

1 Binding Site-enhanced Sequence Pretraining and 2 Out-of-cluster Meta-learning Predict Genome-Wide 3 Chemical-Protein Interactions for Dark Proteins

4 Tian Cai¹, Li Xie², Shuo Zhang¹, Muge Chen³, Di He¹, Amitesh Badkul², Yang
5 Liu², Hari Krishna Namballa⁴, Michael Dorogan⁴, Wayne W. Harding⁴, Cameron
6 Mura⁵, Philip E. Bourne⁵, and Lei Xie^{1,2,6,*}

7 ¹*Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York,
8 10016, USA*

9 ²*Department of Computer Science, Hunter College, The City University of New York, New York, 10065,
10 USA*

11 ³*Master Program in Computer Science, Courant Institute of Mathematical Sciences, New York University*

12 ⁴*Department of Chemistry, Hunter College, The City University of New York, New York, 10065, USA*

13 ⁵*School of Data Science & Department of Biomedical Engineering, University of Virginia, Virginia, 22903,
14 USA*

15 ⁶*Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research
16 Institute, Weill Cornell Medicine, Cornell University, New York, 10021, USA
17 *lei.xie@hunter.cuny.edu*

18 **Keywords**— machine learning, deep learning, out-of-distribution, PortalCG, understudied proteins, compound
19 screening, drug discovery, Opioid Use Disorder, Alzheimer's Disease

20

Abstract

21 Discovering chemical-protein interactions for millions of chemicals across the entire human and
22 pathogen genomes is instrumental for chemical genomics, protein function prediction, drug discovery,
23 and other applications. However, more than 90% of gene families remain dark, i.e., their small molecular
24 ligands are undiscovered due to experimental limitations and human biases. Existing computational
25 approaches typically fail when the unlabeled dark protein of interest differs from those with known ligands
26 or structures. To address this challenge, we developed a deep learning framework PortalCG. PortalCG
27 consists of four novel components: (i) a 3-dimensional ligand binding site enhanced sequence pre-training
28 strategy to represent the whole universe of protein sequences in recognition of evolutionary linkage of
29 ligand binding sites across gene families, (ii) an end-to-end pretraining-fine-tuning strategy to simulate the
30 folding process of protein-ligand interactions and reduce the impact of inaccuracy of predicted structures
31 on function predictions under a sequence-structure-function paradigm, (iii) a new out-of-cluster meta-
32 learning algorithm that extracts and accumulates information learned from predicting ligands of distinct
33 gene families (meta-data) and applies the meta-data to a dark gene family, and (iv) stress model selection
34 that uses different gene families in the test data from those in the training and development data sets to
35 facilitate model deployment in a real-world scenario. In extensive and rigorous benchmark experiments,
36 PortalCG considerably outperformed state-of-the-art techniques of machine learning and protein-ligand
37 docking when applied to dark gene families, and demonstrated its generalization power for off-target
38 predictions and compound screenings under out-of-distribution (OOD) scenarios. Furthermore, in an
39 external validation for the multi-target compound screening, the performance of PortalCG surpassed
40 the human design. Our results also suggested that a differentiable sequence-structure-function deep
41 learning framework where protein structure information serve as an intermediate layer could be superior
42 to conventional methodology where the use of predicted protein structures for predicting protein functions
43 from sequences. We applied PortalCG to two case studies to exemplify its potential in drug discovery:
44 designing selective dual-antagonists of Dopamine receptors for the treatment of Opioid Use Disorder, and
45 illuminating the undruggable human genome for targeting diseases that do not have effective and safe
46 therapeutics. Our results suggested that PortalCG is a viable solution to the OOD problem in exploring
47 the understudied protein functional space.

48

Author Summary

49 Many complex diseases such as Alzheimer’s disease, mental disorders, and substance use disorders do not have
50 effective and safe therapeutics due to the polygenic nature of diseases and the lack of thoroughly validate drug
51 targets and their ligands. Identifying small molecule ligands for all proteins encoded in the human genome will
52 provide new opportunity for drug discovery of currently untreatable diseases. However, the small molecule ligand of
53 more than 90% gene families is completely unknown. Existing protein-ligand docking and machine learning methods
54 often fail when the protein of interest is dissimilar to those with known functions or structures. We develop a
55 new deep learning framework PortalCG for efficiently and accurately predicting ligands of understudied proteins
56 which are out of reach of existing methods. Our method achieves unprecedented accuracy over state-of-the-arts
57 by incorporating ligand binding site information and sequence-to-structure-to-function paradigm into a novel deep
58 meta-learning algorithms. In a case study, the performance of PortalCG surpassed the human design. The proposed
59 computational framework will shed new light into how chemicals modulate biological system as demonstrated by
60 applications to drug repurposing and designing polypharmacology. It will open a new door to developing effective
61 and safe therapeutics for currently incurable diseases. PortalCG can be extended to other scientific inquiries such as
62 predicting protein-protein interactions and protein-nucleic acid recognition.

63

1 Introduction

64 The central aim of scientific inquiry has been to deduce new concepts from existing knowledge or to generalize
65 observations. Numerous such issues exist in the biological sciences. The rise of deep learning has sparked a surge of
66 interest in using machine learning to explore previously unexplored molecular and functional spaces in biology and
67 medicine, ranging from “deorphanizing” G-protein coupled receptors[1] and translating cell-line screens to patient
68 drug responses[2][3], to predicting novel protein structures[4][5][6], to identifying new cell types from single-cell omics
69 data[7]. Illuminating the understudied space of human knowledge is a fundamental problem that one can attempt to
70 address via deep learning—that is, to generalize a “well-trained” model to unseen data that lies Out-of-Distribution
71 (OOD) of the training data, in order to successfully predict outcomes under conditions that the model has never

72 encountered before. While deep learning is capable, in theory, of simulating any function mapping, its generalization
73 power is notoriously limited in the case of distribution shifts[8].

74 The training of a deep learning model starts with a domain-specific model architecture. The final model
75 instance selected and its performance are determined by a series of data-dependent design choices, including model
76 initialization, data used for training/validation/testing, optimization of loss function, and evaluation metrics. Each
77 of these design choices impacts the generalization power of a trained model. The development of several recent
78 deep learning-based approaches—notably transfer learning[9], self-supervised representation learning[10], and meta-
79 learning[11][12]—has been motivated by the OOD challenge. However, each of these methods focuses on only one
80 aspect in the training pipeline of a deep neural network model. Causal learning and mechanism-based modeling (e.g.,
81 based on the first principle of physics) could be a more effective solution to the OOD problem [8], but at present
82 these approaches can be applied only on modest scales because of data scarcity, computational complexity, or limited
83 domain knowledge. Solving large-scale OOD problems in biomedicine via machine learning would benefit from a
84 systematic framework for integrative, beginning-to-end model development and deployment and the incorporation of
85 domain knowledge into the training process.

86 Here, we propose a new deep learning framework, Portal learning of Chemical Genomics (PortalCG) for predicting
87 small-molecule binding to dark proteins whose ligands are unknown and dark gene families in which all protein
88 members do not have known ligands. Here Portal represents multiple training components in an end-to-end deep
89 learning framework used to systematically address OOD challenges. Small molecules act as endogenous or exogenous
90 ligands of proteins, assisting in maintaining homeostasis of a biological system or serving as therapeutics agents
91 to alter pathological processes. Despite tremendous progress in high-throughput screening, the majority of chemical
92 genomics space remains unexplored[13] due to high costs, inherent limitations in experimental approaches, and human
93 biases[14][15]. Even in well-studied gene families such as G-protein coupled receptors (GPCRs), protein kinases,
94 ion channels, and estrogen receptors, a large portion of proteins remain dark[13]. Elucidating dark proteins and
95 gene families can shed light on many essential but poorly understood biological processes, such as microbiome-host
96 interactions mediated by metabolite-protein interactions. Such efforts could also be instrumental for drug discovery.
97 Firstly, although conventional one-drug-one-gene drug discovery process intends to screen drugs against a single target,
98 unrecognized off-target effects are a common occurrence[16]. The off-target is either the cause of undesirable side
99 effects or present unique potential for drug repurposing. Secondly, polypharmacology, i.e., designing drugs that can
100 target multiple proteins, is needed to achieve desired therapeutic efficacy and combat drug resistance for multi-genic
101 diseases[16]. Finally, identifying new druggable targets and discovering their ligands may provide effective therapeutic
102 strategies for currently incurable diseases; for instance, in Alzheimer's disease (AD), many disease-associated genes
103 have been identified through multiple omics studies, but are presently considered undruggable[17].

104 Accurate and robust prediction of chemical-protein interactions (CPIs) across the genome is a challenging OOD
105 problem[1]. If one considers only the reported area under the receiver operating characteristic curve (AUROC),
106 which has achieved 0.9 in many state-of-the-art methods[18][19], it may seem the problem has been solved. However,
107 existing methods have rarely been applied to dark gene families. The performance has been primarily measured in
108 scenarios where the data distribution in the test set does not differ significantly from that in the training set, in
109 terms of similarities between proteins or between chemicals. Few sequence-based methods have been developed and
110 evaluated for an out-of-gene family scenario, where proteins in the test set belong to different (non-homologous) gene
111 families from those in the training set; this sampling bias is even more severe in considering cases where the new gene
112 family does not have any reliable three-dimensional (3D) structural information. Therefore, one can fairly claim that
113 all existing work has been confined to just narrow regions of chemical genomics space for an imputation task, without
114 validated generalizability into the dark proteins for novel discoveries. We have shown that PortalCG significantly
115 outperforms the leading machine learning and protein-ligand docking methods that are available for predicting ligand
116 binding to dark proteins. Thus, PortalCG may shed new light on unknown functions for dark proteins, and empower
117 drug discovery using Artificial Intelligence (AI). To demonstrate the potential of PortalCG, we applied PortalCG
118 to two case studies: designing selective dual-antagonists of Dopamine receptors for Opioid Use Disorder (OUD)
119 with experimental validations, and illuminating the understudied druggable genome for targeting diseases that lack
120 effective and safe therapeutics. The novel genes and their lead compounds identified from PortalCG provide new
121 opportunities for drug discovery to treat currently incurable diseases such as OUDs, and Alzheimer's disease (AD).
122 They warrant further experimental validations.

123 In summary, the contributions of this work are two-fold:

- 124 1. We proposed a new algorithm PortalCG to improve the generalization power of machine learning on OOD
125 problems. Comprehensive benchmark studies demonstrate the promise of PortalCG when applied to exploring
126 the dark gene families in which proteins do not have any known small molecule ligands.
- 127 2. Using PortalCG, we shed new light on unknown protein functions in dark proteins (viz. small molecule-binding
128 properties), and open new avenues in polypharmacology and drug repurposing; as demonstrated by identifying
129 novel drug targets and lead compounds for OUDs and AD.

data split	Common practice	classic scheme applied in OOD	PortalCG	specification
train	IID train	IID train	/	each batch includes data from the same gene family
	/	/	OOD train	data from different gene families are used in each batch
dev	IID-dev	IID-dev	/	from the same gene family as that in the train set
	/	/	OOD-dev	from a different gene family from the training set
test	IID-test	/	/	from the same gene family as that in the training set
	/	OOD-test	OOD-test	from a different gene family from both OOD-dev and training set

Table 1: Data split for stress model instance selection

130 2 Results and Discussion

131 2.1 Overview of PortalCG

132 PortalCG includes four key biology-inspired components: 3-dimensional (3D) binding site-enhanced sequence pre-
133 training, end-to-end sequence-structure-function step-wise transfer learning (STL), out-of-cluster meta-learning (OOC-
134 ML), and stress model selection (see Figure 1).

