Supplementary Information (SI) for Digital Discovery.

Supplementary Information

Fluor-tools: An Integrated Platform for Dye Property Prediction and Structure Optimization

Wenxiang Song,[a] Yuyang Zhang,[b] Le Xiong,[a] Xinmin Li,[a] Jingwei Zhang,[a] Guixia Liu,[a] Weihua Li,[a] Youjun Yang*[b] and Yun Tang*[a]

[a] Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

[b] State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China; Shanghai Key Laboratory of Chemical Biology, Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

E-mail: ytang234@ecust.edu.cn

E-mail: youjunyang@ecust.edu.cn

Content

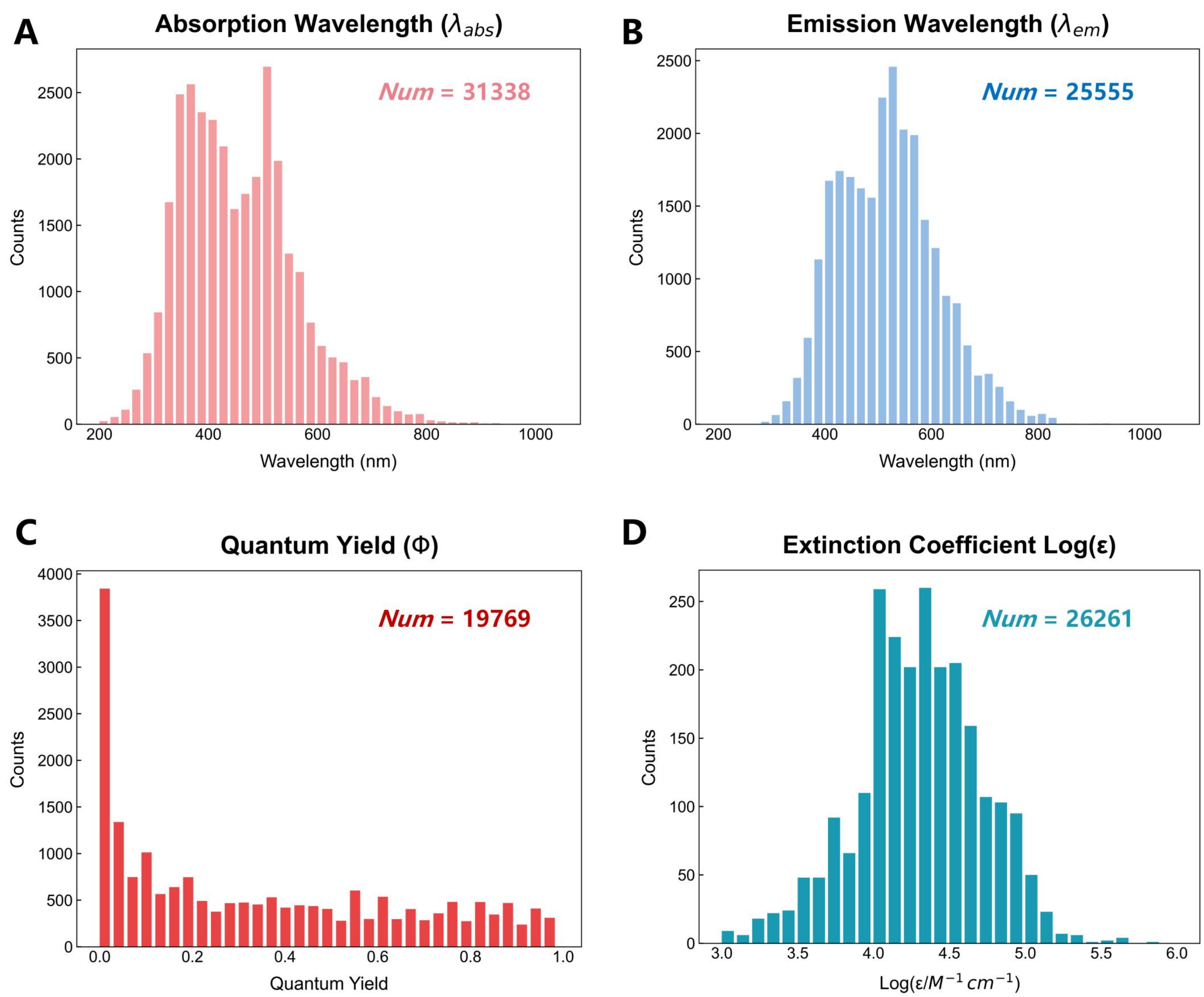
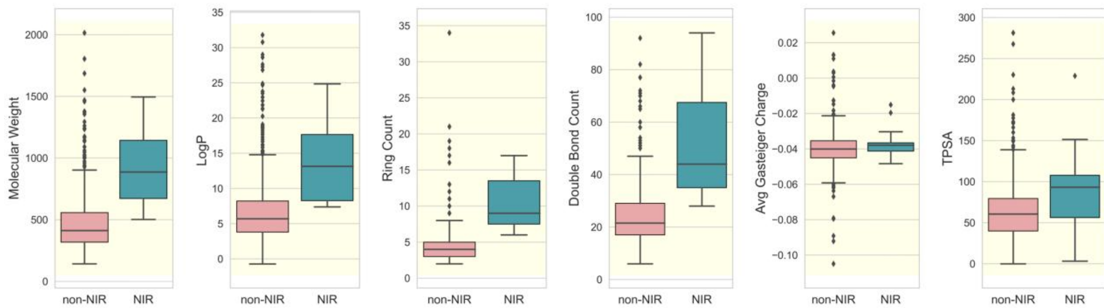
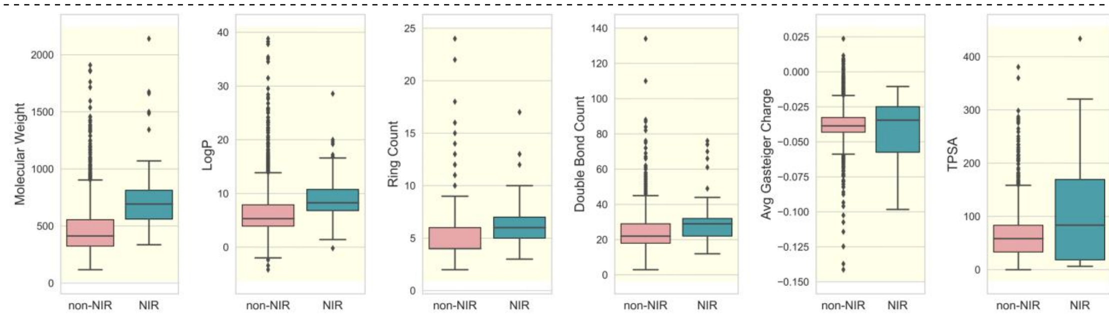


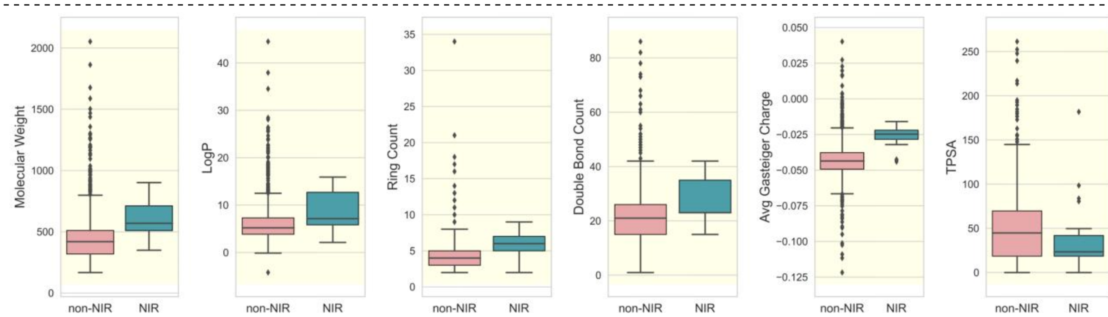
Fig. S1 Property distribution of dyes in the FluoDB dataset.



