Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries

Toni Sivula,† Laxman Yetukuri,‡ Tuomo Kalliokoski,¶ Heikki Käsnänen,¶
Antti Poso,†,§ and Ina Pöhner*,†

†School of Pharmacy, University of Eastern Finland, FI-70211 Kuopio, Finland
‡CSC – IT Center for Science Ltd., FI-02101 Espoo, Finland
¶Computational Medicine Design, Orion Pharma, Orionintie 1A, FI-02101 Espoo, Finland
§Department of Internal Medicine VIII, University Hospital Tübingen, DE-72076 Tübingen, Germany

E-mail: ina.pohner@uef.fi
Abstract

The emergence of ultra-large screening libraries, filled to the brim with billions of readily available compounds, poses a growing challenge for docking-based virtual screening. Machine Learning (ML)-boosted strategies like the tool HASTEN combine rapid ML prediction with the brute-force docking of small fractions of such libraries to increase screening throughput and take on giga-scale libraries. In our case study of an anti-bacterial chaperone and an antiviral kinase, we first generated a brute-force docking baseline for 1.56 billion compounds in the Enamine REAL lead-like library with the fast Glide HTVS protocol. With HASTEN, we observed robust recall of 90% of the true 1000 top-scoring virtual hits in both targets when docking only 1% of the entire library. This reduction of the required docking experiments by 99% significantly shortens the screening time. In the kinase target, the employment of a hydrogen bonding constraint resulted in a major proportion of unsuccessful docking attempts and hampered ML predictions. We demonstrate the optimization potential in the treatment of failed compounds when performing ML-boosted screening and showcase HASTEN as a fast and robust tool in a growing arsenal of approaches to unlock the chemical space covered by giga-scale screening libraries for everyday drug discovery campaigns.
Introduction

Virtual screening (VS) approaches utilizing molecular docking are a common choice in the early stages of structure-based drug discovery projects. Typically, their objective is to find initial small molecule hits predicted to bind to a previously unexplored target, or to discover novel scaffolds in an unbiased way when confirmed binders of the studied target are already known. Especially in the initial screening steps, VS often utilizes large and diverse screening libraries to pool the most promising candidates from a large chemical space.\textsuperscript{1–3}

Chemical libraries of off-the-shelf or readily synthesizable (make-on-demand) compounds are popular to ensure that the docking predictions can be validated in biochemical assays in a timely manner without the need to factor in potentially time-consuming organic synthesis. In recent years, such libraries have continuously grown and today often cover a vast chemical space with compound numbers on the billion scale.\textsuperscript{4} For example, at the beginning of this project, the Enamine REAL library of lead-like compounds had a size of 1.56 billion (March 2021), whereas the current version has 3.53 billion compounds (January 2023).\textsuperscript{5} Structure-based virtual screening of such libraries has produced hits of exceptional quality for several targets (see, e.g., Stein et al.\textsuperscript{3} and Kaplan et al.\textsuperscript{6}).

What is a huge leap forward in terms of access to diverse chemical space and straightforward validation of docking-based hypotheses, on the other hand, becomes a challenge for docking-based VS: Since conventional brute-force docking visits every compound in a chemical library, screening billions of compounds is often no longer feasible, both in terms of the required time, and the computational power. Thus, there is a clear need for faster and more efficient methods.

There are various approaches to tackle the computational expense associated with billion-scale structure-based virtual screening. One strategy is to grow compounds from fragments instead of docking full-size molecules and thus avoiding the enumeration of large numbers of compounds.\textsuperscript{7,8} Alternatively, several strategies that recently gained traction for boosting docking-based screening rely on iterative approaches utilizing machine learning (ML).\textsuperscript{9–11} The idea is simple, yet powerful: A small fraction of a large chemical library is docked by conventional means and used as training data for a ML model, that in turn predicts the docking scores for the entire library. This way, when
not every compound needs to be docked, ML-boosted screening approaches can handle ultra-large libraries in a fraction of the time, providing the opportunity to explore the vast chemical landscape of giga-scale libraries.

One example of such an iterative approach is the tool HASTEN (macHine leArning booSTEd dockiNg).\textsuperscript{9} HASTEN starts from an initial random compound selection from a large chemical library for a first conventional docking run and trains a ML model with the obtained data. Then, scores for the full compound library are predicted and used to rank all compounds. The best-ranked compounds next get selected for docking in the following iteration. This deliberate bias towards top-scoring compounds has been shown to result in excellent recall for datasets on the million-scale.\textsuperscript{9} We aimed to investigate the applicability and performance of the HASTEN approach on even larger datasets, where the training data already reaches the million-scale and predictions on the giga-scale cover billions of compounds.

