LEVERAGING DOMAIN KNOWLEDGE FOR ENHANCED CAUSAL STRUCTURE LEARNING AND OUT-OF-DISTRIBUTION GENERALIZATION IN OBSERVATIONAL DATA

by

Md Hasan Jawad Chowdhury

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing & Information Systems

Charlotte

2024

Approved by:

_____
Dr. Gabriel Terejanu

_____
Dr. Minwoo Lee

_____
Dr. Razvan Bunescu

_____
Dr. Jim Conrad

ABSTRACT

MD HASAN JAWAD CHOWDHURY. Leveraging domain knowledge for enhanced causal structure learning and out-of-distribution generalization in observational data. (Under the direction of DR. GABRIEL TEREJANU)

Causal modeling enables robust counterfactual reasoning and interventional mechanisms to make predictions across different hypothetical scenarios. Nevertheless, uncovering causal relationships from observational data presents a considerable challenge, as unobserved confounders, limited sample sizes, and variations in distributions can give rise to misleading cause-effect associations. Models relying on these relationships may perform poorly when spurious correlations do not hold in test cases. To address these challenges, researchers augment causal learning with known causal relations. This dissertation first investigates the incorporation of domain knowledge in structure learning by introducing additional constraints that convey qualitative knowledge about causal relationships. The experimental designs are specifically equipped to evaluate the role of domain knowledge. Secondly, a concept-driven approach is implemented to determine the advantages of incorporating concept-level prior knowledge. Given the invariant nature of causal relationships, the study then showcases the broader applicability of incorporating domain knowledge by employing a machine learning method for learning adsorption energies, illustrating the advantages of harnessing domain knowledge to obtain invariant molecular representations in catalyst screening. Finally, a novel approach is introduced to enhance robustness and out-of-distribution generalization by leveraging gradient agreement across different environments to identify reliable features. Collectively, these experimental designs advance causal discovery and robust machine learning by utilizing prior knowledge and relational invariances, paving the way for future research on integrating domain knowledge and invariance principles into the learning process.

DEDICATION

To my dear parents, my brother, and sister,

who have supported me throughout my life,

And to my beloved wife, Fahin,

who made countless sacrifices during this PhD journey,

And to our sweet Ayana,

who brought joy and light during the toughest times,

I am forever grateful for your love, strength, and unwavering support!

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my academic advisor, Dr. Gabriel Terejanu. Without his constant support, this dissertation would not have been possible. I am profoundly grateful for the opportunity he provided me to pursue research that aligned with my interests, and for his consistent guidance and mentoring that has not only broadened my intellectual horizons but also prepared me to become an independent researcher. His support beyond research and academia has also helped me navigate the ups and downs of my life.

I extend my sincere thanks to my committee members, Dr. Minwoo Lee, Dr. Razvan Bunescu, and Dr. Jim Conrad, for their invaluable support throughout my PhD journey. I am also grateful to Dr. Anthony Fodor for his contributions as a former committee member. Their insightful feedback and suggestions have significantly shaped my research. I deeply appreciate Dr. Andreas Heyden from the University of South Carolina for his guidance during our weekly meetings, which enhanced my approach to multidisciplinary research. Special thanks to Dr. Dewan Ahmed, Dr. Ahmed Arafa, Dr. Srinivas Akella, Dr. Lisa Russell-Pinson, and many others who shared valuable lessons throughout my PhD. I am also thankful to all my co-authors for their insights and collaboration. Finally, I am deeply thankful for the financial support from the Graduate School at UNC Charlotte, the National Science Foundation under Award Number 2218841, the Lowe's Innovation Fund, and Toyota Racing Development, which have been essential in facilitating my research journey.

During my internships, I had the opportunity to broaden my research perspectives and learn about the applicability of my work both at Lowe's Technology and Toyota Racing Development. I am thankful to Hamidreza Farhidzadeh and Jean-Leah Njoroge at Lowe's for their guidance and support, and to Henri Durand and Jason Ashbrook at Toyota Racing Development for helping me understand real-world challenges and applications of

my research.

To my lab mates and friends — Reza, Tanu, Michael, Pedram, Ouldouz, Touhid, Towfiq, Simon, Tanzeer, Asif, Azim, Mokhtar, Nouf, Giorgio, and many more — your collaboration, friendship, and support have been instrumental in my growth as a researcher and as a person throughout this journey. Sharing this journey with you has broadened my views and improved my outlook on life. I will cherish those memories of fun and laughter that helped me stay human, even during the most stressful times. Your jokes and the good times we shared made the deadlines and the pressures much more bearable.

To my parents, everything I am as a person today—is because of you. I know that even though you're not near me, your blessings are always with me. The memories, your morals, and the teachings you passed on will stay with me every day for the rest of my life. To my brother, Jarjis, and my sister, Jaria—you have been more than siblings; you've been my guardians, my sources of parental guidance who have never let me feel alone, always standing tall like two big trees, providing me shade, comfort, and protection. May the Almighty always keep you in His grace. To my wife, Fahin, and my little angel, Ayana—you both are my champions and my greatest inspiration. Your smiles at the end of each day are my deepest source of refreshment, relief, energy, and motivation, helping me face every new day with renewed enthusiasm and determination. I am deeply grateful for your presence in my life, which has kept me grounded and focused, even during the toughest times. Finally, I would like to thank the Almighty for blessing me with such wonderful people and brilliant minds throughout my journey, and for giving me the strength to keep going.

I want to acknowledge that this list is nowhere near complete. Countless individuals have contributed to my PhD journey, and I apologize to those I have missed to mention in this brief acknowledgment. This journey would not have been possible without the support and contributions of every single one of you. I sincerely thank you all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In recent years, predictive machine learning models have achieved remarkable success across a wide range of scientific domains. These models demonstrate exceptional power when the underlying i.i.d. (independent and identically distributed) assumption holds true in the test environments. Consequently, deep learning approaches have gained widespread adoption in numerous scientific fields [3–9]. However, the primary objective of these models is to achieve accurate predictions, which often leads them to rely on spurious correlations [10, 11] that may not hold true if the data distribution differs between training and application contexts. This dependence on superficial relationships or features can result in poor generalization and subpar performance [12, 13]. Causality, which refers to the relationship between cause and effect, presents a potential solution to address this issue. Causality is the process of establishing a causal connection based on the conditions under which an effect occurs. Naturally, one might wonder what constitutes a causal connection. To address this, let's consider two random variables X and Y as the variables of interest. A causal connection between X and Y, or more specifically, X causing Y, can be inferred if and only if intervening or manipulating X results in changes to Y, while keeping all other variables constant or fixed [14]. In simpler terms, causality can be defined as the relationship between an effect and its underlying cause. Inferring causal structures typically relies on interventions in the system variables, known as Randomized Controlled Trials (RCTs). However, these interventions are often impractical, unethical, or costly. For instance, consider a scenario where we aim to determine whether the positions of the Earth and Moon have a causal effect on ocean tides. Conducting an intervention to reach a conclusion would be impossible. Similarly, if we were to investigate whether smoking

causes lung cancer or not, it would be unethical to conduct experiments on a group of people to observe if they develop lung cancer after smoking. Due to these limitations, we frequently resort to deducing causal implications based on available observations. Discovering causal relationships purely from observational data, known as observational causal discovery, offers a solution to address these challenges.

The significance of understanding data behavior and uncovering the underlying causal structure cannot be overstated, as it plays a crucial role in making informed decisions and analyzing various phenomena. Since causal relations are invariant in nature, they are expected to hold true even when there are distributional changes. Causal analysis allows us to reason under different interventional conditions, answer what-if questions, and provides the foundation for effective action, policymaking, and learning from failures. Relying solely on correlation or statistical association-based studies would not sufficiently address these issues. Learning causal structures from purely observational data is a challenging task, with applications spanning various fields such as genetics [15], machine learning [16, 17], economics [18], and biology [19]. While causality has been a topic of study for many decades, recent advancements in machine learning, complex models, high computational power, large storage systems, and in-depth analytical capabilities, coupled with data availability, have opened new horizons in this research area. However, learning causal links from observational data presents numerous challenges due to limitations such as finite sampling, unobserved confounding factors, selection bias, and measurement errors, which can result in spurious cause-effect relationships [20–22]. To address these issues in practice, researchers frequently enhance causal learning by inducing prior causal knowledge [23–25]. Prior knowledge can come in different forms and can be utilized in various ways to help identify the underlying structure more accurately. This dissertation studies, emphasizes, and investigates various forms of domain knowledge, examines different methods for incorporating these types of expertise, and assesses the impact of integrating such prior knowledge in causal structure learning and out-of-distribution (OOD)

generalization. Furthermore, the application of these approaches is evaluated in a variety of contexts.

1.1 Major Contributions

This dissertation makes significant contributions to the fields of causal structure learning and out-of-distribution (OOD) generalization in machine learning. The key advances are as follows:

- **Evaluation of Induced Expert Knowledge in Causal Structure Learning:** This research extends the state-of-the-art score-based causal learning method, NOTEARS [26], by incorporating domain knowledge. The enhancement allows for more accurate causal structure learning by inducing expert knowledge and reveals key insights in the learning process by evaluating the impact of qualitative domain knowledge. The study assesses how expert-induced knowledge can influence the robustness of causal models. This work is published in the *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)* [27].

- **Concept-Driven Causal Structure Learning:** In this work, the NOTEARS [26] method is adapted to learn causal structures in concept-driven data, which includes both scalar-valued features and multidimensional vector-valued concepts. The approach is particularly beneficial for uncovering causal structures in complex, high-dimensional conceptual data. This work also addresses challenges associated with applying causal constraints in such contexts and explores potential solutions. This research is published in the *International Conference on Machine Learning and Applications (ICMLA 2023)* [28].

- **Invariant Molecular Representations for Heterogeneous Catalysis:** This study explores the use of domain knowledge to exploit the invariance property, fo-

cusing on eliminating spurious correlations and learning invariant or causal relations. It presents a method to develop robust predictors for adsorption energy in chemical reactions by creating invariant representations of molecular species. The research also highlights how these invariant representations can be extended to larger catalysis datasets with multiple surface systems, addressing the challenges posed by the complexity of larger reaction systems. A significant portion of this work is published in the *Journal of Chemical Information and Modeling (2024)* [29].

- **Consistent Gradient-Based Learning for OOD Generalization:** This research introduces CGLearn, a novel method that enhances robustness and generalization in machine learning by learning invariant predictors across different environments. The method relies on gradient agreement across environments to identify reliable features and demonstrates its effectiveness in both linear and nonlinear settings. The study shows the applicability of CGLearn across various regression and classification tasks through experiments on synthetic and real-world datasets. This research is currently under review at the *39th Annual AAAI Conference on Artificial Intelligence (AAAI 2025).*

1.2 Outline

This dissertation is structured as follows:

- **Chapter 1:** Provides an introduction to causal structure learning and delves into the background information pertinent to the study. It covers essential concepts such as graphical and causal terms, building blocks of causal graphs, and commonly used causal assumptions in structure learning methods. This foundational knowledge sets the stage for the subsequent chapters, which present original research and proposed future work.

- **Chapter 2:** Presents the study titled "Evaluation of Induced Expert Knowledge in

Causal Structure Learning by NOTEARS" [27].

- **Chapter 3:** Presents the study titled "CD-NOTEARS: Concept-Driven Causal Structure Learning Using NOTEARS" [28].

- **Chapter 4:** Presents the study titled "Invariant Molecular Representations for Heterogeneous Catalysis" [29].

- **Chapter 5:** Introduces the study titled "CGLearn: Consistent Gradient-Based Learning for Out-of-Distribution Generalization" (under review at AAAI 2025).

- **Chapter 6:** Concludes the dissertation by synthesizing the key contributions and insights from the presented studies, providing a summary of the overall contributions. This chapter also explores promising future work and new research directions opened by the findings. This dissertation primarily emphasizes the significance and explores different pathways of leveraging domain knowledge to enhance causal structure learning and to improve the out-of-distribution (OOD) generalization of machine learning models in observational data.

1.3 Graphical and Causal Terms

In this section, we will review some of the basic concepts and definitions related to graph and causal structure learning. The following terms are expressed considering the graph $G = (V, E)$ over a set of vertices V and a set of edges E connecting these vertices.

Path: A path between two vertices X_1 and X_2 is a sequence of vertices where each subsequent pair of vertices is connected through an edge $E_1 = (X_1, X_2) \in E$ in the graph G.

Undirected path: An undirected path between two vertices X_1 and X_2 is a path or sequence of variables where each subsequent pairs are connected through undirected edges in the graph G.

Directed path: A directed path between two vertices X_1 and X_2 is a sequence of vertices where each subsequent pair of vertices are connected through a directed edge in the graph G.

Acyclic path: An acyclic path is a path in the graph G if it does not contain any vertex X_i more than once in the path. Otherwise, it is a cyclic path.

Undirected graph: An undirected graph is a graph $G = (V, E)$ where all the edges $E = (E_1, \ldots, E_n)$ are undirected.

Directed graph: A directed graph is a graph $G = (V, E)$ which has only directed edges such as all $E_i \in E$ are directed.

DAG: A directed acyclic graph or DAG is a graph where all the edges are directed and all the paths are acyclic.

Mediator: A vertex X_m is a mediator on a directed path between vertex X_s and X_e if the X_m is on the path but it is not the source (the path does not start with X_m) nor the sink (the path does not end with X_m).

Latent variable: A latent variable X_l is a variable that is unmeasured to a graph $G = (V, E)$ where $X_l \notin V$ but it is causally connected to the variables in V.

Confounder: A confounder is a latent variable that is also a common cause of two variables in the vertex set V.

Exogenous variable: A variable X_{ex} is considered exogenous with respect to a set of variables V if it is not influenced by any variable $X_i \in V$. In other words, there are no causal links from any variable $X_i \in V$ to the exogenous variable X_{ex}.

Endogenous variable: A variable X_{en} is considered endogenous with respect to a set of variables V if it is influenced by at least one variable $X_i \in V$. In other words, there exists a causal link from one or more variables in V to the endogenous variable X_{en}.

CGM: A causal graphical model CGM(P_x, G) can be defined as a pair of a graph G and an observational distribution P_x over a set of random variables $X = (X_1, \ldots, X_d)$. The distribution P_x is Markovian with respect to G where $G = (V, E)$ is a DAG that

encodes the conditional dependence structures among the random variables $X_i \in X$ [30]. The node $i \in V$ corresponds to the random variable $X_i \in X$ and the edges $(i, j) \in E$ correspond to the conditional relations encoded by G. In a causal graphical model, the joint distribution P_x can be factorized as $p(x) = \prod_{i=1}^{d} p(x_i|x_{pa_i}^G)$ where $X_{pa_i}^G$ refers to the set of parents for the variable X_i in DAG G and for each $X_j \in X_{pa_i}^G$ there is an edge $(X_j \to X_i) \in E$ and based on additional causal assumptions the directed edges have causal interpretations [30].

Unshielded triples: A triple of variables X_i, X_j, and X_k is considered unshielded if X_i is connected to X_j, and X_j is connected to X_k, but X_i and X_k are not directly connected.

Collider: X_c is considered a collider on a path between vertex X_s and X_t in graph G if there is a substructure $X_i \to X_c \leftarrow X_j$ in the path, for some variables X_i and X_j on that path.

d-separation: Consider two distinct vertices or variables, X and Y, in a graph G, and let Z be a set of vertices such that $X \notin Z$ and $Y \notin Z$. In this scenario, X and Y are considered d-separated given Z if Z blocks all paths between them. A path U is considered blocked if any of the following conditions hold:

1. The path U contains a chain $(A \to B \to C)$ or a fork $(A \leftarrow B \to C)$ structure, where B is an element of Z.

2. The path U contains a collider structure $(A \to B \leftarrow C)$, where B and all of its descendants are not elements of Z.

MEC: A Markov Equivalence Class M is a set of DAGs that encode the same set of conditional independencies.

SEM: A structural equation model (SEM) is defined as a $(S, L(N))$ where $S = (S_1, \ldots, S_d)$ is a collection of d structural equations over a set of random variables $X = (X_1, \ldots, X_d)$ [31].

$$S_j : X_j = f_j(PA_j, N_j), j = 1, \ldots, d \tag{1.1}$$

and here $L(N) = L(N_1, \ldots, N_d)$ is the joint distribution of the noises, $L(N)$ is required to be jointly independent. SEMs are considered mostly for the real valued random variable $X_1, \ldots, X_d$ but they can also be categorical. The corresponding graph of the SEM is obtained simply by drawing direct edges from each parent node to its direct effects where in Eq. 1.1, PA_j defines the set of parents for variable X_j. SEMs are also called functional models.

1.4 3 Building Blocks of Causal Graphs

The fundamental components of a causal structure can be categorized into three distinct types, each with unique statistical implications. These types are as follows:

1.4.1 Chains

A chain is a substructure in a causal path between vertices A, B, and C where the subsequent pairs of vertices are connected with edges that have the same direction.

Figure 1.1: Two possible chain substructures between three variables A, B, and C.

Rule 1 (Conditional Independence in Chains): Variables A and B are conditionally independent when conditioned on C, provided there exists only one directional path U between A and B, with C being any set of variables that intercepts this path.

1.4.2 Forks

A fork is a substructure in a causal pathway between vertices A, B, and C where one of them is a common cause or parent of the remaining two variables.

Rule 2 (Conditional Independence in Forks): When C acts as a common cause

Figure 1.2: A fork substructure between three variables A, B, and C where C is a common cause of variables A and B.

of A and B, and there is a single path U connecting A and B, then A and B are conditionally independent when conditioned on C.

1.4.3 Colliders

A collider is a substructure in a causal path between vertices A, B, and C where one of them is the common effect or child of the remaining two variables.

Figure 1.3: A collider substructure between three variables A, B, and C where C is a common effect of variables A and B.

Rule 3 (Conditional Independence in Colliders): Let C is a colliding node between A and B and there exist only single path U connecting A and B, then the variables A and B are unconditionally independent. But they are dependent conditional on C and any descendants of C.

1.5 Causal Assumptions

Causal discovery based solely on observational data relies on various causal assumptions. The stability and precision of the inferred structure are significantly influenced by the extent to which these assumptions are fulfilled in the observational data. Some of the

most prevalent causal assumptions employed by numerous causal discovery algorithms are outlined below:

1.5.1 Acyclicity

The data distribution P_x is assumed to be generated by a causal graph G that is a directed acyclic graph or DAG.

1.5.2 Causal Markov Assumption (CMA)

The causal Markov assumption assumes that a variable X_i is independent of its non-descendants (non-effects) in the causal graph conditional to its parents (direct causes). Based on this assumption, in a causal graphical model CGM(P_x, G), the joint distribution P_x can be factorized as:

$$p(x) = \prod_{i=1}^{d} p(x_i | x_{pa_i}^G) \tag{1.2}$$

where $X_{pa_i}^G$ refers to the set of parents for the variable X_i in DAG G.

1.5.3 Causal Faithfulness Assumption (CFA)

This assumption refers that the joint distribution P_x is considered faithful to the causal graph G if all conditional independence relationships in P_x are reflected by the structure of the graph G [32].

1.5.4 Causal Sufficiency Assumption (CSA)

The set of variables $X = (X_1, \ldots, X_n)$ is assumed to be causally sufficient. Consideration of the causal sufficiency assumption turns the causal graph complete and also incomplete in two different senses. It assumes that there are no confounders such that the pair $\{X_i, X_j\}$ in X do not share a common cause outside of $X_{\setminus i,j}$, where $X_{\setminus i,j}$ denotes the set of all variables in X except X_i and X_j. However, having an unmeasured intermediate variable in a path that does not satisfy the condition of a confounder is allowed.

CHAPTER 2: EVALUATION OF INDUCED EXPERT KNOWLEDGE IN CAUSAL STRUCTURE LEARNING BY NOTEARS

2.1 Introduction

Machine Learning models have been consistently setting new benchmarks for predictive accuracy. However, out-of-distribution (OOD) generalization continues to be a significant challenge. One approach to address this is by employing causal structures [33] to constrain models and eliminate spurious correlations. The underlying causal knowledge of the problem of interest can significantly help with domain adaptability and OOD generalization [34]. Furthermore, causal models go beyond the capability of correlation-based models to produce predictions. They provide us with powerful counterfactual reasoning and interventional mechanisms to reason under various what-if scenarios [18].

Two of the most prominent approaches in observational causal discovery are constraint-based and score-based methods [35–39]. Although these methods are quite robust if the underlying assumptions are true, they are computationally expensive and their computational complexity increases with the number of system variables due to the combinatorial nature of the DAG constraint. NOTEARS [40] tackles this problem with an algebraic characterization of acyclicity which reduces the combinatorial problem to a continuous constrained optimization. Different approaches [26, 41–43] have been proposed as the nonlinear or nonparametric extensions of this linear continuous optimization, which provides flexibility in modeling different causal mechanisms.

Learning the causal structure purely based on observational data is not a trivial task due to various limitations such as finite sampling, unobserved confounding factors, selection bias, and measurement errors [20–22]. These can result in spurious cause-effect relation-

ships. To address these issues in practice, researchers frequently induce the structure learning process with prior domain knowledge as featured in software packages such as CausalNex[1], causal-learn[2], bnlearn [23], gCastle [24], and DoWhy [25]. Heindorf et al. [44] in their work attempt to construct the first large scale open domain causality graph that can be included in the existing knowledge bases. The work further analyzes and demonstrates the benefits of large scale causality graphs in causal reasoning. Given a partial ancestral graph (PAG), representing the qualitative knowledge of the causal structure, Jaber et al. [45] in their study compute the interventional distribution from observational data. Combining expert knowledge with structural learning further constrains the search space minimizing the number of spurious mechanisms [46] and researchers often leverage this background knowledge by exploiting them as additional constraints for knowledge-enhanced event causality identification [47]. O'Donnell et al. [48] use expert knowledge as prior probabilities in learning Bayesian Network (BN) and Gencoglu and Gruber [49] use the linear NOTEARS model to incorporate knowledge to detect how different characteristics of the COVID-19 pandemic are causally related to each other. Different experts' causal judgments can be aggregated into collective ones [50] and Alrajeh et al. [51] in their work, studied how these judgments can be combined to determine effective interventions. An interesting exploration by Andrews et al. [52] defines tiered background knowledge and shows that with this type of background knowledge the FCI algorithm [35] is sound and complete.

However, understanding how to effectively incorporate and evaluate the impact of induced knowledge is yet to be explored and knowledge regarding this can mitigate some of the challenges of observational causal discovery. Human expertise is crucial in evaluating the learned causal structure [53, 54]. In practice, the process of human assessment and validation typically occurs in an iterative or sequential fashion [55–57]. In structure

[1]https://github.com/quantumblacklabs/causalnex
[2]https://github.com/cmu-phil/causal-learn

learning, this approach is particularly feasible for large causal networks, where the process involves learning, validating, and incorporating new knowledge sequentially through iterative feedback loops. The goal of this study is not to create a new causal discovery algorithm but rather to study this iterative interaction between prior causal knowledge from domain experts that takes the form of model constraints and a state-of-the-art causal structure learning algorithm. Wei et al. [46] have been the first to augment NOTEARS with additional optimization constraints to satisfy the Karush-Kuhn-Tucker (KKT) optimality conditions and Fang et al. [58] in their work leverages the low rank assumption in the context of causal DAG learning by augmented NOTEARS that shows significant improvements. However, none of them have studied the impact of induced knowledge on causal structure learning by augmenting NOTEARS with optimization constraints. For completeness, Section 2.3 provides the formulation of nonparametric NOTEARS [26] with functionality to incorporate causal knowledge in the form of known direct causal and non-causal relations. Nevertheless, this work aims to study the impact of expert causal knowledge on causal structure learning.

Most of the materials presented in this chapter have been published in a conference paper [27]. The main contributions of the current study can be summarized as follows. (1) This work demonstrates an iterative modeling framework to learn causal relations, impose causal knowledge to constrain the causal graphs, and further evaluate the model's behavior and performance. (2) The current study empirically evaluates and demonstrates that: (a) knowledge that corrects the model's mistake can lead to statistically significant improvements, (b) constraints on active edges have a larger positive impact on causal discovery than inactive edges, and (c) the induced knowledge does not correct on average more incorrect active and/or inactive edges than expected. Finally, the impact of additional knowledge in causal discovery is illustrated on a real-world dataset.

This chapter is structured as follows: Section 2.2 introduces the background on causal graphical models (CGMs), score-based structure recovery methods, and a study using

the score-based approach formulated as a continuous optimization and its recent nonparametric extension. Section 2.3 presents the extension of the nonparametric continuous optimization to incorporate causal knowledge in structure learning and detail the proposed knowledge induction process. Section 2.4 shows the empirical evaluations and comparative analyses of the impact of expert knowledge on the model's performance. Finally, in Section 2.5, a summary of the findings and a brief discussion on future work is provided.

2.2 Background

This section reviews the basic concepts related to causal structure learning and briefly covers a recent score-based continuous causal discovery approach using structural equation models (SEMs).

2.2.1 Causal Graphical Model (CGM)

A directed acyclic graph (DAG) is a directed graph without any directed cyclic paths [35]. A causal graphical model CGM$(P_X, \mathcal{G})$ can be defined as a pair of a graph $\mathcal{G}$ and an observational distribution P_X over a set of random variables $X = (X_1, \ldots, X_d)$. With respect to $\mathcal{G}$ this distribution P_X is Markovian where $\mathcal{G} = (V, E)$ is a DAG that encodes the causal structures among the random variables $X_i \in X$ [30]. The node $i \in V$ corresponds to the random variable $X_i \in X$ and edges $(i, j) \in E$ correspond to the causal relations encoded by $\mathcal{G}$. In a causal graphical model, the joint distribution P_x can be factorized as $p(x) = \prod_{i=1}^{d} p(x_i | x_{pa_i}^{\mathcal{G}})$ where $X_{pa_i}^{\mathcal{G}}$ refers to the set of parents (direct causes) for the variable X_i in DAG $\mathcal{G}$ and for each $X_j \in X_{pa_i}^{\mathcal{G}}$ there is an edge $(X_j \to X_i) \in E$ [30].

2.2.2 Score-based Structure Recovery

In a structure recovery method, with n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x_1}| \ldots |\mathbf{x_d}] \in \mathbb{R}^{n \times d}$, the objective is to uncover the causal relationships represented by the DAG $\mathcal{G}$. Most of the approaches follow either a constraint-based or a score-based strategy for observational causal discovery. A score-based approach typically concentrates

on identifying the DAG model $\mathcal{G}$ that fits the observed set of data $\mathbf{X}$ according to some scoring criterion $S(\mathcal{G}, X)$ over the discrete space of DAGs $\mathbb{D}$ where $\mathcal{G} \in \mathbb{D}$ [38]. The optimization problem for structure recovery in this case can be defined as follows:

$$\begin{aligned} \min_{\mathcal{G}} \quad & S(\mathcal{G}, X) \\ \text{subject to} \quad & \mathcal{G} \in \mathbb{D} \end{aligned} \tag{2.1}$$

The challenge with Eq. 2.1 is that the acyclicity constraint in the optimization is combinatorial in nature and scales exponentially with the number of nodes d in the graph. This makes the optimization problem NP-hard [59, 60].

2.2.3 NOTEARS: Continuous Optimization for Structure Learning

NOTEARS [40] is a score-based structure learning approach that reformulates the combinatorial optimization problem to a continuous one through an algebraic characterization of the acyclicity constraint in Eq. 2.1 via trace exponential. This method encodes the graph $\mathcal{G}$ defined over the d nodes to a weighted adjacency matrix $W = [w_1|\dots|w_d] \in \mathbb{R}^{d\times d}$ where $w_{ij} \neq 0$ if there is an active edge $X_i \to X_j$ and $w_{ij} = 0$ if there is not. The weighted adjacency matrix W entails a linear SEM by $X_i = f_i(X) + N_i = w_i^T X + N_i$; where N_i is the associated noise. The authors define a smooth score function on the weighted matrix as $h(W) = \text{tr}(e^{W\circ W}) - \text{d}$ where $\circ$ is the Hadamard product and e^M is the matrix exponential of M. This embedding of the graph $\mathcal{G}$ and the characterization of acyclicity turns the optimization in Eq. 2.1 into its equivalent:

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times d}} \quad & L(W) \\ \text{subject to} \quad & h(W) = 0 \end{aligned} \tag{2.2}$$

where $L(W)$ is the least square loss over W and $h(W)$ score defines the DAG-ness of the graph.

2.2.4 Nonparametric Extension of NOTEARS

A nonparametric extension of the continuous optimization suggested by a subsequent study [26] uses partial derivatives for asserting the dependency of f_j on the random variables. The authors define f_j over the Sobolev space consisting of functions that are square integrable, along with their derivatives. The authors show that f_j can be independent of a random variable X_i if and only if $||\partial_i f_j||_{L^2} = 0$ where ∂_i denotes the partial derivative with respect to the i-th variable. This redefines the weighted adjacency matrix with $W(f) = W(f_1, \ldots, f_d) \in \mathbb{R}^{d \times d}$ where each W_{ij} encodes the partial dependency of f_j on variable X_i. As a result, we can equivalently write Eq. 2.2 as follows:

$$\begin{aligned} \min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} \quad & L(f) \\ \text{subject to} \quad & h(W(f)) = 0 \end{aligned} \tag{2.3}$$

for all $X_j \in X$. Two of the general instances proposed by the study in Ref. [26] are: NOTEARS-MLP and NOTEARS-Sob. A multilayer perceptron having h number of hidden layers and $\sigma : \mathbb{R} \to \mathbb{R}$ activation function can be defined as $M(\mathbf{X}; L) = \sigma(L^{(h)}\sigma(\ldots\sigma(L^{(1)}X))$ where $L^{(l)}$ denotes the parameters associated with l-th hidden layer. The authors in Ref. [26] show that if $||i\text{-th column of } L_j^{(1)}||_2 = 0$ then $M_j(\mathbf{X}; L)$ will be independent of variable X_i which replaces the association of partial derivatives in Eq. 2.3 and redefines the adjacency matrix as $W(\theta)$ with $W(\theta)_{ij} = ||i\text{-th column of } L_j^{(1)}||_2$ where $\theta = (\theta_1, \ldots, \theta_d)$; θ_k denoting the set of parameters for the $M_k(\mathbf{X}; L)$ (k-th MLP). With the usage of neural networks and the augmented Lagrangian method [61] NOTEARS-MLP solves the constrained problem in Eq. 2.3 as follows:

$$\begin{aligned} &\min_{\theta} F(\theta) + \lambda||\theta||_1 \\ &F(\theta) = L(\theta) + \frac{\rho}{2}|h(W(\theta))|^2 + \alpha h(W(\theta)) \end{aligned} \tag{2.4}$$

2.3 Knowledge Induction

In this current work, the multilayer perceptions network of NOTEARS-MLP proposed by Ref. [26] is used as the estimator. This study extends this framework to incorporate causal knowledge by characterizing the extra information as additional constraints in the optimization in Eq. 2.3.