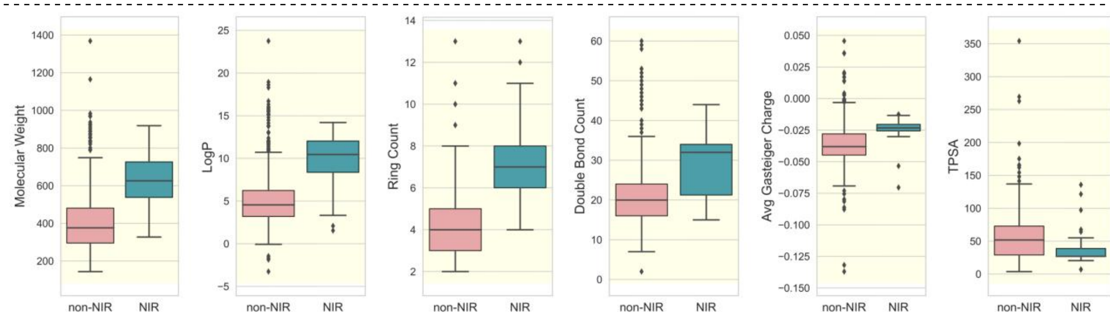
Scaffold: **5n6**



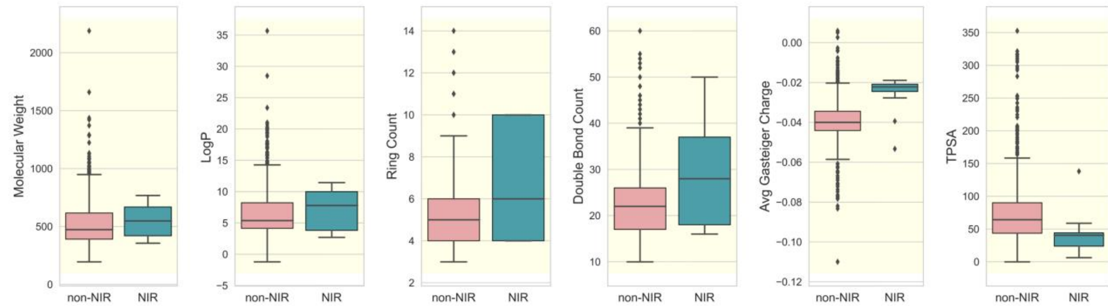
Scaffold: **5p6**



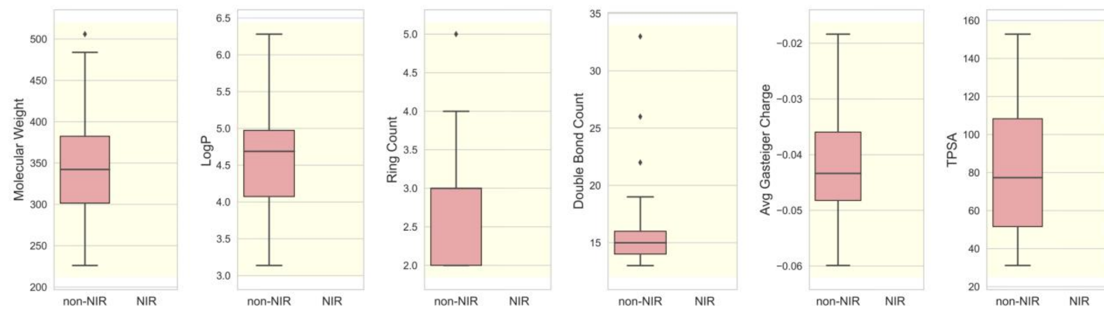
Scaffold: **6n6**



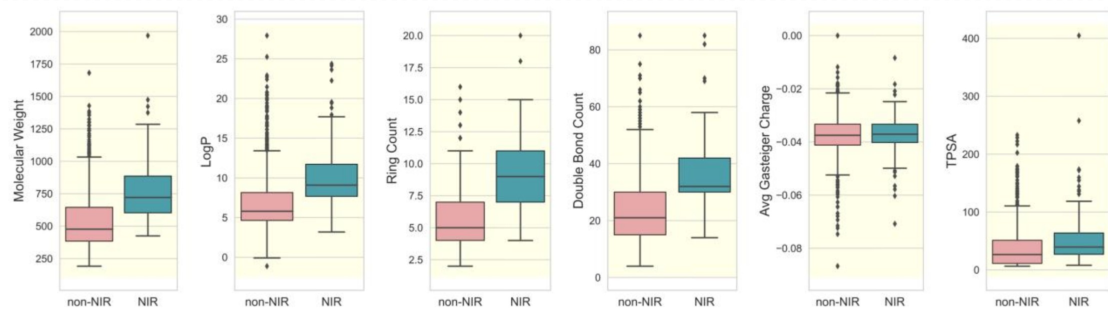
Scaffold: **6p6**



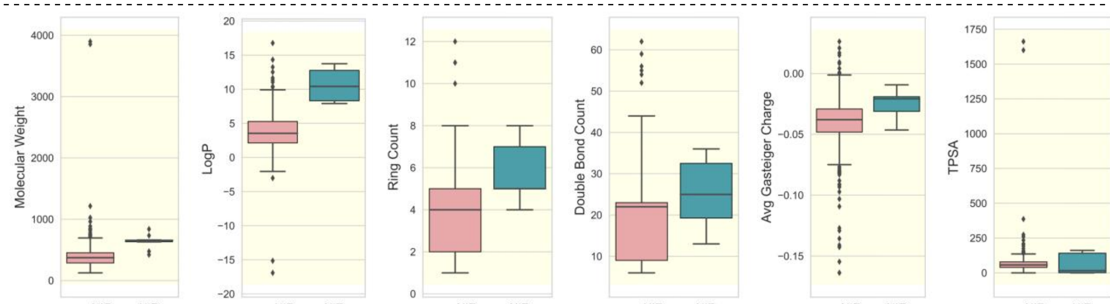
Scaffold: **Acridines**



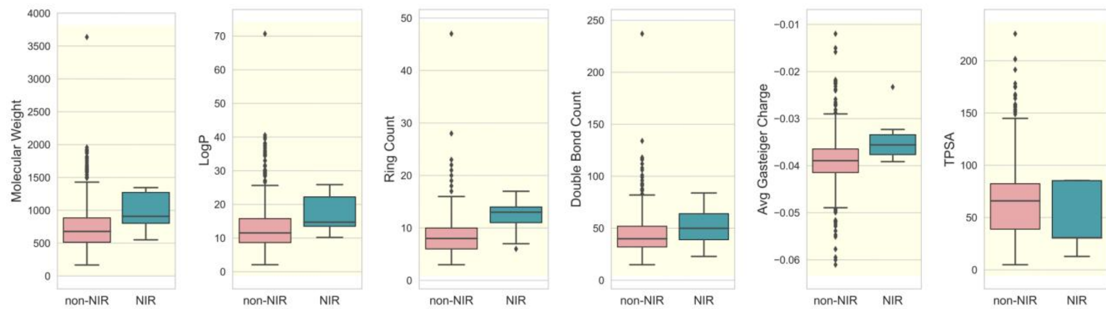
Scaffold: **Azo**



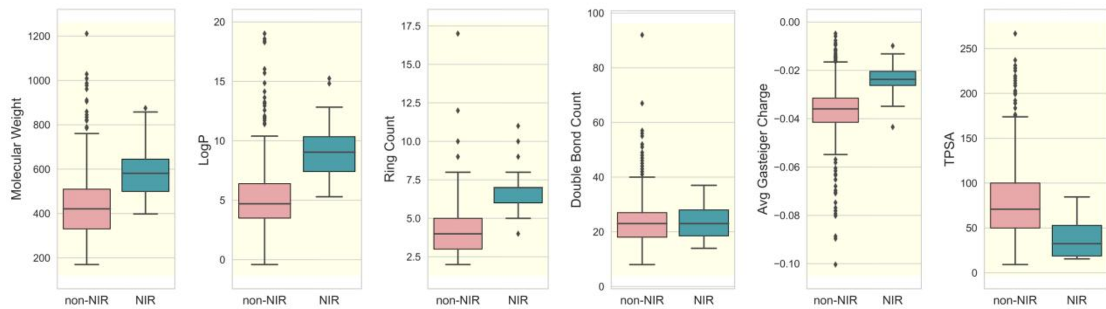
