Accelerating ligand discovery by combining Bayesian optimization with MMGBSA-based binding affinity calculations

Lucas Andersen,[†,§] Max Rausch-Dupont,[‡,§] Alejandro Martínez León,[†] Andrea Volkamer,[¶] Jochen S. Hub,[*,†] and Dietrich Klakow[*,‡]

†*Theoretical Physics and Center for Biophysics, Saarland University, PharmaScienceHub (PSH), 66123 Saarbrücken, Germany*

‡*Spoken Language Systems, Saarland Informatics Campus, Saarland University, PharmaScienceHub (PSH), 66123 Saarbrücken, Germany*

¶*Data Driven Drug Design, Center for Bioinformatics, Saarland Informatics Campus, Saarland University, PharmaScienceHub (PSH), 66123 Saarbrücken, Germany*

§*These authors contributed equally.*

E-mail: jochen.hub@uni-saarland.de; dietrich.klakow@lsv.uni-saarland.de

Abstract

Predicting protein–ligand binding affinity with high accuracy is critical in structure-based drug discovery. While docking methods offer computational efficiency, they often lack the precision required for reliable affinity ranking. In contrast, molecular dynamics (MD)-based approaches such as MMGBSA provide more accurate binding free energy estimates but are computationally intensive, limiting their scalability. To address this trade-off, we introduce an active learning framework that automates molecule selection for docking and MD simulations, replacing manual expert-driven decisions with a data-efficient, model-guided strategy. Our approach integrates fixed — partly pre-trained

deep learning — molecular embeddings (MolFormer, ChemBERTa-2, and Morgan fingerprints) with adaptive regression models (e.g. Bayesian Ridge and Random Forest) to iteratively improve binding affinity predictions. We evaluate this approach retrospectively on a new dataset of 60,000 chemically diverse compounds from ZINC-22 targeting the MCL1 protein using both AutoDock Vina and MMGBSA. Our results show that incorporating MMGBSA scores into the active learning loop significantly enhances performance, recovering 79.9% of the top 1% binders in the whole dataset, compared to only 6.7% when using docking scores alone. Notably, MMGBSA exhibits a stronger correlation with experimental binding affinities than AutoDock Vina on our dataset and enables more accurate ranking of candidate compounds in a runtime efficient way. Furthermore, we demonstrate that a one-at-a-time acquisition active learning strategy consistently outperforms traditional batched acquisition, the latter achieving just 78.4% recovery with MolFormer and Bayesian Ridge. These findings underscore the potential of integrating deep learning-based molecular representations with MD-level accuracy in an active learning framework, offering a scalable and efficient path to accelerate virtual screening and improve hit identification in drug discovery.

Introduction

Drug discovery is a lengthy and costly process, requiring approximately 12 years and an investment of approximately $1.8 billion USD.[1] This process encompasses several challenges along the drug development pipeline, from target identification over hit compound finding, to hit-to-lead optimization and ultimately to (pre)clinical trials. Once a target is identified, promising (bio)active compounds are searched through experimental high-throughput screening (HTS) or its computational counterpart, virtual screening (VS). Although VS is less accurate than HTS, it allows for the rapid evaluation of an extensive number of compounds, thereby sampling a greater portion of the vast chemical space, which is estimated to contain around 10^{60} possible compounds.[2]

To facilitate the identification of promising compounds, several extensive screening libraries are available. The ZINC22[3] database contains over 50 billion chemical compounds that can be ordered from commercial suppliers. Another example is Enamine's Real database[4] featuring building blocks that can be combined to 70 billion compounds, which can be synthesized on demand with a success rate of 80% within two weeks.

Different computational methods have been developed to search the vast chemical space, which vary in their computational cost and in their accuracy in correctly ranking ligands by binding affinity. Molecular docking has been used with the aim to predict the optimal binding conformation of ligands within a protein binding pocket by minimizing an affinity scoring function.[5,6] Docking is computationally efficient and has been used extensively as an enrichment tool to prioritize ligands that are more likely to bind to a target. However, docking protocols typically neglect the flexibility of the protein and rely on highly simplified models for protein-ligand interactions and solvent contributions, which may lead to poor correlations between docking scores and experimental binding affinities.[7] In contrast, methods based on molecular dynamics (MD) simulations can account for protein flexibility and explicit solvent effects, but they come at a significantly higher computational cost compared to docking. Among MD-based methods, free energy perturbation (FEP) may be considered as a gold standard for affinity predictions as it relies on a rigorous statistical framework along an alchemical binding pathway and treats the solvent explicitly throughout the simulations and affinity calculations. Simplified MD-based approaches are given by end-point free energy techniques such as the Molecular Mechanics Generalized-Born Surface Area (MMGBSA) method. MMGBSA involves explicit-solvent simulations of both, the bound and the unbound state, however it approximates the solvent and entropy contributions to the binding affinity with implicit models.[8–10] In this study, we score ligands using MMGBSA, as it provides superior ranking accuracy as compared to docking-based scoring functions such as AutoDock Vina,[11] while being computationally more efficient than FEP.

Active learning approaches, such as Bayesian Optimization,[12] can build upon binding

affinities from MD-based methods such as MMGBSA, enabling binding affinity predictions with MD-level accuracy while maintaining scalability for large-scale VS applications. In active learning, binding affinity predictions are performed iteratively, thereby avoiding the calculation of binding affinities for all available molecules, and instead allowing the algorithm to focus on promising regions of the chemical space. During each iteration, a surrogate model selects a subset of molecules for simulation. The molecules are selected such that the likelihood of finding the molecule with the lowest binding free energy in the entire database is maximized. After each iteration, the surrogate model is updated with the newly computed binding affinities obtained from the simulation, thereby refining its predictions for subsequent rounds.

While the optimization of lead candidates using generative models has gained significant attention in recent years,[13–16] active learning[12] has proven highly effective for screening large datasets in drug design ampaigns.[17–23] Graff *et al.*[21] analyzed several surrogate models in a pool-based active learning setting and found that the best performance is achieved when using multi-layer perceptrons (MLP) and message-passing neural networks[24] (MPNN) as surrogate models. However, training such surrogate models is time-consuming. Cao *et al.*[25] proposed a method in which the entire model, including the pretrained components, is updated during each training iteration. While continuing to train a pretrained model can enhance performance, it is computationally expensive and requires careful handling to prevent the new updates from overwriting previously learned knowledge.[26]

In this work, we demonstrate how combining Bayesian active learning with binding affinities from MD-based MMGBSA calculations can both accelerate the drug discovery process and bring MD-level accuracy to virtual screening applications. To this end, we compiled a large data set of binding affinities for 60.000 chemically diverse compounds from the ZINC-22 database, targeting the binding pocket of the myeloid cell leukemia 1 (MCL1) protein. MCL1 is a promising target for anti-cancer therapies that has been used previously to benchmark binding affinity calculations.

We show that by using embeddings from a pretrained chemical language model without further finetuning, the surrogate model can be a simple model such as ordinary linear regression while maintaining the high level of performance. Additionally, retraining the model speeds up the active learning process during the early stage of learning. Specifically, we find that using a pretrained embedding model does not require retraining the full model, instead updating a simple linear regression model after each iteration yields the same retrieval of top binders and also offers the possibility of one-a-at-a-time molecule acquisition. Using our active learning pipeline, we obtain ∼80% of the top-1% binders according to MMGBSA after querying 6% of the 60.000 compounds. In addition, we find that learning docking scores is easier than learning MMGBSA scores, as shown by retrieving even 97% of the top-1% binders according to docking scores after querying 6% of the compounds. However, since MMGBSA scores exhibit a much stronger correlation with experimental binding affinities compared to docking scores, our results suggest that the combination of active learning with MMGBSA scoring provides a compelling balance between computational efficiency and predictive accuracy for identifying promising ligand candidates.

