JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau
Article

AI-Predicted Full-Spectrum UV Correction Factors: An Analyte-Standard-Free Framework for Reaction Yield Quantification

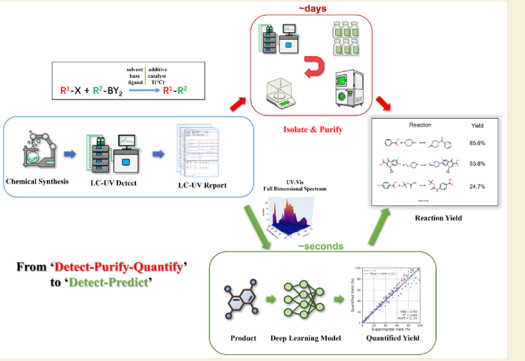
Jiuchuang Yuan,[‖] Mingjun Yang,[‖] Fan Yan, Jing Guo, Lin Zhang, Guosheng Dou, Minjun Liu, Lu Tan, Liwen Fang, Guanfeng Yang, Qun Zeng, Jun Yan, Xuekun Shi, Sarah Trice, Jian Ma, Shuhao Wen, Christopher J. Welch,* and Peiyu Zhang*

Cite This: https://doi.org/10.1021/jacsau.6c00812

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: High-throughput experimentation (HTE) accelerates molecular discovery but struggles to quantify yields without pure analyte standards. The current need to purify standards increases cost, complexity, and cycle time, limiting the scalability of new molecular discovery. Here, we introduce an integrated "detect-predict" framework that leverages artificial intelligence (AI) for analyte standard-free yield quantification. This approach hierarchically trains a prediction model of full-spectrum UV correction factor (CF) by combining quantum chemistry (QC) pre-training on 22,609 compounds for broad chemical space coverage with experimental fine-tuning on 1,845 LC-UV spectra to adapt to real instrumental conditions, thereby mitigating spectral discontinuities and data heterogeneity between computational and experimental domains. These predicted CFs enable direct quantification of analyte concentrations and reaction yields from UV absorbance data. We validated the accuracy of this framework through three independent test cases. First, concentration estimation was tested against 144 FDA compounds with known experimental concentrations. Second, yield quantification for 178 real-world reactions (spanning amide coupling, Suzuki−Miyaura coupling, SN2, and Buchwald-Hartwig transformations) was compared with calibration-curve-derived yields, achieving a mean absolute error of 4.7%. Finally, model-quantified yields were compared with isolated yields for 60 reactions as practical stress tests. The systematic positive bias aligned with the expected difference between crude analytical yield and post-purification isolated yield, enabling chemists to distinguish reaction performance from workup losses. This capability is critical for decision-making in real-world optimization. This AI-driven "detect-predict" framework offers an alternative to the traditional "detect-purify-quantify" workflow, providing a faster and more scalable foundation for HTE-accelerated molecular discovery.



KEYWORDS: *analyte standard-free quantification, correction factor prediction, high-throughput experimentation, full-spectrum UV integration*

INTRODUCTION

High throughput experimentation (HTE) and machine learning (ML) are revolutionizing the synthesis and characterization of new chemical entities (NCEs).[1−7] Whereas molecular discovery in the 20th century involved the synthesis, purification, and archiving of newly prepared compounds in physical libraries at scales sufficient to support future research, modern streamlined approaches typically prepare only enough material for preliminary performance evaluation. Resynthesis at scale is reserved for only a selected few compounds that warrant further investigation.[8] This HTE-driven strategy enables rapid exploration of promising chemical space but introduces a critical challenge: authentic standards of known purity for the vast majority of new molecules are typically not available during the early phases of molecular discovery. Traditional purification and characterization workflows are prohibitively time-intensive and costly in such high throughput contexts. Consequently, developing new strategies to quantify concentrations and reaction yields of newly synthesized compounds represents an active and vital research frontier in high-throughput molecular discovery.

Universal detection methods (*e.g.* CAD, ELSD, CLND, GC-FID, NMR) can offer a degree of structure-agnostic quantification, but their limited range, sensitivity, throughput or cost often precludes use in high-throughput workflows.[9−11] Liquid chromatography (LC) with ultraviolet (UV) spectral detection remains the analytical mainstay due to its speed and accessibility, yet quantification relies on the molar absorptivity

Received: May 25, 2026
Revised: August 18, 2026
Accepted: September 1, 2026

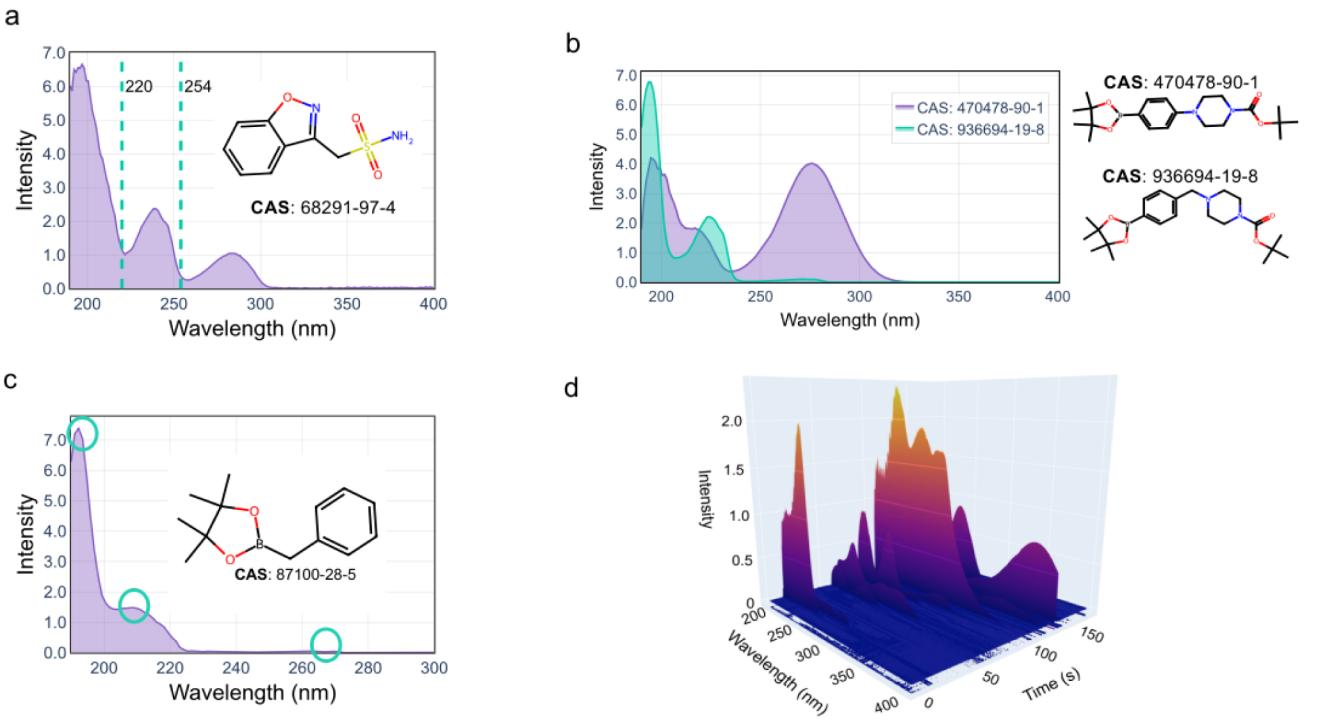


Figure 1. Challenges for quantification of concentration or yield using UV absorbance. (a) traditional method using UV area integrated along the retention time at 220 or 254 nm from LC-UV experiment and comparing with standards; (b) similar compounds can exhibit dramatic differences in UV absorbance, with clifflike transition of ε with small changes in wavelength; (c) selecting the "reddest peak" can sometimes be ambiguous; (d) 3D plot of UV absorbance spectrum (wavelength × intensity) *vs.* HPLC retention time provides a more robust approach for quantification of concentration and yield.

