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

Supporting Information

Dataset

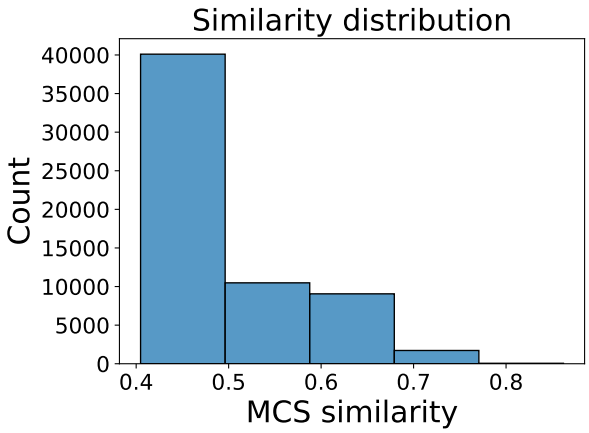

Figure S1: Histogram of MCS-Tanimoto similarity to the query compound. The dataset was screened with a minimum similarity threshold of 0.4. Most compounds show low similarity values, typical for chemical datasets.

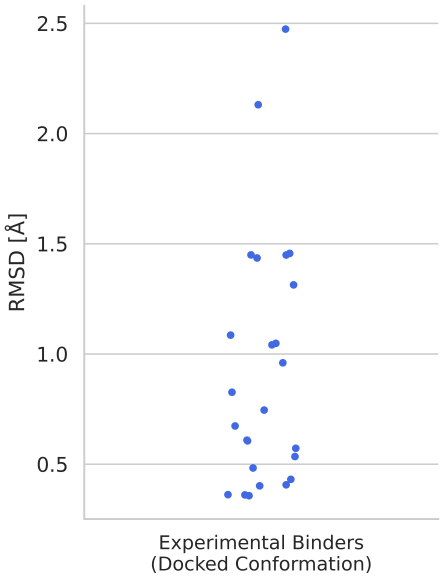

Figure S2: Distribution of root mean-square deviation (RMSD) values for the maximum common substructure (MCS) between docked poses of experimental binders[1] and the reference ligand in the MCL1 crystal structure.[1] Each point represents the RMSD of an individual ligand, calculated using only heavy 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.

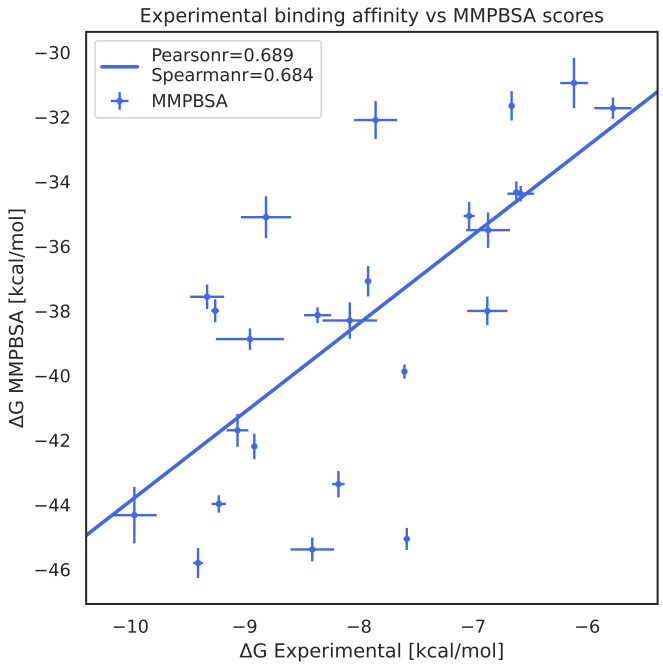

Figure S3: Correlation between MMPBSA (instead of MMGBSA) scores with experimental binding affinities for experimental binders. Pearson and Spearman correlation coefficients are slightly higher compared to results based on MMGBSA (compare with Fig. 2).