Knowledge Type. This study distinguishes between two types of knowledge: (i) *known inactive* is knowledge from the true inactive edges (absence of direct causal relation), and (ii) *known active* is knowledge from the true active edges (presence of direct causal relation).

Knowledge Induction Process. The study adopts an interactive induction process, where the expert knowledge is informed by the outcome of the causal discovery model. Namely, the knowledge is induced to correct the mistakes of the model in the causal structure, in the hope that the new structure is closer to the true causal graph. This process is applied sequentially by correcting the mistakes of the model at each step.

In the following subsections, the formulation of the NOTEARS optimization with constraints and the sequential induction process are presented.

2.3.1 Expert Knowledge as Constraints

An induced knowledge associated with a true active edge, $X_i \rightarrow X_j$ (*known active*) enforces the corresponding cell in the adjacency matrix to be non-zero, $[W(\theta)]_{ij} \neq 0$. This study considers this knowledge as an inequality constraint in the extension of the optimization such that the following statement holds:

$$h^p_{ineq}(W(\theta)) > 0 \tag{2.5}$$

where p enumerates over all the inequality constraints due to induction from the set of *known active* and h_{ineq} is the penalty score associated with the violation of these

inequality constraints. On the other hand, knowledge associated with the true inactive edge, $X_i \nrightarrow X_j$ (*known inactive*) enforces the related cell in $W(\theta)$ to be equal to zero, $[W(\theta)]_{ij} = 0$ if the induction implies there should not be an edge from X_i to X_j. This knowledge is considered as an equality constraint in the optimization such as:

$$h_{eq}^{q}(W(\theta)) = 0 \tag{2.6}$$

where q enumerates over all the equality constraints, induced from the set of *known inactive* and h_{eq} is the penalty score associated with the violation of these equality constraints. With these additional constraints in Eqs. 2.5, 2.6 we can extend Eq. 2.3 to incorporate causal knowledge in the optimization as follows:

$$\begin{aligned} \min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} \quad & L(f) \\ \text{subject to} \quad & h(W(\theta)) = 0, \\ & h_{eq}^{q}(W(\theta)) = 0, \\ & h_{ineq}^{p}(W(\theta)) > 0 \end{aligned} \tag{2.7}$$

NOTEARS uses a thresholding step on the estimated edge weights to reduce false discoveries by pruning all the edges with weights falling below a certain threshold. Because of this, in practice, even the equality constraints in Eq. 2.6 become inequalities to allow for small weights. Finally, slack variables are introduced in the implementation to transform the inequality constraints into equality constraints (see detailed formulation in Appendix A).

By using the similar strategy suggested by Zheng et al. [26] with augmented Lagrangian

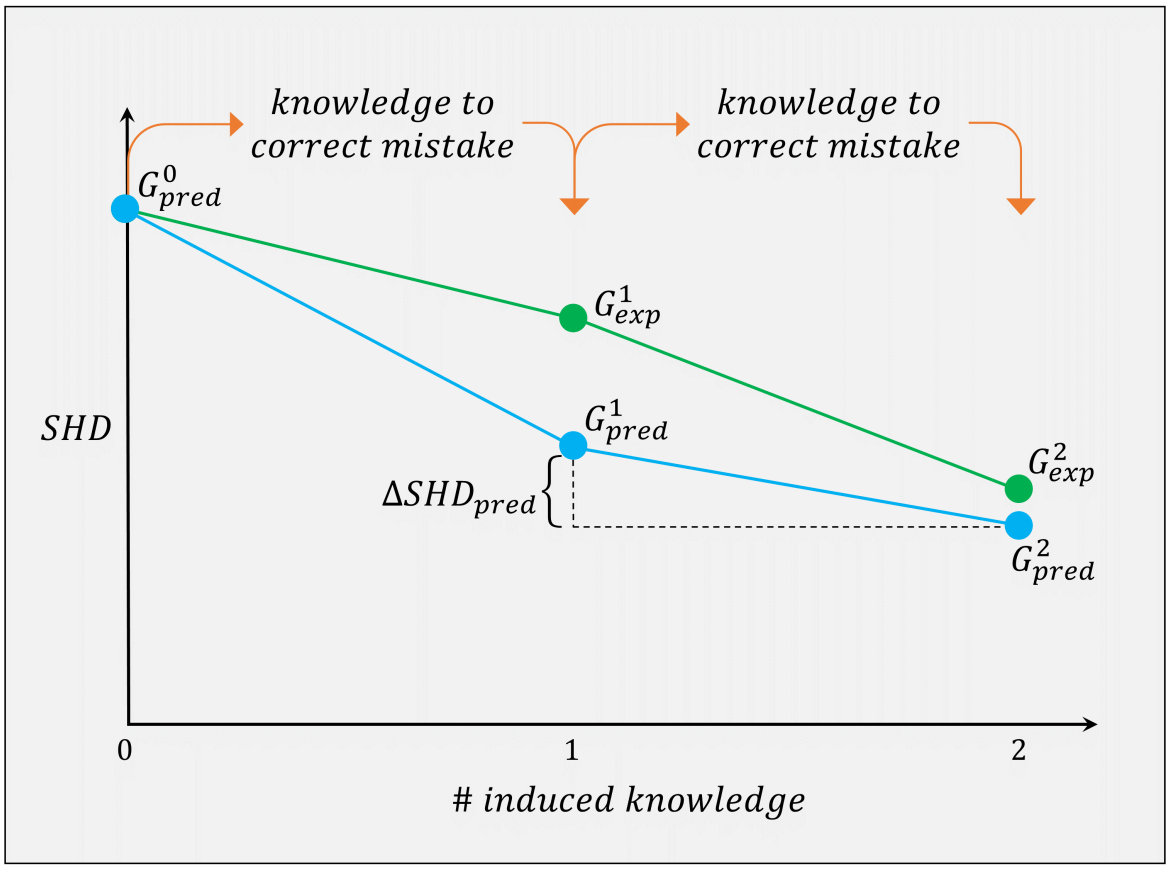


Figure 2.1: Knowledge induction process. Knowledge is induced by carrying over the existing knowledge set along with a new random correction informed by model mistakes.

method the reframed constrained optimization of Eq. 2.4 takes the following form:

$$\begin{aligned} &\min_{\theta} F(\theta) + \lambda ||\theta||_1 \\ F(\theta) &= L(\theta) + \frac{\rho}{2}|h(W(\theta))|^2 + \alpha h(W(\theta)) \\ &+ \sum_p (\frac{\rho_{ineq}}{2}|h^p_{ineq}(W(\theta))|^2 + \alpha_p h^p_{ineq}(W(\theta))) \\ &+ \sum_q (\frac{\rho_{eq}}{2}|h^q_{eq}(W(\theta))|^2 + \alpha_q h^q_{eq}(W(\theta))) \end{aligned} \tag{2.8}$$

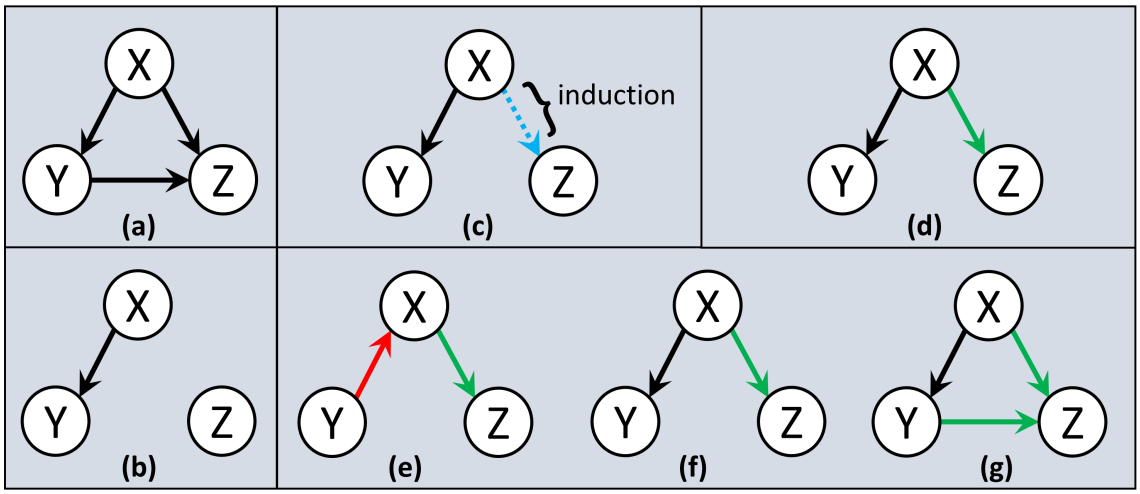


Figure 2.2: Expected graph formulation: (a) true graph, $\mathcal{G}_{true}$, (b) predicted graph by model at step k, $\mathcal{G}^{k}_{pred}$, (c) induced knowledge at step $(k+1)$, (d) expected graph at step $(k+1)$, $\mathcal{G}^{k+1}_{exp}$. Three different examples of many possible predicted graphs at step $(k+1)$, $\mathcal{G}^{k+1}_{pred}$ where the model performs (e) less than expectation, (f) par with expectation, and (g) more than expected.

2.3.2 Sequential Knowledge Induction

In the case of knowledge induction, the optimization is run in a sequential manner where the constraints are informed by the causal mistakes made by the model in the previous step. The process starts with the baseline model without imposing any additional knowledge from the true DAG and get the predicted causal graph denoted by $\mathcal{G}^{0}_{pred}$ in Figure 2.1. Then at each iterative step $(k+1)$, based on the mistakes in the causal graph $\mathcal{G}^{k}_{pred}$ predicted by the NOTEARS-MLP, one additional random piece of knowledge is selected to correct one of the mistakes and add it to the set of constraints identified in the previous k steps, and rerun NOTEARS. A batch of corrections can also be selected, however, this work has focused on estimating the contribution of each piece of knowledge in the form of known active/inactive edge. The major observations are illustrated in Section 2.4.1, Section 2.4.2, Section 2.4.3, and Section 2.4.4.

Expected Causal Graph. The expected causal graph, $\mathcal{G}_{exp}^{k+1}$ at step $(k+1)$ is formed by considering the case where all the knowledge has successfully been induced without impacting any other edges. Figure 2.2d illustrates an example of how to formulate the expected graph for a particular step in the iterative process. It is to be noted that the correction might yield a directed graph (Expected Causal Graph) that is not necessarily a DAG. The objective is to compare the performance between the causal graph predicted by NOTEARS and the expected causal graph. The intuition is that the induced knowledge will probably correct additional incorrect edges, see Figure 2.2g, yielding a performance better than expected.

2.4 Experiments

To empirically evaluate the impact of additional causal knowledge on causal learning and to keep the experimental setup similar to the study in Ref. [26], this study has used an MLP with 10 hidden units and sigmoid activation functions. In all the experimental setups, it is assumed that the prior knowledge is correct (agrees with the true DAG). Despite the known sensitivity of the NOTEARS algorithm to data scaling, as demonstrated in previous study [62], I have conducted experiments using both unscaled and scaled data to ensure the robustness of the findings and I am pleased to report that the conclusions remain unchanged regardless of the scaling of the data, indicating the stability and reliability of the results. While this chapter presents the results using the unscaled data for consistency with the original implementation of NOTEARS [26], it is important to note that the conclusions hold true even when the data is scaled.

Simulation. The performance of the proposed formulation and the impact of induced knowledge is investigated by comparing the DAG estimates with the ground truths. For the simulations with synthetic data, I have considered 16 different combinations following the simulation criteria: two random graph models, Erdos-Renyi (ER) and Scale-Free (SF), number of nodes, $d = \{10, 20\}$, sample size, $n = \{200, 1000\}$, edge density, $s0 = \{1d, 4d\}$.

Table 2.1: Performance metrics considered with their corresponding desirability.

Metric	Desirability
ΔFDR	Lower is better
ΔTPR	Higher is better
ΔFPR	Lower is better
ΔSHD	Lower is better

Table 2.2: Results for inducing redundant knowledge.

Metric	Mean ± Stderr.	Remarks
Δ FDR	-0.00030 ± 0.00017	No harm
ΔTPR	-0.00035 ± 0.00027	No harm
ΔFPR	-0.00097 ± 0.00059	No harm
ΔSHD	-0.00154 ± 0.00167	No harm

For each of these combinations, I have generated 10 different random graphs or true DAGs (as 10 trials for a particular combination) and corresponding data by following a nonlinear data generating process with index models (similar to the study in Ref. [26]) for which the underlying true DAGs are identifiable. The results are summarized over all these 160 random true DAGs and datasets. In the simulations, I have considered the regularization parameter, $\lambda = 0.01$. This study evaluates the performance of causal learning based on the mean and the standard error of different metrics. For statistical significance analysis, I have used t-test with $\alpha = 0.05$ as the significance level.

Metrics. For the comparative analysis, the current study evaluates performance using the following metrics: Structural Hamming Distance (SHD), True Positive Rate (TPR), False Positive Rate (FPR), and False Discovery Rate (FDR). However, since the evaluation is performed over all these 160 random graphs of varying sizes, this study considers Structural Hamming Distance per node (SHD/d) as SHD measure scales with the number of nodes (FDR, TPR, and FPR scale by definition). To evaluate the impact of induced knowledge, the differences in the metrics at different steps are calculated (where we have different sizes of induced knowledge set) and referred to as ΔFDR, ΔTPR, ΔFPR, and ΔSHD, see also Table 2.1. For example, based on the model's prediction I calculate the

Table 2.3: Results for inducing knowledge that corrects model's mistake.

Metric	Knowledge	Mean ± Stderr.	Improvement
ΔFDR	inactive	-0.018 ± 0.002	Significant
ΔFDR	active	-0.008 ± 0.001	Significant
ΔTPR	inactive	-0.007 ± 0.003	Not significant
ΔTPR	active	0.024 ± 0.003	Significant
ΔFPR	inactive	-0.023 ± 0.004	Significant
ΔFPR	active	-0.008 ± 0.003	Significant
ΔSHD	inactive	-0.032 ± 0.012	Significant
ΔSHD	active	-0.071 ± 0.011	Significant

impact of inducing one additional piece of knowledge on the metric SHD (ΔSHD_{pred}) as follows:

$$\Delta\text{SHD}_{pred} = \text{SHD}(\mathcal{G}^{k+1}_{pred}) - \text{SHD}(\mathcal{G}^{k}_{pred}) \tag{2.9}$$

Sanity Check - Redundant Knowledge Does No Harm. As part of the sanity check, this study investigates the impact of induced knowledge that matches the causal relationships successfully discovered by the NOTEARS-MLP. Therefore, in this section, I consider the set of edges that the baseline model correctly classifies as the knowledge source. Here, the study does not distinguish between the edge types of the induced knowledge (*known inactive & active*) since the goal is to investigate whether having redundant knowledge as additional constraints affects the model's performance or not. The results are illustrated in Table 2.2. The empirical evaluation shows that adding redundant knowledge does not deteriorate the performance of NOTEARS-MLP. The performed statistical test reflects that the results after inducing the knowledge from the correctly classified edge set are not statistically different than the results from the model without these knowledge inductions. However, I have noticed that the performance gets worse with highly regularized models. This is consistent with observations by Ng et al. [63] where sparse DAGs result in missing some of the true active edges.

2.4.1 Knowledge that Corrects Model's Mistake

Here, I first investigate the role of randomly chosen knowledge that corrects the model's mistake based on the cause-effect relations of the true graph. Therefore, in this case, I consider the set of misclassified edges from the estimated causal graph as the knowledge source for biasing the model. The results are illustrated in Table 2.3. The empirical result shows statistically significant improvements whenever the induced knowledge corrects misclassified edges in the estimated causal graph except for the case of ΔTPR with *known inactive* edges. However, this behavior is not totally unexpected since knowledge from *known inactive* edges helps to get rid of false discoveries or false positives, which hardly have an impact on true positives.

2.4.2 Known Inactive vs Known Active

In this subsection, I study the impact of different types of induced knowledge on causal discovery to correct the mistakes in the estimated causal graph. As a result, the experimental setup is similar to Section 2.4.1 where I consider the misclassified edge set as the knowledge source. This section considers both *known inactive* and *known active* types of knowledge to induce separately and analyze the differences of their impact on the performance. The results are illustrated in Table 2.4. Based on the statistical test, I have found that inducing *known inactive* is more effective when we compare the performance based on FDR and FPR as misclassification of inactive edges has more impact on these metrics. On the other hand, the results show that inducing *known active* is more effective

Table 2.4: Comparison between the impact of inducing knowledge regarding inactive vs active edges.

Metric	Inactive	Active	Better
ΔFDR	-0.019 ± 0.002	-0.008 ± 0.001	inactive
ΔTPR	-0.007 ± 0.003	0.024 ± 0.003	active
ΔFPR	-0.023 ± 0.004	-0.009 ± 0.004	inactive
ΔSHD	-0.033 ± 0.013	-0.072 ± 0.011	active

Table 2.5: Comparison between the empirical performance vs expectation.

Metric	Knowledge	Empirical	Expected	Remarks
ΔFDR	inactive	-0.019 ± 0.002	-0.016 ± 0.002	No difference
ΔFDR	active	-0.008 ± 0.001	-0.006 ± 0.001	No difference
ΔTPR	inactive	-0.007 ± 0.003	-0.002 ± 0.003	No difference
ΔTPR	active	0.024 ± 0.003	0.022 ± 0.002	No difference
ΔFPR	inactive	-0.023 ± 0.004	-0.021 ± 0.004	No difference
ΔFPR	active	-0.009 ± 0.003	-0.007 ± 0.003	No difference
ΔSHD	inactive	-0.033 ± 0.013	-0.047 ± 0.010	No difference
ΔSHD	active	-0.072 ± 0.011	-0.056 ± 0.010	No difference

on TPR as misclassification of active edges has more impact on this metric. Interestingly, the study has found that *known active* provides a significant improvement over *known inactive* in terms of SHD. This can be attributed to the fact that the induced knowledge based on the true inactive edge (*known inactive*) between two random variables, i.e. from X_i to X_j allows for two extra degrees of freedom since it is still possible to have no edge at all or an active edge from X_j to X_i. However, the induced knowledge based on the true active edge doesn't allow any degrees of freedom. This type of knowledge is more restraining for causal graph discovery and therefore carries more information.

2.4.3 Empirical Performance vs Expectation

In this subsection, I investigate in understanding whether inducing knowledge to correct the model's mistakes exceeds the expected improvement. The experimental setup is similar to Section 2.4.1 and Section 2.4.2 where I consider the misclassified edge set as the knowledge source. I have conducted the experiments using both *known inactive* and *known active* types of knowledge separately. The expected causal graph, $\mathcal{G}_{exp}$ is formulated in a similar manner described in Fig. 2.2. Table 2.5 shows the summary of the performance comparison in these cases with the expected results. The statistical test shows that the induced correct knowledge does not correct on average more incorrect active and/or inactive edges than expected. Therefore, using the information from induced knowledge does not have an additional impact than expected in the global optimization scheme. How-

ever, this is likely due to the fact that the structure of the expected causal graph, $\mathcal{G}_{exp}$ is not well-posed. It's worth noting that $\mathcal{G}_{exp}$ isn't necessarily a DAG since there isn't any constraining mechanism to enforce acyclicity as compared to $\mathcal{G}_{pred}$ (NOTEARS imposes hard acyclicity constraint in the continuous optimization). Although it is to be noted here that solving an acyclicity-constrained optimization problem does not guarantee to return a DAG and Ng et al. [64] in their study illustrates on this behavior and proposes the convergence guarantee with a DAG solution.

2.4.4 Real Data

Here I evaluate the implication of incorporating expert knowledge on the dataset from the study in Ref. [19], which is largely used in the literature of probabilistic graphical models with a consensus network accepted by the biological community. This dataset contains the expression levels of phosphorylated proteins and phospholipids in human cells under different conditions. The dataset has $d = 11$ cell types along with $n = 7466$ samples of expression levels. As for the ground truth of the underlying causal graph, the current study has considered $s_0 = 20$ active edges as suggested by the study [19]. I have opted for ΔTPR, the percentage difference of edges in agreement (higher is better), and the percentage difference of reversed edges (lower is better) as the evaluation metrics since the performance on these metrics would indicate the significance more distinctively. Similar to the synthetic data analysis, this study had 10 trials that I used to summarize the evaluation. The empirical results (Mean $\pm$ Stderr.) show: ΔTPR as 0.020 ± 0.004, the percentage difference of edges in agreement as 0.393 ± 0.086, and the percentage difference of reversed edges as -0.073 ± 0.030. I have found that, with the help of induced knowledge, the model demonstrates statistically significant improvement by correctly identifying more active edges and by reducing the number of edges identified in the reverse direction. Due to the limitation of having access only to a subset of the true active edges, my analyses could not include a comparative study on *known inactive* edges as in the synthetic data

case. It is assumed that the performance could have been improved by fine-tuning the model's parameters but since the main focus of this study is entirely based on the analyses regarding the impact of induced knowledge of different types and from different sources on structure learning, I kept the parameter setup similar for all consecutive steps in the knowledge induction process.

2.5 Summary

In this study, I have investigated the impact of expert causal knowledge on causal structure learning and provided a set of comparative analyses of biasing the model using different types of knowledge. The findings show that knowledge that corrects the model's mistakes yields significant improvements and it does no harm even in the case of redundant knowledge that results in redundant constraints. This suggests that the practitioners should consider incorporating domain knowledge whenever available. More importantly, I have found that knowledge related to active edges has a larger positive impact on causal discovery than knowledge related to inactive edges which can mostly be attributed to the difference between the number of degrees of freedom each case reduces. This finding suggests that the practitioners may want to prioritize incorporating knowledge regarding the presence of an edge whenever applicable. Furthermore, the experimental analysis shows that the induced knowledge does not correct on average more incorrect active and/or inactive edges than expected. This finding is rather surprising to me, as I have expected that every constraint based on a known active/inactive edge to impact and correct more than one edge on average.

This work points to the importance of the human-in-the-loop in causal discovery that can be further explored. Also, I would like to mention that in this study I have adopted hard constraints to accommodate the prior knowledge since I have assumed the priors to be correct. An interesting future direction would be to accommodate continuous optimization with functionality to allow different levels of confidence in the priors.

CHAPTER 3: CD-NOTEARS: CONCEPT DRIVEN CAUSAL DISCOVERY IN HIGH DIMENSIONAL DATA USING NOTEARS

3.1 Introduction

In recent years, the field of causal discovery has gained significant traction, driven by advancements in machine learning models that excel in handling large datasets and approximating intricate relationships. Consequently, numerous methods have emerged to infer causal relationships from observational data. These methods can be categorized into constraint-based algorithms e.g. PC [35], IC [18], and FCI [37], score-based approaches e.g. GES [38] and FGES [39], and functional causal models e.g. LiNGAM [65] and ANMs [66]. Constraint-based methods utilize conditional independence tests and rules to detect edge directions, often pinpointing the Markov equivalence class of the genuine causal graph. Meanwhile, score-based models target causal graph optimization over the DAG space, a process that becomes computationally intensive due to its combinatorial nature. NOTEARS, present in linear [40] and non-parametric [26] forms, adopts an algebraic acyclicity characterization, transforming the combinatorial challenge into continuous constrained optimization. Variants of this continuous optimization approach have surfaced in works Ref. [41,42,67], offering versatile causal mechanism modeling. While NOTEARS stands out for its efficacy across diverse uses, it's not limited to structure learning for continuous or scalar data but extends to feature vectors of conceptual data as well.

For example, consider an IMDb movie dataset with three concepts: revenue (C1), genre (C2), and synopsis (C3). Revenue (C1) represents movie-generated revenue (X1). Assuming our dataset has only thriller and sci-fi genres, we can use one-hot encoding to represent the genre (C2), creating a two-dimensional vector (X2 and X3) for these genres. For the

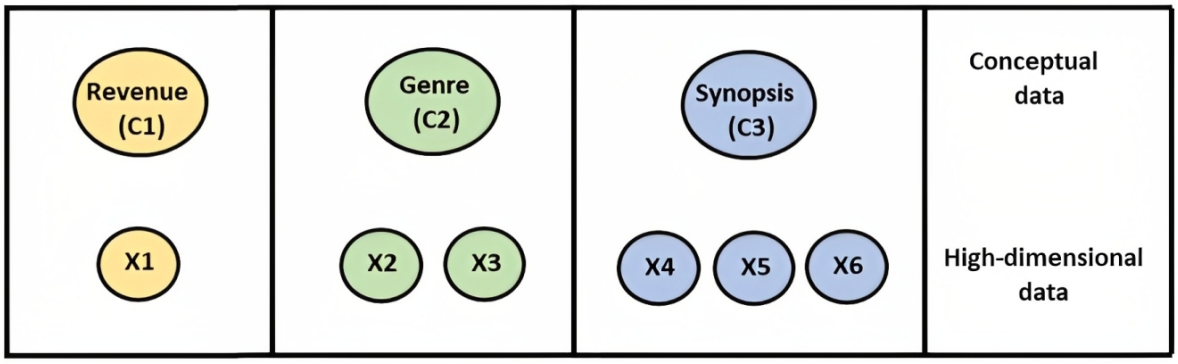


Figure 3.1: Mapping of conceptual data to high-dimensional features for movie dataset. The three main concepts considered are revenue (C1), genre (C2), and synopsis (C3). The one-dimensional feature X1 corresponds to revenue, while the encoding of genre results in two-dimensional features X2 and X3. Synopsis is represented by a three-dimensional embedding with features X4, X5, and X6.

movie synopsis (C3), we can use NLP methods to produce a three-dimensional embedding (X4, X5, and X6). Thus, our dataset has three concepts (C1, C2, C3) leading to a six-dimensional feature vector for each movie (X1 through X6, as depicted in Fig. 3.1). By applying NOTEARS to this vector-valued data, causal relationships within the high-dimensional feature space can be discerned, shedding light on the interconnections between features X1 through X6. However, a general challenge with structure learning is that uncovering the causal structure requires complete coverage of the data distribution. Intuitively, without a comprehensive representation of the data distribution, one can miss latent causal relationships or infer spurious ones due to sample biases. To address this challenge, researchers often provide algorithms with additional knowledge to augment the optimization with prior knowledge, as featured in software packages such as CausalNex [1], causal-learn [2], bnlearn [23], DoWhy [25], and gCastle [24]. Previous studies have shown that incorporating domain knowledge can be beneficial and lead to superior performance. For example, the impact of prior knowledge on score-based causal learning algorithms was evaluated in Ref. [27, 68]. Additionally, another recent study [69] presents KGS, a

[1] https://github.com/quantumblacklabs/causalnex
[2] https://github.com/cmu-phil/causal-learn

novel knowledge-guided greedy score-based causal discovery approach that uses structural priors to constrain the search space and guide the process.

This chapter presents CD-NOTEARS, an extension of the NOTEARS algorithm designed for concept-driven causal structure learning in vector-valued data. This novel approach integrates prior knowledge on relations between concepts and high-dimensional features as meta-information, imposing DAGness on concept-level data, a departure from the original NOTEARS which operates on raw high-dimensional features. Through extensive experiments on varied datasets, this study showcases the proposed method's proficiency in identifying causal relationships, highlighting its enhanced performance compared to the original NOTEARS. The key contributions can be summarized as follows: (1) A novel extension of the NOTEARS algorithm is proposed, that facilitates concept-driven causal structure learning in vector-valued data while incorporating prior relations between different concepts and high-dimensional features, preserving the non-parametric essence of the original NOTEARS algorithm, (2) Departing from traditional methods, this approach emphasizes DAGness at the concept level rather than focusing solely on high-dimensional raw features, and (3) The proposed study illustrates empirical validation through comprehensive experiments on synthetic, benchmark, and real-world datasets.

Most of the materials presented in this chapter are published in Ref. [28]. The remainder of this chapter is organized as follows: Section 3.2 delves into the methodology of CD-NOTEARS, Section 3.3 presents the experimental settings and evaluations. Finally, Section 3.4 encapsulates the conclusions and highlights the significant takeaways.

3.2 Methodology

The proposed CD-NOTEARS method builds on the original nonparametric NOTEARS algorithm [26], specifically the NOTEARS-MLP instance, to infer causal relationships from vector-valued data. In this section, I summarize the background of linear [40] and nonparametric [26] extensions of NOTEARS and then delve into the proposed adaptation:

the CD-NOTEARS approach.

