ARTICLE IN PRESS

Communications Chemistry

https://doi.org/10.1038/s42004-026-01944-5

Article in Press

A conformational benchmark for optical property prediction with solvent-aware graph neural networks

Received: 7 October 2025

Accepted: 5 February 2026

Cite this article as: Potapov, D., Rogovoi, S., Khrabrov, K. *et al.* A conformational benchmark for optical property prediction with solvent-aware graph neural networks. *Commun Chem* (2026). https://doi.org/10.1038/s42004-026-01944-5

Denis Potapov, Sergei Rogovoi, Kuzma Khrabrov, Konstantin Ushenin, Alexey Korovin, Anton Ber, Artur Kadurin & Artem Tsypin

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

ARTICLE IN PRESS

A Conformational Benchmark for Optical Property Prediction with Solvent-aware Graph Neural Networks

Denis Potapov[1,2*†], Sergei Rogovoi[1,3†], Kuzma Khrabrov[1], Konstantin Ushenin[1,4], Alexey Korovin[1], Anton Ber[1], Artur Kadurin[1,4,5,6], Artem Tsypin[1]

[1]AIRI, Moscow, Russia.
[2]Moscow Institute of Physics and Technology, Moscow, Russia.
[3]Lomonosov Moscow State University, Moscow, Russia.
[4] Tomsk State University, Tomsk, Russia.
[5] Kuban State University, Krasnodar, Russia.
[6] ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia.

*Corresponding author(s). E-mail(s): potapov@airi.net;
Contributing authors: rogovoyserg@gmail.com; khrabrov@airi.net; ushenin@airi.net; korovin@airi.net; ber@airi.net; kadurin@airi.net; tsypin@airi.net;
†These authors contributed equally to this work.

ARTICLE IN PRESS

Abstract

Accurately predicting optical spectra of molecules is essential for creating better OLED emitters, solar-cell dyes, and fluorescent probes. Traditional methods, such as time-dependent density-functional theory, are computationally expensive and often inaccurate. Current Graph Neural Network (GNN) approaches for optical properties prediction are faster and offer better performance. Still, they operate on 2D graphs and ignore the 3D geometrical features that control excited-state behavior. We present nablaColors-3D, a rigorously curated dataset for the prediction of optical properties consisting of 26369 chromophore-solvent pairs with three conformations optimized at different levels of quantum theory. Based on this dataset, we establish a scaffold-split benchmark for 3D GNNs and systematically quantify how the fidelity of geometry optimization affects accuracy. Furthermore, we propose a solvent-aware modification for pretrained SE(3)-invariant architectures. Our best model, built on UniMol+, achieves MAE of 15.97 nm on a held-out test set, improving the previous state of the art by more than 30%.

Keywords: UV/Vis spectroscopy, molecular property prediction, optical property prediction, graph neural network, quantum chemistry, organic chromophores

1 Introduction

Accurate prediction of molecular optical properties is fundamental to advancing research across numerous scientific and technological domains. This predictive capability serves critical needs in organic light-emitting diodes (OLEDs) [1–4], organic solar cells [5–7], biosensors and bioimaging [8–11], organic dyes and pigments [12, 13], photocatalysis [14, 15]. The optical characteristics of organic compounds fundamentally determine their applicability across diverse technological domains. For OLEDs, emission peak position and bandwidth directly impact color rendering and device efficiency [2]. For solar energy harvesting, compounds with absorption wavelengths matching the solar emission spectrum are especially valuable, making accurate prediction of this parameter crucial for identifying promising solar cell candidates without extensive experimental screening [16]. Similarly,

ARTICLE IN PRESS

in bioimaging applications, precise control over absorption wavelengths allows researchers to target specific "optical windows" where biological tissues exhibit minimal light attenuation, thereby enhancing imaging depth and resolution [10].

The optical properties of molecules are traditionally determined through experimental techniques such as UV/Vis spectroscopy [17]. While highly informative, these measurements can be time-consuming, resource-intensive, and limited in throughput, particularly when large chemical libraries or unstable compounds are involved. Moreover, the experimental identification of molecular structures with desired optical characteristics often relies on chemical intuition and iterative trial-and-error strategies, which are not always sufficient to explore the vastness of chemical space efficiently.

To address these challenges and to provide a more systematic understanding of molecular excited states, theoretical approaches have been developed to complement experimental methods. Among these, time-dependent density functional theory (TD-DFT) has emerged as a widely adopted tool for predicting the electronic structure of excited states and subsequently estimating absorption and emission wavelengths [18, 19]. However, its accuracy is often limited by the approximate nature of exchange-correlation functionals, resulting in systematic deviations for charge-transfer transitions and double excitations [20, 21]. In addition, the steep computational cost, which scales as $\mathcal{O}(M^3)$ with the number of basis wavefunctions M [22], makes TD-DFT impractical for large chromophores and high-throughput studies. Predicting more complex properties, such as photoluminescence quantum yield (PLQY) or emission lifetimes, requires modeling non-radiative decay and vibronic effects beyond the standard framework[23], while obtaining accurate absorption spectra demands consideration of vibrational broadening and multiple excited states, adding further computational complexity [24].

In light of the limitations of TD-DFT, including accuracy constraints, high computational cost, and limited scalability, machine learning (ML) approaches have emerged as attractive alternatives [25–34]. These methods offer the potential to replicate or even surpass TD-DFT accuracy while enabling rapid screening across vast molecular libraries [25, 30]. Studies employing algorithms such as random forests [35], support vector machines [36], and gradient boosting [37] with features such as circular fingerprints [38], Morgan fingerprints [39], and other SMILES-derived descriptors have shown that even simple representations can capture a substantial portion of the structural information needed to predict absorption properties [25–28]. At the same time, more advanced approaches based on deep residual convolutional neural networks [29] and graph neural networks (GNNs) [30–34] have been developed to leverage structured representations of molecules better to improve these results. For example, Greenman et al. [30] developed the UVVisML framework using a separate directed message passing network for dye and solvent. Sun et al. [32] integrated k-hop subgraphs, edge features, and global chemically intuitive features (e.g., measures of aromaticity, rotatable bonds) of chromophore into the GNN framework. Hung et al. [33] modified the SchNET model [40, 41] by substituting traditional interatomic distance matrices with a bondstep representation derived from SMILES strings, effectively bypassing the collection of 3D conformations of chromophores. Fang et al. [42] utilized multi-modal LLMs and chain-of-thought finetuning. McNaughton et al. [34] addressed a related task of predicting the entire UV-vis absorption spectrum rather than just the peak positions. The study demonstrated that incorporating ab initio simulation data, such as 3D geometries of chromophores and quantum mechanically calculated spectra, alongside 2D molecular graph representations, enhances the prediction accuracy of experimentally obtained spectra. However, the performance boost is mainly due to the incorporation of quantum mechanically calculated spectra.

Recently, Neural Network Potentials (NNPs) [40, 41, 43–53] have emerged as an alternative to costly quantum mechanical simulations. NNPs have been applied, for example, to accelerate molecular dynamics simulations [54], conformational optimization [55, 56], and solubility prediction [57]. To aid in training these models, large and diverse quantum chemical (QC) databases have been collected [58–67]. The abundance of high-quality data enables pretraining of NNPs that can be efficiently finetuned on more complicated datasets or different levels of theory [68, 69], and to solve downstream molecular property prediction tasks [70–72].

Recent advances in quantum-chemically informed NNPs have inspired us to apply a similar approach to the prediction of optical properties. Among these, we focused on absorption peak wavelengths for several reasons. First, absorption data is significantly more abundant and diverse across datasets [25, 73], providing a richer training signal. Second, the majority of existing machine learning models for molecular optoelectronic property prediction have concentrated on absorption [30, 32, 33, 42], establishing it as a widely accepted benchmark. Finally, absorption is often the first experimentally accessible indicator of a molecule's electronic structure and is critical for a

ARTICLE IN PRESS

wide range of applications, from dye design to photovoltaics. Despite our primary focus on absorption peak wavelengths, this work also contains computational experiments on the prediction of other optoelectronic properties.

A major challenge in comparing machine learning models for chromophore optical properties is the lack of a standardized evaluation benchmark. Studies are based on disparate datasets: Lee et al. [31], Hung et al. [33], and Fang et al. [42] rely only on Deep4Chem [73]. Sun et al. [32] utilizes five datasets for testing. Deep4Chem, BODIPYs [74], JCIM_Abs [29], ChemFluor [25], and SMFluo1 [75]. Greenman et al. [30] merge Deep4Chem, ChemFluor, CDEx, DyeAgg [76], and DSSCDB [77] into one large dataset. Evaluation protocols also differ. Lee et al. [31] and Sun et al. [32] use k-fold cross-validation. Greenman et al. [30], Hung et al. [33], and Fang et al. [42] perform a random train/validation/test split. Greenman et al. [30] demonstrates that naive random splitting can introduce data leakage: the same chromophore may appear in both training and test sets when solvent–solute pairs are treated as independent samples. They therefore suggest a scaffold split based on Bemis–Murcko scaffolds [78], which ensures that structurally similar molecules are grouped and restricted to a single subset, thereby preventing overlap between training, validation, and test sets and enabling a more rigorous evaluation of model generalization.

