Deep contrastive feature compression with classical machine learning enables ligand discovery through efficient triage of large chemical libraries

Achal Ajeet Rayakar[1], Chaitanya Kumar Jaladanki[1], Qunxiang Ong[2], Bin Tan[3], Junjun Chen[4], Alicia Qian Ler Tan[2], Ler Ting Rachel Lim[2], Navita Kohaal[5], Yu Chen[5], Jun Wang[3], Weiping Han[2], Jiancheng Hu[4,6], Hwee Kuan Lee[1,7,8,9,10], Hao Fan[1,11,12,13*]

[1]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis St, Matrix, 138671, Singapore
[2]Laboratory of Metabolic Medicine, Institute of Molecular and Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), 11 Biopolis Way, Helios, 138667, Singapore
[3]Department of Medicinal Chemistry, Ernest Mario School of Pharmacy, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, United States
[4]National Cancer Centre Singapore (NCCS), 30 Hospital Blvd, 168583, Singapore
[5]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, 33612, United States
[6]The Cancer and Stem Cell Program, Duke-NUS medical school, 8 College Rd, 169857, Singapore
[7]School of Computing, National University of Singapore, 13 Computing Dr, 117417, Singapore
[8]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Dr, 639798, Singapore
[9]International Research Laboratory on Artificial Intelligence, 15 Computing Drive, 117418, Singapore
[10]Centre for Frontier AI Research (CFAR), 1 Fusionopolis Way, Connexis, 138632, Singapore
[11]Duke-NUS Medical School, 8 College Rd, 169857, Singapore
[12]Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, 117597, Singapore
[13]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, 117596, Singapore
[*]Corresponding author (hfan2006@gmail.com)

Abstract

Improving *in silico* compound-protein interaction (CPI) predictability is critical for productive drug discovery. Current deep learning approaches largely rely on end-to-end models trained on limited labeled CPI data, overlooking preexisting, specialized compound-protein input representations. We present Ligand Extra trees-Accelerated Docking (LEAD), a virtual screening framework that accelerates docking by integrating rapid first-pass CPI prediction via ET-Screen. Unlike end-to-end models, ET-Screen uses seven distinct representations, including embeddings from large-scale protein language models and molecular transformers, as well as an original CPI potential fingerprint. ET-Screen's contrastive compression neural network maps the consolidated 2,457-dimensional compound-protein representation into a compact, discriminative form optimized for CPI classification by an ensemble of decision trees. ET-Screen outperforms state-of-the-art end-to-end approaches by up to 23.39% on diverse retrospective benchmarks while being ~100× faster than standard-precision docking. Its speed enables triaging ~10.8 million prospective drug candidates across five targets, reducing them to 10,000 for docking within the LEAD framework and ultimately yielding novel, experimentally-validated hits.

Introduction

Compounds that bind to proteins can modulate their activity, either enhancing or inhibiting their function. Identifying such compounds is a critical first step in drug discovery, after which they are assessed for efficacy and safety in humans. The multi-stage preclinical-to-market discovery process is marked by a success rate of ~1 in 20,000-30,000[1] and can span over a decade with a cost upwards of USD 1 billion[2–4]. Enhancing the productivity of this process is imperative, especially as multi-billion-compound chemical libraries are now readily accessible for testing against each protein target of interest[5]. Compounding the challenge, the druggable

human genome is estimated to contain approximately 4,479 protein-coding genes[6], each representing a potential therapeutic target. A key avenue for improving this productivity lies in the ability to identify more ligands at a faster pace.

Although routinely used for ligand discovery, high-throughput screening methods that rely on biochemical assays are time- and cost-inefficient for directly screening large libraries. These methods are limited to testing only thousands of compounds per week[7] while consuming substantial amounts of reagents. To overcome this bottleneck, force-field-based molecular docking has conventionally been a key *in silico* approach for prioritizing promising compounds from chemical libraries for experimental validation. However, even advanced docking techniques are too slow for exhaustive screening of modern ultra-large-scale chemical libraries; conducting such screenings within practical timeframes requires extensive parallel computing resources, which can cost hundreds of thousands of USD[8].

Recent advances in machine learning (ML), particularly within deep learning have significantly enhanced virtual screening[9]. For instance, graph convolutional networks (GCNs) excel in capturing complex molecular topologies and electronic interactions without requiring detailed force-field calculations by processing molecules as graphs, where atoms are nodes and bonds are edges. This allows GCNs to efficiently learn feature-rich representations of compounds, enabling rapid screening while bypassing the computational cost of traditional docking approaches[10]. These advancements deliver both high performance and speed, making virtual screening practical for ultra-large chemical libraries.

In parallel with the development of deep learning methodologies for compound-protein interaction (CPI) prediction, sophisticated representations of both protein targets and small molecules have emerged that leverage large-scale datasets and innovative model architectures. Among these, transformer-based models have risen to prominence. By employing self-attention mechanisms, transformers excel at learning complex contextual relationships and dependencies within input data, enabling a deeper understanding of sequential information[11]. For protein sequences, transformers excel at capturing long-range dependencies between amino acids, enabling the extraction of highly informative embeddings that accurately reflect structural, functional, and evolutionary properties[12]. For small molecules, transformers trained on SMILES strings can encode chemical structures into highly organized latent spaces that cluster similar molecules together, capturing subtle relationships between atomic arrangements and chemical properties[13].

Despite these advances, prevailing state-of-the-art CPI prediction models completely or in large part employ end-to-end deep learning pipelines[14–19]. These approaches simultaneously derive compound-protein representations from low-level inputs (e.g., protein sequences and SMILES) and use these representations to predict interactions. While performant, end-to-end models rely on CPI datasets that are orders of magnitude smaller than the vast, unlabeled molecular and structural datasets available to specialized representation-generating models. These models leverage large, task-specific datasets to produce holistic and highly specialized biochemical representations, integrating nuanced context across diverse data. In contrast, end-to-end methods are constrained by the scale and scope of their labeled data, limiting their ability to generalize – particularly for poorly characterized targets with sparse or no experimental interaction data. By failing to incorporate the breadth of contextual information embedded in specialized representations, end-to-end approaches risk generating representations that are less precise, reducing their adaptability and effectiveness in diverse virtual screening tasks.

