PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Contrastive learning in protein language space predicts interactions between drugs and protein targets

Rohit Singh[a,1], Samuel Sledzieski[a,1], Bryan Bryson[b,c], Lenore Cowen[d,2], and Bonnie Berger[a,e,2]

Edited by Barry Honig, Columbia University, New York, NY; received December 6, 2022; accepted April 10, 2023

Sequence-based prediction of drug–target interactions has the potential to accelerate drug discovery by complementing experimental screens. Such computational prediction needs to be generalizable and scalable while remaining sensitive to subtle variations in the inputs. However, current computational techniques fail to simultaneously meet these goals, often sacrificing performance of one to achieve the others. We develop a deep learning model, ConPLex, successfully leveraging the advances in pretrained protein language models ("PLex") and employing a protein-anchored contrastive coembedding ("Con") to outperform state-of-the-art approaches. ConPLex achieves high accuracy, broad adaptivity to unseen data, and specificity against decoy compounds. It makes predictions of binding based on the distance between learned representations, enabling predictions at the scale of massive compound libraries and the human proteome. Experimental testing of 19 kinase-drug interaction predictions validated 12 interactions, including four with subnanomolar affinity, plus a strongly binding EPHB1 inhibitor (K_D = 1.3 nM). Furthermore, ConPLex embeddings are interpretable, which enables us to visualize the drug–target embedding space and use embeddings to characterize the function of human cell-surface proteins. We anticipate that ConPLex will facilitate efficient drug discovery by making highly sensitive in silico drug screening feasible at the genome scale. ConPLex is available open source at ConPLex.csail.mit.edu.

drug discovery | protein language models | contrastive learning | drug–target interaction

Significance

In time and money, one of the most expensive steps of the drug discovery pipeline is the experimental screening of small molecules to determine binding to a protein target of interest. Therefore, accurate high-throughput computational prediction of drug-target interactions would unlock significant value, guiding and prioritizing promising candidates for experimental screening. We introduce ConPLex, a machine learning method for predicting drug-target binding which achieves state-of-the-art accuracy on many types of targets by using a pretrained protein language model. The approach co-locates the proteins and potential drug molecules in a shared feature space while learning to contrast true drugs from similar nonbinding "decoy" molecules. ConPLex is extremely fast, which allows it to rapidly shortlist candidates for deeper investigation.

In the drug discovery pipeline, a key rate-limiting step is the experimental screening of potential drug molecules against a protein target of interest. Thus, fast and accurate computational prediction of drug–target interactions (DTIs) could be extremely valuable, accelerating the drug discovery process. One important class of computational DTI methods, molecular docking, uses 3D structural representations of both the drug and target. While the recent availability of high-throughput accurate 3D protein structure prediction models (1–3) means that these methods can be employed starting only from a protein's amino acid sequence, the computational expense of docking (4) and other structure-based approaches [e.g., rational design (5), active site modeling (6), template modeling (7, 8)] unfortunately remains prohibitive for large-scale DTI screening. An alternative class of DTI prediction methods use 3D structure only implicitly, making rapid DTI predictions when the inputs consist only of a molecular description of the drug [such as the SMILES string (9)] and the amino acid sequence of the protein target. This class of sequence-based DTI approaches enables scalable DTI prediction, but there have been barriers to matching the levels of accuracy obtained by structure-based approaches.

In this paper, we introduce ConPLex, a rapid purely sequence-based DTI prediction method that leverages rich featurizations from pretrained protein language models (PLMs) and show that it can produce state-of-the-art performance on the DTI prediction task at scale. The advance provided by ConPLex comes from two main ideas that together overcome some of the limitations of previous approaches: informative PLM-based representations and contrastive learning. While many methods have been proposed for the sequence-based setting of the DTI problem (10) [e.g., using secure multiparty computation (11), convolutional neural networks (12), or transformers (13)], their protein and drug representations are constructed solely from DTI ground truth data. The high level of diversity among the DTI inputs, combined with the limited availability of DTI training data, limits the accuracy of these methods and their generalizability beyond their training domain. Furthermore, the methods that do generalize often do so by sacrificing fine-grained specificity, i.e., are unable to distinguish true-positive binding compounds from false positives with similar physicochemical properties ("decoys").

Author contributions: R.S., S.S., L.C., and B. Berger designed research; R.S., S.S., B. Bryson, L.C., and B. Berger performed research; R.S., S.S., B. Bryson, L.C., and B. Berger contributed new reagents/analytic tools; R.S., S.S., L.C., and B. Berger analyzed data; and R.S., S.S., B. Bryson, L.C., and B. Berger wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]R.S. and S.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: cowen@cs.tufts.edu or bab@mit.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2220778120/-/DCSupplemental.

Published June 8, 2023.

In contrast, the "PLex" (Pretrained Lexicographic) part of ConPLex helps alleviate the problem of limited DTI training data. As we showed in our preliminary work (14), one way to get around the limited size of DTI datasets that has hampered the quality of the representations learned by previous methods is to transfer learned proteins representations from pretrained PLMs to the DTI prediction task. PLMs learn the distributional characteristics of amino acid sequences over millions of proteins in an unsupervised fashion, generating sequence-based representations that encode deep structural insights. A design paradigm in machine learning is that an informative featurization of the input can enhance the power of even simple models. For DTI, where task-specific data are limited, using PLM-generated representations as the input features allows us to borrow strength from the much larger corpus of single protein sequences (14). Starting with the PLMs, our second insight directly addresses the fine-grained specificity problem in our architecture by using the "Con" (Contrastive learning) part: a protein-anchored contrastive coembedding that colocates the proteins and the drugs into a shared latent space. We show that this coembedding enforces separation between true interacting partners and decoys to achieve both broad generalization and high specificity (Fig. 2).

Putting these two ideas together gives us ConPLex, a representation learning approach that enables both broad generalization and high specificity. We show that ConPLex enables more accurate prediction of DTIs than competing methods while avoiding many of the pitfalls suffered by currently available approaches. Thus, our work constitutes a concrete demonstration of the power of a well-designed transfer learning approach that adapts foundation models for a specific task (15, 16). In particular, we found that the performance of existing sequence-based DTI prediction methods could be sensitive to variation in drug-vs-protein coverage in the dataset, whereas ConPLex performs well in multiple coverage regimes. Indeed, ConPLex performs especially well relative to other methods in the zero-shot prediction setting where no information is available about a given protein or drug at training time. Experimental validation of ConPLex yielded a 63% hit rate (12/19), including four hits with subnanomolar binding affinity, demonstrating the value of ConPLex as an accurate, highly scalable, in silico screening tool.

ConPLex can also be adapted beyond the binary case to make predictions about binding affinity. Furthermore, the shared representation also offers advantages beyond prediction accuracy. The coembedding of both proteins and drugs in the same space offers intepretability, and we show that distances in this space meaningfully reflect protein domain structure and binding function: We leverage ConPLex representations to functionally characterize cell-surface proteins from the Surfaceome database (17), a set of 2,886 proteins localized to the external plasma membrane that participate in signaling and are likely able to be easily targeted by ligands.

ConPLex is extremely fast: As a proof of concept, we make predictions for the human proteome against all drugs in ChEMBL (18) ($\approx 2 \times 10^{10}$ pairs) in just under 24 h using a single NVIDIA A100 GPU. Thus, ConPLex has the potential to be applied for tasks which would require prohibitive amounts of computation for purely structure-based approaches or less efficient sequence-based methods, such as genome-scale side-effect screens, identifying drug repurposing candidates via massive compound libraries searches or in silico deep mutational scans to predict variant effects on binding with currently approved or potential new therapeutics. We note that most DTI methods require significant computation on each drug–target pair (i.e., have quadratic time complexity). Because ConPLex predictions rely only on the distance in the shared space, predictions can be made highly efficiently once embeddings (which have linear time complexity) are computed.

Distinguishing between Low- and High-Coverage DTI Prediction. We benchmark performance of ConPLex and competing methods in two different regimes, which we term low-coverage and high-coverage DTI prediction (Fig. 1*C*). We show that ConPLex outperforms its competitors in both settings, but note that separating the two regimes helps clarify an often-seen issue in the field: methods whose performance varies substantially across different proposed DTI benchmarks. Several prior attempts have been made to standardize DTI benchmarking and develop a consistent framework for model evaluation (19, 20). However, much of this work has overlooked a key aspect of benchmarking that we find to significantly affect model performance—differing per-biomolecule data coverage. We define coverage as the average proportion of drugs or targets for which a data point exists in that dataset, whether that is a positive or negative interaction (*Methods*). Depending on the per-biomolecule data coverage of the benchmark dataset, we claim that these benchmarks are looking at very different problems. In particular, low-coverage