To address the methodological and benchmarking challenges outlined above, our work makes the following key contributions:

- High-quality 3D dataset: We present nablaColors, a rigorously curated dataset of chromophore–solvent pairs with corrected SMILES and nablaColors-3D, a high-quality 3D conformations dataset optimized at multiple levels of theory (xTB, DFT in vacuum and implicit solvent). See section 4.1 for details.
- Benchmark for 3D GNN-based optical properties prediction: We introduce a dedicated benchmark specifically designed for evaluating 3D graph neural networks in molecular optical absorption prediction. This benchmark enables rigorous evaluation of both model architectures and conformational fidelity, as well as reproducible evaluation protocols for spatially informed methods.
- Systematic analysis of the effect of geometry optimization on prediction accuracy: We investigate how the accuracy of absorption prediction depends on the level of theory used for geometry optimization. This analysis enables us to quantify the trade-offs between computational cost and model performance.
- State-of-the-art absorption prediction with pretrained 3D GNNs: We evaluate four equivariant GNNs (PaiNN [79], DimeNet++ [44], GemNet [45], eSCN [48]) and the molecular transformer UniMol+ [80] and show that pretraining on large quantum-chemical datasets is critical for high performance. Our best model, combining pretrained UniMol+ with solvent embeddings, improves MAE by more than 30% compared to the strongest 2D GNN baseline evaluated under the same protocol.
- Multitarget optical property prediction with cross-validation: We extend the best-performing UniProp model to jointly predict absorption, emission, and PLQY in a multitarget setting using an expanded version of the dataset. We perform independent 5-fold scaffold-based cross-validation for both single-target and multitarget setups, reporting strong performance across all properties, and establish the multitarget benchmark for solvent-aware 3D GNNs in optical property prediction.

2 Results

In this section, we present benchmark results for 2D and 3D models trained on the curated nablaColors-3D dataset; full details of data preparation, model architectures, and training protocols are provided in the Methods section.

2.1 Overall Model Performance

We evaluate all models on the proposed nablaColors-3D benchmark. Figure 1 shows the test set MAE (in nm) for all models, grouped by category.

Among the 2D baselines, Chemprop trained with explicit solvent information achieves the lowest error (23.60 nm), while other 2D models such as CGIB and Fluor-Predictor yield slightly higher MAE values. Despite not utilizing geometric information, these models demonstrate solid performance.

A consistent improvement is observed with 3D-based GNNs, which leverage spatial molecular structure. For each 3D model, we report the best result obtained through hyperparameter tuning on

ARTICLE IN PRESS

the validation dataset, using DFT-optimized conformers with implicit solvent and solvent embeddings introduced at the model regression head input. This configuration, providing both high-quality geometries and explicit solvent context, generally led to the lowest MAE and is used in all 3D models in this computational experiment.

Our modified UniProp model, which augments UniMol+ by incorporating a solvent embedding, delivers the best accuracy across all evaluated models, achieving, to the best of our knowledge, a state-of-the-art MAE of 15.97 nm for absorption wavelength prediction. To additionally contextualize these results against a first-principles baseline, we computed vertical absorption wavelength with TD-DFT (B3LYP/def2-TZVPP [81, 82]) on the nablaColors-3D test set, obtaining an MAE of 62 nm. Full training curves for 3D GNNs can be found in Supplementary Figure 1, and additional experiments on model finetuning strategies are in Supplementary Note 1.

The superiority of 3D GNNs over 2D models demonstrates that pretraining 3D GNNs on a large quantum-chemical dataset such as PubChemQC, together with the use of accurate conformers, provides a powerful inductive bias, enabling models to learn physically meaningful representations and substantially improving generalization for spectral property prediction.

2.2 Effect of Conformer Fidelity

We study how the level of quantum theory used for geometry optimization affects optical property prediction by holding all training hyperparameters fixed and varying only the source of conformations. To separate architectural sensitivity from solvent information, we evaluate two regimes: models without any solvent embedding and the same backbones augmented with a Chemprop solvent embedding. Note that for UniMol+ and UniProp, the conformation choice influences only the training phase, not inference. Figure 2 summarizes test MAE across architectures and conformation sources.

Across architectures, higher-fidelity conformers are often associated with lower MAE, but the trend is not consistent for every model and task. In many cases, semi-empirical xTB geometries perform worse than DFT in vacuum geometries. Conformers optimized in implicit solvent further improve accuracy, especially when solvent embeddings are absent. We hypothesize that without explicit solvent embeddings, models extract the information about the solvent from chromophore conformations.

With solvent embeddings enabled, the gaps between different conformation types shrink substantially, suggesting that the embedding supplies much of the solvent-related signal the models require. For UniProp, final accuracy is largely insensitive to conformer fidelity, whereas for UniMol+ without solvent embeddings, higher-fidelity conformers generally yield better results. This yields a practical advantage: UniProp attains strong performance without requiring costly high-level optimizations during training.

However, for UniProp trained on (xTB, DFT implicit) pairs, we still need to generate xTB-optimized conformations for inference, which introduces additional computational burden (see Supplementary Table 4). To mitigate this overhead, we conduct an additional experiment where we train UniProp on (RDKit, DFT implicit) conformation pairs directly. All training hyperparameters and the UniProp backbone are kept fixed. Only the source conformation is changed. The results in Table 1 show that (RDKit, DFT implicit) achieves essentially the same MAE as (xTB, DFT implicit), indicating that UniProp can maintain performance without the xTB optimization step during training.

2.3 Multitarget Learning and Cross-Validation

To further evaluate the generalizability of our UniProp model, we extended the prediction task from single-property regression to multitarget learning. Specifically, we trained the model to jointly predict the maximum absorption wavelength, the maximum emission wavelength, and the photoluminescence quantum yield (PLQY). This experiment was conducted on an extended dataset containing all three target properties. Since not all entries had complete annotations, we adopted a standard multitarget training approach using loss masking for missing values. Multitarget-specific training details, can be found in Supplementary Methods and Supplementary Note 1. We additionally provide cross-validation results for the second best performing model, DimeNet++ with solvent embeddings, in Supplementary Table 3.

We evaluate single-target and multitarget setups using independent 5-fold scaffold-based cross-validation splits. In the single-target case, UniProp achieved an average MAE of 14.8 nm for absorption. In the multitarget setup, the model obtained average errors of 15.3 nm for absorption,

ARTICLE IN PRESS

19.657 nm for emission, and 0.16 for log(PLQY). These results provide a baseline for multitarget optical property prediction with a 3D solvent-aware GNN and facilitate direct comparison with the 2D Chemprop baseline under the same protocol. The cross-validation results are provided in Table 2. During cross-validation, we identified one fold where both UniProp and Chemprop exhibited substantially higher errors. This degradation was linked to multiple measurements of the same solvatochromic probe (betaine dye 36, reported in [83]) across many unique, sparsely represented solvents. When these measurements fall into the test split, a pronounced distribution shift emerges, since most of the corresponding solvents appear only once in the dataset. For fair reporting, the cross-validation metrics in this work are computed with these entries removed from the affected test fold.

3 Discussion

Our results establish that 3D pretrained GNNs provide a substantial improvement in absorption peak prediction compared to both 2D models and TD-DFT baselines. Our extensive benchmarking revealed that utilizing physically informed 3D representations substantially reduces prediction error, achieving more than 30% improvement in mean absolute error (MAE) relative to the strongest 2D baseline. This result is in line with previous research on molecular property prediction, in which 3D models outperform GNNs without explicit spatial features, even with fewer total parameters [80, 84, 85].

One of the factors behind this improvement is the high quality of molecular conformations used as input. Accurate atomic arrangements provided by high-level quantum chemical methods preserve subtle structural features such as bond angles, torsional preferences, and noncovalent interactions, all of which influence excited-state properties [86]. Specifically, the r^{2}SCAN-3c/def2-mTZVPP level of DFT employed here reliably reproduces interatomic distances and bond angles close to ground truth values [87]. By explicitly encoding this geometric information, 3D models generate physically meaningful molecular representations, substantially enhancing prediction accuracy. In contrast, 2D molecular graphs inherently lack this explicit geometric context, limiting their ability to capture geometry-dependent effects.