Here, we present Ligand Extra trees-Accelerated Docking (LEAD), a virtual screening framework that expedites docking with the rapid, accurate triage of large chemical libraries using our ML CPI classifier, ET-Screen. ET-Screen leverages several informative feature sets created by specialized methods, including IP-FP, a protein structure- and SMILES-based interaction potential fingerprint that serves as a context-aware advancement over existing efforts. In doing so, ET-Screen harnesses the diverse expanse of unlabeled biochemical and structural information employed in the creation of these methods. Correspondingly, ET-Screen

also circumvents the significant pitfalls associated with modelling using high-dimensional and heterogeneous feature sets through its use of a feature compression neural network. This network employs a modified contrastive loss function to learn to consolidate its input feature sets into a unified, parsimonious, and discriminative compound-protein representation optimized for use with ET-Screen's classification head, an ensemble of randomized decision trees. With its two-module compression and classification architecture, ET-Screen outperforms end-to-end alternatives by up to 23.39% on two retrospective benchmarking datasets constructed for this work. These datasets, designed to assess models as universal CPI classifiers, feature a broad range of protein families and diverse compounds. Moreover, despite 98% of their compounds being non-interacting, ET-Screen consistently uncovers over five interacting compounds among its top 10 predictions – while being ~100× faster than standard-precision molecular docking. ET-Screen is also applied prospectively to triage a ~10.8 million-compound library across five targets, including two absent in its training data. The 10,000 highest-scoring compounds per target were docked, followed by manual inspection and experimental validation of the most promising candidates. This final, prospective test showed LEAD to be capable of identifying one novel scaffold for each target, demonstrating the framework's ability to accelerate drug discovery.

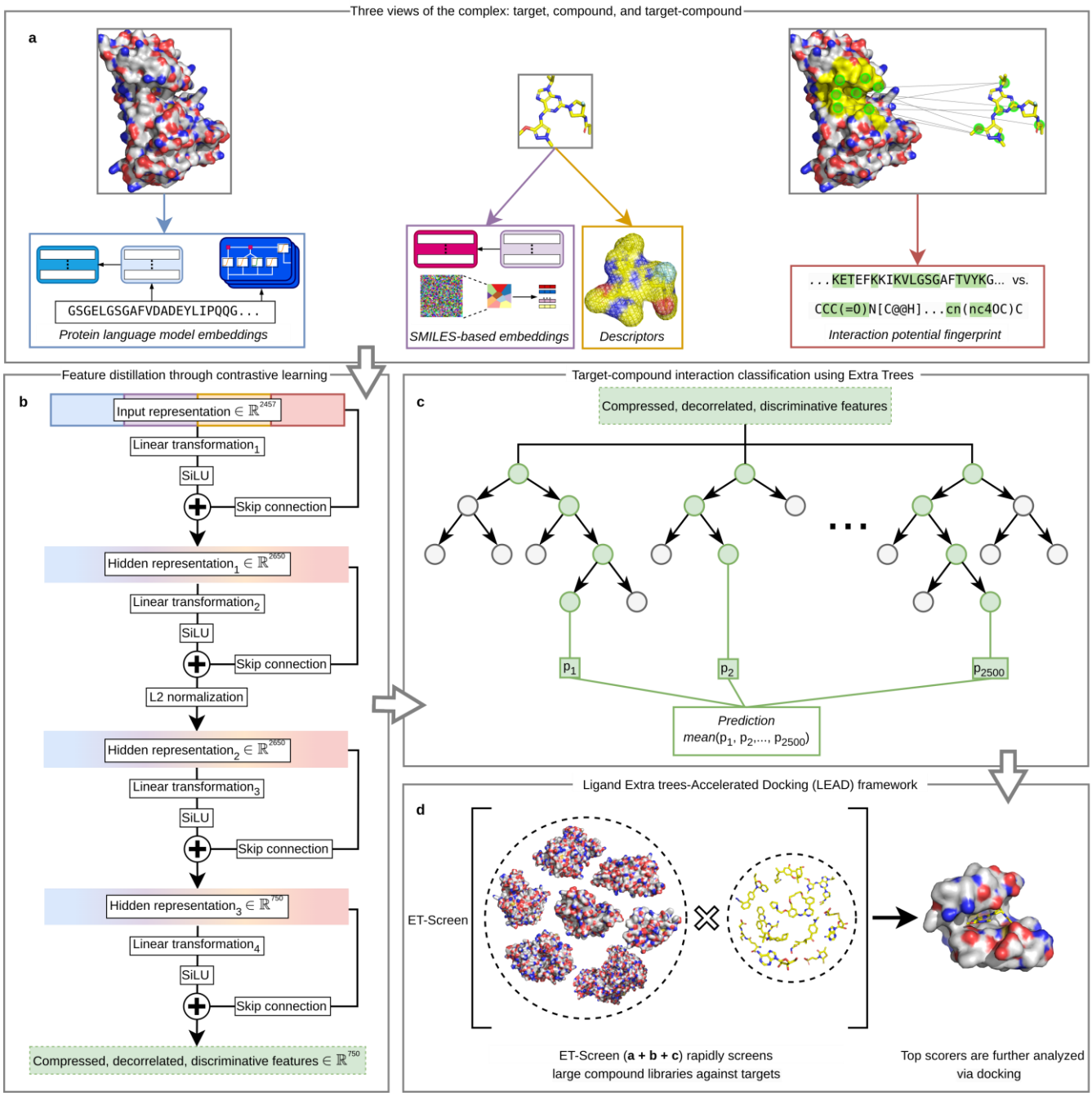

Figure 1: The LEAD framework: ET-Screen followed by docking. a, ET-Screen numerically encodes input compound-protein pairs in three ways: embeddings based on the protein's sequence; the compound's embeddings and descriptors; and an interaction potential fingerprint based on the Merck Molecular Force Field atom types found within the target pocket and compound. **b,** the resulting 2,457-dimensional input representation is compressed into 750 features using supervised contrastive learning via a residual neural network. This neural network combines and compresses input vectors to distinguish those associated with ligands and those with decoys, while decorrelating the output features. **c**, this compressed representation is then passed into an Extra Trees ensemble classifier which produces a pseudoprobability of the input compound binding to its target. **d**, ET-Screen (**a**-**c**) is used to rapidly score large compound libraries to identify the most promising compounds for further investigation via more computationally-intensive molecular docking.

Results

The LEAD framework

LEAD is a four-stage virtual screening framework designed to rapidly predict CPIs across large databases while minimizing the need for molecular docking. The first three stages, collectively referred to as ET-Screen, integrate deep neural network-based feature compression with decision tree-based CPI classification. To predict CPIs, ET-Screen processes 2,457 features that numerically encode the compound-protein pair of interest. These high-dimensional inputs are drawn from seven diverse sets of descriptive feature groups, providing a comprehensive characterization of the protein target and compound (Fig. 1a).

Two of these feature sets are created by protein language models (pLMs) which encode protein sequences into per-amino acid distributed vector representations[20]. Specifically, ET-Screen uses outputs from DLM-LSTM and ESM-2[21]. Both models employ a self-supervised masked language modeling approach[22] to produce representations that reflect proteins' structural and functional characteristics and capture both local and global sequence dependencies. Trained on large-scale unlabeled data, such as the 48 million protein sequences used for ESM-2, these models encode a broad contextual foundation within their representations that may improve the generalization of ET-Screen's downstream classification head. Prior to use by ET-Screen, however, each model's per-amino acid representations are aggregated into a single, fixed-dimensional vector via a weighted mean, with greater weight assigned to residues within the target pocket according to a capped inverse power function.

