.0563
	W_k	0.2039	0.7514	0.9017	0.7350	0.9179	0.0440	1.3272	3.3107	0.0931
C ₃ H ₈ O ₄	W_c	0.1173	0.7832	0.8753	0.7280	0.8919	0.0230	0.9784	1.5122	0.0527
	W_k	0.2091	0.7075	0.9010	0.7331	0.9159	0.0474	1.5128	4.0275	0.1001
C ₃ H ₄ N ₂ O	W_c	0.1210	0.7388	0.8533	0.6993	0.8677	0.0221	1.0281	1.6726	0.0539
	W_k	0.2083	0.7720	0.8829	0.7133	0.9320	0.0469	1.0832	2.1533	0.0992
C ₃ H ₅ NO ₂	W_c	0.1153	0.7759	0.8751	0.7255	0.8881	0.0214	1.0198	1.6396	0.0515
	W_k	0.2426	0.6450	0.8812	0.7071	0.9092	0.0597	1.5169	3.9371	0.1188
C ₂ H ₈ N ₂ O ₃	W_c	0.1204	0.7889	0.8893	0.7459	0.9088	0.0263	1.0602	1.7486	0.0541
	W_k	0.1914	0.7388	0.8961	0.7310	0.9243	0.0381	1.2284	2.6529	0.0879

B.4.2 Optimal Embedding Dimension

Table B.5: Optimal embedding dimension d_e and corresponding normalized stress σ for cMDS with W_c and W_k edge weights. d_e is taken as the number of components at which stress reaches its minimum.

Network	W_c		W_k	
	d_e	σ	d_e	σ
<chem>CH19NO9</chem>	5	0.115	6	0.115
<chem>C2H5NO2</chem>	7	0.123	13	0.116
<chem>C2H10O5</chem>	7	0.122	12	0.108
<chem>C2H6N2O2</chem>	8	0.126	18	0.127
<chem>C2H18O9</chem>	7	0.116	10	0.109
<chem>C3H6O3</chem>	9	0.122	22	0.131
<chem>C3H8O4</chem>	10	0.117	26	0.124
<chem>C3H4N2O</chem>	11	0.118	24	0.131
<chem>C3H5NO2</chem>	10	0.115	29	0.129
<chem>C2H8N2O3</chem>	9	0.119	22	0.126

B.5 Landmark Multidimensional Scaling: Embedding Metrics

Table B.6: LMDS embedding metrics at $d = 10$ for all test networks with both W_c and W_k edge weights. For both W_c and W_k edge weights, results are reported for a random node sampling ratio of $R = 0.3$. All metrics are defined in Section B.3.

Network	Weight	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
CH ₁₉ NO ₉	W_c	0.175	0.721	0.962	0.865	0.967	0.079	1.444	2.681	0.088
	W_k	0.147	0.872	0.973	0.866	0.974	0.052	1.378	3.123	0.069
C ₂ H ₅ NO ₂	W_c	0.162	0.700	0.936	0.818	0.943	0.067	1.253	2.147	0.079
	W_k	0.134	0.870	0.914	0.759	0.935	0.031	0.896	1.492	0.061
C ₂ H ₁₀ O ₅	W_c	0.155	0.744	0.943	0.831	0.951	0.064	1.154	1.863	0.075
	W_k	0.112	0.920	0.951	0.818	0.960	0.025	0.917	1.567	0.053
C ₂ H ₆ N ₂ O ₂	W_c	0.137	0.743	0.897	0.757	0.912	0.038	1.113	1.859	0.062
	W_k	0.174	0.796	0.889	0.723	0.925	0.032	1.031	2.007	0.076
C ₂ H ₁₈ O ₉	W_c	0.143	0.760	0.942	0.826	0.950	0.045	1.247	2.239	0.067
	W_k	0.110	0.920	0.960	0.835	0.964	0.023	1.074	2.195	0.050
C ₃ H ₆ O ₃	W_c	0.123	0.785	0.885	0.742	0.901	0.030	0.958	1.426	0.056
	W_k	0.203	0.753	0.902	0.736	0.919	0.044	1.320	3.283	0.093
C ₃ H ₈ O ₄	W_c	0.117	0.784	0.876	0.729	0.892	0.023	0.976	1.504	0.053
	W_k	0.205	0.719	0.904	0.737	0.919	0.046	1.484	3.875	0.098
C ₃ H ₄ N ₂ O	W_c	0.121	0.739	0.853	0.699	0.868	0.022	1.028	1.671	0.054
	W_k	0.211	0.766	0.882	0.712	0.931	0.048	1.102	2.208	0.101
C ₃ H ₅ NO ₂	W_c	0.115	0.775	0.875	0.725	0.888	0.021	1.021	1.644	0.052
	W_k	0.243	0.642	0.880	0.705	0.908	0.060	1.524	3.966	0.119
C ₂ H ₈ N ₂ O ₃	W_c	0.121	0.788	0.888	0.745	0.908	0.026	1.064	1.760	0.054
	W_k	0.191	0.739	0.896	0.731	0.925	0.038	1.229	2.647	0.088

B.6 Metric Multidimensional Scaling

B.6.1 Embedding Metrics

Table B.7: mMDS embedding metrics at $d = 3$ for all test networks with karc (W_k) edge weights. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
CH ₁₉ NO ₉	0.113	0.924	0.959	0.835	0.965	0.027	0.980	1.857	0.057
C ₂ H ₅ NO ₂	0.175	0.776	0.896	0.732	0.916	0.047	1.181	2.558	0.086
C ₂ H ₁₀ O ₅	0.151	0.855	0.934	0.791	0.941	0.035	1.204	2.820	0.072
C ₂ H ₆ N ₂ O ₂	0.198	0.736	0.867	0.699	0.904	47.72*	1.185	2.595	0.098
C ₂ H ₁₈ O ₉	0.133	0.883	0.948	0.815	0.953	0.029	1.281	3.208	0.065
C ₃ H ₆ O ₃	0.187	0.791	0.909	0.756	0.920	0.169	1.175	2.784	0.089
C ₃ H ₈ O ₄	0.188	0.763	0.903	0.750	0.915	0.275	1.280	3.264	0.090
C ₃ H ₄ N ₂ O	0.179	0.832	0.901	0.746	0.937	814.07*	0.898	1.586	0.090
C ₃ H ₅ NO ₂	0.189	0.783	0.898	0.741	0.922	543.25*	1.110	2.402	0.093
C ₂ H ₈ N ₂ O ₃	0.185	0.755	0.892	0.733	0.916	0.117	1.140	2.487	0.091

* Large NMSE values reflect numerical instability due to very small original distances;
 σ , R^2 , ρ , and τ are the reliable quality indicators for these cases.

Table B.8: mMDS embedding metrics at $d = 10$ for all test networks with karc (W_k) edge weights. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
CH ₁₉ NO ₉	0.068	0.973	0.985	0.902	0.987	0.017	0.596	0.667	0.034
C ₂ H ₅ NO ₂	0.075	0.958	0.973	0.865	0.979	0.204	0.513	0.476	0.036
C ₂ H ₁₀ O ₅	0.070	0.969	0.981	0.890	0.984	0.013	0.559	0.612	0.033
C ₂ H ₆ N ₂ O ₂	0.082	0.955	0.968	0.852	0.977	68.31*	0.481	0.444	0.036
C ₂ H ₁₈ O ₉	0.066	0.971	0.983	0.898	0.986	0.014	0.628	0.795	0.031
C ₃ H ₆ O ₃	0.087	0.955	0.975	0.869	0.977	9.25	0.549	0.604	0.038
C ₃ H ₈ O ₄	0.088	0.948	0.971	0.861	0.974	8.53	0.602	0.714	0.038
C ₃ H ₄ N ₂ O	0.089	0.959	0.967	0.851	0.980	991.38*	0.437	0.390	0.039
C ₃ H ₅ NO ₂	0.093	0.948	0.967	0.849	0.975	1726.40*	0.544	0.578	0.040
C ₂ H ₈ N ₂ O ₃	0.086	0.948	0.967	0.852	0.974	50.23*	0.518	0.531	0.036

* Large NMSE values reflect numerical instability due to very small original distances;

σ , R^2 , ρ , and τ are the reliable quality indicators for these cases.

B.6.2 Optimal Embedding Dimension

Figures B.1 show an example sigmoidal fit to the numerical derivative of $\log_2(\sigma)$ with respect to N_c , and the corresponding stress curves, for the $\text{CH}_{19}\text{NO}_9$ test network embedded using mMDS with W_k edge weights. The inflection point of each fit defines the optimal embedding dimension, d_e . Table B.9 summarizes d_e and the fractal dimension d_f computed by the compact box-burning (CBB) algorithm [23, 81] for all test RNs.

Table B.9: Optimal embedding dimension, d_e , corresponding normalized stress σ , and fractal dimension, d_f , for all ten test networks embedded using mMDS with W_k edge weights.

Network	d_e	σ	d_f	Network	d_e	σ	d_f
$\text{CH}_{19}\text{NO}_9$	4.17	0.091	2.17	$\text{C}_3\text{H}_6\text{O}_3$	7.51	0.091	2.61
$\text{C}_2\text{H}_5\text{NO}_2$	8.06	0.082	2.22	$\text{C}_3\text{H}_8\text{O}_4$	9.96	0.090	2.95
$\text{C}_2\text{H}_{10}\text{O}_5$	7.39	0.080	2.28	$\text{C}_3\text{H}_4\text{N}_2\text{O}$	9.63	0.099	2.71
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	10.43	0.083	2.68	$\text{C}_3\text{H}_5\text{NO}_2$	9.41	0.096	2.82
$\text{C}_2\text{H}_{18}\text{O}_9$	6.54	0.079	2.46	$\text{C}_2\text{H}_8\text{N}_2\text{O}_3$	9.30	0.088	2.95

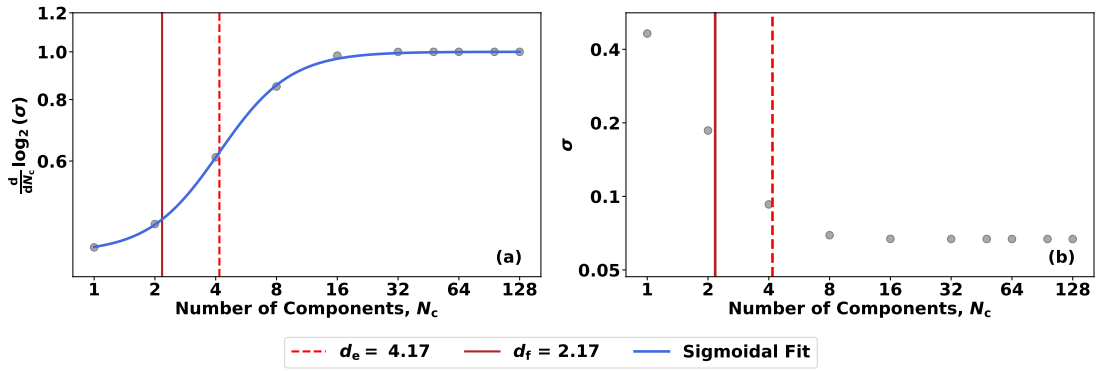


Figure B.1: Sigmoidal fit to $\frac{d}{dN_c} \log_2(\sigma)$ (a) and stress curve σ (b) vs N_c for the $\text{CH}_{19}\text{NO}_9$ RN embedded using metric multidimensional scaling with W_k edge weights. The dashed red vertical line indicates $d_e = 4.17$ and the solid red line indicates $d_f = 2.17$.

B.7 Isomap: Embedding Metrics

Tables B.10 and B.10 report embedding quality metrics for Isomap at embedding dimension $d = 10$ for test RNs with W_k edge weights, across neighborhood sizes $N_n \in \{5, 10, 20, 30\}$. Isomap was evaluated only on the six smaller test networks, as substantial global distortion was observed across all tested networks and neighborhood sizes. Consequently, the four largest RNs ($C_3H_8O_4$, $C_3H_4N_2O$, $C_3H_5NO_2$, and $C_2H_8N_2O_3$) were excluded from testing.

The results are sensitive to the choice of N_n : small values preserve local structure but distort global distances, whereas larger values approach the behavior of cMDS. All metrics are defined in Section B.3. Large NMSE values in some entries reflect the same numerical instability discussed in Section B.6.1; in such cases, σ , R^2 , and τ are more reliable indicators.

Table B.10: Isomap embedding metrics at $d = 10$ for six test networks with W_k edge weights, for neighborhood sizes of $N_n = 5$ and $N_n = 10$. All metrics are defined in Section B.3.

Network	N_n	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$CH_{19}NO_9$	5	0.308	0.435	0.886	0.733	0.880	0.2071	2.480	13.797	0.104
	10	0.851	-3.310	0.779	0.597	0.756	1.3771	8.477	105.200	0.273
$C_2H_5NO_2$	5	1.299	-11.324	0.448	0.312	0.518	3.3314	9.834	141.055	0.353
	10	0.579	-1.448	0.669	0.487	0.688	0.6899	4.267	28.015	0.206
$C_2H_{10}O_5$	5	1.048	-6.002	0.508	0.356	0.533	3.3439	10.138	136.559	0.333
	10	0.812	-3.204	0.674	0.495	0.684	1.4095	7.401	81.990	0.266
$C_2H_6N_2O_2$	5	0.514	-0.784	0.598	0.431	0.585	999.80*	2.843	17.529	0.156
	10	0.672	-2.053	0.599	0.429	0.686	0.8168	4.376	30.003	0.227
$C_2H_{18}O_9$	5	0.286	0.458	0.861	0.693	0.848	0.1550	2.457	14.925	0.092
	10	0.747	-2.696	0.727	0.536	0.718	1.3640	8.634	101.769	0.263
$C_3H_6O_3$	5	1.032	-5.366	0.559	0.394	0.533	3.3233	6.648	84.777	0.269
	10	0.617	-1.279	0.717	0.527	0.701	1.2223	4.116	30.350	0.207

* Large NMSE values reflect numerical instability due to very small original distances.

B.8 T-distributed Stochastic Neighbor Embedding: Embedding Metrics

Tables B.11–B.13 report the best-performing t-SNE configuration for R^2 , τ , and σ for each test RN with W_k edge weights, consolidated across all evaluated hyperparameter combinations ($d \in \{3, 10\}$, $P \in \{50, 75, 150\}$, T , L , M). The parameters were: d (embedding dimension), P (perplexity), T (early exaggeration factor), L (learning rate), and M (maximum iterations). t-SNE was evaluated only on the six smaller test networks. Results are given for $d = 10$.

All metrics were defined in Section B.3. Large NMSE values reflect numerical instability due to very small original distances. In these cases, the reliable indicators are R^2 , ρ , τ , Pearson, and σ .