Extended results

In this section we provide the retrieval rates of all embedding models on the `MCL1-Docking` (S4) and `MCL1-MMGBSA` (S5) datasets. Additionally, we visualize the retrieval curves of the Random Forest in Figure S6.

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

Figure S4: 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.

		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 S5: 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.

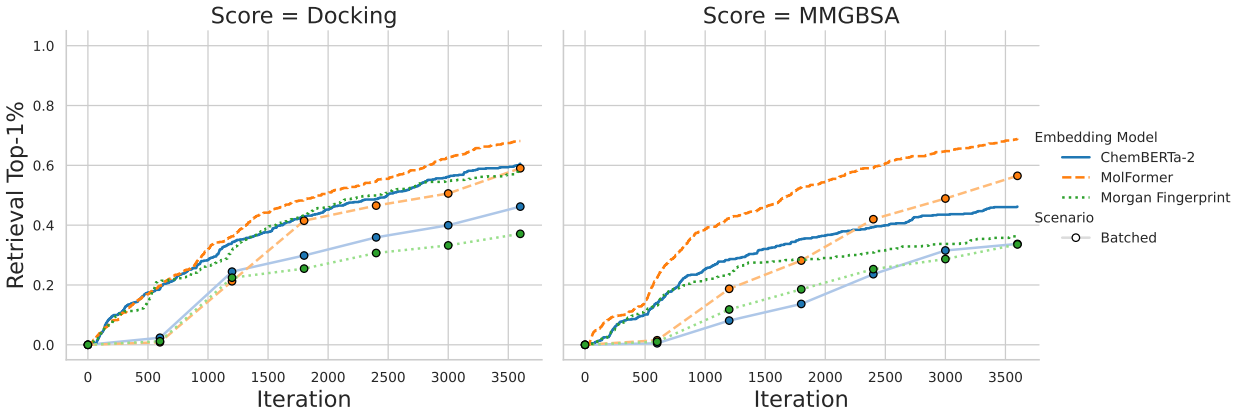

Figure S6: Retrieval of top-1% molecules on docking and MMGBSA scores using a Random Forrest instead of a linear regression model as presented in Fig. 6. in the main text.

MMPBSA

The MD trajectories that were used for the MMGBSA computations, were also used for the MMPBSA computations. For the MMPBSA computations, we used an internal dielectric constant of 1.0, and an external dielectric constant of 80.0 with optimized atomic radii from Tan *et al.*[2]. The Periodic Incomplete Cholesky Conjugate Gradient Descent method was used to solve the PB equations with additional parameters including a mobile ion probe radius of 2.0 Å, a grid spacing of 0.5 Å, and a convergence criterion of 0.001. As for MMGBSA, the entropic effects are not included.

We also run Bayesian optimization using MMPBSA scores instead of MMGBSA scores

(see S7). We use the same settings for this experiment as described in the main text. We find similar trends as for MMGBSA. However, when comparing the overall retrieval performance with MMPBSA scores against the retrieval performance of MMGBSA scores, we find that the retrieval performance using MMPBSA scores is slightly lower compared to that of MMGBSA.