(ε), a property intrinsically linked to molecular structure and solvent environment. Similarly, LC-mass spectrometry (LC-MS), though highly sensitive, requires calibration curves constructed from pure standards, creating bottlenecks for novel compounds in high-diversity screening.[12] The potential of UV spectral prediction has long been recognized to address this "quantitation without standards" challenge. Computational methods have evolved from early empirical rules[13,14] to quantum chemical (QC) calculations[15] to ML approaches for ε prediction,[16,17] but current models still lack the accuracy and robustness needed to enable practical analyte standard-free quantification.[18,19]

Recently, Jensen and coworkers have reported a promising chemical structure-based machine learning approach for estimating ε at a particular wavelength (reddest UV peak) but this model also lacks the degree of transferability and precision that would ideally be needed for new molecule discovery platforms.[20] While it is standard practice to quantify yields by integrating UV absorption peak areas at a single wavelength (e.g., 220 or 254 nm) (Figure 1a),[21] accurately predicting absorbance at a given wavelength can be challenging owing to the large variations in ε between molecules (Figure 1b) and the "clifflike" relationship often seen between absorbance and wavelength, where a wavelength change of a few nanometers can result in order of magnitude changes in absorbance. Furthermore, a clear identification of the reddest peak is often ambiguous, leading to additional challenges (Figure 1c). We therefore shifted focus to full-spectrum UV information (Figure 1d), leveraging three-dimensional LC-UV data (wavelength × intensity vs. HPLC retention time) to enable quantitative and robust yield and concentration determination.

In addition to the selection of modeling tasks for UV absorbance spectra, the development of robust chemical structure-based machine learning models is also constrained by data limitations, which is a challenge in the field. Training data for such models typically derive from two suboptimal sources: (1) small, well-curated datasets collected under standardized conditions, but lacking structural diversity and generalizability; or (2) larger literature-compiled datasets, which suffer from inconsistent protocols, undocumented variables, and absence of negative results. While computational strategies have sought to mitigate these issues, models trained on extensive, diverse, and rigorously curated datasets remain essential for providing accurate yield estimates that enable chemists to differentiate reaction performance from purification-related losses. In this work, we introduce a novel approach that leverages full UV-spectrum correction factor (CF) prediction to enable accurate concentration and yield determination without physical isolation of standards. This model was developed using a two-stage training strategy: pre-training on QC calculated spectra of a large virtual library covering diverse chemical space, followed by fine-tuning with experimental LC-UV data from a smaller set of pure compounds. This hybrid strategy helps to ensure robustness and generality. We validated this approach across three independent test cases: 144 FDA-approved compounds, 178 real-world reactions spanning four transformation types, and practical stress tests. The model achieves high-accuracy quantification that enables informed decision-making in reaction optimization. This capability forms the foundation of a "detect-predict" framework that provides an alternative to traditional "detect-purify-quantify" workflows. By addressing the critical bottlenecks of speed, standardization, and general-

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

izability in high-throughput discovery, this methodology establishes a scalable foundation for quantitative analysis in molecular discovery, with potential applications in pharmaceuticals, agrochemicals, specialty chemicals, catalysts, etc.

COMPUTATIONAL AND EXPERIMENTAL METHODS

Definition of the CF and Its Relationship to Concentration and Yield Quantification

The CF for the target analyte (CF_t) is defined as

$$CF_t = \frac{A_t}{A_{IS}} \times \frac{C_{IS}}{C_t} \tag{1}$$

where A_t and A_{IS} represent the integrated UV absorbance areas over continuous wavelengths for the target (t) analyte and internal standard (IS), respectively; C_t and C_{IS} denote the molar concentration (in mol/L) of the target analyte and internal standard, respectively. Here the internal standard is used as a reference compound that is added to the reaction solution for quantification normalization. For experimental measurements, the UV absorbance area was integrated along the wavelength starting from 190 nm. For QC-calculated UV absorbance area, the starting wavelength was set to 185 nm. This choice was determined by sweeping the QC lower integration limit from 170 to 210 nm in 5 nm steps and evaluating the agreement between QC-derived and experimentally measured CFs; the agreement was optimal at 185 nm, reflecting the systematic wavelength offset between the calculated and experimental spectra. The molecule 4,4'-Di-*tert*-butyl-biphenyl (CAS 1625−91-8) was used as the IS for all experiments and calculated CFs.

From LC-UV analysis of a solution including both target compound and IS (where concentration of IS,C_{IS}, is known prior), A_t and A_{IS} can be derived from the UV absorbance spectrum. Thus, the concentration of target compound can be determined by the following equation:

$$C_t^{pred} = \frac{A_t}{A_{IS}} \times \frac{C_{IS}}{CF_t^{pred}} \tag{2}$$

where CF_t^{pred} is predicted using the SMILES of target compound as input of a predictive machine learning model described below. If the target compound is created from a chemical reaction, then the reaction yield (Y_t^{pred}) can be derived as,

$$Y_t^{pred} = \frac{C_t^{pred}}{C_t^0} \times 100.0\% \tag{3}$$

where C_t^0 is the concentration of the full conversion of the reactants.

To evaluate the performance of the quantification model for CF, concentration and reaction yield, several metrics were employed as follows:

a the coefficient of determination

$$\left(R^2 = 1 - \frac{\sum_{i=1}^{N}(Pred_i - Exp_i)^2}{\sum_{i=1}^{N}\left(Exp_i - \frac{1}{N}\sum_{i=1}^{N}Exp_i\right)2}\right) \tag{4}$$

is calculated, representing the proportion of variance in the measured values explained by the model with a value of 1 indicating perfect fit.

b the mean absolute error

$$\left(\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|Pred_i - Exp_i|\right) \tag{5}$$

and the mean absolute percentage error

$$\left(\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Pred_i - Exp_i|}{Exp_i}\right)$$

were computed to quantify the absolute and relative deviation between predicted and measured values, respectively. Here N represents the number of samples tested.

QC Calculated Dataset for Pre-Training the CF Prediction Model

The virtual pre-training dataset was constructed from two complementary components: a reaction-centered set of synthetically relevant molecules and a diversity-optimized subset from DrugBank, collectively providing 22,609 molecules with broad structural coverage (see "Construction and Diversity Analysis of the Virtual Pre-training Dataset" section in Supplementary Materials). t-SNE analysis confirmed that this combined dataset effectively integrates synthetic relevance with comprehensive chemical diversity (see Table S1 and Figure S1 in Supplementary Materials), establishing a robust foundation for molecular representation learning. The curated dataset exhibits a mean molecular weight of 300 Da (primarily 150−450 Da), a mean heavy atom count of 20, and predominantly 1−4 rings included in the molecules (Figure S2 in Supplementary Materials). The QC CF was derived using eq 1 with known A_t and A_{IS} computed from QC (in this case C_t=1 and C_{IS}=1) by following a five-step calculation workflow described in the "Computed UV Absorbance Spectra using a Five-Step Quantum Chemistry (QC) Workflow" Section of Supplementary Materials.

To assess whether the solvent choice for QC calculations (acetonitrile) introduces systematic bias relative to the experimental mobile phase (water−acetonitrile), we repeated the full QC workflow for 42 randomly selected FDA-approved molecules using water as the implicit solvent. The water- and acetonitrile-derived CFs showed near-perfect agreement (R^2= *0.992*, MAE = 0.018), suggesting that solvent effects are negligible for the integrated full-spectrum CF. Detailed results are provided in the Figure S3 of the Supplementary Materials.

Experimental Dataset for Fine-Tuning the CF Prediction Model

The fine-tuning dataset comprises 1,845 experimentally characterized molecules with measured correction factors (CFs), providing targeted data for model adaptation (see "Fine-tuning Dataset Construction and Diversity Analysis" Section in Supplementary Materials). Structural analysis confirms this dataset is structurally congruent with the pre-training chemical space while introducing experimentally validated CF values, ensuring both domain consistency and task-specific relevance for effective transfer learning (see Tables S2−S3 and Figure S1 in Supplementary Materials). For each compound, LC-UV was detected with the same acetonitrile (ACN)/water gradient as mobile phase (see "Procedure for Determination of Experimental Correction Factors (CFs)" Section in Supplementary Materials). The experimental CFs for these compounds were derived using eq 1 with known A_t, A_{IS}, C_t, and C_{IS} from LC-UV analysis and experiment settings.