Four other feature sets encode compounds based on their SMILES strings[23]: those from MolFormer-XL[24], Mordred[25], mol2vec[26], and MACAW[27]. MolFormer-XL is a large-scale language model that uses masked language modeling on hundreds of millions of SMILES strings to learn to create molecular embeddings that encode both local substructure details and global molecular context. Mordred descriptors are traditional numerical encodings of molecular properties, capturing features such as atom counts, bond connectivity, adjacency matrices, aromaticity, and topological indices. mol2vec improves on traditional molecular fingerprints through aggregating compounds' continuous substructural embeddings, where these embeddings are learned by the method to encode substructures' presence, type, relative importance, co-occurrence patterns, and chemical similarity. Finally, MACAW converts compounds into embeddings within a low-dimensional, continuous space, arranged by the method using reference compounds so as to locate physicochemically- and structurally-similar compounds closer together.

The last feature set, IP-FP, is introduced in this work as a fingerprint that encodes the interaction potential between the target and compound. Preeminent approaches, such as that employed by RF-Score-VS[28], fingerprint intermolecular interactions by counting atom-atom contacts in resolved or docked compound-protein structures. However, since resolved structures for compounds of interest are often unavailable and docking them can be computationally expensive, IP-FP bypasses these requirements by relying solely on the Merck Molecular Force Field (MMFF) atom types[29] of the target and compound. Unlike conventional fingerprints that rely primarily on elemental identity, MMFF types classify atoms based on their hybridization state, bonding environment, and chemical context.

IP-FPs are created in 4 steps. First, an MMFF type is assigned to each atom in the target's binding pocket and in the compound. Next, the MMFF types of the atoms covalently bonded to these atoms are identified. This step enables the creation of an MMFF type pair for each atom in the binding pocket and the compound, where the first element of the pair represents the atom's own type and the second represents the types of its neighboring atoms. In the third step, the frequency of unique combinations of target and compound MMFF type pairs is calculated. Lastly, principal component analysis is applied to consolidate the large number of pair counts into a concise feature set. By integrating both immediate and broader atomic environments, the resulting fingerprint provides a comprehensive representation of the interaction potential between the target and compound.

The seven feature sets are then concatenated into a 2,457-dimensional vector representing the compound-protein pair. While informative, this representation is significantly high-dimensional relative to the size of CPI training datasets, which typically contain only thousands to tens of thousands of pairs. A purely-classical ML approach which trains ET-Screen's Extra Trees[30] classification head directly on this representation would be at risk of overfitting due to the sparsity of training pairs within this high-dimensional feature space[31]. Additionally, the accumulation of noise across many features in high-dimensional settings can obscure discriminative data patterns, further hindering the classifier's ability to predict the interaction of compound-protein pairs not featured in the training data[32]. A contemporary approach of using one deep neural network to learn both feature compression (internally) and interaction classification would face not only these challenges, but also one of entanglement. Namely, the network could risk entangling its compression and classification objectives, making it difficult to ensure that its learned internal representations remain broadly discriminative rather than overly specialized to interaction classification on the training data.

ET-Screen is therefore configured as a two-step compression and classification architecture. This approach assigns each task to a specialized ML module, with compression not guided by classification. First, ET-Screen's residual neural network[33] maps its 2,457-dimensional input onto a compact 750-dimensional feature space (Fig. 1b). This network is trained to compress its input representations by optimizing a modification to the supervised contrastive loss function[34] introduced in this work, which incorporates a correlation penalty. By minimizing this loss over its training pairs, the network learns to cluster the representations of interacting pairs closely together while separating them from non-interacting ones within the new feature space. Skip connections are incorporated into the network to enable incremental adjustments to the learned internal representations, discouraging the network from learning to make unnecessary or overly complex transformations. Simultaneously, the correlation penalty encourages the network to produce features that are orthogonal, which is intended to enhance the performance of the downstream Extra Trees classifier by increasing the diversity of its constituent decision trees (Fig. 1c).

Extra Trees, as an ensemble of independently-trained decision trees, benefits from decorrelated input features because each tree learns to partition the feature space differently, leading to a more diverse set of decision boundaries. Decision trees classify inputs by making hierarchical, axis-aligned splits, and if many trees rely on highly-correlated features, they may construct similar splits which limits the ensemble's ability to capture distinct patterns. This redundancy can reduce the overall effectiveness of the ensemble by decreasing the model variance-reduction benefits typically gained from ensembling[35]. By promoting feature decorrelation, the compression network helps diversify tree structures, leading to more complementary decision boundaries that may improve ET-Screen's generalization. Furthermore, the use of 2,500 trees enhances classifier stability by averaging over a broad set of independent sub-models, reducing susceptibility to individual decision tree variance.

Specifically, ET-Screen's Extra Trees classifier creates an interaction classification for each input compound-protein pair based on a majority vote among its constituent decision trees. Each tree independently classifies the pair as either interacting (output: 1) or non-interacting (0). While this enables a consensus binary classification by the ensemble, the mean of these votes also provides a confidence (or pseudoprobability) score of interaction. These pseudoprobability scores are used within the LEAD framework to prioritize compounds for further evaluation, with only those having the highest scores being selected for downstream investigation and

interpretation via docking (Fig. 1d). LigPrep[36] is used to preprocess these compounds, generating all possible ionization states at pH 7.0 ± 2.0 and multiple conformations per compound. Targets' structures are prepared by assigning hydrogen atoms and partial charges, and receptor grids are generated around bound ligands using Glide's Receptor Grid Generation tool. Docking grids are defined to include entire active sites with sufficient margins to ensure the inclusion of all relevant binding interactions. Prepared compounds are then docked to their target using the Glide in its Standard Precision mode (Glide-SP)[37–39]. Following docking, top compounds (as identified by their docking scores) have their docked poses visually inspected, and the most promising candidates are experimentally validated.

Distinct, diverse retrospective testing datasets

Table 1: datasets used for model training and retrospective testing. On average, D-DUD-E, BindingDB, and ChEMBL have 201, 154, and 51 ligands per target (respectively). The datasets also feature decoys to serve as complementary examples of non-binding compounds. In testing, BindingDB and ChEMBL include 50 times the number of decoys as there are ligands to reflect ligands' practical rarity. In training, D-DUD-E and BindingDB maintain parity between ligands and decoys to aid model learning. While D-DUD-E and ChEMBL are only ever used for model training and testing respectively, BindingDB plays both roles. Its training form is derived by resampling one-fiftieth of the decoys within its testing configuration while retaining the same targets and ligands.

