Supplementary materials:nablaColors: A 3D 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: rogovoi@airi.net; khrabrov@airi.net; ushenin@airi.net; korovin@airi.net; ber@airi.net; kadurin@airi.net; tsypin@airi.net;

S1 Training details

S1.1 Pretraining setup

All 3D models were initialized from pretrained weights obtained on the the HOMO–LUMO gap prediction task. estimating the HOMO–LUMO energy gap. Hydrogen atoms were excluded from all molecular structures during both training and evaluation to reduce model complexity and computational cost.

PaiNN, DimeNet++, eSCN, and GemNet were pretrained by us on the full PCQM4Mv2 dataset, which contains DFT-calculated orbital properties for approximately 3.8 million molecules. UniMol+ was initialized from the official PCQM4Mv2 checkpoint released by the authors.

S1.2 Evaluation protocol

To select optimal hyperparameters and model checkpoints, we used the validation set from the proposed ∇`Colors-3D` benchmark. The validation set was also used for early stopping and to adjust the learning rate (`ReduceLROnPlateau`). To ensure fair comparison, all models—including different architectures and conformer settings—were trained on identical data splits and corresponding molecular conformations.

S1.3 Finetuning setup

The models were originally trained to predict HOMO-LUMO gap. To finetune on the ∇`Colors-3D`, we replace the regression head with a randomly initialized one. We then explore two strategies: 1) finetuning the model with a randomly initialized regression head end-to-end and 2) pretraining the regression head with a frozen backbone and then training the whole model end-to-end. We call these two setups “random” and “staged”.

While most models showed comparable performance under both protocols, the staged approach was crucial for UniMol+, enabling a more stable training process and improved final performance. This effect was consistent across all conformer types.

S1.4 Training Hyperparameters

Table S1: Key hyperparameters for the models used in this study. For the 3D models (PaiNN, DimeNet++, eSCN, GemNet), we report parameters for the `OneCycleLR` scheduler. An alternative `ReduceLROnPlateau` scheduler was also tested. For Chemprop, we report the hyperparameters from our best-performing configuration.

	Chemprop	**PaiNN**	**DimeNet++**	**eSCN**	**GemNet**	**UniMol+**
			Training Hyperparameters			
Optimizer	Adam	AdamW	AdamW	AdamW	AdamW	Adam
Batch Size	128	32	64	32	32	6
Loss Function	MSE	MAE (L1)	MAE (L1)	MAE (L1)	MAE (L1)	UniMol+ loss
LR Scheduler	Cyclical	OneCycleLR	OneCycleLR	OneCycleLR	OneCycleLR	Polynomial Decay
Max LR	1.0×10^{-3}	5.0×10^{-4}	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}	8.0×10^{-5}
Epochs	300	450	450	450	450	—
Total Steps	—	—	—	—	—	300,000
Warmup	20 epochs	30% of steps	30% of steps	30% of steps	30% of steps	30,000 steps
EMA Decay	—	—	—	—	—	0.999
			Model Architecture			
Layers	5 (depth)	3	6	8	4	6
Hidden size	482	128	128	128/256	256	768
Cutoff (Å)	—	5.0	5.0	8.0	12.0	—
Attention Heads	—	—	—	—	—	48

The main hyperparameters for training the models on absorption maximum wavelength prediction are detailed in Table S1. All models were trained with zero weight decay.

For the multi-target setup, we adopted the same hyperparameters but with key modifications to address the different prediction tasks. To handle the prediction of quantum yield (a value between 0 and 1), we applied a **logit transform** to the target values before calculating the loss. This transformation, defined as $\text{logit}(p) = \log(\frac{p}{1-p})$, maps the target to the range of real numbers, which is better suited for regression. To ensure numerical stability, the quantum yield values were clamped within the range $\left[3 \times 10^{-5}, 0.999 + 1 \times 10^{-5}\right]$ prior to transformation. In addition, we introduce loss weight coefficients. The coefficients were set to $[1, 1, 10]$ for the absorption wavelength, the emission wavelength, and the quantum yield, respectively, to align the quantum yield loss values with the absorption and emission loss values.

