RESEARCH PAPER

Integrating Computational Chemistry, Biochemistry, and Machine Learning for Rational Drug Design

Dr.Dhanya Sudheesh[1], Pranit Saraswat[2],Simran[3]
[1]Department of Biochemistry, Indira Gandhi institute of Dental Sciences
sudhidhanya09@gmail.com
[2]Associate Professor, Pharmaceutical Chemistry, Saraswati College of Pharmacy, Gharaun, Mohali
Pranitsaraswat41@gmail.com
[3]Saraswati Institute of Pharmaceutical Education and Research, SGC Group, Gharuan-140413, Mohali, Punjab, India
Orchid ID- 0000-0003-2935-8009
simrankundal55@gmail.com

Abstract
Background: Rational drug design increasingly combines physics-based computational chemistry, biochemical assay data and machine learning (ML), but the statistical gain from integrating these information sources is rarely quantified under controlled conditions in which the true structure–activity relationship is known.
Objective: To evaluate, with a simulation study of known ground truth, whether integrating computational-chemistry and biochemical descriptors improves ML prediction of potency (pIC50), and to compare seven ML algorithms using rigorous statistical tests.
Methods: A virtual library of 1200 compounds in 40 chemical series was simulated with physicochemical, quantum-chemical (HOMO–LUMO gap, dipole moment) and structure-based descriptors (docking score, MM-GBSA binding energy), biochemical descriptors (thermal shift ΔTm, kinetic solubility) and pIC50 generated from a known non-linear model with realistic assay noise (SD 0.40 log units). Seven algorithms (MLR, PLS, kNN, SVR, RF, GBM and MLP) were trained on computational-chemistry (CC), biochemical (BC) and integrated (INT) feature sets and assessed by 10 × repeated, series-grouped 5-fold cross-validation (50 folds), followed by validation on 250 compounds from eight unseen series. Models were compared with the Friedman test and Nemenyi critical difference, Wilcoxon signed-rank tests with Holm correction and the Nadeau–Bengio corrected resampled t-test; robustness was assessed by Y-randomisation, bootstrap confidence intervals and applicability-domain analysis.
Results: Model performance differed significantly (Friedman $\chi^2(6) = 183.9$, $p < 0.001$, Kendall's $W = 0.61$); RF ranked best (mean rank 1.32). Integrated descriptors reduced RMSE compared with CC descriptors in all 7 models (corrected t-test, $p < 0.001$ for RF; Cohen's $d_z = 4.64$). On the external series the integrated RF model achieved $R^2 = 0.62$ (95% CI 0.54–0.68), RMSE = 0.64 and concordance correlation = 0.75, against a theoretical noise ceiling of $R^2 = 0.85$. Y-randomisation confirmed the absence of chance correlation (empirical $p = 0.010$). For identifying actives ($pIC50 \geq 6.5$), the integrated model (AUC = 0.88) outperformed docking scores alone (AUC = 0.77; ΔAUC = 0.11, 95% CI 0.05–0.17). Permutation importance recovered the descriptors that drove potency in the generating model.
Conclusion: Under controlled conditions, integrating biochemical measurements with computational-chemistry descriptors yields statistically and practically meaningful gains in potency prediction and hit identification. Rigorous, group-aware validation and appropriate multiple-comparison statistics are essential to distinguish genuine improvements from noise. These findings should be confirmed with experimental datasets.
Keywords: Rational drug design; machine learning; computational chemistry; QSAR; molecular docking; MM-GBSA; thermal shift assay; statistical model comparison; simulation study
How to cite this article: Dhanya Sudheesh. Integrating Computational Chemistry, Biochemistry, and Machine Learning for Rational Drug Design. Int J Drug Deliv Technol. 2026;16(89s):1734-1742. DOI: 10.25258/ijddt.16.89s.182.

1. Introduction

Rational drug design aims to discover and optimise bioactive molecules using knowledge of the biological target, the physicochemical properties of ligands and the relationship between molecular structure and activity. Computational methods now contribute to almost every stage of this process, from target assessment and virtual screening to lead optimisation [1,2]. Three complementary sources of information are typically available: computational chemistry, which describes molecules through physicochemical, quantum-chemical and structure-based descriptors; biochemistry, which provides direct experimental evidence of target engagement and assay behaviour; and machine learning, which can learn non-linear relationships from these data [3,4].

