*https://doi.org/10.1093/nar/gkac335*

# GenePlexus: a web-server for gene discovery using network-based machine learning

**Christopher A. Mancuso[1], Patrick S. Bills[2], Douglas Krum[2], Jacob Newsted[2], Renming Liu[1] and Arjun Krishnan[1,3,*]**

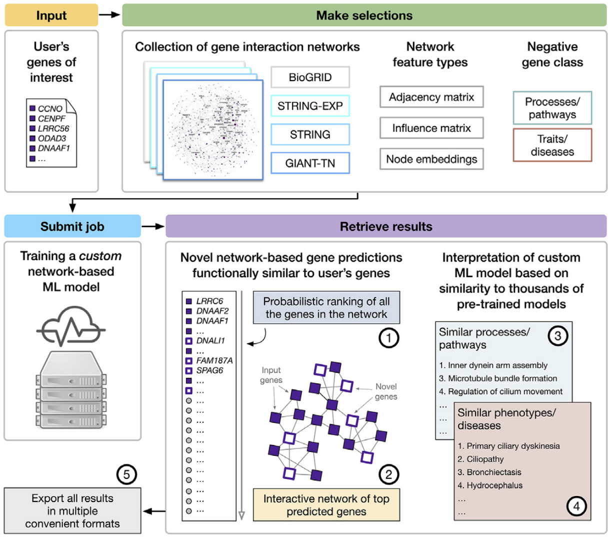
[1] Department Of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, USA, [2] Data Management and Analytics, IT Services, Michigan State University, East Lansing, MI 48824, USA and [3] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

Received March 08, 2022; Revised April 13, 2022; Editorial Decision April 19, 2022; Accepted April 30, 2022

## ABSTRACT

**Biomedical researchers take advantage of high-throughput, high-coverage technologies to routinely generate sets of genes of interest across a wide range of biological conditions. Although these technologies have directly shed light on the molecular underpinnings of various biological processes and diseases, the list of genes from any individual experiment is often noisy and incomplete. Additionally, interpreting these lists of genes can be challenging in terms of how they are related to each other and to other genes in the genome. In this work, we present GenePlexus (https://www.geneplexus.net/), a web-server that allows a researcher to utilize a powerful, network-based machine learning method to gain insights into their gene set of interest and additional functionally similar genes. Once a user uploads their own set of human genes and chooses between a number of different human network representations, GenePlexus provides predictions of how associated every gene in the network is to the input set. The web-server also provides interpretability through network visualization and comparison to other machine learning models trained on thousands of known process/pathway and disease gene sets. GenePlexus is free and open to all users without the need for registration.**

## GRAPHICAL ABSTRACT


Input
Make selections
User's genes of interest
CCNO
CENPF
LRRC56
ODAD3
DNAAF1
...
Collection of gene interaction networks
BioGRID
STRING-EXP
STRING
GIANT-TN
Network feature types
Adjacency matrix
Influence matrix
Node embeddings
Negative gene class
Processes/ pathways
Traits/ diseases
Submit job
Retrieve results
Training a *custom* network-based ML model
Novel network-based gene predictions functionally similar to user's genes
LRRC6
DNAAF2
DNAAF1
...
DNALI1
...
FAM187A
SPAG6
...
Probabilistic ranking of all the genes in the network
1
Input genes
Novel genes
2
Interactive network of top predicted genes
Interpretation of custom ML model based on similarity to thousands of pre-trained models
Similar processes/ pathways
3
1. Inner dynein arm assembly
3. Microtubule bundle formation
4. Regulation of cilium movement
Similar phenotypes/ diseases
1. Primary ciliary dyskinesia
2. Ciliopathy
3. Bronchiectasis
4. Hydrocephalus
4
5
Export all results in multiple convenient formats

## INTRODUCTION

Most complex functions, phenotypes, traits and diseases involve complex interactions between many genes. With the advent of high-throughput, high-coverage technologies (1,2), researchers are able to measure various types of signals pertaining to these phenomena on a genome-wide scale and ultimately generate a list of genes of interest. For instance, differential expression analysis (3,4) of bulk- or single-cell transcriptomes allow researchers to generate gene sets of interest, which provide some initial insight into the molecular underpinnings of the experimental factors being studied. However, these gene sets often suffer from a few drawbacks: (i) the gene sets can be incomplete (i.e. containing false negatives) and noisy (i.e. containing false positives) and (ii) the gene list inherently lacks information about how the individual genes interact with each other and with other genes in the genome.

