Supplementary Materials

Table of content

Supporting Figures

Supporting Tables

DIAMONDScore. *p*-values in parenthesis are calculated between ATGO and other single-based methods and between ATGO+ and other composite methods by two-sided Student's t-test. Specifically, the proposed ATGO and ATGO+ are repeatedly implemented with 10 times on the benchmark dataset to generate the corresponding performance evaluation indices, which are compared with the fixed evaluation index generated by the competing method to calculate *p*-value using two-sided Student's t-test. Bold fonts highlight the best performer in each category.

- Table S5. The ICW-Fmax values of 12 GO prediction methods on the 1068 benchmark proteins. Bold fonts highlight the best performer in each category.
- Table S6. The *p*-values of performance difference between 10 GO prediction methods on 3328 individual CAFA3 targets under post-hoc Nemenyi test at the individual protein level, where the performance of each prediction method is measured by a group of F1-scores, each of which is calculated from the predicted GO terms and native GO annotation in a single test protein. Because the *p*-values can be only approximated in the range from 1.0e-03 to 9.0e-01 under post-hoc Nemenyi test using Python package, the numerical value of 1.0e-03 (or 9.0e-01) means that the *p*-value is below to 1.0e-03 (or upon to 9.0e-01).
- Table S7. The ICW-Fmax values of 10 GO prediction methods on 3328 CAFA3 targets where a sequence identity cut-off $t_1 = 30\%$ between the training and testing proteins was applied to the five in-house methods (ATGO, ATGO+, SAGP, PPIGP and NGP). Bold fonts highlight the best performer in each category.
- Table S8. The performance of ten GO prediction methods under the cut-off $t_1 = 30\%$ on 1177 no-knowledge (NK) and 2151 limited-knowledge (LK) CAFA3 proteins. Bold fonts highlight the best performer in each category.
- Table S9. The numbers of proteins for 20 species in CAFA3 test dataset.
- Table S10. The prediction performance of five GO prediction methods under the cut-off $t_1 = 100\%$ on CAFA3 test proteins. Bold fonts highlight the best performer in each category.
- Table S11. The prediction performance of SAGP and BLAST baseline on our constructed test dataset and CAFA3 test dataset with different cut-off values of sequence identity. Bold fonts highlight the best performer in each category.
- Table S12. The prediction performance of ATGO models via four metric learning methods on two test datasets. Bold fonts highlight the best performer in each category.
- Table S13. The incorrectly predicted GO terms for twelve methods on three proteins in BP aspect.
- Table S14. The numbers of proteins and GO terms in benchmark dataset.
- Table S15. The values of $margin$, c_f, α, and K for three GO aspects.

Supporting Texts

- Text S1. Sequence alignment-based GO prediction.

- Text S2. Protein-protein interaction-based GO prediction.
- Text S3. Naïve-based GO prediction.
- Text S4. Friedman and Nemenyi post-hoc tests at the individual protein level.
- Text S5. An explanation for the difference of *p*-value calculations between Student t-test and Nemenyi post-hoc test.
- Text S6. Information content-weighted maximum F1-score.
- Text S7. Performance comparison between SAGP and BLAST baseline used in CAFA challenge.
- Text S8. Performance comparison between four metric learning methods.
- Text S9. The mathematics formulas for ESM-1b transformer.
- Text S10. The functional similarity between two proteins.

Supporting Figures

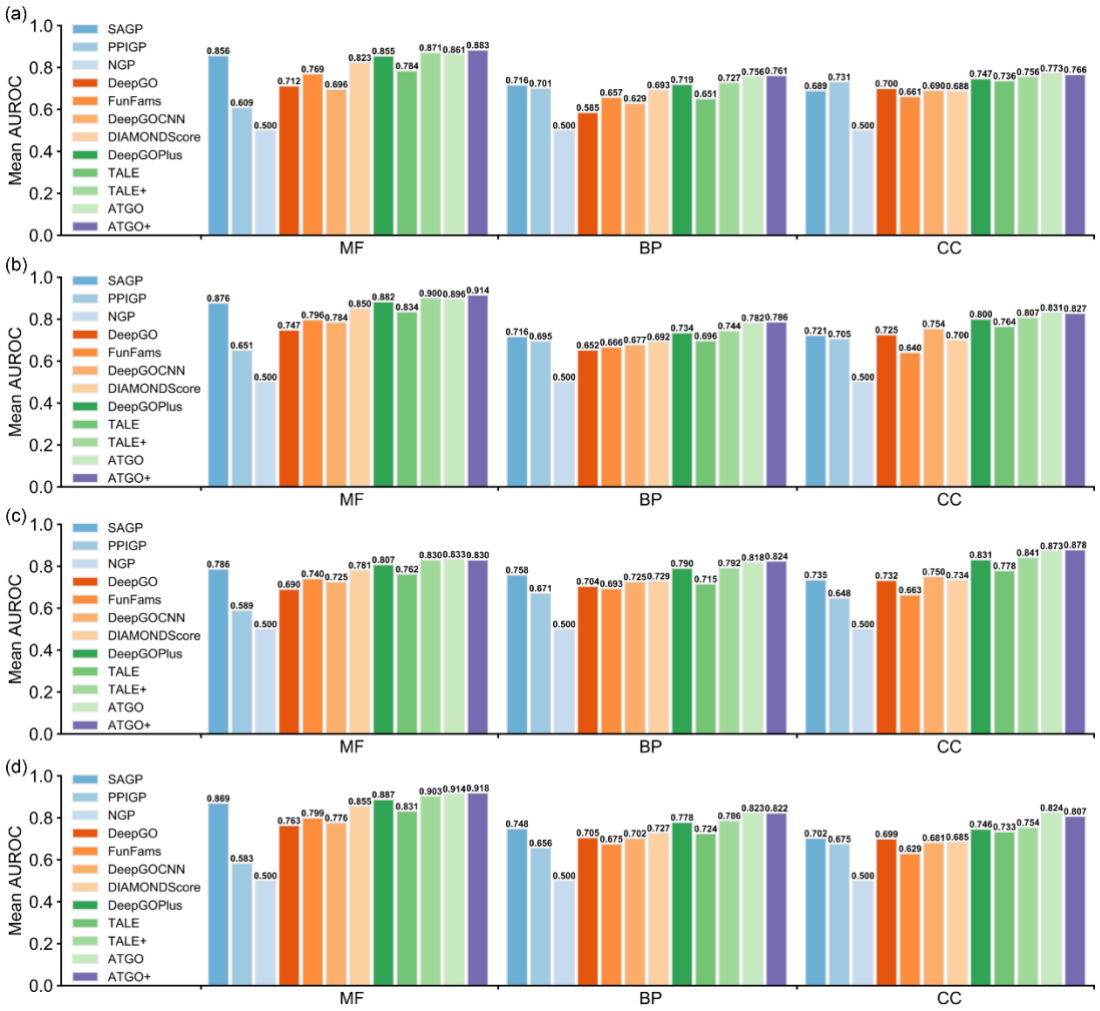

Fig S1. The mean AUROC values of GO terms versus 12 GO prediction methods in four ranges. **(a)** range 5-10. **(b)** range 10-30. **(c)** range 30-50. **(d)** range >50.

