Supporting Information Explainable Active Learning Framework for Ligand Binding Affinity Prediction

E-mail:

Gaussian Process Details and Equations

This section provides a detailed mathematical and conceptual overview of Gaussian Process (GP) Regression, the surrogate model used in our active learning framework. The content here is foundational and follows the standard treatment of GPs in machine learning.[1–3]

Gaussian Process Definition

A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. A GP can be thought of as a distribution over functions, and it is fully specified by its mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{1}$$

The mean function $m(\mathbf{x})$ represents the expected value of the function f at input $\mathbf{x}$. It defines the average of the function before any data is observed. In this study, as is common practice, the mean function was set to zero ($m(\mathbf{x}) = 0$), assuming that any underlying trend in the data is captured by the more flexible covariance function. The covariance function, also known as the kernel, $k(\mathbf{x}, \mathbf{x}')$, models the correlation between the function values at two different input points, $\mathbf{x}$ and $\mathbf{x}'$. It encodes our prior assumptions about the properties of the function we are modeling, such as its smoothness, periodicity, or stationarity. The kernel's role is to measure the "similarity" between data points; points that are "close" in the input space (as defined by the kernel) are expected to have similar output values.

Gaussian Process Regression

In a regression setting, we assume that our observed target values y are related to the latent function $f(\mathbf{x})$ by i.i.d. Gaussian noise:

$$y = f(\mathbf{x}) + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{2}$$

Here, σ_n^2 is the variance of the observation noise, which is a hyperparameter of the model.

Given a training dataset $\mathcal{D} = \{(\mathbf{X}, \mathbf{y})\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of N observations, and a set of test points $\mathbf{X}_*$, the GP prior implies that the observed targets $\mathbf{y}$ and the function values at the test points $\mathbf{f}_*$ are jointly Gaussian:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{pmatrix}\right) \tag{3}$$

where:

- $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is the $N \times N$ covariance matrix where each element $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.
- $\mathbf{K}(\mathbf{X}, \mathbf{X}_*)$ is the covariance matrix between the training and test points.
- $\mathbf{K}(\mathbf{X}_*, \mathbf{X}_*)$ is the covariance matrix of the test points.
- $\mathbf{I}$ is the identity matrix.

Predictive Equations

By conditioning the joint Gaussian prior distribution on the observed training data, we obtain the posterior predictive distribution for the function values $\mathbf{f}_*$ at the test points $\mathbf{X}_*$. This posterior is also a Gaussian distribution, with a predictive mean $\boldsymbol{\mu}(\mathbf{X}_*)$ and predictive covariance $\boldsymbol{\Sigma}(\mathbf{X}_*)$.

For a single test point $\mathbf{x}_*$, the predictive mean and variance are given by:

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} \tag{4}$$

$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* \tag{5}$$

where:

- $\mathbf{K}$ is the $N \times N$ covariance matrix of the training data.
- $\mathbf{k}_*$ is the $N \times 1$ vector of covariances between the test point $\mathbf{x}_*$ and each of the training points.
- $k(\mathbf{x}_*, \mathbf{x}_*)$ is the prior variance at the test point.

The predictive mean $\mu(\mathbf{x}_*)$ provides the model's best estimate of the function value at $\mathbf{x}_*$, while the predictive variance $\sigma^2(\mathbf{x}_*)$ provides a measure of the model's uncertainty in that prediction. This uncertainty is crucial for active learning, as it allows acquisition functions like UCB to balance exploring regions of high uncertainty with exploiting regions of high predicted affinity.

Hyperparameter Optimization

The GP model's behavior is governed by the hyperparameters of the kernel function (e.g., lengthscale ℓ, outputscale s, shape parameter α) and the noise variance σ_n^2. These parameters are learned from the data by maximizing the log marginal likelihood of the observations:

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_{\boldsymbol{\theta}} + \sigma_n^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}} + \sigma_n^2\mathbf{I}| - \frac{N}{2}\log 2\pi \tag{6}$$

where $\boldsymbol{\theta}$ represents the set of all kernel hyperparameters. This objective function naturally balances model fit (the first term) with model complexity (the second term, a log-determinant penalty), thereby providing a principled way to tune the model and avoid overfitting. In this study, gradient-based optimization (using the Adam optimizer) was employed to find the hyperparameters that maximize this objective.

Gaussian Process Kernel

The choice of kernel function is fundamental to the GP's ability to model correlations between data points based on their similarity. In this study, five distinct covariance kernel functions were employed:[1,4]

Tanimoto Kernel Calculated using Equation 7, this kernel is designed for binary feature vectors such as molecular fingerprints. It calculates a ratio between the intersection and union of these feature sets.

$$K(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{7}$$

Here, A and B are binary feature vectors, $|A \cap B|$ denotes the number of common bits set in both vectors, $|A|$ is the number of bits set in A, and $|B|$ is the number of bits set in B.

Linear Kernel Defined by Equation 8, this kernel models linear relationships in the input space.

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2(\mathbf{x} \cdot \mathbf{x}') \tag{8}$$

Here, $\mathbf{x}$ and $\mathbf{x}'$ are input feature vectors, and σ^2 is the variance parameter, which was fixed at 1.0 in this study.

Radial Basis Function (RBF) Kernel Implemented according to Equation 9, the RBF kernel is a stationary kernel that models smooth functions.[4]

$$k(\mathbf{x}, \mathbf{x}') = s \cdot \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2}\right) \tag{9}$$

Here, $\|\mathbf{x}-\mathbf{x}'\|$ denotes the Euclidean distance between the input vectors, s is the outputscale parameter (signal variance), and ℓ is the lengthscale parameter.

Rational Quadratic (RQ) Kernel Calculated using Equation 10, the RQ kernel can be seen as an infinite sum of RBF kernels with different lengthscales, allowing it to model functions with multiple scales of variation.[4]

$$k(\mathbf{x}, \mathbf{x}') = s \cdot \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{10}$$

Here, $\|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance, s is the outputscale, α is the shape parameter that determines the weighting of different lengthscales, and ℓ is the overall lengthscale.

Matérn Kernel (ν=1.5) Defined by Equation 11, this kernel is a generalization of the RBF kernel that allows for control over the smoothness of the resulting function.[5]

$$k(\mathbf{x}, \mathbf{x}') = s \cdot \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{\ell}\right) \tag{11}$$

Here, $\|\mathbf{x} - \mathbf{x}'\|$ is the Euclidean distance, s is the outputscale, ℓ is the lengthscale, ν is the smoothness parameter (fixed at 1.5 in this study), $\Gamma(\cdot)$ is the Gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of the second kind.

