PPIscreenML: Structure-based screening for protein-protein interactions using AlphaFold

Victoria Mischley[1,2], Johannes Maier[3], Jesse Chen[3], and John Karanicolas[1,4*]

[1] Cancer Signaling & Microenvironment Program, Fox Chase Cancer Center, Philadelphia PA 19111

[2] Molecular Cell Biology and Genetics, Drexel University, Philadelphia PA 19102

[3] Triana Biomedicines, Lexington MA 02421

[4] Moulder Center for Drug Discovery Research, Temple University School of Pharmacy, Philadelphia PA 19140

*To whom correspondence should be addressed.

E-mail: **john.karanicolas@fccc.edu**, 215-728-7067

Abstract

Protein-protein interactions underlie nearly all cellular processes. With the advent of protein structure prediction methods such as AlphaFold2 (AF2), models of specific protein pairs can be built extremely accurately in most cases. However, determining the relevance of a given protein pair remains an open question. It is presently unclear how to use best structure-based tools to infer whether a pair of candidate proteins indeed interact with one another: ideally, one might even use such information to screen amongst candidate pairings to build up protein interaction networks. Whereas methods for evaluating quality of modeled protein complexes have been co-opted for determining which pairings interact (e.g., pDockQ and iPTM), there have been no rigorously benchmarked methods for this task. Here we introduce PPIscreenML, a classification model trained to distinguish AF2 models of interacting protein pairs from AF2 models of compelling decoy pairings. We find that PPIscreenML out-performs methods such as pDockQ and iPTM for this task, and further that PPIscreenML exhibits impressive performance when identifying which ligand/receptor pairings engage one another across the structurally conserved tumor necrosis factor superfamily (TNFSF). Analysis of benchmark results using complexes not seen in PPIscreenML development strongly suggest that the model generalizes beyond training data, making it broadly applicable for identifying new protein complexes based on structural models built with AF2.

Introduction

Cellular homeostasis is maintained by a complex protein interaction network that contributes to intricate signaling pathways within the cell [1]. As the human interactome is estimated to contain between 74,000 – 200,000 protein interactions [2], accurate and high-throughput methods to identify these interactions are crucial [3]. Classically, discovery and cataloguing of protein-protein interactions (PPIs) have most commonly employed yeast two-hybrid assays [4] and affinity purification assays [5]. More recently, proximity labeling techniques such as BioID [6] and TurboID [7] have gained popularity. However, each method has its own specific shortcomings, which further compounded due to sensitive dependence on experimental techniques and cellular conditions [1,8]. Collectively, the variabilities in these approaches ultimately lead to high rates of both false positives and false negatives [8,9]. For example, statistical analysis of representative interactome mapping data showed that recovering 65% of true protein interactions would require combining 10 distinct assays [1].

Computational predictions offer a natural means for complementing these experimental techniques. Early methods focused on inferring PPIs based on sequence information, by searching for gene pairs with correlated expression patterns, or distinct genes that can be found fused together in a different organism, or distinct genes that co-evolve with one another [10]. Drawing upon the exciting successes of large language models in many contexts, deep learning methods have also been applied to infer binary PPI pairings using protein sequences [11-13]. Results from such methods are extremely promising, however their lack of explainability make it difficult to determine how such models might perform on PPIs that are dissimilar from the training examples (i.e., their domain of applicability). Moreover, these models are inherently not designed to provide structural information about the bound pose: this information can be

extremely valuable, both to provide biological understanding that spurs further experimentation, and as a starting point for developing new therapeutic agents that modulate the PPI [14].

Broad availability of AlphaFold2 (AF2) [15], especially through the ColabFold framework [16], has provided the research community with unprecedented access to structural information for monomeric proteins, in many cases at resolution comparable to experimentally-determined structures. The startling accuracy of certain AlphaFold2 predictions, in turn, enabled structure-based modeling for numerous new applications in biology and medicine [17,18]. Whereas the original AF2 could be adapted for highly successful predictions of oligomeric assemblies, a subsequent re-trained release dubbed AF2-Multimer provided further enhanced performance in this regime [19].

While AF2 provides often-accurate predictions for interacting protein pairs, there is not a direct measure to evaluate the likelihood that a given query pair indeed interacts with one another – a set of predicted output structures will be returned regardless. AF2 also provides residue-level confidence measures of the predicted structures, specifically the predicted local-distance difference test (pLDDT) [20], the predicted TM-score (pTM) [21], and the predicted aligned error (PAE). The accuracy of structures for predicted protein complexes is often defined using DockQ [22], a metric that combines three complementary and frequently-used quality measures. By combining the number of interface residues with the pLDDT values of interface residues, a confidence measure pDockQ was developed to define the likelihood of an accurate binding mode in protein complexes predicted by AF2 [23]. Interestingly, pDockQ also provided some ability to discriminate between “true” (interacting) versus “decoy” (non-interacting) protein pairs [23] – performance for this task was only modest, however, presumably because development of pDockQ was fundamentally focused elsewhere (recapitulating DockQ values).

In light of interest in using AF2 to distinguish between interacting and non-interacting protein pairs, the AlphaPulldown [24] python package calculates several measures that have been proposed for this task: iPTM [19], PI-score [25], and pDockQ [23]. While each of these measures are intuitively reasonable, none were expressly developed for this classification task.

Here we describe PPIscreenML, a machine learning classifier that uses AF2 models to distinguish interacting versus non-interacting protein pairs. To do so, PPIscreenML uses descriptors drawn from AF2 confidence measures [15] and complements these with energetic terms from the Rosetta scoring function [26]. In contrast to methods such as pDockQ, which was trained as a means of evaluating model quality, PPIscreenML is deliberately and intentionally trained/tested to distinguish between models of interacting proteins and compelling decoy pairs. PPIscreenML exhibits superior performance to other methods for identifying interacting pairs in a retrospective screen. Finally, using the tumor necrosis factor superfamily (TNFSF) as an example of a structurally conserved family of ligand/receptor pairings, we demonstrate that PPIscreenML can accurately recapitulate the selectivity profile in this challenging regime.

Computational Approach

Dataset for training/testing PPIscreenML

Critical to development of a useful classification model is compiling a set of "active" and "decoy" examples that accurately reflect those that will be encountered in future prospective applications. For PPIscreenML, we employed an overarching strategy that mirrors our previous studies leading to a machine learning classifier that identifies active protein / small molecule complexes likely to engage one another [27]. Specifically, our approach seeks to build highly

87 *compelling* decoy complexes, so that the resulting model is challenged to distinguish actives
88 from decoys that cannot be resolved using solely trivial criteria.

89 To build relevant and diverse structures of active protein pairings (**Figure 1a**), we used
90 the DockGround database [28] to assemble the complete set of heterodimeric protein complexes
91 in the PDB excluding homodimers, antibody/antigen complexes, and complexes that were either
92 very large or very small (see *Methods*). Upon filtering for a maximum of 30% sequence identity,
93 there remained 1,481 non-redundant protein-protein complexes with experimentally defined
94 structures. The planned application of PPIscreenML is to distinguish AF2 models of active pairs
95 from AF2 models of inactive pairs – thus, the model should be trained on AF2 actives, rather
96 than experimentally defined structures. Accordingly, we used the sequences of the component
97 proteins to build five AF2 models for each of these 1,481 non-redundant complexes.

98 Active protein-protein pairs that are mis-predicted by AF2 should not lead to predictions
99 of "active" from PPIscreenML, since the specific interactions in the AF2 are not those that lead
100 to binding. In the context of training, then, we sought to use only those models that were
101 correctly predicted by AF2: we therefore filtered the models and retained only those with a
102 DockQ score [22] of at least 0.23 relative to the corresponding experimentally defined structure
103 (0.23 is the quality cutoff defined to be an "acceptable" prediction for the MOAL and CAPRI
104 sets). Filtering on model quality in this manner excluded 932 models from training, leading to a
105 total of 6,473 active models with median DockQ value of 0.88 (**Figure S1a**).

106 In a prospective application of PPIscreenML, however, the DockQ score will not be
107 known *a priori* (since this requires an experimentally defined structure, which itself is evidence
108 of an interaction). To ensure a realistic estimate of the performance that can be expected in future
109 prospective experiments, complexes with structures that were mis-predicted by AF2 were *not*

110 excluded from the test set. Accordingly, active pairs that are not correctly classified by
111 PPIscreenML may occur either because AF2 builds a poor model, or because PPIscreenML fails
112 to recognize a well-built model as an interacting pair. While the test set performance thus reflects
113 more than the success of PPIscreenML, this strategy enhances confidence that the observed test
114 set performance will carry forward to prospective applications in future.

115 With respect to creating a set of decoy complexes (for both training and testing), it is
116 essential that the decoys suitably represent the types of inactive complexes expected to be
117 encountered in prospective applications. In particular, decoys that can be dismissed as inactive
118 based on trivial criteria offer little benefit in developing PPIscreenML: rather, we prioritized
119 building a set of extremely compelling decoys [27], so that training would be impelled to find a
120 more sophisticated and nuanced model for classification.