Independent Benchmark Dataset to Test Quantification Performance for Real-World Reaction Yield and Solution Concentration

To validate the performance of the "detect-predict" method, we constructed three benchmark datasets corresponding to distinct practical laboratory scenarios. First, 144 FDA-approved drug molecules were selected and prepared as solutions of known concentration. The predicted correction factors (CF) (from eq 1) and derived concentrations (from eq 2) were compared against experimental values obtained from LC-UV analysis to assess predictive accuracy. Second, the method was tested for reaction yield quantification (from eq 3) using 178 real-world chemical transformations spanning four common types: amide coupling, Suzuki−Miyaura cross-coupling, SN2 nucleophilic substitution, and Buchwald-Hartwig cross-coupling. Experimental yields for 71 distinct

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

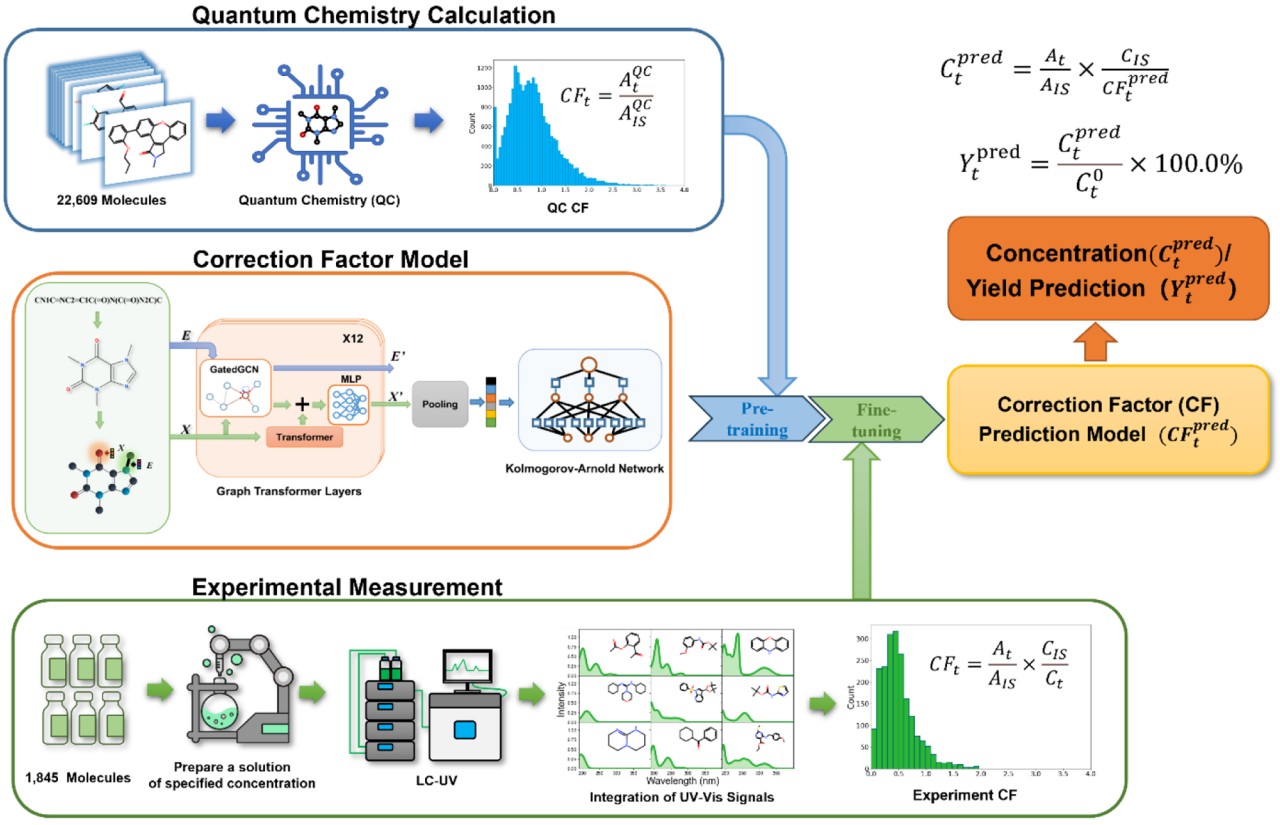


Figure 2. Workflow for CF prediction model training using Graph Transformer + KAN architecture. The model was trained with a two-stage strategy where a QC calculated dataset of 22,609 compounds was first used for pre-training followed by fine-tuning with an LC-UV derived dataset of 1,845 compounds. The solution concentration (C_t^{pred}) or reaction yield (Y_t^{pred}) can be derived from the predicted CF_t and the experimental UV absorbance area (A_t, A_{IS}).

compounds (for which pure standards were available) were determined via calibration curves, providing a benchmark for comparison. Finally, to further evaluate broader applicability, the quantified yields (from eq 3) were compared against isolated yields for another 34 diverse Suzuki–Miyaura and 26 diverse Minisci reactions, validating the method's performance across a wide range of synthetic conditions. This dataset was designed as a practical stress test to assess the method's ability to reflect differences between reaction performance and workup losses, rather than to re-validate accuracy, which had already been established by the first two independent benchmarks. Three complementary analyses including chemical-space visualization, nearest-neighbor fingerprint similarity, and Bemis–Murcko scaffold were performed for comprehensive structural similarity analysis between the pre-training and the other datasets (see Figure S1, Tables S2–S3, and Analysis of Structural Similarity for Molecules in Fine-tuning and Independent Benchmark Dataset with Respect to Pre-training Dataset" Section in Supplementary Materials). It shows that fine-tuning molecules exhibit high scaffold overlap (97.2%) with the pre-training set, reflecting their shared synthetical origin. While the FDA benchmark displays substantial scaffold novelty (only 33.9% overlap) and serves as a strong out-of-distribution test. The real-world reaction products occupy an intermediate position with partial scaffold sharing (77.8%). These results collectively verify that the benchmark datasets probe genuine generalization beyond memorization of training structures.

Model Architecture and Training for CF Prediction

The model architecture was proposed to integrate Graph Transformer layers with a Kolmogorov-Arnold network (KAN[22]) to enable expressive modeling of graph-structured data. As illustrated in Figure 2, the model comprises core components as follows:

Node and Edge Embedding. The molecular graph representation encodes atoms (nodes) and bonds (edges) by using a comprehensive set of chemical features. Each node is characterized by a 65-dimensional feature vector, including one-hot encodings of atomic symbol, chirality, degree, formal charge, number of hydrogens, radical electrons, hybridization state, aromaticity, and ring membership. Edges are represented by 13-dimensional features capturing the bond type, stereochemistry, and conjugation. Raw node features and edge features are linearly projected into a 64-dimensional latent space via trainable weight matrices with tanh activation applied to the resulting embeddings.

Core Graph Processing. 12 stacked GraphGPS layers process the features.[23] Each layer combines a gated message-passing mechanism to capture local structural patterns and a multihead transformer attention (4 heads) to model global graph context. Following graph processing, a graph-level representation is obtained by applying sum pooling to aggregate all node features and then linearly projected to 192 dimensions to prepare features for subsequent KAN processing.

KAN-Based Prediction. Replacing traditional MLPs, a KAN with topology and grid size 3 transforms pooled graph features $\mathbf{h}_G$. Learnable spline-based activation functions enable adaptive nonlinear mappings, with a final Softplus activation ensuring positive-valued predictions: $y = \text{Softplus}(\text{KAN}(\mathbf{h}_G))$.

We implemented a two-stage transfer-learning approach combining QC computations with experimental data.

Pre-Training Stage. The model was initially trained on 21,585 QC-calculated CFs, with two sets of randomly selected 512 compounds used as validation and testing data at the pretraining stage, respectively. The architecture was optimized with Adam (initial lr = 1e-4) and MSE loss. Training employed early stopping (patience = 100) on an NVIDIA H20 GPU with a batch size of 512.

Fine-Tuning Stage. The pre-trained model was subsequently refined using experimental LC-UV data (n = 1,845) with careful dataset partitioning: training (1,760), validation (50), test (35).