For the RBF, Rational Quadratic, and Matérn kernels, all associated hyperparameters (i.e., lengthscale ℓ, shape parameter α, and outputscale s), as well as the model's noise variance σ_n^2, were optimized by maximizing the marginal log-likelihood during model training.[6,7] Models were trained using the Adam optimizer with a learning rate of 0.01 and an exponential decay rate of 0.95 over 100 epochs.

Active Learning Acquisition Strategies

The seven distinct active learning acquisition protocols employed in this study are detailed in table 1. Each protocol began with an initial random batch of 60 compounds, followed by 10 acquisition cycles of 30 compounds each ($b_k = 30$). The exploration-exploitation balance was controlled by varying the α and β parameters in the generalized Upper Confidence Bound (UCB) acquisition function: $s_{\text{acq}}(x) = \alpha \cdot \mu(x) + \beta \cdot \sigma(x)$, where $\mu(x)$ is the predicted mean affinity and $\sigma(x)$ is the predicted standard deviation.

Table 1: Overview of Active Learning Acquisition Protocols and their Parameters

Protocol Name	**Acquisition Protocol Formula (Initial 60 + 10 Cycles of 30)**
Random Baseline	Random(60) + [Random(30)] × 10
UCB-Balanced	Random(60) + [UCB(30)] × 10 *(UCB: $\alpha = 0.5, \beta = 0.5$)*
UCB-Alternate	Random(60) + [Explore(30), Exploit(30)] × 5
UCB-Sandwich	Random(60) + [Explore(30)] × 2 + [Exploit(30)] × 6 + [Explore(30)] × 2
UCB-Explore-heavy	Random(60) + [Explore(30)] × 7 + [Exploit(30)] × 3
UCB-Exploit-heavy	Random(60) + [Exploit(30)] × 7 + [Explore(30)] × 3
UCB-Gradual	Random(60) + [Explore(30)] × 3 + [UCB(30)] × 4 + [Exploit(30)] × 3
Parameter Definitions: *Explore*: $\alpha = 0, \beta = 1$; *Exploit*: $\alpha = 1, \beta = 0$; *UCB*: $\alpha = 0.5, \beta = 0.5$	

Model Validation Analyses

Preprocessing Ablation Study for ChemBERTa To investigate the sensitivity of non-Tanimoto kernels to the scale and dimensionality of high-dimensional ChemBERTa embeddings, we conducted a comprehensive ablation study. The aggregated results, summarized in Table 2, demonstrate that standard preprocessing methods such as scaling and PCA catastrophically degrade model performance for this architecture. This justifies our methodological choice to use the raw, unscaled ChemBERTa embeddings in our main experiments.

Table 2: Aggregated results of the preprocessing ablation study for ChemBERTa embeddings. Performance is averaged across all applicable experimental runs. Metrics shown are the mean final values at cycle 10.

Preprocessing Strategy	**Mean Final $\mathbf{EF_2}$**	**Mean Final BEDROC**	**Mean Final NLPD**
No Preprocessing	**10.07**	**0.213**	**11.27**
StandardScaler	3.56	0.062	12.17
StandardScaler + PCA(50)	4.11	0.063	10.03
StandardScaler + PCA(100)	4.18	0.060	10.00

Uncertainty Calibration Analysis

To address the reviewer's concern regarding uncertainty calibration, we conducted a targeted diagnostic analysis on a representative experimental case (`TYK2`, `LinearKernel`, `ECFP`). Figure 1 compares the calibration profile of the model trained without explicit regularization ("BEFORE") and with weakly informative Gamma priors ("AFTER"). The analysis shows that while the unregularized model was already reasonably well-behaved, the addition of priors resulted in a near-perfectly calibrated reliability diagram, improving the MCE 4.6-fold to 0.05. This provides a sound statistical foundation for our UCB-guided acquisition strategies.

