RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2026, **16**, 21855

Distribution-preserved sampling (DPS) for smarter machine learning assisted ultra-large-scale virtual screening

Alexander Trachtenberg, Alexander Spelkov and Barak Akabayov *

Ultra-large-scale structure-based virtual screening (SBVS) for identifying novel bioactive compounds poses significant computational challenges. These challenges arise from the size of available chemical libraries, which can contain billions of molecules that require exhaustive docking and scoring, placing prohibitive demands on CPU/GPU resources. Small- and mid-sized laboratories often lack access to the high-performance computing clusters or cloud resources necessary to process such workloads in a timely manner. Furthermore, managing and analyzing the resulting terabytes of docking data requires robust data-handling pipelines and expertise that are not universally accessible. Here, we present a data-driven drug development pipeline that leverages a subset of molecules from a database with a common scaffold, reducing the chemical search space by tens to hundreds of orders of magnitude. In this case, the common scaffold that is the key to allowing this reduction is the 2-phenylthiazole moiety, identified through NMR fragment screening. We started with a subset of over 400 000 drug-sized 2-phenylthiazole-containing molecules selected from the zinc database and trained a random forest regression model on about 1% of this data to predict binding scores for the entire library. For this purpose, we used a distribution-preserving sampling approach based on KMeans clustering and binning, and we evaluated its statistical fidelity using KS, Wasserstein, JS, and KL divergence metrics. Our approach preserved the distribution of docking scores, demonstrating the utility of data-driven strategies for scalable virtual screening and establishing a benchmark dataset for machine learning in drug discovery.

Received 11th January 2026
Accepted 20th April 2026

DOI: 10.1039/d6ra00279j

rsc.li/rsc-advances

Introduction

The discovery and optimization of small-molecule drug candidates is a major challenge in drug development due to the vast chemical space,[1] as well as the high operational costs and low efficiency associated with high-throughput screening (HTS).[2] While HTS remains the dominant method for identifying initial hits, it is often constrained by low hit rates, limited scaffold diversity,[3] and reduced efficiency against challenging biological targets.[4]

Virtual screening complements HTS by enabling *in silico* exploration of vastly larger, more diverse chemical spaces, thereby enhancing the efficiency and scope of early-stage drug discovery. In virtual screening, a library of small molecules is evaluated *in silico* to determine the likelihood that one or more will bind to a biological target.[5] Virtual screens are categorized into ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS). In LBVS, screening is guided by the properties of known active compounds and is effective when high-quality activity data is available.[6] However, LBVS methods are often biased toward chemical structures similar to those of known ligands and do not provide information about binding poses.[7] In contrast, SBVS leverages the 3D structure of the target receptor to predict binding free energies and docking poses. SBVS is particularly valuable for revealing novel scaffolds, since it explores chemical space without the bias of known ligands.[8] Linking LBVS with SBVS, therefore, creates a powerful approach that combines the speed and pattern-recognition strengths of LBVS with the structure-level precision of SBVS. Such an integration enhances hit identification by utilizing both ligand features and target structural information, thereby improving accuracy and increasing the likelihood of discovering new, high-affinity binders.[7] However, even though the advantages of an integrated approach are clear, a key gap remains in systematically understanding how best to integrate and weight the outputs from LBVS and SBVS, particularly across diverse target classes and chemical spaces, which limits the generalizability and predictive power of current hybrid approaches.

Among the various techniques used in SBVS, molecular docking is the most widely used due to its computational efficiency. While docking provides a rapid estimate of ligand–receptor interactions, its scoring functions are often inaccurate,[9] leading to a high rate of false positives,[10] particularly in

Department of Chemistry and Data Science Research Center, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel. E-mail: akabayov@bgu.ac.il

screens involving millions of compounds. To address the accuracy problem, virtual screens are often conducted in multi-stage pipelines, in which an initial rapid docking screen evaluates millions of candidates, followed by application to the top-ranked compounds of increasingly accurate (and computationally expensive) methods, such as molecular mechanics/generalized Born surface area (MM/GBSA) rescoring or receptor flexibility modeling.[11] Yet as screening libraries grow to include billions of molecules, the initial stages of docking pose a significant computational challenge that demands extensive CPU and GPU resources.[12] Even when the necessary hardware is available, various factors such as storage and memory constraints, network bandwidth limitations for transferring large datasets, and the need for efficient parallelization create additional hurdles. Furthermore, managing and analyzing the resulting terabytes of docking data requires robust data-handling pipelines and specialized expertise. Consequently, this process is primarily limited to laboratories with access to multi-million-dollar computational facilities.

A promising addition to the virtual screening toolbox is machine learning (ML), which has been incorporated into SBVS workflows to enhance docking prioritization while lowering computational burden. In ML-enhanced SBVS, a model is trained on docking scores obtained from a subset of compounds and then used to predict scores for the remaining library, substantially reducing the number of explicit docking calculations required.[13,14] To ensure that the training subset adequately represents the underlying chemical space, we implemented a distribution-preserving sampling strategy. Conceptually, this approach is an adaptation of classical stratified sampling, designed to maintain the distributions of descriptors and scaffolds rather than introduce a fundamentally new sampling paradigm. What sets our framework apart is not just the sampling principle, but its integration within fragment-based virtual screening (FBVS). By combining distribution-aware subset selection with fragment-informed construction of chemical space, the model is trained on chemically meaningful regions. This synergy improves generalization across scaffold series, reduces bias toward