135 **3D binding site-enhanced sequence pre-training.** Pre-training strategy is a proven powerful approach to
136 boost the generalizability of deep learning models[20]. Pre-trained natural language models have revolutionized
137 Natural Language Processing (NLP)[20]. Significant improvements are also observed when applying the same
138 pre-training strategy to protein sequences for structure[5], function[21][22], and CPI predictions[1]. We begin by
139 performing self-supervised training to map tens of millions of sequences into a universal embedding space, using
140 state-of-the-art *distilled sequence alignment embedding* (DISAE) algorithm [1]. In brief, DISAE first distills the
141 original sequence into an ordered list of amino acid triplets by extracting evolutionarily important positions from
142 a multiple sequence alignment. Then long-range residue-residue interactions can be learned via the Transformer
143 module in ALBERT[10]. A self-supervised masked language modeling (MLM) approach is used where 15% triplets
144 are randomly masked and assumed as unknown and the remaining triplets are used to predict what the masked triplets
145 are. In this way, DISAE leaned the protein sequence representation to capture functional information without the
146 knowledge of their structure and function.

147 3D structural information about the ligand-binding site was used to fine-tune the sequence embedding because
148 it can be evolutionarily related across the fold space and is more informative than the sequence alone for the
149 ligand binding[23]. On the top of pre-trained DISAE embeddings, amino acid residue-ligand atom distance matrices
150 generated from protein-ligand complex structures were predicted from the protein sequence and its ligand. As a result,
151 the original DISAE embedding was re-fined by the 3D ligand binding site information. This structure-regularized
152 protein embedding was used as a hidden layer for supervised learning of cross-gene family CPIs, following an end-to-
153 end sequence-structure-function training process described below.

154 **End-to-end sequence-structure-function STL.** The function of a protein—e.g., serving as a target receptor
155 for ligand binding—stems from its three-dimensional (3D) shape and dynamics which, in turn, is ultimately encoded
156 in its primary amino acid sequence. In general, information about a protein’s structure is more powerful than purely
157 sequence-based information for predicting its molecular function because sequences drift/diverge far more rapidly than
158 do 3D structures on evolutionary timescales. Furthermore, proteins from different gene families may have similar
159 functional sites, thus perform similar functions[23]. Although the number of experimentally-determined structures
160 continues to exponentially increase, and now AlphaFold2 can reliably predict 3D structures of many single-domain
161 proteins, it nevertheless remains quite challenging to directly use protein structures as input for predicting ligand-
162 binding properties of dark proteins. This motivates us to directly use protein sequences to predict ligands of dark
163 proteins in PortalCG. Protein structure information is used as an intermediate layer as trained by the structure-
164 enhanced pre-training to connect a protein sequence and a corresponding protein function (Figure 1A), as inspired
165 by the concept of differentiable biology[24]. By encapsulating the role of structure in this way, inaccuracies and
166 uncertainties in structure prediction are “insulated” and will not propagate to the function prediction. Details of
167 neural network architecture and training methods can be found in section 4.2.

168 **Out-of-cluster meta-learning (OOC-ML).** We designed a new OOC-ML approach to explore dark gene
169 families. Here, predicting ligands of dark gene families can be formulated as the following problem: how can we
170 quickly learn the ligand binding pattern of a new gene family without labeled data from the information obtained
171 from other gene families with a relatively large amount of labeled data? Meta-learning is a general learning strategy
172 that learns a new task without or with few labeled data from outputs (meta-data) generated by multiple other tasks
173 with labeled data, thus naturally fits our purpose. The principle of OOC-ML is first to independently learn the
174 pattern of ligand bindings from each gene family that has labeled data and then to extract the common intrinsic
175 pattern shared by these gene families and apply the learned essential knowledge to dark ones. OOC-ML is similar

176 to ensemble learning that uses a machine learning model at the high level (the second level) to learn how to best
177 combine the predictions from other machine learning models at the low level (the first level), as shown in Figure 1B.
178 Nevertheless, there are three key differences between proposed OOC-ML and classic ensemble learning. First, all
179 low-level models in the ensemble learning use the same training data, and the training data used in the high-level
180 has the same distribution as that used in the low-level. In the OOC-ML, the training data for each low-level model
181 has a different distribution. Specifically, they come from different Pfam families. The training data in the high-level
182 also uses Pfam families that are different from all others used in the low-level. Second, instead of using different
183 machine learning algorithms in the low-level ensemble model, the model architecture for all models in the OOC meta-
184 learning is the same as inspired by Model Agnostic Meta-Learning (MAML)[11]. The difference between models lies
185 in their different parameters (mapping functions) due to the different input data. Finally, ensemble learning uses
186 the predictions from the low-level models as meta data for the input of the high-level model. OOC meta-learning
187 uses gradients of mapping functions of the low-level models as meta data, which represent how the model learns, and
188 retrains the gradients by the high-level model.

189 **Stress model selection.** Finally, training should be stopped at a suitable point in order to avoid overfitting.
190 This was achieved by stress model selection. Stress model selection is designed to basically recapitulate an OOD
191 scenario by splitting the data into OOD train, OOD development, and OOD test sets as listed in Table 1; in this
192 procedure, the data distribution for the development set differs from that of the training data, and the distribution of
193 the test data set differs from both the training and development data. The section 4.2 provides further methodological
194 details, covering data pre-processing, the core algorithm, model configuration, and implementation details.

195 2.2 There are significantly unexplored dark gene families for small molecule 196 binding

197 We inspected the known CPIs between (i) molecules in the manually-curated ChEMBL database, which consists
198 of only a small portion of the chemical space, and (ii) proteins annotated in Pfam-A [25], which represents only
199 a narrow slice of the whole protein sequence space. The ChEMBL26[26] database supplies 1,950,765 chemicals
200 paired to 13,377 protein targets, constituting 15,996,368 known interaction pairs. Even for just this small portion
201 of chemical genomics space, unexplored gene families are enormous, can be seen in the dark region in Figure 2.
202 Approximately 90% of Pfam-A families do not have any known small-molecule binder. Even in Pfam families with
203 annotated CPIs (e.g., GPCRs), there exists a significant number of “orphan” receptors with unknown cognate ligands
204 (Figure 2). Fewer than 1% of chemicals bind to more than two proteins, and < 0.4% of chemicals bind to more than
205 five proteins, as shown in Supplemental Figure S1, S2 and S3. Because protein sequences in the dark gene families
206 could be significantly different from those for the known CPIs, predicting CPIs for dark proteins is an archetypal,
207 unaddressed OOD problem.

208 2.3 PortalCG significantly outperforms state-of-the-art approaches to predicting 209 CPIs of dark gene families

210 Two major categories of approaches have been developed for CPI predictions: machine learning and protein-ligand
211 docking (PLD). Recently published DISAE has been shown to outperform other leading deep learning methods for
212 predicting CPIs of orphan receptors and is explainable[1]. Because the neural network architecture of PortalCG is
213 similar to that of DISAE, we used DISAE as the baseline to evaluate the performance improvement of PortalCG over
214 the state-of-the-art. PortalCG demonstrates superior performance in terms of both Receiver Operating Characteristic
215 (ROC) and Precision-Recall (PR) curves when compared with DISAE, as shown in Figure 3(A). When the ratio
216 of positive and negative cases is imbalanced, the PR curve is more informative than the ROC curve. The PR-
217 AUC of PortalCG and DISAE is 0.714 and 0.603, respectively. In this regard, the performance gain of PortalCG
218 (18.4%) is significant ($p\text{-value} < 1e - 40$). Performance breakdowns for binding and non-binding classes can be found
219 in Supplemental Figure S4. PortalCG exhibits much higher recall and precision scores for positive cases (i.e., a
220 chemical-protein pair that is predicted to bind) versus negative, as shown in Supplemental Figure S4; this is a highly
221 encouraging result, given that there are many more negative (non-binding) than positive cases. The deployment gap,
222 shown in Figure 3(B), is steadily around zero for PortalCG; this promising finding means that we can expect that,
223 when applied to the dark proteins, the performance will be similar to that measured using the development data set.

224 With the advent of high-accuracy protein structural models, predicted by AlphaFold2 [5], it now becomes feasible
225 to use reversed protein-ligand docking (PLD) [27] to predict ligand-binding sites and poses on dark proteins, on a
226 genome-wide scale. In order to compare our method with the reversed protein-ligand docking approach, blind PLD to
227 proteins in the benchmark was performed via Autodock Vina[28] followed by protein-ligand binding affinity prediction
228 using a leading graph neural network-based method SIGN [29] (PLD+SIGN). The binding affinities predicted by
229 SIGN was more accurate than original scores from Autodock Vina (Supplemental Figure S5). The performance

of PLD+SIGN was compared with that of PortalGC and DISAE. As shown in Figure 3(A), both ROC and PR for PLD+SIGN are significantly worse than for PortalGC and DISAE. It is well known that PLD suffers from a high false-positive rate due to poor modeling of protein dynamics, solvation effects, crystallized waters, and other challenges [30]; often, small-molecule ligands will indiscriminately “stick” to concave, pocket-like patches on protein surfaces. For these reasons, although AlphaFold2 can accurately predict many protein structures, the relatively low reliability of PLD still poses a significant limitation, even with a limitless supply of predicted structures [31]. Thus, the direct application of PLD remains a challenge for predicting ligand binding to dark proteins. PortalCG’s end-to-end sequence-structure-function learning could be a more effective strategy in terms of both accuracy and efficacy: protein structure information is not used as a fixed input, but rather as an intermediate layer that can be tuned using various structural and functional information. Furthermore, the inference time of PortalCG for predicting a CPI is several orders of magnitude faster than that needed by the PLD. For example, it takes approximate 1 millisecond for PortalCG to predict a ligand binding to DRD2, while Autodock Vina needs around 10 seconds to dock a ligand to DRD2 excluding the time for defining the binding pocket.

2.4 Both the STL and OOC-ML stages contribute to the improved performance of PortalCG

To gauge the potential contribution of each component of PortalCG to the overall system effectiveness in predicting CPIs for dark proteins, we systematically compared the four models shown in Table 2. Details of the exact model configurations for these experiments can be found in the Supplemental Materials Table S1. As shown in Table 2, Variant 1, with a higher PR-AUC compared to the DISAE baseline, is the direct gain from transfer learning through 3D binding site information, all else being equal; yet, with transfer learning alone and without OOC-ML as an optimization algorithm in the protein function CPI prediction (i.e., Variant 2 versus Variant 1), the PR-AUC gain is minor. Variant 2 yields a 15% improvement while Variant 1 achieves only a 4% improvement over DISAE. PortalCG, in comparison, has the best PR-AUC score. With all other factors held constant, the advantage of PortalCG appears to be the synergistic effect of both STL and OOC-ML. The performance gain measured by PR-AUC under a shifted evaluation setting is significant (p -value < 1e-40), as shown in Supplemental Figure S6.

