Machine Learning Guided AQFEP: A Fast & Efficient Absolute Free Energy Perturbation Solution for Virtual Screening

Jordan E. Crivelli-Decker[1*] **Zane Beckwith**[1*] **Gary Tom**[1,2] **Ly Le**[1] **Sheenam Khuttan**[1,3]
Romelia Salomon-Ferrer[1] **Jackson Beall**[1] **Andrea Bortolato**[1]
[1]SandboxAQ [2]University of Toronto [3]Brooklyn College
[*]These authors contributed equally
{jordan.crivelli.decker,zane.beckwith,ly.le}@sandboxquantum.com
{sheenam,romelia.salomon,jackson,andrea.bortolato}@sandboxquantum.com
{gtom}@cs.utoronto.ca

Abstract

Structure-based methods in drug discovery have become an integral part of the modern drug discovery process. The power of virtual screening lies in its ability to rapidly and cost-effectively explore enormous chemical spaces to select promising ligands for further experimental investigation. Relative Free Energy Perturbation (RFEP) and similar methods are the gold standard for binding affinity prediction in drug discovery hit-to-lead and lead optimization phases, but have high computational cost and the requirement of a structural analog with a known activity. Without a reference molecule requirement, Absolute FEP (AFEP) has, in theory, better accuracy for hit ID, but in practice, the slow throughput is not compatible with VS, where fast docking and unreliable scoring functions are still the standard. Here, we present an integrated workflow to virtually screen large and diverse chemical libraries efficiently, combining active learning with a physics-based scoring function based on a fast absolute free energy perturbation method. We validated the performance of the approach in the ranking of structurally related ligands, virtual screening hit rate enrichment, and active learning chemical space exploration; disclosing the largest reported collection of free energy simulations to date.

1 Introduction

In the last decade, the impact of structure-based methods in drug discovery has increased thanks to improvements in the accuracy of force fields, GPU computing, and machine learning[WKS+23]. A prime example of these advances is the development and implementation of structure-based virtual screening (VS) methods, which have become an integral part of the modern drug discovery process. The effectiveness of VS lies in its capacity to quickly and cost-efficiently examine vast chemical landscapes, reducing potential drug candidates from millions to a more manageable number for subsequent analysis. This capability has profound implications for drug discovery, potentially saving significant amounts of time and resources[FXH+22]. Despite its considerable advantages, virtual screening also poses challenges and limitations, particularly concerning the accuracy of predictions.

One of the biggest limitations of VS is the scoring functions used to predict binding affinity. These scoring functions are often rough approximations and usually do not accurately predict the true binding energy or correct binding pose [BMD23]. They often struggle to balance speed and accuracy, making it difficult to reduce both false positives and negatives in VS applications [BM23]. Additionally, in ligand-based virtual screening, structurally similar compounds are often assumed to have similar

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

activity. However, minor structural changes can sometimes lead to significant changes in activity, a phenomenon known as an “activity cliff", which can dramatically impact predictive accuracy[HLR23].

In that context, binding free energy calculation methods offer a unique advantage in the field of computer-aided drug design and virtual screening. Binding free energy calculations can be more accurate than simple docking scores as they consider not just the static interaction of the ligand and protein, but also the dynamic changes that occur upon binding, including conformational changes and solvation effects[AWH$^+$17]. Relative Free Energy Perturbation (RFEP) and similar methods are the gold standard for binding affinity prediction in drug discovery hit-to-lead and lead optimization phases. However, RFEP's computational cost hampers its application to large libraries and the requirement of a structural analog with a known activity does not allow RFEP to be used in hit identification campaigns[KBD$^+$19]. Without a reference molecule requirement, Absolute FEP (AFEP) has, in theory, better accuracy for hit ID, but in practice, the slow throughput is not compatible with virtual screening, where fast docking and simple, unreliable scoring functions are still the standard[HG21].

To overcome the computational cost and low throughput that have made the application of AFEP to virtual screening campaigns infeasible, we make two advancements. First, we present AQFEP, a novel physics-based function based on a fast absolute free energy perturbation method that shows superior ranking performance when compared to other standard scoring functions. This advancement allows us to screen tens of thousands of molecules in a fraction of the time compared to other absolute free energy methods. Second, we combine this increased *in-silico* screening throughput with Bayesian optimization algorithms, to substantially decrease the computational cost of screening the majority of top-scoring compounds with AQFEP. We analyze various surrogate model architectures and acquisition functions to evaluate their effectiveness in efficiently screening the most promising ligands with AQFEP in the context of a virtual screen for drug discovery. We perform these analyses on both straightforward and challenging systems to demonstrate the flexibility of our approach. Finally, we utilize this workflow on a prospective virtual screen through a 1.17m compound library for hit-finding on a novel protein target.

2 Methods

2.1 System Preparation and Molecular Docking

The published[SBB$^+$20] 3D coordinates of the cMet protein structure and congeneric ligand set were downloaded from `https://github.com/MCompChem/fep-benchmark/tree/master/cmet`. The 3D structure of GLP1R was downloaded from the Protein Data Bank[BBB$^+$22] (pdb ID 7S15), and hydrogens were added using the open-source PyMol software[Sch15] and visually inspected. The GLP1R agonist ligands considered in this study[GEF$^+$22], were drawn using JSME[BE13] to generate the SMILES. Using RDKit[L$^+$13] in Knime[BCD$^+$09] the 3D structures were created, hydrogens added, and protonation states were generated using in-house rule-based SMIRKS and visually inspected.

The commercially available compound library used in the prospective screen, was generated starting from the MCULE in stock database `https://mcule.com/database/`, including more than 5M ligands. The library was deduplicated as described above, and filtered in Knime to ensure drug-like properties, removing reactive groups and unwanted chemical moieties. The library was visually inspected and further curated using Datawarrior[SFvKR15]. The final set including 1.177M ligands was prepared for ligand docking as described above. Part of this library was used for the cMet and GLP1R VS tests. GNINA 1.0[SK21] with the Vinardo[QV16] scoring function was used for all molecular docking studies.

