PINNED: Identifying Characteristics of Druggable Human Proteins Using an Interpretable Neural Network

Michael Cunningham,[1]* Danielle Pins,[2]* Zoltán Dezső,[3] Maricel Torrent,[4] Aparna Vasanthakumar,[1] Abhishek Pandey[2]

*These authors contributed equally to this work

Author affiliations

[1]Genomics Research Center, AbbVie Inc., 1 North Waukegan Rd., North Chicago, IL 60064

[2]Information Research, AbbVie Inc., 1 North Waukegan Rd., North Chicago, IL 60064

[3]Genomics Research Center, AbbVie Inc., 1000 Gateway Boulevard, South San Francisco, CA 94080

[4]Small Molecule Therapeutics and Platform Technologies, AbbVie Inc., 1 North Waukegan Rd., North Chicago, IL 60064

Keywords

Protein, Druggability, Machine learning, Neural network, Interpretable, AlphaFold, Protein pocket, Dark genome, Proteome, Drug target

Abstract

The identification of human proteins that are amenable to pharmacologic modulation without significant off-target effects remains an important unsolved challenge. Computational methods have been devised to identify features which distinguish between “druggable” and “undruggable” proteins, finding that protein sequence, tissue and cellular localization, biological role, and position in the protein-protein interaction network are all important discriminant factors. However, many prior efforts to automate the assessment of protein druggability suffer from low performance or poor interpretability.

We developed a neural network-based machine learning model capable of generating druggability sub-scores based on each of four distinct categories, combining them to form an overall druggability score. The model achieves an excellent performance in separating drugged and undrugged proteins in the human proteome, with an area under the receiver operating characteristic (AUC) of 0.95. Our use of multiple sub-scores allows the assessment of potential protein targets of interest based on distinct contributors to druggability, leading to a more interpretable and holistic model to identify novel targets.

Introduction

The cost of developing new therapeutic drugs has risen significantly in recent years, with the average R&D (Research & Development) cost per new drug ranging between $314 million and $2.8 billion (Wouters et al., 2020). Most of this expense is incurred in the clinical phase, where trial compounds primarily fail due to a poor understanding of the disease process leading to lack of efficacy or toxicity caused either intrinsically by the actual protein being targeted or extrinsically by off-target effects on other proteins (Harrison, 2016). Determining whether a prospective target protein is "druggable," however, is a complex problem without a clearly understood solution. This can lead to a considerable amount of trial and error in the drug development process. A successful method for pre-screening prospective target proteins for druggability could save billions of dollars per year and increase the number of lifesaving drugs reaching the market.

Druggability is a poorly defined term; it can be used narrowly to refer only to a protein's ability to bind an activity-modifying small molecule ligand, or more broadly to refer to a protein's relevance as a therapeutic target in human disease. For this paper's purposes, druggability encompasses the ability of a protein's activity to be modulated for pharmacologic effect by a drug which gains regulatory approval. Undruggable proteins are those that cannot be influenced for therapeutic benefit, either because they

lack disease relevance, are biologically essential, or cannot be targeted through any known drug modality. Throughout our work we will use these definitions.

Recent advances in machine learning offer the potential for *in silico* feature identification of druggable proteins. This can facilitate computational evaluation of prospective targets prior to the initiation of expensive clinical trials. A variety of efforts in this area have taken different approaches, incorporating different predictors of druggability into their feature sets. Several groups have sought to solely use properties derived from the primary protein sequence, achieving impressive results in distinguishing drugged proteins from a select subset of difficult-to-drug proteins (Gong et al., 2021; Jamali et al., 2016; Q. Li & Lai, 2007; Lin et al., 2019; Sikander et al., 2022). However, it is unclear how these models can effectively generalize to the entire proteome. Others have analyzed the position of drugged proteins in the protein-protein interaction network to identify common features (Feng et al., 2017; Z.-C. Li et al., 2015; Mitsopoulos et al., 2015; Viacava Follis, 2021; Yao & Rzhetsky, 2008; Zhu et al., 2009). Although these models successfully extracted network properties of drugged proteins, their effectiveness is undermined by the lack of information about the properties of the proteins themselves, which may be difficult to target chemically.

Given the wide variety of features which may determine whether a protein can be categorized as druggable, it is likely that the most successful approach will incorporate a comprehensive range of properties, including physical and chemical attributes, expression profile, biological functions, and protein-protein interactions. Successful machine learning efforts in this area have utilized features from several of these domains (Bakheet & Doig, 2009; Bull & Doig, 2015; Costa et al., 2010; Ferrero et al., 2017; Yao & Rzhetsky, 2008). A 2020 study by Dezső and Ceccarelli focused exclusively on proteins that were targeted by oncology drugs, generating a feature set including a wide variety of chemical,

expression, biological function, and network properties. Using a random forest-based model, cancer drug targets were capable of being distinguished from the remainder of the proteome with an AUC of 0.89 (Dezső & Ceccarelli, 2020). We utilized this feature set and augmented it with additional protein attributes to build a classifier for property identification of drugged human proteins.

To our knowledge, all previously published machine learning models are trained to discriminate druggable from undruggable proteins with a single druggability score or binary classification. This approach lacks interpretability and wholeness, particularly when many distinct types of features are specifically and uniquely contributing to druggability. For example, a protein's position in the protein-protein interaction network may have major implications for potential off-target effects during clinical trials but does not demonstrate its structural amenability to small molecule modulation. A classifier that separates distinct features into sub-scores prior to obtaining a total druggability score could output multiple types of pertinent information about whether a protein is druggable or undruggable. We created the Predictive Interpretable Neural Network for Druggability (PINNED), a deep learning model which divides its inputs into four distinct groups—sequence and structure, localization, biological functions, and network information—and generates interpretable sub-scores that contribute to a final druggability score.