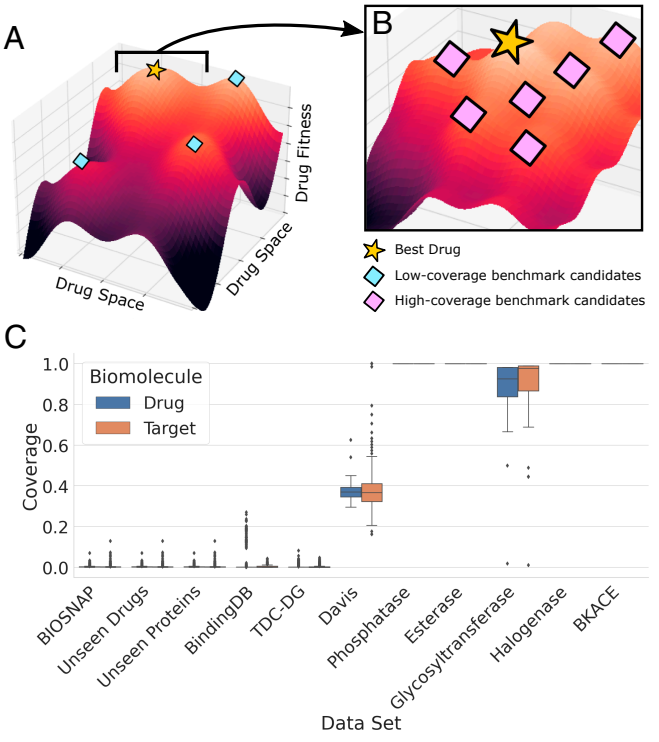

Fig. 1. Drug-target interaction benchmarks display highly variable levels of coverage. Coverage is defined as the proportion of drugs or targets for which a data point (positive or negative) exists in that dataset. High- vs. low-coverage benchmarks tend to reward different types of model performance. (*A*) In this cartoon of an example low coverage dataset, drug candidates cover the full diversity of the space, and no two drugs are highly similar. A successful model can learn a coarse estimate of the fitness landscape, but must accurately model a large part of drug space to generalize to all candidates. (*B*) For high-coverage datasets, drugs tend to be targeted to a specific protein family. Thus, a successful model does not need to generalize nearly as widely but must be able to capture more minor variations in drug fitness to achieve high specificity and differentiate between similar drugs. (*C*) In a review of existing popular DTI benchmark datasets, we find widely varying coverage, from datasets with nearly zero coverage (each drug/target is represented only a few times) to nearly full coverage (all drug-by-target pairs are known in the data).

datasets (Fig. 1*A*) tend to measure the broad strokes of the DTI landscape, containing a highly diverse set of drugs and targets. Such datasets can present a modeling challenge due to the diverse nature of targets covered but allow for a broad assessment of compatibility between classes of compounds and proteins. High-coverage datasets (Fig. 1*B*) represent the opposite trade-off: They contain limited diversity in drug or target type but report a dense set of potential pairwise interactions. Thus, they capture the fine-grained details of a specific subclass of drug–target binding and enable distinguishing between similar biomolecules in a particular context.

The two coverage regimes correspond to different usage cases. The low-coverage regime is relevant when applying DTI models for large-scale scans to predict interactions for a potential target against a large compound library [e.g., for drug repurposing as in Dönertaş et al. (21) and Morselli et al. (22)] or for scanning a candidate drug against an entire proteome to identify potential adverse and off-target effects [as in Huang et al. (23, 24)]. Data at this scale are often low coverage, with only a small number of known interactions for each unique biomolecule. Thus, it is important that DTI models used for these tasks are broadly applicable and can accurately generalize to many different families of proteins and drugs. However, this generalization often comes at the cost of specificity, resulting in models that are unable to distinguish between highly similar drugs or proteins.

The high-coverage regime is relevant when optimizing a particular interaction. Here, models can be trained to be highly specific to a protein family or class of drugs, so much so that a per-drug or per-target model is trained to capture the precise binding dynamics of that biomolecule (25). While such models can be effective for lead optimization, they require high coverage on the biomolecule of interest to make accurate predictions; this may not always be available. Additionally, such models lack the capacity to generalize beyond the training domain and thus cannot be used for genome- or drug bank-scale prediction.

The PLM approach of ConPLex enables strong performance in both regimes. In the low coverage regime, the strength is coming mostly from the "PLex" part, where it can leverage the effective generalization of language models to achieve state-of-the-art performance. On high-coverage datasets, the "Con" part also becomes important, since it becomes feasible to train drug- or target-specific models with high accuracy, and such models often outperform more generic models. We find that while single-task models do perform well given available data, ConPLex is able to achieve extremely high specificity in low-diversity, high-coverage scenarios, while remaining broadly applicable to protein targets with limited data. Thus, ConPLex is applicable for both large-scale compound or target screens and fine-grained, highly specific binding prediction. We discuss the issue of matching the right model to the problem domain with respect to coverage further in the *Discussion*.

Results

Model Overview. To achieve both generalizability and specificity, ConPLex leverages advances in both protein language modeling and metric learning. We start with pretrained representations and learn a nonlinear projection of these representations to a shared space ($\mathbb{R}^{d_h}$). We guide the learning by alternating between two objectives over multiple iterations: a coarse-grained objective of accurately classifying DTIs and a fine-grained objective of distinguishing decoys from drugs. The coarse-grained objective is evaluated over a low-coverage dataset, which trains the model to distinguish between broad classes of drug and target and makes initial predictions in the right "neighborhood" of the DTI space. The fine-grained objective is evaluated over a high-coverage dataset, which fine-tunes the model to distinguish between true and false positive interactions in the same "neighborhood" and achieve high specificity within a class.

To featurize the inputs, here, we use the Morgan fingerprint (26) for small molecules and embeddings from a pretrained ProtBert model (27) for proteins. We investigate other choices for features, including several other foundation PLMs in *SI Appendix*, S2. We note that our framework is flexible to different methods of featurization and make recommendations on the selection of informative representations in the *Discussion*.

ConPLex Achieves State-of-the-Art Performance on Low-Coverage and Zero-Shot Interactions. A key advance of ConPLex is the use of pretrained PLMs for protein representation. As foreshadowed by Scaiewicz and Levitt (28), PLMs have repeatedly been shown to encode evolutionary and structural information (29–31) and to enable broad generalization in low-coverage scenarios (32, 33). Here, we show that ConPLex achieves state-of-the-art performance on three low-coverage benchmark datasets—**BIOSNAP**, **BindingDB**, and **DAVIS**—where it is important to learn the broad strokes of the DTI landscape. In Table 1, we show the average area under the precision–recall curve (AUPR) over five random initializations of each model evaluated on a held-out test set (*Methods*). Here, we compare with several methods which use non-PLM protein features: MolTrans (13), GNN-CPI (34), and DeepConv-DTI (12). In addition, we compare to the EnzPred-CPI model from Goldman et al. (25) (developed simultaneously and independently), which uses a PLM for protein featurization but does not perform a coembedding or utilize a contrastive training step. Finally, we compare with the single-task Ridge regression model described in ref. 25, which

Table 1. ConPLex is highly accurate and generalizes broadly in low coverage settings

Dataset	ConPLex	EnzPred-CPI	MolTrans	GNN-CPI†	DeepConv-DTI†	Ridge
BIOSNAP	0.897 ± 0.001	0.866 ± 0.003	0.885 ± 0.005	0.890 ± 0.004	0.889 ± 0.005	0.641 ± 0.000
BindingDB	0.628 ± 0.012	0.602 ± 0.006	0.598 ± 0.013	0.578 ± 0.015	0.611 ± 0.015	0.516 ± 0.000
DAVIS	0.458 ± 0.016	0.277 ± 0.009	0.335 ± 0.017	0.269 ± 0.020	0.299 ± 0.039	0.320 ± 0.000
Unseen Drugs	0.874 ± 0.002	0.844 ± 0.005	0.863 ± 0.005	–	0.847 ± 0.009	N/A
Unseen Targets	0.842 ± 0.006	0.795 ± 0.004	0.668 ± 0.045	–	0.766 ± 0.022	0.617 ± 0.000

ConPLex outperforms several state-of-the-art methods, including EnzPred-CPI (25), MolTrans (13), GNN-CPI (34), and DeepConv-DTI (12), as well as a simple single-target Ridge regression model, on several low- and zero- coverage benchmark datasets. We report the average and SD of the area under the precision–recall curve (AUPR) for 5 random initializations of each model. Metrics for models with † are taken from ref. 13. Ridge regression cannot be applied for the **Unseen Drugs** dataset since a separate model is trained for each drug in the training set.

trains a different model per drug rather than a single model for the entire benchmark.

Observing the strength of ConPLex to generalize on low-coverage data, we sought to evaluate its performance on fully zero-shot prediction. **Unseen drugs** and **Unseen targets** are variants of the BIOSNAP dataset where drugs/targets in the test set do not appear in any interactions in the training set (*Methods*). Note that for the unseen drugs setting, the Ridge model cannot be applied since a different model must be trained for each drug that appears in the training set. We show that ConPLex achieves the best zero-shot prediction performance (Table 1), further demonstrating the applicability of the model to large-scale, very low-coverage prediction tasks.

Contrastive Learning Enables High-Specificity DTI Mapping. Another key advance of our method is the use of contrastive learning to fine-tune model predictions on high-coverage data to achieve high specificity. Recently, Heinzinger et al. (35) demonstrated the use of semisupervised contrastive learning for effective protein embedding-based annotation transfer. Here, we adapt contrastive learning to a fully supervised setting and demonstrate that the contrastive training is essential to achieving specificity using DTI pairs from the Database of Useful Decoys (**DUD-E**) (36). The DUD-E dataset contains 57 protein targets and drugs which are known to interact with each target. However, it also contains 50 negative "decoy" small molecules for each drug, which have similar physicochemical properties to the truly interacting small molecule but are known to not bind the target. Thus, accurate prediction on DUD-E requires a model to achieve high specificity and to accurately differentiate between highly similar compounds. Additionally, DUD-E contains four different classes of targets—G-protein-coupled receptors (GPCRs), kinases, proteases, and nucleases—so models must generalize across target classes (note that single task models do not have this generalization requirement since a different model is trained per target).

We derive evaluation sets from DUD-E by holding out 50% of proteins in each target class for testing and using the remaining targets for training (full splits are specified in *SI Appendix*, S1). Here, we evaluate a ConPLex model trained on BIOSNAP, both with and without contrastive training on DUD-E, and show that contrastive training is essential to achieving specificity on decoys.