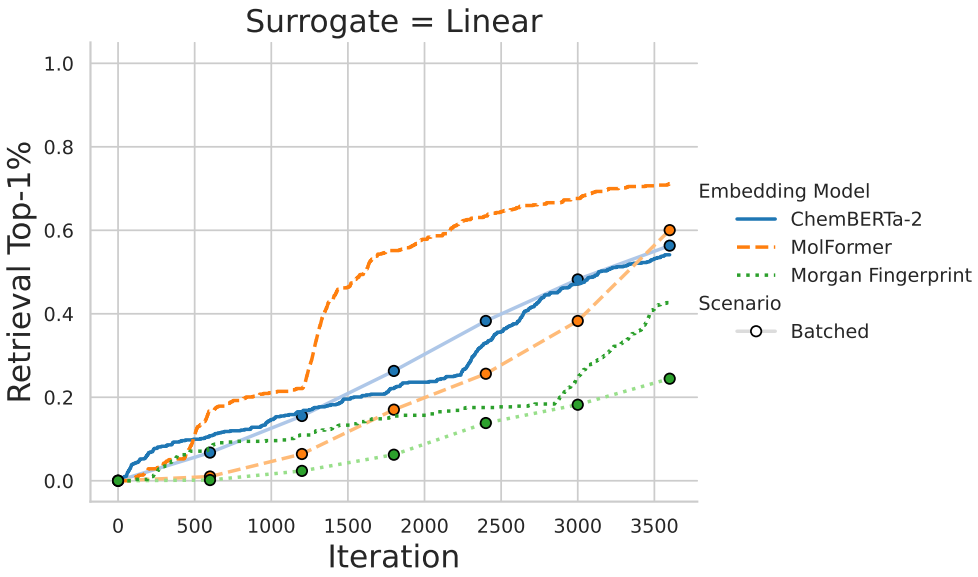

Figure S7: Retrieval of top-1% molecules on MMPBSA scores using a linear regression model.

Initialization robustness

To quantify the dependence of the final retrieval rate on the initial datapoint, we perform two experiments. Initialization with a random datapoint from the dataset and randomly selecting one out of the 50 closest datapoints to the centroid (the default initialization). The second option better reflects a typical drug discovery campaign, where a potential lead compound is known. We observe (S8, S9, S10) that neither choice has a significant influence on the final retrieval rate for the neural representations.

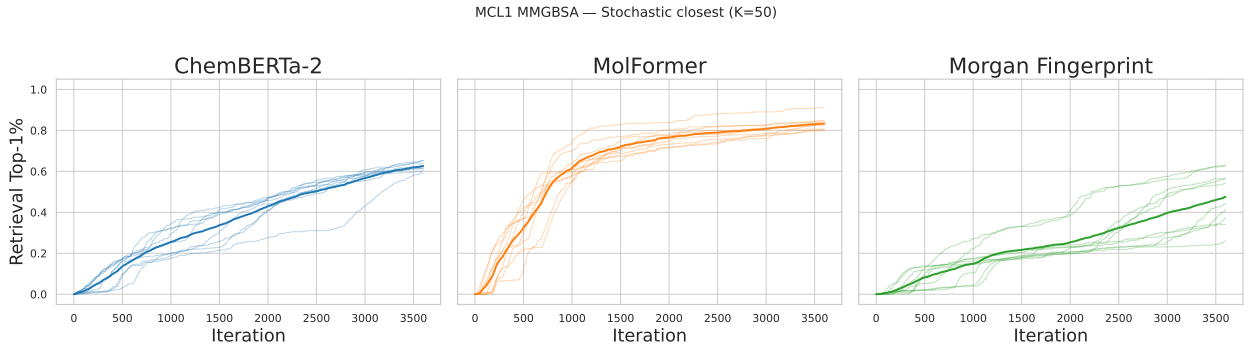

Figure S8: Retrieval of top-1% molecules when randomly choosing one of the 50th closest datapoint to the centroid.

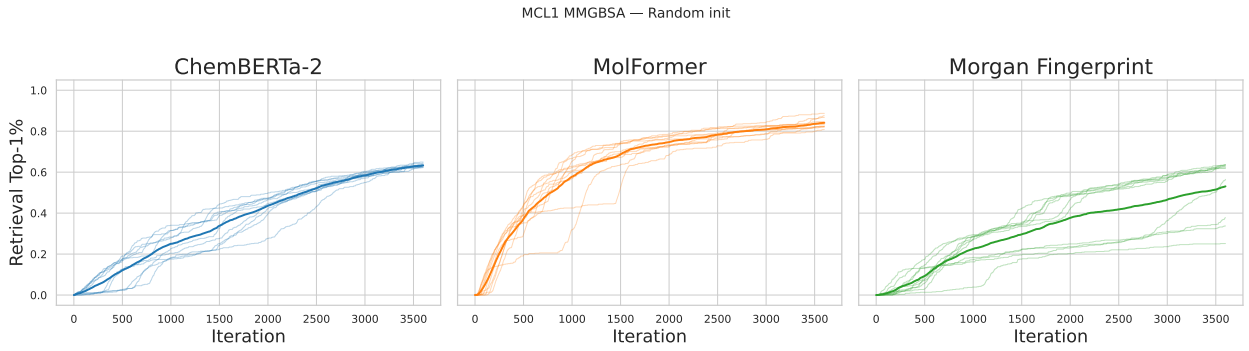

Figure S9: Retrieval of top-1% molecules when randomly choosing from the dataset.