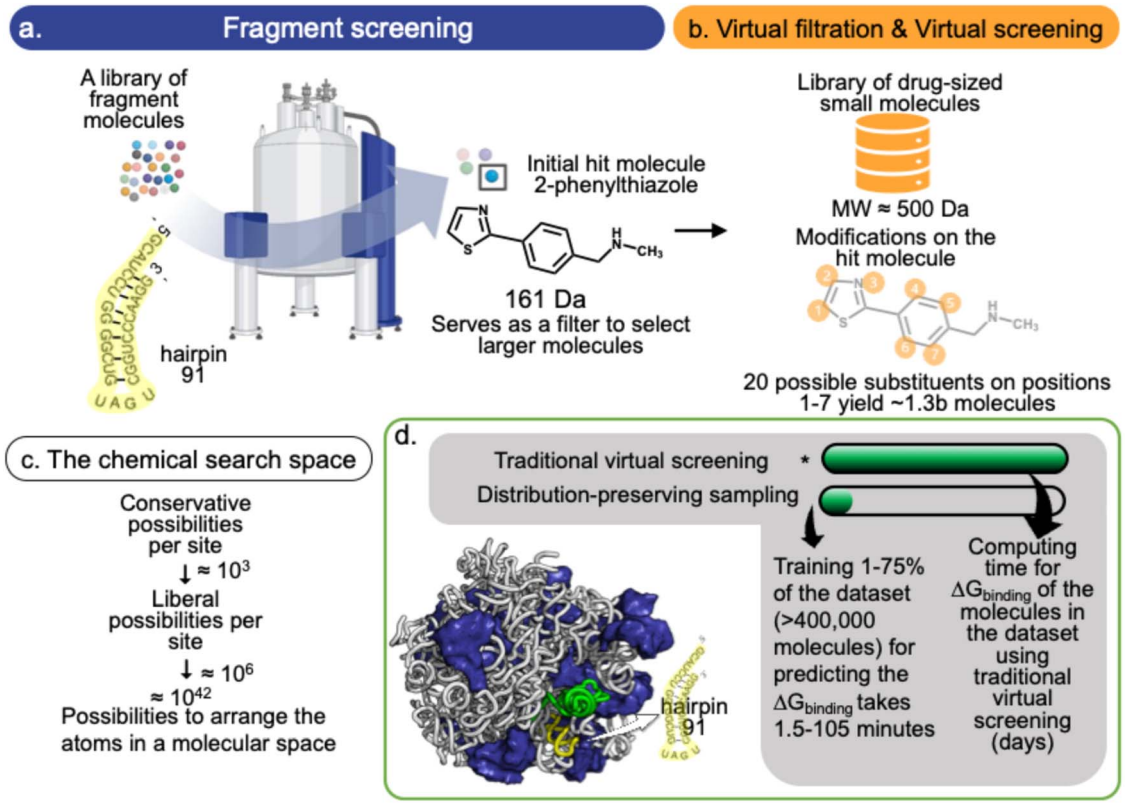

Fig. 1 Generating a subset of molecules with a common molecular scaffold, 2-phenylthiazole, as a strategy to decrease the chemical search space. (a) The scaffold molecule, 2-phenylthiazole, that binds to hairpin 91 was identified by NMR (T2 relaxation) screening from a commercial fragment library.[19] (b) The estimated search space for all possible combinations of 20 substituents across seven positions yielded (20^7) ~1.3 billion new drug-sized molecules. Filtering by molecular weight constraints (*e.g.*, 350 Da) and incorporating additional complexity such as stereochemistry (*e.g.*, three chiral centers) reduced the filtered search space to approximately 2.5 million molecules. (c) This targeted approach reduced the chemical search space from a fragment to a drug-sized molecule by approximately $(10^3)^7 = 21$ orders of magnitude, enabling us to focus on a manageable, chemically relevant subset of candidates. (d) Docking was performed using AutoDock-GPU on a workstation with an NVIDIA RTX 3080 GPU. The compound library, comprising approximately 400 000 compounds, was divided into batches of around 30 000 for concurrent processing. Each batch took around 3.3 days, but due to resource contention, the total time extended to about one week, with final times varying based on GPU utilization and thermal conditions. In contrast, distribution-preserved sampling was performed on a minute time scale for both training and testing on the same computational setup.

overrepresented chemotypes, and stabilizes predictive performance in ultra-large virtual screening (ULVS) campaigns, where maintaining chemical diversity is critical for reliable extrapolation. Notably, ULVS methods achieve high hit accuracy—identifying most true binders—while significantly reducing docking costs by several orders of magnitude. In two representative examples published at the beginning of this decade, application of Bayesian active learning to dock only ~2.4% of a 100-million compound library enabled the identification of ~95% of top hits,[15] and deep learning was used to prioritize 1000 high-quality leads from over a billion ZINC15 molecules without docking them all.[13,14] Recently, a reduction of over 1000-fold in docking calculations was reported, with minimal loss in hit recovery.[13] The above studies demonstrate that ML-guided virtual screening not only accelerates the screening process but also enables access to chemical spaces that would otherwise be computationally intractable. However, in many cases, docking remained supported solely by computational validation, requiring the need for experimental confirmation.

Sparsity—where only a small fraction of the vast chemical space contains active compounds—poses a major challenge in virtual screening. To overcome this, we implemented a two-step fragment-based virtual screening (FBVS)[16] strategy designed to systematically narrow the chemical search space and enhance the efficiency of identifying promising candidates (Fig. 1). In this strategy, we applied virtual filtration to obtain a dataset of molecules with a common scaffold, previously identified by NMR fragment screening, and grew the fragment molecules (~100 Da) into drug-sized molecules by using computational optimization. It may thus be said that FBVS identifies small, low-complexity molecular fragments that bind to a particular target, and then expands them into more complex scaffolds.[16] While FBVS is a promising and rational strategy for identifying small-molecule binders, especially for challenging targets such as RNA, it remains underutilized due to its dependence on expert-driven fragment expansion and the limited availability of fragment libraries.[17] In our case, a 2-phenylthiazole fragment hit, identified by NMR fragment screening, was expanded *via* FBVS into potent inhibitors of the bacterial ribosome.[18,19] In the current in-house study, the hit molecule, 2-phenylthiazole, served as the basis for creating a focused virtual library of over 400 000 derivatives, all sharing the same common molecular scaffold. Docking of the 400 000 molecules (the scaffold-based dataset) to the hairpin 91 RNA target was then performed on this library using the same parameters and docking protocol we had used previously.[19] Importantly, hairpin 91 is a 29-nucleotide RNA hairpin target within the functional core of the ribosome, the peptidyl transferase center (PTC, Fig. 1). Hairpin 91 is relatively small and has a simple 3D structure with defined secondary and tertiary interactions, making it an excellent model for evaluating ligand binding to folded RNA architectures[19]

This scaffold-based dataset (virtual library) provides a unique model that can be leveraged to demonstrate how a chemical constraint (*i.e.*, reduced structural diversity) facilitates efficient ML-based docking-score prediction. Our working hypothesis was that, due to the lower variance in molecular features across this scaffold-constrained chemical space, an ML model using only a small training subset—potentially as little as 1% of the entire dataset—could achieve high predictive accuracy. This approach offers not only computational advantages (reducing model training time by over 100-fold compared to large subsets) but also a practical solution for research groups lacking access to a high-performance computing infrastructure or cloud resources. Moreover, this approach provides a strategy that involves docking only a small subset of the library, with the scores for the remaining library predicted accurately using the trained model, thereby facilitating rapid hit prioritization. In addition to presenting our findings on the benefits of scaffold-based sampling, we introduce the underlying dataset as a potential benchmark for further exploring efficient docking-score prediction in realistic low-resource scenarios.

Methods

Datasets

Scaffold-based dataset. We compiled a library of 450 000 small molecules, all sharing a common phenylthiazole core scaffold, that were sourced from the ZINC15 database by using the filtering function embedded in the ZINC15 website (https://zinc15.docking.org/). These compounds were docked in 3D, by using AutoDock-GPU,[20] against a 29-nucleotide RNA hairpin target within the 23S rRNA of the crystal structure of the *Staphylococcus aureus* ribosome (PDB ID: 4WCE).[21] Importantly, hairpin 91 represents an experimentally resolved structured RNA structural motif within the functional core of the ribosome, the peptidyl transferase center (PTC).[21] Hairpin 91 is relatively small and has a simple 3D structure with defined secondary and tertiary interactions, making it an excellent model for evaluating ligand binding to folded RNA architectures.[19] Such motifs are increasingly being investigated as therapeutic targets because structured RNAs can form ligand-binding pockets analogous to those in proteins.

All ligand structures were prepared and stored in Tripos Mol2 format for compatibility with the docking software. Following docking and subsequent outlier removal (*e.g.*, extremely high docking scores), the final dataset comprised 413 109 molecules with their corresponding docking scores. All structures and docking results for this dataset are available in a public GitHub repository (https://github.com/csbarak/POC).

Benchmark dataset. As a diverse reference, we used the publicly available DOCKSTRING benchmark dataset,[22] which contains ~260 000 drug-like molecules docked against 58 pharmaceutically relevant protein targets (*e.g.*, various kinases and nuclear receptors), making it a versatile resource for benchmarking virtual screening workflows. For our analysis, we selected a target from the enzyme class, 11β-hydroxysteroid dehydrogenase type 1 (HSD11B1), and extracted all ligand-docking score pairs for that target. HSD11B1 is a well-characterized protein target widely used in virtual screening benchmarks due to its pharmacological relevance and extensive docking reference data.

To reduce heterogeneity and minimize the impact of extreme docking scores, we applied an interquartile range (IQR)-based filtering. Molecules with docking scores below Q1 − 2 × IQR or above Q3 + 2.5 × IQR were excluded, along with entries lacking docking values. Although IQR-based filtering removes statistical outliers to stabilize distributional comparisons, such filtering is not always applied in real-world screening pipelines, where extreme values may represent rare but genuine high-affinity hits. Here, filtering was used solely to standardize statistical analysis between datasets rather than to simulate a production virtual screening workflow. Following this filtering step, 255 085 molecules were retained, yielding a narrower, more representative dataset for distribution-preserving subset selection and predictive modeling.