Each source has well-documented limitations. Docking scoring functions are fast but correlate only modestly with measured affinity and are biased by molecular size [5]. End-point free-energy methods such as MM-GBSA usually rank ligands better than docking but are computationally costly and sensitive to the system studied [6,7]. Biochemical measurements such as differential scanning

fluorimetry (thermal shift, ΔTm) provide direct evidence of binding but do not always scale linearly with potency [8], and assay artefacts linked to poor solubility or aggregation can distort apparent activity [9,10]. Experimental potency data are themselves noisy: replicate pKi values in public databases differ by around 0.5 log units [11].

Machine-learning quantitative structure–activity relationship (QSAR) models can combine these heterogeneous descriptors [3,4]. However, published comparisons of algorithms and descriptor sets frequently rely on random rather than series-aware data splits, report single-split metrics without uncertainty, and compare many models without correcting for multiple testing [12–14]. Moreover, with real data the true structure–activity relationship is unknown, so it is impossible to judge how close a model comes to the best achievable performance or whether its interpretation is correct.

Simulation studies, in which data are generated from a known model, overcome this problem and are a standard tool for evaluating statistical methods [15]. This study therefore used a transparent, fully reproducible simulation to address four questions: (1) whether integrating computational-chemistry and biochemical descriptors improves potency prediction compared with either source alone; (2) which of seven commonly used ML algorithms performs best, judged by appropriate non-parametric tests; (3) how close the best model approaches the theoretical noise ceiling and how robust it is to chance correlation and extrapolation; and (4) whether model interpretation recovers the true drivers of potency.

2. Materials and Methods

2.1 Study design

The study was designed and reported following the ADEMP framework for simulation studies (aims, data-generating mechanism, estimands, methods and performance measures) [15]. All data are synthetic; no compounds were synthesised or tested. A single random seed (20261027) was used throughout, and the code and simulated dataset are provided as supplementary material so that every result can be reproduced.

2.2 Data-generating mechanism

A virtual library of 1200 compounds distributed across 40 chemical series (scaffold clusters) was generated. Each compound was assigned an unobserved latent binding propensity $B \sim N(0, 1)$, and each series a potency offset drawn from $N(0, 0.35^2)$ to mimic scaffold-level structure–activity relationships. Descriptors were generated as correlated, noisy functions of B and of each other, with deliberately realistic biases (Table 1). Potency was generated as:

$$pIC50 = 6.0 + 0.80B + s_series - 0.10(cLogP - 3)^2 - 0.004 \cdot max(TPSA - 120, 0) + 0.25 \cdot tanh(gap - 4.6) \cdot B + a + \varepsilon$$

where s_series is the series offset, $a = -0.35$ when $logS < -5.5$ (a solubility-related assay artefact, otherwise 0), and $\varepsilon \sim N(0, 0.40^2)$ represents assay noise consistent with the variability reported for experimental potency data [11]. The model therefore contains a lipophilicity optimum, a polarity penalty, an interaction between electronic structure and binding, and an assay artefact. Two descriptors (MM-GBSA and ΔTm) had 2% of values set to missing at random to mimic incomplete data; these were imputed with the training-set median within each model pipeline.

Table 1: Descriptors, generating mechanism and feature-set assignment

Descriptor (unit)	Generating mechanism (summary)	Set
MW (Da)	Normal (mean 420, SD 70) + series effect; truncated 200–700	CC
cLogP	Linear in MW and B + noise; truncated −1.5 to 7.5	CC
TPSA (Å²)	Linear in MW and cLogP + noise	CC
HBD, HBA, RotB (counts)	Poisson, means dependent on TPSA or MW	CC
HOMO–LUMO gap (eV)	Normal (4.6, 0.55); modifies effect of B on potency	CC
Dipole moment (D)	Gamma (shape 4, scale 1.1); no effect on potency	CC
Docking score (kcal/mol)	$-7.6 - 0.95B - 0.0035(MW - 420) + N(0, 1.15^2)$; size-biased	CC
MM-GBSA ΔG (kcal/mol)	$-44 - 6.2B - 1.5(cLogP - 2.8) + N(0, 5^2)$; lipophilicity-biased	CC
ΔTm, thermal shift (°C)	1.8 + 1.4B + 0.4·series offset + $N(0, 0.9^2)$	BC
logS, kinetic solubility (log mol/L)	Linear in cLogP and MW + noise; triggers assay artefact below −5.5	BC
pIC50 (outcome)	Equation above; assay noise SD 0.40	—

CC, computational chemistry; BC, biochemical; B, latent binding propensity.