Observational causal structure learning aims to learn the causal relationships encoded by a DAG $\mathcal{G}$ from n i.i.d. observations in the data matrix $\mathbf{X} = [\mathbf{x_1}|\ldots|\mathbf{x_d}] \in \mathbb{R}^{n\times d}$. The score-based approach focuses on identifying the DAG model $\mathcal{G}$ that best fits the observed data $\mathbf{X}$ based on a scoring criterion $S(\mathcal{G}, X)$ over the discrete space of DAGs $\mathbb{D}$ where $\mathcal{G} \in \mathbb{D}$ [38]. This optimization problem can be formulated as:

$$\begin{aligned} \min_{\mathcal{G}} \quad & S(\mathcal{G}, X) \\ \text{subject to} \quad & \mathcal{G} \in \mathbb{D} \end{aligned} \tag{3.1}$$

The linear NOTEARS [40] algorithm reformulates the combinatorial optimization in Eq. 3.1 to a continuous one through an algebraic characterization of the acyclicity constraint. This method encodes the graph $\mathcal{G}$ defined over the d nodes into a weighted adjacency matrix $W = [w_1|\ldots|w_d] \in \mathbb{R}^{d\times d}$ where $w_{i,j} \neq 0$ if there is an active edge $X_i \to X_j$ and $w_{i,j} = 0$ otherwise. The weighted adjacency matrix W entails a linear structural equation model (SEM) by $X_i = f_i(X) + N_i = w_i^T X + N_i$; where N_i is the associated noise. The authors define a smooth score function on the weighted matrix as $h(W) = \mathrm{tr}(e^{W\circ W}) - \mathrm{d}$ where $\circ$ is the Hadamard product and e^M is the matrix exponential of M. This reformulates Eq. 3.1 into its equivalent form:

$$\begin{aligned} \min_{W\in\mathbb{R}^{d\times d}} \quad & L(W) \\ \text{subject to} \quad & h(W) = 0 \end{aligned} \tag{3.2}$$

where $L(W)$ is the least square loss over W and $h(W)$ score defines the DAG-ness of the graph. The nonparametric NOTEARS [26] uses partial derivatives on the functional form f_j to determine the dependency of random variable X_j on other random variables. The authors define f_j over the Sobolev space consisting of functions that are square integrable, along with their derivatives, and f_j can be independent of random variable X_i if and only

if $||\partial_i f_j||_{L^2} = 0$ where ∂_i denotes the partial derivative with respect to X_i. This redefines the weighted adjacency matrix as $W(f)$ with each $W_{i,j}$ encoding the partial dependency of f_j on variable X_i and allows us to write Eq. 3.2 equivalently:

$$\begin{aligned} \min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} \quad & L(f) \\ \text{subject to} \quad & h(W(f)) = 0 \end{aligned} \tag{3.3}$$

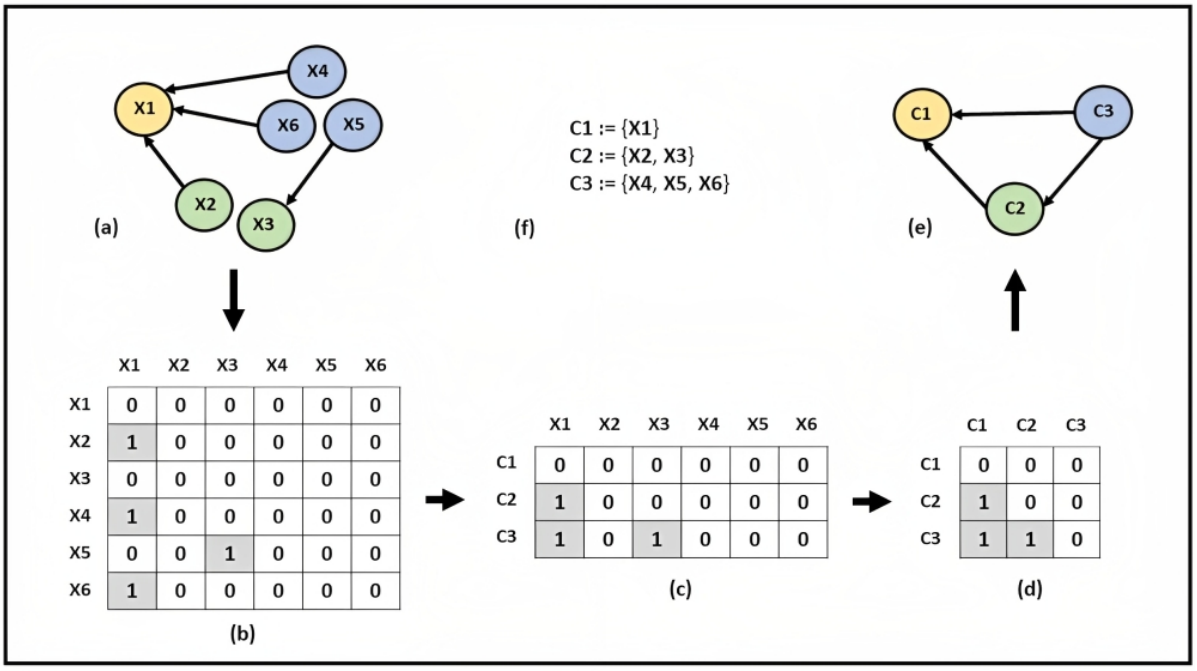


Figure 3.2: Illustration of concept-driven adjacency matrix and graph formulation process from high dimensional data: (a) graphical representation of relations between high dimensional features in raw data, (b) corresponding adjacency matrix for high dimensional graph relations, W, (c) intermediate matrix formulation obtained by applying row aggregation based on the concept-level meta-information, (d) concept-driven adjacency matrix obtained after full transformation using row and column aggregation, W^{con}, (e) graphical representation of the relations between concepts (C1, C2, C3), (f) Prior knowledge or meta-information regarding the concepts and their representations in high dimensional feature space. For the purpose of simplicity, this figure demonstrates the process using binary adjacency matrices.

While NOTEARS deduces causal relationships among features by applying a contin-

uous acyclicity constraint on the high-dimensional adjacency matrix, W, the proposed method, CD-NOTEARS, adopts a concept-driven strategy. Firstly, the method obtains the adjacency matrix similarly to NOTEARS. Instead of directly constraining this matrix, CD-NOTEARS transforms it into an aggregated adjacency matrix, W^{con}, using concept-level prior knowledge. This matrix captures concept-level relationships, with aggregation refining the optimization search space to guide the optimization. CD-NOTEARS imposes acyclicity on the concept-level relations captured in W^{con}. Fig. 3.2 illustrates the approach to derive concept-driven causal relations, W^{con}, from the high-dimensional matrix, W. In order to maintain consistency with the previous example presented in Fig. 3.1, I here demonstrate the matrix transformation using the three concepts introduced earlier. Therefore, C1 refers to the revenue of each movie, represented by a scalar-valued one-dimensional feature X1. Meanwhile, C2 and C3 correspond to the genre and synopsis concepts of the movie, represented by two-dimensional (X2 and X3) and three-dimensional (X4, X5, and X6) feature spaces, respectively. Unlike the original NOTEARS implementation that imposes acyclicity on the raw-level high-dimensional graph as shown in Fig. 3.2(a), CD-NOTEARS imposes acyclicity on the concept-level graph as in Fig. 3.2(e). To achieve the concept-level matrix, this method first generates the high-dimensional adjacency matrix (Fig. 3.2(b)). An intermediate matrix is then formed using row aggregation informed by concept-level meta-information (Fig. 3.2(c)). The final transformation, integrating both row and column aggregation, yields the concept-driven matrix W^{con} (Fig. 3.2(d)), influenced by the relations between concepts and features shown in Fig. 3.2(f). It is to be noted that various matrix transformation or aggregation methods can be employed to get the concept-level relations from the raw relations, as long as they preserve the causal relationships from the raw level to the concept level. Such

transformation or aggregation function should satisfy the following equation:

$$W^{con}_{m,\ n} \ = \ \begin{cases} 0 & \text{if } \forall(X_i \in C_m, X_j \in C_n)\ W_{i,j} = 0 \\ \neq 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Eq. 3.4 allows us to aggregate the raw-level information in W and determine the relationship between concepts such as C_m and C_n. If any of the random variables X_i that belong to concept C_m has a causal link in high-dimensional feature space to any other random variable X_j that belongs to concept C_n, the corresponding cell in the concept-level aggregated matrix, $W^{con}_{m,n}$ should reflect that relationship. Otherwise, the cell in the concept-level matrix is set to zero. After applying the transformation using Eq. 3.4, the optimization problem reformulates to:

$$\begin{aligned} &\min_{f: f_j \in H^1(\mathbb{R}^d), \forall j \in [d]} && L(f) \\ &\text{subject to} && h(W^{con}(f)) = 0 \end{aligned} \tag{3.5}$$

To solve the optimization problem, I have used the augmented Lagrangian method [61], similar to the strategy followed by the original NOTEARS. Therefore, the proposed CD-NOTEARS method preserves the non-parametric nature of the original NOTEARS algorithm while leveraging concept-level meta-information.

3.3 Experiments and Results

To evaluate the extended NOTEARS algorithm, CD-NOTEARS, I conducted case studies comparing its performance against the original NOTEARS model. Given the sensitivity of the NOTEARS algorithm to data scaling, as shown in earlier studies [62, 70], I have scaled the data using the *standardization* method from Python's scikit-learn [71] library. This study ensures consistent model structures by employing an MLP with 10 hidden units and sigmoid activations for both models. While CD-NOTEARS integrates

meta-information during optimization, focusing on concept-level relations, the original NOTEARS first learns the causal graph in the high-dimensional feature space, then post-processes with meta-information. In this study, I have adopted the 'mean' as the aggregation function in both models for concept-level causal graph learning. For comparative analysis, I have utilized two key performance metrics: false discovery rate (FDR) and structural hamming distance (SHD). The FDR, in particular, offers insights into the conservativeness of the method. A lower FDR indicates fewer unwarranted causal claims, addressing the challenge highlighted by previous study [72] regarding non-conservative error trade-offs seen in many causal discovery methods. On the other hand, the SHD, a widely-recognized pattern metric for evaluating causal discovery methodologies [73], provides a holistic view of how closely the predicted graph aligns with the ground truth. To emphasize reliability, I have conducted 50 different random trials for each case study, evaluating the performance of both models based on the mean and standard deviation of the performance metrics. The statistical significance analysis is then performed using a t-test with α level of 0.05.

3.3.1 Synthetic Dataset

To compare the efficacy of CD-NOTEARS against the original NOTEARS method, I first ran simulations on synthetic datasets. This study examined 16 combinations, varying between Erdos-Renyi and Scale-Free graph models (gt = ER, SF), number of nodes (d = 10, 20), sample sizes (n = 200, 1000), and edges (s0 = 1d, 4d), where d indicates node count. Each combination yielded 50 random graphs or true DAGs, generated via the Additive Noise Model (ANM) with MLPs following the methodology in the original work [26]. For the experiments with synthetic datasets, I have considered two different ranges for the dimension of each concept. In the first case, the range was limited to 1 to 3, and in the second case, the range was expanded to 1 to 5. The results are presented in Table 3.1 and Table 3.2, respectively. The evaluation showcases the superiority of CD-

Table 3.1: Performance evaluation of CD-NOTEARS and the original NOTEARS implementation on synthetic data considering random variables as concepts having dimension ranges from 1 to 3.

n	d	s0	gt	fdr		shd	
				CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
200	10	10	ER	**0.86 ± 0.04**	0.89 ± 0.02	**37.84 ± 2.20**	47.36 ± 2.99
		40	SF	0.89 ± 0.04	0.90 ± 0.03	**38.90 ± 2.27**	47.77 ± 2.55
	20	20	ER	**0.48 ± 0.11**	0.56 ± 0.05	**22.04 ± 4.90**	35.78 ± 5.02
		80	SF	**0.58 ± 0.07**	0.66 ± 0.05	**26.20 ± 3.35**	40.30 ± 4.43
	10	10	ER	0.93 ± 0.01	0.94 ± 0.01	**165.64 ± 6.34**	188.09 ± 6.07
		40	SF	0.94 ± 0.02	0.94 ± 0.01	**167.82 ± 6.77**	184.50 ± 6.06
	20	20	ER	**0.75 ± 0.04**	0.77 ± 0.03	**139.42 ± 6.91**	166.78 ± 10.03
		80	SF	**0.78 ± 0.05**	0.81 ± 0.03	**142.90 ± 10.37**	167.21 ± 9.06
1000	10	10	ER	0.83 ± 0.14	0.86 ± 0.07	**22.12 ± 4.91**	33.78 ± 8.07
		40	SF	0.88 ± 0.09	0.86 ± 0.08	**21.42 ± 5.58**	30.70 ± 7.23
	20	20	ER	**0.48 ± 0.15**	0.54 ± 0.09	**30.80 ± 4.41**	33.88 ± 5.13
		80	SF	**0.55 ± 0.18**	0.63 ± 0.10	**27.26 ± 4.57**	32.54 ± 4.61
	10	10	ER	0.93 ± 0.03	0.92 ± 0.02	**122.26 ± 20.65**	152.18 ± 19.26
		40	SF	0.95 ± 0.03	**0.94 ± 0.02**	**124.88 ± 14.90**	149.30 ± 17.91
	20	20	ER	**0.72 ± 0.05**	0.76 ± 0.03	**119.34 ± 9.89**	147.36 ± 11.28
		80	SF	0.76 ± 0.07	0.78 ± 0.05	**115.38 ± 15.17**	140.56 ± 16.82

Table 3.2: Performance evaluation of CD-NOTEARS and the original NOTEARS implementation on synthetic data considering random variables as concepts having dimension ranges from 1 to 5.

n	d	s0	gt	fdr		shd	
				CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
200	10	10	ER	**0.86 ± 0.04**	0.89 ± 0.01	**37.70 ± 1.78**	50.33 ± 2.16
		40	SF	**0.86 ± 0.03**	0.90 ± 0.01	**38.24 ± 1.66**	50.39 ± 2.30
	20	20	ER	**0.48 ± 0.12**	0.57 ± 0.04	**21.92 ± 5.37**	42.50 ± 5.17
		80	SF	**0.59 ± 0.10**	0.67 ± 0.04	**26.64 ± 4.79**	45.16 ± 4.14
	10	10	ER	**0.93 ± 0.01**	0.94 ± 0.01	**161.68 ± 6.44**	194.46 ± 5.92
		40	SF	0.94 ± 0.02	0.95 ± 0.01	**165.14 ± 6.15**	195.42 ± 4.56
	20	20	ER	**0.76 ± 0.04**	0.78 ± 0.01	**139.66 ± 6.65**	179.57 ± 7.04
		80	SF	**0.78 ± 0.05**	0.81 ± 0.02	**139.28 ± 10.15**	180.54 ± 10.21
1000	10	10	ER	0.85 ± 0.06	0.87 ± 0.03	**30.14 ± 4.89**	43.60 ± 4.67
		40	SF	0.89 ± 0.06	0.89 ± 0.03	**29.90 ± 5.14**	43.78 ± 5.33
	20	20	ER	**0.45 ± 0.12**	0.55 ± 0.07	**25.30 ± 5.23**	37.88 ± 6.13
		80	SF	**0.56 ± 0.15**	0.64 ± 0.08	**26.48 ± 5.76**	38.62 ± 6.90
	10	10	ER	0.93 ± 0.02	0.93 ± 0.01	**137.40 ± 13.93**	177.26 ± 9.79
		40	SF	0.94 ± 0.02	0.94 ± 0.02	**136.46 ± 13.81**	174.08 ± 12.48
	20	20	ER	**0.74 ± 0.04**	0.78 ± 0.03	**127.02 ± 10.89**	169.86 ± 10.51
		80	SF	**0.77 ± 0.07**	0.79 ± 0.04	**123.34 ± 16.34**	162.56 ± 15.12

Table 3.3: Comparison of CD-NOTEARS and the original NOTEARS implementation on binary benchmark datasets.

dataset	fdr		shd	
	CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
Lucas	**0.76 ± 0.03**	0.82 ± 0.02	**12.16 ± 0.78**	22.96 ± 1.37
Asia	**0.75 ± 0.01**	0.87 ± 0.04	**9.02 ± 0.14**	16.00 ± 1.13

NOTEARS over the original implementation. By integrating prior knowledge into the graph formulation and imposing acyclicity at the concept level, CD-NOTEARS achieves lower FDR and SHD in most scenarios. This underscores the merit of employing concept-level knowledge for precise causal structure learning.

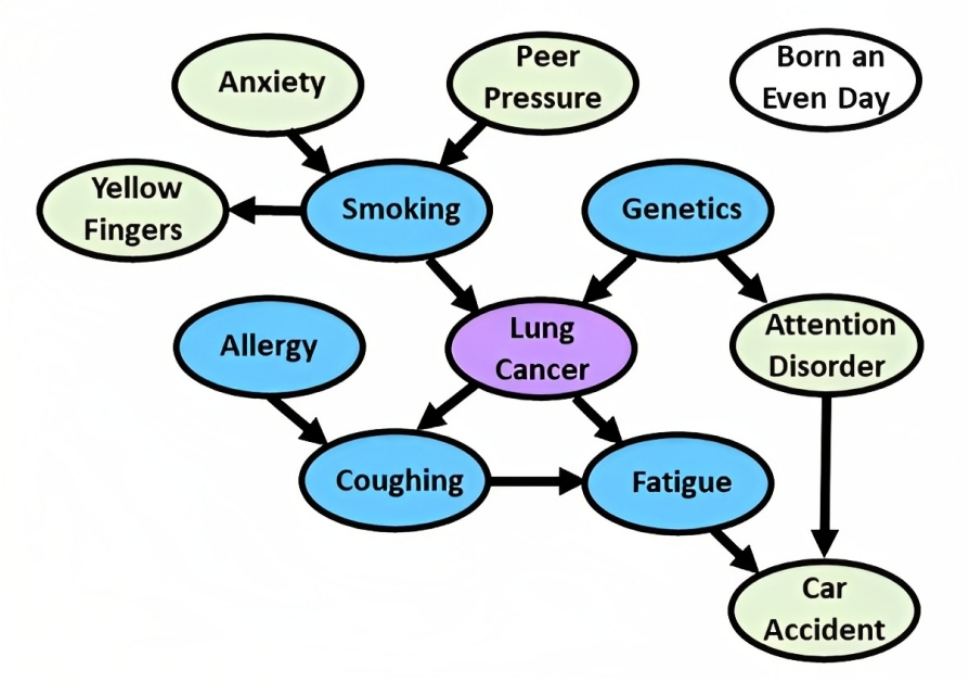


Figure 3.3: Causal graph for unmanipulated distribution of LUCAS0 [2]

3.3.2 Benchmark Dataset

Benchmark Datasets for Binary Variables Next, I compared CD-NOTEARS and the original NOTEARS on two benchmark datasets for categorical variables: LUCAS and ASIA. The LUCAS (LUng CAncer Simple set) dataset [2], sourced from the Causality Workbench project, comprises 2000 instances of 12 binary variables detailing factors af-

Table 3.4: Comparison of the CD-NOTEARS and original NOTEARS implementation on multinary benchmark datasets using PyTorch [1] embedding layer to generate vector-valued data from categorical variables.

dataset	fdr		shd	
	CD-NOTEARS	NOTEARS	CD-NOTEARS	NOTEARS
Dutch	**0.56 ± 0.29**	0.69 ± 0.07	**41.62 ± 1.74**	46.84 ± 2.56
Adult	**0.56 ± 0.20**	0.73 ± 0.07	**38.42 ± 1.34**	44.70 ± 2.23

fecting lung cancer. The data is synthetically created by causal Bayesian networks and in this study, I have used the unmanipulated distribution of the dataset referred to as LUCAS0 [3], as visualized in Fig 3.3. The second dataset, ASIA [74] depicts the interplay between tuberculosis, lung cancer, bronchitis, and Asia visits. Containing 8 binary variables and 5000 samples generated following the causal Bayesian network, its causal graph [75] and dataset [76] are available online. The evaluation, presented in Table 3.3, shows CD-NOTEARS surpassing NOTEARS in terms of FDR and SHD values on both datasets, emphasizing its effectiveness for concept-driven data with binary categorical variables.

Benchmark Datasets for Multinary Variables In the next experimental study, I have assessed CD-NOTEARS on two mixed numeric and multinary datasets: the Dutch Census [77] and the Adult dataset [78]. The Dutch Census has 60,420 entries with 12 attributes utilized for structural learning, such as sex, age, household_position, country_birth, occupation, etc. Among these attributes, sex and occupation are binary, while the remaining attributes can take multiple values. The Adult dataset comprises 32,561 samples with 11 attributes, including a combination of continuous and categorical variables such as age, working_class, sex, hours_per_week, marital_status, income, etc. Age and hours_per_week are continuous variables, while the rest are categorical. Among the categorical variables, sex and income are binary, and the remaining variables are multinary. This study considered the causal graph from a prior study [79] for both datasets.

[3] http://www.causality.inf.ethz.ch/data/LUCAS.html

To process multinary categorical variables, I have used PyTorch's [1] embedding layer to create vector embeddings for each concept. This technique efficiently manages mixed data, leading to a compact dataset. As illustrated in Table 3.4, CD-NOTEARS outperforms NOTEARS in the identification of causal structures from mixed data. By leveraging concept-level understanding and DAG properties, the proposed approach highlights the significance of conceptual insights in high-dimensional causal learning.

3.3.3 Real Data

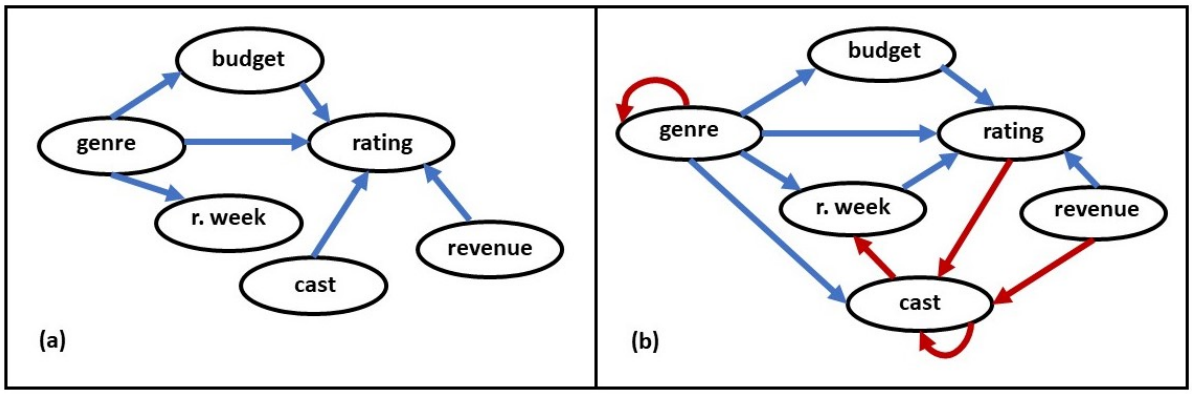


Figure 3.4: Causal relations obtained from the movie datasets using two different models: (a) CD-NOTEARS and (b) the original implementation of NOTEARS. r. week stands for the release week of the movie.

Finally, I have evaluated CD-NOTEARS and the original NOTEARS using the IMDb movie dataset sourced from two Kaggle repositories: IMDB Movie data Analysis [4] and Movie Scripts Corpus [5]. The dataset, after cleaning, had data on 1764 movies, including features like budget, cast, genre, release week, user rating, and revenue. Cast and genre are vector-valued features, while the remaining features are scalar in nature. As each movie sample can have one or more casts and genres, I have applied one-hot encoding to generate embeddings for each sample followed by training an auto-encoder to retain maximum information with lower dimensional features from these concepts. This process was

[4]https://www.kaggle.com/code/robinjrjr/imdb-movie-data-analysis/data
[5]https://www.kaggle.com/datasets/gufukuro/movie-scripts-corpus

applied independently to each of the multi-dimensional concepts in the dataset, namely cast and genre. To ensure a fair comparison, I kept the common settings of both model implementations similar. In total, CD-NOTEARS estimated six edges, which are budget $\rightarrow$ rating, cast $\rightarrow$ rating, genre $\rightarrow$ budget, genre $\rightarrow$ release week, genre $\rightarrow$ rating, and revenue $\rightarrow$ rating. Due to the absence of an established ground truth or consensus within the dataset, in this study, I depended on my own assessment to evaluate the predicted connections. Upon examination, I have discovered that the majority of causal relationships estimated by CD-NOTEARS appeared to be reasonable and coherent. However, the relationship between rating and revenue appears ambiguous as a higher rating of a movie can draw more people to watch the movie, resulting in increased revenue, and conversely, higher revenue could bias viewers to rate the movie higher. Despite this ambiguity, both implementations agreed on the direction of this relationship, suggesting it would not affect the comparative evaluation. Nevertheless, the original implementation of NOTEARS estimated six additional edges, some of which appeared unlikely such as cast $\rightarrow$ release week, rating $\rightarrow$ cast, and revenue $\rightarrow$ cast. Fig 3.4 illustrates the causal relations retrieved by both these models. Notably, NOTEARS applies DAGness to the raw-level high-dimensional features, which resulted in the generation of two self-loops for the concepts cast and genre. While this violates the acyclicity assumption, I found this characteristic intriguing as the selection of one cast may impact the selection of other casts, and a similar phenomenon may apply to genres. Nonetheless, the proposed CD-NOTEARS implementation, which enforces DAGness on the concepts, appears to surpass the original NOTEARS implementation in terms of performance. Although this section lacks a quantitative metric for assessing performance, the analysis of the IMDb movie dataset presents persuasive evidence in favor of CD-NOTEARS.

3.4 Summary

The proposed method, CD-NOTEARS, represents a significant advancement in the field of causal discovery for concept-driven data. By emphasizing acyclicity constraints at the concept level and leveraging prior feature-to-concept knowledge, it refines causal relationship representation, bolstering reliability and accuracy. Through evaluations of diverse datasets, this study has highlighted its efficacy, especially in sectors where conceptual data is prevalent such as healthcare, finance, and social science. This research emphasizes the benefits of integrating prior concept knowledge in causal structure learning, making CD-NOTEARS a valuable addition to the causal discovery repertoire. Looking ahead, there is potential to combine this concept-driven approach with other leading causal discovery methods to further amplify its potency. In conclusion, I firmly believe that the extension of the NOTEARS approach will be a pivotal asset for causal discovery across various domains. I hope that this research will inspire further studies and advancements in the field of causal discovery, ultimately leading to a better understanding of causality in complex systems and guiding effective causal learning methods.

CHAPTER 4: INVARIANT MOLECULAR REPRESENTATIONS FOR HETEROGENEOUS CATALYSIS

4.1 Introduction

The development of efficient and cost-effective catalysts is of great importance in the chemical industry, as catalysts play a crucial role in a wide range of chemical reactions to increase the efficiency and selectivity of these processes [80]. In recent years, computational catalyst screening has gained significant attention as a method for identifying promising catalysts, as it allows for the rapid and efficient evaluation of large numbers of potential catalysts [81,82]. Computational catalysis involves the use of density functional theory (DFT) to compute the energy of the adsorption and transition states of the various elementary chemical processes occurring on a catalyst surface [83]. Macroscopic observables similar to those measured experimentally are then computed using transition state theory and a microkinetic model (MKM), taking the DFT computed adsorption and transition state energies as inputs [84]. However, the computation of adsorption and transition state energies using DFT can be both cost and time-intensive; hence, for most reaction systems of importance involving many intermediates or adsorbate species, the computation of adsorption energy is an arduous and expensive task [85]. DFT employs exchange correlation functional approximations to account for the complex many-body electron-electron interaction terms in the time independent Schrodinger equation (TISE) that describes the stationary state of a quantum mechanical system. Therefore, many DFT functionals exist with each differing in the level of theory and inherent approximations. Hence, DFT-derived adsorption energies and consequently, macroscopic observations depend on the choice of functionals. DFT functionals commonly used within the catalysis

community include PBE-D3, RPBE, BEEF-vdW, and SCAN+rVV10. Adsorption energy, which quantifies a molecule's binding strength to a catalyst's surface, is directly computable via DFT. While not directly indicative of catalysis speed, it provides insights into molecule stability on the catalyst, influencing potential reaction pathways. For an in-depth exploration of the fundamental principles of heterogeneous catalysis, readers are referred to the seminal work, 'Fundamental Concepts in Heterogeneous Catalysis' [83]. Nevertheless, the computation of adsorption energies for a large number of intermediates likely present and kinetically relevant in a chemical process can be computationally costly and even prohibitive due to the expensive nature of these calculations [86].

To address the high computational cost of calculating adsorption and transition-state energies for various active site models, linear scaling relations [86–88] have been developed for surface intermediates and transition states that use few computable descriptors to generate volcano curves on catalyst activity [89]. Nevertheless, the effectiveness of these relations for more complex chemistries remains uncertain. Additionally, the process of selecting descriptors for these calculations often involves a trial-and-error approach. In contrast, a more systematic approach was proposed in a previous study [90], which used Principal Component Analysis (PCA) [91,92] to identify the optimal minimal set of descriptors for the calculations, outperforming conventional descriptor selection methods. To overcome the computational challenge and to predict properties of the chemical entities using machine learning [93,94], the commonly employed approach often takes place in two major steps. Firstly, a suitable and effective descriptor is selected in the initial step, followed by the use of machine learning techniques to predict these chemical properties in the second step. Machine learning models can be trained using data obtained from DFT calculations and can subsequently be used to predict adsorption energies for a broad range of intermediates and catalysts [95,96].

In addition to the computational cost, predicting the adsorption energy of reaction intermediates and learning from multiple functionals can be a daunting task due to the

intricate nature of the interactions between the intermediates and the catalyst surface, as highlighted in previous literature [97, 98]. Machine learning models typically perform well when applied to the same or similar domains or functionals on which they were trained, but their performance can be severely compromised when extrapolated to different functionals [99]. The efficacy of the prediction results is largely dependent on the selection of the functionals used in the calculation and their inherent idiosyncrasies. The utilization of different functionals can result in varying predictions and models, making it challenging to determine the most accurate functional for a particular system. Moreover, the accuracy of the predictions can be further hampered by the quality of the training data, which is often obtained through experiments or DFT calculations. This limited set of training data, combined with the peculiarities of the functionals, results in a high level of uncertainty in the predictions, which poses a significant challenge to accurately predicting the adsorption energy of reaction intermediates from different functionals [100, 101]. Therefore, despite their capability to capture the complex interactions between intermediates and the surface, existing machine-learning strategies are hindered by the differences in functionals and their lack of generalization capability.