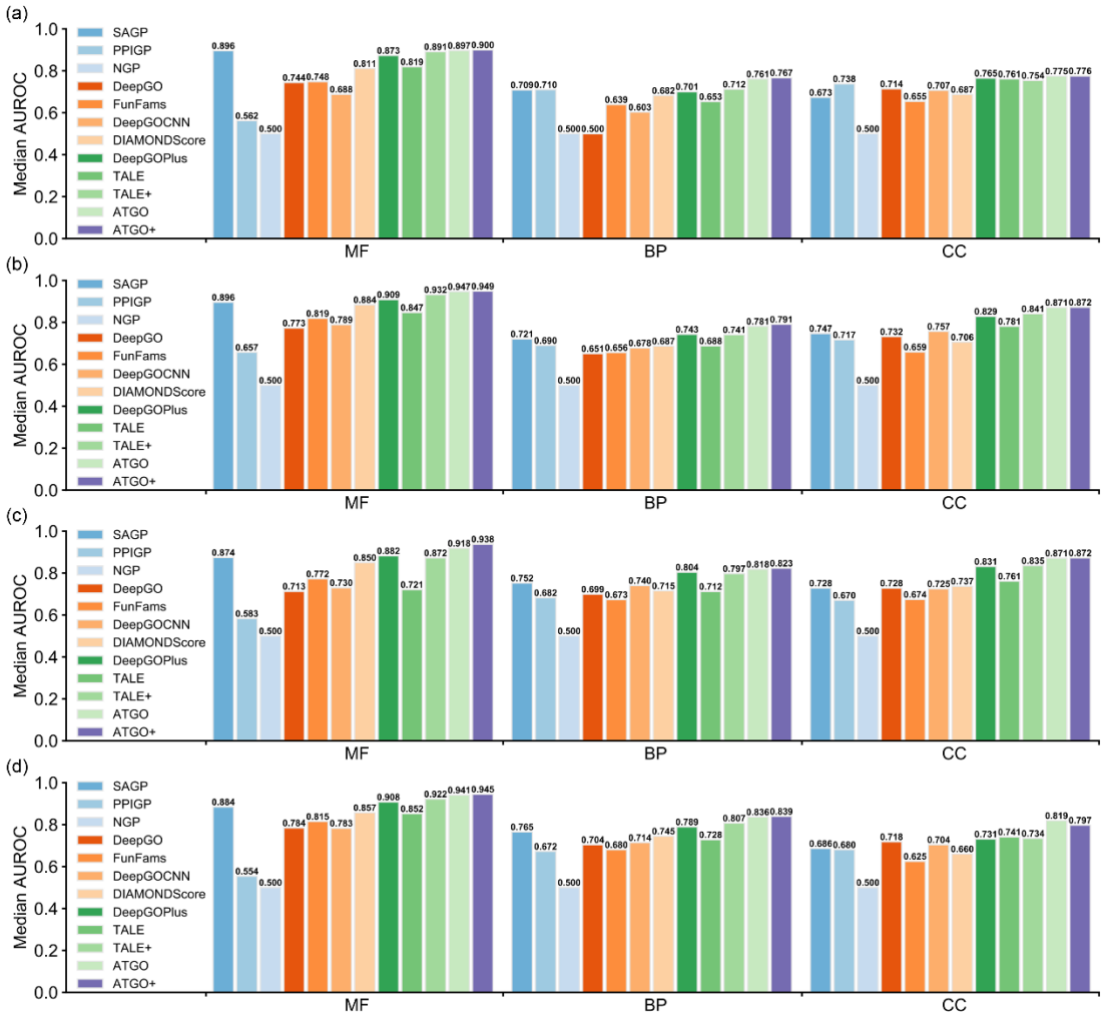

Fig S2. The median AUROC values of GO terms versus 12 GO prediction methods in four ranges. **(a)** range 5-10. **(b)** range 10-30. **(c)** range 30-50. **(d)** range >50.

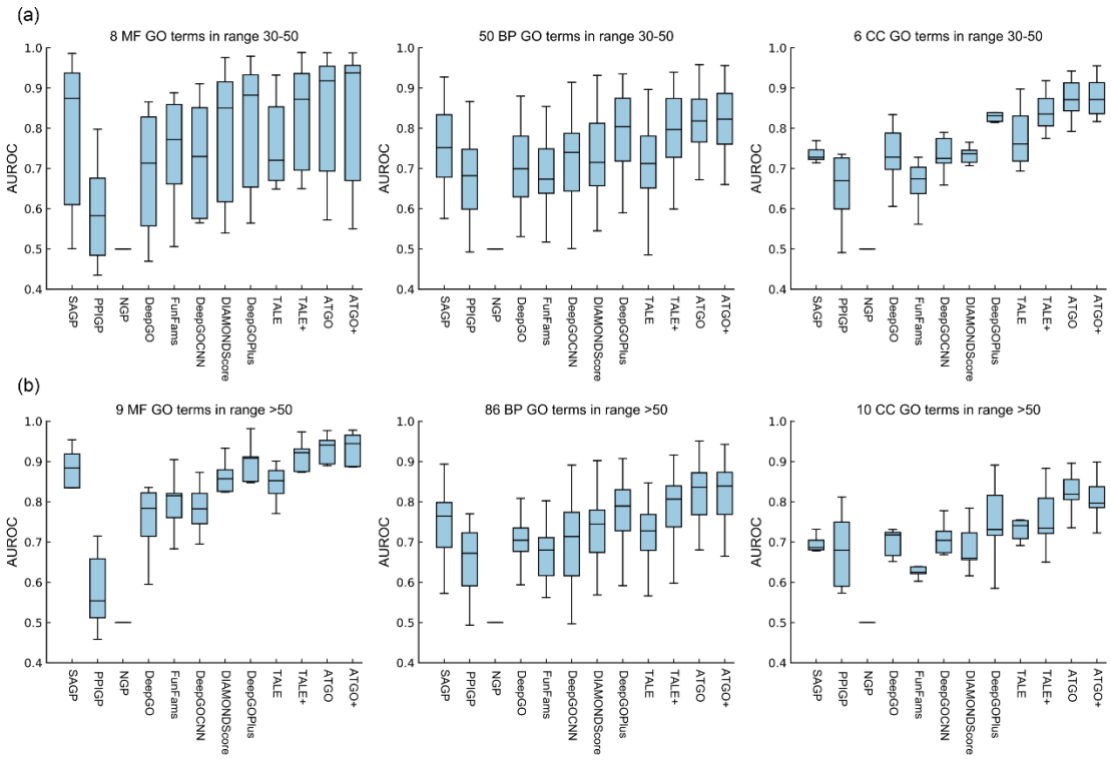

Fig S3. The distributions of AUROC values for GO terms in two ranges versus 12 GO prediction methods, where the median line in the box is the median AUROC value. **(a)** range 30-50. **(b)** range >50.

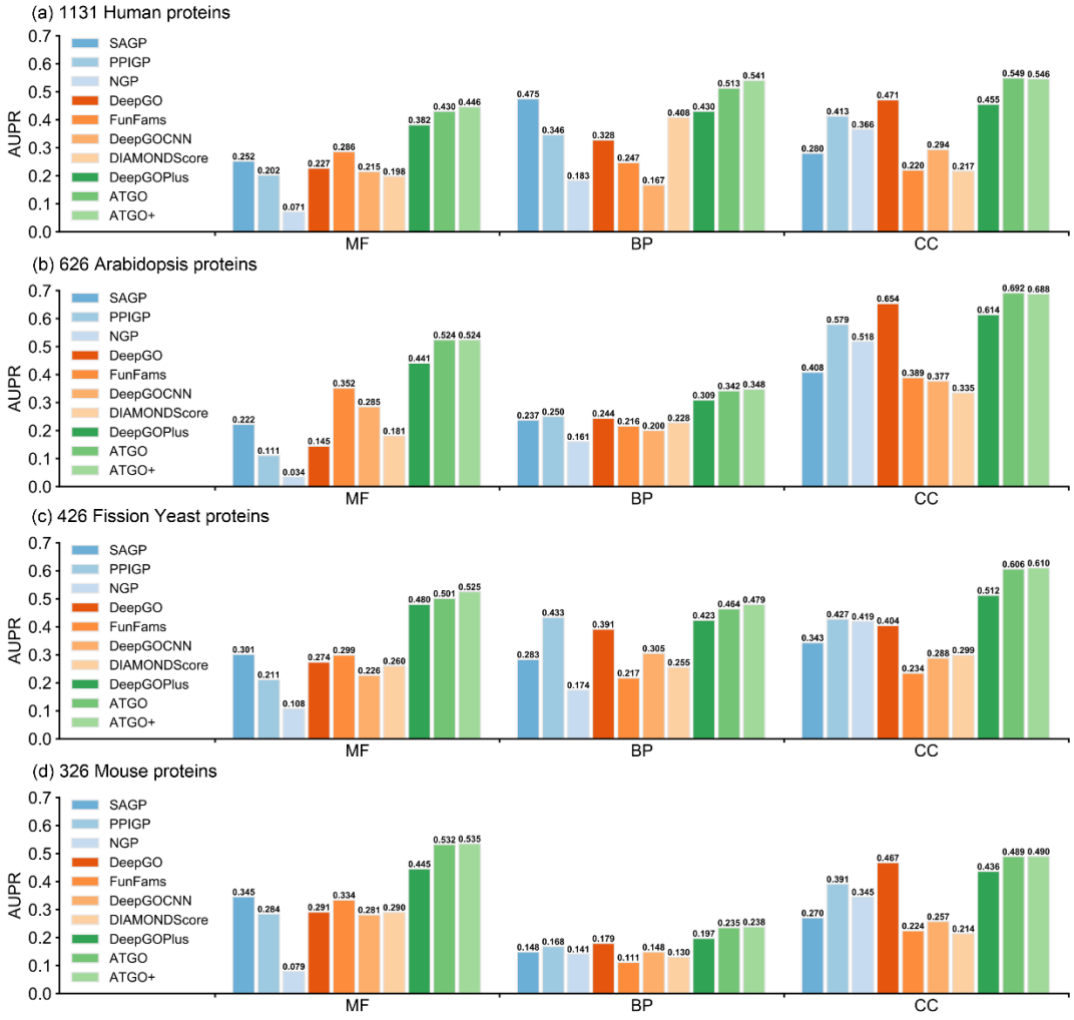

Fig S4. The AUPR values of 10 GO prediction methods under the sequence identity cut-off $t_1 = 30\%$ for three GO aspects on four individual species in CAFA3 test dataset. **(a)** Human **(b)** Arabidopsis **(c)** Fission Yeast **(d)** Mouse.

