Supplementary Material for Zhou et al., Translationally optimal codons associate with structurally sensitive sites in proteins

Table S1: Codon optimality (C_{opt}) and corresponding odds ratio (O_{buried}) of use frequency between buried and exposed sites

		E. coli		*S. cerevisiae*		*D. melanogaster*		*M. musculus*	
Amino acid	Codon	C_{opt}	O_{buried}	C_{opt}	O_{buried}	C_{opt}	O_{buried}	C_{opt}	O_{buried}
Ala	GCT	1.717	0.925	3.919	1.044	0.868	1.003	1.033	1.009
	GCC	0.639	0.974	1.318	1.001	1.351	1.129	1.041	1.059
	GCA	0.986	0.951	0.150	0.950	0.763	0.884	0.960	0.952
	GCG	0.900	1.120	0.122	0.978	0.912	0.896	0.905	0.938
Arg	CGT	2.934	1.069	1.554	1.169	1.090	1.093	1.159	1.103
	CGC	1.189	0.941	0.157	0.800	1.624	0.964	1.092	1.050
	CGA	0.276	1.040	0.042	0.917	0.668	1.044	1.205	1.076
	CGG	0.210	1.007	0.040	1.030	1.135	0.941	1.137	1.020
	AGA	0.144	0.976	4.273	1.005	0.493	0.992	0.782	0.913
	AGG	0.153	0.940	0.122	0.964	0.878	0.960	0.859	0.906
Asn	AAT	0.271	0.774	0.193	0.933	0.562	0.911	0.866	0.874
	AAC	3.693	1.292	5.188	1.072	1.781	1.098	1.154	1.144
Asp	GAT	0.455	0.945	0.485	1.130	0.714	1.082	0.934	0.913
	GAC	2.198	1.058	2.061	0.885	1.401	0.924	1.070	1.095
Cys	TGT	0.622	1.013	3.671	1.255	0.662	1.049	0.944	0.979
	TGC	1.608	0.987	0.272	0.797	1.510	0.953	1.060	1.022
Gln	CAA	0.507	0.855	6.547	1.039	0.582	0.969	0.804	0.832
	CAG	1.972	1.170	0.153	0.962	1.718	1.032	1.245	1.202
Glu	GAA	1.331	1.080	4.168	1.051	0.544	0.978	0.855	0.949
	GAG	0.751	0.926	0.240	0.951	1.838	1.023	1.169	1.054
Gly	GGT	1.840	1.089	10.496	1.404	0.957	1.107	1.018	1.013
	GGC	1.474	0.955	0.304	0.766	1.206	0.981	1.060	1.092
	GGA	0.234	0.905	0.120	0.759	0.901	0.958	0.888	0.950
	GGG	0.327	1.008	0.111	1.037	0.785	0.914	1.040	0.939
His	CAT	0.381	0.973	0.281	1.027	0.694	1.064	0.909	0.934
	CAC	2.622	1.028	3.561	0.974	1.441	0.940	1.100	1.071
Ile	ATT	0.539	1.008	1.261	1.126	0.763	0.957	0.937	1.058
	ATC	3.406	1.036	2.669	0.959	1.788	1.152	1.248	1.081
	ATA	0.130	0.876	0.108	0.867	0.567	0.835	0.736	0.808
Leu	CTT	0.542	0.891	0.280	1.029	0.682	1.002	0.950	0.960
	CTC	0.744	0.855	0.146	0.931	1.235	1.045	1.004	1.041
	CTA	0.285	0.975	0.619	0.979	0.791	0.891	0.950	0.952
	CTG	3.759	1.117	0.286	0.959	1.488	1.103	1.093	1.070
	TTA	0.301	1.017	0.676	0.938	0.475	0.920	0.868	0.895
	TTG	0.527	1.031	4.625	1.098	0.772	0.904	0.972	0.971
Lys	AAA	1.136	1.040	0.208	0.925	0.506	1.093	0.799	0.924
	AAG	0.881	0.962	4.811	1.081	1.975	0.915	1.252	1.082
Phe	TTT	0.285	0.995	0.251	0.930	0.526	0.860	0.906	0.943
	TTC	3.510	1.005	3.985	1.075	1.902	1.163	1.104	1.060
Pro	CCT	0.526	0.957	0.423	0.890	0.714	0.854	1.053	0.984
	CCC	0.278	0.977	0.144	1.001	1.262	1.077	0.998	1.136
	CCA	0.737	1.009	6.153	1.085	0.864	1.139	0.991	0.905
	CCG	2.883	1.033	0.126	1.059	1.083	0.901	0.913	0.971
Ser	TCT	2.360	1.224	2.819	1.109	0.686	1.041	0.983	1.083
	TCC	2.142	1.213	2.565	1.135	1.113	1.202	1.057	1.205
	TCA	0.418	1.038	0.331	1.019	0.712	1.070	0.950	0.986
	TCG	0.581	1.216	0.191	1.062	1.198	1.195	1.063	1.039
	AGT	0.343	0.888	0.364	0.803	0.688	0.887	0.952	0.838
	AGC	1.086	0.682	0.413	0.717	1.329	0.671	1.018	0.880
Thr	ACT	1.963	0.943	2.056	1.035	0.586	0.970	0.909	0.972
	ACC	1.886	1.191	2.640	0.963	1.862	1.065	1.106	1.124
	ACA	0.293	0.849	0.196	1.018	0.601	1.082	0.983	0.970
	ACG	0.413	0.929	0.126	0.971	1.099	0.887	0.988	0.882
Tyr	TAT	0.348	1.093	0.230	0.916	0.558	1.059	0.944	0.932
	TAC	2.876	0.915	4.340	1.092	1.792	0.944	1.060	1.073
Val	GTT	1.668	0.896	1.950	1.069	0.711	0.894	0.910	0.914
	GTC	0.579	1.072	2.696	1.015	1.237	1.005	1.026	1.008
	GTA	1.221	0.934	0.135	0.930	0.640	0.899	0.915	0.922
	GTG	0.707	1.100	0.221	0.936	1.225	1.117	1.079	1.098