For each target in the DUD-E test set, we use t-SNE to visualize the target alongside all drugs and decoys using embeddings learned by both versions of the model. Fig. 3 *A* and *B* shows one such example, the tyrosine kinase *VGFR2*. We also show the distribution of distances in the latent space between the target embedding and the embeddings of the drugs and decoys for each model (Fig. 3 *C* and *D*) (*P*-values from the one-sided *t* test). Without contrastive training, drugs are interspersed with decoys and are far away in space from the target, while ConPLex clusters most true drugs very close to both each other and the *VGFR2* embedding.

In Fig. 3*E*, we show a quantitative analysis of all 31 test-set targets. We compute the effect size (Cohen's *d*) of the difference between predicted drug and decoy scores. We plot these effect sizes for ConPLex trained with and without contrastive training. An increase in the effect size indicates that the coembedding distances learned by the model better represent binding specificity. The effect size increases for every target, and the median effect size between predicted true and decoy compound scores was 0.730 prior to contrastive training compared to 4.716 after. For each class of targets, we also report the median *P*-value (one-sided *t* test) between drug and decoy scores predicted by

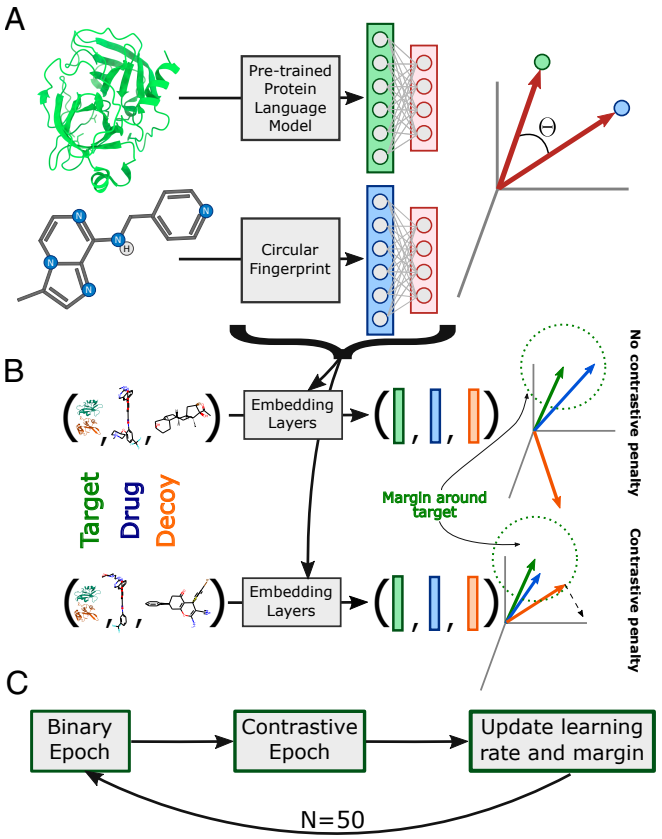

Fig. 2. Outline of the ConPLex model architecture and training framework. ConPLex is trained in two phases, to optimize both generalizability and specificity. (*A*) Protein features are generated using a pretrained PLM [here ProtBert (27)], and drug features are generated using the Morgan fingerprint (26). These features are transformed into a shared latent space by a learned nonlinear projection. The prediction of interaction is based on the cosine distance in this space, and the parameters of the transformation are updated using the binary cross-entropy on a low-coverage dataset. (*B*) In the contrastive phase, triplets of a target, drug, and decoy are transformed in the same way into the shared space. Here, the transformation is treated as a metric learning problem. Parameters are updated using the triplet distance loss on a high-coverage dataset (36) to minimize the target-drug distance while maximizing the target-decoy distance. No additional penalty is applied if the target-decoy distance is greater than the target-drug distance plus some margin. (*C*) ConPLex is trained in alternating epochs of the binary and contrastive phase to simultaneously optimize both objectives. After each round, learning rates and the contrastive margin are updated according to an annealing scheme.

ConPLex. While contrastive training has an extremely large impact on specificity in high-coverage domains, we also show that this additional training does not significantly decrease the model performance on low-coverage benchmarks via an ablation study in *SI Appendix*, S3.

In addition to evaluation on DUD-E, we also evaluate ConPLex on five benchmark datasets derived from family-specific enzyme–substrate screens (*Methods*). These datasets are extremely high coverage, generally including data points for all possible pairs of drugs and targets. We find that in this regime, ConPLex and other PLM-based models like EnzPred-CPI have strong but highly variable performance and are still generally outperformed by a Ridge regression model (*SI Appendix*, S5) as shown previously in ref. 25. However, a fine-scale single-task model is limited in its generalizability beyond the enzyme family on which it was trained (*Discussion*).

ConPLex Discovers DTIs with Subnanomolar Binding Affinity. Since ConPLex exhibited strong performance on several benchmark datasets, we next sought to experimentally validate

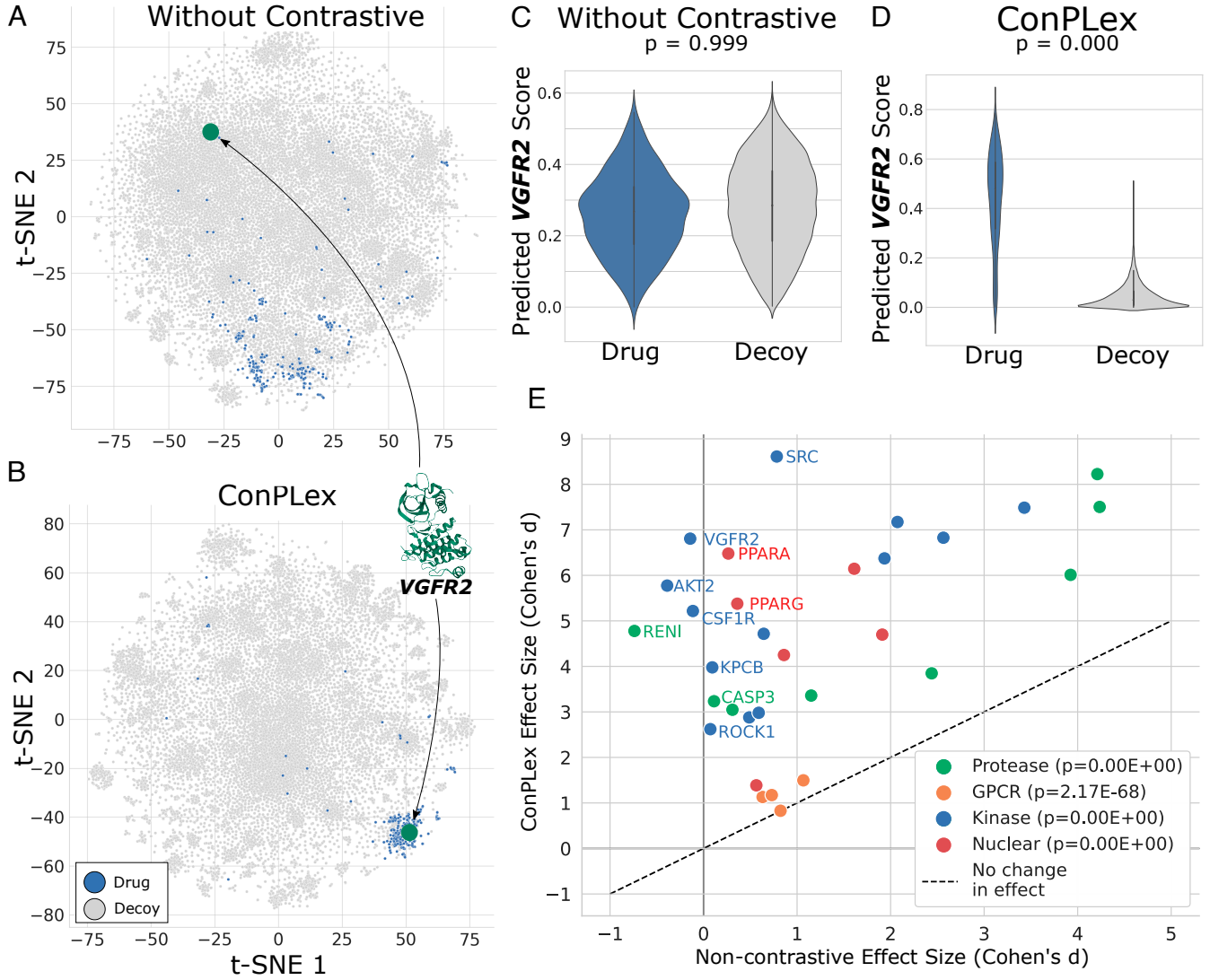

Fig. 3. Contrastive training enables high specificity in discriminating drugs from decoys. We demonstrate that contrastive learning is essential for ConPLex to achieve high specificity using the DUD-E (36) dataset of drugs and decoys (nonbinding small molecules with similar physicochemical properties to the true drugs). (*A* and *B*) Using t-SNE, we show the learned ConPLex latent space for *VGFR2* (green) and known drugs (blue) and decoys (gray). Without contrastive training, drug and decoy representations do not separate, and true drugs are far from their target. With contrastive training, VGFR2 and drugs cluster very tightly compared to decoys. (*C* and *D*) ConPLex predictions significantly differentiate between drugs and decoys after contrastive training ($P = 0.000$ paired *t* test) but do not differ at all without such training ($P = 0.999$). (*E*) We compute the effect size between drug and decoy predictions using Cohen's *d* for all 31 targets in the test set. Targets are classified as proteases (green), GPCRs (orange), kinases (blue), and nuclear proteins (red). This effect is computed for ConPLex both with and without contrastive training. Contrastive training increases the effect size for every target (median 0.730 vs. 4.716). For each class, we report the median *P*-value for ConPLex drug vs. decoy predictions. ConPLex performs particularly well for kinases and nuclear proteins and more poorly for GPCRs.

