

SCOPE: From Shape and Color to Binding Affinity, with Confidence

Jingyi Chen

jlchen@cadence.com

OpenEye, Cadence Molecular Sciences <https://orcid.org/0000-0003-4076-936X>

Chris Neale

OpenEye, Cadence Molecular Sciences <https://orcid.org/0000-0001-9912-078X>

Mireille Krier

OpenEye, Cadence Molecular Sciences <https://orcid.org/0000-0003-4912-2925>

Someina Khor

Isomorphic Labs <https://orcid.org/0000-0001-8578-1835>

Anthony R. Bradley

University of Liverpool <https://orcid.org/0000-0002-0881-3490>

A. Geoffrey Skillman

OpenEye, Cadence Molecular Sciences

Shyamal K. Nath

nath@cadence.com

OpenEye, Cadence Molecular Sciences

Research Article

Keywords:

Posted Date: September 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11130410/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: J.L.C., C.N., M.K., A.G.S., and S.K.N. are employees of OpenEye, Cadence Molecular Sciences and may hold equity or stock in the company. S.K. is an employee of Isomorphic Labs. A.R.B. is affiliated with the University of

Liverpool and Dalton Tx Ltd. These affiliations may constitute potential conflicts of interest. The authors declare no other competing interests.

SCOPE: From Shape and Color to Binding Affinity, with Confidence

Jingyi L. Chen,^[a] Chris Neale,^[a] Mireille Krier,^[a] Someina Khor,^[b] Anthony R. Bradley,^[c,d] A. Geoffrey Skillman,^[a] Shyamal K. Nath^{*[a]}

Three-dimensional quantitative structure–activity relationship (3D-QSAR) is widely used in lead optimization in drug discovery, yet prediction reliability remains underexplored despite the continued expansion of chemical space. Here, we present SCOPE (Shape and Color-based Prediction Ensemble), a confidence-aware ensemble model for binding affinity prediction that incorporates a target-agnostic applicability domain mapping molecular similarity to prediction confidence. SCOPE integrates 3D molecular shape and color representations from ROCS (Rapid Overlay of Chemical Structures) with complementary kernel-based methods, namely kernel partial least squares (kPLS) and Gaussian process regression (GPR). The ensemble combines ROCS-kPLS and ROCS-GPR with a 2D-GPR baseline to produce confidence-weighted predictions. The ROCS-kPLS component enables intuitive interpretation through 3D contour maps highlighting favorable and unfavorable chemical features. Across curated structure–affinity datasets spanning multiple protein targets, SCOPE consistently outperforms the 2D baseline under both random and cluster-based cross-validation. Benchmark comparisons further show that SCOPE achieves strong and competitive performance relative to the best literature models, including classical 3D-QSAR and modern machine learning methods.

Introduction

Three-dimensional quantitative structure–activity relationship (3D-QSAR) is a computational modeling approach that relates biological activity to the three-dimensional properties and spatial arrangements of molecules^[1]. It extends classical QSAR approaches by explicitly incorporating spatial information for activity prediction. The field was pioneered by Comparative Molecular Field Analysis (CoMFA)^[2], which models binding affinity by sampling molecular interaction fields on a 3D grid using various probes and analyzing them with partial least squares (PLS). PLS yields interpretable contour maps that highlight favorable and unfavorable regions for activity. This interpretability makes it particularly well suited for lead optimization (LO), where compounds are typically close structural analogs and design decisions rely on understanding how incremental structural changes

affect activity. To mitigate CoMFA's sensitivity to molecular alignment, Comparative Molecular Similarity Indices Analysis (CoMSIA) was subsequently developed^[3]. CoMSIA employs smoother Gaussian similarity fields and extends the descriptor space to include hydrophobicity and hydrogen bonding interactions, while preserving model interpretability. Together, these methods form the foundation of classical field-based 3D-QSAR methods, and remain widely used in practice, as evidenced by their continued adoption in academic studies^[4,5] as well as commercial^[6,7] and open-source computational platforms^[8–10].

Incorporation of 3D features necessitates the identification of bioactive conformations when constructing predictive models. Fortunately, advances in experimental structural biology have enabled structure-based drug design to become the dominant paradigm in modern drug discovery^[11]. In this context, available co-crystal ligand structures are preferred for model training and prediction. In their absence, prediction tools such as POSIT provide a robust alternative. Specifically, POSIT leverages available target structures and/or co-crystallized ligands to guide pose optimization^[12] and associates a confidence score with each generated pose, providing a natural foundation for confidence-aware predictions employed by SCOPE.

Recent developments in 3D activity prediction have broadly followed three directions. The first builds upon the classical field-based 3D-QSAR paradigm established by CoMFA, primarily through the development of improved field descriptors while preserving the underlying regression framework. Examples include the incorporation of descriptors derived from quantum mechanics (QM)^[13,14]. The second is exemplified by QMOD and QuanSA^[15–17], which employ field-based model induction to jointly optimize ligand alignment and affinity models. The third focuses on the adoption of more advanced machine learning (ML) algorithms, such as random forests (RF), support vector machines (SVM), and modern neural networks (NN)-based models, to better capture non-linear SAR, and in some cases to enable the construction of more robust consensus models^[18–20]. Within this direction, many approaches use alternative 3D descriptors such as protein–ligand interaction features^[19] and molecular properties derived from bioactive conformations^[20]. More recent NN-based models further replace engineered descriptors with learned representations derived directly from protein–ligand structures^[21].

While these approaches improve predictive performance on some of the benchmark datasets, they introduce important trade-offs. QM-derived descriptors are more computationally expensive, and many ML models provide limited model interpretability. NN-based methods, in particular, often benefit from larger datasets than are typically available in LO settings. Importantly, most of these developments do not directly address the challenge of quantifying prediction reli-

[a] OpenEye, Cadence Molecular Sciences, Santa Fe, NM, USA
E-mail: nath@cadence.com

[b] Isomorphic Labs, London, UK

[c] Department of Chemistry, University of Liverpool, Liverpool, UK

[d] Dalton Tx Ltd, London, UK

ability. QMOD and QuanSA represent notable exceptions in this respect, incorporating multiple measures to assess prediction reliability; however, these are not explicitly calibrated against prediction error. More broadly, although the importance of applicability domain (AD) to address prediction reliability is well established in QSAR best practices^[22] and regulatory guidelines^[23], such considerations are rarely incorporated explicitly into 3D activity prediction models. This is partly rooted in their historical use in LO within congeneric series, where domain validity is often implicitly assumed and emphasis is placed on predictive performance and interpretability. However, modern drug discovery increasingly explores broader chemical spaces, and multiple distinct series may be considered in LO^[19], making explicit reliability assessment essential. Moreover, integrating reliability assessment may also broaden the applicability of 3D activity prediction models to earlier discovery stages involving more diverse chemical spaces.

Here, we introduce SCOPE (**S**hape and **C**olor-based **P**rediction **E**nsemble) for predicting the binding affinities of small molecules against protein targets. The key contributions of this work are twofold. First, we extend ROCS^[24] beyond its traditional role in virtual screening. ROCS-based molecular overlap and similarity are integrated with kernel-based methods to enable affinity prediction. Previous studies have incorporated ROCS-derived features into ML models, primarily for classification and virtual screening tasks^[25,26], whereas their use in supervised regression frameworks for affinity prediction remains limited. Because the training of kernel-based methods scales at least quadratically with dataset size, this approach is naturally suited to small- to medium-sized datasets typical of LO. In contrast to CoMFA and CoMSIA, SCOPE avoids explicit grid construction for descriptor generation, reducing performance sensitivity to grid parameters while naturally accommodating nonlinear SAR.

Second, we introduce a target-agnostic AD model that maps molecular similarity to prediction confidence, providing a quantitative measure of prediction reliability. This AD framework enables uncertainty-aware prioritization of compounds and supports more informed decision making in practical workflows. Within the Design–Make–Test–Analyze (DMTA) paradigm, compounds with both high predicted affinity and confidence can be prioritized, while low-confidence predictions identify underrepresented regions of chemical space for targeted exploration to improve model performance. In contrast to existing AD strategies^[27], which typically rely on similarity-based criteria and are implemented as binary decision rules, SCOPE defines domain criteria directly in terms of prediction confidence. This eliminates the need to select similarity metrics and shifts AD definition from similarity thresholds to confidence thresholds. An example of how AD-derived confidence, together with POSIT pose confidence can facilitate decision making in practical LO workflows is given in Section 1.7.

As illustrated in Figure 1, SCOPE is implemented as an ensemble comprising ROCS-kPLS, ROCS-GPR, and 2D-GPR. The ensemble prediction is obtained as confidence-weighted average of individual model predictions, while the ensemble confidence is taken as the median of their confidences. This framework combines two complementary kernel-based methods, namely kernel partial least squares (kPLS) and Gaussian process regression (GPR), reflecting complementary deterministic and probabilistic modeling approaches.

