Supplementary Information (SI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2026

A Low-Cost Automated Platform for Fast and Accurate pH Control via Physics-Informed Active Learning

Quan Yang[1], Zhipeng Xiang[1], Zhiwen Zhu[1], Tairan Yang[1], Qiang Sun[1*]

[1]Materials Genome Institute, Shanghai Engineering Research Center for Integrated Circuits and Advanced Display Materials, Shanghai University, 200444 Shanghai, China

*E-mail: qiangsun@shu.edu.cn

Table of Contents

1. Materials

Unless otherwise specified, all solvents and chemicals were purchased from commercial suppliers and used without further purification. The buffers used in the experiments, including potassium dihydrogen phosphate, citric acid, sodium acetate, and ammonium chloride, were of analytical grade or higher.

The relevant code and datasets for this article have been uploaded to the GitHub repository: https://github.com/YQnb/Low-cost-titration-system

2. Development of the low-cost automated platform

To validate the reliability of the proposed active learning workflow, we designed and constructed a low-cost, modular automated titration platform capable of achieving full closed-loop pH adjustment based on algorithmic decision-making. [1] The core hardware components of the platform include: Arduino Uno microcontroller, a peristaltic pump (DFRobot Gravity Series) with a volume addition accuracy of 0.1 mL, and an analog pH sensor (DFRobot Gravity Series, Arduino-compatible), calibrated via three-point calibration using pH 4.01, 7.00, and 10.01 standard buffers before each experimental run, custom-built magnetic stirring device driven by rotating magnets operated at a constant speed of 600 rpm to ensure sufficient mixing while minimizing the introduction of noise from vortex or bubble formation, initial liquid volume of 30 mL, with titrants being 0.05 mol/L NaOH and HCl solutions.

All hardware control modules are integrated onto a mounting plate using 3D-printed structural components (Figure S1), providing excellent mechanical stability and modular characteristics. The platform's control software, developed in Python, integrates the physics-informed surrogate model and active learning algorithm, enabling full automation from pH measurement and intelligent decision-making to titrant delivery. To ensure measurement accuracy, the electrode was rinsed with deionized water after each pH measurement.

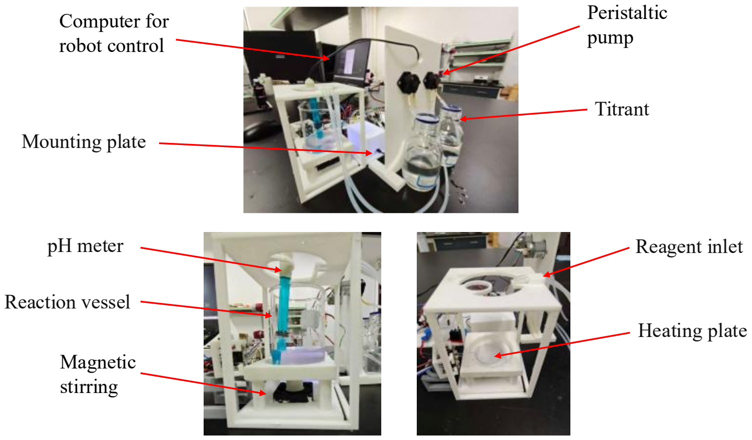

Figure S1 The low-cost automated titration setup.

Table S1. Cost comparison with commercial autotitrators.

Product	Manufacturer	Approximate Price (USD)	Country
ZDJ-4A	INESA	~3155 $	China
HI902C1	HANNA	~10100 $	Italy
Excellence T5	Mettler Toledo	~20200 $	Switzerland
TitroLine 7000	SI Analytics	~11500 $	Germany

Compared with commercial autotitrators (Table S1), our platform offers three distinct advantages: cost-effectiveness at approximately 100 USD—less than 5% of entry-level commercial systems, transparency and modularity with open-source hardware and software enabling full customization, and educational accessibility for hands-on learning in automated experimentation.