[*] To whom correspondence should be addressed. Email: arjun@msu.edu

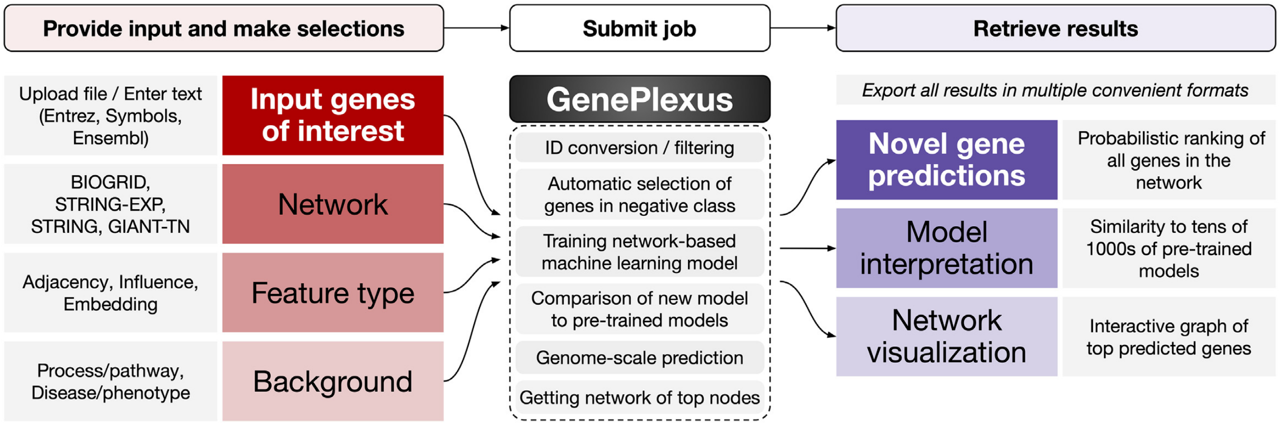

Provide input and make selections
Submit job
Retrieve results
Upload file / Enter text (Entrez, Symbols, Ensembl)
Input genes of interest
BIOGRID, STRING-EXP, STRING, GIANT-TN
Network
Adjacency, Influence, Embedding
Feature type
Process/pathway, Disease/phenotype
Background
GenePlexus
ID conversion / filtering
Automatic selection of genes in negative class
Training network-based machine learning model
Comparison of new model to pre-trained models
Genome-scale prediction
Getting network of top nodes
Export all results in multiple convenient formats
Novel gene predictions
Probabilistic ranking of all genes in the network
Model interpretation
Similarity to tens of 1000s of pre-trained models
Network visualization
Interactive graph of top predicted genes

**Figure 1.** The workflow of the GenePlexus web-server. First the user uploads a gene set of interest and choses the network and representation and how the negative genes should be selected. Next, the data is prepared, the model is trained and the results are created. Finally, the user can retrieve gene predictions, gain insight into the trained model and visualize the network connectivity of the top genes interactively through their browser.

The ability to computationally refine an experimentally-derived gene set by prioritizing genes of interest and predicting other novel genes that may be highly related to the set is a grand challenge in biomedical research (5–11). Although experimental validation is always required, the sheer number of possible novel associations require computational techniques to guide which genes to study next. Over the past few decades, computational methods that incorporate information from genome-wide, context-specific molecular-networks have shown state-of-the-art results (12–21). Recently, we have shown that directly using the connections from genome-wide molecular networks as the features to a supervised machine learning model (referred to as GenePlexus) is a robust, data-driven way to computational predict how associated a gene is to a given input gene set (22).

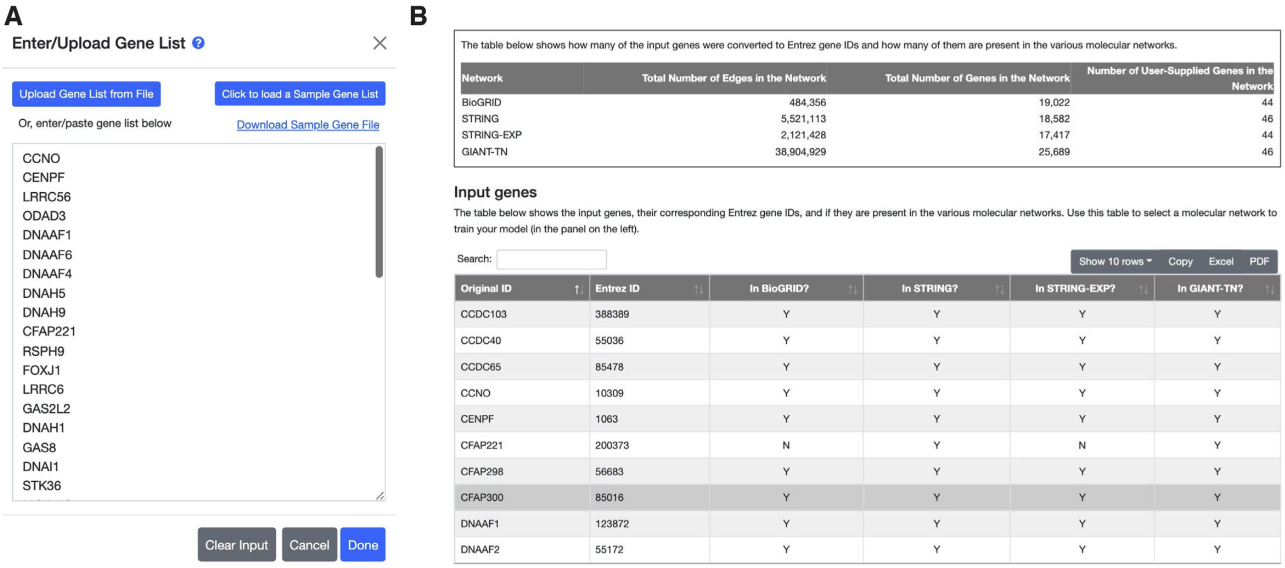
As powerful as these computational methods can be, their impact is fully realized only if they can be put into the hands of biomedical researchers, regardless of programming and computational background. Publicly available web-servers are a great platform for disseminating these results and an ideal web-server would have the following properties:

- Handle gene sets generated across vastly different biological contexts and from different technologies.
- Allow the user to choose from a suite of molecular networks that best fit the biological question.
- Provide predictive insights about additional genes most functionally similar to the user-supplied gene set and provide a confidence level of these predictions in a timely manner.
- Enable the researcher to interpret the underlying computational model and to visualize the connectivity of the top-ranked genes.
- Provide a user interface that is intuitive and easy to use for a biologist regardless of programming skills, provide extensive help/tutorials, and provide open-source code for the predictive model and web-server.

In this work, we present the GenePlexus web-server which addresses all the needs above [Figure 1]. A user can upload a set of genes and choose the desired network properties. Then the web-server trains a *custom* supervised machine learning model using the user-supplied genes as positive labels. Within a few minutes, the user can then retrieve an association probability for every gene in the network, interpret the trained model through a comparison to other models trained on known gene sets that correspond to process/pathways and diseases from the Gene Ontology (23,24) and DisGeNet (25,26) databases, respectively, and visualize the network connectivity of the top-ranked genes. We believe that the GenePlexus web-server will greatly benefit anyone who is looking to determine novel associations to a given gene list in a biologically interpretable manner.