This study proposes a novel approach to address the limitations of current methods for predicting the adsorption energy of reaction intermediates across different functionals. The approach demonstrates that multiple functionals can benefit learning, rather than impede it, and effectively overcomes the difficulties associated with the unique characteristics of individual functionals. The proposed method involves capturing the relative energy differences between pairs of intermediates calculated within the same functional and training the model across all different functionals. This strategy results in a robust, reliable, and generalizable molecular representation across different functionals, representing a significant advancement in this direction.

To achieve this, the proposed implementation involves the extraction of molecular fingerprints and the training of Siamese neural networks on these fingerprints across dif-

ferent functionals, with the aim of learning invariant molecular representations (IMR) for catalysis. Molecular fingerprints provide numerical representations of molecules that encode their chemical properties and can serve as inputs for machine learning models. Several fingerprint generation schemes have been proposed in the literature, including the Coulomb matrix [102] and bag-of-bonds [103] approaches that use distance metrics based on the atomic coordinates of a species, along with atom-centered radial or angular symmetry functions [104–107]. Additionally, non-coordinate-based fingerprints have been developed that utilize molecular features derived from its chemical formula or SMILES notation [108–111]. SMILES, which stands for Simplified Molecular Input Line Entry System, is a compact and intuitive notation system for representing the molecular structure of a compound. SMILES notation can be used to generate fingerprints that capture the structural and chemical features of a molecule. Fingerprints can also be generated from the molecular graph structure, treating atoms and bonds as nodes and edges, respectively [4,112]. They may also be tailored to correspond to a specific property to be learned through back-propagation or other techniques. Notably, SMILES or graph-based fingerprints offer an advantage over coordinate-based descriptors because DFT or semiempirical methods are required only for the training data, unlike coordinate-based methods, which require reliable atomic coordinates even for species in the prediction set, potentially necessitating expensive calculations. Kernel-based models like kernel ridge regression [113], as well as neural network-based approaches such as recurrent neural networks [114], graph convolutional networks [4, 112], and 3D convolutional neural networks [115] have been widely used in this context. Some studies have also employed additive atomic contributions through atomic subnetworks [93, 104].

In this study, I have applied a novel approach by utilizing Siamese neural networks [116, 117], a type of neural network architecture well-suited for comparing pairs of input data and determining their similarity, to learn from the relative comparison of molecular pairs. The aim of the study is to generate molecular representations that capture inherent sim-

ilarities and dissimilarities between pairs of molecules, with the intention to enhance the predictive capability of adsorption energies for reaction intermediates by leveraging additional functionals. Most of the materials of this chapter have been published in a journal paper [29]. "Functionals" in the context of this study, refer to exchange-correlation functionals within DFT. The exchange-correlation functional approximation makes Kohn-Sham DFT a practical method for predicting the energy of a system. And when the study refers to "invariant molecular representations," it indicates representations that are robust to the variations introduced by these different DFT functionals, rather than the traditional invariance associated with rotation, translation, and exchange of atoms often seen in molecular descriptor engineering. This approach allows us to capture the underlying chemistry of the system in a manner that is insensitive to the choice of functional, while being informed by the specific system the model is trained on. To validate the approach, I have applied it to the prediction of adsorption energies for propane dehydrogenation [118] on a platinum catalyst surface and found it to be significantly superior and reliable in its predictive performance across different experimental settings. These results demonstrate the potential of the novel approach in aiding the design and optimization of catalysts for chemical reactions.

4.2 Methodology

The purpose of this section is to present a comprehensive overview of the proposed approach for predicting the adsorption energies of reaction intermediates on a catalyst surface. While the experimental case studies focused on the prediction of adsorption energies for propane dehydrogenation on a platinum catalyst surface using three different types of constant-size molecular fingerprints to generate invariant representations, it is important to note that the proposed method is not limited to these specific choices of data. This approach is more generalizable and can be applied to other types of molecular descriptors as well.

The section is structured as follows: First, I will provide an overview of the data collection and preparation process. Next, I will discuss the species descriptors or fingerprints I have employed, the structure of the proposed model, and the training strategies used to generate molecular representations. Finally, I will outline the process for predictive modeling of adsorption energies from the IMR or invariant molecular representations generated.

4.2.1 Dataset - Data Collection and Preparation

It is well-established that adsorption energies of molecular species can vary significantly depending on the metal surface being examined [83]. For the purpose of the experiments, I utilized data on propane dehydrogenation on a platinum surface model [119]. This work has been done in collaboration with researchers from the Department of Chemical Engineering at the University of South Carolina, Columbia. The calculations for the four DFT functionals, namely PBE-D3 [120, 121], RPBE [122], BEEF-vdW [101], and SCAN+rVV10 [123] were performed using the Vienna Ab initio Simulation Package (VASP) version 5.4.4 [124–126]. Additionally, data for training with random BEEF-vdW ensembles was generated using an ensemble of 2000 non-self-consistent field (NSCF) energies. The NSCF energies are computed using various possible exchange-correlation functionals (within the generalized gradient approximation) while using the RPBE electron densities that are self-consistent only for the RPBE functional. This is the conventional procedure in BEEF-vdW calculations. The data preparation step with training strategies employed in the proposed model will be further elaborated upon in later sections. The dataset consists of 46 intermediate species along with their SMILES notations. SMILES notation is a naming convention for chemical species, with a set of rules allowing for a unique representation of each chemical species. For instance, in SMILES nomenclature, 'C' represents a fully saturated and stable single-carbon molecule, i.e., CH_4 (methane). Unstable molecules or intermediates are denoted within square brackets, such as [CH_3],

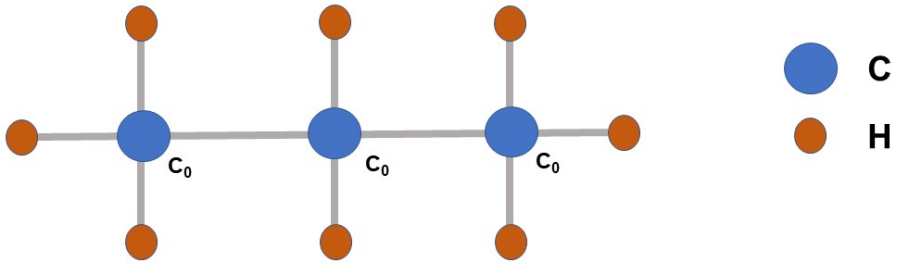
[CH_2], etc. Double and triple bonds are represented as '=' and '#', respectively, while branched species are indicated within parentheses '()'. The dataset also includes the adsorption energies for the Pt(111) metal surface calculated using all four DFT functionals and the 2000 BEEF-vdW ensembles. Additional details on the DFT calculations can be found in Appendix C.

4.2.2 Molecular Fingerprints

Molecular fingerprints can be broadly classified into two categories: 3D fingerprints (coordinate based) and topological/2D fingerprints (non-coordinate-based). 3D fingerprints suffer from the limitation that they require computationally expensive methods such as DFT or other semi-empirical techniques for their generation. In contrast, topological/2D fingerprints offer an alternative means for generating molecular fingerprints from SMILES notations that do not involve such computationally intensive processes. Notably, the molecular fingerprints employed in this study do not include any information regarding the catalyst since all species are adsorbed on the same Pt(111) catalyst surface. In this study, I have utilized three different non-coordinate-based techniques to obtain the fingerprints from SMILES notations of the molecular species. The following section will provide a brief discussion of the three techniques.

4.2.2.1 Flat Molecular Fingerprints (24-length)

Molecular fingerprints are often used to represent the molecular structure of a species. These fingerprints can be generated using various methods, including counting the different bond types present around each atom in a molecular species. In this study, I have employed constant-sized flat molecular fingerprints [127], similar to Ref. study [95], based on the SMILES notation of each species. Specifically, I have generated 24-length flat molecular fingerprints, which capture information about the number of different bond types present in the molecule. An example of the molecular fingerprints generated using this method is shown in Figure 4.1. These fingerprints provide more information than a



H	C	O	O_0	O_1	C_0	C_1	C_2	C_3	C-H	C-O_0	C-O_1
8	3	0	0	0	3	0	0	0	8	0	0

C=O	O-H	C_0-C_0	C_0-C_1	C_0-C_2	C_0-C_3	C_1-C_1	C_1-C_2	C_1-C_3	C_2-C_2	C_2-C_3	C_3-C_3
0	0	2	0	0	0	0	0	0	0	0	0

Figure 4.1: 24-length flat molecular fingerprints for the species $CH_3CH_2CH_3$. In this case, C_0 represents carbon atoms that are fully saturated (no free valence), while C_1, C_2, and C_3 represent carbon atoms with one, two, and three free valencies, respectively. This type of fingerprint contains information based on the number of saturated and unsaturated atoms and the number of bond counts between them.

basic bond count scheme, as they also take into account the number of free valencies and more granular bond count information related to the free valencies. This enables a more comprehensive understanding of the structural features of the molecule.

4.2.2.2 chEMBL Fingerprints (768-length)

In this study, I have adopted the pretrained chEMBL model [128] as an alternative method for generating molecular fingerprints. The chEMBL model is an approximation to the generative recurrent networks for de novo drug design in a prior study [129], primarily intended to capture the syntax of molecular structure in terms of SMILES strings. The resulting learned pattern probabilities can be used for de novo SMILES generation, making these pretrained models widely employed in chemogenomics and de novo drug design. The chEMBL model employed in this study is a masked language model (MLM). An MLM

is basically a neural network trained to predict missing words in a text, enabling it to learn the underlying structure and relationships between the words in the text. The chEMBL model was trained from scratch using 438,552 SMILES notations, and generates 768-length molecular fingerprints for each species, yielding a rich representation of the structural features of the molecule. These fingerprints have a broad range of applications, including the prediction of chemical properties and the design of novel compounds.

4.2.2.3 Morgan Fingerprints (24-length)

Building on the exploration of molecular fingerprints, next I have incorporated the Morgan fingerprints [130] for their unique approach to capturing molecular features. Circular by design, Morgan fingerprints focus on the molecular environment of each atom within a specified radius. By iterating over concentric bonds around each atom, this methodology generates a descriptor that not only identifies the type of atom but also its distinct local environment. In this study, I converted the SMILES notation of each species into 24-length Morgan fingerprints by setting a radius of 2. This ensures the fingerprint captures the molecular environment up to two bonds out from each atom (neighbors and neighbors of neighbors). This choice in the descriptor, combined with the other fingerprint methods, enhances the study's depth, further enriching the comprehension of molecular structures in the dataset.

4.2.3 Molecular Representations

Here in the current study, I have examined three distinct methods for generating molecular representations from the fingerprints: the raw or the original fingerprints (Original), PCA-based (PCA), and Siamese-based (IMR). The Original method involves using the raw fingerprints obtained from SMILES notations without any additional transformation. The second method employs principal component analysis (PCA) [91, 92] to transform the raw fingerprints into lower dimensional molecular representations. This approach is motivated by previous research [90] that found PCA to be effective in obtaining descrip-

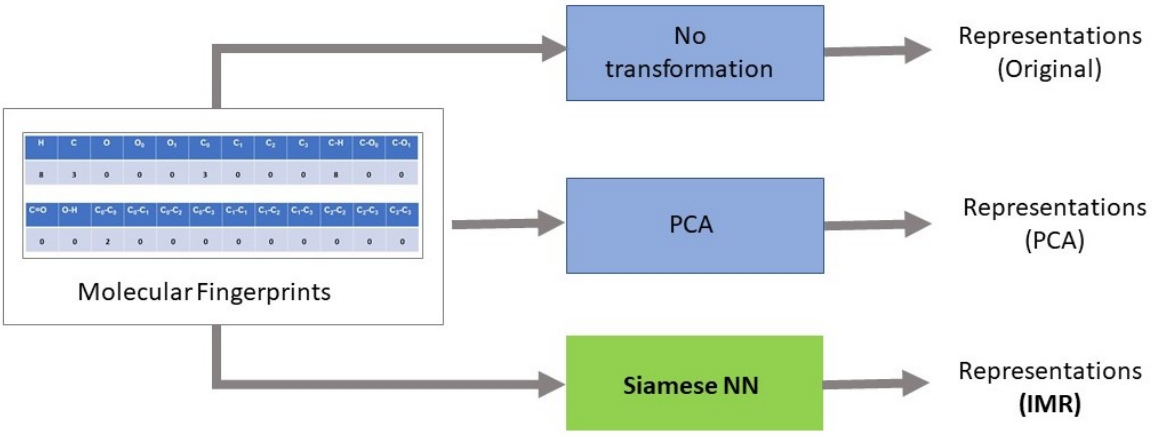


Figure 4.2: Three different methods were adopted to generate molecular representations from molecular fingerprints. Top: (Original) No transformation or modification is done on the original or raw fingerprints, Middle: (PCA) Molecular representations generated based on principal component analysis, and Bottom: (IMR) Molecular representations generated using the trained Siamese neural network model.

tors over other conventional descriptors. For the PCA-based molecular representations, I have used the representation size that explains approximately 98% of the variance for 24-length flat molecular and Morgan fingerprints and 90% of the variance for chEMBL model-based fingerprints. The third method referred to as IMR, utilizes a trained Siamese neural network to generate invariant molecular representations from the raw fingerprints. This allows for a comparison of the performance of different representation generation methods and the identification of those that are most effective for predictive modeling tasks. The process of generating these molecular representations from fingerprints is illustrated in Figure 4.2. Overall, the current study enables the evaluation of the relative effectiveness of various approaches for generating molecular representations, with potential applications in predictive modeling tasks. It is worth noting that the representations obtained in this study are essentially transformed fingerprints, which either aim to reduce the dimensionality of the fingerprints (PCA) or incorporate knowledge about the catalyst (IMR).

4.2.4 Structure of the Proposed Model

The fundamental aim of this study is to develop robust molecular representations for reaction intermediates that are generalizable across multiple functionals. The Siamese neural network [117] is selected for its ability to compare input data pairs and identify their similarities and differences. The hypothesis is that training the network using all functionals in the training set for a given surface will enable it to generate invariant and informative representations of molecular species from raw fingerprints. These representations should be free from information specific to individual functionals but aware of the surface system, allowing the resulting model to be more generalizable.

The Siamese neural network comprises two identical sub-networks with identical weights and architecture, which are used to process and analyze the molecular fingerprints for each pair of molecular species across all different functionals in the training set. These sub-networks generate molecular representations that are then fed into a feedforward neural network. These sub-networks are intended to identify informative information from the raw fingerprints and provide an intermediate representation that the feedforward network can use to learn the relative energy differences. The overall Siamese network is trained to predict the relative difference in adsorption energies of the pair using Mean Absolute Errors (MAEs) as the cost function.

Nonlinearity is incorporated through nonlinear activation functions [131–134] used in the hidden layers of the Siamese networks. The sub-networks are randomly initialized [135, 136] and their weights are updated during training to minimize the MAEs. Figure 4.3(a) illustrates the overall training process of the proposed network. To ensure a fair comparison, I have used the same representation size for the Siamese network to generate representations (IMR) as the number of components used for the corresponding PCA-based molecular representations (PCA).

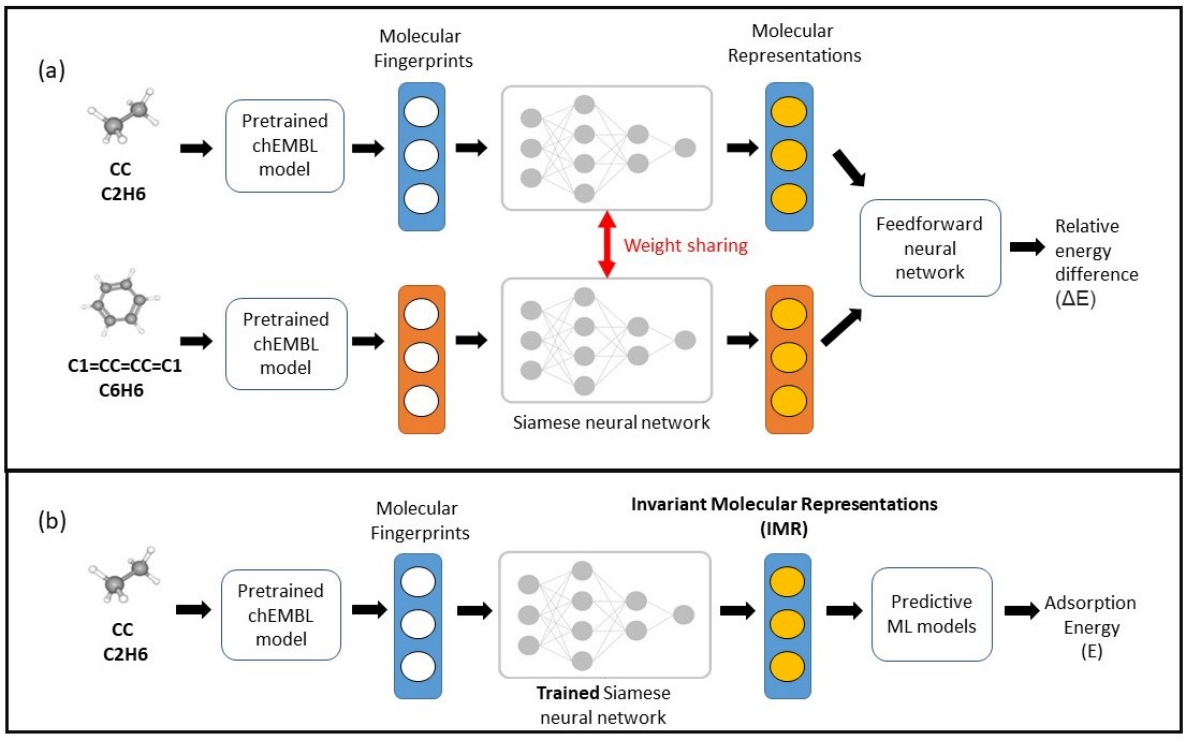


Figure 4.3: Two major steps of the proposed method/pipeline: (a) training Siamese neural network to generate invariant molecular representations (IMR) across different functionals using relative energy difference between species, and (b) predictive modeling of adsorption energies using IMR generated by the Siamese model trained in step (a).

4.2.5 Training Strategies using the Proposed Model

The training of the Siamese neural network in this study involved three distinct strategies. The first two strategies, namely (i) four functional model (FFM) training, and (ii) BEEF-vdW ensemble model (BEM) training, made use of additional functional information to generate more meaningful and concise molecular representations for the predictive modeling tasks. The third training strategy, referred to as functional specific model (FSM) training, served as a validation check for the results obtained. Further details on these training strategies will be presented in the corresponding experimental result sections.

4.2.6 Predictive Modeling with the Proposed Model

In this study, I propose a predictive modeling mechanism for adsorption energies, illustrated in Figure 4.3(b). The first step involves extracting molecular fingerprints based on atomic bond types and counts (flat fingerprints) or utilizing the pretrained chEMBL model (chEMBL fingerprints) or using the Morgan fingerprints. Subsequently, the Siamese neural networks are trained using the training strategies and steps outlined in the previous section, as shown in Figure 4.3(a). The resulting trained sub-network generates IMR for the training samples of molecular species. Finally, I utilize various machine learning algorithms, including ridge, elastic net, kernel ridge, and support vector regression, to train on the IMR generated by the trained sub-network and conduct predictive analysis of the adsorption energies of corresponding molecular samples.

The motivation behind utilizing predictive modeling algorithms on top of the IMR, or in other words, the representations obtained from the trained Siamese network, can be explained as follows. Firstly, with the Siamese network, this work aims to learn molecular representations that are invariant across different functionals. I anticipate that these representations would capture the relationship of the fingerprints to the adsorption energies without relying on the functional but rather considering only the surface system and molecular structures. However, to evaluate the informativeness of IMR toward learning the adsorption energies, I have employed functional-dependent predictive models as the second step. In this study, I have used four machine learning algorithms, namely ridge, elastic net, kernel ridge, and support vector regression, to generate functional-dependent predictive models.

In summary, the input for the predictive ML models was the IMR or the invariant molecular representations of the species generated by the proposed model and training strategy from the raw fingerprints, and the outputs were the actual adsorption energies of those species. Later on, I evaluate the performance of the IMR in comparison to raw or

the original fingerprints (Original) and PCA-based representations (PCA) on predictive analysis.

4.3 Results and Discussion

In the following sections, I have provided a detailed description of the simulation process and results for all the case studies. The observations on the results are also presented.

4.3.1 Simulation

Given the inherent challenges posed by the dataset size of 46 molecules, this study meticulously adopted a robust approach. For each of the experimental case studies, I have conducted 10 random trials, with each trial utilizing a different randomly chosen train and test set at a 2:1 ratio (i.e., approximately 67% of the data for training and 33% for testing), ensuring enhanced variability and consistency in the results.

For each trial, the training data was further subjected to 5-fold cross-validation, where the training set was divided into five subsets. Four of these subsets were employed to train the ML models for predicting adsorption energies, while the fifth subset was designated as the validation set. This process was iterated five times, with each subset taking its turn as the validation set. Model parameters were optimized based on the performance across these validation sets.

I have evaluated the three different molecular representations in predicting the adsorption energies of various species based on the mean and standard deviation of the mean absolute errors (MAEs). To reinforce the reliability of the observations, I have employed a t-test with $\alpha = 0.05$ as the significance level. Furthermore, to safeguard against potential biases and to encompass a broad spectrum of analysis, for each experimental case, this study employed four distinct machine-learning algorithms: ridge regression (ridge), elastic net regression (elastic), support vector regression (svr), and kernel ridge regression (krr).

4.3.2 Four Functional Model (FFM) Training

I first employed a training and testing strategy utilizing data from four distinct DFT functionals: PBE-D3, BEEF-vdW, RPBE, and SCAN+rVV10. I have designed four experimental scenarios, each targeting predictive performance for one of these functionals. For each case, the data from the test functional was reserved, and the Siamese network was trained using molecular species pairs from the remaining three functionals to predict their "relative energy difference". Once trained, this network effectively translates molecular fingerprints into representations.

To assess these representations' predictive capability for adsorption on the test functional, I first randomly partitioned its data into training and test sets for each trial. I then trained various ML models on the training data by using the representations derived from the trained Siamese network. Subsequently, the performance of these ML models was evaluated on the held-out test samples from the test functional. The resulting Mean Absolute Errors (MAEs) indicate the discrepancies between the predicted adsorption energies and the DFT-calculated ones.

To evaluate the method's generalization capacity, I have tested it using four ML models: ridge, elastic, krr, and svr. Each model was trained on three different representations (Original, PCA, and IMR). The Originals are unmodified fingerprints; PCA representations come from principal component analysis applied to these fingerprints, while IMR uses the Siamese network-trained representations.

By evaluating IMR's performance on data excluded from the corresponding functional during Siamese model training, this study assesses its capability to adapt to novel DFT functionals, showcasing the proposed approach's broad applicability.

4.3.2.1 Representations Generated from Flat Fingerprints

I initially examined the experimental results using the FFM training strategy and representations generated from flat molecular fingerprints, which are based on atomic bond

Table 4.1: Evaluation of three molecular representations (Original, PCA, IMR) using 24-length flat molecular fingerprints and FFM strategy. Displayed values are Mean Absolute Errors (MAEs) between predicted and DFT-calculated energies in electron volts (eV). Lower MAEs signify better performance. Bold values are statistically significant based on the t-test.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.31 ± 0.04	0.32 ± 0.04	**0.26 ± 0.05**
	elastic	0.33 ± 0.04	0.34 ± 0.05	**0.27 ± 0.06**
	krr	0.35 ± 0.05	0.33 ± 0.06	0.29 ± 0.03
	svr	0.31 ± 0.06	0.30 ± 0.06	0.29 ± 0.04
BEEF-vdW	ridge	0.31 ± 0.05	0.31 ± 0.05	**0.16 ± 0.03**
	elastic	0.32 ± 0.05	0.33 ± 0.04	**0.16 ± 0.02**
	krr	0.33 ± 0.04	0.34 ± 0.05	**0.15 ± 0.04**
	svr	0.31 ± 0.05	0.31 ± 0.06	**0.14 ± 0.04**
RPBE	ridge	0.31 ± 0.05	0.31 ± 0.05	**0.22 ± 0.06**
	elastic	0.33 ± 0.05	0.33 ± 0.05	**0.20 ± 0.05**
	krr	0.35 ± 0.05	0.36 ± 0.04	**0.22 ± 0.05**
	svr	0.32 ± 0.07	0.34 ± 0.07	**0.20 ± 0.04**
SCAN+rVV10	ridge	0.37 ± 0.05	0.38 ± 0.04	**0.25 ± 0.06**
	elastic	0.40 ± 0.04	0.39 ± 0.04	**0.24 ± 0.03**
	krr	0.42 ± 0.08	0.42 ± 0.07	**0.23 ± 0.05**
	svr	0.38 ± 0.08	0.39 ± 0.10	**0.24 ± 0.06**

types and counts in each molecular species (as mentioned in detail in the earlier section). Table 4.1 presents the corresponding outcomes, which display the test functional used in each experimental case (Test Functional), the machine learning algorithm employed for predictions of adsorption energies (ML Alg.), and the type of input representations used, namely Original, PCA, and IMR. The empirical findings indicate statistically significant improvements in the use of molecular representations learned by the Siamese network (IMR) compared to PCA-based representations (PCA) for three out of four functionals (BEEF-vdW, RPBE, and SCAN+rVV10). For PBE-D3, I also observed improvements when using IMR, although this was statistically significant in the case of the 'ridge' and 'elastic net' as the predictive algorithm.

Table 4.2: Evaluation of three molecular representations (Original, PCA, IMR) using 768-length fingerprints from the chEMBL model and FFM strategy. The values presented are Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, measured in electron volts (eV). Smaller MAEs signify better performance. Bold values are statistically significant via t-test.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.39 ± 0.06	0.37 ± 0.05	**0.24 ± 0.06**
	elastic	0.32 ± 0.07	0.32 ± 0.04	**0.23 ± 0.04**
	krr	0.27 ± 0.07	0.28 ± 0.05	**0.19 ± 0.07**
	svr	0.28 ± 0.06	0.28 ± 0.06	**0.18 ± 0.06**
BEEF-vdW	ridge	0.42 ± 0.07	0.36 ± 0.05	**0.14 ± 0.04**
	elastic	0.34 ± 0.05	0.33 ± 0.04	**0.14 ± 0.05**
	krr	0.34 ± 0.06	0.32 ± 0.05	**0.15 ± 0.07**
	svr	0.32 ± 0.05	0.34 ± 0.04	**0.14 ± 0.05**
RPBE	ridge	0.46 ± 0.08	0.39 ± 0.05	**0.18 ± 0.05**
	elastic	0.37 ± 0.05	0.38 ± 0.04	**0.19 ± 0.06**
	krr	0.44 ± 0.07	0.37 ± 0.04	**0.22 ± 0.06**
	svr	0.38 ± 0.05	0.39 ± 0.04	**0.20 ± 0.05**
SCAN+rVV10	ridge	0.44 ± 0.06	0.39 ± 0.05	**0.23 ± 0.06**
	elastic	0.39 ± 0.05	0.38 ± 0.03	**0.23 ± 0.06**
	krr	0.37 ± 0.04	0.39 ± 0.03	**0.22 ± 0.03**
	svr	0.35 ± 0.05	0.40 ± 0.05	**0.19 ± 0.03**

4.3.2.2 Representations Generated from Transfer Learning

Subsequently, I investigated the performance of the FFM training approach, but this time, with representations generated by using fingerprints from the chEMBL model. The results of the experimental cases are presented in Table 4.2.

Notably, using representations generated from chEMBL fingerprints also demonstrated a clear trend of statistically significant improvements when adopting IMR compared to PCA representations for all four DFT functionals. However, if we consider the performance of the Original representations, the errors varied more widely among the different predictive machine learning algorithms (ridge, elastic, KRR, SVR) compared to those obtained using the flat fingerprints. This indicates that in some instances, the machine learning models struggled to fit properly into the Original representations, as the fea-

Table 4.3: Performance evaluation of three molecular representations (Original, PCA, IMR) using 24-length Morgan fingerprints and FFM training strategy. The values presented are Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, in electron volts (eV). A smaller MAE value indicates better performance. Values highlighted in bold are determined to be statistically significant via the t-test.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.34 ± 0.05	0.33 ± 0.05	0.34 ± 0.11
	elastic	0.32 ± 0.05	0.31 ± 0.05	**0.25 ± 0.04**
	krr	0.31 ± 0.05	0.33 ± 0.07	0.28 ± 0.05
	svr	0.31 ± 0.05	0.31 ± 0.05	0.28 ± 0.06
BEEF-vdW	ridge	0.34 ± 0.04	0.34 ± 0.04	**0.11 ± 0.04**
	elastic	0.33 ± 0.03	0.34 ± 0.04	**0.11 ± 0.04**
	krr	0.33 ± 0.04	0.34 ± 0.04	**0.10 ± 0.03**
	svr	0.33 ± 0.05	0.32 ± 0.05	**0.10 ± 0.03**
RPBE	ridge	0.37 ± 0.06	0.37 ± 0.06	**0.18 ± 0.03**
	elastic	0.37 ± 0.05	0.38 ± 0.05	**0.17 ± 0.03**
	krr	0.38 ± 0.05	0.39 ± 0.05	**0.17 ± 0.04**
	svr	0.38 ± 0.06	0.38 ± 0.06	**0.17 ± 0.04**
SCAN+rVV10	ridge	0.39 ± 0.05	0.39 ± 0.05	**0.22 ± 0.05**
	elastic	0.40 ± 0.03	0.40 ± 0.04	**0.20 ± 0.06**
	krr	0.39 ± 0.04	0.39 ± 0.03	**0.18 ± 0.03**
	svr	0.39 ± 0.05	0.39 ± 0.05	**0.19 ± 0.04**

ture dimension ($d = 768$) in these Original representations was much larger ($d >> n$) compared to the number of samples/species ($n = 46$).

