Hybrid Graph Representation Learning for Molecular Optical Property Prediction in Low-Data Regimes

Michael Montemurri

Department of Mathematics and Statistics

McGill University

Montréal, QC, Canada

April 2026

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of

Master of Science

Abstract

Deep learning has emerged as a powerful tool for molecular property prediction, yet many experimental workflows operate in data-limited regimes where current state-of-the-art graph neural networks (GNNs) are unstable or impractical. This thesis investigates hybrid modelling strategies that decouple molecular representation learning from downstream regression for the prediction of optical properties of organic molecules. By combining directed message passing neural networks (D-MPNNs) with gradient-boosted decision trees, we demonstrate that hybrid architectures outperform end-to-end neural models for absorption wavelength prediction in small-to-mid-sized data regimes, particularly under distribution shift. To understand the origin of these gains, we analyze predictive variability using a nested variance decomposition framework, quantifying the respective contributions of representation learning and downstream readout to model instability. We further examine transfer learning, showing that its effectiveness depends strongly on chemical alignment between pretraining and downstream tasks. Beyond predictive accuracy, we explore the learned latent space using principal component analysis and Shapley Additive Explanations, revealing that the D-MPNN organizes molecular representations along physically meaningful electronic and structural axes without explicit physical supervision. Together, these results indicate that decoupling representation learning from regression can enhance performance in data-limited regimes, offering a principled and computationally practical alternative to fully end-to-end graph neural networks for molecular optical property prediction.

Abrégé

L'apprentissage profond s'est imposé comme un outil puissant pour la prédiction des propriétés moléculaires ; cependant, de nombreux protocoles expérimentaux opèrent dans des régimes à données limitées où les réseaux de neurones sur graphes (GNN) de l'état de l'art s'avèrent instables ou peu pratiques. Cette thèse étudie des stratégies de modélisation hybrides qui découplent l'apprentissage des représentations moléculaires de la régression en aval pour la prédiction des propriétés optiques des molécules organiques. En combinant des réseaux de neurones à passage de messages dirigés (D-MPNN) avec des arbres de décision boostés par gradient, nous démontrons que les architectures hybrides surpassent les modèles de neurones de bout en bout pour la prédiction de la longueur d'onde d'absorption dans des régimes de données de taille restreinte à moyenne, en particulier en présence d'un décalage de distribution. Pour comprendre l'origine de ces gains, nous analysons la variabilité prédictive à l'aide d'un cadre de décomposition imbriquée de la variance, en quantifiant les contributions respectives de l'apprentissage de la représentation et de la couche de lecture en aval à l'instabilité du modèle. Nous examinons en outre l'apprentissage par transfert, en montrant que son efficacité dépend fortement de l'alignement chimique entre le pré-entraînement et les tâches en aval. Au-delà de l'exactitude prédictive, nous explorons l'espace latent appris à l'aide de l'analyse en composantes principales et des explications additives de Shapley, révélant que le D-MPNN organise les représentations moléculaires selon des axes électroniques et structurels physiquement pertinents, et ce sans supervision physique explicite. Dans l'ensemble, ces résultats indiquent que le découplage entre l'apprentissage de

la représentation et la régression peut améliorer les performances dans des régimes à données limitées, offrant une alternative rigoureuse et avantageuse sur le plan calculatoire aux réseaux de neurones sur graphes fonctionnant entièrement de bout en bout pour la prédiction des propriétés optiques moléculaires.

Acknowledgements

I am incredibly grateful to my supervisor, Professor Eric Kolaczyk, for his guidance and steady confidence throughout this project. I especially appreciate the freedom he gave me to follow my curiosity and take this work in directions that excited me; having that level of trust and autonomy made this research not only possible, but genuinely enjoyable.

I would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), whose funding through the NSERC Canada Graduate Scholarships – Master's (CGS-M) program allowed me to focus fully on my research.

I also extend my thanks to my collaborator, Shambhavi Tannir, for her insight, thoughtfulness, and teamwork, and to Gian Favero for his guidance and for introducing me to the deep learning community at Mila.

To my parents and sisters, thank you for your constant love and encouragement. This achievement is built entirely on your support.

And finally, to Sarah: thank you for your patience, your support, and the joy you bring to my life. My partner and my best friend.

Contents

List of Figures

List of Tables

List of Acronyms

ANOVA	Analysis of Variance
D-MPNN	Directed Message Passing Neural Network
DSSCDB	Dye-Sensitized Solar Cell Database
ECFP	Extended-Connectivity Fingerprint
EMS	Expected Mean Squares
GBDT	Gradient-Boosted Decision Trees
GIN	Graph Isomorphism Network
GNN	Graph Neural Network
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
MPNN	Message Passing Neural Network
PCA	Principal Component Analysis
RMSE	Root Mean Squared Error
SHAP	Shapley Additive Explanations
SMILES	Simplified Molecular Input Line Entry System
TD-DFT	Time-Dependent Density Functional Theory
TPSA	Topological Polar Surface Area
UMAP	Uniform Manifold Approximation and Projection
WL	Weisfeiler–Lehman
XGB	Extreme Gradient Boosting

Chapter 1

Introduction

Predicting molecular properties directly from chemical structure *in silico* is a central objective in computational chemistry, with applications spanning materials discovery, drug development, and molecular design. In recent years, machine learning methods operating on molecular graphs or Simplified Molecular Input Line Entry System (SMILES) strings have achieved strong empirical performance. By learning representations directly from molecular structure, these models aim to replace or augment expensive quantum-chemical simulations with scalable, data-driven surrogates.

Despite these advances, many chemically relevant properties remain challenging to predict reliably. Optical absorption and emission, in particular, depend not only on dominant electronic factors such as energy gaps and π-conjugation, but also on subtler structural and environmental effects. Datasets for these properties are often limited in size, heterogeneous in composition, and influenced by experimental variability. This inherent data scarcity limits the statistical efficiency of high-capacity neural models and increases sensitivity to training stochasticity.

At a structural level, molecules are naturally represented as graphs, where atoms correspond to nodes and chemical bonds correspond to edges. Graph-based representations preserve topological relationships while maintaining invariance to atom indexing and molec-

ular orientation. Graph neural networks (GNNs) operationalize this structure by propagating information along bonds through iterative message-passing layers, constructing latent embeddings that summarize local chemical environments and their interactions.

Most molecular machine learning pipelines rely on end-to-end architectures that directly couple these learned embeddings with a neural "head" that maps those embeddings to the predicted response. While effective in data-rich domains and required for jointly training an encoder-head pair via backpropagation, this paradigm implicitly merges representation learning and prediction into a single optimization problem. This coupling makes it difficult to determine whether improvements arise from better structural embeddings, from flexible downstream regression, or from favourable stochastic variation. While theoretically, this jointly optimized encoder-head pair can perfectly fit the data, in practical low-data regimes, the high capacity of these models can lead to overfitting, with regularization becoming nontrivial.

These vulnerabilities become even more evident when there is a distribution shift. Scaffold-based splits, which group molecules by their core chemical structure, provide a better approximation of real-world generalization to new chemical spaces. However, they often show a significant drop in performance compared to random splits. In such cases, relying on average performance metrics can mask the considerable variability observed across different training runs.

Due to the current practical limitations of deep learning, alternative modelling paradigms remain widely used. Fixed molecular descriptors paired with classical regression methods, as well as physical approximations like time-dependent density functional theory (TD-DFT), offer competitive statistical efficiency and strong inductive biases in data-scarce environments. The coexistence of these approaches underscores a need to understand not only which models perform well, but under what specific structural constraints expressive neural modelling actually provides tangible benefits over modular alternatives.

Motivated by these challenges, this thesis explores a representation-centric approach to optical property prediction. Rather than focusing solely on end-to-end predictive accuracy, this work examines the behaviour of molecular representations learned by GNNs, particularly in data-scarce environments. By decoupling the graph encoder from downstream regression tasks and pairing it with implicitly regularized tree-based models, we investigate hybrid modelling strategies for improving performance in data-limited regimes.

The contributions of this thesis are fourfold. First, it systematically compares a broad range of modelling paradigms, including classical descriptor-based models, tree-based methods, physics-informed approaches, and end-to-end graph neural networks, using a Directed Message Passing Neural Network (D-MPNN) as a central reference point. The comparison spans multiple datasets, split strategies, and training set sizes. Second, it evaluates hybrid modelling strategies, demonstrating that decoupling representation learning from downstream prediction provides distinct advantages in low-data regimes. Third, it systematically analyzes the sources of predictive variability using nested variance decomposition, providing a mathematically grounded view of instability. Finally, it investigates the learned latent representations to assess whether the graph encoders capture chemically meaningful structure–property relationships or merely exploit dataset-specific artifacts. These contributions are primarily analytical rather than architectural. Rather than proposing a completely novel model, this work provides a practical characterization of when different modelling approaches are effective for molecular optical property prediction in data-limited settings typical of experimental labs, and identifies the sources of their performance differences in terms of representation quality, downstream regression, and stochastic variability.

The remainder of this thesis proceeds as follows. Chapter 2 situates the present work within the literature on molecular graph learning, molecular property prediction, transfer learning, and evaluation under data scarcity. Chapter 3 formalizes the mathematical framework underlying representation learning with MPNNs, our choice of tree-based regressor, and variance decomposition. Chapter 4 presents the experimental methodology and empirical

findings, with additional supporting analysis provided in Chapter 5. Chapter 6 synthesizes the results and discusses broader implications.

Chapter 2

Related Work

This chapter situates the present work within the broader literature on molecular representation learning and property prediction. We broadly review the development of graph representation learning, the classical approaches to molecular property prediction, the trade-offs between end-to-end and hybrid modelling strategies, and the specific challenges of model evaluation under data scarcity. A more formal discussion of models mentioned in this chapter is left to Chapter 3.

2.1 Graph Representation Learning for Molecular Data

2.1.1 Message Passing and the D-MPNN

The application of deep learning to molecular data has been driven by the Message Passing Neural Network (MPNN) framework, formalized by Gilmer *et al.* [22]. This framework unifies earlier atom-centric approaches [18, 33] by treating representation learning as an iterative process of local information aggregation. While early models demonstrated that neural networks could learn chemically meaningful features directly from structure, their effective depth was limited. As network depth, i.e. the number of message passing steps, increases, standard MPNNs become prone to "oversmoothing", a phenomenon formalized

by Li *et al.* [36] wherein repeated aggregations cause node representations to exponentially converge to indistinguishable vectors. This degradation is exacerbated by the "tottering" problem, first identified in the context of molecular graph kernels (pre-GNNs) by Mahé *et al.* [37]. They observed that random walks on molecular graphs would frequently step back and forth along the same bond, flooding the representation with redundant local information. Dai *et al.* [14] later demonstrated that standard node-centric GNNs suffer from this same inefficiency, as messages continuously cycle between adjacent nodes ($u \to v \to u$).

To address these structural limitations in chemical applications, Yang *et al.* introduced the Directed Message Passing Neural Network (D-MPNN) [63]. By operating on directed bond states rather than atom centers, the D-MPNN explicitly prevents immediate message backtracking. This architectural choice reduces the rapid accumulation of redundant information and has empirically established the D-MPNN as a robust standard baseline for molecular property prediction.

2.1.2 Expressivity and Global Context

Xu *et al.* established that standard MPNNs are limited in expressivity by the 1-dimensional Weisfeiler–Lehman (1-WL) graph isomorphism test [42, 61].

The 1-WL test is a classical algorithm that determines whether two graphs are topologically identical by iteratively aggregating the labels of neighbouring nodes [58]. In a chemical context, this theoretical upper bound implies that standard message-passing architectures cannot perfectly distinguish certain non-isomorphic molecules, such as highly symmetric cyclic structures or specific structural isomers, regardless of network depth or parameterization.

While architectures like the Graph Isomorphism Network (GIN) were developed to provably achieve this maximum 1-WL expressivity [61], overcoming the 1-WL limit entirely requires fundamentally different approaches. To distinguish non-isomorphic graphs that 1-WL cannot, higher-order GNNs and subgraph-based methods have been proposed [3, 42].

Additionally, to capture long-range node interactions that local message passing struggles to model, attention mechanisms and transformer-based architectures originally developed for natural language processing [55] have been adapted to the graph domain. These include Graph Attention Networks (GATs) [56] and transformer-style models such as Graphormer [64]. However, increased theoretical expressivity often comes at the cost of statistical efficiency. Empirical benchmarking suggests that these complex, computation-heavy architectures rarely outperform simpler message-passing schemes on the small, noisy datasets typical of experimental chemistry [46]. This gap highlights the practical preference for relatively simpler architectures, such as the D-MPNN, in low-data regimes.

2.2 Predictive Modelling Paradigms

2.2.1 Classical vs. Neural Approaches

Before the widespread adoption of GNNs, molecular property prediction was typically framed as a supervised learning problem on *fixed* molecular representations and hand-crafted descriptors, paired with classical regression algorithms such as Random Forests and gradient-boosted decision trees (e.g., XGBoost) [4, 12, 20]. A widely used family of fixed representations is the *Extended-Connectivity Fingerprint* (ECFP) [48], also known as the Morgan fingerprint.

Morgan fingerprints represent a molecule as a binary vector encoding the presence of local substructures. Starting from atom-centered identifiers, neighbourhoods are iteratively expanded out to a specified radius (measured in bond hops), producing a multiset of hashed substructure identifiers. These identifiers are then folded into a fixed-length bit vector of dimension d (e.g., $d = 2048$), where each bit indicates the presence of at least one corresponding substructure. In this work, we refer to Morgan fingerprints by their radius and bit-length (e.g., radius 2, 2048 bits).

Classical models were also commonly trained on physicochemical descriptors (e.g., molecular weight, topological polar surface area) [39] and, in some settings, features derived from quantum-chemical simulations. In particular, time-dependent density functional theory (TD-DFT) [9] provides physically grounded estimates of excited-state quantities such as vertical excitation energies, but its computational cost limits its applicability for large-scale screening.

These fixed-representation approaches remain competitive in many low-data regimes due to their strong inductive biases and statistical efficiency, and because tree-based models are often robust to irrelevant or noisy features [26]. In contrast, end-to-end neural models learn both the molecular representation and the predictor jointly from data [59]. While this can enable the model to capture complex, task-specific structure–property relationships [32], it also couples representation learning and regression in a single optimization problem. As a result, end-to-end models can be sensitive to initialization and training stochasticity, leading to higher variance in predictive performance across random seeds [13].

2.2.2 Hybrid and Decoupled Strategies

Hybrid modelling attempts to combine the representational power of GNNs with the robust inference of classical regressors. By decoupling representation learning from prediction, the neural encoder can be used to map discrete molecular graphs into a continuous latent space, which then serves as input features for a downstream model. Deng *et al.* introduced *XGraphBoost*, where they showed that extracting the learned molecular representations from a trained GNN and feeding them into a downstream XGBoost regressor improves upon end-to-end performance across an array of benchmarks [15]. While common in general machine learning (e.g., "frozen encoders" or "linear probing"), this strategy is underexplored in molecular machine learning as a mechanism for improving prediction under data scarcity.

2.3 Learning Under Scarcity and Distribution Shift

2.3.1 The Limits of Pretraining

In natural language processing and computer vision domains, a common strategy to mitigate data scarcity is pretraining on large supervised or unsupervised corpora, and allowing the model to adapt to the nuances of the dataset of interest through fine-tuning [16]. However, the transferability of molecular representations is highly conditional [44, 66]. Hu *et al.* show that naive transfer strategies report frequent "negative transfer", where features learned on a source task degrade performance on a target task due to domain mismatch. Even with their more complex pretraining strategies developed specifically for molecular graphs, they often observe no benefit for regression tasks [29]. Furthermore, fine-tuning pretrained models reintroduces optimization challenges, with performance often depending more on the fine-tuning protocol (e.g., learning rates, freezing schedules) than the pretraining objective itself [43].

2.3.2 Model Stability and Evaluation

Given the heterogeneity of optical property datasets, standard evaluation metrics based solely on mean predictive accuracy can obscure model reliability. In molecular machine learning, performance is often dominated by the choice of data splitting strategy, particularly under scaffold-based splits that explicitly enforce distribution shift [59].

A widely used scaffold split is based on Bemis–Murcko scaffolds [2], which define the *core framework* of a molecule by retaining ring systems and the linkers connecting them while removing peripheral substituents, depicted in Figure 2.1. Molecules are grouped according to this scaffold, and the resulting groups are then assigned to training, validation, and test sets such that scaffolds do not overlap across splits [59]. In contrast to random splits, this protocol prevents the model from benefiting from near-duplicate analogs across sets and

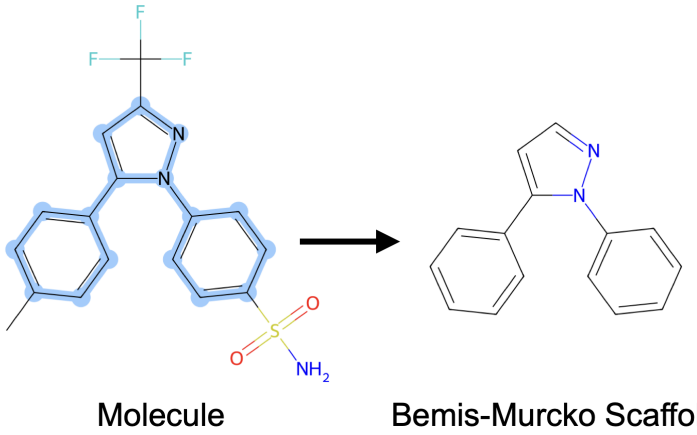

Figure 2.1: Depiction of a molecule and its corresponding Bemis–Murcko scaffold. The left image shows the molecule, and the right image depicts the Bemis–Murcko scaffold representation.

more closely reflects the intended use case of molecular screening: predicting properties for new chemical families rather than minor variants of known structures.

Chapter 3

Modelling Background and Statistical Foundations

3.1 Graph Theory & Invariance

Machine learning on molecular data requires representations that respect the underlying symmetries of the physical system. Unlike images or audio, which are defined on fixed grid structures, molecules are treated as non-Euclidean objects defined by their connectivity and element types. To formalize this, we rely on the framework of geometric deep learning [6].

3.1.1 Graph Representation

We represent a molecule as an undirected graph $G = (V, E)$, where $V = \{v_1, \ldots, v_N\}$ is the set of N nodes (atoms) and $E \subseteq V \times V$ is the set of edges (chemical bonds). The topological structure is encoded by the adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$. Associated with each node v_i is a feature vector $\mathbf{x}_i \in \mathbb{R}^F$ (e.g., atomic number, hybridization), forming a node feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$. Similarly, each edge $(i, j) \in E$ is associated with a bond feature vector $\mathbf{b}_{ij} \in \mathbb{R}^D$ (e.g., bond type or order), forming a bond feature tensor $\mathbf{B} \in \mathbb{R}^{N \times N \times D}$, where $\mathbf{B}_{ij\cdot} = \mathbf{b}_{ij}$ when $(i, j) \in E$ and is zero otherwise. The complete attributed graph is

thus represented as $G = (\mathbf{A}, \mathbf{X}, \mathbf{B})$. While this tensor representation of $\mathbf{B}$ is conceptually convenient, in practice we operate on a sparse edge-based representation. Specifically, we associate each edge $(i, j) \in E$ with its feature vector $\mathbf{b}_{ij}$ and store these features alongside a directed edge index

$$E^{\rightarrow} = \{(i, j), (j, i) \mid (i, j) \in E\}, \tag{3.1}$$

which enumerates all directed bonds in the graph.