2.3 Feature sets and models

Three feature sets were compared: computational chemistry (CC; ten descriptors), biochemical (BC; two descriptors) and integrated (INT; all twelve). Seven algorithms were evaluated: multiple linear regression (MLR), partial least squares (PLS), k-nearest neighbours (kNN), support vector regression (SVR), random forest (RF), gradient boosting machine (GBM), multilayer perceptron (MLP) [16–

19]. Hyperparameters were fixed in advance at widely used values to avoid optimistic bias from tuning on the evaluation folds (PLS, up to 5 components; kNN, k = 7 with distance weighting; SVR, radial basis kernel, C = 3, ε = 0.1; RF, 200 trees, minimum leaf size 2, 50% of features per split; GBM, histogram-based, learning rate 0.05, 300 iterations, 15 leaves, L2 penalty 1; MLP, two hidden layers of 64 and 32 units, L2 penalty 0.01, early stopping). Descriptors were median-imputed and, except for the tree-based models, standardised within each training fold.

2.4 Validation strategy

To mimic prospective prediction of new chemotypes, eight of the 40 series (250 compounds) were randomly held out as an external test set before any modelling; the remaining 32 series (950 compounds) formed the training set [12,13]. Within the training set, models were assessed by series-grouped 5-fold cross-validation repeated 10 times with different random assignments of series to folds (50 folds), so that compounds of the same series never appeared in both training and validation folds. The model ranked best in cross-validation was refitted on the whole training set and evaluated once on the external set.

2.5 Performance measures and statistical analysis

Regression performance was summarised by the coefficient of determination (R^2, or Q^2 in cross-validation), root-mean-square error (RMSE), mean absolute error (MAE), Pearson's r and Lin's concordance correlation coefficient (CCC) [20]. Ninety-five per cent confidence intervals for external metrics were obtained from 2,000 bootstrap resamples [21].

Descriptor distributions were summarised by mean, SD, median and skewness; normality of pIC50 was assessed with the Shapiro–Wilk test. Associations between each descriptor and pIC50 were assessed with Pearson and Spearman correlations, with Holm correction for 12 tests [22]. Multicollinearity was assessed with variance inflation factors (VIF).

Following Demšar's recommendations [14], the seven algorithms were compared on the integrated feature set using the Friedman test on fold-wise RMSE [23], with Kendall's W as effect size and the Nemenyi critical difference (CD) for mean ranks. The best-ranked model was compared with each other model using Wilcoxon signed-rank tests with Holm correction. Because cross-validation folds share training data, feature sets were compared within each algorithm using the Nadeau–Bengio corrected resampled t-test [24], with Cohen's d_z as effect size.

Robustness was examined in three ways. First, a Y-randomisation test with 100 permutations of pIC50 assessed chance correlation, with an empirical p-value [25]. Second, because the true potency was known, a noise ceiling was calculated as the R^2 between noise-free and observed pIC50 in the external set, representing the best achievable performance. Third, an applicability domain (AD) was defined as the mean Euclidean distance of each compound to its five nearest training neighbours in standardised descriptor space, with the 95th percentile of training distances as threshold [26]; absolute errors inside and outside the AD were compared with a one-sided Mann–Whitney U test.

Model interpretation used permutation importance on the external set (30 repeats), whose ranking was compared with the known generating model. Finally, to reflect hit identification, compounds with pIC50 ≥ 6.5 were labelled active and the ability of docking score, MM-GBSA, ΔTm and the integrated model to rank actives was compared by the area under the ROC curve (AUC) with bootstrap 95% CI, the paired bootstrap difference in AUC, and the enrichment factor in the top 5% (EF5%) [27]. Sensitivity, specificity, Matthews correlation coefficient (MCC) and balanced accuracy were calculated for the model at the same threshold. Two-sided α = 0.05 was used throughout. Analyses were performed in Python 3.11 with NumPy, pandas, SciPy, scikit-learn [28], Matplotlib and seaborn.

3. Results

3.1 Descriptive statistics

The simulated library contained 1200 compounds, of which 283 (23.6%) met the activity threshold of pIC50 ≥ 6.5. pIC50 had a mean of 5.78 (SD 1.03; range 1.85–8.92) and was approximately symmetrical (skewness -0.08), although the Shapiro–Wilk test detected a small departure from normality given the large sample (W = 0.997, p = 0.018). Descriptor distributions are summarised in Table 2; physicochemical properties fell within ranges typical of drug-like screening libraries [10].

