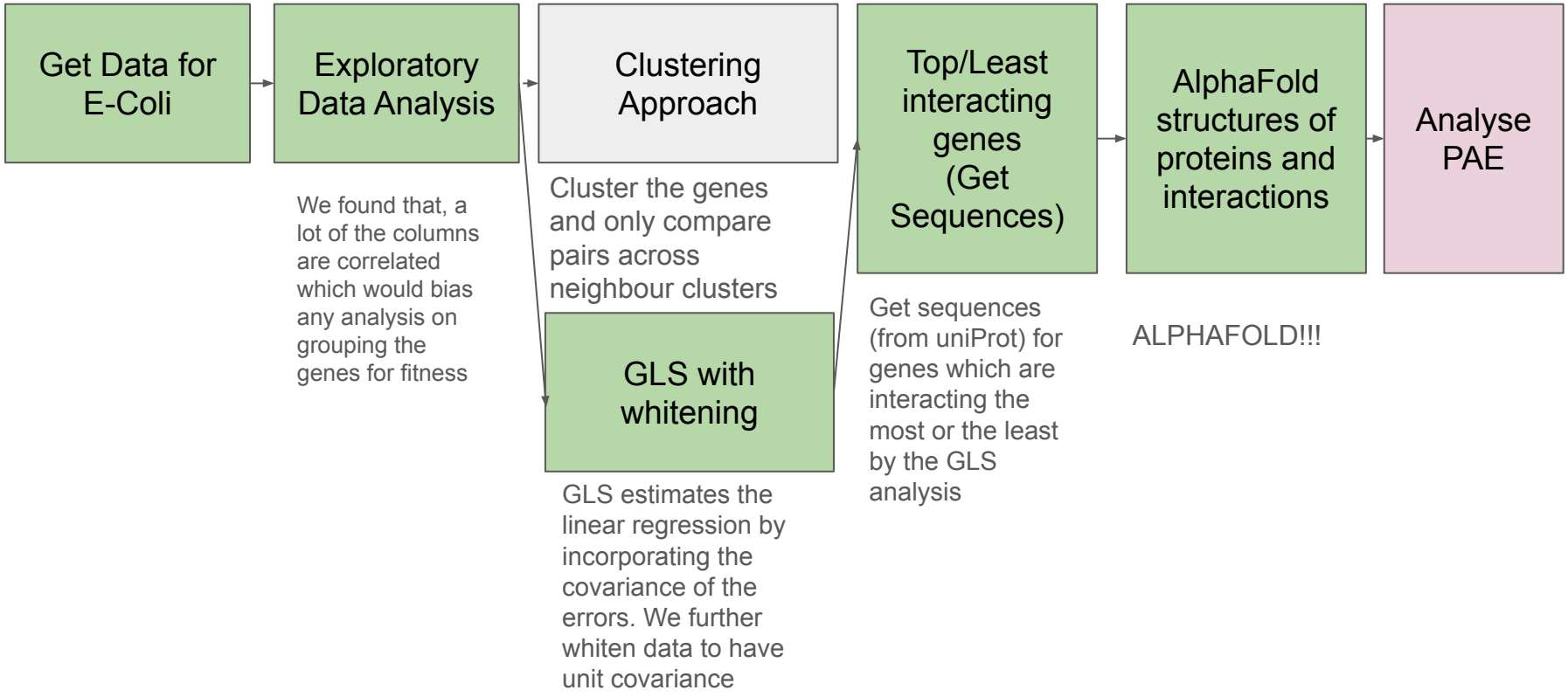
Fitness Map

Predicting protein-protein interactions in E.coli from Transposon sequencing (Tn-Seq) data

Summary of Fitness Map Workflow

1. Exploratory data analysis (EDA)
    a. Clustering analysis
2. Pearson correlation calculation– replaced by Generalized Least Squares
3. Generalized Least Squares (GLS) calculation for every gene pair
4. AlphaFold2-multimer or AlphaFold3 prediction of protein binding
5. Ranking of hits by minimum predicted aligned error of interaction (min PAE)
6. Literature search of the biology of top interacting genes