There are a number of comparable web-servers that analyze a user-supplied gene set in the context of a molecular network, but they all have some limitations. Web-servers for networks such as STRING (27) and GIANT (20) do not provide a predictive element, instead focussing on offering an interactive visualization of the gene set within their networks. Web-servers such as GeneMania (12), HumanNet (28), ToppGene (29) and MaxLink (30) provide predictions using the method of label propagation, a semi-supervised method which our model has been shown to outperform (22). DGLinker (31) is a powerful and comprehensive web-server that trains a supervised machine learning model on the user-supplied gene set. However, the supervised learning model in DGLinker uses three features that are mined from a vast amount of data sources. In comparison, GenePlexus uses the entirety of the network connections as input to the machine learning model. Additionally, DGLinker is specific to gene-disease discovery, whereas GenePlexus is task agnostic.

A key feature that sets GenePlexus apart is how model interpretation is implemented. All the web-servers mentioned above only offer (if offered at all) insights about the input gene set and, possibly, the highly associated novel genes, through the widely-used analysis technique of gene set enrichment. As there exist many excellent open-source tools for enrichment analysis, we chose not to implement that on GenePlexus. Instead, we provide interpretation of the custom trained machine learning model. We do this by comparing the model trained on the user-supplied gene set to


A
Enter/Upload Gene List
Upload Gene List from File
Click to load a Sample Gene List
Download Sample Gene File
Clear Input
Cancel
Done
B
Search:
Show 10 rows
Copy
Excel
PDF

Or, enter/paste gene list below

CCNO
CENPF
LRRC56
ODAD3
DNAAF1
DNAAF6
DNAAF4
DNAH5
DNAH9
CFAP221
RSPH9
FOXJ1
LRRC6
GAS2L2
DNAH1
GAS8
DNAI1
STK36

The table below shows how many of the input genes were converted to Entrez gene IDs and how many of them are present in the various molecular networks.

| Network | Total Number of Edges in the Network | Total Number of Genes in the Network | Number of User-Supplied Genes in the Network |
|---|---|---|---|
| BioGRID | 484,356 | 19,022 | 44 |
| STRING | 5,521,113 | 18,582 | 46 |
| STRING-EXP | 2,121,428 | 17,417 | 44 |
| GIANT-TN | 38,904,929 | 25,689 | 46 |

Input genes

The table below shows the input genes, their corresponding Entrez gene IDs, and if they are present in the various molecular networks. Use this table to select a molecular network to train your model (in the panel on the left).

| Original ID | Entrez ID | In BioGRID? | In STRING? | In STRING-EXP? | In GIANT-TN? |
|---|---|---|---|---|---|
| CCDC103 | 388389 | Y | Y | Y | Y |
| CCDC40 | 55036 | Y | Y | Y | Y |
| CCDC65 | 85478 | Y | Y | Y | Y |
| CCNO | 10309 | Y | Y | Y | Y |
| CENPF | 1063 | Y | Y | Y | Y |
| CFAP221 | 200373 | N | Y | N | Y |
| CFAP298 | 56683 | Y | Y | Y | Y |
| CFAP300 | 85016 | Y | Y | Y | Y |
| DNAAF1 | 123872 | Y | Y | Y | Y |
| DNAAF2 | 55172 | Y | Y | Y | Y |

**Figure 2.** Uploading and validating the gene set. (**A**) The user can either paste gene IDs or upload them from a file. (**B**) Upon clicking the Done button, the genes are converted in Entrez ID space and the overlap of the gene set with the genes in each network is displayed.

thousands of models that were trained on known biological processes and pathways in the Gene Ontology or diseases in DisGeNet.

## WEB-SERVER WALKTHROUGH

The main purpose of the GenePlexus web-server is to discover novel genes that are functionally similar to a user defined set of 'genes of interest'. This purpose is accomplished by training a molecular-network-informed machine learning model specific to the user-supplied genes. The key features of the web-server are:

- Prediction of how every human gene is functionally similar to the user-supplied gene set.
- Allow the user to choose between a number of different molecular networks, varying in interaction source, coverage, and density. The user can also choose how the networks are represented in the machine learning model.
- Interpretation of the model by comparing the model trained using the user-supplied gene set to thousands of models pre-trained using gene sets annotated to biological processes in the Gene Ontology and diseases in DisGeNet.
- Visualization of the network connections for the top-ranked genes.
- The web-server is open source (https://github.com/krishnanlab/geneplexus_app) and has extensive help documentation, both in the form of a help page as well as video tutorials.

## INPUTS

### Adding genes

The first step is for the user to add a set of human genes [Figure 2A]. Users can do this by either entering the genes manually into a box or uploading a file. The genes can be identified using Ensembl IDs (ENSG, ENSP or ENST), Gene Symbols or NCBI Entrez IDs. The web-server uses cookies to allow the added genes to persist as the user navigates across the pages and to allow the user to edit or add additional genes manually or with a file. We emphasize that no tracking cookies are utilized.

### Validating genes

The user will then click the 'Done' button, which will first convert the user-supplied genes into Entrez ID space. This is done because all the networks, gene set collections, and pretrained models are in Entrez ID space. Next, the web-server determines which of the input genes are present in the different molecular networks and returns this information as a brief summary in the form of a searchable, downloadable table [Figure 2B].

## SELECTING PARAMETERS FOR SETTING UP MACHINE LEARNING MODEL

The user will then select a few parameters that will be used to train the model and generate the results:

- **Molecular Network**: The user can choose from four different human networks [Table S1] that vary in interaction source, coverage, and density (BioGRID (v4.2.191) (32,33), STRING (v11.0) (27), STRING-EXP (v11.0) (27) and GIANT-TN (v 1.0) (20)).
- **Feature Type**: This is how the connections in the network are represented in the machine learning model (Adjacency, Influence, Embedding).
- **Negative Gene Class**: This is a gene set collection [Table S2] that will be used to select negatives based on whether the input genes correspond to a process/pathway or a disease.