Table 2: Descriptive statistics of simulated descriptors and potency (N = 1,200)

Variable	n	Mean	SD	Median	Min	Max	Skewness
MW (Da)	1200	434.47	80.47	439.55	200.00	700.00	-0.04
cLogP	1200	2.93	1.32	2.95	-0.90	7.13	-0.05
TPSA (Å²)	1200	84.99	23.27	85.00	20.00	159.90	-0.03
HBD	1200	2.34	1.54	2.00	0.00	9.00	0.69
HBA	1200	5.82	2.48	6.00	0.00	15.00	0.34
RotB	1200	5.34	2.57	5.00	0.00	14.00	0.40

Variable	n	Mean	SD	Median	Min	Max	Skewness
HOMO–LUMO gap (eV)	1200	4.60	0.56	4.59	2.88	6.59	-0.01
Dipole (D)	1200	4.26	2.21	3.83	0.16	17.31	1.18
Docking (kcal/mol)	1200	-7.65	1.51	-7.65	-12.67	-3.04	-0.03
MM-GBSA (kcal/mol)	1176	-44.32	8.52	-44.53	-70.95	-12.46	0.15
ΔTm (°C)	1176	1.82	1.70	1.83	-5.41	6.98	-0.07
logS	1200	-3.86	1.11	-3.86	-7.37	-0.52	0.02
pIC50	1200	5.78	1.03	5.78	1.85	8.92	-0.08

3.2 Correlations and multicollinearity

After Holm correction, 5 of 12 descriptors were significantly correlated with pIC50 (Table 3; Figure 1). The strongest associations were with ΔTm (r = 0.70), MM-GBSA (r = -0.61) and docking score (r = -0.47); negative coefficients for the two energy terms reflect more favourable (more negative) binding energies in more potent compounds. cLogP showed a weak positive linear association (r = 0.15) despite its quadratic role in the generating model, illustrating that univariate correlations can understate non-linear effects. All VIFs were below 3.5 (highest: logS), indicating no problematic multicollinearity.

Table 3: Correlation of each descriptor with pIC50

Descriptor	Pearson r	p (Holm)	Spearman ρ	VIF
MW (Da)	0.009	1.000	0.018	2.04
cLogP	0.152	< 0.001	0.149	3.12
TPSA (Å²)	-0.095	0.008	-0.087	1.25
HBD	-0.028	1.000	-0.029	1.05
HBA	-0.021	1.000	-0.012	1.09
RotB	0.050	0.502	0.042	1.25
HOMO–LUMO gap (eV)	0.016	1.000	0.010	1.00
Dipole (D)	-0.010	1.000	-0.027	1.00
Docking (kcal/mol)	-0.473	< 0.001	-0.436	1.58
MM-GBSA (kcal/mol)	-0.607	< 0.001	-0.569	2.18
ΔTm (°C)	0.704	< 0.001	0.682	1.95
logS	-0.068	0.124	-0.075	3.53

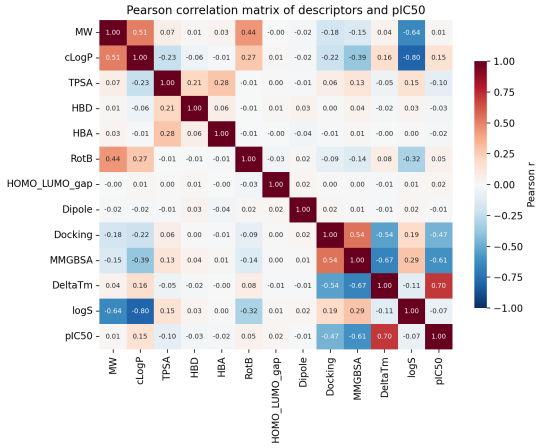


Figure 1: Pearson correlation matrix of the descriptors and pIC50.

3.3 Cross-validated comparison of algorithms

Cross-validated performance for every combination of algorithm and feature set is shown in Table 4 and Figure 2. With integrated descriptors, mean Q² ranged from 0.370 (kNN) to 0.479 (RF), and RMSE from 0.720 to 0.793 pIC50 units.