Scaffold: **BODIPY**



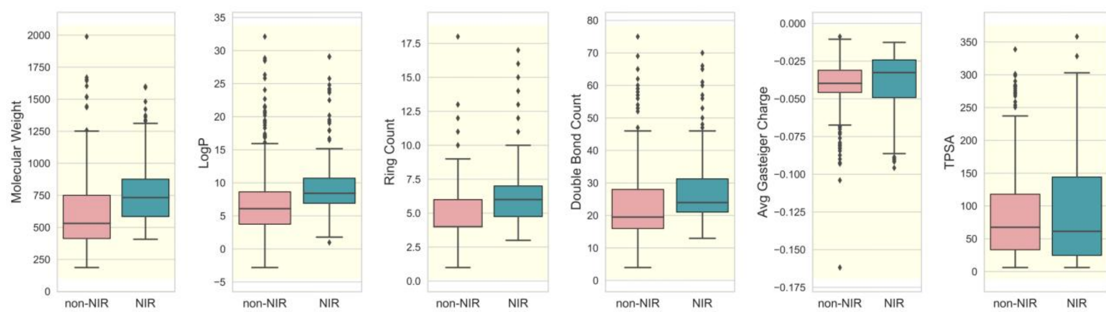
Scaffold: **Benz**



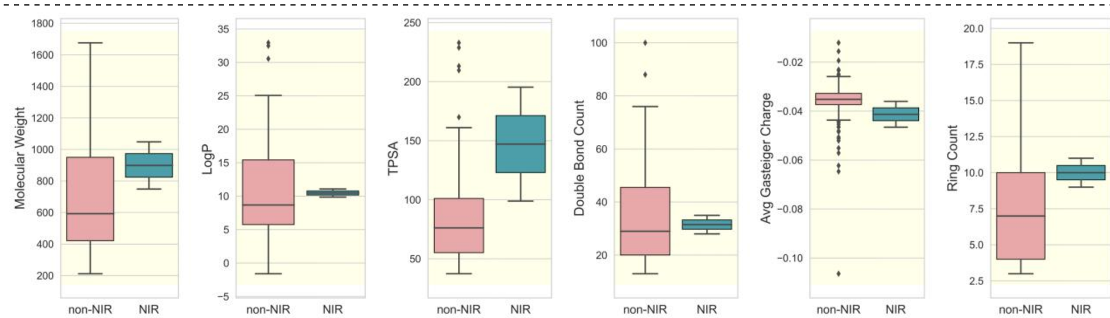
Scaffold: **Carbazole**



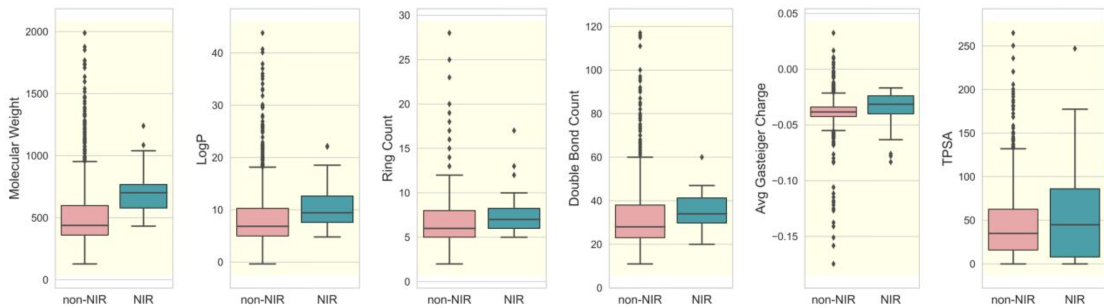
Scaffold: **Coumarin**



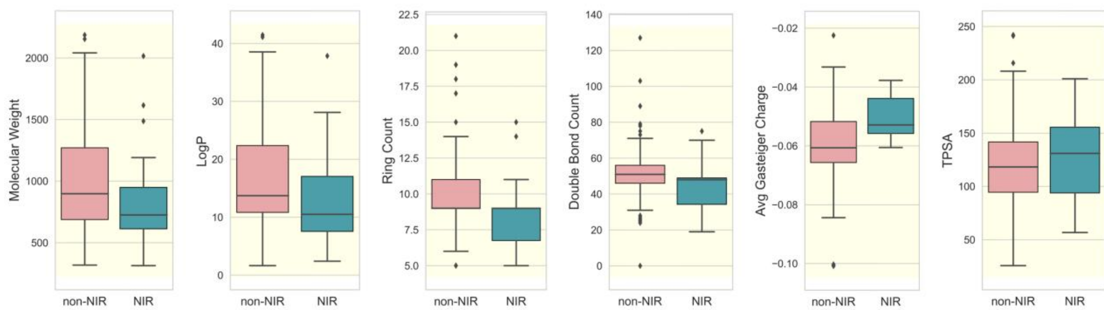
Scaffold: **Cyanine**



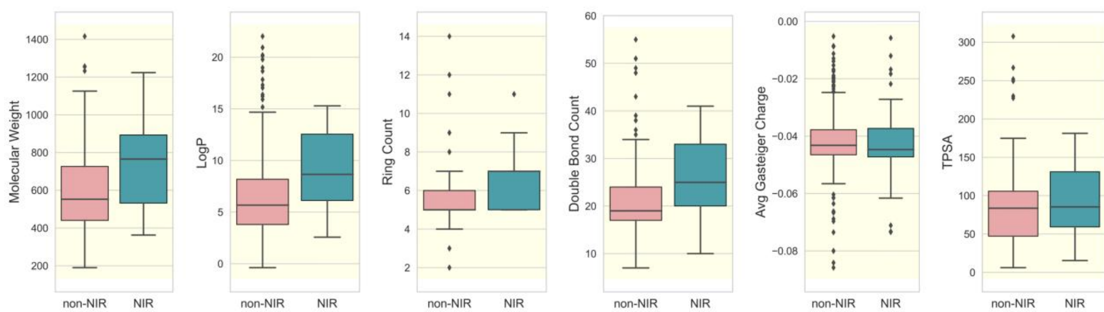
Scaffold: **Naphthalimide**



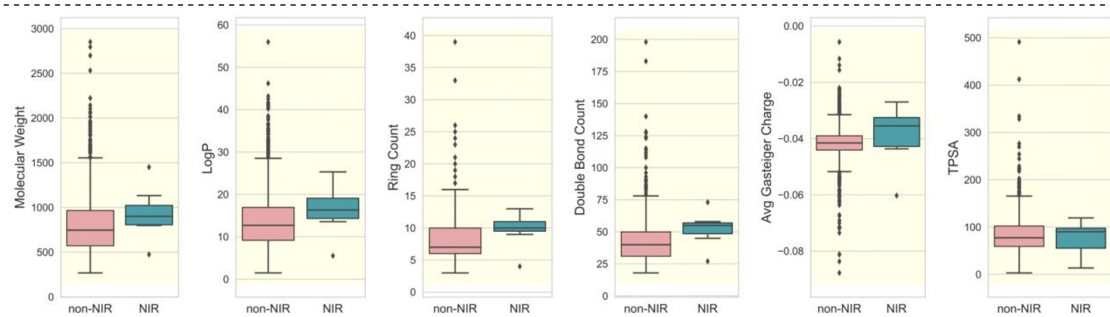