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX–XXX

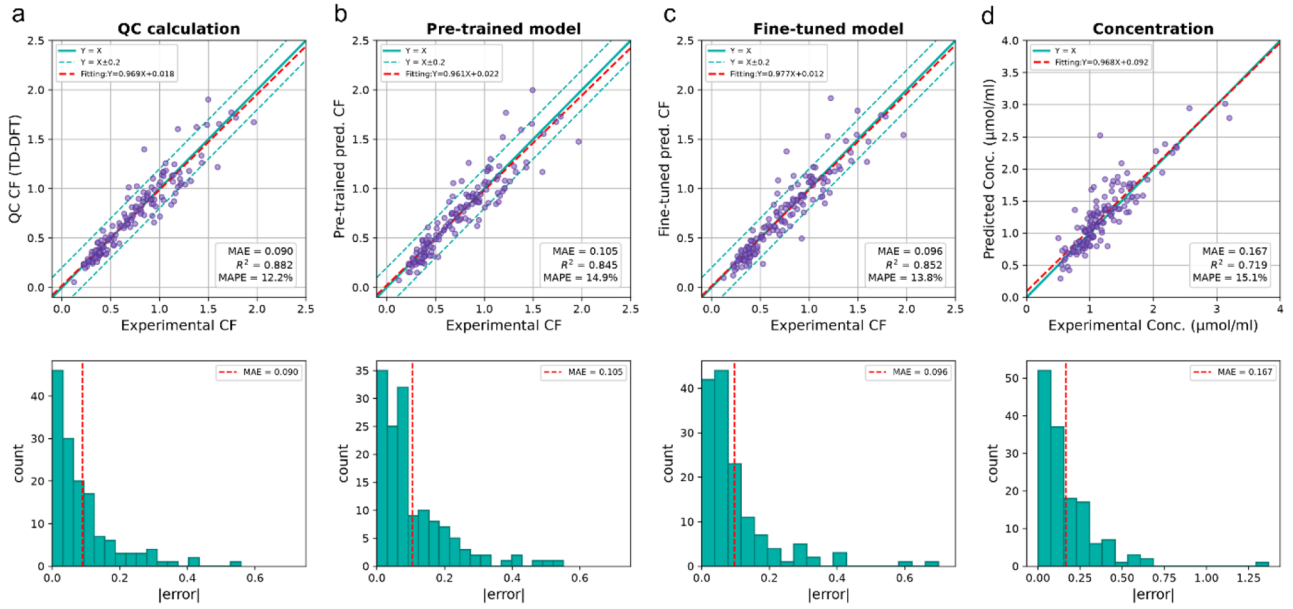


Figure 3. Model performance benchmarked using an independent dataset from 144 FDA approved drug compounds. (a) CFs from QC calculation versus experimental values (top panel) and the error distribution (bottom panel), (b) CFs from pre-trained model prediction versus experimental values (top panel) and the error distribution (bottom panel), (c) CFs from fine-tuned model prediction versus experimental values (top panel) and the error distribution (bottom panel), (d) concentrations from fine-tuned model estimation versus experimental values.

External FDA test set (144) serves as a more comprehensive evaluation across diverse molecular structures, providing rigorous assessment of the model's generalizability to broader chemical space. Fine-tuning used a OneCycleLR scheduler (max_lr = 1e-4, cosine annealing) with momentum cycling (0.85 → 0.95) over 50 epochs, maintaining the same early stopping criteria. Batch processing (size = 512) was accelerated on NVIDIA H20 GPUs.

RESULTS AND DISCUSSION

To enable analyte standard-free quantification, we defined the CF_t in eq 1 as a correction factor to relate UV absorption areas of target analytes and an internal standard (4,4′-di-*tert*-butylbiphenyl) to concentration (eq 2) and yield (eq 3) quantifications. To address UV spectral prediction challenges (e.g., wavelength-sensitive intensity cliffs and noisy public data), we prioritized full-spectrum integration over single-wavelength analysis in a two-stage strategy. A predictive model for CF_t was trained hierarchically by using Graph Transformer + KAN architecture: first on QC derived CF_t values for 22,609 diverse organic molecules (see Figure S2 in Supplementary Materials), then fine-tuned using experimental CF_t of 1,845 compounds from standardized LC-UV analysis (Figure 2). The prediction transferability of the framework was rigorously validated through three test cases: (1) concentration estimation for 144 FDA-approved drugs, (2) reaction yield quantification for 71 unique products across 178 reactions of four widely used reaction types (amide coupling, Suzuki, SN2, Buchwald-Hartwig) under 25 conditions, with reference yields established via calibrated LC-UV, and (3) quantitative comparison between quantified yield and isolated yield over another 34 Suzuki−Miyaura and 26 Minisci reactions.

The CF_t prediction performance and the resulting concentration quantification were first examined against the first benchmarking dataset consisting of 144 FDA-approved drugs. Interestingly the QC calculated CF_t values align very well with the experimentally measured values with R^2 of 0.882, an overall MAE of 0.090 and MAPE of 12.2% (Figure 3a), suggesting robust capture of structure-property relationships from QC calculations and strong validation to use them as pre-training dataset. Figure 3b indicates that the pre-trained model using only QC data already exhibits strong prediction performance for CF_t with R^2 = 0.845, MAE = 0.105 and MAPE = 14.9%. The fine-tuned model using 1845 experimental data shows an overall close R^2 and MAE value with the pre-trained model (Figure 3c). However, the fine-tuned model demonstrates improved performance compared with the pre-trained model within the high-accuracy prediction region with absolute error <0.1, with a population of 69.4% from the fine-tuned model (Figure 3c) compared to 66.0% from the pre-trained model (Figure 3b) and 69.4% from QC calculations alone (Figure 3a). From predicted CF_t^{pred}, the concentration can be derived using eq 2 and shown in Figure 3d, with R^2 = 0.719, MAE = 0.167 μmol/mL and MAPE = 15.1%. The concentrations were predicted with moderate deviations from experimental values (mainly because the range of concentration is very narrow due to the same mass concentration of compound used for solution preparation), though 75.7% fall within the relative error range of 20% and follow a linear fitting $C_t^{pred} = 0.968 \times C_t + 0.092$. In addition to drug-like compounds in FDA dataset, the same conclusion was also observed for a dataset comprised of 1,845 simpler, lead-like compounds (Figure S4). In summary, these results suggest that (1) QC calculations can produce excellent predictions of the experimental CFs and thus can be used for pre-training the prediction model to capture important structure-property relationships; (2) the fine-tuned model demonstrates better performance within the high-accuracy prediction region with absolute error <0.1 and thus can effectively reduce prediction outliers; and (3) the same conclusion applies to complex drug-like molecules as well as structurally simpler lead-like molecules.

To explore the generality of the approach, the model performance was validated by the second benchmark dataset