We find that stress model selection is able to mitigate potential overfitting problems, as expected. Training curves for the stress model selection are in Supplemental Figure S7. As shown in Supplemental Figure S7, the baseline DISAE approach tends to over-fit with training, and IID-dev performances are all higher than PortalCG but deteriorate in OOD-test performance. Hence, the deployment gap for the baseline is -0.275 and -0.345 on ROC-AUC and PR-AUC, respectively, while PortalCG deployment gap is around 0.01 and 0.005, respectively.

Table 2: Ablation study of PortalCG.

Models	Configuration	OOD-test set		Deployment gap	
		ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
PortalCG	PortalCG with all components	0.677±0.010	0.714±0.010	0.010±0.009	0.005±0.010
DISAE	PotalCG w/o STL or OOC-ML	0.636±0.004	0.603±0.005	-0.275±0.016	-0.345±0.012
variant 1	PotalCG w/o OOC-ML	0.661±0.004	0.629±0.005	/	/
variant 2	PotalCG w/o STL	0.654±0.062	0.698±0.015	/	/
PLD+SIGN	/	0.569	0.433	/	/

2.5 PortalCG is competitive on virtual compound screening for novel chemicals

Given that the pretraining, OOC-ML, and stress tests were only applied to proteins, current PortalCG was primarily designed to explore the dark protein space instead of new chemical space. Nevertheless, we evaluated if PortalCG could improve the performance for compound screening for novel chemicals. We employed a widely used DUD-E benchmark that included 8 protein targets along with their active compounds and decoys[32], and compared the performance of PortalCG with that of PLD. We used DUD-E chemicals as testing set. We trained PortalCG by excluding target proteins in the training/validation sets, and have all chemicals in the training/validation set dissimilar to those in the testing set (Tanimoto Coefficient (TC) less than 0.3 or 0.5). Under these chemical similarity thresholds, the false positive rate in the training/validation set was higher than 95.0% assumed that the ratio of actives vs inactives was 1:50 (Supplemental Figure S8).

As shown in Table 3, except targets kif11 and gcr, PortalCG could surprisingly outperform Autodock Vina on other remaining 6 targets in terms of enrichment factors (EFs). Similarly, PortalCG exhibited higher EFs than PLD-SIGN on six proteins. For the EF of 1%, the compound screening performance of PortalCG on 87.5% and

Table 3: Performances of compound screening evaluated using DUD-E benchmark. PortalCG-0.3: the similarities between chemicals in the training/validation set and those in the testing set are less than 0.3 of Tanimoto Coefficient (TC). PortalCG-0.5: the similarities between chemicals in the training/validation set and those in the testing set are less than 0.5 of TC. The best performance is in bold

	EF-1%				EF-20%			
	AutoDock Vina	PLD-SIGN	PotalCG-0.3	PotalCG-0.5	AutoDock Vina	PLD-SIGN	PotalCG-0.3	PotalCG-0.5
akt1	0.00	14.42	1.36	11.24	1.52	3.12	2.61	3.88
ampc	0.00	0.00	2.04	4.08	1.25	0.39	0.31	2.14
cp3a4	0.60	3.03	2.50	10.00	1.65	2.07	0.63	1.38
cxcr4	0.00	1.64	5.00	10.00	0.87	1.89	2.13	2.25
gcr	10.43	2.49	4.65	9.69	1.98	2.03	2.50	1.96
hivpr	4.10	5.02	0.75	13.62	2.31	2.34	1.87	2.84
hivrt	4.77	0.47	1.18	8.28	2.20	1.21	0.15	2.59
kif11	23.15	13.71	1.72	3.45	3.66	3.60	1.60	1.08

273 100.0% of targets is better than random guesses (EF=1.0) when the chemical similarity between the queries and the
 274 training data is 0.3 and 0.5, respectively. In contrast, only 50.0% 75.0% targets are better than the random guess
 275 for Autodock Vina and PLD+SIGN, respectively. It implies that PortalCG has learned certain patterns of CPIs,
 276 although the chemical OOD has not explicitly modeled. Different from PLD, whose EFs varied greatly across targets,
 277 the variance of EFs was relatively small for PortalCG across the targets, suggesting that the model was not biased
 278 to certain proteins. Thus, PortalCG is complementary with PLD, and has a potential to improve the capability of
 279 virtual compound screening, especially, for dark proteins whose reliable structures are not available.

280 2.6 PortalCG is capable of screening selective multi-targeted compounds with 281 novel scaffold for dark proteins

282 Opioid use disorder (OUD) is an overwhelming healthcare and economic burden. Although several pharmaceutical
 283 treatments for OUD exist, they are either restricted in usage or limited in effectiveness. Dopamine D1 and D3 receptors
 284 (DRD1 and DRD3) have been identified as potential drug targets for OUD. DRD1 partial agonists and antagonists
 285 alter the rewarding effects of drugs, while DRD3 antagonists reduce drug incentive and behavioral responses to drug
 286 cues. Moreover, recent evidence suggests that simultaneous targeting of DRD1 and DRD3 may be an effective OUD
 287 therapeutic strategy as the combination of a DRD1 partial agonist and a DRD3 antagonist reduced cue-induced
 288 relapse to heroin in rats[33]. By contrast, dopamine D2 receptor (DRD2) antagonism is associated with cataleptic
 289 side effects which limit the use of DRD2 antagonists as OUD therapeutics. Thus, selective DRD1 and DRD3 dual-
 290 antagonists could be an effective strategy for OUD treatment[34]. Because there are multiple dopamine receptors
 291 (especially, DRD2) that are similar to D1R and D3R, it is challenging to develop a selective dual-antagonist for DDR1
 292 and DRD3. PortalCG may provide new opportunities for OUD polypharmacology.

293 We synthesized 65 compounds based on the scaffold as shown in Figure 4A, which combines structural features of
 294 the DRD1 antagonist (-)-stepholidine with DRD3 antagonist pharmacophore, and determined their binding affinities
 295 to DRD1, DRD2, and DRD3, respectively (supplemental table S2). Tens of thousands of possible chemical structures
 296 could be derived from different combinations of R1, R2, R3, R4, and linker functional groups, as marked in Figure 4A.
 297 We have little knowledge on what is an optimal combination of functional groups for a dual-DRD1/DRD3 antagonist.
 298 If we define an acceptable dual-DRD1/DRD3 antagonist as a compound whose binding affinities are less than 100
 299 nM of Ki to both DRD1 and DRD3 but higher than 100 nM of Ki to DRD2, only 10 compounds satisfied this
 300 condition (successful rate of 15.4%) among the 65 synthesized compounds. It notes that a safe therapeutics for
 301 OUD may need much lower binding affinity than 100 nM of Ki for DRD2. For the 28 DRD1 antagonists with the
 302 Ki less than 100 nM, 15 of them had Ki higher than 100 nM for DRD3, corresponding to a two-class successful
 303 rate of 46.4%. It suggested that our current knowledge is limited for effectively designing selective dual-DRD1/3
 304 antagonists using existing scaffolds, let alone under a novel scaffold. The question is if we can use computational
 305 methods, especially PortalCG, to identify selective dual-DRD1/3 antagonists with a novel scaffold. We performed
 306 a rigorous bind test to validate the performance of PortalCG for this purpose. In the evaluation of PortalCG and
 307 DISAE, all of chemicals in the training data had different scaffolds from 65 test compounds, i.e., an OOD scenario
 308 on the chemical side [35]. Three models were trained with the sequence similarity between DRD1/2/3 and proteins
 309 in the training/validation data ranging from 20% to 60%. The performance was measured by the accuracy of a
 310 three-label classifier. When the sequence identifies between DRD1/2/3 and the proteins in the training/validation
 311 set were less than 40%, PortalCG achieved 20.0% and 50.7% successful rate for the cases where all DRDs and any
 312 two of them were predicted correctly, respectively (Figure 4B). The successful rate for the all DRDs was significantly
 313 higher the human design. Decreasing the sequence identifies between the proteins in the training/validation set and

314 DRD1/2/3 from 40% to 20% only slightly lower the accuracy of PortalCG, as shown in Figure 4C. The performance
315 drops were not statistically significant (p -value > 0.05). Increasing the sequence identities from 40% to 60% also did
316 not significantly change the accuracy. Thus, PortalCG by design was robust to OOD data.

317 We compared PortalCG with three baselines DISAE, PLD+SIGN, and Autodock Vina[28]. The crystal structures
318 of DRD1 (PDB id: 7JOZ), DRD2 (PDB id: 6CM4), and DRD3 (PDB id: 3PBL), which were co-crystallized with
319 ligands, were used for the docking. The 65 compounds were docked to the pre-defined binding pocket based on the
320 co-crystallized ligand. The order of accuracy follows PortalCG>DISAE>PLD-SIGN>Autodock Vina, as shown in
321 Figure 4B. This observation is consistent with our benchmark studies. Note that the complex structure was only
322 used for the baseline PLD models but this information was not used for PortalCG and DISAE.

323 2.7 Illuminating the undruggable human genome for drug repurposing

324 To further demonstrate the potential application of PortalCG, we explored potential drug lead compounds for
325 undrugged disease genes in the dark human genome, and prioritized undrugged genes that can be efficaciously
326 targeted by existing drugs. It is well known that only a small subset of the human genome is considered druggable
327 [36]. Many proteins are deemed “undruggable” because there is no information on their ligand-binding properties
328 or other interactions with small-molecule compounds (be they endogenous or exogenous ligands). Here, we built
329 an “undruggable” human disease protein database by removing the druggable proteins in Pharos [37] and Casas’s
330 druggable proteins [38] from human disease associated genes [17]. A total of 12,475 proteins were included in
331 our disease-associated undruggable human protein list. We applied PortalCG to predict the probability for these
332 “undruggable” proteins to bind to drug-like molecules. Around 6,000 drugs from the Drug Repurposing Hub[39] were
333 used in the screening. The proteins that could bind to a small molecule drug were ranked according to their prediction
334 scores, and 267 of them have a false positive rate lower than 2.18e-05, as listed in the Supplemental Table S3. Table 4
335 shows the statistically significantly enriched functions of these top ranked proteins as determined by DAVID [40]. The
336 most enriched proteins are involved in alternative splicing of mRNA transcripts. Malfunctions in alternative splicing
337 are linked to many diseases, including several cancers [41][42] and Alzheimer’s disease [43]. However, pharmaceutical
338 modulation of alternative splicing process is a challenging task. Identifying new drug targets and their lead compounds
339 for targeting alternative splicing pathways may open new doors to developing novel therapeutics for complex diseases
340 with few treatment options. In addition, several transcription factors and transcription activity related proteins were
341 identified and listed in Table S4, along with their predicted ligands.