Scaffold: **PAHs**



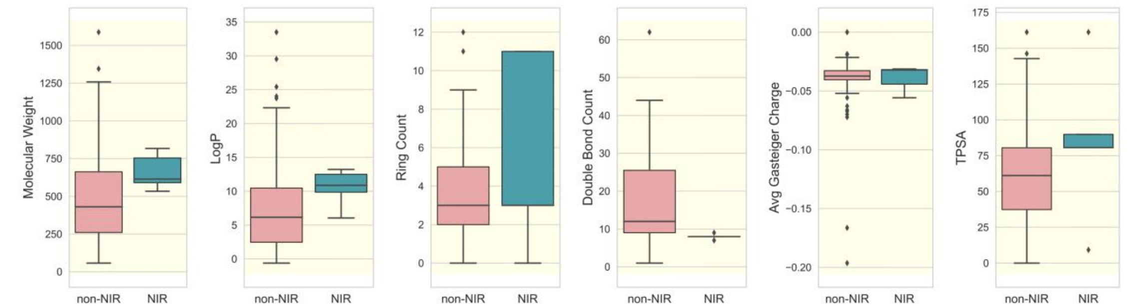
Scaffold: **Porphyrin**



Scaffold: **SquaricAcid**



Scaffold: **Triphenylamine**



Scaffold: **None**

Fig. S2 Comparison of molecular properties and their correlation with NIR across different scaffolds.

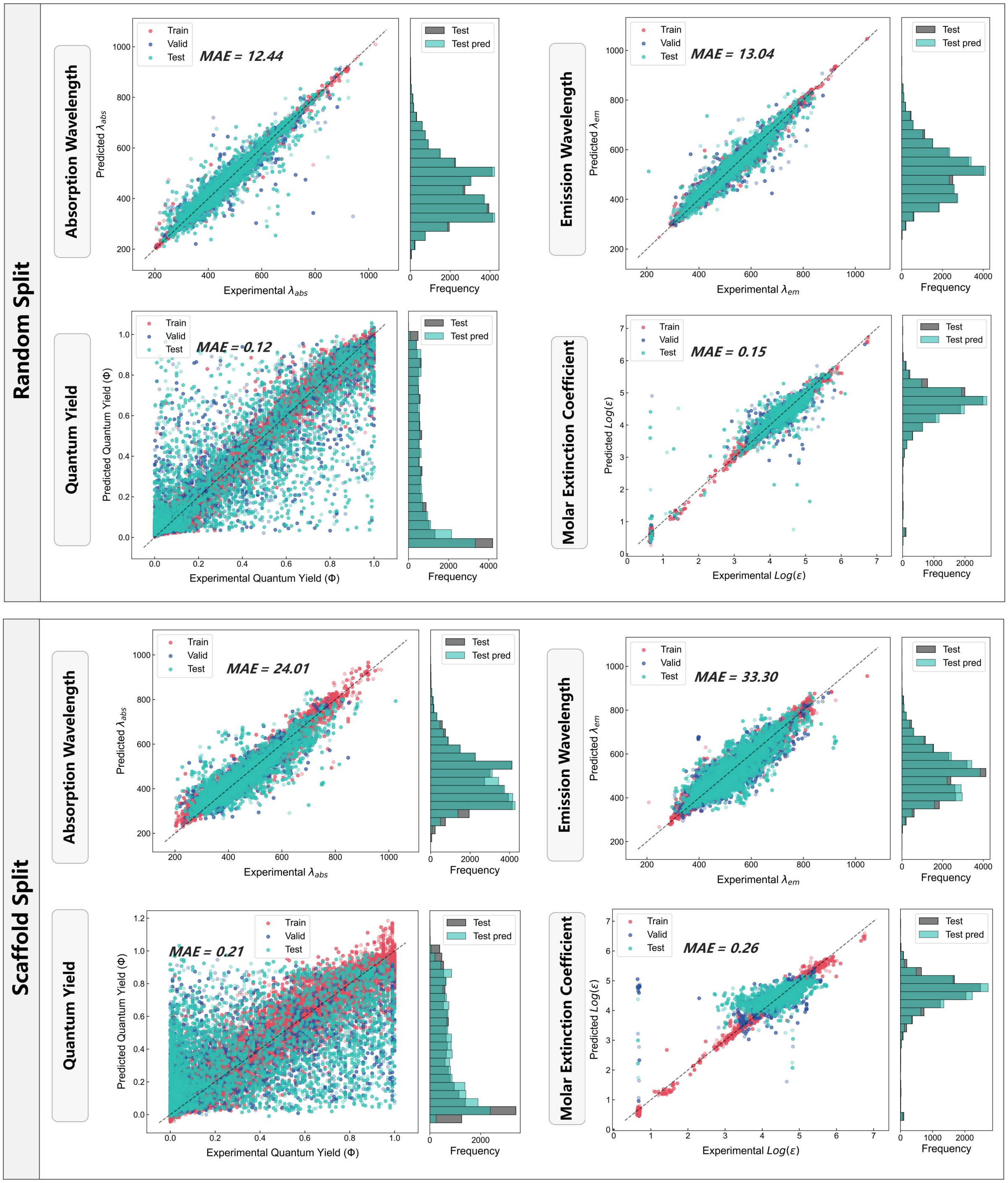


Fig. S3 Comparison between random partitioning and scaffold partitioning.