Table B.11: t-SNE configurations maximising R^2 at $d = 10$ for each test network, selected across $P \in \{50, 75, 150\}$, T , L , M . Best configs: CH₁₉NO₉: P-150 T-24 L-200 M-1000; C₂H₁₀O₅: P-150 T-24 L-200 M-5000; C₂H₁₈O₉: P-50 T-24 M-5000; C₂H₅NO₂: P-150 T-24 L-200; C₂H₆N₂O₂: P-75 T-12 L-200 M-5000; C₃H₆O₃: P-75 T-24 L-500 M-5000. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
CH ₁₉ NO ₉	0.323	0.379	0.651	0.473	0.653	9.228	3.051	15.135	0.149
C ₂ H ₁₀ O ₅	0.419	−0.120	0.464	0.322	0.473	11.309	3.829	21.821	0.184
C ₂ H ₁₈ O ₉	0.463	−0.418	0.391	0.268	0.390	0.187	4.722	39.060	0.207
C ₂ H ₅ NO ₂	0.329	0.207	0.646	0.467	0.602	9511.640*	2.219	9.078	0.126
C ₂ H ₆ N ₂ O ₂	0.364	0.106	0.418	0.289	0.400	6.881	2.117	8.786	0.139
C ₃ H ₆ O ₃	0.367	0.196	0.489	0.341	0.488	4.530	2.397	10.701	0.152

* Large NMSE reflects numerical instability due to very small original distances.

Table B.12: t-SNE configurations maximising τ at $d = 10$ for each test network, selected across $P \in \{50, 75, 150\}$, T , L , M . Best configs: $\text{CH}_{19}\text{NO}_9$: P-150 T-24; $\text{C}_2\text{H}_{10}\text{O}_5$, $\text{C}_2\text{H}_{18}\text{O}_9$, $\text{C}_2\text{H}_5\text{NO}_2$: P-150 T-24 L-200 M-1000; $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$: P-150 T-4 M-1000; $\text{C}_3\text{H}_6\text{O}_3$: P-150 T-4 M-5000. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$\text{CH}_{19}\text{NO}_9$	0.542	-0.748	0.754	0.563	0.737	0.236	5.444	42.674	0.311
$\text{C}_2\text{H}_{10}\text{O}_5$	0.599	-1.285	0.689	0.505	0.688	0.279	5.627	44.570	0.352
$\text{C}_2\text{H}_{18}\text{O}_9$	0.525	-0.825	0.588	0.419	0.577	0.204	5.726	50.253	0.274
$\text{C}_2\text{H}_5\text{NO}_2$	0.592	-1.556	0.731	0.545	0.728	0.284	4.680	29.258	0.359
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	0.433	-0.266	0.500	0.351	0.474	0.112	2.434	12.444	0.170
$\text{C}_3\text{H}_6\text{O}_3$	0.440	-0.156	0.568	0.401	0.536	0.125	2.763	15.393	0.178

Table B.13: t-SNE configurations minimising σ at $d = 10$ for each test network, selected across $P \in \{50, 75, 150\}$, T , L , M . Best configs: $\text{CH}_{19}\text{NO}_9$: P-150 T-24 L-200 M-1000; $\text{C}_2\text{H}_{10}\text{O}_5$: P-150 T-24 L-200 M-5000; $\text{C}_2\text{H}_{18}\text{O}_9$: P-50 T-24 M-5000; $\text{C}_2\text{H}_5\text{NO}_2$: P-150 T-24 L-200; $\text{C}_2\text{H}_6\text{N}_2\text{O}_2$: P-75 T-12 L-200 M-1000; $\text{C}_3\text{H}_6\text{O}_3$: P-75 T-24 L-500 M-5000. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$\text{CH}_{19}\text{NO}_9$	0.323	0.379	0.651	0.473	0.653	9.228	3.051	15.135	0.149
$\text{C}_2\text{H}_{10}\text{O}_5$	0.419	-0.120	0.464	0.322	0.473	11.309	3.829	21.821	0.184
$\text{C}_2\text{H}_{18}\text{O}_9$	0.463	-0.418	0.391	0.268	0.390	0.187	4.722	39.060	0.207
$\text{C}_2\text{H}_5\text{NO}_2$	0.329	0.207	0.646	0.467	0.602	9511.640*	2.219	9.078	0.126
$\text{C}_2\text{H}_6\text{N}_2\text{O}_2$	0.364	0.104	0.417	0.288	0.401	5.148	2.122	8.808	0.139
$\text{C}_3\text{H}_6\text{O}_3$	0.367	0.196	0.489	0.341	0.488	4.530	2.397	10.701	0.152

* Large NMSE reflects numerical instability due to very small original distances.

B.9 Kernel Principal Component Analysis: Embedding Metrics

Tables B.14 and B.15 report embedding quality metrics for KPCA for the `sigmoid` and `linear` kernels at embedding dimension $d = 10$ on RNs with W_k edge weights. Each kernel type is represented by one table, with one row per test network. The linear kernel has no tunable hyperparameters. The sigmoid kernel has two hyperparameters: coefficient (c_0 , an additive constant in the kernel), and scaling parameter (γ , which controls the influence of each training sample).

For all KPCA variants, σ , R^2 , NMSE, MAE, MSE, and $\bar{d}_C$ are strongly affected by numerical instability arising from very small original distances in these networks (see Section B.6.1). The rank-based metrics ρ , τ , and Pearson r are more reliable indicators.

Sigmoid Kernel

Table B.14: KPCA embedding metrics at $d = 10$ for six test networks with W_k edge weights and $K = \text{sigmoid}$, base configuration (default γ and c_0). All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$C_1H_{19}NO_9$	0.862	-3.425	0.424	0.288	0.447	0.688	9.286	108.006	0.708
$C_2H_{10}O_5$	0.835	-3.444	0.462	0.317	0.454	0.622	8.307	86.673	0.652
$C_2H_{18}O_9$	0.877	-4.090	0.433	0.295	0.433	0.723	10.737	140.175	0.741
$C_2H_5N_1O_2$	0.798	-3.651	0.441	0.301	0.416	0.566	6.558	53.234	0.604
$C_2H_6N_2O_2$	0.774	-3.052	0.425	0.293	0.362	0.514	5.557	39.821	0.559
$C_3H_6O_3$	0.798	-2.811	0.436	0.301	0.374	0.540	6.208	50.751	0.581

Linear Kernel

Table B.15: KPCA embedding metrics at $d = 10$ for six test networks with W_k edge weights and $K = \text{linear}$. All metrics are defined in Section B.3.

Network	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$C_1H_{19}NO_9$	319.255	-6.072×10^5	0.931	0.776	0.936	$9.696 \times 10^{4*}$	3406.286	1.482×10^7	0.993
$C_2H_{10}O_5$	206.653	-2.721×10^5	0.903	0.736	0.917	$4.219 \times 10^{4*}$	2062.103	5.306×10^6	0.989
$C_2H_{18}O_9$	597.433	-2.364×10^6	0.900	0.729	0.910	$3.398 \times 10^{5*}$	7181.855	6.511×10^7	0.996
$C_2H_5N_1O_2$	132.450	-1.282×10^5	0.873	0.695	0.896	$1.869 \times 10^{4*}$	1120.625	1.467×10^6	0.984
$C_2H_6N_2O_2$	225.750	-3.447×10^5	0.859	0.680	0.909	$5.328 \times 10^{4*}$	1684.654	3.387×10^6	0.991
$C_3H_6O_3$	245.246	-3.596×10^5	0.873	0.693	0.892	$6.057 \times 10^{4*}$	1947.974	4.789×10^6	0.991

B.10 Distance-preserving Encoder

B.10.1 Embedding Metrics

Table B.16: Distance-preserving Encoder metrics at $d = 3$ embedding dimension for all test networks with W_k edge weights. A landmark and batch sampling ratio of $R = 0.3$ was used. All metrics are defined in Section B.3.

Network	Weight	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
<chem>C1H19NO9</chem>	W_c	0.162	0.760	0.890	0.746	0.902	0.051	1.148	2.306	0.081
	W_k	0.116	0.920	0.958	0.832	0.964	0.028	1.017	1.944	0.059
<chem>C2H5NO2</chem>	W_c	0.196	0.559	0.827	0.669	0.835	0.061	1.341	3.158	0.096
	W_k	0.189	0.738	0.884	0.716	0.896	0.049	1.246	2.996	0.091
<chem>C2H10O5</chem>	W_c	0.189	0.617	0.844	0.691	0.857	0.059	1.269	2.780	0.095
	W_k	0.161	0.834	0.926	0.780	0.930	0.037	1.251	3.230	0.075
<chem>C2H6N2O2</chem>	W_c	0.228	0.284	0.746	0.587	0.771	0.074	1.735	5.172	0.111
	W_k	0.197	0.737	0.866	0.696	0.903	0.056	1.187	2.584	0.098
<chem>C2H18O9</chem>	W_c	0.191	0.571	0.840	0.681	0.852	0.056	1.530	3.991	0.095
	W_k	0.131	0.887	0.949	0.815	0.954	0.029	1.278	3.121	0.064
<chem>C3H6O3</chem>	W_c	0.216	0.341	0.784	0.625	0.802	0.066	1.598	4.380	0.107
	W_k	0.186	0.792	0.910	0.756	0.920	0.047	1.180	2.769	0.090
<chem>C3H8O4</chem>	W_c	0.225	0.204	0.764	0.603	0.782	0.069	1.807	5.549	0.110
	W_k	0.186	0.767	0.907	0.753	0.916	0.047	1.275	3.203	0.090
<chem>C3H4N2O</chem>	W_c	0.237	-0.001	0.729	0.567	0.740	0.074	1.958	6.412	0.116
	W_k	0.175	0.839	0.905	0.750	0.938	0.046	0.884	1.523	0.089
<chem>C3H5NO2</chem>	W_c	0.223	0.158	0.768	0.605	0.778	0.068	1.909	6.158	0.110
	W_k	0.188	0.788	0.899	0.742	0.922	0.049	1.101	2.353	0.093
<chem>C2H8N2O3</chem>	W_c	0.223	0.272	0.759	0.599	0.787	0.071	1.879	6.027	0.110
	W_k	0.185	0.755	0.891	0.732	0.915	0.049	1.140	2.491	0.091

Table B.17: Distance-preserving Encoder metrics at $d = 10$ embedding dimension for all test networks with W_k edge weights. A landmark and batch sampling ratio of $R = 0.3$ was used. All metrics are defined in Section B.3.

Network	Weight	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_C$
$C_1H_{19}NO_9$	W_c	0.086	0.932	0.964	0.869	0.969	0.020	0.638	0.650	0.042
	W_k	0.073	0.968	0.983	0.894	0.985	0.016	0.629	0.777	0.036
$C_2H_5NO_2$	W_c	0.106	0.872	0.940	0.823	0.946	0.020	0.768	0.921	0.048
	W_k	0.105	0.919	0.964	0.843	0.968	0.016	0.689	0.921	0.045
$C_2H_{10}O_5$	W_c	0.095	0.904	0.952	0.845	0.957	0.021	0.670	0.698	0.045
	W_k	0.092	0.946	0.979	0.882	0.981	0.013	0.742	1.055	0.040
$C_2H_6N_2O_2$	W_c	0.095	0.877	0.927	0.803	0.939	0.018	0.753	0.892	0.043
	W_k	0.086	0.950	0.967	0.850	0.975	0.012	0.494	0.492	0.036
$C_2H_{18}O_9$	W_c	0.083	0.919	0.956	0.852	0.961	0.015	0.693	0.758	0.039
	W_k	0.064	0.973	0.984	0.900	0.986	0.010	0.622	0.757	0.030
$C_3H_6O_3$	W_c	0.100	0.860	0.916	0.786	0.927	0.019	0.770	0.932	0.046
	W_k	0.088	0.954	0.974	0.867	0.977	0.012	0.553	0.618	0.039
$C_3H_8O_4$	W_c	0.097	0.852	0.912	0.779	0.923	0.016	0.808	1.033	0.044
	W_k	0.084	0.953	0.973	0.867	0.977	0.011	0.571	0.652	0.037
$C_3H_4N_2O$	W_c	0.103	0.809	0.889	0.746	0.899	0.017	0.881	1.223	0.046
	W_k	0.087	0.960	0.968	0.855	0.980	0.012	0.428	0.377	0.038
$C_3H_5NO_2$	W_c	0.097	0.840	0.907	0.770	0.916	0.016	0.862	1.172	0.044
	W_k	0.094	0.947	0.968	0.852	0.973	0.012	0.533	0.588	0.039
$C_2H_8N_2O_3$	W_c	0.092	0.878	0.924	0.797	0.938	0.015	0.799	1.011	0.042
	W_k	0.081	0.953	0.969	0.857	0.976	0.010	0.497	0.478	0.035

B.10.2 Optimal Embedding Dimension

Table B.18: Optimal embedding dimension, d_e , corresponding normalized stress σ , and fractal dimension, d_f , for all ten test networks embedded using the distance-preserving encoder with W_c and W_k edge weights. d_e is the inflection point of the sigmoidal fit to $\frac{d}{dN_c} \log_2(\sigma)$; d_f is computed using the compact box-burning (CBB) algorithm.

Network	W_c			W_k		
	d_e	σ	d_f	d_e	σ	d_f
<chem>CH19NO9</chem>	6.15	0.098	2.65	4.78	0.091	2.17
<chem>C2H5NO2</chem>	7.20	0.107	2.69	6.10	0.115	2.22
<chem>C2H10O5</chem>	14.60	0.092	2.75	6.01	0.099	2.28
<chem>C2H6N2O2</chem>	10.31	0.098	3.27	9.15	0.089	2.68
<chem>C2H18O9</chem>	7.86	0.089	3.09	7.33	0.072	2.46
<chem>C3H6O3</chem>	9.17	0.104	3.17	8.51	0.098	2.61
<chem>C3H8O4</chem>	10.39	0.099	3.56	9.18	0.089	2.95
<chem>C3H4N2O</chem>	10.96	0.104	3.56	8.18	0.096	2.71
<chem>C3H5NO2</chem>	10.21	0.101	3.55	8.92	0.096	2.82
<chem>C2H8N2O3</chem>	11.08	0.094	3.63	9.30	0.086	2.95

B.11 Randomized tests

Tables B.19 and B.20 report encoder embedding metrics under randomization and Figures B.2–B.7 show scatter plots of original shortest-path distances versus embedded distances for $\text{CH}_{19}\text{NO}_9$ at $d = 3$ and $d = 10$. The original reaction network result (NO) is included for reference. Each randomization entry reports metrics from a single encoder run applied to three independently randomized networks. All metrics are defined in Section B.3.

These randomization schemes probe whether the near-Euclidean geometry observed in reaction networks reflects underlying chemical structure or instead emerges as a generic property of sparse graphs. Topology randomization (RC) rewires edges while preserving symmetric edge weights W , introducing shortcuts that tend to compress shortest-path distances. In contrast, weight randomization schemes (RW, RGW, RWU, RNR) preserve graph topology while modifying W . The NNZ and NA1–NA3 variants perturb RW weights via additive shifts. Symmetry refers to the W of $A \rightarrow B$ being equal to the W of $B \rightarrow A$.

- **NO**: original karc-weighted RN.
- **RC**: edges randomly rewired, symmetry preserved.
- **RW**: W shuffled, symmetry preserved; W range identical to original.
- **RGW**: W sampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$, symmetry enforced; negative values resampled.
- **RWU**: W sampled from $\mathcal{U}(w_{\text{min}}, w_{\text{max}})$ of the original W range, symmetry enforced.
- **RNR**: W sampled from the karc distribution independently per direction, breaking symmetry.
- **NNZ**: RW with zero-valued entries nudged by +0.01.
- **NA1**: RW with all weights shifted by +0.01.
- **NA2**: RW with all weights shifted by +0.1.
- **NA3**: RW with all weights shifted by +1.0.