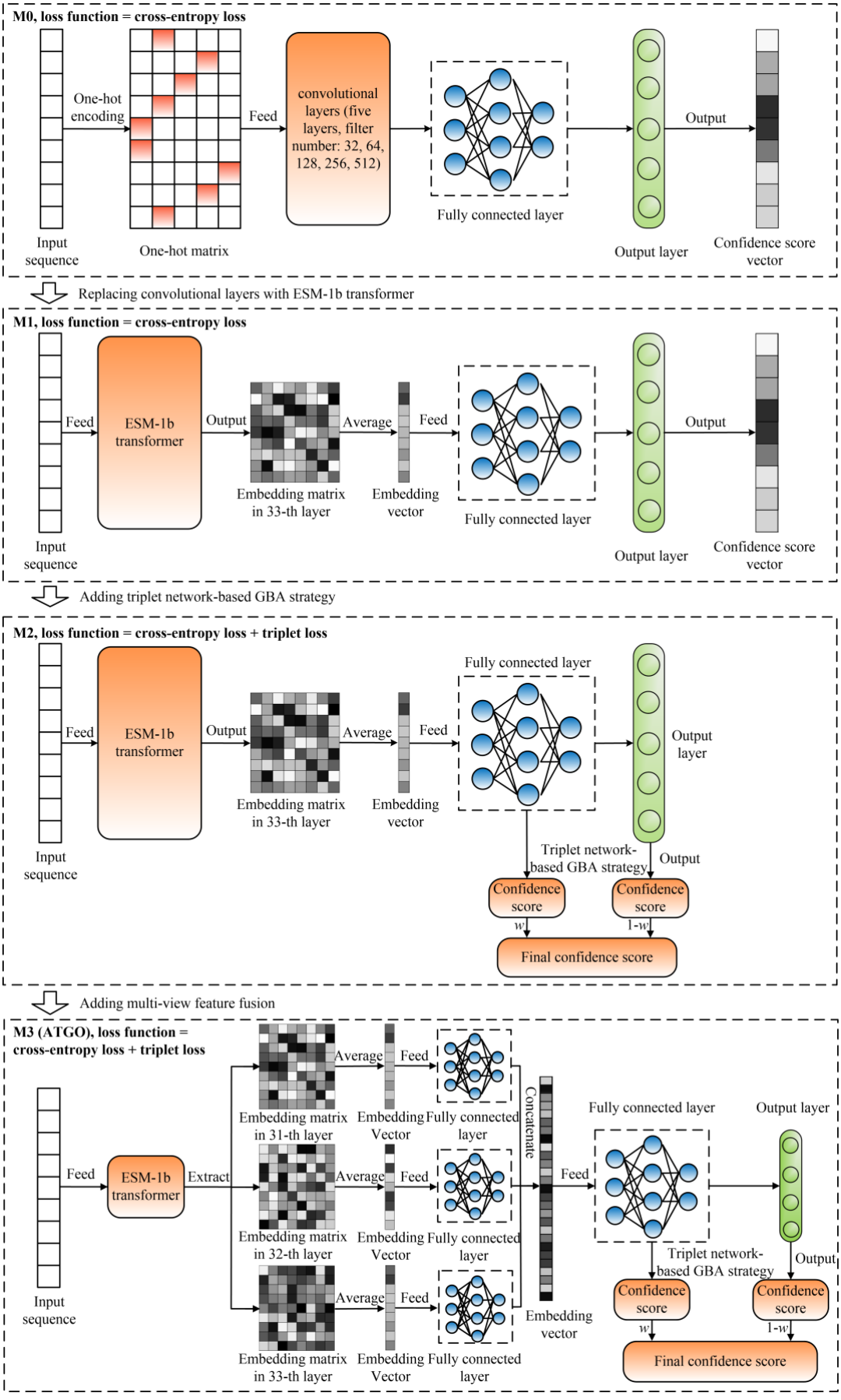

Fig S5. The architectures of different models in ablation study.

Supporting Tables

Table S1. The *p*-values of performance difference between 12 GO prediction methods on 1068 individual test proteins under post-hoc Nemenyi test at the individual protein level, where the performance of each prediction method is measured by a group of F1-scores, each of which is calculated from the predicted GO terms and native GO annotation in a single test protein. Because the *p*-values can be only approximated in the range from 1.0e-03 to 9.0e-01 under post-hoc Nemenyi test using Python package, the numerical value of 1.0e-03 (or 9.0e-01) means that the *p*-value is below to 1.0e-03 (or upon to 9.0e-01).

	Method	SAGP	PPIGP	NGP	DeepGO	FunFams	DeepGOCNN	DIAMONDScore	TALE	ATGO	DeepGOPlus	TALE+	ATGO+
MF	SAGP	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	2.5e-01	9.0e-01	1.0e-03	1.5e-01
	PPIGP	1.0e-03	1.0e+00	8.8e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	NGP	1.0e-03	8.8e-02	1.0e+00	1.0e-03	1.0e-03	6.3e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGO	1.0e-03	1.0e-03	1.0e-03	1.0e+00	9.0e-01	9.0e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	FunFams	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e+00	6.1e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.0e-03	1.0e-03	6.3e-02	9.0e-01	6.1e-01	1.0e+00	1.0e-03	1.9e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DIAMONDScore	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	2.1e-02	9.0e-01	1.0e-03	9.9e-03
	TALE	1.0e-03	1.0e-03	1.0e-03	9.0e-01	9.0e-01	1.9e-01	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	ATGO	2.5e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	2.1e-02	1.0e-03	1.0e+00	3.0e-01	1.0e-03	9.0e-01
	DeepGOPlus	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	3.0e-01	1.0e+00	1.0e-03	1.8e-01
	TALE+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.5e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.9e-03	1.0e-03	9.0e-01	1.8e-01	1.0e-03	1.0e+00
BP	SAGP	1.0e+00	1.0e-03	1.4e-02	8.9e-01	1.0e-03	1.9e-03	3.5e-01	9.0e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	PPIGP	1.0e-03	1.0e+00	1.6e-03	1.0e-03	1.0e-03	1.2e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	NGP	1.4e-02	1.6e-03	1.0e+00	6.7e-01	1.0e-03	9.0e-01	9.0e-01	2.0e-01	1.0e-03	1.1e-03	1.0e-03	1.0e-03
	DeepGO	8.9e-01	1.0e-03	6.7e-01	1.0e+00	1.0e-03	3.3e-01	9.0e-01	9.0e-01	1.0e-03	4.8e-01	1.0e-03	1.0e-03
	FunFams	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.9e-03	1.2e-02	9.0e-01	3.3e-01	1.0e-03	1.0e+00	8.8e-01	5.2e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DIAMONDScore	3.5e-01	1.0e-03	9.0e-01	9.0e-01	1.0e-03	8.8e-01	1.0e+00	9.0e-01	1.0e-03	7.5e-02	1.0e-03	1.0e-03
	TALE	9.0e-01	1.0e-03	2.0e-01	9.0e-01	1.0e-03	5.2e-02	9.0e-01	1.0e+00	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	ATGO	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	9.0e-01
	DeepGOPlus	9.0e-01	1.0e-03	1.1e-03	4.8e-01	1.0e-03	1.0e-03	7.5e-02	9.0e-01	1.0e-03	1.0e+00	1.0e-03	1.0e-03
	TALE+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	1.0e-03	1.0e+00
CC	SAGP	1.0e+00	1.0e-03	9.0e-01	9.0e-01	1.0e-03	1.5e-03	3.1e-01	9.0e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	PPIGP	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	3.2e-01	1.6e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	NGP	9.0e-01	1.0e-03	1.0e+00	9.0e-01	1.0e-03	3.0e-02	8.2e-01	9.0e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	DeepGO	9.0e-01	1.0e-03	9.0e-01	1.0e+00	1.0e-03	8.5e-02	9.0e-01	9.0e-01	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	FunFams	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.5e-03	3.2e-01	3.0e-02	8.5e-02	1.0e-03	1.0e+00	8.8e-01	1.6e-02	1.0e-03	4.4e-02	1.0e-03	1.0e-03
	DIAMONDScore	3.1e-01	1.6e-03	8.2e-01	9.0e-01	1.0e-03	8.8e-01	1.0e+00	7.0e-01	1.0e-03	8.9e-01	1.0e-03	1.0e-03
	TALE	9.0e-01	1.0e-03	9.0e-01	9.0e-01	1.0e-03	1.6e-02	7.0e-01	1.0e+00	1.0e-03	9.0e-01	1.0e-03	1.0e-03
	ATGO	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	9.0e-01
	DeepGOPlus	9.0e-01	1.0e-03	9.0e-01	9.0e-01	1.0e-03	4.4e-02	8.9e-01	9.0e-01	1.0e-03	1.0e+00	1.0e-03	1.0e-03
	TALE+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	1.0e-03	1.0e+00

Table S2. The statistic values between SAGP and ATGO in Group A under Nemenyi post-hoc test on 1068 test proteins for MF aspect versus the increase of K.

K	2	3	4	5	6	7	8	9	10	11	12
DAR	0.1525	0.1802	0.2149	0.2452	0.2539	0.2886	0.3544	0.4133	0.4818	0.4896	0.5641
$\sqrt{K(K+1)/6N}$	0.0416	0.0589	0.0760	0.0931	0.1101	0.1272	0.1442	0.1612	0.1783	0.1953	0.2123
q_α	3.6635	3.0615	2.8274	2.6344	2.3052	2.2689	2.4576	2.5636	2.7029	2.5074	2.6575
p-value	1.0e-03	6.2e-03	2.4e-02	6.4e-02	1.9e-01	2.6e-01	2.1e-01	2.0e-01	1.7e-01	3.0e-01	2.5e-01

Table S3. The prediction performance with including root GO terms for ATGO and TALE on all 1068 test proteins. *p*-values in parenthesis are calculated between ATGO and TALE by two-sided Student's t-test. Specifically, the proposed ATGO is repeatedly implemented with 10 times on the benchmark dataset to generate the corresponding performance evaluation indices, which are compared with the fixed evaluation index generated by TALE to calculate *p*-value using two-sided Student's t-test. Bold fonts highlight the best performer in each category.