To evaluate how the ML-boosted approach performs on giga-scale chemical libraries, such libraries first need to be screened using standard brute-force docking to obtain the baseline data. Such docking data on the giga-scale has so far at least rarely been made publicly available: To date, the only example we are aware of are the Covid-19 screening results obtained with AutoDock-GPU on the Oak Ridge National Laboratory Summit-computer.\textsuperscript{12,13}

Herein, we used the fast Glide HTVS method in one of the largest conventional docking campaigns performed to date, to provide a brute-force docking baseline for analysis and comparison with our ML-boosted approach.\textsuperscript{14} We selected two distinct targets based on ongoing academic drug discovery projects:

The first target, the SurA protein, is a periplasmic chaperone found in Gram-negative bacteria. SurA has prolyl-peptidyl isomerase activity and is involved in the transport and maturation of several outer membrane proteins.\textsuperscript{15–18} Loss of SurA activity has been shown to render resistant bacterial strains sensitive to antibiotics, making SurA an interesting target in combating antibiotic resistance.\textsuperscript{17,19}

Our second target was the Cyclin G-associated kinase (GAK): a serine/threonine kinase, serv-
ing as a regulator of clathrin-mediated endocytosis and clathrin trafficking.\textsuperscript{20–22} GAK represents an important host factor involved in the regulation of viral entry and assembly of different RNA viruses, such as Hepatitis C, Dengue, and Ebola virus, and is of interest as an anti-viral target.\textsuperscript{23–25}

Our study aims to demonstrate the potential of ML-boosted screening with HASTEN for giga-scale applications, showcasing the speed-up, robustness, and successful recall of the majority of top-scoring compounds compared to the brute-force docking results for our two targets. We further demonstrate how large numbers of compounds which fail to dock successfully can hamper the HASTEN approach, and discuss different options to handle such ’failed’ compounds.

Finally, to support future screening approaches, we release a prepared version of the Enamine REAL lead-like screening library used in this study in a Glide-compatible format. Additionally, we release our full giga-scale docking results as benchmarking datasets for the future development and improvement of ML-boosted screening procedures.

**Approach**

**Computational infrastructure.** Computational resources were provided by CSC – IT Center for Science Ltd.\textsuperscript{26} All calculations were performed on the CSC supercomputers Mahti (Atos BullSequana XH2000) and Puhti (Atos BullSequana X400), both running Red Hat Enterprise Linux Server release 7.9.

Ligand preparation and conventional docking steps were done with Mahti. Mahti features 1404 CPU nodes, each equipped with two AMD Rome 7H12 CPUs with 64 physical cores capable of two hardware threads running at 2.6 GHz base frequency, 256 GBs of system memory and Lustre parallel storage system providing a peak file I/O performance of 1.5 GB/s.

Machine learning was conducted on CSC Puhti, which is equipped with 80 GPU nodes, which each have four NVidia Volta V100 GPUs. Puhti GPU nodes further feature two Intel Xeon Cascade Lake 20-core CPUs running at 2.1 GHz, 384 GB of system memory, and a local 3.6 TB NVMe disk.

**Screening database preparation.** The Enamine REAL lead-like (ERLL) library, containing a
total of 1.56 billion compounds (March 2021), was selected for the giga-scale screening. All included compounds have lead-like properties, with molecular weight \( \leq 460 \text{ Da} \), SlogP -4.0 to 4.2, number of hydrogen bond acceptors <10 and donors <5, number of ring systems \( \leq 4 \) and rotatable bonds \( \leq 10 \). The library was obtained in ChemAxon extended SMILES format and converted to regular SMILES using RDKit v2021.03.5, retaining the stereochemical information of the compounds where applicable.

To reduce the variation in docking times between different subsets, compound order first was randomized, and SMILES were then evenly divided into 20 subsets. Ligand 3D structure preparation for docking was carried out with Schrödinger LigPrep (Schrödinger Suite 2021-1). Up to 8 tautomers per compound and 4 stereoisomers per tautomer for a target pH of \( 7.0 \pm 1.5 \) were generated and compound geometries were energy-minimized with the OPLS_2005 forcefield.

Pre-prepared compounds were next collected into Schrödinger Phase databases for use during docking and future studies. Coordinates were stored in the compact internal coordinate representation, and a single conformation was generated with rapid sampling from each input compound during the Phase revise step. To enable parallel processing within the CSC system wall time limits, the 20 input files were further split into a total of 781 individual Phase databases of approx. 4.8 million compounds each.