2.2 AQFEP

AQFEP uses an absolute free energy perturbation calculation based on the double-decoupling alchemical protocol (Fig. 5). The double-decoupling approach is considered the “gold standard" for absolute free energy calculations by ensuring thermodynamic consistency, accurate sampling of the free-energy landscape, and wide applicability to a variety of different systems. For a more detailed description of this method and its advantages, we refer the reader to a comprehensive text on the topic [MABM$^+$20]. By using an absolute free energy calculation, which directly calculates the binding free energy of the given ligand-protein pair, the procedure requires much less human

guidance than the more typical relative free energy calculations. RFEP calculates differences in binding free energies between pairs of ligands and thus requires careful selection of congeneric ligand pairs and specification of the scaffold relationship between the ligands. While absolute free energy calculations are known for being difficult to converge and giving inconsistent results unless run for very long simulation times [CCJ$^+$], the AQFEP method is tuned to reduce simulation noise and allows for significantly reduced simulation time.

In order to achieve the speeds needed for free energy simulations to be applied to VS workflows we made several adjustments to AQFEP. First, the simulation time per lambda window was chosen to be shorter than standard free energy perturbation calculations (usually 5 ns[CCJ$^+$]), in order to evaluate the energy minimum closer to the provided complex conformation. For the method to perform as a scoring function for a given ligand pose, it must ensure it is evaluating a thermodynamic state indicative of that pose. In spirit, this is quite different from attempting to sample the full state space in the hope of calculating the partition function and measuring the binding free energy of the ligand-protein pair, which would ignore the given pose of the ligand. In practice, sufficiently sampling the state space for such a calculation requires far too much simulation time, and running the simulations longer than is done in AQFEP only allows the ligand to briefly sample conformations too dissimilar from the proposed pose, significantly increasing the statistical noise. Due to this design choice, the method is very dependent on the quality of the proposed ligand pose(s) which can impact ranking performance (cf. Section System Preparation and Molecular Docking). Despite this, AQFEP performs a rigorous, multi-step alchemical transformation, unlike end-point methods such as MMGBSA or MMPBSA[BFMM13], and is thus significantly more accurate and physically realistic than such end-point methods.

2.3 Chemical Space Search Strategy

Bayesian optimization[PP05] is a subset of active learning that helps guide the choice of experiments based on some surrogate models' predictions. Formally, we seek to find the set of top-k molecules (M) from a chemical library ($\mathcal{D}$) that maximizes some black-box objective function (here AQFEP of a candidate compound). Top-k scoring molecules $x \in M$ where M is such that: $argmax_{M \subset \mathcal{D}:|M|=k} \sum_{x \in M} f(x)$

We begin by first calculating the objective function $f(x)$ on a set of n points. The evaluations of this function are stored in the dataset L which contains the labeled observations (AQFEP score for a given candidate compound). A surrogate model $\hat{f}(x)$ is trained with this dataset and makes predictions on the remaining unlabeled set of data D'. The model predictions are passed to an acquisition function α that determines the utility of acquiring new data points to be labeled. These selected points are then evaluated and added back into the labeled set L. This process is repeated until some stopping criteria are met. In this work, we use a fixed budget size divided into T steps to determine when the algorithm should stop acquiring labels.

Algorithm 1 Bayesian Optimization

Input: objective function $f(x)$, acquisition function α, surrogate model $\hat{f}(x)$, and some chemical library D
Select random batch $S \subset D$
Evaluate objective $f(x)$ to generate labels y_s for $s \in S$
Initialize L, the labeled set of data (s, y_s) for $s \in S$
for $t \leftarrow 1$ to T **do**
    Train surrogate model $\hat{f}(x)$ using labeled dataset L
    Select new batch $S_t \subset D$ using acquisition function α
    Evaluate objective function $f(x)$ on S_t
    Update L with new labeled batch
end for

To evaluate different surrogate models' performance, we compared two commonly used search strategies: random search and top dockers. In this case, random search is indicative of an exhaustive search strategy where every compound has an equal probability of being evaluated. In addition, we also evaluated the performance of selecting compounds based on their docking scores (Top Dockers)

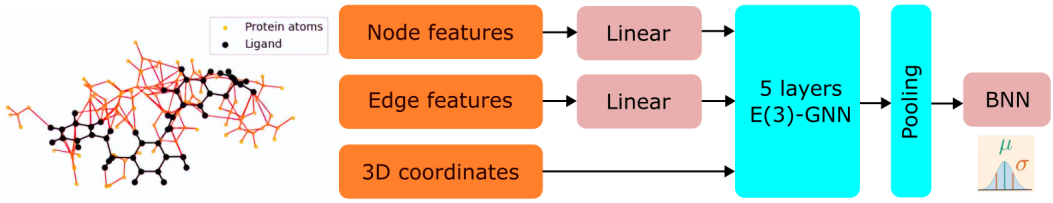

Figure 1: The GraphDock architecture. EGNN take 3 inputs: positions, node embeddings, and edge embeddings. If there are n atoms in an input structure, the position tensor is an n × 3 tensor containing the x, y, and z coordinates of each atom. Edges in the protein-ligand graph are generated using a 3Å cutoff distance. Nodes that fall outside of this 3Å radius are ignored. The protein-ligand graph is then further truncated to include only atoms within 3 nearest-neighbor connections to the ligand atoms in order to reduce the complexity of the graph.

obtained from the GNINA docking software with the Vinardo scoring function. Due to computational limitations, we did not re-dock the compounds with different random seeds and in the figures reported in the paper, there are no error bars reported for this condition.

2.4 Surrogate Models

In this work, we utilized three different surrogate models: Random Forest regression using Morgan fingerprints and D-MPNN as described by [YSJ+19], and an architecture inspired by [SVC+23], GraphDock. Further information regarding surrogate model implementation and details regarding hyperparameters can be found in the Appendix. Below we present a high-level overview of GraphDock.

2.4.1 GraphDock

In this work, we implement a 3D graph neural network, similar to the PointVS model described in [SVC+23](Fig. 1). This model is a lightweight E(n) equivariant graph neural network model, that operates on the 3D protein-ligand complex. The E(n) GNN has demonstrated state-of-the-art performance in regression and classification of chemical datasets [SHW21] while avoiding the use of computationally expensive spherical harmonics. The model is capable of exploiting the symmetries in the protein-ligand complex without the need for augmenting the dataset with translations, rotations, and reflections, as is required for non-equivariant networks such as 3D convolutional neural networks [FMS+20].