Compared to conventional quantum chemistry methods like TD-DFT, our pretrained 3D GNNs achieve comparable or better accuracy while reducing computational cost by several orders of magnitude. TD-DFT calculations scale poorly with molecular size [22] and require hours of CPU time per molecule (see Supplementary Table 4), whereas our models deliver predictions within milliseconds after a single conformer is generated. Among the evaluated 3D architectures, UniMol+ and our solvent-aware variant, UniProp, achieved the most accurate predictions, substantially outperforming PaiNN, DimeNet++, GemNet, and eSCN. While the cost of generating 3D conformations for UniMol+ is higher compared to 2D models operating on SMILES alone, its superior accuracy far outweighs this additional cost, making it a more reliable choice for applications requiring high predictive fidelity. The key advantage of UniMol+ lies in its dual capability to both refine molecular geometries and learn optical properties simultaneously, which most other GNN-based models do not explicitly address. First, the SE(3)-invariant transformer backbone with global attention allows UniMol+ to capture long-range interatomic correlations that are critical for electronic transitions but often missed by GNNs with a limited receptive field. Transformers' global receptive field, coupled with 3D positional encodings (spatial and graph-based), ensures a comprehensive structural representation. Second, the coordinate-denoising pretraining paradigm, where raw conformations from the low-cost computational method are iteratively updated toward low-energy DFT-optimized geometries, provides UniMol+ with an implicit understanding of the energy landscape. The model learns a quasi-optimization trajectory (akin to a few-step surrogate for DFT relaxation), enabling accurate property prediction even from approximate input geometries. This is particularly beneficial when high-level conformations (e.g., from DFT) are unavailable at inference time. The UniMol+ paper [88] shows that this conformation-refinement capability contributes directly to its superior scores on PCQM4MV2 [89] and OC20 [90].

This design has important practical consequences. Since UniMol+ internalizes a correction mechanism for approximate geometries during pretraining, high-fidelity DFT conformations are only required in the training phase. This results in a substantial reduction of computational cost during inference compared to TD-DFT and 3D GNNs tested in this work, as UniMol+ can operate effectively on inexpensive xTB conformers. Such efficiency makes it feasible to perform large-scale virtual screening on datasets containing tens or hundreds of thousands of molecules, which would otherwise be impractical with DFT-level geometries.

ARTICLE IN PRESS

Another substantial aspect influencing model performance was careful data curation. By identifying and correcting problematic entries, such as inconsistent protonation states, invalid SMILES strings, and incorrect metal coordination patterns, we substantially reduced label noise. Data denoising is widely recognized as a crucial technique for improving predictive accuracy and preventing overfitting, as demonstrated in prior studies on molecular property prediction [91–96]. Based on empirical observations during dataset construction and early model development, we found that rigorous data cleaning was essential for stable training and competitive performance, consistent with prior reports emphasizing the importance of high-quality molecular datasets.

Incorporation of explicit solvent embeddings via a lightweight 2D encoder provided additional predictive gains. This solvent-aware module explicitly models molecular interactions with solvents, capturing shifts in molecular geometry and electronic structure induced by the solvent environment. Notably, even with solvent embeddings provided to all models, only those using solvent-optimized geometries generally outperformed others. This underscores the importance of explicitly modeling solvent-induced structural adjustments beyond mere solvent descriptors, suggesting further exploration of solvent-aware geometry optimization for future research.

Beyond absorption prediction, the proposed solvent-embedding architecture offers a general mechanism for learning solvent-specific effects in various property prediction tasks. For the 3D GNNs tested here, the same approach could be extended to predict energies and forces in the context of NNPs, reaction barriers, or solvation-dependent properties. UniMol+, initially trained to refine geometries only in vacuum, can now be adapted to predict solvent-adjusted molecular structures.

The multitarget learning experiments provide compelling evidence for the robustness and versatility of our 3D GNN approach. While we observed a modest increase in absorption prediction error when transitioning from single-target to multitarget learning, this minimal degradation demonstrates that our model maintains its predictive accuracy even when learning multiple optical properties simultaneously. The cross-validation results further validate the stability of our approach, showing consistent performance across different data splits despite the inherent challenges of multitarget learning, including varying data availability across properties and potential interference between learning objectives. Notably, the ability to jointly predict absorption wavelength, emission wavelength, and photoluminescence quantum yield from a single model represents a significant practical advantage for molecular design applications. This multitarget capability transforms our approach from a specialized absorption predictor into a comprehensive optical property prediction platform, enabling researchers to efficiently screen molecular candidates across multiple optical characteristics without requiring separate models for each property. The successful extension to emission wavelength (19.894 nm MAE) and PLQY (0.156 MAE) prediction demonstrates that the 3D geometric representations learned by UniProp capture fundamental molecular features relevant to diverse photophysical processes, not merely absorption-specific electronic transitions.

Despite these promising results, several limitations and opportunities for further development exist. A detailed error analysis (Supplementary Note 2) indicates that residual errors are more strongly associated with out-of-distribution chromophores (as measured by training-set similarity) than with simple size/flexibility descriptors; moreover, a targeted conformer-ensemble test using CREST [97] did not yield systematic improvements for the highest-error molecules. The solvent embedding MPNN block used in this study, while effective, was intentionally kept relatively small due to dataset constraints, including limited solvent diversity and dataset size. Future studies might benefit from larger and more diverse datasets, enabling the exploration of more expressive solvent encoding architectures. Furthermore, while UniMol+ provides robustness and superior predictive accuracy, its computational complexity and memory demands may restrict scalability to extremely large molecules or extensive conformational ensembles. Efforts to optimize these architectures for broader applicability represent a compelling direction for future research.

In conclusion, our findings establish the pronounced advantage of employing 3D GNNs, particularly those incorporating global attention mechanisms and geometry-aware pretraining strategies, for predicting molecular optical properties. By meticulously curating high-quality datasets, explicitly modeling solvent effects, and leveraging multi-level geometric fidelity during model training, we provide a robust and efficient methodology that substantially surpasses traditional methods and allows for accurate virtual screening of optically active molecules.

ARTICLE IN PRESS

4 Methods

4.1 Data preparation

To obtain the proposed nablaColors and nablaColors-3D datasets, we extended the dataset integration strategy of Greenman et al. [30] by incorporating not only maximum absorption wavelengths, but also maximum fluorescence wavelengths and photoluminescence quantum yields (PLQYs). To construct a unified dataset, we combined data from Deep4Chem [73], ChemFluor [25], DyeAgg [76], and DSSCDB [77]. Following Greenman et al. [30], we excluded the CDEx dataset [98] due to concerns about data quality.

During the construction of the combined dataset, we encountered several data quality issues. These included invalid SMILES strings, redundant entries with multiple protonation states assigned to what was nominally the same molecule, and broken π-conjugated systems – particularly in squaraine and cyanine dyes – where incorrect SMILES disrupted the conjugated backbone. A notable source of inconsistency came from entries in [25], where molecules were sourced from online databases in solvated, often charged forms (including counterions). These structures were included in the dataset with varying protonation states but were assigned identical optical property values, despite representing chemically distinct species. Further complications arose when handling metal–organic compounds. Their correct protonation was challenging to determine due to the limitations of the SMILES format, which cannot fully capture coordination environments. Conformer generation for metal–porphyrins also posed a challenge, as SMILES cannot represent donor–acceptor and coordination bonds. In some cases, geometry optimization led to bond rearrangements, producing molecular graphs that were no longer isomorphic to those derived from the original SMILES.

To correct such inconsistencies, we manually verified problematic entries by consulting the original experimental studies cited in the source datasets. This verification allowed us to reconstruct accurate molecular structures where possible and ensure consistency between structural representations and associated optical properties. Through careful inspection of the dataset, we removed 1825 problematic entries and replaced 60 with corrected structures. We refer to the resulting dataset as nablaColors.

Many state-of-the-art machine learning models in chemistry require accurate 3D molecular conformations in addition to the molecular graph structure for reliable property prediction [80, 84, 99]. In this work, we established the following computational pipeline to obtain optimized molecular conformations from the SMILES strings available in our filtered nablaColors dataset:

1. SMILES $\rightarrow$ initial 3D. RDKit's ETKDG2 algorithm [100] generates a 3D Cartesian geometry, which is then optimized with the MMFF94 force field [101].
2. xTB geometry optimization. The MMFF-optimized geometry is further optimized at the GFN2-xTB level of theory in vacuum with the xTB package [102]. The xTB optimization takes 0.37 CPU-hours per molecule on average.
3. DFT optimization. The xTB-optimized structure is refined with density functional theory (DFT) at the r^{2}SCAN-3c/def2-mTZVPP [103] level of theory implemented in ORCA [104]. We perform the DFT optimizations both in vacuum (15 CPU-hours per molecule on average) and with the Conductor-like Polarizable Continuum Model (CPCM) [105] (20 CPU-hours per molecule on average).

The r^{2}SCAN-3c/def2-mTZVPP level of theory was chosen for DFT geometry optimizations following established best-practice recommendations for molecular structure determination. This composite meta-GGA approach has been shown to provide accurate and robust equilibrium geometries, reproducing bond lengths and angles while remaining computationally efficient and well suited for large-scale studies. As summarized by Bursch et al. [106], r^{2}SCAN-3c represents an optimal cost–accuracy compromise for routine geometry optimization tasks.

CPU times required for each optimization stage are reported in Supplementary Table 4.

Throughout conformation preparation, we carefully verified that the molecular graph topology of each optimized structure remained consistent with the original SMILES representations. Additionally, during the curation process, we identified several duplicate entries and erroneous data points, which were either corrected or removed based on examination of the original literature sources. This process resulted in 13 731 individual molecules and 26 369 chromophore-solvent pairs. For each chromophore-solvent pair, the dataset contains 4 conformations calculated at different levels of theory. We refer to this curated set as nablaColors-3D. The data processing pipeline is visualized in Figure 3.