Dataset	Targets	Ligands	Decoys	
			In training	In testing
Deduplicated DUD-E (D-DUD-E)	102	20,456	20,456	NA; only used in training.
BindingDB subset (BindingDB)	49	7,565	7,565	378,250
ChEMBL subset (ChEMBL)	22	1,114	NA; only used in testing.	55,700

Three datasets are used to train and retrospectively test ET-Screen against other CPI classification models: a deduplicated version of the Directory of Useful Decoys, Enhanced[40] (D-DUD-E), and original subsets of the BindingDB[41] and ChEMBL[42] databases (Table 1). Testing follows a cascading approach: first, models are trained on D-DUD-E and tested on BindingDB; next, they are trained on both D-DUD-E and BindingDB and tested on ChEMBL. The BindingDB and ChEMBL subsets introduced in this work are designed to be significantly distinct from each other and from D-DUD-E, ensuring that these tests provide a dual robust assessment of the models' abilities to predict the interaction of diverse compound-protein pairs. Specifically, both datasets have been created to present two challenges: identifying whether novel compounds interact with targets similar to those encountered in model training (i.e., whether they are ligands instead of decoys), and identifying novel ligands for novel targets. The first scenario tests their ability to learn underlying interaction principles rather than overfitting to spurious correlations from target-ligand pairings from training. The second measures models' generalization ability across both protein sequence and chemical diversity when there is little precedent in the training data to rely on. Together, these challenges reward only models capable of both fine-grained interpolation and extrapolation from their training data.

Given the scarcity of experimentally-verified weakly-binding and non-binding compounds, algorithmically-selected decoys must be used as stand-ins during model training and testing. For each dataset, two types of decoys are compiled: matched decoys and randomly-assigned decoys. The matched decoy selection style, as introduced in the creation of DUD-E, selects compounds from the ZINC database[43] that are physicochemically similar but structurally distinct from their corresponding ligands, with a T_c typically between 0.20 and 0.40. This process is designed to create challenging decoys that are unlikely to be false negatives, enabling models to be trained and tested to discern the structural nuances underpinning ligand binding. However, by design, matched decoys may fail to capture the extensive diversity of screening libraries that trained models are practically intended to be applied to[44]. This work therefore introduces and also employs a second, random decoy selection method. In this method, compounds from ZINC are randomly assigned as decoys to each ligand of every target. Namely, ZINC organizes its ligands into tranches based on logP and molecular weight, and since some tranches are larger than others, simple random sampling would inadequately represent compounds from less common tranches. To address this, the random sampling procedure is stratified by tranche to promote physicochemical

diversity among decoys. To also promote structural diversity, decoys are selected to have a T_c of at most 0.30 with any other decoy for the same target. Finally, since decoys are selected independently for each target, any duplicates across targets are replaced with new, randomly chosen compounds. Both matched and random decoys offer a limited and biased approximation of the ideal weak-/non-binder compound sets associated with target, which would be large, diverse, and experimentally-verified. Nonetheless, the limitations associated with both decoy types are distinct: training and retrospectively testing with them, therefore, provides complementary perspectives of models' utility. With random decoys, models learn and are evaluated on their ability to identify ligands amidst a broad, heterogeneous chemical space. While this is a coarser-grained challenge than that posed by matched decoys, considering both sets of tests in combination can help determine which models have both a superior capacity to generalize across diverse chemical environments and to also learn the narrow distinctions inherent in the matched decoy design.

Moreover, existing CPI classification studies typically employ retrospective testing datasets with 1-10 decoys per ligand[14–19,45–47]. While the true ratio of weak- and non-binders to ligands in chemical databases is both practically difficult to ascertain and also varies by both database and target, these conventional ratios fail to capture the scarcity of ligands in real-world screening – where they often represent as little as 0.01% of the database[48]. Accordingly, following the approach used in the creation of DUD-E, a decoy-to-ligand ratio of 50 is adopted for each dataset to better approximate this scarcity, although even this ratio may be optimistic in certain cases.

Given the limitations inherent to retrospective testing, the prospective performance of the LEAD framework is also evaluated on five targets: three human kinases from the D-DUD-E dataset–the serine/threonine-protein kinase B-Raf (BRAF), epidermal growth factor receptor (EGFR), and Mitogen-activated protein kinase (MEK1)–along with the SARS-CoV-2 main protease (M^{pro}) from ChEMBL, and the branched-chain ketoacid dehydrogenase kinase (BCKDK), a target absent from the three datasets with only seven known ligands[49–52]. This is done by first using ET-Screen, trained on D-DUD-E and BindingDB using random decoys, to score (i.e., generate pseudoprobabilities for) ZINC compounds against each target. The 10,000 top-scoring compounds for each target are docked and inspected, with a manually-selected 10-20 then being experimentally validated. Random decoys are selected over matched decoys for ET-Screen's model training to enable it to better distinguish between the hitherto unknown ligands to be identified and the larger, chemically-diverse background of weak-/non-binders from the ~10.8 million-compounds being scored. This experimental setup therefore serves an evaluation of LEAD's ability to identify promising hits for both well- and under-characterized targets.

ET-Screen achieves superior results in retrospective testing

ET-Screen is benchmarked against four ML models and Glide-SP (used as a control) on BindingDB and ChEMBL, using both matched and random decoys. These ML models represent the current state-of-the-art in CPI classification and are selected to be architecturally diverse, employing a variety of convolutional neural network[53], graph neural network[54], and transformer[11] modules. Notably, therefore, none of these representative top performers employ classical machine learning. Moreover, these models learn target and compound representations in an end-to-end fashion, completely or in large part forgoing manual feature engineering.

Each model's performance is primarily compared using the logAUC metric, which is the area under the curve measuring ligand recall as a function of the logarithm of the dataset coverage[55]. Here, dataset coverage refers to the proportion of the dataset's top-ranked compound-protein pairs considered, sorted in descending order by their pseudoprobabilities of interaction as predicted by the model; this axis is transformed in calculating this metric to emphasize models' early enrichment performance. Table 2 demonstrates that ET-Screen consistently outperforms its alternatives across all testing dataset-decoy style configurations. With random decoys, ET-Screen achieves a logAUC improvement of 8.65% over the next-best model on BindingDB and 6.55% on ChEMBL. This lead widens significantly with matched decoys, increasing to 23.39% and 8.92% on BindingDB and ChEMBL respectively. Additionally, ET-Screen outperforms Glide-SP by over 100% while being approximately 100 times faster, as Glide-SP requires about ten seconds to score each compound-protein pair[56].

By narrowing docking efforts to high-confidence candidates identified by ET-Screen, the LEAD framework therefore offers a reliable, scalable, and cost-effective solution to accelerate drug discovery workflows.