342 Diseases associated with these 267 human proteins were also listed in Table 5. Since one protein is always related
343 to multiple diseases, these diseases are ranked by the number of their associated proteins. Most of top ranked diseases
344 are related with cancer development. 21 drugs that are approved or in clinical development are predicted to interact
345 with these proteins as shown in Supplemental Table S5. Several of these drugs are highly promiscuous. For example,
346 AI-10-49, a molecule that disrupts protein-protein interaction between CBFb-SMMHC and tumor suppressor RUNX1,
347 may bind to more than 60 other proteins. The off-target binding profile of these proteins may provide invaluable
348 information on potential side effects and opportunities for drug repurposing and polypharmacology. The drug-target
349 interaction network built for predicted positive proteins associated with Alzheimer’s disease was shown in Figure
350 5. The target proteins in this network were selected according to the threshold of 0.67. The length of the edges in this
351 network was decided by the prediction scores for these drug-target pairs. The longer the edge is, the lower confidence
352 of the prediction is. Thus if a higher threshold was applied, fewer drug-target pairs will appear in this network.
353 In order to validate the binding activity between the drugs and targets in this network, the PLD was performed
354 between the three most promiscuous drugs, AI-10-49, fenebrutinib, PF-05190457 and their predicted targets. Only
355 those targets with known PDB structures or reliable alpha-fold model structures were used in the docking. Docking
356 scores for the 21 drug-target pairs were listed on Supplemental Table S6. For each of the three drugs, the target with
357 the lowest docking score (the highest binding affinity) was selected as a representative. Docking conformations
358 and interactions between the drugs and their representative targets were shown in Figure 5. Functional enrichment,
359 disease associations, and top ranked drugs for the undruggable proteins with well-studied biology (classified as Tbio
360 in Pharos) and those excluding Tbio are list in Supplemental Table S7-S11.

361 3 Conclusion

362 This paper confronts the challenge of exploring dark proteins by recognizing it as an OOD generalization problem in
363 machine learning, and by developing a new deep learning framework to treat this type of problem. We propose
364 PortalCG as a general framework that enables systematic control of the OOD generalization risk. Systematic
365 examination of the PortalCG method revealed its superior performance compared to (i) a state-of-the-art deep learning
366 model (DISAE), and (ii) an AlphaFold2-enabled, GNN-scored, structure-based reverse docking approach. PortalCG
367 showed significant improvements in terms of both sensitivity and specificity, as well as close to zero deployment

David Functional Annotation enrichment analysis				
Enriched terms in UniProtKB keywords	Number of proteins involved	Percentage of proteins involved	P-value	Modified Benjamini p-value
Alternative splicing	171	66.5	7.70E-07	2.00E-04
Phosphoprotein	140	54.5	2.60E-06	3.40E-04
Cytoplasm	91	35.4	1.30E-05	1.10E-03
Nucleus	93	36.2	1.20E-04	8.10E-03
Metal-binding	68	26.5	4.20E-04	2.20E-02
Zinc	48	18.7	6.60E-04	2.90E-02

Table 4: Functional Annotation enrichment for undruggable human disease associated proteins selected by PortalCG

DiseaseName	# of undruggable proteins associated with disease
Breast Carcinoma	90
Tumor Cell Invasion	86
Carcinogenesis	83
Neoplasm Metastasis	75
Colorectal Carcinoma	73
Liver carcinoma	66
Malignant neoplasm of lung	56
Non-Small Cell Lung Carcinoma	56
Carcinoma of lung	54
Alzheimer's Disease	54

Table 5: Top ranked diseases associated with the undruggable human disease proteins selected by PortalCG

368 performance gap. The neural network architecture of PortalCG is similar to DISAE. Its performance improvement
369 over DISAE mainly comes from 3D binding site-enhanced pre-training and OOC-ML optimization. Both PortalCG
370 and DISAE outperform PLD-based methods by getting around the inherent limitations of PLD. Applications of
371 PortalCG to OUD polypharmacology and drug repurposing targeting of hitherto undruggable human proteins affords
372 novel new directions in drug discovery.

373 PortalCG can be further improved along several directions. In terms of protein sequence modeling, additional *a*
374 *prior* knowledge of protein structure and function can be incorporated into the pre-training or supervised multi-task
375 learning. The architecture of PortalCG mainly focuses on addressing the OOD problem of protein space but not
376 chemical space. New methods for modeling chemical structures alone or the joint space of chemicals and proteins
377 will no doubt improve CPI predictions for unseen novel chemicals. Future directions include but not limited to
378 novel representation of 3D chemical structures[44] at the sub-molecular level of scaffold and chemical moieties, pre-
379 training of the chemical space [45], and few-shot learning[46] as well as explicitly modeling amino acid-chemical
380 moiety interactions. The existing PortalCG treats the CPI prediction as a binary classification problem, but can
381 be reformulated as a regression model for predicting binding affinities. By defining domain-specific pre-training
382 and down-stream supervised learning tasks, PortalCG could be a general framework to explore the functions of
383 understudied proteins such as protein-protein interactions and protein-nucleic acid recognition.

384 4 Methods

385 PortalCG as a system level framework involves collaborative new design from data preprocessing, data splitting to
386 model initialization, and model optimization and evaluation. The pipeline of framework is illustrated in Figure 1.
387 Model architecture adopted by PortalCG mostly follows DISAE as shown in Figure 6.

388 4.1 Data sets

389 PortalCG was trained using four major databases, Pfam[25], Protein Data Bank (PDB)[47], BioLp[48] and ChEMBL[26].
390 The data were preprocessed as follows.

- 391 • Protein sequence data. All sequences from Pfam-A families are used to pretrain the protein descriptor following
392 the same setting in DISAE [1]. DISAE distills the original sequence into an ordered list of amino acid triplets
393 by extracting evolutionarily important positions from a multiple sequence alignment.
- 394 • Protein structure. In our protein structure data set, there are 30,593 protein structures, 13,104 ligands, and
395 91,780 ligand binding sites. Binding sites were selected according to the annotation from BioLip (updated to
396 the end of 2020). Binding sites which contact with DNA/RNA and metal ions were not included. If a protein
397 has more than one ligand, multiple binding pockets were defined for this protein. For each binding pocket, the
398 distances between $C\alpha$ atoms of amino acid residues of the binding pocket were calculated. In order to obtain
399 the distances between the ligand and its surrounding binding site residues, the distances between atom i in the
400 ligand and each atom in the residue j of the binding pocket were calculated and the smallest distance was selected
401 as the distance between atom i and residue j. In order to get the sequence feature of the binding site residues
402 in the DISAE protein sequence representation[1], binding site residues obtained from PDB structures (queries)
403 were mapped onto the multiple sequence alignments of its corresponding Pfam family. First, a profile HMM
404 database was built for the whole Pfam families. hmmscan [49] was applied to search the query sequence against
405 this profile database to decide which Pfam family it belongs to. For those proteins with multiple domains, more
406 than one Pfam families were identified. Then the query sequence was aligned to the most similar sequence in
407 the corresponding Pfam family by using phmmmer. Aligned residues on the query sequence were mapped to the
408 multiple sequence alignments of this family according to the alignment between the query sequence and the
409 most similar sequence.
- 410 • Chemical genomics data. CPI classification prediction data is the whole ChEMBL26[26] database where
411 the same threshold for defining positive and negative labels creating as that in DISAE [1] was used. Log-
412 transformation was performed for activities reported in pK_d, pK_i or pIC_50. The activities on a log-scale
413 were then binarized where protein-ligand pairs were considered active if pIC_50 > 5.3, pK_d > 7.3 or pK_i > 7.3.

414 Above data are split into training, validation, and testing sets. Data split statistics are shown in Table 6. Other
415 data statistics are demonstrated in Figure 2 and Supplemental Materials Figure S1, S2, S3.

416 65 compounds were synthesized for testing DRD1/2/3 binding activities. The procedures for the compound
417 synthesis were detailed in Supplemental material Section 1.8, Scheme 1-5. DRD3 binding assays and Ki determinations
418 were performed by the Psychoactive Drug Screening Program (PDSP).

419 For illuminating undruggable human proteins, around 6,000 drugs are collected from CLUE[39]. 12,475 undruggable
420 proteins are collected by removing the druggable proteins in Pharos [37] and Casas's druggable proteins [38] from
421 human disease associated genes [17].

422 4.2 Algorithm

423 4.2.1 Chemical representation

424 A chemical was represented as a graph and its embedding was learned using Graph Isomorphism Network (GIN)[50]
425 which was designed to maximize the representational (or discriminative) power of a Graph Neural Network (GNN)
426 based on Weisfeiler-Lehman (WL) graph isomorphism test. GIN is a common choice as chemical descriptor[35].

427 4.2.2 Protein sequence pre-training

428 Protein descriptor is pretrained from scratch following exactly DISAE [1] on whole Pfam families, making it a universal
429 protein language model. DISAE was inspired by recent success in self-supervised learning of unlabeled data in Nature
430 Language Processing (NLP). It features a novel method, DIstilled Sequence Alignment Embedding (DISAE), for the
431 protein sequence representation. DISAE can utilize all protein sequences to capture functional information without
432 the knowledge of their structure and function. By incorporating biological knowledge into the sequence representation,
433 DISAE can learn functionally important information about protein families that span a wide range of protein space.
434 Different from existing sequence pre-training strategy that uses original protein sequences as input [22], DISAE
435 distilled the original sequence into an ordered list of triplets by extracting evolutionary important positions from a
436 multiple sequence alignment including insertions and deletions. Then long-range residue-residue interactions can be
437 learned via the Transformer module in ALBERT[10]. A self-supervised masked language modeling (MLM) approach
438 was used at this stage. In the MLM, 15% triplets are randomly masked and assumed that they are unknown. Then
439 the remaining triplets are used to predict what the masked triplets are.

440 **4.2.3 Protein structure regularization**

With the protein descriptor pretrained using the sequences from the whole Pfam, chemical descriptors and a distance learner were plugged in to fine-tune the protein representation. The distance learner follows AlphaFold[4] which formulates a multi-way classification on a distrogram. Based on the histogram of distances between amino acids and ligand atoms, a histogram equalization¹ was applied to formulate a 10-way classification on our binding site structure data as in Supplemental material Figure S9. Since protein and chemical descriptors output position-specific embeddings of a distilled protein sequence and all atoms of a chemical, pair-wise interaction features on the binding sites were created with a simple vector operation: a matrix multiplication was used to select embedding vectors of each binding residue and atom; multiply and broadcast the selected embedding vectors into a symmetric tensor as shown in the following, where H is embedding matrix of size (*number_of_residues, embeddingdimension*) or (*number_of_atoms, embeddingdimension*) and A is selector matrix[51],

$$\begin{aligned} H_{\text{binding_site}}^{\text{protein}} &= A^{\text{protein}} * H_{\text{full_distilled}}^{\text{protein}} \\ H_{\text{binding_site}}^{\text{chemical}} &= A^{\text{chemical}} * H_{\text{full_chemical_graph}}^{\text{chemical}} \\ H_{\text{binding_site}}^{\text{interaction}} &= (H_{\text{binding_site}}^{\text{protein}})^T * H_{\text{binding_site}}^{\text{chemical}} \end{aligned}$$

441 This pair-wise interaction feature tensor $H_{\text{binding_site}}^{\text{interaction}}$ was fed into a Attentive Pooling[52] layer followed by feed-
442 forward layer for final 10-way classification. Detailed model architecture configuration could be found in Supplemental
443 Table S1 and Figure6. The intuition for the simplest form of distance learner is to put all stress of learning on
444 the shared protein and chemical descriptors which will carry information across the end-to-end neural network.
445 Again, with standard Adam optimization, shifted evaluation was used to select the “best” instance. Two versions of
446 distance structure prediction were implemented, one formulated as a binary classification, i.e. contact prediction, one
447 formulated as a multi-way classification, i.e. distogram prediction. The performance of the two version are similar,
448 as shown in Figure S9.