Approach


Get Data for E-Coli
Exploratory Data Analysis
Clustering Approach
Top/Least interacting genes (Get Sequences)
AlphaFold structures of proteins and interactions
Analyse PAE
We found that, a lot of the columns are correlated which would bias any analysis on grouping the genes for fitness
Cluster the genes and only compare pairs across neighbour clusters
GLS with whitening
GLS estimates the linear regression by incorporating the covariance of the errors. We further whiten data to have unit covariance
Get sequences (from uniProt) for genes which are interacting the most or the least by the GLS analysis
ALPHAFOLD!!!

EDA - Initial Data

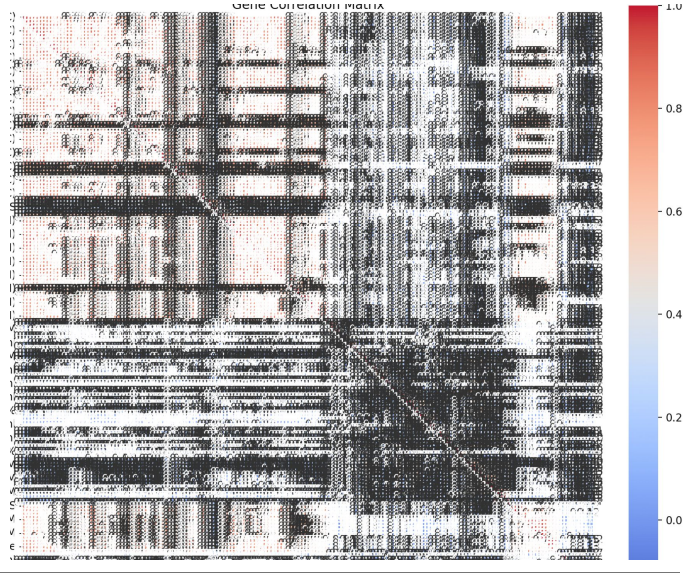
fit_organism_Keio.tsv (E. Coli. from Fitness Browser):

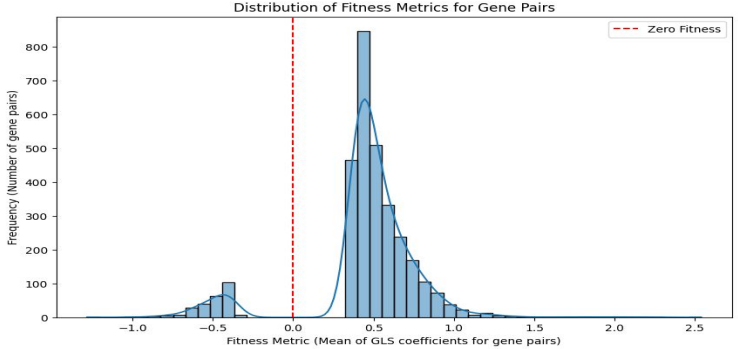
- 132 Experiments (Number of Columns)
- 3790 Genes
- 7,172,655 Gene Pairs