We make several adjustments to our model to make it more effective for small datasets and easier to use in Bayesian optimization algorithms where prediction uncertainties are needed. In the PointVS paper, they use 48 EGNN layers, we reduce the number to 5 layers in this work to reduce computational complexity and improve training time[SVC+23]. In addition, we append a single Bayesian linear layer, with a Gaussian prior, for use in regression tasks and in order to save computational costs associated with estimating uncertainty from the model[KES22, WVB+18]. Protein-ligand graphs were constructed using the same protein and ligand conformation as used for AQFEP scoring. In addition, edges in the protein-ligand graph were generated using a 3Å cutoff distance. This cutoff was selected because it approximates inter-molecular interactions such as hydrogen bonding within the complex. The connected protein-ligand complex is then truncated to only atoms within 3 nearest-neighbor connections to the ligand atoms.For more information on graph featurization and other hyperparameters, readers can turn to the appendix.

3 Results and Discussion

3.1 AQFEP: Superior Ranking Performance

We tested the accuracy of AQFEP on the cMet protein kinase using two common applications of interest for drug discovery: 1) evaluation of the free energy of binding on a set of congeneric ligands; 2) application to virtual screening.

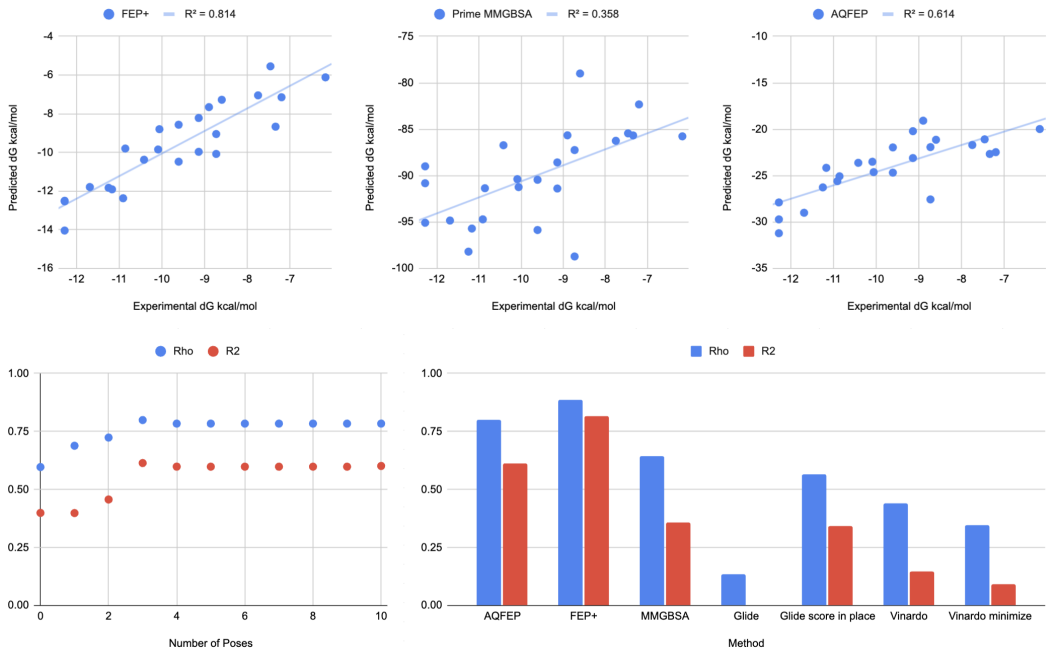

Figure 2: Comparison of the predicted vs experimental free energy of binding for FEP+[SBB$^+$20] (top left), Prime MMGBSA[SBB$^+$20] (top center), and AQFEP (top right). The effect of binding pose prediction on AQFEP is shown in the bottom left plot. The X-axis includes the number of top poses from GNINA/Vinardo scoring function profiled with AQFEP for each ligand. On the Y-axis is reported the corresponding R^2 and Spearman (Rho) correlation based on the best AQFEP scoring poses. Pose 0 corresponds to the ligand position from simple ligand alignment available from[SBB$^+$20]. The bottom right plot includes the comparison of the correlation and ranking performance of AQFEP compared to other commonly used methods. Glide R^2=0.002 is close to the X-axis and therefore not visible in the plot.

3.1.1 Congeneric ligand test

RFEP is the state-of-the-art approach for evaluating the difference in free energy of binding of related ligands, and a large number of benchmark sets are available in the literature[RLS$^+$22]. However, it has several limitations that make it challenging to apply in certain contexts. For example, RFEP is known to be challenging for large perturbations, changes affecting opposite regions of the molecule simultaneously, charge perturbations, large scaffold changes, and perturbations affecting a linker region in the molecule[MAM$^+$20]. AQFEP can be used in all those cases and it does not require an expert user to carefully superimpose the ligand to the reference compound. When evaluating performance on the cMet benchmark set[SBB$^+$20], we find promising ranking performance(Fig. 2), that is close to FEP+, but at a fraction of the computational cost. The results show clear improvements compared to MMGBSA, or molecular docking (Glide and GNINA/Vinardo).

Since AQFEP's MD simulation time is limited for each lambda window to maximize speed, the final prediction will be strongly influenced by the energy minimum closest to the starting conformation. If the starting conformation is far from the physically most important energy well, it is questionable if MD, even with enhanced sampling, will be the best approach to identify the correct lowest energy configuration. To quantify the impact of incorrect ligand conformation on downstream affinity predictions using AQFEP, we re-scored the top 10 poses generated from GNINA/Vinardo in the cMet benchmark described above. Re-scoring with AQFEP the second and third top poses from GNINA/Vinardo (Fig. 2) improved the ranking performance of our method. Considering additional poses beyond 3 did not result in large changes in the ranking performance of the method. For this particular case, using the Vinardo top-scoring pose for each ligand in AQFEP resulted in better performance compared to re-scoring ligand positions generated from ligand alignment (shown as pose 0 in the plot in Fig. 2).

Ligand	IC_{50} (μM)	Vinardo Rank	AQFEP Rank
CHEMBL3402751	2.1	550	13
CHEMBL3402747	3.4	53	20
CHEMBL3402755	4.2	9	7
CHEMBL3402748	5.3	124	9
CHEMBL3402752	30	870	49

Table 1: Ranking performance of the 5 active ligands for AQFEP vs. Docking using Vinardo

3.1.2 VS enrichment test

We created a realistic VS test set including 1 active ligand for every 54000 commercially available drug-like compounds. The 5 low micromolar inhibitors of cMet were selected from the previously discussed congeneric series test set[SBB$^+$20]. For this evaluation we used a classical funnel approach, selecting the top 1000 ligands from molecular docking for AQFEP rescoring. Protein kinases are known to be ideal cases for molecular docking and the Vinardo scoring function was able to rank the 5 ligands in the top 1k compounds (Fig. 6). Rescoring with AQFEP identified 2 actives among the top 10 compounds and all 5 active ligands in the top 50 (Table 1). This analysis showed a clear advantage for AQFEP in early enrichment ranking performance when compared to Vinardo.