121 To build a set of compelling decoys (**Figure 1a**), we began by using each active complex
122 as a template. For each component protein in the complex, we identified its closest structural
123 analog (based on TM-score [29]) amongst all other proteins in our collection. Despite the fact
124 that homologs had been excluded from our dataset on the basis of sequence identity, close
125 structural matches could nonetheless be identified in most cases (**Figure S1b**): 87% of the
126 proteins in our set were paired with a structural analog having TM-score greater than 0.5, a
127 cutoff far beyond that which is observed between random protein pairs [30]. The structural
128 analog for each component protein was then superposed onto the template in the active complex,
129 providing a crude model of these two (presumably) non-interacting proteins that structurally
130 resembles an active complex (**Figure 1b**). As noted earlier, the planned application of
131 PPIscreenML is to distinguish AF2 models of active pairs from AF2 models of inactive pairs: as

for the active pairings, we therefore used the sequences of the component proteins in each of these "decoy" complexes to build five AF2 models for each.

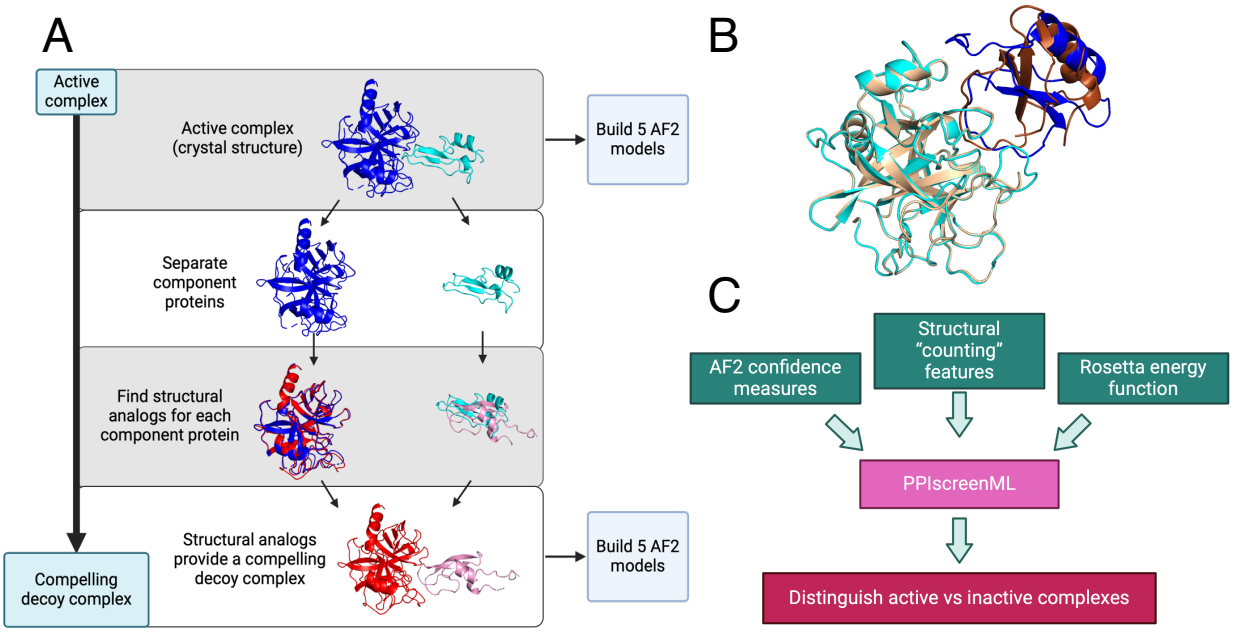

Figure 1: Building a challenging training/testing set for PPIscreenML. (A) A collection of 1,481 non-redundant active complexes with experimentally derived structures were obtained from DockGround, and five AF2 models were built from each of these. To build decoys, the same collection was screened to identify the closest structural matches (by TM-score) for each component protein. The structural homologs for each template protein were aligned onto the original complex, yielding a new (decoy) complex between two presumably non-interacting proteins. Five AF2 models were built for each of these 1,481 decoy complexes. **(B)** An example of a decoy complex (*blue/cyan*) superposed with the active complex from which is was generated (*brown/wheat*). **(C)** A suite of AlphaFold confidence metrics, structural properties, and Rosetta energy terms were used as input features for training PPIscreenML, a machine learning classifier built to distinguish active versus compelling inactive protein pairs.

To train PPIscreenML, data was partitioned using a 60/20/20 split. The data reserved for testing was not used in any phase of training or validation. The dataset includes 5 AF2 models for each (active or decoy) complex: training on one AF2 model of a given complex then using a different model of the same complex in the test set would introduce obvious information leakage. To avoid this, all 5 models for a given complex were *together* placed in either the training set or the validation set or the test set [31]. As noted earlier, all five AF2 models for a given complex

143 were included in the validation and testing sets; however, only those AF2 models of active
144 complexes close to the experimentally derived structures were included in the training set
145 (**Figure S2**).

146 Having partitioned the AF2 models into training, validation, and test sets, a series of
147 quantitative features were extracted from each model. Specifically, three groups of features were
148 calculated from each model (**Figure 1c**): 7 features representing AF2 confidence measures, 33
149 structural "counting" features, and 17 features from the Rosetta energy function. This set of 57
150 features (**Figure S3**) was initially evaluated in development of PPIscreenML.

151 *Developing the PPIscreenML model*

152 Using the compiled collection of features for each complex, we evaluated seven standard
153 machine learning frameworks to build classification models. In each case, models were trained
154 using 5-fold cross validation within the test set, and applied to a validation set that was not
155 included in each fold of training. Models were then compared using the combined cross-fold
156 AUC score (*A*rea *U*nder the *C*urve of a *R*eceiver *O*perating *C*haracteristic (ROC) plot). All seven
157 models had similar performance, with XGBoost (an implementation of gradient-boosted decision
158 trees) yielding the highest AUC value of 0.924 (**Figure S4**).

159 The model was then applied to the test set, that had been completely held-out from model
160 training and evaluation. To better recapitulate a real-world use case, this experiment incorporates
161 two key differences from the earlier evaluation using the validation set (**Figure S2**). First, AF2
162 models of active complexes with poor DockQ scores relative to the corresponding PDB structure
163 are included in the test set, whereas these had been excluded from the validation set. This reflects
164 the reality that active complexes will occasionally be mis-docked by AF2, and these models
165 cannot be discarded in a true prospective application. Second, rather than evaluate performance

using the scores of individual AF2 models, the top-ranked model for each complex is drawn from the scores of all five AF2 models. Here too, in a prospective screening application one would mitigate the possibility that individual AF2 models may be mis-docked by simply using the model that PPIscreenML finds most compelling. Applying the XGBoost model to the held-out test set with these conditions yielded an AUC value of 0.892 (**Figure 2a**).

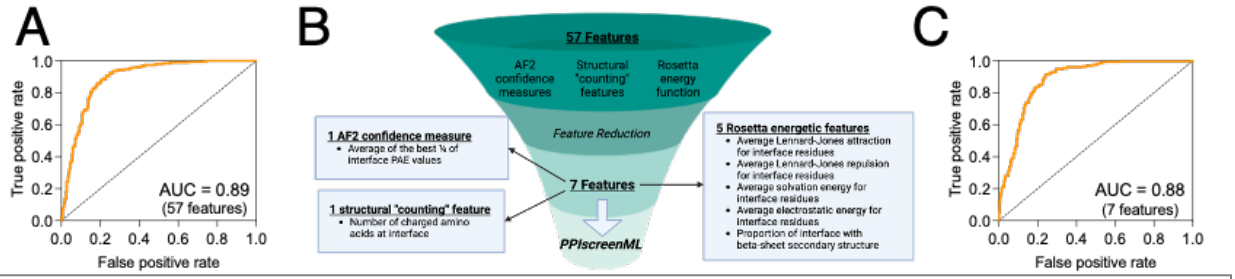

Figure 2: Training and feature reduction for PPIscreenML. (A) Receiver operating characteristic (ROC) plot demonstrating classification performance on a completely held-out test set, for an XGBoost model using 57 features. **(B)** The number of features was reduced using sequential backwards selection, from 57 features to 7 features. **(C)** Classification performance of PPIscreenML (7 features) on the same completely held-out test set.

To mitigate any potential overtraining, sequential backwards selection was applied to reduce the number of features, using the validation set (*not* the test set, which is not used in any model optimization). This approach provided a model with only seven features that maintained the AUC of the model with all features when evaluated using the validation set (**Figure S5**). Specifically, this model uses one feature from the AF2 confidence measures (the average of the top ¼ of the interfacial PAE values), one structural "counting" feature (number of charged amino acids in the interface), and five features from the Rosetta energy function (the average Lennard-Jones attractive score for interfacial residues, the average Lennard-Jones repulsive score for interfacial residues, the average solvation score for interfacial residues, the average electrostatic score for interfacial residues, and the proportion of interfacial residues involved in beta-sheet structures) (**Figure 2b**).

When applied to the test set that was completely held out from model training and optimization, this 7-feature PPIscreenML model yields an AUC of 0.884, comparable to the full 57-feature model (**Figure 2c**).

Results

PPIscreenML out-performs other structure-based methods for identifying interacting protein pairs

Already several structure-based methods have been proposed as a means for distinguishing between models of interacting versus non-interacting protein pairs: these include iPTM [19], PI-score [25], and pDockQ [23]. To evaluate the performance of PPIscreenML relative to these tools, we applied each method to the collection of AF2 models for active pairings and compelling decoys that comprise our test set. As noted earlier, the test set does *not* remove models of true pairings for which the structures have been incorrectly built by AF2, since this would not be possible in a real-world setting. In each case we scored all five AF2 models for a given pairing, and we took the highest score from each method as an indication of whether each pairing is inferred to be active or decoy (**Figure S2**). We note that none of the complexes in this test set have not been used in any phase of PPIscreenML development, but we cannot rule out that some of the active complexes were used in training the other methods.