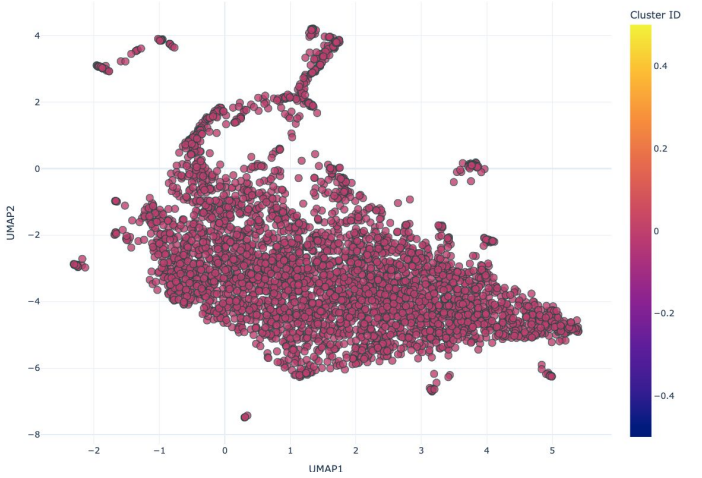
orgId	locusId	sysName	geneName	desc	set1IT003 D-Glucose	set1IT004 D-Glucose	set1IT005 D-Fructos	set1IT006 D-Fructos	set1IT009 D-Maltose	set1IT010 D-Maltose	set1IT011 D-Xylose	set1IT012 D-Xylose (	set1IT013 D-Galacto	set1IT014 D-Galacto	set1IT015 D-Ribose
Keio	10279117	b4699	fnrS	FNR-activated anaer	-0.352	0.662	-0.448	-0.051	-0.107	0.225	-0.323	-0.172	0.475	0.41	0.362
Keio	10279118	b4698	mgrR	sRNA antisense regu	-0.146	-0.032	-0.103	0.136	0.626	0.151	-0.244	-0.276	0.226	-0.299	-0.115
Keio	10279119	b4701	sokX	sok-related sRNA, fur	-0.233	-0.119	0.291	0.01	-0.025	0.44	0.184	-0.328	-0.009	-0.635	0.141
Keio	12785252	b4704	arrS	Antisense sRNA ArrS	-0.058	-0.116	-0.101	-0.535	-1.471	-1.66	-1.188	0.298	-0.397	-1.085	-1.103
Keio	12785254	b4702	mgtL	regulatory leader pep	-3.089	-2.428	-2.041	-2.173	-1.684	-2.951	-4.324	-4.505	-0.617	-1.56	-3.199
Keio	14146	b0001	thrL	thr operon leader pe	-0.564	0.132	-0.086	-0.339	-0.13	-0.278	-0.292	-0.027	-1.278	-1.249	-0.495
Keio	14147	b0002	thrA	bifunctional aspartok	-3.459	-3.633	-3.996	-4.282	-3.913	-3.857	-4.036	-4.002	-4.618	-3.971	-3.701
Keio	14148	b0003	thrB	homoserine kinase (h	-2.971	-3.525	-3.77	-3.258	-3.901	-3.48	-3.821	-5.183	-3.785	-3.686	-4.66
Keio	14149	b0004	thrC	threonine synthase (t	-5.288	-5.037	-5.32	-4.99	-4.922	-5.232	-4.836	-5.108	-4.374	-5.69	-4.091
Keio	14150	b0005	yaaX	hypothetical protein	-0.393	-0.175	0.343	0.393	0.035	-0.319	0.082	-0.117	0.006	-0.118	-0.082
Keio	14151	b0006	yaaA	hypothetical protein	0.165	0.047	-0.154	-0.11	-0.068	0.086	-0.065	-0.007	-0.106	-0.182	-0.169
Keio	14152	b0007	yaaJ	predicted transporter	-0.127	-0.339	-0.109	-0.123	-0.145	-0.033	-0.302	-0.415	-0.017	-0.071	0.581
Keio	14153	b0008	talB	transaldolase B (NCE	-0.001	0.118	-0.046	-0.02	-0.15	-0.263	-0.966	-0.816	-0.335	-0.256	-0.642
Keio	14154	b0009	mog	molybdenum cofacto	-0.144	0.001	0.019	0.168	0.181	0.182	0.295	0.228	0.252	0.158	0.332
Keio	14155	b0010	yaaH	conserved inner mer	0.088	-0.015	-0.194	-0.252	-0.136	-0.077	-0.162	-0.29	0.112	-0.059	0.017
Keio	14156	b0011	yaaW	hypothetical protein	0.151	-0.105	-0.223	-0.012	0.062	0.097	-0.094	-0.006	0.086	-0.051	0.178
Keio	14158	b0013	yaaI	hypothetical protein	0.119	0.225	0.019	0.104	0.464	0.06	0.266	0.178	0.092	0.074	0.243
Keio	14159	b0014	dnaK	molecular chaperon	-0.412	0.216	-0.498	-0.737	-1.882	-1.099	-0.157	-0.146	-1.817	-2.083	-0.187
Keio	14160	b0015	dnaJ	chaperone Hsp40, co	-0.385	-0.221	-0.347	-0.434	-0.48	-0.172	0.015	0.129	-0.692	-0.769	0.139
Keio	14163	b0018	mokC	regulatory protein for	-0.019	0.363	0.316	0.468	0.311	0.006	-1.339	0.376	0.293	0.519	-0.072
Keio	14164	b0019	nhaA	pH-dependent sodiur	0.063	0.139	0.06	0.158	0.052	0.365	0.313	0.359	0.151	0.117	0.053
Keio	14165	b0020	nhaR	DNA-binding transcri	-0.111	0.005	-0.111	-0.031	-0.166	-0.03	-0.049	-0.135	-0.053	-0.033	-0.23
Keio	14173	b0028	fkpB	FKBP-type peptidyl-p	0.133	-0.068	0.049	-0.068	-0.409	0.081	-0.256	-0.147	-0.254	0.226	-0.113