ARTICLE IN PRESS

4.2 3D-GNNs for peak absorption wavelength prediction

State-of-the-art models for optical-property prediction are still dominated by 2D GNNs such as Chemprop [30], SOG [32], CGIB [31] and Fluor-Predictor [107]. However, the 3D molecular conformation modulates the electronic structure [86, 108] and thus the absorption spectrum. Incorporating three-dimensional information into the GNN has therefore been shown to yield systematic performance gains [84, 89, 99].

To explore this advantage, in this work, we benchmark five GNNs: PaiNN [79], DimeNet++ [44], eSCN [48], GemNet-OC [45], and UniMol+ [80]. PaiNN, DimeNet++, eSCN, and GemNet-OC were chosen because they deliver the highest accuracy on the recent ∇^2DFT benchmark [63], whereas UniMol+ is one of the top performers on the PCQM4Mv2 leaderboard [89]. PCQM4Mv2 is a large-scale quantum chemistry dataset that provides DFT-optimized geometries and HOMO–LUMO gaps, which are strongly correlated with absorption maxima [109]. We therefore expect pretraining on this dataset to yield transferable representations for the optical property prediction task. Accordingly, we pretrained PaiNN, DimeNet++, eSCN, and GemNet-OC on PCQM4Mv2 and initialized UniMol+ from its official checkpoint. All five models were then finetuned on nablaColors-3D to predict experimental peak absorption wavelengths. Full details of the pretraining and finetuning procedures, as well as the evaluation protocols, are provided in Supplementary Methods.

The role of the solvent is substantial but secondary relative to the chromophore geometry. Solvent contribution explains only a minor fraction of the overall variance in the dataset. Therefore, we deliberately avoid heavy 3D modeling of the solvent and instead test two complementary strategies: (1) using geometries optimized with an implicit solvent model, and (2) enriching the chromophore representation with a solvent embedding.

For the second approach, we combine a 3D GNN backbone for the chromophore with a lightweight 2D Chemprop encoder [110] for the solvent. This choice is motivated by the relatively small set of unique solvents in the dataset, which does not justify the use of a complex model with a large parameter count. A message-passing architecture like Chemprop provides a sufficiently expressive yet parameter-efficient approximation of solvent effects.

The Chemprop encoder was pre-trained on SMILES strings from the training subset of nablaColors to predict optical properties, using hyperparameters optimized for this task. In this pretraining stage, each sample corresponds to a chromophore–solvent pair: the chromophore SMILES and the solvent SMILES are processed by two Chemprop MPNNs with identical architectures and hyperparameters, but independent weights. The resulting chromophore and solvent latent embeddings are concatenated and passed to a regression head to predict the, peak absorption wavelength. After pretraining, we reuse the solvent Chemprop MPNN as a fixed source of solvent embeddings in our 3D-GNN experiments. During 3D model training and evaluation, only the solvent SMILES is passed through this pretrained Chemprop encoder to obtain a solvent embedding, which is then combined with the chromophore representation produced by the 3D GNN backbone (Figure 4A). This design allows the 3D GNN to focus on capturing geometric and electronic effects of the chromophore, while the Chemprop encoder provides a compact and data-efficient representation of solvent effects. Implementation details, including the Chemprop version used in this study, are provided in Supplementary Methods.

Training is performed in two phases: in the first phase, both pretrained encoders are frozen while the regression head is trained to convergence; in the second phase, the entire network, including both Chemprop and the 3D GNN backbone, is unfrozen for a joint finetune. Full training schedules, architectural variants, optimization hyperparameters, and experimental setup details are provided in Supplementary Information Table S1.

4.3 UniMol+ for peak absorption wavelength prediction

UniMol+ [80] incorporates a geometry-refinement objective during pretraining. Each training instance comprises a low-cost input conformation (e.g. xTB-optimized) and a high-accuracy reference geometry optimized at the DFT level of theory. The network is trained jointly to predict (i) reference coordinates from the input coarse structure (see Lu et al. [80] for details) and (ii) the target property (i.e., peak absorption wavelength). By learning to map cheap input conformers to DFT-quality geometries during training, this approach retains the accuracy associated with DFT-optimized structures while using only low-cost at inference. This feature enables us to avoid the computational bottleneck caused by conformation generation that constrains 3D GNNs, which achieve their best accuracy only with DFT-optimized inputs.

ARTICLE IN PRESS

We propose UniProp, a solvent-aware variant of UniMol+ in which the 2D Chemprop encoder is combined with the UniMol+ transformer backbone to incorporate solvent effects. After projecting the Chemprop-derived solvent embedding onto the atom feature dimension, we sum it with the virtual node embedding, a specialized token analogous to the [CLS] token in BERT [111] (Figure 4B), used as a global molecular representation for property prediction. This allows solvent effects to influence the intermediate molecular representations within the model, rather than being introduced only at the final prediction stage. Such integration aligns with chemical intuition, as solvent interactions impact both the molecule structure and spectral properties of the chromophore, leading to a notable improvement in prediction accuracy. This UniProp architecture delivers the best performance according to our benchmark, outperforming the best 2D model by more than 30% in MAE (see Figure 1).

To further evaluate the robustness of our results, we performed 5-fold scaffold-based cross-validation for UniProp on the single-target absorption task, using a dataset containing only absorption annotations. This cross-validation was run independently from that used for the multitarget setting, ensuring that both evaluations reflect performance under comparable but non-overlapping data splits.

4.4 Benchmarking Setup

Despite significant progress in machine learning for molecular property prediction, the field has lacked a unified and high-quality benchmark for optical and photoluminescence properties. Existing datasets often suffer from corrupted SMILES strings or lack 3D geometries, making it difficult to assess model performance under realistic and controlled conditions.

To address this gap, we introduce a new benchmark based on the nablaColors-3D dataset, which combines experimental measurements of absorption maxima, emission maxima, and photoluminescence quantum yields with optimized molecular geometries across multiple levels of quantum chemical theory, including solvent-aware methods. This benchmark is designed to facilitate rigorous and fair evaluation of both model architectures and conformer fidelity for spectral property prediction.

We follow the scaffold split strategy proposed in Greenman et al. [30], using Bemis–Murcko scaffold partitioning to ensure that structurally similar molecules do not appear in both training and evaluation sets. The dataset is divided into training, validation, and test splits in an 80%/10%/10% ratio. The validation set is used to select hyperparameters and choose the best-performing model checkpoint. Performance is evaluated using Mean Absolute Error (MAE, in nanometers) on the held-out test set.

We benchmark several state-of-the-art 2D graph neural networks (GNNs) for absorption peak prediction, including Chemprop, Chemprop-multifidelity [30], SOG [32], CGIB [31], and FluorPredictor [107]. All 2D models (except for Chemprop, see Section 4.2 for details) were trained from scratch using hyperparameters from their official implementations. As a reference for physically grounded prediction, we include TD-DFT results for the test set computed at the B3LYP/def2-TZVPP [81, 82] level of theory with implicit solvent modeling.

Finally, to investigate how the quality of molecular geometry affects prediction accuracy, we trained a separate 3D GNN for each of the three conformer types in our dataset. Each model was trained using identical splits and hyperparameters, allowing us to isolate the effect of conformer fidelity on model performance.

4.5 Multitarget property prediction

To further evaluate the robustness of our best-performing 3D architecture from the single-target benchmark, we tested its performance in a multitarget setting where the model was trained to jointly predict maximum absorption wavelength, maximum emission wavelength, and photoluminescence quantum yield (PLQY). We used the solvent-aware UniProp variant of UniMol+ with the same architectural configuration and hyperparameters as in the single-target absorption experiments (see Section 4.3); multitarget-specific settings and training schedule are given in Supplementary Methods. Missing labels in the extended dataset were handled via loss masking, ensuring that gradients were computed only for available annotations.

For both the multitarget and single-target cases, we independently applied a 5-fold scaffold-based cross-validation protocol. Performance was evaluated on the held-out folds using MAE for each property reported in the Supplementary Table 3. For reference, we also trained Chemprop [30] on the same multitarget dataset using the same cross-validation setup. Chemprop was chosen as the

ARTICLE IN PRESS

2D model with the strongest performance from previous optical property studies. This comparison serves to contextualize UniProp's performance in the multitarget regime relative to a competitive 2D baseline.

Acknowledgements. This work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

5 Data availability

The dataset and model checkpoints supporting the findings of this study are openly available at https://zenodo.org/records/18061300 [112].

6 Code availability

The associated code, The code supporting the findings of this study is openly available on GitHub at https://github.com/AI4DD/nablaColors [113].

7 Competing interests

The authors declare no competing interests.