3.1.2 Invariance and Equivariance

To be a valid embedding of a physical molecule, a learned function $\phi : \mathcal{G} \to \mathbb{R}^d$ mapping from the space of all valid molecular graphs $\mathcal{G}$ to a d-dimensional continuous embedding space must be invariant to permutation of vertices [6]. That is, for any permutation matrix $\mathbf{P} \in \{0, 1\}^{N \times N}$, the function must satisfy:

$$\phi(\mathbf{PAP}^{\top}, \mathbf{PX}, \mathbf{PBP}^{\top}) = \phi(\mathbf{A}, \mathbf{X}, \mathbf{B}) \tag{3.2}$$

where the permutation acts on the first two indices of the bond feature tensor $\mathbf{B} \in \mathbb{R}^{N \times N \times D}$.

This ensures that the predicted property depends only on the intrinsic molecular structure, not the arbitrary indexing of atoms. Standard GNN architectures achieve this by stacking permutation equivariant layers (which preserve node correspondence) followed by a global invariant pooling function (e.g., summation) [61, 67]:

$$\mathbf{z} = \sum_{v \in V} \mathbf{h}_v^{(T)} \tag{3.3}$$

where $\mathbf{h}_v^{(T)}$ denotes the learned hidden representation of node v at the final GNN layer T, and $\mathbf{z}$ is the graph-level embedding.

3.2 Message Passing Neural Networks

Here we formally describe the Message Passing Neural Network (MPNN) framework developed by Gilmer *et al.* [22], which unifies graph neural networks into a single algebraic form consisting of three phases: *message passing*, *node update*, and *readout*.

3.2.1 General Framework

Following Hamilton's notation of graph representation learning [27], we define the MPNN as an iterative process of neighbourhood aggregation. Let T denote the number of message-passing layers. For $t = 1, \ldots, T$, let $\mathbf{h}_u^{(t)}$ denote the latent representation of node u after t rounds of neighbourhood aggregation, with $\mathbf{h}_u^{(0)} = \mathbf{x}_u$.

The message passing operation can be expressed by defining an aggregated message $\mathbf{m}_{\mathcal{N}(u)}^{(t)}$ that incorporates both neighbouring node states and their corresponding bond features $\mathbf{b}_{vu}$:

$$\mathbf{m}_{\mathcal{N}(u)}^{(t)} = \text{AGGREGATE}^{(t)} \left(\left\{ \left(\mathbf{h}_v^{(t-1)}, \mathbf{b}_{vu} \right) \mid v \in \mathcal{N}(u) \right\} \right) \tag{3.4}$$

$$\mathbf{h}_u^{(t)} = \text{UPDATE}^{(t)} \left(\mathbf{h}_u^{(t-1)}, \mathbf{m}_{\mathcal{N}(u)}^{(t)} \right) \tag{3.5}$$

where AGGREGATE and UPDATE are arbitrary differentiable functions (e.g., neural networks), and $\mathcal{N}(u) = \{v \in V \mid (u, v) \in E\}$ denotes the set of neighbours of node u in the molecular graph $G = (V, E)$.

Intuitively, this two-step process allows each node to compile a summary of its immediate environment and then integrate this contextual information into its own representation. After running T iterations of message passing, the final node embeddings are given by $\mathbf{h}_u^{(T)}$ for all $u \in V$.

Finally, for graph-level tasks such as molecular property prediction, these node-level representations must be aggregated into a single vector representing the entire graph, $\mathbf{z}$.

This *readout* (also called pooling) step is defined by a permutation-invariant function (as discussed in Section 3.1.2), typically a sum or mean pooling over all nodes:

$$\mathbf{z} = \text{READOUT}\left(\{\mathbf{h}_u^{(T)} \mid u \in V\}\right) \tag{3.6}$$

Notice the strict structural bound on the flow of information in this iterative framework. During each message-passing step, a node's "receptive field" expands by one hop. After T iterations, information from nodes at graph distance at most T can influence a given node representation. More generally, for two atoms to jointly influence the hidden state of a shared intermediate node after T iterations, each must lie within T hops of that node. Consequently, the two atoms must be separated by a path of length at most $2T$. In standard implementations where $T = 3$, atoms separated by more than six bonds cannot jointly affect any single node representation during the nonlinear message-passing phase and can only be combined at the final global readout stage. In extended molecular systems, this locality constraint limits the model's ability to capture long-range interactions prior to global aggregation. This limitation motivates the attention-based methods discussed in Chapter 2.

3.2.2 Directed Message Passing Neural Networks

As discussed in Chapter 2, standard MPNNs operating on node states suffer from *tottering*, where a message from node v to u can immediately return to v in the next step ($v \to u \to v$).

To mitigate this, we employ the D-MPNN architecture [63]. The fundamental unit of representation is the **directed bond hidden state** $\mathbf{h}_{vu}^{(t)}$, representing the message sent from atom v to atom u.

For a directed bond from atom v to atom u, the initial message $\mathbf{h}_{vu}^{(0)}$ is computed using the features of the source atom v and the bond (v, u):

$$\mathbf{h}_{vu}^{(0)} = \sigma(\mathbf{W}_i[\mathbf{x}_v \parallel \mathbf{b}_{vu}]), \tag{3.7}$$

where $\mathbf{x}_v$ and $\mathbf{b}_{vu}$ denote atom and bond feature vectors, respectively, $\|$ denotes concatenation, σ denotes the ReLU activation function, and $\mathbf{W}_i$ is a learned initial linear transformation mapping features into a shared d-dimensional hidden space, where d is a chosen hyperparameter.

The update rule prevents backtracking by aggregating messages from all neighbours k of v *except* u (the recipient):

$$\mathbf{m}_{vu}^{(t+1)} = \sum_{k \in \mathcal{N}(v) \setminus \{u\}} \mathbf{h}_{kv}^{(t)} \tag{3.8}$$

The hidden state for the directed bond (v, u) is updated via a learned weight matrix $\mathbf{W}_h \in \mathbb{R}^{d \times d}$ shared across message-passing iterations, followed by activation σ:

$$\mathbf{h}_{vu}^{(t+1)} = \sigma\left(\mathbf{h}_{vu}^{(0)} + \mathbf{W}_h \mathbf{m}_{vu}^{(t+1)}\right) \tag{3.9}$$

After T iterations, the directed bond representations are aggregated back to the atom level by concatenating the original atom features $\mathbf{x}_u$ with the summed incoming bond messages and passed through a final linear layer:

$$\mathbf{h}_u = \sigma\left(\mathbf{W}_a \left[\mathbf{x}_u \,\|\, \sum_{v \in \mathcal{N}(u)} \mathbf{h}_{vu}^{(T)}\right]\right). \tag{3.10}$$

This produces the final node representation $\mathbf{h}_u$, which is then summed over all nodes to produce the global embedding $\mathbf{z}$. This final atom readout follows the Chemprop implementation of Yang *et al.*

Mapped to the general MPNN formalism (Equations 3.4 and 3.5), the D-MPNN employs standard summation as its AGGREGATE operator. The UPDATE function consists of a learned linear transformation of the aggregated message combined with a residual connection to the *initial* bond representation $\mathbf{h}_{vu}^{(0)}$, rather than to the previous hidden state $\mathbf{h}_{vu}^{(t)}$ at every step, followed by a ReLU non-linearity. This helps to prevent the network from "forgetting" the original features during message passing, while also mitigating oversmoothing.

3.3 The Manifold Hypothesis and Hybrid Modelling

A guiding assumption of this thesis is that the D-MPNN encoder $\phi_\theta : \mathcal{G} \to \mathbb{R}^d$ maps molecular graphs into a structured region of latent space. Although the space of all possible adjacency matrices is high-dimensional and combinatorial, chemically valid molecules occupy a constrained subset governed by valence rules and physical stability.

The *manifold hypothesis* in machine learning suggests that high-dimensional data in the real world often concentrate near a lower-dimensional subset of the ambient space [11, 24]. In this context, we hypothesize that the learned molecular representations $\mathbf{z}_{\text{mol}} = \phi_\theta(G)$ and fixed solvent representations $\mathbf{z}_{\text{sol}}$ lie near a structured subset $\mathcal{M} \subset \mathbb{R}^{d+d_{\text{sol}}}$, and that the target property varies smoothly with respect to this representation. That is, we assume there exists a function $f : \mathcal{M} \to \mathbb{R}$ such that

$$y = f(\phi_\theta(G) \parallel \mathbf{z}_{\text{sol}}),$$

and that small changes in the combined latent representation correspond to small changes in the predicted property.

This structure is not imposed explicitly, but emerges during end-to-end training, as the encoder and decoder are jointly optimized to minimize prediction error. In practice, this leads to representations that support stable downstream prediction, even in the absence of an explicit structural constraint.

This perspective motivates the hybrid modelling strategy:

1. Use the D-MPNN encoder to learn an information-dense representation $\phi_\theta(G) = \mathbf{z}_{\text{mol}}$ of molecular structure and concatenate this with a fixed solvent embedding $\mathbf{z}_{\text{sol}}$.

2. Replace the neural decoder with a regularized tree-based regressor that approximates $f(\mathbf{z}_{\text{mol}} \parallel \mathbf{z}_{\text{sol}})$.

By freezing ϕ_θ after training, we decouple representation learning from downstream regression. This allows us to evaluate whether the learned latent representation is sufficiently structured for lower-capacity models to perform competitively, particularly in data-scarce regimes.

3.4 Regularized Gradient Boosting (XGBoost)

While GNNs provide powerful structural representations, standard multi-layer perceptron (MLP) heads are prone to overfitting in low-data regimes due to their flexible parameterization. We therefore first train the encoder jointly with a differentiable MLP head, after which the encoder weights are frozen and the MLP is replaced with XGBoost [12], a tree-based regressor with built-in regularization mechanisms well suited to small datasets.

3.4.1 Additive Tree Ensemble

XGBoost approximates the target y by summing the outputs of K regression trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{z}_i), \quad f_k \in \mathcal{F} \tag{3.11}$$

where $\mathbf{z}_i$ is the frozen latent representation generated by the trained GNN encoder. Each tree f_k partitions the latent space $\mathbb{R}^d$ into disjoint regions (leaves) and assigns a continuous weight w_j to each leaf.

3.4.2 Regularized Objective and Tree Splitting

Unlike standalone Classification and Regression Tree (CART) procedures [5], which greedily select splits by minimizing empirical impurity measures (e.g., mean squared error), classical Gradient Boosting Machines (GBMs) [20] construct additive tree ensembles via first-order functional gradient descent on a specified loss function.

XGBoost extends this framework by incorporating an explicit structural regularization term directly into the boosting objective and by employing a second-order approximation of the loss during tree construction. At boosting iteration t, the objective is

$$\mathcal{L}^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{z}_i)) + \Omega(f_t) \tag{3.12}$$

where l is a differentiable convex loss function. The regularization term is

$$\Omega(f) = \gamma Z + \frac{1}{2}\lambda ||w||^2 \tag{3.13}$$

and penalizes both the number of leaves Z and the magnitude of the leaf weights w.

Using a second-order Taylor expansion of the loss, XGBoost evaluates the quality of a proposed split via the exact loss reduction (Gain):

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{3.14}$$

Here G and H denote the sums of first- and second-order gradients within each candidate child node. The L_2 regularization parameter λ shrinks leaf weights toward zero, discouraging high-variance fits to small partitions, while γ imposes a minimum structural improvement required for a split to occur. Together, these terms introduce an explicit bias toward simpler trees, potentially improving stability in low-sample regimes.

3.5 Nested Random-Effects Variance Decomposition

To quantify the source of predictive variability in the hybrid model, we adopt a two-factor nested random-effects model. Let y_{ijk} denote a performance metric obtained from the k-th replicate of the j-th regression head trained on the i-th encoder. The hierarchical design is modeled as

$$y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk}, \tag{3.15}$$

where $A_i \sim \mathcal{N}(0, \sigma_A^2)$ represents encoder-level variability, $B_{j(i)} \sim \mathcal{N}(0, \sigma_{B|A}^2)$ represents head-level variability nested within encoders, and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ captures residual stochasticity.

Under a balanced design, classical ANOVA Method of Moments estimators provide closed-form estimates of these variance components. More detailed derivations and implementation details are provided in Chapters 4 and 5.

Chapter 4

Hybrid Graph Neural Networks for Optical Property Prediction

Preamble to Chapter 4

The following chapter reproduces the manuscript entitled "*Hybrid GNN–Tree Models for Optical Property Prediction: Decoupling Representation and Regression in Low-Data Regimes*" by Michael Montemurri, Shambhavi Tannir, and Eric Kolaczyk. This work is intended for submission to the *Journal of Chemical Information and Modeling*.

The research presented in this chapter constitutes the primary original contribution of this thesis. The candidate was responsible for the conceptual development, implementation, experimental design, empirical analysis, and preparation of the manuscript. Collaborators provided guidance on interpretation and scientific framing.

The manuscript is reproduced here in full, with minor formatting adjustments for consistency with thesis style (e.g., figure numbering and section references).

This chapter builds on the theoretical and methodological foundations established in Chapters 2 and 3 and presents the empirical evaluation and comparative analysis of the classical, D-MPNN, and hybrid modelling frameworks.

4.1 Introduction

Machine learning models are now widely used for molecular property prediction (MPP), offering a data-driven alternative to quantum chemical simulation and experimental screening in areas such as drug discovery, materials science, and photophysical materials design [8, 30, 49]. Deep learning approaches that learn structure–property relationships directly from molecular graphs have achieved state-of-the-art performance across a wide array of MPP benchmarks [21, 28, 64].

Despite these successes, the practical utility of deep models is frequently constrained by data availability. Experimental research groups often operate in a "low-data" regime, generating task-specific datasets ranging from tens to a few thousand molecules. A variety of deep learning architectures have been applied to MPP tasks, including sequence-based models operating on string-based representations and graph-based models operating directly on molecular structure. Among these, graph neural networks (GNNs) have emerged as a dominant paradigm because they incorporate molecular topology as an explicit inductive bias. However, in low-data settings, end-to-end GNNs often exhibit unstable training dynamics, high sensitivity to random initialization, and poor generalization under distribution shift [13, 26]. This raises critical questions about the efficacy of GNN-based deep learning relative to simpler baselines when labeled data is scarce.

These challenges are particularly prevalent in the prediction of optical properties, such as absorption and emission wavelengths, which are vital for organic electronics, photonic materials, and fluorescent probes [19]. Unlike ground-state properties (e.g., solubility or toxicity), optical responses are governed by excited-state electronic structure, which depends sensitively on π-conjugation length, solvent environment, and conformational flexibility [35, 38, 47]. Small structural modifications can yield large, non-intuitive spectral shifts, creating "activity cliffs" that are difficult for smooth, continuous models to capture without dense sampling [54].

While quantum chemical methods such as Time-Dependent Density Functional Theory (TD-DFT) provide physically grounded estimates of excited-state properties, they scale poorly with both molecule and dataset size, requiring substantial computational resources [10, 17]. Tannir *et al.* demonstrate that descriptor-based gradient-boosted ensembles can achieve strong performance on curated optical datasets when augmented with handcrafted TD-DFT–derived features [53]. However, such approaches depend on computing quantum-chemical descriptors for each molecule at inference time, limiting their applicability in large-scale screening. To retain the informational benefits of quantum calculations without incurring inference-time costs, recent work by Greenman *et al.* has explored multi-fidelity strategies that train machine learning surrogates jointly on experimental and computational datasets to avoid the need for DFT calculations at inference [25, 45].

Conversely, in purely experimental low-data regimes, classical models based on fixed molecular descriptors (e.g., Morgan fingerprints) often outperform deep learning approaches due to their strong inductive biases and statistical efficiency [48, 59, 60]. However, these approaches may fail to capture the higher-order structural effects and long-range dependencies relevant to optical response.

Transfer learning has been proposed as a remedy for the lack of labeled data, but its effectiveness for molecular regression tasks is highly regime-dependent. Recent studies indicate that without strict chemical alignment between pretraining and downstream tasks, transfer learning can yield negligible or even negative transfer, particularly in regression tasks involving distinct physical mechanisms [7, 29, 51].

To balance the expressive power of GNNs with the statistical stability of classical regression, researchers are increasingly adopting hybrid modelling strategies. For instance, Deng *et al.* introduced XGraphBoost, a framework that utilizes GNN-learned representations as input features for Gradient-Boosted Decision Trees (GBDTs). Their comprehensive benchmark evaluation identified the combination of eXtreme Gradient Boosting (XGB) with a Directed Message Passing Neural Network (D-MPNN) as the top-performing hybrid strat-

egy [12, 15, 63]. While this work establishes the promise of hybrid architectures, it does not address the regimes most relevant to experimental photophysics, where datasets are small, domain-specific, and often subject to distribution shift. Moreover, prior hybrid studies have largely focused on aggregate benchmark performance, without examining how predictive accuracy, variance, and learned representations evolve as a function of dataset size.

Hybrid models explicitly decouple representation learning from downstream regression, making the structure of the learned embedding space a central object of study rather than a hidden intermediate. In low-data regimes, where models are prone to memorizing noise rather than learning generalizable features, it is imperative to ensure that learned representations capture chemically meaningful structure rather than overfitting to the limited training data.

In this work, we systematically evaluate a suite of classical, D-MPNN, and hybrid modelling strategies for optical property prediction across multiple datasets, data regimes, and train–test splits. We compare end-to-end D-MPNNs with hybrid models that use D-MPNN embeddings to train lightweight tree-based heads. We demonstrate that hybrid models broadly outperform end-to-end D-MPNNs on peak absorption wavelength prediction in small-data regimes. Through controlled transfer learning experiments, we show that the benefits of pretraining are strongly regime- and coverage-dependent, yielding substantial gains on chemically aligned downstream datasets but limited or negative returns under significant distribution shift. Furthermore, using a nested experimental design, we decompose predictive variability into contributions arising from representation learning and downstream prediction; we find that while predictive instability is dominated by representation learning across regimes, the downstream head contributes a substantial fraction in data-limited settings. Finally, we analyze the structure of learned molecular embeddings and demonstrate that these representations retain chemically meaningful organization aligned with known determinants of optical response, providing practical guidance for modelling optical properties in small and specialized chemical datasets.

4.2 Background and Methods

This section presents the modelling framework and analytical procedures used throughout this work, focusing on methodological components that are agnostic to dataset and experimental setting. Dataset-specific design choices, training regimes, and evaluation protocols are treated separately.