Computational Background

Binding affinity calculations with MMGBSA

MMGBSA is an end-point free energy method that estimates binding affinities based on MD simulations of only the bound and unbound states. MMGBSA employs a thermodynamic cycle to estimate the binding affinity ΔG (Fig. 1c), leading to the definition:

$$\Delta G = \Delta G^{\mathrm{PL}}_{\mathrm{solvation}} - \Delta G^{\mathrm{P}}_{\mathrm{solvation}} - \Delta G^{\mathrm{L}}_{\mathrm{solvation}} + \Delta G_{\mathrm{gas}}$$

Here, $\Delta G^{\mathrm{PL}}_{\mathrm{solvation}}$, $\Delta G^{\mathrm{P}}_{\mathrm{solvation}}$, and $\Delta G^{\mathrm{L}}_{\mathrm{solvation}}$ denote the solvation free energies of the protein-ligand complex, the protein, and the ligand, respectively, computed with the Gen-

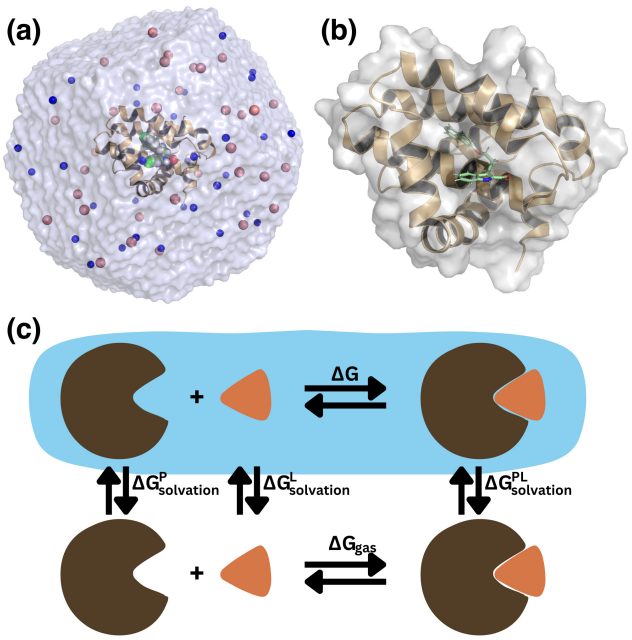

Figure 1: (a) MD simulations system of MCL1–ligand complex solvated in explicit solvent. The protein is shown as cartoon representation, the ligand as spheres, water as transparent surface, and NaCl ions as blue/red spheres. (b) Close-up view on the protein with bound ligand. (c) Thermodynamic cycle used for computing the binding free energy ΔG for a ligand (orange) binding to a protein (brown) using MMGBSA. The blue area illustrates states solvated in water. For definition of the the mathematical symbols, see text.

eralized Born (GB) method or the Poisson-Boltzmann (PB) method,[27] together with the solvent-accessible surface area (SA). These methods operate within the framework of implicit solvent models, simplifying the representation of solvation effects.

ΔG_{gas} denotes the binding affinity in the gas phase, which is computed as the difference in expected values of potential energy, derived from a molecular mechanics force field, between the products (protein-ligand complex) and the reactants (protein and ligand). This term may optionally include an entropic correction.[28,29] While the MD simulations are typically conducted in explicit solvent (Fig. 1a) to ensure realistic conformational sampling, the solvation free energy terms ($\Delta G^{\text{X}}_{\text{solvation}}$) are estimated using implicit solvation models (GB or PB combined with SA). Although absolute binding free energy values from MMGBSA often deviate significantly from experimental binding affinities, the method has been shown to provide reasonably accurate ranking of ligands binding to the same target protein.[30–32]

Bayesian Optimization

Bayesian Optimization (BO)[12] is a technique for efficiently identifying the global optimum of a computationally expensive black-box function f, where neither gradient information nor an analytical expression is available. Given a dataset of labeled inputs $\mathcal{D}$, BO mitigates the cost of labeling the entire domain by constructing a probabilistic surrogate model $P(\hat{f}|\mathcal{D})$ of the objective function f, which subsequently evaluates the utility of labeling each unlabeled data point. Since the objective is to identify the optimum of f, the utility is typically defined as a measure of improvement relative to the currently labeled data points. Gaussian Processes (GP)[33] have traditionally been a popular choice for surrogate modeling, while Bayesian Neural Networks (BNN)[34,35] have emerged as a more recent alternative.

Surrogate models are typically Bayesian for two main reasons. First, in theory, Bayes' theorem allows efficient model updates when new labels are acquired, eliminating the need for costly full-model retraining. Second, Bayesian models provide access to the full posterior distribution, enabling more robust selection of the next evaluation point by considering, for example, the expected prediction rather than only the most likely prediction.

The utility of a new point can be assessed by acquisition functions such as *expected improvement* $\mathrm{EI}(x)$.[36] This criterion assigns zero utility to points that are not expected to improve upon the best observed value, $f^\star$, while assigning the expected improvement to all other points. In the case where a looking for a minimizer of f, EI is given by

$$\mathrm{EI}(x) = \mathbb{E}\left[[\![\hat{f}(x) < f^\star]\!] \cdot (f^\star - \hat{f}(x))\right], \tag{1}$$

where $[\![\cdot]\!]$ denotes the Iverson bracket, which evaluates to 1 if its argument is true and otherwise to 0. The expectation is computed over all surrogate models $\hat{f}$ that have a non-zero posterior probability.

Assuming a Gaussian posterior distribution, the expected improvement permits the fol-

lowing closed-form expression:

$$\mathrm{EI}(x) = \sigma^2(x)\left[\delta \cdot \Phi(\delta) + \phi(\delta)\right] \tag{2}$$

$$\delta = \frac{f^\star - \mu(x)}{\sigma^2(x)} \tag{3}$$

Here, Φ and ϕ denote the cumulative distribution and probability density function of the standard normal distribution, respectively, and μ, σ denote the mean and standard deviation of the posterior distribution at point x. Essentially, the first term assigns high values to points that are likely to improve (exploit), while the second term assigns higher values to points associated with a large uncertainty (explore). Thus, EI automatically balances exploitation and exploration.

Bayesian Optimization with foundation models

In recent years, self-supervised pretraining of deep neural networks, such as transformer models,[37] has proven effective for learning rich input representations, particularly in fields like natural language processing,[38,39] computer vision,[40–43] and cheminformatics.[44–48] Additionally, these pretrained models are highly effective for transfer learning, as their parameters already capture general data features, allowing them to adapt to downstream tasks with few labeled examples.[48–51]

The main difficulty in adapting such models to BO is that they are not Bayesian by default. Simple approaches, such as Monte Carlo Dropout,[52] cannot be easily applied to pretrained models that were not originally trained with dropout, as these models lack the necessary dropout layers to enable uncertainty estimation. The uncertainty estimates are used to compute the expected improvement (Equation 2), guiding the selection of ligands with high informational value.

Dataset

To simulate a realistic early-stage drug discovery scenario and evaluate our active learning approach, we construct a dataset as a representative subset of a large molecular library, such as ZINC.[3] Starting from a few known binders, a screening set is then assembled by selecting molecules from the database that are structurally similar to the initial candidate, enabling focused exploration while conserving computational resources.

Protein target: As a case study, we select the MCL1 protein, which has been shown being a promising target for anti-cancer therapy.[53] The crystal structure was taken from the protein data bank (PDB code 4HW3).[54] Friberg *et al.*[54] provide experimental measurements of the inhibitory constant for 41 MCL1 binders.

Small molecule data set: The ZINC22 compound collection, comprising more than 54 billion molecules, serves as screening library in our study.[55] We compare each of these 54 billion molecules with the strongest experimental binder according to Friberg *et al.*[54] using a Tanimoto similarity coefficient. We measure similarity using the maximum common substructure (MCS) metric.[56] We leverage RDKit[57] to calculate the MCS and include all molecules above a similarity threshold of 0.4. Afterwards, to obtain an initial pose for the MMGBSA computation we dock all molecules using AutoDock Vina.[58] For each ligand, we carry out MD simulations to obtain the binding free energy using MMGBSA (see Methods). The final size of our dataset comprises $59,356$.

Figures 2a/b show the correlation of the experimental binding affinities with MMGBSA and AutoDock Vina for the set of experimentally known binders. Evidently, the Vina score exhibits poor Pearson correlation with the experimental binding affinities of 0.177 and poorly ranks the ligands as shown by a Spearman correlation coefficient of 0.174 (Fig. 2a). MMGBSA greatly overestimates the absolute binding affinity as common for MMGBSA calculations, however it reveals greatly improved correlation with the experimental

data shown by Pearson and Spearman correlations of 0.610 and 0.607, respectively (Fig. 2b). Using MMPBSA instead of MMGBSA further improves the correlation with experimental data to 0.689 and 0.684, respectively (Fig. S4). These findings supports our assumption that the computationally more demanding MMGBSA method outperforms AutoDock Vina for ranking of ligands. Notably, among the 59,356 compounds of our augmented data set, docking and MMGBSA scores correlate poorly (Fig. 2c), demonstrating that the MMGBSA and Vina scores are sensitive to different molecular properties. The orange and pink areas of Fig. 2c highlight the top 1% of molecules based on the MMGBSA score and the Vina score, respectively, demonstrating that the set of top-1% binders according to MMGBSA and Vina are nearly non-overlapping. Only 1.6% of the top-1% binders according to MMGBSA also belong to the top-1% binders according to the Vina score. Hence, Vina scores do not identify the top binders as determined by MMGBSA. Taken together, MMGBSA shows a stronger correlation with experimental binding data and therefore provides more realistic binding affinity estimates.

Datasets in the literature

To compare our active learning pipeline with previous work, we benchmark in addition against the dataset utilized by Graff *et al.*[21] and Cao *et al.*[25]. The dataset comprises docking scores for the 50,240 molecules taken from the Enamine's Discovery Diversity Collection (`Enamine50k`), docked against thymidylate kinase using AutoDock Vina, as described by Graff *et al.*[21]. In contrast to our work, these previous studies did not consider MMGBSA scores.