However, several limitations should be acknowledged. Commercial instruments such as the SI Analytics TitroLine 7000 offer superior pH resolution (0.001 pH vs. 0.01 pH) and potential resolution (0.1 mV), along with higher precision in titrant delivery (0.07% reproducibility vs. 0.1 mL stepwise addition). These systems also provide advanced capabilities including multi-mode titration (pH, mV, μA, photometric), automated sample handling, GLP compliance, and robust chemical resistance. Our platform currently lacks long-term stability under continuous operation, and compatibility with industrial automation standards.

3. Generation of the theoretical titration curve

The theoretical titration curves, serving as training data for the physics-informed surrogate model, were generated by numerically solving the fundamental chemical equilibrium equations governing acid-base systems. Our approach rigorously accounts for multi-buffer systems and ionic strength effects, ensuring physicochemical consistency throughout the entire dataset. The theoretical titration curves were generated using a custom Python script that calculates pH as a function of titrant volume by solving the charge balance equation for aqueous systems. The model is based on the following fundamental principles:

For each buffer component, the equilibrium between the two species is described by:

$$[HA] = C_{total} \cdot \frac{[H^+]}{[H^+] + K_a} \qquad [A^-] = C_{total} \cdot \frac{K_a}{[H^+] + K_a} \qquad Eqn.S1$$

where C_{total} is the initial concentration of the component.

Accounting for volume changes during titration, the real-time concentration of each component is calculated as:

$$C_{current} = C_{initial} \times \frac{V_{initial}}{V_{total}} \quad Eqn. S2$$

At any point during titration, the total positive charge equals the total negative charge:

$$[\mathrm{H}^{+}] - \frac{\mathrm{K_w}}{[\mathrm{H}^{+}]} + \sum_{i} \mathrm{z_{c,i}C_{c,i}} - \sum_{\mathrm{j}} |\,\mathrm{z_{a,j}}|\mathrm{C_{a,j}} + \sum_{\mathrm{k}} (\mathrm{z_{HA,k}[HA]_k} + \mathrm{z_{A,k}[A^-]_k}) = 0 \quad Eqn. S3$$

where:$\mathrm{C_{c,i}}$、$\mathrm{C_{a,j}}$ represent the concentrations of additional cations and anions (including those introduced by titrants and initial counterions), in mol·L^{-1}. $\mathrm{[HA]_k}$、$\mathrm{[A^-]_k}$ represent the concentrations of acidic and basic species of the k-th buffer system. z denotes the charge (with sign) of the corresponding ion. The equation is solved by expressing all unknown terms as functions of hydrogen ion concentration $[\mathrm{H}^{+}]$.

The numerical solution of Eqn. S3 was implemented using the scipy.optimize.brentq method in Python to ensure robust computation across the pH range of 0–14. The Brent method was applied over the interval $[10^{-14}, 1]$ to locate the root corresponding to $[\mathrm{H}^{+}]$. The computational procedure includes the following steps: 1. Set the pK_a values, initial concentrations, initial protonation states (HA/A$^-$), and counterion charges for all components. 2. Add titrant in 0.1 mL increments, updating the total system volume and the concentration of each component accordingly. 3. At each titration point, solve for the hydrogen ion concentration $[\mathrm{H}^{+}]$ and compute the corresponding pH value. 4. The calculation terminates when the pH falls outside the range of 2–12 or when the total titrant volume exceeds 30 mL.

4. Neural network architectures

To achieve efficient learning and rapid prediction of theoretical titration behavior, we designed a specialized neural network architecture capable of effectively handling the compositional nature of chemical systems and the continuous variation of the titration process. The network employs a dual path encoding structure that separately processes system composition features and process variables, aligning with the physical essence of the titration problem. The system composition is handled using DeepSet architecture,[2] enabling effective encoding of variably sized, unordered sets of buffer components. The process variable (titrant volume) is processed by a dedicated volume encoder. The outputs are then integrated through a fusion network to make predictions.

