Explainabilty Comparison between Random Forests and Neural Networks— Case Study of Amino Acid Volume Prediction: supplemental document

This is a supplementary document for inclusion of material with submission to Information journals. This document, which may include supplementary information such as expanded descriptions of materials and figures and tables, will be published as a PDF linked to the primary article. The supplemental file should only present information that would be useful and worthwhile for the reader, for example, details that would be necessary to reproduce an experiment. The article, however, must be coherent without the supplemental PDF file.

1. DATA-SET

In this section it's possible to find the original data-set, downloaded from https://www.rcsb.org/. The original data-set is composed by 446 files belonging to myoglobins and 6 file about spike proteins. The Spike files will be only preprocessed and inserted into Rosy and Roberta Database (DBR^2) but not analyzed. The reason of this choice is due to a preliminary analysis of the data-set: after an alignment among the primary sequences, it has been clarified that the definition of the structures of spike proteins of the COVID-19 weren't, up to date, not fully assessed and the number of proteins deposited and catalogued were small, reducing the variance. The files attached, show the PDB codes of the relative proteins downloaded with other information-i.e. Entity ID, Asym ID, Auth Asym ID, Database Name, Accession Code(s), Sequence, Polymer Entity Sequence Length, Entity Macromolecule Type, Molecular Weight- that could be useful for a preliminary analysis of the data-set, giving a general overview of the case study.

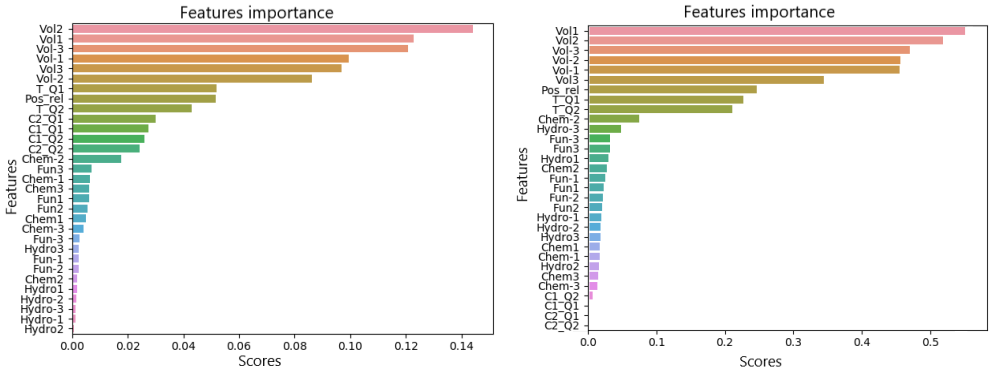
2. RESULTS

As described in the paper, the analysis is composed by 3 different parts:

- Analysis on the most frequent amino acid found in the myoglobins, as described in the Section 6.1 of the paper: lysine (LYS), whose results are graphs, representing the feature importance for the prediction obtained with both Random Forest (RF) and Multilayer Perceptron Neural Network (MLP). On the y-axis there is the name of the feature, while on the x-axis there is the relative score obtained in the prediction, sorted by descending order S1a and S1b. The analysis continues with an extension of the data-set, considering a wider amino acid's around, Figure S2a, S2b, S3a and S3b

- Analysis on all the amino acids but processed separately- described in Section 6.2-, retraining the two predictors on 20 csv files, and storing the results obtained in tables. The tables S1, S2, S3, S4, S5 and S6, show on the row the 20 different amino acids, and on the columns the 6 best features obtained, with their scores and the MSE, RMSE, MAE. Then the best features obtained have been dropped from the data-set, and retraining the algorithms, we obtained the results shown in the tables S7, S8, S9, S10, S11 and S12

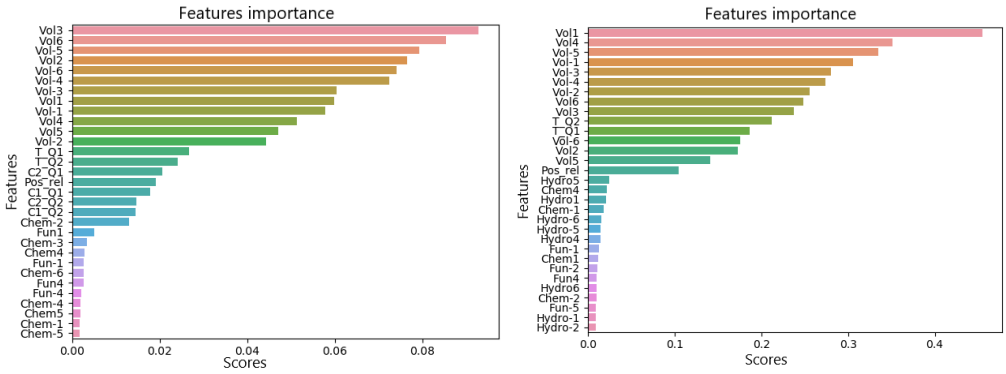
- Analysis on all the amino acids but processed together, -Section 6.3- training the algorithms on a unique csv file with all the data detected for every occurrence and storing Features Importance (FI) graphs S4a and S4b . The same analysis is replicated but dropping the best one feature from the data-set, obtaining the results shown in S5a and S5b