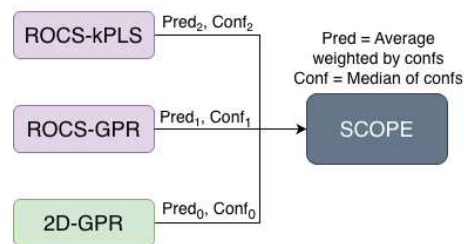


Figure 1. SCOPE as an ensemble of ROCS-kPLS, ROCS-GPR, and baseline 2D-GPR models. Conf denotes prediction confidence.

Pairwise molecular shape and color overlaps are directly used as kernels in kPLS, preserving 3D interpretability and enabling 3D contour map visualization to guide chemical design. For the GPR models, we adopt a similarity-embedding strategy in which molecules are represented by their Tanimoto similarity vectors to the training set followed by a linear kernel, where nonlinearity is introduced through the similarity-based representation.

The validation of SCOPE consists of two complementary components. First, the model design is evaluated with an emphasis on statistical significance using curated structure–affinity datasets, assessing the benefits of ensembling and incorporation of 3D information for affinity prediction. Second, SCOPE is benchmarked against third-party 3D-QSAR and ML methods on selected literature datasets. In the Results section, we first discuss two complementary cross-validation (CV) protocols used to construct the target-agnostic AD. Then the model design is validated through Friedman tests with post hoc Conover–Holm pairwise comparisons among the 2D-GPR, ROCS-GPR, ROCS-kPLS, and ensemble models. We then present external benchmarking against third-party methods, followed by comparison with a physics-based approach. Finally, we illustrate model interpretability for ROCS-kPLS and ROCS-GPR and conclude with a representative practical LO workflow with SCOPE.

Methods

0.1. ML models

0.1.1. ROCS-kPLS

kPLS extends PLS through the "kernel trick", which implicitly maps the input data into a high-dimensional feature space where nonlinear relationships can be modeled using linear latent variables. Assuming a training set of size n and an explicit feature vector space of dimension m , kPLS replaces matrix operations of $n \times m$ in PLS with more compact ones involving $n \times n$ kernel matrices.

In this work, we employ a composite molecular overlap kernel built on the ROCS framework. Each heavy atom is represented by an atom-centered Gaussian function, and molecular shape is defined as their sum, yielding a smooth scalar field that approximates the overall molecular shape. In parallel, chemical "color" features are encoded using additional Gaussian functions centered on color atoms, assigned according to the Implicit Mills–Dean color force field^[28]. The kernel is expressed as

$$K(\text{mol}_i, \text{mol}_j) = K_{\text{shape}}(\text{mol}_i, \text{mol}_j) + \alpha K_{\text{color}}(\text{mol}_i, \text{mol}_j) \quad (1)$$

where the weighting factor α modulates the relative contribution of shape and chemical color features, and is computed as the ratio between the maximum self shape overlap and the maximum self color score across all molecules in the training set.

The primary hyperparameter in kPLS is the number of latent variables d , which balances model complexity and generalization: too many components lead to overfitting, whereas too few cause underfitting. The optimal value is determined by CV on the training set^[29]. Specifically, values ranging from 2 to 15 are evaluated, and the final choice maximizes CV performance, assessed here using Pearson's correlation coefficient squared (r^2). An upper bound of 15 components was found sufficient based on prior experience.

Due to the additive nature of kPLS, the predicted affinity can be decomposed into atomic contributions, enabling grid-based 3D interpretation via contour maps. A uniformly spaced grid is defined over the ligand domain, and probe atom contributions are evaluated at each point. Here, a carbon probe is used to capture shape contributions, while six color probes, namely hydrogen bond (H-bond) acceptor, H-bond donor, hydrophobe, rings, cation, and anion, defined by the Implicit Mills–Dean color force field, are used to sample chemical features. The resulting contour maps highlight spatial “hotspots” where molecular shape or chemical features correlate positively or negatively with activity, providing a structural guide for LO.

All kPLS models were trained using an internal implementation^[30] that sequentially extracts supervised latent components from the kernel matrix via deflation^[31].

0.1.2. 2D-GPR and ROCS-GPR

In contrast to the deterministic latent-variable formulation of kPLS, GPR places a Bayesian prior over functions and defines the model through a kernel-induced covariance structure, yielding both predictions and uncertainty estimates that can be interpreted as an alternative approach for defining the AD. However, these uncertainties may be miscalibrated in practice and do not necessarily provide a reliable quantitative measure of prediction error without additional calibration^[32,33], as discussed in Section 1.2.

Molecular fingerprints are used for descriptor generation in the 2D-GPR model. However, 2D fingerprints do not admit a physically meaningful notion of overlap and are typically compared using the Tanimoto coefficient. While the Tanimoto coefficient defines a valid kernel for binary fingerprints^[34], this validity does not generally hold for continuous 3D overlap measures. To enable a unified and well-defined GPR framework for both 2D and 3D representations, a similarity-embedding strategy was adopted in which each molecule is represented by an explicit feature vector of similarities to the training set. In the 2D-GPR model, this consists of n Tanimoto similarities computed from 2D circular fingerprints using the OpenEye GraphSim Toolkit^[35]. In the ROCS-GPR model, an additional set of n 3D Tanimoto similarities based on shape and color is concatenated, yielding a $2n$ -dimensional feature vector. GPR is then applied using a dot-product (linear) kernel with an additive noise term, equivalent to Bayesian linear regression in similarity space. The kernel takes the form

$$K(\text{mol}_i, \text{mol}_j) = \sigma_0^2 + \phi(\text{mol}_i)^\top \phi(\text{mol}_j) + \sigma_n^2 \delta_{ij}, \quad (2)$$

where the first two terms represent the dot-product kernel with bias, and the final term represents noise. $\phi(\text{mol})$ represents the feature vectors in similarity space. For a test molecule, the posterior mean prediction for ROCS-GPR can be expressed in linear form as

$$\hat{y}_{\text{ROCS-GPR}}(\text{mol}) = \hat{b} + \hat{w}_{2D}^\top \phi_{2D}(\text{mol}) + \hat{w}_{3D}^\top \phi_{3D}(\text{mol}). \quad (3)$$

where $\hat{b}$ is a global bias and $\hat{w}$ denotes the effective regression weights. Equation 3 enables interpretation of the ROCS-GPR model at the level of 2D and 3D feature contributions.

In this work, we used the scikit-learn implementation of GPR with default hyperparameters for both the model and kernel components^[36].

0.2. Curation of structure–affinity datasets

To evaluate the benefits of ensembling and incorporating 3D information, structure–affinity datasets were curated. As with any ML approach, predictive performance depends strongly on data quality. To provide the 3D models with a well-controlled evaluation setting, datasets were curated from the RCSB Protein Data Bank (PDB)^[37] spanning a diverse set of human protein targets. Each dataset consists of ligands with experimentally resolved poses and affinity data for a single target, aligned into a high-quality reference structure via protein sequence-based alignment.

Given the inherent noise in heterogeneous public data, several precautions were taken from both the structure and affinity perspectives. For example, target-specific criteria such as protein conformational state (e.g., kinase structures restricted to DFG-in) was applied to ensure binding environment comparability. While single-lab noise for experimental affinity data is estimated at approximately 0.3 log units^[38], reported variability increases to approximately 0.5–0.7 from heterogeneous sources^[39,40]. To mitigate this issue, the median value was used when multiple measurements were available. Full details of the curation workflow are provided in Section 1.1 of the Supporting Information (SI).

Table 1 summarizes the seventeen curated datasets. The first fourteen were used for AD construction and model design validation, while the remaining three served as external validation sets for the AD. All datasets exhibit broad potency ranges, providing sufficient signal-to-noise ratio for learning meaningful SAR^[41].

Before undertaking any ML studies, exploratory data analysis was performed to characterize data distributions. Figure S2 (a–c) in the SI summarizes potency as well as pairwise 2D and 3D Tanimoto similarity distributions, along with corresponding potency differences to relate structural and activity similarity. The 3D similarity distributions exhibit a broader spread, whereas 2D similarities are strongly skewed toward low values (< 0.2), indicating limited substructural overlap in circular fingerprint space. In contrast to 2D representations, compounds dissimilar in fingerprint space can exhibit higher 3D similarity, and a clearer relationship between structural and activity similarity is also observed using 3D measures, suggesting improved discriminative power.

Although these datasets deviate from typical LO settings, their diversity supports more robust SAR learning and broader AD coverage. While 3D activity models have

Table 1. Details of seventeen curated structure–affinity datasets ordered by descending size.