4.3.2.3 Representations Generated from Morgan Fingerprints

Moving on to the results derived from Morgan's fingerprints, I have applied the same FFM training strategy. The details of this analysis can be found in Table 4.3. Similar to the previous findings, the molecular representations processed by the Siamese network (IMR) stood out, demonstrating superior performance over the PCA-based representations, particularly for the BEEF-vdW, RPBE, and SCAN+rVV10 functionals. As for the PBE-D3, while there was an evident improvement with the use of IMR, it was statistically significant only in the case when I employed the 'elastic net' as the regression algorithm.

4.3.3 BEEF-vdW Ensemble Model (BEM) Training

Next, I have adopted a similar approach to the FFM training, with a few key differences. To train the predictive machine learning models namely ridge, elastic, krr, and svr, I have utilized data from one of the four DFT functionals in each case focusing on learning the predictive performance on that specific functional. However, instead of using the remaining three DFT functionals to train the Siamese network (as we have seen in FFM training), I used the BEEF-vdW ensembles, and the Siamese network was trained using molecular species pairs from these BEEF-vdW ensembles to predict their "relative energy difference".

The BEEF-vdW functional produces an ensemble of 2000 NSCF energies for each species. In every trial, I randomly select an ensemble of 50 BEEF-vdW functional energies from the available 2000. This allows us to treat each BEEF-vdW ensemble functional as distinct, leveraging them as different functionals—a novel approach in physical chemistry.

Upon training, the Siamese network adeptly converts the molecular fingerprints into representations. To assess the predictive power of these representations for adsorption on a test functional, I partitioned its data into training and test sets for each trial. Different ML models are trained on the training set by using the representations generated from the trained Siamese network. These ML models' performance is then evaluated on the held-out test samples of the test functional, with the resulting Mean Absolute Errors (MAEs) denoting the difference between predicted and DFT-calculated adsorption energies.

The primary objective in training the Siamese network on these ensemble functional energies is to learn molecular representations that capture the inherent similarities and variances between pairs of molecules across different BEEF-vdW ensemble energies. The effectiveness of these learned representations is validated by using them to train ML models that predict the adsorption energies on distinct DFT functionals.

Table 4.4: Evaluation of three molecular representations (Original, PCA, IMR) generated using 24-length flat molecular fingerprints with the BEM training strategy. Presented values are Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, in electron volts (eV). A smaller MAE value indicates better performance. Values that are statistically significant based on the t-test are represented in bold.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.31 ± 0.04	0.32 ± 0.04	**0.25 ± 0.04**
	elastic	0.33 ± 0.04	0.34 ± 0.05	**0.26 ± 0.05**
	krr	0.35 ± 0.05	0.33 ± 0.06	**0.23 ± 0.05**
	svr	0.31 ± 0.06	0.30 ± 0.06	**0.24 ± 0.04**
BEEF-vdW	ridge	0.31 ± 0.05	0.31 ± 0.05	**0.19 ± 0.03**
	elastic	0.32 ± 0.05	0.33 ± 0.04	**0.19 ± 0.03**
	krr	0.33 ± 0.04	0.35 ± 0.04	**0.13 ± 0.04**
	svr	0.31 ± 0.05	0.31 ± 0.06	**0.14 ± 0.05**
RPBE	ridge	0.31 ± 0.05	0.31 ± 0.05	**0.25 ± 0.07**
	elastic	0.33 ± 0.05	0.33 ± 0.04	**0.23 ± 0.06**
	krr	0.35 ± 0.05	0.36 ± 0.04	**0.19 ± 0.04**
	svr	0.32 ± 0.07	0.34 ± 0.07	**0.21 ± 0.07**
SCAN+rVV10	ridge	0.37 ± 0.05	0.38 ± 0.04	**0.24 ± 0.05**
	elastic	0.40 ± 0.04	0.39 ± 0.05	**0.23 ± 0.05**
	krr	0.42 ± 0.08	0.41 ± 0.07	**0.20 ± 0.04**
	svr	0.38 ± 0.08	0.39 ± 0.10	**0.22 ± 0.04**

4.3.3.1 Representations Generated from Flat Fingerprints

To assess the performance of the approach in the context of BEEF-vdW Ensemble Model (BEM) training, I first evaluated the performance of the model using flat molecular fingerprints. Table 4.4 presents the corresponding outcomes for this experiment. In this table, I present the test functional used in the experimental case (Test Functional), the machine learning algorithm utilized for predicting adsorption energies (ML Alg.), and the type of representations (Original, PCA, and IMR) that were used as input for the predictive models.

As we can see, the molecular representations generated by the Siamese network (IMR) exhibit noteworthy improvements for all testing functionals and across all predictive machine learning algorithms, as evidenced by the results. However, when the analysis was

Table 4.5: Performance evaluation of three molecular representations (Original, PCA, and IMR) using 768-length fingerprints derived from the chEMBL model, combined with the BEM training strategy. The values are given in terms of Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, expressed in electron volts (eV). A lower MAE value signifies superior performance. Bolded values indicate statistical significance as determined by the t-test.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.39 ± 0.06	0.35 ± 0.05	**0.26 ± 0.04**
	elastic	0.32 ± 0.07	0.32 ± 0.04	**0.22 ± 0.04**
	krr	0.27 ± 0.07	0.29 ± 0.04	**0.22 ± 0.05**
	svr	0.28 ± 0.06	0.29 ± 0.06	**0.20 ± 0.06**
BEEF-vdW	ridge	0.42 ± 0.07	0.37 ± 0.06	**0.13 ± 0.05**
	elastic	0.34 ± 0.05	0.33 ± 0.04	**0.13 ± 0.05**
	krr	0.34 ± 0.06	0.33 ± 0.04	**0.11 ± 0.05**
	svr	0.32 ± 0.05	0.34 ± 0.05	**0.15 ± 0.07**
RPBE	ridge	0.46 ± 0.08	0.38 ± 0.03	**0.17 ± 0.04**
	elastic	0.37 ± 0.05	0.38 ± 0.04	**0.17 ± 0.03**
	krr	0.44 ± 0.07	0.37 ± 0.04	**0.18 ± 0.04**
	svr	0.38 ± 0.05	0.39 ± 0.04	**0.16 ± 0.06**
SCAN+rVV10	ridge	0.44 ± 0.06	0.38 ± 0.04	**0.25 ± 0.10**
	elastic	0.39 ± 0.05	0.39 ± 0.03	**0.23 ± 0.07**
	krr	0.37 ± 0.04	0.39 ± 0.03	**0.18 ± 0.05**
	svr	0.35 ± 0.05	0.39 ± 0.05	**0.20 ± 0.04**

compared to the findings presented in the section on FFM training with 24-length flat fingerprints, I found no statistically significant difference in the IMR performance between the FFM and BEM training strategies for the majority of cases; specifically, this was true in 12 out of 16 cases.

4.3.3.2 Representations Generated from Transfer Learning

Next, in Table 4.5, I present the empirical results of the case study in which I used a similar training strategy (BEM training), but with representations generated by fingerprints from pretrained chEMBL model. Consistent with the earlier experimental findings, the current results confirm the superiority of IMR over PCA-based representations across all experimental cases. Notably, IMR showed exceptional performance specific to the BEEF-vdW functional, compared to the other three DFT functionals. These phenomena

Table 4.6: Performance evaluation of three molecular representations (Original, PCA, and IMR) using 24-length Morgan fingerprints with the BEM training strategy. The values are presented in terms of Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, expressed in electron volts (eV). A lower MAE value denotes superior performance. Values highlighted in bold are determined to be statistically significant via the t-test.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.34 ± 0.05	0.33 ± 0.04	0.31 ± 0.07
	elastic	0.32 ± 0.05	0.31 ± 0.04	**0.22 ± 0.04**
	krr	0.31 ± 0.05	0.34 ± 0.06	**0.27 ± 0.06**
	svr	0.31 ± 0.05	0.32 ± 0.06	**0.25 ± 0.05**
BEEF-vdW	ridge	0.34 ± 0.04	0.34 ± 0.05	**0.11 ± 0.02**
	elastic	0.33 ± 0.03	0.33 ± 0.04	**0.12 ± 0.03**
	krr	0.33 ± 0.04	0.34 ± 0.04	**0.10 ± 0.05**
	svr	0.33 ± 0.05	0.33 ± 0.05	**0.10 ± 0.07**
RPBE	ridge	0.37 ± 0.06	0.37 ± 0.07	**0.18 ± 0.09**
	elastic	0.37 ± 0.05	0.39 ± 0.05	**0.19 ± 0.13**
	krr	0.38 ± 0.05	0.38 ± 0.05	**0.18 ± 0.06**
	svr	0.38 ± 0.06	0.38 ± 0.06	**0.18 ± 0.05**
SCAN+rVV10	ridge	0.39 ± 0.05	0.39 ± 0.05	**0.23 ± 0.06**
	elastic	0.40 ± 0.03	0.39 ± 0.04	**0.22 ± 0.05**
	krr	0.39 ± 0.04	0.39 ± 0.03	**0.19 ± 0.03**
	svr	0.39 ± 0.05	0.39 ± 0.05	**0.22 ± 0.05**

can be attributed to the training strategy of the Siamese model, which is based on the random BEEF-vdW ensembles.

4.3.3.3 Representations Generated from Morgan Fingerprints

Diving into the results from the Morgan fingerprints, as presented in Table 4.6, the application of the BEM training strategy again demonstrated patterns in line with the prior observations. Much like before, the IMR consistently outperformed the PCA-based representations across all the experimental cases. Specifically, when using the BEEF-vdW functional for testing, the performance spike was evident. As mentioned in the earlier section, this enhanced performance particularly in the BEEF-vdW test functional can likely be traced back to the Siamese model's training approach, which heavily leverages the random BEEF-vdW ensembles.

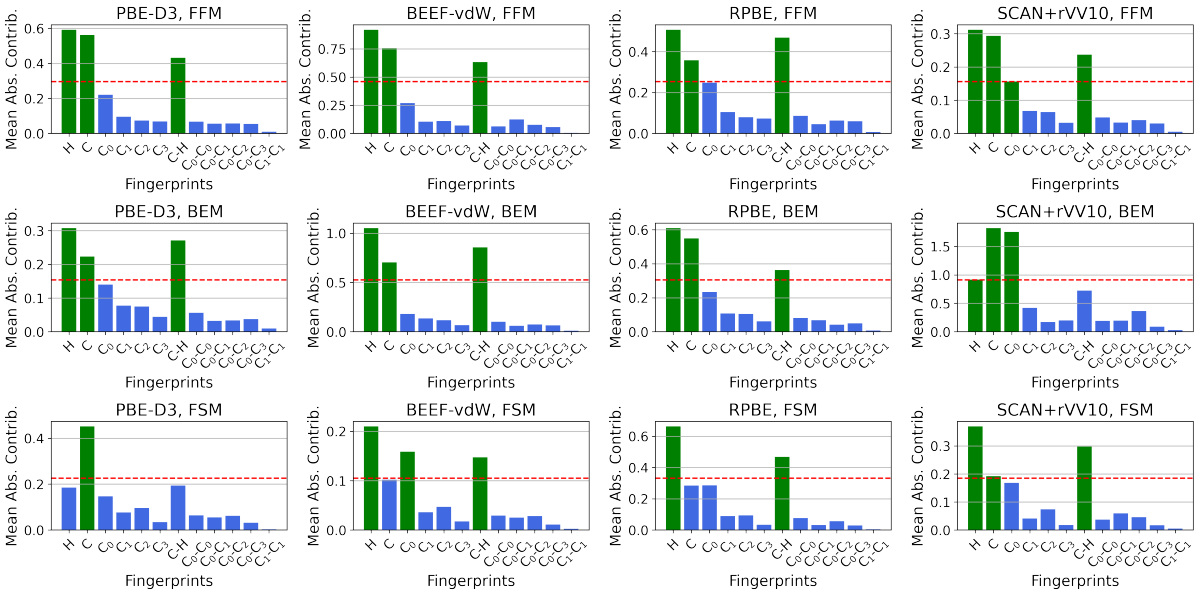


Figure 4.4: Feature contribution analysis across different training strategies and DFT functionals. This figure illustrates the mean absolute contribution of various molecular fingerprints (e.g., H, C, C_0) in a matrix format, where rows and columns represent training strategies for the Siamese network and DFT functionals, respectively. The 1st, 2nd, and 3rd rows represent the FFM, BEM, and FSM training strategies, while the 1st to 4th columns correspond to PBE-D3, BEEF-vdW, RPBE, and SCAN+rVV10 functionals, respectively. A dotted red line in each plot marks a threshold set at 50% of the maximum contribution value for that specific scenario, delineating the top contributing fingerprints. Fingerprints with negligible contributions were omitted for clarity. This analysis underscores the significant fingerprints contributing to adsorption energies across various training strategies and functionals.

4.3.4 Sanity Check

As a validation of the proposed approach, I have conducted four experimental case studies using the four DFT functionals (similar to FFM and BEM) but these case studies are designed to simulate the scenario of model training without additional functionals, and the results are presented in detail in Appendix B, which includes information on the experimental setup and the corresponding findings.

4.3.5 Fingerprint Contribution Analysis

In this section, I present a comprehensive feature contribution analysis to examine the impact of different molecular fingerprints on adsorption energy predictions. The analysis includes all three training strategies for the Siamese network namely- the Four Functional Model (FFM), the BEEF-vdW Ensemble Model (BEM), and the Functional Specific Model (FSM), across all four distinct DFT functionals: PBE-D3, BEEF-vdW, RPBE, and SCAN+rVV10. Figure 4.4 illustrates the mean absolute contribution of the fingerprints (such as H, C, and C_0) for each training strategy and functional. The attribution values shown here were determined using integrated gradients, a method implemented in the Python-based Captum library [137]. Notably, for functionals like PBE-D3, BEEF-vdW, and RPBE, both FFM and BEM consistently identify the same top contributing fingerprints (e.g., number of hydrogen atoms, number of carbon atoms, and carbon-hydrogen bonds), indicating a shared understanding of key features. This agreement highlights the robustness of the proposed training strategy to learn IMR that captures essential characteristics from the original fingerprints, leading to superior predictive performance. However, the FSM strategy, which lacks the benefit of leveraging additional functionals, shows divergent feature contributions, underlining the effectiveness of FFM and BEM in exploiting functional invariances. Interestingly, when testing on the SCAN+rVV10 functional, we can see weaker agreement between FFM and BEM. This can be attributed to the distinct characteristics of the SCAN+rVV10 as a meta-GGA functional, in contrast to the other three GGA (Generalized Gradient Approximation) functionals (such as PBE-D3, BEEF-vdW, and RPBE) used in obtaining the invariant molecular representations, highlighting the importance of using diverse functionals in generating the invariant and robust representations. Notably, this aligns with findings by a prior study [138], that demonstrated through Mahalanobis distance analysis that the SCAN+rVV10 is comparatively less accurately captured by methods typically effective for GGA functionals, underscoring

its unique functional properties. The fingerprint contribution analysis emphasizes the capability of the proposed strategies to extract meaningful essence and insights from original molecular fingerprints, thereby generating superior representations that ultimately enhance the predictive accuracy of adsorption energies. A species-based breakdown of fingerprint contributions is provided in Appendix D for one of the experimental case scenarios.

4.3.6 Goodness-of-Fit Analysis (using D^2-score)

Finally, I conducted a comprehensive goodness-of-fit analysis using the D^2-score, implemented in Python's Scikit-learn [71] library. The D^2-score is a metric that assesses how well a model explains variance compared to a null model, where the null model is based on using the median from the training samples as the predictions for the test samples. Additional details on how this score is defined can be found in Appendix E. This approach was utilized to evaluate the performance of the models across all three training strategies, namely FFM, BEM, and FSM, along with the molecular representations used in the study- Original, PCA, and IMR. The findings reveal that models based on Original and PCA representations were no better than the null model, a behavior observed irrespective of molecular fingerprints used, such as 24-length flat molecular fingerprints, 768-length chEMBL fingerprints, and 24-length Morgan fingerprints, for generating molecular representations. This phenomenon highlights the challenges posed by the dataset size and complexity of the molecular interactions considered in this study. In contrast, models trained with the IMR consistently led to significantly better-fitted models, in the case of FFM and BEM training strategies, underscoring the benefit of leveraging additional functionals to learn robust representations. Furthermore, aligning with the prior findings, this study observed that models especially demonstrate a better fit to the BEEF-vdW functional when tested, on both the FFM and BEM training strategies. Detailed results, including the D^2-scores for each case study, are available in Appendix E for further

reference.

4.4 Learning IMRs with Multiple Surface Systems

In this section, I proposed an extension of the Invariant Molecular Representation (IMR) learning method to larger datasets and multiple surface systems, addressing the limitations of the previous method [29] which used only one surface system - Pt (111) and 46 reaction samples.

4.4.1 Datasets

For this study, I have utilized the datasets from large repositories such as the Catalysis Hub [139] and the Open Catalyst Project [140] to demonstrate the robustness and scalability of the proposed extension.

Catalysis Hub: I have selected the "High-Throughput Calculations of Catalytic Properties of Bimetallic Alloy Surfaces" [141] from Catalysis Hub as one of the datasets for the current study. The dataset includes chemisorption properties for key adsorbates across 2,035 bimetallic alloy surfaces, each in distinct stoichiometric ratios. This extensive collection covers a wide and varied chemical space of importance for catalytic applications, enabling the creation of machine learning-driven predictive models to accelerate theoretical catalysis research and discovery. Notably, the dataset uses the BEEF-vdW DFT functional for calculating the adsorption energies, which is particularly suited for capturing van der Waals interactions.

Open Catalyst 2020 (OC20): The OC20 dataset [98] is another comprehensive resource that provides adsorption energies for various reaction intermediates on different catalyst surfaces. In this study, I have employed this dataset, particularly for the Initial Structure to Relaxed Energy (IS2RE) task data, to further validate the efficacy of the extension of the IMR learning method in diverse catalytic environments. The dataset consists of 1,281,040 Density Functional Theory (DFT) relaxations and includes characteristics related to bulk, adsorbate, and reaction site properties. The IS2RE task involves

predicting the energy difference between an initial and a relaxed structure, which is crucial for understanding the stability and reactivity of the adsorbates on catalyst surfaces.

4.4.2 Limitations of Previous Study

This section identifies a few limitations of the initial work presented previously when learning invariant molecular representations with the datasets mentioned above:

1. **Single Surface System**: The original work on the IMR learning method [29] focused on a single surface system - Pt (111). Therefore, it required generating descriptors only for the adsorbate since all the reaction samples used the same surface system. This approach is insufficient for reactions involving multiple surface systems. With multiple surface systems, it is crucial to generate descriptors for both the adsorbate and the surface to capture the interaction dynamics accurately.

2. **Large Number of Reaction Samples**: The previous study [29] used a Siamese neural network, generating pairwise data by combining each pair of molecules and training the model based on the relative energy differences for those pairs across all environments or functionals. However, with more reaction samples in larger datasets (e.g., OC20 data), generating pairwise samples would require training on billions of samples, which is computationally prohibitive.

4.4.3 Methodology

I have gathered adsorption energy data from the selected datasets, ensuring a wide range of surface systems and adsorbates. To address the limitations, the proposed approach is implemented with the following steps:

1. **Descriptor Generation**: I have generated descriptors for both the adsorbates and the surface systems as both contain invaluable information required to define the characteristics of adsorption reactions properly. For the adsorbate, I have adopted

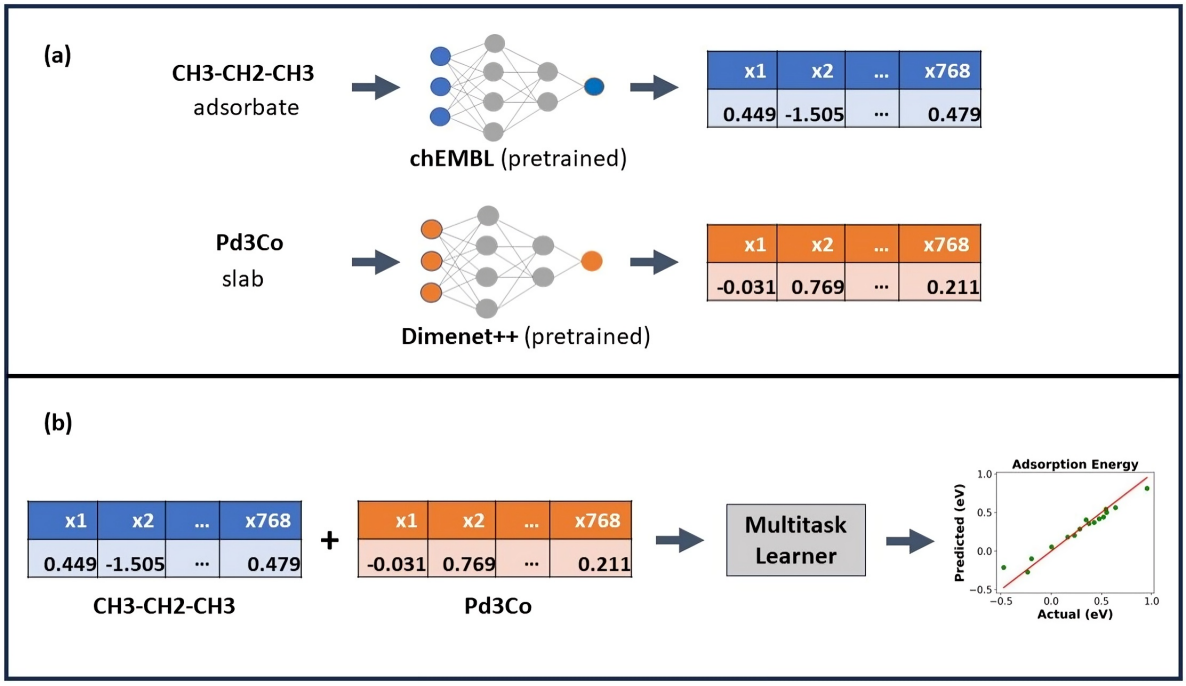


Figure 4.5: Pipeline of the proposed study: (a) Generating descriptors for adsorbate and surface, (b) Multitask learner for generating invariant representations and predicting adsorption energies.

the pretrained chEMBL model [128] to generate 768-length molecular descriptors, similar to the initial IMR method [29]. For the surface system, I utilized the pretrained DimeNet++ model from the Open Catalyst Project repository [142, 143]. For this study, I have generated similar 768-length surface descriptors to define the surface. Figure 4.5 demonstrates the pipeline used to generate the surface and adsorbate descriptors that were used in the machine learning model to predict the adsorption energies.

2. **Model Structure**: As the large number of samples restricted us from forming pairwise data, I have used a multitask learner where a common shared encoder is used to learn the invariant molecular representations that were shared by both tasks of predicting energies for both functionals. Separate headers are used to learn functional-specific characteristics. The trained encoder generates richer represen-

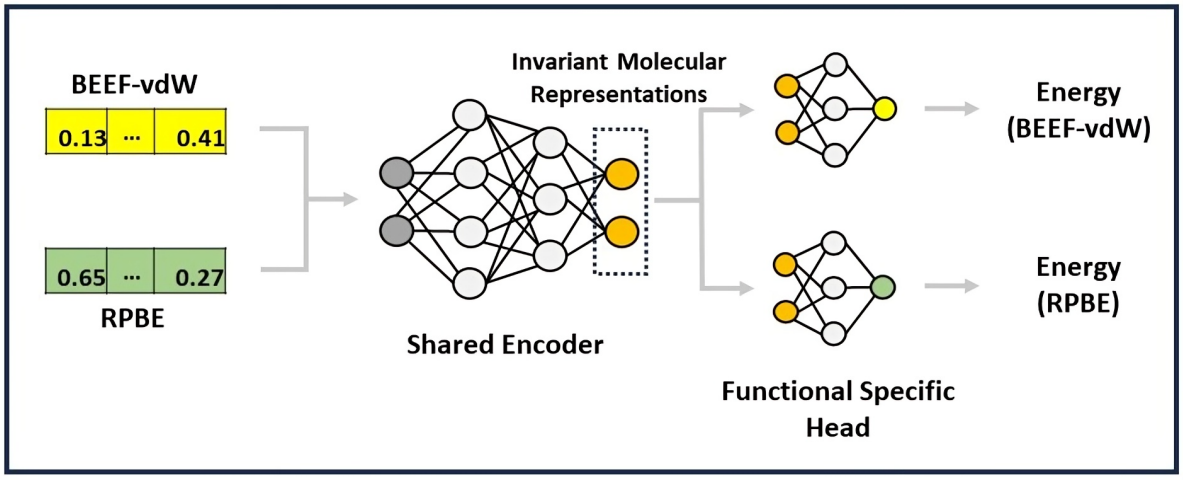


Figure 4.6: Multitask learner architecture: The encoder is shared for both tasks that learn invariant molecular representations, and functional-specific heads are used to learn functional-specific characteristics.

Table 4.7: Performance evaluation of three molecular representations (Original, PCA, and IMR) in larger datasets with multiple surface systems. The section reports the mean and standard deviation of the corresponding metric over 10 random trials. Statistically significant results are in bold.

Metric	Dataset	Functional	Original	PCA	IMR
R^2	Catalysis Hub	BEEF-vdW	0.894 ± 0.003	0.900 ± 0.004	**0.931 ± 0.005**
	OC20	RPBE	0.634 ± 0.003	0.627 ± 0.003	**0.678 ± 0.003**
MAE	Catalysis Hub	BEEF-vdW	0.522 ± 0.006	0.505 ± 0.005	**0.388 ± 0.016**
	OC20	RPBE	0.993 ± 0.002	1.000 ± 0.002	**0.906 ± 0.002**

tations by learning from the invariances of both functionals used. Finally, I have used the linear regression model implemented in Python's Scikit-learn [71] library on top of Original, PCA, and IMR (learned and generated by the shared encoder) to predict the adsorption energies. The two functionals I used to impose and learn the invariances are BEEF-vdW for the data from Catalysis Hub, and RPBE for OC20 data. The architecture is illustrated in Figure 4.6.

4.4.4 Results

The performance of the extended IMR learning method was evaluated using metrics such as R^2 (the higher the better) and Mean Absolute Error (MAE) (the lower the better).

Table 4.7 presents the results of the performance evaluation for three molecular representations (Original, PCA, and IMR) on larger datasets with multiple surface systems. The evaluation is reported based on the mean and standard deviation of the corresponding metrics over 10 random trials. Each trial utilized a different randomly chosen train and test set at a 3:1 ratio (i.e., 75% of the data for training and 25% for testing), ensuring enhanced variability and consistency in the results. For statistical significance tests, I have used a t-test with $\alpha = 0.05$ as the significance level.

The findings show that the proposed approach to learning invariant molecular representations achieves substantially better performance than the Original and PCA-based representations in predicting adsorption energies across various datasets and surface systems. By proposing a method that avoids the need for pairwise data, thus allowing the IMR learning method to scale to large datasets, and accommodating diverse surface systems, this study addresses two major limitations of the original IMR work [29]. I believe that this method of generating robust and reliable molecular representations holds great promise for predictive modeling in catalysis research, offering substantial impact in the design and optimization of catalysts for chemical reactions.

4.5 Summary

In this chapter, initially, I have introduced a novel learning method for predicting the adsorption energy of reaction intermediates. The proposed approach demonstrates the efficacy of learning from multiple functionals to overcome the challenges posed by the idiosyncrasies of different functionals by capturing the relative energy difference between pairs of intermediates calculated within the same functional and training the model across all different functionals. This study has reached several key conclusions: (i) by incorporating additional functionals, the proposed method generates superior representations (IMR) that lead to significantly improved performance compared to the Original and PCA-based representations; (ii) throughout numerous test cases, the Siamese model consistently per-

formed more effectively with BEEF-vdW as the test functional. Interestingly, this boost in performance persisted not only in the BEM training strategy where I used BEEF-vdW ensembles to train the Siamese model but also when the model was trained on the other three DFT functionals (FFM training strategy); (iii) the performance of two different training strategies with additional functionals (FFM and BEM) was found to be similar, despite the fact that FFM employed only three training DFT functionals, whereas the BEM employed 50 random ensembles or functionals. This can be attributed to the fact that even though the number of BEEF-vdW ensembles employed in the training process is large, they are less diverse compared to the DFT functionals, resulting in each DFT functional being more informative for learning the invariant molecular representations (IMR) of the species; and finally, (iv) a novel extension of the IMR learning method is proposed to address the limitations of the initial pair-wise model. This extension generates descriptors for both the adsorbate and the surface, enabling accurate modeling of interactions across multiple surface systems, and employs an efficient approach to handle the large number of reaction samples in extensive datasets, thereby facilitating the learning of invariant molecular representations from diverse functionals.

The proposed approach represents a significant advancement in the field of predicting the adsorption energy of reaction intermediates, as it enables the capture of the underlying chemistry of the system in a manner that is insensitive to the choice of functional and aware of the system the models are trained on. The findings of the study highlight the potential of the proposed approach to generating informative and robust molecular representations, which can result in improved performance in predictive modeling tasks.

Moving forward, I anticipate the general applicability of the method to a broad spectrum of functionals. In addition, a key area of potential work is predicting reliable transition state energies through machine learning, as these are often the most time-consuming steps for generating accurate chemical reaction models on catalysts. Incorporating these transition states and reaction energies to generate better predictive models can help to

generate accurate inputs with quantified uncertainty into microkinetic models, to better predict key experimental data, and allow for accurate calibration of existing experimental data into kinetic models. The objective is to learn molecular representations that are invariant, robust, and reliable, which can inform the design and optimization of catalysts for chemical reactions. I expect that the proposed method holds great promise in the field of predictive modeling for the adsorption energy of reaction intermediates and has the potential to make a substantial impact in this area.