4.2.1 Graph Neural Network Encoder

Molecules are represented as graphs $G = (V, E)$, where nodes $v \in V$ correspond to atoms and edges $(v, u) \in E$ correspond to chemical bonds. Following Yang *et al.* [63], we initialize our feature vectors with the standard RDKit atom- and bond-level features listed explicitly in Tables S1 and S2.

We employ a D-MPNN as the molecular graph encoder using the Chemprop open-source package (version 1.6.1) [63]. In contrast to atom-centered message passing, the D-MPNN propagates messages along directed bonds to prevent message backtracking. For a directed bond from atom v to atom u, the initial message $\mathbf{h}_{vu}^{(0)}$ is computed using the features of the source atom v and the bond (v, u):

$$\mathbf{h}_{vu}^{(0)} = \sigma(\mathbf{W}_i[\mathbf{x}_v \,\|\, \mathbf{b}_{vu}]), \tag{4.1}$$

where $\mathbf{x}_v$ and $\mathbf{b}_{vu}$ denote atom and bond feature vectors, respectively, $\|$ denotes concatenation, σ denotes the ReLU activation function, and $\mathbf{W}_i$ is a learned initial linear transformation mapping features into a shared d-dimensional hidden space, where d is a chosen hyperparameter.

In each message-passing step $t \in \{1, \ldots, T\}$, the hidden state $\mathbf{h}_{vu}^{(t)}$ is updated by aggregating messages from all incoming directed bonds (k, v) except for the reverse bond (u, v):

$$\mathbf{h}_{vu}^{(t+1)} = \sigma\left(\mathbf{h}_{vu}^{(0)} + \mathbf{W}_h \sum_{k \in \mathcal{N}(v) \setminus \{u\}} \mathbf{h}_{kv}^{(t)}\right), \tag{4.2}$$

where $\mathcal{N}(v)$ denotes the neighbours of atom v, and $\mathbf{W}_h \in \mathbb{R}^{d \times d}$ is a learned weight matrix shared across all steps.

After T iterations, the bond-level messages are aggregated to form atom-level representations $\mathbf{h}_u$. The original atom features $\mathbf{x}_u$ are concatenated with the summed incoming bond messages and passed through a final linear layer:

$$\mathbf{h}_u = \sigma\left(\mathbf{W}_a \left[\mathbf{x}_u \,\|\, \sum_{v \in \mathcal{N}(u)} \mathbf{h}_{vu}^{(T)}\right]\right). \tag{4.3}$$

The global molecular embedding $\mathbf{z}_{\text{mol}} \in \mathbb{R}^d$ is then computed via global sum pooling, yielding a permutation-invariant representation of the molecular graph:

$$\mathbf{z}_{\text{mol}} = \sum_{u \in V} \mathbf{h}_u. \tag{4.4}$$

Let $\mathbf{z}_{\text{sol}}$ denote the fixed solvent representation. The full representation used for downstream prediction is

$$\mathbf{z} = [\mathbf{z}_{\text{mol}} \,\|\, \mathbf{z}_{\text{sol}}]. \tag{4.5}$$

Throughout this work, we refer to the mapping from this combined representation to the scalar property prediction as the head. In the end-to-end setting, this head is a neural feedforward network (FFN), whereas in the hybrid setting it is replaced by a separate gradient-boosted tree model.

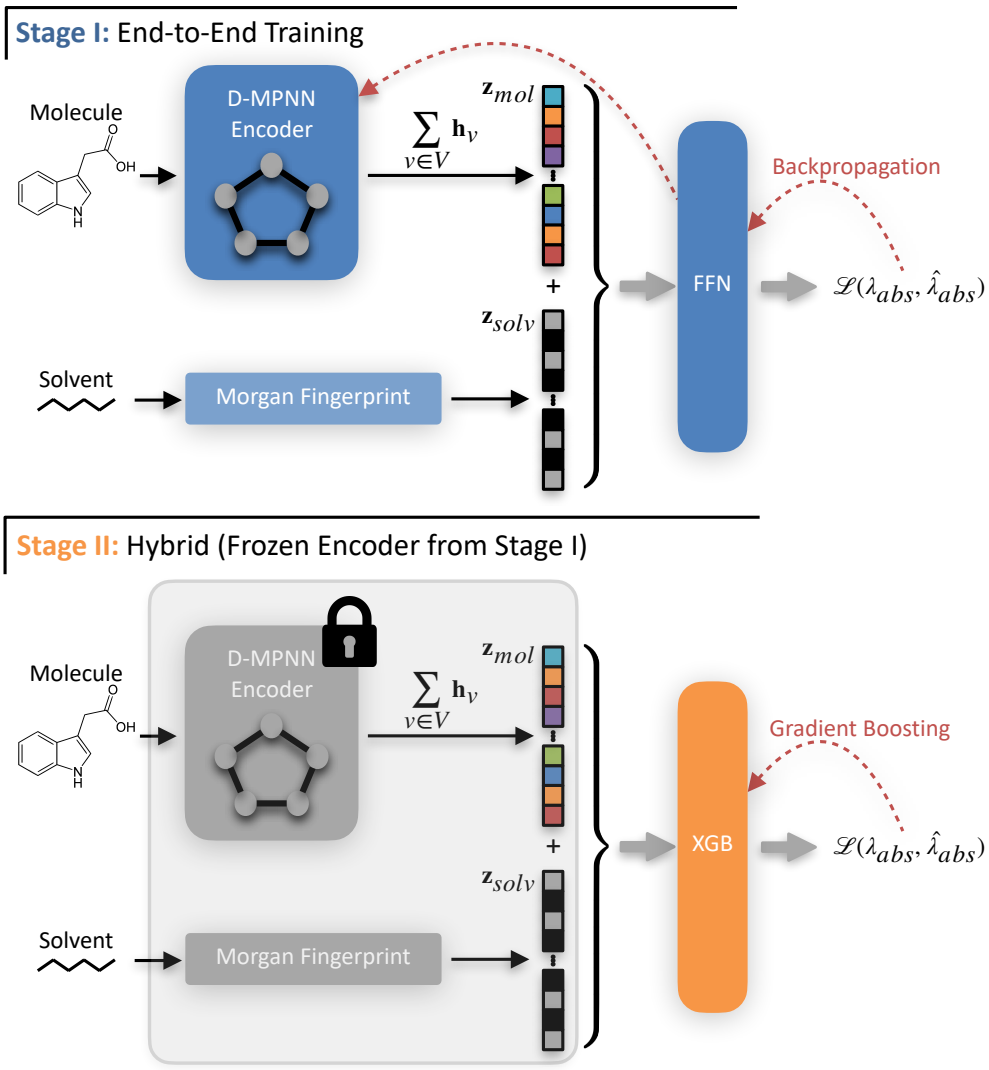

Figure 4.1: Two-stage hybrid modelling framework. Stage I (Top): A D-MPNN encoder maps the molecular graph to per-atom hidden states, which are aggregated via sum pooling to form a molecular representation ($\mathbf{z}_{\text{mol}}$). This representation is concatenated with fixed solvent Morgan fingerprints ($\mathbf{z}_{\text{sol}}$) and used to train a feedforward neural network (FFN) in an end-to-end manner, with gradients of the training loss propagated through both the FFN and encoder. **Stage II (Bottom):** The pretrained D-MPNN encoder from Stage I is frozen and reused to generate molecular representations ($\mathbf{z}_{\text{mol}}$), which are again concatenated with solvent fingerprints ($\mathbf{z}_{\text{sol}}$) and used as fixed features for an XGBoost head.

4.2.2 End-to-End and Hybrid Modelling Paradigms

We consider two modelling paradigms built upon the same D-MPNN encoder ϕ_θ with parameters θ, depicted in Figure 4.1.

In the end-to-end paradigm, the molecular embedding $\mathbf{z}_{\text{mol}} = \phi_\theta(G)$ and solvent embedding $\mathbf{z}_{\text{sol}}$ are concatenated and passed to a feedforward neural network (FFN) head $g_\psi(\cdot)$, and all parameters $\{\theta, \psi\}$ are optimized jointly:

$$\hat{y}_{\text{E2E}} = g_\psi(\mathbf{z}) = g_\psi(\phi_\theta(G), \mathbf{z}_{\text{sol}}). \tag{4.6}$$

Given training data $\{(G_i, \mathbf{z}_{\text{sol},i}, y_i)\}_{i=1}^N$, parameters are learned by minimizing the empirical risk

$$\mathcal{L}(\theta, \psi) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, g_\psi(\phi_\theta(G_i), \mathbf{z}_{\text{sol},i}))\,, \tag{4.7}$$

where $\ell(\cdot, \cdot)$ denotes the task-specific loss function (mean squared error for regression in our experiments). Gradients of $\mathcal{L}$ are computed with respect to both encoder parameters θ and head parameters ψ.

This approach allows the learned representation to adapt directly to the prediction task but may introduce instability in low-data regimes due to the high-dimensional parameter space.

In the hybrid paradigm, we train the full encoder–head model, $g_\psi(\phi_\theta(G), \mathbf{z}_{\text{sol}})$, end-to-end on the downstream task. After learning the parameters, we discard the FFN head $g_\psi(\cdot)$ and freeze the encoder $\phi_{\theta^*}(G)$, where θ^* indicates that the parameters are frozen. Predictions are then generated by a separate regression head h_ω:

$$\hat{y}_{\text{hybrid}} = h_\omega(\phi_{\theta^*}(G), \mathbf{z}_{\text{sol}}), \tag{4.8}$$

where only the head parameters ω are optimized in the second-stage regression step. Decoupling representation learning from regression enables the use of models with different

inductive biases, such as gradient-boosted trees, while ensuring that all comparisons are performed on the same frozen molecular embeddings augmented with identical solvent covariates.

4.2.3 Downstream Regression via Gradient Boosting

In our hybrid framework, we utilize XGBoost [12] to serve as a regularized non-linear head for the frozen D-MPNN embeddings. Unlike the multi-layer perceptrons typically used in end-to-end GNNs, XGBoost constructs an additive ensemble of decision trees that incorporates explicit regularization on tree complexity (e.g., minimum split loss (γ) and leaf weights (λ)). This structural regularization is particularly advantageous in the data-limited regimes characteristic of experimental photophysics, where overparameterized neural heads are prone to overfitting.

4.2.4 Baseline Models

We evaluated several baseline approaches common to molecular property prediction. All baseline models were trained and evaluated using the same canonical data splits, fixed training subsamples, and evaluation protocols as the D-MPNN-based models.

Fingerprint-based baselines. As a classical cheminformatics baseline, molecules were represented using fixed-length Morgan fingerprints (radius 2, 2048 bits) computed with RDKit. Solvents were represented using separate Morgan fingerprints (radius 2, 1024 bits) which were concatenated with the molecular fingerprints. This concatenation strategy mirrors the solvent integration used in the D-MPNN and hybrid architectures, ensuring that all models have access to equivalent environmental context. To capture global physicochemical properties often missed by substructure fingerprints, we additionally explored augmenting these vectors with nine scalar descriptors: molecular weight (MolWt), octanol–water partition coefficient (MolLogP), topological polar surface area (TPSA), hydrogen bond acceptor and

donor counts (HBA/HBD), aromatic and aliphatic ring counts, rotatable bond count, and the fraction of sp^3-hybridized carbon atoms (f_{sp^3}). These tabular representations were used to train a suite of regression models: ridge and polynomial regressions, Random Forests, and XGBoost models.

Physics-informed baselines. We evaluated the predictive utility of low-fidelity quantum-chemical (QC) descriptors derived from density functional theory calculations reported by Greenman *et al.* [25]. In that work, molecular geometries were generated from SMILES strings and optimized using a multistage pipeline including semi-empirical tight-binding (GFN2-xTB), followed by DFT geometry optimization at the BP86-D3/def2-SVP level. TD-DFT calculations were then performed using the Tamm–Dancoff approximation at the ωB97X-D3/def2-SVPD level of theory to estimate vertical excitation energies. For a subset of dye–solvent pairs, additional solvent-corrected TD-DFT calculations were carried out using the IEFPCM continuum solvation model.

While these descriptors provide valuable physical signal, they require nontrivial QC calculations and are therefore substantially more expensive than purely data-driven representations. Accordingly, these baselines are included to contextualize the performance of learned representations relative to low-fidelity physics, rather than as a practical alternative in large-scale or low-latency settings.

Motivated by Tannir *et al.* [53], we utilize the resulting HOMO–LUMO gap estimates as physics-informed descriptors. For molecules where this information was not available ($\leq 2\%$), we employ median imputation. These QC features were combined with fingerprint-based representations and used as inputs to the same downstream regression models, enabling direct comparison between learned representations, purely empirical features, and hybrid physics-informed baselines.

Finally, we evaluated a minimal parametric model motivated by Frontier Molecular Orbital (FMO) theory. Approximating the dominant optical transition energy by the energy

gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), and invoking the Planck–Einstein relation ($E = hc/\lambda$), we fit an inverse-gap regression of the form:

$$\lambda_{\max} \approx \beta_0 + \beta_1(\Delta\epsilon_{\text{H–L}})^{-1}, \tag{4.9}$$

to establish a predictive baseline rooted in first-principles physics. While this approximation neglects vibronic structure, excited-state relaxation, and explicit solvent reorganization, it provides a useful low-capacity benchmark against which more expressive learned representations can be evaluated.

For standard fingerprint- and descriptor-based baselines, we performed hyperparameter optimization using Optuna, utilizing the validation set for model selection. In contrast, the DFT-informed baselines were trained using default hyperparameters; consequently, we combined the training and validation sets for these models to maximize data utilization during fitting. Regardless of the training protocol, the test set was strictly reserved for final evaluation to ensure fair comparison across all architectures.

4.2.5 Variance Decomposition Framework

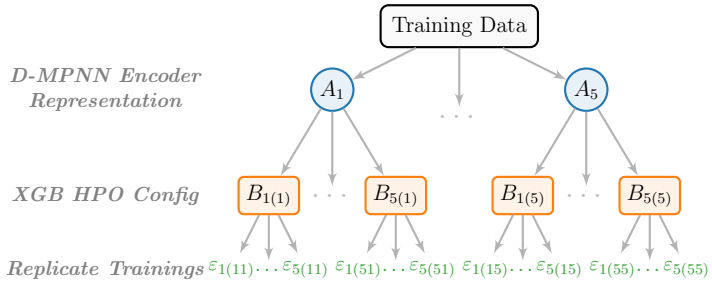

Figure 4.2: Nested variance decomposition of predictive performance. Schematic of the two-factor experimental design together with estimated variance components.

To quantify the sources of predictive variability in the hybrid encoder–head architecture, we adopt a two-factor nested random-effects variance decomposition. This framework

partitions total performance variability into contributions arising from stochasticity in representation learning (encoder training), stochasticity in model selection at the head level (Bayesian hyperparameter optimization for XGBoost), and residual stochasticity associated with downstream regression given fixed hyperparameters.

We employ a balanced nested experimental design. Specifically, we train $I = 5$ independent D-MPNN encoders with different random initializations. For each frozen encoder, we perform $J = 5$ independent Bayesian hyperparameter optimizations for the XGBoost head, yielding distinct heads conditioned on that encoder. Finally, for each encoder–head pair, we train and evaluate the XGBoost model across $K = 5$ random seeds using the fixed hyperparameter set, capturing stochasticity arising from subsampling and tree construction. This design yields a total of $N = I \times J \times K = 125$ evaluations per dataset split.

Let y_{ijk} denote the test-set RMSE obtained from the i-th encoder initialization ($i = 1, \ldots, I$), the j-th regression head configuration nested within that encoder ($j = 1, \ldots, J$), and the k-th replicate seed ($k = 1, \ldots, K$). We model performance as

$$y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk}, \tag{4.10}$$

where $A_i \sim \mathcal{N}(0, \sigma_A^2)$ captures encoder-level variability due to stochastic representation learning, $B_{j(i)} \sim \mathcal{N}(0, \sigma_{B|A}^2)$ captures variability induced by hyperparameter selection for the XGBoost head conditioned on a fixed encoder, and $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ represents residual variability associated with stochastic optimization and tree construction given fixed hyperparameters.

Variance components $(\sigma_A^2, \sigma_{B|A}^2, \sigma_\varepsilon^2)$ were estimated using the ANOVA Method of Moments estimator for balanced designs (see Supporting Information for exact Expected Mean Squares formulas) [50]. From these estimates, the proportion of variance attributable to the

learned representation is computed as:

$$\eta^2_{encoder} = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_{B|A} + \sigma^2_\varepsilon}. \tag{4.11}$$

4.2.6 Ensemble Modelling

Following Greenman *et al.* [25] we employ ensemble predictors to improve predictive stability and mitigate stochasticity arising from random initialization and training dynamics. Ensemble predictions are formed by averaging the outputs of M independently trained models:

$$\hat{y}_{\text{ens}} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}^{(m)}, \tag{4.12}$$

where each ensemble member $\hat{y}^{(m)}$ corresponds to a model trained with a distinct random seed.

In the hybrid paradigm, ensemble members correspond to independently trained encoder–head pipelines. In the end-to-end paradigm, ensemble members correspond to independently trained joint encoder–head models. Unless otherwise stated, ensemble predictions are used for all reported results.

4.2.7 Transfer Learning and Fine-Tuning Protocols

To evaluate the utility of learned representations in data-scarce regimes, we compare four distinct initialization and adaptation strategies. We report results for the following protocols:

1. **Random Initialization (Baseline):** The D-MPNN encoder and prediction head are initialized randomly and trained end-to-end on the downstream dataset.

2. **Zero-Shot Transfer:** The encoder and prediction head, pretrained on the largest dataset, are applied directly to the downstream dataset without any parameter up-

dates. This serves as a reference for the off-the-shelf generalizability of the source model.

3. **Two-Stage Fine-Tuning:** The encoder is initialized from weights pretrained on the source task and adapted to the target using a gradual unfreezing schedule.

4. **Hybrid on Fine-Tuned Embeddings:** The D-MPNN is first fine-tuned via the two-stage protocol. The resulting adapted encoder is then frozen, and a gradient-boosted tree (XGBoost) head is trained on the extracted embeddings.

Two-Stage Fine-Tuning Schedule. For the fine-tuning strategies, immediate end-to-end optimization was found to cause training instability and catastrophic forgetting in ultra-low-data regimes ($N < 100$). To mitigate this, we employ a two-stage "gradual unfreezing" schedule.

Let θ denote encoder parameters and ψ denote head parameters. Recall that $\mathcal{L}(\theta, \psi)$ denotes the empirical risk, which in this case can be written as

$$\mathcal{L}(\theta, \psi) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, g_\psi(\phi_\theta(G_i), \mathbf{z}_{\text{sol},i})) \,.$$

Starting from a pretrained checkpoint (θ_0, ψ_0), we first train only the head while the encoder remains frozen:

$$\textbf{Stage A (Warm-Up):} \quad \theta \leftarrow \theta_0, \qquad \psi \leftarrow \arg\min_{\psi} \mathcal{L}(\theta_0, \psi), \tag{4.13}$$

for E_A epochs using a learning rate scaled by α_A. This allows the head to align with the existing feature space before modifying the representation itself.