Table S2: List of optimal codons

Amino acid	*E. coli*	*S. cerevisiae*	*D. melanogaster*	*M. musculus*
Ala	GCT	GCT, GCC	GCC	-
Arg	CGT, CGC	CGT, AGA	CGC	CGT, CGC, CGA, CGG
Asn	AAC	AAC	AAC	AAC
Asp	GAC	GAC	GAC	GAC
Cys	TGC	TGT	TGC	-
Gln	CAG	CAA	CAG	CAG
Glu	GAA	GAA	GAG	GAG
Gly	GGT, GGC	GGT	GGC	GGC
His	CAC	CAC	CAC	CAC
Ile	ATC	ATT, ATC	ATC	ATC
Leu	CTG	TTG	CTC, CTG	CTG
Lys	-	AAG	AAG	AAG
Phe	TTC	TTC	TTC	TTC
Pro	CCG	CCA	CCC	CCT
Ser	TCT, TCC	TCT, TCC	TCC, TCG, AGC	TCC
Thr	ACT, ACC	ACT, ACC	ACC	ACC
Tyr	TAC	TAC	TAC	-
Val	GTT, GTA	GTT, GTC	GTC, GTG	GTG

Table S3: Odds ratio of optimal codon usage between buried and exposed sites, for different solvent accessibility cutoffs

	E. coli			*S. cerevisiae*			*D. melanogaster*			*M. musculus*		
AA	5%	15%	35%	5%	15%	35%	5%	15%	35%	5%	15%	35%
Ala	0.94	0.96	$0.90^{*(*)}$	1.04	1.04	1.06	1.12^{***}	1.13^{***}	1.15^{***}	–	–	–
Arg	1.18	1.06	1.03	$1.46^{*(*)}$	$1.23^{(*)}$	$1.28^{**(*)}$	1.03	1.02	0.89^{**}	1.26^{***}	1.13^{**}	1.22^{***}
Asn	1.60^{***}	1.45^{***}	1.29^{***}	0.94	1.02	1.04	1.05	1.09	1.06	1.25^{***}	1.19^{***}	1.16^{***}
Asp	1.07	$1.12^{*(*)}$	1.06	$0.79^{*(*)}$	0.89	$0.89^{(*)}$	0.93	0.90^{*}	$0.93^{(*)}$	$1.12^{*(*)}$	1.10^{**}	$1.10^{**(*)}$
Cys	1.04	1.01	1.12	$1.35^{(*)}$	$1.45^{(*)}$	1.25	1.03	1.00	0.89	–	–	–
Gln	1.35^{***}	1.24^{***}	$1.18^{**(*)}$	0.96	1.01	1.04	0.98	1.08	1.01	1.42^{***}	1.28^{***}	1.20^{***}
Glu	1.11	1.10	1.09^{*}	1.10	1.04	1.06	0.97	1.01	1.02	1.09	1.08^{*}	1.06^{*}
Gly	1.04	1.07	1.05	1.64^{***}	1.49^{***}	1.47^{***}	1.03	1.01	0.99	1.13^{***}	1.12^{***}	1.08^{**}
His	1.14	1.06	1.06	1.04	1.01	0.92	$0.86^{(*)}$	0.97	0.95	1.09	1.11^{*}	1.11^{*}
Ile	0.99	1.04	1.04	1.08	$1.16^{(*)}$	1.17	1.01	$1.12^{**(*)}$	1.30^{***}	$1.10^{**(*)}$	$1.11^{**(*)}$	1.10^{*}
Leu	1.10^{***}	1.12^{***}	1.10^{*}	1.03	1.08	1.12	$1.09^{*(*)}$	$1.08^{*(*)}$	1.13^{**}	1.00	1.06^{*}	$1.05^{(*)}$
Lys	–	–	–	1.07	1.12	1.05	1.07	0.93	0.93	1.08	1.14^{**}	$1.10^{**(*)}$
Phe	1.06	1.04	0.99	1.02	1.13	1.01	1.28^{***}	$1.15^{*(*)}$	$1.24^{**(*)}$	$1.06^{(*)}$	1.08^{*}	1.07
Pro	1.02	1.08	1.08	1.08	1.12	1.05	1.04	1.11^{*}	$1.09^{(*)}$	1.06	1.01	1.00
Ser	1.49^{***}	1.49^{***}	1.32^{***}	1.25^{***}	1.24^{***}	1.25^{***}	1.03	1.01	0.98	1.19^{***}	1.27^{***}	1.26^{***}
Thr	1.22^{***}	1.17^{***}	$1.13^{*(*)}$	0.92	0.98	1.02	1.06	1.04	1.05	1.13^{***}	1.11^{***}	1.12^{***}
Tyr	0.97	0.97	0.93	0.95	1.02	0.98	0.93	0.88^{*}	0.98	–	–	–
Val	0.90^{**}	0.86^{***}	$0.88^{*(*)}$	$1.16^{*(*)}$	1.08	1.09	1.05	$1.14^{**(*)}$	1.15^{**}	1.04	$1.09^{**(*)}$	1.14^{***}
Overall	1.07^{***}	1.09^{***}	1.07^{***}	1.10^{***}	1.11^{***}	1.09^{***}	1.05^{***}	1.05^{***}	1.04^{***}	1.09^{***}	1.11^{***}	1.11^{***}