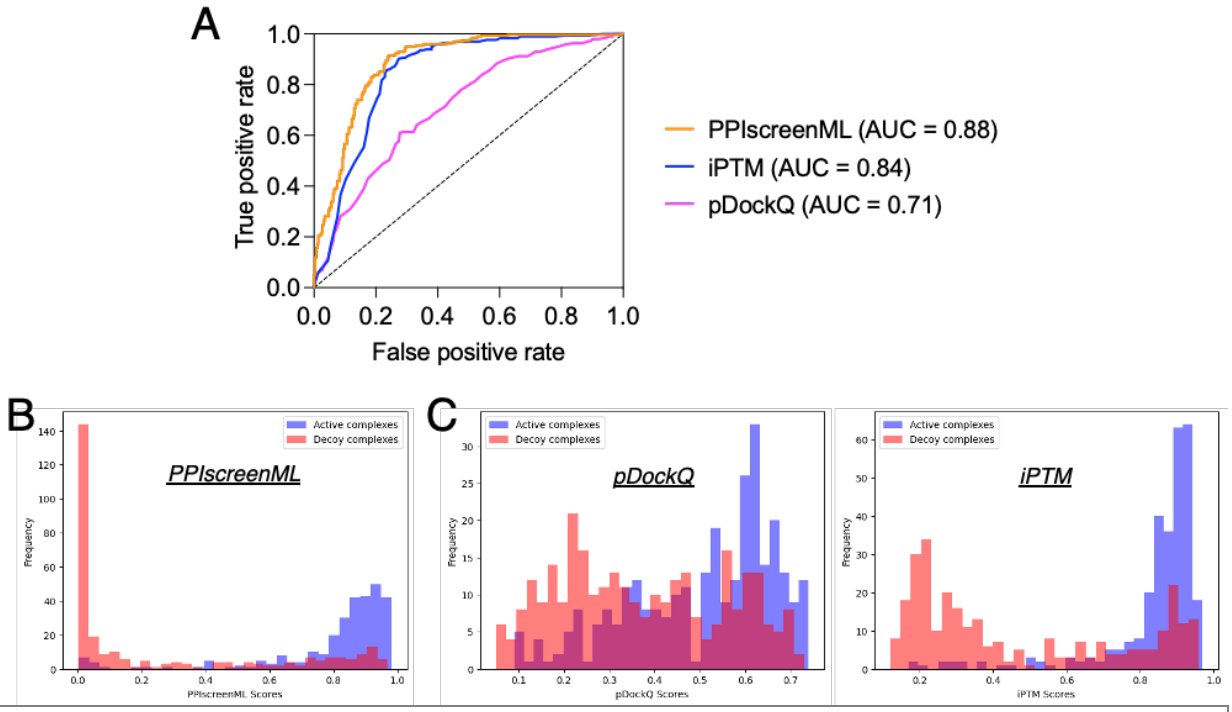

Figure 3: Classification performance of PPIscreenML relative to pDockQ and iPTM. The same test set is used here. These complexes were not seen in any phase of developing PPIscreenML, but may have been used in developing pDockQ or iPTM. **(A)** Receiver operating characteristic (ROC) plot shows superior performance of PPIscreenML relative to these other two methods. **(B)** Overlaid histograms show clear separation of actives and decoys scored using PPIscreenML. **(C)** Overlaid histograms show overlapping distributions when models are scored with pDockQ or iPTM.

Applying each method to the same test set, we find that PPIscreenML out-performs the other methods: its AUC value is 0.884, compared to AUC values of 0.843 and 0.710 for iPTM and pDockQ, respectively (**Figure 3a**). The basis for these AUC values is evident from the distribution of raw scores assigned to active / decoy complexes by each method. PPIscreenML assigns (correct) scores very close to either 1 or 0 for most of the actives or decoys, with clear separation between them (**Figure 3b**); on the other hand, the alternate methods both assign more scores with intermediate values, and they also include many incorrect assignments (**Figure 3c**).

An advantage of using *R*eceiver *O*perating *C*haracteristic (ROC) plots for defining the performance of binary classifiers is that one need not define a threshold score at which a given input would be assigned the label of active or inactive; rather, all points in the test set are simply

ranked, and the ROC plot reflects the preponderance of actives that score ahead of decoys. To facilitate adoption of PPIscreenML in different setting, we use the test set results to calculate various performance measures (e.g. false positive rate, false negative rate) as a function of threshold score used to assign the label (**Table 1**). The use of a high (stringent) threshold value yields high precision but poor recall: this would be preferable in a very broad screening context in which one seeks (for example) to identify new candidate interactions at proteome-wide scale, and one wishes to test only the more confidently predicted pairings. Conversely, a lower (less stringent) threshold value yields poorer precision but improved recall: this would be preferable when screening in a more restricted space such as (for example) when seeking to identify additional pairings for members of a protein family known to interact with one another.

PPIscreenML Threshold	FPR	TPR	FNR	TNR	Precision	Recall	F1 Score
0.98	0.00	0.00	1.00	1.00	1.00	0.00	0.01
0.95	0.01	0.16	0.84	0.99	0.96	0.16	0.27
0.9	0.07	0.39	0.61	0.93	0.85	0.39	0.54
0.85	0.11	0.60	0.40	0.89	0.85	0.60	0.70
0.8	0.15	0.74	0.26	0.85	0.83	0.74	0.78
0.75	0.18	0.80	0.20	0.82	0.82	0.80	0.81
0.7	0.21	0.85	0.15	0.79	0.80	0.85	0.83
0.64	0.23	0.88	0.12	0.77	0.80	0.88	0.84
0.6	0.24	0.89	0.11	0.76	0.79	0.89	0.84
0.56	0.24	0.92	0.08	0.76	0.79	0.92	0.85
0.5	0.27	0.93	0.07	0.73	0.78	0.93	0.85
0.47	0.27	0.93	0.07	0.73	0.77	0.93	0.84
0.4	0.29	0.95	0.05	0.71	0.76	0.95	0.85
0.38	0.30	0.95	0.05	0.70	0.76	0.95	0.85
0.28	0.33	0.95	0.05	0.67	0.75	0.95	0.84
0.21	0.35	0.95	0.05	0.65	0.73	0.95	0.83
0.18	0.35	0.96	0.04	0.65	0.73	0.96	0.83
0.08	0.43	0.96	0.04	0.57	0.69	0.96	0.81
0.05	0.48	0.97	0.03	0.52	0.67	0.97	0.79
0	1.00	1.00	0.00	0.00	0.50	1.00	0.67

Table 1: Performance of PPIscreenML using various threshold values. By adjusting the threshold score at which a test complex is assigned as active/decoy PPIscreenML can be used in regimes that prioritize returning only the most confident pairings (a high threshold score yields high precision but poor recall) or in exploratory regimes that return more speculative pairings as well (a lower threshold score yields high recall but poorer precision).

226 *PPIscreenML recapitulates selectivity within a protein superfamily*

227 As a challenging further benchmark for PPIscreenML performance, we sought to
228 evaluate whether it could recapitulate the pattern of binding interactions across a structurally
229 conserved protein superfamily. The tumor necrosis factor superfamily (TNFSF) comprises 18
230 protein ligands and 28 receptors in humans, leading to 504 potential pairings. However, only 36
231 of these pairings are reported to show productive binding to one another, as observed in a
232 comprehensive flow cytometry experiment [32].

233 The TNFSF ligands share a conserved fold, and they assemble into trimers. The TNFSF
234 receptors also share a conserved fold, and three copies of the cell-surface receptor independently
235 bind around the edges of the trimeric ligand. While AF2 typically provides highly accurate
236 structures of cognate ligand/receptor complexes, it also builds analogous models for
237 ligand/receptor pairings that do not actually interact (**Figure 4a**). Accordingly, this system
238 represents an ideal benchmark to evaluate whether PPIscreenML can use structural cues to
239 distinguish active pairings from close analogs that do not bind one another. Importantly for this
240 benchmark, no TNFSF ligands or receptors were included in the development of PPIscreenML
241 thus far, because the training and testing sets were restricted to dimeric complexes.

242 We used AF2 to build models of all 504 potential TNFSF ligand/receptor pairings, and
243 then evaluated these models using PPIscreenML. Overall, the ligand/receptor pairings that
244 scored most favorably with PPIscreenML indeed correspond to those which have been described
245 experimentally (**Figure 4b**). Of note, for 14 of the 18 TNFSF ligands, the receptor with the
246 highest PPIscreenML score is indeed a true interactor; of the two receptors with the highest
247 PPIscreenML score, a true interactor is found for 17 of the 18 ligands.

248 Analysis of these data via ROC plot confirms the outstanding performance of
249 PPIscreenML for identifying interacting pairs (**Figure 4c**). Further, the AUC value of 0.93 in this

new benchmark is similar to that observed when applying PPIscreenML to the held-out test set of heterodimeric complexes (0.88), supporting the generality of PPIscreenML for application to new data beyond the training regime.