UniProt	Protein	Size	Activity type	Activity range	Reference PDB
P56817	BACE1	136	pIC ₅₀	7.3	5HU1
O60885	BRD4	128	pIC ₅₀	4.6	4Z1Q
P24941	CDK2	123	pIC ₅₀	6.1	6GUK
P00918	CAII	109	pK _i	9.5	6RH4
P00742	Factor Xa	79	pK _i	5.0	2Y5H
O14757	CHK1	77	pIC ₅₀	10.1	4HYI
Q16539	p38 α	73	pIC ₅₀	6.4	2RG6
Q9H2K2	TNKS2	69	pIC ₅₀	5.4	5C5R
P07900	HSP90 α	61	pIC ₅₀	5.4	4NH8
Q13627	DYRK1A	58	pIC ₅₀	3.3	6S14
P11309	PIM1	57	pIC ₅₀	5.9	5DWR
O60674	JAK2	56	pIC ₅₀	5.8	5CF8
P00797	Renin	53	pIC ₅₀	5.4	3Q5H
P43405	Syk	53	pIC ₅₀	4.4	5CXH
Q9NWZ3	IRAK4	49	pIC ₅₀	6.8	8TX0
P34913	sEH	48	pIC ₅₀	6.4	5ALY
P27487	DPP-IV	46	pIC ₅₀	5.5	3C45

traditionally been applied to congeneric series in LO, it is not inherently restricted to such settings, and modern campaigns increasingly explore multiple distinct chemical series^[19]. In contrast, the literature datasets in Section 1.4 reflect more project-specific LO settings.

0.3. Validation of model design

The first fourteen structure–affinity datasets were used to validate the proposed model design through systematic statistical testing, assessing whether one model consistently outperforms another across multiple performance metrics under different CV schemes. Figure S3 (SI) illustrates the procedure.

Starting from fourteen datasets and four ML models, two complementary CV schemes were used: 5×2 CV based on random splits^[42] and leave-one-cluster-out (LOCO) CV. The 5×2 CV scheme, recommended for smaller datasets^[43], involves five random half-splits with alternating training and test sets, yielding ten splits per dataset. For LOCO CV, each dataset was clustered into five clusters using K-medoids clustering where pairwise distance is defined as 1.0–2D Tanimoto (cluster distributions in Table S3). Each cluster was held out once, producing five splits per dataset. K-medoids was chosen for its more balanced cluster sizes and fewer small clusters compared to alternatives such as hierarchical clustering. Clustering was implemented using the scikit-learn-extra package (method="pam", init="k-medoids++")^[44].

Performance metrics were aggregated across CV splits prior to statistical testing. Under 5×2 CV, mean values were used since variability was small (Figure S4). In contrast, LOCO CV showed higher variability making split-level estimates unreliable; therefore, pooled metrics were computed by concatenating predictions across splits. Aggregated values (Pearson's r^2) under both CV schemes are reported in Table S2 (SI).

Statistical testing followed Rainio *et al.*^[45] using the Friedman test^[46] to detect global differences among models. This non-parametric test was selected primarily because ML met-

rics are often bounded and skewed. When significant ($p < 0.05$), Conover post hoc tests^[47] with Holm correction^[48] were applied for pairwise comparisons. Implementation used scipy.stats^[49] and scikit_posthocs^[50].

This procedure was applied across three performance metrics: (1) Pearson's r^2 , which quantifies linear correlation; (2) Kendall's τ , which measures rank-order agreement; and (3) median absolute error (MedAE), which reflects the typical magnitude of prediction error in the same unit as modeled endpoint. In practical LO settings, correlation and ranking are often more directly relevant for compound prioritization and decision-making, with absolute error serving as a secondary metric.

0.4. Construction and validation of target-agnostic AD

A target-agnostic AD was constructed separately for each individual model. The training data of the AD was derived by pooling the validation sets from 5×2 and LOCO CV across the first fourteen structure–affinity datasets, yielding 6762 data points, with approximately 80% from 5×2 CV.

This AD relates a molecule's maximum similarity to the corresponding training set, normalized to the [0, 1] interval, to an associated prediction confidence, defined as the empirical fraction of accurate predictions. Prediction confidence can thus be interpreted as the empirical probability that a model prediction is accurate at a given level of similarity to the training data.

Specifically, by aggregating predictions across all held-out test sets from CV, the maximum similarity of a query molecule x to the corresponding training set $s_{\max}(x)$ was computed for each model. Similarity metrics were chosen to be consistent with each model's descriptor representation: 2D Tanimoto similarity for the 2D-GPR model, ROCS Tanimoto similarity for the ROCS-kPLS model, and the sum of 2D and ROCS Tanimoto similarities for the ROCS-GPR model.

The resulting similarity values were normalized on a

per-model basis and subsequently binned into intervals of width 0.1. Within each similarity bin, the fraction of accurate predictions was computed to estimate empirical prediction confidence as a function of similarity, effectively approximating the conditional probability $P(|\varepsilon(x)| \leq \delta \mid s_{\max}(x) = s)$, where $\varepsilon(x)$ denotes the prediction error and δ is the chosen accuracy tolerance. In this work, δ was set at 0.7, such that a prediction is considered accurate when the absolute deviation is within 0.7 log units, which roughly corresponds to the upper range of experimental noise observed in heterogeneous public datasets (see Section 0.2).

To obtain a smooth prediction confidence calibration curve, bins containing fewer than 30 molecules (less than 0.5% of total data points) were excluded to remove sparse regions for each model. Remaining accuracy values were smoothed using weighted isotonic regression, with bin counts as weights. Boundary points at (0.0, 0.0) and (1.0, 1.0) were added to ensure full coverage. Prediction confidence for a query molecule is then obtained by interpolating this similarity–accuracy curve based on its maximum similarity to the training set, representing the empirical probability of achieving the specified accuracy. For the ensemble, predictions are combined as a confidence-weighted average, with ensemble confidence defined as the median of individual model confidences.

In order to validate the target-agnostic AD, including the ensemble, the last three structure–affinity datasets, corresponding to distinct protein targets were reserved as external validation sets. Because prediction confidence is defined probabilistically, the same 5×2 and LOCO CV protocols were applied to generate resampled external predictions, yielding 858 data points.

Results and Discussion

1.1. Two complementary cross-validation protocols

Although temporal splits provide the most realistic estimate of prospective performance in LO, such information is typically unavailable. In their absence, random splits (5×2 CV) provide an optimistic estimate due to chemotype mixing, while cluster-based splits (LOCO CV) provide a pessimistic estimate reflecting generalization to unseen chemical series. Together, these protocols estimate practical upper and lower bounds on expected performance.

As K-medoids clustering does not explicitly control inter-cluster separation, we quantified chemical separation by comparing nearest-neighbor test–train similarity distributions under both clustering and random splits. Under cluster-based splits, median nearest-neighbor similarities were consistently low (0.12–0.18) with narrow interquartile ranges and upper-quartile values typically below 0.25–0.30, indicating effective suppression of close neighbors. In contrast, random splits produced higher similarities (medians 0.25–0.45), broader distributions, and upper-quartile values frequently exceeding 0.40–0.50 (Tables S3, S4).

These results demonstrate that LOCO CV enforces stronger chemical separation than random 5×2 CV, yielding a more extrapolative and chemically challenging evaluation regime.

1.2. Applicability domain and prediction confidence

The left panel of Figure 2 shows the fraction of accurate predictions as a function of maximum similarity to the training set for all individual models, including both raw accuracy values and the smoothed calibration curve. Overall, accuracy increases with increasing maximum similarity to the training set, with occasional fluctuations near the bin boundaries attributable to limited sample sizes in those regions.

Separate analyses for 5×2 and LOCO CV (Figure S5) indicate that differences primarily reflect shifts in similarity coverage rather than systematic change in accuracy at fixed similarity. This suggests that prediction reliability is largely governed by test–train similarity, motivating inclusion of both CV schemes in AD construction.

After obtaining calibration curves for individual models, the ensemble confidence is defined as the median of individual model confidences. The right panel of Figure 2 shows the validation results for the AD, i.e., the observed fraction of accurate predictions as a function of AD-derived prediction confidence, including the ensemble. Overall, all models exhibit an approximately linear relationship between confidence and empirical accuracy, indicating that the AD functions as a global estimator of predictive reliability across these protein targets. Among various models, the ensemble model shows a smoother trend than ROCS-kPLS and 2D-GPR, suggesting improved stability.