449 **4.2.4 Out-of-cluster Meta Learning (OOC-ML)**

450 With fine-tuned protein descriptor in the protein function space, a binary classifier is plugged on, which is a ResNet[53]
451 layered with two linear layers as shown in Supplemental Table S1 and Figure6. What plays the major role in this
452 phase is the optimization algorithm OOC-ML as shown in pseudocode **Algorithm 1** and Figure 1. The first level
453 (low leverl) model training is reflected in line 4-9, and line 10 shows ensemble training of the second level (high level)
454 models. Note that more variants could be derived from changing sampling rule (line 3 and 5) and the second level
455 ensemble rule.

456 **4.2.5 Stress model instance selection**

457 In classic training scheme common practice, there are 3-split data sets, “train set”, “dev set” and “test set”. Train
458 set as the name suggested is used to train model. Test set as commonly expected is used to set an expectation of
459 performance when applying the trained model to unseen data. Dev set is to select the preferred model instance. In
460 OOD setting, data is split (see Table 1) such that dev set is a OOD from train set and test set is a OOD from both
461 train and dev set. Deployment gap is calculated by deducting ODD-dev performance with OOD-test performance.

462 **4.3 Baseline models**

463 Machine learning methods for CPI predictions have been widely explored by many paradigms. As summarized
464 in the survey [54], in addition to deep learning methods, there are similarity/distance based methods, matrix
465 factorization, network-based, and feature-based methods. For CPI predictions with OOD generalization challenge, the
466 similarity/distance-based, matrix factorization, and network-based methods have major obstacles. The Similarity/distance
467 based methods rely on drug-drug similarity matrix and target-target similarity matrix as input. Because the
468 similarities between dark proteins and proteins with known ligands are low, no reliable predictions can be made.
469 Matrix Factorization is popular for its high efficiency but the cold-start nature of the dark proteins doesn't fit
470 matrix factorization paradigm. Network-based methods usually utilize protein-protein interactions. Such methods
471 have advantages to predict the functional associations of ligand binding, but not the direct physical interactions.
472 Furthermore, these methods are not scalable to millions of proteins and millions of chemicals. Almost of all studies
473 based on these methods only focus on thousands of targets and thousands of drugs. PortalCG belongs to the category
474 of the feature-based method. In our recently published work[1], we have shown that DISAE outperforms other state-
475 of-the-art feature-based methods. Thus, we only compare PortalCG with DISAE in this paper.

¹Histogram equalization: https://en.wikipedia.org/wiki/Histogram_equalization

Algorithm 1: PortaCG, Out-of-cluster Meta-learning

```
input :  $p(\mathbf{D})$ , CPI data distribution over whole Pfams, where each  $D_i \in \mathbf{D}$  is a set of CPI pairs  
for the  $p_{fam_i}$ ;  
 $\alpha, \beta$ , learning step size hyperparameters;  
 $L$ , number of optimization steps in each round of the first level training;  
 $T$ , number of the second level training steps;  
 $K$ , number of points sampled from a local neighborhood  
output:  $\theta$  the whole model weights  
1 initialization whole model weights  $\theta$  (with weight transferred from portal for protein and chemical  
descriptors and random initialized weights for binary classifier)  
2 for  $t$  in  $T$  do  
3   Sample a  $D_i \sim p(D)$ ;  
4   for  $l$  in  $L$  do  
5     Sample a positive-negative balanced mini-batch of  $K$  pairs  $neighborhood_m \sim D_i$ ;  
6     for  $point_j$  in  $neighborhood_m$  do  
7       Evaluate  $\nabla_{\theta} L_{point_j}(f_{\theta})$  with respect to  $K$  examples;  
8       Compute adapted parameters with gradient descent:  $\theta'_i = \theta - \alpha \nabla_{\theta} L_{point_j}(f_{\theta})$ ;  
9     end  
10    Update  $\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{D_i \sim p(D)} L_{point_j}(f'_{\theta})$ ;  
11  end  
12 end
```

476 Besides the machine learning method, protein-ligand docking (PLD) is a widely used approach to predict CPIs.
477 We evaluate the performance of PLD based on Autodock Vina[28] and AlphaFold2 predicted structures[5] followed
478 by SIGN re-scoring[29]. Structure-aware Interactive Graph Neural Networks (SIGN) [29] is a graph neural network
479 proposed for the prediction of protein-ligand binding affinity. SIGN builds directional graphs to model the structures
480 and interactions in protein-ligand complexes. Both distances and angles are integrated in the aggregation processes.
481 SIGN is trained on PDBbind [55], which is a well-known public dataset containing 3D structures of protein-ligand
482 complexes together with experimentally determined binding affinities. Similar to SIGN [29], we used the PDBbind
483 v2016 dataset and the corresponding refined set, which contains 3767 complexes, to perform the training. We followed
484 SIGN [29] for training and testing. For the directional graph used in SIGN, we constructed them with cutoff-threshold
485 $\theta_d = 5\text{\AA}$. The number of hidden layers is set to 2. All of the other settings are kept the same as those used in the
486 original paper of SIGN. We randomly split the PDBbind refined set with a ratio of 9:1 for training and validation.

dataset	usage in PortalCG	count	sample size	note
Pfam 33.1	STL, the first pretraining step to train DISAE	# Pfam families # sequences	17,772 54,409,760	random split in training and testing
PDB	STL, the second pretraining step to learn contact map between amino acid residues and ligand atoms at binding sites	train # Pfam families # proteins # binding sites (protein-ligand pairs) # chemical	319 5,926 6,896 3,168	
		test # Pfam families # proteins # binding sites (protein-ligand pairs) # chemical	733 1,497 1,573 670	Pfam families in OOD-dev and OOD-test are held out from PDB pre-training.
ChEMBL 26	OOC-ML	OOD-train # protein-ligand pairs # chemical # Pfam families	1,672,277 478,939 333	within each split (OOD-train/IID-dev/OOD-dev/OOD-test), the data is random split into support and query sets in a ratio of 5:1 for each Pfam family unless there are only one class (binding or not) of data
		IID-dev # protein-ligand pairs # chemical # Pfam families	6,536 6,096 333	
		OOD-dev # protein-ligand pairs # chemical # Pfam families	165,655 98,975 701	
		OOD-test # protein-ligand pairs # chemical # Pfam families	0 162,354 104,299	
			700	
			0	
			0	
			0	

Table 6: Data statistics for each training stage

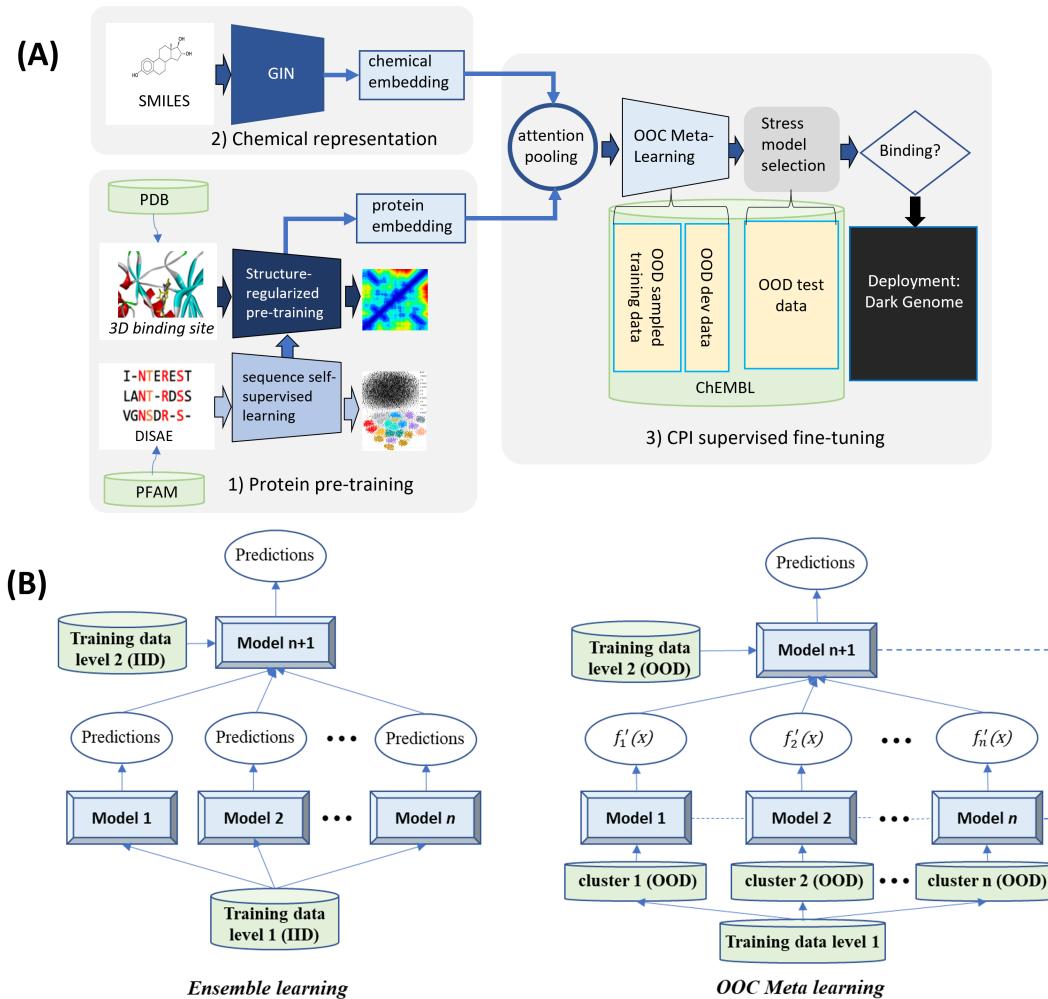


Figure 1: **(A) Scheme of PortalCG:** PortalCG enables chemical protein interactions (CPIs) prediction for dark proteins across gene families with four key components: ligand binding site enhanced sequence pretraining, end-to-end transfer learning following sequence-structure-function paradigm, out-of-cluster meta-learning (OOC-ML), and stress model selection. **(B) Illustration of OOC-ML with the comparison with classic ensemble learning:** OOC-ML follows the same spirit as the ensemble learning, but different in data split, model architecture, and optimization schema.