EDA - Gene Fitness and Correlation

- Features are highly correlated and suggested failure of regular regression techniques if applied
- Initial clustering revealed no significant decision boundaries in the embeddings that could help identify interesting gene-pairs
- With gls:



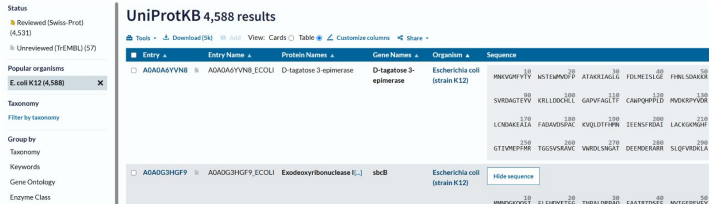




From GLS to AlphaFold Prediction

- **Obtained protein sequences for gene pairs**

1. Fitness Browser fasta file with protein sequences
2. UniProtKB data downloaded as .tsv, used for missing sequences in Fitness Browser file



- **Added amino acid sequences to data frame and eliminated unknown sequences**
  - 8 genes with unknown amino acid sequences among the top 50 most correlated pairs: many seem to code for RNAs

small RNA SraB

sRNA regulating ompA/C translation

	gene1	gene1_locusId	gene1_seq	gene2	gene2_locusId	gene2_seq	coefficient
9725021	Gene_2567	17245	MTVFNKFARTFKSHWLLYLCVIVFGITNLVASSGAHMVQRLLFFVL...	Gene_1225	15651	MKSTSDLFNEIIPLGRLIHMVNQKKDRLLNEYLSPLDITAAQFKVL...	5.033844
10544578	Gene_2783	17552	MNIYIGWLFKLIPLIMGLICIALGGFVLESSGQSEYFVAGHVLISL...	Gene_2574	17261	MENNEIQSVLMNALSLQEVHVSGDGSHFQVIAVGELFDGMSRVKKQ...	4.562073
10543821	Gene_2783	17552	MNIYIGWLFKLIPLIMGLICIALGGFVLESSGQSEYFVAGHVLISL...	Gene_1817	16347	MLSIFKPAPHKARLPAAEIDPTYRRLRWQIFLGIFFGYAAYYLVRK...	4.355901
12025962	Gene_3174	18059	MNTQYNSSYIFSITLVATLGGLLFGYDTAVISGTVESLNTVFVAPQ...	Gene_2850	17626	MQAYFDQLDRVRYEGSKSSNPLAFRHYNPDELVLGKRMEEHLRFAA...	4.136822
11273043	Gene_2975	17780	MFRRNLITSAILLMAPLAFSAQSLAESLTVEQRLELLEKALRETQS...	Gene_3744	3446183	Unknown	4.094814
393677	Gene_103	14285	MTIEYTKNYHHLTRIATFCALLYCNTAFSAELVEYDHTFLMGQNAS...	Gene_3514	1936407	Unknown	4.055026
10545240	Gene_2783	17552	MNIYIGWLFKLIPLIMGLICIALGGFVLESSGQSEYFVAGHVLISL...	Gene_3237	18130	MHLSTHPTSYPTRYQEIAAKLEQELRQHYRCGDYLPAEQQLAARFE...	4.050859