The inclusion of both RNA and protein targets enables assessment of whether subset-selection strategies behave consistently across structurally distinct biomolecular classes.

Feature engineering. We computed a broad set of molecular descriptors for each compound in the above two datasets using the 2025.03.2 RDKit cheminformatics toolkit (https://github.com/rdkit/rdkit). In total, 217 molecular descriptors (covering physicochemical, topological, and structural properties) were computed using the standard descriptor set implemented in RDKit. Specifically, these are all the descriptors that were returned by the built-in descriptor list function in RDKit. This approach ensures comprehensive and reproducible feature coverage without manual feature selection or bias toward any predefined descriptor subset. A full list with brief explanations for each descriptor is provided in the SI. The following two filtering and one power transformation steps were then applied to refine the feature set: (1) variance filtering: descriptors with near-constant values were removed. Specifically, any descriptor with a variance below 0.001 (after min–max normalization) across the dataset was discarded as uninformative. (2) Correlation filtering: To reduce redundancy, we dropped one descriptor from any pair with a high pairwise Pearson correlation (>0.90), which is used as a threshold value to remove highly collinear features that do not contribute independent information. (3) Skewness correction: feature skewness was calculated for each descriptor, and variables with positive skewness >0.75 were log-transformed, provided all the descriptor's values were non-negative. This threshold targeted strongly right-skewed distributions that could distort scaling during normalization. The transformation reduced extreme tails while preserving rank relationships among values, improving numerical stability of subsequent preprocessing without materially altering descriptor relationships.

Feature selection *via* mutual information. We evaluated the predictive power of each remaining feature by computing its mutual information (MI) with respect to the docking score (the regression target). We then selected the top 50 descriptors with the highest MI scores for use in model training and analysis. This filter focuses the modeling on the most informative features. The method mutual_info_regression from the scikit-learn library (https://scikit-learn.org) was used for MI computation and feature selection. Fig. S1 and S2 show the top 50 descriptors selected and their MI scores for the scaffold-based and benchmark datasets, respectively.

Subset generation

To determine the amount of data necessary to maintain model performance and preserve data distributions, we generated progressively smaller subsets for each dataset, namely, our scaffold-based dataset of 450 000 molecules and the DOCKSTRING benchmark dataset, for the following fractions of the full data: 1%, 5%, 10%, 25%, 50%, and 75%. Two sampling strategies were employed to create these subsets: random sampling and distribution-preserving sampling. For the random sampling, we used the train_test_split method from the scikit-learn library to obtain independent and identically distributed (i.i.d.) samples for each fraction. This method preserves samples from the same probability distribution and provides an unbiased random subset of the desired size. Each subset was selected without replacement from the full dataset. For the distribution-preserving sampling, we employed a hybrid sampling approach to ensure that each subset preserved both the feature space distribution and the target (docking score) distribution of the original dataset. In other words, this "preserved" sampling aimed to maintain the characteristics common to the subset and the full dataset.

Distribution-preserving sampling strategy

To obtain the distribution-preserving subsets, we combined clustering in feature space with stratification in target space, in the following three steps: (1) clustering in feature space: using the selected 50 descriptors, we grouped molecules with similar descriptor profiles into 30 clusters *via* the KMeans method from the scikit-learn library (with n_clusters = 30). This initial clustering was selected to provide sufficient resolution across the 50 selected descriptors while avoiding over-fragmentation of the feature space. To assess the n_clusters parameter sensitivity, we compared several values in the range of 10–50 clusters. Across both datasets and different subset sizes, 30 clusters provided the best overall balance between preserving descriptor-level and target-level distributions, while higher values produced increasingly sparse cluster–bin strata and lower values yielded coarser chemical grouping. Moderate deviations from this value did not materially alter the overall divergence or model performance trends. (2) Binning in target space: Within each of the 30 clusters, we further binned the molecules based on their docking scores. We created 50 bins per cluster by using quantiles of the score distribution in that cluster. Importantly, 50 quantile-based bins were found to be an optimized value, balancing granularity and data per bin (deviations from this value did not materially alter divergence trends). This step thus stratified compounds by activity (docking score) level within chemically similar groups. (3) Proportional sampling: we then sampled molecules from each cluster-bin stratum proportionally to the prevalence of that stratum in the full dataset. In practice, for a given target fraction (*e.g.*, 10% of the data), we sampled ∼10% of the molecules from each bin of each cluster. This methodology ensured that the selected subset preserved the proportional representation of each chemical cluster and the full range of docking scores. By preserving both cluster membership and score quantile distribution, the subset

retained the joint distribution of chemical diversity and activity of the original dataset. This hybrid procedure ensured that the selected subset mirrored the structural diversity and activity distribution of the full dataset. In essence, the subset was a miniature version of the original data, with both feature and target distributions closely aligned.

Comparison metrics for target distribution

Because the RDKit descriptors encompass heterogeneous feature types (both continuous traits, such as molecular weight, which may be roughly normally distributed, and discrete counts, such as the number of rotatable bonds, which yield skewed or heavy-tailed distributions), no single metric is sufficient to capture all possible distribution shifts. Thus, to quantitatively assess how well the feature and target distributions were preserved in the generated subsets, we computed four complementary divergence metrics for each molecular descriptor (and for the docking score), where each metric captures a different aspect of the difference between two distributions.[23,24] The four metrics, namely, the Kolmogorov–Smirnov (KS) statistic, the Wasserstein distance, the Kullback–Leibler (KL) divergence, and the Jensen–Shannon (JS) divergence, may be described in brief as follows: (1) the KS statistic is a non-parametric measure that quantifies the maximum difference between the empirical cumulative distribution functions of two samples. It detects discrepancies in distribution location and shape without assuming a specific form, making it useful for various types of feature. (2) The Wasserstein distance measures the minimum cost of transforming one distribution into another. It is finite, even for non-overlapping distributions, and effectively captures differences in location (means) and scale (variances). (3) The KL divergence measures the amount of information lost when approximating the true distribution, based on the full dataset, with a distribution derived from a subset of that data. Although it is more sensitive to differences in the tails of the distributions and becomes undefined in cases of zero-probability mismatches, KL divergence remains a fundamental metric in information theory for assessing how two distributions differ from one another. (4) The JS divergence is a symmetric and smoothed version of the KL divergence that quantifies differences in probability mass allocation across bins. Its bounded nature makes it interpretable and robust, especially for comparing probability distributions with partial overlap. We used these four metrics to evaluate distributional similarity by comparing descriptor and docking score distributions in each subset to the full dataset. This assessment is crucial for ensuring that our distribution-preserving sampling creates statistically representative subsets, unlike random sampling, which can distort distributions.

Statistical summary of divergences

We summarized the above divergence metrics across features with the aim of comparing distribution preservation under different sampling strategies. We calculated the mean and standard deviation for each metric—KS, Wasserstein, KL, and JS—across the 50 selected descriptors. Lower mean values indicate better similarity between the subset and the full dataset, while higher standard deviations suggest poorer preservation for some features. We also used violin plots to visualize feature-wise divergence values, showing the distribution (median, IQR, and kernel density) of the metrics. These plots enable rapid comparison of how closely the resulting feature distributions produced by each strategy match the originals; narrower violins indicate smaller divergence across most features.