		Retrieval (mean)	Retrieval (min)	Retrieval (max)	± 95% CI
Random	ChemBERTa-2	0.633	0.622	0.649	0.005
	MolFormer	0.841	0.808	0.887	0.016
	Morgan FP	0.531	0.251	0.637	0.087
Stochastic closest	ChemBERTa-2	0.626	0.592	0.654	0.013
	MolFormer	0.834	0.798	0.914	0.022
	Morgan FP	0.476	0.261	0.629	0.073

Figure S10: Retrieval of top-1% molecules for different initialization strategies. The confidence interval is computed using bootstrapping with 10k resampling draws.

Additional investigations on the plateau

In this section we carry out multiple investigations on how to overcome the plateau, which we observed when using MolFormer embeddings in combination with MMGBSA scoring. Namely, we compare upper confidence bound (UCB) acquisition and initialization with a diverse set of compounds to the default choices in the main text. Additionally, we provide results on how the choice of the batch size affects the final retrieval performance. We find all three choices to be independently helpful in avoiding the plateau.

Upper confidence bound acquisition

The upper confidence bound selects compounds by maximizing

$$\mathrm{UCB}(x) = \mu(x) + \kappa\sigma(x),$$

where $\mu(x)$ and $\sigma(x)$ denote the posterior and mean standard deviation respectively. Compared to expected improvement (EI), this allows manually favoring exploration strength by increasing κ. By tuning κ we find that we are able to avoid the plateau, which we observed using the MolFormer embedding (Figure S11). However, the impact on the final retrieval rate is negligible.

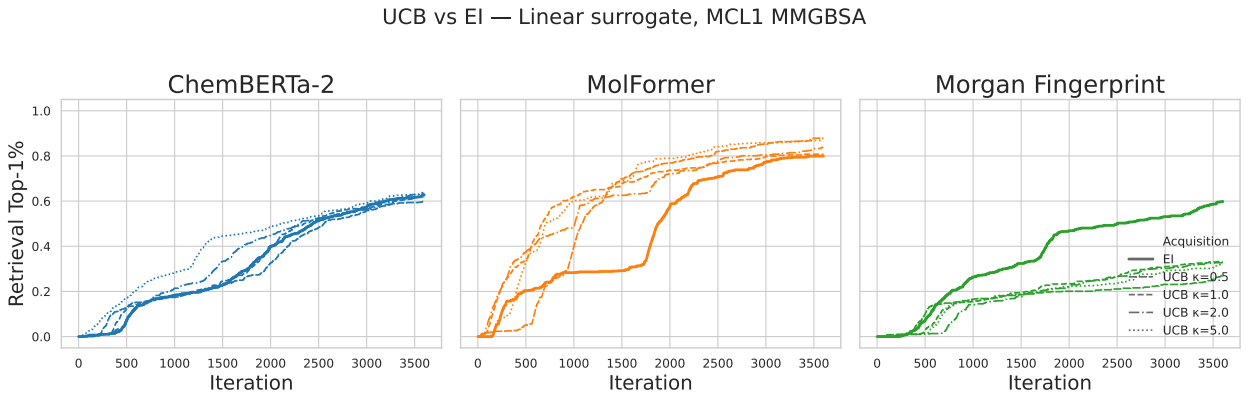

Figure S11: Retrieval of top-1% molecules with the linear model using UCB with multiple exploration values.