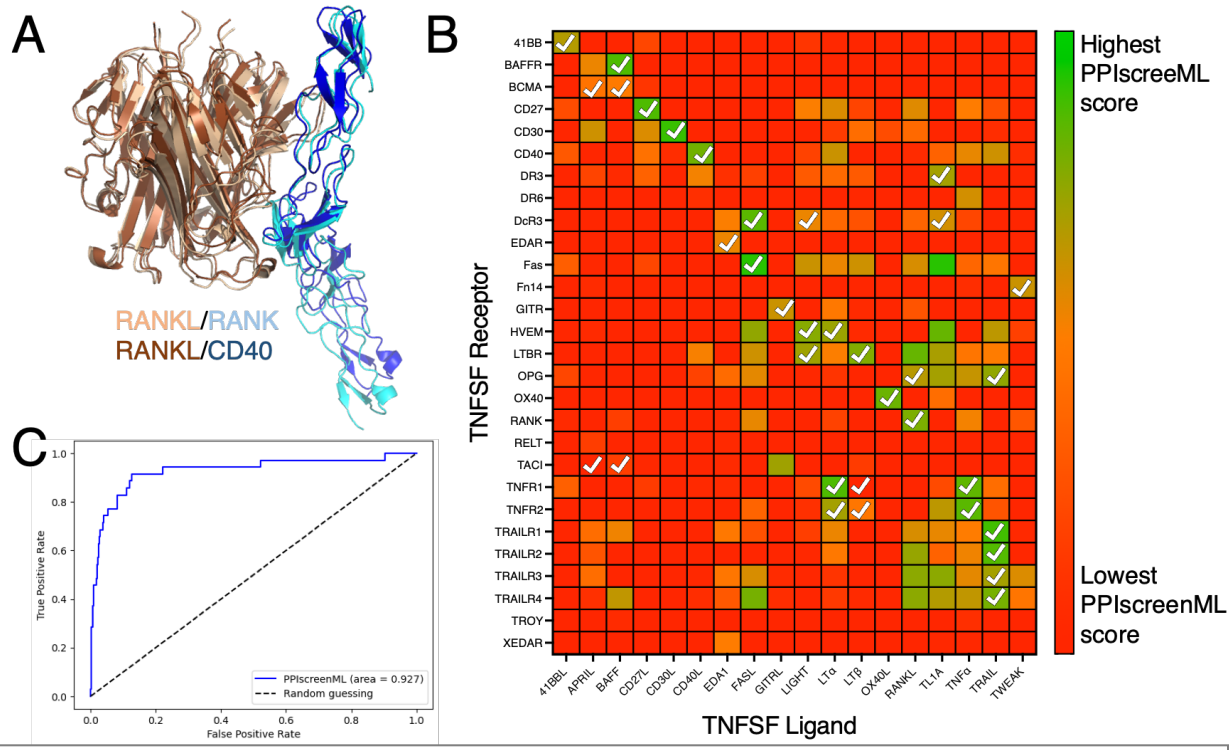

Figure 4: Application of PPIscreenML to identify active pairings within the tumor necrosis factor superfamily (TNFSF). (A) Structurally-conserved TNFSF ligands bind to structurally-conserved TNFSF receptors; AF2 builds models of these complexes in the canonical pose for cognate pairings (RANKL/RANK are shown in *wheat/cyan*) but also in some cases for non-cognate pairings (RANKL/CD40 are shown in *brown/blue*). **(B)** Each ligand/receptor pairing was built with AF2 and scored with PPIscreenML (heatmap colored from low score in *red*, to high score in *green*). Ligand/receptor pairings observed in a comprehensive cellular assay are indicated with *white* checkmarks. **(C)** Receiver operating characteristic (ROC) plot demonstrating PPIscreenML classification of TNFSF ligand/receptor pairings.

Discussion

The availability of AlphaFold2 and related methods [15,16,19,33], has enabled structure-based analysis of proteins and protein complexes to an extent that has rapidly transformed many

259 parts of biology and medicine [17,18]. Numerous studies have incorporated AF2 models to
260 justify or rationalize the structural basis for a particular protein-protein interaction, and in many
261 cases use either iPTM or pDockQ as further support for the relevance of the proposed
262 interaction. Such studies – which range from predicting protein quaternary structures to defining
263 biological functions of peptides and intrinsically disordered regions [34-40] – highlight the value
264 of using AF2 to nominate new candidate protein-protein interactions (i.e., in a screening setting),
265 extending it beyond simply providing structures for fully validated protein pairings.

266 While the early adoption of iPTM or pDockQ has spurred their use in this screening,
267 neither one was intended for this use-case. Rather, both were developed with the goal of
268 assessing model quality for AF2 structures, and the ability to distinguish between active and
269 inactive structures on the basis of model quality was a convenient secondary benefit [19,23]. Due
270 to this history, however, neither method was rigorously benchmarked in a manner that would set
271 expectations for their performance in real-world applications.

272 To lay the groundwork for this study, we began by building a set of non-redundant active
273 complexes, and from these a set of highly compelling decoy complexes. The overarching
274 philosophy behind construction of this set was to realistically mimic conditions that would be
275 encountered in future (prospective) screening applications. Notably, distinguishing actives from
276 decoys in this set is markedly more challenging than in other (less realistic) benchmarks: for
277 example, pDockQ provided an AUC value of 0.87 in its first-reported classification experiment
278 [23], but only 0.71 when applied to this newer set (**Figure 3a**).

279 Developing this challenging and realistic set allowed us to train PPIscreenML, a tool
280 expressly designed and optimized to be used in screening for protein-protein interactions. As
281 expected, performance of PPIscreenML is superior to that of iPTM or pDockQ for this screening

282 regime (**Figure 3a**). Moreover, PPIscreenML also makes use of subtle structural cues to provide
283 its discriminative power, as evidenced from its impressive performance on the TNFSF selectivity
284 benchmark (**Figure 4**).

285 Contemporaneously with our study, other groups have developed alternate methods for
286 predicting pairs of interacting proteins. For example, the AF2Complex method also shows
287 superiority over iPTM [41], albeit through testing on arbitrary collections of protein pairings
288 (rather than on compellingly-built decoys).

289 The natural extension of using AF2 and related methods to carry out structure-based
290 screening for protein-protein interactions, naturally, is using these methods to define whole
291 interactomes. Already certain studies have applied these tools either within targeted protein
292 families [42,43] for groups of proteins thought to belong together in large complexes [41,44,45].
293 Importantly, careful consideration of the statistical measures needed for productively scaling
294 assessments to whole proteomes must complement these studies; for example, this will allow for
295 defining thresholds that ideally balance precision and recall (**Table 1**), and also ensure that the
296 expected results justify these larger-scale analyses.

297 There is also ample reason for optimism that performance of PPIscreenML will further
298 improve as the underlying AF2 models improve. Specifically, older versions of the AF2 software
299 yielded slightly worse performance from PPIscreenML (**Figure S6, S7**), but the difference in
300 performance could be attributed to the fact that a (slightly) smaller proportion of the active
301 complexes were correctly built. Thus, the true bottleneck in extending PPIscreenML may not lie
302 with distinguishing correct versus incorrect interfaces. Rather, to screening for interactions at
303 whole-proteome scale may currently by instead limited by the fact that not all protein pairs are

correctly modeled by AF2, and active pairings cannot be reliably identified if their structures are incorrectly built.

Methods

Model availability

All protein structures, code, and machine learning models described in this paper are available on GitHub (https://github.com/victoria-mischley/PPIScreenML.git).

Set of active complexes

Heterodimeric protein complexes were downloaded from the DockGround database [28]. Complexes that were downloaded were published in the PDB [46] between up to 01/26/23 with a maximum of 3 Å resolution. No homodimers or antibody/antigen complexes were included in the dataset. The dataset was further filtered for a maximum of 30% sequence identity. Any heterodimeric complex with a total length greater than 1,550 amino acids was filtered out of the dataset, due to our modeling capabilities at the time. Finally, the dataset was separated based on protein size. If a heterodimeric complex had at least one protein less than 50 amino acids, this complex was included in the protein-peptide dataset and excluded from the protein-protein dataset. After filtering, there were a total of 1,481 protein-protein complexes and 456 protein-peptide complexes.

The sequences for these protein complexes were then downloaded directly from the PDB [46]. Sequences were then used to model the complexes with ColabFold Batch [16] alphafold-

323 multimer v2 [19]. Templates were used in building AF2 models, each model was recycled 12
324 times, and the AMBER relax step was included.
325 To avoid including inaccurate structural models in the set of actives, DockQ score [22]
326 was used to filter out incorrect models. DockQ [22] is a continuous score that incorporates
327 information from three scoring metrics, allowing for a holistic view of the complex's accuracy:
328 the fraction of native interfaces (Fnat), the Ligand Root Mean Square Deviation (LRMS), and
329 iRMS. While DockQ is a continuous measure, a DockQ score of 0.23 or greater is considered at
330 least "acceptable" quality [22]. Therefore, we only included in the training set models with
331 DockQ scores greater than 0.23. Out of the five models provided by AlphaFold for each
332 heterodimeric complex, every model with a DockQ score greater than 0.23 was kept in the
333 training set (**Figure S1**). Keeping all models with a DockQ score greater than 0.23 rather than
334 taking only the model with the highest DockQ score allowed us to introduce some noise into our
335 training dataset. However, our test set included all five AF models, as we would not know the
336 correct multimeric structure a priori (**Figure S2**). Hence, utilizing all five models in the training
337 set provides a more meaningful readout of expected performance in future prospective
338 applications.