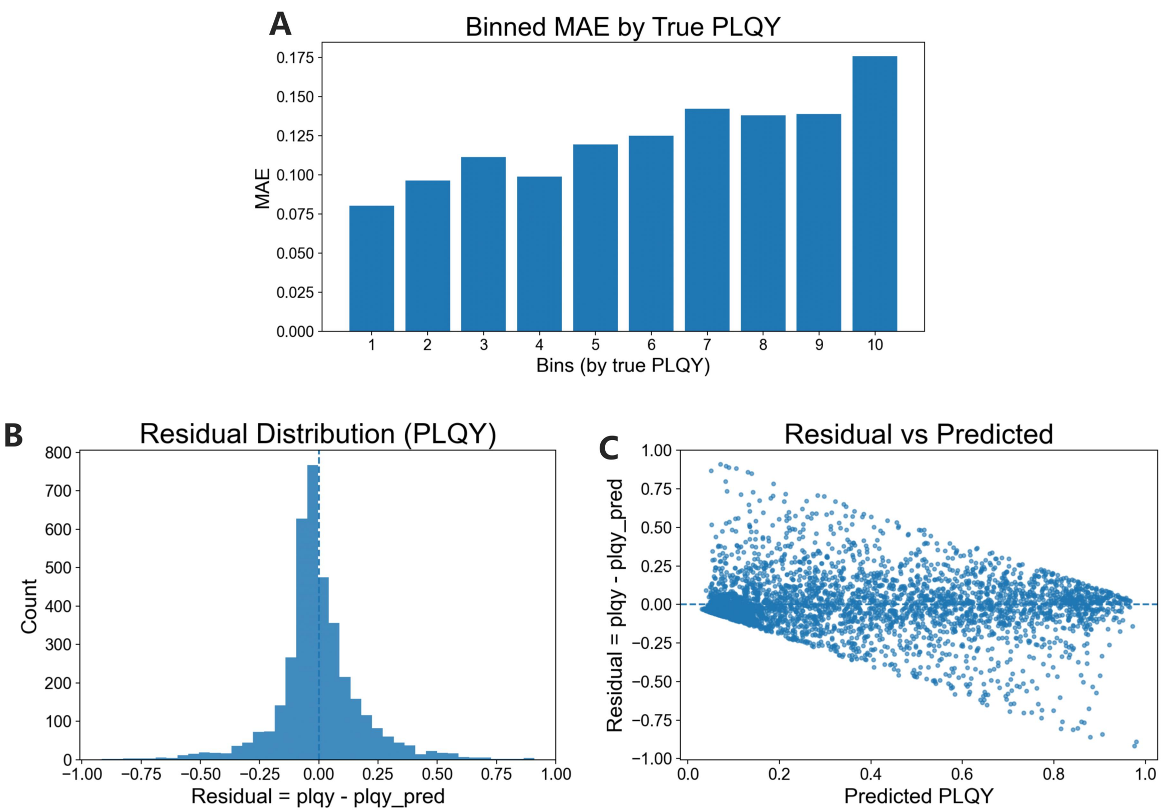


Fig. S4 Quantum yield error analysis. (A) Binned MAE analysis: the quantum yield was divided into 10 bins, and the MAE was calculated for each interval; (B) residual histogram; (C) scatter plot of residuals versus predicted values.

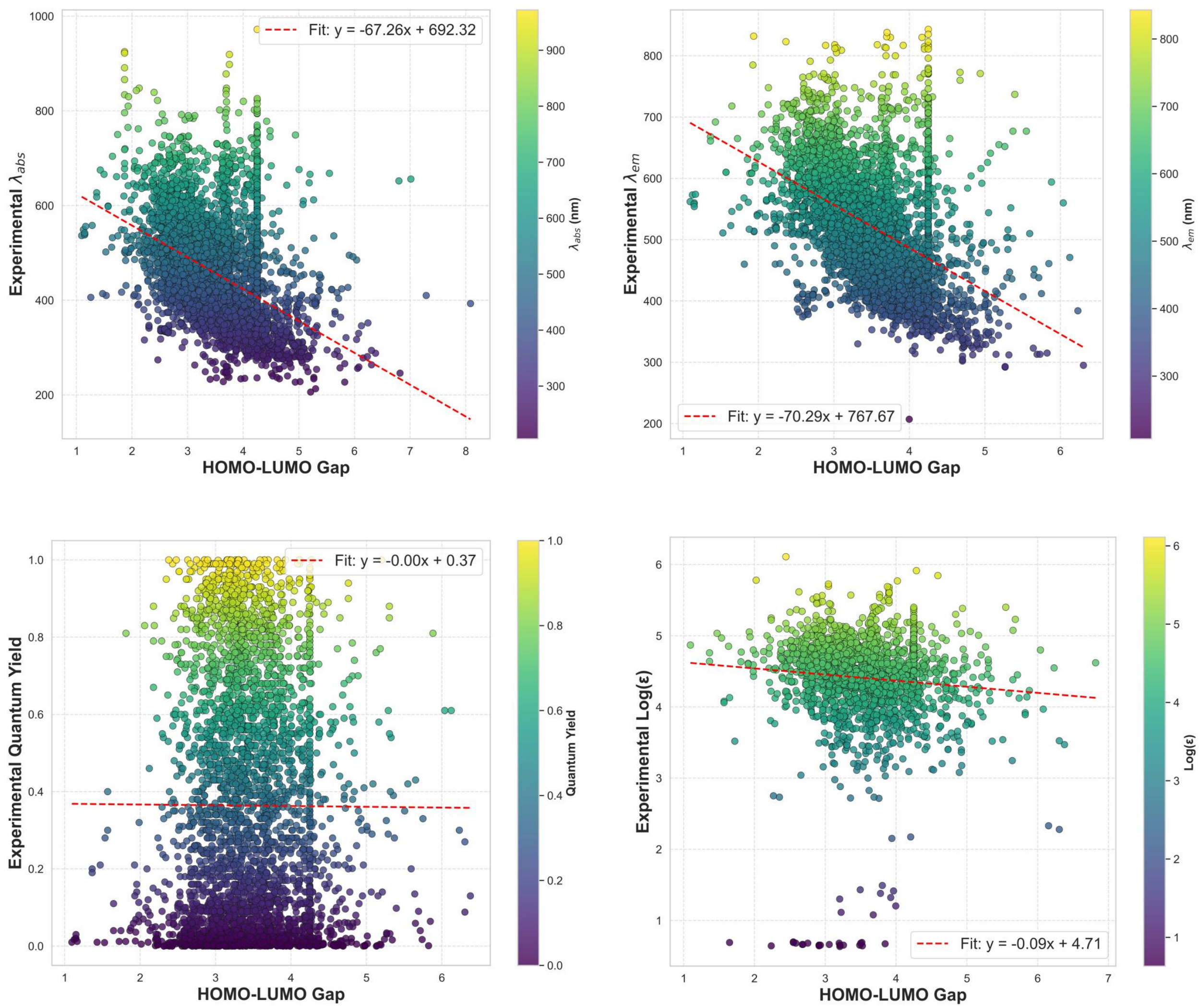


Fig. S5 Relationship between Unimol+ predicted HOMO-LUMO gap and dye properties. It can be observed that there is an obvious linear relationship between the HOMO-LUMO gap and the dye's absorption wavelength as well as emission wavelength.