CHAPTER 5: CGLEARN: CONSISTENT GRADIENT-BASED LEARNING FOR OUT-OF-DISTRIBUTION GENERALIZATION

5.1 Introduction

The advent of large datasets, sophisticated algorithms, and highly advanced complex models has propelled machine learning to achieve remarkable success across various domains. Despite these advancements, the performance of these models is heavily reliant on the assumption that the test data distribution is identical to the training data distribution. However, this dependency often leads to overfitting, as models become overparameterized and inadvertently learn spurious correlations from the training data [10–12]. Traditional models prioritize predictive accuracy without accounting for the causal relationships within the data, which becomes problematic when there are discrepancies between the training and test distributions. Consequently, models that rely on these spurious correlations exhibit significant performance degradation when faced with out-of-distribution (OOD) test data, undermining their robustness and generalization capabilities [13, 144].

Understanding causal relationships is essential for model interpretability, as well as enhancing generalization and robustness [145–147]. While Randomized Controlled Trials (RCTs) are ideal for learning causal structures, they are often expensive, unethical, or impractical. As a result, various methods for causal discovery have been developed. Constraint-based methods utilize conditional independence tests to discern causal directions [18, 37, 148], often yielding the Markov Equivalence Class (MEC) of causal structures. Score-based methods aim to optimize causal graphs over Directed Acyclic Graphs (DAGs) [38, 39, 149], but the search space's combinatorial nature can be computationally intensive. Methods like NOTEARS [40] have converted this combinatorial problem into

continuous optimization, resulting in various effective adaptations [26, 41, 42, 46, 63, 67]. Nonetheless, learning causal structures from observational data remains challenging due to different issues like selection bias, measurement errors, and confounders [22,150]. Furthermore, models solely based on empirical risk optimization can become reliant on spurious correlations.

To address these issues in practice, studies often leverage prior domain knowledge to enhance causal discovery [27,28,47–49,52]. Unfortunately, many causal discovery methods depend on specific assumptions (e.g., linearity, non-Gaussian noise) that do not always hold in real-world data. In addition to that some of these methods exploit variance scales e.g. var-sortability to identify causal orderings, performing well on unstandardized data but poorly after standardization [62, 151–153]. A recent line of study focuses on exploiting the invariance property of causal relationships across different environments. Methods like Invariant Causal Prediction (ICP) [154] aim to identify causal predictors by ensuring the conditional distribution of the target given these predictors remains stable across environments. This method leverages the invariance of causal relationships under different interventions, iterating over feature subsets to find those invariant across environments, considering them as potential causal parents of the target variable. Another study, IRM [13] optimizes a penalty function to achieve OOD generalization for predictive models, ensuring robust performance across environments. These methods significantly reduce the absorption of spurious correlations by focusing on stable and invariant relationships. The invariant learning framework provides a promising strategy to enhance model robustness and generalization in the presence of distribution shifts, with various domains exploiting invariance to learn better predictors and robust models [29, 155–157].

Motivated by this line of work and the current drawbacks of existing methods in structure learning and OOD generalization, this chapter presents, CGLearn, a general framework designed to enhance the generalization of ML models by leveraging gradient consistency across different environments. The materials presented in this chapter are

submitted as a conference paper and are currently under review. CGLearn does not require extensive domain knowledge or assumptions over data linearity or noise, making it a versatile and practical approach for learning robust predictive models. By focusing on feature invariance, emphasizing on reliable features, and reducing dependence on spurious correlations, CGLearn enhances the reliability and robustness of the models. The main contributions of this study are stated as follows:

- A novel general framework is proposed, which improves consistency in learning robust predictors by focusing on features that show consistent behavior across environments.
- This study provides both linear and nonlinear implementations of CGLearn, demonstrating its versatility and applicability across different model architectures.
- The current work demonstrates that CGLearn achieves superior predictive power and generalization, even without multiple environments, unlike most state-of-the-art methods that require diverse environments for effective generalization.
- The empirical evaluations on synthetic and real-world datasets, covering both linear and nonlinear settings, as well as regression and classification tasks, validate the effectiveness and robustness of the proposed method.

The remainder of this chapter is organized as follows: Section 5.2 delves into the methodology of CGLearn, detailing its linear and nonlinear implementations. Section 5.3 presents the experimental settings and evaluations. Finally, Section 5.4 encapsulates the conclusions, highlights the significant takeaways, and discusses future directions.

5.2 Methodology

This section presents the methodology of CGLearn, detailing both its linear and nonlinear implementations. The section starts by explaining the regular Empirical Risk Minimization (ERM) approach and then introduces the concept of gradient consistency used

in CGLearn. The primary concept of CGLearn is to enforce gradient consistency for each factor of our variable of interest across multiple environments to identify and utilize invariant features, thereby enhancing generalization and reducing dependence on spurious correlations.

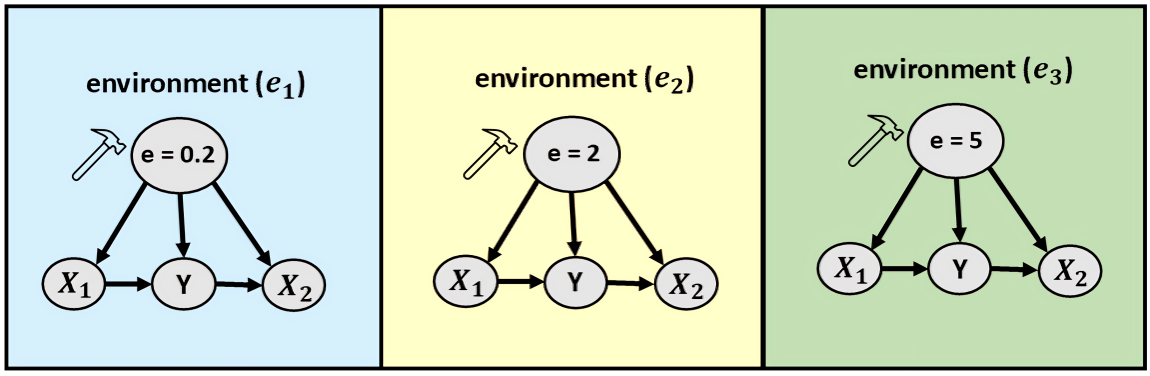


Figure 5.1: Illustration of three environments generated by intervening on the variable e, which takes distinct values $e = 0.2$, $e = 2$, and $e = 5$ in environments e_1, e_2, and e_3, respectively. In each environment, X_1 acts as a causal factor for the target variable Y, while X_2 is a spurious (non-causal) factor with respect to Y. This figure exemplifies how different interventions on e create distinct environments.

5.2.1 Empirical Risk Minimization (ERM)

Let's consider a simple linear problem where the goal is to predict the target variable Y using two features X_1 (causal) and X_2 (spurious) across multiple environments. Let $e_1, e_2, \ldots, e_m$ represent different environments. Environments can be considered as distinct distributions generated by different interventions, all of which share similar underlying causal mechanisms (see Fig. 5.1).

In the ERM framework, the weights for the features are updated by minimizing the empirical risk or the cost function (L), which is typically the mean squared error (MSE) between the predicted and actual values for a regression problem and cross-entropy loss for a classification task. Suppose the weights for the features at step t are w_1^t for X_1 and w_2^t for X_2. The gradient of the loss with respect to these weights in environment e_i is given by $\nabla L_j^{e_i}$, where $j \in \{1, 2\}$ and $i \in \{1, \ldots, m\}$ in our problem of interest.

The aggregated gradient across all environments can be calculated as the mean of the gradients:

$$\mu_j^{\text{grad}} = \frac{1}{m} \sum_{i=1}^{m} \nabla L_j^{e_i} \quad \text{for} \quad j \in \{1, 2\} \tag{5.1}$$

Using this aggregated gradient, the weights are updated as follows:

$$w_j^{t+1} = w_j^t - \eta \mu_j^{\text{grad}} \quad \text{for} \quad j \in \{1, 2\} \tag{5.2}$$

where η is the learning rate. In this setup of a standard Empirical Risk Minimization, the weights for both X_1 and X_2 get updated in each step regardless of their consistency across environments.

5.2.2 Linear Implementation of CGLearn

CGLearn modifies this approach by introducing a consistency check for the gradients. The idea is to update the weights only if the gradients are consistent across the available environments. This strategy focuses on invariant features and ignores spurious ones, expecting better generalization.

First, the proposed approach calculates the gradient of each feature in every environment as shown in Section 5.2.1. The mean of the gradients can be calculated as described in Eq. 5.1. Next, it computes the standard deviation of the gradients for each feature across all environments as follows:

$$\sigma_j^{\text{grad}} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(\nabla L_j^{e_i} - \mu_j^{\text{grad}} \right)^2} \tag{5.3}$$

The proposed method then calculates the consistency ratio, which is the absolute value of the ratio of the mean gradient to the standard deviation of the gradients:

$$C_j^{\text{ratio}} = \left| \frac{\mu_j^{\text{grad}}}{\sigma_j^{\text{grad}}} \right| \tag{5.4}$$

The consistency ratio, C_j^{ratio} defined in Eq. 5.4, is considered to be an indicator of the invariance of the gradient of variable X_j across all the training environments. A relatively larger mean compared to the standard deviation would indicate more similar or invariant gradients across the environments for the feature X_j, resulting in a higher value of C_j^{ratio}. On the other hand, a larger standard deviation indicates more diversity across the environments for X_j. Finally, this method formulates a consistency mask based on a predefined threshold C^{thresh}:

$$C_j^{\text{mask}} = \begin{cases} 1 & \text{if } C_j^{\text{ratio}} \geq C^{\text{thresh}} \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

The weights are updated only for the feature that has a nonzero mask and remains unchanged otherwise as per the following equation:

$$w_j^{t+1} = w_j^t - \eta \left(\mu_j^{\text{grad}} \cdot C_j^{\text{mask}} \right) \quad \text{for} \quad j \in \{1, 2\} \tag{5.6}$$

Considering our motivating example, where X_1 is causal and X_2 is spurious, we expect C_1^{mask} to be 1 and C_2^{mask} to be 0, indicating that the gradients of the causal variable X_1 are more consistent across the environments while they are not for the spurious variable X_2. Therefore, the weight for X_1 is mostly updated throughout the training steps while the weight for X_2 is not. The model thus focuses on the features that show consistency for learning the predictors of the target. This implementation strategy ensures to emphasis on reliable, invariant features while minimizing the impact of unreliable features by keeping their weights unchanged (or keeping the changes to a minimum). As a result, the contributions of the spurious features remain constant in the context of the model updates. The next section extends the CGLearn method to a nonlinear setting using multilayer perceptron (MLP) as an instance.

5.2.3 Nonlinear Implementation

For the nonlinear implementation using a multilayer perceptron (MLP), this study focuses on the gradients in the first hidden layer (h_1), where feature contributions can be distinctly identified. By controlling the contribution of spurious features at the first hidden layer, this method ensures they do not influence the final output. The process involves calculating the L^2-norm of the gradients for each feature in each environment, followed by determining the consistency ratio and mask to impose the consistency constraint.

$\|\nabla L^{e_i}_{jh_1}\|_2$ denotes the L^2-norm of the gradients of the j-th feature X_j in the i-th environment e_i at the first hidden layer h_1. We can compute the mean and standard deviation of the L^2-norm of the gradients across all environments as follows:

$$\mu_j^{\text{grad}} = \frac{1}{m}\sum_{i=1}^{m} \|\nabla L^{e_i}_{jh_1}\|_2 \tag{5.7}$$

$$\sigma_j^{\text{grad}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left(\|\nabla L^{e_i}_{jh_1}\|_2 - \mu_j^{\text{grad}}\right)^2} \tag{5.8}$$

The consistency ratio, C_j^{ratio} and the consistency mask, C_j^{mask} for feature X_j can be calculated by following Eq. 5.4 and 5.5 respectively. All the weights that belong to a particular feature, X_j in the first hidden layer h_1, are updated by following a similar strategy to Eq. 5.6. This updating strategy that depends on the consistency ratio, ensures that only the features that show consistency across the environments are considered to be updated. Otherwise, the weights remain unchanged, effectively treating them as constants similar to the linear implementation. For weights corresponding to the rest of the model other than the first hidden layer are updated as similar to ERM.

Fig. 5.2 illustrates a simple demonstration of the nonlinear MLP implementation of CGLearn. In this figure, X_1 and X_2 represent causal and spurious features, respectively, in accordance with our earlier motivating example. The gradient consistency is checked in

the first hidden layer (h_1), and weights are updated only if the consistency ratio exceeds the threshold, ensuring that features that behave consistently across environments are utilized.

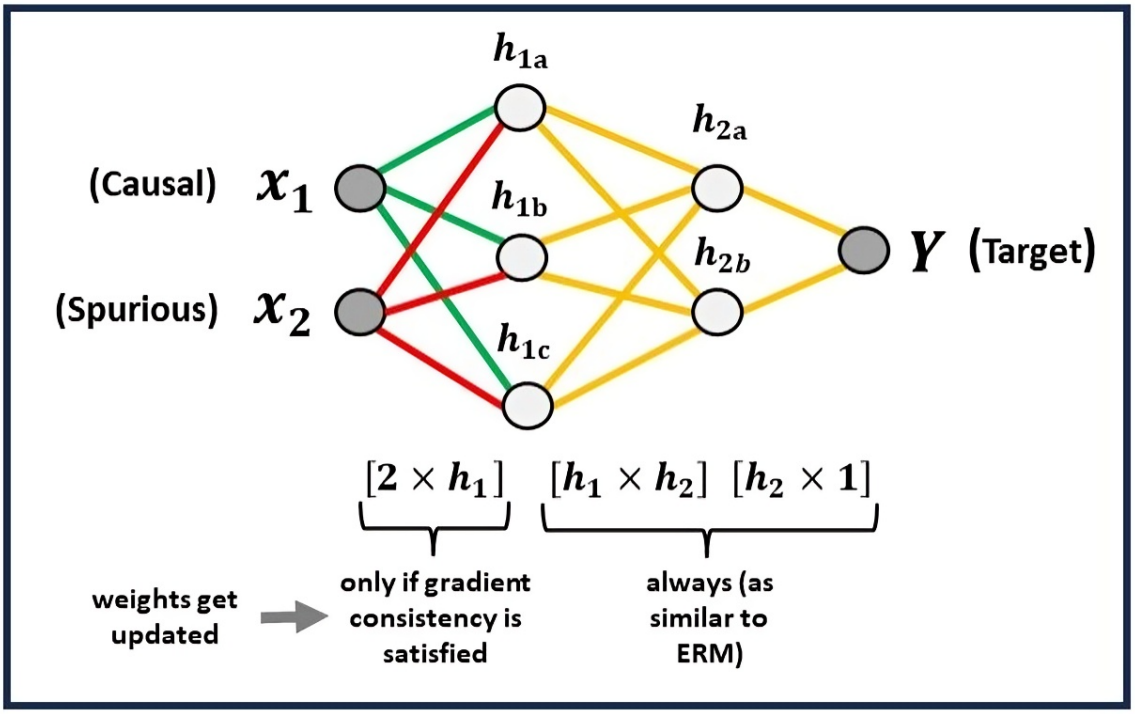


Figure 5.2: Nonlinear MLP implementation of CGLearn. X_1 (causal) and X_2 (spurious) feed into the first hidden layer h_1. Weight updates in h_1 are performed based on gradient consistency (using L^2-norm) for each feature across all training environments. The rest of the weights such as weights in h_2, are updated similarly to ERM (without imposing any consistency constraints).

In both implementations, the goal is to ensure that the model relies on features that show invariance across different environments. This leads to more robust and generalizable models by reducing dependency on spurious correlations.

5.3 Experiments and Results

In this study, I have considered three different major scenarios to assess the predictivity, robustness, and generalization capabilities of CGLearn. The first two scenarios are the ones where I considered linearly generated dataset-based experiments and in the last experimental case I used the nonlinear implementation of CGLearn using multilayer

perceptron (MLP) and applied it to different real-world regression and classification tasks.

For all evaluations, I have reported the mean and standard deviation of the performance metrics considered. For statistical significance tests, t-test with $\alpha = 0.05$ as the significance level is used.

5.3.1 Linear Multiple Environments

To evaluate the performance of the proposed CGLearn method, I first generated synthetic linear datasets inspired by the approach used in the Invariant Risk Minimization (IRM) framework [13]. The goal was to create diverse environments to test the robustness of the model under varying conditions.

This case study considers eight different experimental setups based on three key factors. Each setup included datasets with one target variable Y and ten feature variables X_1 to X_{10}. Features X_1 to X_5 acted as causal parents of Y, while X_6 to X_{10} were influenced by Y (non-causal). First, the study distinguished between scrambled (S) and unscrambled (U) observations by applying an orthogonal transformation matrix S for scrambled data and using the identity matrix I for unscrambled data. This scrambling ensures that the features are not directly aligned with their original scales, making the learning task more challenging. Secondly, it considers fully-observed (F) scenarios where hidden confounders did not directly affect the features (i.e., no hidden confounder effects on features), and partially-observed (P) scenarios where hidden confounders influenced the features with Gaussian noise. Third, I have incorporated two types of noise for the target variable Y: homoskedastic (O) noise, where the noise variance remained constant across different environments, and heteroskedastic (E) noise, where the noise variance varied depending on the environment, increasing with higher values of e. This distinction captures different real-world scenarios where noise may or may not depend on external factors. For each of these eight configurations (combinations of S/U, F/P, and O/E), I have generated datasets corresponding to three distinct environments defined by the values $e \in \{0.2, 2, 5\}$. Each

dataset consisted of 1000 samples. To ensure consistency with the IRM methodology and experimental setup, I have used $e = 5$ as the validation environment and determined the optimal consistency threshold (C^{thresh}) for the proposed method using the performance based on this validation data. We selected the threshold C^{thresh} from the candidate values $\{0.25, 1, 4, 16, 64\}$ based on validation performance. This threshold is critical for identifying the invariant and most reliable features across different environments. For more details on the data generation process, readers are referred to the IRM paper [13].

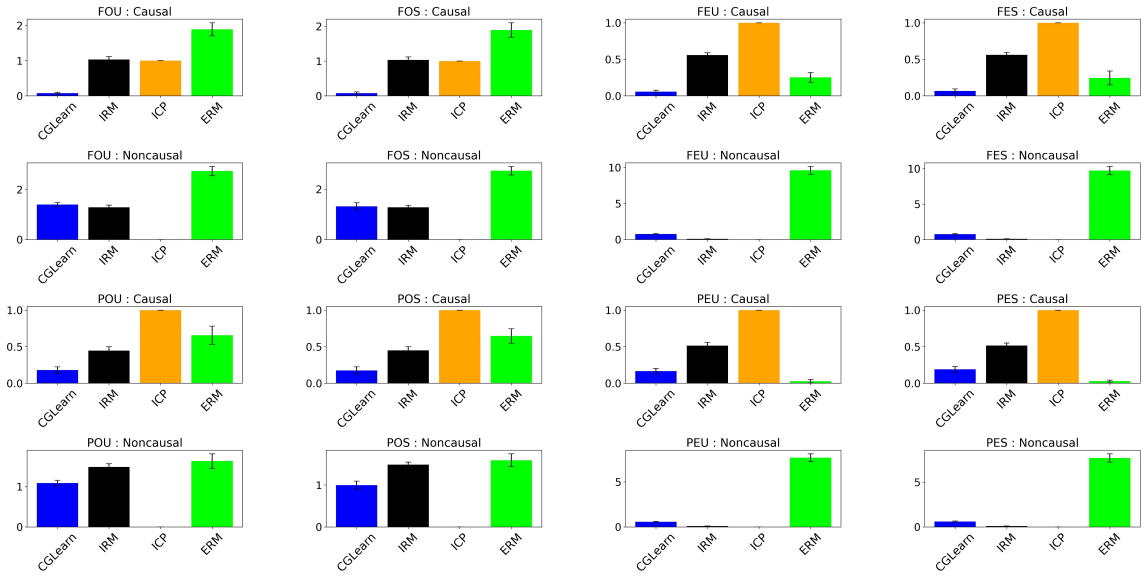


Figure 5.3: Performance comparison of CGLearn, IRM, ICP, and ERM across various linear multiple environment setups. Each subplot represents different configurations of the data, showing the mean squared error (MSE) for causal and noncausal variables over 50 trials.

The study compares the performance of CGLearn with Empirical Risk Minimization (ERM), Invariant Causal Prediction (ICP) [154], and IRM [13]. I have considered 50 random trials and reported the results in Fig. 5.3. In most of the cases, the proposed CGLearn approach achieves the lowest mean squared error (MSE), demonstrating superior performance across various test cases to distinguish the causal and noncausal factors of the target by exploiting invariance across environments. IRM performs better than ERM

but does not match the accuracy of CGLearn. ERM shows the highest errors in most cases, as it fails to differentiate between causal and noncausal features, relying on spurious correlations. Interestingly, ICP performs well in noncausal scenarios but poorly in causal ones. This observation aligns with the findings from the IRM study [13], which noted that ICP's conservative nature leads it to reject most covariates as direct causes, resulting in high causal errors.

5.3.2 Linear Single Environment

To evaluate the performance of the proposed method in scenarios with only one environment, I have generated synthetic linear datasets without relying on multiple environments as in previous experiments. For each of the eight cases, I have used a single setting with $e = 2$. The data generation process was similar to the previous section, with each dataset consisting of 1000 samples and ten feature variables, X_1 to X_{10}. The first five features (X_1 to X_5) acted as causal parents of the target variable Y, while the remaining five features (X_6 to X_{10}) were influenced by Y. Given the single environment setup, this experimental case could not consider IRM and ICP methods, as they require multiple environments to distinguish between causal and noncausal factors. Therefore, I compared the results of CGLearn solely with Empirical Risk Minimization (ERM).

To impose invariance in the proposed study, I created multiple batches, with $b = \{3, 5\}$ representing the number of batches created from the dataset. The last batch was used as the validation batch to determine the optimal consistency threshold parameter (C^{thresh}). We selected the threshold C^{thresh} from the candidate values $\{0.25, 1, 4, 16, 64\}$ based on validation performance. The gradient consistency is imposed across different batches to learn consistent and reliable factors of the target.

Table 5.1 demonstrates the evaluation in the single environment setup. Considering the causal error across all eight cases, CGLearn consistently achieves significantly lower mean squared errors (MSE) compared to ERM. For the noncausal error, CGLearn also

Table 5.1: Performance evaluation of CGLearn and ERM in linear single environmental setups. The table shows the Mean Squared Errors (MSE) for causal and noncausal variables across 50 trials for each configuration. Bold values indicate statistical significance.

Cases	Causal Error (MSE)		Noncausal Error (MSE)	
	CGLearn	ERM	CGLearn	ERM
FOU	**1.28 ± 0.40**	1.57 ± 0.13	0.61 ± 0.19	**0.54 ± 0.05**
FOS	**1.40 ± 0.43**	1.61 ± 0.10	0.53 ± 0.17	0.52 ± 0.06
FEU	**0.13 ± 0.05**	0.20 ± 0.04	**7.22 ± 2.15**	8.28 ± 0.28
FES	**0.16 ± 0.06**	0.20 ± 0.04	**7.47 ± 2.23**	8.36 ± 0.30
POU	**0.28 ± 0.11**	0.37 ± 0.08	0.51 ± 0.18	0.48 ± 0.11
POS	**0.34 ± 0.13**	0.39 ± 0.07	0.46 ± 0.17	0.48 ± 0.10
PEU	**0.24 ± 0.10**	0.32 ± 0.07	**5.11 ± 1.57**	5.83 ± 0.43
PES	**0.26 ± 0.10**	0.31 ± 0.06	**5.21 ± 1.58**	5.81 ± 0.36

outperforms ERM in most cases, suggesting the superiority of CGLearn. Even in the absence of multiple environments, the optimization strategy based on gradient consistency across different batches enables CGLearn to achieve better predictive power than standard ERM.

5.3.3 Nonlinear Multiple Environments

For the nonlinear experimental setups, in this study, I have considered two types of supervised learning tasks: regression and classification, both on real-world datasets. This approach allows us to evaluate the performance and robustness of the proposed CGLearn method in different real-world contexts. Recent work has highlighted limitations in the original Invariant Risk Minimization (IRM) framework, particularly in nonlinear settings where deep models tend to overfit [158]. To address this, the current section includes Bayesian Invariant Risk Minimization (BIRM) as a baseline, which has been shown to alleviate overfitting issues by incorporating Bayesian inference and thereby improving generalization in nonlinear scenarios [159].

Regression Tasks. For the regression tasks, I have used the Boston Housing dataset [160] and the Yacht Hydrodynamics dataset [161] for the comparative analysis of the nonlinear implementation of CGLearn with other baseline methods. The Boston Housing

Table 5.2: Performance comparison in nonlinear experimental setups for regression tasks. The table shows the RMSE for training and test environments across 10 trials. In test cases, the statistically significant values are marked in bold.

Dataset	# Optimal Envs.	Method	RMSE (Train)	RMSE (Test)
Boston	7	ERM	3.57 ± 0.11	6.43 ± 0.45
		IRM	3.79 ± 0.33	6.99 ± 0.74
		BIRM	3.77 ± 0.50	7.70 ± 0.52
		CGLearn	1.91 ± 0.26	**5.49 ± 0.28**
Yacht	5	ERM	0.21 ± 0.04	3.47 ± 1.15
		IRM	2.90 ± 0.03	4.36 ± 0.38
		BIRM	0.71 ± 0.19	3.15 ± 0.75
		CGLearn	0.48 ± 0.23	**2.29 ± 0.42**

dataset consists of 506 instances and 13 continuous attributes. It concerns housing values in suburbs of Boston, with the task being to predict the median value of owner-occupied homes (MEDV) based on attributes such as per capita crime rate (CRIM), proportion of residential land zoned for large lots (ZN), average number of rooms per dwelling, and etc. The Yacht Hydrodynamics dataset consists of 308 instances and 6 attributes. The task is to predict how much resistance a yacht experiences in the water, relative to its weight, based on different factors related to the shape of the yacht's hull and a specific speed-related measurement. Since real-world datasets do not naturally come with different environments, here I have followed a similar approach to the study by Ge et al. [162]. I used the K-Means [163] clustering algorithm to generate diverse environments and determined the optimal number of environments (between 3 to 10) using the Silhouette [164] method. For each dataset, I created all possible test cases where each environment was considered as for the test purpose once, and the rest were used for training. The evaluation is done by averaging the results over all possible test cases and repeating the process for 10 random trials. The models were evaluated based on RMSE, with the results shown in Table 5.2. For the Boston Housing dataset, I have found the optimal number of environments was 7, while for the Yacht Hydrodynamics dataset, it was 5. From Table 5.2, we can observe that all four methods perform better in the training environments than

the test environments, as expected. However, CGLearn shows significantly lower error in the testing or unseen environments compared to the other methods, demonstrating that imposing gradient consistency leads to less dependence on spurious features and thus better generalization.

Classification Tasks. For the classification tasks, I have evaluated the performance on two real-world classification datasets: the Wine Quality dataset for red and white wines from the UCI repository [165]. The Wine Quality dataset for red wine has 1599 instances and 11 attributes, while the dataset for white wine has 4898 instances and 11 attributes. The goal is to model wine quality based on physicochemical tests, such as fixed acidity, volatile acidity, citric acid, residual sugar, pH, and etc. Similar to the regression tasks, I have used K-means clustering to generate diverse environments and determined the optimal number of environments using the Silhouette method, finding 4 as the optimal number of environments for both classification datasets. I then generated all possible test cases where each environment was considered for the test purpose once, and the rest were used for training (as similar to the regression tasks). The performances were then averaged over all possible test cases and the process was conducted for 10 random trials. I have used accuracy and F1-score as evaluation metrics, with the results shown in Table 5.3. As expected, all methods performed better in training environments compared to test environments. However, the study finds that CGLearn achieved higher accuracy and F1-scores, which are desirable, and the superior performance was statistically significant for the F1-score on the Wine Quality Red dataset. It also had significantly better accuracy on the Wine Quality White dataset. Similar to the regression tasks, CGLearn demonstrated better predictive power and generalization over ERM, IRM, and BIRM for the classification tasks.

Limitations of CGLearn with Invariant Spurious Features. For this section, I have evaluated CGLearn on the Colored MNIST dataset, a synthetic binary classification task derived from MNIST [166] and proposed in the IRM study [13]. This dataset intro-

Table 5.3: Performance comparison in nonlinear setups for classification tasks. The table shows accuracy and F1-score for training and test environments across 10 trials. The statistically significant values are in bold for the test cases. WQR and WQW represent the Wine Quality Red and Wine Quality White datasets respectively. # Opt. Envs. indicates the number of optimal environments for each dataset determined using the K-Means clustering algorithm.

Dataset	# Opt. Envs.	Method	Accuracy (Train)	Accuracy (Test)	F1-score (Train)	F1-score (Test)
WQR	4	ERM	62.07 ± 0.34	58.08 ± 1.72	0.692 ± 0.004	0.535 ± 0.010
		IRM	63.68 ± 0.19	58.70 ± 1.54	0.644 ± 0.003	0.542 ± 0.014
		BIRM	64.94 ± 0.37	57.97 ± 0.93	0.626 ± 0.004	0.536 ± 0.011
		CGLearn	61.59 ± 0.44	59.60 ± 0.46	0.638 ± 0.008	**0.553 ± 0.007**
WQW	4	ERM	58.73 ± 0.22	51.15 ± 0.34	0.590 ± 0.002	0.447 ± 0.008
		IRM	58.82 ± 0.26	51.60 ± 0.40	0.566 ± 0.003	0.450 ± 0.013
		BIRM	58.04 ± 0.18	51.87 ± 0.32	0.530 ± 0.006	0.460 ± 0.026
		CGLearn	58.23 ± 0.38	**52.33 ± 0.32**	0.555 ± 0.005	0.460 ± 0.007

duces color as a spurious feature that strongly correlates with the label in the training environments but has the correlation reversed in the test environment. I have applied the nonlinear implementation of CGLearn and compared it with the results of ERM and IRM as reported in Ref. [13]. Over 10 trials, ERM achieved a training accuracy of 87.4 ± 0.2 and a test accuracy of 17.1 ± 0.6, while IRM achieved a training accuracy of 70.8 ± 0.9 and a test accuracy of 66.9 ± 2.5. In my experimental study, CGLearn achieved a training accuracy of 93.1 ± 0.8 and a test accuracy of 29.1 ± 0.8. While CGLearn slightly outperformed ERM in the test environment, it still struggled to generalize. This limitation arises because CGLearn imposes gradient consistency on the training environments to distinguish invariant features from spurious ones. However, in the Colored MNIST setup, the spurious feature (color) is consistent across both training environments, leading CGLearn to erroneously treat it as an invariant feature. Consequently, CGLearn relies on color and performs poorly in the test environment. To improve CGLearn's generalization, future work should focus on adapting the method to account for the varying nature of spurious features, even when they appear consistent across training environments.