(a) Analysis with RF

(b) Analysis with MLP

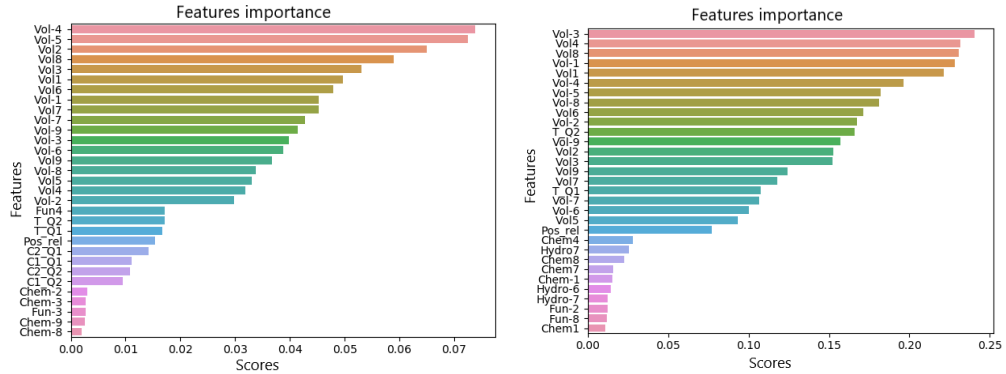
Fig. S1. Feature importance obtained for Lysine volume prediction, trained using data about the 3 amino acids preceding and following the Amino acid 0 (AA_0) in the primary sequence of the protein.



(a) Analysis with RF

(b) Analysis with MLP

Fig. S2. Feature importance obtained for Lysine volume prediction, trained using data about the 6 amino acids preceding and following the AA_0 in the primary sequence of the protein.



(a) Analysis with RF

(b) Analysis with MLP

Fig. S3. Feature importance obtained for Lysine volume prediction, trained using data about the 9 amino acids preceding and following the AA_0 in the primary sequence of the protein.