Table 4: Cross-validated performance (mean ± SD over 50 series-grouped folds)

Model	CC: Q²	CC: RMSE	BC: Q²	BC: RMSE	INT: Q²	INT: RMSE
MLR	0.310 ± 0.097	0.830 ± 0.079	0.416 ± 0.079	0.763 ± 0.065	0.451 ± 0.088	0.740 ± 0.069
PLS	0.310 ± 0.097	0.830 ± 0.079	0.416 ± 0.079	0.763 ± 0.065	0.451 ± 0.088	0.739 ± 0.069
kNN	0.209 ± 0.095	0.889 ± 0.073	0.356 ± 0.085	0.800 ± 0.055	0.370 ± 0.090	0.793 ± 0.075
SVR	0.214 ± 0.122	0.884 ± 0.078	0.419 ± 0.072	0.762 ± 0.065	0.406 ± 0.085	0.769 ± 0.064

Model	CC: Q^2	CC: RMSE	BC: Q^2	BC: RMSE	INT: Q^2	INT: RMSE
RF	0.337 ± 0.099	0.814 ± 0.080	0.378 ± 0.084	0.788 ± 0.062	0.479 ± 0.085	0.720 ± 0.073
GBM	0.262 ± 0.110	0.857 ± 0.074	0.349 ± 0.084	0.805 ± 0.062	0.420 ± 0.091	0.759 ± 0.067
MLP	0.231 ± 0.116	0.875 ± 0.072	0.425 ± 0.079	0.758 ± 0.068	0.400 ± 0.093	0.773 ± 0.067

CC, computational chemistry (10 descriptors); BC, biochemical (2); INT, integrated (12). Q^2, cross-validated R^2; RMSE in pIC50 units.

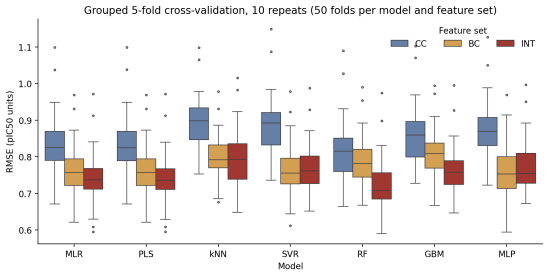


Figure 2: Distribution of fold-wise RMSE for each algorithm and feature set (50 folds each).

The Friedman test showed significant differences between algorithms on the integrated set ($\chi^2(6) = 183.9$, $p < 0.001$) with a large effect (Kendall's $W = 0.61$). Mean ranks were RF 1.32, PLS 2.66, MLR 3.22, GBM 4.42, MLP 5.02, SVR 5.08, kNN 6.28 (Figure 3). With a Nemenyi CD of 1.27, RF ranked significantly better than all six other algorithms; the smallest rank difference was with PLS (1.34). Pairwise Wilcoxon tests with Holm correction confirmed that RF had lower RMSE than every other algorithm (all adjusted $p < 0.001$; Table 5). The practical margin over the linear models was, however, small: the median fold-wise RMSE advantage of RF over MLR and PLS was only 0.020 and 0.020 pIC50 units, compared with 0.071 over kNN.

Table 5: Comparison of the best-ranked model with other algorithms (integrated set, fold-wise RMSE)

Comparison	Mean Friedman rank	Median ΔRMSE	Folds favouring RF (%)	Wilcoxon W	Holm-adjusted p
RF vs MLR	1.32 vs 3.22	0.020	88	65	< 0.001
RF vs PLS	1.32 vs 2.66	0.020	88	65	< 0.001
RF vs kNN	1.32 vs 6.28	0.071	100	0	< 0.001
RF vs SVR	1.32 vs 5.08	0.050	100	0	< 0.001
RF vs GBM	1.32 vs 4.42	0.041	100	0	< 0.001
RF vs MLP	1.32 vs 5.02	0.052	92	26	< 0.001

ΔRMSE = RMSE(comparator) − RMSE(RF); positive values favour RF.

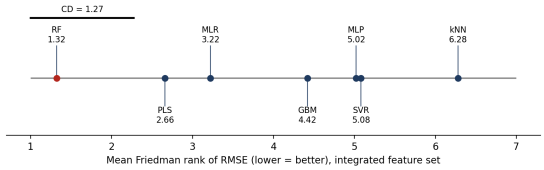


Figure 3: Mean Friedman ranks of the seven algorithms (integrated set). The bar shows the Nemenyi critical difference ($\alpha = 0.05$).

3.4 Effect of integrating descriptor sources

Integrating biochemical with computational-chemistry descriptors significantly reduced RMSE compared with computational-chemistry descriptors alone for all 7 algorithms (corrected resampled t-test, all $p < 0.001$), with large effect sizes (d_z 2.86–4.69; Table 6). Compared with biochemical descriptors alone, integration improved RMSE significantly for MLR, PLS, RF, GBM, but not for kNN, SVR, MLP. For RF, mean RMSE fell by 0.093 relative to CC and by 0.067 relative to BC. The two biochemical descriptors alone gave lower RMSE than the ten computational-chemistry descriptors for 7 of 7 algorithms (significantly for 6), reflecting the more direct link between thermal shift and binding in the generating model.