Results

Many factors influence a protein's druggability, including its effectiveness as a disease-modifying target and its propensity for causing undesired side-effects. A protein's physical and chemical properties, such as amino acid composition, secondary structure, post-translational modification, and others, can determine whether it can be readily liganded by a drug-like molecule. Its position in the complex network of protein-protein interactions which occur within the human body can influence its role in

disease and its potential for off-target effects. The biological function of a protein plays a significant role in whether it is a useful drug target; however, many proteins are involved in multiple different processes, disturbance of any of which can lead to unanticipated consequences for homeostasis and thus to off-target effects. Additionally, a protein's expression profile across target and non-target tissues can have implications for its efficacy and safety.

To incorporate all these contributions to druggability, we generated a feature set that contains a variety of data for 20,404 human proteins, including properties extracted from the protein sequence, tissue specificity, subcellular localization, biological functions, and position in the protein-protein interaction network (Dezső & Ceccarelli, 2020). The features were divided into four feature groups: sequence and structure, localization, biological functions, and network information. Each category was then augmented with additional features obtained from the protein sequence, Gene Ontology (GO) knowledgebase (Ashburner et al., 2000), and the protein's 3-dimensional structure as estimated by the artificial intelligence system AlphaFold (Jumper et al., 2021) (Table 1).

Sequence and structure	**Localization**	**Biological functions**	**Network information**
52 physiochemical features	Predicted subcellular localization	Enzyme classification	Signaling maps
Grouped Dipeptide Composition (GDPC)	Tissue specificity	Essentiality of mouse homolog	Network features
Pseudo Amino Acid Composition (PAAC)	Gene Ontology (GO) cellular components	Biological processes (MetaCore)	
fpocket data from AlphaFold models		Molecular functions (MetaCore)	
		Biological processes (Gene Ontology)	
		Gene Ontology (GO) molecular functions	

Table 1. All features used to train the model, divided into the four feature groups

Sequence and structure properties

Sequence and structure properties included information about 52 physiochemical features, such as protein molecular weight and amino acid residues, charge and isoelectric points, extinction coefficients, predicted post-translational modifications, secondary structure, and solvent accessibility. Previous works indicate that the grouped dipeptide composition (GDPC) and pseudo amino acid composition (PAAC) of a protein may be useful characteristics in determining its druggability (Gong et al., 2021; Lin et al., 2019; Sikander et al., 2022). GDPC represents the relative composition of all the amino acid 2-mers in a protein's sequence, with the 20 amino acids being reduced to five groups according to their physical properties. PAAC is an algorithm designed to reduce the sequence characteristics of a protein to a defined-length vector while incorporating information about their sequence order (Chou, 2001). GDPC and PAAC were generated for each of the proteins in our dataset and included in the sequence and structural properties.

AlphaFold is a deep learning network developed by DeepMind that can predict a protein's structure from its three-dimensional amino acid sequence. The AlphaFold Protein Structure Database was established between AlphaFold and EMBL-EBI (David et al., 2022). This database contains the predicted protein structure models of nearly the full UniProt human proteome. It is available as an open-source database. Fpocket is an open-source software package able to automatically detect and provide pocket descriptors in a protein's 3-dimensional structure (Le Guilloux et al., 2009). It enables the identification of potential drug binding sites and provides relevant properties based on each pocket detected. The pockets are ranked according to their ability to bind to small molecules as a cavity prediction algorithm. Fpocket was utilized to identify druggable and undruggable protein cavities based on the trajectories produced by the simulation. AlphaFold models of each protein were collected from the AlphaFold database and pocket information was generated using Fpocket.

Localization

The Subcellular Localization Predictive System (CELLO) was used to predict subcellular localization for each protein in the dataset (Yu et al., 2004). We included this prediction, in addition to tissue specificity data obtained from the Genotype-Tissue Expression (GTEx) and the Human Protein Atlas (HPA) (GTEx Consortium, 2017; Uhlén et al., 2015). The GO Knowledgebase was used to retrieve Cellular Component annotations for each protein. These labels are manually assigned based on published literature and represent the cellular structures in which the protein performs its functions.

Biological functions

Gene essentiality, assessed by lethality of mouse homozygous loss-of-function mutations (Georgi et al., 2013) and enzyme classifications obtained from the Swiss-prot database (Bairoch & Boeckmann, 1991), were included in the biological functions score. Scores were generated by Dezső et al. for each gene ontology in the MetaCore database based on their 102-protein target enrichment set of cancer drugs. The highest three ontology scores in the categories— "Biological Functions," "Molecular Process," and "Maps" (signaling pathways)—were included in that protein's feature set. "Biological Functions" and "Molecular Process" were used as inputs to the "biological functions" sub-score, while "Maps" was included in the "network information" sub-score (see below). It should be noted that the Biological Functions score generated by Dezső et al. represents only one feature input into the biological functions network.

Network information

The signaling pathways ("Maps") score generated by Dezső et al. was included in the network information features. Degree, closeness, betweenness, eigen centrality, and PageRank of each protein in

the protein-protein interaction network were calculated using information from the STRING database (Szklarczyk et al., 2019). These features were incorporated into the network information input.

