Synthesis Lectures on
Data Mining and Knowledge Discovery

SYNTHESIS
COLLECTION OF TECHNOLOGY

Gang Liu · Eric Inae ·
Meng Jiang

Deep Learning for Polymer Discovery

Foundation and Advances

Springer

Synthesis Lectures on Data Mining and Knowledge Discovery

The series focuses on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Gang Liu · Eric Inae · Meng Jiang

Deep Learning for Polymer Discovery

Foundation and Advances

Springer

Gang Liu
Department of Computer Science
and Engineering
University of Notre Dame
Notre Dame, IN, USA

Eric Inae
Department of Computer Science
and Engineering
University of Notre Dame
Notre Dame, IN, USA

Meng Jiang
Department of Computer Science
and Engineering
University of Notre Dame
Notre Dame, IN, USA

ISSN 2151-0067 ISSN 2151-0075 (electronic)
Synthesis Lectures on Data Mining and Knowledge Discovery
ISBN 978-3-031-84731-8 ISBN 978-3-031-84732-5 (eBook)
https://doi.org/10.1007/978-3-031-84732-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Modern polymeric materials have revolutionized various aspects of our lives, driving advancements in sustainability, biomaterials, biosensors, energy solutions, and aerospace technologies. The ever-evolving engineering and environmental challenges of our time demand materials with unconventional properties, such as high temperature stability, exceptional thermal conductivity, and biodegradability.

Polymers are composed of molecules represented as graph structures, where atoms act as nodes and bonds as edges. This inherent structure makes deep learning techniques—such as Transformers and Graph Neural Networks (GNNs)—indispensable for discovering new polymers that fulfill modern requirements.

Deep learning paradigms, namely prediction and generation, are integral to material virtual screening and inverse design, respectively. This book offers a systematic exploration of deep learning techniques tailored to polymer discovery, bridging the disciplines of materials science and artificial intelligence. It equips researchers and practitioners with foundational concepts and state-of-the-art methods for predicting polymer properties and designing novel polymers using advanced neural network architectures.

The content spans a broad spectrum of topics, progressing from fundamental concepts to advanced methodologies. It begins with polymer data representations and neural network architectures (Chap. 1) before delving into frameworks for property prediction (Chap. 2) and inverse polymer design (Chap. 3). Key approaches include sequence-based and graph-based techniques, leveraging neural network models such as LSTMs, GRUs, GCNs, and GINs. Advanced discussions encompass interpretable graph deep learning with environment-based augmentation (Chap. 4), semi-supervised methods for addressing label imbalance (Chap. 5), and data-centric transfer learning using generative methods like diffusion models (Chap. 6). Each topic is presented with detailed problem definitions, method descriptions, and experimental validations.

The book tackles pressing issues in polymer discovery, such as accurate property prediction, efficient design of polymers with desired traits, model interpretability, handling

imbalanced and limited labeled data, and leveraging unlabeled data for enhanced predictions. Practical examples and experiments on real-world datasets demonstrate the efficacy of the proposed methodologies.

This book is designed for researchers, graduate students, and professionals in materials science, chemistry, and computer science who are interested in harnessing deep learning for polymer discovery and design. It serves as a primer, practical guide, and reference for those seeking to integrate artificial intelligence into materials research and development, inspiring innovation at the intersection of science and technology.

Notre Dame, USA
December 2024

Gang Liu
Eric Inae
Meng Jiang

Competing Interests The authors declare the following as potential competing interests: This work was supported by NSF IIS-2142827, IIS-2146761, IIS-2234058, CBET-2332270, and ONR N00014-22-1-2507.

Contents

Polymer Data and Deep Neural Networks

1

1.1 Polymer Data Representations

Polymers can be represented in multiple forms such as sequences, graphs, and vectors, as illustrated in Fig. 1.1. Neural networks have demonstrated their effectiveness in science by learning intricate relationships from data and leveraging this understanding to inform decision-making processes. When applying machine learning models (e.g., neural networks) to polymer tasks, the initial step involves determining the most suitable representation for the polymer data to be utilized by the models.

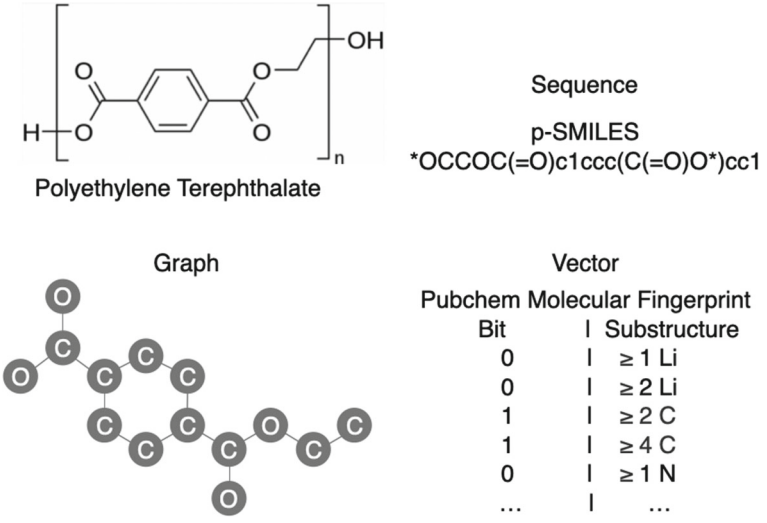

Fig. 1.1 Visualization of a polymer's representations: **a** polyethylene terephthalate (PET); **b** sequence; **c** graph; **d** feature vector

G. Liu et al., *Deep Learning for Polymer Discovery*, Synthesis Lectures on Data Mining and Knowledge Discovery, https://doi.org/10.1007/978-3-031-84732-5_1