In the second stage, we unfreeze the encoder and fine-tune the full model end-to-end:

$$\begin{aligned} \textbf{Stage B (End-to-End):} \quad & (\theta, \psi) \leftarrow (\theta_A, \psi_A), \\ & (\theta, \psi) \leftarrow \arg\min_{\theta,\psi} \mathcal{L}(\theta, \psi), \end{aligned} \tag{4.14}$$

for E_B additional epochs using a reduced learning rate scale $\alpha_B \ll \alpha_A$. Unless otherwise noted, we use $E_A = 30$ and $E_B = 170$.

4.2.8 Analysis of Learned Molecular Representations

To analyze the structure of the learned molecular embeddings, we perform principal component analysis (PCA) on the frozen D-MPNN encoder outputs. PCA provides an orthogonal basis that captures dominant modes of variation in the learned embedding space while remaining invariant to rotations of the original coordinate system, which is essential given the non-identifiability of latent embedding axes across training runs.

All analyses are conducted on post hoc projections of the learned embeddings and do not influence model training. To account for stochastic variation across training seeds, PCA is performed independently for each encoder instance, followed by sign alignment of principal axes based on their correlation with an external physical reference quantity. This enables consistent aggregation and comparison of latent directions across seeds.

To probe chemical interpretability, aligned principal components are correlated with the independently computed physicochemical descriptors in Table 4.1 and target properties using Spearman rank correlation. These descriptors are not provided to the model during training and are used solely for interpretive analysis of the learned representation.

4.3 Experimental Setup

This section describes the datasets, targets, data splitting strategies, training protocols, and evaluation procedures used in all experiments.

Table 4.1: Physicochemical descriptors computed via RDKit for latent space correlation analysis. These features were not used during model training but serve as independent probes for interpreting the learned representations.

Descriptor	Description
NumAromaticRings	Number of aromatic ring systems
FractionCSP3	Fraction of sp^3-hybridized carbon atoms
NumConjugatedBonds	Total number of conjugated bonds
LongestConjChain	Diameter (in bonds) of the conjugated-bond subgraph
TPSA	Topological polar surface area
NumHBA	Number of hydrogen bond acceptors
NumHBD	Number of hydrogen bond donors
ExactMolWt	Exact molecular weight
HeavyAtomCount	Number of non-hydrogen atoms
NumRotatableBonds	Number of rotatable single bonds
NumEtherO	Number of ether oxygen atoms
NumMethoxy	Number of methoxy substituents
NumMethylEster	Number of methyl ester groups

4.3.1 Datasets and Targets

We evaluate all models on multiple datasets of organic chromophores and fluorophores with experimentally measured optical properties. Across all datasets, the prediction targets are peak absorption and/or emission wavelengths reported in nanometers (nm). Each dataset differs in size, chemical diversity, and application domain, enabling systematic evaluation across data regimes.

The largest dataset, Deep4Chem [31], contains experimentally measured absorption and emission wavelengths for a chemically diverse collection of organic molecules compiled from the literature. To probe transfer learning and robustness under reduced data availability, we additionally consider smaller, domain-specific datasets, including the Dye-Sensitized Solar Cell Database (DSSCDB) [57], the ChemFluor dataset [32], and the Jeffries-EL dataset [53] of blue-emitting benzobisoxazole-based OLEDs. The Jeffries-EL dataset is substantially smaller and is primarily used as a qualitative case study rather than for full model benchmarking.

To provide a qualitative overview of the chemical diversity and overlap across datasets, we visualize the global organization of molecular space using a two-dimensional Uniform Manifold Approximation and Projection (UMAP) embedding (Figure 4.3). Each point corresponds to a molecule, coloured by dataset.

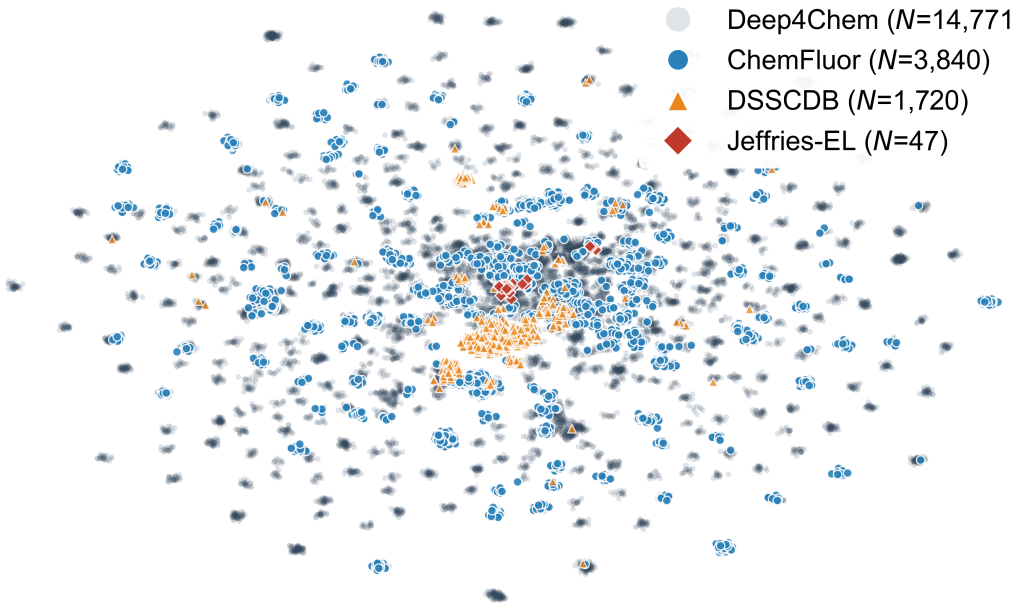

Figure 4.3: Qualitative comparison of dataset chemical space. Two-dimensional UMAP projection of molecules from all datasets computed from 2048-bit Morgan fingerprints (radius 2) using the Jaccard distance. To reduce overplotting due to repeated structures (e.g., identical fingerprints across solvents), points are displayed with a small random jitter.

All datasets were preprocessed following the protocol of Greenman *et al.* [25] without modification. Molecular structures were canonicalized using RDKit, and only solution-phase measurements were retained; solid-state measurements were excluded. For molecules with multiple reported measurements in the same solvent, entries were filtered based on consistency: measurements differing by more than 5 nm were discarded, while remaining duplicate measurements were averaged. Measurements corresponding to the same molecular structure in different solvents were treated as distinct datapoints through solvent-specific feature representations.

For datasets containing both absorption and emission measurements, the number of usable datapoints differs between targets due to incomplete experimental coverage. All analyses were therefore conducted on target-specific subsets derived from the same underlying molecular pool. Unless otherwise noted, dataset sizes N refer to datapoints rather than unique molecules. A summary of dataset sizes and target availability is provided in Table 4.2.

Table 4.2: Summary of datasets used for optical property prediction after preprocessing. Counts are reported as Absorption / Emission where applicable.

Dataset	Target(s)	Datapoints	Unique Molecules
Deep4Chem	Abs., Emis.	14,771 / 14,378	5,841 / 5,597
DSSCDB	Abs., Emis.	1,720 / 690	1,647 / 674
ChemFluor	Abs., Emis.	3,840 / 3,953	2,622 / 2,617
Jeffries-EL	Emis.	47	47

4.3.2 Data Splitting and Training Regimes

To assess generalization under both interpolation and distribution shift, we employ two splitting strategies. Random splits are used to approximate in-distribution generalization, while Bemis–Murcko scaffold splits are used to evaluate extrapolation to unseen chemical scaffolds. We adopt the same random and Bemis–Murcko scaffold splitting procedures as Greenman *et al.* [25] and verify split equivalence by matching scaffold assignments and split indices. To quantify the effective degree of distribution shift induced by each splitting strategy, we compute the maximum test-to-train Tanimoto similarity for all test molecules under both random and scaffold splits (Figure 4.4). As expected, scaffold splitting substantially reduces train–test chemical overlap for Deep4Chem and ChemFluor, while DSSCDB exhibits comparatively high similarity even under scaffold splits, consistent with its narrower chemical diversity.

Fixed subsampling for data-regime experiments. To evaluate performance as a function of training set size N, we constructed fixed subsampled training sets from each canonical

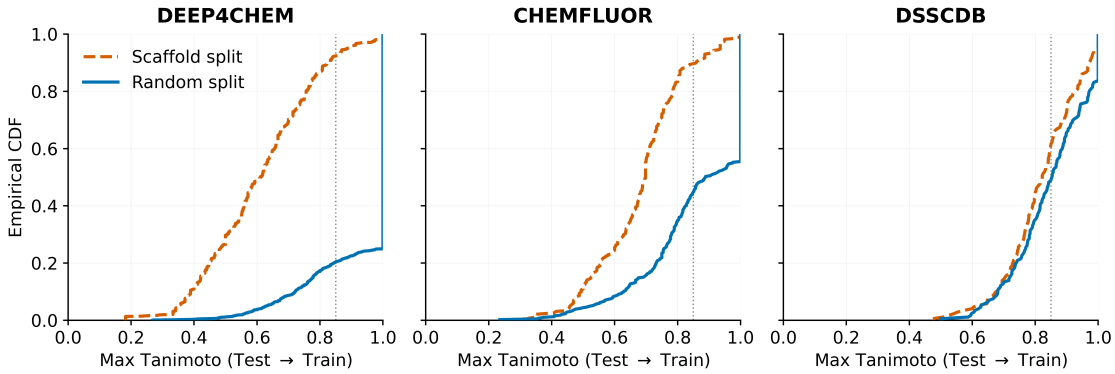

Figure 4.4: Train–test chemical similarity under random and scaffold splits. Empirical cumulative distribution functions (CDFs) of the maximum Tanimoto similarity between each test molecule and the training set, computed using Morgan fingerprints, for random and Bemis–Murcko scaffold splits across Deep4Chem, ChemFluor, and DSSCDB. Vertical dotted lines indicate a commonly used similarity threshold ($T = 0.85$), above which molecules are typically considered highly similar. Scaffold splits substantially reduce train–test chemical overlap for Deep4Chem and ChemFluor, while DSSCDB exhibits higher intrinsic similarity across splits, reflecting its more chemically homogeneous composition.

split. For a given dataset, target, and split type, the original validation and test sets were held fixed, and training subsets were generated by sampling without replacement from the canonical training pool at sizes $N \in \{50, 100, 250, 1000, 4000\}$, as well as the full training set where available. For datasets with smaller training pools, only the feasible subset sizes were considered. Subsampling was performed using a fixed random seed, and the same subsampled training sets were reused across all model architectures and random initialization runs to ensure that differences across N reflect model behaviour rather than variation in the sampled training examples. For hybrid models, downstream heads were trained on embeddings extracted from the corresponding subsampled training set, with validation and test evaluation performed on the fixed canonical splits.

4.3.3 Model Training Protocols

End-to-end D-MPNN training. For each data regime, we conducted a 200-trial Optuna Bayesian hyperparameter search over architectural and optimization parameters, including network depth, FFN capacity, dropout rate, batch size, and learning-rate schedule. Optimization was performed using the Adam optimizer with a learning-rate warmup schedule and mean squared error (MSE) loss. We fixed the hidden dimension of the encoder at 482, consistent with Greenman *et al.* [25] to ensure a consistent representational capacity across all data regimes and to mitigate numerical instability observed in ultra-small N settings when the hidden dimension was allowed to vary.

Unless explicitly tuned, all remaining parameters were set to Chemprop defaults. Models were trained for a fixed 200 epochs without early stopping. This protocol was chosen to strictly isolate the impact of dataset size on training dynamics, avoiding the confounding influence of varying stopping criteria across data regimes. For each data regime, the optimal hyperparameter configuration identified by Optuna was reused across all ensemble members. To verify that fixed training horizons do not induce late-stage overfitting, we inspected training and validation trajectories across representative small- and large-data regimes and observed stable convergence without divergence in validation error (see Figure S1).

Hybrid model training. For the hybrid paradigm, D-MPNN encoders were pretrained as described above, then frozen. The downstream XGBoost heads were tuned via a separate 100-trial Optuna search for each N, optimizing tree depth, learning rate (η), and regularization terms (λ, α).

Ensembling and Reproducibility. To mitigate stochasticity arising from random initialization and training dynamics, all reported performance metrics are based on ensemble predictions unless otherwise noted.

- **Deep and Hybrid Models:** Ensembles consist of $M = 5$ independently trained models (distinct random seeds) sharing the same hyperparameter configuration. For hybrid models, this involves pairing 5 independently trained encoders with 5 corresponding downstream heads.

- **Gradient-Boosted Baselines:** Due to their low computational cost, tree-based baselines are evaluated using larger ensembles of $M = 30$ models.

4.3.4 Evaluation Metrics

Model performance is evaluated using Root Mean Squared Error (RMSE) computed on the held-out test set. All reported results correspond to the RMSE of the ensemble-averaged prediction. Where shown, error bars denote 95% percentile bootstrap confidence intervals for the test RMSE, computed by resampling test molecules with replacement (2000 replicates). These intervals reflect sampling variability of the evaluation set and do not quantify variability across random initializations or ensemble members.

All deep learning models were implemented using PyTorch (v1.12) within the Chemprop framework (v1.6.1) [63]. Chemical informatics operations and featurization were performed using RDKit (v2022.03.1), and gradient-boosted tree models were implemented using XGBoost (v1.6.2). Training and evaluation were conducted on NVIDIA H100 GPUs.

To ensure reproducibility, all code, canonical data splits, and trained model checkpoints used in this thesis are available at `https://github.com/michaelmontemurri/uvvis-hybrid-gnn`.

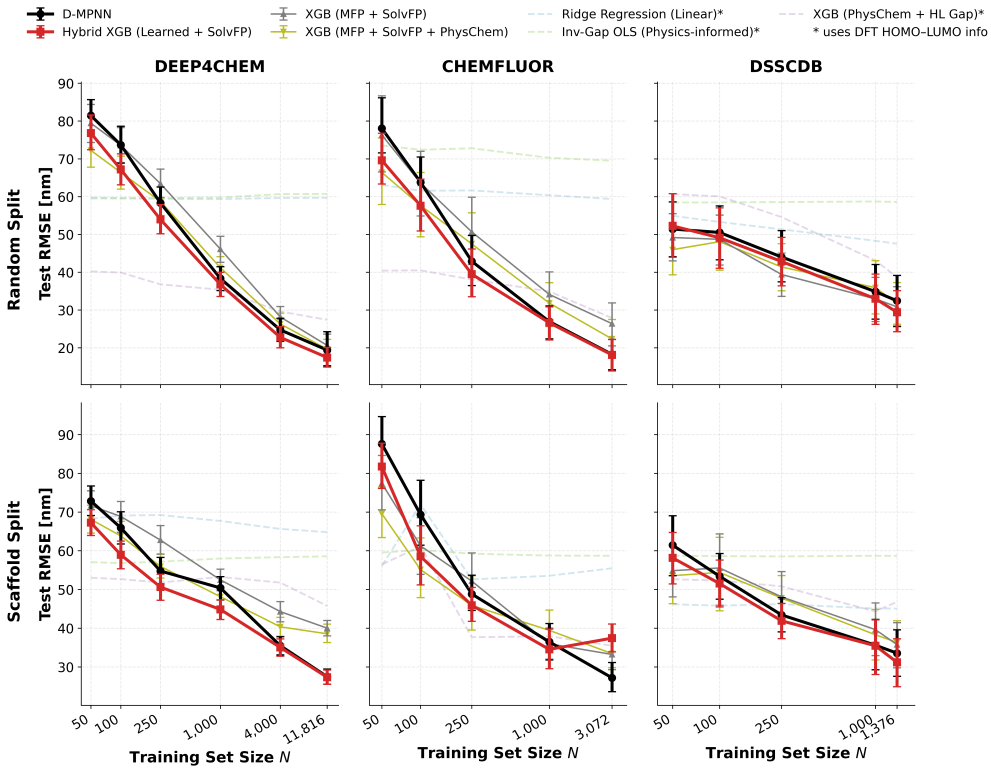

Figure 4.5: Peak absorption wavelength ($\lambda_{\max}$) prediction across Deep4Chem, ChemFluor, and DSSCDB. Test RMSE versus training set size N for random (top row) and Bemis–Murcko scaffold (bottom row) splits. Each column corresponds to a dataset and uses dataset-specific N values. Error bars indicate 95% percentile bootstrap confidence intervals for the test RMSE, computed by resampling test molecules with replacement (2000 replicates). Models marked with an asterisk incorporate TD-DFT–derived HOMO–LUMO gap information.

4.4 Results and Discussion

4.4.1 Absorption Wavelength Prediction Across Data Regimes

Consistent hybrid advantage in low-to-mid data regimes. Figure 4.5 reports test RMSE as a function of training set size N for peak absorption wavelength ($\lambda_{\max}$) prediction. Across datasets and split types, the hybrid D-MPNN→XGBoost model generally outperforms the end-to-end D-MPNN in low-to-mid data regimes ($N \leq 1000$), with the strongest gains appearing under scaffold splits and in the smaller training-size settings. The mag-

nitude of this effect is summarized in Table 4.3; for Deep4Chem, we observe an average RMSE reduction of **6.5% for random splits** and **9.2% for scaffold splits** within the $50 \leq N \leq 1000$ range.

We attribute this overall advantage to the structural regularization of gradient-boosted trees. In data-limited regimes, high-dimensional FFN heads are prone to high-variance estimation and overfitting. In contrast, the tree-based structure of XGBoost enforces simpler decision boundaries within the learned embedding space, acting as a robust regularization prior. This effect is most pronounced under scaffold splits (Deep4Chem), where the distribution shift is largest (see Figure 4.4). As data availability increases ($N > 1000$) or distribution shift decreases (DSSCDB), the end-to-end model receives sufficient signal to stabilize the head, and the performance gap closes.

Table 4.3: Percentage RMSE improvement of the Hybrid model (frozen D-MPNN embeddings → XGBoost) relative to the end-to-end D-MPNN. Positive values indicate lower RMSE for the Hybrid model.

	DEEP4CHEM		**CHEMFLUOR**		**DSSCDB**	
N	Rand (%)	Scaf (%)	Rand (%)	Scaf (%)	Rand (%)	Scaf (%)
50	**+5.7**	**+7.7**	**+10.8**	**+9.8**	-1.6	**+5.4**
100	**+8.8**	**+10.6**	**+9.8**	**+15.6**	**+2.8**	**+3.5**
250	**+7.5**	**+7.6**	**+8.0**	**+6.0**	**+2.7**	**+3.5**
1000	**+4.1**	**+11.0**	+1.2	**+5.4**	**+5.4**	+0.4
Max $N^{\dagger}$	**+8.0**	+1.3	+0.8	-28.8	**+9.3**	**+7.0**

† Max N: DEEP4CHEM (11,816), CHEMFLUOR (3,072), DSSCDB (1,376).

Benchmarking against physical and classical baselines. Fingerprint-based baselines exhibit the expected hierarchy: augmenting Morgan fingerprints with physicochemical descriptors improves performance, but both graph-based models (End-to-End and Hybrid) generally outperform these fixed-feature approaches once $N \geq 250$.