**Receptor preparation and SiteMap analysis.** Structures of SurA (PDB-ID 1m5y, chain A) and GAK (PDB-ID 4y8d, chain A) were retrieved from the RCSB Protein Data Bank. The Schrödinger Protein Preparation Wizard was used for structure preparation, hydrogen addition, and bond order assignment. Missing sidechains were added with Prime. Crystallographic agents and water molecules were deleted. State generation for the original crystallographic ligand in the GAK structure was performed with Epik for pH 7.0±2.0. Amino acid protonation states for pH 7.0 were assigned with PROPKA and the hydrogen bonding network was optimized. Receptors were then subjected to a restrained energy minimization in the OPLS_2005 forcefield until the heavy atom RMSD compared to the previous minimization step fell below 0.3 Å.

For a comparison of the pocket properties of the two chosen targets, binding pocket proper-
ties were computed with Schrödinger SiteMap.\textsuperscript{29,35,36} The prepared GAK structure was directly subjected to SiteMap analysis, using only the site defined by the crystallographic ligand. SiteMap was run with default parameters (at least 15 site points per site, a more restrictive definition of hydrophobicity, standard grid and cropping at 4 Å from the nearest site point).

The SurA apo-structure was first subjected to a 1 μs molecular dynamics simulation with Desmond and frames from the last 200 ns were analyzed with SiteMap to identify probable and druggable small molecule binding sites (calculating up to 5 top-ranked sites per run; data not shown). The selected site was located in the crevice between N- and C-terminal core and P1 domains of SurA (for further discussion of the SurA domain architecture and pockets, see Calabrese et al.\textsuperscript{37}).

**Receptor grid generation and docking.** Grid generation for SurA used the frame with the most druggable and consistently identified site as described above (kindly provided by T. Kronenberger), and the receptor grid with a size of 30 Å\textsuperscript{3} was centered on the site centroid. For GAK, the grid center was defined as the centroid of the crystallographic ligand. Both grids were prepared with the OPLS\textsubscript{2005} forcefield. For GAK, additionally, a hydrogen-bonding constraint on the hinge-region amide (Cys126 backbone amide hydrogen) was set up.

Conventional docking of the 1.56 billion Enamine REAL lead-like library was carried out with Schrödinger Glide v9.0 in High-Throughput Virtual Screening (HTVS) mode.\textsuperscript{14,29} Van-der-Waals radii of nonpolar ligand atoms were scaled to 0.8 with a charge-cutoff of 0.15 e and nonplanar amide conformations were penalized in both targets. Additionally, for GAK, the hydrogen-bonding constraint on the hinge-region amide of Cys126 was used. With HTVS mode, the OPLS\textsubscript{2005} forcefield was used, and a single pose per ligand was collected after subjecting 5 poses to post-docking minimization.

**Simulated ML-boosted docking with HASTEN.** The machine learning-boosted docking was simulated using the simu-dock mode in a local implementation of HASTEN v0.2 (optimized for CSC Puhti).\textsuperscript{9} For ML, Chemprop v1.3.1 (with Python 3.8.12) inside a singularity-container was
used.\textsuperscript{38} Briefly, simu-dock allows the use of pre-generated docking data instead of actual docking in the HASTEN procedure. Whenever a compound is selected for docking by the algorithm, the pre-generated results of the conventional docking study will be loaded. Since scores are only added to the training data when compounds were selected for docking, the system exhibits identical behavior to screening with HASTEN when run including the brute-force docking steps.

In the first iteration, training was initialized with a random selection of 0.1\% of the full library (1.56 million compounds). The selected compound subset was split randomly into training, validation, and test sets amounting to 80\%, 10\%, and 10\% of the selected compounds, respectively. When no docking score was obtained during the conventional docking run (i.e. the compound did not dock successfully or failed to satisfy the constraint in GAK), an arbitrary failed score of +5.0 or 0.0 was applied, or failed compounds were excluded entirely from the training data. We performed one round of HASTEN with each treatment of failed compounds for each target. For SurA, experiments with a failed score of +5.0, and for GAK, experiments with the exclusion of failed compounds, were repeated in triplicate.

We used default parameters for regression in Chemprop to predict the docking score, except for the batch\_size parameter, which was increased to 250 (from default: 50) to speed up training (see SI for a list of parameters). Once the training was completed, the scores for all 1.56 billion compounds in the library were predicted from their SMILES strings. Compounds were then ordered by the predicted score, and docking scores of the top-ranked 0.1\% of compounds, that were not previously selected for docking, were loaded to simulate their conventional docking. All loaded docking scores were used to train the ML model from scratch during the next iteration, and the procedure was repeated nine times, which corresponds to docking 1\% of the 1.56 billion input library by conventional means (compare also the schematic workflow in Figure 1).