Finally, to examine how GPR uncertainty relates to predictive reliability, calibration analysis was performed on the ROCS-GPR model using 5×2 CV predictions for the BACE1 dataset. The calibration curve (Figure S5, left) indicates overconfident and miscalibrated GPR uncertainty estimates, consistent with prior findings^[32,33]. In contrast, AD-derived prediction confidence (Figure S5, right) shows improved correspondence with empirical accuracy, with only slight underestimation at low confidence levels.

1.3. Validation of model design

Having established the AD and ensemble framework, we next validate the model design by assessing (1) whether ensembling improves performance over individual models and (2) whether incorporating 3D information improves affinity prediction.

The resulting Friedman test p-values comparing the ensemble, ROCS-kPLS, ROCS-GPR, and 2D-GPR models across both CV schemes and three performance metrics are summarized in Table S5 (SI). In all cases, p-values were well below the 0.05 significance threshold, indicating statistically significant differences in model rankings. Accordingly, Conover–Holm post hoc tests were performed, with paired significance results shown in Figure 3.

Overall, the ensemble consistently outperforms the 2D-GPR baseline across different CV protocols and performance metrics. However, in contrast to random 5×2 CV, fewer significant differences are detected under cluster-based LOCO CV, and the advantage of the ensemble over the individual 3D models diminishes in this setting. Under random 5×2 CV, higher test–train similarity leads to higher prediction confidence (Figure S7) with balanced model contributions (Figure S8), allowing differences between models to be effectively combined. In contrast, LOCO CV emphasizes extrapolation regions, where lower and more heterogeneous confidence reflects domain shift. In this regime, model errors may become

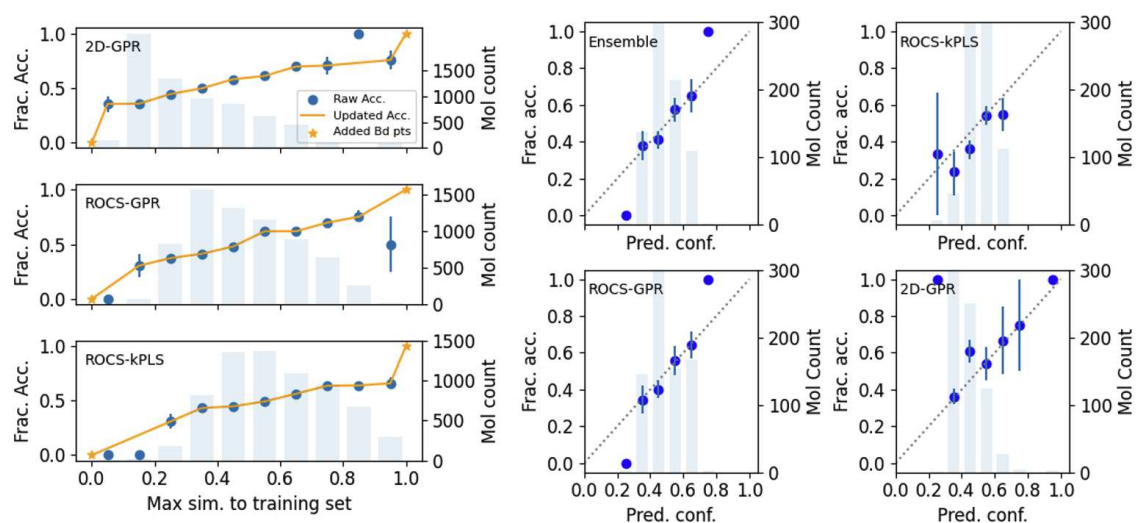


Figure 2. Left (AD): Fraction accurate as a function of normalized maximum similarity to the training set. Uncertainties for raw accuracy values are estimated using 95% confidence interval through bootstrapping. Updated accuracy with added boundary points is used as final calibration curve for prediction confidence. Blue bars correspond to the number of molecules within each similarity bin. Right (validation of AD): Observed fraction accurate as a function of prediction confidence. Uncertainties are estimated using 95% confidence interval through bootstrapping. Blue bars correspond to the number of molecules within each confidence bin.

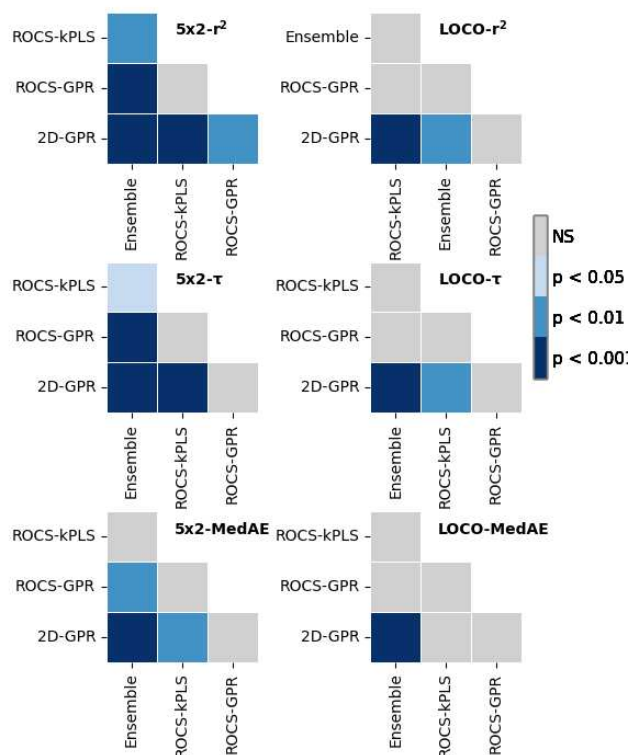


Figure 3. Paired significant plots from Conover-Holm post hoc tests from 5x2 and LOCO CV using Pearson's r^2 , Kendall's τ , and MedAE as performance metrics. NS means not significant. Models are ordered from the highest average rank to the lowest.

more systematic, and the ensemble places greater weight on the 3D models, reducing effective diversity among models and limiting the benefit of ensembling^[51].

To complement the rank-based non-parametric tests, we further examined descriptive effect sizes for the primary comparison between SCOPE and 2D-GPR (Figure S9). Absolute paired Cohen's d ranges from 0.9 to 2.0, with larger effects under random splitting and more moderate but still substantial effects under cluster-based splitting. Consistent improvements are observed for both r^2 and τ , with mean differences of 0.09–0.20. However, average improvements in MedAE are limited to approximately 0.1 log units, which is below typical experimental variability.

These results indicate that the primary benefit of ensembling lies in improved correlation and ranking performance rather than reductions of absolute prediction error. Similar limits on absolute error reduction in ML models have been reported in prior studies^[19]. Furthermore, even rigorous alchemical free energy methods typically achieve absolute errors on the order of 1 kcal/mol (≈ 0.7 log units), approaching experimental reproducibility limits^[52].

Alternative GPR configurations were evaluated but did not outperform the current implementation (Figure S10). A simple k -nearest neighbors (kNN, $k = 3$) model was also compared with 2D-GPR (Figure S11). While kNN often achieves lower MedAE, potentially due to local averaging, it exhibits overall inferior correlation and ranking performance.

1.4. SCOPE vs. third-party methods

In this section, we benchmark SCOPE against various 3D-QSAR and ML methods reported in the literature across multiple datasets, which are predominantly drawn from LO contexts. To ensure a fair and consistent comparison, we adopt the same data-splitting protocols and input 3D conformers used in the original studies. Unless a three-way training/validation/test split was explicitly defined, the latent space dimension d for the ROCS-kPLS model was optimized using leave-one-out CV on the training set.

1.4.1. Industrial dataset PDE10A

Tosstorff *et al.* (2022) released a high-quality industrial dataset comprising 1162 PDE10A (Phosphodiesterase 10A) inhibitors spanning several chemical series with experimentally determined pIC_{50} values^[19]. Ligand poses were generated using a multi-template docking workflow that leveraged 77 PDE10A X-ray protein–ligand co-crystal structures. Using this dataset, the authors designed experiments that reflect realistic LO scenarios and systematically evaluated a wide range of 2D, 3D, and hybrid scoring functions, as well as ML approaches, for binding affinity prediction.

Model performance was assessed under three data-splitting strategies, each employing training, validation, and test sets: (1) A single random split was used to measure interpolation performance; (2) Three rolling temporal splits annotated by the test year were designed to evaluate prospective performance across project timelines, where models were trained exclusively on data generated prior to the test year; (3) Three binding mode splits were used to assess extrapolation capability across protein–ligand interaction patterns. In this setting, models were trained on two binding mode classes and evaluated on the held-out third class.