3.2 ML Driven Chemical Space Exploration Maximizes Efficiency and Hit Rate

3.2.1 Retrospective Test - GLP1R

Despite large improvements in computational efficiency, AQFEP in practice is still too slow to be applied to the ultra-large virtual screens that have now become the norm. For example, ZINC, a popular open-source database of commercially available compounds, now has close to 30 billion entries [TTC$^+$23]. These libraries are challenging to screen against even with standard structure-based drug design tools (e.g. docking) and were previously thought to be unattainable for free energy calculations. Search optimization strategies exist to efficiently search through chemical libraries for compounds that have desirable properties, such as active learning or Bayesian Optimization (BO). However, due to the computational cost of free energy methods, generating sufficiently large quantities of data needed to train models that are broadly generalizable to vast chemical space is not possible. Bayesian optimization and active learning have been previously applied to virtual screening campaigns using molecular docking [BKK$^+$21, GSC21] and RFEP [TWF$^+$22, KBD$^+$19], but to our knowledge, we disclose its application to the largest set of AFEP calculations to date.

To mitigate the impact of growing library size, we employ a Bayesian optimization strategy to perform a model-guided search and seek to find a set of top-k molecules that have the lowest binding affinity as measured with AQFEP. We tested this approach (c.f. Section Bayesian Optimization) on the GLP1R receptor, a challenging Family B GPCR membrane protein related to type 2 diabetes[JRB$^+$17]. We included two weak agonists as actives[GEF$^+$22] in a library of more than 12,720 commercially available ligands with drug-like properties. As an initial evaluation to demonstrate how AQFEP and Bayesian optimization algorithms can be used together, we calculated AQFEP scores for all compounds in this library. We also included in this library 2 known active molecules to evaluate the ability of our method to screen known actives in the presence of a large number of decoys. Data acquisition was simulated with an initial random 2% selection followed by sequential 2% acquisitions for 5 iterations totaling a 12% total library screen. We test several surrogate models that are typically used in Bayesian optimization strategies including Random Forest (RF) with Morgan fingerprints and D-MPNN [YSJ$^+$19]. In addition, we also compare these to a 3D Graph Neural Network architecture (GraphDock). To quantify the advantages of model-guided search, we benchmarked our approach against random search, indicative of an exhaustive search through the library, and the top-scoring compounds from the docking program GNINA using the Vinardo scoring function (see Methods).

Bayesian optimization using any of the surrogate models tested, yields clear improvements over other heuristic search algorithms. The RF model operating on molecular fingerprints showed similar performance across acquisition functions in the top-k task (Fig. 3 and Table 4). A similar pattern emerges when examining the D-MPNN surrogate with both greedy strategies performing best. The GraphDock model performed best when compared to other surrogates with UCB finding 322 of the top 500, on average.

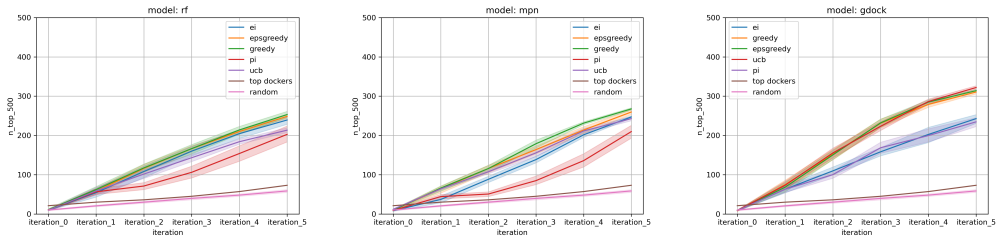

Figure 3: Bayesian optimization performance on GLP1R retrospective test as measured by the number of top 500 scores found as a function of total compounds evaluated. Each trace represents the performance of a given surrogate with a specific acquisition function. Each experiment began with 2% random selection followed by 5 iterations of 2%, for a total of 12% (1530 compounds). The total library size was 12,720. Error bars reflect $\pm$ one standard deviation across 10 runs.

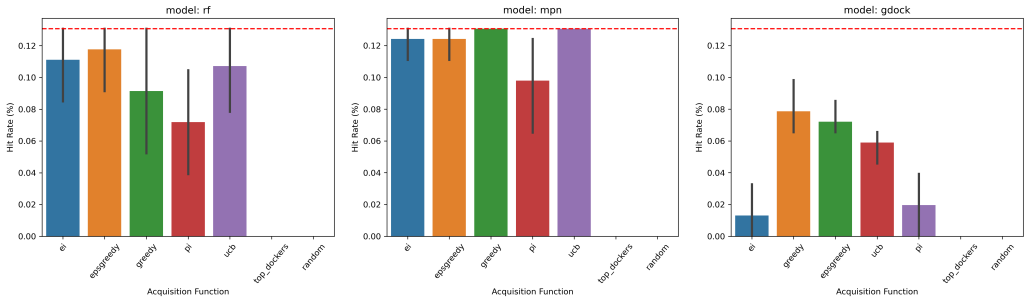

Figure 4: Bayesian optimization performance on GLP1R retrospective test as measured by the average hit rate across 10 runs. Each bar represents the performance of a given surrogate with a specific acquisition function and indicates how often a given surrogate-acquisition pairing can find both known actives in the library. The red dashed line indicates the theoretical max performance (e.g. finding both actives in all 10 runs). The red dashed line is 0.13% (2/1530). Error bars reflect $\pm$ one standard deviation across 10 runs.

An alternative way of understanding the performance of the surrogate model is to calculate the number of times each screening workflow identified one of the known actives placed within this library (Fig 4 and Table 5). The RF model utilizing Morgan fingerprints performed remarkably well with the ϵ-greedy acquisition function having an average hit rate of 0.12 $\pm$ 0.04. When using the D-MPNN as a surrogate, two different greedy acquisition functions found both known actives in 10/10 runs (max hit rate is 0.13% 2/1530). Interestingly, while GraphDock performed quite well in the top-k task, this model's hit rate was low when compared to other surrogates. This effect may be due to the structural similarity of the two active compounds helping a 2D approach (D-MPNN and RF) to identify both simultaneously. In contrast, GraphDock, is affected by the quality of the docked poses used to train the model. To further investigate, we examined the individual poses of the known actives that were generated with GNINA/Vinardo. We found that one of the two active compounds had a predicted pose that was quite different from the X-ray conformation of related agonist compounds. In addition, AQFEP scores for the active compounds in this library had relatively low ranks and both active compounds did not place in the top 500 in the precomputed library of 12,720 (Active 1 FEP Rank=1102, Active 2 FEP Rank=1737). It is likely that this wrong pose, combined with relatively low AQFEP scores for the actives, hampered the ability of GraphDock to identify this active ligand. Finally, both random search and docking fail to screen any known active ligands.