Alphafold2-Multimer and Alphafold3

- **Adapted ColabFold (Nature Methods, 2022) script to predict protein interaction for each gene pair using Alphafold2-Multimer and save outputs as a zip file titled with gene locus IDs.** ~1-4 hours per prediction

- **AlphaFold3 Server (https://alphafoldserver.com/) used for proteins with long sequences surpassing memory limits on ColabFold**

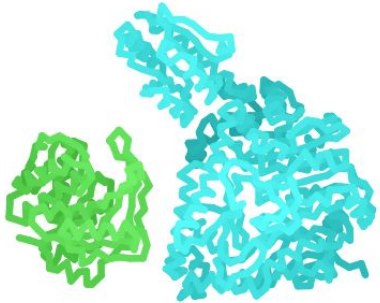
Limitation: two different AlphaFold versions for predictions that, while robust, cannot be directly compared

Predicted Structure Examples

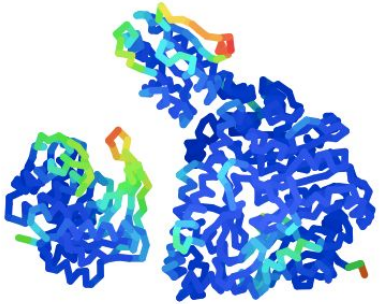
GLS: 5.0338

Min PAE: 6.00, GLS: 3.15
Proteins blue and yellow

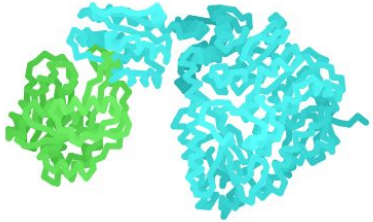
colored by chain



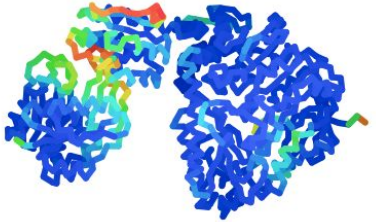
colored by pLDDT



colored by chain



colored by pLDDT

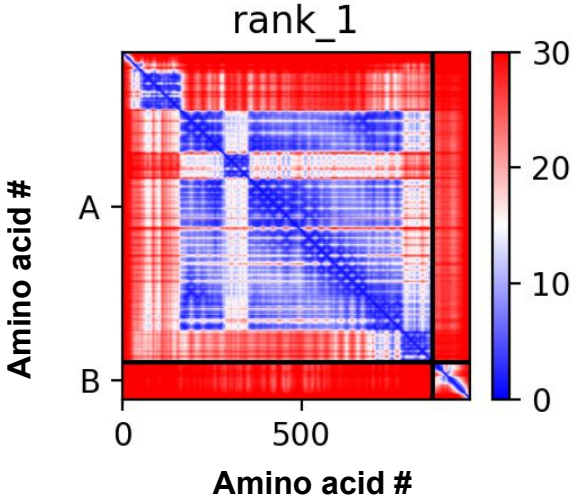


Calculating Minimum Predicted Aligned Error of Interaction (Min PAE)

Min PAE is the standard metric for evaluating AlphaFold protein interaction predictions.