Analysis of the results. Recall values of the top 100, 1000, and 10 000 compounds were computed after every iteration. We define recall herein as the fraction of the true top-scoring 100, 1000, and 10 000 compounds when ranked by brute-force docking results, that were found using the ML model and selected for docking by HASTEN up until the current iteration.
To assess the consistency of results obtained in repeated predictions, three replicates with a different random selection of compounds for the initial training set were used. The overlap of the compounds selected for docking by each model was calculated by counting which compounds among the top-scoring 100, 1000, and 10,000 compounds were selected by one, two, or three of the replicates after each iteration.

**Results and Discussion**

The chosen targets SurA and GAK have distinct binding pocket properties. One objective in selecting targets for our case study was to ensure their distinct binding pocket properties. Since docking and, in particular, scoring success is target-dependent, any ML approach trained on docking scores will inherently reflect the same target dependence. However, having two targets with distinct properties assessed on the giga-scale would indicate whether the ML-boosted HASTEN performs equally well for both targets when compared against the brute-force docking backdrop and thus highlight potential additional dependencies arising from the choice of the ML method. To assess the pocket properties, we first computed pocket descriptors with Schrödinger SiteMap, as summarized in Table 1.

Table 1: SurA and GAK binding pocket properties computed with SiteMap: DScore, druggability score (values of 1.0 or greater are generally considered druggable); hydrophobic, hydrophilic, hydrophobic, and hydrophilic character of the site, respectively (a value of 1.0 represents the average for tight-binding sites); don/acc, the ratio of hydrogen bond donors to hydrogen bond acceptors.

<table>
<thead>
<tr>
<th>Target</th>
<th>DScore</th>
<th>Volume [Å³]</th>
<th>hydrophobic</th>
<th>hydrophilic</th>
<th>don/acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SurA</td>
<td>1.15</td>
<td>363</td>
<td>1.52</td>
<td>0.80</td>
<td>1.25</td>
</tr>
<tr>
<td>GAK</td>
<td>1.03</td>
<td>507</td>
<td>0.66</td>
<td>1.02</td>
<td>0.91</td>
</tr>
</tbody>
</table>

The selected binding pocket of SurA was smaller than the GAK pocket (363 vs. 507 Å³, Table 1) and displayed a high hydrophobicity (hydrophobic 1.52, Table 1). Furthermore, the binding site features more hydrogen bond donors than acceptors (donor acceptor ratio 1.25, Table 1). The GAK target was, on the other hand, more hydrophilic (hydrophilic 1.02, Table 1) and had a higher
proportion of acceptors in the binding pocket (donor acceptor ratio 0.91, Table 1).

In conclusion, the two targets chosen for our case study display diverging binding pocket properties and can consequently be expected to favorably interact with different chemical scaffolds, thereby allowing us to investigate the method’s performance on the giga-scale in two distinct screening scenarios.

A ready-to-use Glide-compatible screening library and giga-scale brute-force docking to SurA and GAK.

The first step towards the generation of a brute-force docking baseline was the preparation of ligand 3D structures from the original 1.56 billion input SMILES. Relevant tautomers and stereoisomers increased the library size to approx. 3.8 billion structures, which we distributed in 781 ligand databases for parallel docking. Preparation of this screening library took around 30 days (457 600 CPU hours) when utilizing 640 CPUs on the CSC Mahti supercomputer.

Next, we performed the brute-force docking of the complete prepared 3.8 billion structures to the SurA and GAK targets using Schrödinger Glide in HTVS mode to generate the docking baseline for later comparison. With an approximate processing capacity of 40 compounds per minute per CPU core, Glide HTVS proved to be the fastest available method to generate a giga-scale docking dataset. Using 640 CPUs of the CSC Mahti supercomputer, we spent 85 days (1 305 216 CPU hours) on SurA docking, and the GAK conventional screening was completed in 53 days (809 216 CPU hours).