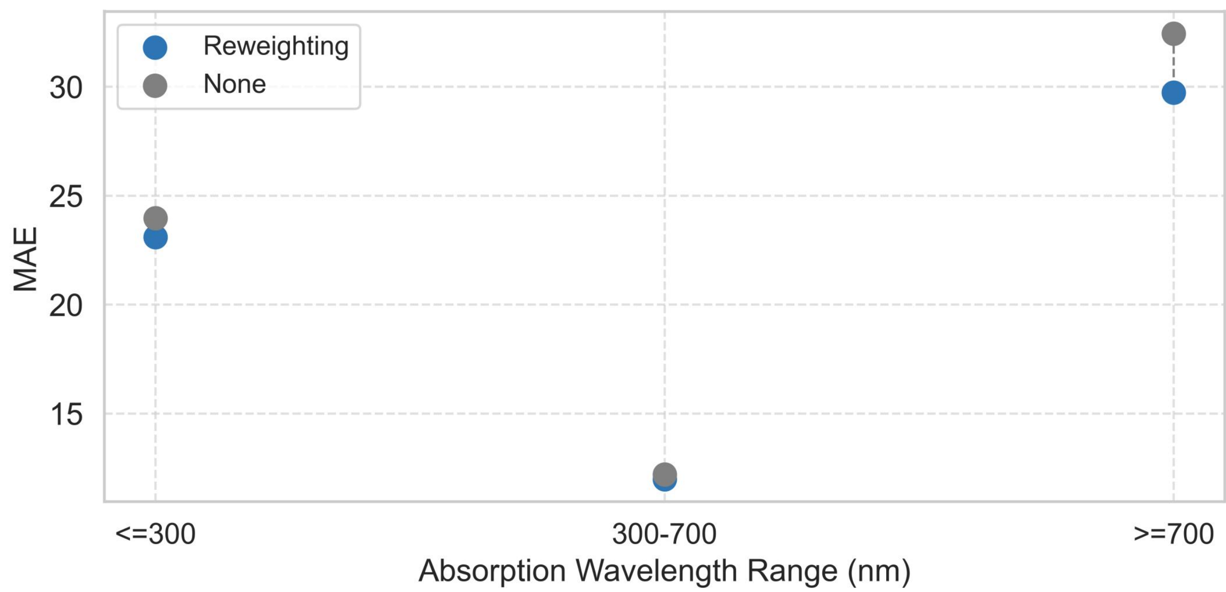


Fig. S6 The specific impact of LDS reweighting on the three absorption wavelength intervals.

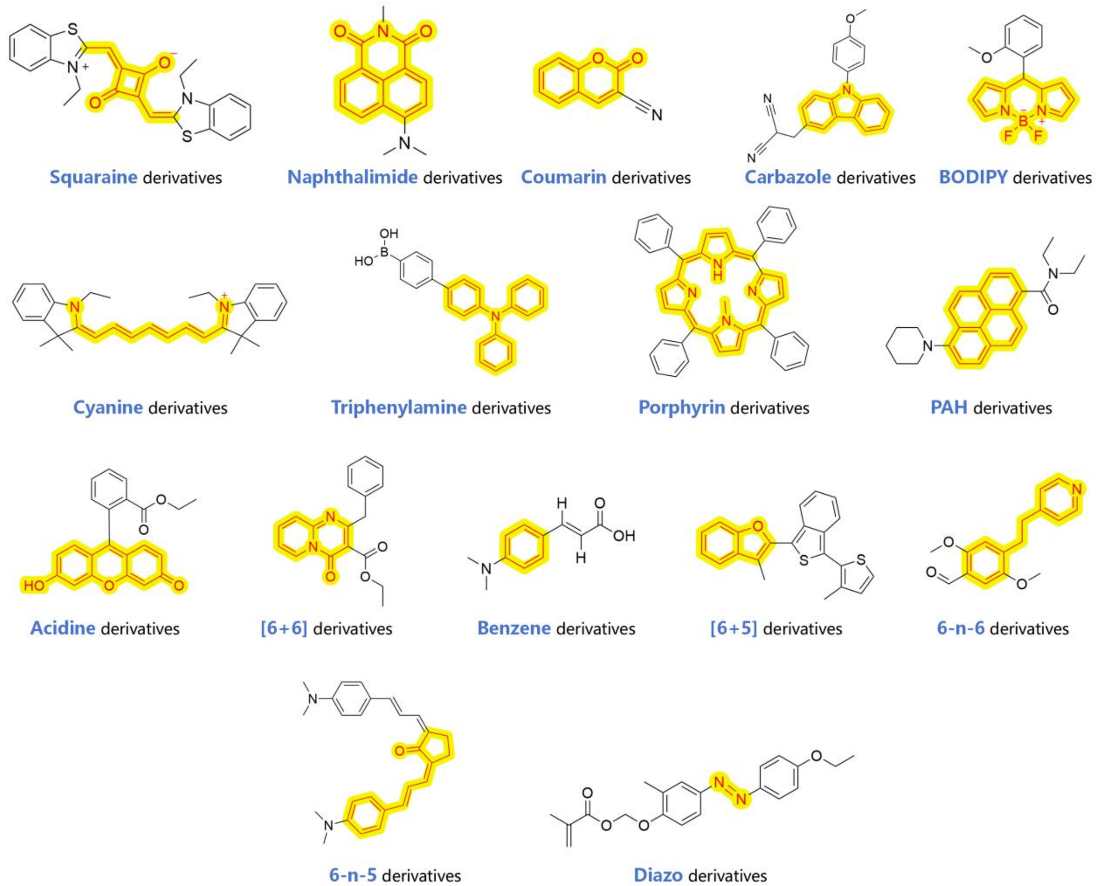


Fig. S7 Representative members of each fluorescent scaffold.

Table S1 Performance of different molecular scaffolds in Fluor-pred across four properties

Scaffold	***MAE (Number of data)***			
	λ_{abs} (nm)	**λ_{em} (nm)**	**Φ_{PL}**	**ε_{max}**
None	27.055 (56)	19.387 (34)	0.105 (22)	0.131 (28)
SquaricAcid	11.359 (83)	12.574 (30)	0.129 (13)	0.068 (11)
Naphthalimide	12.541 (87)	19.189 (70)	0.245 (65)	0.136 (37)
Coumarin	10.609 (272)	13.830 (216)	0.117 (152)	0.126 (119)
Carbazole	12.270 (388)	15.773 (255)	0.111 (179)	0.135 (90)
Cyanine	11.530 (310)	12.125 (258)	0.105 (182)	0.091 (156)
BODIPY	6.784 (1412)	8.809 (1308)	0.113 (1216)	0.178 (374)
Triphenylamine	19.104 (641)	19.106 (312)	0.155 (202)	0.147 (142)
Porphyrin	41.671 (31)	12.293 (20)	0.058 (18)	0.254 (8)
PAHs	13.269 (630)	15.778 (557)	0.127 (437)	0.162 (205)
Acridines	15.43 (242)	15.363 (176)	0.123 (81)	0.217 (77)
5p6	15.878 (688)	15.595 (483)	0.113 (347)	0.144 (251)
6p6	16.157 (259)	16.745 (196)	0.107 (148)	0.167 (116)
5n6	12.385 (361)	12.374 (277)	0.102 (222)	0.136 (135)
6n6	13.084 (591)	12.376 (489)	0.107 (383)	0.136 (209)
Azo	34.337 (10)	/	/	0.059 (8)
Benz	17.097 (197)	14.190 (158)	0.1078 (122)	0.234 (76)

Table S2 Performance comparison of normal sampling strategies for the binary classification model of NIR dyes