Note.—AA: amino acid; 5%: sites with solvent accessibility $< 5\%$ are considered as buried; 15%: sites with solvent accessibility $< 15\%$ are considered as buried; 35%: sites with solvent accessibility $< 35\%$ are considered as buried; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S4: Odds ratio of optimal codon usage between buried and exposed sites at evolutionary conserved sites only

AA	*E. coli*	*S. cerevisiae*	*D. melanogaster*	*M. musculus*
Ala	0.95	1.08	$1.09^{*(*)}$	–
Arg	0.99	1.16	0.92	1.12^{**}
Asn	1.39^{***}	1.05	1.09	1.17^{***}
Asp	1.08	$0.89^{(*)}$	0.90^{*}	1.12^{**}
Cys	0.86	$1.62^{(*)}$	1.05	–
Gln	1.12	0.99	0.99	1.20^{***}
Glu	1.07	1.02	1.09	1.06
Gly	1.07	1.45^{***}	0.99	1.10^{**}
His	1.02	0.97	$0.82^{*(*)}$	1.03
Ile	1.05	1.01	1.09	1.05
Leu	$1.14^{**(*)}$	1.09	$1.12^{*(*)}$	1.02
Lys	–	1.12	$0.86^{*(*)}$	1.10^{**}
Phe	1.03	1.08	1.16^{*}	1.08
Pro	1.03	$1.14^{(*)}$	1.11^{*}	1.00
Ser	1.49^{***}	$1.16^{(**)}$	1.05	1.22^{***}
Thr	1.21^{***}	0.99	1.03	1.14^{***}
Tyr	0.97	1.09	0.93	–
Val	$0.89^{*(*)}$	1.09	1.24^{***}	1.08^{*}
Overall	1.08^{***}	1.09^{***}	1.03^{**}	1.10^{***}

Note.—AA: amino acid; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S5: Odds ratio of optimal codon usage between buried and exposed sites, calculated after excluding all ribosomal proteins from the data set

AA	*E. coli*	*S. cerevisiae*	*D. melanogaster*	*M. musculus*
Ala	0.91^{*}	1.06	1.14^{***}	–
Arg	1.01	$1.23^{*(*)}$	0.96	1.19^{***}
Asn	1.36^{***}	1.09	$1.13^{*(*)}$	1.17^{***}
Asp	1.07	$0.87^{(*)}$	0.92^{*}	1.11^{***}
Cys	0.97	1.35	0.92	–
Gln	1.21^{***}	1.05	1.04	1.26^{***}
Glu	$1.09^{(*)}$	1.06	1.02	$1.06^{(*)}$
Gly	1.06	1.46^{***}	0.98	1.11^{***}
His	1.03	0.97	$0.92^{*(*)}$	$1.09^{(*)}$
Ile	1.05	$1.20^{(*)}$	1.20^{***}	1.10^{*}
Leu	$1.13^{**(*)}$	$1.11^{(*)}$	1.16^{***}	1.08^{**}
Lys	–	1.09	0.90^{*}	1.10^{**}
Phe	1.01	1.11	1.19^{**}	1.07
Pro	1.04	1.08	1.09^{*}	0.98
Ser	1.42^{***}	1.27^{***}	1.02	1.25^{***}
Thr	1.20^{***}	1.00	$1.07^{(*)}$	1.14^{***}
Tyr	$0.88^{(*)}$	1.10	0.93	–
Val	0.85^{***}	1.12	1.19^{***}	1.11^{***}
Overall	1.08^{***}	1.11^{***}	1.05^{***}	1.12^{***}

Note.—AA: amino acid; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

Table S6: Odds ratio of optimal codon usage between structurally important and unimportant sites