Table B.19: Encoder metrics under randomization for $\text{CH}_{19}\text{NO}_9$ at $d = 3$. Randomization schemes are described at the beginning of Section B.11; metric definitions in Section B.3. NO denotes the original reaction network; all other entries report mean $\pm$ standard deviation over three runs.

Randomization	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_c$
NO	0.116	0.920	0.958	0.832	0.964	0.028	1.017	1.944	0.059
RC	3.4414 ± 1.9441	-20.3259 ± 22.9794	0.9916 ± 0.0007	0.9518 ± 0.0023	0.4141 ± 0.1109	5.95 ± 6.6001	0.7092 ± 0.2730	25.5564 ± 28.5802	0.1242 ± 0.0174
RW	1.5579 ± 0.2512	-2.9077 ± 1.2903	0.9889 ± 0.0010	0.9373 ± 0.0039	0.5344 ± 0.0397	0.9117 ± 0.0863	0.3934 ± 0.0220	5.2301 ± 1.6302	0.1379 ± 0.0018
RGW	0.1713 ± 0.0040	0.6622 ± 0.0188	0.8919 ± 0.0042	0.7210 ± 0.0049	0.9004 ± 0.0032	0.0444 ± 0.0008	1.4589 ± 0.0226	3.6396 ± 0.1100	0.0859 ± 0.0014
RWU	0.1842 ± 0.0044	0.6436 ± 0.0281	0.8808 ± 0.0058	0.7093 ± 0.0076	0.8903 ± 0.0080	0.0483 ± 0.0014	1.7585 ± 0.0158	5.4846 ± 0.1305	0.0909 ± 0.0025
RNR	0.9836 ± 0.3177	-0.8326 ± 1.0114	0.5547 ± 0.0182	0.4194 ± 0.0198	0.4647 ± 0.0598	119.68 ± 46.8492	0.6792 ± 0.0724	2.1941 ± 1.3135	0.5002 ± 0.0026
NNZ	1.2466 ± 0.9563	-3.1215 ± 4.7040	0.9690 ± 0.0045	0.8587 ± 0.0086	0.6567 ± 0.2385	0.5180 ± 0.5210	0.2847 ± 0.1198	3.7778 ± 4.2536	0.1260 ± 0.0106
NA1	1.0648 ± 0.5307	-1.5288 ± 2.3324	0.9691 ± 0.0040	0.8607 ± 0.0080	0.6199 ± 0.1668	0.3387 ± 0.2405	0.2312 ± 0.0227	2.3249 ± 2.1538	0.1251 ± 0.0209
NA2	0.1924 ± 0.0026	0.8051 ± 0.0084	0.9241 ± 0.0070	0.7759 ± 0.0113	0.9209 ± 0.0046	0.0492 ± 0.0017	0.3570 ± 0.0066	0.2696 ± 0.0111	0.0926 ± 0.0022
NA3	0.1682 ± 0.0015	0.6752 ± 0.0107	0.8955 ± 0.0024	0.7280 ± 0.0047	0.9035 ± 0.0025	0.0438 ± 0.0014	1.6903 ± 0.0284	5.0574 ± 0.0713	0.0838 ± 0.0018

Table B.20: Encoder metrics under randomization for $\text{CH}_{19}\text{NO}_9$ at $d = 10$. Randomization schemes are described at the beginning of Section B.11; metric definitions in Section B.3. NO denotes the original reaction network; all other entries report mean $\pm$ standard deviation over three runs.

Randomization	σ	R^2	ρ	τ	Pearson	NMSE	MAE	MSE	$\bar{d}_c$
NO	0.073	0.968	0.983	0.894	0.985	0.016	0.629	0.777	0.036
RC	4.6634 ± 1.7095	-34.2159 ± 24.3947	0.9660 ± 0.0229	0.9081 ± 0.0382	0.3404 ± 0.0775	6.20 ± 3.9285	0.9092 ± 0.2530	41.7204 ± 30.7955	0.1887 ± 0.0655
RW	0.8871 ± 0.5549	-0.5493 ± 1.6149	0.9785 ± 0.0121	0.9151 ± 0.0253	0.7169 ± 0.1745	0.3931 ± 0.2305	0.3434 ± 0.0808	2.0633 ± 2.1605	0.1597 ± 0.0453
RGW	0.0747 ± 0.0043	0.9356 ± 0.0082	0.9677 ± 0.0033	0.8466 ± 0.0076	0.9696 ± 0.0031	0.0104 ± 0.0005	0.6480 ± 0.0236	0.6921 ± 0.0561	0.0350 ± 0.0016
RWU	0.0863 ± 0.0009	0.9218 ± 0.0037	0.9604 ± 0.0019	0.8326 ± 0.0041	0.9629 ± 0.0020	0.0127 ± 0.0002	0.8239 ± 0.0026	1.2048 ± 0.0145	0.0389 ± 0.0004
RNR	1.0976 ± 0.6356	-1.5604 ± 2.5322	0.5562 ± 0.0467	0.4195 ± 0.0381	0.4497 ± 0.1047	125.68 ± 48.7228	0.6929 ± 0.0979	3.0895 ± 3.3588	0.4973 ± 0.0062
NNZ	2.4934 ± 2.3987	-17.9781 ± 26.0077	0.9808 ± 0.0034	0.8884 ± 0.0097	0.5511 ± 0.3052	1.72 ± 2.2041	0.3865 ± 0.2761	17.8215 ± 24.8341	0.0998 ± 0.0075
NA1	1.6784 ± 1.3661	-6.6489 ± 10.2492	0.9793 ± 0.0025	0.8840 ± 0.0063	0.5607 ± 0.2092	0.8499 ± 1.1261	0.2888 ± 0.1084	6.8874 ± 9.0647	0.1171 ± 0.0101
NA2	0.1287 ± 0.0029	0.9128 ± 0.0041	0.9729 ± 0.0014	0.8662 ± 0.0041	0.9670 ± 0.0005	0.0168 ± 0.0001	0.2217 ± 0.0049	0.1206 ± 0.0071	0.0493 ± 0.0002
NA3	0.0702 ± 0.0019	0.9433 ± 0.0041	0.9714 ± 0.0012	0.8552 ± 0.0032	0.9729 ± 0.0011	0.0094 ± 0.0002	0.7342 ± 0.0218	0.8818 ± 0.0464	0.0332 ± 0.0009

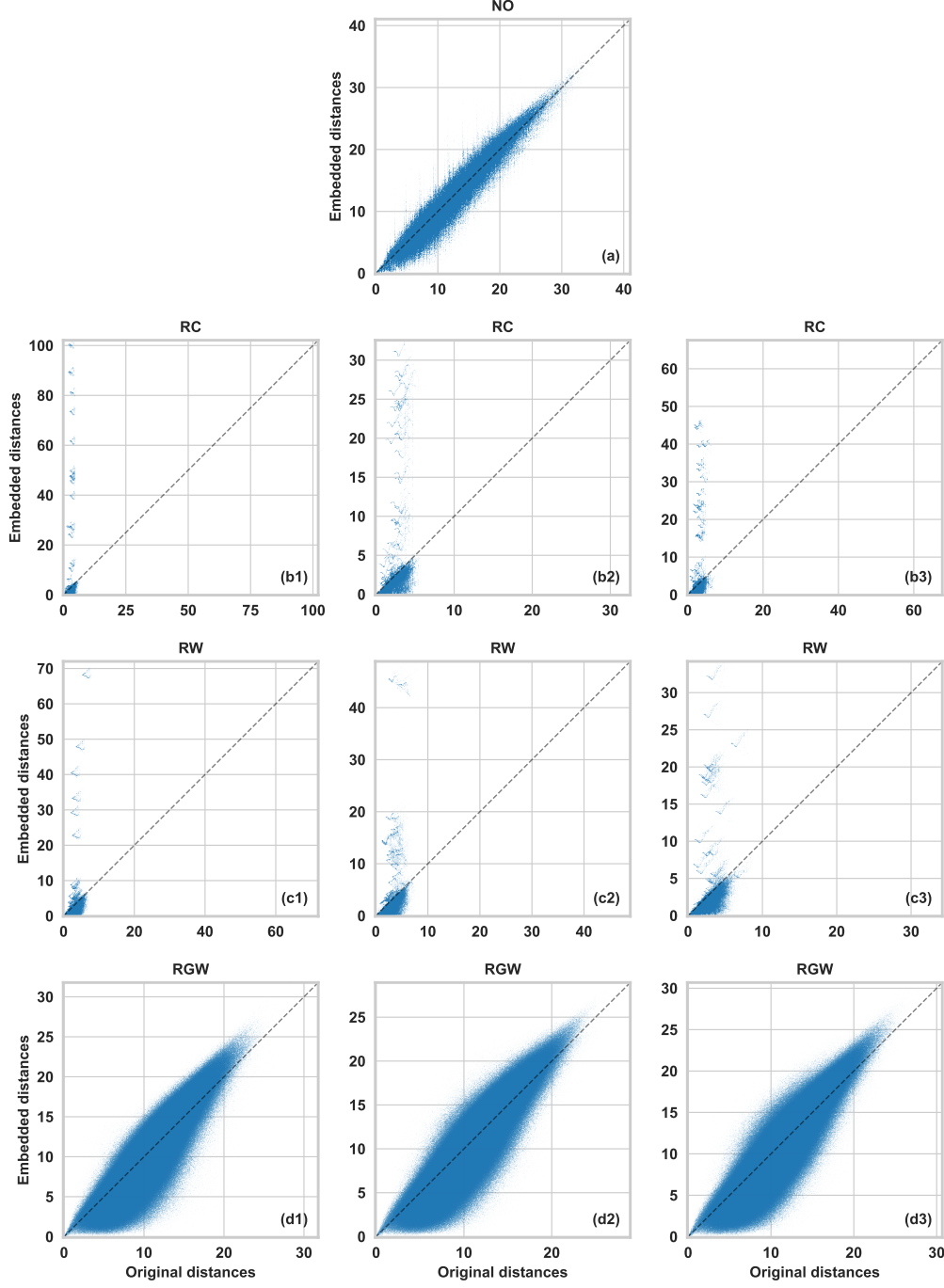


Figure B.2: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 3$, for three independent randomization runs each. The dashed line indicates perfect distance preservation. (a) Original karc-weighted network (NO). (b1)–(b3) Rewired-topology variant with symmetric edge weights retained (RC). (c1)–(c3) Symmetrically shuffled edge weights drawn from the original weight distribution (RW). (d1)–(d3) Symmetric edge weights resampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$, with negative and zero samples resampled until positive (RGW).

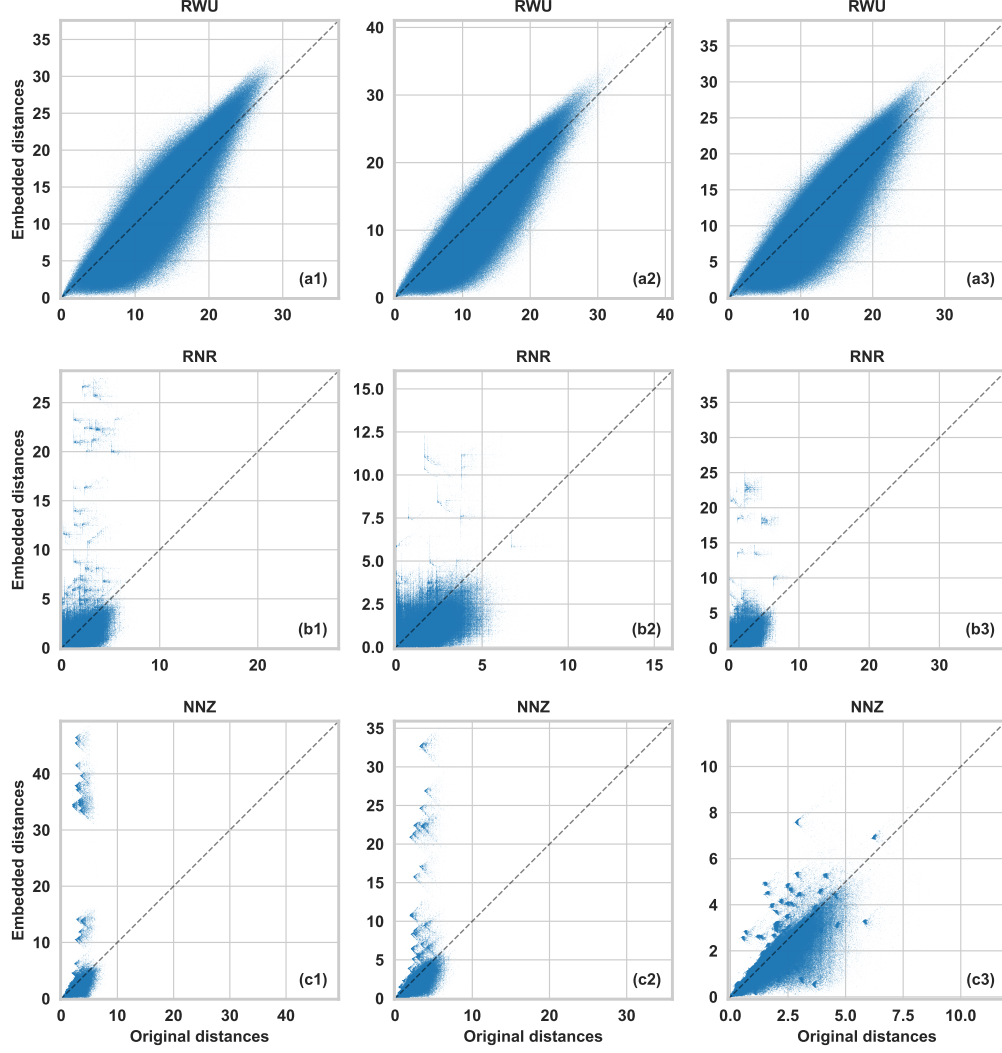


Figure B.3: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 3$, continued from Figure B.2. (a1)–(a3) Edge weights resampled from a uniform distribution $\mathcal{U}(w_{\min}, w_{\max})$ of the original weight range, symmetry enforced (RWU). (b1)–(b3) Non-reversibly randomized edge weights, with forward and reverse edges assigned independent random values, breaking the detailed-balance condition (RNR). (c1)–(c3) Symmetrically shuffled edge weights with zero-valued entries nudged to $+0.01$ (NNZ).