For more information about these choices, see the 'Running the model' section and the *Supplemental Material*. Additionally, the Help page of the web-server provides some guidance on which choice of parameters would be appropriate for a given user-supplied gene set.

The user also has the *option* to provide two additional pieces of information:

- **Jobname**: GenePlexus automatically creates an eight digit random alpha-numeric job name. In addition, the user is able to supply a memorable prefix to this name.
- **Email**: The user can add their email address to receive a message when the job has started and when the job is completed. These emails include the job parameters and the url to retrieve the job results.

## RUNNING THE MODEL

When the user hits the 'Submit' button, they will be directed to the jobs page where they can see the status of all recent jobs that were submitted. Once the job is completed, the link to display the results will become active. After the job is submitted, the user's gene set and selected parameters will be passed to the computational model that performs the following steps

- **Positive and negative gene selection**: The positive examples in the machine learning model are all the genes in the user-supplied gene list that could be mapped to Entrez IDs and were present in the chosen network. The negative examples are chosen in the following way:
  - Consider all genes that are present in the selected gene set collection ('Negative Gene Class') as negative examples.
  - Remove from this negative set any genes in the positive set.
  - Additionally, remove genes annotated to gene sets in the collection that significantly overlap with the positive set.
- **Training a machine learning model**: The machine learning model is a regularized (l2-norm) logistic regression model with a regularization parameter of 1. The features used in the model are created based on the user's selection of the network and feature type, where the feature types are:
  - **Adjacency**: Features are the connections in the original network.
  - **Influence**: Features are generated by using a random-walk with restart diffusion kernel on the original network.
  - **Embedding**: Features are generated using the node embedding algorithm *node2vec* (34).
- **Evaluating the machine learning model**: If the positive set contains at least 15 genes, the model is internally evaluated using 3-fold cross validation. The cross validation results provide a very useful measure of whether the model worked well on the gene set of interest, and can be used to help a user choose the optimal network, feature, and negative selection options.
- **Generation of the results**: Once the model is trained, it is used to predict a score for every gene in the network. The results also include the similarity of thousands of pre-trained process/pathway- and disease-gene models to the custom-trained model, and a visualization of the network connectivity of the top-ranked genes.

For more detailed information on the machine learning model see (22) and the *Supplemental Material*.

## RESULTS

The GenePlexus web-server returns a number of useful results that can either be downloaded as one compressed (zip) file or individually in multiple useful formats. The results page can be navigated through a number of tabs, and the parameters used to generate the results as well as the cross-validation results are displayed at the top of each tab.

### Predicted gene associations

The main result returned is the predicted functional similarity of every gene contained in the selected network to the user-supplied input list. The first three columns give details on the genes with hyperlinks to NCBI gene pages with more information. The last four columns provide the following information:

- **Probability**: The predicted probability from the logistic regression model.
- **Training-Label**: The label of the gene used during training with P: positive, N: negative and U: unused.
- **Known/Novel**: Positive genes are considered 'Known' and negative genes and genes unused during training are considered 'Novel'.
- **Rank**: Provides the ranking of the gene based on its predicted probability.

These results are returned as a searchable table that is originally sorted by predicted probability. For example, the user can display just the scores for the 'Novel' genes by typing 'Novel' into the search box above the table.

### Model interpretability

In web-servers that provide a similar service, if interpretability of the results is offered, it is done so in the form of biological processes (or other curated gene sets) enriched in the user-supplied gene set with the addition of predicted genes. As gene set enrichment is easily accessible through dozens of web-servers and software packages, we instead provide the user with some interpretation of the custom-trained machine learning model.

We accomplish this by comparing the model trained on the user-supplied gene set to thousands of models pre-trained using known gene sets corresponding to biological processes from the Gene Ontology and diseases from DisGeNet. These pre-trained models are built using the same network, feature, and negative gene set collection used to train the custom model trained on the user-supplied gene set. For detailed information on how the similarity score is calculated, see the *Supplemental Material*. We highlight that this feature is unique to GenePlexus. It provides a very network-specific interpretation of the trained model that relies on data generated using >10 000 computational hours on high-memory nodes.

There are two tabs that show the most similar models trained on biological process (Gene Ontology) and disease (DisGeNet) gene sets. In each tab, the first two columns in the table are the ID and long-form name of the known gene set along with a hyperlink to more information. The last two columns contain the similarity to the user's custom-trained model and the rank.

The tables that contain the predicted gene association scores or the similarity to pre-trained models only include the top 500 entries. This is done to increase the speed in which the user can load and interact with these tables. At any point, the user can easily download the full results, which contain >17k gene predictions and similarities to thousands of known gene sets.

### Network graph

Finally, the top gene predictions are also visualized in the context of the original network that was used to train the model. The user can directly change the number of nodes (up to a maximum of 50) or view nodes based on a prediction probability threshold. The set of edges that are displayed can also be changed by setting an edge weight threshold. Individual nodes can be dragged to specific positions and the entire network can be panned and zoomed. Upon clicking on a node, a list of information about that node is supplied.

## WEB-SERVER IMPLEMENTATION

The web-server is implemented using services on the Microsoft Azure cloud platform. The front end is a low resource service that allows the user to upload genes and select parameters, and was written using the microframework Flask. When a job is submitted, the web-server automatically creates a high-resource containerized instance in Azure that contains a Docker version of the source code needed to train the model and generate the results. Once the job is complete, this container is automatically deleted. The ability of the web-server to automatically create and delete these high-resource containers on-demand allows it to simultaneously train numerous machine learning models at minimal costs.

A key feature of the GenePlexus web-server is the ability to allow a user to choose from a variety of networks. Although the networks contained in the current version vary greatly in interaction source, coverage, and density, our implementation can add new networks based on user feedback. We have designed the backend data formats and structure to be flexible so that it is easy to incorporate new networks and gene set collections.