predictions using an in vitro biochemical binding assay. We selected 51 kinases from the Surfaceome database (17) with commercially available assays from the DiscoveryX company and used ConPLex to scan against a set of 4715 compounds from the ZINC database (37) purchasable from the Cayman Chemical Company (*Methods*) (38). We selected 19 interactions spanning 5 kinases and 14 compounds in an unbiased manner. (These pairs were chosen based solely on top scoring ConPLex predictions, without any use of prior knowledge from experimental results or in the literature.) We determined K_D values for each of the 19 interactions (Table 2), finding that 12/19 pairs tested had K_D values less than 100 nM. Of these, four bound with subnanomolar affinity, all of which recapitulate known interactions in the literature. Weglicki et al. identified AG-1478 as an *EGFR* inhibitor but noted that its therapeutic use may be limited due to triggering hypomagnesemia and cardiac dysfunction (39). Sordella et al. (40) described the downstream impact in lung cancer when Gefitinib inhibits *EGFR*. In a review of Nintedanib discovery, Roth et al. (41) noted it as an *FLT3* inhibitor, and Wang et al. (42) described Linifanib inhibition of *FLT3*.

We also identify an interaction between *EPHB1* and PD-166326 with nearly subnanomolar affinity ($K_D = 1.30$). Wolff et al. (43) previously identified PD-166326 as a tyrosine-kinase inhibitor but did not report any binding to *EPHB1*, and DrugBank (44) lists only *ABL1* as a known target (DrugBank ID: DB08339). *EPHB1* has been implicated in chronic pain (45, 46); at the time of publication, there are no known inhibitors of *EPHB1* listed in the Protein Kinase Inhibitor Database (PKIDB) (47), and our findings indicate that PD-166326 may act as a binder to *EPHB1*. Future work could involve further characterization of this interaction, its impact on *EPHB1* function, and possible therapeutic outcomes. In Fig. 4*B*, we show that PD-166326 is the only compound from our screen close to *EPHB1* in coembedding space.

Table 2. We selected and tested 19 potential binding interactions, where the selection of tests was done based solely on ConPLex-predicted interaction and without consulting previous experiments or literature

	EGFR	EPHB1	FLT3	KIT	TGFBR2
AG-1478	0.33*	–	–	–	–
Gefitinib	0.60*	–	–	–	–
Janex 1	26.00	–	–	–	–
SB-431542	>1e4	–	–	–	–
AG-1296	>1e4	–	62.00	27.00	–
ZM 447439	>1e4	–	–	>1e4	–
PD-166326	–	**1.30**	–	–	–
Nintedanib	–	–	0.17*	–	–
Linifanib	–	–	0.72*	1.70	–
Sorafenib	–	–	7.20	36.00	–
Imatinib	–	–	–	6.00	–
Wortmannin	–	–	–	–	>1e4
Pluripotin	–	–	–	–	>1e4
Monorden	–	–	–	–	>1e4

We determined the K_D values for each interaction via an in vitro biochemical assay (*Methods*), and we show here the K_D in nM units. Twelve exhibited binding affinity in the nanomolar range, including four (denoted with *) binding with subnanomolar affinity. The only target for which we incorrectly predicted there would be hits was *TGFBR2*, which has no known inhibitors in PKIDB (47), suggesting that it may be difficult to target. We recapitulate several known interactions (*Results*) and find a tightly binding interaction between *EPHB1* and *PD 166326* (**bold**), which to our knowledge has not been previously characterized.

Notably, all three of the compounds that we predicted to interact with *TGFBR2* were false positives (Wortmannin, Pluripotin, and Monorden). Despite its significance in cancer signaling (48), there are no known inhibitors of *TGFBR2* in PKIDB, suggesting that it may be difficult to target via small-molecule drugs.

Additionally, we found that ConPLex predictions were well calibrated. Using varying thresholds, we can compute a precision–recall curve (over 19 data points, AUPR = 0.91). For high-precision screening, we recommend using a ConPLex-predicted threshold of 0.923 (*SI Appendix*, S7).

Incorporating Drug Binding Information Improves Protein Representations. One of the advantages of the coembedding approach that our model takes is the ability to visualize and investigate the shared embedding space. For instance, we show in Fig. 4 *A–C* that kinases and their inhibitors tend to colocalize within the space. Seeking to expand our analysis, we subsequently mapped all 2,716 predicted surface proteins from the Surfaceome database into ConPLex embedding space and investigated their representations. In Fig. 4*D*, we show the projections all Surfaceome proteins, colored by their classification into one of five functional categories [from Almén et al. (49)]—transporters, receptors, enzymes, miscellaneous, and those that are unclassified. ConPLex projections of surface proteins cluster in embedding space by functional type, with transporters and receptors especially separating from other classes.

However, the Almén functional classification is quite broad and may group proteins with vastly different functions and binding properties. We further demonstrate the link between ConPLex projections and protein function, by evaluating how the learned DTI embedding space separates proteins by domains contained therein. We identified Pfam domains (50) for each protein in the Surfaceome database using HMMscan (51) and compared the projections of proteins that share the same domains. We identified 780 unique domains across all proteins, of which 126 domains were represented in at least 10 proteins. To quantitatively evaluate the coherence of ConPLex embeddings, we trained separate logistic regression classifiers for each domain to separate proteins with that domain from others and used the model's confidence ($log(\frac{p}{1-p})$) for in-sample proteins as a measure of separation for the domain. We find that for all 126 domains, the model more confidently discriminated domains when trained on ConPLex representations than the baseline ProtBert embeddings.

Fig. 4*F* shows the change in confidence scores for all 126 domains, where the dotted line represents equal confidence using either ConPLex or ProtBert. We find that prediction of all domains was improved using ConPlex. However, proteins with kinase domains (PF14575, PF01404, PF00069, and PF07714) separated especially well, whereas 7-transmembrane (7TM) domains characteristic of GPCRs (PF00001, PF00002, PF00003, and PF13853) showed more modest improvement (*SI Appendix*, S6). In Fig. 4*E*, we show the same visualization of projections as in Fig. 4*D* but colored by another top-differentiated domain, PKinase (PF00069). As discussed previously, ConPLex was trained contrastively with several kinase targets and excels at kinase prediction on DUD-E (Fig. 3*E*), so it is unsurprising that proteins with these domains separate well. In fact, one of the top differentiated domains is the Ephrin ligand binding domain (PF01404), which is responsible for binding to the ephrin ligand (52). While the model was also trained contrastively with 7TM GPCR targets, many fewer training data samples were available. In addition to the dearth of training examples, GPCRs are less soluble than kinases and tend to exhibit more dynamic behavior—all of which contribute to difficulty predicting ligand binding. Future work in this area might adjust distances in this landscape to account for the low metric entropy of biological sequences, as demonstrated by Berger et al. (53).

Adapting ConPLex for Affinity Prediction. While we have to this point been using the model to predict probabilities of interaction and perform binary classification, we show that ConPLex can be easily adapted to perform binding affinity prediction and that this model too achieves state-of-the-art performance. The final step of our binary interaction predictor is converting the cosine distance between the projections in the DTI space to a probability using a sigmoid activation (*Methods*). However, it is completely natural to replace this activation with a dot product between the two projections, which enables the model to make real-valued predictions, which can then be interpreted as a binding affinity. We evaluated ConPLex trained for affinity prediction on the Therapeautics Data Commons (TDC) DTI Domain Generalization (**TDC-DG**) benchmark. The TDC-DG benchmark contains binding affinity (IC_{50}) data from interactions patented between 2013 and 2018, with the test set drawn from interactions patented in 2019 and 2021 (*Methods*). Thus, these data require out-of-domain generalization and correspond to the real-life scenario of training on interactions up to a known point and predicting interactions which are yet to be documented. We trained ConPLex to predict binding affinity with five random train/validation splits and achieve an average Pearson correlation (PCC) coefficient between the true and predicted affinity of $0.538(\pm 0.008)$ on the held-out test set. At submission, ConPLex is the top-performing method on the TDC-DG benchmark on TDC (Table 3).