339 *Set of compelling decoy complexes*

340 The set of active complexes were separated into their individual component proteins. For
341 each protein, the TM-score [29] was calculated between that protein and every other protein. In
342 each complex, the two component proteins were replaced by their closest structural analogs: this
343 provided a simple means to generate new complexes that resemble known complexes, but that
344 are built using pairs of (presumably) non-interacting proteins. In cases for which pair of
345 structural analogs happened to come from the same alternate complex (i.e. the newly generated

346 pairing corresponds to one of the active pairings), the analogs with the second-best TM-scores
347 were used to generate a new decoy pairing.
348 After selection of the two proteins that would comprise a "decoy" pairing, their
349 sequences were used to build a set of AF2 models using the same settings as for the active
350 complexes.
351 We did note that this method for building compelling decoys, did give a slight bias
352 towards larger proteins being used (**Figure S8**): this arises because larger decoy proteins are
353 (slightly) more likely to have higher TM-score to the templates, and thus are slightly favored in
354 building decoy complexes. Because this artifact could be leveraged by features that explicitly
355 scale with the total size of the component proteins, such features were excluded from our study.

356 *Feature extraction*

357 Structural features were extracted from AF2 models using three python scripts (all
358 available on GitHub): one for AF2-based features, one for Rosetta-based features (implemented
359 using pyRosetta), and one for structural "counting" features (implemented using python package
360 BioPandas). Residues were classified as interface residues if the Cβ atom (or Cα for glycine)
361 was within 12 Å of a Cβ atom on the other chain.
362 To calculate iPAE top ¼, PAE values were calculated for all interacting residue pairs
363 then the highest ¼ of these values were averaged. To calculate tPAE top ¼, all PAE values for
364 interchain residues-residue pairs were calculated then the best (lowest) ¼ of these values were
365 averaged.

366 *Data partitioning*

367 Each protein pairing was randomly assigned to either training set / validation set / test set,
368 with an 60/20/20 split. To avoid information leakage, all AF2 models for a given protein pairing
369 were *together* placed in either training set / validation set / test set (rather than assign individual
370 AF2 models to different sets).

371 *Training and validation sets*

372 The active complexes in the training set were filtered to include only AF2 models with
373 greater than 0.23 DockQ score (relative to the corresponding crystal structure). Stratified kfold 5-
374 fold cross validation was used for the training and validation sets, such that all AF2 models for a
375 given complex were together placed in either the training set or the validation set.

376 *Test set*

377 Complexes used in the test set were separated from those in the training set, and not used
378 in any part of developing PPIscreenML.

379 Evaluation of PPIscreenML on the test set differed from the training/validation set in two
380 ways (**Figure S2**), both intended to ensure that model performance on the test set will extend to
381 future prospective applications. First, since experimentally derived structures will not be
382 available, AF2 models in the test set were not filtered based on DockQ. Thus, the test set
383 includes some active complexes for which the AF2 models are incorrectly built, as would occur
384 in a prospective context. Second, rather than evaluate PPIscreenML scores for individual AF2
385 models, performance was evaluated by using PPIscreenML to classify each complex.
386 Specifically, PPIscreenML was used to define scores for all five AF2 models, then the inferred
387 class for the complex was assigned based on the maximum value from the five models.

388 *Machine learning classifier*

389 Code for model training and optimization was adapted from a standard source [47].

390 PPIscreenML was built using the XGBoost framework [48]. Other machine learning
391 frameworks evaluated included decision trees, random forest, K-nearest neighbors, logistic
392 regression, AdaBoost, and Bagging (**Figure S4**). Sequential backwards selection was then used
393 to reduce the number of features from 57 to 7 (**Figure 2**). Finally, the model was hyperparameter
394 tuned on the following parameters: maximum depth, number of estimators, learning rate, gamma,
395 regularized lambda, and minimum child weight.

396 All model optimization (selecting the framework, feature reduction, and hyperparameter
397 tuning) was carried out using the validation set to define the best model, and not ever the test set.

398 *Impact of AF2 version*

399 Development of PPIscreenML and all performance benchmarking used models built with
400 AF-multimer v2.3. Performance is similar when PPIscreenML is applied to models built with
401 AF-ptm (**Figure S6**) or with AF-multimer v2.2 (**Figure S7**). In both cases the AUC values were
402 slightly worse than for AF-multimer v2.3, but analysis of incorrect predictions shows that this
403 difference can be traced to the fact that AF-ptm and AF-multimer v2.2 build mis-docked models
404 of active complexes slightly more often than AF-multimer v2.3 does.

405 *TNFSF case study*

406 All 18 TNFSF ligands were modeled against all 28 TNFSF receptors, for a total of 504
407 pairings (using sequences obtained from Uniprot). Given the shared arrangement of subunits in
408 all complexes described to date, a 3:3 stoichiometry was assumed in all cases. Because three

409 receptor subunits separately engage the outside edges of the trimeric ligand, AF2 models
410 included three copies of the ligand with one copy of the receptor.
411 Five models for each ligand/receptor pairing were built using the CF-batch alphafold-
412 multimer v2.3, with templates, 12 recycles and AMBER relax step. Features from each structural
413 model were extracted and used as input to PPIscreenML. Because PPIscreenML was trained on
414 heterodimeric complexes, the three copies of the trimeric ligand were considered as the first
415 entity and the receptor was considered as the second entity.
416 PPIscreenML results were compared to TNFSF binding specificity data drawn from a
417 previously published comprehensive survey using flow cytometry [32].

418 *Computational resources*
419 Approximately 7,700 GPU hours were used to build the AF2 models in this work. All
420 models were built on ACCESS San Diego Supercomputer's Expanse NVIDIA V100 GPU
421 processors [49].

422 Acknowledgements

423 Parts of several figures were created using BioRender.
424 This work was supported by the W.M. Keck Foundation, by the NIH National Institute of
425 General Medical Sciences (R01GM141513), and by the National Institute of Biomedical
426 Imaging and Bioengineering (through fellowship 1F30EB034594). This research was also
427 funded in part through the NIH/NCI Cancer Center Support Grant P30CA006927.
428 This work used the Extreme Science and Engineering Discovery Environment (XSEDE)
429 allocation MCB130049, which is supported by National Science Foundation grant number

1548562. This work also used computational resources through allocation MCB130049 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296.

References

1. Choi SG, Olivet J, Cassonnet P, Vidalain PO, Luck K, Lambourne L, Spirohn K, Lemmens I, Dos Santos M, Demeret C, Jones L, Rangarajan S, Bian W, Coutant EP, Janin YL, van der Werf S, Trepte P, Wanker EE, De Las Rivas J, Tavernier J, Twizere JC, Hao T, Hill DE, Vidal M, Calderwood MA, Jacob Y. Maximizing binary interactome mapping with a minimal number of assays. *Nat Commun*. 2019; 10:3907.

2. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M. An empirical framework for binary interactome mapping. *Nat Methods*. 2009; 6:83-90.

3. Elhabashy H, Merino F, Alva V, Kohlbacher O, Lupas AN. Exploring protein-protein interactions at the proteome level. *Structure*. 2022; 30:462-75.

4. Fields S, Song O. A novel genetic system to detect protein-protein interactions. *Nature*. 1989; 340:245-6.

5. Dunham WH, Mullin M, Gingras AC. Affinity-purification coupled to mass spectrometry: basic principles and strategies. *Proteomics*. 2012; 12:1576-90.

459 6. Roux KJ, Kim DI, Burke B, May DG. BioID: A Screen for Protein-Protein Interactions. *Curr*
460 *Protoc Protein Sci*. 2018; 91:19 23 1-19 23 15.
461
462 7. Cho KF, Branon TC, Udeshi ND, Myers SA, Carr SA, Ting AY. Proximity labeling in mammalian
463 cells with TurboID and split-TurboID. *Nat Protoc*. 2020; 15:3971-99.
464
465 8. Qin W, Cho KF, Cavanagh PE, Ting AY. Deciphering molecular interactions by proximity
466 labeling. *Nat Methods*. 2021; 18:133-43.
467
468 9. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment
469 of the reliability of high throughput observations. *Mol Cell Proteomics*. 2002; 1:349-56.
470
471 10. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for
472 genome-wide prediction of protein function. *Nature*. 1999; 402:83-6.
473
474 11. Zhao N, Zhuo M, Tian K, Gong X. Protein-protein interaction and non-interaction predictions using
475 gene sequence natural vector. *Commun Biol*. 2022; 5:652.
476
477 12. Jha K, Karmakar S, Saha S. Graph-BERT and language model-based framework for protein-protein
478 interaction identification. *Sci Rep*. 2023; 13:5663.
479
480 13. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: Boosting Prediction of Protein-Protein
481 Interactions with Deep Neural Networks. *J Chem Inf Model*. 2017; 57:1499-510.
482

483 14. Lu H, Zhou Q, He J, Jiang Z, Peng C, Tong R, Shi J. Recent advances in the development of
484 protein-protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target*
485 *Ther*. 2020; 5:213.
486
487 15. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates
488 R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-
489 Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M,
490 Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW,
491 Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold.
492 *Nature*. 2021; 596:583-9.
493
494 16. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making
495 protein folding accessible to all. *Nature Methods*. 2022; 19:679-+.
496
497 17. Akdel M, Pires DEV, Pardo EP, Janes J, Zalevsky AO, Meszaros B, Bryant P, Good LL, Laskowski
498 RA, Pozzati G, Shenoy A, Zhu W, Kundrotas P, Serra VR, Rodrigues CHM, Dunham AS, Burke
499 D, Borkakoti N, Velankar S, Frost A, Basquin J, Lindorff-Larsen K, Bateman A, Kajava AV,
500 Valencia A, Ovchinnikov S, Durairaj J, Ascher DB, Thornton JM, Davey NE, Stein A, Elofsson A,
501 Croll TI, Beltrao P. A structural biology community assessment of AlphaFold2 applications. *Nat*
502 *Struct Mol Biol*. 2022; 29:1056-67.
503
504 18. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and
505 medicine. *Signal Transduct Target Ther*. 2023; 8:115.
506
507 19. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S,
508 Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A,

Tunyasuvunakool K, Jain R, Clancy E, Kohli P, Jumper J, Hassabis D. Protein complex prediction with AlphaFold-Multimer. In press 2021.

20. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013; 29:2722-8.

21. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005; 33:2302-9.

22. Basu S, Wallner B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One*. 2016; 11:e0161879.

23. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 (vol 13, 1265, 2022). *Nature Communications*. 2022; 13.

24. Yu D, Chojnowski G, Rosenthal M, Kosinski J. AlphaPulldown-a python package for protein-protein interaction screens using AlphaFold-Multimer. *Bioinformatics*. 2023; 39.

25. Malhotra S, Joseph AP, Thiyagalingam J, Topf M. Assessment of protein-protein interfaces in cryo-EM derived assemblies. *Nat Commun*. 2021; 12:3399.

26. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, Basanta B, Bender BJ, Blacklock K, Bonet J, Boyken SE, Bradley P, Bystroff C, Conway P, Cooper S, Correia BE, Coventry B, Das R, De Jong RM, DiMaio F, Dsilva L, Dunbrack R, Ford AS, Frenz B, Fu DY, Geniesse C, Goldschmidt L, Gowthaman R, Gray JJ, Gront D, Guffy S, Horowitz S, Huang PS, Huber T, Jacobs TM, Jeliazkov JR, Johnson DK, Kappel

535 K, Karanicolas J, Khakzad H, Khar KR, Khare SD, Khatib F, Khramushin A, King IC, Kleffner R,
536 Koepnick B, Kortemme T, Kuenze G, Kuhlman B, Kuroda D, Labonte JW, Lai JK, Lapidoth G,
537 Leaver-Fay A, Lindert S, Linsky T, London N, Lubin JH, Lyskov S, Maguire J, Malmstrom L,
538 Marcos E, Marcu O, Marze NA, Meiler J, Moretti R, Mulligan VK, Nerli S, Norn C, O'Conchuir
539 S, Ollikainen N, Ovchinnikov S, Pacella MS, Pan X, Park H, Pavlovicz RE, Pethe M, Pierce BG,
540 Pilla KB, Raveh B, Renfrew PD, Burman SSR, Rubenstein A, Sauer MF, Scheck A, Schief W,
541 Schueler-Furman O, Sedan Y, Sevy AM, Sgourakis NG, Shi L, Siegel JB, Silva DA, Smith S, Song
542 Y, Stein A, Szegedy M, Teets FD, Thyme SB, Wang RY, Watkins A, Zimmerman L, Bonneau R.
543 Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods*.
544 2020; 17:665-80.
545
546 27. Adeshina YO, Deeds EJ, Karanicolas J. Machine learning classification can reduce false positives
547 in structure-based virtual screening. *Proc Natl Acad Sci U S A*. 2020; 117:18477-88.
548
549 28. Collins KW, Copeland MM, Kotthoff I, Singh A, Kundrotas PJ, Vakser IA. Dockground resource
550 for protein recognition studies. *Protein Sci*. 2022; 31:e4481.
551
552 29. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template
553 quality. *Proteins*. 2004; 57:702-10.
554
555 30. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5?
556 *Bioinformatics*. 2010; 26:889-95.
557
558 31. Ong WJG, Kirubakaran P, Karanicolas J. Poor Generalization by Current Deep Learning Models
559 for Predicting Binding Affinities of Kinase Inhibitors. *bioRxiv*. 2023.
560

561 32. Bossen C, Ingold K, Tardivel A, Bodmer JL, Gaide O, Hertig S, Ambrose C, Tschopp J, Schneider
562 P. Interactions of tumor necrosis factor (TNF) and TNF receptor family members in the mouse and
563 human. *J Biol Chem*. 2006; 281:13964-71.
564
565 33. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch
566 LN, Schaeffer RD, Millan C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH,
567 Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-
568 Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD,
569 Read RJ, Baker D. Accurate prediction of protein structures and interactions using a three-track
570 neural network. *Science*. 2021; 373:871-6.
571
572 34. Lim Y, Tamayo-Orrego L, Schmid E, Tarnauskaite Z, Kochenova OV, Gruar R, Muramatsu S,
573 Lynch L, Schlie AV, Carroll PL, Chistol G, Reijns MAM, Kanemaki MT, Jackson AP, Walter JC.
574 In silico protein interaction screening uncovers DONSON's role in replication initiation. *Science*.
575 2023; 381:eadi3448.
576
577 35. Bret H, Gao J, Zea DJ, Andreani J, Guerois R. From interaction networks to interfaces, scanning
578 intrinsically disordered regions using AlphaFold2. *Nat Commun*. 2024; 15:597.
579
580 36. Danneskiold-Samsoe NB, Kavi D, Jude KM, Nissen SB, Wat LW, Coassolo L, Zhao M, Santana-
581 Oikawa GA, Broido BB, Garcia KC, Svensson KJ. AlphaFold2 enables accurate deorphanization
582 of ligands to single-pass receptors. *bioRxiv*. 2023.
583
584 37. Marciano S, Dey D, Listov D, Fleishman SJ, Sonn-Segev A, Mertens H, Busch F, Kim Y, Harvey
585 SR, Wysocki VH, Schreiber G. Protein quaternary structures in solution are a mixture of multiple
586 forms. *Chem Sci*. 2022; 13:11680-95.

587

588 38. Martin J. AlphaFold2 Predicts Whether Proteins Interact Amidst Confounding Structural
589 Compatibility. *J Chem Inf Model*. 2024.

590

591 39. Teufel F, Refsgaard JC, Kasimova MA, Deibler K, Madsen CT, Stahlhut C, Gronborg M, Winther
592 O, Madsen D. Deorphanizing Peptides Using Structure Prediction. *J Chem Inf Model*. 2023;
593 63:2651-5.

594

595 40. Pogozheva ID, Cherepanov S, Park SJ, Raghavan M, Im W, Lomize AL. Structural Modeling of
596 Cytokine-Receptor-JAK2 Signaling Complexes Using AlphaFold Multimer. *J Chem Inf Model*.
597 2023; 63:5874-95.

598

599 41. Gao M, Nakajima An D, Parks JM, Skolnick J. AF2Complex predicts direct physical interactions
600 in multimeric proteins with deep learning. *Nat Commun*. 2022; 13:1744.

601

602 42. Weeratunga S, Gormal RS, Liu M, Eldershaw D, Livingstone EK, Malapaka A, Wallis TP,
603 Bademosi AT, Jiang A, Healy MD, Meunier FA, Collins BM. Interrogation and validation of the
604 interactome of neuronal Munc18-interacting Mint proteins with AlphaFold2. *J Biol Chem*. 2024;
605 300:105541.

606

607 43. Baryshev A, La Fleur A, Groves B, Michel C, Baker D, Ljubetic A, Seelig G. Massively parallel
608 protein-protein interaction measurement by sequencing (MP3-seq) enables rapid screening of
609 protein heterodimers. *bioRxiv*. 2023.

610

611 44. Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS,
612 Albanese P, Keller A, Scheltema RA, Bruce JE, Leitner A, Kundrotas P, Beltrao P, Elofsson A.

613 Towards a structurally resolved human protein interaction network. *Nat Struct Mol Biol*. 2023;
614 30:216-25.
615
616 45. Humphreys IR, Pei J, Baek M, Krishnakumar A, Anishchenko I, Ovchinnikov S, Zhang J, Ness TJ,
617 Banjade S, Bagde SR, Stancheva VG, Li XH, Liu K, Zheng Z, Barrero DJ, Roy U, Kuper J,
618 Fernandez IS, Szakal B, Branzei D, Rizo J, Kisker C, Greene EC, Biggins S, Keeney S, Miller EA,
619 Fromme JC, Hendrickson TL, Cong Q, Baker D. Computed structures of core eukaryotic protein
620 complexes. *Science*. 2021; 374:eabm4805.
621
622 46. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, Craig PA, Crichlow GV, Dalenberg K,
623 Duarte JM, Dutta S, Fayazi M, Feng Z, Flatt JW, Ganesan S, Ghosh S, Goodsell DS, Green RK,
624 Guranovic V, Henry J, Hudson BP, Khokhriakov I, Lawson CL, Liang Y, Lowe R, Peisach E,
625 Persikova I, Piehl DW, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Vallat B, Voigt M, Webb
626 B, Westbrook JD, Whetstone S, Young JY, Zalevsky A, Zardecki C. RCSB Protein Data Bank
627 (RCSB.org): delivery of experimentally-determined PDB structures alongside one million
628 computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids*
629 *Res*. 2023; 51:D488-D508.
630
631 47. Raschka S, Liu Y, Mirjalili V. Machine Learning with PyTorch and Scikit-Learn: Packt Publishing
632 Ltd; 2022.
633
634 48. Chen T, Guestrin C, editors. XGBoost. Proceedings of the 22nd ACM SIGKDD International
635 Conference on Knowledge Discovery and Data Mining; 2016 2016-08-13: ACM.
636
637 49. Strande S, Cai H, Tatineni M, Pfeiffer W, Irving C, Majumdar A, Mishin D, Sinkovits R, Norman
638 M, Wolter N, Cooper T, Altintas I, Kandes M, Perez I, Shantharam M, Thomas M, Sivagnanam S,

Hutton T. Expanse: Computing without Boundaries. Practice and Experience in Advanced Research Computing2021. p. 1-4.