4.1 DeepSet Encoder

The composition encoder processes chemical information from each buffer component through a structured feature extraction pipeline. Input features for each component comprise its pKa value, concentration, initial protonation state (HA/A$^-$), and counterion charge, with the system supporting up to four distinct components. The encoding process follows these stages: Each component's feature vector x_i is first transformed by a shared component-encoding multilayer perception (MLP) ϕ that projects individual components into a latent representation space. These encoded component representations are then aggregated through permutation-invariant sum pooling to form a unified system-level embedding. Finally, a system-level MLP ρ processes this aggregated representation to produce the final compositional descriptor z_{system}:

$$z_{system} = \rho\left(\sum_{i=1}^{N} \phi(x_i)\right) \quad Eqn.S4$$

where x_i denotes the feature vector of the i-th component, ϕ represents the component-encoding MLP, and ρ signifies the system-level MLP. This hierarchical encoding scheme ensures invariance to component ordering while effectively capturing emergent properties arising from component interactions.

4.2 Volume Encoder

The titrant volume is first processed using sinusoidal positional encoding to capture its sequential nature during titration. The encoded volume is then mapped to a 64-dimensional feature space via MLP.

4.3 Fusion and Prediction

The 256-dimensional system descriptor and the 64-dimensional volume feature are concatenated and passed through an MLP to predict the final pH value (Figure S2).

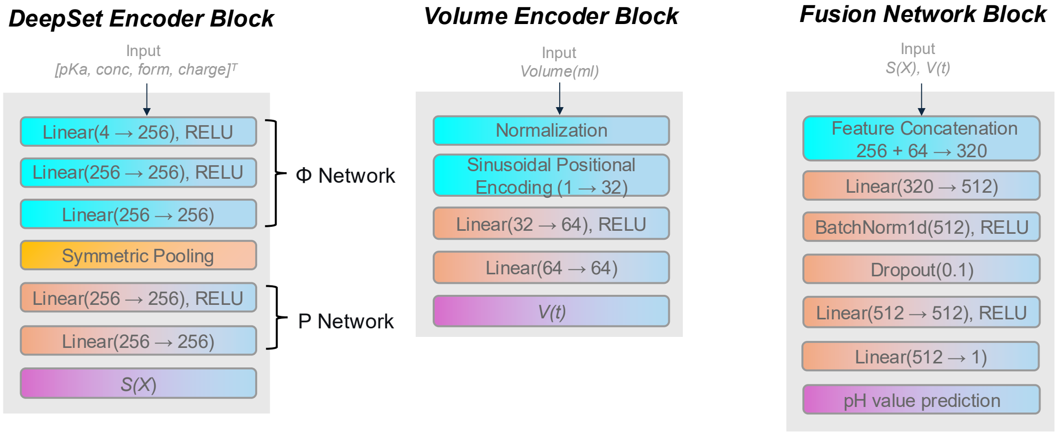

Figure S2 Architecture of various network modules.

To validate the architectural advantages of our approach, we constructed a baseline model comprising a standard fully-connected neural network with comparable parameter count. This baseline was designed to predict pH from titrant volume alone, as preliminary tests showed that incorporating system composition features severely impaired its performance and resulted in unphysical volume–pH relationships.

4.4 CVFNet vs Baseline model

CVFNet's architecture enables seamless integration of both input modalities when available. When only the titrant volume is provided, CVFNet utilizes its dedicated VolumeEncoder module, which employs sinusoidal positional encoding to capture nonlinear volume-pH relationships. Simultaneously, its chemical component encoder remains active but receives zero-padded placeholder inputs, allowing the fusion network to maintain a consistent forward pass while relying primarily on volume-derived features. To demonstrate the effectiveness of our CVFNet model, we constructed a five-layer neural network with an equivalent number of parameters as a baseline model for comparison. As shown in Figure S3a-b, both CVFNet and the baseline model used only titrant volume as input, yet CVFNet demonstrated superior prediction accuracy due to its inherent architectural inductive biases. In Figure S3c-d, CVFNet used titrant volume and component features as input, while the baseline model remained unchanged with volume-only inputs, as it fundamentally lacks the capacity to process compositional features without performance degradation.