## BENCHMARKING THE GENEPLEXUS METHOD

The supervised network-based machine learning model that forms the backbone of GenePlexus has been extensively benchmarked in (22). In that work, the supervised model was shown to outperform the widely-used, state-of-the-art method of label propagation (12–15,35–48). The comparison included a number of different tasks (function-, disease-, trait-gene prediction), networks (BioGRID, STRING, InBioMap, GIANT-TN, STRING-EXP), validation schemes (temporal holdout, study-bias holdout, 5-fold CV), and evaluation metrics (auPRC, P@topK, auROC).

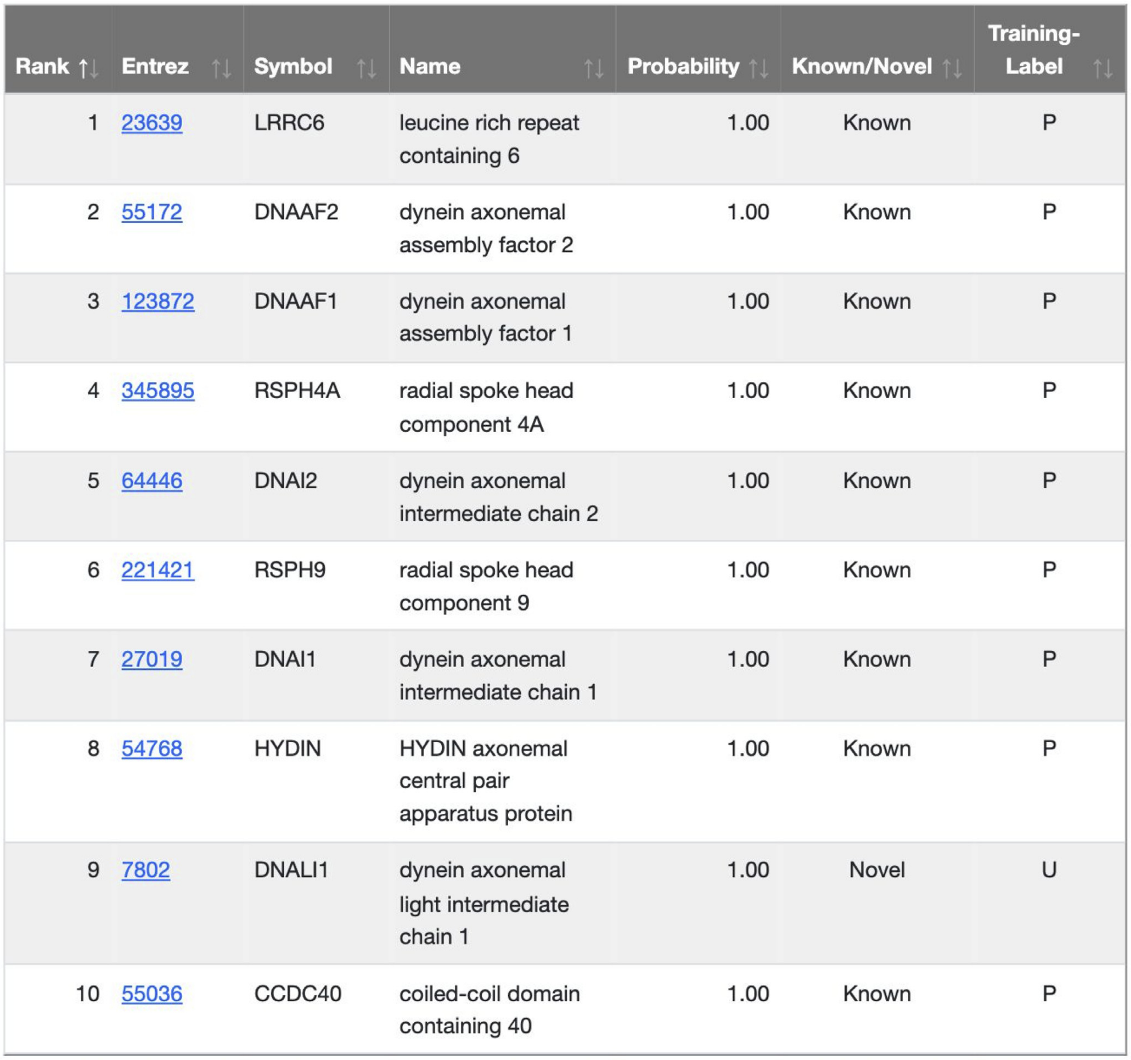
## ILLUSTRATIVE EXAMPLE

Throughout this work, we demonstrate the utility and features of GenePlexus by applying it to discover genes associated with primary ciliary dyskinesia (PCD). PCD is a genetic condition in which the microscopic organelles (cilia) in the respiratory system have defective function. While a few genes associated with PCD are already known, the genetic cause of the disorder is unknown in many individuals with PCD, making it critical to continue identifying novel PCD genes. Here, we used GenePlexus to predict novel genes associated with PCD based on a gene interaction network, starting with a set of 46 known PCD genes, obtained from the DisGeNet database. The PCD model was trained using the adjacency matrix representation of the STRING network, and the negative genes were determined based on other similar diseases in DisGeNet. This is the same example gene set available to a user to explore on the GenePlexus web-server.

Typically in the GenePlexus web-server, as expected, the top-ranked genes consist of many genes included in the user-supplied gene set, and this can be seen for PCD [Figure 3]. A number of these genes belong to the family of axonemal dyneins that cause sliding of microtubules in the axonemes of cilia and flagella (49,50). With the above stated network choices, GenePlexus predicts that *DNALI1*, dynein axonemal light intermediate chain 1, is functionally similar to this input set and is highly connected to known positive genes in the network [Figure 4C]. Additionally, there exists experimental evidence that *DNALI1* is associated with PCD (51,52).