	E. coli		*S. cerevisiae*		*D. melanogaster*		*M. musculus*	
AA	Whole	Nonlinker	Whole	Nonlinker	Whole	Nonlinker	Whole	Nonlinker
Ala	1.00	1.03	1.07	1.03	1.00	0.99	–	–
Arg	0.98	0.94	1.13	1.02	1.02	0.96	$1.14^{**(*)}$	$1.17^{**(*)}$
Asn	$1.17^{*(*)}$	$1.15^{(*)}$	1.02	1.13	$1.13^{(*)}$	$1.18^{*(*)}$	$1.09^{(*)}$	$1.14^{*(*)}$
Asp	1.01	1.01	0.98	0.95	1.00	1.03	1.02	1.02
Cys	0.84	0.92	$1.49^{*(*)}$	1.31	$1.22^{(*)}$	1.13	–	–
Gln	$1.12^{(*)}$	1.04	$1.22^{(*)}$	$1.28^{(*)}$	1.01	0.92	1.03	1.07
Glu	$1.13^{*(*)}$	1.11	$1.15^{(*)}$	1.15	$1.09^{(*)}$	1.06	1.00	1.03
Gly	0.99	0.97	1.33^{***}	$1.21^{(**)}$	1.04	1.06	1.06	1.03
His	0.97	1.02	1.09	1.07	1.06	1.05	1.02	1.03
Ile	1.00	1.00	1.10	1.06	1.06	1.11	$1.13^{**(*)}$	1.15^{**}
Leu	1.04	1.00	$1.16^{*(*)}$	$1.26^{**(*)}$	1.06	$1.12^{*(*)}$	1.07^{*}	$1.08^{(*)}$
Lys	–	–	1.01	0.96	0.92	0.97	1.11^{**}	1.13^{**}
Phe	1.14	1.03	1.09	1.09	$1.18^{(*)}$	1.07	$1.14^{*(*)}$	1.18^{**}
Pro	1.03	1.02	1.10	1.06	1.01	0.91	1.05	1.09
Ser	1.30^{***}	$1.26^{**(*)}$	$1.14^{(*)}$	1.13	1.04	1.06	$1.10^{*(*)}$	1.12^{*}
Thr	$1.13^{(*)}$	$1.15^{(*)}$	0.93	0.93	$1.09^{(*)}$	1.02	1.05	1.02
Tyr	0.88	0.87	0.96	0.97	0.97	0.97	–	–
Val	$0.90^{*(*)}$	$0.92^{(*)}$	1.07	1.00	1.24^{***}	$1.16^{*(*)}$	1.07^{*}	$1.08^{(*)}$
Overall	1.04^{**}	1.03	1.09^{***}	1.07^{***}	1.05^{***}	1.04^{***}	1.07^{***}	1.08^{***}

Note.—AA: amino acid; Whole: odds ratio for whole protein sequences; Nonlinker: odds ratio for sequences without domain boundary region; -: no optimal codon. Significance levels: $^{***}P < 0.001$; $^{**}P < 0.01$; $^{*}P < 0.05$. Significance levels in parentheses disappear after correction for multiple testing.

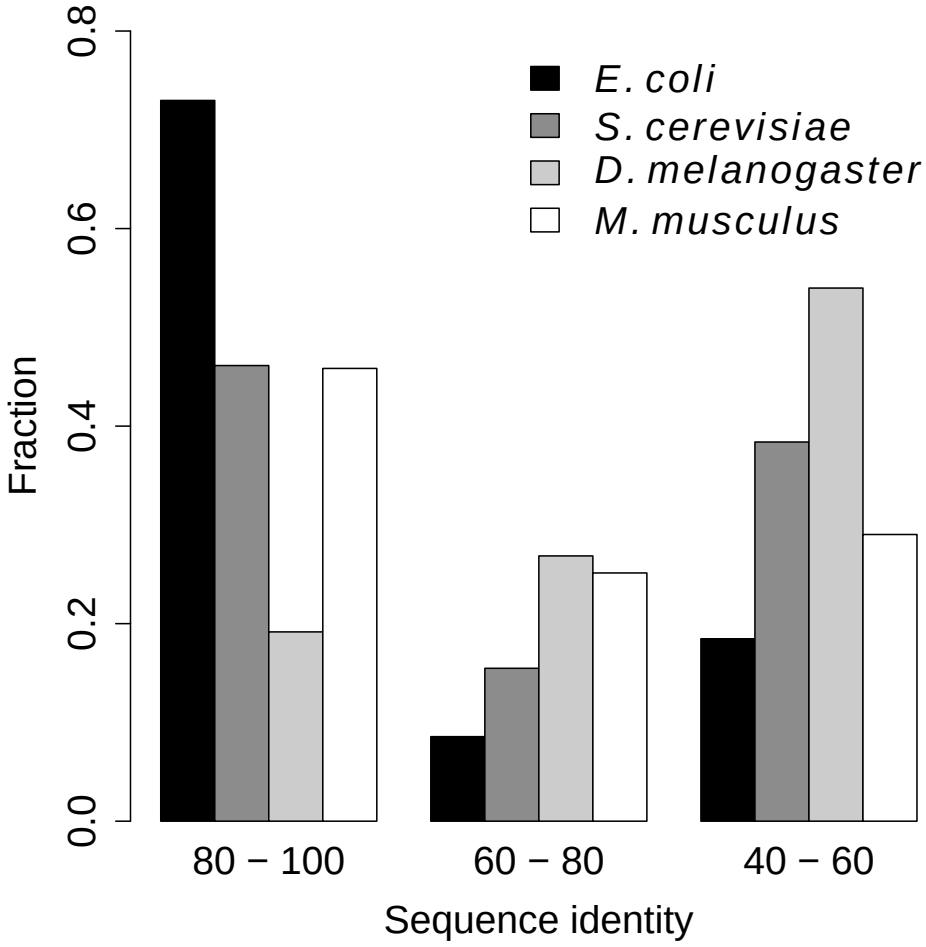

Figure S1: Distribution of sequence identities between gene sequences and associated sequences of protein crystal structures for each species. *E. coli* has the largest fraction of sequences with closely-related crystal structures, while fly has the smallest fraction.