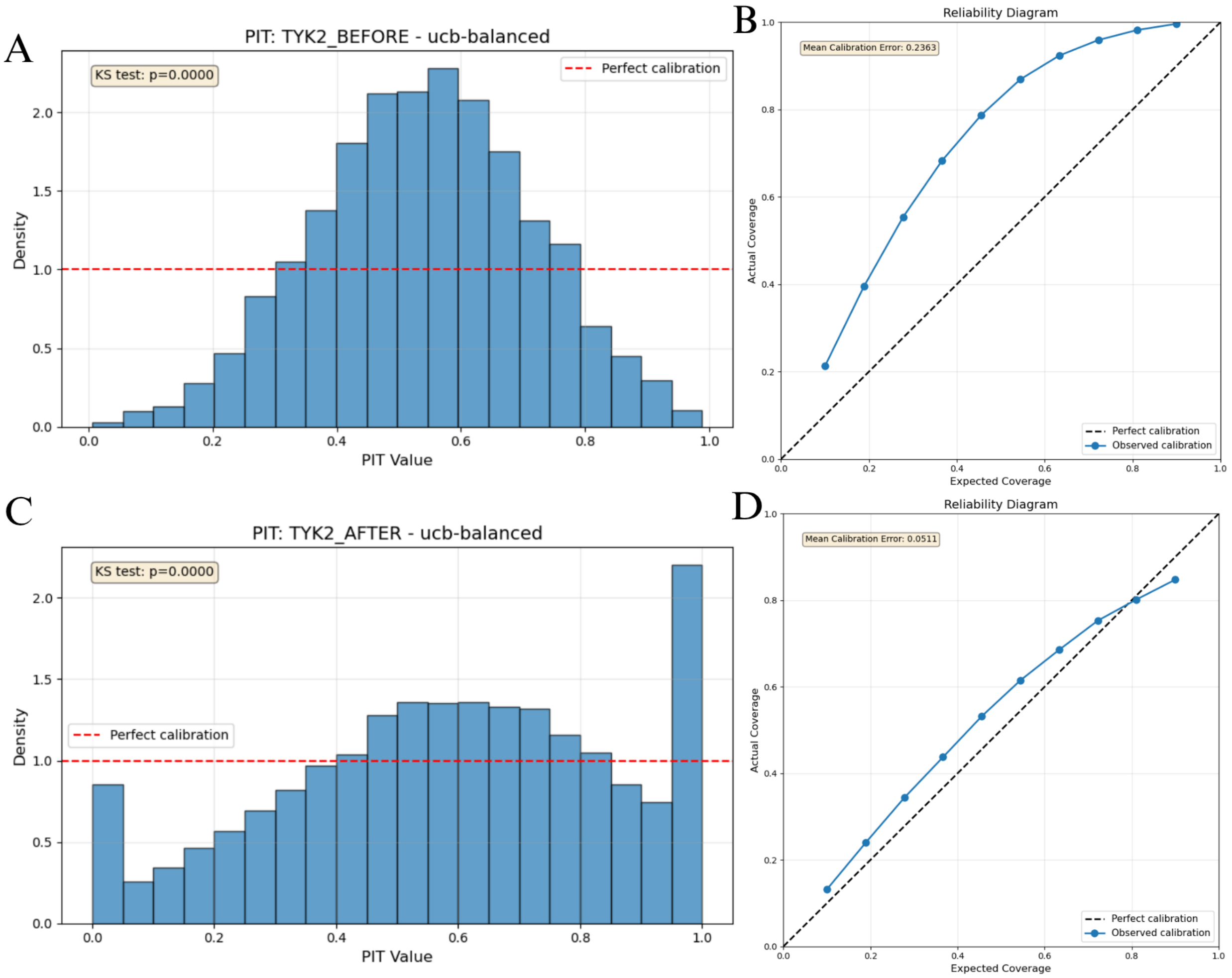


Figure 1: **Validation of GP Model Calibration.** This figure compares the calibration profile of a model trained without regularization (A, B) and with hyperparameter priors (C, D). **(A, B)** The ”BEFORE” model exhibits a slightly underconfident profile (MCE = 0.24). **(C, D)** The ”AFTER” model is excellently calibrated, with a near-perfect reliability diagram (MCE = 0.05).

Area Under the Learning Curve (AULC) Summaries

To provide an integrated measure of each strategy's overall efficiency, we calculated the Area Under the Learning Curve (AULC) based on the Recall@2% performance across all 10 cycles. A higher AULC indicates a more efficient strategy.

Table 3: Mean AULC by Protocol and Dataset.

Protocol	**D2R**	**MPRO**	**TYK2**	**USP7**
random	0.07 ± 0.00	0.07 ± 0.00	0.02 ± 0.00	0.06 ± 0.00
ucb-alternate	0.10 ± 0.07	0.38 ± 0.31	0.07 ± 0.07	0.49 ± 0.30
ucb-balanced	0.09 ± 0.05	0.18 ± 0.16	0.05 ± 0.07	0.19 ± 0.18
ucb-exploit-heavy	0.11 ± 0.07	0.40 ± 0.33	0.06 ± 0.07	0.49 ± 0.30
ucb-explore-heavy	0.10 ± 0.06	0.38 ± 0.30	0.04 ± 0.04	0.48 ± 0.29
ucb-gradual	0.10 ± 0.05	0.40 ± 0.32	0.05 ± 0.06	0.48 ± 0.30
ucb-sandwich	0.11 ± 0.07	0.40 ± 0.33	0.07 ± 0.08	0.49 ± 0.31

Table 4: Mean AULC by Kernel and Dataset.

Kernel	**D2R**	**MPRO**	**TYK2**	**USP7**
LinearKernel	0.16 ± 0.08	0.09 ± 0.05	0.12 ± 0.08	0.23 ± 0.14
MaternKernel	0.07 ± 0.02	0.51 ± 0.28	0.01 ± 0.01	0.56 ± 0.31
RBFKernel	0.07 ± 0.01	0.55 ± 0.26	0.01 ± 0.01	0.60 ± 0.29
RQKernel	0.10 ± 0.02	0.11 ± 0.03	0.06 ± 0.04	0.14 ± 0.06

Statistical Analysis Details

To rigorously quantify the impact of methodological choices on active learning performance, we conducted a multi-factor Analysis of Variance (ANOVA). This section details the statistical formulation and effect size calculations used in the main text.

ANOVA Design

We employed a fixed-effects, four-factor ANOVA model using Type II Sums of Squares (SS). The model predicts the final Recall of Top Compounds (R_k) at Cycle 10 based on the following factors:

- **Dataset** (4 levels: TYK2, USP7, D2R, MPRO)
- **Kernel** (5 levels: Linear, RBF, Matérn, RQ, Tanimoto)
- **Representation** (3 levels: ECFP, MACCS, ChemBERTa)
- **Protocol** (6 levels: All UCB variants)

The ‘Random‘ protocol was excluded from the ANOVA to focus on the variance within active learning strategies.

Omega-Squared (ω^2) Effect Size

While Partial Eta-Squared (η_p^2) is commonly reported, it tends to overestimate effect sizes in small samples and sums to $> 100\%$ across factors. To address the reviewer's comment regarding "unique variance," we calculated **Omega-Squared (ω^2)**, which provides a less biased estimate of the proportion of the total population variance explained by each factor.

The formula used for ω^2 for a specific factor effect is:

$$\omega^2 = \frac{SS_{\text{effect}} - (df_{\text{effect}} \times MS_{\text{error}})}{SS_{\text{total}} + MS_{\text{error}}} \tag{12}$$

where:

- SS_{effect} is the Sum of Squares for the factor.
- df_{effect} is the degrees of freedom for the factor.
- MS_{error} is the Mean Square Error of the residuals.
- SS_{total} is the Total Sum of Squares.

Values of ω^2 typically range from 0 to 1, where 0.01 indicates a small effect, 0.06 a medium effect, and 0.14 a large effect. Negative values (possible due to the unbiased estimator formula) were treated as zero.

Bootstrap Confidence Intervals

To assess the stability of these effect size estimates, we computed 95% Confidence Intervals (CIs) using non-parametric bootstrapping: 1. **Resampling:** The dataset of experimental results ($N = 1,080$) was resampled with replacement 1,000 times. 2. **Recalculation:** For each resampled dataset, the full ANOVA model was refitted, and ω^2 values were recalculated for all factors. 3. **Interval Construction:** The 95% CI was derived from the 2.5th and 97.5th percentiles of the distribution of recalculated ω^2 values.

Detailed Dataset Characteristics

This section provides a detailed characterization of the four protein target datasets used in this study: TYK2, USP7, D2R, and MPRO. We detail their procurement, chemical properties, and diversity profiles, which provide essential context for interpreting the active learning performance results presented in the main manuscript.