S2 Additional experiments

Table S2: Impact of regression-head initialization on *dft implicit-solvent* conformers. Columns show Train/Test MAE (nm) for **Random** vs. **Pretrained** heads

Model	Random		Pretrained	
	Train MAE ↓	Test MAE ↓	Train MAE ↓	Test MAE ↓
DimeNet++	3.522	25.062	2.848	18.457
ESCN	2.651	21.949	2.419	22.147
GemNet	1.790	20.640	1.785	21.316
Unimol+	5.447	22.019	5.491	18.358
UniProp	1.771	19.682	1.788	17.744

S2.1 Impact of Regression Head Initialization

We found that the effectiveness of head pretraining depends on the expressivity of the regression head.

UniMol+, for example, utilizes a simple regression heads that is composed of two linear layers with GELU activation and layer normalization. UniMol+-based models consistently benefited from head pretraining, particularly when trained with lower-quality input geometries.

In contrast, GemNet's default regression head consists of an initial dense layer followed by multiple residual layers. Despite its expressive capacity, this configuration leads to overfitting when the backbone is frozen: the model achieves ∼5 nm MAE on the training set, but ∼30 nm on validation and test sets. This suggests that decoder expressiveness can harm generalization in transfer settings.

Less expressive regression heads used in models like DimeNet++, PaiNN, and eSCN demonstrated more stable behavior, but gained less from head pretraining. We hypothesize that this is due to their inability to leverage pretrained features.

Together, these findings highlight that the benefit of head pretraining depends on a balance between capacity and regularization. Pretrained heads are most helpful when the decoder is expressive enough to benefit from pretrained features, but not so complex as to overfit in low-data or frozen-backbone regimes.

S2.2 Combined Cross-Validation Results

Table S3: Consolidated cross-validation MAE metrics.

Model	Target	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean Val Loss
Single Target							
UniProp	absorption	15.797 (*29.077*)	15.899	14.528	15.217	14.962	15.280 (*17.936*)
Chemprop	absorption	26.034 (*41.780*)	22.109	19.118	20.475	20.970	21.741 (*24.890*)
Multitarget							
	absorption	15.282 (*29.806*)	14.258	14.871	16.251	17.192	15.570 (*18.475*)
UniProp	emission	20.334	17.863	20.250	19.709	21.314	19.894
	PLQY	0.165	0.159	0.162	0.147	0.146	0.155
	absorption	21.938 (*36.900*)	20.144	21.621	22.550	23.787	22.008 (*25.000*)
Chemprop	emission	29.567	26.022	28.741	27.595	29.904	28.366
	PLQY	0.184	0.175	0.177	0.158	0.176	0.174

This section reports full five-fold cross-validation metrics for the best 2D baseline and for our proposed UniProp model. We first train both models in a single-task setting to predict only the peak absorption wavelength. We then train them in a multitask setting in which, for each chromophore–solvent pair, the models jointly predict the peak absorption wavelength, the peak emission wavelength, and the photoluminescence quantum yield (PLQY). For **Fold 1**, we report two MAE values due to the betaine dye36 issue described in Section 3.3. The italicized MAE is computed on the full fold. The non-italicized MAE is computed on the same fold after excluding the ∼ 200 samples in which the chromophore is betaine dye36.

S2.3 Conformer calculation times

In this section, we provide details on time it takes to calculate optimized geometries using different levels of quantum theory.

Table S4: CPU times for conformer calculations using different methods. Times are in seconds, minutes and hours.

Method	Mean	Median	Max	Total time
Orca Solvent	1493.98 s (24.90 min)	373.20 s (6.22 min)	6554.51 min	10940.93 h
Orca Vacuum	1127.32 s (18.79 min)	342.13 s (5.70 min	1939.08 min	4301.04 h
XTB Vacuum	27.54 s (0.46 min)	9.01 s (0.15 min)	25.32 min	118.65 h

S2.4 Training curves

In this section, we provide full training plot for the solvent-aware variants of 3D GNNs. Each row corresponds to a single model. The left subplot in each row depicts training loss, whereas the right subplot depicts the validation loss. In each subplot, we provide three curves, corresponding to different conformation types.