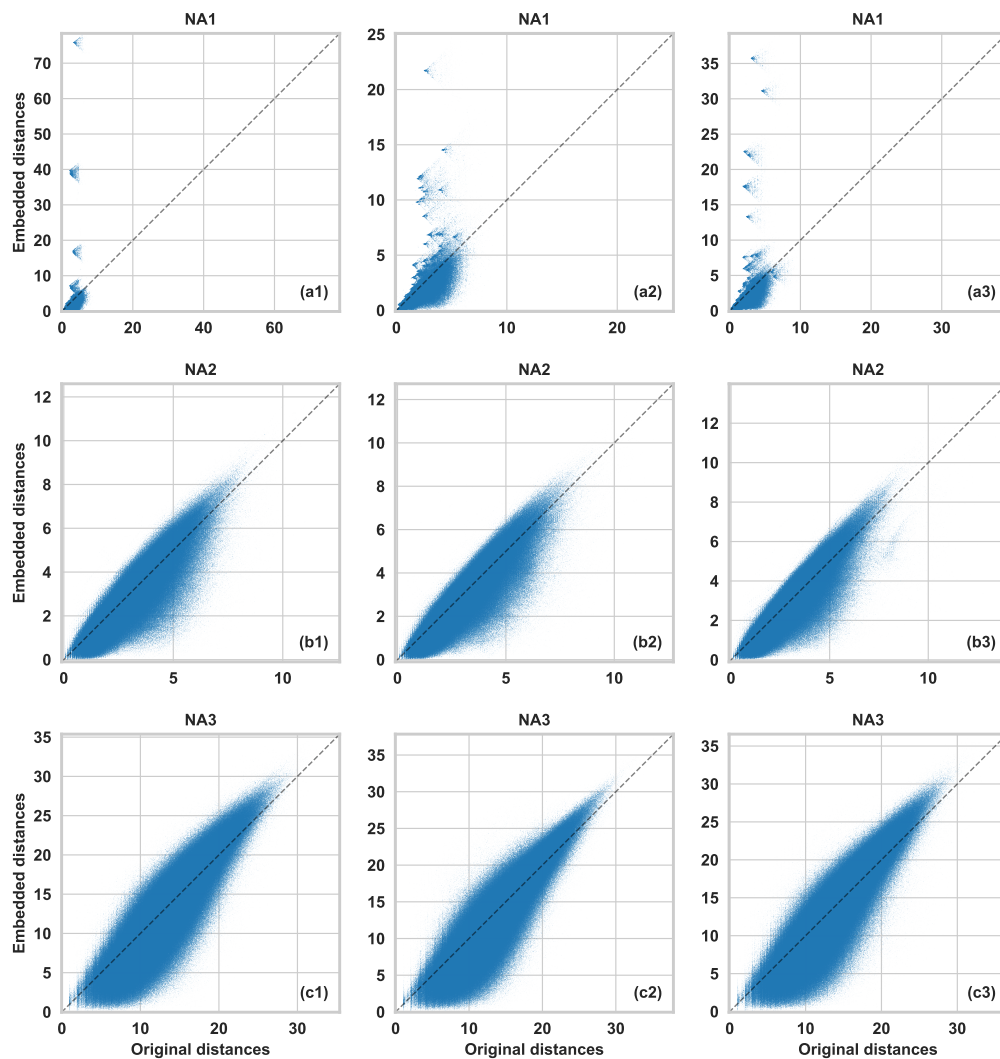


Figure B.4: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 3$, continued from Figure B.3. All panels show symmetrically shuffled edge weights with additive uniform shifts applied to all weights. (a1)–(a3) Shift of $+0.01$ (NA1). (b1)–(b3) Shift of $+0.1$ (NA2). (c1)–(c3) Shift of $+1.0$ (NA3).

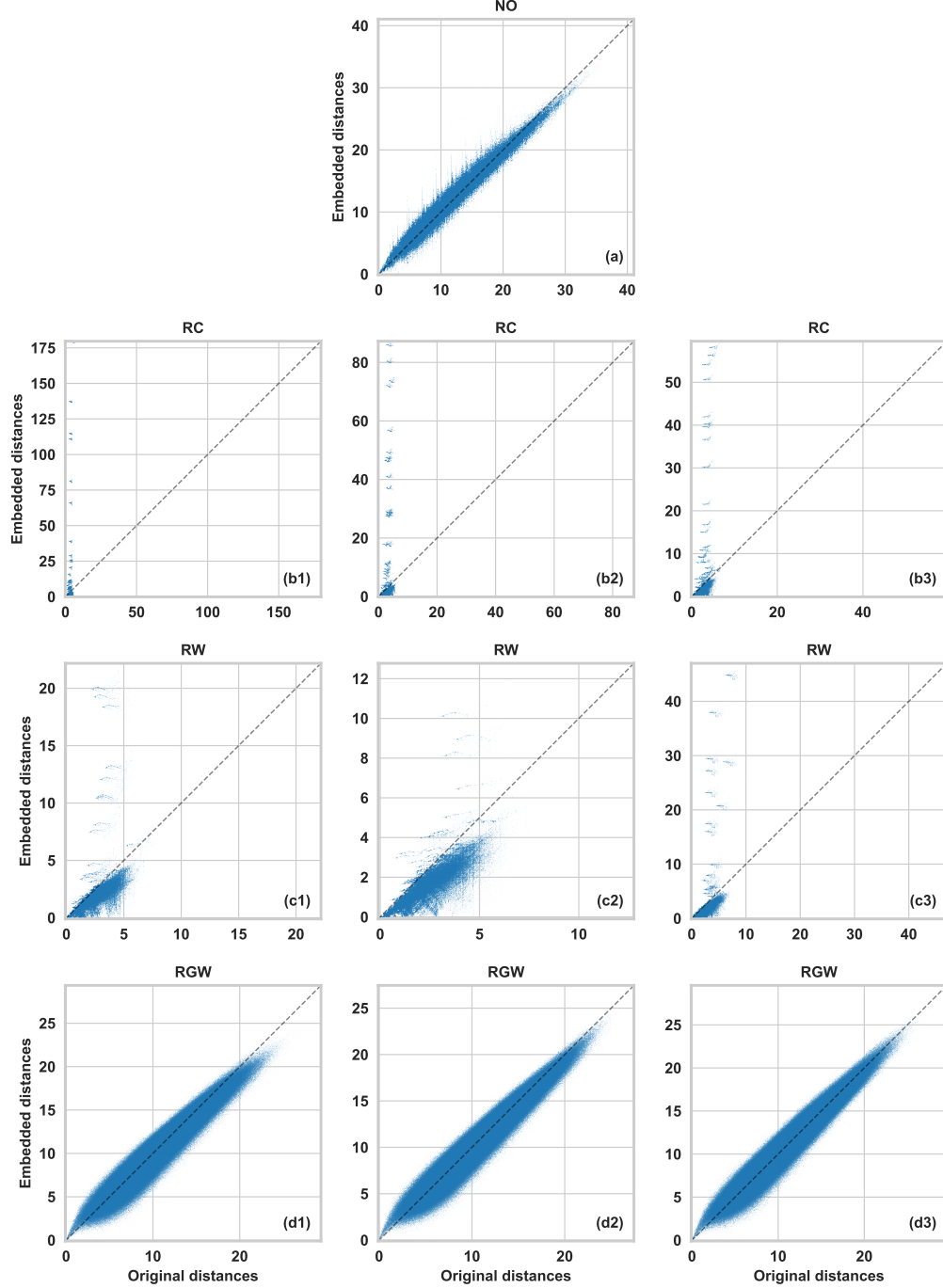


Figure B.5: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 10$, for three independent randomization runs each. The dashed line indicates perfect distance preservation. (a) Original karc-weighted network (NO). (b1)–(b3) Rewired-topology variant with symmetric edge weights retained (RC). (c1)–(c3) Symmetrically shuffled edge weights drawn from the original weight distribution (RW). (d1)–(d3) Symmetric edge weights resampled from $\mathcal{N}(\mu_{\text{orig}}, \sigma_{\text{orig}}^2)$, with negative and zero samples resampled until positive (RGW).

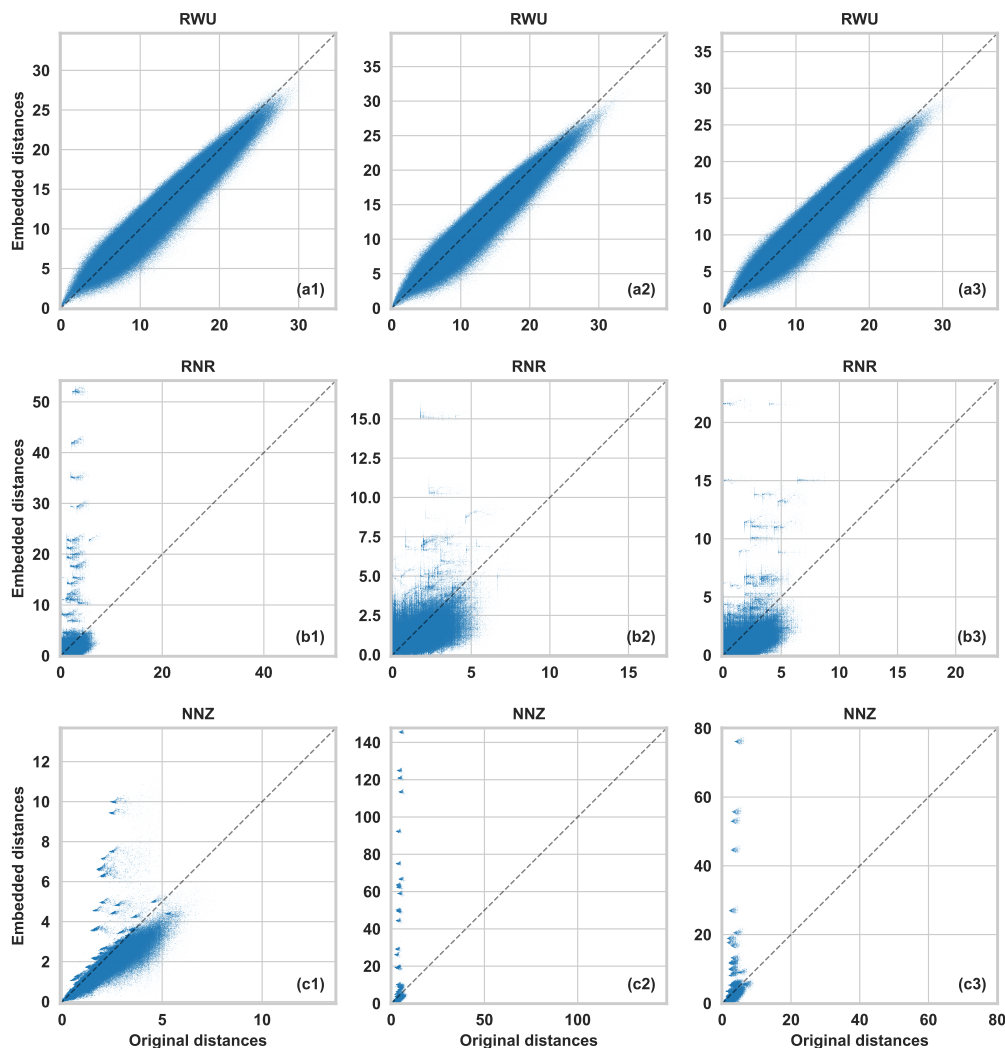


Figure B.6: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 10$, continued from Figure B.5. (a1)–(a3) Edge weights resampled from a uniform distribution $\mathcal{U}(w_{\min}, w_{\max})$ of the original weight range, symmetry enforced (RWU). (b1)–(b3) Non-reversibly randomized edge weights, with forward and reverse edges assigned independent random values, breaking the detailed-balance condition (RNR). (c1)–(c3) Symmetrically shuffled edge weights with zero-valued entries nudged to $+0.01$ (NNZ).

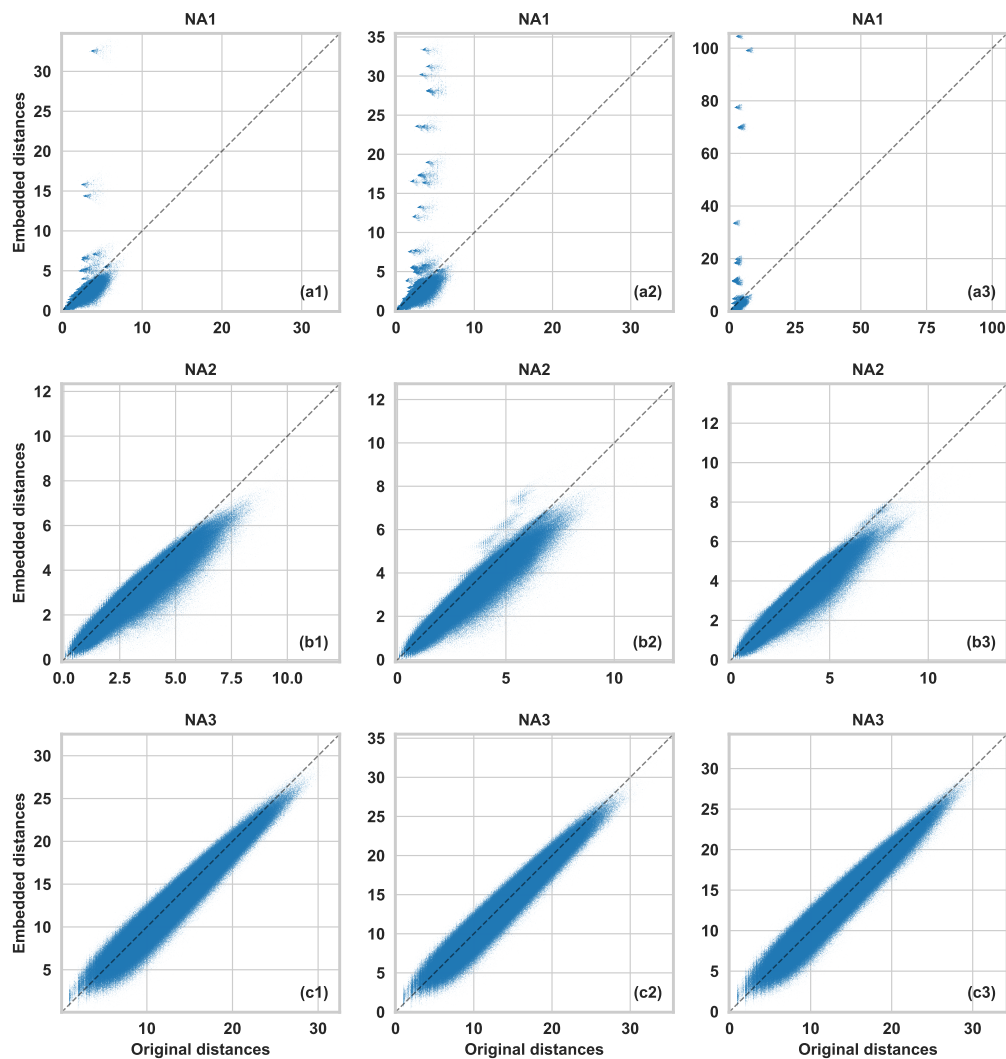


Figure B.7: Scatter plots of embedded versus original shortest-path distances for the $\text{CH}_{19}\text{NO}_9$ reaction network at $d = 10$, continued from Figure B.6. All panels show symmetrically shuffled edge weights with additive uniform shifts applied to all weights. (a1)–(a3) Shift of $+0.01$ (NA1). (b1)–(b3) Shift of $+0.1$ (NA2). (c1)–(c3) Shift of $+1.0$ (NA3).

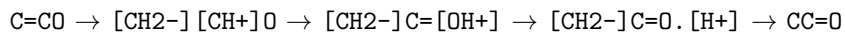
B.12 Pathway Discovery Costs

Pathway costs in the original reaction network (C) and embedded space ($\hat{C}$) are reported for all reference pathways across the five benchmark reactions. Within each reaction, pathways are ordered by increasing C . Multiple pathways sharing the same cost and same overall mechanism are represented by a single pathway with multiplicity m . The embedded rank gives the position of the pathway when all source-to-target paths are sorted by $\hat{C}$.