Table 6: Comparison of feature sets within each algorithm (Nadeau–Bengio corrected resampled t-test, df = 49)

Model	INT vs CC: ΔRMSE	t	P	d_z	INT vs BC: ΔRMSE	t	p	d_z
MLR	0.091	7.05	< 0.001	3.66	0.024	2.73	0.009	1.42
PLS	0.091	7.13	< 0.001	3.70	0.024	2.77	0.008	1.44
kNN	0.096	9.03	< 0.001	4.69	0.008	0.38	0.708	0.20
SVR	0.115	8.15	< 0.001	4.23	-0.007	-0.48	0.633	-0.25
RF	0.093	8.94	< 0.001	4.64	0.067	5.05	< 0.001	2.62
GBM	0.098	8.32	< 0.001	4.33	0.046	2.96	0.005	1.54
MLP	0.102	5.50	< 0.001	2.86	-0.015	-0.74	0.465	-0.38

ΔRMSE = mean RMSE(comparator set) − mean RMSE(INT); positive values favour the integrated set.

3.5 External validation and noise ceiling

Refitted on all 950 training compounds and applied to the 250 compounds of the eight unseen series, the integrated RF model achieved R^2 = 0.62 (95% CI 0.54–0.68), RMSE = 0.64 (0.59–0.70), MAE = 0.50, r = 0.79 and CCC = 0.75 (Table 7; Figure 4). The integrated MLR model performed almost identically (R^2 = 0.62), whereas RF models restricted to CC or BC descriptors performed markedly worse (R^2 = 0.51 and 0.48).

Because the noise-free potency was known, the best achievable performance could be quantified. The R^2 between true and observed pIC50 in the external set, the noise ceiling, was 0.85 (RMSE 0.41). The integrated model therefore captured 73% of the explainable variance; its predictions correlated with the true potency at r = 0.84 (R^2 = 0.70).

Table 7: External validation on eight unseen chemical series (n = 250)

Model	R^2	RMSE	MAE	Pearson r	CCC
RF, integrated (95% CI)	0.62 (0.54–0.68)	0.64 (0.59–0.70)	0.50 (0.45–0.55)	0.79 (0.74–0.83)	0.75 (0.70–0.79)
MLR, integrated	0.62	0.64	0.50	0.79	0.76
RF, CC only	0.51	0.73	0.59	0.71	0.66
RF, BC only	0.48	0.75	0.59	0.70	0.65
Noise ceiling (true vs observed)	0.85	0.41	—	—	—

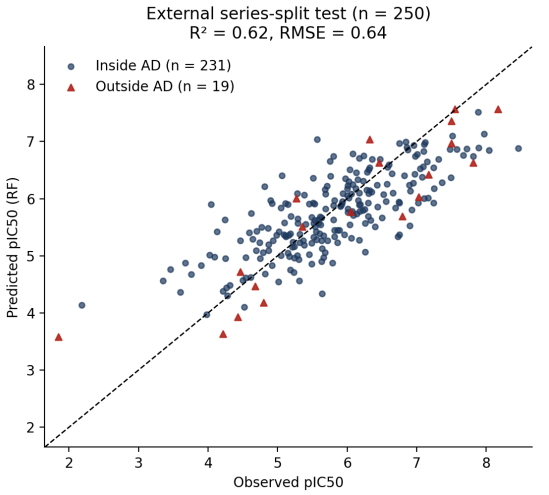


Figure 4: Observed versus predicted pIC50 for the external test set, coloured by applicability domain. Dashed line: identity.

3.6 Y-randomisation and applicability domain

In the Y-randomisation test, the cross-validated Q^2 of the RF model with true pIC50 was 0.49, whereas the 100 models built on permuted pIC50 had a mean Q^2 of -0.05 (SD 0.017; maximum -0.01), giving an empirical p = 0.010, the smallest value attainable with 100 permutations (Figure 6). The model was therefore not the product of chance correlation.

Of the external compounds, 231 (92%) fell inside the applicability domain. Errors were larger outside the domain (MAE 0.59 vs 0.50; RMSE 0.73 vs 0.64), but the difference was not statistically significant with only 19 compounds outside (Mann–Whitney U = 1905, p = 0.170). Distance to the training set was weakly but significantly correlated with absolute error (Spearman ρ = 0.13, p = 0.037).