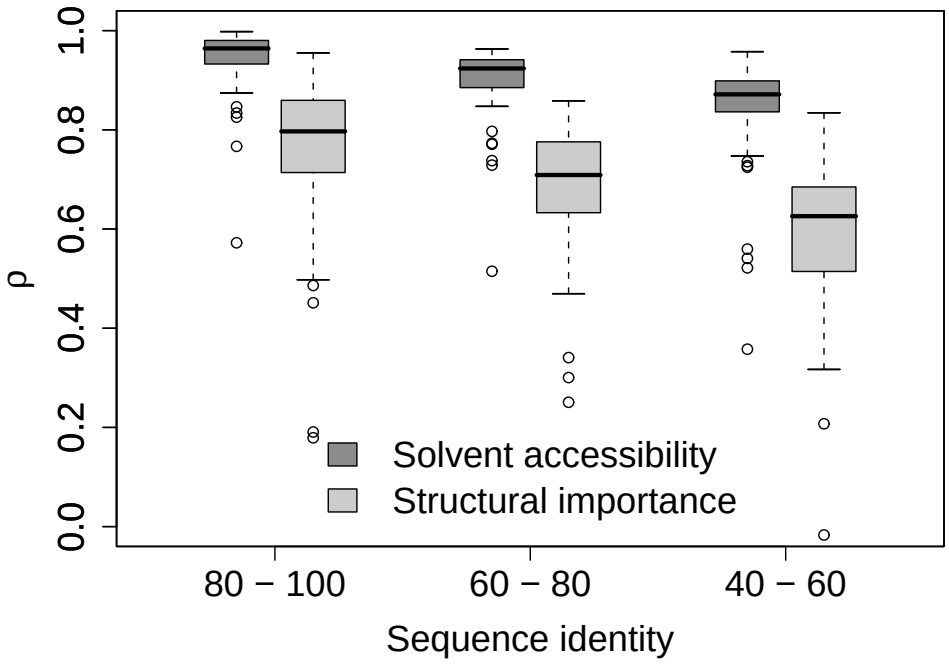

Figure S2: Conservation of solvent accessibility and structural importance as a function of sequence diversity. In each sequence identity group, for each quantity (solvent accessibility or structural importance), and for each representative protein, we calculated the Spearman correlation ρ of the quantity between the homologous sites of the representative protein and those of its homolog. The box plots show the distribution of ρ values in each sequence-identity group and for each quantity.