Dataset Procurement and Preprocessing

The active learning framework was evaluated using four diverse protein target datasets (Table 5). These targets represent a range of biological functions, including a kinase (TYK2), a deubiquitinase (USP7), a G protein-coupled receptor (D2R), and a viral protease (MPRO).

TYK2 This dataset was derived from Thompson et al.[8] and comprises 9,997 compounds from a congeneric series built around an amino-pyrimidine core scaffold. Affinity values were provided as pKi, obtained from Relative Binding Free Energy (RBFE) calculations. For details on the dataset generation pipeline, including enumeration, filtering, docking, and final selection, please refer to the original publication.[8]

MPRO This dataset was sourced from the COVID Moonshot project[9] and contains 2,062 compounds targeting the SARS-CoV-2 main protease. Affinity is represented by experimental pIC50 values, amalgamated from single enantiomers and racemic mixtures. It exhibits moderate-to-high scaffold diversity.

D2R This dataset is a subset of the ACNet dataset curated from ChEMBL (v28),[10] containing 2,502 compounds targeting the Dopamine D2 Receptor.[11] Affinity is represented by pKi values, with duplicate SMILES entries averaged. It displays high scaffold diversity.

USP7 This dataset was curated by Shen et al.[12] from ChEMBL[10] and includes 1,799 compounds targeting the deubiquitinase USP7. Original experimental affinities (Ki, Kd, IC50) were standardized to pIC50 values, with duplicates aggregated. It features high scaffold diversity.

For all datasets, preprocessing confirmed valid SMILES representations for all compounds. Affinity values were used as provided by the curated sources, with TYK2 and D2R employing pKi, and MPRO and USP7 employing pIC50.

Table 5: Summary of Dataset Properties. This table is also presented in the main manuscript.

Property	**TYK2**	**USP7**	**D2R**	**MPRO**
Binding Assay	pKi	pIC50	pKi	pIC50
Ligands (Total)	9997	1799	2502	2062
Scaffolds (Unique)	104	770	1034	934
Std Dev (Affinity)	1.36	1.31	1.44	0.91
N/M ratio	0.0104	0.428	0.413	0.452

Physicochemical Property Distributions

The datasets exhibit distinct physicochemical property profiles. TYK2, as a congeneric series, has a very narrow distribution of properties, with a mean molecular weight (MW) of approximately 369 Da and a low standard deviation. In contrast, USP7 and D2R contain significantly larger and more diverse molecules, with mean MWs of 578 Da and 446 Da, respectively, and a wide range of values for properties like LogP, TPSA, H-Bond Acceptors/Donors, and Rotatable Bonds. MPRO occupies an intermediate position in terms of property distributions.

Binding Affinity Distributions

The distribution of binding affinity values also varies significantly across datasets (Figure **??**). TYK2 and D2R, which use pKi values, show approximately normal distributions centered around mean affinities of 7.51 and 6.92, respectively. The pIC50 distributions for USP7 and MPRO are more skewed, with MPRO exhibiting a particularly narrow distribution of high-affinity compounds. It is important to note that pKi and pIC50 values are distinct measurement types and are not directly comparable on the same scale.

Active Learning Analysis Platform (Interactive Web Tool)

Given the multi-faceted nature of our results-spanning four datasets, three molecular representations, five kernels, and seven protocols-presenting all findings exhaustively within a static manuscript is impractical. For example, 2 displays the performance of six distinct protocols for the TYK2 dataset alone, and this represents only a fraction of the generated data.

To enhance transparency, reproducibility, and allow for deeper exploration of our findings, we have developed the *Unified Drug Discovery Analysis Platform*, an interactive web tool. This platform provides open access to our complete dataset and analysis workflows, bridging the gap between our high-level performance metrics and the deep, mechanistic insights from our interpretability analyses. The tool is accessible at: https://shapanalysis.streamlit.app/

Key Functionalities

The platform is organized into several modules, each designed to investigate a different aspect of the study. Users can either load the complete demo dataset from our study with a single click or upload their own data (in the specified format) for comparative analysis.

- **Protocol Performance:** This module allows for a detailed, interactive comparison of discovery strategies. Users can generate custom learning curves and final performance bar charts, filtering by dataset, kernel, fingerprint, and protocol to understand context-dependent efficacy.
- **Feature Evolution:** Here, users can visualize how the SHAP importance of key chemical features changes throughout the active learning simulation. This provides insight into the model's learning dynamics and how feature priorities shift as more data is acquired.
- **Chemical Fragments:** This is the core interpretability module. It translates abstract feature importances (e.g., fingerprint bits) from the GP models into tangible, chemically meaningful fragments. For any given feature, the tool displays the corresponding molecular substructure, its prevalence, and highlights it within the parent molecules that exhibit high binding affinity.

- **Molecular Analysis:** This module provides tools for exploring the chemical datasets. Functionalities include generating interactive 2D chemical space projections (PCA, t-SNE, UMAP), visualizing top-ranked compounds, and plotting model-predicted affinity against experimental values to assess model calibration.

- **Advanced Analytics & Publication Figures:** This area contains high-level, cross-dataset analyses and enables the one-click generation of the summary figures presented in our manuscript. This includes analyses of chemical diversity, property distributions, and comprehensive performance heatmaps.

Example Workflow

A user interested in investigating the most effective strategy for the TYK2 target could, for example:

1. Use the **Protocol Performance** tab to identify the best-performing protocol for TYK2 (e.g., UCB-Exploit-heavy).

2. Navigate to the **Chemical Fragments** tab and select the corresponding SHAP analysis file.

3. Execute the fragment analysis to identify the top-ranked chemical features that the model learned were crucial for TYK2 binding.

4. Generate a publication-ready figure summarizing these key fragments using the **Publication Figures** tab.

This workflow allows any researcher to move seamlessly from high-level performance metrics to specific, chemically-interpretable insights, thereby bridging the gap between benchmarking and mechanistic understanding.