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

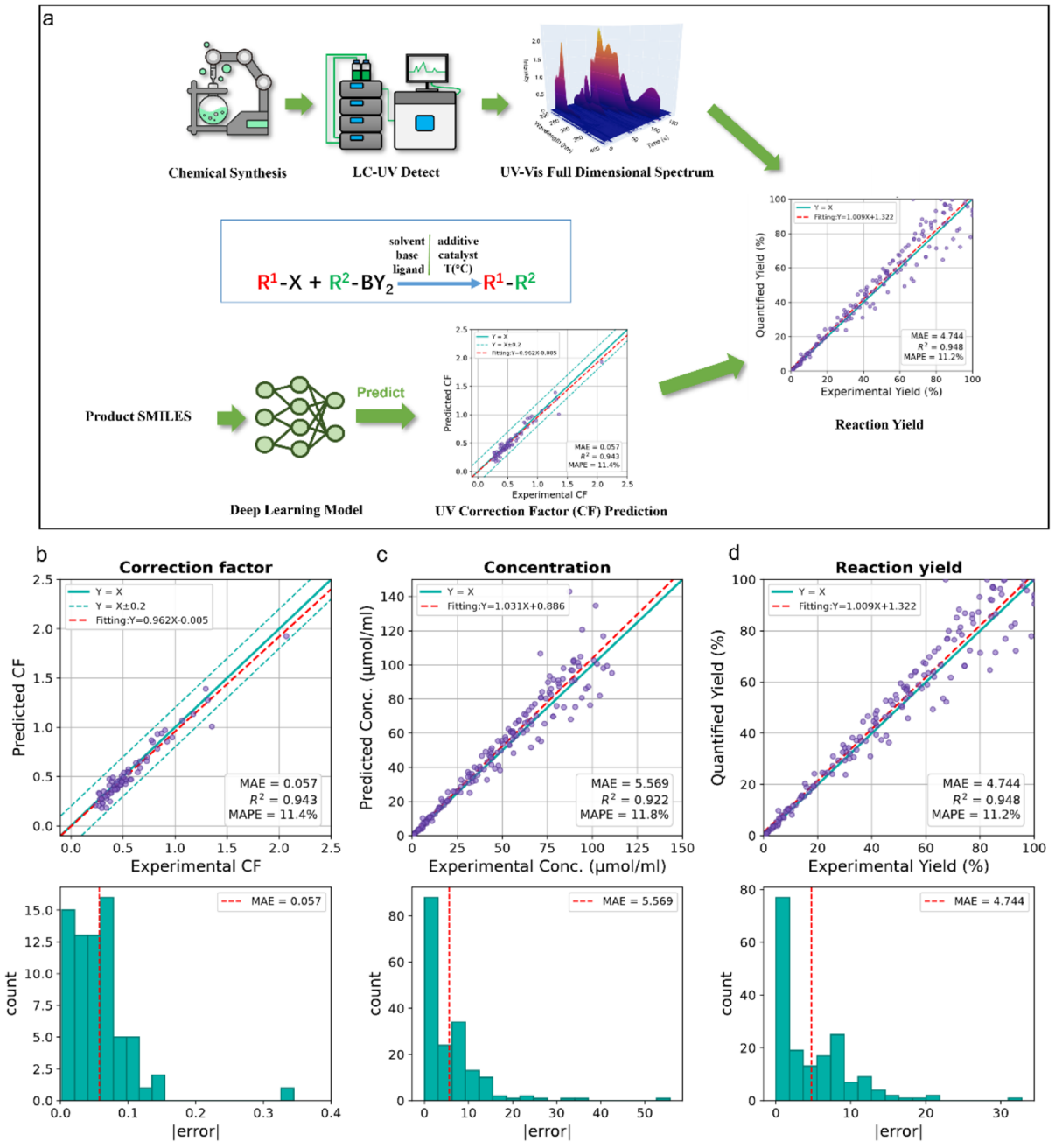


Figure 4. Model performance benchmarked with yield quantification from 178 real-world chemical reactions. (a) the "detect-predict" workflow for reaction yield quantification by combining experimental UV spectra and predicted CFs, (b) CFs from fine-tuned model prediction compared to experimental values (top panel) and the error distribution (bottom panel), (c) concentration of target product calculated from eq 2 using predicted CFs compared to experimental values (top panel) and the error distribution (bottom panel), (d) reaction yield calculated from eq 3 using predicted CFs compared to experimental values (top panel) and the error distribution (bottom panel).

for reaction yield quantification. To this end, 178 real-world reactions spanning four key transformations (amide coupling, Suzuki cross-coupling, SN2, Buchwald-Hartwig) under 25 distinct conditions were conducted and experimental yields were derived using the calibration curve established with multiple concentrations of pure standards. The "detect-predict" workflow for reaction yield quantification was established with chemical synthesis followed by LC-UV signal analysis and CF_t prediction for both the product and IS (Figure 4a). For each target product of these reactions, the compound-specific calibration curves were constructed using different concentrations of standard samples, from which the reference yields were derived by interpolating experimental UV areas into these curves (Figure S5). The predicted CF_t values show strong alignment with experimental measurements across all reaction types (Figure 4b), with R^2 = 0.943, MAE = 0.057, MAPE = 11.4%, and linear fitting $CF_t^{pred} = 0.962 \times CF_t^{exp} - 0.005$, corresponding to calculated concentrations using eq 2 as $C_t^{pred} = 1.031 \times C_t^{exp} - 0.886$ with R^2 = 0.922, MAE = 5.569 μmol/mL, and MAPE = 11.8% (Figure 4c). Crucially, the quantified yields (in percentage) by eq 3 demonstrate strong correlation with reference values (Figure 4d), with $Y_t^{pred} = 1.009 * Y_t^{exp} + 1.322$ and R^2 =

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

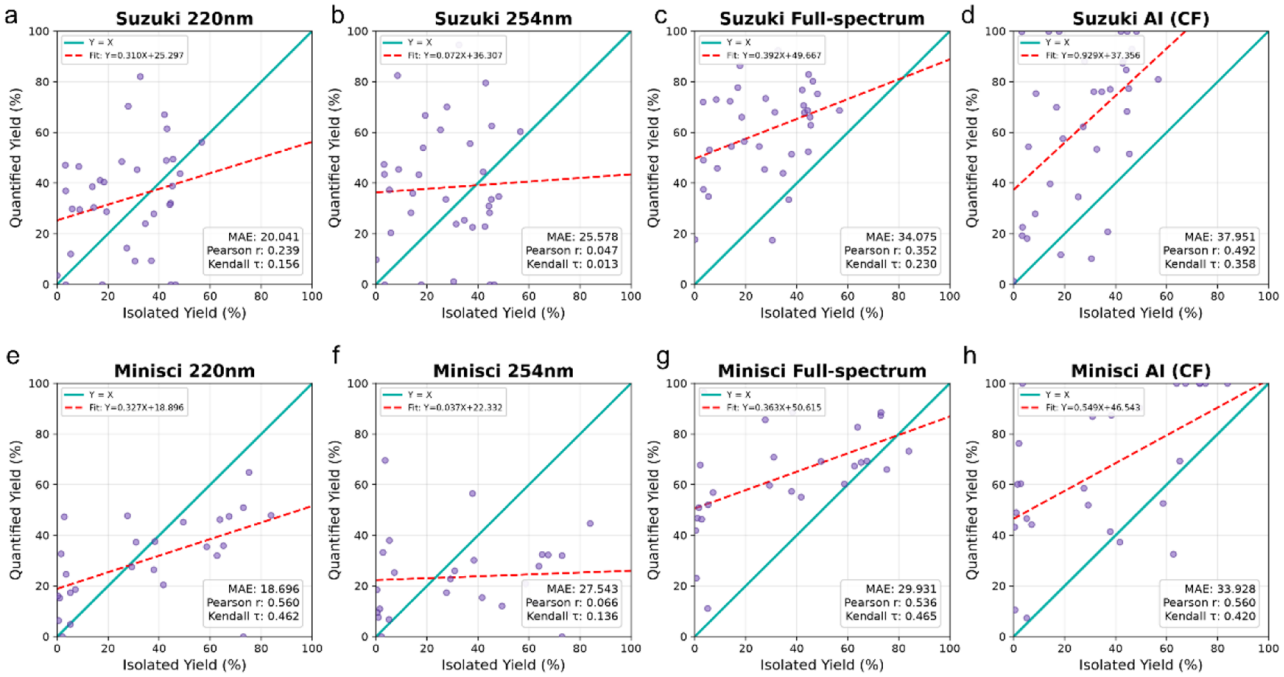


Figure 5. Model performance benchmarked against the isolated yield from (a–d) 34 diverse Suzuki–Miyaura cross-coupling and (e–h) 26 diverse Minisci reactions followed by isolation via preparative reversed-phase chromatography. The reaction yield was quantified using the UV area-based normalization method at 220 nm (a and e), 254 nm (b and f), full spectrum (c and g) and the AI predicted method based on predicted CF_t values (d and h).

0.948, MAE = 4.744, MAPE = 11.2% with 59.0% of reactions determined within an error of 5% and 88.8% within an error of 10%. In particular, the absolute yield error is smaller in the low-yield region than in the high-yield region. Interestingly, we found the prediction error for CF_t distributes uniformly from low to high (Figures 3a–c and 4b) while the prediction error for solution concentration C_t and reaction yield Y_t increases with increasing C_t (Figure 3 4c) and Y_t (Figure 4d) value, respectively. This observation can be explained by eq 2 (eq 3) for reaction yield), where higher concentration corresponds to larger UV area A_t obtained from LC-UV analysis and the C_{IS}/A_{IS} remains almost the same for low or high concentration values, thus resulting in larger prediction error for the higher concentration region.

To verify that the reported performance is not inflated by repeated measurements of the same product, we re-analyzed the reaction benchmark at the product level. Product-level MAE = 4.72% and R^2 = 0.938 were nearly identical to the reaction-level values (4.74% and 0.948), confirming that the accuracy reflects genuine molecular generalization. Breakdown by reaction type showed robust performance across all four transformations, with Suzuki–Miyaura reactions achieving the highest accuracy (MAE = 2.69%) and amide coupling showing a moderately higher MAE (6.42%) due to greater structural diversity (Table S4). Collectively, these results demonstrate the workflow's capacity to allow direct, reliable yield assessment without requiring purified standards under realistic reaction procedures, suggesting that this "detect-predict" workflow has the potential to provide an alternative to the time- and resource-intensive "detect-purify-quantify" approach commonly used in yield quantification.

To further evaluate the applicability of the "detect-predict" approach in a practical laboratory setting, we conducted a final stress test study comparing quantified yields from eq 3 against traditional isolated yields. This validation comprises 34 diverse Suzuki-Miyaura cross-coupling and 260 diverse Minisci reactions conducted at 0.2 and 0.25 mmol respectively, with products isolated via preparative reversed-phase chromatography. The 34 reactions were selected from a HTE database of 78,000 in-house Suzuki–Miyaura cross-coupling reactions, each of which had been quantitatively assessed by the widely-used UV peak-area normalization and CF-based quantification model developed in this study. The selection criterion was that the quantified yields for these reactions differed substantially across the two types of quantification methods. The products of these 34 reactions were isolated and purified, and the experimentally derived isolated yields were systematically compared with the yields determined by various quantification methods to evaluate the accuracy and reliability of each quantification approach. Besides Suzuki–Miyaura cross-coupling reactions, the Minisci reaction was also selected as another test case due to its pivotal role in the late-stage C–H functionalization of heteroarenes in drug discovery. In this dataset, we designed 260 reactions starting from readily available alcohols as alkyl radical precursors, therefore representing fast and economical access to diverse molecular scaffolds. Of them, 26 reactions with high quantified yields and clean profiles were experimentally isolated and purified for comparison to different yield quantification methods. Thus, these two datasets represent practical laboratory assessments of the utility of the yield quantification workflow for decision making.