In the ultra-low data regime ($N \leq 100$), physics-informed baselines utilizing TD-DFT–derived HOMO–LUMO gaps (dashed lines) remain the most robust, particularly under scaf-

fold splits. This highlights the necessity of strong physical inductive biases when experimental supervision is scarce. However, as N increases, the D-MPNN-based models rapidly overtake the DFT-augmented baselines. This crossover indicates that the learned graph representations eventually capture higher-order structural information that exceeds the predictive content of a single vacuum-phase orbital gap. The graph models achieve this accuracy without the computational overhead of calculating DFT descriptors at inference time.

Comparison to Multi-Fidelity Frameworks. We benchmark our approach against the multi-fidelity D-MPNN ensemble of Greenman *et al.* [25]. Using identical preprocessing and scaffold splits on Deep4Chem, our Hybrid model achieves a test RMSE (27.5 nm) comparable to their reported ChempropMultiFidelity result (27.47 nm).

The Greenman framework relies on auxiliary pretraining with TD-DFT excitation energies. Generating this auxiliary data for the ∼5,800 molecules in Deep4Chem requires an estimated 10^5 core-hours of quantum-chemical calculations (RDKit → xTB → DFT → TD-DFT) (K. Greenman, personal communication, Feb 11, 2026). Our results demonstrate that such expensive pretraining may not be necessary for comparable performance on this benchmark. By using a hybrid architecture with implicit regularization, we achieve comparable accuracy using purely experimental supervision, without the need for a quantum-chemical pipeline.

Summary: Regime-dependent tradeoffs and practical workflows. These results reveal a clear regime-dependent tradeoff. In the data-scarce limits ($N \lesssim 1000$) typical of experimental screening, the hybrid architecture offers meaningful accuracy gains; as dataset size increases, this advantage diminishes and fully joint optimization becomes competitive. The computational cost of fitting an XGBoost head on fixed embeddings is negligible compared to the cost of training the D-MPNN encoder itself. Consequently, given a trained D-MPNN, practitioners can extract the learned embeddings and fit a lightweight XGBoost head with virtually no overhead. This ensures optimal performance in the low-data limit

without sacrificing the scalability of the end-to-end model as data grows. We note that this hybrid advantage is most pronounced for absorption; as detailed in the Supporting Information, results for emission prediction were less systematic, consistent with the greater physical complexity of excited-state relaxation and solvent reorganization effects, which are harder to capture in a static representation.

4.4.2 Variance Decomposition

Figure 4.5 shows that model hybridization is most beneficial in low-N regimes; we next ask which sources of stochasticity drive this behaviour. We use the variance decomposition framework introduced in Methods to identify which sources of stochasticity dominate predictive performance across data regimes, and to explain the observed convergence of hybrid and end-to-end models at scale.

Head sensitivity and selection. To isolate variability attributable solely to the regression head, we evaluated three models (MLP, RF, and XGBoost) on a fixed, pretrained D-MPNN encoder. As shown in the Supporting Information (Figure S2), XGBoost typically provided the strongest mean performance with minimal variance, motivating its adoption as the standard head for the hybrid architecture used throughout this work.

Regime-dependent variance decomposition. Figure 4.6 presents the variance decomposition on Deep4Chem as a function of training set size. In small-data regimes ($N \leq 250$), a substantial proportion of predictive variability ($\sim$20–30%) is attributable to the downstream head. This implies that under limited supervision, the mapping from molecular representation to property is underdetermined; the choice of regression model and its optimization significantly impacts generalization.

As N increases, the contribution of the head diminishes, and predictive variability becomes dominated almost entirely by stochasticity in encoder training. This explains the convergence of hybrid and end-to-end models at scale: once the dataset is large enough to

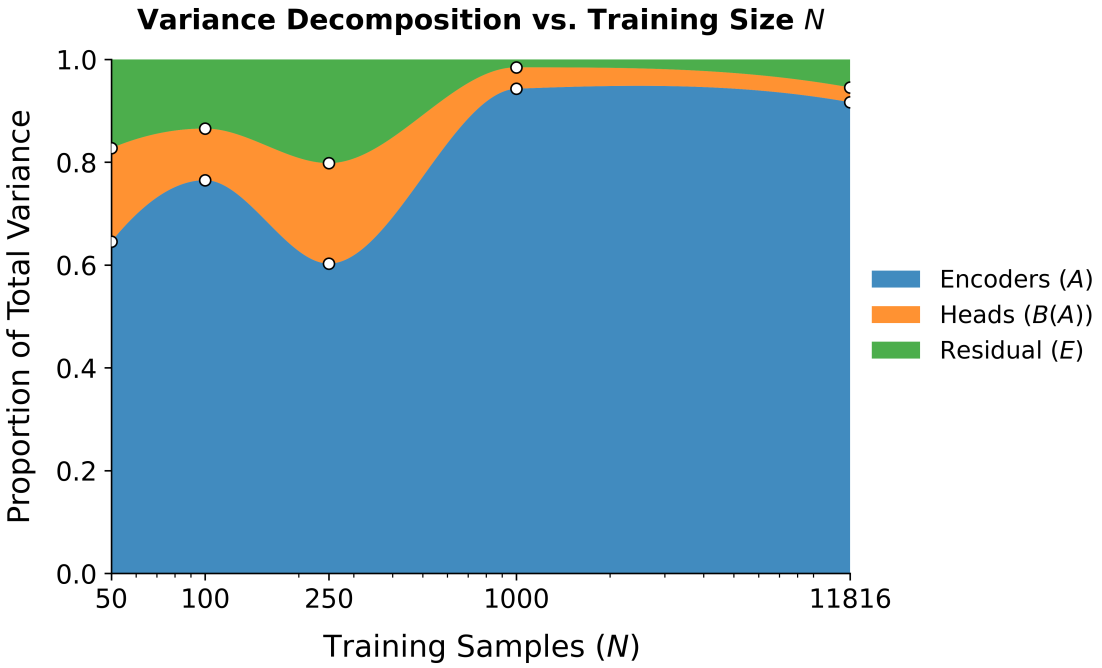

Figure 4.6: Proportion of total test-set RMSE variance attributable to encoder-level stochasticity, head-level variability, and residual noise. The breakdown is shown as a function of training set size N on Deep4Chem, with variance proportions estimated using a two-factor nested random-effects decomposition.

constrain the representation, the specific choice of head becomes less critical. The hybrid advantage in low-data regimes, therefore, stems from XGBoost's ability to efficiently navigate the high-variance "head" component of the error budget.

Uncertainty estimates for the variance components, obtained via bootstrap resampling, are reported in the Supporting Information (Table S3); while confidence intervals are wide at small training set sizes, the qualitative dominance of encoder-level stochasticity is preserved across regimes.

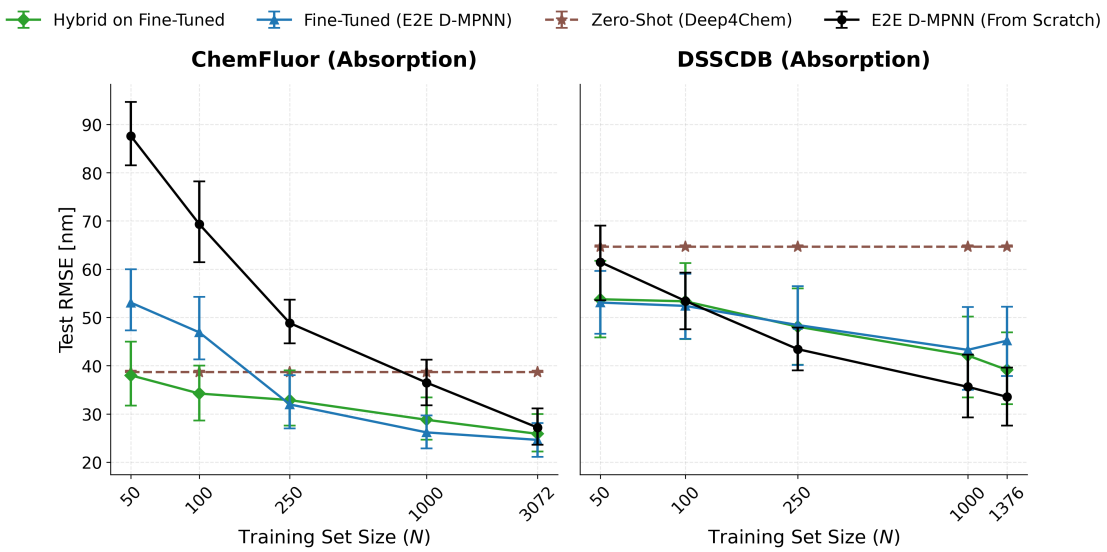

Figure 4.7: Regime-dependent benefits of transfer learning. Test RMSE (scaffold split) versus training set size N for **(A) ChemFluor** and **(B) DSSCDB**. Strategies include training from scratch (black circles), two-stage fine-tuning of a Deep4Chem-pretrained encoder (blue triangles), and hybrid XGBoost head applied to fine-tuned embeddings (green diamonds). The zero-shot performance of the Deep4Chem model (brown stars) is provided as a baseline. Error bars indicate 95% percentile bootstrap confidence intervals for the test RMSE, computed by resampling test molecules with replacement (2000 replicates).

4.5 Transfer Learning and Pretraining Effects

We next evaluate whether pretraining on the large Deep4Chem corpus can mitigate data scarcity in smaller, domain-specific tasks. We compare four strategies defined in Methods: Random Initialization, Zero-Shot Pretrained, Two-Stage Fine-Tuning, and Hybrid Transfer.

Transfer to ChemFluor (High Chemical Overlap). Figure 4.7A demonstrates that for ChemFluor, which shares significant structural similarity with Deep4Chem, transfer learning appears particularly effective. Fine-tuning a pretrained encoder yields substantial RMSE reductions over training from scratch in the $N \leq 1000$ regime. Notably, the Hybrid Transfer strategy on these fine-tuned embeddings provides even further improvement. This confirms

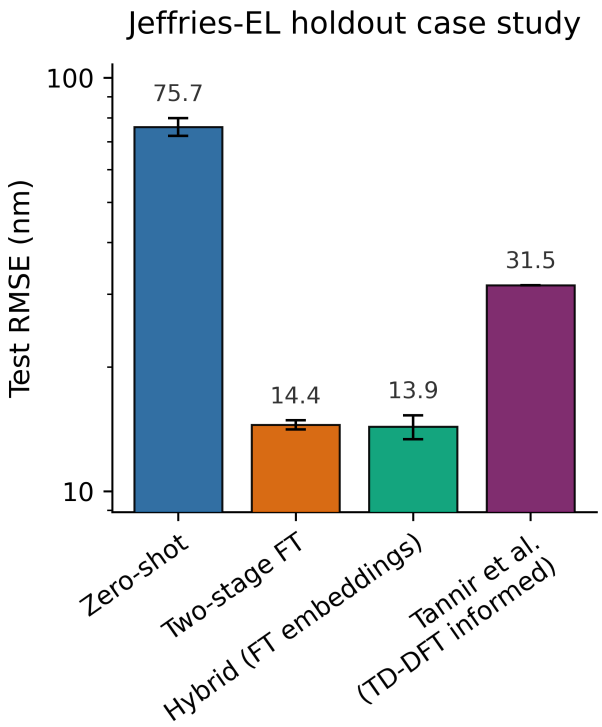

Figure 4.8: Jeffries-EL Case Study ($N = 43$). Test RMSE on the holdout set. The Hybrid model trained on fine-tuned embeddings achieves the lowest error, outperforming both standard fine-tuning and domain-specific physics-informed baselines (Tannir *et al.*).

that even when transfer is successful, the regularization benefits of the hybrid model persist in the ultra-low-data limit.

Transfer to DSSCDB (Low Chemical Overlap). In contrast, transfer to the DSSCDB dataset (Figure 4.7B) illustrates the limitations of pretraining under distribution shift. DSSCDB contains distinct dye classes poorly represented in Deep4Chem. Consequently, the zero-shot baseline performs poorly, and fine-tuning offers only negligible gains at $N = 50$. As data availability increases, models trained from scratch rapidly surpass the transfer-learning models. This likely reflects representational bias from pretraining.

Case Study: Jeffries-EL (Ultra-Small N). Based on the transfer behaviours observed on ChemFluor and DSSCDB, we hypothesized that the Jeffries-EL dataset ($N = 43$) would

benefit significantly from pretraining due to its chemical alignment with Deep4Chem. Moreover, we anticipated that the hybrid architecture would further enhance performance by stabilizing the head in this ultra-low-data limit. The results are consistent with these expectations. Figure 4.8 shows that the hybrid transfer model achieves the lowest test RMSE, outperforming both standard fine-tuning and the domain-specific TD-DFT-informed baselines reported by Tannir *et al.* on the exact same holdout set [53]. This reinforces our earlier findings: when downstream chemistry aligns with the pretraining corpus, hybrid transfer learning can leverage prior knowledge to outperform even the physics-based descriptors.

4.5.1 Interpretation of Learned Molecular Representations

The D-MPNN encoder serves as a nonlinear feature extractor, mapping complex molecular graphs into high-dimensional latent representations. While such high-dimensional vectors are necessary to provide the representational capacity required for state-of-the-art accuracy, they risk becoming "black boxes" that obscure the underlying chemical logic [23, 63]. A critical question for any scientific ML model, particularly under data scarcity, is whether it learns physically meaningful features or merely exploits dataset statistics. To address this, we analyze the structure of the learned D-MPNN embeddings on the Deep4Chem and ChemFluor datasets using PCA and interpretable descriptors.

Stability of the Latent Space. Despite the stochasticity of training, the D-MPNN tends to converge to a stable low-dimensional manifold across encoder seeds. The PCA reveals that the resulting embedding space is relatively low-dimensional: the top 24 components capture $\sim$90% of the variance in the 482-dimensional embedding space. By aligning the principal components across independent training seeds (using the HOMO–LUMO gap to resolve sign ambiguity), we find that the dominant axis (PC_0) is reproduced with high consistency across runs.

This dominant variance is closely aligned with predictive utility. Five independent auxiliary XGBoost heads trained on the seed-dependent PCA-transformed embeddings consistently identify PC_0 as the most influential feature, accounting for an average of 48.1% of total feature importance, with remaining importance distributed across secondary axes. Complementary SHAP (Shapley Additive Explanations) analysis supports this hierarchy, indicating that the model relies primarily on the PC_0 axis for global spectral shifts while using higher-order components for localized adjustments (see Supporting Information Figure S8).

We observed similar semantic trends across datasets (see Supporting Information Figure S5). This replication supports the claim that the D-MPNN often learns a chemically meaningful organization aligned with known determinants, regardless of stochastic initialization or specific dataset composition.

Correlation with Physical and Chemical Descriptors. To interpret these latent directions, we examine the Spearman rank correlations between the aligned principal components and independent RDKit descriptors (Figure 4.9). These descriptors (Table 4.1) were not provided to the model during training; the observed correlations emerge solely from the model's learned representation of the molecular graph.

Across encoder seeds, the aligned principal components exhibit highly stable correlations with physically meaningful quantities. The dominant component (PC_0) is strongly correlated with both the HOMO–LUMO gap and absorption wavelength ($\rho = -0.820 \pm 0.007$ and $\rho = 0.942 \pm 0.003$, respectively). This correlation is far stronger than that of simple topological descriptors like *NumConjugatedBonds* ($\rho = 0.24$). Full seed-level statistics are reported in Supporting Information Table S4.

4.5.2 Chemical Interpretation of Principal Components

PC_0: The Effective Conjugation Axis. As visualized in Figure 4.10, PC_0 can be interpreted as a learned coordinate for **"effective conjugation length"**. It separates molecules

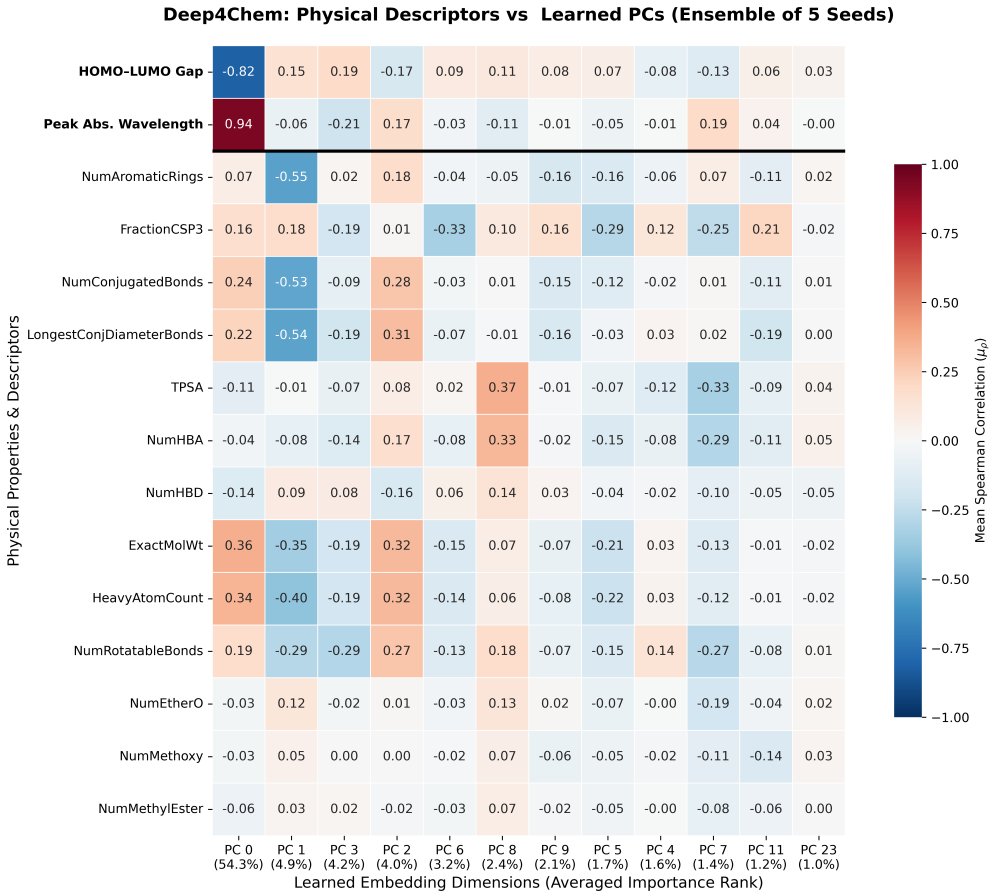

Figure 4.9: Consensus Interpretability Map: Learned PCs vs Physics. Aligned Spearman rank correlations between the top 12 predictive principal components (PCs) and physicochemical descriptors across an ensemble of 5 seeds. PCs are ordered by averaged XGB feature importance. The horizontal divider separates structural descriptors from the photophysical targets, HOMO–LUMO gap and peak absorption wavelength.

not just by the number of double bonds, but by their ability to delocalize electrons. Low PC_0 values correspond to torsionally twisted or interrupted π-systems, while high PC_0 values correspond to extended, planar architectures that promote efficient orbital overlap. The saturation of $\lambda_{\max}$ at high PC_0 values mirrors the physical "saturation of redshift" observed in long conjugated polymers [41], suggesting that the model has learned latent axes exhibiting behaviour consistent with known non-linear physics of delocalization without explicit supervision.