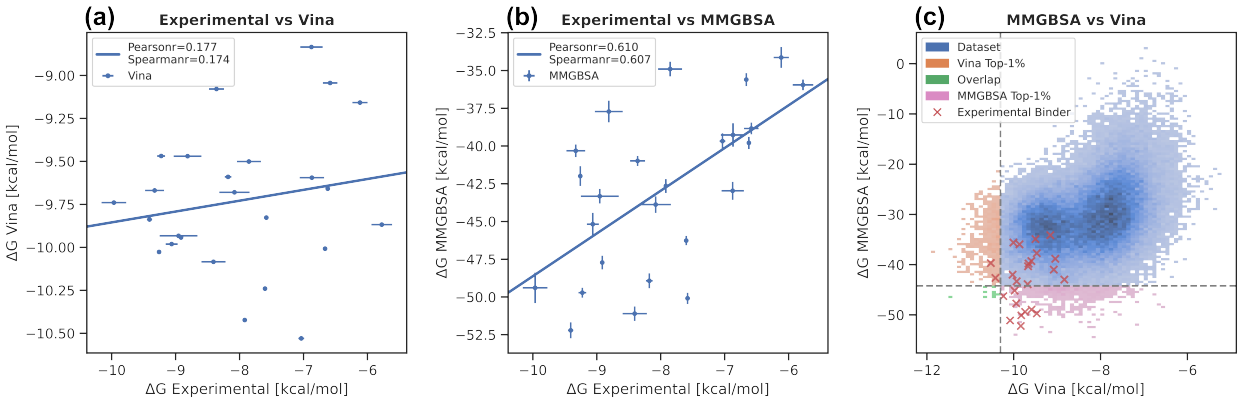

Figure 2: Comparison of binding affinities from MMGBSA or AutoDock Vina with experimental data. (a) Correlation between experimental binding affinity and AutoDock Vina scores for the set of experimentally known binders. Each ligand was docked inside of the protein with the search box spanning the MCL1 binding pocket. The RMSD values computed from the maximum common substructure between each ligand and the reference ligand of the crystal structure are provided in the Supplementary Data. (b) Correlation between experimental and MMGBSA-based binding affinity for the set of experimentally known binders. The docked poses as identified for (a) were used as starting structures for the MD simulations associated with the MMGBSA computation. (c) Correlation between AutoDock Vina and MMGBSA for the augmented dataset highlighting the top 1% of molecules according to the respective computational method. The red crosses indicate the computed score for the experimental binders. Each ligand was docked following the same procedure as in (a).

Methods

Binding free energy calculations using MMGBSA

Protein-ligand complex generation: The crystal structure of MCL1 was taken from the Protein Data Bank (PDB number 4HW3[54]). Starting with the SMILES representation of the ligand, RDKit[57] and EasyDock[59] were used to obtain the ligand in PDBQT format. The ligand was protonated using Dimorphite-DL[60] at pH 7.4. The ligand was docked into the MCL1 binding pocket with AutoDock Vina,[5] using a cubic box centered at the reference ligand location at $(67.9, -32.6, 27.2)$ Å and using a box length of 20 Å. In case that Dimorphite-DL proposed more than one protonation state, the state with strongest affinity according to the Vina score was used for follow-up simulations. An example conformation of a ligand in the MCL1 pocket is shown in Fig. 1a/b.

MD simulation: The protein-ligand complex was placed in an octahedral simulation box using a distance of at last 1.5 nm between the protein and the box surface. The system was solvated with water and neutralized with 150 mM of NaCl. Protein interactions were described by the Amber99sb-ildn force field[61] and the TIP3P water model was applied.[62] Ligand interactions were described with the Open Force Field, version 2.0.0.[63] Ions were described with the Amber parameters.[61] The ligand topology file was generated with TOFF.[64] Hydrogen mass repartitioning was used with a hydrogen mass factor of 3. Genheden and Ryde[27] recommend an equilibration time of 100 to 200 ps for MMGBSA calculations. In this study, we adopted a multi-step approach with a total equilibration time of 107.5 ps. Following energy minimization, the system was equilibrated through the following steps: (i) NVT-equilibration with an integration time step of $\Delta t = 2$ fs for 10 ps; (ii) NVT-equilibration with $\Delta t = 3$ fs for 15 ps; (iii) NPT-equilibration with $\Delta t = 3$ fs for 22.5 ps; (iv) NPT-equilibration with $\Delta t = 4$ fs for 60 ps. Position restraints on the heavy atoms using a force constant of 2500 kJ mol^{-1}nm^{-2} were activated from steps (i) to (iii), and removed for subsequent steps. Dispersive interactions and short-range repulsion were describe with a Lennard-Jones potential with a cutoff at 1 nm. Electrostatic interactions were computed with the particle-mesh Ewald (PME) method using a real-space cutoff at 1 nm.[65,66] Bond lengths involving hydrogen atoms were constrained with LINCS.[67] The pressure was controlled using the Parrinello-Rahman barostat with a time constant of 2.0 ps.[68] The temperature was controlled at 298.15 K using the Langevin thermostat with an inverse friction constant of 2.0 ps. During production simulations, a time step of 4 fs was used.

MMGBSA calculations: After equilibration, the system was simulated for 950ps. Frames were extracted every 50 ps, yielding 20 starting conformations for 20 production runs of 100 ps each. Both the 950 ps simulation and the 100 ps production runs were conducted under the same conditions as step (iv). MMGBSA-based binding affinities were computed from 20 frames extracted from each production run using gmx_MMPBSA.[28,69] The final binding

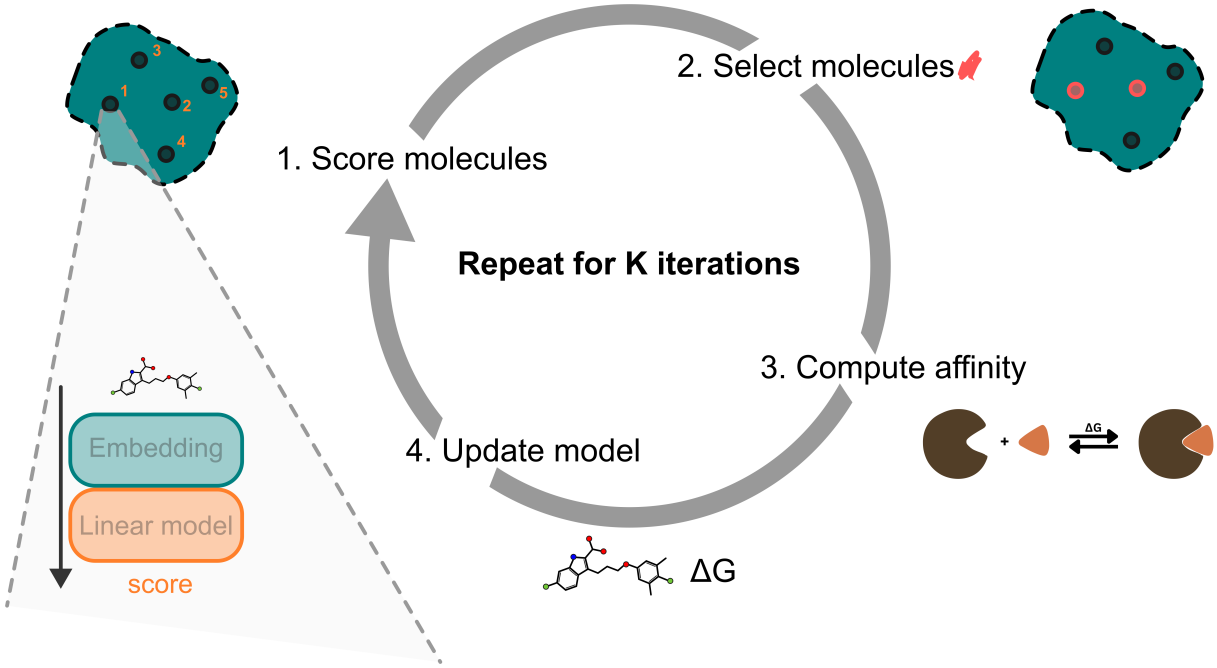

Figure 3: Pipeline of the active learning process:
(1) All molecules are initially scored using the current surrogate model.
(2) A subset of molecules is then selected with the acquisition function for binding affinity evaluation.
(3) Binding affinities are computed using either classical docking methods or MD-based techniques such as MMGBSA.
(4) The surrogate model is retrained with the additional data to refine predictions.

affinity is reported as the average and standard error across these 20 simulations. The GB-OBC2 model[70] was used with an internal dielectric constant of 1.0 and external dielectric constant of 78.5, the mbondi2 radii set, a surface tension of 0.0072 kcal/mol/Å^2, and a salt concentration of 150 mM. Following recommendations by Su *et al.*[71], the entropy term was not included. An example MD simulation system including topology and parameter files is provided on GitHub at `https://github.com/maxrdu/bayesian-optimization-mmgbsa`.