Method	Fmax			AUPR		
	MF	**BP**	**CC**	**MF**	**BP**	**CC**
TALE	0.549 (7.3e-16)	0.361 (2.4e-16)	0.600 (1.1e-16)	0.383 (2.6e-18)	0.254 (1.4e-18)	0.438 (4.7e-17)
ATGO	**0.688**	**0.465**	**0.695**	**0.689**	**0.413**	**0.654**

Table S4. The summary of the proposed ATGO/ATGO+ and other ten competing GO prediction methods on a subset of 562 test proteins which have available templates or interaction partners in all of SAGP, PPIGP, FunFams, and DIAMONDScore. *p*-values in parenthesis are calculated between ATGO and other single-based methods and between ATGO+ and other composite methods by two-sided Student's t-test. Specifically, the proposed ATGO and ATGO+ are repeatedly implemented with 10 times on the benchmark dataset to generate the corresponding performance evaluation indices, which are compared with the fixed evaluation index generated by the competing method to calculate *p*-value using two-sided Student's t-test. Bold fonts highlight the best performer in each category.

	Methods	**Fmax**			**AUPR**		
		MF	**BP**	**CC**	**MF**	**BP**	**CC**
Single algorithms	SAGP	0.637 (1.2e-06)	0.418 (4.6e-08)	0.598 (1.9e-10)	0.412 (5.2e-17)	0.274 (2.0e-15)	0.422 (8.5e-19)
	PPIGP	0.332 (2.9e-17)	0.387 (1.2e-12)	0.590 (4.1e-11)	0.209 (1.5e-19)	0.294 (1.9e-14)	0.537 (1.4e-15)
	NGP	0.246 (3.5e-18)	0.272 (1.2e-17)	0.525 (2.6e-14)	0.121 (2.7e-20)	0.174 (2.3e-18)	0.404 (4.0e-19)
	DeepGO	0.383 (1.3e-16)	0.347 (3.9e-15)	0.547 (1.8e-13)	0.318 (2.1e-18)	0.250 (2.3e-16)	0.495 (4.5e-17)
	FunFams	0.512 (3.9e-14)	0.343 (2.4e-15)	0.482 (1.5e-15)	0.325 (2.6e-18)	0.176 (2.5e-18)	0.293 (1.0e-20)
	DeepGOCNN	0.352 (5.2e-17)	0.309 (1.4e-16)	0.494 (3.1e-15)	0.282 (8.1e-19)	0.212 (1.7e-17)	0.366 (9.6e-20)
	DIAMONDScore	0.629 (7.3e-08)	0.405 (1.1e-10)	0.580 (8.6e-12)	0.322 (2.4e-18)	0.232 (6.1e-17)	0.316 (2.0e-20)
	TALE	0.397 (2.2e-16)	0.318 (2.7e-16)	0.527 (3.0e-14)	0.338 (3.9e-18)	0.243 (1.3e-16)	0.499 (6.0e-17)
	ATGO	0.662	0.439	0.645	0.647	0.373	0.633
Composite algorithms	DeepGOPlus	0.641 (9.8e-09)	0.412 (2.0e-10)	0.580 (6.9e-14)	0.581 (2.4e-13)	0.335 (4.5e-16)	0.543 (1.3e-16)
	TALE+	0.640 (7.1e-09)	0.423 (1.2e-08)	0.611 (1.7e-11)	0.588 (6.5e-13)	0.346 (6.4e-15)	0.621 (5.1e-10)
	ATGO+	**0.666**	**0.445**	**0.648**	**0.651**	**0.383**	**0.643**

Table S5. The ICW-Fmax values of 12 GO prediction methods on the 1068 benchmark proteins. Bold fonts highlight the best performer in each category.

Methods		ICW-Fmax		
		MF	**BP**	**CC**
Single algorithms	SAGP	0.562	0.329	0.393
	PPIGP	0.199	0.240	0.350
	NGP	0.195	0.167	0.250
	DeepGO	0.315	0.232	0.324
	FunFams	0.435	0.255	0.322
	DeepGOCNN	0.273	0.213	0.190
	DIAMONDScore	0.560	0.325	0.387
	TALE	0.351	0.217	0.257
	ATGO	0.590	0.347	0.486
Composite algorithms	DeepGOPlus	0.569	0.340	0.350
	TALE+	0.569	0.350	0.439
	ATGO+	**0.595**	**0.367**	**0.488**

Table S6. The *p*-values of performance difference between 10 GO prediction methods on 3328 individual CAFA3 targets under post-hoc Nemenyi test at the individual protein level, where the performance of each prediction method is measured by a group of F1-scores, each of which is calculated from the predicted GO terms and native GO annotation in a single test protein. Because the *p*-values can be only approximated in the range from 1.0e-03 to 9.0e-01 under post-hoc Nemenyi test using Python package, the numerical value of 1.0e-03 (or 9.0e-01) means that the *p*-value is below to 1.0e-03 (or upon to 9.0e-01).

	Method	SAGP	PPIGP	NGP	DeepGO	FunFams	DeepGOCNN	DIAMONDScore	ATGO	DeepGOPlus	ATGO+
MF	SAGP	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.8e-03	9.0e-01	1.0e-03	9.0e-01	1.0e-03
	PPIGP	1.0e-03	1.0e+00	1.0e-03	8.1e-01	9.0e-01	4.0e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	NGP	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGO	1.0e-03	8.1e-01	1.0e-03	1.0e+00	3.4e-01	8.3e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	FunFams	1.0e-03	9.0e-01	1.0e-03	3.4e-01	1.0e+00	2.7e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.8e-03	4.0e-02	1.0e-03	8.3e-01	2.7e-03	1.0e+00	1.0e-02	1.0e-03	1.0e-03	1.0e-03
	DIAMONDScore	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-02	1.0e+00	1.0e-03	9.0e-01	1.0e-03
	ATGO	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	9.0e-01
	DeepGOPlus	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	1.0e+00
BP	SAGP	1.0e+00	1.0e-03	1.0e-03	1.2e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03
	PPIGP	1.0e-03	1.0e+00	7.9e-03	1.0e-03	1.0e-03	1.0e-03	8.6e-02	1.0e-03	1.0e-03	1.0e-03
	NGP	1.0e-03	7.9e-03	1.0e+00	1.0e-03	1.0e-03	1.5e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGO	1.2e-02	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	4.9e-01	1.0e-03	4.8e-02	1.0e-03
	FunFams	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.0e-03	1.0e-03	1.5e-02	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DIAMONDScore	1.0e-03	8.6e-02	1.0e-03	4.9e-01	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03
	ATGO	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	3.9e-01
	DeepGOPlus	9.0e-01	1.0e-03	1.0e-03	4.8e-02	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	3.9e-01	1.0e-03	1.0e+00
CC	SAGP	1.0e+00	9.0e-01	9.0e-01	1.0e-03	1.0e-03	1.0e-03	2.9e-01	1.0e-03	8.9e-01	1.0e-03
	PPIGP	9.0e-01	1.0e+00	9.0e-01	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	8.5e-02	1.0e-03
	NGP	9.0e-01	9.0e-01	1.0e+00	1.0e-03	1.0e-03	1.0e-03	9.0e-01	1.0e-03	8.9e-02	1.0e-03
	DeepGO	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03	3.8e-03	1.0e-03
	FunFams	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.8e-01	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DeepGOCNN	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.8e-01	1.0e+00	1.0e-03	1.0e-03	1.0e-03	1.0e-03
	DIAMONDScore	2.9e-01	9.0e-01	9.0e-01	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	3.0e-03	1.0e-03
	ATGO	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e+00	1.0e-03	7.4e-01
	DeepGOPlus	8.9e-01	8.5e-02	8.9e-02	3.8e-03	1.0e-03	1.0e-03	3.0e-03	1.0e-03	1.0e+00	1.0e-03
	ATGO+	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	1.0e-03	7.4e-01	1.0e-03	1.0e+00

Table S7. The ICW-Fmax values of 10 GO prediction methods on 3328 CAFA3 targets where a sequence identity cut-off $t_1 = 30\%$ between the training and testing proteins was applied to the five in-house methods (ATGO, ATGO+, SAGP, PPIGP, and NGP). Bold fonts highlight the best performer in each category.

	Methods	ICW-Fmax		
		MF	**BP**	**CC**
Single algorithms	SAGP	0.439	0.413	0.355
	PPIGP	0.226	0.303	0.343
	NGP	0.143	0.175	0.314
	DeepGO	0.256	0.279	0.357
	FunFams	0.455	0.375	0.391
	DeepGOCNN	0.291	0.211	0.163
	DIAMONDScore	0.431	0.391	0.356
	ATGO	0.479	0.399	0.426
Composite algorithms	DeepGOPlus	0.438	0.389	0.353
	ATGO+	**0.487**	**0.437**	**0.426**

Table S8. The performance of 10 GO prediction methods under the cut-off $t_1 = 30\%$ on 1177 no-knowledge (NK) and 2151 limited-knowledge (LK) CAFA3 proteins. Bold fonts highlight the best performer in each category.