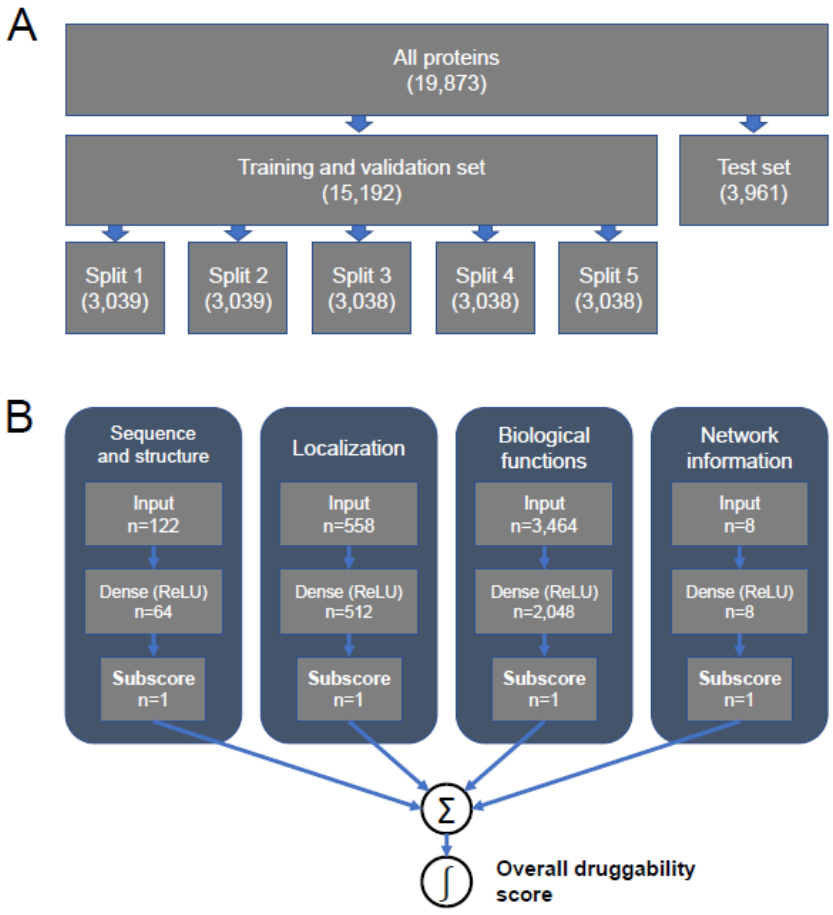

Figure 1. Design of the PINNED model and dataset. **A** Division of the data into training, validation, and test sets. **B** PINNED architecture including the four constituent subnetworks

Protein set

The National Center for Biotechnology Information (NCBI) Pharos database, a data repository of human protein properties and drugged status, identifies proteins as confirmed drug targets if they are "protein drug targets via which approved drugs act" ("Tclin") (Sheils et al., 2021). As of October 2022, 704 of the 20,412 proteins in Pharos are categorized as Tclin.

All other proteins are classified as one of three other categories: undrugged proteins which bind small molecules with high potency ("Tchem"), proteins with well-studied biology ("Tbio"), and proteins not meeting the criteria for any of the other categories ("Tdark") (Sheils et al., 2021). Of these proteins, 19,873 were represented in both Dezso et al.'s dataset and the AlphaFold database, including 696 of the 704 Tclin proteins in Pharos. We used the 696 Tclin proteins as our positive "drugged" set, and the remaining 19,177 proteins in the other categories as our negative "undrugged" set (Fig. 1A). It is likely that the undrugged set contains many potentially druggable proteins which have not yet been targeted by approved therapeutics.

PINNED model

The model architecture consisted of four separate deep neural networks, designated "sequence and structure," "localization," "biological functions," and "network information." Each network contained an input layer, a hidden layer with ReLU activation, and a single output neuron representing the network sub-score. The four sub-scores were summed, producing a logit which was passed through a sigmoid function to generate the final probability of druggability (Fig. 1B).

Prior to model tuning, 20% of the dataset was held out to form a separate test set, which was used to evaluate the model after the optimal architecture had been determined. The remaining data was divided into five equal groups, one of which was held out as a validation set, while the remaining four were combined to form the training set (5-fold cross-validation) (Fig. 1A). It was necessary to oversample the positive set to prevent the model from converging towards a naïve negative classifier due to the significant imbalance between drugged and undrugged proteins. Within the training set, drugged proteins were separated from the validation set, then randomly oversampled with replacement until the number of drugged and undrugged proteins was equal. The feature matrix was then divided

into sequence and structure, localization, biological functions, and network information matrices. These matrices served as inputs to their respective networks.

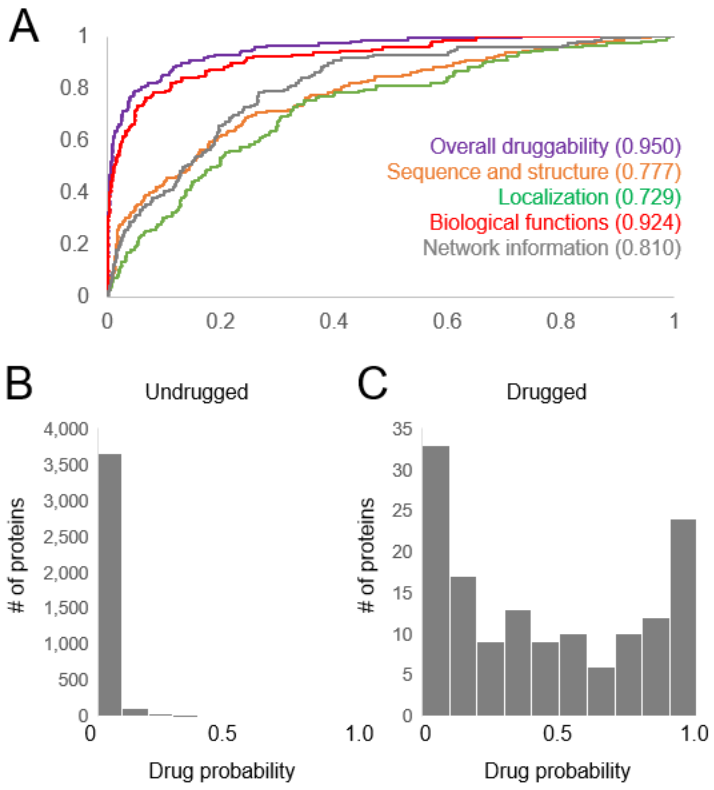

Figure 2. Performance of PINNED on the test set. **A** AUC curve of the model and each subnetwork for distinguishing between drugged and undrugged proteins. **B** Histogram showing the distribution of druggability probabilities for undrugged proteins in the test set. **C** Histogram showing the distribution of druggability probabilities for drugged proteins in the test set

After hyperparameter optimization, a model was trained on the full training/validation set, with the held-out test data used as the final validation set. The complete model achieved an excellent AUC of 0.950 on the test set (Fig. 2A), with the scores from each subnetwork attaining a lower AUC. Although the biological functions sub-score performed by far the best at an AUC of 0.924, the other networks still successfully classified proteins as drugged or undrugged with reasonable discriminatory power. The full

model consistently scored undrugged proteins in the test set as having low druggability due to the substantial number of negative examples (undrugged proteins) to learn from (Fig. 2B). Druggability scores were more variable for the drugged proteins, reflecting the difficulty of identifying a consistent "druggable" profile from a small number of positives (Fig. 2C). However, PINNED's high AUC demonstrates its ability to successfully distinguish between proteins with high and low druggability potential.