B.12.1 Enol Tautomerization ($\text{C}_2\text{H}_4\text{O}$)

ID	m	C	$\hat{C}$	Embedded rank
ET-0	5	3.751	4.348	1

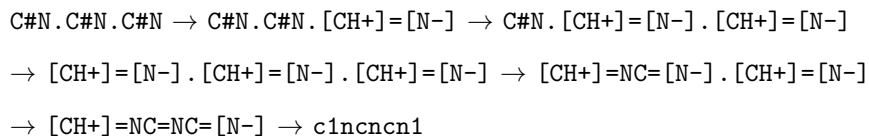
ET-0:



B.12.2 HCN Trimerization ($\text{C}_3\text{H}_3\text{N}_3$)

ID	m	C	$\hat{C}$	Embedded rank
HT-0	9	5.647	6.068	1

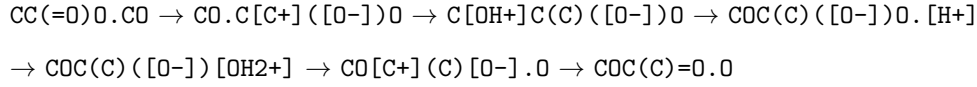
HT-0:



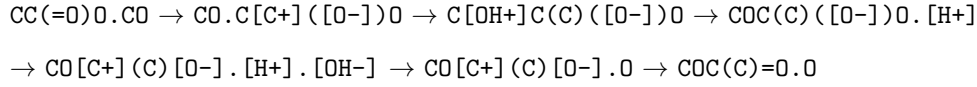
B.12.3 Esterification (C₃H₈O₃)

ID	m	C	$\hat{C}$	Embedded rank
ES-0	1	4.205	10.529	1
ES-1	2	5.064	12.307	7
ES-2	2	5.162	11.092	4
ES-3	3	5.808	10.789	2
ES-4	4	6.021	12.869	10
ES-5	3	6.218	11.930	6

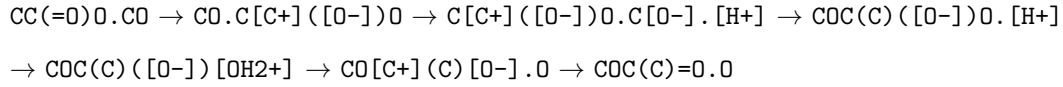
ES-0:



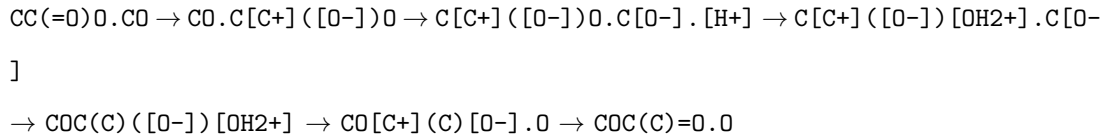
ES-1:



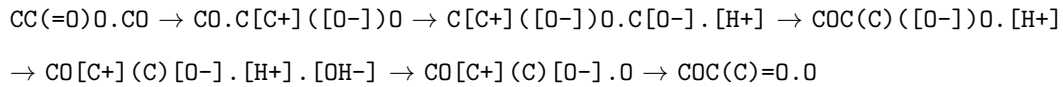
ES-2:



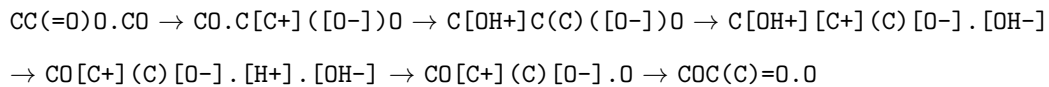
ES-3:



ES-4:



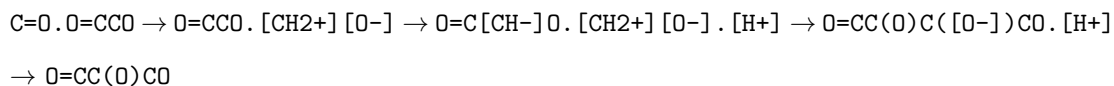
ES-5:



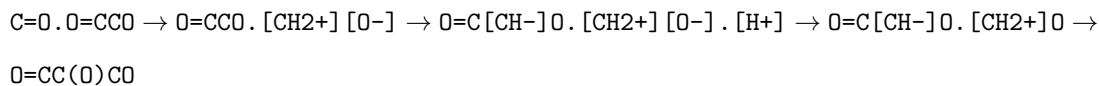
B.12.4 C₃ Aldol Addition (C₃H₆O₃)

ID	m	C	$\hat{C}$	Embedded rank
A3-0	2	4.567	6.522	2
A3-1	3	4.898	6.180	1

A3-0:



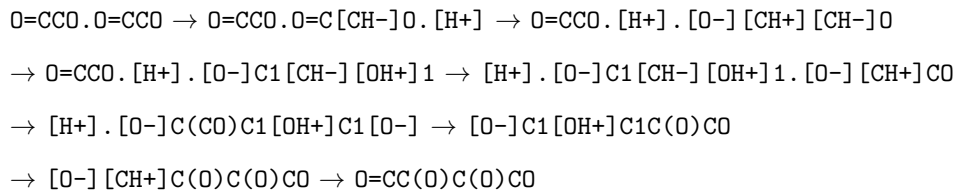
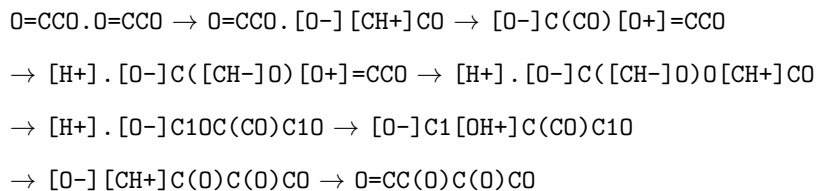
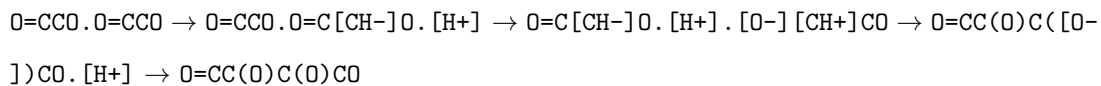
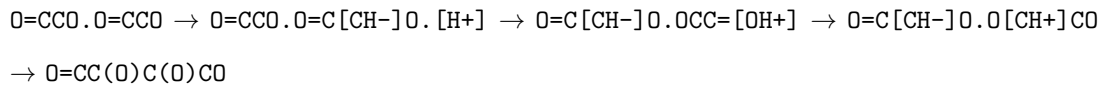
A3-1:



B.12.5 C₄ Aldol Addition (C₄H₈O₄)

Pathways A4-0 and A4-1 are the length-4 reference pathways used for recall evaluation. Several length-8 and length-6 pathways ranked lower than the reference pathways. The first two lower-cost pathways (A4-L0 and A4-L1) are shown here for reference.

ID	m	C	$\hat{C}$	Original rank	Embedded rank
A4-L0	7	4.333	N/A	1	N/A
A4-L1	3	4.365	N/A	2	N/A
A4-0	2	4.553	4.450	5	1
A4-1	3	4.922	5.595	23	2

A4-L0:**A4-L1:****A4-0:****A4-1:**

Appendix C

Supplementary Material for Chapter 4

Contents

C.1	Reaction Rule Set	201
C.2	Equivalence Rules Set	207
C.3	Energy Calculation Details	208
C.4	Bounded Reaction Network Workflow	209
C.4.1	Infrastructure	209
C.4.2	Stage 1: <code>initialize5</code>	210
C.4.3	Stage 2: <code>pipeline5</code>	211
C.4.4	Stage 3: <code>export5</code>	213
C.5	Exhaustive Reaction Network Statistics	213
C.6	Bounded Reaction Network Scaling	214
C.6.1	O_3 Bounded RN with $N_s = [\text{O-}] [\text{O+}] = 0$	214
C.6.2	O_4H_2 Bounded RN with $N_s = [\text{O-}] [\text{O+}] = 0.0$	216
C.6.3	CH_2O_2 Bounded RN with $N_s = \text{C} = [\text{O+}] [\text{O-}]$	218
C.6.4	CH_4O_3 Bounded RN with $N_s = \text{C} \text{O} \text{O} \text{O}$	220
C.6.5	CH_4O_3 Bounded RN with $N_s = \text{C} \text{O} [\text{O}] . [\text{OH}]$	222

C.6.6	CH_6O_4 Bounded RN with $N_s = \text{C000.0}$	224
C.6.7	CH_6O_4 Bounded RN with $N_s = \text{C0[0].[OH].0}$	225
C.6.8	$\text{C}_2\text{H}_4\text{O}_3$ Bounded RN with $N_s = \text{C=C.0=[O+][O-]}$	226
C.6.9	$\text{C}_2\text{H}_6\text{O}_4$ Bounded RN with $N_s = \text{C=C.0=[O+][O-].0}$	227
C.7	Exhaustive Reaction Network Scaling with $\Delta E_{\max}$	227

C.1 Reaction Rule Set

The 43,526 SMARTS-encoded [75] transformation rules are partitioned into four categories: 17,795 unimolecular transformations involving a single reactant, 5855 bimolecular transformations involving bond formation between two reactants, 4969 cyclization transformations corresponding to ring formation, and 14,907 ring-opening transformations. Tables C.2–C.6 summarize representative rule families rather than listing all instantiated SMARTS rules explicitly. The full SMARTS rule set is provided in the project repository [218].

The rules were generated through systematic enumeration of allowed atom types, bond orders, formal charge states, and radical assignments. Here, radical count refers to the number of atom-centered unpaired electrons represented within `RDKit`'s graph-based electron model [77] rather than physical spin multiplicity. Rule generation was constrained such that: (i) the sum of absolute formal charges on either side of a transformation does not exceed two; (ii) each side of a transformation contains at most four radical electrons in total; (iii) the change in total radical-electron count between the two sides of a transformation does not exceed two; (iv) the change in total formal charge between the two sides of a transformation does not exceed two; and (v) for transformations involving hydrogen atoms, radical-electron changes are distributed across atoms such that both electrons are not assigned to the same hydrogen center, whereas multiple radical electrons may be assigned to a single carbon or oxygen center. Enumerated atom types include carbon (formal charges 0, +1, -1, +2, -2), oxygen (0, +1, -1, -2), and hydrogen (0, +1, -1), with maximum bond counts and maximum radical-electron counts defined for each charge state (Table C.1).

Table C.1: Atom types enumerated in the CHO rule set, with the maximum number of bonds (*max bond*) and maximum number of radical electrons (*max radical*) permitted for each charge state.

Element	Charge state	Max bonds	Max radicals
Carbon	+2	2	2
	+1	3	3
	0	4	4
	-1	3	3
	-2	1	2
Oxygen	+1	3	3
	0	2	2
	-1	1	1
	-2	0	0
Hydrogen	+1	0	0
	0	1	1
	-1	1	1

All reaction rules preserve atomic stoichiometry and operate through redistribution of bond order, formal charge, radical-electron localization, and molecular connectivity. In generalized form, the transformations follow SMARTS templates such as `[A:1]-[B:2]>>[A:1].[B:2]` for bond dissociation and `[A:1].[B:2]>>[A:1]-[B:2]` for bond formation, with exhaustive enumeration over atom identities, bond orders, formal charges, and radical states. All transformations were ensured to be reversible, with each bond-breaking transformation paired to a corresponding bond-forming rule. The SMARTS patterns employ extended RDKit-compatible reaction SMARTS syntax [217] that allows explicit radical specification. Radical electrons are encoded through the extended SMARTS trailing-block notation `|^n:i,j,...|`, appended after the SMARTS string, where each clause `^n:i,j,...` specifies that atoms at zero-indexed positions `i,j,...` in the SMARTS pattern carry `n` unpaired electrons. For atoms on the reactant side of a rule, this notation acts as a matching criterion that selects molecules with the specified radical count. For atoms on the product side, it specifies the radical count assigned to the generated atom and prevents automatic implicit hydro-

gen assignment during SMARTS product generation by fixing the intended valence state. This is necessary because molecule sanitization in `RDKit` [77] can otherwise alter reaction product stoichiometry for nonstandard valence states through implicit hydrogen addition. These radical counts reflect `RDKit`'s graph-based electron-bookkeeping representation rather than physically observed spin multiplicities. Consequently, radical counts larger than those found in common stable species may appear in the enumerated rule set, such as an atomic carbon center being assigned four radical electrons rather than the two unpaired electrons expected from its quantum-mechanical ground state. Explicit atom-mapping labels (e.g., :1, :2) preserve atom identity between reactant and product patterns.

Unimolecular rules encode bond dissociation and bond-order interconversion for all CHO atom pairs (C-C, C-O, C-H, H-H, H-O, and O-O) in all charge and radical states permitted by the constraints above. Bond-order interconversions occur only between adjacent bond orders (single $\leftrightarrow$ double and double $\leftrightarrow$ triple). Both homolytic and heterolytic transformations are included. Homolytic transformations produce radical pairs through symmetric electron redistribution, whereas heterolytic transformations produce polarized, charge-separated products through asymmetric electron redistribution.

Bimolecular rules encode intermolecular bond formation including both homolytic radical recombination and heterolytic charge-neutralizing association processes.

Cyclization rules are formally identical to the bimolecular rules but use component-grouped SMARTS syntax, $([A:1].[B:2])>>[A:1]-[B:2]$, which restricts rule application to reactive centers within the same molecular graph and thereby captures intramolecular ring closure. Ring-closure patterns involving hydrogen atoms are excluded because they are uncommon in stable cyclic structures.

Ring-opening rules require separate treatment because the acyclic bond-dissociation rules do not apply to cyclic bonds within `RDKit`'s reaction SMARTS matching framework. The 14,907

ring-opening rules therefore employ explicit ring SMARTS patterns with wildcard bridging atoms ($[*:n]$) and wildcard bond variables ($[~]$). These rules enumerate ring opening in three- to five-membered heavy-atom rings across all permitted charge and radical states. Ring-opening rules involving hydrogen atoms are excluded because they are uncommon in stable cyclic structures.

Table C.2: Representative C–C transformation rule families illustrating homolytic and heterolytic bond dissociation, association, and bond-order interconversion processes.

Homolytic dissociation	Homolytic association	Heterolytic dissociation	Heterolytic association
$C - C \rightarrow C^\bullet + C^\bullet$	$C^\bullet + C^\bullet \rightarrow C - C$	$C - C \rightarrow C^- + C^+$	$C^- + C^+ \rightarrow C - C$
$C = C \rightarrow C - C^{\bullet\bullet}$	$C - C^{\bullet\bullet} \rightarrow C = C$	$C = C \rightarrow C^- - C^+$	$C^- - C^+ \rightarrow C = C$
$C \equiv C \rightarrow C = C^{\bullet\bullet}$	$C = C^{\bullet\bullet} \rightarrow C \equiv C$	$C \equiv C \rightarrow C^- = C^+$	$C^- = C^+ \rightarrow C \equiv C$
$C^- - C^- \rightarrow C^{\bullet-} + C^{\bullet-}$	$C^{\bullet-} + C^{\bullet-} \rightarrow C^- - C^-$	$C^- - C^- \rightarrow C^{2-} + C$	$C^{2-} + C \rightarrow C^- - C^-$

Each row represents a rule family systematically expanded across all allowed charge, radical, and bond-order states consistent with the electronic-state constraints described in the text. Additional variants, including charged radicals and symmetry-related electronic configurations, are omitted for brevity.