Because individual molecule-level predictions were not reported, we followed the literature and evaluated test-set

performance using Spearman's ρ and root-mean-square error (RMSE). Performance was then compared across multiple modeling approaches: (1) Mean of combined training and validation set (baseline for RMSE); (2) Vanilla GOLD.PLP: (PLP docking score from GOLD, baseline for Spearman's ρ); (3) Template GOLD.PLP (multi-template docking score); (4) RF-PLP (scoring function with parameters fitted to training set); (5) AttentiveFP (2D NN-based ML); (6) 2D3D hybrid (ensemble average of RF-PLP and AttentiveFP); (7) ACNN (3D NN-based ML); (8) PotentialNet (3D NN-based ML).

Figure 4 compares the predictive performance of SCOPE and 2D-GPR against third-party approaches across three temporal splits and three binding mode splits using Spearman's ρ . SCOPE demonstrates consistently strong performance across all splits in terms of ρ , achieving results comparable to the best-performing models in the literature. In contrast, literature 3D ML models, including ACNN and PotentialNet, exhibit substantial variability across different data splits.

RMSE results (Figure S12) show that none of the evaluated methods exhibits a clear advantage over the simple baseline, defined as the mean value of the combined training and validation set, which yields an RMSE of approximately 1 log unit. This is consistent with the earlier findings where improvements in rank-based metrics do not necessarily translate to reduced absolute error in affinity prediction. For the temporal splits, SCOPE achieves RMSE comparable to the baseline. In contrast, it underperforms the baseline in two of three binding mode splits, a trend also observed for literature 3D ML models.

Performance under the random split (Figure S13) shows that SCOPE achieves comparable performance with the best-performing literature models for both Spearman's ρ and RMSE.

Having established the comparative performance of SCOPE across all splits, we next examine the relationship between predicted and experimental values directly for each split. Figure 5 presents predicted versus experimental pIC_{50} values obtained with SCOPE across the three temporal splits and three binding mode splits.

For the temporal splits, a clear positive trend along the parity line is observed, consistent with the strong Spearman's ρ and lower RMSE values. In particular, alignment is noticeably improved for temporal_split_2013 compared to temporal_split_2011, reflecting cumulative training data that expands chemical coverage. This is further supported by the prevalence of higher-confidence predictions in temporal_split_2013. In contrast, the binding mode splits exhibit systematic deviations from the parity line, with a flattened slope indicating reduced sensitivity to experimental potency. Downward deviations in the first two splits align with higher RMSE values, while a greater proportion of low-confidence predictions suggests limited training coverage of the relevant chemical space.

To further examine individual model behavior under binding mode splits, 2D-GPR, ROCS-GPR, and ROCS-kPLS predictions were compared in Figure S14 (SI). Among these models, 2D-GPR shows the most pronounced prediction flattening, whereas ROCS-kPLS preserves the broadest dynamic range. Reduced flattening in ROCS-GPR compared to 2D-GPR suggests that 3D similarity is more discriminative, mitigating regression-to-the-mean behavior. An alternative 2D-GPR formulation using a Tanimoto kernel yields similarly poor performance (Figure S15), indicating that this

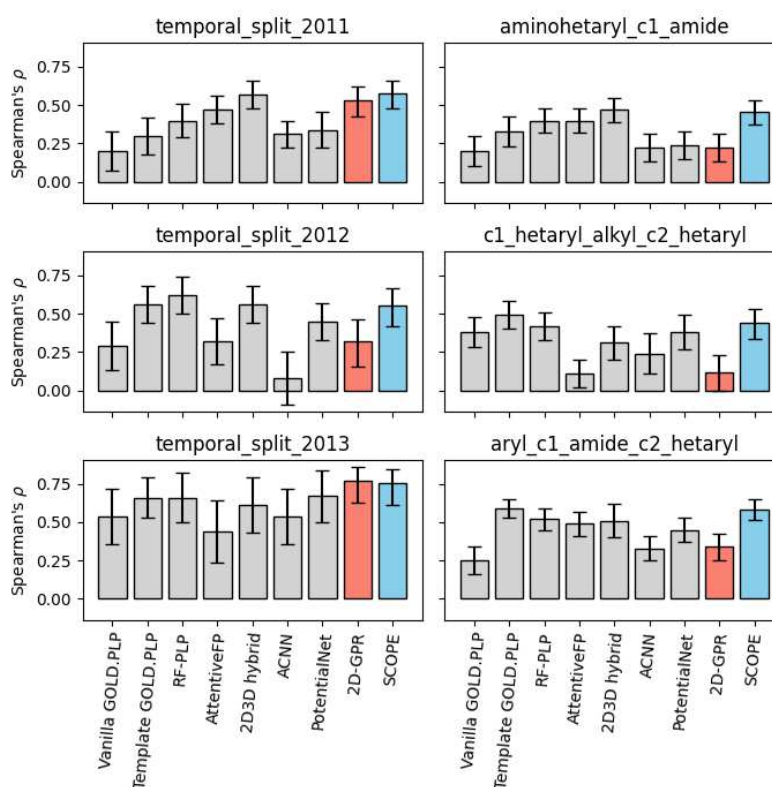


Figure 4. Performance comparison between SCOPE (blue) and 2D-GPR (red) against third-party methods for temporal splits (left) and binding mode splits (right) using Spearman's ρ . Prediction error is estimated using 95% confidence interval.

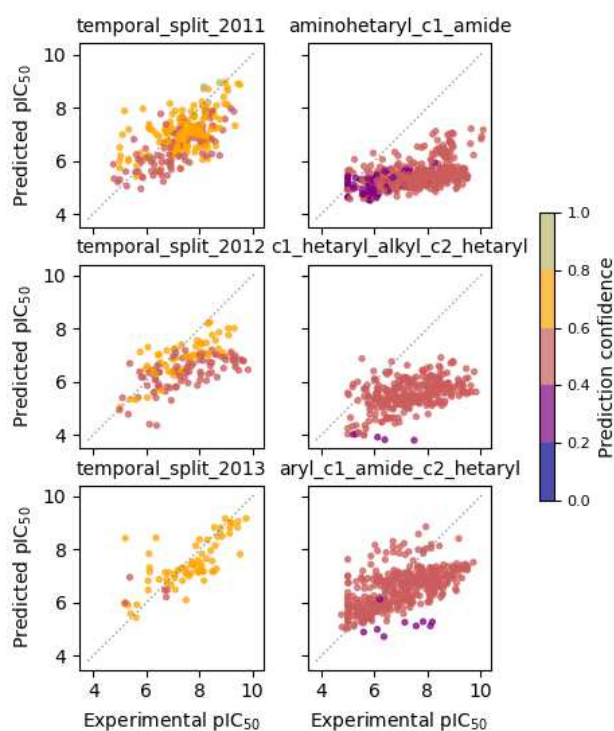


Figure 5. Predicted pIC_{50} versus experimental pIC_{50} under three temporal and three binding mode splits. Data points are colored by their corresponding prediction confidence level.

limitation is not driven by the choice of similarity representation.

Taken together, SCOPE shows strong and competitive performance relative to the best literature models across random, temporal, and binding mode splits in ranking and correlation. Results under binding mode splits further reinforce QSAR best practices emphasizing explicit definition and interpretation of the AD, particularly in extrapolative regimes.

1.4.2. Eight Sutherland datasets

The eight Sutherland datasets^[53] have become widely adopted benchmarks for evaluating 3D-QSAR methods. They comprise eight pre-aligned target-specific ligand sets with experimentally determined binding affinity data. For each dataset, a single training/test split was defined, with approximately one-third assigned to the test set and selected to maximize structural dissimilarity with training set. Details of these datasets are provided in Table S6 (SI).

Benchmarking comparisons include test set Coefficient of Determination (COD) results for CoMFA, CoMSIA-basic (steric and electrostatic fields), and CoMSIA-extra (including hydrophobic and H-bond descriptors), as well as several recently developed 3D-QSAR methods/frameworks: (1) Open3DQSAR^[54], which is an open-source 3D-QSAR framework adopting CoMFA- and CoMSIA-like analysis; (2) COSMOsar3D^[13], which employs field descriptors based on QM-derived local σ -profiles; (3) QMFA^[14], which constructs field descriptors from QM-derived local properties; (4) QMOD^[16], which jointly determines ligand alignment during model optimization.

For Open3DQSAR, two variants were reported: one using the original Sutherland alignments and another incorporating Open3DALIGN^[55], denoted as Open3DQSAR and Open3DALIGN/Open3DQSAR, respectively. In the QMFA study, H-bond interaction fields were generated using water and formamide probes, labeled as QMFA-water and QMFA-formamide. QMOD results were originally reported as Pearson's r^2 , with no expected substantial differences from COD noted by the authors. Results for SCOPE were obtained using the original Sutherland alignment.

Figure 6 shows the distribution of COD values obtained using SCOPE along with the collection of third-party methods described above. Dataset-specific COD values for all methods are reported in Table S7 (SI).