3.2.2 Prospective Test - Novel Protein Target

We next turned to evaluate the effectiveness of this workflow applied to a prospective search in a much larger chemical space. Here, we screened a 1.17m compound library of commercially available compounds (MCULE) against a novel protein target. In this experiment, we performed two internal controls to show that our method was intelligently sampling the chemical space. First, we randomly selected selected 10,000 compounds to evaluate the number of compounds that have low AQFEP

scores. Second, we docked the entire library (1.17m) with GNINA and selected 10,000 compounds with the lowest docking scores. From here we can establish a baseline on how conventional heuristic selection strategies score with AQFEP. To ensure the highest quality labels for our algorithm, we removed AQFEP scores from this set if the simulation had a convergence error of greater than 1 kcal/mol. This resulted in 7,542 and 8,930 ligands in the Random and Top Dockers selection respectively.

We used this heterogeneous sample (random selection plus top-dockers) to serve as our initial seed set of labeled data in our active learning workflow. We performed 3 iterations of active learning with a 10,000 compound acquisition size. In each iteration, we also examined the convergence error for our AQFEP scores, and scores with greater than 1 kcal/mol error were removed. This resulted in 6,792, 3,317, and 3,527 compounds profiled in tranches 3 through 5 respectively. In total, we performed 30,108 free energy simulations throughout our workflow, representing to our knowledge the largest reported free energy screens to date. We can understand how well our algorithm is performing by looking for changes in the median AQFEP score across iterations. If our algorithm is performing as intended, we should see the median AQFEP score decrease across Bayesian optimization iterations. Utilizing a D-MPNN surrogate with a greedy acquisition function we can see in Figure 7, that across iterations, model-guided selection selects consistently lower-scoring compounds across iterations, and is substantially better than both internal controls. Based on the analysis of the AQFEP free energy of binding distribution for the random sample we defined a promising ligand to have a predicted score of -20 kcal/mol. The number of such promising ligands identified in each iteration (Table 2) showcases the efficiency of the proposed method.

Search Strategy	**Number of Ligands identified with AQFEP<-20 kcal/mol**
Random	16
Top Dockers	46
Tranche 3 (D-MPNN)	305
Tranche 4 (D-MPNN)	351
Tranche 5 (D-MPNN)	267

Table 2: Number of ligands with low AQFEP scores across tranches

4 Conclusion

In this work, we describe AQFEP, a physics-based approach to evaluating the free energy of binding for diverse ligands to protein targets that is superior to molecular docking scoring functions in predicting ligand binding free energy. It balances speed and ligand ranking accuracy. The speed of AQFEP enables the virtual screening of libraries of tens of thousands of ligands at unprecedented accuracy in a target-independent way. Compared to RFEP, AQFEP does not require a similar compound with known activity, it can be used to score a pose directly after docking without the need for careful alignment to a congeneric compound. It is significantly faster, with one ligand typically taking 1-2 hours on one Nvidia T4 GPU, approximately 10 times faster than RFEP and 40-70 times faster than other AFEP solutions on comparable hardware.

Most importantly, the speed of AQFEP also enables the generation of high-quality labeled datasets large enough for the training of a variety of supervised machine learning models. These models achieve speeds comparable to standard docking scoring functions. A pose-dependent 3D-EGNN using AQFEP labels can be used as a surrogate model in a Bayesian optimization framework enabling efficient search through libraries of millions of ligands. This potentially reduces the need for profiling additional ligands with AQFEP to only those that are structurally novel.

This ML-guided search of chemical space using AQFEP as an objective function shows improvements in both hit rate and percent of top compounds screened when compared to random and other heuristic search algorithms (top dockers). Across Bayesian optimization iterations in a prospective search, a greedy D-MPNN surrogate model selects lower-scoring compounds than other selection strategies (top dockers and random search).

Together, this work demonstrates the effectiveness of the unification of a fast and accurate physics-based scoring function with BO algorithms to unlock the capability to perform large virtual screens using free energy calculations.

References

[AWH+17] Robert Abel, Lingle Wang, Edward D Harder, BJ Berne, and Richard A Friesner. Advancing drug discovery through enhanced free energy calculations. *Accounts of chemical research*, 50(7):1625–1632, 2017.

[BBB+22] Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Jose M Duarte, Shuchismita Dutta, Maryam Fayazi, Zukang Feng, et al. Rcsb protein data bank: Celebrating 50 years of the pdb with new tools for understanding and visualizing biological macromolecules in 3d. *Protein Science*, 31(1):187–208, 2022.

[BCD+09] Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Kilian Thiel, and Bernd Wiswedel. Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1):26–31, November 2009.

[BE13] Bruno Bienfait and Peter Ertl. Jsme: a free molecule editor in javascript. *Journal of cheminformatics*, 5(1):1–6, 2013.

[BFMM13] Andrea Bortolato, Marco Fanton, Jonathan S Mason, and Stefano Moro. Molecular docking methodologies. *Biomolecular Simulations: Methods and Protocols*, pages 339–360, 2013.

[BKK+21] Flavio Ballante, Albert J Kooistra, Stefanie Kampen, Chris de Graaf, and Jens Carlsson. Structure-based virtual screening for ligands of g protein–coupled receptors: what can molecular docking do for you? *Pharmacological Reviews*, 73(4):1698–1736, 2021.

[BM23] Davide Bassani and Stefano Moro. Past, present, and future perspectives on computer-aided drug design methodologies. *Molecules*, 28(9):3906, 2023.

[BMD23] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *arXiv preprint arXiv:2308.05777*, 2023.

[CCJ+] Wei Chen, Di Cui, Steven V Jerome, Mayako Michino, Eelke B Lenselink, David J Huggins, Alexandre Beautrait, Jeremie Vendome, Robert Abel, Richard A Friesner, et al. Enhancing hit discovery in virtual screening through absolute protein–ligand binding free-energy calculations. *Journal of Chemical Information and Modeling*.