These results clearly demonstrate that conventional architectures fail to learn meaningful physicochemical relationships even from basic volume-pH data and cannot utilize compositional information when available. In contrast, CVFNet's structured encoding strategy, featuring a dedicated composition encoder and separate volume pathway, enables robust learning across both input configurations. This specialized design captures complex composition–property mappings that standard architectures cannot represent, providing an efficient model for subsequent active learning cycles.

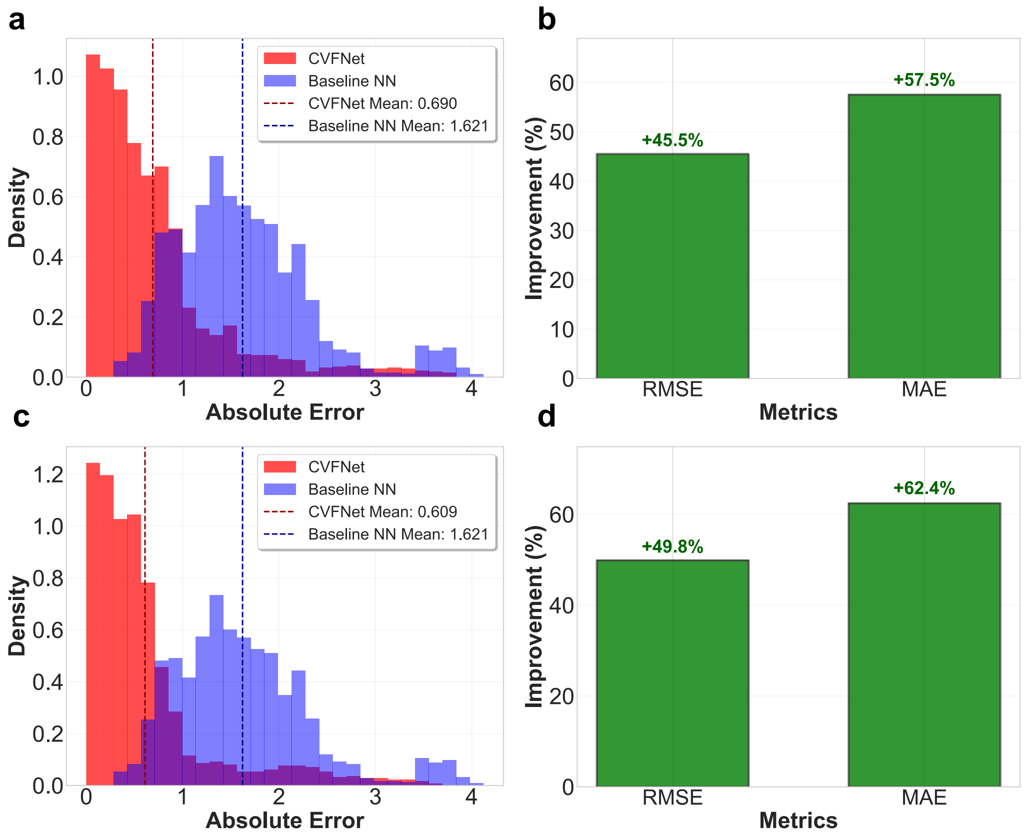

Figure S3 Performance comparison of CVFNet against baseline model under different input configurations. (a, b) CVFNet using only titrant volume as input *vs* baseline model using only volume as input; (c, d) CVFNet using titrant volume and component features as input *vs* baseline model using only volume as input.