Active Learning pipeline

The pipeline is illustrated in Figure 3. The pipeline features the interplay of three central components: the surrogate model, the acquisition function, and the binding affinity computation technique. The surrogate model is the base AI model that is used, on the one hand, to predict binding affinities and, on the other hand, to estimate the uncertainty of the model prediction (Step (1) in Fig. 3). The surrogate model comprises two parts. The first part is an embedding method that transforms the molecules represented as a SMILES string to a high-dimensional vector, whereas the second part is a regression model that predict the binding affinity from the high-dimensional vector. We explored three different techniques to map the SMILES representation of a molecule to a high-dimensional vector: Morgan fingerprints,[72] ChemBERTa-2,[46] and MolFormer.[47] The latter two methods are transformer-based language models that have been pretrained using self-supervised learning.

The acquisition function is a mathematical function that computes the informational value of a molecule based on the model prediction and uncertainty estimation. Step (2) of Fig. 3 uses the acquisition function to compute the informational value of each molecule and selects a predefined number of most valuable molecules for which the binding affinity is computed. After a selection of promising molecules, Step (3) computes the respective binding affinities, for instance using an MD-based technique such as MMGBSA, thereby expanding our dataset. Step (4) of Fig. 3 uses the expanded dataset to retrain our surrogate model for the next iteration. The process is repeated starting at Step (1).

In summary, in each iteration, we score all unlabeled molecules using the surrogate model and use the acquisition function to select the set of molecules with the highest utility value. In case we acquire not only one but a batch of b data points per iteration, we update the model only every b iterations. After the label is queried, the regression head of the surrogate model is retrained. The initial set of computed binding affinities comprises a single molecule taken as the molecule closest to the centroid in the embedding space. The pseudo-code of the active learning pipeline is shown in algorithm 1.

Algorithm 1 Bayesian Optimization with Surrogate Model

Require: Embedding model Embed, set of molecules $\mathcal{M}$, labeler f, number of iterations K

$\mathcal{X} \leftarrow \{\text{Embed}(m) \mid m \in \mathcal{M}\}$ ▷ Create embeddings
$x_0 \leftarrow \text{Centroid}(\mathcal{X})$ ▷ Find molecule closest to Centroid
$y_0 \leftarrow f(x_0)$ ▷ Query initial label
Initialize regression model $\hat{f}$ with (x_0, y_0) ▷ Train initial surrogate model
$\mathcal{L} \leftarrow \{x_0\}$ ▷ Set of labeled points
for $i \leftarrow 1$ to K **do**
  $x_i \leftarrow \arg\max_{x \in \mathcal{X} \setminus \mathcal{L}} \text{EI}(x, \hat{f})$ ▷ Select next unlabeled point
  $y_i \leftarrow f(x_i)$ ▷ Query label
  Update $\hat{f}$ with (x_i, y_i) ▷ Update surrogate model
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{x_i\}$ ▷ Add to set of labeled points
end for

Bayesian linear regression

Bayesian Optimization requires access to the predictive posterior distribution of the surrogate model. We choose Bayesian linear regression as a simple model that fulfills this requirement. In combination with a pretrained embedding model, this is essentially equivalent to a Gaussian Process (GP) with a fixed basis function (the embedding model) and a dot product kernel[33]. Compared to usual linear regression, Bayesian linear regression assumes a prior distribution over the weights. In particular let $\boldsymbol{w} \in \mathbb{R}^d$ be the weights of the model, we assume $p(\boldsymbol{w}|\alpha) = \mathcal{N}(0, \alpha^{-1}\mathbf{I})$, with α^{-1} being the variance. As is standard in regression tasks, we assume that the target is corrupted by independent Gaussian noise, i.e. $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$, where $p(\epsilon) = \mathcal{N}(0, \beta^{-1})$, $f(\mathbf{x}) = \boldsymbol{w}^T\mathbf{x}$, and $\mathbf{x}$ is the embedding vector of the ligand. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be the training inputs with associated targets $\mathbf{y} \in \mathbb{R}^n$. By Bayes theorem, the posterior distribution of the weights is a Gaussian with mean $\hat{\boldsymbol{w}} = \beta\boldsymbol{\Sigma}\boldsymbol{X}^T\mathbf{y}$ and inverse covariance matrix $\boldsymbol{\Sigma}^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{X}^T\boldsymbol{X}$. Marginalizing the posterior distribution over the weights yields the predictive posterior distribution as a Gaussian with mean $\hat{\boldsymbol{w}}^T\mathbf{x}$ and variance $\sigma^2(\mathbf{x}) = \beta^{-1} + \mathbf{x}^T\boldsymbol{\Sigma}\mathbf{x}$. For BO applications, the true values of α and β are typically not known. We infer them by maximizing the marginal likelihood[73]. For more information we refer to Bishop[73] or Murphy[74]. Notably, the combination with a pretrained model is essentially the approach taken by Snoek *et al.*[35], where the weights of the embedding model

may be considered as hyperparameters of the basis function.

Results

As discussed in the dataset section, MMGBSA and docking scores differ considerably in their ability to predict top binders. MMGBSA scores show a stronger correlation with experimental binding affinities compared to Vina scores, suggesting that MMGBSA provides a more reliable binding affinity estimate among the whole dataset (Fig. 2a/b). These differences are likely explained by the account for protein and ligand flexibility by the MD simulation. Critically, the sets of top-1% binders identified by MMGBSA and AutoDock Vina differ substantially, suggesting that identifying top binders based only on the docking scores yields only with low probability good binders according to MMGBSA or according to experiment .

Updating only the regression model is sufficient for rapid retrieval of top binders

Prior work by Graff *et al.*[21] and Cao *et al.*[25] benchmarked the performance of a surrogate model in finding top-1% ligands among the `Enamine50k`. As scores, these authors used docking scores against thymidylate kinase obtained by Autodock Vina (see Datasets). To compare our pipeline with previous work, we follow the experimental setting by querying a batch of b molecules, where b is approximately 1% of the dataset size, corresponding to $b = 500$ for the `Enamine50k` dataset. We perform six rounds, such that only 6% of the whole dataset is screened. The results are summarized in Table 1 (first column).

To select new molecules for evaluation, we score the molecule's utility using the Expected Improvement (EI) criterion.[36] Prior work[21,25] has also considered alternative acquisition functions such as the Upper Confidence Bound (UCB) and surrogate model predictions (greedy selection). In contrast to UCB, EI does not require tuning of additional hyperparameters. Following Cao *et al.*[25], we assess the model performance using the retrieval rate of the

Table 1: Top-1% retrieval rate on the `Enamine50k` dataset based on Autodock Vina docking scores against thymidylate kinase. The first column shows the retrieval rate upon querying a batch of 500 molecules per iteration for six iterations, totaling 3000 molecules. The second column presents the retrieval rate upon querying 3000 molecules individually, one after another. We report results for both the greedy and upper confidence bound (UCB) acquisition strategies from Cao *et al.*[25] (rows 1-4) and compare to the results obtained by our method when using expected improvement (rows 5-10).

	Top-1% Retrieval Rate (Batched Query)	Top-1% Retrieval Rate (Single Molecule Query)
MolFormer (Greedy)[25]	78.36	✗
MolFormer (UCB)[25]	79.24	✗
RF (Greedy)[25]	54.52	✗
RF (UCB)[25]	37.08	✗
ChemBERTa-2 + Linear	78.23	81.85
ChemBERTa-2 + RF	65.73	73.99
MolFormer + Linear	74.39	78.02
MolFormer + RF	40.52	58.26
Morgan fingerprints + Linear	65.12	73.79
Morgan fingerprints + RF	39.91	52.62

top-1% of molecules according to the binding affinity score. Cao *et al.*[25] report a top-1% retrieval rate of 79.24% on `Enamine50k` using the MolFormer model after querying 6% of molecules in the dataset.[47] In comparison, our approach combining MolFormer with a linear model (`MolFormer+Linear`) achieves a slightly lower rate of 74.39%. However, substituting MolFormer with ChemBERTa-2 embeddings (`ChemBERTa-2+Linear`) improves performance, yielding a retrieval rate of 78.23%. The similar performance of the batched acquisition and one-at-a-time acquisition models indicates that updating the regression model instead of updating the entire pipeline including the regression model and the much larger embedding model has minimal impact on the overall retrieval rate. Thus, it is not required to further train the embedding model. Furthermore, we find that using a Random Forest model instead of the linear regression model significantly worsens performance in all cases (cf. Table 1). A similar trend is observed with Morgan fingerprints, which exhibit comparatively lower retrieval performance than models based on pretrained embeddings. Consequently, we focus our further analysis on results obtained using the ChemBERTa-2 and MolFormer embed-

ding models in combination with the linear regression model. Additional results involving the RF model and Morgan fingerprints are presented in Tables S1 and S2. In conclusion, these findings suggest that full model finetuning is not necessary; instead, updating only the regression model after each iteration is sufficient to achieve strong retrieval performance.

Querying one-at-at-time molecule instead of querying batches of molecules improves top-1% retrieval rate

Previous work[21,25] has employed deep neural architectures, for which finetuning until convergence is computationally expensive. As a result, prior approaches queried large batches, typically around 1% of the dataset, in each iteration to reduce the number of training steps. By updating only the regression head of our surrogate models, training becomes significantly faster, making it feasible to use batch sizes as small as a single molecule per iteration.