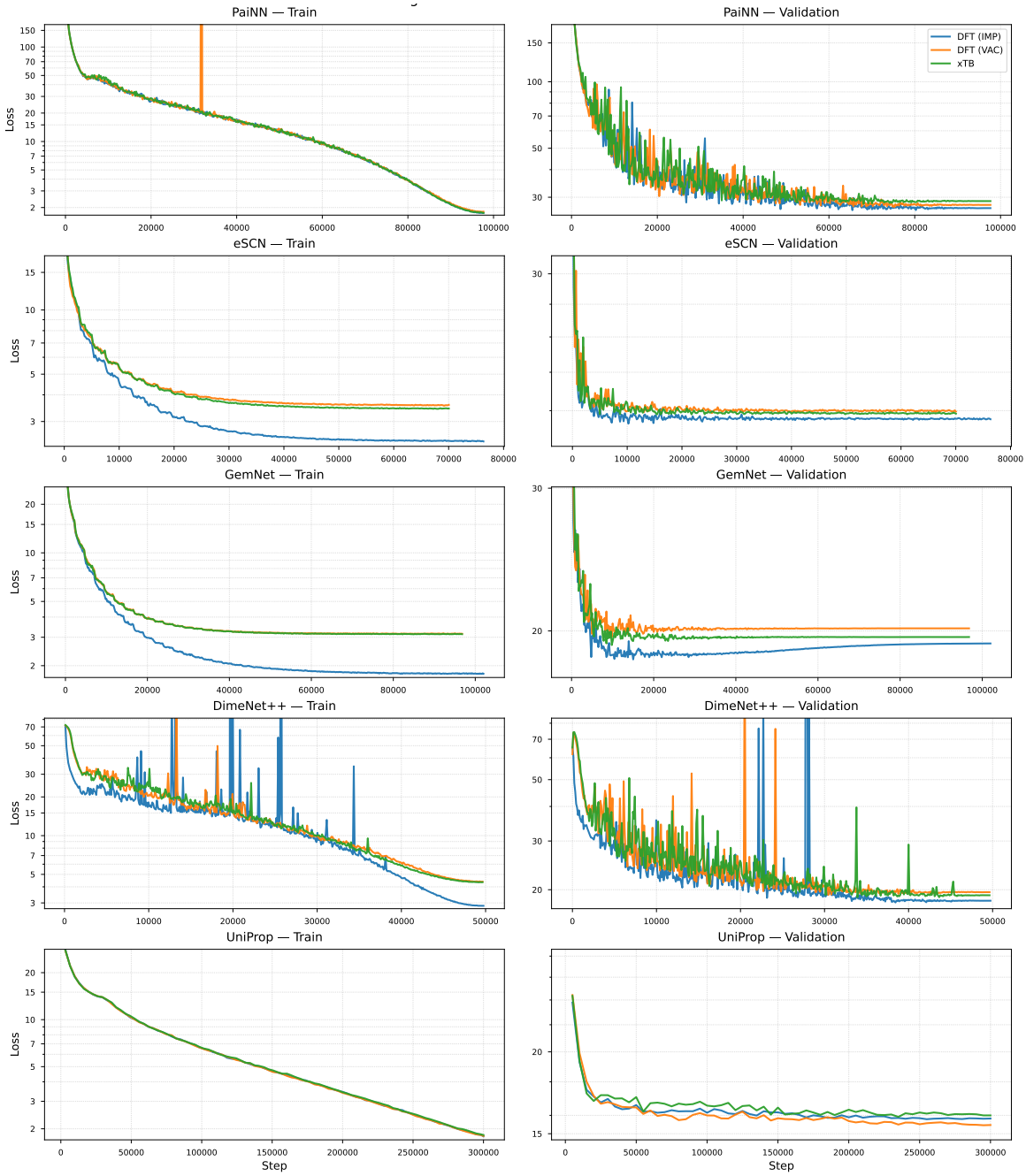

Fig. S1: Training curves for the solvent-aware variants of 3D GNNs. The y-axis is log-scaled.