Dataset	Method	Fmax			AUPR			Coverage		
		MF	BP	CC	MF	BP	CC	MF	BP	CC
NK proteins	SAGP	0.467	0.351	0.472	0.248	0.189	0.280	0.84	0.85	0.83
	PPIGP	0.286	0.316	0.476	0.176	0.212	0.440	0.85	0.83	0.84
	NGP	0.184	0.260	0.467	0.083	0.159	0.380	1.00	1.00	1.00
	DeepGO	0.302	0.332	0.502	0.230	0.233	0.501	1.00	1.00	1.00
	FunFams	0.461	0.356	0.430	0.282	0.181	0.250	0.63	0.63	0.61
	DeepGOCNN	0.267	0.304	0.428	0.203	0.193	0.297	1.00	1.00	1.00
	DIAMONDScore	0.463	0.350	0.462	0.196	0.171	0.225	0.78	0.79	0.78
	ATGO	0.513	0.393	0.557	0.472	0.314	**0.559**	1.00	1.00	1.00
	DeepGOPlus	0.473	0.373	0.473	0.385	0.269	0.472	1.00	1.00	1.00
	ATGO+	**0.523**	**0.396**	**0.557**	**0.482**	**0.316**	0.555	1.00	1.00	1.00
LK proteins	SAGP	0.461	0.548	0.479	0.241	0.392	0.322	0.81	0.93	0.88
	PPIGP	0.224	0.422	0.421	0.138	0.357	0.394	0.92	0.92	0.83
	NGP	0.142	0.339	0.416	0.055	0.175	0.348	1.00	1.00	1.00
	DeepGO	0.259	0.423	0.469	0.176	0.333	0.468	1.00	1.00	1.00
	FunFams	0.481	0.472	0.508	0.320	0.243	0.332	0.67	0.76	0.75
	DeepGOCNN	0.342	0.284	0.392	0.251	0.191	0.275	1.00	1.00	1.00
	DIAMONDScore	0.452	0.518	0.469	0.200	0.344	0.256	0.74	0.89	0.84
	ATGO	0.498	0.564	0.523	0.468	0.465	0.528	1.00	1.00	1.00
	DeepGOPlus	0.449	0.524	0.478	0.394	0.401	0.470	1.00	1.00	1.00
	ATGO+	**0.511**	**0.574**	**0.525**	**0.473**	**0.488**	**0.534**	1.00	1.00	1.00

Table S9. The numbers of proteins for 20 species in CAFA3 test dataset.

Species name	Taxonomy ID	Sample number
Human	9606	1131
Arabidopsis	3702	626
Fission Yeast	284812	426
Mouse	10090	326
Escherichia Coli	83333	224
Fly	7227	209
Rat	10116	97
Bacillus Subtilis	224308	76
Dictyostelium Discoideum	44689	49
Zebrafish	7955	46
Budding Yeast	559292	32
Candida Albicans	237561	27
Salmonella Enterica	99287	16
Xenopus Laevis	8355	14
Methanocaldococcus Jannaschii	243232	7
Pseudomonas Putida	160488	7
Helicobacter Pylori	85962	5
Saccharolobus Solfataricus P2	273057	4
Mycoplasma Genitalium	243273	3
Pseudomonas Aeruginosa	208963	3

Table S10. The prediction performance of 5 GO prediction methods under the cut-off $t_1 = 100\%$ on CAFA3 test proteins. Bold fonts highlight the best performer in each category.

Dataset	Method	Fmax			AUPR		
		MF	BP	CC	MF	BP	CC
All 3328 proteins	SAGP	0.520	0.515	0.504	0.328	0.366	0.350
	PPIGP	0.253	0.390	0.473	0.160	0.312	0.461
	NGP	0.166	0.302	0.445	0.065	0.170	0.366
	ATGO	0.548	0.520	0.555	0.504	0.445	0.551
	ATGO+	**0.551**	**0.540**	**0.559**	**0.514**	**0.470**	**0.546**
1177 no-knowledge proteins	SAGP	0.494	0.387	0.509	0.297	0.230	0.337
	PPIGP	0.299	0.335	0.491	0.189	0.236	0.471
	NGP	0.192	0.260	0.467	0.082	0.160	0.380
	ATGO	**0.533**	0.400	0.569	0.476	0.338	0.562
	ATGO+	0.532	**0.419**	**0.570**	**0.490**	**0.346**	**0.553**
2151 limited-knowledge proteins	SAGP	0.537	0.602	0.498	0.347	0.473	0.366
	PPIGP	0.221	0.435	0.456	0.140	0.366	0.447
	NGP	0.147	0.339	0.416	0.055	0.175	0.348
	ATGO	0.562	0.602	0.539	0.526	0.529	0.535
	ATGO+	**0.566**	**0.622**	**0.542**	**0.532**	**0.564**	**0.537**

Table S11. The prediction performance of SAGP and BLAST baseline on our constructed test dataset and CAFA3 test dataset with different cut-off values of sequence identity**.** Bold fonts highlight the best performer in each category.

Dataset	Method	Fmax			AUPR		
		MF	BP	CC	MF	BP	CC
1068 test proteins constructed in this work	BLAST baseline	0.440	0.292	0.375	0.315	0.166	0.269
	SAGP	**0.597**	**0.400**	**0.534**	**0.351**	**0.242**	**0.322**
3328 CAFA3 targets under the cut-off $t_1 = 30\%$	BLAST baseline	0.352	0.248	0.325	0.204	0.128	0.210
	SAGP	**0.463**	**0.465**	**0.473**	**0.244**	**0.302**	**0.298**
3328 CAFA3 targets under the cut-off $t_1 = 100\%$	BLAST baseline	0.388	0.322	0.391	0.256	0.198	0.259
	SAGP	**0.520**	**0.515**	**0.504**	**0.328**	**0.366**	**0.350**

Table S12. The prediction performance of ATGO models via four metric learning methods on two test datasets. Bold fonts highlight the best performer in each category.

Dataset	Method	Fmax			AUPR		
		MF	BP	CC	MF	BP	CC
1068 test proteins constructed in this work	F1	0.627	0.425	0.623	0.603	0.361	**0.600**
	JS	**0.629**	0.423	0.622	0.600	0.355	0.557
	WF1	0.628	**0.426**	0.623	**0.606**	**0.364**	0.579
	WJS	0.628	**0.426**	**0.624**	0.587	0.358	0.592
3328 CAFA3 targets under the cut-off $t_1 = 30\%$	F1	**0.501**	0.495	0.542	**0.469**	0.397	0.546
	JS	0.498	0.491	0.544	0.441	0.401	0.543
	WF1	0.500	0.492	**0.545**	0.429	0.404	**0.550**
	WJS	0.497	**0.497**	0.544	0.429	**0.410**	0.549

Table S13. The incorrectly predicted GO terms for 12 methods on three proteins in BP aspect.

Method	A6XMY0	E7CIP7	F4I082
SAGP		GO:0044419 GO:0009607 GO:0009605 GO:0043207 GO:0050896 GO:0009617 GO:0006952 GO:0006950 GO:0051707	GO:0032502 GO:0042335 GO:0006869 GO:0006810 GO:0071702 GO:0051234 GO:0051179 GO:0048856
PPIGP			GO:0032502 GO:0009628 GO:0044238 GO:0009987 GO:0044237 GO:0071704 GO:0009058 GO:0006807 GO:0065007 GO:0016043 GO:0008152 GO:1901576 GO:0042221 GO:0044249 GO:0050789 GO:0071840
NGP	GO:0032502 GO:0044238 GO:0019222 GO:0060255 GO:0048856 GO:0044237 GO:0050794 GO:0071704 GO:0006807 GO:0065007 GO:0016043 GO:0008152 GO:0048518 GO:0043170 GO:0050789 GO:0071840	GO:0032502 GO:0048856 GO:0019222 GO:0060255 GO:0050896 GO:0009987 GO:0044237 GO:0050794 GO:0006807 GO:0065007 GO:0016043 GO:0048518 GO:0050789 GO:0071840	GO:0032502 GO:0044238 GO:0048856 GO:0019222 GO:0060255 GO:0009987 GO:0044237 GO:0050794 GO:0071704 GO:0006807 GO:0065007 GO:0016043 GO:0008152 GO:0048518 GO:0043170 GO:0071840 GO:0050789
DeepGO	GO:0080090 GO:0019222 GO:0031326 GO:0031323 GO:0050789 GO:0071704 GO:2000112 GO:0060255 GO:0065007 GO:0048518 GO:0048519 GO:0065008 GO:0010468 GO:0019219 GO:0048583 GO:0009889 GO:1903506 GO:0050794 GO:0051171 GO:0008152 GO:2001141 GO:0051704 GO:0044238 GO:0051252 GO:0044237 GO:0051239 GO:0010556 GO:0048523 GO:0048522	GO:2001141 GO:0009987 GO:0080090 GO:0019222 GO:0019219 GO:0009889 GO:0050896 GO:0051252 GO:0031323 GO:1903506 GO:0050794 GO:0006355 GO:0010556 GO:0065007 GO:0051171 GO:0031326 GO:0060255 GO:0010468 GO:2000112 GO:0050789	GO:0048856 GO:0009892 GO:0080090 GO:0019222 GO:0009891 GO:0031327 GO:0031326 GO:0031325 GO:2000241 GO:0031323 GO:0048580 GO:0010629 GO:0032501 GO:0007165 GO:0050789 GO:0009893 GO:0009755 GO:0010605 GO:0009890 GO:0033993 GO:0031328 GO:2000112 GO:2000113 GO:0009737 GO:0060255 GO:0065007 GO:0048518 GO:0048519 GO:0010468 GO:0032502 GO:0009719 GO:0031324 GO:0048608 GO:0050793 GO:0009987 GO:0009889 GO:1903506 GO:0050794 GO:0003006 GO:0001101 GO:0051239 GO:2001141 GO:0042221 GO:0010033 GO:1901700 GO:0022414