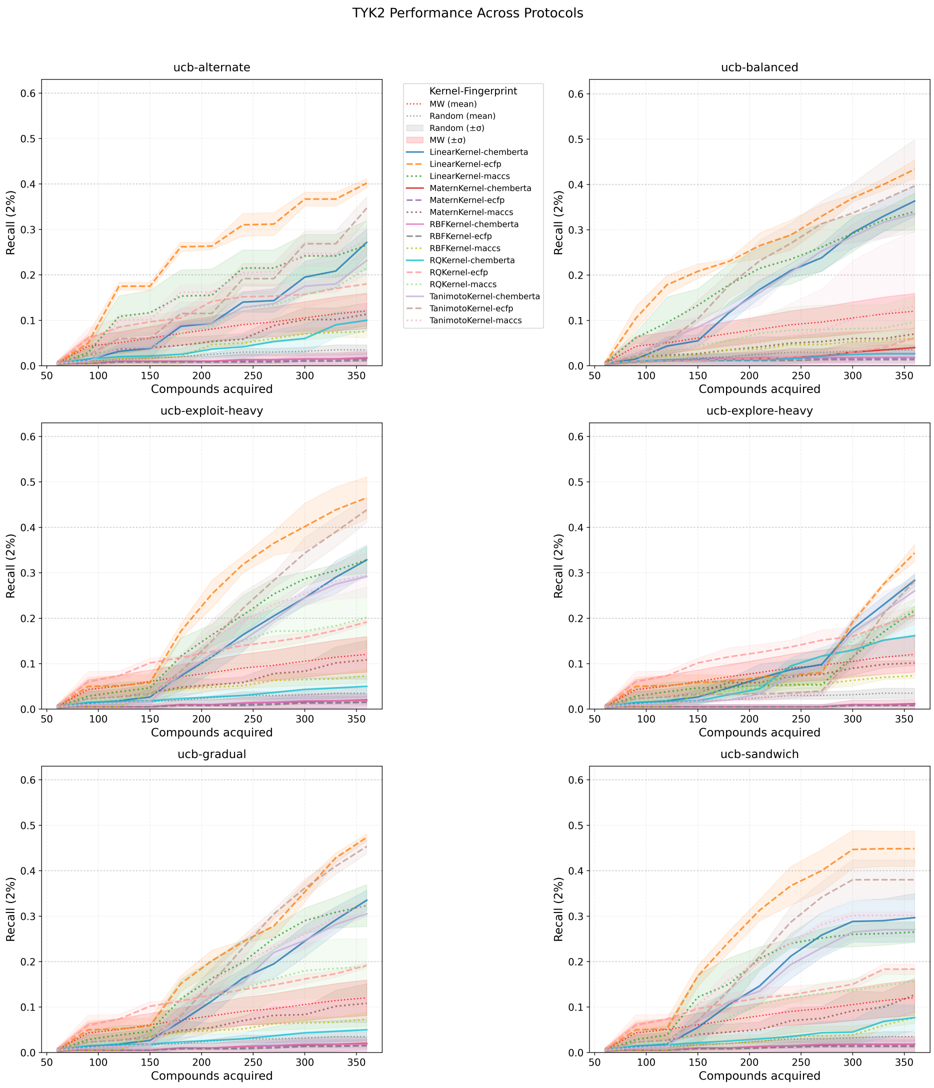
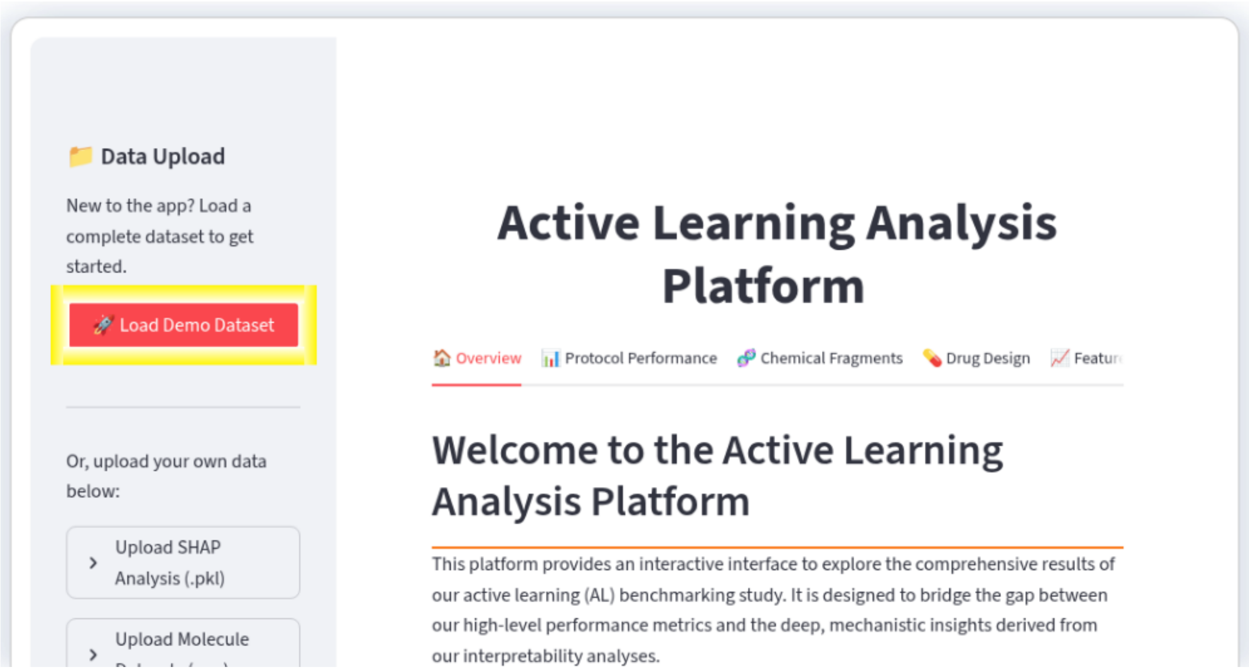


Figure 2: Performance of six active learning protocols on the TYK2 dataset, illustrating the complexity of results for a single target. The web tool allows for dynamic generation and comparison of such plots across all experimental conditions.



New to the app? Load a complete dataset to get started.

Or, upload your own data below:

Active Learning Analysis Platform

Welcome to the Active Learning Analysis Platform

This platform provides an interactive interface to explore the comprehensive results of our active learning (AL) benchmarking study. It is designed to bridge the gap between our high-level performance metrics and the deep, mechanistic insights derived from our interpretability analyses.