Diverse Initialization

Apart from changing the acquisition function to favor exploration, we also have the option to initialize the model with a selection of compounds, potentially overcoming the bias of starting with a single compound. In order to estimate the effect of this choice, we select a diverse set of compounds using a greedy min-max strategy. Pseudo-code for the initial selection can be found in Algorithm S1. Based on the retrieval process in Figure S12 we can observe that this strategy is also able to avoid the plateau in some settings. However, choosing to many datapoints again results in being stuck for some iterations. We suppose that with to many datapoints, the suboptimal greedy strategy biases towards certain regions. Visual inspection of the selected datapoints in a UMAP projection (Figure S14) supports this hypothesis. While for 10 or 50 compounds, the datapoints are nicely distributed across the space, with 100 samples, most of the new compounds are sampled from the upper cluster (last row, middle plot).

Effect of batch size

As noted in the discussion section, there is a trade-off between one-at-a-time acquisition and large batch sizes. In this section we investigate the effect of the batch size k in more detail. We choose $k \in \{1, 5, 10, 50, 100, 600\}$ and run Bayesian optimization with the linear model

Algorithm 1 Greedy MaxMin Diversity Selection

Require: Set of unlabeled molecules $\mathcal{U}$, selection size s

$\mathcal{S} \leftarrow \{\text{molecule closest to centroid of } \mathcal{U}\}$ ▷ Seed selection
while $|\mathcal{S}| < s$ **do**
  for each candidate $m \in \mathcal{U} \setminus \mathcal{S}$ **do**
    $\delta(m) \leftarrow \min_{m' \in \mathcal{S}} \|m - m'\|$ ▷ Distance to nearest selected molecule
  end for
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{\arg\max_m \ \delta(m)\}$ ▷ Select the furthest molecule
end while
return $\mathcal{S}$

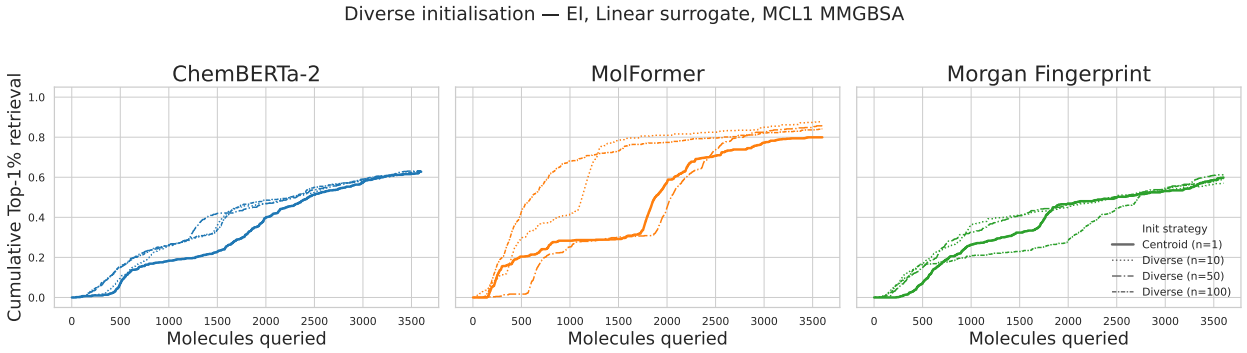

Figure S12: Retrieval of top-1% molecules with the linear model starting with a set of n diverse compounds from the dataset.

on the `MCL1-Docking` and `MCL1-MMGBSA` dataset (Figure S13). We find that comparably small batch sizes (such as 5, 10, 50 or 100) result in a similar final retrieval rate. At the same time, we observe that the batch size can have a regularizing effect, avoiding the plateau in case of the MolFormer embeddings, when using MMGBSA scores.