- **Wrote script to calculate Min PAE from per-residue PAE matrix in AlphaFold2 outputs, and from .json file in AlphaFold3 outputs**
- **Added Min PAE values to data frame**

Blue diagonal line is each residue compared to itself



The predicted error of each residue in protein A and B is compared to all other residues in A and B

Min PAE for top correlated gene pairs

27 of the gene pairs with the top 50 highest GLS coefficients

	gene1_locusId	gene2_locusId	coefficient	Min_PAE_of_Interaction
22	1936269	12785254	3.145769	6.00
1	17552	17261	4.562073	7.26
20	17914	1937234	3.218183	9.08
23	17780	1937234	3.138881	9.97
7	17552	17788	3.846149	10.05
10	15023	1937202	3.460303	10.47
16	17552	18132	3.382060	12.64
26	14285	17261	2.884668	14.19
4	17552	18130	4.050859	14.21
17	14285	17904	3.355063	14.96
13	1937043	12785254	3.433179	15.10
6	17552	15217	3.862169	15.39
15	18059	17630	3.398975	18.77
0	17245	15651	5.033844	18.81
19	14285	15399	3.236867	19.92
24	15147	17111	3.121541	20.28
25	15023	16386	3.100637	20.35
9	17937	15399	-3.501067	21.34
21	17552	17785	3.168216	21.81
5	18059	17625	3.873120	22.20
12	18332	15978	3.435351	22.61
2	17552	16347	4.355901	23.25
14	14285	18129	3.422726	24.92
18	17724	16404	3.260949	24.97
11	17628	17625	3.449925	25.03
3	18059	17626	4.136822	25.34
8	16738	18132	3.618890	25.91

Controls: min PAE for least correlated pairs

	gene1_locusId	gene2_locusId	coefficient	Min_PAE_of_Interaction
28	15045	14698	4.703083e-09	17.53
27	17437	14332	4.520093e-09	25.03

Values that confidently indicate binding are typically **<1.5** for AlphaFold3, but **<10** may still be a possible binder

No Clear Relationship between GLS and Protein Interaction Prediction

Top interacting gene pairs Min PAE Summary

Number of gene pairs	Mean	Max	Min	Standard Deviation
27	17.59	25.91	6.00	6.11

- From our current data, there is so far no clear relationship between GLS coefficient and likelihood of protein binding.
- Limitations include a small sample size of AlphaFold predictions for gene pairs, for which we could not compare possibility of binding over a range of GLS coefficient magnitudes.

gene1_locusId	count
17552	9
14285	7
16738	4
17628	3
18059	3
18184	2
15023	2
17245	1

gene2_locusId	count
16347	3
17261	2
17626	2
18130	2
15399	2
17625	2
15217	2
18132	2
1937234	2
17630	2
18129	2
12785254	2
17111	2
14288	1

Redundant Gene Fitness Effects?

In the top 50 interacting gene pairs, 20 genes appeared more than once. yhiM (gene with locus ID 17552) appeared 9 times, and 14285 appeared 7 times.

Genes with fitness scores that match yhiM: many are impacted by phosphate levels

	LocusID	Gene_name	Description
0	17261	yrbA	orf, hypothetical protein (VIMSS)
1	16347	glpT	sn-glycerol-3-phosphate transporter (NCBI)
2	18130	phnF	predicted DNA-binding transcriptional regulator of phosphonate uptake and biodegradation (NCBI)
3	15217	fabF	3-oxoacyl-(acyl carrier protein) synthase (NCBI)
4	17788	pstS	phosphate transporter subunit (NCBI)
5	14288	folK	2-amino-4-hydroxy-6-hydroxymethyldihyropteridine pyrophosphokinase (NCBI)
6	18132	phnE	phosphonate/organophosphate ester transporter subunit (NCBI)
7	17785	pstB	phosphate transporter subunit (NCBI)
8	18129	phnG	carbon-phosphorus lyase complex subunit (NCBI)

Future Directions

- Increase number of AlphaFold predictions to represent a range of GLS coefficients
- Decide on consistent AlphaFold modeling method (AlphaFold3 is superior)
- Analyze frequency of protein-protein interactions in GLS top scoring gene pairs vs. lower scoring pairs– is this statistically significant?
- Obtain a deeper understanding of the fitness data and brainstorm how to best leverage it for finding hidden connections between genes
- Literature search of top-interacting pairs and connection back to their fitness results