Figure 3: The Load Demo Dataset button will populate the entire application with our pre-analyzed dataset, allowing you to interactively explore all our findings immediately

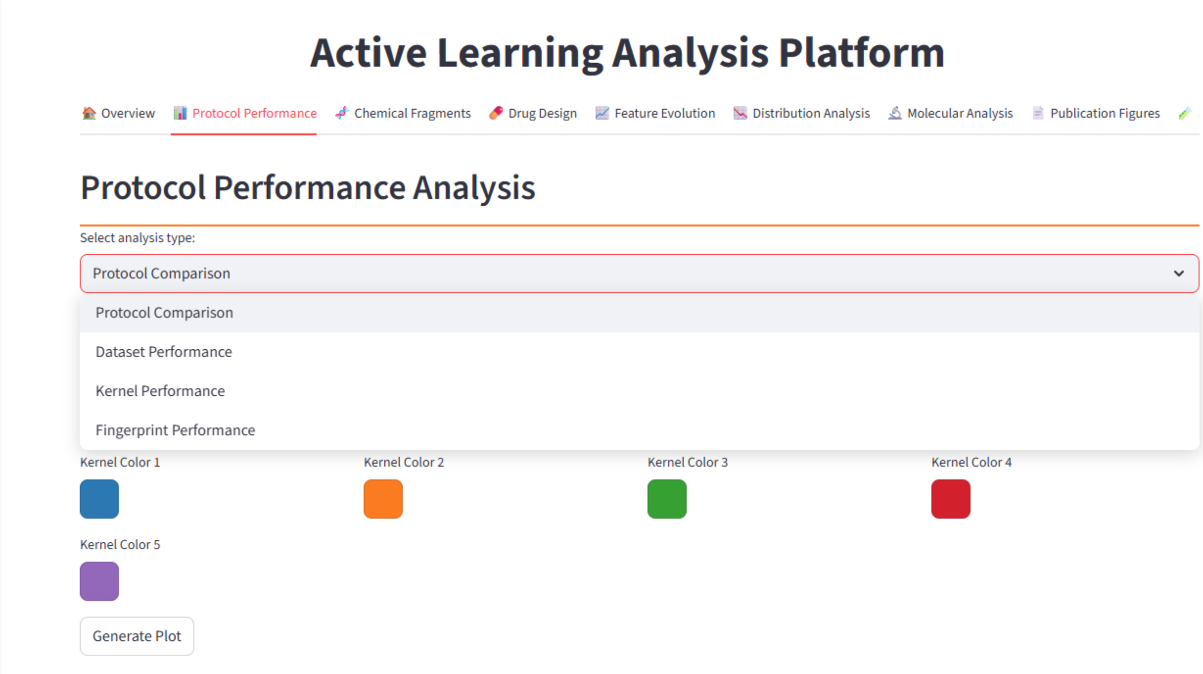
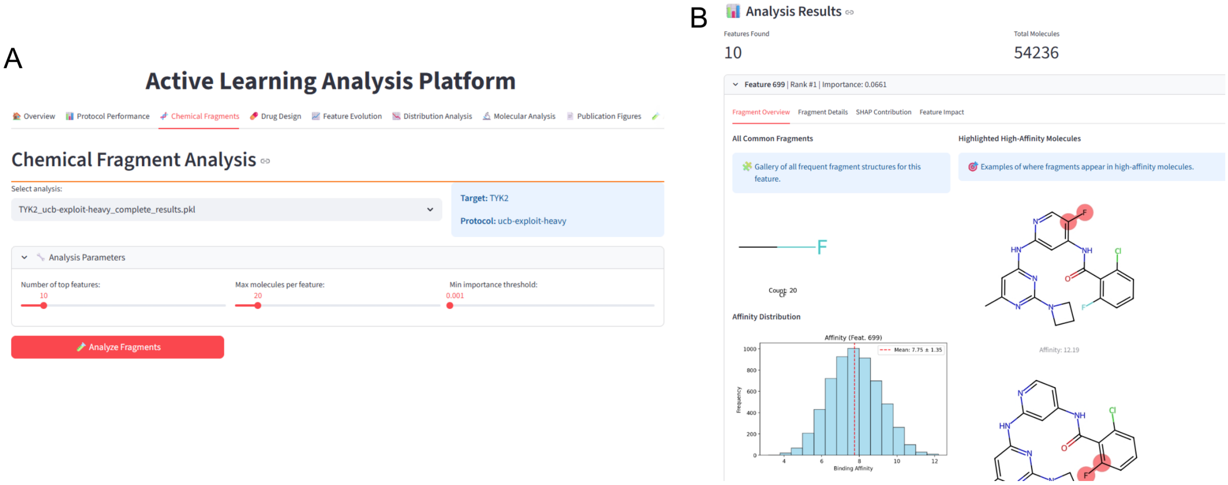

Select analysis type: Protocol Comparison

Figure 4: Highlighted here is the **Protocol Performance** tab, which allows for detailed, interactive comparison of discovery strategies. Within this module, users can select from multiple analysis types via the dropdown menu, including: **Protocol Comparison**, to compare the performance of different acquisition strategies; **Dataset Performance**, to analyze how protocols perform on a specific dataset; **Kernel Performance**, to evaluate different GP kernels; and **Fingerprint Performance**, to assess the impact of molecular representations. This level of granularity empowers researchers to explore our comprehensive dataset and interactively investigate the context-dependent nature of active learning performance.



Active Learning Analysis Platform

Chemical Fragment Analysis

Select analysis: TYK2_ucb-exploit-heavy_complete_results.pkl

Target: TYK2

Protocol: ucb-exploit-heavy

Analysis Parameters

Number of top features: 10

Max molecules per feature: 20

Min importance threshold: 0.001

Analysis Results

Features Found: 10

Total Molecules: 54236

Feature 699 | Rank #1 | Importance: 0.0661

All Common Fragments

Gallery of all frequent fragment structures for this feature.