8 Author Contributions

Denis Potapov: Data curation, Validation, Conceptualization, Investigation, Software, Formal analysis, Visualization, Writing—original draft. Sergei Rogovoi: Data curation, Formal analysis, Investigation, Software, Writing—original draft. Kuzma Khrabrov: Data curation, Investigation, Writing—review & editing. Konstantin Ushenin: Visualization, Writing—review & editing. Alexey Korovin: Writing—review & editing. Anton Ber: Data curation. Artur Kadurin: Supervision, Resources, Writing—review & editing. Artem Tsypin: Supervision, Conceptualization, Formal analysis, Software, Visualization, Writing—original draft.

ARTICLE IN PRESS

References

[1] Chen, J., Liu, H., Guo, J., Wang, J., Qiu, N., Xiao, S., Chi, J., Yang, D., Ma, D., Zhao, Z., *et al.*: Robust luminescent molecules with high-level reverse intersystem crossing for efficient near ultraviolet organic light-emitting diodes. Angewandte Chemie International Edition **61**(10), 202116810 (2022)

[2] Kim, H.J., Yasuda, T.: Narrowband emissive thermally activated delayed fluorescence materials. Advanced Optical Materials **10**(22), 2201714 (2022)

[3] Uoyama, H., Goushi, K., Shizu, K., Nomura, H., Adachi, C.: Highly efficient organic light-emitting diodes from delayed fluorescence. Nature **492**(7428), 234–238 (2012)

[4] Yang, Z., Mao, Z., Xie, Z., Zhang, Y., Liu, S., Zhao, J., Xu, J., Chi, Z., Aldred, M.P.: Recent advances in organic thermally activated delayed fluorescence materials. Chemical Society Reviews **46**(3), 915–1016 (2017)

[5] Venkatraman, V., Alsberg, B.K.: A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes. Dyes and Pigments **114**, 69–77 (2015)

[6] Tortorella, S., De Angelis, F., Cruciani, G.: Quantitative structure-property relationship modeling of small organic molecules for solar cells applications. Journal of Chemometrics **32**(2), 2957 (2018)

[7] Wen, Y., Fu, L., Li, G., Ma, J., Ma, H.: Accelerated discovery of potential organic dyes for dye-sensitized solar cells by interpretable machine learning models and virtual screening. Solar Rrl **4**(6), 2000110 (2020)

ARTICLE IN PRESS

[8] Liu, R., Cai, Y., Park, J.-M., Ho, K.-M., Shinar, J., Shinar, R.: Organic light-emitting diode sensing platform: Challenges and solutions. Advanced Functional Materials **21**(24), 4744–4753 (2011)

[9] Shinar, J., Shinar, R.: Organic light-emitting devices (oleds) and oled-based chemical and biological sensors: an overview. Journal of Physics D: Applied Physics **41**(13), 133001 (2008)

[10] Su, Y., Yu, B., Wang, S., Cong, H., Shen, Y.: Nir-ii bioimaging of small organic molecule. Biomaterials **271**, 120717 (2021)

[11] Li, B., Zhao, M., Feng, L., Dou, C., Ding, S., Zhou, G., Lu, L., Zhang, H., Chen, F., Li, X., *et al.*: Organic nir-ii molecule with long blood half-life for in vivo dynamic vascular imaging. Nature communications **11**(1), 3102 (2020)

[12] Arenas-Vivo, A., Beltrán, F., Alcázar, V., La Orden, M., Urreaga, J.M.: Fluorescence labeling of high density polyethylene for identification and separation of selected containers in plastics waste streams. comparison of thermal and photochemical stability of different fluorescent tracers. Materials Today Communications **12**, 125–132 (2017)

[13] Resch-Genger, U., Grabolle, M., Cavaliere-Jaricot, S., Nitschke, R., Nann, T.: Quantum dots versus organic dyes as fluorescent labels. Nature methods **5**(9), 763–775 (2008)

[14] Zhang, B., Sun, L.: Artificial photosynthesis: opportunities and challenges of molecular catalysts. Chemical Society Reviews **48**(7), 2216–2264 (2019)

[15] Gao, Y., Nie, W., Wang, X., Fan, F., Li, C.: Advanced space-and time-resolved techniques for photocatalyst studies. Chemical Communications **56**(7), 1007–1021 (2020)

[16] Tang, J., Xiao, Z., Xu, K.: Ultra-thin metamaterial absorber with extremely bandwidth for solar cell and sensing applications in visible region. Optical Materials **60**, 142–147 (2016)

[17] Picollo, M., Aceto, M., Vitorino, T.: Uv-vis spectroscopy. Physical sciences reviews **4**(4), 20180008 (2019)

[18] Casida, M.E., Huix-Rotllant, M.: Progress in time-dependent density-functional theory. Annual review of physical chemistry **63**(1), 287–323 (2012)

[19] Liang, J., Feng, X., Hait, D., Head-Gordon, M.: Revisiting the performance of time-dependent density functional theory for electronic excitations: Assessment of 43 popular and recently developed functionals from rungs one to four. Journal of chemical theory and computation **18**(6), 3460–3473 (2022)

[20] Loos, P.-F., Scemama, A., Blondel, A., Garniron, Y., Caffarel, M., Jacquemin, D.: A mountaineering strategy to excited states: Highly accurate reference energies and benchmarks. Journal of chemical theory and computation **14**(8), 4360–4379 (2018)

[21] Tripathy, V., Flood, A.H., Raghavachari, K.: Accelerated computer-aided screening of optical materials: Investigating the potential of δ-scf methods to predict emission maxima of large dye molecules. The Journal of Physical Chemistry A **128**(39), 8333–8345 (2024)

[22] Whitfield, J.D., Yung, M.-H., Tempel, D.G., Boixo, S., Aspuru-Guzik, A.: Computational complexity of time-dependent density functional theory. New Journal of Physics **16**(8), 083035 (2014)

[23] Lin, Z., Kohn, A.W., Van Voorhis, T.: Toward prediction of nonradiative decay pathways in organic compounds ii: Two internal conversion channels in bodipys. The Journal of Physical Chemistry C **124**(7), 3925–3938 (2020)

[24] Laurent, A.D., Jacquemin, D.: Td-dft benchmarks: a review. International Journal of Quantum Chemistry **113**(17), 2019–2039 (2013)

ARTICLE IN PRESS

[25] Ju, C.-W., Bai, H., Li, B., 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), 1053–1065 (2021)

[26] Zhao, Y., Chen, K., Zhu, L., Huang, Q.: Data-driven machine learning models for quick prediction of the stokes shift of organic fluorescent materials. Dyes and Pigments **220**, 111670 (2023)

[27] Kang, B., Seok, C., Lee, J.: A benchmark study of machine learning methods for molecular electronic transition: Tree-based ensemble learning versus graph neural network. Bulletin of the Korean Chemical Society **43**(3), 328–335 (2022)

[28] Mahato, K.D., Kumar, U.: Optimized machine learning techniques enable prediction of organic dyes photophysical properties: Absorption wavelengths, emission wavelengths, and quantum yields. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **308**, 123768 (2024)

[29] Jung, S.G., Jung, G., Cole, J.M.: Automatic prediction of peak optical absorption wavelengths in molecules using convolutional neural networks. Journal of Chemical Information and Modeling **64**(5), 1486–1501 (2024)

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

[31] Lee, N., Hyun, D., Na, G.S., Kim, S., Lee, J., Park, C.: Conditional graph information bottleneck for molecular relational learning. In: International Conference on Machine Learning, pp. 18852–18871 (2023). PMLR

[32] Sun, M., Fu, C., Su, H., Xiao, R., Shi, C., Lu, Z., Pu, X.: Enhancing chemistry-intuitive feature learning to improve prediction performance of optical properties. Chemical Science **15**(42), 17533–17546 (2024)

[33] Hung, S.-H., Ye, Z.-R., Cheng, C.-F., Chen, B., Tsai, M.-K.: Enhanced predictions for the experimental photophysical data using the featurized schnet-bondstep approach. Journal of Chemical Theory and Computation **19**(14), 4559–4567 (2023)

[34] McNaughton, A.D., Joshi, R.P., Knutson, C.R., Fnu, A., Luebke, K.J., Malerich, J.P., Madrid, P.B., Kumar, N.: Machine learning models for predicting molecular uv–vis spectra with quantum mechanical properties. Journal of Chemical Information and Modeling **63**(5), 1462–1471 (2023)

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

[36] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**, 273–297 (1995)

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

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

[39] Morgan, H.L.: The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation **5**(2), 107–113 (1965)

[40] Schütt, K., Kindermans, P.-J., Sauceda Felix, H.E., Chmiela, S., Tkatchenko, A., Müller, K.-R.: Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems **30** (2017)

[41] Schütt, K.T., Sauceda, H.E., Kindermans, P.-J., Tkatchenko, A., Müller, K.-R.: Schnet–a deep

ARTICLE IN PRESS

learning architecture for molecules and materials. The Journal of Chemical Physics **148**(24) (2018)

[42] Fang, J., Zhang, S., Wu, C., Yang, Z., Liu, Z., Li, S., Wang, K., Du, W., Wang, X.: Moltc: Towards molecular relational modeling in language models. In: Findings of the Association for Computational Linguistics ACL 2024, pp. 1943–1958 (2024)