Comparison of random *vs.* preserved sampling

To directly evaluate the effectiveness of the distribution-preserving sampling strategy, we performed a paired comparison between each preserved subset and a correspondingly sized random subset: For each subset fraction (1%, 5%, 10%, 25%, 50%, 75%), we generated five independent random subsets of that size using different random seeds *via* train_test_split (preliminary tests showed that variance estimates stabilized with $\geq$5 replicates, and additional repeats produced negligible changes in mean divergence metrics while substantially increasing runtime). Each random subset of the specified size was drawn as a training set, and the remaining molecules (the complement of that subset in the full dataset) served as the test set. This approach (akin to repeated hold-out validation) allowed us to assess the variability of results from standard random sampling while still utilizing the entire dataset in each evaluation (training + test = 100% of data). We aggregated the divergence metrics for these five random trials to obtain an average and variance for how well random sampling preserves distributions for each subset size. For the distribution-preserving strategy, we constructed, for each fraction, a single preserved subset using the clustering + binning procedure described above. This subset was used as the training set, with the remainder of the data serving as the test set (analogous to the random subset case). We did not repeat preserved sampling multiple times for a given fraction, since the procedure is largely deterministic once clustering is fixed, as 50 quantile-based bins were found to be an optimized value and "greedy k-means++" algorithm for centroid initialization was chosen (giving the same initialization coordinates for the centroids, in each run). We then compared the distributional divergence metrics of preserved *vs.* random subsets of equal size. We plotted summary line charts for each divergence metric, showing how the mean divergence (averaged across features) changed as the subset fraction increased, for both sampling methods. These trend lines illustrate whether preserved subsets achieved lower divergence (closer to the full distribution) than random subsets, especially for small fractions. Additionally, we compared feature-wise divergence *via* violin plots for random *vs.* preserved subsets side by side for each fraction. This allowed us to observe not only the average behavior but also the spread of divergences: for example, whether preserved sampling dramatically reduces the tail of high-divergence features compared to random sampling. Collectively, these visual comparisons directly highlighted the advantages (if any) of the

distribution-preserving approach – relative to conventional random sampling – in maintaining the characteristics of the original data.

Evaluation of model performance

Finally, we assessed how well each subset (from both sampling strategies) supported predictive modeling of docking scores. Using the RandomForestRegressor method from the scikit-learn library with default values (100 trees in the forest), we trained a random forest regression model on each subset (on an Intel i7-1255U 1.70 GHz CPU) and then evaluated its performance at predicting docking scores on the remaining data. A Random Forest regressor was selected because it performs well on structured descriptor data with minimal preprocessing, is robust to multicollinearity, and requires relatively little hyperparameter tuning compared to deep learning models, which typically demand larger training sets and extensive optimization. In this study, default hyperparameters were intentionally used to evaluate whether the proposed subset-selection strategy improves predictive performance independently of model optimization. This strategy isolates the effect of sampling methodology rather than confounding it with architecture-specific tuning.

We used the following two stage evaluation setup, designed to simulate a fivefold cross-validation-like analysis for each dataset fraction: (1) random subsets: for each fraction f (1%, 5%,…, 75%), we used the five independent random training subsets (described above) as five different training sets. Each model was trained on a subset of size f and tested on the complementary portion (the remaining 99%, 95%, *etc.*, of the data not seen during training). We recorded the performance metrics for each of these five runs. (2) Preserved subset: for each fraction f, we trained a model on a single preserved subset of size f and tested it on the remaining data. (We used the same full test set as in the random sampling case for that fraction to enable fair comparison). Model performance was evaluated using root mean squared error (RMSE) and the coefficient of determination (R^2) on the test set, which provided measures of the magnitude of the prediction error and of the explained variance, respectively. We also recorded the training time (in minutes) for each model as an indicator of computational efficiency relative to the size of the training set. For the random sampling strategy, the five runs for each fraction allowed us to compute an average RMSE and R^2, as well as their variability (standard deviation), to account for the uncertainty due to random subset selection. We report these average performance metrics and compare them with the single-run performance of the distribution-preserving subset of the same size.

This experimental design ensured a fair, size-matched comparison between the two sampling methods. By always performing the evaluation on the full complement of data not used for training, we maintained a consistent evaluation set for a given fraction f. Comparing the accuracy of the random forest technique when trained on a preserved subset *vs.* on multiple random subsets revealed whether distribution-preserving sampling leads to better predictive performance (*i.e.*, lower RMSE or higher R^2) for a given training set size. In addition, tracking the model performance as the training fraction increased provided insight into how quickly each method approached the performance that would have been obtained by using the full dataset. The distributional fidelity of the subsets was evaluated along with their practical utility in modeling tasks. By examining the distributional divergence and model predictive performance side by side, we could determine whether preserving the data's intrinsic structure translated into tangible improvements in learning outcomes, compared to traditional random sampling.

Results

Comparison of docking score distributions across datasets

The distributions of the docking scores varied between the two datasets: the benchmark dataset showed a unimodal, approximately Gaussian, distribution, whereas the scaffold dataset revealed a bimodal distribution (Fig. 2). We note that the scaffold dataset contained about 1.6 times as many molecules as the benchmark dataset, providing a larger training pool for the random forest regression model. This greater volume of data was crucial for accurately capturing the complex bimodal distribution of the full dataset and for enhancing prediction accuracy for both molecular subpopulations in the scaffold dataset. Despite differences in the molecular diversity and the shape of the binding score distributions, both datasets are well-suited for analysis using the same pipeline outlined in the Methods section. The range and scale of docking scores were comparable across datasets, allowing consistent application of preprocessing, distribution-preserving or random sampling, and model training strategies. Additionally, both datasets used the widely used AutoDock software suite to compute docking

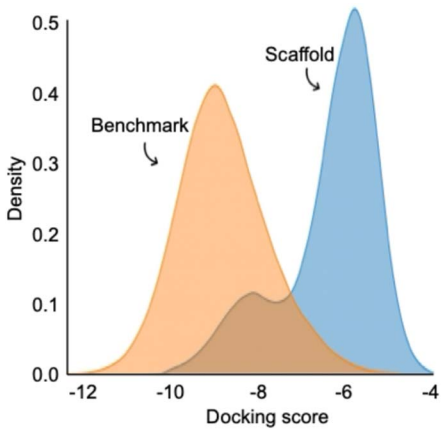

Fig. 2 Kernel density estimates (KDE) of docking score distributions for the scaffold and benchmark datasets. The benchmark dataset displays a unimodal, approximately Gaussian distribution, while the scaffold dataset shows a bimodal pattern, reflecting the presence of two distinct molecular subpopulations. The docking score ranges are similar for the two datasets, supporting their comparability within a unified analysis pipeline. Each KDE curve is independently normalized and reflects the relative frequency distribution within its respective dataset.

scores, specifically AutoDock-GPU for the scaffold dataset and AutoDock Vina[25] for the benchmark dataset, ensuring methodological coherence in score generation.

The bimodal distribution observed for the scaffold dataset may reflect two dominant ligand–target interaction regimes. In structured RNA systems, ligands can bind either within compact groove-like pockets or along exposed surface regions, resulting in distinct score distributions. Because in the scaffold dataset all ligand molecules share the same core scaffold, their binding poses may cluster into a limited number of interaction modes, which can manifest as separated score peaks. In contrast, the benchmark dataset contains structurally diverse molecules lacking a common scaffold, leading to a broader and more continuous distribution of docking scores.