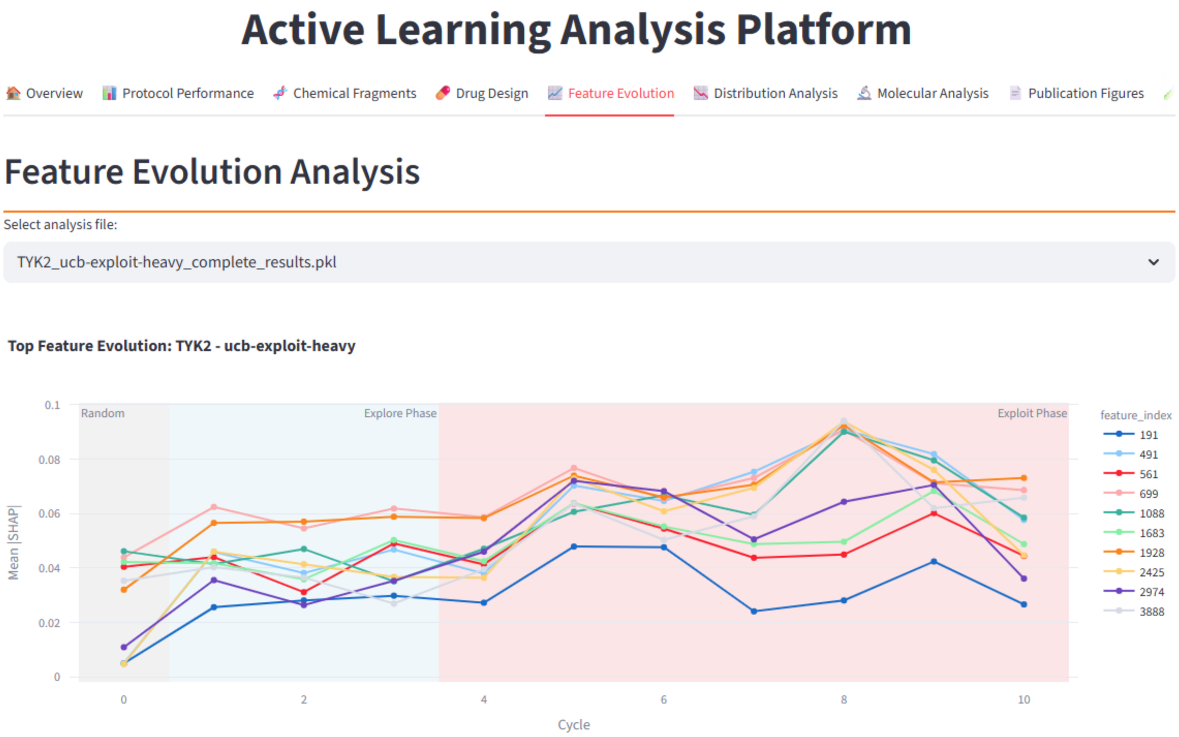
Count: 20
CF

Affinity Distribution

Highlighted High-Affinity Molecules

Examples of where fragments appear in high-affinity molecules.

Figure 5: This figure illustrates the process of translating abstract SHAP feature importance into chemically meaningful insights using our platform. **(A) Analysis Setup:** The user selects a specific experimental condition (e.g., TYK2 with the ucb-exploit-heavy protocol) from the pre-computed SHAP analysis files. They can then adjust analysis parameters, such as the number of top features to investigate, and initiate the analysis. **(B) Detailed Analysis Results:** The platform generates an interactive report for each high-importance feature. Shown here for Feature 699, the ”Fragment Overview” tab displays a gallery of common fragments (in this case, ‘cF‘), an affinity distribution histogram for compounds containing this feature, and examples of high-affinity parent molecules with the fragment highlighted. Further tabs allow for deeper exploration of fragment details and SHAP contributions.



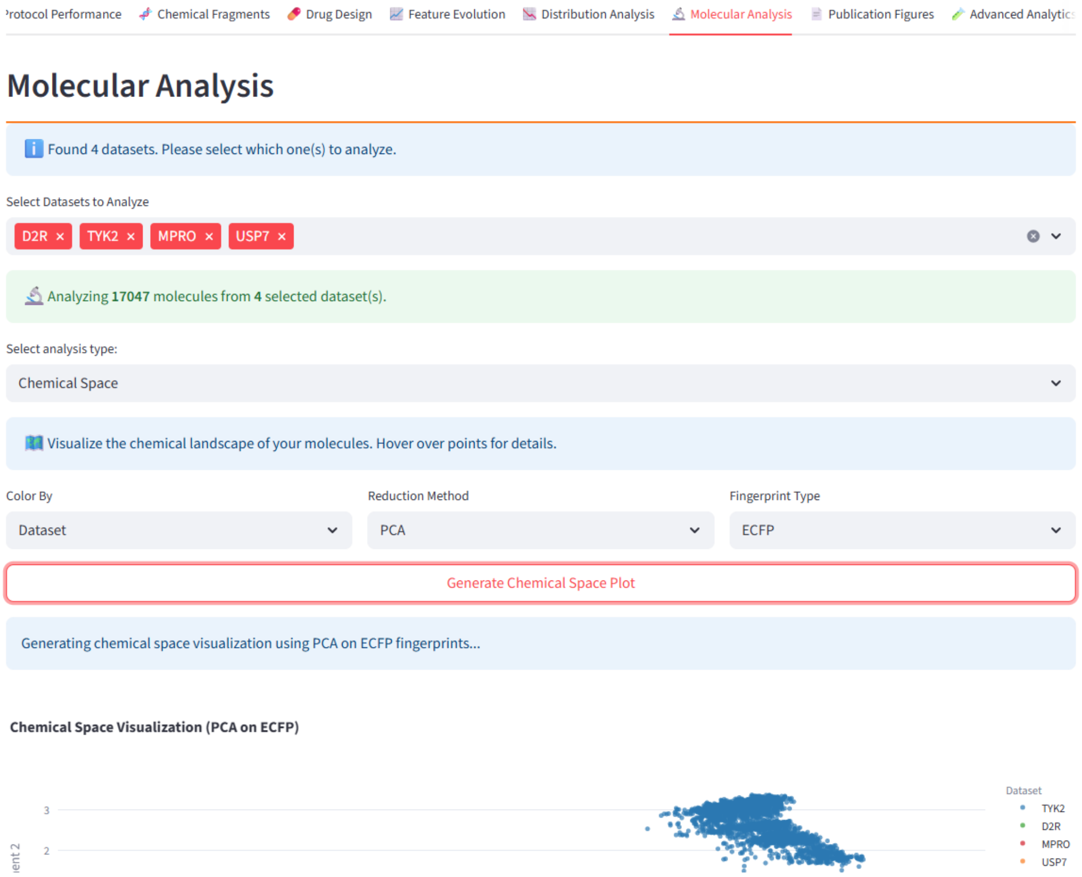
Active Learning Analysis Platform

Feature Evolution Analysis

Select analysis file:

TYK2_ucb-exploit-heavy_complete_results.pkl

Figure 6: **Feature Evolution Analysis Module in the Interactive Web Tool.** This screenshot showcases the ”Feature Evolution” tab of our platform, which allows for the dynamic visualization of how the model’s understanding of structure-activity relationships (SARs) evolves over the active learning process. Users can select a specific experimental condition (e.g., TYK2 with the ucb-exploit-heavy protocol) to generate a plot that tracks the Mean Absolute SHAP Importance of the top-ranked features across all 10 acquisition cycles. The plot includes color-coded backgrounds to indicate the different acquisition phases of the protocol (e.g., Random, Explore Phase, Exploit Phase), providing a clear visual link between the active learning strategy and the resulting feature prioritization dynamics.