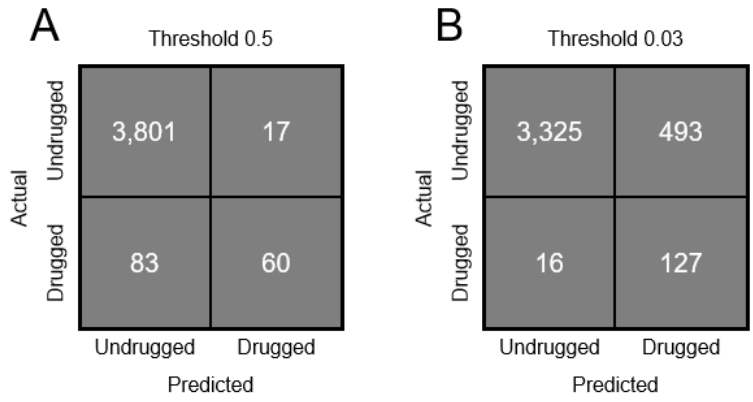

Figure 3. Confusion matrices of PINNED on the test set. **A** Confusion matrix with threshold for druggability set at 0.5. **B** Confusion matrix with threshold set at 0.03 to balance sensitivity and specificity

Reducing the druggability score required to consider a protein "druggable" can increase the sensitivity of the predictor. By default, this value was set as 0.5 during training, but may be changed to any arbitrary value during inference. At a threshold of 0.5, PINNED achieves excellent specificity but low sensitivity, with many drugged proteins in the test set being mistakenly classed as undruggable (Fig. 3A, Table 2). At a reduced threshold of 0.03, chosen to balance sensitivity and specificity, all the drugged proteins are properly classed, while many undrugged proteins are now considered "druggable" (Fig. 3B, Table 2). This cohort of undrugged proteins with high druggability scores represents potential opportunities for pharmaceutical targeting.

	Sensitivity	Specificity	Accuracy	AUC
0.5 threshold	0.420	0.996	0.975	0.950
0.03 threshold	0.888	0.871	0.871	

Table 2. Comparison of PINNED's test performance at different druggability thresholds. A threshold of 0.5 is used for training, while requiring a lower score of 0.03 allows a closer balance between sensitivity and specificity

Comparing PINNED's performance to prior machine learning efforts to assess protein druggability is challenging due to the wide variety of datasets used and metrics reported. Many previous works exclude proteins with significant homology to drugged proteins from their undrugged sets (Bakheet & Doig, 2009; Feng et al., 2017), even though there may be significant differences between these proteins' properties which alter their utility as drug targets. Similarly, some construct an idealized set of "undruggable" proteins, making it difficult to generalize to the whole proteome (Charoenkwan et al., 2022; Gong et al., 2021; Jamali et al., 2016; Q. Li & Lai, 2007; Lin et al., 2019; Sikander et al., 2022; Sun et al., 2018; Zhu et al., 2009). Others only focus on a specific target or indication, such as oncology (Bazaga et al., 2020; de Falco et al., 2021; Dezső & Ceccarelli, 2020; Jeon et al., 2014), or ion channels (Huang et al., 2010). Restricting our focus to models which seek to assess the druggability of the entire proteome, we find that PINNED comfortably outperforms much of the prior literature in sensitivity, specificity, and AUC (Bull & Doig, 2015; Costa et al., 2010; Ferrero et al., 2017; Yao & Rzhetsky, 2008) (Additional file 1). A recent publication by Raies et al. achieved a higher AUC, but without the constituent sub-scores PINNED generates (Raies et al., 2022). The interpretability of our model is a unique advantage which enhances its value to the target selection process.

	Feature	Category	Change in test loss
1	MetaCore Molecular Function 3	Biological functions	0.00560
2	Enzyme classification—non-enzyme	Biological functions	0.00292
3	Essentiality—unknown	Biological functions	0.00277
4	Degree (STRING interactions)	Network information	0.00232
5	Mitochondrial respiratory chain complex I assembly	Biological functions	0.00208
6	ATP binding	Biological functions	0.00181
7	Transmembrane helices	Sequence and structure	0.00171
8	Voltage-gated potassium channel activity	Biological functions	0.00133
9	PageRank (STRING interactions)	Network information	0.00125
10	Potassium ion transmembrane transport	Biological functions	0.00121

Table 3. Most notable features, as ranked by change in test loss after random permutation of the feature

After training the model and assessing it on the test set, we ablated each feature by randomly shuffling (permuting) the values among the protein test set and assessed the increase in test loss induced by the change. As loss is inversely related to the network's performance, more prominent features will result in a higher increase in loss after being permuted. We found that features belonging to the biological function's subnetwork comprised seven of the top 10 (Table 3), consistent both with the substantial number of features in that network and the fact that it was by far the most significant in contributing to PINNED's performance. Many of the features, including essentiality, degree, transmembrane helices, and PageRank, overlapped with the most notable features selected by Dezső et al. 2020. This indicates a similarity between the properties of oncology targets and other drugged proteins. Additionally, several of the top features derived from GO annotations—ATP binding, voltage-gated potassium channel activity, and potassium ion transmembrane transport—are known to be relevant factors in druggability (Chène, 2002; Wulff et al., 2009).