Descriptor variance and dimensionality analysis between datasets

A comparative analysis was conducted to examine differences in the descriptors of the scaffold and benchmark datasets prior to feature selection (Fig. 3). After analyzing the distribution of docking scores (Fig. 2), the focus of the analysis shifted to the distribution of chemical features, derived from the RDKit. The variances of all common descriptors between the two datasets (210 out of 217 descriptors) were calculated (Fig. 3A). The log-scaled *y*-axis showed significant differences in feature variance between the two datasets. For the benchmark dataset, the descriptor variance reached 10^{27}, while that for the scaffold dataset reached only 10^{11}. This notable disparity of 16 orders of magnitude highlighted the much greater chemical diversity in the benchmark dataset. In contrast, the scaffold dataset, which consists of molecules sharing a common phenylthiazole core, exhibited a much more compact and homogeneous chemical feature space. A complementary perspective using principal component analysis (PCA) was used to explain the differences in variance between the two datasets (Fig. 3B). To explain 90% of the cumulative variance in descriptor space, 72 principal components were needed for the benchmark dataset, whereas only 51 components were required for the scaffold dataset. This difference further emphasized the lower intrinsic dimensionality and redundancy in the scaffold feature space. These results justify using the same modeling pipeline for both datasets and support the premise that, for chemically homogeneous datasets like scaffold, predictive models can be effectively trained on much smaller subsets without compromising performance. This method, which focuses on a dataset of scaffold-derived molecules, enables more efficient, cost-effective virtual screening.

Feature-wise distribution divergence

The divergence distributions between random and distribution-preserving subsets across all molecular descriptors were evaluated using four metrics, KS statistic, Wasserstein distance, JS divergence, and KL divergence. The results are displayed as split violin plots for each subset fraction (Fig. 4a and b). These plots reflect the variability in divergence values across all features and allow direct visual comparison between the sampling strategies.

For the KS statistic, the preserved subsets consistently showed markedly narrower distributions of divergence values than their randomly sampled counterparts across all subset sizes, even for the smallest fraction (1%). This observation held for both datasets and indicated that preserved subsets more reliably retain the cumulative structure of the original feature distributions.

For the Wasserstein distance, distributions of preserved subsets had similar widths to their randomly sampled counterparts across all subset sizes (except for the 25% subset in the scaffold-based dataset and the 1% and 10% subsets in the benchmark dataset, where randomly sampled subsets had narrower distributions than preserved). Nevertheless, the scale of the divergence values was notably different across the two

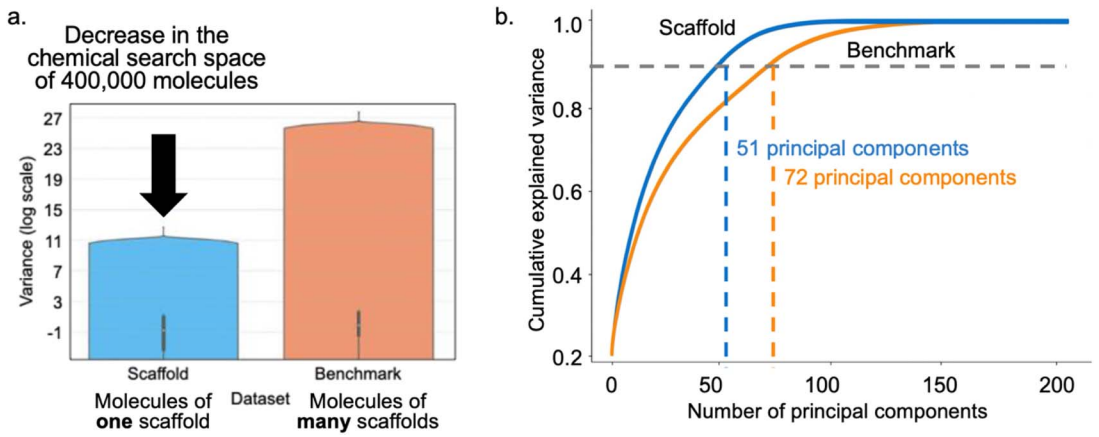

Fig. 3 Comparison of descriptor variances between the benchmark and scaffold datasets. (a) Variance distributions plots for extracted RDKit descriptors common to both datasets, shown on a logarithmic scale. The benchmark dataset exhibits substantially higher descriptor variability (up to 10^{27}) compared to the scaffold-based dataset (up to 10^{11}), reflecting its greater chemical diversity. (b) Cumulative explained variance from PCA. To capture 90% of the variance, 72 components are needed for the benchmark dataset, while only 51 are required for the scaffold-based dataset. These findings highlight the reduced intrinsic dimensionality of the scaffold-based dataset, consistent with its constrained chemical space due to a shared molecular scaffold.