When comparing the user's custom-trained model to models pre-trained on known disease gene sets from DisGeNet [Figure 4A], unsurprisingly the closest models are 'PCD' (since this was trained using the same input genes and parameters as the user trained model) and 'ciliopathy', which is PCD's parent term in the Disease Ontology (53). Other top associated diseases are bronchiectasis, hydrocephalus and joubert syndrome, which are other ciliopathy related diseases.

GenePlexus also compares the user's custom-trained model to models pre-trained on known biological process gene sets from Gene Ontology [Figure 4B]. It can be seen that most of the top models are related to either dynein assembly or cilia movement, which have been shown to be underlying mechanisms deregulated in PCD (52). Though the above two analyses may seem similar to gene set enrichment analysis, we note a key difference. Instead of providing interpretation of the list of genes directly, GenePlexus interprets the machine learning model custom-built for PCD to identify diseases and processes that have highly similar patterns of network connectivity compared to PCD. Thus, even if a relevant process/pathway/disease has few (if any) genes overlapping with known PCD genes, it will be considered



| Rank | Entrez | Symbol | Name | Probability | Known/Novel | Training-Label |
|---|---|---|---|---|---|---|
| 1 | 23639 | LRRC6 | leucine rich repeat containing 6 | 1.00 | Known | P |
| 2 | 55172 | DNAAF2 | dynein axonemal assembly factor 2 | 1.00 | Known | P |
| 3 | 123872 | DNAAF1 | dynein axonemal assembly factor 1 | 1.00 | Known | P |
| 4 | 345895 | RSPH4A | radial spoke head component 4A | 1.00 | Known | P |
| 5 | 64446 | DNAI2 | dynein axonemal intermediate chain 2 | 1.00 | Known | P |
| 6 | 221421 | RSPH9 | radial spoke head component 9 | 1.00 | Known | P |
| 7 | 27019 | DNAI1 | dynein axonemal intermediate chain 1 | 1.00 | Known | P |
| 8 | 54768 | HYDIN | HYDIN axonemal central pair apparatus protein | 1.00 | Known | P |
| 9 | 7802 | DNALI1 | dynein axonemal light intermediate chain 1 | 1.00 | Novel | U |
| 10 | 55036 | CCDC40 | coiled-coil domain containing 40 | 1.00 | Known | P |

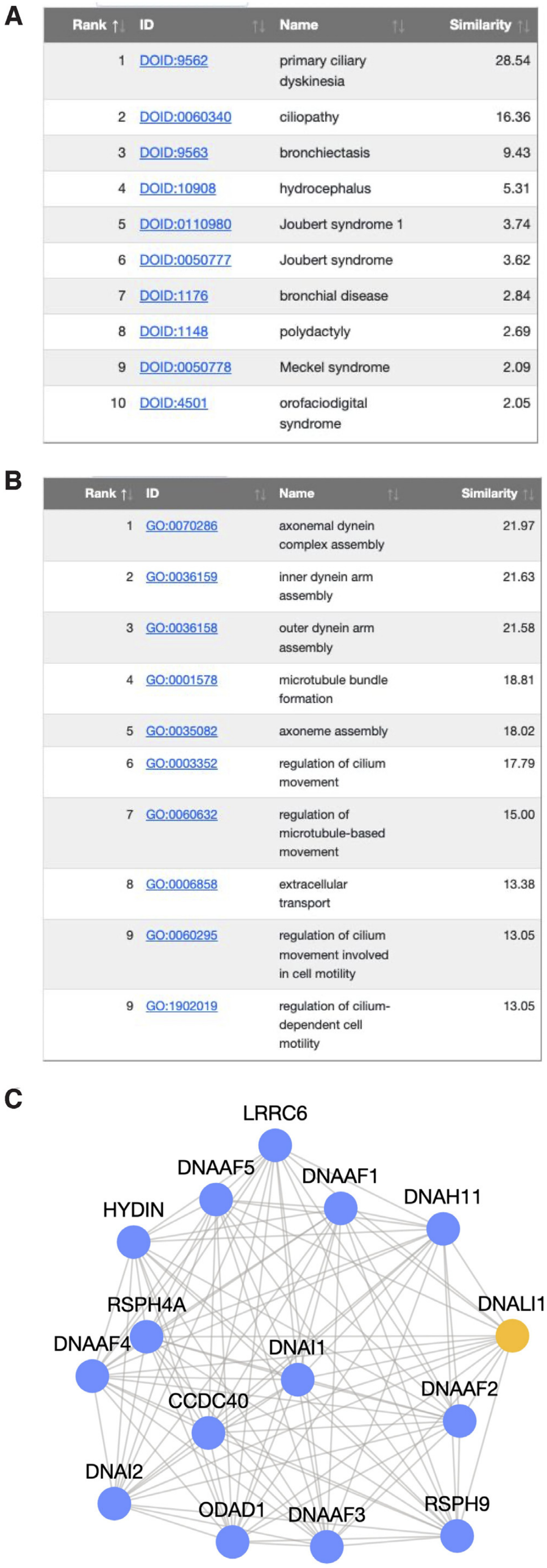
**Figure 3.** Genome-wide Prediction. For every gene in the genome-scale molecular network that was used to train the model, a score is calculated of how associated it is to the user-supplied gene set and displayed as an interactive table.

similar to PCD if the two sets of genes have overlapping network neighborhoods. This model interpretation approach takes advantage of pre-training thousands of models that took >10 000 computational hours, and is an analysis that is unique to GenePlexus.