	UniProt ID	Gene name	Protein	Pharos class	Score
1	P21917	DRD4	D(4) dopamine receptor	Tchem	0.9994
2	P50052	AGTR2	Type-2 angiotensin II receptor	Tchem	0.9991
3	Q9Y6Q6	TNFRSF11A	Tumor necrosis factor receptor superfamily member 11A	Tbio	0.9970
4	P34972	CNR2	Cannabinoid receptor 2	Tchem	0.9969
5	P33032	MC5R	Melanocortin receptor 5	Tchem	0.9967
6	P32241	VIPR1	Vasoactive intestinal polypeptide receptor 1	Tchem	0.9953
7	Q9HCR9	PDE11A	Dual 3',5'-cyclic-AMP and -GMP phosphodiesterase 11A	Tchem	0.9953
8	P21918	DRD5	D(1B) dopamine receptor	Tchem	0.9952
9	P23416	GLRA2	Glycine receptor subunit alpha-2	Tchem	0.9927
10	P41968	MC3R	Melanocortin receptor 3	Tchem	0.9925

Table 4. Highest scoring undrugged proteins

To generate druggability scores for the entire proteome, we split our entire dataset, including the training/validation and test sets into five parts. Each part was held out and the remaining four were used to train a classifier model. The scores for the held-out set were designated as the final druggability scores for the protein set. This process was repeated with each of the sets being held out once to generate scores for the proteins in the entire proteome. Of the 10 highest-scoring undrugged proteins in the proteome, all except TNFRSF11A are listed by Pharos as Tchem, having validated high-potency small molecule ligands (Table 4). The mechanism of action for many drugs is not entirely clear, as they may interact with multiple proteins in the same family, making conclusive classification of proteins as targets or non-targets challenging. We cross-referenced all top 10 scoring proteins with the Therapeutic Targets Database (TTD) and Open Targets, two other curated databases of drug-target interactions (Koscielny et

al., 2017; Zhou et al., 2022). Of these, five were listed by TTD and two by Open Targets as already being the targets of approved therapeutics, while two were listed by TTD and two by Open Targets as clinical trial targets (Additional file 2). This discrepancy between databases reflects the difficulty of conclusively classifying proteins as mechanism of action drug targets. However, the high prevalence of likely interactors of approved drugs demonstrates that PINNED successfully generalizes the properties of drugged proteins to previously unseen data.

	UniProt ID	**Gene name**	**Protein**	**Score**
1	Q8TAA3	PSMA8	Proteasome subunit alpha-type 8	0.9277
2	A6NHL2	TUBAL3	Tubulin alpha chain-like 3	0.7725
3	P01880	IGHD	Immunoglobulin heavy constant delta	0.6276
4	Q86T26	TMPRSS11B	Transmembrane protease serine 11B	0.6195
5	Q5TAH2	SLC9C2	Sodium/hydrogen exchanger 11	0.6038
6	Q9Y2U2	KCNK7	Potassium channel subfamily K member 7	0.5257
7	A6NNS2	DHRS7C	Dehydrogenase/reductase SDR family member 7C	0.4923
8	P0DPH8	TUBA3D	Tubulin alpha-3D chain	0.4891
9	Q5I0G3	MDH1B	Putative malate dehydrogenase 1B	0.4547
10	P01780	IGHV3-7	Immunoglobulin heavy variable 3-7	0.4105

Table 5. Highest scoring Tdark proteins

Of the 20,412 proteins in the Pharos database, 5,679 (28%) are designated as “Tdark”—having extremely limited data about their properties and functions. Considerable interest exists in exploring these understudied parts of the genome, particularly to discover novel therapeutic targets which have previously been overlooked (Oprea, 2019). At least one of the top scoring Tdark proteins in our model has been investigated as a drug target (Table 5). Transmembrane protease serine 11B (TMPRSS11B) was identified as upregulated in lung squamous cell carcinomas, serving as a poor prognostic marker. Inhibition of the protein *in vitro* reduced transformation and proliferation (Updegraff et al., 2018).

TMPRSS11B's sub-scores for sequence and structure, localization, biological functions, and network information, compared to the Tclin (drugged) proteins, were respectively in the 84th, 97th, 29th, and 1st percentiles (Additional file 3). The high score for sequence and structure is consistent with the observation that transmembrane helices are highly indicative of druggability (Table 3). Similarly, for the localization subnetwork, permutation importance suggests three of the five most notable features are GO annotations related to localization to the plasma membrane (Additional file 4). Although TMPRSS11B attains a lower score in the biological functions network, it is higher than 95% of undrugged proteins. Its network information score, however, is low even among undrugged proteins, at the 7th percentile. This may indicate that TMPRSS11B lacks the network centrality to have a significant impact on cellular homeostasis. Overall, our results indicate that TMPRSS11B may be structurally amenable to drugging and demonstrates localization and biological activity consistent with other drug targets but could suffer from difficulties with efficacy in clinical trials. The use of multiple sub-scores to characterize a protein's druggability profile enables a more detailed analysis of its potential strengths and weaknesses rather than a single unified score.

Discussion

The implementation of a pre-screening methodology that differentiates druggable and undruggable targets can help ameliorate the difficulty of target selection in pharmaceutical development and aid in allocating R&D investments to promising targetable proteins. Consequently, it is imperative that an interpretable model can accurately identify novel druggable targets. We developed a neural network-based machine learning model able to produce druggability sub-scores based on separate feature categories spanning multiple factors in druggability. These allow the analysis of each category individually and its contribution to an overall druggability score.

PINNED attained excellent results in its ability to distinguish drugged from undrugged proteins with an AUC of 0.95. Importantly, this was achieved on the entire proteome, indicating that the model can handle cases generated by family members of drugged proteins. Notably, PINNED was far better at assigning low druggability scores to undrugged proteins than assigning high scores to drugged proteins (Fig. 2), consistent with the large imbalance between the two classes. By reducing the score required to designate a protein as "druggable," it is possible to increase the sensitivity of the classifier in positively labeling drugged proteins, at the expense of also designating as druggable many currently undrugged proteins (Fig. 3). However, these may represent proteins which are already the targets of approved drugs but have not been formally labeled due to insufficient evidence, or potential new targets which merit further investigation (Table 4).

Among our sub-scores, the biological functions network achieved the best performance with a standalone AUC of 0.924. This is potentially due to it being the largest subnetwork, with 3,464 inputs, allowing it to incorporate a large amount of information about protein function. The network information sub-score attained the second-highest performance at 0.810, despite being by far the smallest network, suggesting that the relationship between number of inputs and classification value is complex. Sequence and structure were the lowest-performing subnetworks, achieving AUCs of 0.777 and 0.729. However, these scores were still competitive with previous efforts at using machine learning to assess protein druggability (Additional file 1). This result indicates that our druggability sub-scores are useful not just as inputs to the overall score, but as standalone estimates of each protein's druggability within that subdomain.