In contrast to the high concordance observed with the internal calibration-based benchmarks, the model-quantified yields exhibited a systematic positive bias relative to the isolated yields (Figure 5d and h). The MAE between yields estimated from eq 3 and isolated yields was 38.0% and 33.9% for the 34 Suzuki and 26 Minisci reactions, respectively. This

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX–XXX

substantial discrepancy, however, carries a clear physical rationale. In small-scale millimole synthesis, the isolation process incurs significant and unavoidable purification losses arising from retention on the stationary phase and inefficiencies in sample injection, fraction collection, solution transfer, and evaporation. The isolated yield is therefore a convolution of the true reaction's chemical output and the efficiency of physical recovery. This AI model represents the yield quantification of the crude mixture during HPLC analysis stage, providing an estimate of the chemical yield prior to these material losses. Consequently, the systematic overprediction versus isolated yield is consistent with material losses during the traditional isolation process, suggesting a potential bias when using isolated yield to assess reaction performance in high-throughput exploration. Despite the absolute offset, this AI-driven method demonstrated competitive or superior trend-capturing performance for the underlying chemical trends. As shown in Figure 5a−d, it achieved the highest Pearson correlation (0.492) with isolated yields for the Suzuki-Miyaura cross-coupling reactions, as compared to 0.239, 0.047, and 0.352 for the UV area-based method at 220 nm, 254 nm, and full spectrum, respectively. For the Minisci reactions (Figure 5e−h), the Pearson correlation coefficient is 0.560, 0.560, 0.066, and 0.536 for this AI quantified method and for the UV area-based methods with 220 nm, 254 nm and full spectrum, respectively. Although comparable correlation coefficients were observed for the AI quantification and UV area-based method at 220 nm, most AI quantified yields are larger than the isolated yield in Figure 5h, which indicates most of the model quantified reaction yields represent the reaction yield before material attrition during workup, isolation and purification processes. For miniaturized, high-throughput experimentation, isolated yield is a convoluted metric that merges chemical reactivity with purification efficiency. By providing a theoretically grounded yield that tracks the crude analytical mixture prior to isolation while bypassing the noise of purification, this AI-driven LC-UV analysis establishes a more robust and cost-effective foundation for data-driven reaction optimization and machine learning in large-scale HTE.

Domain of Applicability (DoA)

To provide users with quantitative criteria for assessing prediction reliability, a systematic error analysis was conducted along two dimensions: structural similarity to the training set and the CF distribution with different magnitude values. The results inform a tiered decision-making framework for deployment.

For structural similarity analysis, Tanimoto similarity (ECFP4, radius = 2, 2048-bit) was computed for each benchmark molecule against both the experimental fine-tuning (FT) and QC pre-training sets. The results reveal a consistent and monotonic degradation pattern that defines the model's applicability domain. Across all benchmark and training-set combinations, prediction error increases gradually with decreasing similarity, and the model never collapses even at the farthest distances. This smooth, cliff-free degradation indicates that the model has learned continuous structure−property relationships rather than memorizing discrete training scaffolds.

FDA Benchmark versus the FT Training Set (Table S5). The FT set contains 1,845 compounds and represents the tighter reference for evaluating out-of-distribution generalization. Among the 144 FDA molecules, 49% have a maximum Tanimoto similarity below 0.30, indicating low fingerprint similarity with any molecule in the FT training set. This subset constitutes a stringent out-of-distribution test. Performance degrades monotonically as similarity decreases: the MAE increases from 0.057 for molecules with T_{max} in the bin of 0.50−0.70 to 0.122 for those with T_{max} in the bin of 0.10−0.20, while the R^2 declines from 0.932 to 0.664. Importantly, even in the most distant bin (0.10−0.20), the model retains an R^2 of 0.664, demonstrating meaningful predictive power far from the training distribution.

FDA Benchmark versus the QC Pre-Training Set (Table S5). The QC set, comprising 22,609 compounds, provides substantially broader chemical space coverage. Only 7% of FDA molecules fall below a T_{max} of 0.30, compared to 49% for the FT set, and the minimum observed is 0.24. The most distant bin (0.20−0.30) yields an R^2 of 0.669, which is nearly identical to the value of 0.664 obtained for the corresponding FT comparison. This agreement provides an internal consistency check on the degradation rate and confirms that the model's generalization behavior is robust regardless of the reference training set used.

Reaction Products Benchmark (Table S5). All 71 reaction products have a T_{max} of at least 0.30 against both training sets. Within the well-populated similarity regime (T_{max} = 0.30−0.70), the CF mean absolute error is stable at 0.05−0.06 and the R^2 ranges from 0.93 to 0.97. These results suggest that the model operates squarely within its trained domain for synthetically relevant molecules, which are the primary targets of the proposed framework.

In summary, the Tanimoto similarity analysis reveals a monotonic degradation pattern: when the maximum similarity to the fine-tuning set is larger than 0.30, nominal accuracy applies with a CF MAE of approximately 0.05−0.10 , making predictions suitable for quantitative yield estimation. When similarity falls between 0.20 and 0.30, moderate degradation occurs with a CF MAE of 0.11 and an R^2 of about 0.80, and predictions remain useful for ranking and trend analysis but require wider error margins. When similarity drops below 0.20, predictions should be treated as semi-quantitative, and orthogonal verification by a single quantum chemical calculation or qNMR measurement is recommended. This tiered framework is further corroborated by examining the three largest outliers in the FDA benchmark (see Figure S6 in the Supplementary Materials). Consistent with the similarity-based guidance, all three have a maximum Tanimoto similarity below 0.30 to the FT set, placing them in the moderate-to-high error regime where predictions warrant cautious interpretation. However, two of these outliers (Cefditoren Pivoxil and Levothyroxine) have a Tanimoto similarity above 0.53 to the QC pre-training set yet remain poorly predicted, revealing a key limitation of 2D fingerprint similarity. ECFP4 fingerprints encode local atomic connectivity within a four-bond diameter but cannot capture features that dominate UV absorption: heavy-atom quantum effects, such as spin−orbit coupling in iodine, or the electronic consequences of extended heteroatom-rich conjugated systems, where the collective influence on the excited state is not reflected in the topological representation. For the third molecule (Selamectin), the conformational complexity of this macrocycle leads to underestimation by both the QC and the AI model, as the limited TD-DFT conformer sampling may not capture the solution-phase conformational ensemble governing its UV

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

spectrum. Thus, Tanimoto similarity alone is insufficient to identify all failure modes, and molecules with unusual electronic or conformational characteristics may evade detection by this metric.

Stratifying prediction errors by the experimental CF value reveals a multiplicative error structure. The MAE scales with the CF magnitude, increasing from 0.032 for CF values between 0.10 and 0.30 to 0.186 for CF values between 1.50 and 2.00 on the FDA benchmark (Table S6). In contrast, the MAPE remains stable at 9.4−16.4% across all CF bins. This behavior is expected for a multiplicative quantity and underscores the importance of reporting percentage-based metrics for practitioners. Two practical considerations arise from this analysis. First, a low CF value, typically below 0.3, indicates weak UV absorption, in which case experimental noise may dominate the measurement. Users should therefore verify the chromatographic integration quality before attributing any discrepancy to model error. Second, predictions at extreme CF values, specifically below 0.2 or above 2.0, lack sufficient statistical support due to data sparsity in both the training and benchmark sets. Such predictions should be treated as qualitative estimates, and orthogonal verification is recommended where feasible. A similar conclusion was also observed for the reaction dataset (Table S6). Together with the Tanimoto similarity result, these analyses offer a tiered, actionable framework for deploying the model in high-throughput experimentation workflows, enabling users to calibrate trust in individual predictions and determine when orthogonal verification is needed.

Robustness and Limitations

Real-world variability can affect the reliability of the proposed AI-based LC-UV yield estimation framework, which is summarized in the following discussions.

Internal-Standard Design. The CF is defined as a ratio of the analyte UV area to that of a coinjected internal standard, which cancels first-order effects of gradient shape, detector drift, injection volume variability, and day-to-day or instrument-to-instrument variation. This is a standard advantage of internal standard-based quantification. However, second-order effects such as differential detector nonlinearity at very different analyte-to-standard concentrations may introduce a residual bias. Formal interlaboratory reproducibility studies will be the subject of future work, though consistent performance across 178 reactions under 25 conditions (MAE = 4.7%) supports within-laboratory robustness.