SCOPE achieves one of the highest median COD while maintaining a comparatively tight interquartile range, indicating both strong average predictive accuracy and enhanced robustness across datasets. Although COSMOsar3D and two QMFA-based variants show slightly higher median performance, they exhibit at least one outlier, reflecting reduced robustness. Open3DQSAR attains comparable median COD but displays greater variability, as indicated by a larger interquartile range.

Predicted versus experimental potency values across all eight datasets indicate that these datasets are predominantly evaluated in an interpolative regime with high confidences (0.6–0.8) where test compounds are well represented within the training domain (Figure S16).

1.4.3. BACE1 dataset

Subramanian *et al.* (2016) published a comprehensive dataset of human BACE1 inhibitors and evaluated various regression

models under a challenging, data-limited scenario intended to mimic realistic LO conditions^[56]. Models were trained on 205 ligands with experimentally resolved X-ray binding poses and validated on 1273 structurally related analogs without structural data, all with experimental pIC₅₀ values. Binding poses for validation set were generated by propagating alignments from corresponding reference X-ray ligands. The study systematically compared 2D ML models, including multiple linear regression (MLR), random forest (RF), and deep neural networks (DNN) against 3D-QSAR approaches, including CoMFA, CoMSIA, atom-based QSAR (ABM), and variants using force-field and Gaussian formulations (FQSAR_ff and FQSAR_gau).

Because per-ligand predictions for all models were provided, we used Pearson's r^2 , Kendall's τ , and MedAE as performance metrics. Figure 7 shows the validation set performance comparison among SCOPE, 2D-GPR, and the third-party approaches described above.

Across all three performance metrics, SCOPE demonstrates consistently strong performance, achieving the highest Pearson's r^2 and Kendall's τ . Its performance in MedAE is also comparable to the strongest approaches from the literature, namely 2D ML methods RF and DNN. Although 2D-GPR underperforms relative to RF and DNN in terms of MedAE, it achieves competitive performance in terms of ranking and correlation.

Similar to the Sutherland datasets, predicted versus experimental potency values show that most predictions are associated with relatively high confidence (> 0.6) (Figure S17), consistent with the strong overall predictive performance observed for SCOPE.

1.5. Comparison with <MM/PBSA>

Here we extend the comparison to a physics-based approach, namely ensemble-averaged Molecular Mechanics/Poisson–Boltzmann Surface Area (<MM/PBSA>), derived from molecular dynamics (MD) simulations using the first fourteen structure–affinity datasets. Fully rigorous alchemical free energy methods were not included, as they are typically applied to small, highly optimized congeneric ligand series that comprise only tens of compounds per target^[52]. In contrast, MM/PBSA is used as a computationally efficient, physics-informed ranking method^[57,58], placing it closer in scope to 3D activity models. Details of MD simulations and <MM/PBSA> calculations are provided in Section 1.4 of SI.

Kendall's τ was used as the ranking metric. A small number of ligands were excluded when computing Kendall's τ for <MM/PBSA>, primarily arising from force field application issues. Specifically, 45 ligands were excluded for CAII, 7 for JAK2, and 5 for CDK2, while all remaining datasets lost at most two ligands.

Figure 8 compares Kendall's τ between <MM/PBSA> and SCOPE evaluated under random and cluster-based splits. For SCOPE, results across CV splits are shown as box plots to illustrate the full range of observed performance.

Under random splits, SCOPE performs consistently higher or comparable with <MM/PBSA>. In practice, however, this evaluation regime presupposes the availability of an initial training set that is sufficiently large and diverse within the chemical space of interest. Under this assumption, SCOPE can reliably achieve ranking performance comparable to, and often exceeding <MM/PBSA> across a broad range of targets.

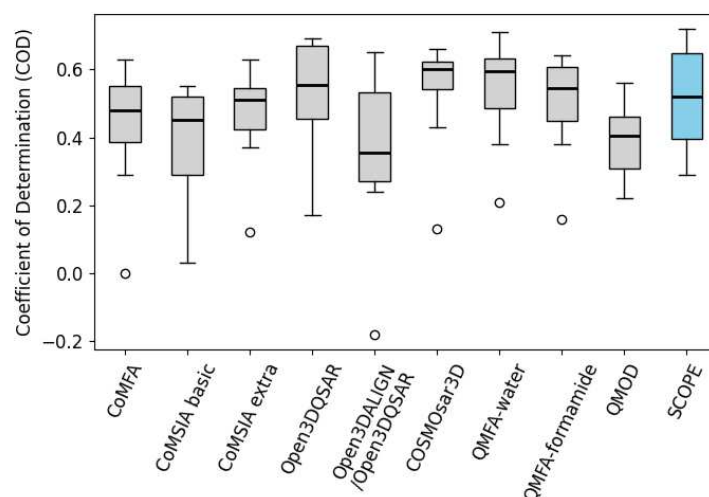


Figure 6. Performance comparison between SCOPE (blue) against third-party 3D-QSAR methods across the eight Sutherland datasets using COD.

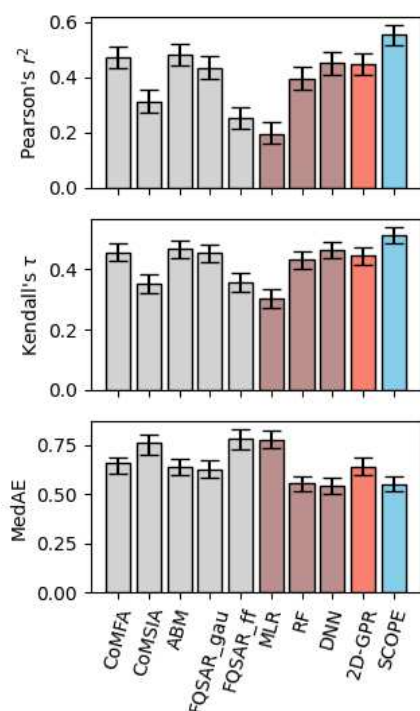


Figure 7. Performance comparison among SCOPE (blue) and 2D-GPR (red) against third-party methods, including classical field-based approaches (gray) and 2D ML methods (brown) for BACE1 using Pearson's r^2 (top), Kendall's τ (middle), and MedAE (bottom). Prediction error is estimated using 95% confidence interval.

In contrast, SCOPE performance evaluated under cluster-based splits exhibits substantially broader distributions of Kendall's τ , highlighting pronounced variability across held-out clusters. These results indicate that extrapolative performance depends strongly on the chemical distinctiveness of the new scaffold. In such cases, a more practical strategy is to first acquire a modest amount of experimental data for the new scaffold to establish baseline coverage of the relevant chemical space, and subsequently retrain or update the model to support further optimization.

1.6. Model interpretation

Beyond predictive performance, interpretability is critical for practical deployment; here we examine how SCOPE can be interpreted to provide actionable design guidance as well as mechanistic insights.

Using temporal_split_2011 of the PDE10A dataset as an illustrative example, the ROCS-kPLS model component was trained on the combined training and validation data comprising compounds prior to 2011. Model interpretation was performed by first constructing a 3D domain grid enclosing all ligands, using a grid spacing of 1 Å. The grid was sampled using multiple probe atoms: a carbon atom and six color atoms defined according to the Implicit Mills–Dean color force field. Figure 9 presents a representative 3D contour map derived from the ROCS-kPLS model highlighting favorable regions for H-bond acceptor.

Importantly, these contour maps capture the collective contribution of all molecules used to build the model, reflecting shared spatial interaction patterns learned across the dataset. As such, they provide actionable guidance for medicinal chemists by highlighting regions where introducing, strengthening, or repositioning a chemical feature is likely to improve activity, as well as regions where such features may be less favorable.

Interpretation of ROCS-GPR was carried out by analyzing the relative contributions of 2D and 3D similarity features across the test set, which serves as a model diagnostic tool. Figure S18 (SI) shows how 3D contributions become more dominant under binding mode split compared to temporal

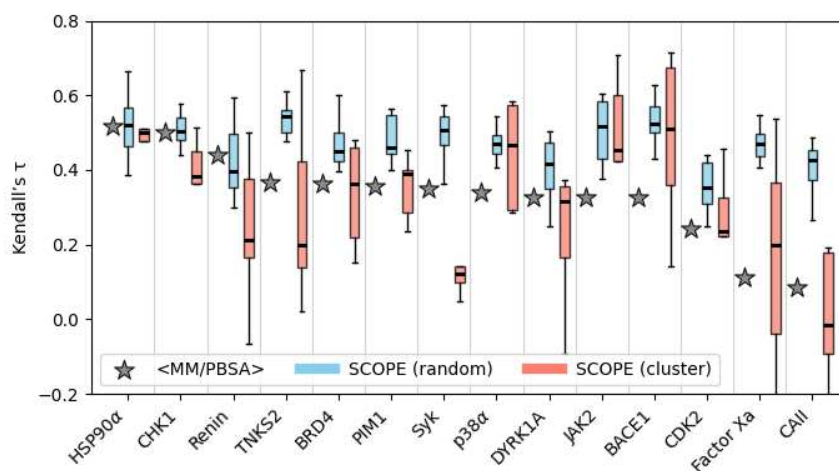


Figure 8. Performance comparison between SCOPE under random and cluster splits against <MM/PBSA> using Kendall's τ . Datasets are in descending order of <MM/PBSA> performance.