Table C.3: Representative C–H and H–H transformation rule families illustrating homolytic and heterolytic bond dissociation and association processes.

Homolytic dissociation	Homolytic association	Heterolytic dissociation	Heterolytic association
$C - H \rightarrow C^\bullet + H^\bullet$	$C^\bullet + H^\bullet \rightarrow C - H$	$C - H \rightarrow C^- + H^+$	$C^- + H^+ \rightarrow C - H$
$C^- - H \rightarrow C^{\bullet-} + H^\bullet$	$C^{\bullet-} + H^\bullet \rightarrow C^- - H$	$C - H \rightarrow C^+ + H^-$	$C^+ + H^- \rightarrow C - H$
$C^+ - H \rightarrow C^{\bullet+} + H^\bullet$	$C^{\bullet+} + H^\bullet \rightarrow C^+ - H$	$C^- - H \rightarrow C + H^-$	$C + H^- \rightarrow C^- - H$
$H - H \rightarrow H^\bullet + H^\bullet$	$H^\bullet + H^\bullet \rightarrow H - H$	$H - H \rightarrow H^- + H^+$	$H^- + H^+ \rightarrow H - H$

Additional charge-separated, radical-containing, and reverse-association variants are generated systematically but are not shown explicitly.

Table C.4: Representative C–O and H–O transformation rule families illustrating bond dissociation and bond-order interconversion processes across multiple charge and radical states.

Homolytic dissociation	Homolytic association	Heterolytic dissociation	Heterolytic association
$C-O \rightarrow C^\bullet + O^\bullet$	$C^\bullet + O^\bullet \rightarrow C-O$	$C-O \rightarrow C^- + O^+$	$C^- + O^+ \rightarrow C-O$
$C-O^+ \rightarrow C^\bullet + O^{\bullet+}$	$C^\bullet + O^{\bullet+} \rightarrow C-O^+$	$C-O \rightarrow C^+ + O^-$	$C^+ + O^- \rightarrow C-O$
$C=O \rightarrow C-O^{\bullet\bullet}$	$C-O^{\bullet\bullet} \rightarrow C=O$	$C=O \rightarrow C^- - O^+$	$C^- - O^+ \rightarrow C=O$
$C=O^+ \rightarrow C-O^{\bullet\bullet+}$	$C-O^{\bullet\bullet+} \rightarrow C=O^+$	$C=O \rightarrow C^+ - O^-$	$C^+ - O^- \rightarrow C=O$
$C\equiv O^+ \rightarrow C=O^{\bullet\bullet+}$	$C=O^{\bullet\bullet+} \rightarrow C\equiv O^+$	$C\equiv O^+ \rightarrow C^- = O^{2+}$	$C^- = O^{2+} \rightarrow C\equiv O^+$
$H-O \rightarrow H^\bullet + O^\bullet$	$H^\bullet + O^\bullet \rightarrow H-O$	$H-O \rightarrow H^+ + O^-$	$H^+ + O^- \rightarrow H-O$
$H-O^+ \rightarrow H^\bullet + O^{\bullet+}$	$H^\bullet + O^{\bullet+} \rightarrow H-O^+$	$H-O^+ \rightarrow H^+ + O$	$H^+ + O \rightarrow H-O^+$

The complete rule set additionally includes systematically enumerated electronic-state variants involving charged radicals, hypervalent oxygen centers, and all symmetry-distinct charge distributions permitted by the generation constraints.

Table C.5: Representative O–O transformation rule families illustrating peroxide-, superoxide-, and ozonide-type bond transformations across multiple bond orders, charge states, and radical configurations.

Homolytic dissociation	Homolytic association	Heterolytic dissociation	Heterolytic association
$O-O \rightarrow O^\bullet + O^\bullet$	$O^\bullet + O^\bullet \rightarrow O-O$	$O-O \rightarrow O^- + O^+$	$O^- + O^+ \rightarrow O-O$
$O-O^+ \rightarrow O^\bullet + O^{\bullet+}$	$O^\bullet + O^{\bullet+} \rightarrow O-O^+$	$O^- - O^- \rightarrow O^{2-} + O$	$O^{2-} + O \rightarrow O^- - O^-$
$O^- - O \rightarrow O^{\bullet-} + O^\bullet$	$O^{\bullet-} + O^\bullet \rightarrow O^- - O$	$O^- - O^+ \rightarrow O + O$	$O + O \rightarrow O^- - O^+$
$O^+ - O^+ \rightarrow O^{\bullet+} + O^{\bullet+}$	$O^{\bullet+} + O^{\bullet+} \rightarrow O^+ - O^+$		
$O=O \rightarrow O-O^{\bullet\bullet}$	$O-O^{\bullet\bullet} \rightarrow O=O$	$O=O \rightarrow O^- - O^+$	$O^- - O^+ \rightarrow O=O$
$O=O^+ \rightarrow O-O^{\bullet\bullet+}$	$O-O^{\bullet\bullet+} \rightarrow O=O^+$	$O^+ = O^+ \rightarrow O-O^{2+}$	$O-O^{2+} \rightarrow O^+ = O^+$
$O^+ \equiv O^+ \rightarrow O^+ = O^{\bullet\bullet+}$	$O^+ = O^{\bullet\bullet+} \rightarrow O^+ \equiv O^+$	$O^+ \equiv O^+ \rightarrow O=O^{2+}$	$O=O^{2+} \rightarrow O^+ \equiv O^+$

Additional rule variants include all permitted combinations of oxygen charge states, radical multiplicities, and adjacent bond-order interconversions generated under the electronic-state constraints.

Table C.6: Representative ring-opening and cyclization transformation rule families. Ring-opening rules use explicit cyclic SMARTS patterns with wildcard atoms to enumerate three- to five-membered heavy-atom rings in all permitted charge, radical, and bond-order states. Cyclization rules are the intramolecular analogs of the bimolecular bond-formation rules and are restricted to reactive centers within the same molecular graph.

Ring opening (228 rules)	Ring closing – cyclization (117 rules)
Homolytic single-bond cleavage: $A^{\bullet}(-X-)B^{\bullet} \rightarrow A^{\bullet}-X-B^{\bullet}$	Homolytic cyclization: $A^{\bullet}-X-B^{\bullet} \rightarrow A^{\bullet}(-X-)B^{\bullet}$
Heterolytic single-bond cleavage: $A^{-}(-X-)B^{+} \rightarrow A^{-}-X-B^{+}$	Heterolytic cyclization: $A^{-}-X-B^{+} \rightarrow A^{-}(-X-)B^{+}$
Homolytic bond-order decrease: $A(=X-)B^{\bullet\bullet} \rightarrow A=X-B^{\bullet\bullet}$	Homolytic bond-order increase: $A=X-B^{\bullet\bullet} \rightarrow A(=X-)B^{\bullet\bullet}$
Heterolytic bond-order decrease: $A^{-}(=X-)B^{+} \rightarrow A^{-}=X-B^{+}$	Heterolytic bond-order increase: $A^{-}=X-B^{+} \rightarrow A^{-}(=X-)B^{+}$
3–5 membered ring topologies and all permitted charge/radical variants are enumerated systematically. Here, A, B, X, Y, and Z denote arbitrary heavy atoms (C or O), corresponding to wildcard SMARTS atoms ([*]).	Cyclization rules apply the same transformations intramolecularly, whereas the corresponding bimolecular rules apply to reactive centers on separate molecular fragments.

Each schematic row represents a generalized SMARTS transformation family expanded systematically across allowed ring sizes, atom identities, charge states, radical configurations, and adjacent bond-order changes. The table therefore summarizes representative rule families rather than explicitly listing all instantiated ring-opening and cyclization SMARTS rules.

C.2 Equivalence Rules Set

The 11,940 SMARTS-encoded [75] equivalence rules form a subset of the unimolecular transformation rules described in Section C.1. These rules preserve atomic connectivity and molecular stoichiometry while redistributing bond order, formal charge, and radical-electron localization. Unlike the dissociation and association rules, equivalence rules do not create or destroy bonds.

The equivalence rules enumerate electronically allowed interconversions for bonded C–C, C–O, and O–O atom pairs consistent with the electronic-state constraints described in Section C.1. These transformations represent resonance-like valence reorganizations, including adjacent bond-order interconversions (single $\leftrightarrow$ double and double $\leftrightarrow$ triple), charge redistribution, and radical delocalization.

As in the full reaction-rule set, the equivalence rules employ extended RDKit-compatible SMARTS syntax [217] with explicit atom mapping, formal charge specification, and radical-electron annotations. Representative equivalence-rule families are summarized in Table C.7. The complete set of 11,940 instantiated SMARTS equivalence rules is provided in the project repository [218]. All equivalence zero-energy transformations were ensured to be reversible, with each bond order decrease transformation paired to a corresponding bond order increase rule.

Table C.7: Representative equivalence-rule families illustrating bond-order redistribution, charge transfer, and radical delocalization transformations that preserve atomic connectivity and molecular stoichiometry.

Bond-order redistribution	Charge redistribution	Radical delocalization
$C - C^{\bullet\bullet} \leftrightarrow C = C$	$C^- - O^+ \leftrightarrow C = O$	$C^\bullet - O \leftrightarrow C - O^\bullet$
$C = C^{\bullet\bullet} \leftrightarrow C \equiv C$	$O^- - O^+ \leftrightarrow O = O$	$O^\bullet - O \leftrightarrow O - O^\bullet$
$C^- = O^+ \leftrightarrow C \equiv O^+$	$C^- = C^+ \leftrightarrow C \equiv C$	$C^{\bullet-} - C \leftrightarrow C - C^{\bullet-}$

Each row represents a generalized equivalence-rule family systematically expanded across allowed atom identities, bond orders, charge states, and radical configurations consistent with the electronic-state constraints. The table summarizes representative rule families rather than explicitly listing all instantiated SMARTS equivalence rules.

C.3 Energy Calculation Details

Node energies are the sum of representative molecular energies assigned to each equivalence class, as described in the main text. The 11,940 unimolecular equivalence rules covering all pairwise bond-order, formal charge, and radical interconversions for C-C, C-O, and O-O bonded pairs are described in Section C.2. Nodes for which any molecular energy calculation failed were removed together with their associated edges.

Molecular energies were computed from SMILES strings as follows. Three-dimensional geometries were generated using RDKit [77], with 1000 conformers produced by the ETDG algorithm [248] (maximum 5000 embedding iterations). All conformers were optimized with the UFF force field [249] (maximum 3000 steps), and the lowest-energy conformer was submitted to MOPAC [155] for geometry optimization and single-point energy evaluation. MOPAC calculations used the PM7 semiempirical Hamiltonian with solvent effects described by COSMO [250] at $\varepsilon = 78.4$, a gradient norm convergence threshold of 0.0, an SCF

convergence criterion of 10^{-9} , and a maximum of 1000 optimization cycles. If the selected conformer yielded an unphysically positive molecular energy, up to four additional conformers were resubmitted to MOPAC before the calculation was marked as failed. Further details of the energy calculation implementation in `colibri2` are given in previous work [24].

C.4 Bounded Reaction Network Workflow

The bounded generation workflow in `colibri2` consists of three main stages. They are `initialize5`, `pipeline5`, and `export5`. This section describes the underlying infrastructure, the logic of each worker process, and the data structures used for inter-worker communication. In the implementation, a reaction-network *node* is a flask (superformula) stored as a `Supergeometry` row, and *edges* are reaction *transformations* recorded in the transformation table together with reaction-rule identifiers.

C.4.1 Infrastructure

Persistent network data are stored in a PostgreSQL [73] relational database. The database contains four primary tables, consisting of valid nodes (`Supergeometry`), invalid nodes (`InvalidSuperformula`), edges (`transformations`), and reaction rules. Relational storage normalizes repeated data across tables, reducing disk footprint relative to flat-file representations. Optional metadata may also be stored alongside each node.

Inter-worker communication uses Redis [72] as both a message broker and an in-memory store. Three queues and three molecule/flask caches are central to generation, and separate Redis *registries* track in-flight work and timeouts (Table C.8). The flask-transformation queue holds items consisting of a flask, the transformation that produced it (`None` for the seed flask), a timestamp used for lease-based retries, and a generation index. The equiva-

lence queue holds formula SMILES awaiting equivalence-class processing by `equivalence5`. The geometry queue holds representative SMILES awaiting energy calculation by `compute5`. The flask cache stores per-flask aggregated energies and status (`CachedSupergeometry`). The equivalence-formula cache stores the representative energy and status for each formula SMILES (`CachedEquivalenceFormula`). The geometry cache stores per-representative energies and calculation status (`CachedGeometry`). The running-geometry registry records in-flight energy jobs (with expiring keys) so `compute5` does not duplicate work and so `reaper5` can detect stalls. `network5` additionally uses registries for lease-timeouts and for a heartbeat on active network expansion.

Table C.8: Primary Redis queues, caches, and registries during `pipeline5` execution.

Name	Type	Purpose
Flask–transformation queue	Queue	Flasks and producing transformation pending expansion by <code>network5</code>
Equivalence queue	Queue	Formulas pending equivalence processing by <code>equivalence5</code>
Geometry queue	Queue	Representatives pending energy calculation by <code>compute5</code>
Flask cache	Cache	Per-flask aggregated energies and status
Equivalence cache	Cache	Per-formula representative energy and status
Geometry cache	Cache	Per-representative energy and calculation status
Running-geometry registry	Registry	In-flight energy jobs; expiry drives <code>reaper5</code>

C.4.2 Stage 1: `initialize5`

`initialize5` performs one-time setup of the workflow state. The relevant PostgreSQL tables are cleared. All Redis queues, caches, and registries used by the workflow are flushed. The seed flask at generation $g = 0$, paired with a `None` transformation, is appended to the flask–transformation queue. Each distinct formula SMILES from the seed flask is simultaneously inserted into the equivalence queue so that equivalence-level energies are available before the first expansion step.

C.4.3 Stage 2: pipeline5

pipeline5 runs proton5 to completion, then starts network5, equivalence5, compute5, and reaper5 in parallel. If any of these processes exits with a non-zero status, the wrapper terminates the remaining workers.

proton5 computes the $[H^+]$ reference molecular energy as the difference of hydronium and water molecular energies at the configured electronic-structure level, and writes the result to the equivalence cache (and to the geometry cache under the $[H^+]$ SMILES). This is separated from the main workflow as this is a unique case of defining a molecular energy as a difference of two other molecular energies, and this only needs to be done once per workflow execution.

network5 is the central expansion worker. For each iteration, a flask-transformation item is dequeued. For the multiset of formula SMILES in that flask, the equivalence cache is checked. If any entry is missing, those formulas are added to the equivalence queue and the flask item is requeued for later processing. Lease timers and a pending-flask registry avoid indefinite waits when no geometry jobs are running. Repeated failures can mark the flask as timed out.