			GO:0009725 GO:0051252 GO:0019219 GO:0006355 GO:0010556 GO:2000026 GO:0048522 GO:0097305 GO:0010558 GO:0048523 GO:0051171
FunFams		GO:0002376 GO:0009607 GO:0009605 GO:0043207 GO:0050896 GO:0009617 GO:0006952 GO:0006950 GO:0098542 GO:0042742 GO:0044419 GO:0050829 GO:0051707 GO:0006955	
DeepGOCNN	GO:0048583 GO:0023052 GO:0007165 GO:0023051 GO:0050829 GO:0010646 GO:0050789 GO:0051716 GO:0009966 GO:0051179 GO:0065007 GO:0065009 GO:0003008 GO:0007186 GO:0032502 GO:0032501 GO:0006810 GO:0050794 GO:0050830 GO:0007267 GO:0007154 GO:0051234 GO:0055085 GO:0051704 GO:0010469 GO:0048856	GO:0032501 GO:0009605 GO:0050896 GO:0009987 GO:0071554 GO:0006950 GO:0065007 GO:0051179 GO:0051704	GO:0044238 GO:0009987 GO:0044237 GO:0050794 GO:0071704 GO:0065007 GO:0016043 GO:0008152 GO:0051179 GO:0050789 GO:0071840
DIAMONDScore		GO:0044419 GO:0009607 GO:0009605 GO:0043207 GO:0050896 GO:0009617 GO:0006952 GO:0006950 GO:0051707	
TALE	GO:0050829 GO:0032501 GO:0065007 GO:0050789 GO:0050794	GO:0071554 GO:0009987	GO:0044238 GO:0009987 GO:0044237 GO:0050794 GO:0071704 GO:0006807 GO:0065007 GO:0008152 GO:0050789
ATGO		GO:0044419 GO:0009987	GO:0032502 GO:0009628 GO:0003006 GO:0009987 GO:0022414 GO:0065007 GO:0050789
DeepGOPlus	GO:0032501 GO:0065007	GO:0044419 GO:0009617 GO:0009607 GO:0009605 GO:0043207 GO:0050896	GO:0009987

		GO:0009987 GO:0006952 GO:0006950 GO:0051707	
TALE+		GO:0009617 GO:0009607 GO:0009605 GO:0043207 GO:0050896 GO:0009987 GO:0044419 GO:0006950 GO:0051707	GO:0009987
ATGO+		GO:0044419 GO:0050896	GO:0032502 GO:0042335 GO:0048856

Table S14. The numbers of proteins and GO terms in benchmark dataset.

Benchmark dataset	N_{MF}^{P}	N_{BP}^{P}	N_{CC}^{P}	N_{ALL}^{P}	N_{MF}^{T}	N_{BP}^{T}	N_{CC}^{T}	N_{ALL}^{T}
Training dataset	49135	79491	71982	109132	6581	20882	2782	30245
Validation dataset	515	860	664	1089	818	3894	417	5129
Test dataset	577	839	586	1068	876	3469	382	4727

$N_{MF}^{P}/N_{BP}^{P}/N_{CC}^{P}/N_{ALL}^{P}$: The number of proteins for MF/BP/CC/all three aspects.

$N_{MF}^{T}/N_{BP}^{T}/N_{CC}^{T}/N_{ALL}^{T}$: The number of GO terms for MF/BP/CC/all three aspects.

Table S15. The values of $margin$, c_f, α, and K for three GO aspects

GO aspect	$margin$	c_f	α	K
MF	0.1	0.8	5	30
BP	0.1	0.8	5	100
CC	0.1	0.8	5	100

Supporting Texts

Text S1. Sequence alignment-based GO prediction (SAGP)

In SAGP, we select the function templates, which share high sequence similarity with the query, to annotate its function. Specifically, for a query sequence, BLAST software [1] is used to scan the corresponding templates with an e-value cutoff of 0.1. The confidence score of the GO term q by SAGP is calculated by

$$S(q)_{SAGP} = \frac{\sum_{k=1}^{n} b_k \cdot I_k(q)}{\sum_{k=1}^{n} b_k} \quad \text{(S1)}$$

where n is the number of templates identified, b_k is the bit-score of k-th template by BLAST; $I_k(q) = 1$, if the k-th template is associated with q in the experimental function annotation; otherwise, $I_k(q) = 0$.

Text S2. Protein-protein interaction-based GO prediction (PPIGP)

For a query, we search its interaction partners from the STRING database [2] for functional annotation. Then, we remove the interaction partners which are not found in the training dataset. Finally, the remaining partners are used to annotate the query. The confidence score is calculated using the same scoring function as in SAGP (i.e., Eq. S1), where b_k is the score assigned by STRING as confidence of interaction between the query and the k-th partner.

Text S3. Naïve-based GO prediction (NGP)

In NGP, the confidence score that a query is associated with GO term q is calculated by the frequency of q in the training dataset:

$$\mathrm{S}(q)_{NGP} = N(q)/N_{GO} \quad \text{(S2)}$$

where $N(q)$ is the number of proteins associated with q, and N_{GO} is the number of proteins with at least one annotation for the same GO aspect as q. This predictor can be thought of as a prior arising from the overall abundance of a particular annotation in the training dataset.

Text S4. Friedman and Nemenyi post-hoc tests at the individual protein level

We use Friedman test [3], one of the most used approaches in analysis of variants, to identify whether there is a significant performance difference among a group of GO prediction methods. If the significance factor (i.e., p-value) is below to 0.05 in Friedman test, the Nemenyi post-hoc test [4] is further performed to identify the performance difference between pairwise prediction methods.

It is noted that most of competing methods, such as SAGP, PPIGP, FunFams, and DIAMONDScore, are performed with constant evaluation indices (i.e., Fmax and AUPR values) in the entire dataset. Therefore, there is no proper statistical test to identify the performance difference between the above-mentioned methods in the entire dataset. In view of this, we perform Friedman and Nemenyi post-hoc tests at the individual protein level rather than the entire dataset level. Specifically, the performance of each prediction method is measured by a set of F1-scores, each of which is calculated from the predicted GO terms and native GO annotation in a single protein (see Eq. S32). Moreover, the predicted GO terms of different methods are determined

by their own cut-off setting to achieve the highest Fmax value. Finally, the Friedman and Nemenyi post-hoc tests are performed on the F1-scores of individual test proteins to calculate the *p*-values of performance difference among prediction methods.

Text S5. An explanation for the difference of *p*-value calculations between Student t-test and Nemenyi post-hoc test.

It can be found that there exists a big gap of *p*-values between Student t-test at *the entire dataset level* and Nemenyi post-hoc test at *the individual protein level*. To explain this observation, we firstly introduce the procedures of the above-mentioned two statistical tests.

A. Student t-test. We select one-sample t-test [5] to calculate the *p*-values between ATGO and other 10 competing methods, because each competing method is performed with constant evaluation indices (i.e., Fmax and AUPR values) in the entire dataset. Specifically, ATGO is repeatedly implemented on the benchmark dataset in N times to generate a set of evaluation indices, i.e., $X = \{x_1, x_2, \dots, x_N\}$, where x_i is the evaluation index of ATGO in the i-th time, while the evaluation index of the competing method is defined as u. The *p*-value between ATGO and a competing method with single implementation is calculated by the following two steps:

(1) The statistic value t_X is defined as:

$$t_X = \frac{u_X - u}{\sigma_X / \sqrt{N}} \tag{S3}$$

where u_X and σ_X are mean and standard deviation for X, respectively, and t_X obeys t-distribution with degree of freedom = $N - 1$ (In this work, $N = 10$).

(2) The significance factor (i.e., *p*-value) can be approximately calculated from the probability dense function of t-distribution using integral in ranges of (-∞, -$|t_X|$) and ($|t_X|$, +∞).

We select two-sample t-test [6] to calculate the *p*-value between ATGO and ATGO+, because they are both performed with variable evaluation indices in the entire dataset. Specifically, ATGO and ATGO+ are repeatedly implemented on the benchmark dataset in N_1 and N_2 times, respectively, to generate two sets of evaluation indices, i.e., $X = \{x_1, x_2, \dots, x_{N_1}\}$ and $Y = \{y_1, y_2, \dots, y_{N_2}\}$, where x_i and y_j are the evaluation indices of ATGO and ATGO+ in the i-th time and j-th time, respectively. The *p*-value between ATGO and ATGO+ is calculated by the following two steps:

(1) The statistic value t_{XY} is defined as:

$$t_{XY} = \frac{u_X - u_Y}{\sqrt{\frac{(N_1-1)S_X + (N_2-1)S_Y}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}} \tag{S4}$$

where u_X and u_Y are mean values for X and Y, respectively; S_X and S_Y are variances for X and Y, respectively; and t_{XY} obeys t-distribution with degree of freedom = $N_1 + N_2 - 2$ (In this work, $N_1 = N_2 = 10$).