[43] Gasteiger, J., Groß, J., Günnemann, S.: Directional message passing for molecular graphs. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=B1eWbxStPH

[44] Gasteiger, J., Giri, S., Margraf, J.T., Günnemann, S.: Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In: Machine Learning for Molecules Workshop, NeurIPS (2020)

[45] Gasteiger, J., Becker, F., Günnemann, S.: Gemnet: Universal directional graph neural networks for molecules. Advances in Neural Information Processing Systems **34**, 6790–6802 (2021)

[46] Liao, Y.-L., Smidt, T.: Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=KwmPfARgOTD

[47] Liao, Y.-L., Wood, B.M., Das, A., Smidt, T.: Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=mCOBKZmrzD

[48] Passaro, S., Zitnick, C.L.: Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In: International Conference on Machine Learning, pp. 27420–27438 (2023). PMLR

[49] Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C.J., Kornbluth, M., Kozinsky, B.: Learning local equivariant representations for large-scale atomistic dynamics. Nature Communications **14**(1), 579 (2023)

[50] Batatia, I., Kovacs, D.P., Simm, G., Ortner, C., Csányi, G.: Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in neural information processing systems **35**, 11423–11436 (2022)

[51] Wang, Y., Wang, T., Li, S., He, X., Li, M., Wang, Z., Zheng, N., Shao, B., Liu, T.-Y.: Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. Nature Communications **15**(1), 313 (2024)

[52] Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C.J., Ceder, G.: Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence **5**(9), 1031–1041 (2023)

[53] Qu, E., Krishnapriyan, A.: The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. Advances in Neural Information Processing Systems **37**, 139030–139053 (2024)

[54] Lahey, S.-L.J., Rowley, C.N.: Simulating protein–ligand binding with neural network potentials. Chemical science **11**(9), 2362–2368 (2020)

[55] Tsypin, A., Ugadiarov, L.A., Khrabrov, K., Telepov, A., Rumiantsev, E., Skrynnik, A., Panov, A., Vetrov, D.P., Tutubalina, E., Kadurin, A.: Gradual optimization learning for conformational energy minimization. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=FMMF1a9ifL

[56] Tayfuroglu, O., Zengin, I.N., Koca, M.S., Kocak, A.: Deepconf: Leveraging ani-ml potentials for exploring local minima with application to bioactive conformations. Journal of Chemical Information and Modeling **65**(6), 2818–2833 (2025)

ARTICLE IN PRESS

[57] Moore, J.H., Cole, D.J., Csanyi, G.: Computing hydration free energies of small molecules with first principles accuracy. arXiv preprint arXiv:2405.18171 (2024)

[58] Rupp, M., Tkatchenko, A., Müller, K.-R., Von Lilienfeld, O.A.: Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters **108**(5), 058301 (2012)

[59] Ramakrishnan, R., Dral, P.O., Rupp, M., Von Lilienfeld, O.A.: Quantum chemistry structures and properties of 134 kilo molecules. Scientific data **1**(1), 1–7 (2014)

[60] Eastman, P., Behara, P.K., Dotson, D.L., Galvelis, R., Herr, J.E., Horton, J.T., Mao, Y., Chodera, J.D., Pritchard, B.P., Wang, Y., *et al.*: Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. Scientific Data **10**(1), 11 (2023)

[61] Eastman, P., Pritchard, B.P., Chodera, J.D., Markland, T.E.: Nutmeg and spice: models and data for biomolecular machine learning. Journal of chemical theory and computation **20**(19), 8583–8593 (2024)

[62] Khrabrov, K., Shenbin, I., Ryabov, A., Tsypin, A., Telepov, A., Alekseev, A., Grishin, A., Strashnov, P., Zhilyaev, P., Nikolenko, S., *et al.*: nabladft: Large-scale conformational energy and hamiltonian prediction benchmark and dataset. Physical Chemistry Chemical Physics **24**(42), 25853–25863 (2022)

[63] Khrabrov, K., Ber, A., Tsypin, A., Ushenin, K., Rumiantsev, E., Telepov, A., Protasov, D., Shenbin, I., Alekseev, A.M., Shirokikh, M., Nikolenko, S., Tutubalina, E., Kadurin, A.: $\nablaˆ2$DFT: A universal quantum chemistry dataset of drug-like molecules and a benchmark for neural network potentials. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024). https://openreview.net/forum?id=ElUrNM9U8c

[64] Isert, C., Atz, K., Jiménez-Luna, J., Schneider, G.: Qmugs, quantum mechanical properties of drug-like molecules. Scientific Data **9**(1), 273 (2022)

[65] Nakata, M., Shimazaki, T.: Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. Journal of chemical information and modeling **57**(6), 1300–1308 (2017)

[66] Smith, J.S., Isayev, O., Roitberg, A.E.: Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. Scientific data **4**(1), 1–8 (2017)

[67] Smith, J.S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A.E., Isayev, O., Tretiak, S.: The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. Scientific data **7**(1), 134 (2020)

[68] Zhang, D., Bi, H., Dai, F.-Z., Jiang, W., Liu, X., Zhang, L., Wang, H.: Pretraining of attention-based deep learning potential model for molecular simulation. npj Computational Materials **10**(1), 94 (2024)

[69] Shoghi, N., Kolluru, A., Kitchin, J.R., Ulissi, Z.W., Zitnick, C.L., Wood, B.M.: From molecules to materials: Pre-training large generalizable models for atomic property prediction. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=PfPnugdxup

[70] Urquhart, R.J., Teijlingen, A., Tuttle, T.: Ani neural network potentials for small molecule p k a prediction. Physical Chemistry Chemical Physics **26**(36), 23934–23943 (2024)

[71] Shiota, T., Ishihara, K., Mizukami, W.: Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. Digital Discovery **3**(9), 1714–1728 (2024)

ARTICLE IN PRESS

ARTICLE IN PRESS

[72] Raja, A., Zhao, H., Tyrchan, C., Nittinger, E., Bronstein, M.M., Deane, C., Morris, G.M.: On the effectiveness of quantum chemistry pre-training for pharmacological property prediction. In: ICML 2024 AI for Science Workshop (2024)

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

[74] Ksenofontov, A.A., Lukanov, M.M., Bocharov, P.S., Berezin, M.B., Tetko, I.V.: Deep neural network model for highly accurate prediction of bodipys absorption. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **267**, 120577 (2022)

[75] Shao, J., Liu, Y., Yan, J., Yan, Z.-Y., Wu, Y., Ru, Z., Liao, J.-Y., Miao, X., Qian, L.: Prediction of maximum absorption wavelength using deep neural networks. Journal of Chemical Information and Modeling **62**(6), 1368–1375 (2022)

[76] Venkatraman, V., Kallidanthiyil Chellappan, L.: An open access data set highlighting aggregation of dyes on metal oxides. Data **5**(2), 45 (2020)

[77] Venkatraman, V., Raju, R., Oikonomopoulos, S.P., Alsberg, B.K.: The dye-sensitized solar cell database. Journal of Cheminformatics **10**, 1–9 (2018)

[78] Bemis, G.W., Murcko, M.A.: The properties of known drugs. 1. molecular frameworks. Journal of medicinal chemistry **39**(15), 2887–2893 (1996)

[79] Schütt, K., Unke, O., Gastegger, M.: Equivariant message passing for the prediction of tensorial properties and molecular spectra. In: International Conference on Machine Learning, pp. 9377–9388 (2021). PMLR

[80] Lu, S., Gao, Z., He, D., Zhang, L., Ke, G.: Data-driven quantum chemical property prediction leveraging 3d conformations with uni-mol+. Nature communications **15**(1), 7104 (2024)

[81] Becke, A.D.: Density-functional thermochemistry. iii. the role of exact exchange. The Journal of chemical physics **98**(7), 5648–5652 (1993)

[82] Weigend, F., Ahlrichs, R.: Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. Physical Chemistry Chemical Physics **7**(18), 3297–3305 (2005)

[83] Reichardt, C.: Solvatochromic dyes as solvent polarity indicators. Chemical reviews **94**(8), 2319–2358 (1994)

[84] Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., Wang, H.: Geometry-enhanced molecular representation learning for property prediction. Nature Machine Intelligence **4**(2), 127–134 (2022)

[85] Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., Liò, P.: 3d infomax improves gnns for molecular property prediction. In: International Conference on Machine Learning, pp. 20479–20502 (2022). PMLR

[86] He, Z., Bi, H., Liang, B., Li, Z., Zhang, H., Wang, Y.: Frontier molecular orbital weighted model based networks for revealing organic delayed fluorescence efficiency. Light: Science & Applications **14**(1), 75 (2025)

[87] Grimme, S., Hansen, A., Ehlert, S., Mewes, J.-M.: r2scan-3c: A "swiss army knife" composite electronic-structure method. The Journal of Chemical Physics **154**(6) (2021)

[88] Lu, S., Gao, Z., He, D., Zhang, L., Ke, G.: Highly accurate quantum chemical property prediction with uni-mol+. arXiv preprint arXiv:2303.16982 (2023)