Let c_m be the stoichiometric count of formula m in the flask. Once all molecular energies are available, the flask energy is computed as $E_K = \sum_m c_m \mu_m$, where μ_m is the cached representative energy for m . If any μ_m has a non-success status, E_K is set to null and the failure status is propagated. When a transformation is present, the energy difference $|E_{K'} - E_K|$ between the product flask K' and the reactant flask K is evaluated against $\Delta E_{\max}$ when that cutoff is configured. If the criterion fails, the product flask and the transformation are not stored and the flask is not queued for further expansion (avoiding disconnected growth), but it may be expanded later if a valid edge is generated. If the criterion is satisfied

(or no cutoff is set), the flask and transformation are stored in PostgreSQL as appropriate.

If the current generation g has not yet reached $G_{\max}$, reaction rules are applied to K to generate product flasks $\{K'\}$, and each child is assigned generation $g + 1$. Products that fail `RDKit` [77] validation are stored in the invalid-node table and are not expanded further. Products with undefined energies can be excluded from further expansion depending on configuration (e.g. skipping database storage when energies are missing).

equivalence5 processes formulas from the equivalence queue. For each formula, the configured zero-energy equivalence rules (Section C.2) are applied to enumerate its equivalence class. Representatives are chosen to maximize total bond order (among equivalent structures) and, on ties, to minimize the sum of absolute formal charges. For each tied representative, the geometry cache is checked. Representatives with cached successful energies are used directly, and missing representatives are queued on the geometry queue for **compute5**. Once all required representative energies are available, the lowest energy among successful results is assigned to the equivalence class and written to the equivalence cache.

compute5 processes entries on the geometry queue. Before starting a calculation, the running-geometry registry is checked to avoid duplicate submissions. Molecular geometries are built and optimized using the configured programs (e.g. MOPAC [155] or XTB [251, 252]), and molecular energies (or configured property energies) are written with status (success, failure, or timeout) to the geometry cache.

reaper5 listens for Redis expiry events on running-geometry keys. When a job key expires, the corresponding formula is requeued up to a maximum number of retries if the geometry cache still holds no successful result. Once retries are exhausted, the geometry cache is updated with a timeout status. The process also monitors overall queue and registry idle

state and exits when the workflow is inactive.

C.4.4 Stage 3: `export5`

Once `pipeline5` is complete, `export5` reads stored nodes, edges, and reaction rules from PostgreSQL and writes the network to GraphML [76]. Stream, multi-stream, and NetworkX-based writers are available. When the output path uses a `.bz2` suffix, writers emit bzip2-compressed GraphML.

C.5 Exhaustive Reaction Network Statistics

Table C.9: Exhaustive reaction networks (RNs) before and after energy validation. N_f : full-graph nodes; E_f : full-graph edges; N_v : energy-validated nodes with failed-energy nodes removed; E_v : energy-validated edges with failed-energy edges removed; M_f : unique molecules in the full network; M_v : molecules with validated energies.

Stoichiometry	N_f	E_f	N_v	E_v	M_f	M_v
O ₃	13	40	13	40	10	14
O ₄ H ₂	2922	26,702	760	5352	970	192
CH ₂ O ₂	2460	17,952	878	5324	791	398
CH ₄ O ₃	281,005	3,702,400	23,527	239,332	58,821	3047

C.6 Bounded Reaction Network Scaling

C.6.1 O₃ Bounded RN with $N_s = [0-] [0+] = 0$

Table C.10: Network statistics for O₃ bounded RN with starting node $N_s = [0-] [0+] = 0$ at varying maximum generation depths $G_{\max}$ and allowed molecular energy difference between connected nodes of $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. Notation: n : nodes; e : edges; M : unique molecules; Q_M : molecule equivalence classes; N_N : neutral nodes; Q_N : neutral nodes collapsed by equivalence.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	3	4	4	3	2	1	1	3	4	4	3	2	1
2	5	10	7	5	3	2	2	5	10	7	5	3	2
3	9	18	11	7	7	4	3	9	18	11	7	7	4
4	11	30	14	10	8	5	4	11	30	14	10	8	5
5	11	30	14	10	8	5	5	12	32	14	10	8	5

Table C.11: Network statistics for O₃ bounded RN with $N_s = [0-] [0+] = 0$ and unrestricted $\Delta E_{\max}$ at varying $G_{\max}$). Notation: N_v : energy-validated nodes with failed-energy nodes removed; E_v : energy-validated edges with failed-energy edges removed; N_f : full-graph nodes before energy validation; E_f : full-graph edges before energy validation. See Table C.10 for notation.

$G_{\max}$	N_v	E_v	N_f	E_f	M	Q_M	N_N	Q_N
1	3	4	3	4	4	3	2	1
2	5	10	5	10	7	5	3	2
3	11	22	11	22	12	8	8	5
4	13	40	13	40	14	10	9	6
5	13	40	13	40	14	10	9	6

Table C.12: Network statistics for O_3 bounded RN with $N_s = [0-] [0+] = 0$ and unlimited $G_{\max}$ at varying $\Delta E_{\max}$). See Table C.10 for notation.

$\Delta E_{\max}$ (eV)	n	e	M	Q_M	N_N	Q_N
1	2	2	2	1	2	1
1.5	2	2	2	1	2	1
2	2	2	2	1	2	1
2.5	2	2	2	1	2	1
3	11	30	14	10	8	5
3.5	11	30	14	10	8	5
4	12	32	14	10	8	5
4.5	13	36	14	10	9	6
5	13	36	14	10	9	6

C.6.2 O₄H₂ Bounded RN with $N_s = [0-] [0+] = 0.0$

Table C.13: Network statistics for O₄H₂ bounded RN with $N_s = [0-] [0+] = 0.0$ at varying $G_{\max}$ and $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	4	6	7	6	2	1	1	4	6	7	6	2	1
2	10	24	12	10	3	2	2	10	24	12	10	3	2
3	22	60	22	16	10	7	3	24	64	24	18	10	7
4	47	150	38	32	22	15	4	55	172	45	38	23	16
5	104	364	65	58	41	30	5	133	478	75	65	48	35
6	170	734	79	72	55	42	6	221	1090	89	78	68	50
7	228	1062	91	80	70	54	7	290	1574	98	85	83	60
8	273	1286	101	89	80	61	8	341	1878	104	91	91	67
9	315	1480	103	89	89	64	9	384	2110	112	98	93	68
10	346	1620	105	91	92	67	10	415	2290	119	102	94	68
11	379	1752	112	98	93	68	11	436	2402	124	105	94	68
12	401	1864	119	102	94	68	12	446	2470	126	105	94	68
13	426	1982	123	104	94	68	13	452	2504	128	106	94	68
14	438	2030	126	105	94	68	14	453	2516	129	107	94	68
15	441	2046	126	105	94	68	15	454	2520	130	107	94	68

Table C.14: Network statistics for O₄H₂ bounded RN with $N_s = [0-] [0+] = 0.0$ and unrestricted $\Delta E_{\max}$ at varying $G_{\max}$). See Table C.10 and Table C.11 for notation.

$G_{\max}$	N_v	E_v	N_f	E_f	M	Q_M	N_N	Q_N
1	5	8	5	8	9	8	3	2
2	14	38	18	50	14	12	5	3
3	50	134	80	248	37	31	18	14
4	144	506	249	1006	73	63	45	35
5	320	1486	639	3304	121	102	81	62
6	492	2938	1208	7640	145	117	101	77
7	633	4092	1854	13,570	167	136	117	85
8	726	4984	2368	19,080	178	143	122	88
9	780	5468	2700	22,956	186	150	122	87
10	805	5686	2848	25,076	192	153	122	88
11	808	5768	2895	25,880	192	153	122	87
12	810	5776	2910	26,236	192	153	122	86
13	814	5792	2916	26,322	192	153	122	87
14	814	5794	2920	26,208	192	153	122	86
15	814	5794	2922	26,432	192	153	122	87

Table C.15: Network statistics for O₄H₂ bounded RN with $N_s = [0-] [0+] = 0.0$ and unlimited $G_{\max}$ at varying $\Delta E_{\max}$). See Table C.10 for notation.

$\Delta E_{\max}$ (eV)	n	e	M	Q_M	N_N	Q_N
1	2	2	3	2	2	1
1.5	2	2	3	2	2	1
2	2	2	3	2	2	1
2.5	313	1256	105	90	92	66
3	446	2078	127	105	94	68
3.5	450	2294	127	105	94	68
4	454	2520	130	107	94	68
4.5	455	2722	130	107	95	69
5	456	2962	131	108	96	70

C.6.3 CH₂O₂ Bounded RN with $N_s = \mathbf{C} = [0+] [0-]$

Table C.16: Network statistics for CH₂O₂ bounded RN with $N_s = \mathbf{C} = [0+] [0-]$ at varying $G_{\max}$ and $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	6	10	9	7	4	2	1	6	10	9	7	4	2
2	15	36	18	11	6	3	2	16	40	20	13	7	4
3	31	84	34	22	13	10	3	36	106	41	27	17	12
4	61	184	65	43	26	19	4	78	236	83	56	33	22
5	103	320	96	55	49	28	5	145	464	125	74	65	34
6	160	524	143	71	79	37	6	227	824	184	92	101	42
7	214	790	178	81	103	43	7	295	1192	227	107	122	50
8	247	982	199	86	119	49	8	341	1466	254	119	134	54
9	265	1072	207	90	128	50	9	375	1632	270	122	140	55
10	287	1150	220	100	136	52	10	407	1764	289	130	142	56
11	296	1180	225	104	139	54	11	425	1908	294	132	142	56
12	310	1230	232	108	142	56	12	433	1954	298	135	143	57
13	321	1270	239	111	143	56	13	436	1982	300	136	145	58
14	335	1326	246	116	144	57	14	436	1986	300	136	145	58
15	351	1396	255	117	143	56	15	436	1986	300	136	145	58

Table C.17: Network statistics for CH₂O₂ bounded RN with $N_s = \mathbf{C}=[0+] [0-]$ and unrestricted $\Delta E_{\max}$ at varying $G_{\max}$). See Table C.10 and Table C.11 for notation.

$G_{\max}$	N_v	E_v	N_f	E_f	M	Q_M	N_N	Q_N
1	7	12	7	12	11	9	4	2
2	24	60	24	60	30	21	8	5
3	72	212	75	220	77	54	25	18
4	190	678	228	794	172	102	60	37
5	382	1672	555	2332	270	137	105	50
6	592	3080	1123	5732	348	172	155	65
7	727	4122	1700	9950	376	187	177	74
8	793	4774	2048	13,758	388	192	190	79
9	830	4980	2260	15,470	394	195	198	83
10	853	5138	2365	16,446	398	198	203	87
11	862	5204	2402	16,456	397	197	207	91
12	874	5296	2428	16,728	397	197	212	93
13	877	5332	2441	16,704	397	197	214	94
14	877	5338	2449	17,214	398	198	214	95
15	878	5338	2453	17,184	398	198	214	95

Table C.18: Network statistics for CH₂O₂ bounded RN with $N_s = \mathbf{C}=[0+] [0-]$ and unlimited $G_{\max}$ at varying $\Delta E_{\max}$). See Table C.10 for notation.

$\Delta E_{\max}$ (eV)	n	e	M	Q_M	N_N	Q_N
1	41	130	44	16	32	10
1.5	83	302	88	35	60	22
2	212	778	164	69	118	42
2.5	350	1320	245	113	137	54
3	409	1618	285	129	145	58
3.5	431	1804	296	135	145	58
4	436	1986	300	136	145	58
4.5	460	2208	315	146	149	61
5	503	2574	325	155	156	66

C.6.4 CH₄O₃ Bounded RN with $N_s = \text{C000}$

Table C.19: Network statistics for CH₄O₃ bounded RN with $N_s = \text{C000}$ at varying $G_{\max}$ and $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	14	26	23	23	9	9	1	17	32	29	29	9	9
2	69	196	69	67	30	30	2	100	306	93	91	36	36
3	215	694	137	129	82	78	3	344	1262	197	187	110	105
4	507	1862	264	244	187	170	4	911	3760	392	352	252	227
5	971	4072	454	369	336	283	5	1969	9186	698	551	452	354
6	1686	7384	691	491	527	365	6	3705	18,652	1128	786	667	443
7	2538	11,428	953	614	697	416	7	5791	31,804	1588	976	879	513
8	4272	20,956	1404	822	904	487	8	8591	51,900	2167	1231	1046	565
9	5440	27,706	1635	929	951	504	9	10,681	69,214	2541	1415	1165	628
10	6777	36,842	1869	1054	1054	554	10	12,099	80,898	2773	1534	1237	679
11	8253	47,388	2138	1244	1191	640	11	12,983	87,052	2866	1572	1262	694
12	9154	52,514	2284	1331	1216	657	12	13,640	91,676	2931	1605	1281	706
13	10,133	57,602	2453	1415	1228	667	13	14,040	94,334	2968	1627	1287	711
14	10,990	62,640	2618	1474	1235	670	14	14,350	96,230	2991	1643	1292	714
15	11,624	66,632	2717	1519	1236	671	15	14,567	97,220	3002	1651	1295	717

Table C.20: Network statistics for CH₄O₃ bounded RN with $N_s = \text{C000}$ and unrestricted ΔE_{max} at varying G_{max}). See Table C.10 and Table C.11 for notation.

G_{max}	N_v	E_v	N_f	E_f	M	Q_M	N_N	Q_N
1	20	38	20	38	35	35	9	9
3	649	2866	693	3002	338	311	148	142
4	1891	10,276	2804	14,440	722	597	350	307
5	4440	27,726	9109	55,998	1301	948	686	516
6	8649	61,892	26,367	188,036	1987	1248	1124	740
7	13,957	113,772	58,160	481,210	2567	1493	1578	937
8	18,815	172,082	108,894	1,023,576	2887	1634	1992	1108
9	21,806	216,408	168,633	1,776,580	2993	1662	2230	1181
10	23,174	237,846	219,493	2,540,754	3032	1679	2333	1221
11	23,810	245,976	249,891	3,068,660	3045	1682	2366	1231
12	24,063	249,258	266,503	3,378,428	3048	1682	2375	1240
13	24,143	250,384	274,498	3,523,232	3048	1682	2377	1248
14	24,180	250,736	278,309	3,577,074	3048	1682	2377	1233
15	24,202	250,918	279,831	3,601,482	3048	1682	2377	1236

Table C.21: Network statistics for CH₄O₃ bounded RN with $N_s = \text{C000}$ and unlimited G_{max} at varying ΔE_{max}). See Table C.10 for notation.