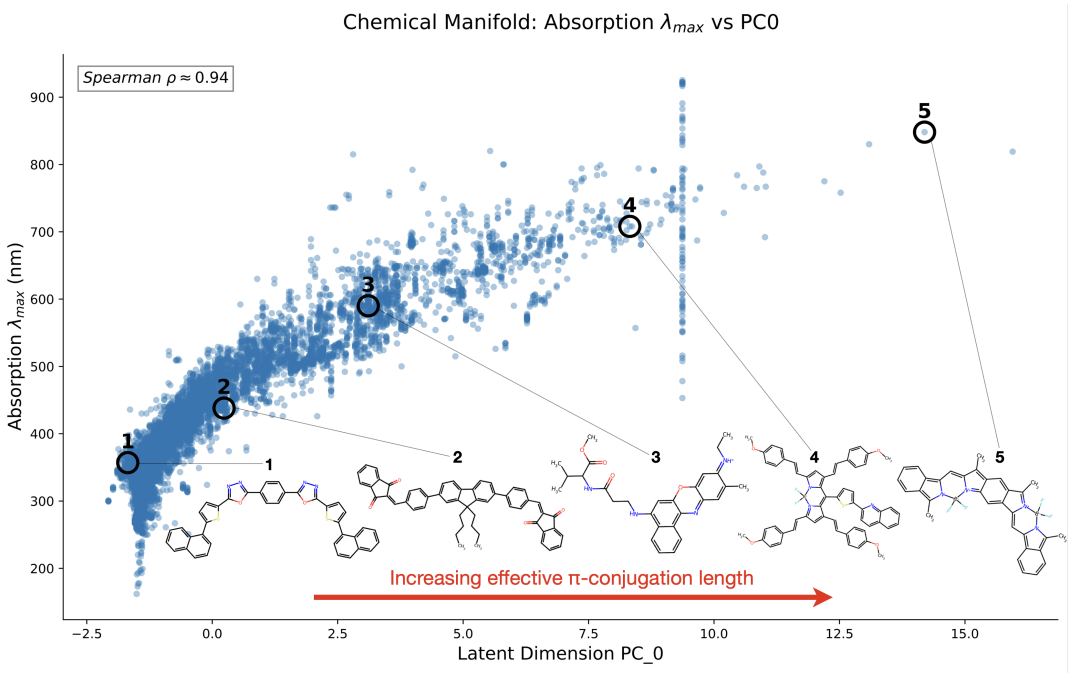

Figure 4.10: Chemical manifold visualization showing the relationship between absorption $\lambda_{\max}$ and the dominant latent dimension PC_0. The axis corresponds to "effective conjugation length": low values indicate interrupted or localized π-systems, while high values indicate extended, planar delocalization.

Secondary Structural and Environmental Axes. Beyond the dominant electronic axis, several secondary components encode orthogonal structural and environmental factors. PC_1 primarily reflects aromaticity and conjugated topology, exhibiting negative correlations with *NumAromaticRings* ($\rho = -0.55$) and *LongestConjChain* ($\rho = -0.54$), while PC_2 tracks overall molecular size and flexibility, correlating with *ExactMolWt* ($\rho = 0.32$), *HeavyAtomCount* ($\rho = 0.32$), and *NumRotatableBonds* ($\rho = 0.27$).

Additionally, PC_8 appears to reflect polarity and solvation effects, correlating with topological polar surface area and hydrogen-bonding capacity. These dimensions adjust the predicted absorption wavelength around the dominant shift captured by PC_0, suggesting that the model distinguishes electronic delocalization from substituent-driven polarity and solvation effects (see Supporting Information, Figures S6 and S7).

While several principal components have clear and chemically meaningful interpretations, not all latent directions exhibit strong correlations with the descriptors examined here. This is expected: PCA is a linear projection of a highly nonlinear learned representation, and individual components may encode interactions or higher-order structure not captured by standard physicochemical descriptors. The fact that multiple dominant components align so strongly with fundamental electronic and structural quantities is therefore notable. Additional analyses and tentative interpretations of lower-correlation components are provided in the Supporting Information.

Orthogonality of Electronic and Structural Factors. To examine the separation of electronic and structural effects, we perform a counterfactual analysis (Figure 4.11). By projecting pairs of molecules (in a fixed solvent) that differ by single structural edits onto the aligned global PC_0 axis, we observe two distinct trajectories:

- **Conjugation Extension and Aromatic Fusion (e.g., Bibenzyl → Stilbene; "Broken Pentacene" → Pentacene):** Structural edits that physically extend the conjugated network (linking double bonds or fusing aromatic rings) result in large displacements along the PC_0 axis, driving the expected redshift [1, 62].

- **Electronic Push-Pull (e.g., *para*-Nitro-substitution):** Edits that induce charge transfer without extending the π-chain (e.g., donor-acceptor substitutions) produce vertical shifts in predicted $\lambda_{\max}$ (red or blue depending on the solvent), with minimal displacement along PC_0 [40, 52, 65].

These trends suggest that the latent representation separates wavelength shifts associated with the extent of π-conjugation (captured by PC_0) from those associated with substituent-induced charge transfer (captured by secondary dimensions), without requiring explicit physical supervision.

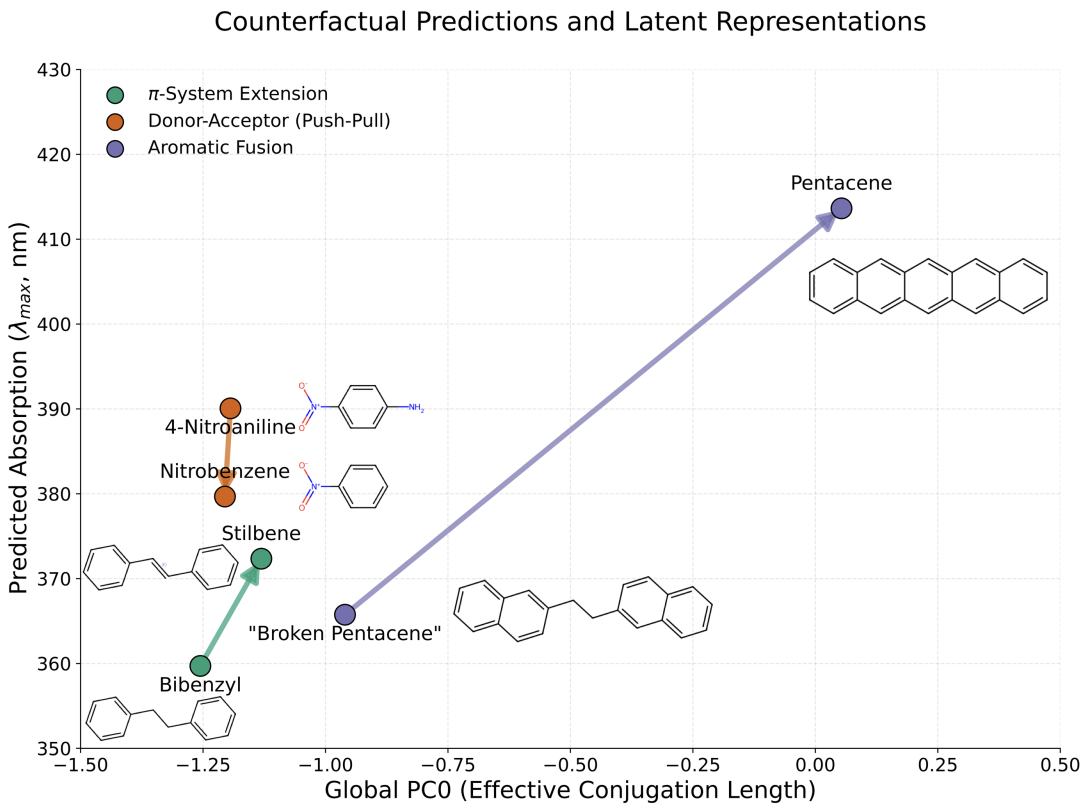

Figure 4.11: Counterfactual interpretation of the learned latent space. Latent embeddings for chemically matched pairs are projected onto the global PC_0 axis. Each arrow denotes a structural modification while keeping solvent constant. The horizontal axis (PC_0) tracks the effective conjugation length, while the vertical axis shows predicted $\lambda_{\max}$.

4.6 Conclusion

Data scarcity remains the primary bottleneck for applying deep learning to experimental photophysics. In this work, we showed that hybrid architectures can improve performance in this setting. By replacing the standard neural head of a trained D-MPNN with a post-hoc XGBoost model, we achieved relatively consistent improvements in absorption wavelength prediction across low-to-mid data regimes ($N \lesssim 1000$), particularly under rigorous scaffold splits that simulate realistic discovery scenarios.

Mechanisms of Stability and Interpretability. Our analysis identifies two distinct mechanisms driving this advantage. First, variance decomposition revealed that in small-data regimes, the mapping from learned representations to molecular properties is underdetermined; the hybrid model appears to stabilize this mapping by imposing the structural regularization of decision trees, effectively pruning the high-variance error component associated with neural heads. Second, we showed that the D-MPNN encoder, even when trained on limited data, learns a latent space that aligns with physically meaningful structure. The dominant principal component (PC_0) spontaneously aligns with the concept of "effective conjugation length," separating extended π-conjugation from substituent-driven electronic effects. This suggests that the model's predictive performance is driven in part by representations aligned with known physical factors, rather than relying solely on dataset-specific correlations.

Practical Implications. These findings suggest a practical workflow for molecular screening in low-data settings. Since fitting an XGBoost head on fixed embeddings incurs virtually negligible computational cost, practitioners can leverage the hybrid approach to maximize predictive stability in data-scarce regimes without sacrificing the scalability of end-to-end learning as datasets grow. Furthermore, our transfer learning results indicate that this strategy is particularly effective when adapting pretrained models to chemically similar downstream tasks (e.g., Deep4Chem $\rightarrow$ ChemFluor), though its utility diminishes under severe domain shift (e.g., Deep4Chem $\rightarrow$ DSSCDB).

Limitations and Future Directions. Several open challenges point toward future research.

- **Excited-State Physics:** The hybrid advantage was less systematic for emission prediction. This likely reflects the fact that ground-state graphs do not explicitly encode excited-state relaxation or solvent reorganization energies. Future work should ex-

plore integrating low-cost conformational sampling or solvent-aware message passing to better constrain these dynamic effects.

- **Domain Adaptation:** While transfer learning succeeded for structurally overlapping datasets, it failed under significant distribution shift. Addressing this will likely require more principled domain adaptation techniques under data scarcity, such as adversarial pretraining or uncertainty-weighted fine-tuning, rather than simple parameter transfer.

- **Feature Integration:** An additional direction is to evaluate the integration of learned embeddings with physicochemical descriptors within the hybrid framework. Tree-based models naturally accommodate heterogeneous feature sets, suggesting that hybrid approaches may provide a practical mechanism for incorporating complementary domain knowledge in low-data regimes. A controlled comparison with end-to-end models trained on the same augmented inputs is required to determine whether this flexibility translates into improved predictive performance.

- **Scope of Properties:** Extending this analysis to intensive properties like quantum yield or fluorescence lifetime, which depend heavily on non-radiative decay pathways, will further clarify the boundaries of graph-based representation learning.

Ultimately, this work supports the use of the hybrid D-MPNN→XGBoost framework as a strong and practical baseline for optical property prediction in low-data regimes. By combining the representational power of graph neural networks with the robust inference of classical machine learning, it provides a scalable and computationally efficient approach that remains grounded in chemically meaningful structure.

Chapter 5

Supporting Information for Chapter 4

This chapter presents the Supporting Information corresponding to the manuscript in Chapter 4. It includes additional experimental details, extended analyses, and supplementary figures and tables supporting the results presented in the main text.

5.1 Model Specification and Training Details

5.1.1 Atom and Bond Features

All categorical features are one-hot encoded. Atomic mass is included as a real-valued feature scaled to unit order.

Table S1: Atom features used in the D-MPNN encoder.

Feature	**Description**	**Size**
Atom type	Atomic number–based encoding	100
Number of bonds	Total bonded neighbors	6
Formal charge	Integer formal charge	5
Chirality	Tetrahedral CW/CCW or unspecified	4
Number of hydrogens	Bonded hydrogen count	5
Hybridization	sp, sp^2, sp^3, sp^3d, sp^3d^2	5
Aromaticity	Aromatic atom indicator	1
Atomic mass	Scaled atomic mass	1

Table S2: Bond features used in the D-MPNN encoder.

Feature	Description	Size
Bond type	Single, double, triple, aromatic	4
Conjugation	Conjugated bond indicator	1
Ring membership	Bond in ring indicator	1
Stereo	None, any, E/Z, cis/trans	6

5.1.2 Training Convergence Under Fixed Horizons

Figure S1 presents training and validation loss curves for $N = 100$ and $N = 11{,}816$ to assess whether fixed training horizons induce late-epoch validation divergence.

5.2 Variance Component Estimation

5.2.1 Model Specification

To quantify the sources of predictive variability, we employ a fully nested ANOVA design. Let y_{ijk} denote the performance metric (RMSE) for the k-th replicate of the j-th regression head trained on the i-th encoder. The experimental design consists of 125 independent training runs structured as follows:

- $i = 1, \ldots, a$: Number of independent Encoder training runs ($a = 5$).
- $j = 1, \ldots, b$: Number of independent Head training runs nested within each encoder ($b = 5$).
- $k = 1, \ldots, n$: Number of independent random seed replicates per head ($n = 5$).

The statistical model is:

$$y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk} \tag{5.1}$$

where:

- μ is the grand mean.

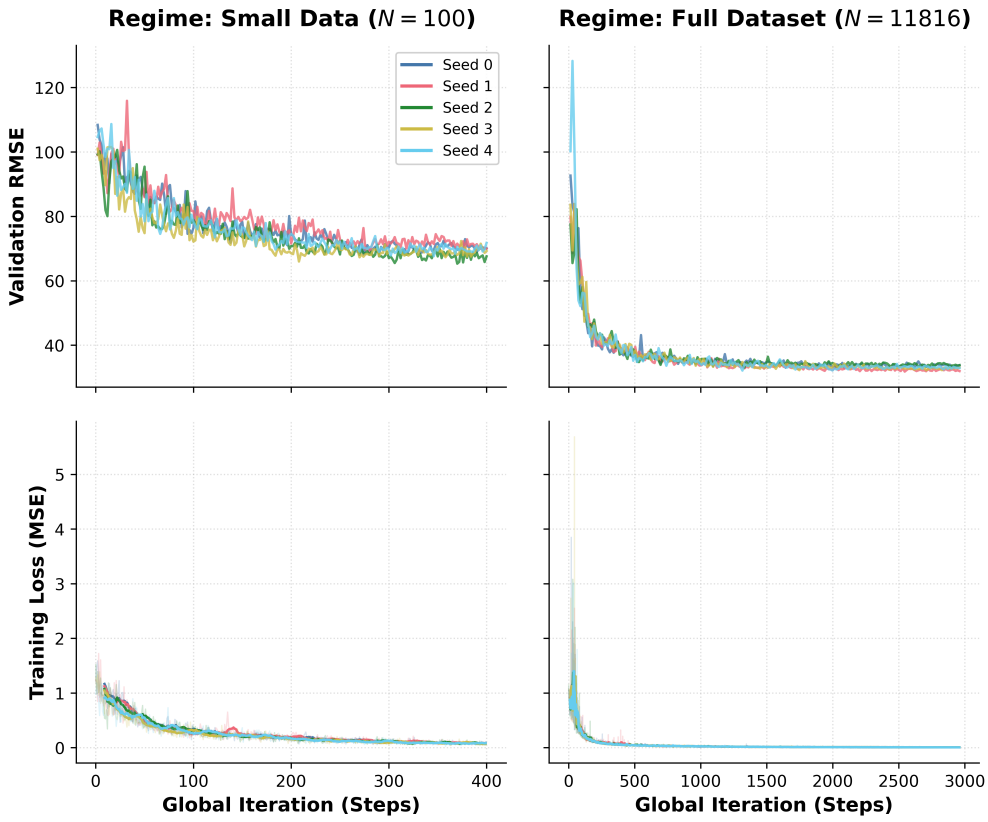

Figure S1: Training and validation convergence under fixed training horizons. Validation RMSE (top) and training MSE (bottom) across five random seeds for the Deep4Chem dataset in a small-data regime ($N = 100$, left) and the full dataset ($N = 11{,}816$, right). Using a fixed y-axis for direct comparison, the plots show rapid stabilization for large N and expected stochasticity for small N. The absence of late-epoch divergence in either regime supports the use of a fixed training horizon for cross-model comparisons.

- $A_i \sim \mathcal{N}(0, \sigma_A^2)$ is the random effect of the **Encoder**.
- $B_{j(i)} \sim \mathcal{N}(0, \sigma_{B|A}^2)$ is the random effect of the **Head** (nested within Encoder).
- $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is the residual error (seed variability).

5.2.2 Sum of Squares and Mean Squares

We compute the Sum of Squares (SS) and Mean Squares (MS) as follows:

1. Encoders (Factor A)

$$SS_A = bn \sum_{i=1}^{a} (\bar{y}_{i..} - \bar{y}_{...})^2 \tag{5.2}$$

$$MS_A = \frac{SS_A}{a-1} \tag{5.3}$$

2. Heads within Encoders (Factor B nested in A)

$$SS_{B(A)} = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij.} - \bar{y}_{i..})^2 \tag{5.4}$$

$$MS_{B(A)} = \frac{SS_{B(A)}}{a(b-1)} \tag{5.5}$$

3. Residual / Error

$$SS_E = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \bar{y}_{ij.})^2 \tag{5.6}$$

$$MS_E = \frac{SS_E}{ab(n-1)} \tag{5.7}$$

5.2.3 Expected Mean Squares (EMS) and Estimators

Assuming a balanced design and random effects, the Expected Mean Squares are:

$$\mathbb{E}[MS_A] = \sigma_\varepsilon^2 + n\sigma_{B|A}^2 + bn\sigma_A^2 \tag{5.8}$$

$$\mathbb{E}[MS_{B(A)}] = \sigma_\varepsilon^2 + n\sigma_{B|A}^2 \tag{5.9}$$

$$\mathbb{E}[MS_E] = \sigma_\varepsilon^2 \tag{5.10}$$

By solving this system of linear equations, we obtain the Method of Moments estimators used in the main text:

$$\hat{\sigma}_\varepsilon^2 = MS_E \tag{5.11}$$

$$\hat{\sigma}_{B|A}^2 = \frac{MS_{B(A)} - MS_E}{n} \tag{5.12}$$

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{B(A)}}{bn} \tag{5.13}$$

In cases where an estimate is negative it is clamped to zero.