Table 2: **model performance under each testing dataset and decoy style configuration (best**, second best**).** To assess variability in ML model training due to random initialization of parameters and the order of training data, testing was conducted four times with different random seeds. The mean and standard deviation of these runs are calculated and presented below. In this CPI scenario with 50 decoys per ligand, a random classifier yields a logAUC of ~14.50 (equivalent to an AUROC of 0.50 and AUPRC of ~0.02), while a perfect classifier achieves a logAUC of ~71.30 (equivalent to an AUROC and AUPRC of 1.00).

Model	Matched decoys			Random decoys		
	LogAUC	**AUROC**	**AUPRC**	**LogAUC**	**AUROC**	**AUPRC**
BindingDB testing						
ET-Screen	**49.69 ± 0.33**	**0.91 ± 0.00**	**0.38 ± 0.01**	**61.02 ± [illegible]**	**0.96 ± 0.00**	**0.65 ± 0.00**
BridgeDPI[15]	40.27 ± 1.35	0.86 ± 0.01	0.18 ± 0.01	56.16 ± [illegible]	0.93 ± 0.00	0.52 ± 0.01
DrugBAN[17]	37.76 ± 2.12	0.83 ± 0.02	0.17 ± 0.02	52.92 ± [illegible]	0.92 ± 0.01	0.44 ± 0.04
PSICHIC[19]	37.26 ± 0.85	0.83 ± 0.01	0.16 ± 0.02	54.62 ± [illegible]	0.91 ± 0.01	0.50 ± 0.02
MolTrans[14]	32.24 ± 0.83	0.78 ± 0.01	0.09 ± 0.01	44.22 ± [illegible]	0.86 ± 0.01	0.25 ± 0.03
Glide	23.50	0.61	0.11	26.47	0.64	0.12
ChEMBL testing						
ET-Screen	**55.44 ± 0.26**	**0.93 ± 0.00**	**0.52 ± 0.01**	**56.11 ± [illegible]**	**0.93 ± 0.00**	**0.54 ± 0.01**
BridgeDPI	49.51 ± 0.99	0.90 ± 0.01	0.36 ± 0.02	52.06 ± [illegible]	0.92 ± 0.01	0.42 ± 0.01
DrugBAN	45.67 ± 1.81	0.88 ± 0.02	0.28 ± 0.03	49.60 ± [illegible]	0.90 ± 0.02	0.36 ± 0.05
PSICHIC	50.90 ± 0.89	0.90 ± 0.01	0.39 ± 0.03	52.66 ± [illegible]	0.90 ± 0.01	0.45 ± 0.04
MolTrans	34.26 ± 1.39	0.79 ± 0.02	0.12 ± 0.02	41.70 ± [illegible]	0.84 ± 0.01	0.21 ± 0.01
Glide	25.76	0.66	0.10	25.72	0.64	0.10

Only ET-Screen and Glide show themselves able to perform approximately as well with matched decoys as they do with random decoys. On ChEMBL testing, ET-Screen and Glide exhibit minimal logAUC reductions of 1.21% and -0.15% (respectively) with matched as opposed to random decoys, while the four competing ML models experience larger reductions of 3.46-21.72%. On BindingDB testing, however, these models show logAUC reductions of 37.16-46.59% from random to matched decoy styles; in contrast, ET-Screen's and Glide's logAUC values decrease far less – by (only) 22.80% and 11.22% respectively. Relative to ChEMBL testing, BindingDB testing exacerbates the challenge matched decoys pose over random ones. ET-Screen's substantial lead over all other models on BindingDB testing with matched decoys is therefore a demonstration of its greater generalizability. The specific contribution of ET-Screen's distinctive two-module compression and classification architecture to the model's robustness against overfitting and lead over alternative approaches will be further considered in Table 3.

While the rank order of models is nearly identical in terms of the more commonly-used area under the recall-false positive rate curve (AUROC), the metric tends to minimize performance differences between models. On BindingDB testing, ET-Screen's AUROC improvement over the next-best model is 3.23% with random decoys and 5.81% with matched decoys. On ChEMBL testing, these figures further reduce to 1.09% with random decoys and 3.33% with matched decoys. ET-Screen's AUROC improvements are far less pronounced than those observed in terms of logAUC as a consequence of AUROC's replacement of database coverage with the false positive rate. This results in an inflated perception of every model's performance in this imbalanced CPI classification context; for example, on ChEMBL with random decoys, ET-Screen and BridgeDPI both achieve near-perfect mean AUROC scores of 0.93 and 0.92, respectively. However, their mean logAUC scores, 56.11 for ET-Screen and 52.06 for BridgeDPI (out of ~71.30), present both a more moderate and differentiated picture of their relative performance. Given non-binders significantly outnumber ligands both in retrospective testing and prospective ligand discovery, practically useful models need to achieve both high ligand recall and precision among top-scoring compound-protein pairs.

While the area under the precision-recall curve (AUPRC) metric, also tabulated, accounts for precision, it does not over-weight models' early enrichment but rather quantifies their enrichment across the entire ranked database. Consequently, the metric exaggerates performance differences between models relative to logAUC: with random decoys, ET-Screen achieves an AUPRC improvement of 25.00% over the next-best model on BindingDB and 20.00% on ChEMBL. This lead widens dramatically with matched decoys, increasing to 111.11% on BindingDB and 33.33% on ChEMBL. However, in prospective virtual screening applications, where only a small subset of top-ranked ligand candidates can practically be experimentally tested, AUPRC differences overstate differences in models' practical utility.