Supplemental Figures

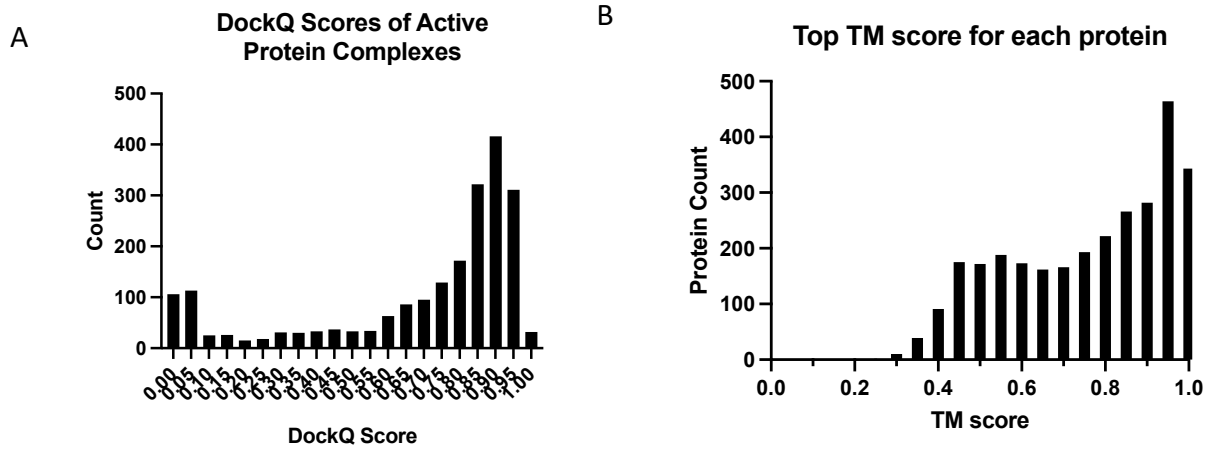

Figure S1: Composition of the active and decoy datasets. (A) For the set of actives, the distribution of DockQ scores are shown (calculated relative to the PDB structure for each complex). Complexes are included in the dataset only if their DockQ scores are at least 0.23, but most DockQ scores for these active pairings are much higher. **(B)** Compelling decoy complexes are assembled by drawing an alternate protein with high TM score relative to the components of the parental (active) complex. In most cases decoy complexes are assembled using component proteins with TM-scores of at least 0.5, corresponding to close structural matches.

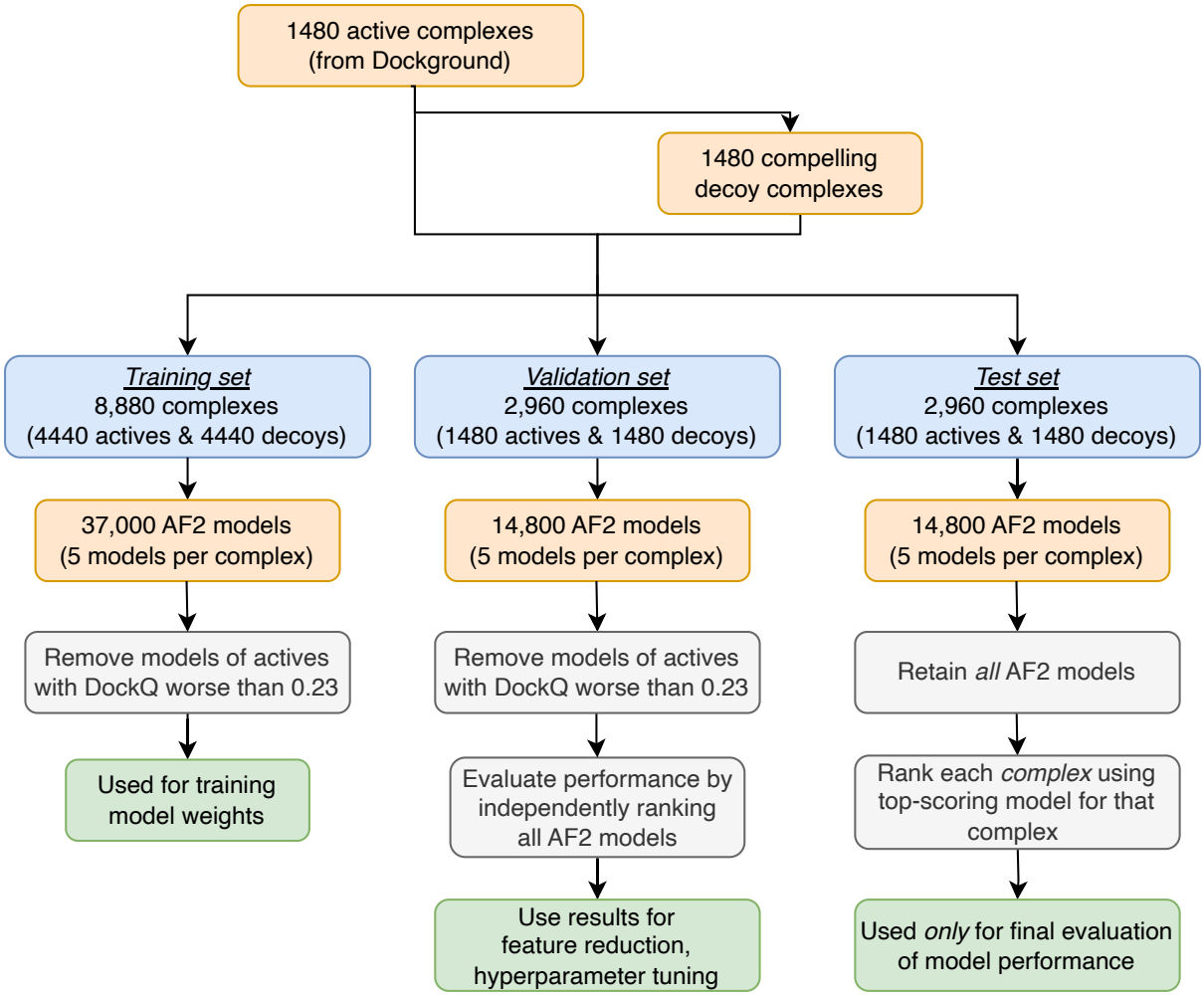

Figure S2: Overview of dataset construction and evaluation. Each active complex was used to generate one compelling decoy complex. Complexes were divided into training / validation / test sets, and 5 AF2 models were built for each complex. Key differences in how the sets were constructed / evaluated are highlighted in *white*. First, the AF2 models of active complexes in the training and validation sets were filtered to keep only correctly-docked models (DockQ>0.23); by contrast, *all* models were retained in the test set with no filtering. Second, performance of the validation set was evaluated by ranking all AF2 models independently; by contrast, performance of the test set was evaluated by ranking each *complex*, using the best-scoring AF2 model for each complex.

AF2 confidence measures

- PTM score
- plDDT
- iplDDT
- iPAE
- iPAE top ¼
- tPAE
- tPAE top ¼

Structural "counting" features

- Number of interface (IF) contacts
- Number of residues comprising the IF
- Number polar uncharged amino acids in IF
- Number nonpolar amino acids in IF
- Number of negative amino acids in IF
- Number of positive amino acids in IF
- Number of charged amino acids in IF
- Number of polar-polar contacts in IF
- Number of polar uncharged-polar uncharged contacts in IF
- Number of charged-charged contacts in IF
- Number of nonpolar-nonpolar contacts in IF
- Number of positive-negative contacts in IF
- Number of negative-positive contacts in IF
- Number of each individual amino acid type (20 features)

Rosetta energy function

- Total bound – unbound energy
- LJatr contribution to bound-unbound
- LJrep contribution to bound-unbound
- Solvation contribution to bound-unbound
- Electrostatic contribution to bound-unbound
- Lk_ball_wtd contribution to bound-unbound
- Number of interfacial hydrogen bonds
- Average LJatr of IF residues
- Average LJrep of IF residues
- Average sol of IF residues
- Average elec of IF residues
- Average fa_dun of IF residues
- Average hydrogen bonds energy of backbone side chains
- Proportion of IF secondary structure in helix
- Proportion of IF secondary structure in beta sheet
- Proportion of IF secondary structure in loop
- Interfacial buried solvent accessible surface area (bound – unbound)

Figure S3: List of 57 total features considered when training PPIscreenML. A list of features extracted from the structural models are shown and grouped into three categories: features from AF2 confidence measures, structural "counting" features extracted using the python package Biopandas, and features from the Rosetta energy function.