The 10 most relevant features fed into PINNED, in terms of impact on accuracy, span three of the four subnetworks, with the majority coming from biological functions, but none from localization (Table 3).

While this finding is consistent with the fact that the localization subnetwork achieves the lowest standalone AUC, the "transmembrane helices" feature in the sequence and structure network can be assumed to be a strong indicator of whether a protein is localized to the plasma membrane, which dominates the most important localization features (Table S3). Some collinearity exists between the feature inputs between the different networks. This is an inevitable result of the proteins' functions, structures, and interactions being closely interrelated. However, the observation that many proteins score highly on some subnetworks but poorly on others demonstrates that they capture distinct information about a protein's druggability. Many of the top features overlap with those identified in previous publications (Bull & Doig, 2015; de Falco et al., 2021; Dezső & Ceccarelli, 2020; Kim et al., 2017). This suggests that machine learning models trained to predict protein druggability converge on a common set of important contributors.

The "dark genome" encompasses the proteins in the human proteome which have not been extensively studied, especially as prospective drug targets, and has thus become of particular interest to the pharmaceutical industry (Oprea, 2019). Our work indicates that a substantial number of proteins in the dark genome may have drug-like properties. For instance, we found transmembrane serine protease TMPRSS11B, a dark genome protein, is similar in structure, localization, and function to many successfully drugged targets. Our model enables dark genome proteins with disease associations to be investigated for druggability potential.

Conclusions

We established a neural network-based machine learning model, termed PINNED, able to assess proteins' druggability based on their sub-scores across four distinct categories. We have demonstrated that our proposed methodology is a highly predictive network (test AUC 0.95) with the ability to

estimate the druggability of over 20,000 proteins spanning the entire human proteome. PINNED can be used as a pre-screening tool to determine a protein's amenability to drugging prior to the initiation of pre-clinical programs and identify weaknesses in the form of low sub-scores of top targets that do not necessarily score high in all four areas, providing room for insight and early remediation. This methodology enables the exploration of novel targets cost-effectively while improving the clinical phase success rate.

Materials and methods

Drug targets

Drugged and undrugged proteins and sequences were obtained from the Pharos database on October 12, 2022. Proteins categorized as Tclin were labeled as drugged, while proteins categorized as Tchem, Tbio, or Tdark were labeled as undrugged. Protein features were obtained from Dezső et al.'s features (Dezső & Ceccarelli, 2020) and the AlphaFold database (David et al., 2022). A protein list was generated from the intersection of these three databases. Proteins not found in all the databases were removed, leaving the final protein set used to train the model as the intersection of the three sets.

All features generated by Dezső et al. were incorporated into our feature set and divided between the four subnetworks. These include characteristics calculated or predicted from the amino acid sequence, such as posttranslational modifications, enzyme classification, localization, secondary structure, and sequence motifs. Details on the generation of these features can be found in Dezső et al. 2020. All numeric features were standardized to a mean of 0 and standard deviation of 1 ("standard scaled"), while all categorical features were one-hot encoded.

Sequence and structure properties sub-score

Information about protein molecular weight and amino acid residues, charge and isoelectric points, extinction coefficients, predicted post-translational modifications, secondary structure, and solvent accessibility from Dezső et al.'s feature set were included as sequence and structure properties.

Grouped dipeptide composition (GDPC) and pseudo amino acid composition (PAAC) were calculated using the iFeature toolkit (Chen et al., 2018). All selenocysteine (U) residues in the protein sequences were converted to cysteine (C) for the calculations. A lambda of 3 was chosen for PAAC.

Human protein structure predictions were acquired from AlphaFold (last modification on 05/05/2022). The structures were curated to run through Fpocket. Fpocket is an open-source protein prediction algorithm based on the Voronoi tessellation and the alpha sphere theory (Le Guilloux et al., 2009). Fpocket begins by filtering the vertices and finding the correlated alpha spheres dependent on their minimum and maximum size. Alpha spheres that are clustered together equate to a recognized pocket. The pockets are further reduced based on the zones of compacted atom packing. The alpha spheres are labeled based on their contact to atoms, then ranked based on their prospective binding capabilities towards small molecules. All features were standard scaled.

Localization sub-score

Protein localization and tissue specificity data obtained from Dezső et al. was included in the localization data.

GO terms were downloaded from the Target Central Resource Database (TCRD) on July 29, 2022, and separated into GO terms categorized as Components, Functions, or Processes. They were used to

generate a one-hot encoded GO terms matrix that mapped each protein in the dataset. Terms mapped to less than 10 proteins were excluded. GO Components were included in the localization data, while Functions and Processes were included in the biological functions data (see below).

Biological functions sub-score

Scores generated for each protein by Dezső et al. from the MetaCore database for "Biological Function," and "Molecular Process" were standard-scaled and included in the "biological functions" sub-score. The enzyme classification and gene essentiality feature from Dezső et al. were included in the biological functions data.

GO Functions and Processes were obtained and processed as described above and included in biological functions.

Network information sub-score

The "Maps" (signaling pathways) scores from Dezső et al. and calculated protein-protein interaction network features were used as the input to the network information subnetwork.

Model

Features for all four sub-scores were combined into a single feature matrix. 20% of the proteins were selected at random prior to model development and held out as a test set. Prior to training, the drugged proteins in the training set were randomly oversampled with replacement until the quantity was equal to the quantity of undrugged proteins. Oversampling by SMOTE, ADASYN, or applying different weights to positive and negative samples were evaluated, but performance was not improved.

Our model was implemented in Python 3.7.13 using TensorFlow 2.11.0 and consisted of four densely connected neural networks, corresponding to the four sub-scores. Each consisted of a single input layer of size n inputs, a hidden layer with size 2^i, where i is the largest integer such that $2^i \leq n$, and an output layer of size 1, representing that network's sub-score. ReLU activation was applied to the hidden layers, and an L2 penalty of 0.001 was applied to both the hidden and output layers. The four subnetwork output layers were summed to generate the logits of the overall druggability score. Different numbers of hidden layers, dropout for the input and hidden layers, learning rates, and L2 coefficients were tested, and the above values were found to lead to optimal AUC scores on validation sets.