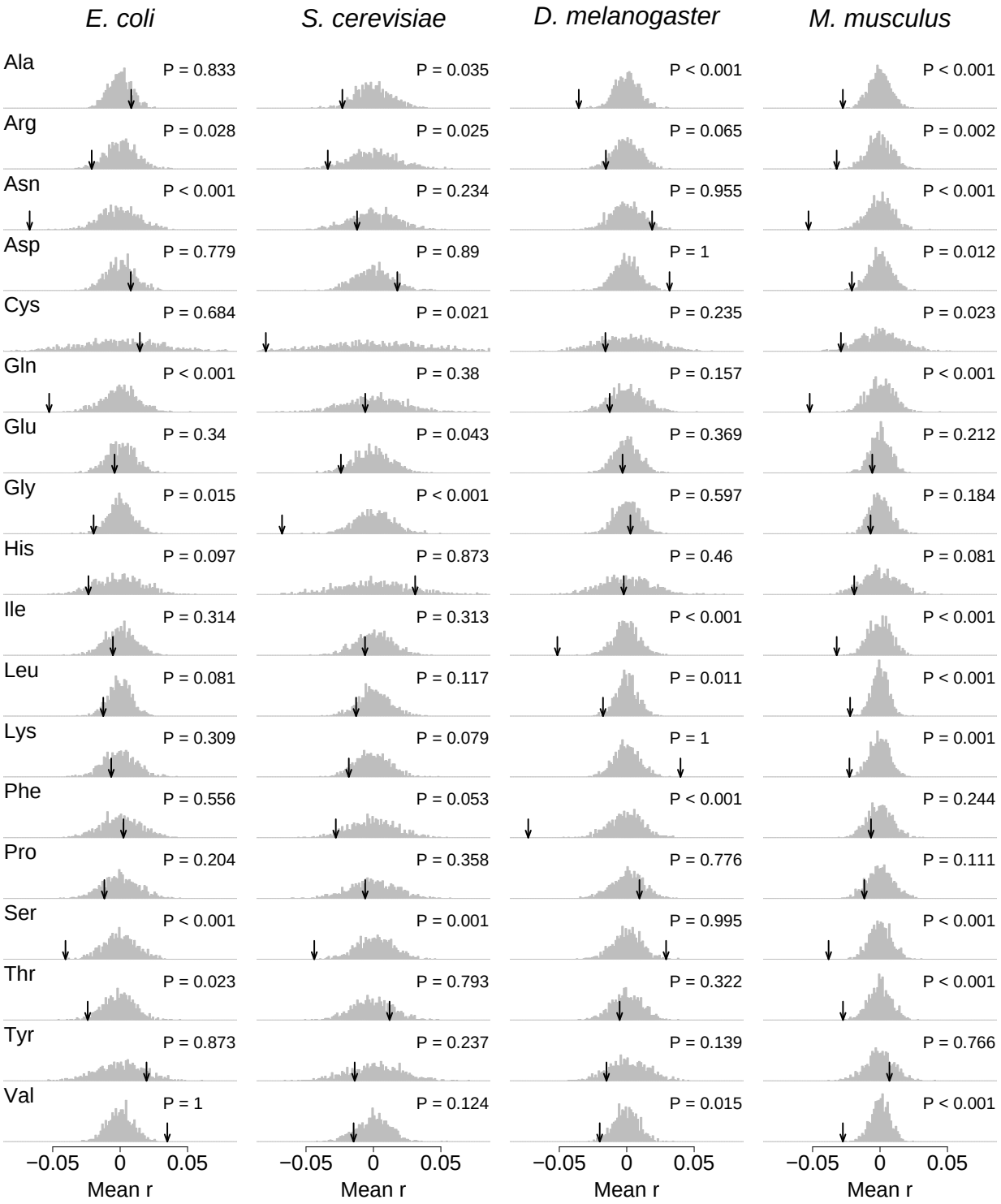

Figure S3: Test for association between codon optimality and solvent accessibility, separately for each amino acid. The black arrows indicate the mean correlation coefficient between the two quantities over all genes. The gray histograms show the sampling distribution of the same quantity under the null hypothesis of no association. P values are one-sided and not corrected for multiple testing.

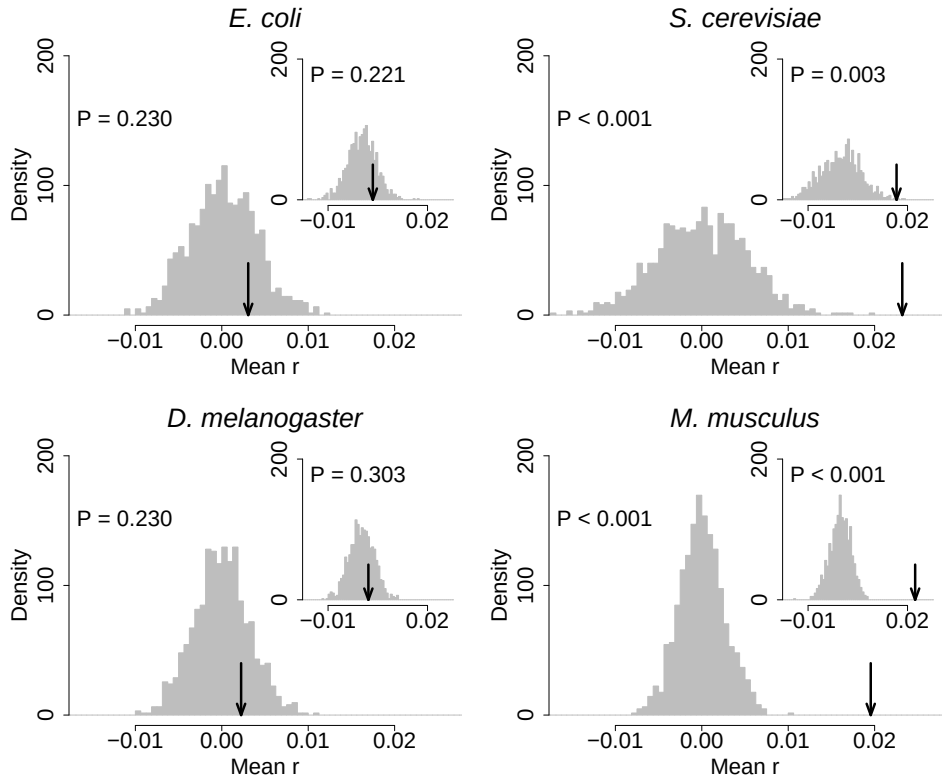

Figure S4: Test for association between codon optimality and structural importance. The black arrows indicate the mean correlation coefficient between these two quantities over all amino acids and all genes. The gray histograms show the sampling distribution of the same quantity under the null hypothesis of no association. The main figure of each panel shows the results for complete genes, and the inset shows the results when interdomain linker regions are excluded.