[FMS+20] Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, 2020. PMID: 32865404.

[FXH+22] Elissa A Fink, Jun Xu, Harald Hübner, Joao M Braz, Philipp Seemann, Charlotte Avet, Veronica Craik, Dorothee Weikert, Maximilian F Schmidt, Chase M Webb, et al. Structure-based discovery of nonopioid analgesics acting through the α2a-adrenergic receptor. *Science*, 377(6614):eabn7065, 2022.

[GEF+22] David A Griffith, David J Edmonds, Jean-Philippe Fortin, Amit S Kalgutkar, J Brent Kuzmiski, Paula M Loria, Aditi R Saxena, Scott W Bagley, Clare Buckeridge, John M Curto, et al. A small-molecule oral agonist of the human glucagon-like peptide-1 receptor. *Journal of Medicinal Chemistry*, 65(12):8208–8226, 2022.

[GSC21] DE Graff, EI Shakhnovich, and CW Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning, chem, 2021.

[HG21] Germano Heinzelmann and Michael K Gilson. Automation of absolute protein-ligand binding free energy calculations for docking refinement and compound evaluation. *Scientific reports*, 11(1):1116, 2021.

[HLR23] Sophia MN Hönig, Christian Lemmen, and Matthias Rarey. Small molecule superposition: A comprehensive overview on pose scoring of the latest methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 13(2):e1640, 2023.

[HSY+20] Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020. PMID: 32702986.

[JRB+17] Ali Jazayeri, Mathieu Rappas, Alastair JH Brown, James Kean, James C Errey, Nathan J Robertson, Cédric Fiez-Vandal, Stephen P Andrews, Miles Congreve, Andrea Bortolato, et al. Crystal structure of the glp-1 receptor bound to a peptide agonist. *Nature*, 546(7657):254–258, 2017.

[KBD+19] Kyle D Konze, Pieter H Bos, Markus K Dahlgren, Karl Leswing, Ivan Tubert-Brohman, Andrea Bortolato, Braxton Robbason, Robert Abel, and Sathesh Bhat. Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin-dependent kinase 2 inhibitors. *Journal of chemical information and modeling*, 59(9):3782–3793, 2019.

[KES22] Ranganath Krishnan, Pi Esposito, and Mahesh Subedar. Bayesian-torch: Bayesian neural network layers for uncertainty estimation. `https://github.com/IntelLabs/bayesian-torch`, January 2022.

[L+13] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

[MABM+20] Antonia S. J. S. Mey, Bryce K. Allen, Hannah E. Bruce McDonald, John D. Chodera, Davi d F. Hahn, Maxmillian Kuhn, Julien Michel, David L. Mobley, Levi N. Naden, Samarjeet Prasad, Andrea Rizzi, Jenke Scheen, Michael R. Shirts, Gary Tresadern, and Huafeng Xu. Best practices for alchemical free energy calculations [article v1.0]. *Living Journal of Computational Molecular Science*, 2(1):18378, Dec. 2020.

[MAM+20] Antonia SJS Mey, Bryce K Allen, Hannah E Bruce Macdonald, John D Chodera, David F Hahn, Maximilian Kuhn, Julien Michel, David L Mobley, Levi N Naden, Samarjeet Prasad, et al. Best practices for alchemical free energy calculations [article v1. 0]. *Living journal of computational molecular science*, 2(1), 2020.

[NW94] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. 1:55–60 vol.1, 1994.

[PP05] Martin Pelikan and Martin Pelikan. Bayesian optimization algorithm. *Hierarchical Bayesian Optimization Algorithm: Toward a New Generation of Evolutionary Algorithms*, pages 31–48, 2005.

[QV16] Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

[RLS+22] Gregory Ross, Chao Lu, Guido Scarabelli, Steven Albanese, Evelyne Houang, Robert Abel, Edward Harder, and Lingle Wang. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. 2022.

[SBB+20] Christina EM Schindler, Hannah Baumann, Andreas Blum, Dietrich Böse, Hans-Peter Buchstaller, Lars Burgdorf, Daniel Cappel, Eugene Chekler, Paul Czodrowski, Dieter Dorsch, et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *Journal of Chemical Information and Modeling*, 60(11):5457–5474, 2020.

[Sch15] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[SFvKR15] Thomas Sander, Joel Freyss, Modest von Korff, and Christian Rufener. Datawarrior: an open-source program for chemistry aware data visualization and analysis. *Journal of chemical information and modeling*, 55(2):460–473, 2015.

[SHW21] Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. 139:9323–9332, 18–24 Jul 2021.

[SK21] Jocelyn Sunseri and David Ryan Koes. Virtual screening with gnina 1.0. *Molecules*, 26(23):7369, 2021.

[SLT+03] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003. PMID: 14632445.

[SSW+16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1):148–175, January 2016.

[SVC+23] Jack Scantlebury, Lucy Vost, Anna Carbery, Thomas E Hadfield, Oliver M Turnbull, Nathan Brown, Vijil Chenthamarakshan, Payel Das, Harold Grosjean, Frank von Delft, et al. A small step toward generalizability: Training a machine learning scoring function for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 2023.

[TTC+23] Benjamin I. Tingle, Khanh G. Tang, Mar Castanon, John J. Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S. Moroz, and John J. Irwin. Zinc-22–a free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of Chemical Information and Modeling*, 63(4):1166–1176, 2023. PMID: 36790087.

[TWF+22] James Thompson, W Patrick Walters, Jianwen A Feng, Nicolas A Pabon, Hongcheng Xu, Brian B Goldman, Demetri Moustakas, Molly Schmidt, and Forrest York. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences*, 2:100050, 2022.

[VSP+23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[WKS+23] Ke Wu, Eduard Karapetyan, John Schloss, Jaydutt Vadgama, and Yong Wu. Advancements in small molecule drug design: A structural perspective. *Drug Discovery Today*, page 103730, 2023.

[WVB+18] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches, 2018.

[YSJ+19] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

5 Appendix

5.1 AQFEP

AQFEP double-decoupling alchemical protocol is shown in Fig. 5. A general comparison of computational methods to predict ligand-protein free energy of binding is reported in Table 3.