[89] Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., Leskovec, J.: Ogb-lsc: A large-scale challenge for machine learning on graphs. arXiv preprint arXiv:2103.09430 (2021)

ARTICLE IN PRESS

[90] Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., *et al.*: Open catalyst 2020 (oc20) dataset and community challenges. Acs Catalysis **11**(10), 6059–6072 (2021)

[91] Fourches, D., Muratov, E., Tropsha, A.: Trust, but verify: on the importance of chemical structure curation in cheminformatics and qsar modeling research. Journal of chemical information and modeling **50**(7), 1189 (2010)

[92] Kim, Y., Cho, J., Naser, N., Kumar, S., Jeong, K., McCormick, R.L., John, P.C.S., Kim, S.: Physics-informed graph neural networks for predicting cetane number with systematic data quality analysis. Proceedings of the Combustion Institute **39**(4), 4969–4978 (2023)

[93] Haghighatlari, M., Li, J., Heidar-Zadeh, F., Liu, Y., Guan, X., Head-Gordon, T.: Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. Chem **6**(7), 1527–1542 (2020)

[94] Li, S.-C., Wu, H., Menon, A., Spiekermann, K.A., Li, Y.-P., Green, W.H.: When do quantum mechanical descriptors help graph neural networks to predict chemical properties? Journal of the American Chemical Society **146**(33), 23103–23120 (2024)

[95] Körner, P., Glüge, J., Glüge, S., Scheringer, M.: Critical insights into data curation and label noise for accurate prediction of aerobic biodegradability of organic chemicals. Environmental Science: Processes & Impacts **26**(10), 1780–1795 (2024)

[96] Gadaleta, D., Lombardo, A., Toma, C., Benfenati, E.: A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. Journal of cheminformatics **10**(1), 60 (2018)

[97] Pracht, P., Bohle, F., Grimme, S.: Automated exploration of the low-energy chemical space with fast quantum chemical methods. Physical Chemistry Chemical Physics **22**(14), 7169–7192 (2020)

[98] Beard, E.J., Sivaraman, G., Vázquez-Mayagoitia, Á., Vishwanath, V., Cole, J.M.: Comparative dataset of experimental and computational attributes of uv/vis absorption spectra. Scientific data **6**(1), 307 (2019)

[99] Zou, Z., Zhang, Y., Liang, L., Wei, M., Leng, J., Jiang, J., Luo, Y., Hu, W.: A deep learning model for predicting selected organic molecular spectra. Nature Computational Science **3**(11), 957–964 (2023)

[100] Wang, S., Witek, J., Landrum, G.A., Riniker, S.: Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. Journal of chemical information and modeling **60**(4), 2044–2058 (2020)

[101] Tosco, P., Stiefl, N., Landrum, G.: Bringing the mmff force field to the rdkit: implementation and validation. Journal of Cheminformatics **6**(1), 37 (2014) https://doi.org/10.1186/s13321-014-0037-3

[102] Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., Grimme, S.: Extended tight-binding quantum chemistry methods. Wiley Interdisciplinary Reviews: Computational Molecular Science **11**(2), 1493 (2021)

[103] Gasevic, T., Stückrath, J.B., Grimme, S., Bursch, M.: Optimization of the r2scan-3c composite electronic-structure method for use with slater-type orbital basis sets. The Journal of Physical Chemistry A **126**(23), 3826–3838 (2022)

[104] Neese, F.: The orca program system. Wiley Interdisciplinary Reviews: Computational Molecular Science **2**(1), 73–78 (2012)

[105] Barone, V., Cossi, M.: Quantum calculation of molecular energies and energy gradients in

ARTICLE IN PRESS

solution by a conductor solvent model. The Journal of Physical Chemistry A **102**(11), 1995–2001 (1998)

[106] Bursch, M., Mewes, J.-M., Hansen, A., Grimme, S.: Best-practice dft protocols for basic molecular computational chemistry. Angewandte Chemie **134**(42), 202205735 (2022)

[107] Song, W., Xiong, L., Li, X., Zhang, Y., Wang, B., Liu, G., Li, W., Yang, Y., Tang, Y.: Fluorpredictor: An interpretable tool for multiproperty prediction and retrieval of fluorescent dyes. Journal of Chemical Information and Modeling **65**(6), 2854–2867 (2025)

[108] Bilgili, H., Yakalı, G., Kaya, N., Karaman, M., Demiç, S.: Structure–property relationship on aggregation-induced enhanced emission of the spirobifluorene derivatives including herringbone, t, and helical aggregation mode in solid-phase: synthesis, crystallography, density functional theory, optical, and thermal properties. The Journal of Physical Chemistry C **125**(8), 4893–4908 (2021)

[109] Lupo Pasini, M., Mehta, K., Yoo, P., Irle, S.: Two excited-state datasets for quantum chemical uv-vis spectra of organic molecules. Scientific Data **10**(1), 546 (2023)

[110] Heid, E., Greenman, K.P., Chung, Y., Li, S.-C., Graff, D.E., Vermeire, F.H., Wu, H., Green, W.H., McGill, C.J.: Chemprop: a machine learning package for chemical property prediction. Journal of Chemical Information and Modeling **64**(1), 9–17 (2023)

[111] Devlin, J., Chang, M.-W., Lee, K., 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), pp. 4171–4186 (2019)

[112] Potapov, D., Rogovoi, S., Khrabrov, K., Ushenin, K., Korovin, A., Ber, A., Kadurin, A., Tsypin, A.: Dataset for the Article: nablaColors: A 3D Benchmark for Optical Property Prediction with Solvent-aware Graph Neural Networks. https://doi.org/10.5281/zenodo.18061300 . https://zenodo.org/records/18061300

[113] Potapov, D., Rogovoi, S., Khrabrov, K., Ushenin, K., Korovin, A., Ber, A., Kadurin, A., Tsypin, A.: Code For: nablaColors: A 3D Benchmark for Optical Property Prediction with Solvent-aware Graph Neural Networks. https://doi.org/10.5281/zenodo.18327022 . https://doi.org/10.5281/zenodo.18327022

ARTICLE IN PRESS

ARTICLE IN PRESS

Fig. 1: Absorption MAE on the test split of our proposed nablaColors-3D benchmark for 2D and 3D GNNs. For 2D GNNs we evaluate Chemprop variants [30], SOG [32], CGIB [31], and FluorPredictor [107]. For the TD-DFT baseline, the MAE is 62 nm. All models except the UniProp use geometries optimized with DFT and CPCM both during training and inference. UniProp model uses pairs of xTB-optimized, and DFT-optimized with CPCM model conformations during training, and xTB-optimized conformations during inference.

Fig. 2: Test MAE (nm) for five 3D GNN backbones across three conformer sources and two solvent regimes. Colors encode conformer source: xTB, DFT in vacuum (DFT-vac), and DFT with implicit solvent via CPCM (DFT-imp). Hatching indicates models without a solvent embedding, whereas solid bars denote models with a Chemprop solvent embedding. All models except the UniProp use the specified type of conformations in both training and inference. UniProp only uses the specified conformation type in training. During inference, UniProp-xTB uses RDKit conformations, and UniProp-DFT-vac/DFT-imp use xTB conformations.

Fig. 3: Data processing pipeline. First, the UV/Vis dataset of absorption spectra is curated to obtain the nablaColors dataset. Then we generate chromophore conformations for each unique chromophore in nablaColors. After that, the chromophore conformations are sequentially optimized with MMFF first, then with xTB (I), and finally with DFT in both vacuum (II) and implicit solvent (III). We refer to the resulting collection of 3D conformations as nablaColors-3D