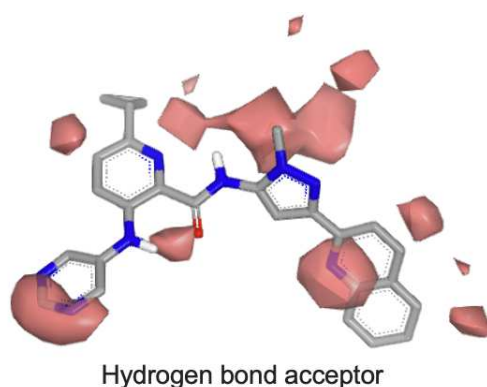


Figure 9. 3D Contour map highlighting favorable regions for H-bond acceptor derived from the ROCS-kPLS model. A high contour value of 0.19 is chosen. The most potent compound from the combined training and validation set is shown for structural reference.

split, reinforcing the benefit of 3D information in extrapolative regimes where 2D similarity is less informative.

1.7. An example lead optimization workflow using SCOPE

Building on the characterization of SCOPE, we demonstrate its use in a representative LO workflow based on the PDE10A dataset. The 2011 temporal split is treated as a retrospective prospective-like scenario, with pre-2011 compounds used for training and 2011 compounds as an independent test set, allowing assessment of how SCOPE would support decision making in practice.

Ligand binding poses were generated using POSIT^[12] against all 77 PDE10A co-crystal structures. One training compound failed pose generation, yielding 622 training, 70 validation, and 250 test compounds. In addition to pose generation, POSIT also provides pose-confidence estimates, which can be used as a quality-control criterion during model construction or incorporated into prediction confidence during prediction.

Figure S19 (SI) shows that both training+validation and test sets are dominated by high pose confidence (0.8–1.0), reflecting the extensive structural coverage in the present setting. Accordingly, no pose-based filtering was applied. In more practical settings with fewer co-crystal structures, pose confidences are expected to be more heterogeneous, and it can be used to filter unreliable poses during model construction to preserve training-set quality, with the choice of threshold depending on the specific dataset and requiring a balance between pose reliability and chemical coverage.

SCOPE was trained on the training+validation set and applied to the temporally held-out test set. At prediction time, prediction confidence was updated as the product of the AD-derived confidence and the corresponding pose confidence, ensuring that predictions are considered confident only when both the learned SAR and the underlying binding pose are well supported.

Spearman's ρ and RMSE are reported for consistency with prior validation analyses. Following the practical-significance framework of Ash *et al.*^[43], regression predictions were converted into post hoc classification decisions using an activity threshold of $\text{pIC}_{50} \geq 7$, and precision and recall were used to assess reliability of prioritized compounds and the fraction of active compounds successfully identified.

To reflect how AD can support decision making, the impact of applying confidence thresholds on predictive performance and decision-oriented metrics is illustrated in Figure S20 (SI). Applying a confidence threshold of 0.6 increases ρ from 0.50 to 0.59, reduces RMSE from 0.96 to 0.86, improves recall from 0.64 to 0.82, and maintains high precision (0.85 to 0.86). We note that the magnitude of these effects depends on the chosen threshold; in practice, threshold selection reflects a balance between reliability and coverage. By comparison, literature-reported poses evaluated on the full set yield slightly higher Spearman's ρ (0.58 [0.48, 0.66]) and lower RMSE (0.94 [0.86, 1.0]) than POSIT-generated poses (Spearman's ρ =0.50 [0.40, 0.60]; RMSE=0.96 [0.88, 1.0]). Importantly, the 95% confidence intervals for these metrics overlap, suggesting that the observed differences are modest.

In this example, filtering out low-confidence predictions improves predictive performance by restricting evaluations

to in-domain regions. In DMTA contexts, however, selectively including low-confidence compounds may be desirable to probe underrepresented regions of chemical space and improve model performance over successive cycles^[59]. Overall, this example demonstrates how SCOPE can support practical LO workflows by integrating AD-based prediction confidence and POSIT-derived pose confidence for informed decision making.

Conclusion

SCOPE incorporates a target-agnostic AD framework that maps similarity to an empirically calibrated prediction confidence. Building on this foundation, it is constructed as a confidence-weighted ensemble of ROCS-kPLS, ROCS-GPR, and 2D-GPR models, with the overall prediction confidence defined as the median of individual model confidences. In prospective applications, this confidence information can be used in several complementary ways. Compounds predicted to have both high affinity and high confidence can be prioritized as reliable candidates for immediate progression. Conversely, compounds assigned low prediction confidence identify regions of chemical space that are poorly represented by existing training data. Strategically selecting and experimentally evaluating a subset of such low-confidence compounds provides a rational, data-driven approach for expanding chemical-space coverage and improving downstream model performance as new data are incorporated.

Validation results using curated structure–affinity datasets demonstrate that SCOPE consistently outperforms the 2D-GPR baseline across multiple performance metrics under both random-split-based and cluster-based CV schemes. These results validate the architectural design of SCOPE and support the conclusion that access to high-quality 3D structural information can substantially enhance ranking- and correlation-based predictive performance, particularly in extrapolative regimes.

Across the PDE10A, eight Sutherland, and BACE1 benchmarks, SCOPE exhibits strong and robust performance relative to field-based 3D-QSAR and modern ML models. On PDE10A, it matches top literature models under temporal and binding mode splits in ranking. On the Sutherland datasets, it ranks among the best COD performers with lower variability than higher-cost QM-based methods. On BACE1, it achieves the strongest correlation and ranking performance with competitive MedAE with the best literature models.

Comparison with (MM/PBSA) highlights the complementary roles of data-driven and physics-based approaches: ML excels once sufficient high-quality data are available, whereas physics-based MD methods remain valuable in data-sparse regimes and offer the additional advantage of mechanistic insight into protein–ligand interactions.

A key strength of SCOPE lies in its interpretability. The 3D contour maps generated from the ROCS-kPLS component provide spatially resolved insight into favorable and unfavorable molecular features learned collectively across the training set, complementing predictive performance with chemically intuitive explanations. These maps offer actionable guidance for ligand design that cannot be inferred from ranking metrics alone, reinforcing the practical value of 3D activity models in medicinal chemistry workflows.

In the representative workflow, we further demonstrate how SCOPE can be applied in a practical LO setting to

guide compound prioritization by jointly incorporating the model's AD and POSIT-derived pose confidence into decision making. Overall, these results position SCOPE as a practical and scalable modeling framework that effectively balances predictive accuracy, interpretability, and computational efficiency. The method is particularly well suited for discovery programs in which reliable binding poses and moderate to large amounts of ligand data are available, while its AD supports responsible use in extrapolative settings.

For future work, incorporating pose ensembles may improve prediction stability and accuracy by reducing sensitivity to the choice of a single predicted pose, for example by aggregating predictions across the top-ranked POSIT poses. Active learning would enable confidence-guided selection of new compounds, allowing systematic expansion of chemical space specifically where the model is most uncertain. Finally, tighter integration with physics-based refinement, for example, applying MD selectively to low-confidence or high-impact compounds, would support a hybrid workflow in which SCOPE provides rapid triaging while physics-based methods offer mechanistic validation.

Acknowledgments

The authors gratefully acknowledge Feng Yu for supporting the workflow for curation of the structure–affinity datasets, and Amir Daghighi for contributing to the construction of the ensemble model during their previous internships at OpenEye, Cadence Molecular Sciences. We also thank Ali Mozaffari for assistance with the early development of these tools, Katarina Stanciakova and Hari Muddana for valuable feedback. Microsoft Copilot was used to assist with language refinement, and all content was reviewed and verified by the authors.