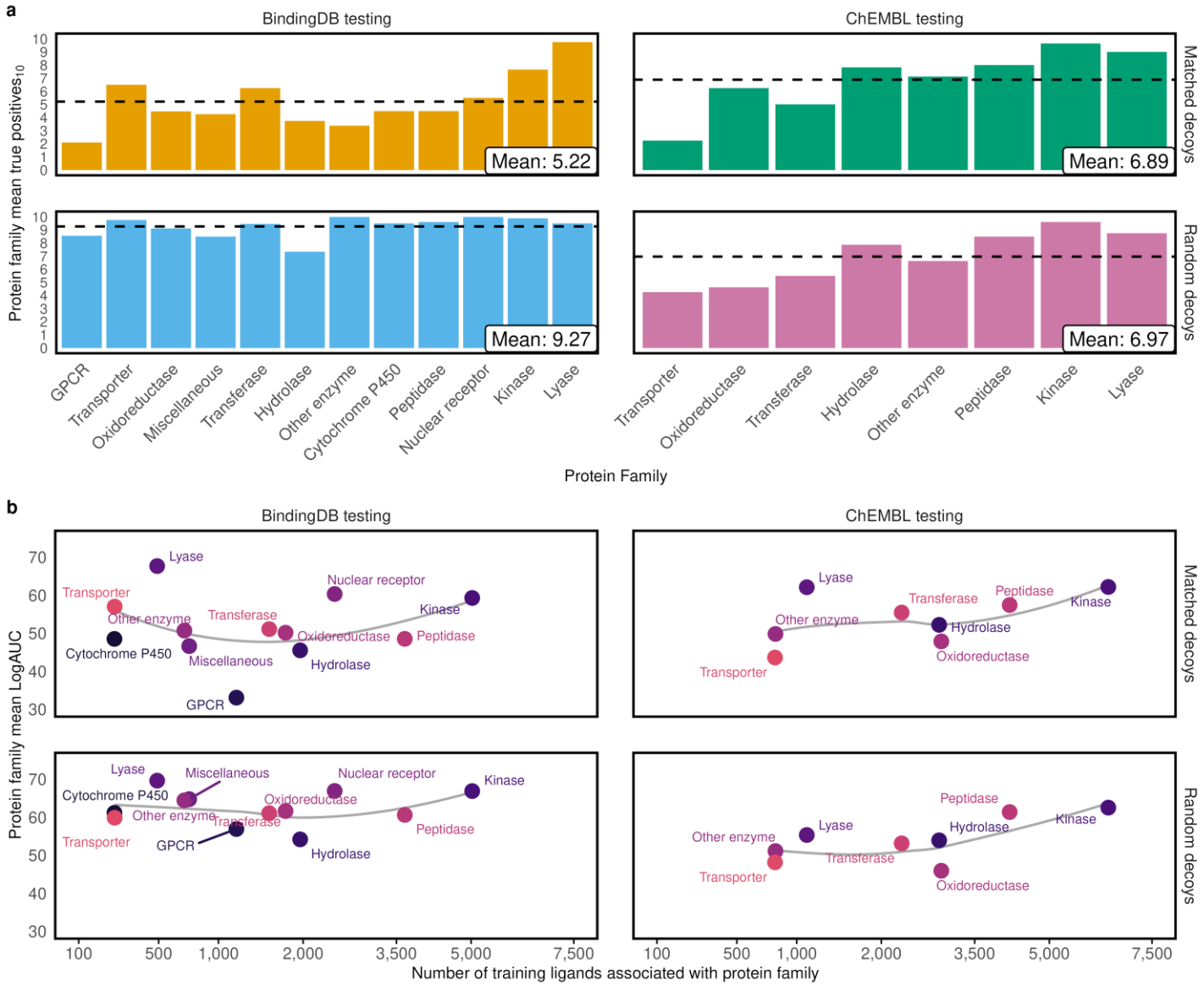

Figure 2: ET-Screen's performance on each protein family in BindingDB and ChEMBL. a, the number of ligands that ET-Screen recovers among the top ten highest-scored compounds (true positives$_{10}$) under every testing dataset-decoy style configuration. True positives$_{10}$ values are calculated on each target individually before their mean is taken and displayed by protein family. **b,** ET-Screen's mean logAUC on each protein family plotted against the family's representation in the training data for each configuration. While some of ET-Screen achieves some of its highest logAUC values on the best-represented protein families (peptidases and kinases), the relationship between performance and training representation is otherwise inconsistent.

Along with its logAUC, ET-Screen's true positives$_{10}$, which counts the number of ligands among the ten highest-scoring (highest-pseudoprobability) predictions for a given target, further evidences its high early enrichment ability: the mean true positives$_{10}$ on each protein family is usually five or greater (Fig. 2a). In

particular, ET-Screen achieves high enrichment even on protein families underrepresented in its training data; Fig. 2b shows that its logAUC scores on protein families constituting less than 2% of the training data (e.g., lyases and transporters) are comparable and sometimes even greater than those on kinases which represent about a quarter of ET-Screen's training. Moreover, variation in familywise logAUC usually occurs only within a 10-point range.

LEAD identifies new hits for five targets

Following filtering of the ~10.8 million-compound ZINC subset via ET-Screen and docking, 8-26 compounds were manually chosen for *in vitro* experimental testing against each target. This approach identified at least one previously untested ligand for each target, demonstrating inhibitory activities ranging from 4.10-124.60 μM, with a hit rate of 8-10%. *In vitro* testing of the three human kinases from D-DUD-E was carried out using immunoblotting, while M^{pro} activity was evaluated with a FRET-based enzymatic assay. For BCKDK, cell proliferation assays were performed using Hep3B wildtype and BCKDK knockdown cell lines, confirming the specificity and potency of the identified compounds. For BCKDK, an underexplored protein target with potential implications in cancer therapeutics and metabolic diseases, the identified ligand features a novel scaffold. Likewise, novel scaffolds were discovered for ligands of BRAF and M^{pro}, offering promising opportunities for further optimization and development. The identified hits for EGFR and MEK1, while novel to these targets, have been previously reported to bind other proteins, highlighting their potential for repurposing and further investigation.

Each ET-Screen module is necessary and optimal

While ET-Screen consistently outperforms its competition, Table 3 considers whether its feature compression and classifier head modules are both necessary and optimal for its performance. This is done by replacing them with other modules, both independently and in combination, and also testing configurations where feature compression is disabled (Configuration 5) or directly guided end-to-end by the classification loss (Configuration 4). Configurations employing gradient boosting[57] classification heads are also considered; gradient boosting serves as an alternate decision tree ensembling approach to Extra Trees and has recently been used with success for binding affinity prediction[58]. Also in comparison are configurations where TabNet replaces ET-Screen's Extra Trees classification head. TabNet is a neural network specifically designed to address deep learning's limitations with tabular data relative to decision tree-based methods, such as overparameterization and the absence of an inductive bias aligned with tabular decision manifolds[59].

ET-Screen achieves a 1.87-3.12% higher logAUC than the next-best alternative configuration across all scenarios except for BindingDB testing with matched decoys. In this case, it trails the top-performing configuration – Extra Trees classification without any feature compression – by 6.60%. Notably, Configurations 8 and 11, which use Extra Trees classification with PCA and feature agglomeration (FA)[60]-based feature compression, respectively, also outperform ET-Screen on this test by similar margins. This pattern suggests that ET-Screen may have slightly overfit to the D-DUD-E training data for this scenario due to the additional complexity introduced by its use of neural network-based feature compression. However, in all other scenarios, neural network-based feature compression approaches (Configurations 1-4) are approximately on par with or outperform the alternate Extra Trees configurations, with ET-Screen specifically achieving a 5.07-11.93% greater logAUC than they do. This underscores the overarching benefit of ET-Screen's two-module architecture.

Table 3: **logAUC performance of alternative ET-Screen architectures (best**, second best**).** To assess variability in ML model training due to random initialization of parameters and the order of training data, testing was conducted four times with different random seeds. The mean and standard deviation of these runs are calculated and presented below.