(2) The *p*-value can be approximately calculated from the probability dense function of t-distribution using integral in ranges of (-∞, -$|t_{XY}|$) and ($|t_{XY}|$, +∞).

In this work, we use Python package "scipy" to implement Student t-test to calculate the *p*-values in the range from 0 to 1.

B. Nemenyi post-hoc test. Given a group of K GO prediction methods, we use N samples (test proteins) to evaluate their performance, where the performance of an individual prediction method is measured by a set of F1-scores, each of which is

calculated from predicted GO terms and native GO terms on a single sample. For a pair of methods (M_i, M_j) in this group, the corresponding p-value of F1-scores in Nemenyi post-hoc test is calculated by the following three steps.

(1) The difference of average rank between M_i and M_j, denoted as DAR_{ij}, is calculated:

$$DAR_{ij} = \frac{1}{N}\left|\sum_{n=1}^{N} rank_{i,n} - \sum_{n=1}^{N} rank_{j,n}\right| \quad \text{(S5)}$$

where $rank_{i,n}$ is the rank of M_i among K methods on the n-th sample from the view of F1-score.

(2) The statistic value q_α is defined as:

$$q_\alpha = DAR_{ij} / \sqrt{\frac{K(K+1)}{6N}} \quad \text{(S6)}$$

where q_α obeys studentized range distribution [7] with degree of freedom = infinity and sample number = K. The higher value of q_α means the higher significance of performance difference between M_i and M_j, and α is the corresponding significance factor, i.e., p-value.

(3) The p-value can be approximately calculated from the probability dense function of studentized range distribution with preset statistic thresholds using Gleason's approach [8].

In this work, we use Python package "scikit posthocs" to implement Nemenyi post-hoc test, where the p-values are approximated using Gleason's approach. Because the minimal and maximal preset statistic thresholds in Gleason's approach are 0.001 and 0.900, respectively, the p-values can be only approximated in range (0.001, 0.900). If the p-value is below to 0.001 (or upon 0.900), "scikit posthocs" package will directly output 0.001 (or 1.000).

By reviewing Student t-test and Nemenyi post-hoc test, the big gap between their p-values is mainly attributed to the following two aspects. (1) Student t-test can approximate the p-values in the range from 0 to 1, while Nemenyi post-hoc test can only approximate the p-values in a much smaller range, i.e., (0.001, 0.9), in our programs. (2) The significance of performance difference between two GO prediction methods (M_i and M_j) may be decreased with the increase of the number of methods in a group under Nemenyi post-hoc test. Specifically, we suspect that the significant difference between M_i and M_j (i.e., the value of q_α) may be decreased with the increase of the value of K, because the increase rate of DAR_{ij} is lower than that of $\sqrt{K(K+1)/6N}$ in Eq. S6. To further demonstrate this point, we designed the following test.

Starting from a group of GO prediction methods (named Group A) including SAGP and ATGO, we incrementally add PPIGP, NGP, DeepGO, FunFams, DeepGOCNN, DIAMONDScore, TALE, DeepGOPlus, TALE+, and ATGO+ to Group A and then perform Nemenyi post-hoc test for Group A on our constructed 1068 test proteins in MF aspect. Table S2 lists the statistic values between SAGP and ATGO versus the increase of K. It can be found that the p-value between SAGP and ATGO is consistently increased from 1.0e-03 to 2.6e-01 (when $2 \leq K \leq 7$) and then fluctuates in the range from 1.7e-01 to 3.0e-01 (when $8 \leq K \leq 12$). This observation can be explained as follows. By reviewing Table 1 in the main text, we find that the first five GO prediction methods (i.e., PPIGP, NGP, DeepGO, FunFams, and DeepGOCNN) shows much lower

Fmax values both than SAGP and ATGO in MF aspect, indicating the rank difference between SAGP and ATGO from the view of F1-score cannot be changed on most test proteins after adding these five methods into Group A. As a result, the increase rate of DAR is lower than that of $\sqrt{K(K+1)/6N}$, leading to the continuous decrease of q_α. As for other five methods (DIAMONDScore, TALE, DeepGOPlus, TALE+, and ATGO+), most of them achieve the comparable Fmax values with both SAGP and ATGO, indicating that the value of DAR will be dramatically increased after adding these methods into Group A. Therefore, the increase of DAR can keep up with that of $\sqrt{K(K+1)/6N}$, leading to the fluctuation of q_α in a fixed range.

This experiment has demonstrated two points. Frist, the significance of performance difference between two GO prediction methods in a group is not only dependent on their performance but also associated with the performance of other prediction methods under Nemenyi post-hoc test. Second, this significance may be decreased with the increase of the number of GO prediction methods in a group.

In light of the above data and insight, we prefer to use Student t-test to identify the performance difference between two GO prediction methods at the entire dataset level, because the corresponding *p*-value can be approximated in a more precise range and not be affected by the performances of other prediction methods in the same group.

Text S6. Information content-weighted maximum F1-score

The information content-weighted maximum F1-score (ICW-Fmax) is defined as:

$$\text{ICW-Fmax} = \max_{0\le t\le 1}\left[\frac{2\cdot icwpr(t)\cdot icwrc(t)}{icwpr(t)+icwrc(t)}\right] \tag{S7}$$

where t is a cut-off value of confidence score; $iwcpr(t)$ and $icwrc(t)$ are IC-weighted precision and IC-weighted recall, respectively, with confidence score $\geq t$:

$$\begin{cases} icwpr(t) = \dfrac{\sum_{GO_i\in GOSET\_TP(t)} IC(GO_i)}{\sum_{GO_j\in(GOSET\_TP(t)\cup GOSET\_FP(t))} IC(GO_j)} \\ icwrc(t) = \dfrac{\sum_{GO_i\in GOSET\_TP(t)} IC(GO_i)}{\sum_{GO_j\in(GOSET\_TP(t)\cup GOSET\_FN(t))} IC(GO_j)} \end{cases} \tag{S8}$$

$$IC(GO_i) = -log_2(1/p(GO_i|\text{parents of } GO_i \text{ in GO})) \tag{S9}$$

where $GOSET\_TP(t)$ is the set of correctly predicted GO terms, $GOSET\_TP(t) \cup GOSET\_FP(t)$ is the set of all predicted GO terms, $GOSET\_TP(t) \cup GOSET\_FN(t)$ is the set of experimentally annotated GO terms, $IC(GO_i)$ is the information content for the GO term GO_i, $p(GO_i|\text{parents of } GO_i \text{ in GO})$ is the conditional probability of GO_i given its parents of the GO structure (see details in [9]).

Text S7. Performance comparison between SAGP and BLAST baseline used in CAFA challenge

In BLAST baseline of CAFA challenge, the confidence score that a query is associated with GO term q is calculated by:

$$\text{S}(q)_{BLAST-baseline} = \max\{s_1\cdot I(q)_1, s_2\cdot I(q)_2, \dots, s_n\cdot I(q)_n\} \tag{S10}$$

$$s_i = N_i^{id}/N_i^{al} \tag{S11}$$

where n is the number of templates in BLAST search, s_i is the local sequence identity between query and the i-th template, N_i^{id} is the number of identical residues in the

local alignment region, and N_i^{al} is the length of local alignment region; if the i-th template is associated with GO term q in the native annotation, $I(q)_i = 1$; otherwise, $I(q)_i = 0$.

Table S11 summarizes the performance of SAGP and BLAST baseline on our constructed test dataset and CAFA3 test dataset with different cut-off values of sequence identity. It can be found that SAGP achieves much better performance than BLAST baseline in all three GO aspects. Taking CAFA3 test dataset as an example, SAGP gains 54.9% and 65.8% average improvements of Fmax and AUPR values, respectively, in three GO aspects under the cut-off $t_1 = 30\%$. In addition, BLAST baseline shows much worse performance than most of competing methods, such as FunFams and DeepGOPlus.

Text S8. Performance comparison between four metric learning methods

We separately use four metric learning methods, including F1-score (F1, see Eq. S32), Jaccard similarity (JS) [10], weighted F1-score (WF1), and weighted Jaccard similarity (WJS), to measure the functional similarity in triplet loss, where the weights of GO terms are measured by information content [11]. The formulas of JS, WF1, and WJS are described as follows.

$$\mathrm{JS} = \frac{|GOSET_A \cap GOSET_B|}{|GOSET_A \cup GOSET_B|} \tag{S12}$$

$$\mathrm{WF1} = 2(\mathrm{pre}_w \times \mathrm{rec}_w)/(\mathrm{pre}_w + \mathrm{rec}_w) \tag{S13}$$

$$\mathrm{pre}_w = \frac{\sum_{GO_i \in (GOSET_A \cap GOSET_B)} w(GO_i)}{\sum_{GO_j \in GOSET_A} w(GO_j)},\ \mathrm{rec}_w = \frac{\sum_{GO_i \in (GOSET_A \cap GOSET_B)} w(GO_i)}{\sum_{GO_j \in GOSET_B} w(GO_j)} \tag{S14}$$

$$\mathrm{WJS} = \frac{\sum_{GO_i \in (GOSET_A \cap GOSET_B)} w(GO_i)}{\sum_{GO_j \in (GOSET_A \cup GOSET_B)} w(GO_j)} \tag{S15}$$

$$w(GO_i) = -log_2(1/p(GO_i|\text{parents of } GO_i \text{ in GO})) \tag{S16}$$

where $GOSET_A$ and $GOSET_B$ are sets of GO terms in native annotations for proteins A and B, respectively, $|.|$ is the number of elements in a set, $w(GO_i)$ is the weight (measured by information content) of GO_i, and $p(GO_i|\text{parents of } GO_i \text{ in GO})$ is the conditional probability of GO_i given its parents of the GO structure (see details in [9]). Two proteins are considered to have the same function if their functional similarity is larger than a cut-off value c_f. The values of c_f are 0.8, 0.5, 0.8 and 0.5 for F1, JS, WF1, and WJS, respectively, in each GO aspect.