To investigate the strengths and limitations of ConPLex for affinity prediction, we evaluated performance by target type. Targets were annotated with Pfam domains (50) using HMMscan (51), and the PCC between predicted and true IC_{50} was computed over targets in each family (full details *SI Appendix*, S8). We observed especially strong performance on

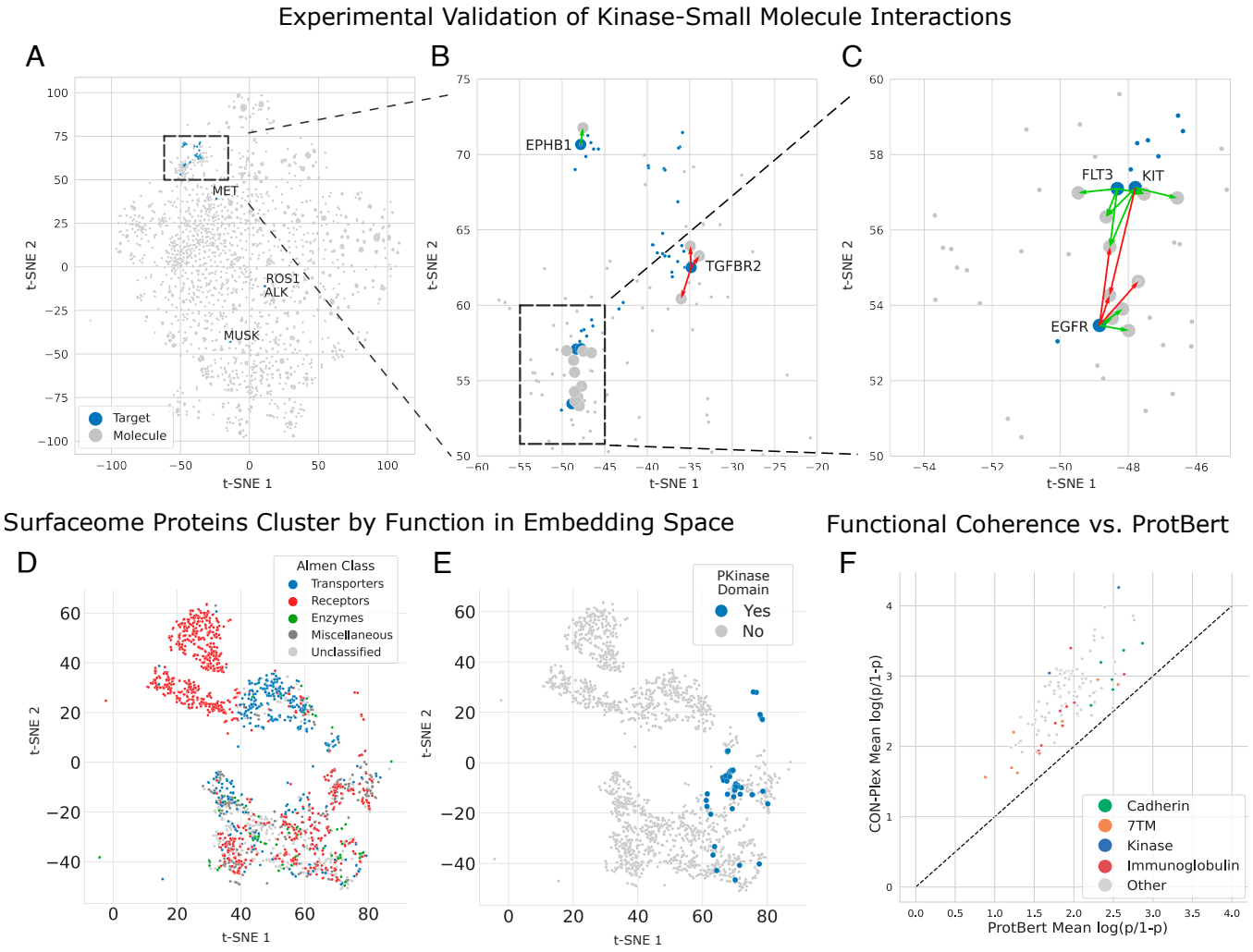

Fig. 4. The shared representation space learned by ConPLex captures DTI and protein function. (*A*) We show that 51 kinases from the Surfaceome (17) database cluster together in ConPLex embedding space but occupy just a small section of the entire space when coembedded with the compounds from the ZINC (37) Cayman-purchasable library. (*B* and *C*) Zooming in on the full embedding space highlights drug–target pairs chose for experimental validation. *EPHB1* has only a single compound nearby in embedding space, PD-166326, which was confirmed to bind with single-digit nanomolar affinity. *FLT3* and *KIT* are neighbors in embedding space and tightly bind many of the same compounds; both bind to Linifanib with $< 2nM$ affinity. *EGFR* was not found to bind to any of the compounds also tested with *FLT3* and *KIT* but binds three other drugs nearby in the embedding space, two of which bind with subnanomolar affinity. On the other hand, none of the three compounds we tested nearby *TGFBR2* (Wortmannin, Pluripotin, and Monorden) were found to bind. (*D*) ConPLex representations of all cell surface proteins from the Surfaceome (17) cluster by functional class as assigned in Almén et al. (49). (*E*) These representations also cluster by several functional Pfam domains (50), such as the PKinase domain (PF00069) shown in blue. (*F*) We evaluated the coherence of representations for each domain by training a logistic regression classifier and report the model's average confidence for proteins containing that domain as $log(\frac{p}{1-p})$. ConPLex separates all 126 domains better than the untransformed ProtBert embeddings (*SI Appendix*, S6, $P = 4.85 \times 10^{-54}$, paired t test), discriminating kinase domains (blue) especially well. We have also highlighted other classes of domains, including cadherins (green), 7-transmembrane proteins (orange), and immunoglobulins (red).

12 immunoglobulin targets (Pfam domains PF13927, PF13895, PF07679, PF00047), where we observed a PCC of 0.803. In keeping with our previous finding of ConPLex's relative strength on kinases over GPCRs (Fig. 3*E*), we observed a correlation of 0.578 on 94 protein targets with PKinase domains (PF07714, PF00069), including targets with SH3 domains (PF00018, PF07653; 13 targets; PCC = 0.705) and PI3K domains (PF00613, PF00792; 6 targets; PCC = 0.633). However, we observed substantially weaker performance on 7TM domains (PF00001; 14 targets; PCC = 0.254) and GPCR domains (PF10320; 8 targets; PCC = 0.176). To assess ConPLex's variability in its accuracy, we computed a 95% prediction interval based on a linear regression between the true and predicted IC_{50} (*SI Appendix*, S8). While the correlations were strong, we found substantial variability around the true IC_{50}, with the width of the prediction interval around the true $ln(IC_{50})$ being ± 4.89 $ln(nM)$. Altogether, the variability in ConPLex's performance across domains makes it important to understand the target of interest when using ConPLex to predict binding affinity.

Discussion

Much previous work has recognized the value of meaningful drug representations (54, 55) for DTI prediction, yet relatively little work has focused on the target protein representation. As a method to use pretrained PLMs for DTI prediction, ConPLex is yet another example of the power of transferring learned representations for biology (13, 31, 32, 56, 57). This approach enables broad generalization to unseen proteins as well as extremely fast model inference (>10× speed-up even over other sequence-based approaches *SI Appendix*, S4). This speed is

Table 3. ConPLex can be adapted for state-of-the-art affinity prediction

Model	PCC
ConPLex	0.538 ± 0.008
MMD	0.433 ± 0.010
CORAL	0.432 ± 0.010
ERM	0.427 ± 0.012
MTL	0.425 ± 0.010
GroupDRO	0.384 ± 0.006
AndMASK	0.288 ± 0.019
IRM	0.284 ± 0.021

By replacing the cosine distance in the final step of ConPLex with a dot product between the projections, ConPLex can be used for affinity prediction rather than binary classification. The TDC-DG dataset contains IC_{50} values for patented drug–target pairs, where training/testing data are split from before/after 2018. We report the average and SD of the PCC coefficient between true and predicted values across five train/validation splits. Metrics for all methods other than ConPLex come from the TDC leaderboard (19, 58), where at the time of submission, ConPLex is the best-performing method.

particularly valuable for drug repurposing and iterative screening, where large compound libraries are evaluated against hitherto-uncharacterized proteins implicated in a disease of interest. The coembedding approach which enables this speedup could also be effective for integrative multistructure models [e.g., the IMP framework (59)] where efficient scanning of possible combinations is important. Recent methods have also demonstrated the power of PLMs for transferring knowledge between species (32), and our framework may enable more accurate transfer of DTI from the model organisms on which drugs are initially tested, to their eventual use in human patients. Skolnick and Zhou (60) have reported the importance of considering small molecule binding pockets for protein–protein interaction prediction; thus, our DTI-informed protein representations may also be useful in that context. While structural similarity is often implicitly learned by PLMs, future work could explicitly incorporate structure where such data are available, perhaps by incorporating a more advanced projection architecture like the Geoformer (3).

It has been shown in previous work that the performance of different PLMs varies on different tasks and that there is not one clearly "best" language model (14, 61, 62). While we have chosen to use ProtBert here, it is likely that other existing or newly developed language models may yield better performance for certain types of drugs or targets. Likewise, advancements in drug representation may improve performance—the ConPLex framework is flexible to different input features, and it remains important to experiment with different feature choices for the task at hand (*SI Appendix*, S2).

ConPLex approaches the DTI decoy problem from the perspective of adversarial machine learning, where the model must act as a discriminator for adversarial examples from the decoy database. This approach is directly enabled by the coembedding architecture—to compute the triplet distance loss, the protein and drugs must be coembedded, and the distance between them must be meaningful and simply computed. Such an approach would not be feasible using a model which concatenates features up front, nor for a model which has significant computation defining the probability of interaction after the coembedding. Thus, the shared lexicographic space in which we embed the proteins, targets, and decoys is key. Future work could explore adapting molecular generation methods such as JT-VAE or HierG2G (63, 64) to directly act as a generator for decoys. High-specificity DTI prediction is valuable beyond decoy detection—greater specificity of inference can help improve personalized medicine or the modeling of drug effects against rare variants from underrepresented populations.

It is also important to consider the coverage of the problem to select an appropriate method. While we recommend the use of PLM-based features in all cases, if enough data are available, for specific enzyme-family prediction tasks, we still recommend the use of single-task models (25). To verify individual interactions, energy-based molecular docking will likely be more accurate, although at the cost of being substantially slower (4). Different classes of computational tools for DTI prediction each have varying strengths, and the highest quality predictions can be achieved by leveraging all of these methods together where each is most fit.

