Amino acid	All *k*-mers	Top 1000	Increase / decrease (percentage)
W	1.14 %	0.01 %	**-99.14 %**
C	1.28 %	0.08 %	**-93.78 %**
M	2.09 %	0.18 %	**-91.17 %**
Y	3.03 %	0.45 %	**-85.06 %**
H	2.34 %	0.51 %	**-78.28 %**
F	3.88 %	1.19 %	**-69.39 %**
N	3.97 %	1.54 %	**-61.36 %**
Q	3.60 %	1.96 %	-45.49 %
R	5.83 %	4.67 %	-19.79 %
P	4.84 %	3.96 %	-18.00 %
K	5.30 %	4.60 %	-13.18 %
I	6.02 %	5.45 %	-9.47 %
S	6.28 %	5.79 %	-7.68 %
T	5.18 %	4.98 %	-3.72 %
D	5.68 %	5.82 %	2.49 %
G	7.32 %	9.40 %	28.53 %
V	6.92 %	9.02 %	30.26 %
L	9.77 %	13.73 %	40.45 %
E	6.47 %	9.51 %	46.93 %
A	9.07 %	17.14 %	**89.00 %**

Table S4. Comparison between the distribution of amino acid content in all *k*-mers in the dataset and in the top 1000 *k*-mers in the predictor. The most significant variations are highlighted in bold.