For each metric learning method, we re-trained the corresponding GO prediction model using the ATGO framework, which was further benchmarked on our constructed test dataset and CAFA3 test dataset, as summarized in Table S12. We found that there is no significant performance difference among four metric learning methods for each GO aspect. Specifically, the absolute increases of Fmax values between the best and worst performers are both less than 0.01 for all three aspects in each test dataset, suggesting that the effectiveness of the proposed ATGO framework is not sensitive to the choices of different metric learning methods.

Text S9. The mathematics formulas for ESM-1b transformer

A. Masking

For an input sequence, the masking strategy [12] is performed on the corresponding tokens (i.e., amino acids). Specifically, we randomly sample 15% tokens, each of which is changed as a special "masking" token with 80% probability, a randomly chosen alternate amino acid with 10% probability, and the original input token (i.e., no change) with 10% probability.

B. One-hot encoding

The masked sequence is represented as a $L \times 28$ matrix using one-hot encoding [13], where 28 is the types of tokens, including 20 common amino acids, 6 non-common amino acids (B, J, O, U, X and Z), 1 gap token, and 1 "masking" token.

C. Embedding with positions

The one-hot coding matrix X of the masked sequence is multiplied by an embedding weight matrix W_E to generate an embedding matrix H_E:

$$H_E = XW_E,\ X \in R^{L\times 28}, W_E \in R^{28\times D}, H_E \in R^{L\times D} \tag{S17}$$

where L is the length of the masked sequence, 28 is the types of tokens in the masked sequence, and D is the embedding dimension.

Then, the position embedding strategy is used to record to position of each token in the masked sequence to generate a position embedding matrix H_P:

$$H_P = \begin{bmatrix} h_1 \\ h_2 \\ \dots \\ h_L \end{bmatrix}, h_i = \left(v_{i,1}, v_{i,2}, \dots, v_{i,D}\right),\ H_P \in R^{L\times D} \text{, and } h_i \in R^D \tag{S18}$$

$$v_{i,2k} = \sin\left(\frac{i}{10000^{2k/D}}\right), v_{i,2k+1} = \cos\left(\frac{i}{10000^{(2k+1)/D}}\right),\ k = 0,\ 1,\ ..,\ (D-1)/2 \tag{S19}$$

where h_i is the embedding vector for the i-th position in the masked sequence.

Finally, two embedding matrices are added as an initial combination embedding matrix H_1:

$$H_1 = H_E + H_P,\ H_1 \in R^{L\times D} \tag{S20}$$

D. Self-attention

The initial embedding matrix H_1 is fed to self-attention block with n layers, each of which consists of m attention heads, a linear unit, and a feed-forward network (FFN). In each attention head, the scale dot-product attention is performed as follows:

$$A_{i,j} = softmax(M_{i,j}^Q {M_{i,j}^K}^T/\sqrt{d_{ij}})\ M_{i,j}^V \tag{S21}$$

$$M_{i,j}^Q = H_i W_{i,j}^Q,\ M_{i,j}^K = H_i W_{i,j}^K,\ M_{i,j}^V = H_i W_{i,j}^V \tag{S22}$$

$$d_{ij} = D/m,\ W_{i,j}^Q, W_{i,j}^K, W_{i,j}^V \in R^{D\times(\frac{D}{m})},\ M_{i,j}^Q,\ M_{i,j}^K, M_{i,j}^V,\ A_{i,j} \in R^{L\times(\frac{D}{m})} \tag{S23}$$

where $A_{i,j}$ is the attention matrix in the (i-th layer, j-th head), $M_{i,j}^Q$, $M_{i,j}^K$, and $M_{i,j}^V$ are

Query, Key, and Value matrices in the (i-th layer, j-th head), H_i is the input matrix in the i-th layer, $W_{i,j}^Q$, $W_{i,j}^K$, and $W_{i,j}^V$ are the weight matrices to be trained, and d_{ij} is the scale parameter.

The outputs of all m attention heads in i-th layer are concatenated as a new matrix A_i, which is further fed to a linear unit to output the matrix U_i:

$$A_i = A_{i,1}(c)A_{i,2} \ldots (c)A_{i,m} \quad \text{(S24)}$$

$$U_i = A_i W_i^1 + b_i^1,\ W_i^1 \in R^{D\times D},\ A_i, b_i^1, U_i \in R^{L\times D} \quad \text{(S25)}$$

where W_i^1 and b_i^1 are the weight matrix and bias, respectively, in the linear unit.

E. Feed-forward network with shortcut connections

The U_i is added by H_i to generate a new matrix F_i, which is further fed to the FFN to output the matrix T_i:

$$F_i = H_i + U_i \quad \text{(S26)}$$

$$T_i = \text{gelu}(F_i W_i^2 + b_i^2)W_i^3 + b_i^3,\ W_i^2, W_i^3 \in R^{D\times D},\ b_i^2, b_i^3, T_i \in R^{L\times D} \quad \text{(S27)}$$

$$\text{gelu}(x) = \text{x}\emptyset(x) \quad \text{(S28)}$$

where W_i^2 and W_i^3 are weight matrices in the FFN, b_i^2 and b_i^3 are bias in the FFN, and $\emptyset(x)$ is the integral of Gaussian Distribution for x.

The F_i is added by T_i as the output the i-th attention layer:

$$H_{i+1} = F_i + T_i,\ H_{i+1} \in R^{L\times D} \quad \text{(S29)}$$

The output of the last attention layer is fed to a fully connected layer with SoftMax function to generate a $L \times 28$ probability matrix:

$$P = SoftMax(H^n W^n + b^n),\ P \in R^{L\times 28} \quad \text{(S30)}$$

where the element value (l-th, c-th) in P indicates the probability that the l-th token in the masked sequence is predicted as the c-th type of amino acid, W^n and b^n are weight matrix and bias, respectively.

F. Loss function

The loss function is designed as:

$$Loss_{esm} = E_{x\sim X}\left[\sum_{l\in x(M)}\left(-\frac{logP_{l,c(l)}}{|x(M)|}\right)\right] \quad \text{(S31)}$$

where x is a sequence in training protein set X, $x(M)$ is a set of masking position in x, $|x(M)|$ is the number of elements in $x(M)$, $c(l)$ is the type index of amino acid for the l-th token in x before masking, -$logP_{l,c(l)}$ is negative log likelihood of the true amino acid x_l under condition of masking, and $E_{x\sim X}[.]$ indicates the mean operation on the function.

Text S10. The functional similarity between two proteins

The functional similarity of two proteins is measured by the F1-score between their GO terms:

$$\text{F1} - \text{score} = 2(\text{pre} \times \text{rec})/(\text{pre} + \text{rec}),\ \text{pre} = ns/n_a,\ \text{rec} = ns/n_b \quad \text{(S32)}$$

where ns is the number of same GO terms between two proteins, n_a and n_b are the numbers of GO terms for proteins a and b, respectively.

References

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389-3402.

2. Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) STRING: a database of predicted functional associations between proteins. Nucleic acids research 31: 258-261.

3. Sheldon MR, Fillyaw MJ, Thompson WD (1996) The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. Physiotherapy Research International 1: 221-228.

4. Hilton A, Armstrong RA (2006) Statnote 6: post-hoc ANOVA tests. Microbiologist 2006: 34-36.

5. Crawford J, Howell DC, Garthwaite PH (1998) Payne and Jones revisited: estimating the abnormality of test score differences using a modified paired samples t test. Journal of clinical and experimental neuropsychology 20: 898-905.

6. Heeren T, D'Agostino R (1987) Robustness of the two independent samples t-test when applied to ordinal scaled data. Statistics in medicine 6: 79-90.

7. Kokoska S, Nevison C. Critical values for the studentized range distribution. Statistical tables and formulae: Springer; 1989. p. 64-66.

8. Gleason JR (1999) An accurate, non-iterative approximation for studentized range quantiles. Computational statistics & data analysis 31: 147-158.

9. Clark WT, Radivojac P (2013) Information-theoretic evaluation of predicted ontological annotations. Bioinformatics 29: i53-i61.

10. Bag S, Kumar SK, Tiwari MK (2019) An efficient recommendation generation using relevant Jaccard similarity. Information Sciences 483: 53-64.

11. Hayn C (1995) The information content of losses. Journal of accounting and economics 20: 125-153.

12. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

13. Buckman J, Roy A, Raffel C, Goodfellow I, editors. Thermometer encoding: One

hot way to resist adversarial examples. International Conference on Learning Representations; 2018.