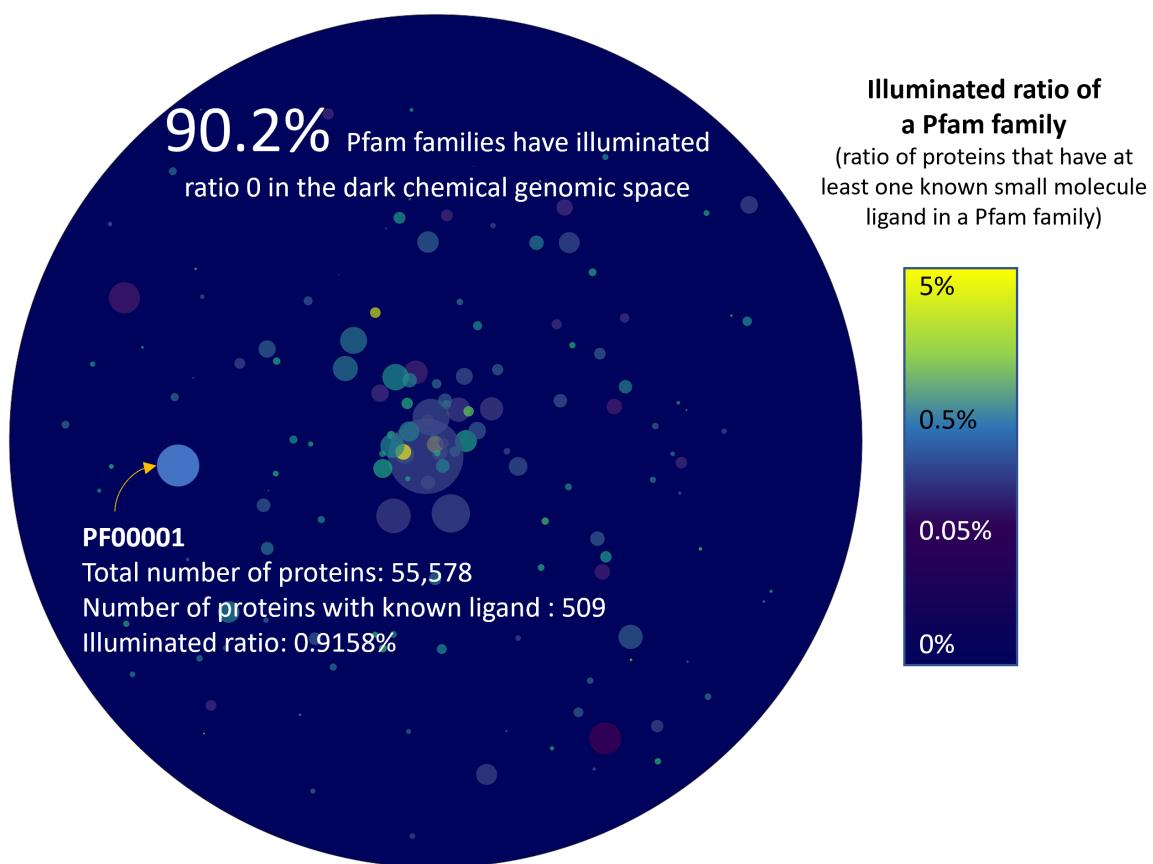


Figure 2: **Chemical genomics space in statistics: The ratio of proteins that have at least a known ligand in each Pfam family.** Each color bubble represents a Pfam family. The size of a bubble is proportional to the total number of proteins in the Pfam family. 1,734 Pfam families have at least one known small molecule ligand. Most of these Pfam families have less than 1% proteins with known ligands. Furthermore, around 90.2% of total 17,772 Pfam families remain dark without any known ligand information.

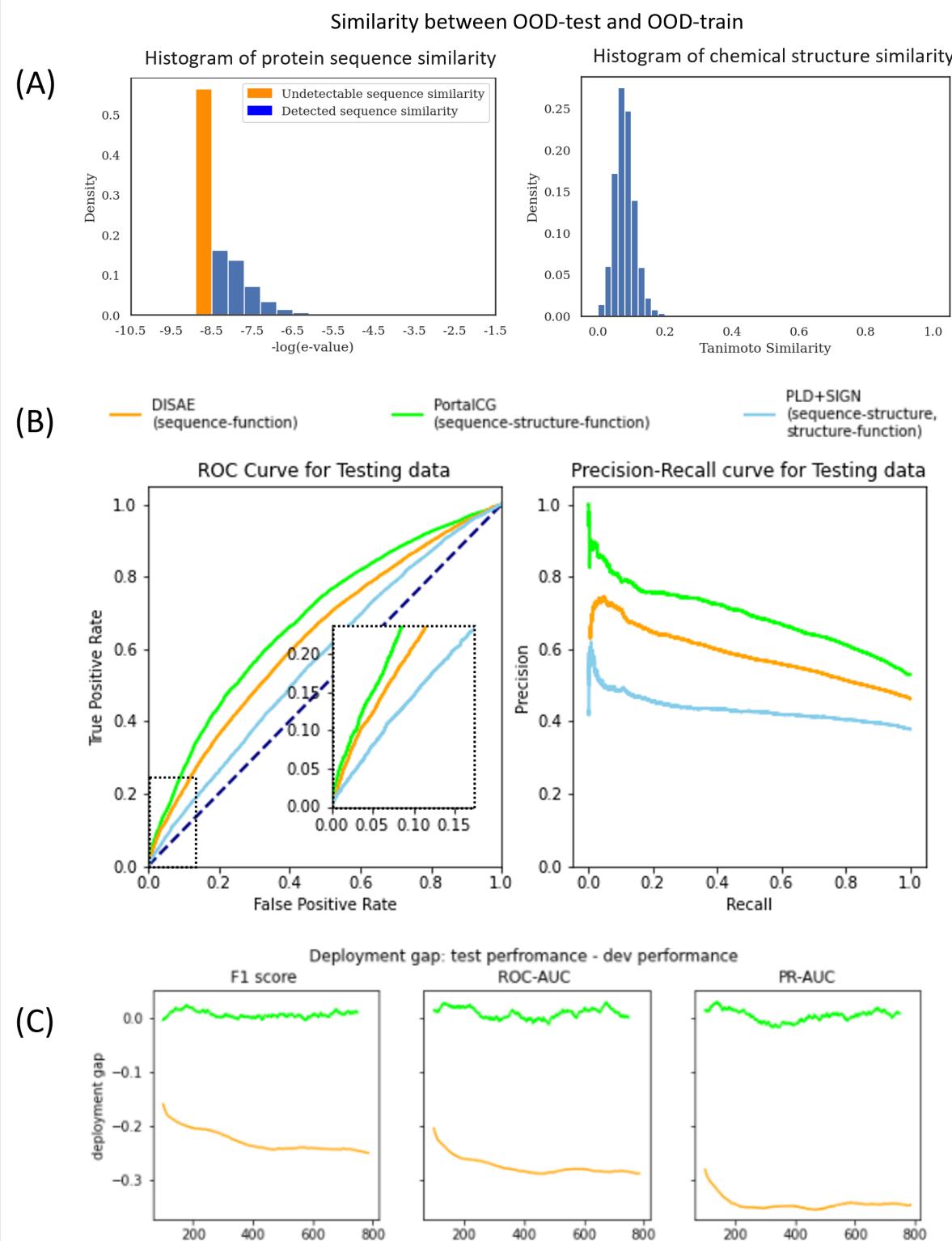


Figure 3: Comparison of PortalCG with the state-of-the-art methods DISAE and PLD+SIGN as baselines using the OOD test where the proteins in the testing data come from different Pfam families from the proteins in the training and validation data. (A) Histograms of protein sequence and chemical structure similarities between OOD-train and OOD-test. The majority of protein sequences in the training set does not have detectable similarity to the proteins in the testing set. (B) Receiver Operating Characteristic (ROC) and Precision-Recall curves for the “best” model instance selected by the stress test. Due to the imbalanced active/inactive data, the Precision-Recall (PR) curve is a more reliable measure than the ROC curve. (C) Deployment gaps of PortalCG and DISAE.¹⁶ The deployment gap of PortalCG is steadily around zero as training step increases while the deployment performance of DISAE deteriorates.

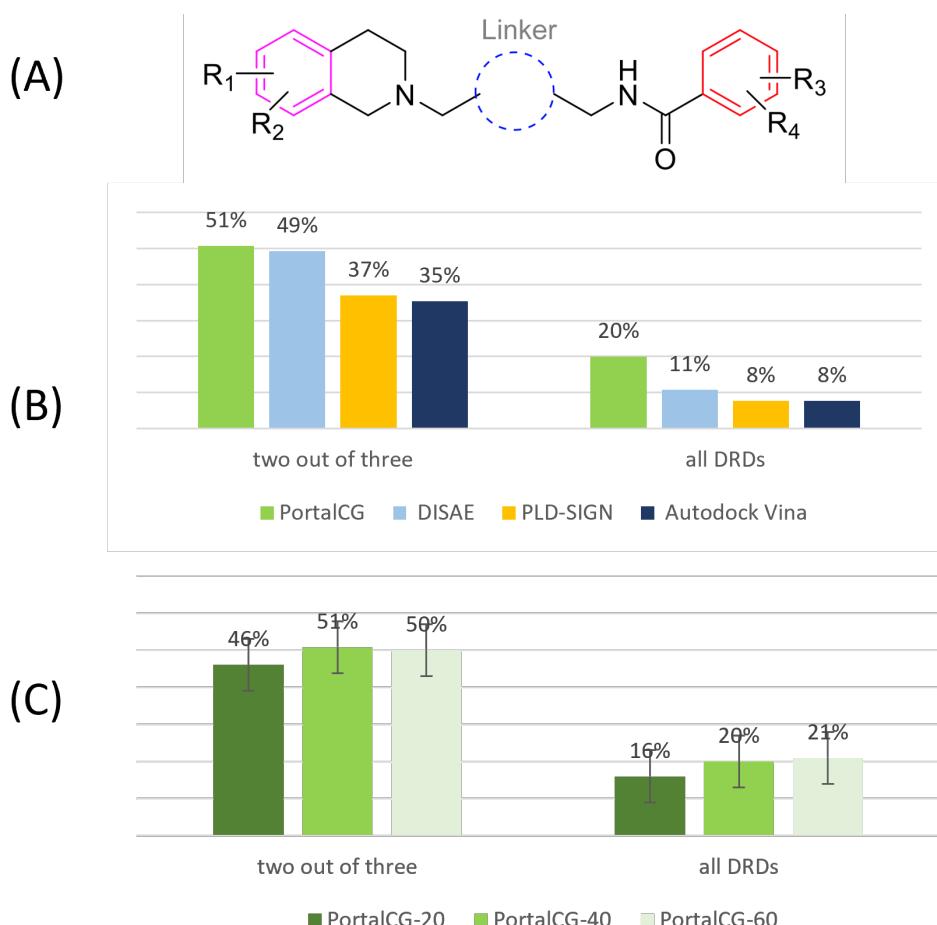


Figure 4: (A) The chemical scaffold on which 65 compounds were synthesized for potential selective dual-DRD1/DRD2 antagonists. Tens of thousands of chemicals can be generated from the different combination of four functional groups R1, R2, R3, and R4 and a linker group. (B) The prediction accuracy of DRD binding profile classification. (C) The performance of PortalCG when the sequence similarities between the proteins in the training/validation set and DRD1/DRD2/DRD3 were less than 20%, 40%, and 60%, respectively. The performance was measured by the accuracy of a three-label classifier. "Two out of three" and "all DRDs" represented the accuracy when two labels and all three labels were predicted correctly.