Drug discovery is a fundamental task for human health yet remains both extremely expensive and time consuming, with the median drug requiring over 1 billion dollars (65) and 10 y (66) from development to approval and distribution. While experimental results will remain the gold standard for validating drug functionality, in silico prediction of drug–target binding remains much faster and cheaper and so will continue to play an important role in early screening of therapeutic candidates (67). To address this step in the drug design pipeline, we have introduced ConPLex. DTI prediction methods should be able to generalize to unseen types of drugs and targets, while also discriminating between highly similar molecules with different binding properties. ConPLex tackles both of these challenges through its dual use of PLMs and contrastive learning. We hope that its broad applicability, specificity on decoys, and ability to scale to massive data will allow ConPLex to be a critical step in this pipeline and contribute to the efficient discovery of effective therapeutics.

Materials and Methods

Computing Dataset Coverage. Let $1_{(i,j)}$ be the indicator variable, meaning there exists an observation of drug i and target j. For a dataset with m unique drugs and n unique targets, we can define the coverage for drug d as $C_d = \frac{1}{n}\sum_{j=0}^{n} 1_{(d,j)}$ and for a target t as $C_t = \frac{1}{m}\sum_{i=0}^{m} 1_{(i,t)}$. Then, for a given dataset, we can evaluate the median drug and target coverage. A dataset with maximum coverage would have a single data point for each drug–target pair and, thus, a median coverage of 1 for both drugs and targets. Conversely, each drug and target would be represented only a single time in a minimum coverage dataset, resulting in drug and target coverages of $\frac{1}{n}$ and $\frac{1}{m}$, respectively. We report the median drug and target coverage for each benchmark dataset in Table 4. Since the DUD-E dataset is separated out by targets, we instead report the median number of drugs against each target.

Benchmarks Overview.

Low coverage benchmarks. We evaluate our framework on three broad-scale, low-coverage benchmark datasets. Two datasets, **DAVIS** (68) and **BindingDB** (69), consist of pairs of drugs and targets with experimentally determined dissociation constants (K_D). Following ref. 13, we treat pairs with $K_D < 30$ as positive DTIs, while larger K_D values are negative. The third dataset, ChG-Miner from **BIOSNAP** (70), consists of only positive DTIs. We create negative DTIs by randomly sampling an equal number of protein–drug pairs, making the assumption that a random pair is unlikely to be positively interacting. The DAVIS dataset represents a few-shot learning setting: It contains only 2,086 training interactions, compared to 12,668 for BindingDB and 19,238 for BIOSNAP. The rest of the data preparation follows (13). The datasets are split into 70% for training, 10% for validation, and the remaining 20% for testing. Training data are artificially subsampled to have an equal number of positive and negative interactions, while validation and test data are left at the ratio originally in the dataset.

Table 4. Full specification of benchmark datasets

Dataset	Drugs	Targets	Median Coverage	# Training	# Validation	# Test
BIOSNAP	4,510	2,181	0.0023/0.0020	9,670/9,568	1,396/1,352	2,770/2,727
Unseen Drugs				9,535/9,616	1,383/1,353	2,918/2,675
Unseen Targets				9,876/9,499	1,382/1,386	2,578/2,762
BindingDB	7,165	1,254	0.0008/0.0010	6,334/6,334	927/5,717	1,905/11,384
DAVIS	68	379	0.3707/0.3676	1,043/1,043	160/2,846	303/5,708
TDC-DG	140,746	477	0.0021/0.0005	146,891	36,539	49,028
Phosphatase	165	218	1.0/1.0	5,054/27,286	—	370/3,260
Esterase	96	146	1.0/1.0	2,150/10,426	—	926/514
Glycosyltransferase	89	54	0.9259/0.9778	725/3,042	—	113/417
Halogenase	62	42	1.0/1.0	303/1,991	—	20/290
BKACE	17	161	1.0/1.0	255/2,193	—	19/270
DUD-E †				8,996/406,208	—	11,430/521,132
GPCR	99,671	5	18,563			
Kinase	315,399	26	15,409			
Protease	286,089	15	9,271			
Nuclear	151,133	11	16,257			

We report the number of unique drugs and targets, the median (drug/target) coverage, and the number of training, validation, and test samples in each dataset. The numbers of pairs are shown as (positive/negative), except for TDC-DG (19, 58), which is a regression task; thus, the total number of pairs is shown. We consider BIOSNAP (70), BindingDB (69), DAVIS (68), and TDC-DG as low-coverage, while Phosphatase (71), Esterase (72), Glycosyltransferase (73), Halogenase (74), BKACE (75), and DUD-E (36) are considered high-coverage. † Because DUD-E is a decoy dataset, we report as coverage the median number of true drugs or decoys for each target.

Zero-shot benchmarks. We evaluate our framework on two zero-shot prediction modifications of BIOSNAP. Following ref. 13, the **Unseen proteins** set was created by selecting 20% of proteins from the full set and selecting any interactions including these proteins for the test set. Thus, there are no proteins which appear in both the training and test set. The corresponding process was used to create the **Unseen drugs** dataset. The training set was then further split using 7/8 of the interactions for training and 1/8 of the interactions for testing. As above, data are subsampled so that training is balanced.

Continuous benchmarks. Continuous affinity prediction data come from the TDC-DG (19). The TDC-DG consists of 140,746 unique drugs and 477 unique targets derived from BindingDB (69) interactions that have patent information. Each interaction is labeled with an experimentally determined dissociation constant (IC_{50}). Interactions are temporally split, so that training pairs are from patents filed between 2013 and 2018, and test pairs are from between 2019 and 2021. In addition, 20% of the training pairs are randomly set aside as a validation set. We train five different models with the train/validation splits determined by the TDC benchmarking framework and report the average PCC coefficient of predictions on the test set.

High coverage benchmarks. The Database of Useful Decoys: Enhanced (**DUD-E**) (36) consists of 102 protein targets and known binding partners (average 224 molecules per target). For each binding partner, there are 50 "decoys," or physiochemically similar compounds that are known not to bind with the target. Of note, 57 of the targets are classified as either GPCRs, kinases, nuclear proteins, or proteases. We generate train–test splits by splitting targets within classes, so that there are representative members of each class in both the training and test sets, but no target appears in both the training and test set (26 train and 31 test). These data are by definition high coverage since there are several true and decoy compounds available for each target. We provide the full target splits in *SI Appendix*, S1.

We also evaluate several protein-family-specific datasets from various different sources compiled by Goldman et al. (25). These include DTI data on β-ketoacid cleavage (**BKACE**) (75), **Esterase** (72), **Glycosyltransferases** (73), **Halogenase** (74), and **Phosphatase** (71) enzymes. These data are uniformly very high coverage, with a known data point for nearly every drug–target pair. Following ref. 25, we performed a 10-fold cross-validation where the data were split into train–test sets by target, so that all drugs appear in both the training and test set, but no target does.

ConPLex Model.

Target featurization. We generate protein target features using pretrained PLMs: These models generate a protein embedding $E_{full} \in \mathbb{R}^{n \times d_t}$ for a protein of length n, which is then mean-pooled along the length of the protein resulting in a vector $E \in \mathbb{R}^{d_t}$. Specifically, we investigate the pretrained models Prose (30), ESM (76), and ProtBert (27), with default dimensions $d_t = 6165, 1280, and 1024$, respectively (*SI Appendix*, S2). Elnaggar et al. recommend the use of ProtT5XLUniref50, but we found that it did not perform as well as ProtBert for the DTI prediction task. We emphasize that the language and projection models are used exclusively to generate input features–their weights are kept unchanged and are not updated during DTI training.

Drug featurization. We featurize the drug molecule by its Morgan fingerprint (26), an encoding of the SMILES string of the molecular graph as a fixed-dimension embedding $M \in \mathbb{R}^{d_m}$ (we chose $d_m = 2{,}048$) by considering the local neighborhood around each atom. The utility of the Morgan fingerprint for small molecule representation has been demonstrated in refs. 25 and 77. We additionally investigated the use of molecule embeddings from Mol2Vec (78) and MolR (79) and found that they failed to perform as well as the Morgan fingerprint (*SI Appendix*, S2).

Transformation into a shared latent space and prediction. Given a target embedding $T \in \mathbb{R}^{d_t}$ and small molecule embedding $M \in \mathbb{R}^{d_m}$, we transform them separately into $T^*, M^* \in \mathbb{R}^h$ using a single fully connected layer with a ReLU activation. These layers are parameterized with weight matrices $W_t \in \mathbb{R}^{h \times d_t}$, $W_m \in \mathbb{R}^{h \times d_m}$, and bias vectors $b_t, b_m \in \mathbb{R}^h$.

$$T^* = ReLU(W_t T + b_t) \quad [1]$$

$$M^* = ReLU(W_m M + b_m) \quad [2]$$

Given the latent embeddings T^*, M^*, we compute the probability of a DTI $\hat{p}(T^*, M^*)$ as the cosine similarity between the embedding vectors, followed by a sigmoid activation. Thus, we compute the predicted probability as:

$$\hat{p}(T^*, M^*) = \sigma\left(\frac{T^* \cdot M^*}{||T^*||_2 \cdot ||M^*||_2}\right) \quad [3]$$

When predicting compound binding affinity $\hat{y}(T^*, M^*)$, we substitute the sigmoid and cosine similarity (Eq. **3**) with a dot product followed by a ReLU activation, which gives a nonnegative distance in the embedding space (Eq. **4**).

$$\hat{y}(T^*, M^*) = ReLU(T^* \cdot M^*) \quad [4]$$

Training. The model is trained both for broad and fine predictions, with the loss computed depending on the training dataset. Broad-scale training data use the binary cross-entropy loss (L_{BCE}) between the true labels y and the predicted

interaction probabilities $\hat{p}$. When the model is trained to predict binding affinity, we substitute the binary cross-entropy loss with the mean squared error loss (L_{MSE}) during supervision.