Normal Sampling									
Model	*ACC*	*MCC*	*F1 Score*	*Recall*	*Precision*	*SP*	*BA*	*AUC*	*PRAUC*
LGBM	0.981± 0.0018	0.831± 0.0161	0.838± 0.0152	0.778± 0.0238	**0.909± 0.0245**	**0.995± 0.0016**	0.886± 0.0117	0.985± 0.0036	0.919± 0.0139
GBRT	0.981± 0.0037	0.834± 0.0349	0.841± 0.0341	0.786± 0.0473	0.906± 0.0280	0.994± 0.0017	0.890± 0.0239	0.982± 0.0085	0.916± 0.0230
XGBoost	0.980± 0.0026	0.828± 0.0243	0.772± 0.0378	0.772± 0.0378	0.910± 0.0347	0.995± 0.0023	0.883± 0.0185	**0.987± 0.0028**	0.914± 0.0158
RF	0.978± 0.0026	0.803± 0.0252	0.808± 0.0253	0.730± 0.0401	0.908± 0.0321	0.995± 0.0020	0.862± 0.0197	0.983± 0.0060	0.899± 0.0227
SAM	0.978± 0.0020	0.806± 0.0178	0.814± 0.0171	0.754± 0.0300	0.887± 0.0287	0.993± 0.0020	0.874± 0.0146	0.982± 0.0062	0.898± 0.0123
GT	0.979± 0.0034	0.821± 0.0275	0.832± 0.0267	0.815± 0.0424	0.852± 0.0411	0.990± 0.0036	0.902± 0.0205	0.979± 0.0080	0.887± 0.0269
GIN	0.977± 0.0045	0.818± 0.0362	0.829± 0.0342	**0.839± 0.0582**	0.823± 0.0495	0.987± 0.0047	0.913± 0.0282	0.980± 0.0109	0.894± 0.0383
GCN	0.971± 0.0038	0.765± 0.0291	0.779± 0.0274	0.787± 0.0386	0.774± 0.0445	0.984± 0.0043	0.885± 0.0186	0.974± 0.0075	0.844± 0.0283
GAT	0.972± 0.0044	0.773± 0.0366	0.785± 0.0370	0.783± 0.0731	0.796± 0.0603	0.985± 0.0058	0.884± 0.0348	0.975± 0.0084	0.865± 0.0255
AFP	0.976± 0.0048	0.799± 0.0379	0.812± 0.0356	0.809± 0.0383	0.816± 0.0503	0.987± 0.0043	0.898± 0.0194	0.974± 0.0064	0.865± 0.0255
Stacking	**0.981± 0.0035**	**0.839± 0.0321**	**0.848± 0.0308**	0.815± 0.0473	0.885± 0.0242	0.993± 0.0016	**0.904± 0.0237**	0.985± 0.0083	**0.921± 0.0251**

Table S3 Performance comparison of normal undersampling strategies for the binary classification model of NIR dyes

Undersampling									
Model	ACC	MCC	F1 Score	Recall	Precision	SP	BA	AUC	PRAUC
LGBM	0.927± 0.0039	0.634± 0.0110	0.624± 0.0122	0.946± 0.0125	0.465± 0.0141	0.926± 0.0045	0.936± 0.0056	0.977± 0.0032	0.799± 0.0233
GBRT	0.936± 0.0084	0.662± 0.0240	0.655± 0.0277	0.945± 0.0179	0.502± 0.0345	0.936± 0.0095	0.940± 0.0077	0.977± 0.0071	0.817± 0.0329
XGBoost	0.919± 0.0108	0.608± 0.0312	0.597± 0.0334	0.932± 0.0251	0.440± 0.0359	0.918± 0.0116	0.925± 0.0133	0.972± 0.0064	0.769± 0.0316
RF	0.929± 0.0051	0.638± 0.0207	0.629± 0.0200	0.940± 0.0231	0.473± 0.0196	0.928± 0.0051	0.934± 0.0120	0.977± 0.0047	0.793± 0.0300
SAM	0.933± 0.008	0.658± 0.029	0.647± 0.029	**0.960± 0.023**	0.488± 0.029	0.931± 0.007	**0.946± 0.014**	0.978± 0.007	0.801± 0.049
GT	**0.979± 0.003**	**0.822± 0.028**	**0.832± 0.027**	0.815± 0.042	**0.852± 0.041**	**0.990± 0.004**	0.903± 0.021	0.979± 0.008	0.888± 0.027
GIN	0.978± 0.005	0.819± 0.036	0.829± 0.034	0.839± 0.058	0.824± 0.050	0.987± 0.005	0.913± 0.028	**0.980± 0.011**	**0.895± 0.038**
GCN	0.972± 0.004	0.765± 0.029	0.780± 0.027	0.787± 0.039	0.775± 0.045	0.984± 0.004	0.886± 0.019	0.975± 0.008	0.845± 0.028
GAT	0.973± 0.004	0.773± 0.037	0.785± 0.037	0.784± 0.073	0.796± 0.060	0.986± 0.006	0.885± 0.035	0.975± 0.008	0.865± 0.026
AFP	0.976± 0.005	0.800± 0.038	0.812± 0.036	0.809± 0.038	0.817± 0.050	0.987± 0.004	0.898± 0.019	0.975± 0.006	0.867± 0.025

Table S4 Performance comparison of normal oversampling strategies for the binary classification model of NIR dyes

Oversampling									
Model	ACC	MCC	F1 Score	Recall	Precision	SP	BA	AUC	PRAUC
LGBM	**0.981± 0.0018**	0.831± 0.0161	**0.838± 0.0152**	0.778± 0.0238	0.909± 0.0245	**0.995± 0.0016**	0.886± 0.0117	**0.985± 0.0036**	**0.919± 0.0139**
GBRT	0.981± 0.0037	**0.834± 0.0349**	0.841± 0.0341	0.786± 0.0473	0.906± 0.0280	0.994± 0.0017	0.890± 0.0239	0.982± 0.0085	0.916± 0.0230
XGBoost	0.980± 0.0026	0.828± 0.0243	0.772± 0.0378	0.772± 0.0378	**0.910± 0.0347**	0.995± 0.0023	0.883± 0.0185	0.987± 0.0028	0.914± 0.0158
RF	0.978± 0.0026	0.803± 0.0252	0.808± 0.0253	0.730± 0.0401	0.908± 0.0321	0.995± 0.0020	0.862± 0.0197	0.983± 0.0060	0.899± 0.0227
SAM	0.978± 0.0020	0.806± 0.0178	0.814± 0.0171	0.754± 0.0300	0.887± 0.0287	0.993± 0.0020	0.874± 0.0146	0.982± 0.0062	0.898± 0.0123
GT	0.979± 0.0034	0.821± 0.0275	0.832± 0.0267	0.815± 0.0424	0.852± 0.0411	0.990± 0.0036	0.902± 0.0205	0.979± 0.0080	0.887± 0.0269
GIN	0.977± 0.0045	0.818± 0.0362	0.829± 0.0342	**0.839± 0.0582**	0.823± 0.0495	0.987± 0.0047	**0.913± 0.0282**	0.980± 0.0109	0.894± 0.0383
GCN	0.971± 0.0038	0.765± 0.0291	0.779± 0.0274	0.787± 0.0386	0.774± 0.0445	0.984± 0.0043	0.885± 0.0186	0.974± 0.0075	0.844± 0.0283
GAT	0.972± 0.0044	0.773± 0.0366	0.785± 0.0370	0.783± 0.0731	0.796± 0.0603	0.985± 0.0058	0.884± 0.0348	0.975± 0.0084	0.865± 0.0255
AFP	0.976± 0.0048	0.799± 0.0379	0.812± 0.0356	0.809± 0.0383	0.816± 0.0503	0.987± 0.0043	0.898± 0.0194	0.974± 0.0064	0.865± 0.0255

Table S5 Optimal parameters for the binary classification model of near-infrared dyes