Molecular Analysis

Found 4 datasets. Please select which one(s) to analyze.

Select Datasets to Analyze: D2R, TYK2, MPRO, USP7

Analyzing **17047** molecules from **4** selected dataset(s).

Select analysis type: Chemical Space

Visualize the chemical landscape of your molecules. Hover over points for details.

Color By: Dataset

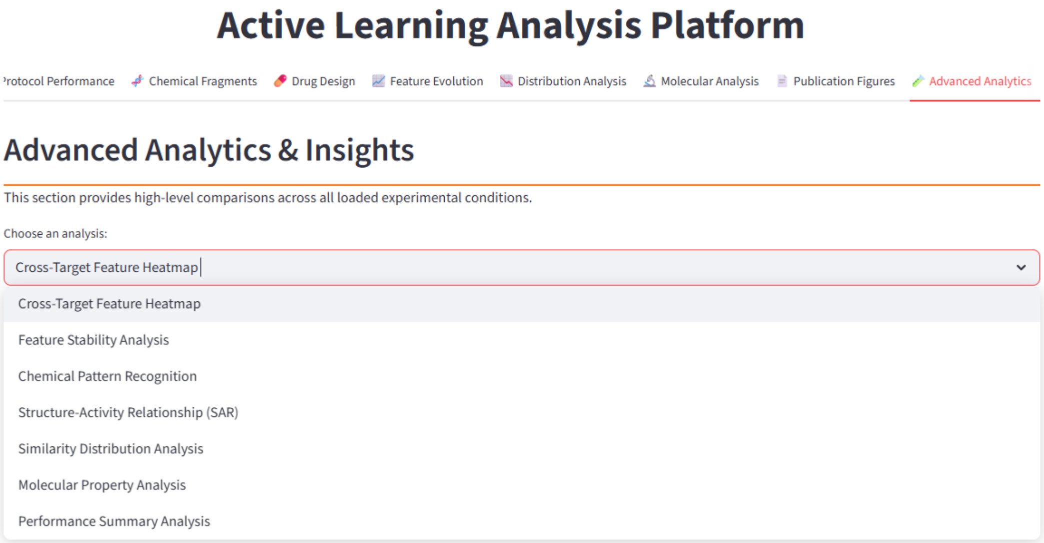
Reduction Method: PCA

Fingerprint Type: ECFP

Generating chemical space visualization using PCA on ECFP fingerprints...

Chemical Space Visualization (PCA on ECFP)

Figure 7: **Molecular Analysis Module in the Interactive Web Tool.** This screenshot showcases the setup for the ”Molecular Analysis” tab, which provides tools for exploring the chemical datasets. Users can select one or more datasets for analysis and then choose from various functionalities, such as generating interactive 2D chemical space projections (PCA, t-SNE, UMAP). The tool allows for customization of the visualization, including the choice of molecular representation (e.g., ECFP, MACCS), the dimensionality reduction method, and the property used for coloring the data points (e.g., Dataset, affinity), enabling a flexible and in-depth exploration of the chemical landscape.



Active Learning Analysis Platform

rotocol Performance
Chemical Fragments
Drug Design
Feature Evolution
Distribution Analysis
Molecular Analysis
Publication Figures
Advanced Analytics

Advanced Analytics & Insights

This section provides high-level comparisons across all loaded experimental conditions.

Choose an analysis: Cross-Target Feature Heatmap

Cross-Target Feature Heatmap
Feature Stability Analysis
Chemical Pattern Recognition
Structure-Activity Relationship (SAR)
Similarity Distribution Analysis
Molecular Property Analysis
Performance Summary Analysis

Figure 8: **Advanced Analytics Module in the Interactive Web Tool.** This screenshot showcases the ”Advanced Analytics & Insights” tab, which provides high-level comparative analyses across all loaded experimental conditions. Users can select from a dropdown menu of advanced analytical tools, including: **Cross-Target Feature Heatmap**, to compare SHAP feature importances across different targets and protocols; **Feature Stability Analysis**, to assess the consistency of feature importance over AL cycles; **Chemical Pattern Recognition**, to identify common scaffolds across datasets; as well as broader analyses of SAR, similarity distributions, and performance summaries. This module enables a comprehensive, high-level exploration of the data beyond individual experimental conditions.

References

(1) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press, 2006.

(2) Schulz, E.; Speekenbrink, M.; Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **2018**, *85*, 1–16.

(3) Krause, A.; Hübotter, J. *Probabilistic Artificial Intelligence*; arXiv preprint arXiv:2502.05244, 2025.

(4) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press, 2002.

(5) Matérn, B. Spatial Variation. *Meddelanden från Statens Skogsforskningsinstitut* **1960**, *49*.

(6) MacKay, D. J. C. In *Maximum Entropy and Bayesian Methods: Seattle, 1991*; Smith, C. R., Erickson, G. J., Neudorfer, P. O., Eds.; Springer Netherlands: Dordrecht, 1992; pp 39–66.

(7) Sundararajan, S.; Keerthi, S. S. Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. *Neural Computation* **2001**, *13*, 1103–1118.

(8) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *Artif. Intell. Life Sci.* **2022**, *2*, 100050.

(9) Achdout, H.; Aimon, A.; Bar-David, E.; Morris, G. M. COVID moonshot: open science discovery of SARS-CoV-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv* **2020**,

(10) Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **2018**, *47*, D930–D940.

(11) Zhang, Z.; Zhao, B.; Xie, A.; Bian, Y.; Zhou, S. Activity Cliff Prediction: Dataset and Benchmark. 2023; arXiv:2302.07541 (accessed Sep 10, 2023).

(12) Shen, W.-f.; Tang, H.-w.; Li, J.-b.; Li, X.; Chen, S. Multimodal data fusion for supervised learning-based identification of USP7 inhibitors: a systematic comparison. *J. Cheminform.* **2023**, *15*, 1–16.