3.7 Model interpretation

Permutation importance on the external set ranked ΔTm, MM-GBSA, cLogP, docking score, HOMO–LUMO gap as the most influential descriptors (Figure 5). This ordering agrees with the generating model, in which potency depended on latent binding (best reflected by ΔTm and MM-GBSA), a lipophilicity optimum and an interaction with the HOMO–LUMO gap. Descriptors with no or negligible direct effect in the generating model (dipole moment, TPSA, HBD, HBA, rotatable bonds and MW) had importances close to zero. The importance of logS (0.002) was small, consistent with the assay artefact affecting only the least soluble compounds.

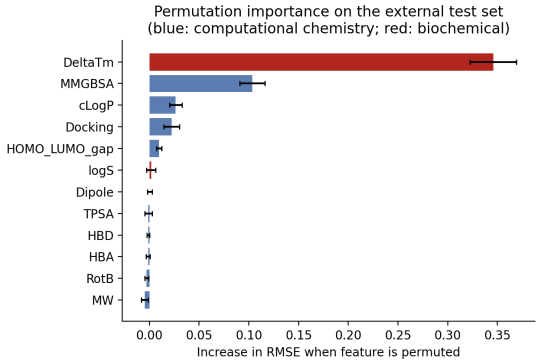


Figure 5: Permutation importance of descriptors for the integrated model on the external test set (mean ± SD of 30 repeats).

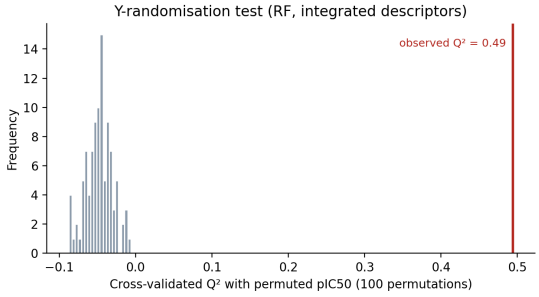


Figure 6: Y-randomisation test: distribution of cross-validated Q^2 for 100 models built on permuted pIC50 values, compared with the observed Q^2.

3.8 Identification of active compounds

The external set contained 62 actives (24.8%). The integrated model ranked actives best (AUC = 0.88), followed by ΔTm (0.87), MM-GBSA (0.86) and docking score (0.77) (Table 8; Figure 7). The paired bootstrap difference in AUC between the integrated model and docking alone was 0.11 (95% CI 0.05–0.17; $p < 0.001$). Early enrichment was also higher (EF5% 3.36 vs 2.69 for docking). Using a predicted pIC50 ≥ 6.5 as the classification threshold, the model had a sensitivity of 0.53, specificity of 0.93, precision of 0.72, balanced accuracy of 0.73 and MCC of 0.52. The low sensitivity reflects regression towards the mean in predictions near the threshold.

Table 8: Ranking of active compounds (pIC50 ≥ 6.5) in the external set

Score used for ranking	AUC	95% CI	EF5%
Docking score	0.77	0.71–0.83	2.69
MM-GBSA	0.86	0.81–0.91	3.02
ΔTm	0.87	0.82–0.91	3.36
RF (integrated)	0.88	0.83–0.92	3.36

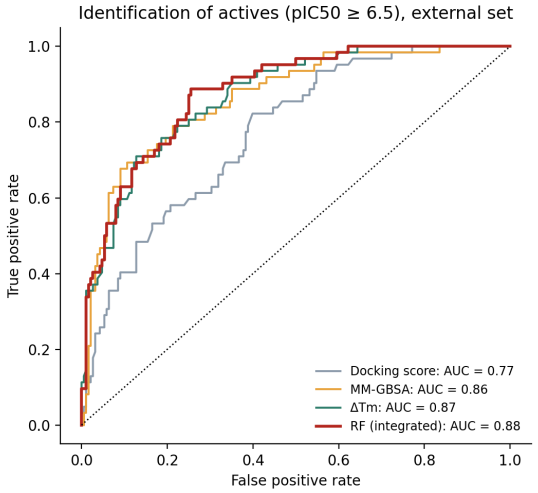


Figure 7: Receiver operating characteristic curves for identifying actives in the external set.

4. Discussion

4.1 Principal findings

This simulation study shows that integrating biochemical measurements with computational-chemistry descriptors produces consistent, statistically robust improvements in potency prediction. The gain over computational-chemistry descriptors alone was significant for every algorithm, with large effect sizes, and was confirmed on unseen chemical series. Against a known noise ceiling, the best model explained about 73% of the explainable variance, so a substantial part of the remaining error reflected irreducible assay noise rather than model deficiencies. Because assay noise of around 0.4–0.5 log units is typical of real potency data [11], external R^2 values much above 0.8 should be viewed with suspicion in comparable settings.