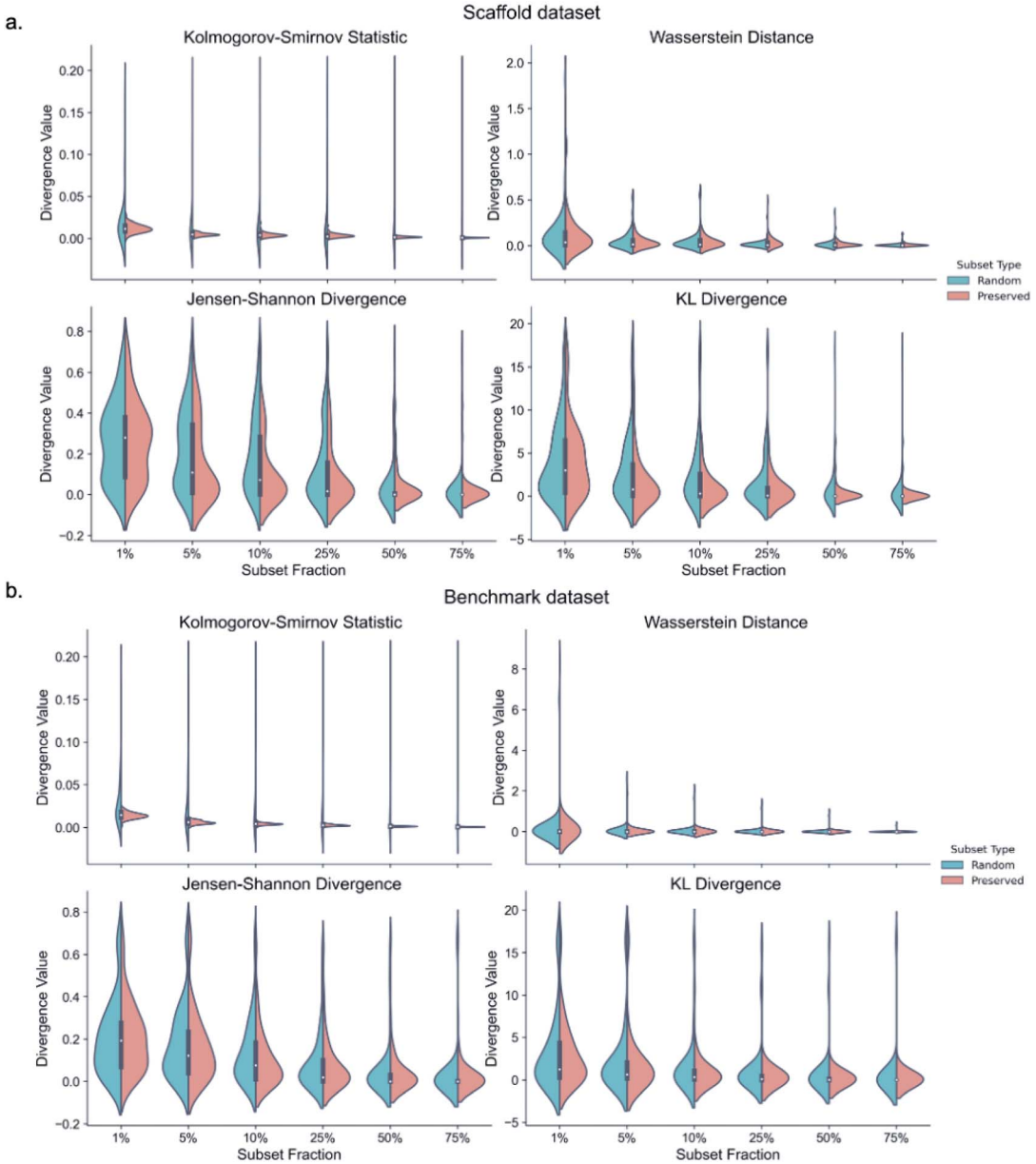

Fig. 4 Feature-wise distribution of divergence metrics between full dataset and sampled subsets. Split violin plots of the results for (a) the scaffold-based dataset and (b) the benchmark dataset showing the distribution of divergence values computed for each molecular descriptor between the full dataset and subset samples, across different subset fractions (1–75%). Results are presented for four divergence metrics: the Kolmogorov–Smirnov (KS) statistic, the Wasserstein distance, the Jensen–Shannon (JS) divergence, and the Kullback–Leibler (KL) divergence. In each plot, the left half of each violin corresponds to random sampling, and the right half, to distribution-preserving sampling. Narrower violins indicate more consistent (lower variance) divergence across features. Preserved subsets exhibit lower variability in divergence values, particularly for the KS and KL metrics, highlighting the effectiveness of the distribution-preserving sampling strategy in maintaining the statistical structure of molecular descriptors.

datasets: the *y*-axis range for the benchmark dataset was approximately four times larger than that of the scaffold-based dataset. This wider spread in the benchmark data indicated greater heterogeneity and less stability in feature distribution shifts across subsets, likely due to the chemical diversity of the molecules and the higher overall descriptor variance.

For JS divergence, the divergence distributions were overall quite similar between the two datasets. Nonetheless, for the scaffold-based dataset, random subsets tended to exhibit slightly more variability, especially at lower subset fractions. Starting from the 50% fraction, preserved subsets began to show a modest narrowing of the distribution, reinforcing the sampling method's effect at larger data sizes.

For the KL divergence, a more dataset-specific trend was evident. In the scaffold-based dataset, random subsets displayed somewhat broader divergence distributions than preserved subsets across most fractions. Notably, for fractions of 50% or higher, the preserved subsets became distinctly

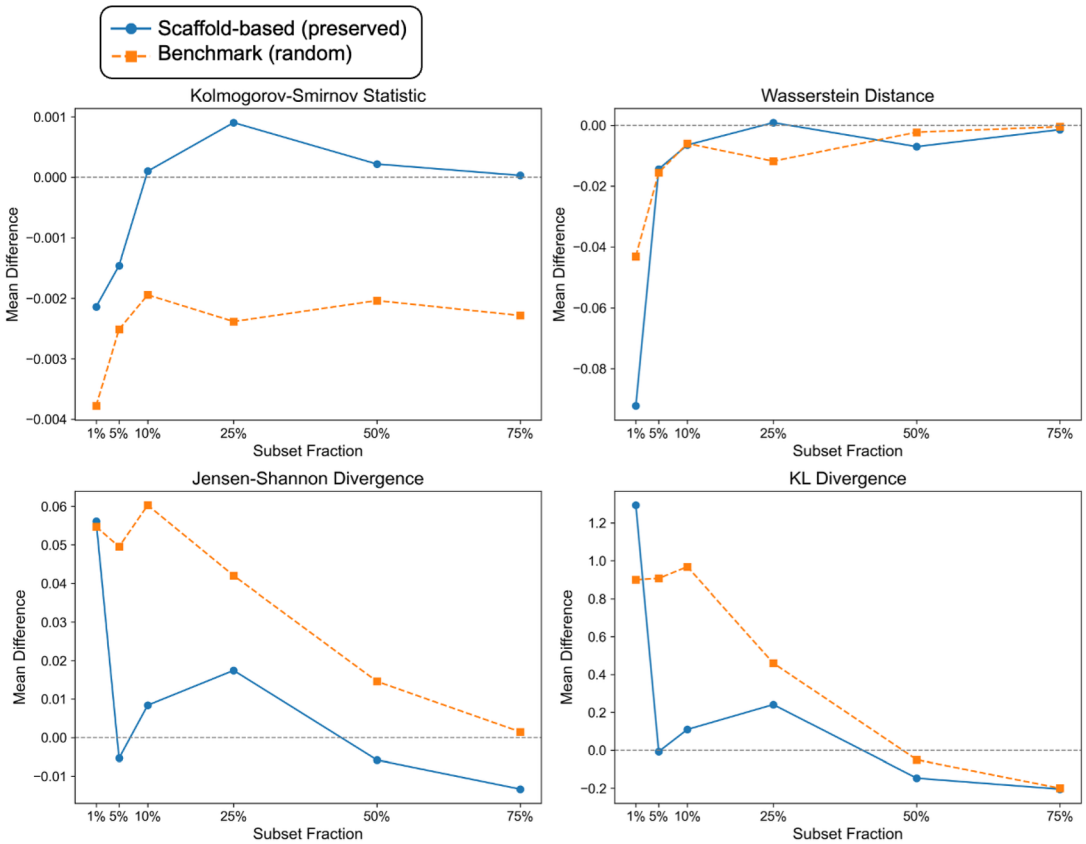

Fig. 5 Mean divergence differences between scaffold-based and benchmark datasets across four distribution comparison metrics. Each panel shows the difference in mean divergence values between the scaffold-based and benchmark datasets across subset fractions (1% to 75%), separately for the preserved and random sampling strategies. Positive values indicate that scaffold-based subsets diverge more from their full dataset than benchmark subsets, while negative values indicate the opposite. The KS statistic and Wasserstein distance showed mean divergence differences close to zero across most fractions and sampling methods, indicating broadly comparable representativeness between datasets, with minor fluctuations depending on the metric and subset size. For JS and KL divergences, patterns varied with subset size and sampling method: scaffold-based subsets diverged to a greater extent at most lower fractions, while benchmark subsets diverged at higher fractions. This analysis highlights how the dataset structure and sampling strategy influence the representativeness of the subset.

narrower. In contrast, for the benchmark dataset, the difference between the preserved and random subsets was less pronounced, possibly due to its greater chemical diversity and baseline variability in descriptor distributions.

Taken together, these results indicate that the distribution-preserving sampling method, based on KMeans clustering and target-value binning, effectively retained the statistical structure of the full dataset, particularly for the scaffold-based dataset. The narrower distributions of divergence values across multiple metrics and subset sizes provide strong evidence that the method worked as intended. This robustness in distributional preservation is a critical foundation for downstream ML tasks, as it ensures that the training data remains representative of the original chemical and biological space.

Comparing divergence differences across datasets

The mean divergence difference between the scaffold-based and benchmark datasets across the four divergence metrics and two sampling strategies (random and preserved) was determined (Fig. 5). Here, positive values indicate that scaffold subsets diverged to a greater extent from their full dataset than benchmark subsets, while negative values indicate the opposite—that benchmark subsets diverged to a greater extent from their full dataset than scaffold-based subsets. Across all four metrics, divergence differences were small in magnitude but nonetheless informative. Negative values observed in the mean difference calculated by the KS statistic for random sampling across all subsets indicated that benchmark subsets diverged to a greater extent than scaffold-based subsets. The trend of these differences was practically repeated for preserved sampling with positive or near-zero values for 10–75% fractions and negative values for the smaller subset sizes (1–5%). For Wasserstein distance, all values were slightly negative or near zero, indicating broadly similar divergence levels between subsets and their full datasets (except for small fluctuation observed at 1% and 25% fractions).