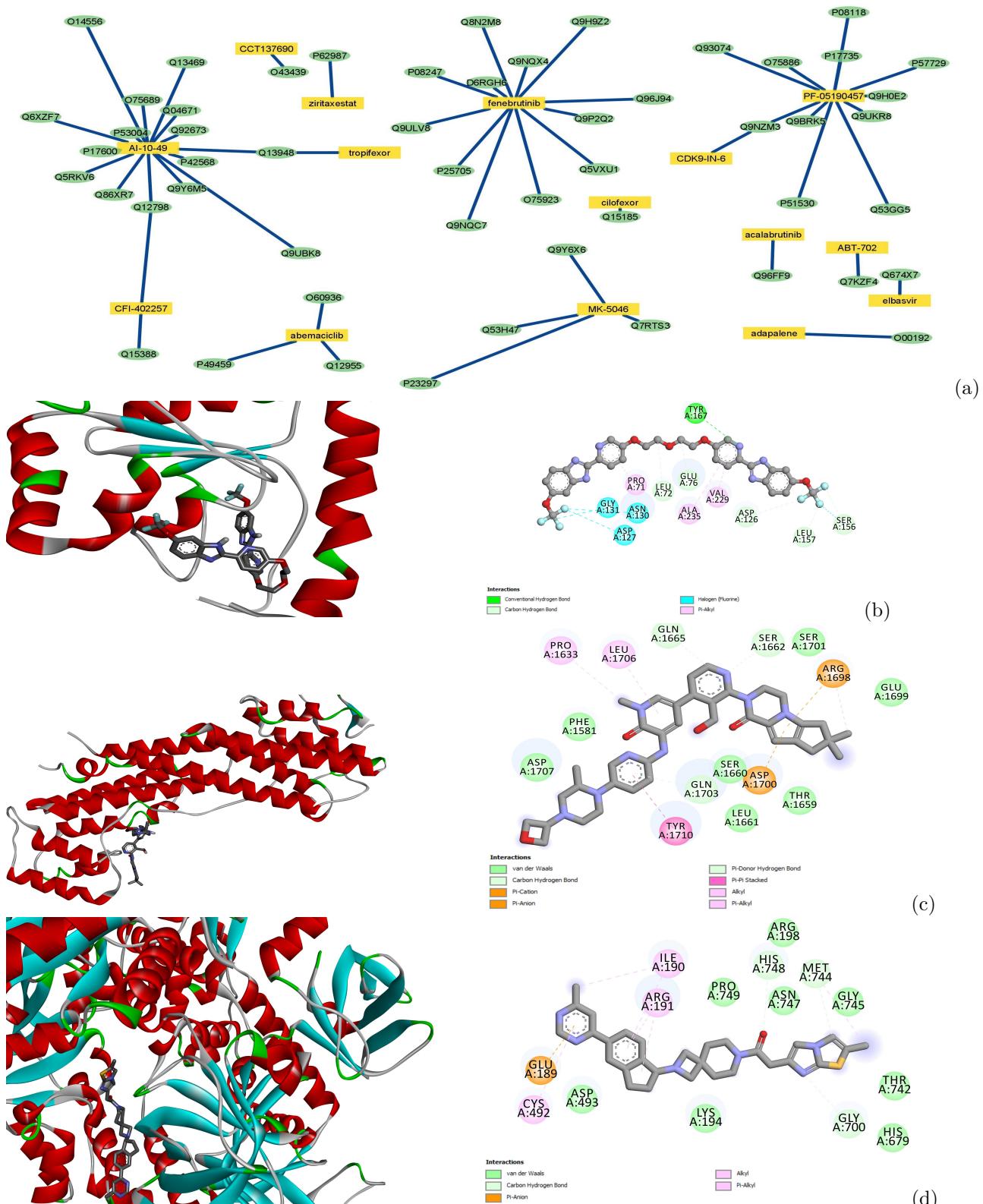


Figure 5: Drug-target interaction network for proteins associated with Alzheimer's disease and docking poses for representative drug-target pairs calculated by Autodock Vina. (a) Drug-target interaction network predicted by PortalCG. Yellow rectangles and green ovals represent drugs and targets, respectively. (b) Docking pose and ligand binding interactions between protein TIR domain-containing adapter molecule 2 (Uniprot: Q86XR7) and AI-10-49. (c) Docking pose and ligand binding interactions between protein Unconventional myosin-Vc (Uniprot: Q9NQX4) and fenebrutinib. (d) Docking pose and ligand binding interactions between DNA replication ATP-dependent helicase/nuclease (Uniprot: P51530) and PF-05190457.

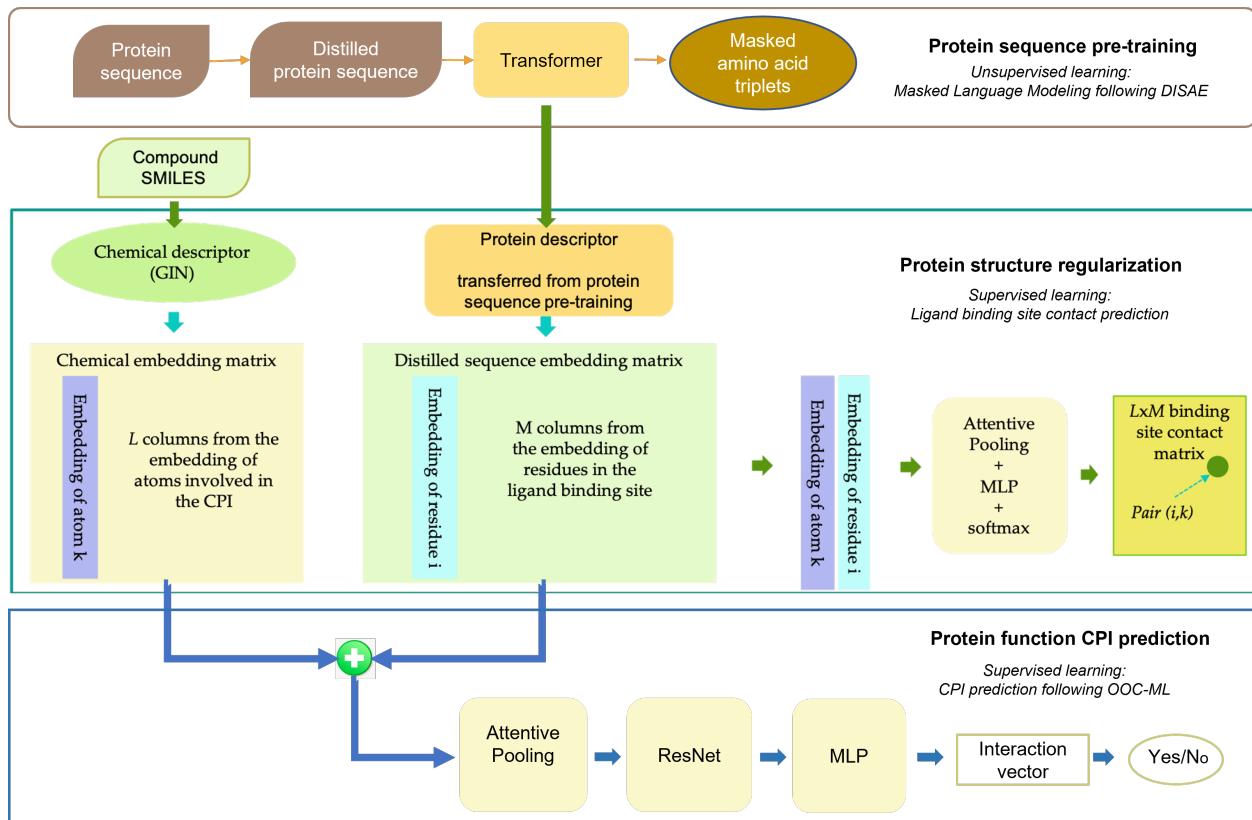


Figure 6: Illustration of PortalCG architecture for three stages of training. The architecture of protein sequence pre-training used a transformer architecture and masked language modeling as detailed in ref[1]. Pretrained protein descriptor was then used in binding site enhanced sequence pre-training. In this stage, the task was to predict amino acid residue and ligand atom distance matrices. Finally, protein descriptors that were pretrained and regularized in the previous two stages were concatenated with chemical descriptors via an attention network to predict CPIs. Chemical structures were represented by GIN[50], a graph neural network model. The second and third stages had same model architecture but the model parameters were transferred from the second to the third stages. OOC-ML as an optimization algorithm was not a model architecture component, and only used in the CPI prediction.

487 Author Contributions

488 TC conceived the concept of PortalCG, implemented the algorithms, performed the experiments, and wrote the
489 manuscript; Li Xie and SZ prepared data, performed the experiments, and wrote the manuscript; MC, DH, and
490 YL implemented algorithms; KN, MD, and WWH designed and synthesized compounds; CM and PEB refined the
491 concepts and wrote the manuscript; Lei Xie conceived and planned the experiments, wrote the manuscript.

492 Data and software availability

493 Data used are described in section 4.3 and can be downloaded from public resource. Trained PortalCG model and
494 PortalCG codes can be found in the Code Ocean.
495 <https://github.com/XieResearchGroup/PortalLearning>.

496 Acknowledgement

497 This project has been funded with federal funds from the National Institute of General Medical Sciences of National
498 Institute of Health (R01GM122845) and the National Institute on Aging of the National Institute of Health (R01AD05755).
499 We appreciate that Hansaim Lim helped with proof reading and provided constructive suggestions. Ki determinations,
500 and receptor binding and activity profiles were generously provided by the National Institute of Mental Health's
501 Psychoactive Drug Screening Program, Contract #HHSN-271-2008-00025-C (NIMH PDSP). The NIMH PDSP is
502 directed by Bryan L. Roth MD, PhD at the University of North Carolina at Chapel Hill and Project Officer Jamie
503 Driscoll at NIMH, Bethesda MD, USA.