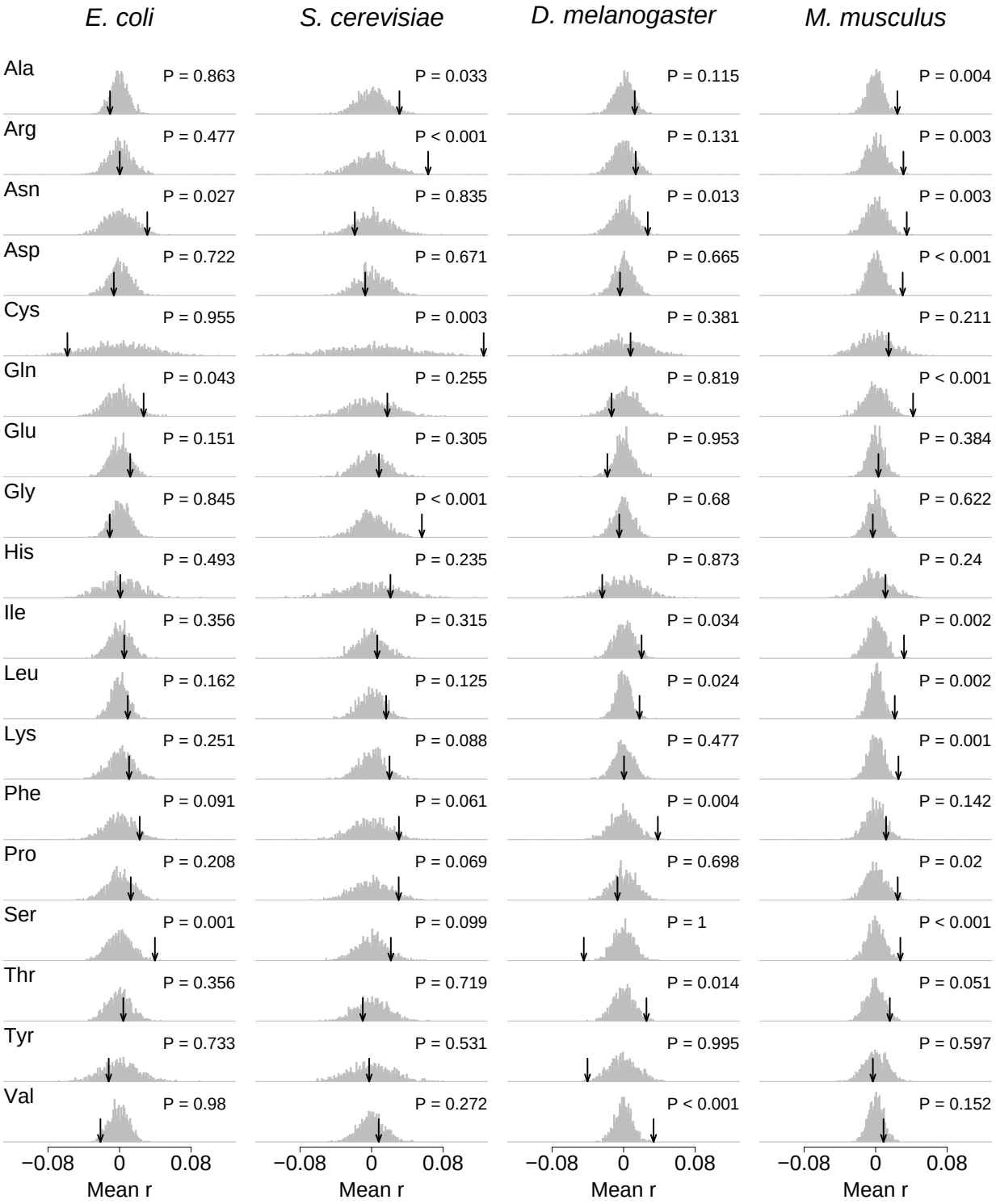

Figure S5: Test for association between codon optimality and structural importance, separately for each amino acid. The black arrows indicate the mean correlation coefficient between the two quantities over all genes. The gray histograms show the sampling distribution of the same quantity under the null hypothesis of no association. P values are one-sided and not corrected for multiple testing.