RF ranked best across the 50 cross-validation folds, but its practical advantage over the linear MLR and PLS models was small, and on the external set MLR performed almost identically. Flexible methods that depend on distances in descriptor space (kNN, SVR and MLP) did not benefit from adding weakly informative descriptors to the strong biochemical ones; for these algorithms, integrated and biochemical-only sets were statistically indistinguishable. This illustrates that greater model complexity or more descriptors do not guarantee better performance, and that algorithm choice

should be justified with appropriate statistics rather than single-split metrics [14].

4.2 Implications for rational drug design

Three practical lessons follow. First, orthogonal biochemical data such as thermal shift should be treated as informative descriptors in ML models, not only as filters, because they carry information on binding that scoring functions capture poorly [5,8]. Second, docking scores alone are limited for ranking potency; in this simulation, combining them with MM-GBSA, physicochemical and biochemical descriptors raised the AUC for active identification by 0.11, consistent with the known size bias of docking and the better ranking ability of end-point free-energy methods [5,6]. Third, model evaluation should use series-aware validation, report uncertainty, correct for multiple comparisons and include Y-randomisation and applicability-domain analysis, as recommended in QSAR best practice [4,12,25,26].

The interpretation analysis is also reassuring: permutation importance recovered the drivers of potency built into the generating model and assigned near-zero importance to irrelevant descriptors. With real data, interpretation cannot be checked against a known truth; simulation provides a way to verify that the chosen interpretation method behaves as expected before applying it to experimental datasets.

4.3 Strengths and limitations

The main strength of the study is the known ground truth, which allowed estimation of the noise ceiling and verification of model interpretation, combined with rigorous, series-grouped validation and non-parametric multiple-comparison statistics. All analyses are reproducible from the supplied code and seed.

The study also has important limitations. The data are simulated, so the absolute performance values depend on the assumptions of the generating model, including the strength of the link between each descriptor and binding, the noise levels and the functional form. Real structure–activity relationships include activity cliffs, conformational effects and target-specific interactions that this simple model does not represent. Descriptors were generated directly rather than computed from molecular structures, so the study does not assess molecular representations such as fingerprints or graph neural networks. Hyperparameters were fixed rather than tuned by nested cross-validation, which may have disadvantaged some algorithms. Only one simulated dataset was analysed; varying the data-generating parameters across many replicates would give a fuller picture of when integration helps most. The conclusions should therefore be regarded as methodological and confirmed with experimental datasets, for example public benchmark sets [29].

5. Conclusion

In a controlled simulation with known ground truth, integrating biochemical measurements with computational-chemistry descriptors significantly improved ML prediction of potency and the identification of active compounds compared with either source alone or with docking scores. RF performed best overall, but simpler linear models were close behind, and assay noise set a clear upper limit on achievable accuracy. Rigorous statistical comparison, series-aware validation and robustness checks are essential when integrating computational chemistry, biochemistry and machine learning for rational drug design.

References

1. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr. Computational methods in drug discovery. Pharmacol Rev. 2014;66(1):334-95.
2. Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303(5665):1813-8.
3. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77.
4. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR modeling: where have you been? Where are you going to? J Med Chem. 2014;57(12):4977-5010.
5. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, et al. A critical assessment of docking programs and scoring functions. J Med Chem. 2006;49(20):5912-31.
6. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10(5):449-61.
7. Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model. 2011;51(1):69-82.
8. Niesen FH, Berglund H, Vedadi M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nat Protoc. 2007;2(9):2212-21.
9. Shoichet BK. Screening in a spirit haunted by artifacts. Drug Discov Today. 2006;11(13-14):607-15.
10. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23(1-3):3-25.
11. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The experimental uncertainty of heterogeneous public Ki data. J Med Chem. 2012;55(11):5165-73.

12. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform. 2010;29(6-7):476-88.
13. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model. 2013;53(4):783-90.
14. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1-30.
15. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-102.
16. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta. 1986;185:1-17.
17. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-97.
18. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
19. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-232.
20. Lin LI. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255-68.
21. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
22. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65-70.
23. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc. 1937;32(200):675-701.
24. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn. 2003;52(3):239-81.
25. Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model. 2007;47(6):2345-57.
26. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. Molecules. 2012;17(5):4791-810.
27. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007;47(2):488-508.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.
29. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci. 2018;9(2):513-30.