504 References

- [1] T. Cai, H. Lim, K. A. Abbu, Y. Qiu, R. Nussinov, and L. Xie, "Msa-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: Application to gpcrome deorphanization," *Journal of Chemical Information and Modeling*, vol. 61, no. 4, pp. 1570–1582, 2021.
- [2] J. Ma, S. H. Fong, Y. Luo, C. J. Bakkenist, J. P. Shen, S. Mourragui, L. F. Wessels, M. Hafner, R. Sharan, J. Peng, *et al.*, "Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients," *Nature Cancer*, vol. 2, no. 2, pp. 233–244, 2021.
- [3] D. He, Q. Liu, Y. Wu, and L. Xie, "A context-aware deconfounding autoencoder for robust prediction of personalized clinical drug response from cell-line compound screening," *Nature Machine Intelligence*, pp. 1–14, 2022.
- [4] N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, and D. Baker, "Improved protein structure refinement guided by deep learning based accuracy estimation," *Nature communications*, vol. 12, no. 1, pp. 1–11, 2021.
- [5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, pp. 1–11, 2021.
- [6] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a 3-track network," *bioRxiv*, 2021.
- [7] Y. Li, P. Luo, Y. Lu, and F.-X. Wu, "Identifying cell types from single-cell data based on similarities and dissimilarities between cells," *BMC bioinformatics*, vol. 22, no. 3, pp. 1–18, 2021.
- [8] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
- [9] W. Chen, Z. Yu, Z. Wang, and A. Anandkumar, "Automated synthetic-to-real generalization," in *International Conference on Machine Learning*, pp. 1746–1756, PMLR, 2020.
- [10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.
- [11] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," *CoRR*, vol. abs/1703.03400, 2017.
- [12] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *CoRR*, vol. abs/2004.05439, 2020.
- [13] T. I. Oprea, "Exploring the dark genome: implications for precision medicine," *Mammalian Genome*, vol. 30, no. 7, pp. 192–200, 2019.
- [14] G. Kustatscher, T. Collins, A.-C. Gingras, T. Guo, H. Hermjakob, T. Ideker, K. S. Lilley, E. Lundberg, E. M. Marcotte, M. Ralser, *et al.*, "Understudied proteins: opportunities and challenges for functional proteomics," *Nature Methods*, pp. 1–6, 2022.
- [15] G. Kustatscher, T. Collins, A.-C. Gingras, T. Guo, H. Hermjakob, T. Ideker, K. S. Lilley, E. Lundberg, E. M. Marcotte, M. Ralser, *et al.*, "An open invitation to the understudied proteins initiative," *Nature Biotechnology*, pp. 1–3, 2022.
- [16] L. Xie, L. Xie, S. L. Kinnings, and P. E. Bourne, "Novel computational approaches to polypharmacology as a means to define responses to individual drugs," *Annual review of pharmacology and toxicology*, vol. 52, pp. 361–379, 2012.
- [17] J. Piñero, J. M. Ramírez-Anguita, J. Saüch-Pitarch, F. Ronzano, E. Centeno, F. Sanz, and L. I. Furlong, "The disgenet knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, vol. 48, p. D845–D855, 1 2020.
- [18] M. Karimi, D. Wu, Z. Wang, and Y. Shen, "Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks," *Bioinformatics*, vol. 35, no. 18, pp. 3329–3338, 2019.
- [19] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, 2018.
- [20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

- 556 [21] S. Sledzieski, R. Singh, L. Cowen, and B. Berger, “Sequence-based prediction of protein-protein interactions: a
557 structure-aware interpretable deep learning model,” *bioRxiv*, 2021.
- 558 [22] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, “Biological
559 structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” *Proceedings
560 of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021.
- 561 [23] L. Xie and P. E. Bourne, “Detecting evolutionary relationships across existing fold space, using sequence order-
562 independent profile-profile alignments,” *Proceedings of the National Academy of sciences*, vol. 105, no. 14,
563 pp. 5441–5446, 2008.
- 564 [24] M. AlQuraishi and P. K. Sorger, “Differentiable biology: using deep learning for biophysics-based and data-driven
565 modeling of molecular mechanisms,” *Nature methods*, vol. 18, no. 10, pp. 1169–1180, 2021.
- 566 [25] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. Sonnhammer, S. C. Tosatto, L. Paladin,
567 S. Raj, L. J. Richardson, *et al.*, “Pfam: The protein families database in 2021,” *Nucleic Acids Research*, vol. 49,
568 no. D1, pp. D412–D419, 2021.
- 569 [26] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J.
570 Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos,
571 I. Smit, and A. R. Leach, “The ChEMBL database in 2017,” *Nucleic Acids Research*, vol. 45, pp. D945–D954,
572 11 2016.
- 573 [27] H. Huang, G. Zhang, Y. Zhou, C. Lin, S. Chen, Y. Lin, S. Mai, and Z. Huang, “Reverse screening methods to
574 search for the protein targets of chemopreventive compounds,” *Frontiers in chemistry*, vol. 6, p. 138, 2018.
- 575 [28] O. Trott and A. J. Olson, “Autodock vina: improving the speed and accuracy of docking with a new scoring
576 function, efficient optimization and multithreading,” *Journal of Computational Chemistry*, vol. 31, pp. 455–461,
577 2010.
- 578 [29] S. Li, J. Zhou, T. Xu, L. Huang, F. Wang, H. Xiong, W. Huang, D. Dou, and H. Xiong, “Structure-aware
579 interactive graph neural networks for the prediction of protein-ligand binding affinity,” in *Proceedings of the 27th
580 ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 975–985, 2021.
- 581 [30] S. Z. Grinter and X. Zou, “Challenges, applications, and recent advances of protein-ligand docking in structure-
582 based drug design,” *Molecules*, vol. 19, no. 7, pp. 10150–10176, 2014.
- 583 [31] M. Jaiteh, I. Rodríguez-Espigares, J. Selent, and J. Carlsson, “Performance of virtual screening against gpcr
584 homology models: Impact of template selection and treatment of binding site plasticity,” *PLoS computational
585 biology*, vol. 16, no. 3, p. e1007680, 2020.
- 586 [32] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, “Directory of useful decoys, enhanced (dud-e):
587 better ligands and decoys for better benchmarking,” *Journal of medicinal chemistry*, vol. 55, no. 14, pp. 6582–
588 6594, 2012.
- 589 [33] S. T. Ewing, C. Dorcely, R. Maudi, G. Paker, E. Schelbaum, and R. Ranaldi, “Low-dose polypharmacology
590 targeting dopamine d1 and d3 receptors reduces cue-induced relapse to heroin seeking in rats,” *Addiction Biology*,
591 vol. 26, no. 4, p. e12988, 2021.
- 592 [34] E. Galaj, S. Ewing, and R. Ranaldi, “Dopamine d1 and d3 receptor polypharmacology as a potential treatment
593 approach for substance use disorder,” *Neuroscience & Biobehavioral Reviews*, vol. 89, pp. 13–28, 2018.
- 594 [35] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, “Strategies for pre-training graph
595 neural networks,” 2020.
- 596 [36] C. Finan, A. Gaulton, F. A. Kruger, R. T. Lumbers, T. Shah, J. Engmann, L. Galver, R. Kelley, A. Karlsson,
597 R. Santos, *et al.*, “The druggable genome and support for target identification and validation in drug
598 development,” *Science translational medicine*, vol. 9, no. 383, 2017.
- 599 [37] T. K. Sheils, S. L. Mathias, K. J. Kelleher, V. B. Siramshetty, D.-T. Nguyen, C. G. Bologa, L. J. Jensen,
600 D. Vidović, A. Koleti, S. C. Schürer, A. Waller, J. J. Yang, J. Holmes, G. Bocci, N. Southall, P. Dharkar,
601 E. Mathé, A. Simeonov, and T. I. Oprea, “Uterd and pharos 2021: mining the human proteome for disease
602 biology,” *Nucleic Acids Research*, vol. 49, pp. D1334–D1346, 1 2021.
- 603 [38] C. Finan, A. Gaulton, F. A. Kruger, R. T. Lumbers, T. Shah, J. Engmann, L. Galver, R. Kelley, A. Karlsson,
604 R. Santos, J. P. Overington, A. D. Hingorani, and J. P. Casas, “The druggable genome and support for target
605 identification and validation in drug development,” *Science Translational Medicine*, vol. 9, p. eaag1166, 3 2017.
- 606 [39] S. M. Corsello, J. A. Bittker, Z. Liu, J. Gould, P. McCarren, J. E. Hirschman, S. E. Johnston, A. Vrcic, B. Wong,
607 M. Khan, *et al.*, “The drug repurposing hub: a next-generation drug library and information resource,” *Nature
608 medicine*, vol. 23, no. 4, pp. 405–408, 2017.

- 609 [40] X. Jiao, B. T. Sherman, D. W. Huang, M. W. B. Robert Stephens, H. C. Lane, and R. A. Lempicki, “David-ws:
610 a stateful web service to facilitate gene/protein list analysis,” *Bioinformatics*, vol. 28, p. 1805–1806, 7 2012.
- 611 [41] D. O. Bates, J. C. Morris, S. Oltean, and L. F. Donaldson, “Pharmacology of modulators of alternative splicing,”
612 *Pharmacological reviews*, vol. 69, no. 1, pp. 63–79, 2017.
- 613 [42] K.-q. Le, B. S. Prabhakar, W.-j. Hong, and L.-c. Li, “Alternative splicing as a biomarker and potential target
614 for drug discovery,” *Acta Pharmacologica Sinica*, vol. 36, no. 10, pp. 1212–1218, 2015.
- 615 [43] J. E. Love, E. J. Hayden, and T. T. Rohn, “Alternative splicing in alzheimer’s disease,” *Journal of Parkinson’s
616 disease and Alzheimer’s disease*, vol. 2, no. 2, 2015.
- 617 [44] S. Zhang, Y. Liu, and L. Xie, “Efficient and accurate physics-aware multiplex graph neural networks for 3d small
618 molecules and macromolecule complexes,” *arXiv preprint arXiv:2206.02789*, 2022.
- 619 [45] Y. Liu, H. Lim, and L. Xie, “Exploration of chemical space with partial labeled noisy student self-training and
620 self-supervised graph embedding,” *BMC bioinformatics*, vol. 23, no. 3, pp. 1–21, 2022.
- 621 [46] Y. Liu, Y. Wu, X. Shen, and L. Xie, “Covid-19 multi-targeted drug repurposing using few-shot learning,”
622 *Frontiers in Bioinformatics*, vol. 1, 2021.
- 623 [47] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne,
624 “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, pp. 235–242, 01 2000.
- 625 [48] J. Yang, A. Roy, and Y. Zhang, “Biolip: a semi-manually curated database for biologically relevant ligand–protein
626 interactions,” *Nucleic acids research*, vol. 41, no. D1, pp. D1096–D1103, 2012.
- 627 [49] S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, and R. D. Finn, “Hmmer web server: 2018 update,”
628 *Nucleic acids research*, vol. 46, no. W1, pp. W200–W204, 2018.
- 629 [50] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?,” *arXiv preprint arXiv:1810.00826*, 2018.
- 630 [51] S. Boyd and L. Vandenberghe, *Introduction to applied linear algebra: vectors, matrices, and least squares*.
631 Cambridge university press, 2018.
- 632 [52] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, “Attentive pooling networks,” *arXiv preprint arXiv:1602.03609*,
633 2016.
- 634 [53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE
635 conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- 636 [54] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, “Machine learning
637 approaches and databases for prediction of drug–target interaction: a survey paper,” *Briefings in bioinformatics*,
638 vol. 22, no. 1, pp. 247–269, 2021.
- 639 [55] R. Wang, X. Fang, Y. Lu, and S. Wang, “The pdbsbind database: Collection of binding affinities for protein–
640 ligand complexes with known three-dimensional structures,” *Journal of medicinal chemistry*, vol. 47, no. 12,
641 pp. 2977–2980, 2004.