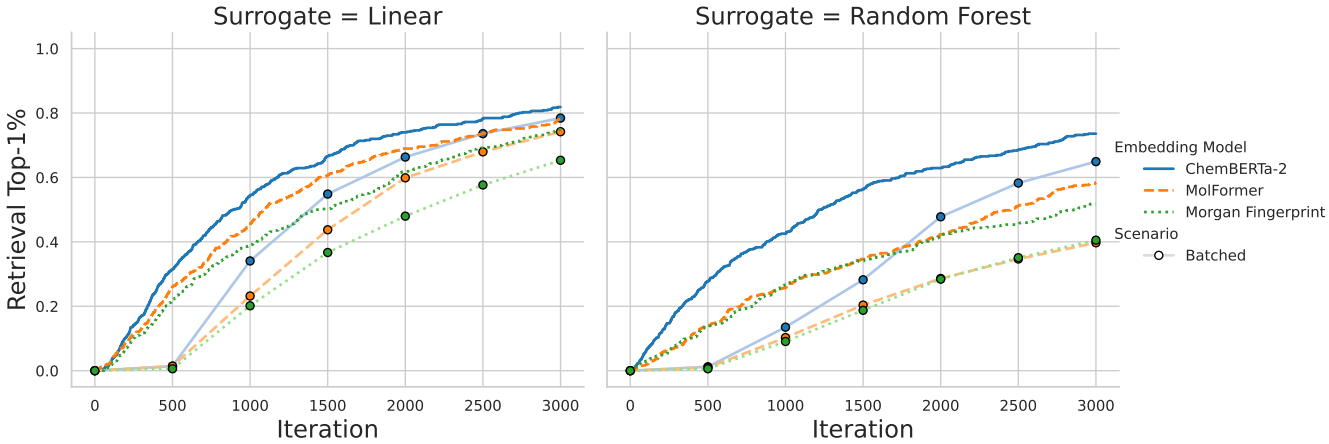

Figure 4: Top-1% retrieval rate based on docking scores of the `Enamine50k` dataset as a function of the number of screened molecules. Lines without markers represent one-at-a-time molecule acquisition, while lines with markers correspond to batch acquisitions of 500 molecules per iteration. Color codes and line styles indicate choice of the embedding model (see legend). Left: based a linear regression model. Right: Based on a Random Forest model.

By querying a single molecule per iteration, we achieve a top-1% retrieval rate of 81.85% using `ChemBERTa-2+Linear` on the `Enamine50k` dataset, surpassing the previous best perfor-

Table 2: Top-1% retrieval rate of the linear regression model based on docking and MMGBSA scores for the MCL1 dataset. Full results, including retrieval rates using Morgan fingerprints and a random forest regression model, are provided in the supplementary material.

	MCL1-Docking		MCL1-MMGBSA	
	Top-1% Retrieval Rate (Bateched Query)	Top-1% Retrieval Rate (Single Molecule Query)	Top-1% Retrieval Rate (Bateched Query)	Top-1% Retrieval Rate (Single Molecule Query)
ChemBERTa-2	94.60	97.13	59.86	62.73
MolFormer	93.76	95.78	78.41	79.93

mance of 79.24% (Table 1, compare first with second column). The advantage of querying single molecules is more evident during the early stages of the active learning process (Fig. 4): here, top-1% retrieval rate increases more rapidly with a one-at-at-time acquisition as compared to querying batches of 500 molecules, irrespective of the choices for the surrogate and embedding model.

Active learning of MMGBSA scores is more challenging than learning of docking scores

Having established our active learning pipeline and compared it to previous work on the `Enamine50k` dataset, we next apply our pipeline to our newly derived dataset comprising 59,356 compounds taken from the ZINC22 database binding to MCL1 (see Datasets). Accordingly, we analyze the performance of active learning on both the docking scores (`MCL1-Docking`) and the MD-based MMGBSA binding affinity (`MCL1-MMGBSA`), while again comparing batched-query active learning approach with the one-at-a-time variant. Focusing first at the `MCL1-Docking` dataset, the batched model results in comparable scores between the MolFormer and ChemBERTa-2 models, both achieving over 90% retrieval rate (Table 2). Querying a single molecule, instead of a batch, greatly improves the top-1% retrieval during the early stages of the active learning process (Fig. 5, left) and, after 3600 iterations, improves the top-1% retrieval rate in both cases, with `MolFormer+Linear` reaching 96% and `ChemBERTa-2+Linear` reaching 97%. These advantages of the one-at-a-time acquisition are in line with the findings on the docking-based `Enamine50k` dataset presented above. Fur-

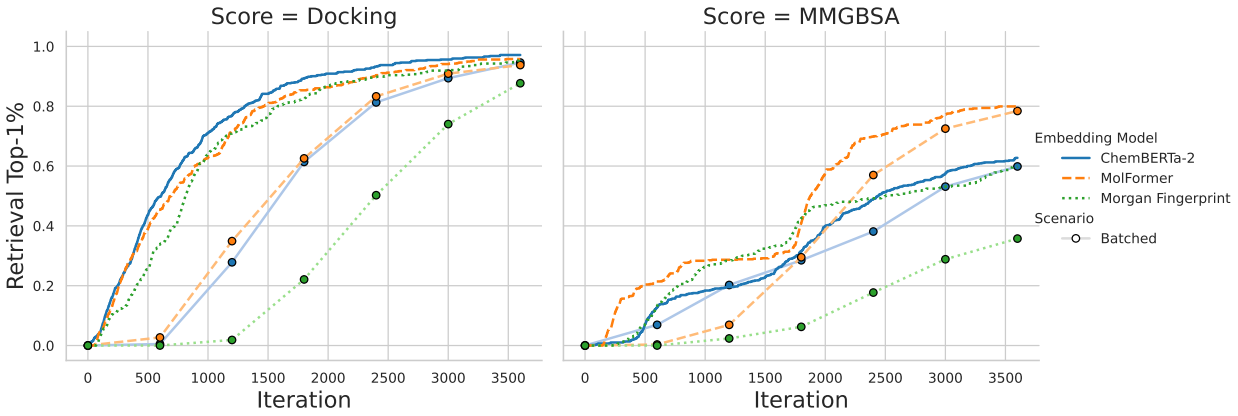

Figure 5: Top-1% retrieval rate of molecules based on docking and MMGBSA scores using a linear regression model. Lines without markers represent one-at-a-time molecule acquisition, while lines with markers correspond to batch acquisitions of 500 molecules per iteration. Results for the retrieval process using the random forest (RF) model are provided in the supplementary materials.

thermore, the finding that using MolFormer and ChemBERTa-2 yields similar learning rates aligns with the findings on the `Enamine50k` dataset.

Applying our pipeline to `MCL1-MMGBSA` dataset shows, as a key finding of our study, successful learning of MMGBSA scores, achieving top-1% retrieval rates up to 80% upon querying 6% of the compounds.

However, active learning of the `MCL1-MMGBSA` scores revealed also striking differences relative learning of the `MCL1-Docking` scores. We obtained lower top-1% retrieval rates when using MMGBSA scores (63% and 80%) as compared to using docking scores (97% and 96%; Table 2; Fig. 5, compare left with right panel). In addition, while the top-1% retrieval rate based on docking scores increases rapidly during early stages of the process, top-1% retrieval rate based on MMGBSA increases irregularly and requires by far more queries before reaching significant values. These findings suggest that learning docking scores is far easier as compared to learning MMGBSA scores. We attribute these findings to the relative simplicity of the docking scoring algorithm, which we hypothesize renders the energy landscape smoother and its patterns easier for the model to learn. In contrast, MMGBSA scores may have a more complex structure owing to the physically more detailed

representation of the ligand–protein interactions, rationalizing the need for a larger number of queries to achieve good top-1% retrieval rates.

Moreover, choice of the embedding plays a different role during learning of MMGBSA scores as compared to docking scores (Fig. 5, right, blue and orange curves). Upon learning docking scores with the linear surrogate model, learning based on ChemBERTa-2 slightly outperforms learning based on MolFormer. In contrast, upon learning MMGBSA scores, using MolFormer leads to 20% better learning rates compared to using ChemBERTa-2. To test whether these findings are consistent across independent active learning campaigns, we repeated training of our surrogate model, yet starting from different initial molecules (Fig. 6). Evidently, learning rates differ considerably upon using different initial molecules. However, top-1% learning rates after 3600 interations using Molformer consistently outperform the learning rates using ChemBERTa-2 or Morgan fingerprints, as shown by top-1% retrieval rates averaged over seven campaigns of 84%, 62%, and 50%, respectively. We hypothesize that the structural context embedded in MolFormer by the use of masked language modeling as a pretraining objective helps our surrogate model in learning MMGBSA scores. Thus, systematic testing of different embedding strategies for learning MMGBSA scores may be insightful in future studies.