Detector Linearity and Saturation. All data were acquired within the validated linear range of the diode-array detector; samples exceeding this range were diluted and reanalyzed. Full-spectrum integration further reduces the saturation risk by distributing the signal across wavelengths.

Low-UV Compounds and Strong Chromophores. Molecules lacking chromophores above 190 nm (e.g., saturated hydrocarbons) yield poor signal-to-noise ratios and are unsuitable for this method. Conversely, compounds with strong chromophores (e.g., polyenes, porphyrins) are well handled as long as detector response remains linear, as demonstrated by accurate predictions for FDA drugs containing such motifs.

Co-Elution and Peak Assignment. Co-elution with UV-absorbing byproducts can bias the integrated area and thus enlarge the deviation between estimated and experimental yields. The target signal is recovered as follows. First, the full UV spectrum is integrated over wavelength to yield a single integrated-absorbance trace as a function of retention time, which carries the peaks of all components in the mixture. The target analyte's peak is then located on this trace by LC-MS-based peak identification using its expected molecular ion. When the target peak partially overlaps with a co-eluting neighbor, peak fitting is applied in the retention-time dimension to attribute the integrated area to the target compound. This deconvolution is reliable for partially overlapping peaks. It fails when (i) co-eluting species share the same nominal mass (e.g., stereoisomers, regioisomers), or (ii) peaks overlap too heavily to be resolved by fitting; such cases are flagged and excluded from yield quantification, and manual inspection or orthogonal separation is required.

Product Degradation. Degradation during LC analysis (e.g., hydrolysis and oxidation) leads to underestimated yields. This limitation is shared by all LC-based quantification methods; users should verify the stability for reactive functional groups.

Solution-Phase Speciation. The model inputs the neutral SMILES structure, but the actual absorbing species may differ because of salt counterions, tautomerism, or protonation state. Under acidic reversed-phase conditions, basic nitrogen atoms are protonated, shifting UV bands by tens of nanometers. Full-spectrum integration mitigates these shifts but does not eliminate them entirely. The experimental fine-tuning set (1 845 compounds) includes diverse protonation states and tautomeric forms, providing implicit exposure to this variability.

Model Comparison and Ablation Study

To evaluate the performance and architectural contributions of this proposed framework, we compared the GT+KAN model against multiple baseline methods and conducted a critical ablation study (see "Performance Comparison of GT+KAN and Other Prediction Models" Section in Supplementary Materials). Compared to conventional fingerprint-based models, the GT+KAN model used in this study provides substantial performance gains, particularly on out-of-distribution FDA molecules (see Table S7 in Supplementary Materials). Crucially, replacing KAN with a parameter-matched MLP confirmed that the KAN module further enhances cross-distribution generalization. In addition, the performance of GT+KAN model was compared to Chemprop model[24,25] which was trained following the same TD-DFT pretraining and experimental fine-tuning protocol used for GT+KAN. Under a parameter-matched configuration, GT+KAN achieves performance comparable to the state-of-the-art (SOTA) Chemprop MPNN after fine-tuning, with test-set accuracy slightly favoring GT+KAN (R^2 = 0.921 vs 0.908) and FDA accuracy slightly favoring Chemprop (R^2 = 0.852 vs 0.860), both within narrow margins. This demonstrates that GT+KAN is competitive not only with the simple ECFP-based baselines, but also with SOTA graph neural network architectures of equivalent capacity, supporting its applicability as the predictor in this detect−predict framework.

CONCLUSION

This work addresses the longstanding trade-off between scalability and accuracy in HTE-driven molecular discovery by introducing an integrated "detect-predict" framework. In contrast to the widely used UV area normalization method (Figure S7), this new workflow can deliver robust yield

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

quantification for reactions conducted under real-world conditions, resulting in 59.0% reaction yields estimated with absolute error <5% and 88.8% with error <10% (Figure 4). This capability can be attributed to three contributions: (1) prediction of the CF by leveraging LC-UV data across the entire UV spectrum (rather than the use of a specific wavelength or UV peak maximum); (2) use of QC calculated UV spectra of a virtual library of structurally diverse compounds for pre-training to effectively capture key structure-property relationships; and (3) collection of experimental data using unified protocols for model fine-tuning to mitigate prediction outliers and thus enhance the prediction robustness. Unlike single-wavelength based approaches, fragment-based QC ε predictions, or structure-based ML models sensitive to spectral cliffs, this full-spectrum approach leverages multidimensional LC-UV data to overcome the accuracy limitations of prior methods. By hierarchically training a predictive model with pre-training on QC data (22,609 compounds) for chemical space coverage and fine-tuning on experimental HTE data (1,845 compounds) for real-world fidelity, we close the "data diversity-quality gap" for yield quantification based on UV spectrum. This gap has historically undermined the performance of literature-curated models. The concentration benchmark with 144 FDA-approved drugs shows an MAE of 0.167 μmol/mL, and the 178 reactions spanning four widely used chemical reaction types with 71 unique products and 25 distinct conditions show a yield quantification MAE of 4.74%. These results suggest that the model generalizes robustly across real-world complexity, achieving accurate yield quantification without requiring analyte standards. The final practical validation against 60 diverse cross-coupling and Minisci reactions revealed that AI-quantified yields exhibit a systematic, physically explainable positive bias relative to isolated yields, consistent with the expected material losses incurred during small-scale purification. This work thus shifts the focus from isolated yield to a crude yield metric that reflects the reaction outcome prior to purification losses, providing a more informative basis for data-driven reaction optimization and molecular discovery in both large-scale and miniaturized HTE workflows. This methodology offers a practical alternative to traditional "detect-purify-quantify" workflows by enabling faster, more scalable "detect-predict" quantification in high-throughput experimentation. Through AI-driven LC-UV analysis, it enables direct yield quantification without physical isolation, accelerating optimization cycles from weeks to days while eliminating solvent-intensive purification and aligning with Green Chemistry principles.[26] The framework applies broadly to time-sensitive applications including high-throughput direct-to-biology experiments,[27] pharmaceutical lead optimization,[28] catalyst screening,[3,29] and continuous manufacturing,[30] where real-time prediction enhances sustainability while maintaining analytical rigor.

To validate the practical utility and physical plausibility of the machine learning prediction approach, we conducted complementary computational cost and interpretability analyses. A quantitative comparison reveals that while QC calculations require an average of 121.5 core-hours per molecule, the GT+KAN model achieves inference in milliseconds, enabling a 10^5−10^6-fold acceleration with more reliable prediction performance (see Table S8 in Supplementary Materials). Combined with this substantial speed advantage, the machine learning model offers a practical alternative to QC calculation: it delivers accuracy comparable to QC at millisecond cost, and within the reaction-centered domain for which it was designed, it achieves superior performance (Figures 3 and S4). Furthermore, perturbation-based interpretability analysis[31] demonstrates that the model correctly identifies conjugated π-systems as primary contributors to absorption spectral CFs, aligning with established quantum mechanical principles and confirming that the learned structure-property relationships are physically meaningful rather than correlational artifacts (see Figure S8 in Supplementary Materials). Together, these analyses substantiate that this approach provides both the computational efficiency and accuracy required for practical deployment and the interpretable insights necessary for new molecular discovery.

The reliability of chemical quantification, particularly in UV-based methods, is contingent upon high-quality data as cumulative experimental errors can significantly impact predictive model performance. This analysis demonstrates that model accuracy scales directly with both dataset size and data quality (Figure S9), underscoring the critical need for standardized and well-curated experimental data. While the proposed detect-predict strategy effectively resolves the scalability-accuracy trade-off inherent in conventional methods, its full potential is best realized with access to large, consistent datasets. Future work will therefore focus on developing automated high-throughput experimental workflows and standardized data capture protocols. This path forward promises to further enhance the accuracy and broad applicability of this AI-enabled reaction yield quantification approach, establishing it as a robust and efficient alternative to resource-intensive purification-based methods for molecular discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

Readers can access the data and codes used to reproduce this study's predicted CFs on GitHub (https://github.com/xtalpi-inc/lcms2yield). The shared data and codes are as follows: 1. The SMILES and QC-calculated CFs for 22,609 compounds used at the pretraining stage; 2. The SMILES and LC-UV-derived experimental CFs for 1,845 compounds used at the fine-tuning stage; 3. The SMILES and LC-UV-derived experimental CFs for 144 FDA approved drug compounds used for the first independent benchmark; 4. The SMILES and LC-UV-derived experimental CFs for target products from 178 real-world reactions used for the second independent benchmark; 5. The SMILES and LC-UV-derived experimental CFs for target products from another 34 Suzuki-Miyaura cross-coupling reactions and 26 Minisci reactions used for the third independent benchmark; 6. The CF prediction model accepts SMILES as input and outputs the predicted CFs.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.6c00812.