In our setting, the library preparation step contributed approx. 30% to the total required time for the full brute-force docking study. To support future screening efforts and enable a time reduction during the ligand preparation step, we release the entire prepared and randomized Enamine REAL lead-like library (March 2021) in 781 Glide-compatible, ready-to-use Phase databases (https://doi.org/10.23729/2de314bb-59af-452a-955c-c2ff0c5ea57f). Moreover, we acknowledge that brute-force approaches to screening efforts on this scale remain elusive in most settings, even when using the fastest available docking methods. At the same time, giga-scale libraries are becoming more and more common and novel approaches to, e.g., ML-boosted docking,
should thus be evaluated on giga-scale datasets. In the hope of providing a useful benchmarking
dataset for such future applications, we further release our full giga-scale docking results for the
two targets SurA and GAK.

(https://doi.org/10.23729/2170dc9c-4905-43c3-aeee-a574d360737f)

**ML-boosted giga-scale screening of SurA and GAK.**

For accelerating the screening process with the help of ML, we used the tool HASTEN, which has
been previously validated on million-scale datasets. HASTEN aims to identify the top-scoring
compounds rather than attempting a generalized prediction of docking scores for a given target. By
iteratively selecting compounds with the best predicted scores, the training data will be progres-
sively enriched in both true positives (already ranked correctly by the model) and false positives
(ranked highly by the model, although the compounds dock poorly). This will improve HASTEN’s
capability of identifying true top-scoring compounds with every iteration. Figure 1 summarizes our
adapted procedure for the giga-scale screening.

We first started from an initial random selection of 0.1% of the 1.56 billion compounds. Previ-
ous work with HASTEN for million-scale data typically involved the selection of 1% of the library
on each iteration, but since our training data at 0.1% already exceeds one million compounds, we
decided to instead aim for a final total docking of 1% of the giga-scale library.

The brute-force docking step was simulated by utilizing the pre-generated docking data for
the complete screening library: Docking scores of compounds that were selected for docking by
the algorithm were loaded directly. Only scoring data of selected compounds was considered,
which ensures HASTEN to run as if the brute-force docking step had been performed as part of
the workflow. Docking scores and corresponding compound SMILES were used as input data for
training a ML model with Chemprop. Next, with the generated model, docking scores for the
full ERLL library were predicted, and compounds were ranked by their predicted scores. During
the next iteration, docking results for the top-ranked 0.1% of the compounds were added to the
training data. This process was repeated nine times to end with a total training dataset amounting
to 1% of the full giga-scale dataset.
Figure 1: Overview of the HASTEN workflow for the ML-boosted giga-scale screening against SurA and GAK targets. In the conventional docking step (highlighted in orange), brute-force docking was simulated by loading the docking scores of selected compounds from the pre-generated brute-force docking dataset. The ML procedure (highlighted in blue) involves a training step with all docking data of compounds that were selected for docking until the current iteration, followed by full giga-scale library prediction.

Reducing the required number of compounds to dock to 1% of the full library lowered the total time spent on ligand preparation and SurA docking to around one day and 4 hours when utilizing 640 CPUs of the CSC Mahti supercomputer, and 20 hours for the GAK target (including ligand preparation; 17628 and 12668 CPU hours for SurA and GAK, respectively).

The ML steps of the HASTEN protocol (Figure 1, blue boxes) consumed an additional time of 203-335 hours for ML model training and prediction. Training took 143 hours on a single NVidia Volta V100 GPU of CSC Puhti. Prediction steps were parallelized to utilize 10 GPUs, resulting in a total prediction time of 52-184 hours (depending on whether multiple predictions were run in parallel on one GPU or each GPU ran only a single instance of Chemprop).

Removing the necessity of docking 99% of the entire giga-scale library allows the HASTEN procedure to complete the screening in around 10-14 days for each of the two targets. Additionally,
this approach enables a user to shift the computational load associated with a giga-scale screening project from CPUs (most docking tools) to GPUs.

**Adjusting the failed score or excluding failed compounds from the training data can improve the recall.**

When docking was unable to produce a docking score, e.g. because all sampled poses were energetically unfavorable, the original HASTEN protocol associated an arbitrary positive docking score with the affected SMILES strings to mimic a positive and therefore unfavorable energy. We started out with such a ‘failed score’ of +5.0 for the SurA target and observed excellent recall values.

Herein, we define recall as the number of true 100, 1000, and 10 000 top-scoring compounds according to the brute-force docking approach, that had also been selected for docking by HASTEN. With SurA and a failed score of +5.0, we were able to recall around 95%, 90%, and 85% of the top-scoring 100, 1000, and 10 000 compounds, respectively (see Figure 2 for top 1000 and Table S1). Using the same approach with the GAK target, on the other hand, resulted in recalls of only 70%, 67%, and 59% of the top-scoring 100, 1000, and 10 000 compounds, respectively (Figure 2 and Table S2).