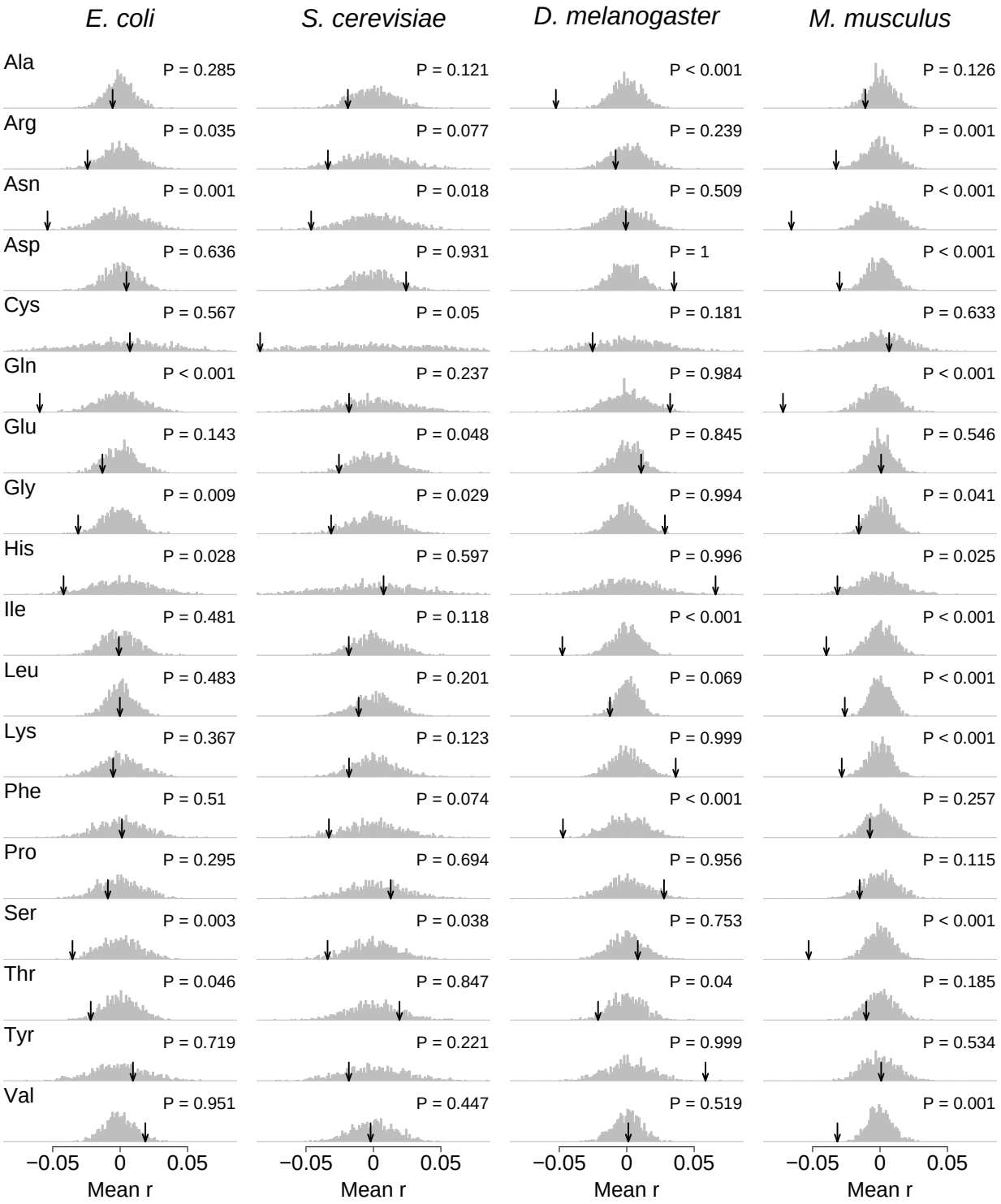

Figure S6: Test for association between codon optimality and solvent accessibility, separately for each amino acid and excluding inter-domain linker regions. The black arrows indicate the mean correlation coefficient between the two quantities over all genes. The gray histograms shows the sampling distribution of the same quantity under the null hypothesis of no association. P values are one-sided and not corrected for multiple testing.

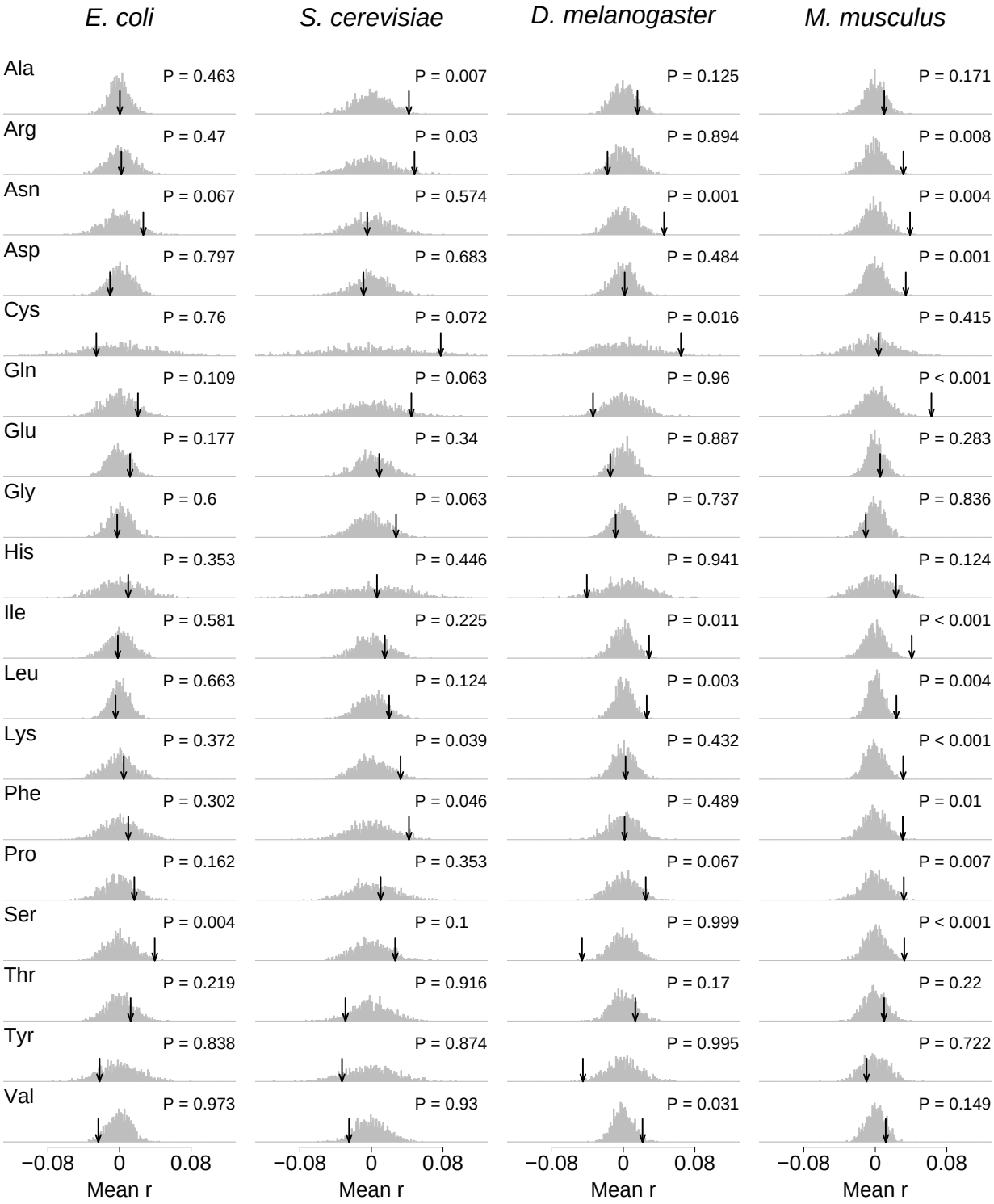

Figure S7: Test for association between codon optimality and structural importance, separately for each amino acid and excluding inter-domain linker regions. The black arrows indicate the mean correlation coefficient between the two quantities over all genes. The gray histograms show the sampling distribution of the same quantity under the null hypothesis of no association. P values are one-sided and not corrected for multiple testing.