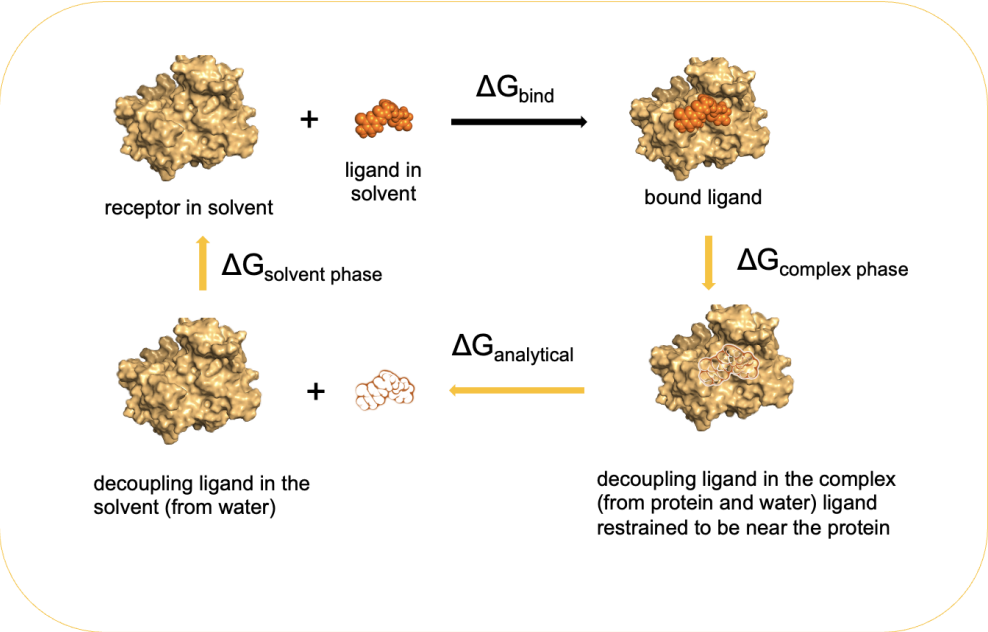

Figure 5: Double-decoupling alchemical protocol

Method	Time per ligand on 1 T4 GPU	Congeneric ligand required
Docking	seconds/minutes	No
MMGBSA	minutes/hours	No
AQFEP	**1-2 hours**	**No**
RFEP	1-2 days	Yes
AFEP	1 week	No

Table 3: General computational time comparison

5.1.1 Speed Comparison to Other Methods

To increase the ligand throughput of AQFEP, so that it could be used within a VS workflow, several changes were implemented in AQFEP. Most notably, the MD simulation time for each lambda window was chosen to be shorter than standard free energy perturbation calculations. This adjustment allows AQFEP to be used at a throughput much higher than traditional free energy perturbation methods, allowing the virtual screening of libraries of tens of thousands of ligands. For instance, using 20k NVIDIA T4 GPUs, AQFEP can reach 10-20k ligands profiled per hour (Table 3). However, this design choice causes the method to be very dependent on the quality of the proposed ligand pose(s) which can impact ranking performance. In the sections that follow, we highlight this effect and show its impact on downstream scoring with AQFEP.

5.2 Bayesian Optimization

5.2.1 Acquisition Functions

Below we provide brief descriptions of the acquisition functions used in this work. We refer the reader to other works that describe these acquisition functions in more detail[SSW+16].

In this work, we evaluated the following acquisition functions:

$$Random(x) \sim U(0,1) \tag{1}$$

$$Greedy(x) = \hat{\mu}(x) \tag{2}$$

Greedy is a naive method where the top-scoring molecules ranked by the model are always selected for evaluation, without taking into account uncertainty.

$$EpsilonGreedy(x) = \begin{cases} \hat{\mu}(x), \text{ with probability } 1-\epsilon \\ U(0,1), \text{ with probability } \epsilon \end{cases} \tag{3}$$

Epsilon greedy is analogous to the greedy strategy except that it makes a random selection from the library with probability ϵ. In experiments, we report, a value of $\epsilon = 0.1$ was used.

$$UCB(x) = \hat{\mu}(x) + \beta\hat{\sigma}(x) \tag{4}$$

The upper confidence bound acquisition policy is an 'optimistic' method that selects molecules based on their potential to yield optimal values. The utility of acquiring a given molecule is calculated by summing the predicted mean value with its predicted standard deviation. It attempts to balance exploration and exploitation by enabling molecules to be selected that have moderate mean predicted scores but large standard deviations. The β parameter can be adjusted to more heavily weight the standard deviation term. In the experiments we report, a value of $\beta = 2$ was used.

$$PI(x) = \begin{cases} \Phi(z),\ \hat{\sigma}(x) > 0 \\ 1,\ \hat{\sigma}(x) = 0 \text{ and } \gamma(x) > 0 \\ 0,\ \hat{\sigma}(x) = 0 \text{ and } \gamma(x) <= 0 \end{cases} \tag{5}$$

The probability of improvement acquisition policy aims to select the molecule that has the highest probability of improving upon the currently identified best score. The PI score for a molecule is computed utilizing the standard deviation associated with that molecule to compute the amount of probability mass that molecule has above the current best solution. It is important to note that PI does not consider the magnitude of the improvement.

$$EI(x) = \begin{cases} \gamma(x)\Phi(z) + \hat{\sigma}(x)\phi(z),\ \hat{\sigma} > 0 \\ \gamma(x),\ \hat{\sigma} = 0 \end{cases} \tag{6}$$

The expected improvement policy is analogous to the PI policy but considers the magnitude of the improvement. The value is computed for a molecule by calculating the expected value of the probability density that the molecule has above the current best solution. The EI method can be augmented with a parameter that can encourage more exploration.

For PI and EI: $\gamma(x) := \hat{\mu}(x) - f^* + \xi$; $z(x) := \frac{\gamma(x)}{\hat{\sigma}(x)}$; $\hat{\mu}(x)$ and $\hat{\sigma}^2$ are the surrogate models predicted mean and uncertainties for point x, respectively. Φ and ϕ are the CDF and PDF of the standard normal distribution and f^* is the current maximum objective function value. In experiments we report using EI and PI, we use a value of $\xi = 0.01$.

5.2.2 Random Forest

Random forest regression is an ensemble learning technique that utilizes a set of decision trees. Each individual tree is fit with a random subset of the training features and observations in an attempt to de-correlate the trees [SLT+03]. During inference, uncertainty estimates can be derived by examining the mean value of the ensemble of trees and the variance of the predictions from the ensemble. Chemical libraries were featurized with molecular fingerprints. There are a variety of molecular fingerprints that differ in their specific implementations, however, they all can be broadly understood as representing the presence or absence of a specific sub-structure within a molecule into a vector of fixed length. In this work, we utilized the Morgan Fingerprint with a bit length of 2048 and a radius of 3. The RF surrogate used was fit with $n\_estimators = 100$, $max\_depth = 8$.