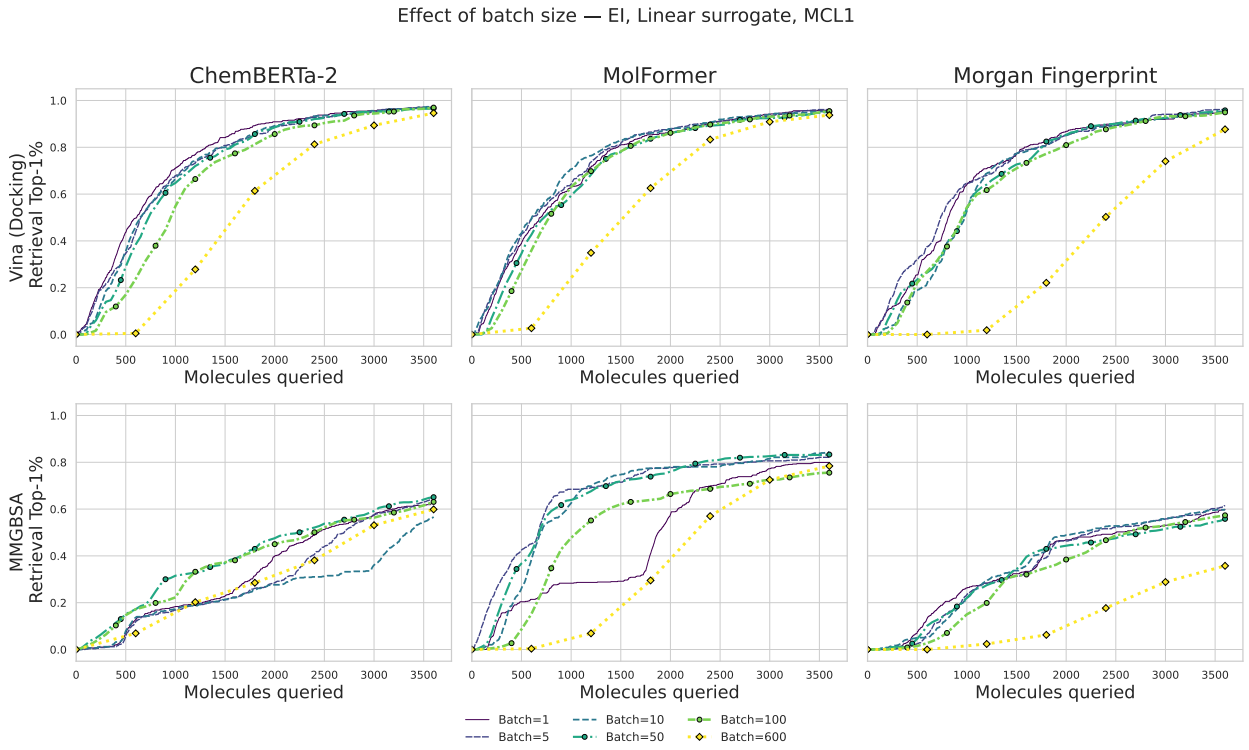

Figure S13: Retrieval of top-1% molecules with the linear model, varying the batch size between 1 and 600.

Low-dimensional visualizations

In this section we provide UMAP visualizations (`n_neighbors=30`, `min_dist`=0.05, PCA initialization) of the embedding spaces. Datapoints are colored by their MMGBSA scores.