Table S1. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 3 amino acids preceding and following AA_0 in the primary sequence of the protein.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Vol-1	0,1834	Vol1	0,1433	Vol2	0,1418	Vol-3	0,1146	Vol-2	0,0986	Vol3	0,0886	0,5385	1,0601	1,0296
ALA	7694	Vol1	0,1926	Vol-2	0,1664	Vol-3	0,1513	Vol-1	0,1087	Vol2	0,1003	Vol3	0,0741	0,7402	2,4231	1,5566
VAL	3572	Vol-3	0,1964	Vol3	0,1649	Vol2	0,1461	Vol-2	0,1087	Vol-1	0,0944	Vol1	0,0925	1,8636	11,9905	3,4627
LEU	8229	Vol-3	0,1423	Vol1	0,1414	Vol3	0,1320	Vol2	0,1259	Vol-1	0,1127	Vol-2	0,0952	2,8407	22,0444	4,6951
ILE	4101	Vol3	0,1533	Fun2	0,1480	Vol1	0,1404	Vol2	0,1218	Vol-3	0,0887	Vol-2	0,0765	2,6804	23,1006	4,8063
MET	1105	Vol2	0,3989	Vol-2	0,1980	Vol1	0,1033	Vol-1	0,0746	Vol3	0,0658	Vol-3	0,0570	3,8676	43,4884	6,5946
PRO	1879	Vol1	0,3227	Vol-1	0,1904	Vol3	0,1225	Vol2	0,0906	Vol-3	0,0613	Vol-2	0,0479	1,6259	7,8896	2,8088
SER	2723	Vol1	0,1933	Vol-3	0,1618	T_Q2	0,1341	Vol-2	0,1130	Vol3	0,0795	Vol2	0,0756	1,0704	3,8359	1,9586
THR	2575	Vol2	0,2915	Vol-1	0,2099	Vol-2	0,1409	Vol1	0,0693	Vol3	0,0497	Vol-3	0,0493	1,6654	7,3203	2,7056
CYS	12	Vol-1	0,0988	Chem-1	0,0791	Vol2	0,0772	T_Q2	0,0665	Hydro1	0,0656	Vol3	0,0600	4,3423	20,7714	4,5576
ASN	850	Vol1	0,2073	Vol2	0,2048	T_Q2	0,0737	Vol-3	0,0659	Vol3	0,0649	Vol-2	0,0579	2,4039	14,6888	3,8326
GLN	2504	Vol2	0,1905	Vol-1	0,1839	Pos_rel	0,1160	Vol1	0,1028	Vol-3	0,0776	Vol3	0,0715	5,9150	90,3575	9,5057
PHE	2928	Vol-3	0,2460	Vol-2	0,1669	Vol3	0,1348	Vol1	0,0921	Vol-1	0,0864	Vol2	0,0603	4,6112	62,6633	7,9160
TYR	1275	Vol2	0,2504	Vol-3	0,1506	Vol-2	0,1411	Vol-1	0,1132	Vol3	0,0805	Vol1	0,0575	5,1439	84,7859	9,2079
TRP	947	Vol-1	0,3870	Vol3	0,2328	Vol2	0,0798	Vol1	0,0582	Vol-2	0,0537	Fun2	0,0477	4,8437	129,4846	11,3791
LYS	8821	Vol2	0,1441	Vol1	0,1229	Vol-3	0,1208	Vol-1	0,0994	Vol3	0,0969	Vol-2	0,0861	9,1833	186,7161	13,6644
HIS	5258	Vol-1	0,2123	Vol3	0,1802	Vol2	0,1481	Vol-2	0,1170	Vol-3	0,1014	Vol1	0,0608	3,3231	32,2098	5,6754
ARG	1572	Vol1	0,2031	Pos_rel	0,1527	Vol2	0,1064	Vol-3	0,0884	Vol3	0,0781	Vol-2	0,0726	10,5045	291,1569	17,0633
ASP	3293	Chem-2	0,2049	Vol3	0,1764	Vol-2	0,1586	Vol1	0,1174	Vol-3	0,0749	Vol2	0,0523	2,8097	20,4244	4,5193
GLU	6404	Vol-2	0,1971	Vol1	0,1445	Vol-1	0,1183	Vol3	0,0856	Vol2	0,0854	Chem3	0,0694	5,2578	69,3417	8,3272

Table S2. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 3 amino acids preceding and following AA_0 in the primary sequence of the protein.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Pos_rel	0,1315	T_Q1	0,1271	T_Q2	0,1196	Hydro-6	0,0853	Chem4	0,0770	Hydro3	0,0692	1,6121	4,9393	2,2225
ALA	7694	T_Q2	0,0918	T_Q1	0,0826	Pos_rel	0,0532	Hydro-4	0,0403	Hydro-5	0,0315	Chem-5	0,0152	1,8956	8,2349	2,8697
VAL	3572	Chem8	0,1403	Hydro-8	0,0864	Chem-1	0,0454	Hydro-4	0,0360	T_Q1	0,0357	Chem-4	0,0257	4,1212	36,6524	6,0541
LEU	8229	T_Q2	0,1240	T_Q1	0,1189	Chem-6	0,0663	Hydro5	0,0639	Hydro-2	0,0391	Chem5	0,0372	6,7312	93,1337	9,6506
ILE	4101	T_Q1	0,0211	Pos_rel	0,0182	Hydro5	0,0157	Hydro9	0,0128	T_Q2	0,0105	Hydro-8	0,0103	6,8182	93,1517	9,6515
MET	1105	Chem9	0,0771	Hydro-2	0,0678	Chem-1	0,0470	Pos_rel	0,0342	Chem-3	0,0321	T_Q2	0,0298	7,6145	124,5784	11,1615
SER	2723	Pos_rel	0,2545	T_Q2	0,1299	T_Q1	0,0989	Chem5	0,0587	Chem9	0,0378	Fun-5	0,0267	2,8237	16,9648	4,1188
PRO	1879	T_Q1	0,0614	T_Q2	0,0240	Pos_rel	0,0237	Chem-4	0,0175	Hydro3	0,0119	Chem-7	0,0105	3,7529	48,0048	6,9286
THR	2575	T_Q1	0,1608	T_Q2	0,1483	Chem7	0,0364	Fun-5	0,0134	Hydro-4	0,0133	Hydro1	0,0132	3,7133	30,9326	5,5617
CYS	12	T_Q1	0,0461	Pos_rel	0,0084	Chem-9	0,0022	Chem3	0,0013	Fun-6	0,0012	Hydro-6	0,0011	2,0410	5,0489	2,2470
ASN	850	T_Q2	0,1876	Pos_rel	0,1855	Chem-9	0,0669	Hydro-3	0,0644	Hydro-9	0,0525	Chem2	0,0451	3,9476	25,3953	5,0394
GLN	2504	Pos_rel	0,1150	T_Q1	0,0434	Hydro-9	0,0407	Chem-9	0,0287	Fun-9	0,0286	Chem-6	0,0224	10,7624	211,8817	14,5562
PHE	2928	Chem-1	0,1004	Hydro7	0,0528	Chem-7	0,0439	T_Q1	0,0433	T_Q2	0,0312	Chem7	0,0219	12,1458	261,9531	16,1850
TYR	1275	T_Q1	0,0296	Pos_rel	0,0202	Fun4	0,0162	Hydro4	0,0140	Chem4	0,0061	Chem6	0,0044	14,0738	328,5074	18,1248
TRP	947	Hydro1	0,0123	Fun1	0,0079	Chem-2	0,0025	Chem-6	0,0007	Chem7	0,0004	Hydro8	0,0001	8,6417	204,3007	14,2934
LYS	8821	T_Q2	0,2676	T_Q1	0,1934	Pos_rel	0,1222	Hydro-9	0,0480	Chem-5	0,0471	Chem4	0,0396	15,6154	451,5171	21,2489
HIS	5258	T_Q2	0,3339	T_Q1	0,2945	Hydro8	0,0734	Chem6	0,0733	Chem3	0,0508	Pos_rel	0,0467	7,5599	122,2233	11,0555
ARG	1572	Pos_rel	0,1728	T_Q2	0,1509	T_Q1	0,1340	Chem4	0,1005	Chem1	0,0110	Hydro3	0,0099	15,9363	591,0136	24,3108
ASP	3293	T_Q2	0,1385	Pos_rel	0,0930	T_Q1	0,0785	Chem-4	0,0656	Chem6	0,0423	Chem-6	0,0224	5,1691	46,9522	6,8522
GLU	6404	T_Q1	0,2719	T_Q2	0,1062	Hydro-5	0,0477	Hydro-1	0,0395	Chem-1	0,0308	Pos_rel	0,0306	9,7335	160,3813	12,6642