One major difference between the datasets is the number of failed compounds: While for SurA less than 3% of all compounds in the ERLL library fail to dock (total failed: 42 149 150 compounds), for GAK, 45% are not docked successfully (704 564 272 compounds, compare also the overall distribution of docking scores in Figure S1). Consequently, a total of 46 846 failed compounds were selected into the HASTEN training data for SurA (final training data size >15 million compounds, of which 0.3% had the failed score). In contrast, for GAK, 1 471 958 failed compounds were part of the training data, i.e. around 9% of the total training data had the failed score of +5.0.

A reason for the high number of failing compounds may lie in the treatment of hydrogen-bonding constraints in Glide: With the active hinge region amide constraint, any compound with no hydrogen bond accepting group will be directly excluded from evaluation and receive the failed
Figure 2: Recall curves of the 1000 top-scoring virtual hits in the HASTEN approach for SurA (left) and GAK (right) with different treatment of failed compounds: The curves represent the resulting recall per iteration for runs using a failed score of +5.0 (orange, diamonds), 0.0 (blue, squares), and excluding failed compounds from the training data (yellow, circles). The data is also summarized in Tables S1 and S2 in the Supporting Information.

score. For any other compound, the initial placement will depend on its hydrogen bond acceptors: With constraint fulfillment being the first objective, rather than optimizing compound orientations for enclosure in the pocket, partially exposed compound poses can occur, since they fulfill the constraint, albeit being energetically unfavorable and thus, likewise, receiving the failed score.

We hypothesize that the large resulting fraction of failed compounds in the training data drives the learning process in the ML step towards primarily identifying failed compounds to minimize the chosen metric (RMSE) rather than picking up smaller differences between the successfully docked compounds.

To improve the recall for the GAK dataset, we next attempted adjusting the failed score parameter: Motivated by our hypothesis, we set the failed score to 0.0 and thus closer to the dataset mean, which improved the recall of the top-scoring 100, 1000, and 10 000 compounds by 9-13% when compared to the failed score of +5.0 (compare Figure 2 and Table S2). For SurA, we tested the same approach, which, however, did not consistently improve the recall (Table S1).

Finally, we assessed the complete removal of all failed compounds from the training data. Notably, the exclusion of all failed compounds resulted in a recall of 94%, 90%, and 84% of the top 100, 1000, and 10 000 true virtual hits for GAK, which is similar to the results achieved initially.
with SurA (compare Figure 2 and Tables S1 and S2). Moreover, using the same approach for SurA also improved the recall, albeit only slightly by about 1-2% (Table S1).

Comparing the number of failed compounds selected for docking during each iteration between the initial GAK screen with a failed score of +5.0 and the screen where failed compounds were excluded supports our hypothesis that the model learns to recognize compounds that will fail: The initial random selection of both runs includes around 708 000 failed compounds. In later iterations, only about 10% of this initial number get selected when failed compounds were considered with a failed score of +5.0 (see Figure S2). On the other hand, when excluding failed compounds completely and thus not providing the model with any training reference to recognize features of failed compounds, around 30-40% of the initial number of failed compounds get added during each iteration (see Figure S2).

When lowering the failed score or excluding failed compounds, the improvement of recall suggests that the learning process is instead driven by the identification of good-scoring compounds to minimize the RMSE metric. Furthermore, the validation and test set RMSE values per iteration suggest that a failed score closer to the mean or dropping of the failed compounds in both cases overall improves the model (see Figure S5).

In summary, we showed that in certain docking scenarios, the complete removal of failed compounds from the training data appears to improve the model quality. Depending on the case, it can also greatly enhance the recall (GAK) or show only a minor impact on the recalled compounds (SurA). In particular, dropping failed compounds is likely beneficial when a large proportion of evaluated compounds fail to dock successfully. This protocol modification can also add to the speed-up of model training, as there is less data to process. Our case study identified the treatment of failed compounds or their assigned score as factors that can improve recall in ML-boosted screening campaigns with HASTEN/Chemprop.

**Prospects of further speed-up by hyperparameter choice and protocol modification.**

A key objective of our case study was to investigate the trade-off between maximizing speed and recall in a screening campaign with HASTEN. It is of note that the results presented herein were
achieved with Chemprop default hyperparameters, and no hyperparameter optimization was performed. Given the size of our training data, any optimization would have to be done with similarly sized data and thus take a significant amount of additional time for every alternative explored.

While it is reasonable to assume that even better models could be achieved by careful optimization of the hyperparameters, the current work confirms that model quality and recall are good when utilizing the default Chemprop settings. Thus, default hyperparameters represent a suitable choice to harness the time-saving potential of ML-boosted docking with HASTEN.