## DISCUSSION

The GenePlexus web-server provides a powerful tool that any researcher can use to understand and expand any list of human genes generated from an experimental/empirical study including omics profiling, phenotypic assay, association study or drug screen. The web-server leverages the strengths of genome-scale human gene networks and machine learning to help users discover additional novel genes that are functionally similar to their list of input genes. Given a list of input genes, GenePlexus predicts novel genes that have similar network neighborhoods with the input genes, even when these novel genes are not directly connected to the input genes. The web-server achieves this by building a custom machine learning model that finds patterns of connectivity in the network that are distinctive to the input genes and then using this model to find other genes that have similar network patterns. In addition to expanding the original gene set, this analysis illuminates the functional relationships between the known and novel genes by placing them in the context of a molecular network. GenePlexus also helps the user peer into the blackbox by revealing similar machine learning models built for other biological gene sets. We have implemented features so that, at every stage of the analysis, users can export both top-ranked and full results in various convenient formats including plain-text tables, PDFs and image files (as appropriate).

While the GenePlexus web-server contains a number of useful features and options to choose from, we are open to and encourage users to suggest new features. Future versions of the web-server will likely include a larger variety of networks (both human and model species), a public and searchable database of anonymized results from machine learning models trained by other users, the ability to predict novel genes in model species based on human gene networks, and the *option* for users to create accounts to help them keep track of past jobs. We plan on updating the web-server on a yearly basis, which includes adding additional features as well as updating the data used by the web-server. We will continue to have older versions of the web-server publicly available through links on the most current version.



| Rank | ID | Name | Similarity |
|---|---|---|---|
| 1 | DOID:9562 | primary ciliary dyskinesia | 28.54 |
| 2 | DOID:0060340 | ciliopathy | 16.36 |
| 3 | DOID:9563 | bronchiectasis | 9.43 |
| 4 | DOID:10908 | hydrocephalus | 5.31 |
| 5 | DOID:0110980 | Joubert syndrome 1 | 3.74 |
| 6 | DOID:0050777 | Joubert syndrome | 3.62 |
| 7 | DOID:1176 | bronchial disease | 2.84 |
| 8 | DOID:1148 | polydactyly | 2.69 |
| 9 | DOID:0050778 | Meckel syndrome | 2.09 |
| 10 | DOID:4501 | orofaciodigital syndrome | 2.05 |

| Rank | ID | Name | Similarity |
|---|---|---|---|
| 1 | GO:0070286 | axonemal dynein complex assembly | 21.97 |
| 2 | GO:0036159 | inner dynein arm assembly | 21.63 |
| 3 | GO:0036158 | outer dynein arm assembly | 21.58 |
| 4 | GO:0001578 | microtubule bundle formation | 18.81 |
| 5 | GO:0035082 | axoneme assembly | 18.02 |
| 6 | GO:0003352 | regulation of cilium movement | 17.79 |
| 7 | GO:0060632 | regulation of microtubule-based movement | 15.00 |
| 8 | GO:0006858 | extracellular transport | 13.38 |
| 9 | GO:0060295 | regulation of cilium movement involved in cell motility | 13.05 |
| 9 | GO:1902019 | regulation of cilium-dependent cell motility | 13.05 |

**Figure 4.** Interpretability features of GenePlexus. (A) The model trained using the user-supplied gene set is compared to thousands of models pre-trained on known gene sets from the (**A**) GeneOntology and (**B**) DisGeNet databases. (**C**) The network connectivity of the top associated genes are displayed as an interactive graph.

Though many modern web-servers implement an API that allows users to access the web-server programmatically, implementing this feature in a predictive web-server like GenePlexus presents unique challenges. Using an API, a single user could submit many jobs in a short amount of time. Though this could be very useful for a computational biologist looking to expand many gene sets, each submitted job launches a high-resource container on the cloud, which becomes an expensive endeavor. To this end, we have released an open source python package (https://pypi.org/project/geneplexus/) that could be used on its own without the web-server.

## DATA AVAILABILITY

The GenePlexus web-server is freely available at https://www.geneplexus.net/ and the code base for creating the web-server is freely available at https://github.com/krishnanlab/geneplexus_app.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank members of the Krishnan Lab for valuable discussions and feedback on the manuscript and the web-server. We also thank Dr Janani Ravi for feedback on the web-server.

## FUNDING

US National Institutes of Health (NIH) [R35 GM128765 to A.K., F32 GM134595 to C.A.M.]. Funding for open access charge: MSU Startup Funds.

*Conflict of interest statement.* None declared.

## REFERENCES

1. Heller,M.J. (2002) DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.*, **4**, 129–153.
2. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
3. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
4. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
5. Sharan,R., Ulitsky,I. and Shamir,R. (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.
6. Peña-Castillo,L., Tasan,M., Myers,C.L., Lee,H., Joshi,T., Zhang,C., Guan,Y., Leone,M., Pagnani,A., Kim,W.K. *et al.* (2008) A critical assessment of Mus musculusgene function prediction using integrated genomic evidence. *Genome Biol.*, **9**, S2.

7. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
8. Bernardes,J.S. and Pedreira,C.E. (2013) A review of protein function prediction under machine learning perspective. *Recent Pat. Biotechnol.*, **7**, 122–141.
9. Piro,R.M. and Cunto,F.D. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.
10. Yang,J., Lee,S.H., Goddard,M.E. and Visscher,P.M. (2011) GCTA: a tool for Genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
11. Jiang,Y., Oron,T.R., Clark,W.T., Bankapur,A.R., D'Andrea,D., Lepore,R., Funk,C.S., Kahanda,I., Verspoor,K.M., Ben-Hur,A. *et al.* (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.*, **17**, 184.
12. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
13. Köhler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.
14. Vanunu,O., Magger,O., Ruppin,E., Shlomi,T. and Sharan,R. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol.*, **6**, e1000641.
15. Leiserson,M., Vandin,F., Wu,H.-T., Dobson,J., Eldridge,J., Thomas,J., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M. *et al.* (2015) Pan-Cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
16. Guan,Y., Ackert-Bicknell,C.L., Kell,B., Troyanskaya,O.G. and Hibbs,M.A. (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.*, **6**, e1000991.
17. Park,C.Y., Wong,A.K., Greene,C.S., Rowland,J., Guan,Y., Bongo,L.A., Burdine,R.D. and Troyanskaya,O.G. (2013) Functional knowledge transfer for High-accuracy prediction of Under-studied biological processes. *PLoS Comput. Biol.*, **9**, e1002957.
18. Lanckriet,G.R.G., Deng,M., Cristianini,N., Jordan,M.I. and Noble,W.S. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput.*300–311.
19. Barutcuoglu,Z., Schapire,R.E. and Troyanskaya,O.G. (2006) Hierarchical multi-label prediction of gene function. *Bioinformatics*, **22**, 830–836.
20. Greene,C.S., Krishnan,A., Wong,A.K., Ricciotti,E., Zelaya,R.A., Himmelstein,D.S., Zhang,R., Hartmann,B.M., Zaslavsky,E., Sealfon,S.C. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.
21. Krishnan,A., Zhang,R., Yao,V., Theesfeld,C.L., Wong,A.K., Tadych,A., Volfovsky,N., Packer,A., Lash,A. and Troyanskaya,O.G. (2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.*, **19**, 1454–1462.
22. Liu,R., Mancuso,C.A., Yannakopoulos,A., Johnson,K.A. and Krishnan,A. (2020) Supervised-learning is an accurate method for network-based gene classification. *Bioinformatics*, **36**, 3457–3465.
23. 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. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
24. The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
25. Piñero,J., Queralt-Rosinach,N., Bravo,À., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028.
26. Piñero,J., Bravo,À., Queralt-Rosinach,N., Gutiérrez-Sacristán,A., Deu-Pons,J., Centeno,E., García-García,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
27. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
28. Kim,C.Y., Baek,S., Cha,J., Yang,S., Kim,E., Marcotte,E.M., Hart,T. and Lee,I. (2022) HumanNet v3: an improved database of human gene networks for disease research. *Nucleic Acids Res.*, **50**, D632–D639.
29. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37**, W305–W311.
30. Guala,D., Sjölund,E. and Sonnhammer,E.L.L. (2014) MaxLink: network-based prioritization of genes tightly linked to a disease seed set. *Bioinformatics*, **30**, 2689–2690.
31. Hu,J., Lepore,R., Dobson,R.J.B., Al-Chalabi,A., Bean,D.M. and Iacoangeli,A. (2021) DGLinker: flexible knowledge-graph prediction of disease–gene associations. *Nucleic Acids Res.*, **49**, W153–W161.
32. Stark,C., Breitkreutz,B.-J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
33. Oughtred,R., Stark,C., Breitkreutz,B.-J., Rust,J., Boucher,L., Chang,C., Kolas,N., O'Donnell,L., Leung,G., McAdam,R. *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
34. Grover,A. and Leskovec,J. (2016) node2vec: Scalable Feature Learning for Networks. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. pp.855–864.
35. Wang,X., Gulbahce,N. and Yu,H. (2011) Network-based methods for human disease gene prediction. *Brief. Funct. Genomics*, **10**, 280–293.
36. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257.
37. Murali,T.M., Dyer,M.D., Badger,D., Tyler,B.M. and Katze,M.G. (2011) Network-based prediction and analysis of HIV dependency factors. *PLOS Comput. Biol.*, **7**, e1002164.
38. Cowen,L., Ideker,T., Raphael,B.J. and Sharan,R. (2017) Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.*, **18**, 551–562.
39. Zhou,D., Bousquet,O., Lal,T.N., Weston,J. and Schölkopf,B. (2003) Learning with local and global consistency. In: *Proceedings of the 16th International Conference on Neural Information Processing Systems, NIPS'03*. MIT Press, Cambridge, MA, USA, pp. 321–328.
40. Zhu,X., Ghahramani,Z. and Lafferty,J. (2003) Semi-supervised learning using gaussian fields and harmonic functions. In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML'03*. AAAI Press, pp.912–919.
41. Tsuda,K., Shin,H. and Schölkopf,B. (2005) Fast protein classification with multiple networks. *Bioinformatics*, **21**, ii59–ii65.
42. Komurov,K., White,M.A. and Ram,P.T. (2010) Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput. Biol.*, **6**, e1000889.
43. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
44. Karaoz,U., Murali,T.M., Letovsky,S., Zheng,Y., Ding,C., Cantor,C.R. and Kasif,S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 2888–2893.
45. Deng,M., Chen,T. and Sun,F. (2004) An integrated probabilistic model for functional prediction of proteins. *J. Comput. Biol.*, **11**, 463–475.
46. Nabieva,E., Jim,K., Agarwal,A., Chazelle,B. and Singh,M. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**, i302–i310.
47. Leone,M. and Pagnani,A. (2005) Predicting protein functions with message passing algorithms. *Bioinformatics*, **21**, 239–247.
48. Cáceres,J.J. and Paccanaro,A. (2019) Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput. Biol.*, **15**, e1007078.

49. King,S.M. (2009) Chapter 3 - Purification of axonemal dyneins and dynein-associated components from chlamydomonas. In: King,S.M. and Pazour,G.J. (eds). *Methods in Cell Biology*. Academic Press, Vol. **92**, pp. 31–48.

50. King,S.M. (2016) Axonemal dynein arms. *Cold Spring Harb. Perspect. Biol.*, **8**, a028100.

51. Olcese,C., Patel,M.P., Shoemark,A., Kiviluoto,S., Legendre,M., Williams,H.J., Vaughan,C.K., Hayward,J., Goldenberg,A., Emes,R.D. *et al.* (2017) X-linked primary ciliary dyskinesia due to mutations in the cytoplasmic axonemal dynein assembly factor PIH1D3. *Nat. Commun.*, **8**, 14279.

52. Loges,N.T., Olbrich,H., Becker-Heck,A., Häffner,K., Heer,A., Reinhard,C., Schmidts,M., Kispert,A., Zariwala,M.A., Leigh,M.W. *et al.* (2009) Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. *Am. J. Hum. Genet.*, **85**, 883–889.

53. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R. *et al.* (2019) Human disease ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.