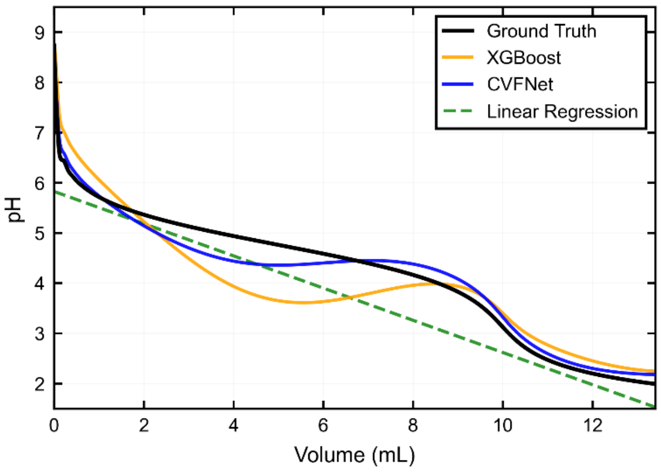

Figure S4. Comparison of Linear Regression, XGBoost, and CVFNet on a representative titration curve. The x-axis represents the volume of HCl added (mL), and the y-axis represents the measured pH value.

Linear Regression completely fails to capture the nonlinear titration behavior. XGBoost shows improved performance but still deviates near sharp pH transitions. CVFNet most closely follows the ground truth across the entire range, achieving the best predictive accuracy. This demonstrates that neither linear models nor tree-based methods are sufficient for accurate titration curve modeling, justifying our selection of a specialized neural network architecture.

5. Modified Expected Improvement acquisition function

Our system employs a hybrid architecture combining a neural network (NN) and a Gaussian process (GP),[3] designed to provide the acquisition function with both accurate predictions and reliable uncertainty estimates. This architecture first utilizes a pre-trained neural network to generate baseline predictions of the titration curve, leveraging the physicochemical priors learned from large-scale theoretical data:

$$u_{NN}(x) = NN(x;\theta) \quad Eqn.S5$$

where θ represents the parameters of the pre-trained network.

To account for potential discrepancies between the theoretical model and the actual experimental system, we introduce Gaussian process regression to learn the systematic residual between experimental observations and neural network predictions:

$$r(x) = pH_{obs} - u_{NN}(x) \quad u_{GP}(x) = GP(x;r) \qquad Eqn.S6$$

The final prediction function $u(x)$ integrates both the baseline prediction from the neural network and the residual correction from the Gaussian process, thereby combining physical priors with experimental specificity:

$$u(x) = u_{NN}(x) + u_{GP}(x) \quad Eqn.S7$$

The uncertainty of the model prediction is quantified by the Gaussian process:

$$\sigma(x) = \sigma_{GP}(x) \quad Eqn.S8$$

For each candidate titration volume x, we compute predicted error: $E_{pred}(x) =| u(x) - pH_{target} |$, current best error: $E_{best} = min_i \mid pH_{obs}^{(i)} - pH_{target} \mid$.

The improvement for candidate x is defined as:

$$I(x) = E_{best} - (E_{pred}(x) - \xi) \quad Eqn.S9$$

where $\xi = 0.01$ is a small exploration incentive parameter.

The acquisition function follows the classical EI expression:

$$\alpha_{EI(x)} = I(x) \cdot \Phi(\frac{I(x)}{\sigma(x)}) + \sigma(x) \cdot \phi(\frac{I(x)}{\sigma(x)}) \quad Eqn.S10$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution and probability density functions, respectively.

To ensure chemical plausibility, we impose directional constraints:

If $pH_{current} > pH_{target}$: only consider $x \leq x_{current}$ (acid addition)

If $pH_{current} < pH_{target}$: only consider $x \geq x_{current}$ (base addition)

Additionally, step size constraints $\Delta x_{min} \leq| x - x_{current} |\leq \Delta x_{max}$ ensure practical titration increments.