5.2.3 D-MPNN

In this work, we utilized the directed message-passing neural network (D-MPNN) implemented by [YSJ+19]. MPNNs treat the molecule as a connected graph and construct a feature vector for that graph, de-novo. This is in contrast to fixed fingerprints which do not have the flexibility to adjust their embeddings. Broadly, MPNNs operate in two stages, the message-passing phase and the readout

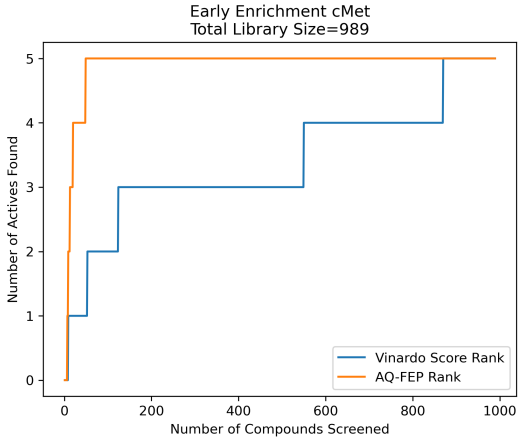

Figure 6: Ranking performance on cMet enrichment test set with 5 known actives. Enrichment using a standard docking workflow performs well and is improved further using AQFEP.

phase. In the message-passing phase, "messages" are passed between atoms and/or bonds and their direct neighbors, and incoming messages are used to update the "hidden state" of each atom and/or bond. The message passing phase is repeated over multiple (e.g., 3) iterations, at which point the hidden states of each atom are aggregated (e.g., summed) to produce a molecule-level feature vector. By training this model at the same time as an FFNN operating on the feature vector, MPNNs are able to learn a task-specific representation of an input molecular graph. For more details on the D-MPNN model and the specific implementation of this architecture, we refer the reader to [YSJ$^+$19].

The message-passing neural network used here utilized standard settings from the molecule model class in the Chemprop library [YSJ$^+$19]: messages passed on directed bonds, messages subjected to ReLU activation, a learned encoded representation of dimension 300, and the output of the message passing phase fully connected to an output layer of size 1. The model was trained using the Adam optimization algorithm, a Noam learning rate scheduler (initial, maximum, and final learning rates of 0.1, 0.001, and 0.0001, respectively), and a root-mean-squared error loss function over 50 epochs with a batch size of 50. For more details on the Noam learning rate scheduler, see [VSP$^+$23]. The model was trained with early stopping tracking the validation score using a patience value of 10. When uncertainty values were needed for metric function calculation, an MVE model based on the work done by [HSY$^+$20] was used. This model featured an output size of two and was trained using the loss function defined by Nix and Weigend [NW94].

5.2.4 GraphDock

GraphDock is composed of 5 layers of E(3)GNN, with 100 hidden dimensions, and the same parameters for chemical prediction tasks specified in [SHW21]. Node and edge representations first pass through separate feed-forward neural networks, each with hidden dimensions 32. The output graph is pooled to give an embedding with 64 dimensions (See figure 1).

The graph featurization step is the same as is implemented in the Chemprop library [YSJ$^+$19], in addition to the 3D coordinates, and the identity of a node (atom in protein or in ligand). The GraphDock model was trained using the Adam optimization algorithm with a batch size of 64 and a fixed learning rate of 5e-4 for 500 epochs. Early stopping was used to prevent over-fitting with a patience of 50.

5.3 cMet Virtual Screen

A comparison of the virtual screening enrichment between the Vinardo scoring function and AQFEP is provided in Fig. 6.

surrogate	acq	mean	std
gdock	ei	242.8	19.7
	epsgreedy	314.0	5.9
	greedy	310.6	8.2
	pi	234.6	18.5
	ucb	**322.0**	7.3
mpn	ei	247.0	5.0
	epsgreedy	259.6	11.2
	greedy	**267.5**	5.0
	pi	209.7	27.4
	ucb	242.7	10.0
rf	ei	239.3	16.9
	epsgreedy	248.6	9.8
	greedy	**253.2**	13.5
	pi	202.7	35.0
	ucb	213.4	10.1
top dockers	top dockers	73.0	NA
random	random	58.6	5.9

Table 4: Average top-k performance on GLP1R retrospective test with a 2% acquisition size as measured by the average number (n) of top smiles evaluated in the pre-computed library across 10 experiment runs. Bolded numbers indicate the best performing method within each surrogate model.

surrogate	acq	mean	std
theoretical max	NA	**0.130**	NA
gdock	ei	0.013	0.028
	epsgreedy	0.072	0.021
	greedy	0.079	0.028
	pi	0.020	0.032
	ucb	0.060	0.021
mpn	ei	0.124	0.021
	epsgreedy	0.124	0.021
	greedy	**0.130**	0.000
	pi	0.098	0.046
	ucb	**0.130**	0.000
rf	ei	0.111	0.044
	epsgreedy	0.118	0.041
	greedy	0.092	0.063
	pi	0.072	0.057
	ucb	0.107	0.044
top dockers	top dockers	0.000	NA
random	random	0.000	0.000

Table 5: Average hit-rate performance on GLP1R retrospective test

5.4 GLP1R Retrospective Test

The average top-k performance on GLP1R retrospective test is included in Table 4. The average hit rate for each of the surrogates and acquistion functions can be found in Table 5

5.4.1 Prospective Screen - Novel Protein Target

Below we illustrate changes in the median AQFEP score across tranches of our active-learning virtual screening workflow applied to a novel protein target.

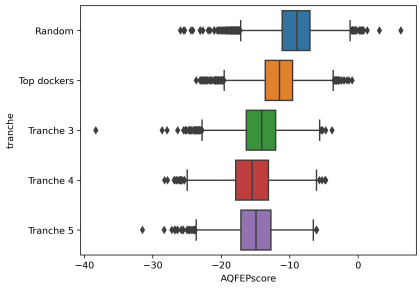

Figure 7: Model guided search results in profiling of compounds with better (lower) AQFEP scores across tranches against a novel protein target. In total 30,108 compounds were profiled which is comprised of 7,542 compounds from Random selection, 8,930 from Top Dockers selection, and 13,636 selected by a greedy D-MPNN surrogate model.