Considering that MMGBSA scores correlate more strongly with experimental binding affinities than docking scores and that its top binders differ significantly from those identified by docking scores (Fig. 2), we infer that MMGBSA-based selection may better identify candidates with stronger experimental binding affinities. While models may achieve higher top-1% retrieval rates when using docking scores, the binders selected according to the MMGBSA scores are likely of higher binding affinity, making them more promising candidates for further development in the drug discovery process.

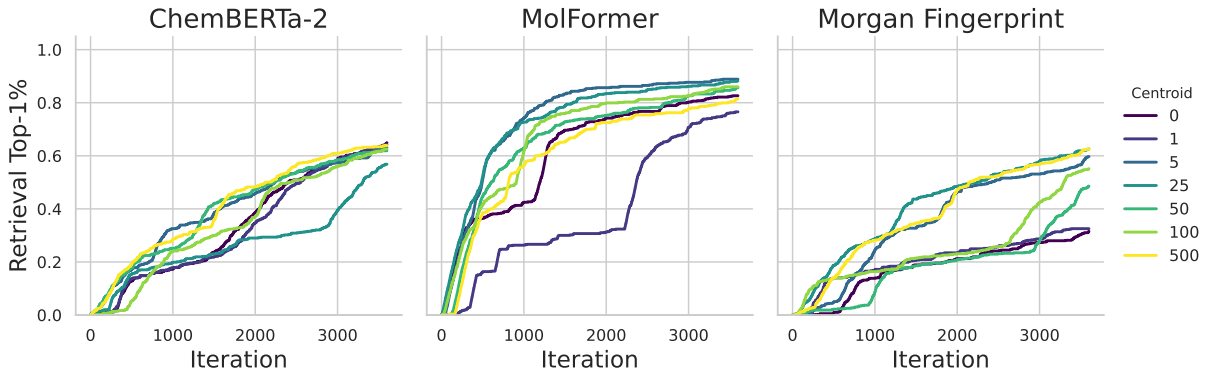

Figure 6: Top-1% retrieval rate of molecules based on MMGBSA scores using a linear regression, one-at-a-time acquisition, and different embedding models: from left to right, ChemBERTa-2, MolFormer, and Morgan fingerprints. Training of the surrogate model was initiated with different molecules highlighted by color taken as the centroid molecule (0), as well as the 1st, 5th, 25th, 50th, 100th and 500th nearest molecule to the centroid. Learning rates may differ considerably when using different initial molecules. However, Molfomer consistently outperforms both ChemBERTa-2 and Morgan fingerprints.

Discussion

Influence of the embedding

Our results demonstrate that the choice of embedding model has a substantial impact on top-1% retrieval performance. Different retrieval rates upon using ChemBERTa-2 compared to using MolFormer may stem from the different sizes of their pretraining datasets and from different pretraining objectives. MolFormer was trained on 100 million SMILES-encoded molecules, sourced from the complete PubChem database and from a random 10% subset of ZINC[3]. In comparison, ChemBERTa-2 was pretrained on 77 million molecules exclusively from PubChem. Additionally, MolFormer employs the standard masked language modeling (MLM) objective,[38] whereas ChemBERTa-2 is available in two variants: one trained using MLM and another using a multi-task regression (MTR) objective, which involves predicting RDKit-derived molecular properties from SMILES representations. In this study, we have utilized the MTR version of ChemBERTa-2.

Recently, Sultan *et al.*[48] suggested that the MTR objective achieves superior results in

property prediction tasks. The binding affinity depends, in addition to properties such as ligand hydrophobicity, also on structural information such as the linker length between two chemical subgroups[54] and packing properties of interface atoms.[75] We hypothesize that subtle differences between the structural properties of molecules are not necessarily reflected in some chemical properties and therefore can lead to similar embeddings of these molecules. As a result, accurate prediction of binding free energies using simple linear regression is nearly impossible. The MLM objective on the other hand is not biased towards embeddings that encode domain specific knowledge (such as physicochemical properties of molecules), but rather implicitly learns the structure of molecules and possibly information relevant for the task at hand. It is important to note that additional protein targets must be tested to draw a definitive conclusion, as the observed results may also originate from MolFormer embeddings more effectively capturing ligand characteristics that may be particularly relevant to the MCL1 protein target and thereby resulting in the observed significantly higher top-1% retrieval rates.

Linear model versus Random Forest

Our findings indicate that a simple Bayesian linear regression model consistently outperforms the Random Forest (RF) model across all datasets and embedding models. Following the approach of Graff *et al.*[21] and Cao *et al.*[25], the RF model estimates predictive uncertainty as the standard deviation across the outputs of its individual decision trees. As previously reported in both studies, the RF model yields better performance when its mean prediction is used directly as the acquisition function, rather than relying on uncertainty-based strategies such as Expected Improvement (EI) or Upper Confidence Bound (UCB). We therefore attribute the weaker performance of the RF model to its poorly calibrated uncertainty estimates. In contrast, the standard deviation provided by the Bayesian linear regression model captures both the uncertainty due to measurement noise and the uncertainty in the model parameters.[73] We assume that this dual consideration acts as an implicit regularizer, helping

to mitigate overfitting.

Batch acquisition versus one-at-a-time acquisition

Across all datasets, we observe that selecting a single molecule per iteration consistently results in higher retrieval rates, regardless of the applied embedding or regression model, in particular during the early stages of learning.

Across all datasets, selecting a single molecule per iteration consistently yields higher retrieval rates, especially during early learning, regardless of the embedding or regression model used (Figs. 4 and 5). This is likely due to more frequent model updates, which enhance predictive performance. Unlike batched acquisition, one-at-a-time acquisition allows the model to retrain after each new data point, which is particularly advantageous when training data is still scarce.

While one-at-a-time acquisition improves early retrieval, it limits parallelization because each selection depends on the outcome of the previous simulation. In contrast, batched strategies enable concurrent evaluations, improving resource utilization on parallel architectures and overall throughput. Future work could explore dynamic batching, where new samples are incorporated as soon as computations are complete, to balance frequent updates with parallel efficiency.

Irregular Retrieval Patterns in MMGBSA Compared to Docking

When relying on docking scores, the top-1% retrieval rate improves steadily, plateauing after approximately 2000 iterations (cf. Figure 5) and having reached a high retrieval rate. In contrast, MMGBSA scores often exhibit an early plateau in retrieval rate, sometimes as soon as after 1,000 iterations, before showing a sudden improvement in later stages and eventually stabilizing around 3,000 iterations (cf. Figure 5). It is to note, that such intermediate plateaus tend to occur only for some starting molecules (cf. Figure 6). We find however, that this trend is highly dependent on the initial molecule. By choosing different initial

datapoints, we can recover similarly steady improvement (cf. Figure 6). This suggests that, instead of relying on a single initial sample, starting with a more diverse sample, potentially trading the increased initial cost for potentially avoiding plateaus in later iterations.

This plateau suggests that, upon learning MMGBSA scores, the model may get trapped in a specific region of the chemical space before eventually identifying a new set of molecular properties that help to escape to other promising regions. Notably, this initial plateau is not observed when using the batched acquisition strategy, presumably because the sampling is too coarse.

Although higher top-1% retrieval rates are achieved when using docking scores, it is important to consider how well these scores actually reflect true binding affinity. To assess this, we have compared predicted binding affinities, based on both docking and MMGBSA scores, with experimentally measured values for a set of known binders (see Figure 2 (a) and Figure 2 (b)). Our analysis reveals that MMGBSA scores exhibit a much stronger correlation with the experimental data (Pearson-R: 0.61) compared to docking scores (Pearson-R: 0.18). Furthermore, Figure 2 (c) shows no apparent correlation between MMGBSA and docking scores across our dataset. Notably, the sets of top-1% binders identified by each scoring method differ significantly, with only a 1.6% overlap (represented by the green region in the figure). This comes as a result of MMGBSA accounting for the dynamics of both protein and ligand, in contrast AutoDock Vina relies on simplified empirical scoring functions.[58] We hypothesisze that the account for dynamics makes the relationship between ligand structure and predicted binding affinity more complex, making it more difficult to capture by machine learning models. This added complexity may help explain the observed irregular improvement of top-1% retrieval rates over the iterations.

Moreover, the embedding models used in this study work solely with the SMILES representation of molecules as input and therefore lack explicit information about their three-dimensional structures and spatial orientation within the protein's binding pocket. Such structural context may be critical for accurately capturing the interactions that influence

MMGBSA scores. The absence of this information could be a key factor limiting the model's ability to achieve top-1% retrieval rates with MMGBSA scores comparable to those observed with docking scores.

Conclusion and Future Work

We presented an active learning approach that combines pretrained embedding models with lightweight Bayesian regression models to predict binding affinities. In contrast to previous methods that require fine-tuning of the entire model for each new batch of data, our strategy greatly reduces computational overhead by updating only the parameters of the regression head. This design enables the use of smaller query batches, facilitating a more responsive feedback loop and ultimately leading to improved retrieval of top binders.