Comprehensive computational and experimental details, including the procedure for determining experimental correction factors (CFs), the five-step QC workflow for computing UV absorbance spectra, the construction and diversity analysis of the virtual pre-training and fine-tuning datasets, and an analysis of the structural similarity between the fine-tuning/benchmark molecules

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX

and the pre-training set; a performance comparison of the GT+KAN model against other prediction models, a computational cost analysis justifying the machine learning approach, and an interpretability analysis of molecular structure contributions to predicted CFs; eight tables and nine figures, including the chemical space distribution of datasets (Figures S1−S2 and Tables S1−S3), the effect of solvent choice on QC-calculated CF (Figure S3), hierarchical performance decomposition of the reaction yield benchmark (Table S4), domain of applicability analysis (Tables S5−S6 and Figure S6), calibration curves for reaction yield calculation (Figure S5), model performance on the benchmark set (Figure S4 and Tables S7−S8), a comparison of predicted reaction yields from different models (Figure S7), interpretability analysis of molecular structure contributions to predicted CFs (Figure S8), and learning curves analyzing the effects of training data size and quality (Figure S9) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Christopher J. Welch − *Welch Innovation, LLC, Fishers, Indiana 46040, United States;* orcid.org/0000-0002-8899-4470; Email: christopher@welchinnovation.com

Peiyu Zhang − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*; Email: peiyu.zhang@xtalpi.com

Authors

Jiuchuang Yuan − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Mingjun Yang − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China;* orcid.org/0000-0003-0928-8541

Fan Yan − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Jing Guo − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Lin Zhang − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Guosheng Dou − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Minjun Liu − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Lu Tan − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Liwen Fang − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Guanfeng Yang − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Qun Zeng − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Jun Yan − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Xuekun Shi − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Sarah Trice − *XtalPi, Inc, Somerville, Massachusetts 02143, United States*

Jian Ma − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China*

Shuhao Wen − *Shenzhen Jingtai Technology Co., Ltd. (XtalPi), Shenzhen 518045, China; XtalPi, Inc, Somerville, Massachusetts 02143, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.6c00812

Author Contributions

‖J.Y. and M.Y. contributed equally to this work. P.Z. originated the idea and conceived the study; J.Y.(1), M.Y., C.W. designed the study, analyzed the data, and wrote the manuscript; F.Y. developed the target product identification algorithm from LC-UV spectrum; J.G. designed the benchmark case for Minisci reactions; L.Z. and G.D. designed the data collection protocol; M.L. and G.Y. conducted the reactions and LC-UV experiments; L.T., L.F., and Q.Z. proposed the Q_C. workflow and performed partial calculations; J.Y.(2) supervised the reaction procedure and developed the workflow; X.S., S.T., J.M., and S.W. discussed and designed the study; all authors revised and approved the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was partially funded by the Shenzhen Science and Technology Program (KJZD20240903100304007 and KTD20210811090114013).

REFERENCES

(1) Rinehart, N. I.; et al. A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C−N couplings. *Science* **2023**, *381*, 965−972.

(2) Perera, D.; et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **2018**, *359*, 429−434.

(3) Angello, N. H.; et al. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **2022**, *378*, 399−405.

(4) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360*, 186−190.

(5) Tom, G.; et al. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124*, 9633−9732.

(6) Reizman, B. J.; Wang, Y.-M.; Buchwald, S. L.; Jensen, K. F. Suzuki−Miyaura cross-coupling optimization enabled by automated feedback. *React. Chem. Eng.* **2016**, *1*, 658−666.

(7) Mennen, S. M.; et al. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23*, 1213−1242.

(8) Santanilla, A. B.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015**, *347*, 49−53.

(9) Schilling, K.; Holzgrabe, U. Recent applications of the Charged Aerosol Detector for liquid chromatography in drug quality control. *J. Chromatogr. A* **2020**, *1619*, 460911.

(10) Zhang, K.; Kurita, K. L.; Venkatramani, C.; Russell, D. Seeking universal detectors for analytical characterizations. *J. Pharm. Biomed. Anal.* **2019**, *162*, 192−204.

(11) Grainger, R.; Whibley, S. A Perspective on the Analytical Challenges Encountered in High-Throughput Experimentation. *Org. Process Res. Dev.* **2021**, *25*, 354−364.

(12) Hu, M.; et al. Continuous collective analysis of chemical reactions. *Nature* **2024**, *636*, 374−379.

(13) Fieser, L. F.; Fieser, M.; Rajagopalan, S. ABSORPTION SPECTROSCOPY AND THE STRUCTURES OF THE DIOSTEROLS. *J. Org. Chem.* **1948**, *13*, 800−806.
(14) Woodward, R. B. Structure and the Absorption Spectra of α,β-Unsaturated Ketones. *J. Am. Chem. Soc.* **1941**, *63*, 1123−1126.
(15) McNaughton, A. D.; et al. Machine Learning Models for Predicting Molecular UV−Vis Spectra with Quantum Mechanical Properties. *J. Chem. Inf. Model.* **2023**, *63*, 1462−1471.
(16) Fitch, W. L.; et al. Prediction of Ultraviolet Spectral Absorbance Using Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 830−840.
(17) Urbina, F.; et al. UV-adVISor: Attention-Based Recurrent Neural Networks to Predict UV−Vis Spectra. *Anal. Chem.* **2021**, *93*, 16076−16085.
(18) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; et al. Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **2018**, *361*, No. eaar6236.
(19) Gold, E. R.; Cook-Deegan, R. AI drug development's data problem. *Science* **2025**, *388*, 131−131.
(20) McDonald, M. A.; Koscher, B. A.; Canty, R. B.; Jensen, K. F. Calibration-free reaction yield quantification by HPLC with a machine-learning model of extinction coefficients. *Chem. Sci.* **2024**, *15*, 10092−10100.
(21) Zhong, H.; Liu, Y.; Sun, H.; Liu, Y.; Zhang, R.; Li, B.; Yang, Y.; Huang, Y.; Yang, F.; Mak, F. S.; et al. Towards global reaction feasibility and robustness prediction with high throughput data and bayesian deep learning. *Nat. Commun.* **2025**, *16* (1), 4522.
(22) Liu, Z; Wang, Y; Vaidya, S; Ruehle, F; Halverson, J; Soljacic, M; Hou, T; Tegmark, M., et al. KAN: Kolmogorov-Arnold Networks. *arXiv*. **2024**.
(23) Rampášek, L.; Galkin, M.; Dwivedi, V.P.; Luu, A.T.; Wolf, G.; Beaini, D. et al.Recipe for a General, Powerful, Scalable Graph Transformer. *arXiv*. **2022**
(24) Heid, E.; et al. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9−17.
(25) Greenman, K. P.; Green, W. H.; Gómez-Bombarelli, R. Multi-fidelity prediction of molecular optical peaks with deep learning. *Chem. Sci.* **2022**, *13*, 1152−1162.
(26) Welch, C. J.; et al. Greening analytical chromatography. *Green. Anal. Chem.* **2010**, *29*, 667−680.
(27) Thomas, R. P.; et al. A direct-to-biology high-throughput chemistry approach to reactive fragment screening. *Chem. Sci.* **2021**, *12*, 12098−12106.
(28) Nippa, D. F.; et al. Enabling late-stage drug diversification by high-throughput experimentation with geometric deep learning. *Nat. Chem.* **2024**, *16*, 239−248.
(29) Nie, W.; Wan, Q.; Sun, J.; Chen, M.; Gao, M.; Chen, S. Ultra-high-throughput mapping of the chemical space of asymmetric catalysis enables accelerated reaction discovery. *Nat. Commun.* **2023**, *14*, 6671.
(30) Hartman, R. L.; McMullen, J. P.; Jensen, K. F. Deciding Whether To Go with the Flow: Evaluating the Merits of Flow Reactors for Synthesis. *Angew. Chem. Int. Ed.* **2011**, *50*, 7502−7519.
(31) Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C.-Y.; et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat. Commun.* **2023**, *14*, 2585.

https://doi.org/10.1021/jacsau.6c00812
JACS Au XXXX, XXX, XXX−XXX