At each iteration, the next experimental point is selected as:

$$x_{next} = \underset{x}{argmax}\ \alpha_{EI}(x) \quad Eqn. S11$$

This approach automatically balances exploitation (selecting points with high expected improvement) and exploration (investigating regions with high uncertainty), while respecting the fundamental thermodynamics of acid-base systems through the physically informed prior.

Table S1 shows the number of iterations required to reach target pH (7.0 ± 0.1) for five different acquisition functions: Random sampling, Uncertainty sampling, Upper Confidence Bound (UCB), standard Expected Improvement (EI), and our modified EI. Each value represents the average of three independent experimental measurements.

Table S2 Comparison of Iteration Counts for Different Active Learning Strategies

Strategy	**Standard EI**	**UCB**	**Modified EI**	**Random**	**Uncertainty**
Iterations	6.1	6.3	4.9	8	6

6. Active learning process

This example demonstrates the sequential refinement of the hybrid model through the iterative cycle of prediction, experimentation, and Bayesian updating. The Gaussian process residual model effectively bridges the simulation-to-reality gap, enabling convergence to the target pH within five iterations while maintaining physicochemical consistency through the pre-trained neural network prior.

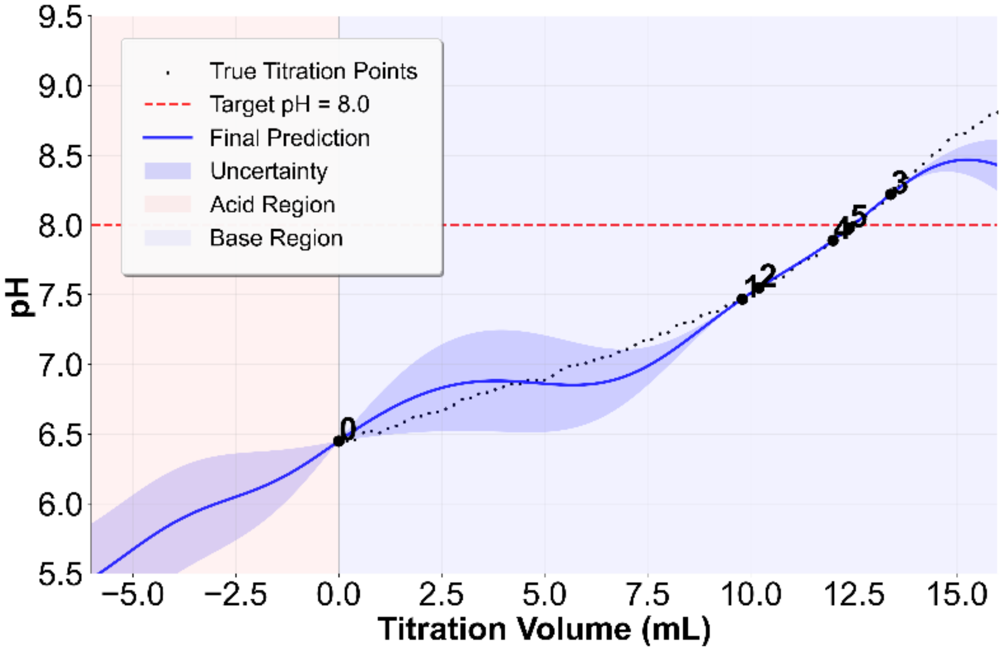

Figure S5 Active learning process for pH adjustment in a phosphate-citrate mixed buffer

system (target pH = 8.0).

7. Reference

(1) A. Pomberger, N. Jose, D. Walz, J. Meissner, C. Holze, M. Kopczynski, P. Müller-Bischof and A. A. Lapkin, *Chemical Engineering Journal*, 2023, **451**, 139099.

(2) M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Póczos, R. Salakhutdinov and A. J. Smola, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA, 2017, pp. 3394–3404.

(3) B. Shahriari, K. Swersky, Z. Wang, R. P. Adams and N. de Freitas, *Proceedings of the IEEE*, 2016, **104**, 148–175.