In the case of JS divergence, preserved scaffold-based subsets showed greater divergence than benchmark subsets at lower fractions (1–25%, except for 5%, which was slightly negative), but this trend was reversed at higher fractions (50–75%). For random subsets, the scaffold-based subsets diverged to

a greater extent for nearly all fractions, as indicated by having positive values. For preserved subsets, small fluctuations around zero, with both positive and negative values, were observed across subset sizes, indicating that the relative divergence between datasets depended on the fraction rather than following a single monotonic trend. For KL divergence, scaffold-based subsets initially diverged to a greater extent than benchmark subsets for small to moderate subset sizes (1–25%), but benchmark divergence exceeded scaffold-based divergence for larger fractions (50–75%). To summarize, mean differences between the two datasets were metric-dependent and typically small in magnitude, suggesting that neither dataset consistently exhibited greater divergence across all conditions. Importantly, these plots summarize mean divergence values; as shown in Fig. 4, feature-level distributions reveal that preserved sampling generally produces narrower divergence distributions across descriptors, supporting the idea that scaffold-constrained chemical space enables more stable and representative subsampling, especially at lower fractions. This observation supports the broader premise of the study that for chemically consistent datasets, such as the scaffold-based dataset, smaller training subsets can still maintain important distributional properties, potentially resulting in more accurate and computationally efficient modeling.

Subset size and random forest model performance

To evaluate how subset size influences model accuracy, we trained random forest regressors on progressively larger subsets (1%, 5%, 10%, 25%, 50%, 75%) of both the scaffold-based and benchmark datasets by using both random and distribution-preserving sampling methods. As expected, performance improved consistently with increasing subset size for both datasets (Fig. 6). Notably, for the scaffold-based dataset, the random forest model achieved high predictive accuracy, even when trained on only 1% of the data (~4100 molecules), yielding an R^2 of ~0.82 and an RMSE of ~0.48.

In practical virtual screening terms, RMSE reflects how accurately predicted docking scores approximate computed scores used for ranking compounds. Because screening decisions typically depend on relative ranking rather than absolute score values, modest RMSE values can still be sufficient for prioritizing top candidates, provided ranking consistency is preserved. These findings demonstrate that, within a chemically constrained scaffold-defined space, a small and well-sampled subset can be sufficient for accurate prediction of docking scores. In comparison, the benchmark dataset, which exhibits much higher structural diversity, required substantially larger training sets to achieve comparable performance.

A key advantage of using smaller subsets is the reduction in computational cost, the results of which are depicted in Table S1. For example, training a model on only 1% of the scaffold-based dataset required approximately 1.5 minutes, while training on 75% took over 105 minutes, representing a more than 70-fold speedup. Similarly, for the benchmark dataset, training time increased from just over 1 minute at 1% to more than 115 minutes at 75%, resulting in a more than 100-fold difference. This substantial improvement in computational efficiency underscores the value of intelligently selected small subsets in large-scale virtual screening pipelines.

Notably, model performance was generally comparable between random and preserved subsets, especially for the scaffold-based dataset. This finding can be attributed to the robustness of random forests, which are ensemble models that benefit from decorrelated decision trees and tend to perform well even on randomly drawn data, particularly when the underlying distribution is relatively homogeneous, as is the case in the scaffold-constrained dataset. Nonetheless, distribution-preserving sampling still offers advantages in maintaining diversity and reducing feature divergence, especially for very small fractions.

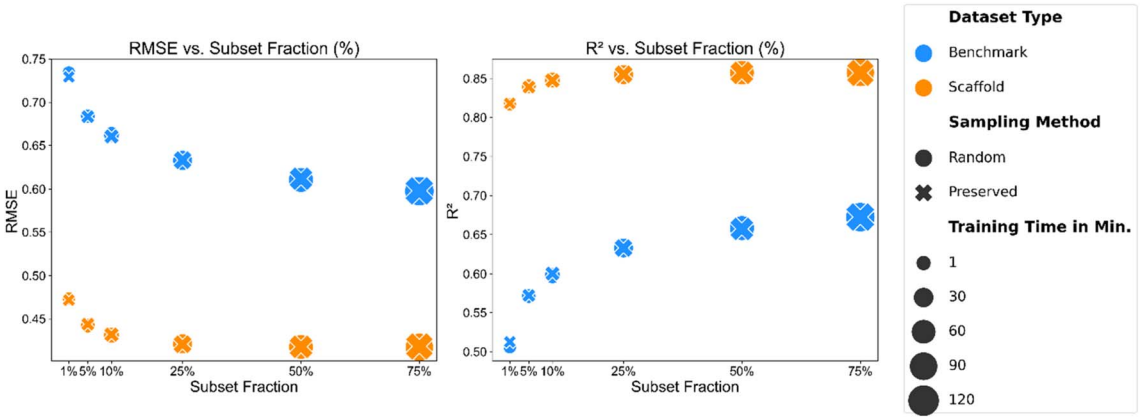

Fig. 6 Subset size *vs.* model performance for the scaffold-based and benchmark datasets. RMSE and R^2 scores for random forest models trained on increasing subset fractions (1% to 75%) of the scaffold-based and benchmark datasets. Each point represents the mean performance across simulated fivefold splits, with marker size proportional to training time in minutes. Both random and distribution-preserving ("Preserved") sampling strategies were evaluated. For the scaffold-based dataset, accurate predictions were achieved with as little as 1% of the data, demonstrating the efficiency of modeling within a constrained chemical space. In contrast, the benchmark dataset required significantly more data to reach comparable accuracy due to its structural diversity. The distribution-preserving strategy maintained or slightly improved model performance, particularly for lower subset sizes.

Discussion

The development of efficient computational strategies to explore large chemical libraries is a critical aspect of modern drug discovery. In the current study, we demonstrate that a scaffold-based approach combined with ML enables accurate prediction of docking scores when using only a small fraction of a full dataset, thereby substantially reducing computational costs.

Our pipeline centered on a focused library of over 400 000 molecules, all derived from a 2-phenylthiazole scaffold identified through FBVS and experimentally validated. The focused library was compared to a widely used benchmark dataset from DOCKSTRING, which was generated using the same docking software family (AutoDock Vina and AutoDock-GPU) and shared similar docking score ranges. Notably, the scaffold-based dataset exhibited markedly lower variance in RDKit molecular descriptors (Fig. 3). This reduction in feature variance is a direct consequence of the common scaffold architecture, which constrains chemical diversity and reduces the degrees of freedom in the feature space. These properties suggest that the scaffold-based dataset is particularly well suited for ML tasks, where lower variance may lead to simpler and more generalizable models.

To assess the representativeness of training subsets selected from the full dataset, we compared random sampling against a distribution-preserving strategy based on KMeans clustering and binning of the docking score target. Feature-wise divergence metrics (KS, Wasserstein, JS, KL) showed that the preserved subsets generally produced narrower distributions of divergence values across most subset sizes (Fig. 4), indicating that this method effectively maintains the original feature distribution. This conservation of the original feature distribution was particularly evident for smaller subset fractions (*e.g.*, 1–10%), which are the most relevant subsets for reducing computational effort. A comparison of mean divergence values between the two datasets (Fig. 5) suggested that benchmark subsets diverged more than scaffold-based subsets at small subset sizes, especially under random sampling, although differences between strategies and datasets were often minimal and occasionally reversed. These findings suggest that the scaffold-based dataset may be more amenable to accurate model training on minimal data, while also highlighting that preserved sampling does not universally outperform random sampling under all conditions and dataset configurations.

Model performance analysis (Fig. 6) further confirmed the utility of the scaffold-based dataset. A random forest regressor trained on just 1% of the scaffold-based dataset achieved an RMSE of ~0.47 and R^2 of ~0.82, with a training time of only 1.5 min. Notably, training on 75% of the data improved the RMSE only marginally (~0.42) but required 70 times more training time (~105 min).