Configuration number	Feature compression	Classifier	BindingDB testing		ChEMBL testing	
			Matched decoys	Random decoys	Matched decoys	Random decoys
1 (ET-Screen)	Contrastive learning	Extra Trees	49.69 ± 0.33	**61.02 ± 0.14**	**55.44 ± 0.26**	**56.11 ± 0.11**
2	Contrastive learning	Gradient boosting	47.04 ± 0.15	59.09 ± 0.35	54.42 ± 0.37	54.25 ± 0.75
3	Contrastive learning	TabNet	42.63 ± 0.95	56.38 ± 0.71	51.38 ± 0.43	52.29 ± 1.08
4	End-to-end compression and		42.54 ± 0.31	58.99 ± 0.07	53.47 ± 0.25	54.41 ± 0.11

	classification neural network					
Classifier alone						
5	None	Extra Trees	**52.97 ± 0.02**	57.52 ± 0.02	49.53 ± 0.04	53.40 ± 0.03
6	None	Gradient boosting	49.86 ± 0.02	59.78 ± 0.01	51.19 ± 0.01	54.08 ± 0.03
7	None	TabNet	41.72 ± 0.83	57.94 ± 0.24	51.59 ± 0.62	52.07 ± 0.70
Alternative feature compression						
8	PCA	Extra Trees	51.99 ± 0.05	57.34 ± 0.06	48.67 ± 0.10	51.17 ± 0.08
9	PCA	Gradient boosting	50.55 ± 0.00	56.86 ± 0.00	48.85 ± 0.00	52.35 ± 0.01
10	PCA	TabNet	39.27 ± 1.08	53.33 ± 1.36	46.95 ± 1.09	49.51 ± 2.55
11	FA	Extra Trees	51.98 ± 0.06	57.80 ± 0.02	48.91 ± 0.04	52.96 ± 0.05
12	FA	Gradient boosting	48.10 ± 0.03	58.71 ± 0.01	50.32 ± 0.00	53.32 ± 0.01
13	FA	TabNet	39.80 ± 2.37	54.85 ± 0.79	48.89 ± 0.64	49.70 ± 1.52

Nonetheless, most alternative configurations still remain competitive with, and can outperform, the end-to-end alternatives BridgeDPI, DrugBAN, PSICHIC, and MolTrans. This highlights the soundness of ET-Screen's core design principle: to leverage feature sets made by specialized representation-creating methods. Augmented with deep feature compression, ET-Screen shows itself to be a strong alternative to contemporary end-to-end deep learning approaches.

Discussion

This work introduces LEAD, a virtual screening framework that employs ET-Screen to rapidly and accurately triage large chemical databases before targeted investigation with docking and experimental validation. LEAD is successfully used to identify hits for targets that are both present in and significantly different from those encountered during its training. ET-Screen employs deep contrastive neural network-based feature compression to consolidate a diverse set of input representations, including protein language model embeddings, molecular fingerprints, and an interaction potential fingerprint, into a parsimonious, discriminative form optimized for use with its decision tree-based interaction classification head. This novel two-module approach differs from the current state-of-the-art, which relies solely on deep learning and, in at least large part, forgoes manual feature engineering in favor of learning compound-protein representations for interaction classification in an end-to-end fashion.

ET-Screen shows itself to consistently be superior to these alternatives over two diverse, distinct benchmarking datasets introduced with this work. These compound-protein datasets are paired both with conventional physicochemically-matched decoys and randomly-selected decoys following a selection regime designed to address biases inherent in the former. Both this regime and the datasets serve two purposes: to be directly usable in future research, and to demonstrate how such data can be effectively constructed and compared to existing resources. Similarly, ET-Screen is not only performant as-is, but its feature compression process offers a flexible design that allows for straightforward incorporation of new representations and connection with alternative classifier heads. This adaptability positions LEAD as a forward-compatible platform that can evolve with advancements in both machine learning methodologies and bioinformatics.

References

1. Yamaguchi, S., Kaneko, M. & Narukawa, M. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clin Transl Sci* **14**, 1113 (2021).
2. Wouters, O. J., McKee, M. & Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **323**, 844 (2020).
3. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ* **47**, 20–33 (2016).
4. Hughes, J., Rees, S., Kalindjian, S. & Philpott, K. Principles of early drug discovery. *Br J Pharmacol* **162**, 1239–1249 (2011).
5. Kuan, J., Radaeva, M., Avenido, A., Cherkasov, A. & Gentile, F. Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* **13**, e1678 (2023).
6. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**, (2017).
7. Szymański, P., Markowicz, M. & Mikiciuk-Olasik, E. Adaptation of high-throughput screening in drug discovery-toxicological screening tests. *Int J Mol Sci* **13**, 427–452 (2012).