5.4 Summary

This chapter presents, CGLearn, a novel approach for developing robust and predictive machine learning models by leveraging gradient consistency across multiple environments. By focusing on the agreement of gradients, CGLearn effectively identifies and utilizes invariant features, leading to superior generalization and reduced reliance on spurious correlations. The extensive experiments on both synthetic and real-world datasets, including regression and classification tasks, demonstrated that CGLearn outperforms traditional ERM and state-of-the-art invariant learners like ICP and IRM, achieving lower errors and better generalization in diverse scenarios. Notably, even in the absence of predefined environments, this study demonstrated that CGLearn can be effectively applied to different subsamples of data, leading to better predictive models than regular ERM. This flexibility enhances the applicability of the proposed approach in a wide range of real-world scenarios where many state-of-the-art methods require diverse and defined environments for OOD generalization.

Despite its strengths, CGLearn has limitations, particularly in scenarios where spurious features are invariant across environments, as observed in the Colored MNIST experiments. Such cases violate our assumption as generally we expect and observe causal features to be stable and invariant in nature whereas spurious features do not [146, 167]. CGLearn erroneously considers these invariant but spurious features as reliable, impacting its generalization performance. Addressing this limitation and adapting CGLearn to better handle such cases is a promising direction for future research.

Overall, the proposed method provides a significant step forward in the field of robust machine learning by effectively harnessing causal invariance. This work opens new avenues for developing models that are not only highly predictive but also resilient to distribution shifts, paving the way for more reliable applications in real-world settings.

CHAPTER 6: CONCLUSIONS AND FUTURE WORK

The goal of this dissertation is to highlight the different types of prior knowledge humans possess that can be utilized to aid causal learning, how to integrate this knowledge to enhance the learning process and to improve out-of-distribution (OOD) generalization, and to provide novel tools and methods for this purpose. This chapter summarizes the key findings and contributions of the dissertation, drawing conclusions on its impact. Finally, I explore potential future research directions and possibilities that emerge from the presented works.

6.1 Summary of the Dissertation and Contributions

This dissertation has explored various methods to enhance causal structure learning and improve OOD generalization by integrating domain knowledge. By addressing the challenges associated with inferring causal relationships from observational data, significant advancements have been made in developing robust and accurate predictive models. First, Chapter 1 provides the foundational knowledge necessary for understanding causal structure learning, covering essential concepts such as graphical and causal terms, the building blocks of causal graphs, and commonly used causal assumptions. This foundation set the stage for the original research presented in the subsequent chapters.

Chapter 2 focuses on a study that extends the original NOTEARS model to incorporate domain knowledge, enhancing its ability to learn causal structures. Key contributions include demonstrating that expert knowledge can significantly improve causal discovery, particularly when correcting active edges, and providing a comprehensive analysis of how different types of knowledge impact the model's performance. An interesting observation is that even redundant knowledge does no harm, suggesting that practitioners should

induce knowledge whenever available.

Chapter 3 adds another dimension to the causal learning and human-in-the-loop repertoire. It presents "CD-NOTEARS," a concept-driven structure learning method that incorporates conceptual knowledge into the learning process. The key contributions of this chapter include developing a novel extension that outperforms the original NOTEARS implementation and demonstrating its effectiveness in uncovering causal structures in high-dimensional conceptual data.

Chapter 4 focuses on utilizing the invariance property to develop causal and robust models. Since causal relations are invariant in nature, this research leverages prior knowledge of environments or data distributions to exploit this property, eliminating spurious correlations and learning robust causal relations. Significant contributions include the development of a novel method for predicting adsorption energy by creating invariant molecular representations, demonstrating superior performance over traditional approaches or molecular representations. This chapter also introduces an extension to handle multiple surface systems and large datasets effectively to learn the invariant representations.

Finally, Chapter 5 presents "CGLearn," a novel and general framework that enhances robustness and generalization in machine learning models by learning invariant predictors through gradient agreement across different environments. Key contributions include the proposed linear and nonlinear implementation of the approach, and demonstration of CGLearn's superior performance, its robust applicability across various tasks, and its effectiveness in leveraging gradient agreement for causal invariance.

To summarize, this dissertation has made substantial contributions to the field of causal structure learning and robust machine learning by integrating domain knowledge and employing advanced and novel techniques to enhance the learning process. The findings and methodologies presented not only deepen our understanding of causal inference from observational data but also open new avenues for future research. By paving the way for more reliable and accurate predictive models, I believe this work lays a solid foundation

for ongoing advancements in the intersection of causal learning and machine learning modeling, ultimately contributing to the development of more resilient AI systems.

6.2 Looking Forward: New Research Directions and Future Work

This dissertation opens several promising avenues for future research in enhancing causal structure learning and OOD generalization by integrating domain knowledge. One significant direction is the development of techniques that incorporate varying levels of confidence in prior knowledge, blending hard constraints with continuous optimization. This approach could enhance model flexibility and robustness. Additionally, integrating concept-driven methods with other causal discovery techniques and applying them to diverse domains like healthcare, finance, and social sciences could yield valuable insights and improvements. Automated methods to derive conceptual knowledge from data could also broaden the applicability of these approaches.

Expanding robust molecular representation methods to diverse reaction systems and more complex processes, such as predicting reliable transition state energies through machine learning for chemical reactions, presents another future research direction. Addressing the limitations of current models, particularly in scenarios where spurious features are invariant across environments, is crucial. Developing techniques to better distinguish between causal and spurious features can enhance generalization performance. Exploring the application of these models in various real-world scenarios without predefined environments will further broaden their impact. Overall, this dissertation provides a strong basis for advancing causal learning and machine learning modeling and I believe this work will facilitate the creation of dependable, adaptable, and causally informed predictive models for better generalization.

REFERENCES

[1] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. P. Pellet, P. Spirtes, and A. Statnikov, "Causality workbench," in *Causality in the Sciences*, Oxford University Press, 2011.

[3] A. Agrawal and A. Choudhary, "Deep materials informatics: Applications of deep learning in materials science," *Mrs Communications*, vol. 9, no. 3, pp. 779–792, 2019.

[4] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, and P. Riley, "Molecular graph convolutions: moving beyond fingerprints," *Journal of computer-aided molecular design*, vol. 30, pp. 595–608, 2016.

[5] D. Bourilkov, "Machine and deep learning applications in particle physics," *International Journal of Modern Physics A*, vol. 34, no. 35, p. 1930019, 2019.

[6] M. Ge, F. Su, Z. Zhao, and D. Su, "Deep learning analysis on microscopic imaging in materials science," *Materials Today Nano*, vol. 11, p. 100087, 2020.

[7] O. Koteluk, A. Wartecki, S. Mazurek, I. Kołodziejczak, and A. Mackiewicz, "How do machines learn? artificial intelligence as a new era in medicine," *Journal of Personalized Medicine*, vol. 11, no. 1, p. 32, 2021.

[8] A. Mosavi, Y. Faghan, P. Ghamisi, P. Duan, S. F. Ardabili, E. Salwana, and S. S. Band, "Comprehensive review of deep reinforcement learning methods and applications in economics," *Mathematics*, vol. 8, no. 10, p. 1640, 2020.

[9] J. Shen, J. Chowdhury, S. Banerjee, and G. Terejanu, "Machine fault classification using hamiltonian neural networks," *arXiv preprint arXiv:2301.02243*, 2023.

[10] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, "An investigation of why overparameterization exacerbates spurious correlations," in *International Conference on Machine Learning*, pp. 8346–8356, PMLR, 2020.

[11] T. Wang, R. Sridhar, D. Yang, and X. Wang, "Identifying and mitigating spurious correlations for improving robustness in nlp models," *arXiv preprint arXiv:2110.07736*, 2021.

[12] Y. Ming, H. Yin, and Y. Li, "On the impact of spurious correlation for out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10051–10059, 2022.

[13] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[14] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, 2009.

[15] E. C. Neto, M. P. Keller, A. D. Attie, and B. S. Yandell, "Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes," *The annals of applied statistics*, vol. 4, no. 1, p. 320, 2010.

[16] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[17] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips tutorial*, vol. 1, p. 2017, 2017.

[18] J. Pearl, *Causality*. Cambridge university press, 2009.

[19] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[20] G. Cooper, "Causal discovery from data in the presence of selection bias," in *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 140–150, 1995.

[21] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, pp. 973–978, Lawrence Erlbaum Associates Ltd, 2001.

[22] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proceedings of the twenty-first international conference on Machine learning*, p. 114, 2004.

[23] M. Scutari, "Learning bayesian networks with the bnlearn r package," *arXiv preprint arXiv:0908.3817*, 2009.

[24] K. Zhang, S. Zhu, M. Kalander, I. Ng, J. Ye, Z. Chen, and L. Pan, "gcastle: A python toolbox for causal discovery," *arXiv preprint arXiv:2111.15155*, 2021.

[25] A. Sharma and E. Kiciman, "Dowhy: An end-to-end library for causal inference," *arXiv preprint arXiv:2011.04216*, 2020.

[26] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, "Learning sparse nonparametric DAGs," in *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425, PMLR, 2020.

[27] J. Chowdhury., R. Rashid., and G. Terejanu., "Evaluation of induced expert knowledge in causal structure learning by notears," in *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, pp. 136–146, INSTICC, SciTePress, 2023.

[28] J. Chowdhury and G. Terejanu, "Cd-notears: Concept driven causal structure learning using notears," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 808–813, IEEE, 2023.

[29] J. Chowdhury, C. Fricke, O. Bamidele, M. Bello, W. Yang, A. Heyden, and G. Terejanu, "Invariant molecular representations for heterogeneous catalysis," *Journal of Chemical Information and Modeling*, vol. 64, no. 2, pp. 327–339, 2024.

[30] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[31] J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *Journal of Machine Learning Research*, 2014.

[32] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3, pp. 1–28, SpringerOpen, 2016.

[33] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, 2017.

[34] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain adaptation by using causal inference to predict invariant conditional distributions," *arXiv preprint arXiv:1707.06422*, 2017.

[35] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[36] J. Pearl and T. S. Verma, "A theory of inferred causation," in *Studies in Logic and the Foundations of Mathematics*, vol. 134, pp. 789–811, Elsevier, 1995.

[37] D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson, "Learning high-dimensional directed acyclic graphs with latent and selection variables," *The Annals of Statistics*, pp. 294–321, 2012.

[38] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[39] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International journal of data science and analytics*, vol. 3, no. 2, pp. 121–129, 2017.

[40] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, "DAGs with no tears: Continuous optimization for structure learning," *arXiv preprint arXiv:1803.01422*, 2018.

[41] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: DAG structure learning with graph neural networks," in *International Conference on Machine Learning*, pp. 7154–7163, PMLR, 2019.

[42] S. Lachapelle, P. Brouillard, T. Deleu, and S. Lacoste-Julien, "Gradient-based neural DAG learning," *arXiv preprint arXiv:1906.02226*, 2019.

[43] I. Ng, Z. Fang, S. Zhu, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," *arXiv preprint arXiv:1910.08527*, 2019.

[44] S. Heindorf, Y. Scholten, H. Wachsmuth, A.-C. Ngonga Ngomo, and M. Potthast, "Causenet: Towards a causality graph extracted from the web," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3023–3030, 2020.

[45] A. Jaber, J. Zhang, and E. Bareinboim, "Causal identification under Markov equivalence," *arXiv preprint arXiv:1812.06209*, 2018.

[46] D. Wei, T. Gao, and Y. Yu, "Dags with no fears: A closer look at continuous optimization for learning bayesian networks," *arXiv preprint arXiv:2010.09133*, 2020.

[47] J. Liu, Y. Chen, and J. Zhao, "Knowledge enhanced event causality identification with mention masking generalizations," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3608–3614, 2021.

[48] R. T. O'Donnell, A. E. Nicholson, B. Han, K. B. Korb, M. J. Alam, and L. R. Hope, "Causal discovery with prior information," in *Australasian Joint Conference on Artificial Intelligence*, pp. 1162–1167, Springer, 2006.

[49] O. Gencoglu and M. Gruber, "Causal modeling of twitter activity during Covid-19," *Computation*, vol. 8, no. 4, p. 85, 2020.

[50] R. Bradley, F. Dietrich, and C. List, "Aggregating causal judgments," *Philosophy of Science*, vol. 81, no. 4, pp. 491–515, 2014.

[51] D. Alrajeh, H. Chockler, and J. Y. Halpern, "Combining experts' causal judgments," *Artificial Intelligence*, vol. 288, p. 103355, 2020.

[52] B. Andrews, P. Spirtes, and G. F. Cooper, "On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge," in *International Conference on Artificial Intelligence and Statistics*, pp. 4002–4011, PMLR, 2020.

[53] D. Bhattacharjya, T. Gao, N. Mattei, and D. Subramanian, "Cause-effect association between event pairs in event datasets," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 1202–1208, 2021.

[54] Z. Li, X. Ding, T. Liu, J. E. Hu, and B. Van Durme, "Guided generation of cause and effect," *arXiv preprint arXiv:2107.09846*, 2021.

[55] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[56] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, "Accelerating human-in-the-loop machine learning: Challenges and opportunities," in *Proceedings of the second workshop on data management for end-to-end machine learning*, pp. 1–4, 2018.

[57] Y. Yang, E. Kandogan, Y. Li, P. Sen, and W. S. Lasecki, "A study on interaction in human-in-the-loop machine learning for text analytics," in *IUI Workshops*, 2019.

[58] Z. Fang, S. Zhu, J. Zhang, Y. Liu, Z. Chen, and Y. He, "Low rank directed acyclic graphs and causal structure learning," *arXiv preprint arXiv:2006.05691*, 2020.

[59] D. M. Chickering, "Learning Bayesian networks is NP-complete," in *Learning from data*, pp. 121–130, Springer, 1996.

[60] M. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of Bayesian networks is NP-hard," *Journal of Machine Learning Research*, vol. 5, 2004.

[61] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[62] A. Reisach, C. Seiler, and S. Weichwald, "Beware of the simulated dag! causal discovery benchmarks may be easy to game," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27772–27784, 2021.

[63] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and DAG constraints for learning linear DAGs," *arXiv preprint arXiv:2006.10201*, 2020.

[64] I. Ng, S. Lachapelle, N. R. Ke, S. Lacoste-Julien, and K. Zhang, "On the convergence of continuous constrained optimization for structure learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 8176–8198, PMLR, 2022.

[65] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery.," *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[66] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in neural information processing systems*, vol. 21, pp. 689–696, 2008.

[67] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," in *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 424–432, SIAM, 2022.

[68] A. C. Constantinou, Z. Guo, and N. K. Kitson, "The impact of prior knowledge on causal structure learning," *arXiv preprint arXiv:2102.00473*, 2021.

[69] U. Hasan and M. O. Gani, "KGS: Causal discovery using knowledge-guided greedy equivalence search," *arXiv preprint arXiv:2304.05493*, 2023.

[70] M. Kaiser and M. Sipos, "Unsuitability of NOTEARS for causal graph discovery," *arXiv preprint arXiv:2104.05441*, 2021.

[71] 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," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[72] A. H. Petersen, J. Ramsey, C. T. Ekstrøm, and P. Spirtes, "Causal discovery for observational sciences using supervised machine learning," *arXiv preprint arXiv:2202.12813*, 2022.

[73] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama, "Methods and tools for causal discovery and causal inference," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 12, no. 2, p. e1449, 2022.

[74] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988.

[75] "causal-graph-asia." https://www.bnlearn.com/bnrepository/discrete-small.html#asia.

[76] "dataset-asia." https://github.com/AnaRitaNogueira/Methods-and-Tools-for-Causal-Discovery-and-Causal-Inference.

[77] Dutch Central Bureau for Statistics, "Volkstelling, 2001," 2001.

[78] M. Lichman, "UCI machine learning repository." http://archive.ics.uci.edu/ml, 2013.

[79] L. Zhang, Y. Wu, and X. Wu, "Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 11, pp. 2035–2050, 2018.

[80] C. R. Catlow, M. Davidson, C. Hardacre, and G. J. Hutchings, "Catalysis making the world a better place," 2016.

[81] K. Reuter, C. P. Plaisance, H. Oberhofer, and M. Andersen, "Perspective: On the active site model in computational catalyst screening," *The Journal of Chemical Physics*, vol. 146, no. 4, p. 040901, 2017.

[82] J. Jover and N. Fey, "The computational road to better catalysts," *Chemistry–An Asian Journal*, vol. 9, no. 7, pp. 1714–1723, 2014.

[83] J. K. Nørskov, F. Studt, F. Abild-Pedersen, and T. Bligaard, *Fundamental concepts in heterogeneous catalysis*. John Wiley & Sons, 2014.

[84] A. H. Motagamwala and J. A. Dumesic, "Microkinetic modeling: a tool for rational catalyst design," *Chemical Reviews*, vol. 121, no. 2, pp. 1049–1076, 2020.

[85] C. Bo, F. Maseras, and N. López, "The role of computational results databases in accelerating the discovery of catalysts," *Nature catalysis*, vol. 1, no. 11, pp. 809–810, 2018.

[86] J. K. Nørskov, F. Abild-Pedersen, F. Studt, and T. Bligaard, "Density functional theory in surface chemistry and catalysis," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 937–943, 2011.

[87] M. Busch, M. D. Wodrich, and C. Corminboeuf, "Linear scaling relationships and volcano plots in homogeneous catalysis–revisiting the suzuki reaction," *Chemical science*, vol. 6, no. 12, pp. 6754–6761, 2015.

[88] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skulason, T. Bligaard, and J. K. Nørskov, "Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces," *Physical review letters*, vol. 99, no. 1, p. 016105, 2007.

[89] J. Greeley, "Theoretical heterogeneous catalysis: scaling relationships and computational catalyst design," *Annual review of chemical and biomolecular engineering*, vol. 7, pp. 605–635, 2016.

[90] A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, and G. A. Terejanu, "Prediction of adsorption energies for chemical species on metal catalyst surfaces using machine learning," *The Journal of Physical Chemistry C*, vol. 122, no. 49, pp. 28142–28150, 2018.

[91] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[92] D. Lopez-Paz, S. Sra, A. Smola, Z. Ghahramani, and B. Schölkopf, "Randomized nonlinear component analysis," in *International conference on machine learning*, pp. 1359–1367, PMLR, 2014.

[93] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Physical review letters*, vol. 98, no. 14, p. 146401, 2007.

[94] F. Pereira, K. Xiao, D. A. Latino, C. Wu, Q. Zhang, and J. Aires-de Sousa, "Machine learning methods to predict density functional theory b3lyp energies of homo and lumo orbitals," *Journal of chemical information and modeling*, vol. 57, no. 1, pp. 11–21, 2017.

[95] A. J. Chowdhury, W. Yang, K. E. Abdelfatah, M. Zare, A. Heyden, and G. A. Terejanu, "A multiple filter based neural network approach to the extrapolation of adsorption energies on metal surfaces for catalysis applications," *Journal of Chemical Theory and Computation*, vol. 16, no. 2, pp. 1105–1114, 2020.

[96] A. J. Chowdhury, W. Yang, A. Heyden, and G. A. Terejanu, "Comparative study on the machine learning-based prediction of adsorption energies for ring and chain species on metal catalyst surfaces," *The Journal of Physical Chemistry C*, vol. 125, no. 32, pp. 17742–17748, 2021.

[97] A. Kolluru, M. Shuaibi, A. Palizhati, N. Shoghi, A. Das, B. Wood, C. L. Zitnick, J. R. Kitchin, and Z. W. Ulissi, "Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery," *ACS Catalysis*, vol. 12, no. 14, pp. 8572–8581, 2022.

[98] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, *et al.*, "Open catalyst 2020 (oc20) dataset and community challenges," *Acs Catalysis*, vol. 11, no. 10, pp. 6059–6072, 2021.

[99] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, pp. 151–175, 2010.

[100] A. J. Medford, J. Wellendorff, A. Vojvodic, F. Studt, F. Abild-Pedersen, K. W. Jacobsen, T. Bligaard, and J. K. Nørskov, "Assessing the reliability of calculated catalytic ammonia synthesis rates," *Science*, vol. 345, no. 6193, pp. 197–200, 2014.

[101] J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, and K. W. Jacobsen, "Density functionals for surface science: Exchange-correlation model development with bayesian error estimation," *Physical Review B*, vol. 85, no. 23, p. 235149, 2012.

[102] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058301, 2012.

[103] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *The journal of physical chemistry letters*, vol. 6, no. 12, pp. 2326–2331, 2015.

[104] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of chemical physics*, vol. 134, no. 7, p. 074106, 2011.

[105] T. Morawietz and J. Behler, "A density-functional theory-based neural network potential for water clusters including van der waals corrections," *The Journal of Physical Chemistry A*, vol. 117, no. 32, pp. 7356–7366, 2013.

[106] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *The Journal of chemical physics*, vol. 145, no. 17, p. 170901, 2016.

[107] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, *et al.*, "Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for co2 reduction," *Acs Catalysis*, vol. 7, no. 10, pp. 6600–6608, 2017.

[108] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[109] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug discovery today*, vol. 23, no. 8, pp. 1538–1546, 2018.

[110] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science*, vol. 361, no. 6400, pp. 360–365, 2018.

[111] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[112] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," *Advances in neural information processing systems*, vol. 28, 2015.

[113] M. Rupp, R. Ramakrishnan, and O. A. Von Lilienfeld, "Machine learning for quantum mechanical properties of atoms in molecules," *The Journal of Physical Chemistry Letters*, vol. 6, no. 16, pp. 3309–3313, 2015.

[114] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS central science*, vol. 4, no. 1, pp. 120–131, 2018.

[115] W. Torng and R. B. Altman, "3d deep convolutional neural networks for amino acid environment similarity analysis," *BMC bioinformatics*, vol. 18, no. 1, pp. 1–23, 2017.

[116] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546, IEEE, 2005.

[117] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[118] C. H. Fricke, O. H. Bamidele, M. Bello, J. Chowdhury, G. Terejanu, and A. Heyden, "Modeling the effect of surface platinum–tin alloys on propane dehydrogenation on platinum–tin catalysts," *ACS Catalysis*, vol. 13, no. 16, pp. 10627–10640, 2023.

[119] C. Fricke, B. Rajbanshi, E. A. Walker, G. Terejanu, and A. Heyden, "Propane dehydrogenation on platinum catalysts: Identifying the active sites through bayesian analysis," *ACS Catalysis*, vol. 12, no. 4, pp. 2487–2498, 2022.

[120] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters*, vol. 77, no. 18, p. 3865, 1996.

[121] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu," *The Journal of chemical physics*, vol. 132, no. 15, p. 154104, 2010.

[122] B. Hammer, L. B. Hansen, and J. K. Nørskov, "Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals," *Physical review B*, vol. 59, no. 11, p. 7413, 1999.

[123] H. Peng, Z.-H. Yang, J. P. Perdew, and J. Sun, "Versatile van der waals density functional based on a meta-generalized gradient approximation," *Physical Review X*, vol. 6, no. 4, p. 041005, 2016.

[124] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational materials science*, vol. 6, no. 1, pp. 15–50, 1996.

[125] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical review B*, vol. 54, no. 16, p. 11169, 1996.

[126] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," *Physical review B*, vol. 47, no. 1, p. 558, 1993.

[127] C. R. Collins, G. J. Gordon, O. A. Von Lilienfeld, and D. J. Yaron, "Constant size descriptors for accurate machine learning models of molecular properties," *The Journal of chemical physics*, vol. 148, no. 24, p. 241718, 2018.

[128] "chembl pretrained model." https://huggingface.co/mrm8488/chEMBL_smiles_v1.

[129] A. Gupta, A. T. Müller, B. J. Huisman, J. A. Fuchs, P. Schneider, and G. Schneider, "Generative recurrent networks for de novo drug design," *Molecular informatics*, vol. 37, no. 1-2, p. 1700111, 2018.

[130] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service.," *J. Chem. Doc.*, vol. 5, no. 2, pp. 107–113, 1965.

[131] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized mlp architectures of neural networks," *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111–122, 2011.

[132] T. G. Tan, J. Teo, and P. Anthony, "A comparative investigation of non-linear activation functions in neural controllers for search-based game ai engineering," *Artificial Intelligence Review*, vol. 41, pp. 1–25, 2014.

[133] A. L. Maas, A. Y. Hannun, A. Y. Ng, *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, p. 3, Atlanta, Georgia, USA, 2013.

[134] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[135] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, PMLR, 2013.

[136] B. Hanin and D. Rolnick, "How to start training: The effect of initialization and architecture," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[137] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[138] N. A. Szaro, M. Bello, C. H. Fricke, O. H. Bamidele, and A. Heyden, "Benchmarking the accuracy of density functional theory against the random phase approximation for the ethane dehydrogenation network on pt (111)," *J. Phys. Chem. Lett.*, vol. 14, pp. 10769–10778, 2023.

[139] "Catalysis hub." https://www.catalysis-hub.org/.

[140] "Open catalyst project." https://opencatalystproject.org/.

[141] O. Mamun, K. T. Winther, J. R. Boes, and T. Bligaard, "High-throughput calculations of catalytic properties of bimetallic alloy surfaces," *Scientific data*, vol. 6, no. 1, p. 76, 2019.

[142] J. Gasteiger, S. Giri, J. T. Margraf, and S. Günnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules," *arXiv preprint arXiv:2011.14115*, 2020.

[143] "Fair chemistry." https://github.com/FAIR-Chem/fairchem.

[144] Y. He, Z. Shen, and P. Cui, "Towards non-iid image classification: A dataset and baselines," *Pattern Recognition*, vol. 110, p. 107383, 2021.

[145] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai," *International journal of human-computer studies*, vol. 146, p. 102551, 2021.

[146] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and S. Y. Philip, "Generalizing to unseen domains: A survey on domain generalization," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[147] B. G. Santillan, "A step towards the applicability of algorithms based on invariant causal learning on observational data," *arXiv preprint arXiv:2304.02286*, 2023.

[148] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2001.

[149] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560, 2018.

[150] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, 2011.

[151] M. Kaiser and M. Sipos, "Unsuitability of notears for causal graph discovery when dealing with dimensional quantities," *Neural Processing Letters*, vol. 54, no. 3, pp. 1587–1595, 2022.

[152] A. Reisach, M. Tami, C. Seiler, A. Chambaz, and S. Weichwald, "A scale-invariant sorting criterion to find a causal order in additive noise models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[153] W. Ormaniec, S. Sussex, L. Lorch, B. Schölkopf, and A. Krause, "Standardizing structural causal models," *arXiv preprint arXiv:2406.11601*, 2024.

[154] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference by using invariant prediction: identification and confidence intervals," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 78, no. 5, pp. 947–1012, 2016.

[155] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. Lilienfeld, and K.-R. Müller, "Learning invariant representations of molecules for atomization energy prediction," *Advances in neural information processing systems*, vol. 25, 2012.

[156] Q. Wang, Y. Zheng, G. Yang, W. Jin, X. Chen, and Y. Yin, "Multiscale rotation-invariant convolutional neural networks for lung texture classification," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 184–195, 2017.

[157] R. Bose and A. M. Roy, "Invariance embedded physics-infused deep neural network-based sub-grid scale models for turbulent flows," *Engineering Applications of Artificial Intelligence*, vol. 128, p. 107483, 2024.

[158] E. Rosenfeld, P. K. Ravikumar, and A. Risteski, "The risks of invariant risk minimization," in *International Conference on Learning Representations*, 2021.

[159] Y. Lin, H. Dong, H. Wang, and T. Zhang, "Bayesian invariant risk minimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16021–16030, 2022.

[160] D. Harrison and D. L. Rubinfeld, "Hedonic prices and the demand for clean air." J. Environ. Econ. and Management, 1978. UCI Machine Learning Repository, http://lib.stat.cmu.edu/datasets/boston.

[161] J. Gerritsma, R. Onnink, and A. Versluis, "Yacht Hydrodynamics." UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5XG7R.

[162] Y. Ge, S. Ö. Arik, J. Yoon, A. Xu, L. Itti, and T. Pfister, "Invariant structure learning for better generalization and causal explainability," *arXiv preprint arXiv:2206.06469*, 2022.

[163] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[164] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[165] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Wine Quality." UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

[166] Y. LeCun, L. D. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, no. 276, p. 2, 1995.

[167] J. Woodward, *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[168] M. Di Ventra and S. T. Pantelides, "Hellmann-feynman theorem and the definition of forces in quantum time-dependent and transport problems," *Phys. Rev. B*, vol. 61, no. 23, p. 16207, 2000.

[169] H. J. Monkhorst and J. D. Pack, "Special points for brillouin-zone integrations," *Phys. Rev. B*, vol. 13, no. 12, p. 5188, 1976.

[170] M. Methfessel and A. Paxton, "High-precision sampling for brillouin-zone integration in metals," *Phys. Rev. B*, vol. 40, no. 6, p. 3616, 1989.