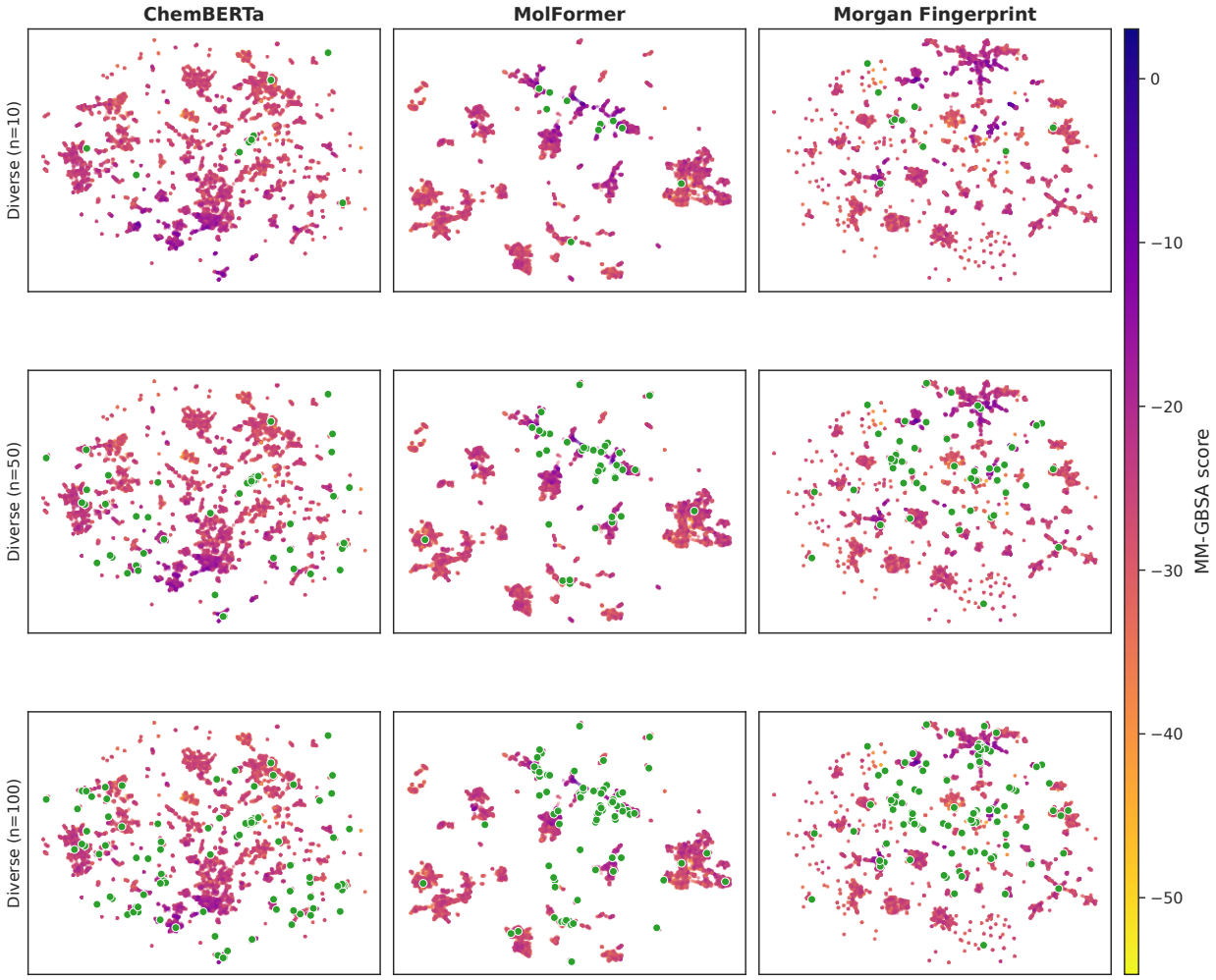

Figure S14: UMAP visualization of the dataset. The n compounds selected for diverse initialization are highlighted by green markers.