Conflict of Interest

J.L.C., C.N., M.K., A.G.S., and S.K.N. are employees of OpenEye, Cadence Molecular Sciences and may hold equity or stock in the company. S.K. is an employee of Isomorphic Labs. A.R.B. is affiliated with the University of Liverpool and Dalton Tx Ltd. These affiliations may constitute potential conflicts of interest. The authors declare no other competing interests.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

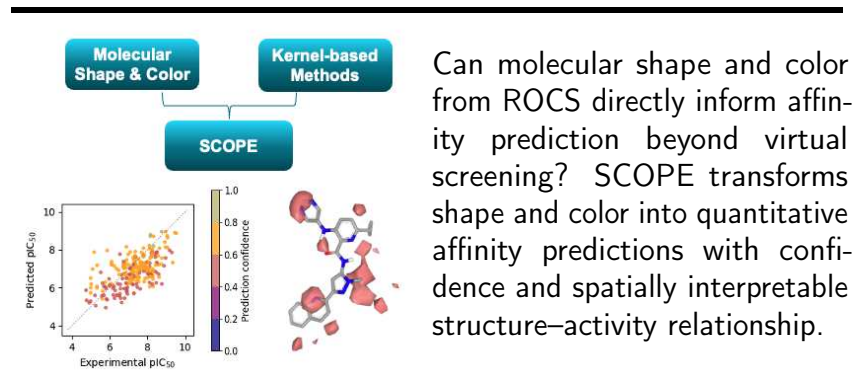
Keywords: SCOPE • ROCS • Prediction confidence • 3D Contour map • Lead optimization

References

- [1] J. Verma, V. M. Khedkar, E. C. Coutinho, *Curr. Top. Med. Chem.* **2010**, *10*, 95.
- [2] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959.
- [3] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130.

- [4] N. Ashraf, A. Asari, N. Yousaf, M. Ahmad, M. Ahmed, A. Faisal, M. Saleem, M. Muddassar, *Front. Chem.* **2022**, *10*, 1003816.
- [5] T. Kim, K.-C. Chung, H. Park, *Pharmaceuticals* **2023**, *16*, 1509.
- [6] Certara, L.P., *SYBYL-X*, Certara, Princeton, NJ, USA.
- [7] Pharmacelera, PharmQSAR: 3D QSAR Software Package.
- [8] P. Tosco, T. Balle, *J. Mol. Model.* **2011**, *17*, 201.
- [9] R. Ragno, *J. Comput.-Aided Mol. Des.* **2019**, *33*, 855.
- [10] C. L. Haga, C. N. Le, X. D. Yang, D. G. Phinney, *Pharmaceuticals* **2025**, *18*, 440.
- [11] S. K. Burley, A. Wu-Wu, S. Dutta, S. Ganesan, S. X. F. Zheng, *Oncogene* **2024**, *43*, 2229.
- [12] B. P. Kelley, S. P. Brown, G. L. Warren, S. W. Muchmore, *J. Chem. Inf. Model.* **2015**, *55*, 1771.
- [13] A. Klamt, M. Thormann, K. Wichmann, P. Tosco, *J. Chem. Inf. Model.* **2012**, *52*, 2157.
- [14] A. El Kerdawy, S. Güßregen, H. Matter, M. Hennemann, T. Clark, *J. Chem. Inf. Model.* **2013**, *53*, 1486.
- [15] A. N. Jain, *J. Comput.-Aided Mol. Des.* **2010**, *24*, 865.
- [16] A. E. Cleves, A. N. Jain, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 127.
- [17] A. E. Cleves, A. N. Jain, *J. Comput.-Aided Mol. Des.* **2018**, *32*, 731.
- [18] G. Floresta, V. Abbate, *RSC Adv.* **2021**, *11*, 14587.
- [19] A. Tosstorff, M. G. Rudolph, J. C. Cole, M. Reutlinger, C. Kramer, H. Schaffhauser, A. Nilly, A. Flohr, B. Kuhn, *J. Comput.-Aided Mol. Des.* **2022**, *36*, 753.
- [20] M. S. Bahia, O. Kaspi, M. Touitou, I. Binayev, S. Dhail, J. Spiegel, N. Khazanov, A. Yosipof, H. Senderowitz, *Mol. Inf.* **2023**, *42*, 2200186.
- [21] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar, V. S. Pande, *ACS Central Sci.* **2018**, *4*, 1520.
- [22] A. Tropsha, *Molecular Informatics* **2010**, *29*, 476.
- [23] Organisation for Economic Co-operation and Development, *Guidance document on the validation of (quantitative) structure-activity relationship [(Q) SAR] models*, Organisation for Economic Co-operation and Development **2014**.
- [24] P. C. D. Hawkins, A. G. Skillman, A. Nicholls, *J. Med. Chem.* **2007**, *50*, 74.
- [25] T. Sato, H. Yuki, D. Takaya, S. Sasaki, A. Tanaka, T. Honma, *J. Chem. Inf. Model.* **2012**, *52*, 1015.
- [26] S. Kearnes, V. S. Pande, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 609.
- [27] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, *Molecules* **2012**, *17*, 4791.
- [28] J. E. J. Mills, P. M. Dean, *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607.
- [29] S. Wold, M. Sjöström, L. Eriksson, *Chemometrics and Intelligent Laboratory Systems* **2001**, *58*, 109.
- [30] OpenEye, Cadence Molecular Sciences, Santa Fe, NM, *OEKernelPLS* **2026**.
- [31] R. Rosipal, L. J. Trejo, *J. Mach. Learn. Res.* **2001**, *2*, 97.
- [32] A. Capone, S. Hirche, G. Pleiss, *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 36579.
- [33] M. Holzenkamp, D. Lyu, U. Kleinekathöfer, P. Zaspel, *Mach. Learn.: Sci. Technol.* **2025**, *6*, 045019.
- [34] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, *Neural Netw.* **2005**, *18*, 1093.
- [35] OpenEye, Cadence Molecular Sciences, Santa Fe, NM, *GraphSim TK* **2026**.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
- [37] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.
- [38] S. P. Brown, S. W. Muchmore, P. J. Hajduk, *Drug Discov. Today* **2009**, *14*, 420.
- [39] C. Kramer, T. Kalliokoski, P. Gedeck, A. Vulpetti, *J. Med. Chem.* **2012**, *55*, 5165.
- [40] T. Kalliokoski, C. Kramer, A. Vulpetti, P. Gedeck, *PLoS One* **2013**, *8*, e61007.
- [41] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2014**, *57*, 4977.
- [42] T. G. Dietterich, *Neural Comput.* **1998**, *10*, 1895.
- [43] J. R. Ash, C. Wognum, R. Rodríguez-Pérez, M. Aldeghi, A. C. Cheng, D.-A. Clevert, O. Engkvist, C. Fang, D. J. Price, J. M. Hughes-Oliver, W. P. Walters, *J. Chem. Inf. Model.* **2025**.
- [44] scikit-learn-extra contributors, *scikit-learn-extra: Additional tools for scikit-learn* **2020**.
- [45] O. Rainio, J. Teuhon, R. Klén, *Sci. Rep.* **2024**, *14*, 6086.
- [46] M. Friedman, *J. Am. Stat. Assoc.* **1937**, *32*, 675.
- [47] W. J. Conover, *Practical Nonparametric Statistics*, Wiley **1999**.
- [48] S. Holm, *Scand. J. Stat.* **1979**, pages 65–70.
- [49] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, *Nat. Methods* **2020**, *17*, 261.
- [50] M. A. Terpilowski, *J. Open Source Softw.* **2019**, *4*, 1169.
- [51] D. Wood, T. Mu, A. M. Webb, H. W. J. Reeve, M. Luján, G. Brown, *J. Mach. Learn. Res.* **2023**, *24*, 1.
- [52] G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, E. D. Harder, L. Wang, *Communications Chemistry* **2023**, *6*, 222.
- [53] J. J. Sutherland, L. A. O'Brien, D. F. Weaver, *J. Med. Chem.* **2004**, *47*, 5541.
- [54] P. Tosco, T. Balle, *J. Chem. Inf. Model.* **2012**, *52*, 302.
- [55] P. Tosco, T. Balle, F. Shiri, *J. Comput.-Aided Mol. Des.* **2011**, *25*, 777.
- [56] G. Subramanian, B. Ramsundar, V. Pande, R. A. Denny, *J. Chem. Inf. Model.* **2016**, *56*, 1936.
- [57] S. Genheden, U. Ryde, *Expert Opin. Drug Discov.* **2015**, *10*, 449.
- [58] C. Wang, P. H. Nguyen, K. Pham, D. Huynh, T.-B. N. Le, H. Wang, P. Ren, R. Luo, *J. Comput. Chem.* **2016**, *37*, 2436.
- [59] P. Mas, B. Filoche-Rommé, M. Bianciotto, R. Vuilleumier, *Digital Discovery* **2026**.

Entry for the Table of Contents



Can molecular shape and color from ROCS directly inform affinity prediction beyond virtual screening? SCOPE transforms shape and color into quantitative affinity predictions with confidence and spatially interpretable structure–activity relationship.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Sl.pdf](#)
- [datasets.zip](#)