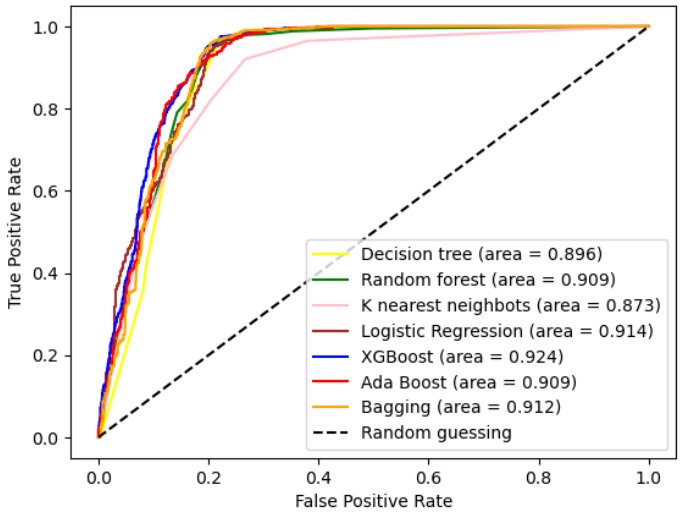

Figure S4: Comparisons of different machine learning classifiers. ROC plots for classifiers built using different machine learning frameworks (each uses all 57 features).

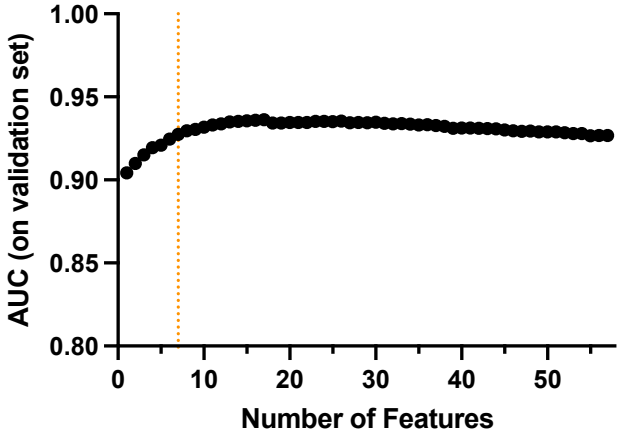

Figure S5: Feature reduction for PPIscreenML. Sequential backwards selection was used to characterize models with diminishing numbers of features. Performance of each candidate model was evaluated on the validation set (drawn from the training set). The vertical dashed line indicates the model selected for PPIscreenML (7 features).

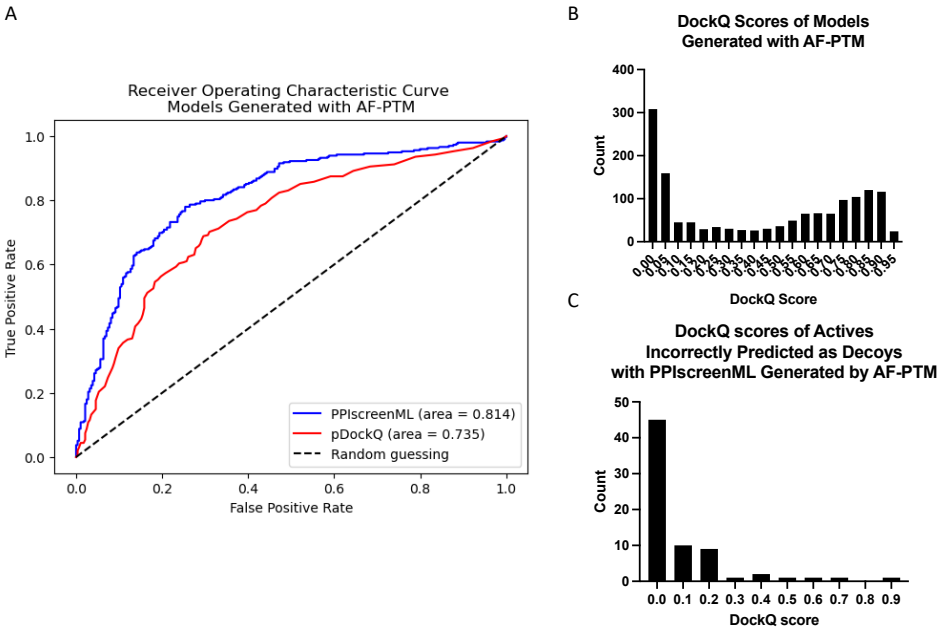

Figure S6: PPIscreenML performance on models generated with AF-ptm. (A) ROC curve of PPIscreenML and pDockQ tested on models built with AF-ptm. iPTM is not included because it is not available for this version of AF. **(B)** DockQ distribution for actives in the test set. **(C)** Among actives that were incorrectly classified by PPIscreenML as "not interacting", most were mis-docked by AF-ptm.

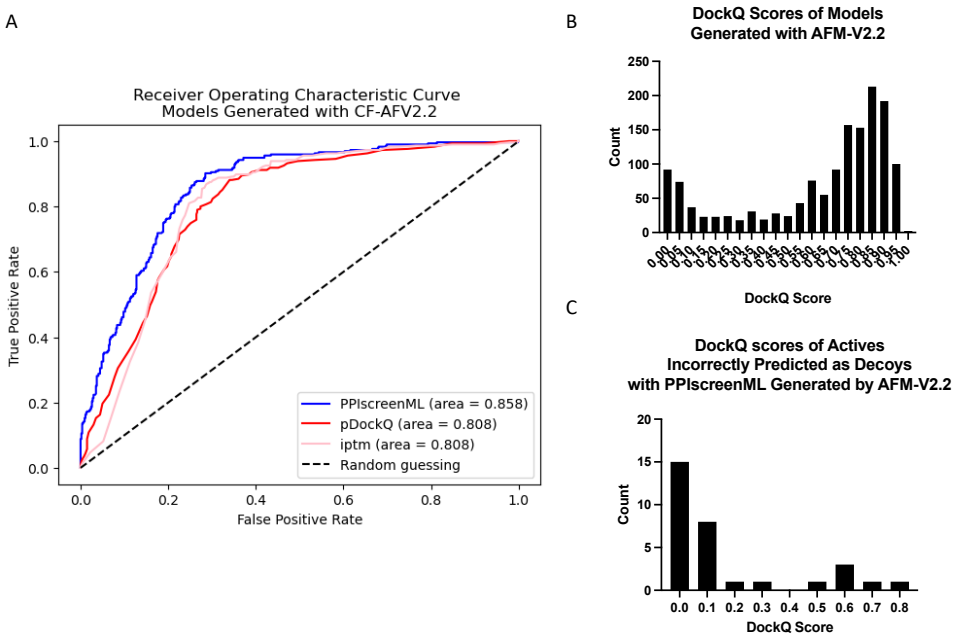

Figure S7: PPIscreenML performance on models generated with AF-multimer version 2.2. **(A)** ROC curve of PPIscreenML and pDockQ tested on models built with **AF-multimer-2.2**. **(B)** DockQ distribution for actives in the test set. **(C)** Among actives that were incorrectly classified by PPIscreenML as "not interacting", most were mis-docked by AF-multimer-2.2.

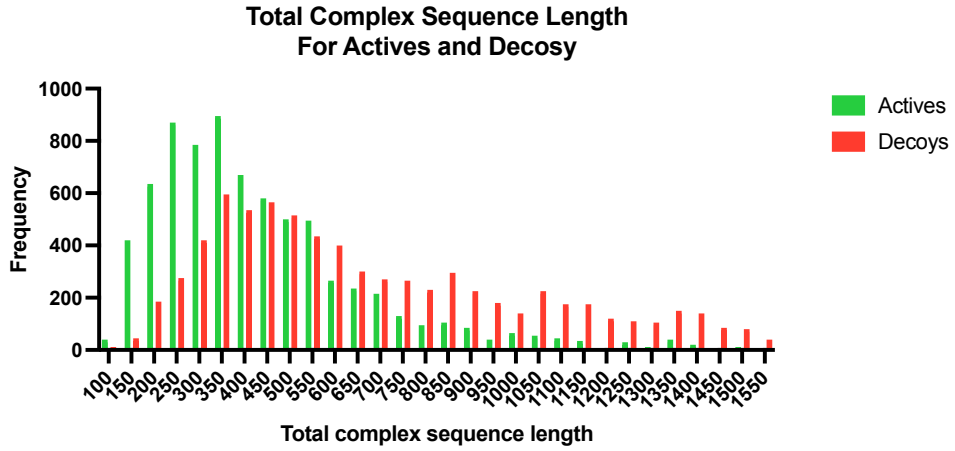

Figure S8: Overlaid histogram of total sequence length between actives and decoys. The method we describe for building compelling decoy complex does exhibit a slight bias for building decoys pairs from component proteins larger than the starting template. This arises from the use of TM-align to define structural analogs, because large query proteins are slightly more likely to yield high scores than small query proteins (for a given template protein). This artifact can allow a model to "cheat" if it includes any features that can serve as a proxy for the total number of residues in a model; accordingly, we ensured that no such features were included in developing PPIscreenML. Importantly, this artifact does not affect the composition or structural features of the generated decoy complexes, which are not systematically different from the active complexes.