Support vector machine, logistic regression, XGBoost, and random forest models were also evaluated and found to deliver performance comparable or inferior to neural network.

The model was trained using the Adam optimizer with TensorFlow default parameters at a learning rate of $10^{-3.5}$, with a batch size of 32 and the binary cross entropy loss function.

Abbreviations

ADASYN: Adaptive Synthetic

AUC: Area Under the Receiver Operating Characteristic

CELLO: Cellular Localization

GDPC: Grouped Dipeptide Composition

GO: Gene Ontology

GTEx: Genotype-Tissue Expression

HPA: Human Protein Atlas

NCBI: National Center for Biotechnology Information

PAAC: Pseudo Amino Acid Composition

PINNED: Predictive Interpretable Neural Network for Druggability

R&D: Research and Development

ReLU: Rectified Linear Unit

SMOTE: Synthetic Minority Over-sampling Technique

STRING: Search Tool for the Retrieval of Interacting Genes/Proteins

TCRD: Target Central Resource Database

TMPRSS11B: Transmembrane Serine Protease 11B

TTD: Therapeutic Target Database

Declarations

Availability of data and materials

Our code is available at https://github.com/abbvie-external/Predictive-Interpretable-Neural-Network-for-Druggability-PINNED-

Competing interests

The authors declare no competing interests.

Funding

All funding was provided by AbbVie, Inc.

Authors' contributions

MC developed and tested the machine learning model. DP generated the fpocket data. MC and DP wrote the manuscript. AP supervised the project. ZD generated protein features and developed an

earlier model which was the inspiration for this work. All authors reviewed the manuscript and gave feedback.

Acknowledgements

The authors would like to thank Keith Kelleher of the NIH, as well as Maria Argiriadi, Marlon Cowart, Felix DeAnda, Jacob Degner, Phil Hajduk, Howard Jacob, Jozsef Karman, Xavier Langlois, Frank Oellien, Ahmad Sheikh, and Jeff Waring of AbbVie, Inc. for their assistance with this project.

Funding and authors' information

All authors are employees of AbbVie. The design, study conduct, and financial support for this research were provided by AbbVie. AbbVie participated in the interpretation of data, review, and approval of the publication.

Ethical approval

Not applicable.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Bairoch, A., & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic Acids Research*, *19*(suppl), 2247–2249. https://doi.org/10.1093/nar/19.suppl.2247

Bakheet, T. M., & Doig, A. J. (2009). Properties and identification of human protein drug targets. *Bioinformatics*, *25*(4), 451–457.

Bazaga, A., Leggate, D., & Weisser, H. (2020). Genome-wide investigation of gene-cancer associations for the prediction of novel therapeutic targets in oncology. *Scientific Reports*, *10*(1), 10787. https://doi.org/10.1038/s41598-020-67846-1

Bull, S. C., & Doig, A. J. (2015). Properties of Protein Drug Target Classes. *PLOS ONE*, *10*(3), e0117955. https://doi.org/10.1371/journal.pone.0117955

Charoenkwan, P., Schaduangrat, N., Lio', P., Moni, M. A., Shoombuatong, W., & Manavalan, B. (2022). Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework. *IScience*, *25*(9), 104883. https://doi.org/10.1016/j.isci.2022.104883

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Webb, G. I., Smith, A. I., Daly, R. J., Chou, K.-C., & Song, J. (2018). iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, *34*(14), 2499–2502. https://doi.org/10.1093/bioinformatics/bty140

Chène, P. (2002). ATPases as drug targets: learning from their structure. *Nature Reviews Drug Discovery*, *1*(9), 665–673. https://doi.org/10.1038/nrd894

Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Genetics*, *43*(3), 246–255. https://doi.org/10.1002/prot.1035

Costa, P. R., Acencio, M. L., & Lemke, N. (2010). A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics*, *11*(S5). https://doi.org/10.1186/1471-2164-11-S5-S9

David, A., Islam, S., Tankhilevich, E., & Sternberg, M. J. E. (2022). The AlphaFold Database of Protein Structures: A Biologist's Guide. *Journal of Molecular Biology*, *434*(2), 167336. https://doi.org/10.1016/j.jmb.2021.167336

de Falco, A., Dezso, Z., Ceccarelli, F., Cerulo, L., Ciaramella, A., & Ceccarelli, M. (2021). Adaptive one-class Gaussian processes allow accurate prioritization of oncology drug targets. *Bioinformatics*, *37*(10), 1420–1427. https://doi.org/10.1093/bioinformatics/btaa968

Dezső, Z., & Ceccarelli, M. (2020). Machine learning prediction of oncology drug targets based on protein and network properties. *BMC Bioinformatics*, *21*(1), 104. https://doi.org/10.1186/s12859-020-3442-9

Feng, Y., Wang, Q., & Wang, T. (2017). Drug Target Protein-Protein Interaction Networks: A Systematic Perspective. *BioMed Research International*, *2017*, 1–13. https://doi.org/10.1155/2017/1289259

Ferrero, E., Dunham, I., & Sanseau, P. (2017). In silico prediction of novel therapeutic targets using gene–disease association data. *Journal of Translational Medicine*, *15*(1). https://doi.org/10.1186/s12967-017-1285-6

Georgi, B., Voight, B. F., & Bućan, M. (2013). From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genetics*, *9*(5), e1003484. https://doi.org/10.1371/journal.pgen.1003484

Gong, Y., Liao, B., Wang, P., & Zou, Q. (2021). DrugHybrid_BS: Using Hybrid Feature Combined With Bagging-SVM to Predict Potentially Druggable Proteins. *Frontiers in Pharmacology*, *12*, 771808. https://doi.org/10.3389/fphar.2021.771808

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), 204–213. https://doi.org/10.1038/nature24277