APPENDIX A: THRESHOLD INCORPORATION AND SLACK VARIABLES

In Eq. 2.5, we have seen that our inequality constraint takes the following form:

$$h^{p}_{ineq}(W(\theta)) > 0$$

where p enumerates over each induced knowledge associated with a true active edge (*known active*) $X_i \rightarrow X_j$ imposing $[W(\theta)]_{ij} \neq 0$. NOTEARS uses a thresholding step that reduces false discoveries where any edge weight below the threshold value, w_{thresh} in its absolute value is set to zero. Thus, for any induction from true active edges ($X_i \rightarrow X_j$) we have the following constraint:

$$[W(\theta)]^2_{ij} \geq W^2_{thresh}.$$

The current study converts inequality constraints in the optimization to equality by introducing a set of slack variables y_p such that:

$$-[W(\theta)]^2_{ij} + W^2_{thresh} + y_p = 0 \qquad \text{s.t.} \quad y_p \geq 0 \tag{A.1}$$

In a similar manner, using the threshold value, W_{thresh} the equality constraints (associated with *known inactive* edges) take the form as:

$$[W(\theta)]^2_{ij} - W^2_{thresh} + y_q = 0 \qquad \text{s.t.} \quad y_q \geq 0 \tag{A.2}$$

where q enumerates over each induction associated with true inactive edge $X_i \nrightarrow X_j$ imposing $[W(\theta)]_{ij} = 0$.

APPENDIX B: SANITY CHECK - FUNCTIONAL SPECIFIC MODEL (FSM)

In this section, I have conducted four experimental case studies for the four DFT functionals (similar to the FFM and BEM). However, in all these cases, I have used both the training samples for the Siamese network and samples for the predictive analysis from the same functional. This simulates the scenario in which we do not have access to information from additional functionals but the Siamese model training is specific to the particular functional where it is tested.

Representations Generated from Flat Fingerprints

Table B.1: Performance evaluation of three molecular representation types (Original, PCA, and IMR) derived from 24-length flat molecular fingerprints using the FSM training approach. Values are given as Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, expressed in electron volts (eV). A lower MAE signifies enhanced performance.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.31 ± 0.04	0.32 ± 0.04	0.33 ± 0.04
	elastic	0.33 ± 0.04	0.34 ± 0.05	0.33 ± 0.04
	krr	0.35 ± 0.05	0.33 ± 0.05	0.38 ± 0.10
	svr	0.31 ± 0.06	0.30 ± 0.06	0.36 ± 0.08
BEEF-vdW	ridge	0.31 ± 0.05	0.31 ± 0.05	0.32 ± 0.04
	elastic	0.32 ± 0.05	0.33 ± 0.04	0.32 ± 0.04
	krr	0.33 ± 0.04	0.35 ± 0.04	0.37 ± 0.10
	svr	0.31 ± 0.05	0.31 ± 0.06	0.37 ± 0.09
RPBE	ridge	0.31 ± 0.05	0.31 ± 0.05	0.32 ± 0.05
	elastic	0.33 ± 0.05	0.33 ± 0.04	0.33 ± 0.05
	krr	0.35 ± 0.05	0.36 ± 0.04	0.37 ± 0.11
	svr	0.32 ± 0.07	0.34 ± 0.07	0.34 ± 0.10
SCAN+rVV10	ridge	0.37 ± 0.05	0.38 ± 0.04	0.38 ± 0.04
	elastic	0.40 ± 0.04	0.39 ± 0.05	0.39 ± 0.04
	krr	0.42 ± 0.08	0.42 ± 0.07	0.41 ± 0.13
	svr	0.38 ± 0.08	0.39 ± 0.10	0.42 ± 0.12

The results of our experiments using FSM training and representations generated from flat molecular fingerprints are presented first. These results are illustrated in Table B.1. The empirical findings show no significant difference in the performance of molecular rep-

resentations using our proposed model (IMR) compared to the PCA-based representations (PCA). This suggests that even in the absence of information from additional functionals, the performance of IMR is on par with that of baseline representations.

Representations Generated from Transfer Learning

Table B.2: Performance assessment of three molecular representation types (Original, PCA, and IMR) derived from fingerprints of the pretrained chEMBL model via FSM training. Values are presented as Mean Absolute Errors (MAEs) between predicted and DFT-calculated adsorption energies, in electron volts (eV). A lower MAE suggests superior accuracy.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.39 ± 0.06	0.35 ± 0.05	0.31 ± 0.05
	elastic	0.32 ± 0.07	0.33 ± 0.06	0.31 ± 0.05
	krr	0.27 ± 0.07	0.28 ± 0.04	0.30 ± 0.05
	svr	0.28 ± 0.06	0.29 ± 0.06	0.31 ± 0.05
BEEF-vdW	ridge	0.42 ± 0.07	0.37 ± 0.04	0.34 ± 0.04
	elastic	0.34 ± 0.05	0.33 ± 0.03	0.34 ± 0.04
	krr	0.34 ± 0.06	0.32 ± 0.04	0.34 ± 0.04
	svr	0.32 ± 0.05	0.34 ± 0.04	0.34 ± 0.05
RPBE	ridge	0.46 ± 0.08	0.39 ± 0.05	0.38 ± 0.04
	elastic	0.37 ± 0.05	0.37 ± 0.04	0.37 ± 0.03
	krr	0.44 ± 0.07	0.37 ± 0.04	0.38 ± 0.05
	svr	0.38 ± 0.05	0.40 ± 0.04	0.38 ± 0.05
SCAN+rVV10	ridge	0.44 ± 0.06	0.38 ± 0.04	0.36 ± 0.06
	elastic	0.39 ± 0.05	0.39 ± 0.03	0.36 ± 0.06
	krr	0.37 ± 0.04	0.39 ± 0.03	0.35 ± 0.06
	svr	0.35 ± 0.05	0.39 ± 0.05	0.36 ± 0.06

In Table B.2, the results of the experimental cases with representations generated by using FSM training and chEMBL fingerprints are presented. Again, we can see no statistical difference between the IMR and the PCA representations for any of the cases. Analogous to the scenario with flat molecular fingerprints, we can conclude that the proposed method generates molecular representations (IMR) that perform equally well for predictive modeling compared to the baseline methods (Original, PCA), even when the training data is specific to only one functional.

Representations Generated from Morgan Fingerprints

Table B.3: Evaluation of three molecular representation types (Original, PCA, and IMR) derived from 24-length Morgan fingerprints through FSM training. The values are given as Mean Absolute Errors (MAEs) between the predicted and DFT-calculated adsorption energies, expressed in electron volts (eV). A lower MAE suggests enhanced accuracy.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	0.34 ± 0.05	0.33 ± 0.04	0.36 ± 0.05
	elastic	0.32 ± 0.05	0.31 ± 0.04	0.34 ± 0.05
	krr	0.31 ± 0.05	0.32 ± 0.05	0.36 ± 0.07
	svr	0.31 ± 0.05	0.31 ± 0.05	0.36 ± 0.07
BEEF-vdW	ridge	0.34 ± 0.04	0.34 ± 0.05	0.34 ± 0.04
	elastic	0.33 ± 0.03	0.33 ± 0.04	0.34 ± 0.05
	krr	0.33 ± 0.04	0.34 ± 0.04	0.34 ± 0.05
	svr	0.33 ± 0.05	0.33 ± 0.05	0.33 ± 0.05
RPBE	ridge	0.37 ± 0.06	0.38 ± 0.06	0.39 ± 0.07
	elastic	0.37 ± 0.05	0.39 ± 0.05	0.39 ± 0.07
	krr	0.38 ± 0.05	0.39 ± 0.05	0.42 ± 0.06
	svr	0.38 ± 0.06	0.38 ± 0.06	0.39 ± 0.08
SCAN+rVV10	ridge	0.39 ± 0.05	0.39 ± 0.05	0.40 ± 0.05
	elastic	0.40 ± 0.03	0.39 ± 0.04	0.41 ± 0.04
	krr	0.39 ± 0.04	0.39 ± 0.03	0.41 ± 0.05
	svr	0.39 ± 0.05	0.40 ± 0.05	0.40 ± 0.05

Turning attention to Morgan's fingerprints, the findings from the FSM training are documented in Table B.3. Again we can see the trend persists. The comparative evaluation between our proposed IMR model and PCA representations reveals no significant differences. Thus, it's evident that, even when utilizing Morgan fingerprints, the IMR continues to deliver performance on par with established baselines in the absence of additional functionals.

APPENDIX C: DFT CALCULATION DETAILS

All DFT calculations were performed using the Vienna Ab initio Simulation Package (VASP) version 5.4.4. For geometric optimization, electron-exchange correlation was described by the Perdew-Burke-Ernzerhof (PBE) [120] functional coupled with dispersion corrections based on the D3-technique [121].

The Pt(111) catalyst site model, cleaved from an optimized Pt bulk crystal, comprised of 4 layers of 4x4 atoms with a vacuum space of 20 Å between periodic slabs. Pt(111) slab and intermediate species were relaxed until the Hellmann-Feynman force [168] per atom was less than 0.03 eV $Å^{-1}$. The Brillouin zone integration was sampled using a 5 $\times$ 5 $\times$ 1 Monkhorst-Pack [169] k-mesh with Methfessel-Paxton smearing [170] width (σ) of 0.2 eV. Frequency calculations were performed on the optimized structures to obtain the intermediate's entropic properties and free energy. For the other functionals, single-point calculations were done to obtain the corresponding energy of the PBE-D3 optimized structures, and these VASP energies were combined with PBE-D3-based vibrational frequencies to compute their free energies.

For all four functionals, the free energies were referenced to the bare catalyst slab, gas-phase propane, and gas-phase hydrogen energies.

APPENDIX D: MOLECULAR SPECIES SPECIFIC FINGERPRINT CONTRIBUTION

Figure D.1 illustrates an in-depth species-based fingerprint contribution breakdown employing the Four Functional Model (FFM) strategy with the PBE-D3 functional. This heatmap visualization elucidates the contributions of individual fingerprints to the prediction of adsorption energies across various molecular species. The attribution analysis presented in this heatmap has been calculated using integrated gradients, as implemented in the Python-based Captum library [137]. Through this analysis, we can observe significant insights into the model's behavior; for instance, the C_2 fingerprint shows increased attribution in species such as CCC and $CHCH_2CH$, where the presence of carbon atoms with two free valencies is crucial. This analysis exemplifies how the proposed strategy effectively capitalizes on functional invariances to learn invaluable patterns, thereby enhancing our understanding of molecular interactions within the dataset.

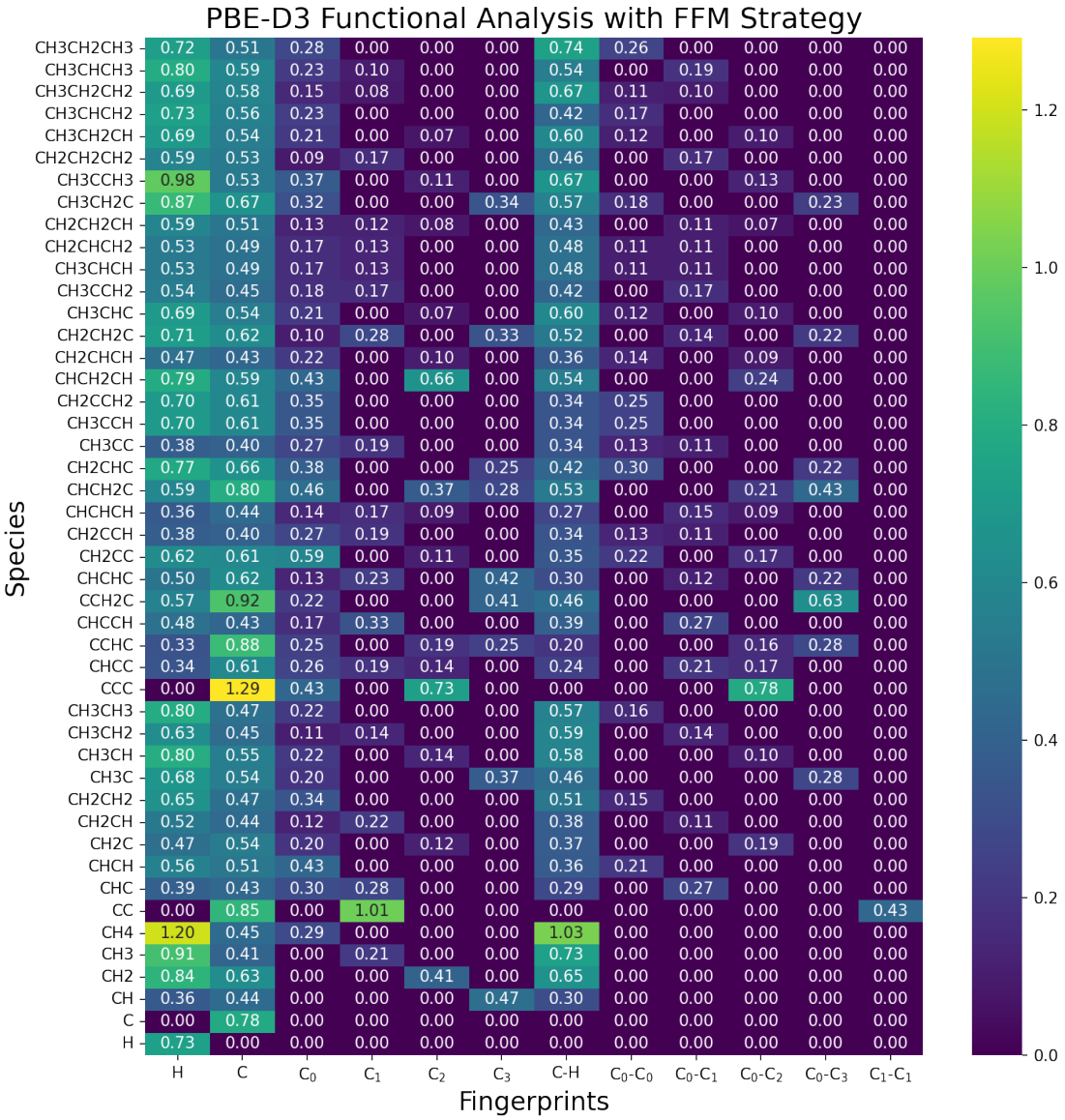


Figure D.1: Species-based breakdown of fingerprint contribution for FFM training strategy and PBE-D3 functional. Each cell in the heatmap signifies the contribution of a specific fingerprint to the adsorption energy prediction for a particular molecular species. Fingerprints with negligible contributions have been omitted for clarity. The color gradient indicates the magnitude of contribution, emphasizing the impact of specific fingerprints.

APPENDIX E: GOODNESS-OF-FIT ANALYSIS ACROSS EXPERIMENTAL CASES

Table E.1: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 24-length flat molecular fingerprints via FFM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.01 ± 0.15	-0.02 ± 0.14	0.13 ± 0.21
	elastic	-0.07 ± 0.14	-0.09 ± 0.15	0.13 ± 0.17
	krr	-0.12 ± 0.19	-0.06 ± 0.11	0.06 ± 0.16
	svr	0.00 ± 0.16	0.04 ± 0.14	0.06 ± 0.14
BEEF-vdW	ridge	0.05 ± 0.14	0.05 ± 0.14	0.51 ± 0.08
	elastic	0.02 ± 0.11	0.00 ± 0.08	0.51 ± 0.06
	krr	-0.02 ± 0.12	-0.04 ± 0.15	0.55 ± 0.08
	svr	0.05 ± 0.18	0.05 ± 0.20	0.59 ± 0.10
RPBE	ridge	0.17 ± 0.16	0.17 ± 0.15	0.43 ± 0.14
	elastic	0.13 ± 0.10	0.12 ± 0.09	0.49 ± 0.11
	krr	0.09 ± 0.10	0.05 ± 0.14	0.42 ± 0.15
	svr	0.14 ± 0.21	0.09 ± 0.19	0.46 ± 0.12
SCAN+rVV10	ridge	0.08 ± 0.09	0.07 ± 0.08	0.40 ± 0.10
	elastic	0.01 ± 0.07	0.03 ± 0.06	0.40 ± 0.07
	krr	-0.03 ± 0.19	-0.03 ± 0.19	0.44 ± 0.09
	svr	0.05 ± 0.22	0.05 ± 0.17	0.40 ± 0.15

Finally, I have utilized the D^2-score, implemented in Python's Scikit-learn [71] library, to quantify the goodness-of-fit for the models in all experimental cases. This score measures the fraction of deviance explained by the model relative to an intercept-only model and is defined as follows:

$$D^2(y, \hat{y}) = 1 - \frac{\text{dev}(y, \hat{y})}{\text{dev}(y, y_{null})} \tag{E.1}$$

In this equation, y represents the true target values, $\hat{y}$ denotes the predictions made by the model, and y_{null} is the median of the targets calculated on the training samples. Therefore, the term $\text{dev}(y, \hat{y})$ refers to the deviation of the model predictions from the true target values which is the mean absolute error of the model. Similarly, $\text{dev}(y, y_{null})$ signifies the mean absolute deviation of the true target values from the median calculated, serving as the baseline comparison for the model's predictive power. The best possible

D^2-score is 1.0, indicating a perfect prediction. Here, I present the mean and standard deviation of the D^2-score calculated across all 10 trials for each experimental case scenario as in Table E.1, E.2, E.3, E.4, E.5, E.6, E.7, E.8, and E.9.

Table E.2: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 768-length chEMBL fingerprints via FFM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.28 ± 0.22	-0.20 ± 0.17	0.19 ± 0.33
	elastic	-0.02 ± 0.13	-0.04 ± 0.13	0.24 ± 0.18
	krr	0.14 ± 0.14	0.10 ± 0.14	0.38 ± 0.18
	svr	0.12 ± 0.08	0.10 ± 0.07	0.41 ± 0.18
BEEF-vdW	ridge	-0.28 ± 0.20	-0.11 ± 0.15	0.58 ± 0.15
	elastic	-0.03 ± 0.15	0.01 ± 0.04	0.57 ± 0.16
	krr	-0.03 ± 0.21	0.02 ± 0.12	0.56 ± 0.18
	svr	0.02 ± 0.11	-0.04 ± 0.12	0.58 ± 0.16
RPBE	ridge	-0.21 ± 0.20	-0.04 ± 0.09	0.50 ± 0.18
	elastic	0.01 ± 0.15	0.01 ± 0.05	0.47 ± 0.21
	krr	-0.17 ± 0.26	0.02 ± 0.10	0.42 ± 0.14
	svr	0.00 ± 0.12	-0.04 ± 0.08	0.46 ± 0.12
SCAN+rVV10	ridge	-0.10 ± 0.18	0.03 ± 0.14	0.43 ± 0.14
	elastic	0.04 ± 0.09	0.05 ± 0.10	0.43 ± 0.15
	krr	0.08 ± 0.07	0.03 ± 0.08	0.46 ± 0.07
	svr	0.13 ± 0.10	0.01 ± 0.10	0.54 ± 0.09

Table E.3: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 24-length Morgan fingerprints via FFM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.08 ± 0.11	-0.07 ± 0.14	-0.08 ± 0.32
	elastic	-0.01 ± 0.06	-0.01 ± 0.08	0.19 ± 0.13
	krr	-0.01 ± 0.10	-0.04 ± 0.15	0.10 ± 0.11
	svr	-0.01 ± 0.11	0.01 ± 0.08	0.11 ± 0.14
BEEF-vdW	ridge	-0.03 ± 0.10	-0.02 ± 0.10	0.66 ± 0.08
	elastic	0.01 ± 0.04	-0.02 ± 0.05	0.66 ± 0.08
	krr	-0.01 ± 0.06	-0.02 ± 0.06	0.69 ± 0.06
	svr	0.01 ± 0.11	0.02 ± 0.11	0.70 ± 0.06
RPBE	ridge	0.02 ± 0.09	0.03 ± 0.09	0.53 ± 0.09
	elastic	0.02 ± 0.03	-0.01 ± 0.03	0.56 ± 0.07
	krr	0.00 ± 0.06	-0.03 ± 0.06	0.54 ± 0.08
	svr	0.01 ± 0.08	0.00 ± 0.07	0.56 ± 0.11
SCAN+rVV10	ridge	0.03 ± 0.14	0.04 ± 0.14	0.45 ± 0.12
	elastic	0.01 ± 0.06	0.03 ± 0.06	0.52 ± 0.11
	krr	0.04 ± 0.06	0.04 ± 0.07	0.57 ± 0.04
	svr	0.03 ± 0.06	0.05 ± 0.07	0.54 ± 0.06

Table E.4: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 24-length flat molecular fingerprints via BEM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.01 ± 0.15	-0.02 ± 0.14	0.21 ± 0.11
	elastic	-0.07 ± 0.14	-0.10 ± 0.15	0.18 ± 0.08
	krr	-0.12 ± 0.19	-0.06 ± 0.11	0.25 ± 0.14
	svr	0.00 ± 0.16	0.04 ± 0.14	0.22 ± 0.12
BEEF-vdW	ridge	0.05 ± 0.14	0.05 ± 0.14	0.42 ± 0.11
	elastic	0.02 ± 0.11	0.00 ± 0.09	0.42 ± 0.12
	krr	-0.02 ± 0.12	-0.06 ± 0.15	0.59 ± 0.11
	svr	0.05 ± 0.18	0.05 ± 0.20	0.58 ± 0.12
RPBE	ridge	0.17 ± 0.16	0.17 ± 0.16	0.33 ± 0.18
	elastic	0.13 ± 0.10	0.13 ± 0.09	0.39 ± 0.15
	krr	0.09 ± 0.10	0.04 ± 0.13	0.50 ± 0.09
	svr	0.14 ± 0.21	0.09 ± 0.19	0.46 ± 0.13
SCAN+rVV10	ridge	0.08 ± 0.09	0.07 ± 0.08	0.42 ± 0.10
	elastic	0.01 ± 0.07	0.03 ± 0.06	0.43 ± 0.11
	krr	-0.03 ± 0.19	-0.03 ± 0.19	0.51 ± 0.07
	svr	0.05 ± 0.22	0.05 ± 0.17	0.47 ± 0.09

Table E.5: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 768-length chEMBL fingerprints via BEM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.28 ± 0.22	-0.16 ± 0.21	0.18 ± 0.10
	elastic	-0.02 ± 0.13	0.00 ± 0.10	0.29 ± 0.12
	krr	0.14 ± 0.14	0.09 ± 0.09	0.21 ± 0.14
	svr	0.12 ± 0.08	0.06 ± 0.11	0.14 ± 0.18
BEEF-vdW	ridge	-0.28 ± 0.20	-0.13 ± 0.16	0.60 ± 0.12
	elastic	-0.03 ± 0.15	-0.01 ± 0.04	0.59 ± 0.14
	krr	-0.03 ± 0.21	-0.01 ± 0.12	0.67 ± 0.12
	svr	0.02 ± 0.11	-0.03 ± 0.12	0.56 ± 0.18
RPBE	ridge	-0.21 ± 0.20	-0.01 ± 0.10	0.55 ± 0.09
	elastic	0.01 ± 0.15	0.00 ± 0.04	0.57 ± 0.06
	krr	-0.17 ± 0.26	0.02 ± 0.10	0.54 ± 0.09
	svr	0.00 ± 0.12	-0.03 ± 0.08	0.58 ± 0.14
SCAN+rVV10	ridge	-0.10 ± 0.18	0.05 ± 0.13	0.40 ± 0.23
	elastic	0.04 ± 0.09	0.03 ± 0.09	0.43 ± 0.13
	krr	0.08 ± 0.07	0.04 ± 0.09	0.57 ± 0.09
	svr	0.13 ± 0.10	0.03 ± 0.11	0.52 ± 0.08

Table E.6: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 24-length Morgan fingerprints via BEM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.08 ± 0.11	-0.06 ± 0.12	0.01 ± 0.19
	elastic	-0.01 ± 0.06	-0.01 ± 0.08	0.28 ± 0.12
	krr	-0.01 ± 0.10	-0.08 ± 0.19	0.15 ± 0.13
	svr	-0.01 ± 0.11	-0.03 ± 0.15	0.20 ± 0.13
BEEF-vdW	ridge	-0.03 ± 0.10	-0.03 ± 0.09	0.66 ± 0.05
	elastic	0.01 ± 0.04	0.00 ± 0.06	0.64 ± 0.08
	krr	-0.01 ± 0.06	-0.03 ± 0.07	0.71 ± 0.13
	svr	0.01 ± 0.11	0.00 ± 0.11	0.70 ± 0.15
RPBE	ridge	0.02 ± 0.09	0.02 ± 0.10	0.52 ± 0.20
	elastic	0.02 ± 0.03	-0.02 ± 0.04	0.49 ± 0.34
	krr	0.00 ± 0.06	0.00 ± 0.06	0.52 ± 0.17
	svr	0.01 ± 0.08	0.01 ± 0.07	0.52 ± 0.10
SCAN+rVV10	ridge	0.03 ± 0.14	0.03 ± 0.13	0.44 ± 0.13
	elastic	0.01 ± 0.06	0.03 ± 0.07	0.45 ± 0.14
	krr	0.04 ± 0.06	0.03 ± 0.07	0.53 ± 0.06
	svr	0.03 ± 0.06	0.03 ± 0.06	0.45 ± 0.13

Table E.7: Mean and standard deviation based on D^2-scores using Original, PCA, and IMR representations derived from 24-length flat molecular fingerprints via FSM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.01 ± 0.15	-0.02 ± 0.14	-0.08 ± 0.14
	elastic	-0.07 ± 0.14	-0.09 ± 0.15	-0.07 ± 0.14
	krr	-0.12 ± 0.19	-0.05 ± 0.10	-0.22 ± 0.25
	svr	0.00 ± 0.16	0.04 ± 0.14	-0.13 ± 0.15
BEEF-vdW	ridge	0.05 ± 0.14	0.05 ± 0.14	0.04 ± 0.11
	elastic	0.02 ± 0.11	0.01 ± 0.10	0.02 ± 0.11
	krr	-0.02 ± 0.12	-0.06 ± 0.16	-0.11 ± 0.27
	svr	0.05 ± 0.18	0.05 ± 0.20	-0.11 ± 0.27
RPBE	ridge	0.17 ± 0.16	0.17 ± 0.15	0.15 ± 0.15
	elastic	0.13 ± 0.10	0.13 ± 0.09	0.14 ± 0.12
	krr	0.09 ± 0.10	0.04 ± 0.14	0.03 ± 0.28
	svr	0.14 ± 0.21	0.09 ± 0.19	0.09 ± 0.27
SCAN+rVV10	ridge	0.08 ± 0.09	0.07 ± 0.08	0.07 ± 0.07
	elastic	0.01 ± 0.07	0.03 ± 0.06	0.04 ± 0.05
	krr	-0.03 ± 0.19	-0.03 ± 0.19	0.00 ± 0.25
	svr	0.05 ± 0.22	0.05 ± 0.17	-0.02 ± 0.24

Table E.8: Mean and standard deviation of D^2-scores for models using Original, PCA, and IMR representations derived from 768-length chEMBL fingerprints via FSM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.28 ± 0.22	-0.15 ± 0.16	0.00 ± 0.11
	elastic	-0.02 ± 0.13	-0.06 ± 0.14	0.00 ± 0.12
	krr	0.14 ± 0.14	0.10 ± 0.06	0.02 ± 0.09
	svr	0.12 ± 0.08	0.09 ± 0.07	0.00 ± 0.09
BEEF-vdW	ridge	-0.28 ± 0.20	-0.12 ± 0.14	-0.03 ± 0.12
	elastic	-0.03 ± 0.15	0.00 ± 0.03	-0.02 ± 0.12
	krr	-0.03 ± 0.21	0.02 ± 0.12	-0.04 ± 0.13
	svr	0.02 ± 0.11	-0.04 ± 0.11	-0.04 ± 0.14
RPBE	ridge	-0.21 ± 0.20	-0.03 ± 0.10	0.00 ± 0.12
	elastic	0.01 ± 0.15	0.02 ± 0.05	0.01 ± 0.12
	krr	-0.17 ± 0.26	0.02 ± 0.10	0.00 ± 0.16
	svr	0.00 ± 0.12	-0.05 ± 0.07	-0.01 ± 0.15
SCAN+rVV10	ridge	-0.10 ± 0.18	0.05 ± 0.13	0.10 ± 0.16
	elastic	0.04 ± 0.09	0.04 ± 0.10	0.11 ± 0.15
	krr	0.08 ± 0.07	0.04 ± 0.09	0.13 ± 0.14
	svr	0.13 ± 0.10	0.03 ± 0.10	0.12 ± 0.15

Table E.9: Mean and standard deviation of D^2-scores for models using Original, PCA, and IMR representations derived from 24-length Morgan fingerprints via FSM training, across 10 trials. Higher scores indicate better-fitted models.

Test Functional	ML Alg.	Original	PCA	IMR
PBE-D3	ridge	-0.08 ± 0.11	-0.08 ± 0.13	-0.15 ± 0.16
	elastic	-0.01 ± 0.06	0.00 ± 0.04	-0.10 ± 0.22
	krr	-0.01 ± 0.10	-0.02 ± 0.11	-0.21 ± 0.39
	svr	-0.01 ± 0.11	0.01 ± 0.08	-0.26 ± 0.39
BEEF-vdW	ridge	-0.03 ± 0.10	-0.03 ± 0.10	-0.02 ± 0.13
	elastic	0.01 ± 0.04	0.00 ± 0.06	-0.04 ± 0.18
	krr	-0.01 ± 0.06	-0.02 ± 0.08	-0.04 ± 0.16
	svr	0.01 ± 0.11	0.01 ± 0.12	-0.02 ± 0.16
RPBE	ridge	0.02 ± 0.09	0.02 ± 0.09	-0.01 ± 0.11
	elastic	0.02 ± 0.03	-0.02 ± 0.03	-0.03 ± 0.12
	krr	0.00 ± 0.06	-0.02 ± 0.08	-0.12 ± 0.16
	svr	0.01 ± 0.08	0.01 ± 0.07	-0.04 ± 0.22
SCAN+rVV10	ridge	0.03 ± 0.14	0.03 ± 0.13	0.00 ± 0.19
	elastic	0.01 ± 0.06	0.03 ± 0.07	-0.02 ± 0.14
	krr	0.04 ± 0.06	0.03 ± 0.07	-0.01 ± 0.15
	svr	0.03 ± 0.06	0.03 ± 0.06	0.00 ± 0.15