Given the high computational demands of MD-based binding free energy estimations, existing datasets typically contain only a few thousand data points. In this study, computed binding free energies with MMGBSA for approximately 60.000 ligands targeting MCL1. To test our active learning framework, we compared its performance upon using MMGBSA scores relative to using docking scores. After querying only 6% of the dataset, the model successfully identified 80% of the top-1% ligands according to MMGBSA, laying the ground for large-scale virtual screenings with MMGBSA-level accuracy. Additinaly, our results indicate that docking scores are easier to predict with a linear model, compared to MMGBSA scores, suggesting that MMGBSA scores rely on more complex structural and dynamic properties compared to docking scores, which represented by the SMILES representation alone. We hypothesize that enriching SMILES-based embeddings with additional information, particularly regarding ligand pose within the binding pocket, may help address this information gap and alleviate the observed performance difference. While MMGBSA scores are more challenging for the model to learn, they exhibit a stronger correlation with experimental binding affinities than docking scores, suggesting that top binders selected based on MMGBSA more

promising compounds for downstream drug discovery tasks. Based on our results, this active learning pipeline is now ready to incorporate additional scores into the learning loop such as affinities from free energy perturbation simulations or from wet lab experiments.

Acknowledgments

We thank Afnan Sultan for helpful discussions. This research was funded in part by the European Union's Horizon 2020 Research and Innovation Program under Marie Skłodowska Curie Grant 860592 and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; grant INST 256/539-1).

References

(1) Morgan, S.; Grootendorst, P.; Lexchin, J.; Cunningham, C.; Greyson, D. The cost of drug development: A systematic review. *Health Policy* **2011**, *100*, 4–17.

(2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Medicinal Research Reviews* **1996**, *16*, 3–50.

(3) Tingle, B. I.; Tang, K. G.; Castanon, M.; Gutierrez, J. J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y. S.; Irwin, J. J. ZINC-22 - A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *Journal of Chemical Information and Modeling* **2023**, *63*, 1166–1176, PMID: 36790087.

(4) Enamine enamine.net. `https://enamine.net/compound-collections/real-compounds`, [Accessed 17-10-2024].

(5) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemi-*

cal Information and Modeling **2021**, *61*, 3891–3898, Publisher: American Chemical Society.

(6) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry* **2004**, *47*, 1739–1749, PMID: 15027865.

(7) Pantsar, T.; Poso, A. Binding Affinity via Docking: Fact and Fiction. *Molecules* **2018**, *23*, 1899.

(8) Tuccinardi, T. What is the current value of MM/PBSA and MM/GBSA methods in drug discovery? *Expert Opinion on Drug Discovery* **2021**, *16*, 1233–1237, PMID: 34165011.

(9) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews* **2019**, *119*, 9478–9508.

(10) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and Use of the MM-PBSA Approach for Drug Discovery. *Journal of Medicinal Chemistry* **2005**, *48*, 4040–4048.

(11) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *Journal of Computational Chemistry* **2011**, *32*, 866–877.

(12) Frazier, P. I. A Tutorial on Bayesian Optimization. **2018**,

(13) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.;

Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **2018**, *4*, 268–276.

(14) Tripp, A.; Daxberger, E.; Hernández-Lobato, J. M. Sample-Efficient Optimization in the Latent Space of Deep Generative Models via Weighted Retraining. Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020; pp 11259–11272.

(15) Stanton, S.; Maddox, W.; Gruver, N.; Maffettone, P.; Delaney, E.; Greenside, P.; Wilson, A. G. Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders. Proceedings of the 39th International Conference on Machine Learning. 2022; pp 20459–20478.

(16) Maus, N.; Jones, H. T.; Moore, J. S.; Kusner, M. J.; Bradshaw, J.; Gardner, J. R. Local Latent Space Bayesian Optimization over Structured Inputs. 2023; `http://arxiv.org/abs/2201.11872`.

(17) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. **2003**, *43*, 667–673.

(18) De Grave, K.; Ramon, J.; De Raedt, L. Active Learning for High Throughput Screening. Discovery Science. 2008; pp 185–196.

(19) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. **2015**, *20*, 458–465.

(20) Pyzer-Knapp, E. O. Bayesian optimization for accelerated drug discovery. *IBM Journal of Research and Development* **2018**, *62*, 2:1–2:7.

(21) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual

screening through molecular pool-based active learning. *Chemical Science* **2021**, *12*, 7866–7881.

(22) Van Tilborg, D.; Grisoni, F. Traversing Chemical Space with Active Deep Learning for Low-Data Drug Discovery — Nature Computational Science. 2024; `https://www.nature.com/articles/s43588-024-00697-2`.

(23) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences* **2022**, *2*, 100050.

(24) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **2019**, *59*, 3370–3388.

(25) Cao, Z.; Sciabola, S.; Wang, Y. Large-Scale Pretraining Improves Sample Efficiency of Active Learning-Based Virtual Screening. **2024**, *64*, 1882–1891.

(26) Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K. Q.; Artzi, Y. Revisiting Few-sample BERT Fine-tuning. International Conference on Learning Representations. 2020.

(27) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* **2015**, *10*, 449–461, PMID: 25835573.

(28) Valdés-Tresanco, M. S.; Valdés-Tresanco, M. E.; Valiente, P. A.; Moreno, E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *Journal of Chemical Theory and Computation* **2021**, *17*, 6281–6291, PMID: 34586825.

(29) Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; van der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation* **2017**, *13*, 1034–1043, PMID: 28245118.

(30) Alibay, I.; Magarkar, A.; Seeliger, D.; Biggin, P. C. Evaluating the use of absolute binding free energy in the fragment optimisation process. *Communications Chemistry* **2022**, *5*, 105.

(31) Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V. Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *Journal of Chemical Theory and Computation* **2014**, *10*, 1228–1241, PMID: 24683369.

(32) Genheden, S. MM/GBSA and LIE estimates of host–guest affinities: dependence on charges and solvation model. *Journal of Computer-Aided Molecular Design* **2011**, *25*, 1085–1093.

(33) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2005.

(34) Springenberg, J. T.; Klein, A.; Falkner, S.; Hutter, F. Bayesian Optimization with Robust Bayesian Neural Networks. Advances in Neural Information Processing Systems. 2016.

(35) Snoek, J.; Rippel, O.; Swersky, K.; Kiros, R.; Satish, N.; Sundaram, N.; Patwary, M.; Prabhat, M.; Adams, R. Scalable Bayesian Optimization Using Deep Neural Networks. Proceedings of the 32nd International Conference on Machine Learning. pp 2171–2180.

(36) Jones, D. R. A Taxonomy of Global Optimization Methods Based on Response Surfaces. *Journal of Global Optimization* **2001**, *21*, 345–383.

(37) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems. 2017.

(38) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019; pp 4171–4186.

(39) Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018.

(40) Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021; http://arxiv.org/abs/2010.11929.

(41) Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. 2021; http://arxiv.org/abs/2104.14294.

(42) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations.

(43) He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 9726–9735.

(44) Honda, S.; Shi, S.; Ueda, H. R. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. 2019.

(45) Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular Representation Learning with Language Models and Domain-Relevant Auxiliary Tasks. 2020; http://arxiv.org/abs/2011.13230.

(46) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. 2022.

(47) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. **2022**, *4*, 1256–1264.

(48) Sultan, A.; Rausch-Dupont, M.; Khan, S.; Kalinina, O.; Volkamer, A.; Klakow, D. Transformers for molecular property prediction: Domain adaptation efficiently improves performance. 2025; http://arxiv.org/abs/2503.03360, arXiv:2503.03360 [cs].

(49) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. 2019.

(50) Brown, T. *et al.* Language Models Are Few-Shot Learners. Advances in Neural Information Processing Systems. 2020; pp 1877–1901.

(51) Hernandez, D.; Kaplan, J.; Henighan, T.; McCandlish, S. Scaling Laws for Transfer. 2021; http://arxiv.org/abs/2102.01293.

(52) Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2016; http://arxiv.org/abs/1506.02142.

(53) Widden, H.; Placzek, W. J. The multiple mechanisms of MCL1 in the regulation of cell fate. *Communications Biology* **2021**, *4*, 1029.

(54) Friberg, A.; Vigil, D.; Zhao, B.; Daniels, R. N.; Burke, J. P.; Garcia-Barrantes, P. M.; Camper, D.; Chauder, B. A.; Lee, T.; Olejniczak, E. T.; Fesik, S. W. Discovery of

Potent Myeloid Cell Leukemia 1 (Mcl-1) Inhibitors Using Fragment-Based Methods and Structure-Based Design. *Journal of Medicinal Chemistry* **2013**, *56*, 15–30.

(55) Tingle, B.; Tang, K.; Castanon, J.; Gutierrez, J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y.; Irwin, J. ZINC-22 - A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. 2022; `https://chemrxiv.org/engage/chemrxiv/article-details/634f2185dfbd2bbe525b876a`.

(56) Zhang, B.; Vogt, M.; Maggiora, G. M.; Bajorath, J. Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 937–950.