Harrison, R. K. (2016). Phase II and phase III failures: 2013–2015. *Nature Reviews Drug Discovery*, *15*(12), 817–818. https://doi.org/10.1038/nrd.2016.184

Huang, C., Zhang, R., Chen, Z., Jiang, Y., Shang, Z., Sun, P., Zhang, X., & Li, X. (2010). Predict potential drug targets from the ion channel proteins based on SVM. *Journal of Theoretical Biology*, *262*(4), 750–756. https://doi.org/10.1016/j.jtbi.2009.11.002

Jamali, A. A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., & Ebrahimie, E. (2016). DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. *Drug Discovery Today*, *21*(5), 718–724. https://doi.org/10.1016/j.drudis.2016.01.007

Jeon, J., Nim, S., Teyra, J., Datti, A., Wrana, J. L., Sidhu, S. S., Moffat, J., & Kim, P. M. (2014). A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, *6*(7), 57. https://doi.org/10.1186/s13073-014-0057-7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kim, B., Jo, J., Han, J., Park, C., & Lee, H. (2017). In silico re-identification of properties of drug target proteins. *BMC Bioinformatics*, *18*(S7). https://doi.org/10.1186/s12859-017-1639-3

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., Hasan, S., Karamanis, N., Maguire, M., Papa, E., Pierleoni, A., Pignatelli, M., Platt, T., Rowland, F., Wankar, P., Bento, A. P., Burdett, T., Fabregat, A., Forbes, S., … Dunham, I. (2017). Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Research*, *45*(D1), D985–D994. https://doi.org/10.1093/nar/gkw1055

Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, *10*(1), 168. https://doi.org/10.1186/1471-2105-10-168

Li, Q., & Lai, L. (2007). Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics*, *8*(1), 353. https://doi.org/10.1186/1471-2105-8-353

Li, Z.-C., Zhong, W.-Q., Liu, Z.-Q., Huang, M.-H., Xie, Y., Dai, Z., & Zou, X.-Y. (2015). Large-scale identification of potential drug targets based on the topological features of human protein–protein interaction network. *Analytica Chimica Acta*, *871*, 18–27. https://doi.org/10.1016/j.aca.2015.02.032

Lin, J., Chen, H., Li, S., Liu, Y., Li, X., & Yu, B. (2019). Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier. *Artificial Intelligence in Medicine*, *98*, 35–47. https://doi.org/10.1016/j.artmed.2019.07.005

Mitsopoulos, C., Schierz, A. C., Workman, P., & Al-Lazikani, B. (2015). Distinctive Behaviors of Druggable Proteins in Cellular Networks. *PLOS Computational Biology*, *11*(12), e1004597. https://doi.org/10.1371/journal.pcbi.1004597

Oprea, T. I. (2019). Exploring the dark genome: implications for precision medicine. *Mammalian Genome*, *30*(7–8), 192–200. https://doi.org/10.1007/s00335-019-09809-0

Raies, A., Tulodziecka, E., Stainer, J., Middleton, L., Dhindsa, R. S., Hill, P., Engkvist, O., Harper, A. R., Petrovski, S., & Vitsios, D. (2022). DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-04245-4

Sheils, T. K., Mathias, S. L., Kelleher, K. J., Siramshetty, V. B., Nguyen, D.-T., Bologa, C. G., Jensen, L. J., Vidović, D., Koleti, A., Schürer, S. C., Waller, A., Yang, J. J., Holmes, J., Bocci, G., Southall, N., Dharkar, P., Mathé, E., Simeonov, A., & Oprea, T. I. (2021). TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Research*, *49*(D1), D1334–D1346. https://doi.org/10.1093/nar/gkaa993

Sikander, R., Ghulam, A., & Ali, F. (2022). XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. *Scientific Reports*, *12*(1), 5505. https://doi.org/10.1038/s41598-022-09484-3

Sun, T., Lai, L., & Pei, J. (2018). Analysis of protein features and machine learning algorithms for prediction of druggable proteins. *Quantitative Biology*, *6*(4), 334–343. https://doi.org/10.1007/s40484-018-0157-2

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., … Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220), 1260419. https://doi.org/10.1126/science.1260419

Updegraff, B. L., Zhou, X., Guo, Y., Padanad, M. S., Chen, P.-H., Yang, C., Sudderth, J., Rodriguez-Tirado, C., Girard, L., Minna, J. D., Mishra, P., DeBerardinis, R. J., & O'Donnell, K. A. (2018). Transmembrane Protease TMPRSS11B Promotes Lung Cancer Growth by Enhancing Lactate Export and Glycolytic Metabolism. *Cell Reports*, *25*(8), 2223-2233.e6. https://doi.org/10.1016/j.celrep.2018.10.100

Viacava Follis, A. (2021). Centrality of drug targets in protein networks. *BMC Bioinformatics*, *22*(1), 527. https://doi.org/10.1186/s12859-021-04342-x

Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, *323*(9), 844. https://doi.org/10.1001/jama.2020.1166

Wulff, H., Castle, N. A., & Pardo, L. A. (2009). Voltage-gated potassium channels as therapeutic targets. *Nature Reviews Drug Discovery*, *8*(12), 982–1001. https://doi.org/10.1038/nrd2983

Yao, L., & Rzhetsky, A. (2008). Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Research*, *18*(2), 206–213. https://doi.org/10.1101/gr.6888208

Yu, C.-S., Lin, C.-J., & Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n* -peptide compositions. *Protein Science*, *13*(5), 1402–1406. https://doi.org/10.1110/ps.03479604

Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., Qiu, Y., & Chen, Y. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research*, *50*(D1), D1398–D1407. https://doi.org/10.1093/nar/gkab953

Zhu, M., Gao, L., Li, X., Liu, Z., Xu, C., Yan, Y., Walker, E., Jiang, W., Su, B., Chen, X., & Lin, H. (2009). The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network. *Journal of Drug Targeting*, *17*(7), 524–532. https://doi.org/10.1080/10611860903046610

Supplementary Information

Additional file 1. Comparison to previous druggability classifiers. **Additional file 2.** Target classification of highest scoring undrugged proteins. **Additional file 3.** All protein scores. **Additional file 4**. Feature importance scores.