5.2.4 Bootstrap Uncertainty Analysis

Bootstrap uncertainty analysis. To assess the robustness of the variance component estimates, we performed a nonparametric bootstrap by resampling encoder–head–replicate triplets and recomputing variance components for each resample. As shown in Table S3, confidence intervals are wide in low-data regimes, reflecting the limited number of encoders and heads available for estimation. Nevertheless, across all dataset sizes, the qualitative ordering of variance sources is preserved, with encoder-level stochasticity consistently contributing the largest share of predictive variability.

Table S3: Bootstrap confidence intervals for variance component proportions. Mean estimates and 95% bootstrap confidence intervals for the proportion of total predictive variance attributable to encoder stochasticity (A), head stochasticity (B(A)), and residual seed noise (E) across training set sizes on the Deep4Chem dataset. Variance components were estimated using the ANOVA Method of Moments under a balanced nested design ($I = 5$, $J = 5$, $K = 5$). Due to the limited number of encoders and heads, confidence intervals are wide, particularly in low-data regimes; estimates should therefore be interpreted qualitatively.

Subset	**Encoders (A)**			**Heads (B(A))**			**Residual (E)**		
	Mean	CI_{low}	CI_{high}	Mean	CI_{low}	CI_{high}	Mean	CI_{low}	CI_{high}
N00050	0.646	0.028	0.853	0.182	0.022	0.731	0.172	0.058	0.459
N00100	0.765	0.184	0.894	0.100	0.031	0.561	0.134	0.049	0.335
N00250	0.603	0.000	0.791	0.195	0.082	0.691	0.202	0.068	0.536
N01000	0.943	0.083	0.977	0.042	0.015	0.619	0.015	0.003	0.369
N11816	0.917	0.516	0.960	0.029	0.007	0.226	0.054	0.021	0.274

5.3 Head Sensitivity and Fixed-Encoder Ablations

As shown in Figure S2, holding the encoder fixed isolates variability attributable solely to the regression head. We observe differences in both average performance and variability across heads, with the dashed line marking the end-to-end D-MPNN reference.

5.3.1 Fixed-Encoder Head Comparison

In Figure S2, we fix a trained D-MPNN encoder and evaluate different regression heads, reporting mean performance over 30 seeds with standard deviation. This isolates the contribution of the downstream mapping from the learned representation. Tree-based models, particularly XGBoost, exhibit improved stability and competitive performance relative to neural heads, motivating their use in the hybrid framework.

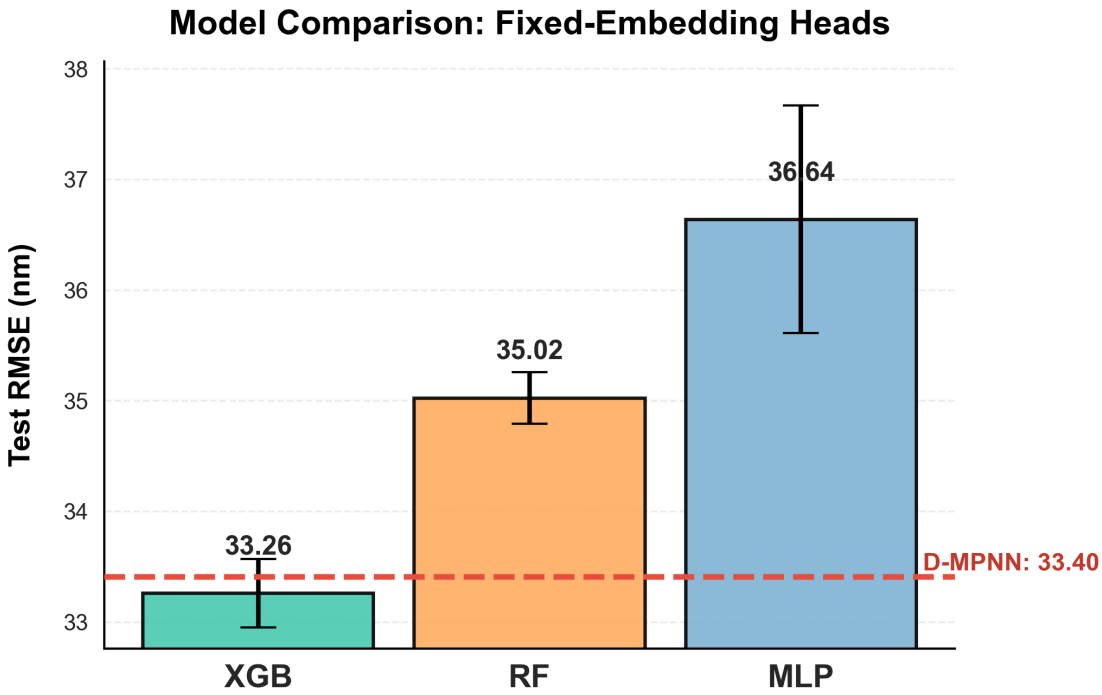

Figure S2: Test-set RMSE for different regression heads trained on a fixed D-MPNN encoder representation. Values are the mean ± SD over 30 seeds. The dashed line represents the performance of the end-to-end D-MPNN model.

5.3.2 Ensembling Effects Across Regimes

Ensembling effects and stability across regimes. Figure S3 compares individual models to ensemble predictions obtained by averaging independently trained instances. Ensembling consistently reduces test-set RMSE for both hybrid and end-to-end models, with the largest gains observed in low-data regimes. The improvement is most pronounced where performance is most variable, indicating that averaging helps mitigate run-to-run instability.

This pattern is consistent with the variance decomposition results. In low-data settings ($N \leq 250$), predictive performance is limited by both variability in the learned representation and sensitivity of the downstream mapping. The hybrid model primarily reduces the latter, while ensembling further improves performance by averaging over variability in the representation.

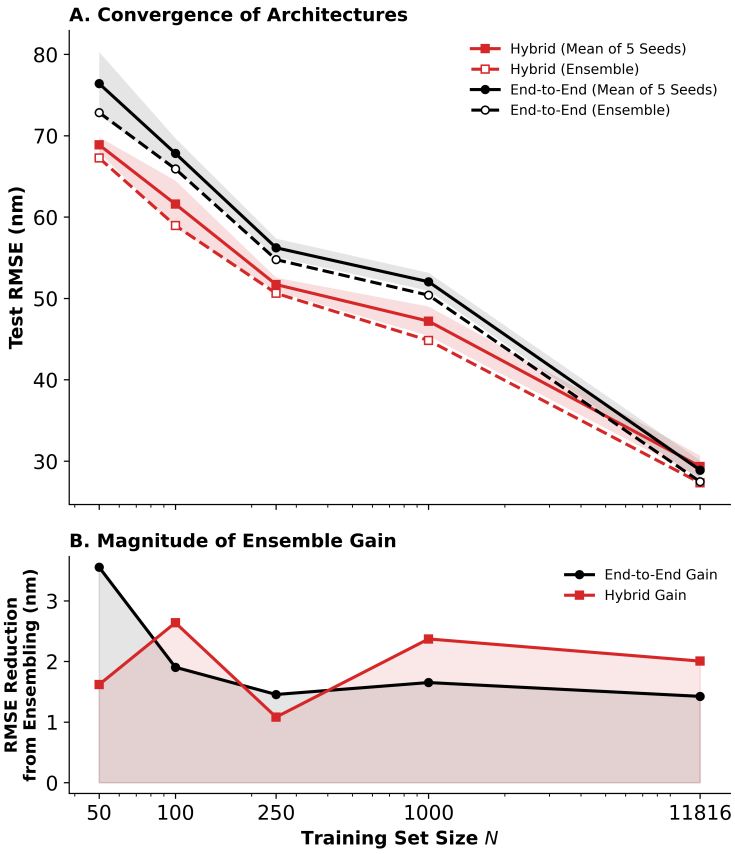

Figure S3: Stability and ensemble analysis. Test RMSE on Deep4Chem across training set sizes for the Hybrid (red) and End-to-End (black/gray) models. Solid lines show the mean of five independent training seeds (error bars: one standard deviation), while dashed lines indicate 5-member ensembles.

Implications of Ensemble Performance on Representation Diversity. Ensemble theory shows that averaging predictions reduces error when individual models make partially uncorrelated mistakes [34]. The performance gains observed for the hybrid ensemble relative to individual D-MPNNs therefore suggest that the models capture different aspects of the chemical space, leading to prediction errors that are not fully correlated.

5.4 Emission Prediction and Physical Limitations

To assess generalization, we extend our analysis to peak emission wavelength prediction. This task is more challenging than absorption prediction, as emission depends on excited-state relaxation and solvent reorganization (Stokes shift), which are not captured by the ground-state molecular graphs used as input.

Figure S4 shows test RMSE for emission prediction under scaffold splits. Unlike absorption, where the hybrid advantage is consistent, emission exhibits a regime-dependent trade-off. In the Deep4Chem and DSSCDB datasets, hybrid and end-to-end models perform similarly, with the end-to-end model showing a slight advantage at larger N.

In ChemFluor, performance varies more across N, and neither architecture shows a consistent advantage over the full range. Overall, these results suggest that hybridization is most effective when the learned representation captures the dominant drivers of the target property. For emission, additional excited-state and solvent reorganization effects are likely not captured by the current inputs, and may require either richer features or end-to-end adaptation.

5.5 Latent Representations

5.5.1 Cross-Seed Stability of Principal Components

Table S4 shows that the dominant latent component (PC_0) exhibits highly consistent correlations with both the HOMO–LUMO gap and the target absorption wavelength across independently trained encoder seeds.

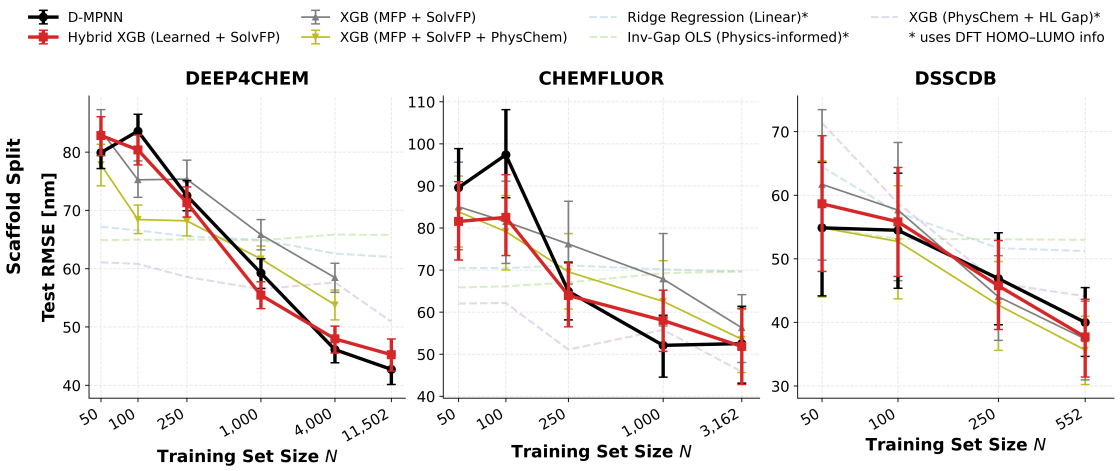

Figure S4: Peak emission wavelength prediction across Deep4Chem, DSSCDB, and ChemFluor. Test RMSE versus training set size N (scaffold split). Each panel uses the dataset-specific N values. Error bars indicate 95% bootstrap confidence intervals for the test RMSE, computed by resampling test molecules with replacement (2000 replicates).

Table S4: Cross-seed stability of correlations between the dominant latent component (PC_0) and key physical properties on the Deep4Chem dataset. Values represent the Spearman correlation coefficient (ρ) for each independently trained encoder seed.

Encoder Seed	PC_0 vs. HOMO–LUMO Gap	PC_0 vs. λ_{max}
0	−0.817	0.942
1	−0.819	0.941
2	−0.826	0.941
3	−0.811	0.939
4	−0.827	0.947
Aggregate	$\mathbf{-0.820 \pm 0.007}$	$\mathbf{0.942 \pm 0.003}$

5.5.2 Interpretation of Higher-Order Principal Components

While the dominant principal components admit relatively clear chemical interpretations, several higher-order dimensions exhibit weaker or more diffuse correlations with the selected descriptor set. Below, we summarize the most prominent trends visible in Figure 4.9 of the main text, noting that these interpretations are necessarily tentative.

PC_3: Weak Size and Flexibility Signal. Unlike PC_2, which captures global molecular size in a coherent manner, PC_3 displays comparatively weak negative correlations with *NumRotatableBonds* ($\rho = -0.29$), *ExactMolWt* ($\rho = -0.19$), and *HeavyAtomCount* ($\rho = -0.19$). These magnitudes suggest only a mild association with molecular compactness or flexibility, and do not support a strong one-dimensional physical interpretation.

PC_4 and PC_5: Low-Amplitude Structural Variation. Both PC_4 and PC_5 exhibit generally weak correlations ($|\rho| < 0.3$) across the descriptor panel. PC_5 shows a modest negative association with *FractionCSP3* ($\rho = -0.29$), while PC_4 lacks a dominant alignment with any individual descriptor. These components likely reflect distributed structural effects or composite interactions rather than a single chemically interpretable axis.

PC_6: Hybridization Sensitivity. PC_6 shows a moderate negative correlation with *FractionCSP3* ($\rho = -0.33$), indicating sensitivity to the degree of sp^3 character. This pattern is consistent with a distinction between more planar, conjugated systems and more saturated, three-dimensional scaffolds, though the effect size remains secondary relative to the dominant axes.

PC_7: Polarity-Associated Variation. PC_7 exhibits moderate negative correlations with *TPSA* ($\rho = -0.33$) and *NumHBA* ($\rho = -0.29$), alongside smaller associations with flexibility-related descriptors. This suggests sensitivity to polarity-related substitution patterns, though again without a uniquely defining descriptor signature.

PC_9 and Higher Components: Diffuse Effects. Higher-order components such as PC_9 show no strong monotonic correlations with the evaluated descriptors ($|\rho| \leq 0.16$). Nevertheless, some of these dimensions are ranked as moderately important by the downstream XGBoost model. This is consistent with the fact that PCA is a linear projection applied to a highly nonlinear learned representation; individual components need not correspond to simple, one-dimensional chemical quantities. Instead, these higher-order directions likely encode nonlinear interactions, composite structural motifs, or dataset-specific variation not captured by standard descriptor panels.

5.5.3 ChemFluor Correlation Map

Having established that the dominant principal component is stable across encoder seeds for Deep4Chem, here we examine whether similar structure emerges on the ChemFluor dataset.

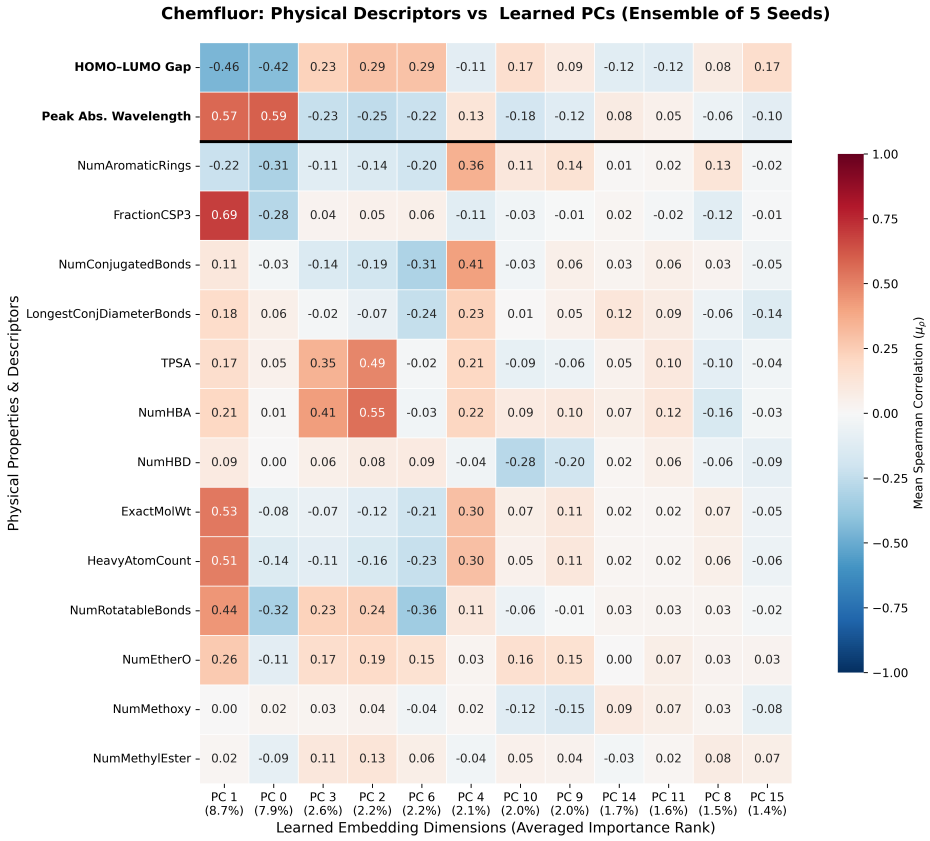

Figure S5: Consensus Interpretability Map: Learned PCs vs Physics. Aligned Spearman rank correlations between the top 12 predictive principal components (PCs) and physicochemical descriptors across an ensemble of 5 seeds on the ChemFluor dataset. PCs are ordered by averaged XGB feature importance. The horizontal divider separates structural descriptors from the photophysical targets, HOMO–LUMO gap and peak absorption wavelength.

5.5.4 Latent Density Analyses

Figure S6 presents hexbin density plots illustrating the relationship between selected latent principal components and independent physicochemical descriptors. The top principal components exhibit structured alignment with chemically meaningful axes. In particular, PC_0 correlates strongly with effective conjugation length and primary optical targets, PC_2 tracks molecular size through molecular weight and heavy atom count, PC_8 reflects polarity and hydrogen-bonding capacity, and PC_1 captures aromatic core complexity. These relation-

ships support the interpretation that the learned representation organizes molecules along physically meaningful dimensions rather than arbitrary latent directions.

