

Gang Liu · Eric Inae ·
Meng Jiang

Deep Learning for Polymer Discovery

Foundation and Advances

Synthesis Lectures on Data Mining and Knowledge Discovery

Series Editors

Jiawei Han, University of Illinois at Urbana-Champaign, Urbana, USA

Lise Getoor, University of California, Santa Cruz, USA

Johannes Gehrke, Microsoft Corporation, Redmond, USA

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

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>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer
Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

If disposing of this product, please recycle the paper.

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

1	Polymer Data and Deep Neural Networks	1
1.1	Polymer Data Representations	1
1.1.1	Sequences	2
1.1.2	Graphs	2
1.1.3	Vectors	3
1.2	Neural Networks	4
1.2.1	Modeling Tasks	4
1.2.2	Basic Neural Network Components	6
1.2.3	Advanced Neural Network Architecture	8
1.3	Foundation: Neural Network-Based Frameworks for Polymer Modeling	10
1.3.1	Deep Learning for Polymer Property Prediction	10
1.3.2	Deep Learning for Inverse Polymer Design	12
1.4	Advances: Deep Learning with Interpretable, Imbalance-Robust, and Generative Graph Methods	13
1.4.1	Interpretable Learning: Graph Rationalization with Environment-Based Augmentation	13
1.4.2	Imbalanced Learning: Semi-supervised Graph Imbalanced Regression	14
1.4.3	Generative Modeling: Data-Centric Learning from Unlabeled Graphs with Diffusion Model	14
	References	15
2	Deep Learning for Polymer Property Prediction	17
2.1	Problem Definition and Datasets	17
2.1.1	Polymer Property Classification	17
2.1.2	Polymer Property Regression	18
2.1.3	Dataset, Task Formulation and Evaluation	18

2.2	Sequence(SMILES)-Based Prediction	20
2.2.1	Recurrent Neural Networks	20
2.2.2	LSTM	20
2.2.3	GRU	21
2.2.4	Transformer	22
2.3	Graph-Based Prediction	23
2.3.1	Graph Neural Networks	23
2.3.2	Graph Convolutional Networks (GCNs)	24
2.3.3	GraphSAGE	25
2.3.4	GAT	26
2.3.5	GIN	26
2.3.6	Graph Transformers	27
2.4	Specific Techniques	28
2.4.1	Hyperparameter Tuning	28
2.4.2	Data Augmentation	30
2.4.3	Other Learning Paradigms	31
2.5	Summary	33
	References	34
3	Deep Learning for Inverse Polymer Design	37
3.1	Problem Definition and Datasets	37
3.1.1	Unconditional Generation Without Constraints	37
3.1.2	Conditional Generation With Constraints	38
3.1.3	Datasets, Task Formulation and Evaluation	38
3.2	Generative Neural Network Architectures	39
3.2.1	Generative Adversarial Networks	40
3.2.2	Variational Autoencoders	40
3.2.3	Diffusion Models	41
3.3	Unconstrained Polymer Generation	42
3.3.1	Sequence-Based Generation	43
3.3.2	Graph-Based Generation	45
3.4	Constrained Polymer Generation	47
3.4.1	Constraint Types	47
3.4.2	Adding Constraints as Features	50
3.5	Summary	51
	References	52
4	Interpretable Learning: Graph Rationalization	
	with Environment-Based Augmentation	55
4.1	Introduction	55
4.2	Problem Definition	58
4.2.1	Graph Property Prediction	58
4.2.2	Graph Rationalization	59

4.3	Interpretable Graph Neural Networks: GREASE	60
4.3.1	Rationale-Environment Separation	60
4.3.2	Environment-Based Augmentations	61
4.3.3	Optimization	62
4.4	Experiments	63
4.4.1	Experimental Settings	63
4.4.2	RQ1: Results on Effectiveness	65
4.4.3	RQ2: Ablation Study on GREASE	65
4.4.4	RQ3: Case Study on Polymer Data	65
4.4.5	RQ4: Results on Efficiency	71
4.4.6	RQ5: Sensitivity Analysis	72
4.5	Conclusion	72
	References	74
5	Imbalanced Learning: Semi-Supervised Graph Imbalanced Regression	77
5.1	Introduction	77
5.2	Problem Definition	79
5.3	Self-Training Framework: SGIR	80
5.3.1	A Self-Training Framework for Iteratively Balancing Scalar Label Data	80
5.3.2	Balancing with Confidently Predicted Labels	81
5.3.3	Balancing with Augmentation via Label-Anchored Mixup	82
5.3.4	Optimization	83
5.4	Theoretical Motivations	84
5.5	Experiments	88
5.5.1	Experimental Settings	88
5.5.2	RQ1: Effectiveness on Property Prediction	91
5.5.3	RQ2: Ablation Studies and Sensitivity Analysis	96
5.6	Conclusion	98
	References	100
6	Generative Modeling: Data-Centric Learning from Unlabeled Graphs with Diffusion Model	103
6.1	Introduction	103
6.2	Problem Definition	106
6.3	Data-Centric Transfer Framework: DCT	107
6.3.1	Overview of Developed Framework	107
6.3.2	Learning Data Distribution from Unlabeled Graphs	107
6.3.3	Generating Task-Specific Labeled Graphs	109

6.4	Experiments	113
6.4.1	Experimental Settings	113
6.4.2	RQ1: Outstanding Property Prediction Performance	115
6.4.3	RQ2: Ablation Studies and Performance Analysis	118
6.4.4	RQ3: Case Study for the Interpretability of Visible Knowledge Transfer	120
6.5	Conclusion	121
	References	121

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.

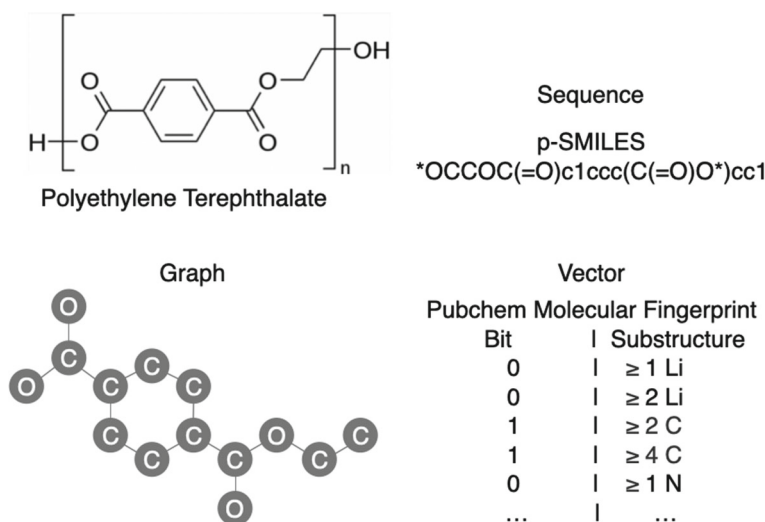


Fig. 1.1 Visualization of a polymer's representations: **a** polyethylene terephthalate (PET); **b** sequence; **c** graph; **d** feature vector