(57) Landrum, G. RDKit: Open-source cheminformatics. 2010; `https://www.rdkit.org/`.

(58) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455–461.

(59) Minibaeva, G.; Ivanova, A.; Polishchuk, P. EasyDock: customizable and scalable docking tool. *Journal of Cheminformatics* **2023**, *15*, 1–11, Number: 1 Publisher: BioMed Central.

(60) Ropp, P. J.; Kaminsky, J. C.; Yablonski, S.; Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **2019**, *11*, 14.

(61) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 1950–1958.

(62) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L.

Comparison of Simple Potential Functions for Simulating Liquid Water. *JCP* **1983**, *79*, 926–935.

(63) Boothroyd, S. *et al.* Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **2023**, *19*, 3251–3275, PMID: 37167319.

(64) Martínez-León, A. ale94mleon/TOFF: 0.1.0 - 2023.07.27. 2023; `https://doi.org/10.5281/zenodo.8189649`.

(65) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98*, 10089–10092.

(66) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995**, *103*, 8577–8593.

(67) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **1997**, *18*, 1463–1472, Relation: http://www.rug.nl/informatica/organisatie/overorganisatie/iwi Rights: University of Groningen. Research Institute for Mathematics and Computing Science (IWI).

(68) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(69) Miller, B. R. I.; McGee, T. D. J.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* **2012**, *8*, 3314–3321, PMID: 26605738.

(70) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **2004**, *55*, 383–394.

(71) Su, P.-C.; Tsai, C.-C.; Mehboob, S.; Hevener, K. E.; Johnson, M. E. Comparison of radii sets, entropy, QM methods, and sampling on MM-PBSA, MM-GBSA, and QM/MM-GBSA ligand binding energies of F. tularensis enoyl-ACP reductase (FabI). *Journal of Computational Chemistry* **2015**, *36*, 1859–1873.

(72) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107–113.

(73) Bishop, C. M. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer-Verlag New York, 2006.

(74) Murphy, K. P. *Probabilistic Machine Learning: An introduction*; MIT Press, 2022.

(75) Marillet, S.; Boudinot, P.; Cazals, F. High-resolution crystal structures leverage protein binding affinity predictions. *Proteins: Structure, Function, and Bioinformatics* **2016**, *84*, 9–20.

Supplementary Material for:

Accelerating ligand discovery by combining Bayesian optimization with MMGBSA-based binding affinity calculations

Table S1: Retrieval rates of all embedding-regression model combinations on the `MCL1-Docking` data. (B) indicates batched querying of 1% of the dataset for 6 iterations.

Regression Model	Embedding	Top-1% retrieval rate (B)	Top-1% retrieval rate
Linear	ChemBERTa-2	94.60	97.13
	MolFormer	93.76	95.78
	Morgan fingerprint	87.68	95.95
Random Forest	ChemBERTa-2	46.20	60.53
	MolFormer	59.02	68.12
	Morgan fingerprint	37.09	58.17

Table S2: Retrieval rates of all embedding-regression model combinations on the `MCL1-MMGBSA` data. (B) indicates batched querying of 1% of the dataset using 6 iterations.

		Top-1% retrieval rate (B)	Top-1% retrieval rate
Linear	ChemBERTa-2	59.86	62.73
	MolFormer	78.41	79.93
	Morgan fingerprint	35.75	59.86
Random Forest	ChemBERTa-2	33.72	46.20
	MolFormer	56.49	68.97
	Morgan fingerprint	33.55	36.25

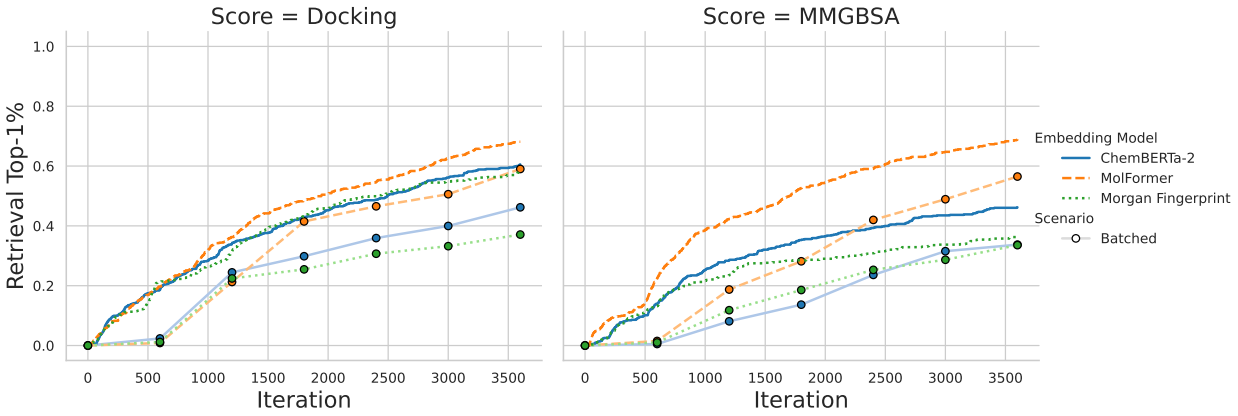

Figure S1: Retrieval of top-1% molecules on docking and MMGBSA scores using a Random Forrest instead of a linear regression model as presented in Fig. 5.

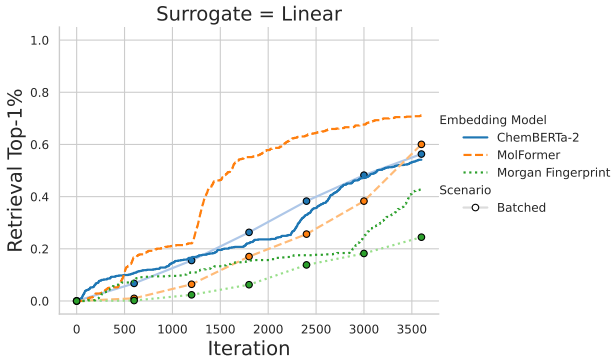

Figure S2: Retrieval of top-1% molecules on MMPBSA scores using a linear regression model.

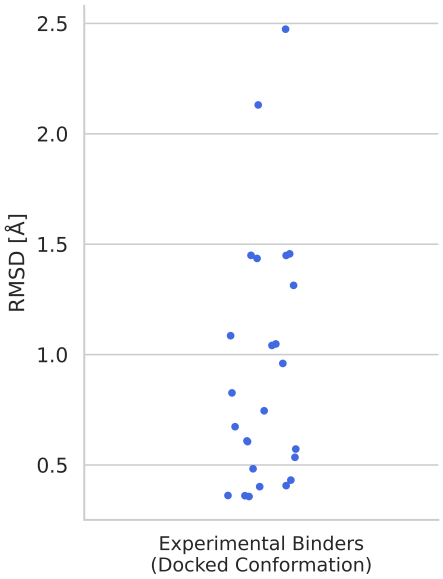

Figure S3: Distribution of root mean-square deviation (RMSD) values for the maximum common substructure (MCS) between docked poses of experimental binders[54] and the reference ligand in the MCL1 crystal structure.[54] Each point represents the RMSD of an individual ligand, c alculated using only heavy atoms. All ligands were docked into the same MCL1 protein structure, and RMSD values were computed without explicit alignment to the reference ligand. The generally low RMSD values indicate that AutoDock Vina consistently positioned the experimental binders in conformations closely resembling that of the crystallographic reference ligand.

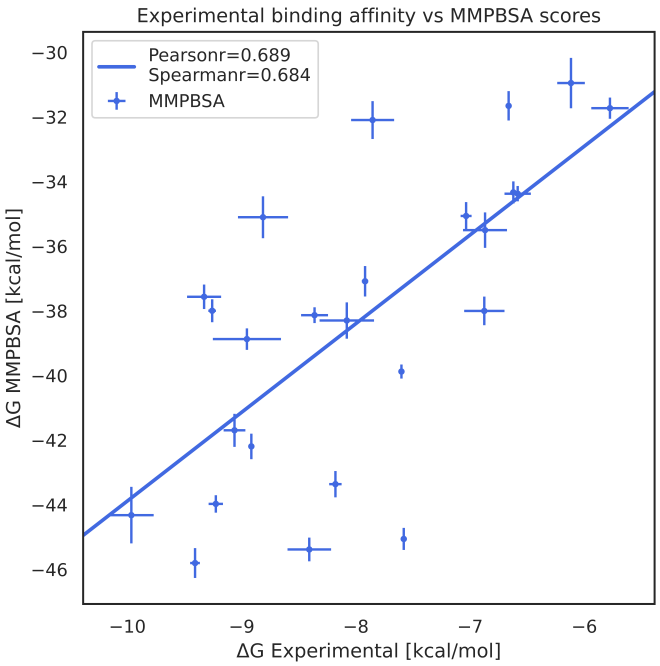

Figure S4: Correlation between MMPBSA (instead of MMGBSA) scores with experimental binding affinities for experimental binders. Pearson and Spearman correlation coefficients are slightly higher compared to results based on MMGBSA (compare with Fig. 2).