In our test cases, we achieved highly similar RMSE values for validation and test sets (see Figure S5 and Tables S5 and S6), suggesting good generalizability. It can further be seen that depending on the protocol and target, RMSE values of individual ML iterations converge well before the final iteration 10. Similarly, recall curves (see Figure 2) also often show an earlier convergence.

Thus, if speed is of the essence, our results underline that HASTEN could, for example, be stopped already at iteration 5, which doubles the speed-up achieved by ML-boosted docking while often sacrificing less than 5% in recall of the top-scoring compounds. The speed gain could likely be increased even more by utilizing smaller training dataset sizes. This is further supported by a recent screening campaign performed by Orion Pharma, combining HASTEN with Glide SP (Standard Precision) to screen a 4.1 billion compound version of Enamine REAL against an oncology target. In this case, only two 0.1% iterations were performed and yielded approximately 100,000 high-scoring virtual hits (estimated recall based on the initial random sampling around 0.5, data not shown).

We conclude that good predictive models can be obtained with Chemprop default hyperparameters and major time investment in parameter optimization appears not strictly necessary. Taken together, our data support early stopping as a viable option to achieve even larger screening speed-up. Depending on the emphasis of the screening campaign and the desired outcome, HASTEN runs can thus be tweaked to either maximize the speed or instead focus on maximally improving
HASTEN robustly identifies the same compounds irrespective of the initial random selection. As a final step in our giga-scale assessment, we aimed to verify that HASTEN robustly recalls the majority of the top-scoring compounds irrespective of the initial random compound selection. To study the overlap of recalled compounds, we performed our ML-boosted screening experiment in triplicate, with each run starting from a different random set of compounds.

For SurA, we repeated the run with the original protocol and a failed score of +5.0, and for GAK, the run with failed compounds excluded from the training data. As can be seen in Figure 3, some variation occurs during the first iterations, especially for the top 100 and top 1000 virtual hits. However, recalls converged in later iterations. The final recall of the top scoring 100, 1000, and 10 000 SurA virtual hits was on average 94%, 90%, and 85%, respectively. For GAK, average recalls were 94%, 90% and 84% (see Tables S3 and S4). Thus, in conclusion, all three runs for both targets had a highly similar recall.

Figure 3: Recall curves for HASTEN runs performed in triplicate for the targets SurA (left, failed score +5.0) and GAK (right, failed compounds dropped). The recall is defined as the percentage of 100, 1000, and 10 000 top-scoring compounds according to the conventional docking, that were also selected for docking by the HASTEN approach. The percentage of top 100 virtual hits is shown in yellow (circles), top 1000 in blue (squares), and top 10 000 in orange (diamonds). The curves represent the average recall with error bars indicating the standard deviation. The individual results are listed in Tables S3 and S4 in the Supporting Information.

We further investigated the overlap in compound selection: As shown in Figure 4 for the top
1000 virtual hits, the different HASTEN runs initially have no overlap, and from iteration 2 rapidly converge into largely the same final selection of compounds. Of the 90% recalled top 1000 virtual hits, around 30% were selected by all three replicates already on iteration 2 and around 87-88% were recalled consistently during the HASTEN runs by all three replicates (see Figure 4, black bars and Venn diagrams for iterations 2-10 in Figures S3 and S4 in the Supporting Information). Thus, our case study indicates that a single run of HASTEN is sufficient and no major recall benefit could be gained from repeating experiments with a different initial random selection, at least in a setting with training data on the million-scale. Additionally, the swift convergence into the same selection of compounds indicates that the robustness of the HASTEN approach can still be assumed when stopping on an earlier iteration.

Figure 4: Overlap of top-scoring 1000 compounds recalled in three replicates of HASTEN runs for SurA (left) and GAK (right). Stacked bars indicate the percentage of recalled compounds per iteration, with black segments: compounds recalled in all three replicates; orange segments: compounds that were recalled by any two replicates; gray segments: compounds that were only identified in a single run. The data is also visualized in Venn diagrams in Figures S3 and S4 in the Supporting Information for SurA and GAK, respectively.

**Conclusion**

Our case study of the two targets SurA and GAK confirmed the applicability of the ML-boosted docking tool HASTEN on the so far unprecedented giga-scale. Our comparison with the corre-
sponding brute-force docking results demonstrated comparable recall for the two distinct targets, identifying, for example, 90% of the true top-scoring 1000 virtual hits, although a modified protocol was necessary in the case of GAK.