8. Tingle, B. I. & Irwin, J. J. Large-Scale Docking in the Cloud. *J Chem Inf Model* **63**, 2735–2741 (2023).
9. Oliveira, T. A. de, Silva, M. P. da, Maia, E. H. B., Silva, A. M. da & Taranto, A. G. Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. *Drugs and Drug Candidates 2023, Vol. 2, Pages 311-334* **2**, 311–334 (2023).
10. Koutroumpa, N. M., Papavasileiou, K. D., Papadiamantis, A. G., Melagraki, G. & Afantitis, A. A Systematic Review of Deep Learning Methodologies Used in the Drug Discovery Process with Emphasis on In Vivo Validation. *Int J Mol Sci* **24**, 6573 (2023).
11. Vaswani, A. *et al.* Attention Is All You Need. *Adv Neural Inf Process Syst* **2017-December**, 5999–6009 (2017).
12. Chandra, A., Tünnermann, L., Löfstedt, T. & Gratz, R. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife* **12**, (2023).
13. McGibbon, M. *et al.* From intuition to AI: evolution of small molecule representations in drug discovery. *Brief Bioinform* **25**, (2023).
14. Huang, K., Xiao, C., Glass, L. M. & Sun, J. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics* **37**, 830–836 (2021).
15. Wu, Y., Gao, M., Zeng, M., Zhang, J. & Li, M. BridgeDPI: a novel Graph Neural Network for predicting drug–protein interactions. *Bioinformatics* **38**, 2571–2578 (2022).
16. Zheng, S., Li, Y., Chen, S., Xu, J. & Yang, Y. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence 2020 2:2* **2**, 134–140 (2020).
17. Bai, P., Miljković, F., John, B. & Lu, H. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nature Machine Intelligence 2023 5:2* **5**, 126–136 (2023).
18. Yazdani-Jahromi, M. *et al.* AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Brief Bioinform* **23**, (2022).
19. Koh, H. Y., Nguyen, A. T. N., Pan, S., May, L. T. & Webb, G. I. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence 2024 6:6* **6**, 673–687 (2024).
20. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst* **12**, 654-669.e3 (2021).
21. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)* **379**, 1123–1130 (2023).
22. Sinha, K. *et al.* Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* 2888–2913 (2021) doi:10.18653/v1/2021.emnlp-main.230.
23. Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* **28**, 31–36 (1988).
24. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence 2022 4:12* **4**, 1256–1264 (2022).
25. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J Cheminform* **10**, 1–14 (2018).
26. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J Chem Inf Model* **58**, 27–35 (2018).
27. Blay, V., Radivojevic, T., Allen, J. E., Hudson, C. M. & Garcia Martin, H. MACAW: An Accessible Tool for Molecular Embedding and Inverse Molecular Design. *J Chem Inf Model* **62**, 3551–3564 (2022).
28. Wójcikowski, M., Ballester, P. J. & Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports 2017 7:1* **7**, 1–10 (2017).
29. Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J Comput Chem* **20**, 720–729 (1999).
30. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63**, 3–42 (2006).
31. Keogh, E. & Mueen, A. Curse of Dimensionality. in *Encyclopedia of Machine Learning and Data Mining* 314–315 (Springer US, Boston, MA, 2017). doi:10.1007/978-1-4899-7687-1_192.
32. Elman, M. R., Minnier, J., Chang, X. & Choi, D. Noise Accumulation in High Dimensional Classification and Total Signal Index. *Journal of Machine Learning Research* **21**, 1–23 (2020).
33. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-December**, 770–778 (2015).
34. Khosla, P. *et al.* Supervised Contrastive Learning. (2020).
35. Schapire, R. E., Freund, Y., Bartlett, P. & Lee, W. S. Boosting the margin: a new explanation for the effectiveness of voting methods. *https://doi.org/10.1214/aos/1024691352* **26**, 1651–1686 (1998).
36. LigPrep LLC New York NY 2022, S. LigPrep, Schrödinger, LLC, New York, NY 2022. (2022).
37. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J Med Chem* **47**, 1739–1749 (2004).
38. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J Med Chem* **47**, 1750–1759 (2004).
39. Glide LLC New York NY 2022, S. Glide, Schrödinger, LLC, New York, NY, 2022.

40. Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J Med Chem* **55**, 6582–6594 (2012).
41. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* **35**, D198–D201 (2007).
42. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* **40**, D1100 (2011).
43. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J Chem Inf Model* **55**, 2324–2337 (2015).
44. Li, H., Sze, K. H., Lu, G. & Ballester, P. J. Machine-learning scoring functions for structure-based virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* **11**, e1478 (2021).
45. Lee, I., Keum, J. & Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol* **15**, e1007129 (2019).
46. Chen, L. *et al.* TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **36**, 4406–4414 (2020).
47. Yazdani-Jahromi, M. *et al.* AttentionSiteDTI: an interpretable graph-based model for drug-target interaction prediction using NLP sentence-level relation classification. *Brief Bioinform* **23**, 1–14 (2022).
48. Shun, T. Y., Lazo, J. S., Sharlow, E. R. & Johnston, P. A. Identifying actives from HTS data sets: Practical approaches for the selection of an appropriate HTS data-processing method and quality control review. *J Biomol Screen* **16**, 1–14 (2011).
49. Bhattacharya, S. K. *et al.* BCKDK inhibitors. (2020).
50. Acevedo, A. *et al.* The BCKDK inhibitor BT2 is a chemical uncoupler that lowers mitochondrial ROS production and de novo lipogenesis. *Journal of Biological Chemistry* **300**, (2024).
51. Li, C., Yang, Q. & Zhang, L. Identification of putative allosteric inhibitors of BCKDK via virtual screening and biological evaluation. *J Enzyme Inhib Med Chem* **39**, (2024).
52. Roth Flach, R. J. *et al.* Small molecule branched-chain ketoacid dehydrogenase kinase (BDK) inhibitors with opposing effects on BDK protein levels. *Nature Communications 2023 14:1* **14**, 1–14 (2023).
53. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv Neural Inf Process Syst* **25**, (2012).
54. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The graph neural network model. *IEEE Trans Neural Netw* **20**, 61–80 (2009).
55. Mysinger, M. M. & Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model* **50**, 1561–1573 (2010).
56. Docking and scoring - Schrödinger. https://www.schrodinger.com/life-science/learn/white-papers/docking-and-scoring/.
57. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front Neurorobot* **7**, 63623 (2013).
58. Luttens, A. *et al.* Rapid traversal of vast chemical space using machine learning-guided docking screens. *Nature Computational Science 2025* 1–12 (2025) doi:10.1038/s43588-025-00777-x.
59. Arık, S. & Pfister, T. TabNet: Attentive Interpretable Tabular Learning. *35th AAAI Conference on Artificial Intelligence, AAAI 2021* **8A**, 6679–6687 (2019).
60. sklearn.cluster.FeatureAgglomeration — scikit-learn 1.3.2 documentation. https://scikit-learn.org/1.3/modules/generated/sklearn.cluster.FeatureAgglomeration.html.
61. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **50**, 742–754 (2010).
62. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 1–13 (2015).
63. Assay and Activity Questions | ChEMBL Interface Documentation. https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/chembl-data-questions.
64. Rice, P., Longden, L. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277 (2000).
65. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
66. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830 (2011).
67. Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations* 38–45 (2020) doi:10.18653/V1/2020.EMNLP-DEMOS.6.
68. Landrum, G. *et al.* rdkit/rdkit: 2024_03_5 (Q1 2024) Release. Preprint at https://doi.org/10.5281/ZENODO.12782092.
69. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst* **32**, (2019).
70. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). (2016).
71. Ruder, S. An overview of gradient descent optimization algorithms. (2016).
72. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2014).
73. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *International Conference on Artificial Intelligence and Statistics* (2010).
74. Protein Preparation Wizard LLC NewYork NY 2022, S. Protein Preparation Wizard, Schrödinger, LLC, New York, NY, 2022.

75. Sacco, M. D. *et al.* Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against Mpro and cathepsin L. *Sci Adv* **6**, (2020).
76. Ma, C. *et al.* Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Research 2020 30:8* **30**, 678–692 (2020).
77. Yap, J. *et al.* Characterize Disease-related Mutants of RAF Family Kinases by Using a Set of Practical and Feasible Methods. *J Vis Exp* **2019**, (2019).
78. Yap, J. *et al.* The stability of R-spine defines RAF inhibitor resistance: A comprehensive analysis of oncogenic BRAF mutants with in-frame insertion of αC-β4 loop. *Sci Adv* **7**, (2021).
79. Ericksen, R. E. *et al.* Loss of BCAA Catabolism during Carcinogenesis Enhances mTORC1 Activity and Promotes Tumor Development and Progression. *Cell Metab* **29**, 1151-1165.e6 (2019).