Training on fine-scale data (DUD-E) was performed using contrastive learning. Contrastive learning uses triplets of training points rather than pairs, denoted the **anchor**, **positive**, and **negative**, and aims to minimize the distance between the anchor and positive examples while maximizing the distance between the anchor and the negative examples. In the DTI setting, the natural choice for a triplet is the protein target as the anchor, the true drug as the positive, and decoy as the negative example, respectively. We derive a training set of triplets in the following manner: For each known interacting drug–target pair (T, M^+), we randomly sample $k = 50$ noninteracting pairs (T, M^-) and generate the triplets (T, M^+, M^-), where M^- is drawn from the set of all decoys against T. We map these to latent space embeddings as described above. Since all the entities are now comparable to each other, we can compute the triplet margin-distance loss (L_{TRM}).

$$L_{TRM}(a, p, n) = \frac{1}{N}\sum_{i=1}^{N} max(D(a, p) - D(a, n) + m, 0) \quad [5]$$

where

$$D(u, v) = 1 - \hat{p}(u, v) \quad [6]$$

The margin m sets the maximum required delta between distances, above which the loss is zero.

Margin annealing. The margin m sets the maximum required delta between distances, above which the loss is zero. Initially, a large margin requires the decoy to be much further from the target than the drug to avoid a penalty, resulting in larger weight updates. As training progresses, lower margins relax this constraint, requiring only that the drug be closer than the decoy as $m \to 0$. Here, the margin is initialized at $M_{max} = 0.25$ according to a tanh decay with restarts schedule. Every $E_{max} = 10$ contrastive epochs, the margin is reset to the initial M_{max}, for a total of 50 epochs. At epoch i, the margin is set to

$$m(i) = M_{max}(1 - \tanh(\frac{2(i \bmod E_{max})}{E_{max}})) \quad [7]$$

Implementation. Model weights were initialized using the Xavier method from a normal distribution (80). Weights were updated with error backpropagation using the AdamW optimizer (81) for a total of 50 epochs. For the binary classification task, the learning rate was initially set to 10^{-4} and adjusted according to a cosine annealing schedule with warm restarts (82) every 10 epochs. For the contrastive task, the learning rate was initially set to 10^{-5}, and the same annealing schedule was followed. The margin for the contrastive loss was initially set to 0.25 and decreased to a minimum of 0 over 50 epochs according to a tanh decay schedule with restarts every 10 epochs. We used a latent dimension $d = 1{,}024$ (results were robust to even with lower dimensions, and much higher dimensions may overfit or be subject to topological restrictions) and a batch size of 32. The model was implemented in PyTorch version 1.11. Model training, and inference was performed on a machine with a 112-core Intel Xeon Gold 6258R CPU and using a single NVIDIA A100 GPU. We compare training and inference run times in *SI Appendix*, S4.

Surfaceome Analysis. We evaluate the functional use of ConPLex embeddings using data from the Surfaceome database (17), which contains 2,886 cell-surface proteins. We identified Pfam domains using HMMscan from HMMER3 (51) with default settings. We analyzed domains hit in >10 proteins. For each domain, we trained a logistic regression classifier from sklearn with balanced class weights. We also evaluated domain coherence using spectral clustering with $k = 10$ clusters and evaluated the adjusted mutual information (AMI) between true clusters (protein has/does not have domain) and predicted clusters (*SI Appendix*, S6).

Experimental Determination of Kinase Binding Affinity. From the Surfaceome (17) database, we selected 51 kinases which were available by the KdELECT assay from DiscoveryX. From the ZINC database (37), we selected 4715 compounds purchasable from the Cayman Chemical Company. Using a ConPLex model trained on BindingDB and fine-tuned on DUD-E, we predicted all pairwise interactions between kinases and small molecule drugs. Without previously consulting the literature on kinases or drugs, we selected 5 kinases which were highly represented in the top predictions (*EGFR*, *EPHB1*, *FLT3*, *KIT*, and *TGFBR2*). We then selected 19 binding pairs to test, covering 14 drugs with high ConPLex-predicted interactions. The full list of ConPLex predictions can be found in *SI Appendix*, Data S1.

We performed K_D determination using the KdELECT assay from the DiscoveryX company, following the procedure from Hie et al. (83). KdELECT measures competition between test compounds and an immobilized, active site-directed ligand. Ligands are tagged with DNA oligomers, and competition is measured by qPCR of this barcode. BL21-derived *E. coli* were infected with T7 phase strains tagged with each kinase target and incubated with shaking at 32°C. Streptavidin-coated magnetic beads were treated with a biotinylated ligand at room temperature for 30 min, following which the beads were blocked with excess biotin and washed with blocking buffer [SeaBlock (Pierce), 1% bovine serum albumin (BSA), 0.05% Tween 20, and 1 mM dithiothreitol (DTT)] to remove unbound ligand. Test compounds were prepared as 111X stocks in 100% DMSO. An 11-point, threefold compound dilution series was created, with a top test compound concentration of 10,000 nM. Three DMSO control points were also used. Test compounds are distributed by acoustic transfer (noncontact dispensing) in 100% DMSO and then diluted into the assays for a final DMSO concentration of 0.9%.

Kinases, ligand-bound affinity beads, and test compounds were combined in 1X binding buffer [20% SeaBlock, 0.17X phosphate-buffered saline (PBS), 0.05% Tween 20, and 6 mM DTT] in a 384-well plate, with a final volume of 0.02 mL for each reaction. Plates were incubated for 1 h at room temperature with shaking. Affinity beads were washed with wash buffer (1x PBS, 0.05% Tween 20), resuspended in elution buffer (1x PBS, 0.05% Tween 20, 0.5 mM nonbiotinylated affinity ligand), and incubated for 30 min at room temperature with shaking. The concentration of kinases was measured using qPCR. To compute K_D of the binding, a standard dose–response curve was fit to the Hill equation curves using the Levenberg–Marquardt algorithm (Hill slope = −1).

Genome-wide ChEMBL Scan. We trained a ConPLex model using BindingDB and DUD-E and used it to make predictions for all pairs of human proteins against all drugs in ChEMBL. Human protein sequences were taken from the STRING database and processed following ref. 32, resulting in 15,816 proteins between 50 and 800 amino acids long. Small molecule structures were downloaded from ChEMBL 30 (18), resulting in 1,533,652 compounds. Prediction took just under a day, accounting for embedding time.

Data, Materials, and Software Availability. Dataset data have been deposited in Github (https://github.com/samsledje/ConPLex_dev) (38). Previously published data were used for this work (19, 36, 68–75).

ACKNOWLEDGMENTS. R.S. and B. Berger were supported by the NIH grant R35GM141861. S.S. was supported by the NSF Graduate Research Fellowship under Grant No. 2141064. B. Bryson was supported by the Terry and Susan Ragon Foundation. L.C. was supported by NSF grant CCF-1934553. We thank Kapil Devkota, Mert Erden, Tristan Bepler, and Tim Truong for helpful discussions.

Author affiliations: [a]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139; [c]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Department of Computer Science, Tufts University, Medford, MA 02155; and [e]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

1. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
2. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
3. R. Wu *et al.*, High-resolution de novo structure prediction from primary sequence. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.07.21.500999 (Accessed 7 December 2022).
4. L. Pinzi, G. Rastelli, Molecular docking: Shifting paradigms in drug discovery. *Int. J. Mol. Sci.* **20**, 4331 (2019).