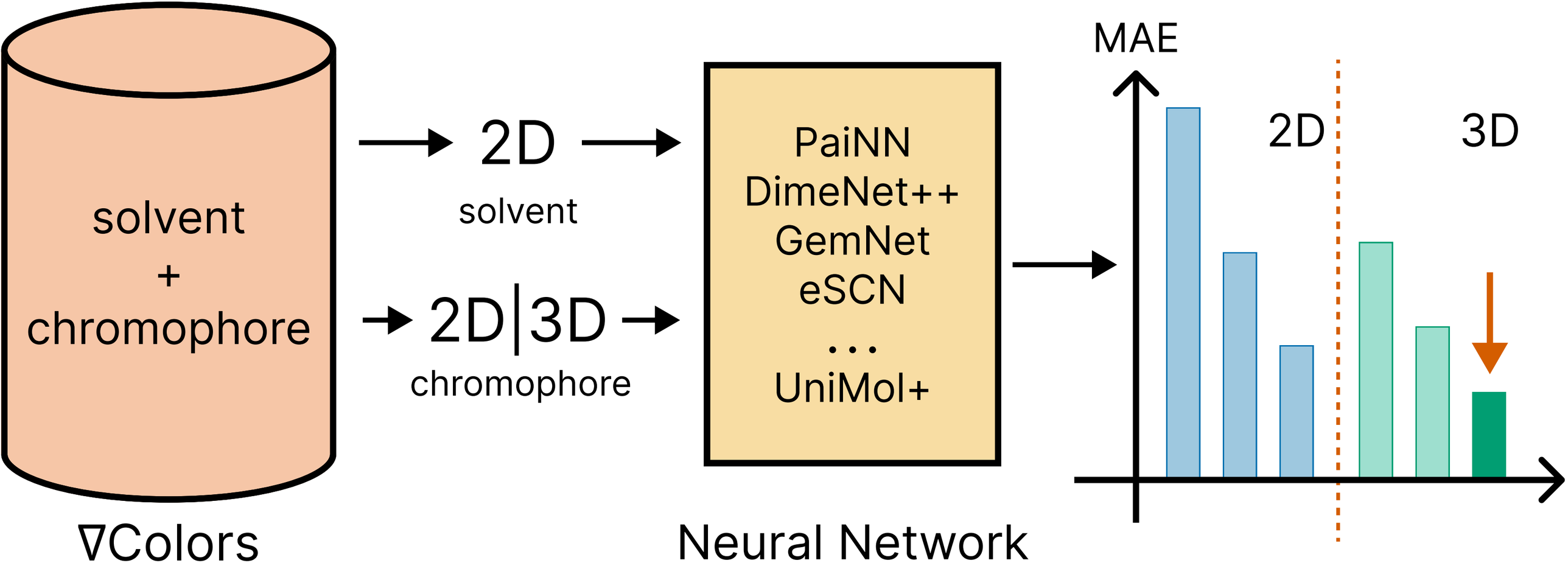
Fig. 4: Architecture of 3D GNNs for optical properties prediction. The solvent is represented as SMILES and processed with a Chemprop model, and A) solvent embeddings are concatenated with final chromophore embeddings obtained for PaiNN, DimeNet++, GemNet, and eSCN, B) solvent embedding added to the virtual node for UniMol+ prior to the self-attention blocks. The predicted target, λ_{abs}, denotes the wavelength corresponding to the maximum of the absorption spectrum (absorption maximum).

Table 1: UniProp performance for different conformation-pair choices used during training. We report row-wise MAE and RMSE (in nm) and the coefficient of determination R^2 on the test set.

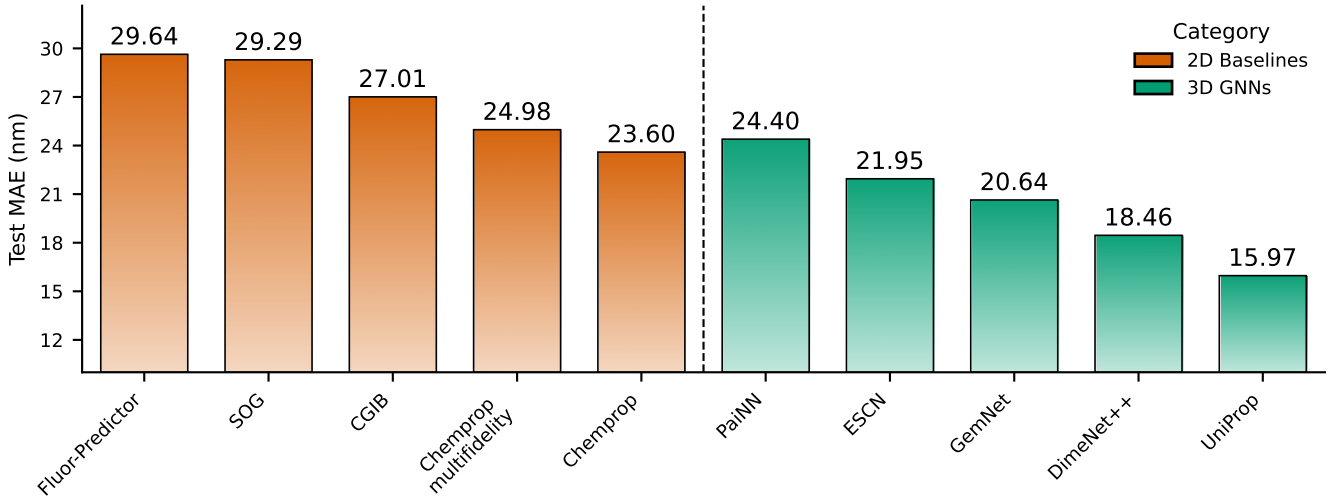
Training pair	MAE	RMSE	R^2
xTB → DFT implicit	16.0	27.2	0.929
RDKit → DFT implicit	15.9	27.6	0.927
RDKit → xTB	16.3	27.8	0.926
xTB → DFT vacuum	16.0	27.8	0.926

Table 2: Five-fold cross-validation MAE for UniProp and Chemprop. Values are mean ± standard deviation across folds; absorption and emission in nm, PLQY unitless.

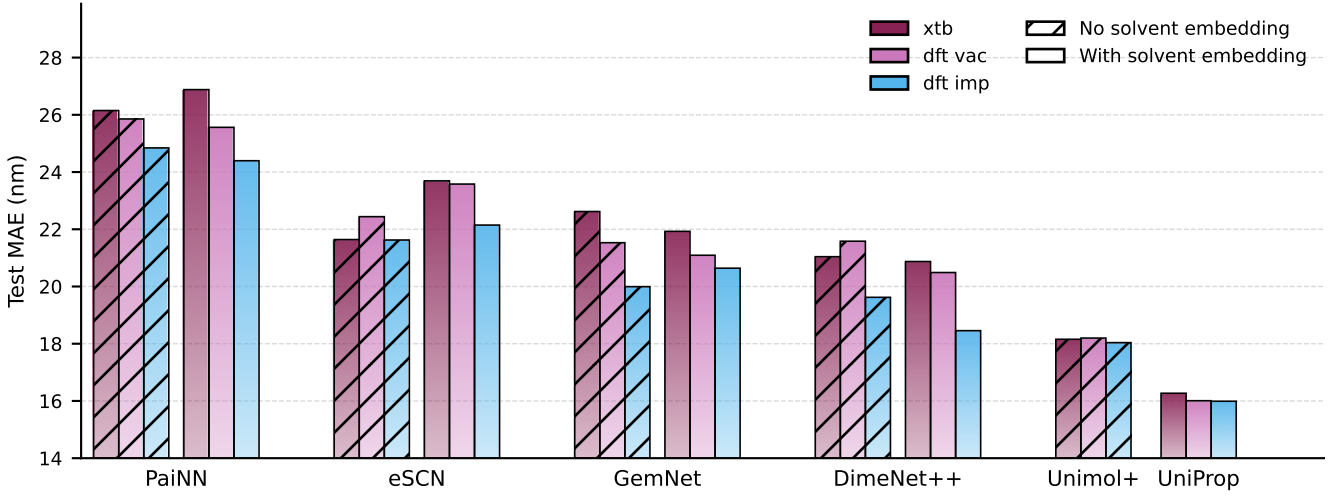
Model	Absorption only	Multitarget		
	Abs MAE	Abs MAE	Ems MAE	PLQY MAE
Chemprop	21.7 ± 2.3	22.0 ± 1.2	28.4 ± 1.4	0.17 ± 0.01
UniProp	14.8 ± 0.7	15.3 ± 1.0	19.7 ± 1.0	0.16 ± 0.01


solvent
+
chromophore
2D
solvent
2D|3D
chromophore
PaiNN
DimeNet++
GemNet
eSCN
...
UniMol+
MAE
2D
3D
∇Colors
Neural Network

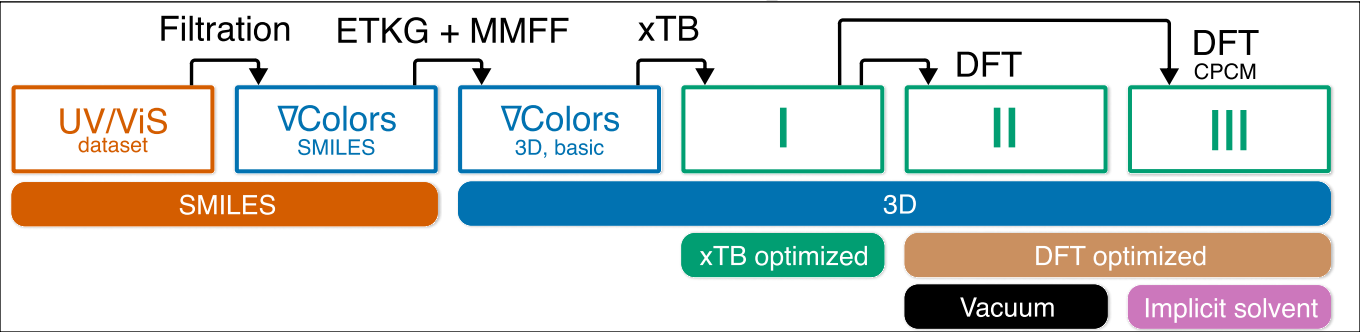
ARTICLE IN PRESS



ARTICLE IN PRESS



ARTICLE IN PRESS



ARTICLE IN PRESS