In virtual screening workflows, predictive models are primarily used to prioritize top-scoring candidates rather than reproduce exact docking values. Therefore, acceptable prediction error depends on whether highly ranked compounds remain near the top of the list. While enrichment metrics were not explicitly evaluated here, the observed predictive accuracy and preserved score distributions suggest that the approach is suitable for shortlist generation prior to downstream validation. Importantly, these findings demonstrate the practical advantage of scaffold-constrained libraries, namely, accurate predictions can be obtained rapidly and with minimal resources. Interestingly, the performance of random and distribution-preserving splits was comparable for the two datasets. This finding may be attributed to the robustness of the random forest model, which sample features and instances during training and is less sensitive to skewed distributions than other models.

Notably, meaningful predictive accuracy was achieved without hyperparameter optimization for the Random Forest model, indicating that the observed performance gains arise primarily from dataset design and subset representativeness rather than model-specific tuning. This supports the general applicability of the proposed pipeline, since it does not depend on computationally intensive model selection or parameter searches. Specifically, our pipeline allows researchers to conduct docking calculations on a small subset of a scaffold-based library and then to use the resulting model to predict scores across the entire library. This advantage is particularly impactful for research groups lacking access to high-performance computing resources, as it offers a scalable alternative to exhaustive docking campaigns. Furthermore, the scaffold dataset introduced in this study may serve as a new community benchmark for evaluating sampling strategies and predictive models in scaffold-focused virtual screening. Overall, our findings highlight the synergistic benefits of combining FBVS-derived scaffold libraries with ML and distribution-preserving sampling. This approach offers a cost-effective pathway for exploring ultra-large chemical libraries, enhancing hit discovery, and democratizing access to structure-based virtual screening.

However, it must be noted that this study evaluated only a single scaffold (2-phenylthiazole), and therefore generalization to other scaffolds cannot be assumed. The magnitude of performance gains may depend on scaffold rigidity, substituent flexibility, or intrinsic descriptor variance, which can differ substantially across chemical classes. Because the scaffold was experimentally identified and generating large-scale additional scaffold-restricted datasets is resource-intensive, systematic validation across multiple scaffolds was beyond the scope of this work. Future studies should therefore examine diverse scaffolds using the same methodology to determine which structural families benefit most from subset-based learning and distribution-preserving sampling strategies. This can eventually lead to prospective applications, in which top-ranked compounds predicted by the model are prioritized for experimental validation *via* synthesis or purchase, followed by biochemical or biophysical assays. Such validation would provide a direct test of whether subset-trained models can successfully recover true binders while substantially reducing computational cost.

Author contributions

A. T. designed and executed the computational workflow and carried out the structural analyses. A. S. contributed to data preprocessing and assisted in the analysis. B. A. conceptualized the project, validated the analytical framework, and supervised the project. A. T. and B. A. wrote the manuscript. All authors contributed to the content and approved the final version.

Conflicts of interest

There are no conflicts to declare.

Data availability

Due to the large amount of data related to this paper, all structures, docking results, data analysis scripts, and all data described in this manuscript are available in a public GitHub repository (https://github.com/csbarak/POC).

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6ra00279j.

Acknowledgements

This work was supported by the Israel Science Foundation (ISF) to B. A. Grant number: 2414/25 and the Ministry of Science and Technology to B. A. Grant number: 8117002.

References

1 J. L. Reymond, *Acc. Chem. Res.*, 2015, **48**, 722–730.
2 G. Schneider, *Nat. Rev. Drug Discovery*, 2018, **17**, 97–113.
3 C. Gorgulla, *Annu. Rev. Biomed. Data Sci.*, 2023, **6**, 229–258.
4 R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer and G. S. Sittampalam, *Nat. Rev. Drug Discovery*, 2011, **10**, 188–195.
5 E. Lionta, G. Spyrou, D. K. Vassilatis and Z. Cournia, *Curr. Top. Med. Chem.*, 2014, **14**, 1923–1938.
6 A. J. Banegas-Luna, J. P. Ceron-Carrasco and H. Perez-Sanchez, *Future Med. Chem.*, 2018, **10**, 2641–2658.
7 J. Vazquez, M. Lopez, E. Gibert, E. Herrero and F. J. Luque, *Molecules*, 2020, **25**, 4723.
8 B. K. Shoichet, *Nature*, 2004, **432**, 862–865.
9 S. Y. Huang, S. Z. Grinter and X. Zou, *Phys. Chem. Chem. Phys.*, 2010, **12**, 12899–12908.
10 R. S. Ferreira, A. Simeonov, A. Jadhav, O. Eidam, B. T. Mott, M. J. Keiser, J. H. McKerrow, D. J. Maloney, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2010, **53**, 4891–4905.
11 J. R. Wallen, H. Zhang, C. Weis, W. Cui, B. M. Foster, C. M. W. Ho, M. Hammel, J. A. Tainer, M. L. Gross and T. Ellenberger, *Structure*, 2017, **25**, 157–166.
12 C. Gorgulla, A. Boeszoermenyi, Z. F. Wang, P. D. Fischer, P. W. Coote, K. M. Padmanabha Das, Y. S. Malets, D. S. Radchenko, Y. S. Moroz, D. A. Scott, K. Fackeldey, M. Hoffmann, I. Iavniuk, G. Wagner and H. Arthanari, *Nature*, 2020, **580**, 663–668.
13 A. Luttens, I. Cabeza de Vaca, L. Sparring, J. Brea, A. L. Martinez, N. A. Kahlous, D. S. Radchenko, Y. S. Moroz, M. I. Loza, U. Norinder and J. Carlsson, *Nat. Comput. Sci.*, 2025, **5**, 301–312.
14 F. Gentile, J. C. Yaacoub, J. Gleave, M. Fernandez, A. T. Ton, F. Ban, A. Stern and A. Cherkasov, *Nat. Protoc.*, 2022, **17**, 672–697.
15 D. E. Graff, E. I. Shakhnovich and C. W. Coley, *Chem. Sci.*, 2021, **12**, 7866–7881.
16 M. Singh, B. Tam and B. Akabayov, *Molecules*, 2018, **23**(2), 233.
17 D. A. Erlanson, R. S. McDowell and T. O'Brien, *J. Med. Chem.*, 2004, **47**, 3463–3482.
18 H. Grimberg, V. S. Tiwari, B. Tam, L. Gur-Arie, D. Gingold, L. Polachek and B. Akabayov, *J. Cheminf.*, 2022, **14**(1), 4.
19 B. Tam, D. Sherf, S. Cohen, S. A. Eisdorfer, M. Perez, A. Soffer, D. Vilenchik, S. R. Akabayov, G. Wagner and B. Akabayov, *Chem. Sci.*, 2019, **10**, 8764–8767.
20 D. Santos-Martins, L. Solis-Vasquez, A. F. Tillack, M. F. Sanner, A. Koch and S. Forli, *J. Chem. Theory Comput.*, 2021, **17**, 1060–1073.
21 Z. Eyal, D. Matzov, M. Krupkin, I. Wekselman, S. Paukner, E. Zimmerman, H. Rozenberg, A. Bashan and A. Yonath, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, E5805–E5814.
22 M. Garcia-Ortegon, G. N. C. Simm, A. J. Tripp, J. M. Hernandez-Lobato, A. Bender and S. Bacallado, *J. Chem. Inf. Model.*, 2022, **62**, 3486–3502.
23 L. Pardo, *Statistical Inference Based on Divergence Measures*, Chapman and Hall/CRC, 2018.
24 M. Vogt, A. M. Wassermann and J. Bajorath, *Information*, 2010, **1**, 60–73.
25 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.