5. B. M. Bonk, Y. Tarasova, M. A. Hicks, B. Tidor, K. L. Prather, Rational design of thiolase substrate specificity for metabolic engineering applications. *Biotechnol. Bioeng.* **115**, 2167–2182 (2018).
6. R. C. de Melo-Minardi, K. Bastard, F. Artiguenave, Identification of subfamily-specific sites based on active sites modeling and clustering. *Bioinformatics* **26**, 3075–3082 (2010).
7. S. J. Trudeau *et al*., PrePCI: A structure- and chemical similarity-informed database of predicted protein compound interactions. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.09.17.508184 (Accessed 7 December 2022).
8. R. Singh, D. Park, J. Xu, R. Hosur, B. Berger, Struct2Net: A web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Res.* **38**, W508–W515 (2010).
9. E. Anderson, G. D. Veith, D. Weininger, *SMILES, A Line Notation and Computerized Interpreter for Chemical Structures* (Environmental Research Laboratory, US Environmental Protection Agency, 1987).
10. M. Bagherian *et al*., Machine learning approaches and databases for prediction of drug–target interaction: A survey paper. *Brief. Bioinf.* **22**, 247–269 (2021).
11. B. Hie, H. Cho, B. Berger, Realizing private and practical pharmacological collaboration. *Science* **362**, 347–350 (2018).
12. I. Lee, J. Keum, H. Nam, DeepConv-DTI: Prediction of drug–target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **15**, e1007129 (2019).
13. K. Huang, C. Xiao, L. M. Glass, J. Sun, MolTrans: Molecular interaction transformer for drug–target interaction prediction. *Bioinformatics* **37**, 830–836 (2021).
14. S. Sledzieski, R. Singh, L. Cowen, B. Berger. "Adapting protein language models for rapid DTI prediction in *Machine Learning for Structural Biology Workshop (MLSB) at NeurIPS* (2021).
15. R. Bommasani *et al*., On the opportunities and risks of foundation models. arXiv [Preprint] (2021). http://arxiv.org/abs/2108.07258 (Accessed 7 December 2022).
16. S. Gururangan *et al*., Don't stop pretraining: Adapt language models to domains and tasks. arXiv [Preprint] (2020). http://arxiv.org/abs/2004.10964 (Accessed 7 December 2022).
17. D. Bausch-Fluck *et al*., The in silico human surfaceome. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10988–E10997 (2018).
18. D. Mendez *et al*., ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
19. K. Huang *et al*., Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. arXiv [Preprint] (2021). http://arxiv.org/abs/2102.09548 (Accessed 7 December 2022).
20. N. Zong *et al*., Beta: A comprehensive benchmark for computational drug–target prediction. *Brief. Bioinf.* **23**, bbac199 (2022).
21. H. M. Dönertaş, M. Fuentealba Valenzuela, L. Partridge, J. M. Thornton, Gene expression-based drug repurposing to target aging. *Aging Cell* **17**, e12819 (2018).
22. D. Morselli Gysi *et al*., Network medicine framework for identifying drug-repurposing opportunities for Covid-19. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025581118 (2021).
23. K. Huang *et al*., DeepPurpose: A deep learning library for drug–target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2020).
24. Y. Huang *et al*., A framework for identification of on- and off-target transcriptional responses to drug treatment. *Sci. Rep.* **9**, 1–9 (2019).
25. S. Goldman, R. Das, K. K. Yang, C. W. Coley, Machine learning modeling of family wide enzyme-substrate specificity screens. *PLoS Comput. Biol.* **18**, e1009853 (2022).
26. H. L. Morgan, The generation of a unique machine description for chemical structures–A technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
27. A. Elnaggar *et al*., ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv [Preprint] (2020). http://arxiv.org/abs/2007.06225 (Accessed 7 December 2022).
28. A. Scaiewicz, M. Levitt, The language of the protein universe. *Curr. Opin. Genet. Dev.* **35**, 50–56 (2015).
29. T. Bepler, B. Berger, "Learning protein sequence embeddings using information from structure" in *7th International Conference on Learning Representations, ICLR 2019* (2019).
30. T. Bepler, B. Berger, Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
31. M. Heinzinger *et al*., Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf.* **20**, 1–17 (2019).
32. S. Sledzieski, R. Singh, L. Cowen, B. Berger, D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein–protein interactions. *Cell Syst.* **12**, 1–14 (2021).
33. R. Singh, K. Devkota, S. Sledzieski, B. Berger, L. Cowen, Topsy-Turvy: Integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38**, i264–i272 (2022).
34. M. Tsubaki, K. Tomii, J. Sese, Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**, 309–318 (2019).
35. M. Heinzinger *et al*., Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genom. Bioinf.* **4**, lqac043 (2022).
36. M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet, Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **55**, 6582–6594 (2012).
37. J. J. Irwin, B. K. Shoichet, Zinc–A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
38. R. Singh, S. Sledzieski, B. Bryson, L. Cowen, B. Berger, surfaceome_cayman_validation_scan.csv. Github. https://github.com/samsledje/ConPLex_dev/blob/main/dataset/surfaceome_cayman_validation_scan.csv. Deposited 20 March 2023.
39. W. B. Weglicki, J. H. Kramer, C. F. Spurney, J. J. Chmielinska, I. T. Mak, The EGFR tyrosine kinase inhibitor tyrphostin AG-1478 causes hypomagnesemia and cardiac dysfunction. *Can. J. Physiol. Pharmacol.* **90**, 1145–1149 (2012).
40. R. Sordella, D. W. Bell, D. A. Haber, J. Settleman, Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science* **305**, 1163–1167 (2004).
41. G. J. Roth *et al*., Nintedanib: From discovery to the clinic. *J. Med. Chem.* **58**, 1053–1063 (2015).
42. E. S. Wang *et al*., Phase 1 trial of linifanib (ABT-869) in patients with refractory or relapsed acute myeloid leukemia. *Leuk. Lymphoma* **53**, 1543–1551 (2012).
43. N. C. Wolff *et al*., PD166326, a novel tyrosine kinase inhibitor, has greater antileukemic activity than imatinib mesylate in a murine model of chronic myeloid leukemia. *Blood* **105**, 3995–4003 (2005).
44. D. S. Wishart *et al*., DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
45. V. Cibert-Goton *et al*., Involvement of EphB1 receptors signalling in models of inflammatory and neuropathic pain. *PLoS ONE* **8**, e53673 (2013).
46. S. Liu *et al*., Blocking EphB1 receptor forward signaling in spinal cord relieves bone cancer pain and rescues analgesic effect of morphine treatment in rodents EphB1 receptor is critical to bone cancer pain. *Cancer Res.* **71**, 4392–4402 (2011).
47. F. Carles, S. Bourg, C. Meyer, P. Bonnet, PKIDB: A curated, annotated and updated database of protein kinase inhibitors in clinical trials. *Molecules* **23**, 908 (2018).
48. Y. Drabsch, P. Ten Dijke, TGF-β signalling and its role in cancer progression and metastasis. *Cancer Metastasis Rev.* **31**, 553–568 (2012).
49. M. S. Almén, K. J. Nordström, R. Fredriksson, H. B. Schiöth, Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.* **7**, 1–14 (2009).
50. S. El-Gebali *et al*., The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
51. S. R. Eddy, Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
52. J. P. Himanen, M. Henkemeyer, D. B. Nikolov, Crystal structure of the ligand-binding domain of the receptor tyrosine kinase EphB2. *Nature* **396**, 486–491 (1998).
53. B. Berger, M. S. Waterman, Y. W. Yu, Levenshtein distance, sequence comparison and biological database search. *IEEE Trans. Inf. Theory.* **67**, 3287–3294 (2020).
54. K. Huang *et al*., DeepPurpose: A deep learning library for drug–target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2021).
55. B. Ramsundar, "Molecular machine learning with DeepChem," PhD thesis (Stanford University, 2018).
56. B. Hie, E. D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
57. M. Littmann, M. Heinzinger, C. Dallago, K. Weissenow, B. Rost, Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.* **11**, 1–15 (2021).
58. I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization. arXiv [Preprint] (2020). http://arxiv.org/abs/2007.01434 (Accessed 7 December 2022).
59. D. Russel *et al*., Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10**, e1001244 (2012).
60. J. Skolnick, H. Zhou, Implications of the essential role of small molecule ligand binding pockets in protein–protein interactions. *J. Phys. Chem. B* **126**, 6853–6867 (2022).
61. B. L. Hie, K. K. Yang, P. S. Kim, Evolutionary velocity with protein language models. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.06.07.447389 (Accessed 7 December 2022).
62. C. Hsu, H. Nisonoff, C. Fannjiang, J. Listgarten, Combining evolutionary and assay-labelled data for protein fitness prediction. bioRxiv [Preprint] (2021). https://doi.org/10.1101/2021.03.28.437402.
63. W. Jin, R. Barzilay, T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation" in *International Conference on Machine Learning* (PMLR, 2018), pp. 2323–2332.
64. W. Jin, R. Barzilay, T. Jaakkola, "Hierarchical generation of molecular graphs using structural motifs" in *International Conference on Machine Learning* (PMLR, 2020), pp. 4839–4848.
65. O. J. Wouters, M. McKee, J. Luyten, Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *J. Am. Med. Assoc.* **323**, 844–853 (2020).
66. G. A. Van Norman, Drugs, devices, and the FDA. Part 1: An overview of approval processes for drugs. *JACC: Basic Transl. Sci.* **1**, 170–179 (2016).
67. V. T. Sabe *et al*., Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* **224**, 113705 (2021).
68. M. I. Davis *et al*., Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
69. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, M. K. Gilson, BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).
70. M. Zitnik, R. Sosič, S. Maheshwari, J. Leskovec, BioSNAP Datasets: Stanford biomedical network dataset collection (2018). http://snap.stanford.edu/biodata.
71. H. Huang *et al*., Panoramic view of a superfamily of phosphatases through substrate profiling. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1974–E1983 (2015).
72. M. Martínez-Martínez *et al*., Determinants and prediction of esterase substrate promiscuity patterns. *ACS Chem. Biol.* **13**, 225–234 (2017).
73. M. Yang *et al*., Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.* **14**, 1109–1117 (2018).
74. B. F. Fisher, H. M. Snodgrass, K. A. Jones, M. C. Andorfer, J. C. Lewis, Site-selective C–H halogenation using flavin-dependent halogenases identified via family-wide activity profiling. *ACS Cent. Sci.* **5**, 1844–1856 (2019).
75. K. Bastard *et al*., Revealing the hidden functional diversity of an enzyme family. *Nat. Chem. Biol.* **10**, 42–49 (2014).
76. A. Rives *et al*., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
77. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
78. S. Jaeger, S. Fulle, S. Turk, Mol2vec: Unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).
79. H. Wang *et al*., Chemical-reaction-aware molecule representation learning. arXiv [Preprint] (2021). http://arxiv.org/abs/2109.09888 (Accessed 7 December 2022).
80. X. Glorot, Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks" in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (JMLR Workshop and Conference Proceedings, 2010), pp. 249–256.
81. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv [Preprint] (2017). http://arxiv.org/abs/1711.05101 (Accessed 7 December 2022).
82. I. Loshchilov, F. Hutter, SGDR: Stochastic gradient descent with warm restarts. arXiv [Preprint] (2019). http://arxiv.org/abs/1608.03983 (Accessed 7 December 2022).
83. B. Hie, B. D. Bryson, B. A. Berger, Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell Syst.* **11**, 461–477.e9 (2020).