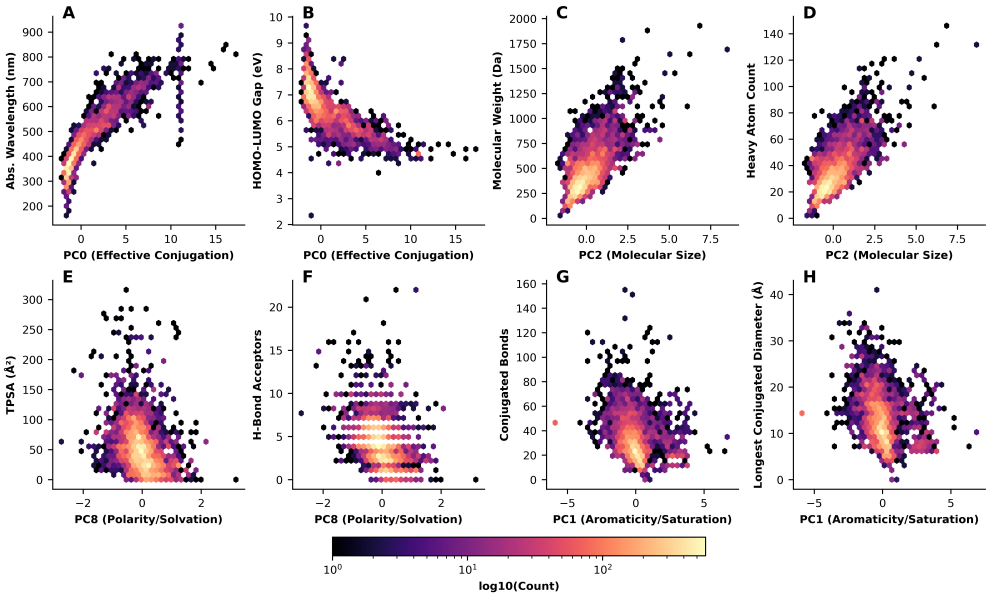

Figure S6: Density distribution of latent dimensions against physicochemical descriptors. Hexbin plots (log-scaled density) demonstrate the alignment of the top Principal Components (PCs) with independent physical properties. **(A-B)** PC_0 captures the effective conjugation length, showing strong correlation with the primary targets. **(C-D)** PC_2 functions as a molecular size axis, tracking weight and heavy atom count. **(E-F)** PC_8 encodes polarity and hydrogen-bonding capacity (TPSA and NumHBA). **(G-H)** PC_1 distinguishes aromatic core complexity, as evidenced by its relationship with conjugated bond count and diameter.

While marginal density plots reveal pairwise relationships, here we visualize the joint latent space of two important PCs to observe how orthogonal axes organize molecular space.

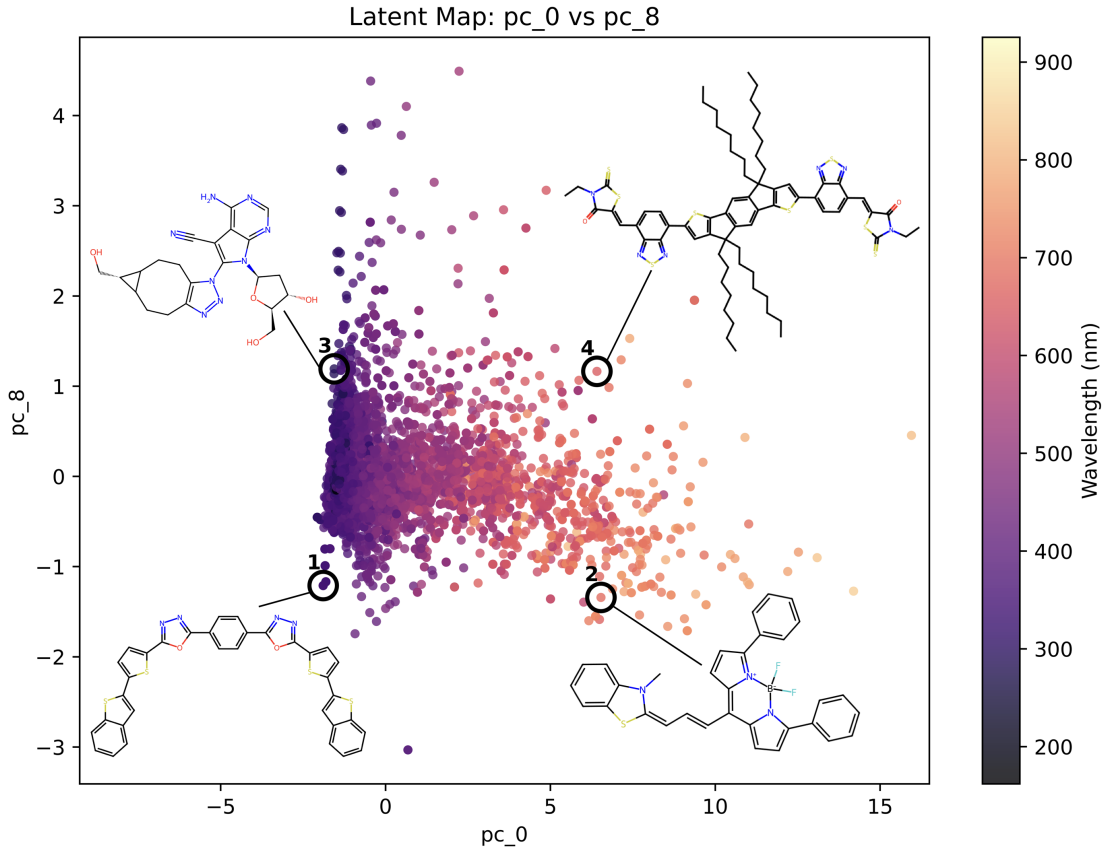

Figure S7: Latent manifold visualization of PC_8 (Polarity and Solvation axis) versus PC_0 (Effective Conjugation axis). The embedding successfully separates the electronic drivers of absorption (horizontal shift) from the structural ballast governing polarity and solvation (vertical shift). Representative molecules highlight the transition from rigid, non-polar systems (bottom) to highly substituted, polar, or branched architectures (top).

5.5.5 SHAP Attribution Analysis

Finally, to connect geometric structure with predictive behaviour, we analyze feature attributions using SHAP values. The SHAP analysis aligns with the principal component interpretation. PC_0 dominates the predictions and shows a smooth, monotonic relationship with $\lambda_{\max}$, consistent with its role as the main driver of spectral shift. Higher-order components contribute smaller, more localized effects, suggesting that they capture secondary

factors such as polarity and substituent-driven perturbations. Overall, the model appears to rely on a small number of structured latent features rather than diffuse correlations.

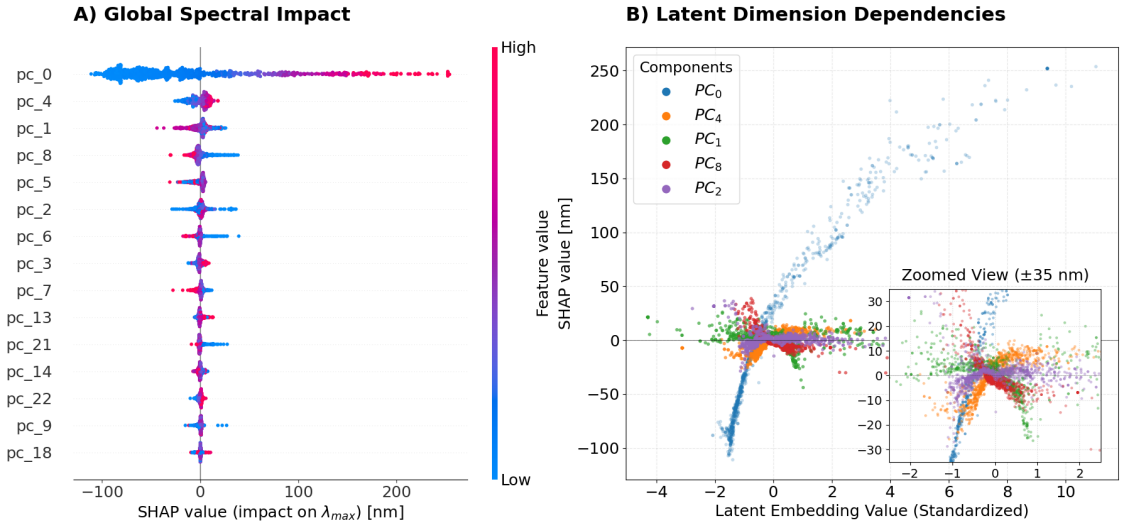

Figure S8: Interpretation of the learned latent representation via SHAP values. (A) SHAP summary plot showing the relative contribution of principal components to predicted $\lambda_{\max}$, with PC_0 dominating the global spectral shift. (B) SHAP dependence plots illustrating how latent dimensions are used by the model: PC_0 exhibits a smooth, monotonic relationship with absorption wavelength, while higher-order components contribute smaller, localized, and non-linear adjustments (inset, ±35 nm).

Chapter 6

Conclusion and Summary

This thesis investigated whether decoupling representation learning from downstream regression improves molecular optical property prediction in data-scarce and distribution-shifted settings. Specifically, we evaluated hybrid pipelines in which a directed message passing neural network (D-MPNN) encoder is first trained end-to-end on the downstream task, then frozen and paired with a post-hoc gradient-boosted tree regressor. We compared this approach against end-to-end fine-tuning, training from scratch, and several classical and physics-informed baselines across multiple datasets and split strategies. Several broad patterns emerged.

First, in small-N regimes for absorption prediction, training a post-hoc XGBoost model on learned representations yields measurable improvements over end-to-end training at negligible additional computational cost. This effect is particularly pronounced under scaffold-based distribution shift, where training and test molecules differ structurally. In these regimes, the explicit structural regularization of the tree-based model can mitigate the instability of the high-capacity neural decoder, which is otherwise prone to overfitting. This hybrid advantage was less systematic for emission prediction, suggesting that static ground-state graphs may lack the necessary information to reliably capture excited-state relaxation and solvent reorganization without end-to-end adaptation.

Second, variance decomposition analyses isolated the mechanisms driving this hybrid advantage. While encoder initialization and training dynamics account for the dominant portion of performance variance across all data regimes, the variability introduced by the downstream regression head becomes substantial in small-N settings. By replacing the highly variable neural head with a regularized ensemble of trees, the hybrid architecture appears to learn a smoother mapping from the unstable learned representations to the response.

Third, transfer learning experiments clarified the boundaries of pretraining utility. Fine-tuning pretrained encoders provides substantial performance gains when the downstream task aligns chemically with the pretraining corpus. However, under severe distribution shift, pretraining offers negligible or even negative returns. In successful transfer scenarios, freezing the fine-tuned encoder and applying a hybrid readout preserves the adapted latent structure while maximizing downstream statistical efficiency.

Beyond predictive performance, analysis of the learned latent space demonstrates that the D-MPNN encoder captures relatively stable, chemically meaningful structure without explicit supervision. The dominant embedding dimension strongly correlates with effective π-conjugation length and the HOMO–LUMO gap, while secondary dimensions disentangle structural and environmental factors like molecular size and polarity.

These results suggest a practical modelling guideline. In experimental photophysics regimes where data are scarce or distribution shift is anticipated, decoupling representation learning from regression is a beneficial strategy. Freezing a task-trained graph encoder and fitting a gradient-boosted tree on the embeddings provides competitive or improved performance for virtually no additional overhead once the encoder is trained.

Conversely, in large-N settings, fully joint optimization becomes sufficiently constrained and the benefit of decoupling diminishes. Future work may extend this analysis to properties more strongly governed by excited-state relaxation, alternative encoder architectures, or domain adaptation techniques specifically designed for severe distribution shift. Nonetheless, the central finding remains clear: the physically grounded representations learned by graph

neural networks can be effectively leveraged by simpler, regularized regressors, offering a practical modelling strategy for data-limited chemical domains.

Bibliography

[1] Adachi, M., and Murata, Y. Relationship between π-conjugation size and electronic absorption spectrum: Novel π-conjugation size dependence of indoaniline dyes. *The Journal of Physical Chemistry A 102*, 5 (1998), 841–845.

[2] Bemis, G. W., and Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry 39*, 15 (1996), 2887–2893.

[3] Bodnar, C., Frasca, F., Otter, N., Wang, Y., Lio, P., Montufar, G. F., and Bronstein, M. Weisfeiler and lehman go cellular: CW networks. *Advances in neural information processing systems 34* (2021), 2625–2640.

[4] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[5] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[6] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine 34*, 4 (2017), 18–42.

[7] Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D., and Lió, P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature communications 15*, 1 (2024), 1517.

[8] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. Machine learning for molecular and materials science. *Nature 559*, 7715 (2018), 547–555.

[9] Casida, M. E. Time-dependent density functional response theory for molecules. In *Recent Advances In Density Functional Methods: (Part I)*. World Scientific, 1995, pp. 155–192.

[10] Casida, M. E. Time-dependent density-functional theory for molecules and molecular solids. *Journal of Molecular Structure: THEOCHEM 914*, 1-3 (2009), 3–18.

[11] Cayton, L., et al. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep 12*, 1-17 (2005), 1.

[12] Chen, T., and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 785–794.

[13] Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. Debiased contrastive learning. *Advances in neural information processing systems 33* (2020), 8765–8775.

[14] Dai, H., Dai, B., and Song, L. Discriminative embeddings of latent variable models for structured data. In *International conference on machine learning* (2016), PMLR, pp. 2702–2711.

[15] Deng, D., Chen, X., Zhang, R., Lei, Z., Wang, X., and Zhou, F. Xgraphboost: extracting graph neural network-based features for a better prediction of molecular properties. *Journal of chemical information and modeling 61*, 6 (2021), 2697–2705.

[16] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (2019), pp. 4171–4186.

[17] Dreuw, A., and Head-Gordon, M. Single-reference ab initio methods for the calculation of excited states of large molecules. *Chemical reviews 105*, 11 (2005), 4009–4037.

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

[19] Facchetti, A. π-conjugated polymers for organic electronics and photovoltaic cell applications. *Chemistry of Materials 23*, 3 (2011), 733–758.

[20] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[21] Gasteiger, J., Gross, J., and Günnemann, S. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123* (2020).

[22] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning* (2017), Pmlr, pp. 1263–1272.

[23] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science 4*, 2 (2018), 268–276.

[24] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[25] Greenman, K. P., Green, W. H., and Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chemical science 13*, 4 (2022), 1152–1162.

[26] Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems 35* (2022), 507–520.

[27] Hamilton, W. L. *Graph representation learning*. Morgan & Claypool Publishers, 2020.

[28] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems 33* (2020), 22118–22133.

[29] Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).

[30] Jha, D., Ward, L., Paul, A., Liao, W.-k., Choudhary, A., Wolverton, C., and Agrawal, A. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports 8*, 1 (2018), 17593.

[31] Joung, J. F., Han, M., Jeong, M., and Park, S. Experimental database of optical properties of organic compounds. *Scientific data 7*, 1 (2020), 295.

[32] Ju, C.-W., Bai, H., Li, B., and Liu, R. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission

wavelengths and quantum yields. *Journal of Chemical Information and Modeling 61*, 3 (2021), 1053–1065.

[33] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design 30*, 8 (2016), 595–608.

[34] Krogh, A., and Vedelsby, J. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems* (1995), vol. 7, MIT Press, pp. 231–238.

[35] Lakowicz, J. R. *Principles of fluorescence spectroscopy*. Springer, 2006.

[36] Li, Q., Han, Z., and Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence* (2018), vol. 32.

[37] Mahé, P., Ueda, N., Akutsu, T., Perret, J.-L., and Vert, J.-P. Extensions of marginalized graph kernels. In *Proceedings of the twenty-first international conference on Machine learning* (2004), p. 70.

[38] Martin, R. M. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.

[39] Mauri, A. alvadesc: A tool to calculate and analyze molecular descriptors and fingerprints. In *Ecotoxicological QSARs*. Springer, 2020, pp. 801–820.

[40] Máximo-Canadas, M., and Borges Jr, I. Absorption spectra of p–nitroaniline derivatives: charge transfer effects and the role of substituents. *Journal of Molecular Modeling 30*, 5 (2024), 120.

[41] Meier, H., Stalmach, U., and Kolshorn, H. Effective conjugation length and uv/vis spectra of oligomers. *Acta Polymerica 48*, 9 (1997), 379–384.

[42] Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 4602–4609.

[43] Mosbach, M., Andriushchenko, M., and Klakow, D. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884* (2020).

[44] Qiu, J., Chen, Q., Dong, Y., Zhang, J., Yang, H., Ding, M., Wang, K., and Tang, J. GCC: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (2020), pp. 1150–1160.

[45] Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the δ-machine learning approach. *Journal of chemical theory and computation 11*, 5 (2015), 2087–2096.

[46] Rampášek, L., Galkin, M., Dwivedi, V. P., Luu, A. T., Wolf, G., and Beaini, D. Recipe for a general, powerful, scalable graph transformer. *Advances in Neural Information Processing Systems 35* (2022), 14501–14515.

[47] Reichardt, C., and Welton, T. *Solvents and solvent effects in organic chemistry*. John Wiley & Sons, 2011.

[48] Rogers, D., and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling 50*, 5 (2010), 742–754.

[49] Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow Jr, R. A., Fisher, J., Jansen, J. M., Duca, J. S., Rush, T. S., et al. Rethinking drug design in the artificial intelligence era. *Nature reviews drug discovery 19*, 5 (2020), 353–364.

[50] Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*. Wiley, New York, 1992.

[51] Sun, R., Dai, H., and Yu, A. W. Does GNN pretraining help molecular representation? *Advances in Neural Information Processing Systems 35* (2022), 12096–12109.

[52] Szatylowicz, H., Jezuita, A., Ejsmont, K., and Krygowski, T. M. Classical and reverse substituent effects in meta-and para-substituted nitrobenzene derivatives. *Structural Chemistry 28*, 4 (2017), 1125–1132.

[53] Tannir, S., Pan, Y., Josephs, N., Cunningham, C., Hendrick, N. R., Beckett, A., McNeely, J., Beeler, A., Jeffries-El, M., and Kolaczyk, E. D. Predicting emission wavelengths in benzobisoxazole-based oleds with gradient boosted ensemble models. *The Journal of Physical Chemistry A 128*, 30 (2024), 6116–6123.

[54] Van Tilborg, D., Alenicheva, A., and Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling 62*, 23 (2022), 5938–5951.

[55] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[56] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[57] Venkatraman, V., Raju, R., Oikonomopoulos, S. P., and Alsberg, B. K. The dye-sensitized solar cell database. *Journal of Cheminformatics 10*, 1 (2018), 18.

[58] Weisfeiler, B., and Lehman, A. A. A Reduction of a Graph to a Canonical Form and an Algebra Arising During This Reduction. *Nauchno-Technicheskaya Informatsia Ser. 2*, N9 (1968), 12–16.

[59] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science 9*, 2 (2018), 513–530.

[60] Xia, J., Zhang, L., Zhu, X., Liu, Y., Gao, Z., Hu, B., Tan, C., Zheng, J., Li, S., and Li, S. Z. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. *Advances in Neural Information Processing Systems 36* (2023), 64774–64792.

[61] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

[62] Yamaguchi, Y., Matsubara, Y., Ochi, T., Wakamiya, T., and Yoshida, Z.-i. How the π conjugation length affects the fluorescence emission efficiency. *Journal of the American Chemical Society 130*, 42 (2008), 13867–13869.

[63] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling 59*, 8 (2019), 3370–3388.

[64] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? *Advances in neural information processing systems 34* (2021), 28877–28888.

[65] Yokoyama, T., Taft, R., and Kamlet, M. J. Effects of 4-substituents on electronic spectra of some 2-nitroaniline derivatives. *Spectrochimica Acta Part A: Molecular Spectroscopy 40*, 7 (1984), 669–673.

[66] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., and Shen, Y. Graph contrastive learning with augmentations. *Advances in neural information processing systems 33* (2020), 5812–5823.

[67] Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. *Advances in neural information processing systems 30* (2017).