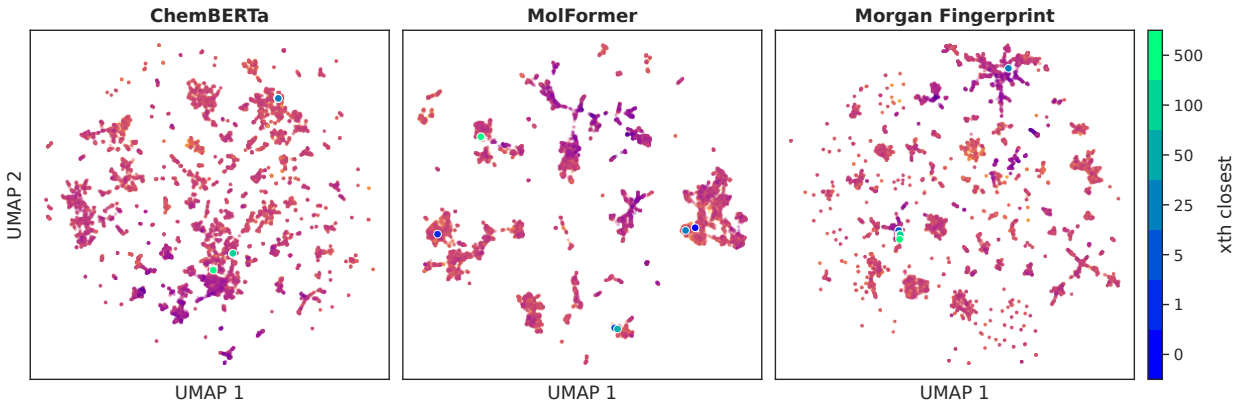

Figure S15: UMAP visualization of the dataset. The kth nearest datapoints to the medoid are highlighted.

Smoothness

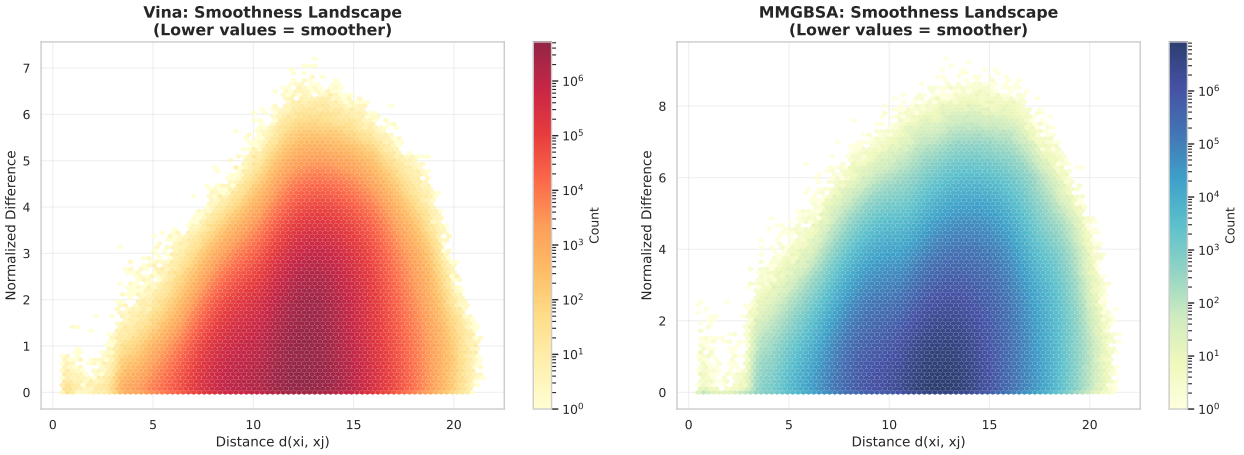

Figure S16: Hexbin plots of the normalized activity difference binned by distance in Mol-Former embedding space. For the same distance, the difference in the activity value for MMGBSA is generally larger than for Vina, indicating a less smooth energy surface.

Computational Cost

The primary factor determining the computational cost of the dataset preparation is the MD simulation. For each of the $59,356 \approx 60,000$ ligands, a total of 3 ns of MD simulations were conducted. For the MD simulations, we used an Nvidia GTX 1070 TI GPU, which needed approximately 30 minutes to complete. A total of $60,000 \cdot 0.5\,\text{h} = 30.000\,\text{h}$ GPU hours were required for the MD simulations. In addition, the preparation of the ligands including the protonation and initial docking within the pocket, as well as the final MMGBSA and MMPBSA computations were conducted on CPU-only compute nodes since these workflows do not utilize a GPU. The MMGBSA computations were conducted in approximately 2 minutes, while the MMPBSA computations were conducted in 10 minutes taking 5 times longer than MMGBSA. Protonation required negligible time per ligand, whereas docking times varied significantly with ligand size.

References

(1) 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.

(2) Tan, C.; Yang, L.; Luo, R. How Well Does PoissonBoltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis. *The Journal of Physical Chemistry B* **2006**, *110*, 18680–18687, PMID: 16970499.