ΔE_{max} (eV)	n	e	M	Q_M	N_N	Q_N
1	2362	8538	901	492	739	336
1.5	9794	43,226	2641	1361	1044	525
2	10,840	52,610	2844	1536	1158	610
2.5	12,554	65,996	2920	1586	1230	665
3	13,518	76,100	2971	1625	1237	672
3.5	14,429	87,352	2983	1634	1250	682
4	15,360	101,412	3011	1657	1298	719
4.5	15,715	112,550	3021	1661	1307	728
5	15,851	121,600	3027	1667	1313	734

C.6.5 CH₄O₃ Bounded RN with $N_s = \text{CO}[0] . [\text{OH}]$

Table C.22: Network statistics for CH₄O₃ bounded RN with $N_s = \text{CO}[0] . [\text{OH}]$ at varying $G_{\max}$ and $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	7	12	11	11	5	5	1	9	16	15	15	5	5
2	34	88	38	36	19	17	2	50	130	54	52	22	20
3	117	370	95	91	53	50	3	194	652	129	125	72	67
4	312	1068	188	176	124	117	4	592	2270	274	247	185	166
5	685	2616	342	297	258	219	5	1422	6304	522	428	361	293
6	1286	5448	569	425	449	332	6	2881	13,830	900	654	599	414
7	2208	10,002	849	559	644	390	7	5094	27,172	1383	890	844	491
8	3270	15,182	1122	680	790	441	8	7321	42,562	1807	1054	965	528
9	5147	26,728	1565	894	957	506	9	9654	62,184	2276	1290	1123	596
10	6078	31,690	1741	999	1038	549	10	11,354	75,464	2590	1449	1201	655
11	7833	44,508	2059	1195	1151	615	11	12,590	84,414	2794	1549	1246	686
12	8448	48,290	2154	1249	1180	632	12	13,426	90,314	2900	1597	1274	702
13	10,091	57,710	2480	1406	1220	660	13	13,925	93,602	2946	1617	1287	712
14	10,948	62,786	2629	1472	1229	667	14	14,232	95,654	2978	1634	1290	712
15	11,568	66,632	2734	1522	1235	670	15	14,431	96,716	2993	1644	1292	714

Table C.23: Network statistics for CH_4O_3 bounded RN with $N_s = \text{CO}[\text{O}] \cdot [\text{OH}]$ and unrestricted ΔE_{max} at varying G_{max} . See Table C.10 and Table C.11 for notation.

G_{max}	N_v	E_v	N_f	E_f	M	Q_M	N_N	Q_N
1	13	24	13	24	23	23	6	6
2	89	252	92	264	81	76	31	29
3	428	1640	553	2072	238	218	112	104
4	1436	7168	2315	11,002	549	449	305	256
5	3700	21,878	8283	46,922	1082	787	603	435
6	7756	54,080	25,788	174,244	1770	1137	1043	667
7	12,978	104,666	62,916	494,268	2403	1416	1523	869
8	17,868	162,292	109,072	1,014,760	2811	1590	1905	1031
9	21,204	209,130	172,658	1,824,972	2976	1658	2169	1156
10	22,864	233,954	218,844	2,547,190	3035	1679	2300	1194
11	23,624	243,650	247,564	3,022,048	3046	1682	2349	1221
12	24,006	248,266	264,780	3,360,932	3048	1682	2372	1240
13	24,124	250,076	273,330	3,508,848	3048	1682	2377	1238
14	24,164	250,626	277,587	3,582,004	3048	1682	2377	1238
15	24,197	250,836	279,525	3,617,292	3048	1682	2377	1239

Table C.24: Network statistics for CH_4O_3 bounded RN with $N_s = \text{CO}[\text{O}] \cdot [\text{OH}]$ and unlimited G_{max} at varying ΔE_{max} . See Table C.10 for notation.

ΔE_{max} (eV)	n	e	M	Q_M	N_N	Q_N
1	2362	8538	901	492	739	336
1.5	9794	43,226	2641	1361	1044	525
2	10,840	52,610	2844	1536	1158	610
2.5	12,554	65,996	2920	1586	1230	665
3	13,518	76,100	2971	1625	1237	672
3.5	14,429	87,352	2983	1634	1250	682
4	15,360	101,412	3011	1657	1298	719
4.5	15,715	112,550	3021	1661	1307	728
5	15,851	121,600	3027	1667	1313	734

C.6.6 CH₆O₄ Bounded RN with $N_s = \text{C000.0}$

Table C.25: Network statistics for CH₆O₄ bounded RN with $N_s = \text{C000.0}$ at varying $G_{\max}$ and $\Delta E_{\max} = 3$ and $\Delta E_{\max} = 4$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	15	28	24	24	9	9	1	18	34	30	30	9	9
2	92	268	78	76	32	32	2	128	394	103	101	38	38
3	375	1354	163	155	100	96	3	578	2252	229	219	129	124
4	1092	4648	322	302	254	237	4	1899	8850	470	429	331	306
5	2503	12,066	569	478	504	447	5	5033	26,630	904	747	688	584
6	4969	26,142	981	755	894	708	6	11,571	66,822	1684	1303	1228	954
7	9080	50,184	1636	1182	1434	1043	7	23,878	149,472	3034	2242	1969	1431
8	15,917	91,150	2632	1844	2129	1471	8	43,430	299,524	4939	3478	2837	1953
9	26,746	159,786	4000	2737	2982	1977	9	69,758	522,372	7254	4830	3812	2494
10	41,443	264,640	5576	3690	3946	2501	10	100,796	813,160	9671	6129	4835	3016
11	58,343	396,488	7032	4546	4861	2954	11	130,595	1,118,396	11,931	7381	5743	3470
12	76,787	540,338	8397	5431	5607	3342	12	154,504	1,360,228	13,533	8281	6373	3788
13	96,826	692,434	9939	6384	6150	3622	13	174,755	1,559,002	14,666	8930	6727	3977

C.6.7 CH₆O₄ Bounded RN with $N_s = \text{C0[0] . [OH] . 0}$

Table C.26: Network statistics for CH₆O₄ bounded RN with $N_s = \text{C0[0] . [OH] . 0}$ at varying $G_{\max}$. See Table C.10 for notation.

$\Delta E_{\max} = 3$							$\Delta E_{\max} = 4$						
$G_{\max}$	n	e	M	Q_M	N_N	Q_N	$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	7	12	12	12	5	5	1	9	16	16	16	5	5
2	39	98	41	39	21	19	2	58	144	59	57	24	22
3	165	528	105	101	64	61	3	277	922	140	136	84	79
4	571	2158	218	206	168	161	4	1075	4342	313	286	247	228
5	1581	6874	421	376	388	349	5	3258	15,752	634	536	559	484
6	3513	17,304	731	580	735	601	6	8161	44,960	1175	904	1070	834
7	6788	36,344	1233	905	1185	869	7	17,878	108,344	2067	1506	1749	1223
8	12,408	69,912	2023	1403	1796	1253	8	34,077	228,746	3477	2500	2604	1730
9	21,294	125,408	3050	2108	2524	1690	9	56,596	419,918	5354	3740	3548	2271
10	32,394	197,968	4302	2907	3330	2133	10	83,772	661,372	7659	5069	4454	2756
11	49,372	329,510	5899	3889	4386	2683	11	113,832	954,158	10,066	6359	5429	3258
12	66,894	464,256	7314	4750	5196	3100	12	140,724	1,228,606	12,227	7537	6086	3618
13	86,637	614,566	8802	5699	5798	3429	13	163,274	1,451,988	13,777	8426	6547	3865

C.6.8 C₂H₄O₃ Bounded RN with $N_s = \mathbf{C=C.O}=[0+] [0-]$

Table C.27: Network statistics for C₂H₄O₃ bounded RN with $N_s = \mathbf{C=C.O}=[0+] [0-]$ and $\Delta E_{\max} = 3$ at varying $G_{\max}$). See Table C.10 for notation.

$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	6	10	8	7	5	4
2	18	54	17	15	12	8
3	54	170	44	36	31	20
4	177	546	142	108	89	57
5	598	1956	442	336	261	186
6	2014	6966	1329	921	738	488
7	6258	24,270	3644	2286	2078	1311
8	15,971	68,386	8186	4568	4614	2608
9	35,661	171,882	15,933	8203	8851	4469
10	67,274	351,134	26,480	12,992	14,158	6634

C.6.9 C₂H₆O₄ Bounded RN with $N_s = \text{C}=\text{C}.0=[0+] [0-] .0$

Table C.28: Network statistics for C₂H₆O₄ bounded RN with $N_s = \text{C}=\text{C}.0=[0+] [0-] .0$ and $\Delta E_{\max} = 3$ at varying $G_{\max}$. See Table C.10 for notation.

$G_{\max}$	n	e	M	Q_M	N_N	Q_N
1	7	12	11	10	5	4
2	27	82	22	20	13	9
3	94	330	58	48	36	24
4	334	1220	178	144	119	82
5	1261	4816	568	455	402	292
6	4659	19,124	1818	1322	1202	846
7	15,960	70,122	5437	3786	3413	2357
8	48,863	232,192	14,055	9282	8612	5671
9	133,506	696,606	31,724	20,317	19,212	12,122
10	319,737	1,807,582	63,635	39,525	37,916	22,562

C.7 Exhaustive Reaction Network Scaling with $\Delta E_{\max}$

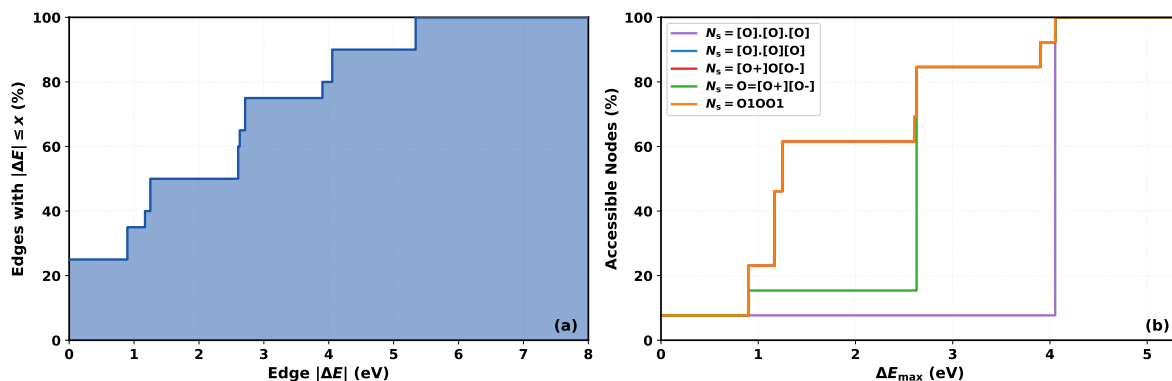


Figure C.1: Cumulative percentage of (a) edges with $|\Delta E| \leq x$ for the exhaustive O₃ RN and (b) accessible nodes retained as a function of the bounded-generation energy threshold $\Delta E_{\max}$ for different N_s .

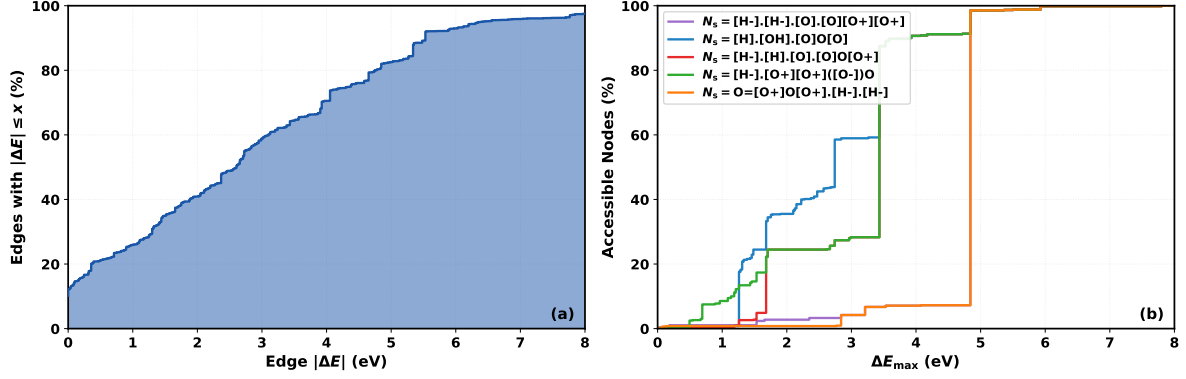


Figure C.2: Cumulative percentage of (a) edges with $|\Delta E| \leq x$ for the exhaustive O_4H_2 RN and (b) accessible nodes retained as a function of the bounded-generation energy threshold $\Delta E_{\max}$ for different N_s .

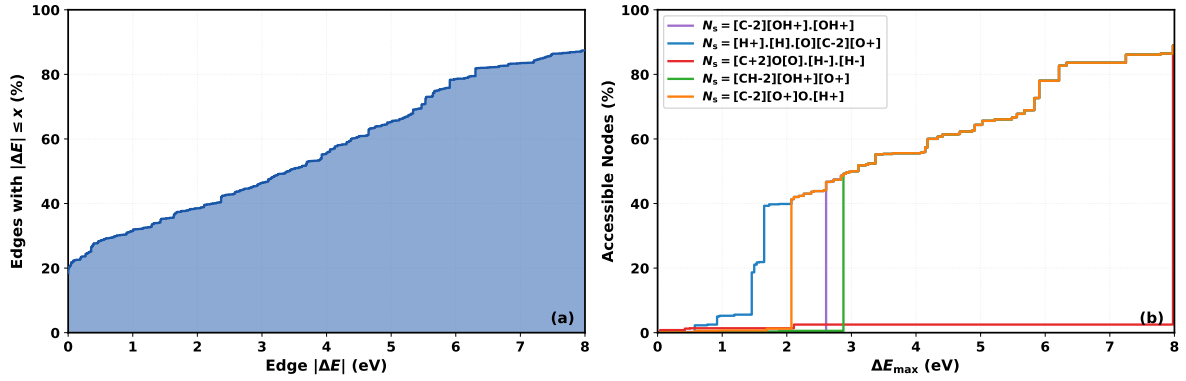


Figure C.3: Cumulative percentage of (a) edges with $|\Delta E| \leq x$ for the exhaustive CH_2O_2 RN and (b) accessible nodes retained as a function of the bounded-generation energy threshold $\Delta E_{\max}$ for different N_s .

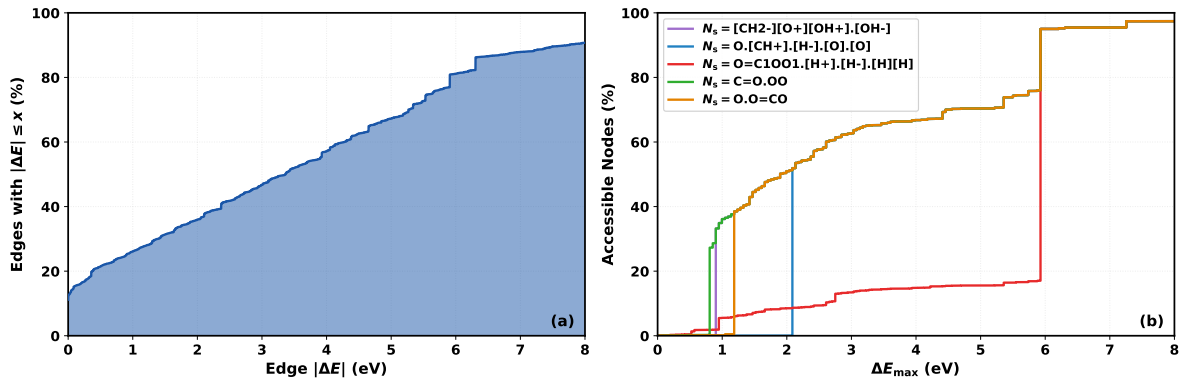


Figure C.4: Cumulative percentage of (a) edges with $|\Delta E| \leq x$ for the exhaustive CH_4O_3 RN and (b) accessible nodes retained as a function of the bounded-generation energy threshold $\Delta E_{\max}$ for different N_s .