LGBM	GBRT	XGBoost	RF	SAM
'boosting_type': 'gbdt' 'objective': 'binary' 'class_weight': 'balanced' 'colsample_bytree': 1 'importance_type': 'split' 'learning_rate': 0.1 'max_depth': -1 'min_child_samples': 30 'min_child_weight': 0.001 'n_estimators': 200 'num_leaves': 31 'reg_alpha': 0.1 'reg_lambda': 0.0 'subsample': 1 'subsample_freq': 0	'learning_rate': 0.1 'max_depth': 10 'min_samples_leaf': 1 'min_samples_split': 10 'n_estimators': 300 'subsample': 1.0, 'max_features': 'sqrt'	'objective': 'binary:logistic' 'eval_metric': 'logloss' 'booster': 'gbtree' 'colsample_bylevel': 1 'colsample_bynode': 1 'colsample_bytree': 1 'gamma': 0 'max_depth': 6 'min_child_weight': 1 'n_estimators': 200 'reg_alpha': 0 'reg_lambda': 0 'scale_pos_weight': 1 'subsample': 1	'bootstrap': True 'criterion': 'gini' 'max_depth': None 'max_features': 'auto' 'min_samples_leaf': 1 'min_samples_split': 2 'n_estimators': 600 'n_jobs': -1 'oob_score': False 'random_state': seed 'verbose': 0 'warm_start': False	'C': 100.0 kernel: 'rbf' 'gamma': 0.01 probability: True
GT	**GIN**	**GCN**	**GAT**	**AFP**
"hidden_channels": 256 "num_layers": 3 "dropout": 0.5 "lr": 0.0001 "weight_decay": 1e-05 "gamma": 1 "pos_weight": 1.079 "batch_size": 64 "n_heads": 5	"hidden_channels": 128 "num_layers": 5 "dropout": 0.2 "lr": 0.0001 "weight_decay": 1e-05 "gamma": 1 "pos_weight": 1.808 "batch_size": 128	"hidden_channels": 64 "num_layers": 2 "dropout": 0.2 "lr": 0.001 "weight_decay": 1e-05 "gamma": 1 "pos_weight": 1.015 "batch_size": 100	"hidden_channels": 256 "num_layers": 2 "dropout": 0.2 "lr": 0.0001 "weight_decay": 1e-05 "gamma": 1 "pos_weight": 1.08 "batch_size": 32	"hidden_channels": 256 "num_layers": 5 "dropout": 0.3 "lr": 0.001 "weight_decay": 1e-05 "gamma": 1 "pos_weight": 1.006 "batch_size": 32 "num_timesteps": 5

Table S6 Display of some forward transformation rules

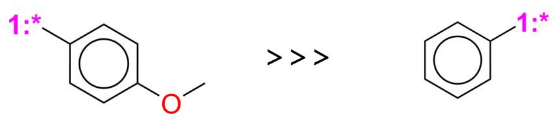
Index	NIR transformation rules	Frequency (non-NIR -->-- NIR)	label_value
0	 N#CC(=C[1:*])C(=O)O --->>>--- O=C(O)[1:*]	62	0.025
1	 N#CC(=Cc1ccc([1:*])cc1)C(=O)O --->>>--- O=C(O)c1ccc([1:*])cc1	61	0.043
2	 COc1ccc([1:*])cc1 --->>>--- c1ccc([1:*])cc1	7	0.0014
3	 c1ccc([1:*])cc1 --->>>--- CN(C)c1ccc([1:*])cc1	65	0.1391
4	 N#CC(=Cc1ccc([1:*])s1)C(=O)O --->>>--- O=C(O)c1ccc([1:*])cc1	59	0.1612
5	 O=C(O)c1ccc([1:*])s1 --->>>--- O=C(O)c1ccc([1:*])cc1	60	0.1812
6	 CN1C(=C[1:*])C(C)(C)c2ccccc21 --->>>--- CN1C(=C[1:*])C(C)(C)c2cc(C(=O)O)ccc21	295	1

Table S7 Fluor-pred Features Description

Object	Characterization	Description [a]	Methods
Dye	Molecular Graph	Including 39 node features and 10 edge features	Attentive FP
	Dye Scaffold Type	A total of 17 scaffold types	Zhu's script
	MMP Fingerprint	A total of 136 dimensions	Custom script
	Morgan Fingerprint	A total of 1024 dimensions	RDKit
	Quantum Chemistry Information	A total of 1 dimension, including HOMO-LUMO gap	Uni-mol+
	Molecular Properties	A total of 5 dimensions, including molecular weight, LogP, number of rings, number of double bonds, and TPSA	RDKit
Solvent	Solvent Type	A total of 71 solvent types	Custom script
	Morgan Fingerprint	A total of 1024 dimensions	RDKit

[a] For more detailed technical specifications, please refer to the open-source code of Fluor-pred.

Table S8 Model evaluation metrics and their corresponding formulas

Index	Name	Equation
1	*ACC*	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$
2	*MCC*	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
3	*F1 Score*	$F1 = \frac{2TP}{2TP + TP + FP}$
4	*Recall*	$Recall = \frac{TP}{TP + FN}$
5	*Precision*	$Precision = \frac{TP}{TP + FP}$
6	*SP (Specificity)*	$SP = \frac{TN}{TN + FP}$
7	*BA (Balanced Accuracy)*	$BA = \frac{Recall + SP}{2}$
8	*AUC*	$AUC = \int_0^1 TPR\,(x)dx$
9	*PRAUC*	$PRAUC = \int_0^1 Precision\,(r)dr$
10	*MAE*	$MAE = \frac{1}{m}\sum_{i=1}^{m} \lvert(y_i - \hat{y}_i)\rvert$
11	*RMSE*	$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$
12	R^2	$R^2 = 1 - \frac{\sum_i \ (\hat{y}_i - y_i)^2}{\sum_i \ (\overline{y}_i - y_i)^2}$

Table S9 Statistics on the number of molecules for each scaffold type

Scaffold	Count	Count (*non-NIR*)	Count (*NIR*)
5n6 (6-n-5)	931	912	19
5p6 ([6+5])	2546	2431	115
6n6 (6-n-6)	1463	1406	57
6p6 ([6+6])	887	845	42
Acridines	954	933	21
Azo (Diazo)	38	38	0
BODIPY	1526	1324	202
Benz (Benzene)	699	689	10
Carbazole	1257	1240	17
Coumarin	779	696	83
Cyanine	980	584	396
Naphthalimide	257	255	2
PAHs	1683	1631	52
Porphyrin	321	297	24
SquaricAcid (Squaraine)	320	280	40

List of Abbreviations

ACC, accuracy; AD, applicability domain; AI, artificial intelligence; BA, balanced accuracy; CNN, convolutional neural networks; DL, deep learning; DM, distance-based method; GCN, graph convolutional networks; GNN, graph neural network; KDE, kernel density estimation; LDS, label distribution smoothing; MCC, matthews correlation coefficient; ML, machine learning; MMPA, matched molecular pair analysis; NIR, near-infrared; non-NIR, non-near-infrared; QSAR-MMP, quantitative structure-activity relationship-molecular matching pair; RLAT, residual lightweight attention; SARs, structure-activity relationships; SP, specificity; TD-DFT, time-dependent density functional theory; TPSA, topological polar surface area; λ_{abs}, absorption wavelength; λ_{em}, emission wavelength; Φ_{PL}, fluorescence quantum yield; ε_{max}, molar absorption coefficient