Herein, our primary objective was to investigate whether HASTEN can be applied to speed up screening efforts on the giga-scale. Our ‘ground truth’ and reference in this work were brute-force docking results generated with the particularly fast Glide HTVS protocol. Increased docking speed is typically achieved by limiting conformational sampling and is thus associated with a scoring accuracy trade-off. The more robust a scoring protocol, the easier it should be for the ML model to associate a previously unseen compound structure with an appropriate score. We thus hypothesize that equally powerful models to the ones achieved in this work could be generated with smaller training data stemming from more robust docking approaches, as also underlined by previous work with HASTEN. Importantly, this allows the approach to be tailored to the available resources: A user can either generate more training data with a less robust, but faster, docking methodology or rely on fewer docking results for training when utilizing more robust and computationally expensive methods.

One should keep in mind that ligand preparation and explicit brute-force docking steps remain the most time-consuming part of the procedure and should be kept to the required minimum. To support and potentially speed up future screening efforts involving brute-force docking campaigns, we release the entire prepared ERLL library as Glide-compatible, randomized ready-to-use screening databases.

Our results indicated additional time-saving potential in strategies such as early stopping, which we found often associated with only a minor drop in the recall of virtual hits. Depending on the available resources, as well as the desired outcome of the screening campaign, we herein outlined possible modifications to the HASTEN protocol for maximum speed-up or maximum recall of top-scoring hits.

It is important to note that we herein compare against brute-force docking as our baseline. Docking scores have known limitations in their ability to rank compounds correctly and identify
true actives. As such, the HASTEN protocol is, by definition, most useful when the chosen docking approach successfully enriches in true actives.

In our case study, an excellent recall was achieved when using the default parameters of the ML engine Chemprop. It is however possible that both speed and recall could be further improved by performing hyperparameter optimization. While Chemprop provided excellent results in our case study, we acknowledge that it may not always be the ML engine of choice for every screening scenario.

With the growth trend of chemical libraries in mind, novel approaches to facilitate giga-scale screening will continue to be developed. Benchmarking such approaches on ultra-large datasets can be challenging since the generation of such reference data is prohibited by time and resource consumption in many settings. We thus also release the full giga-scale docking results to provide potential benchmarking datasets for future (ML) approaches, that seek to predict docking scores from SMILES.

In conclusion, HASTEN represents a robust approach to identify the bulk of the top-scoring virtual hits of a brute-force giga-scale docking campaign by reducing the required docking calculations by 99% (or more). Using HASTEN for ML-boosted docking is thus one viable strategy in a growing arsenal of methods designed to tackle the challenges associated with screening giga-scale libraries in everyday drug discovery.
Acknowledgement

The authors would like to thank Schrödinger LLC for their generous supply of academic licenses. We acknowledge CSC for generous computational resources and thank Atte Sillanpää of CSC for continued support and many helpful discussions. Thales Kronenberger is acknowledged for supplying input data for the SurA target and helpful feedback along the way. T.S., A.P. and I.P. gratefully acknowledge the support of the Jane and Aatos Erkko foundation. L.Y. acknowledges the Academy of Finland for funding. This work was supported by the Academy of Finland grants #336473 and #333191 to A.P.

Associated supplementary material

The following Supporting Information is available:

The full prepared ready-to-use Enamine REAL lead-like screening library (March 2021) in Schrödinger Phase database format is made available free of charge at https://doi.org/10.23729/2de314bb-59af-452a-955c-c2ff0c5ea57f.

The final docking results for the two targets SurA and GAK are made available as giga-scale benchmarking datasets free of charge at https://doi.org/10.23729/2170dc9c-4905-43c3-aece-a574d360737f.

Author contributions

References


(5) Enamine REAL lead-like compounds. [https://enamine.net/compound-collections/real-compounds/real-compound-libraries](https://enamine.net/compound-collections/real-compounds/real-compound-libraries), [Online; accessed 2023-01-05].


(9) Kalliokoski, T. Machine Learning Boosted Docking (HASTEN): An Open-source Tool To


(26) CSC – IT Center for Science Ltd. https://www.csc.fi, [Online; accessed 2023-02-06].

24
(27) ChemAxon Extended SMILES and SMARTS - CXSMILES and CXSMARTS. ChemAxonExtendedSMILESandSMARTS-Documentation, [Online; accessed 2021-03-30].


Enamine REAL lead-like 1.56 billion compounds

Brute-force docking with Glide

ML-boosted docking with HASTEN

100% >90%
identified virtual top hits

For TOC only.