Table S3. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 6 amino acids preceding and following AA_0 in the primary sequence of the protein

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Vol1	0,1698	Vol-1	0,1033	Vol5	0,1001	Vol2	0,0916	Vol-5	0,0889	Vol4	0,0693	0,5151	1,0495	1,0244
ALA	7694	Vol1	0,1384	Vol-3	0,0851	Vol3	0,0735	Vol-2	0,0721	Vol-6	0,0699	Vol-1	0,0694	0,6254	1,4676	1,2114
VAL	3572	Vol4	0,2384	Vol-6	0,1291	Vol-3	0,0920	Vol-4	0,0801	Vol2	0,0751	Vol5	0,0718	1,4999	8,4438	2,9058
LEU	8229	Vol-5	0,1410	Vol1	0,1047	Vol-3	0,0874	Vol3	0,0855	Vol4	0,0683	Vol5	0,0649	2,5713	19,2820	4,3911
ILE	4101	Vol4	0,1442	Fun2	0,1438	Vol1	0,0908	Vol-5	0,0699	Vol5	0,0650	Vol-4	0,0612	2,4010	16,3103	4,0386
MET	1105	Vol2	0,3211	Vol-2	0,1964	Vol6	0,1458	Vol-4	0,0649	Vol-1	0,0536	Vol4	0,0333	2,8892	33,9461	5,8263
PRO	1879	Vol1	0,2198	Vol-6	0,1313	Vol5	0,1036	Vol-1	0,0823	Vol3	0,0810	Vol2	0,0545	1,4459	7,1697	2,6776
SER	2723	Vol1	0,1513	Vol-5	0,0994	Vol6	0,0880	Vol-2	0,0819	Vol2	0,0818	T_Q2	0,0802	1,0899	3,3742	1,8369
THR	2575	Vol2	0,2530	Vol-1	0,1098	Vol-2	0,0965	Vol5	0,0864	Vol6	0,0651	Vol4	0,0526	1,4738	6,0494	2,4595
CYS	12	Vol-1	0,0692	Vol-3	0,0544	Vol6	0,0500	Vol4	0,0497	Chem-6	0,0495	Fun2	0,0459	3,4113	14,6685	3,8299
ASN	850	Vol6	0,3460	Vol1	0,1414	Vol4	0,0697	Vol2	0,0499	Chem-2	0,0450	Vol-4	0,0414	2,0666	9,8000	3,1305
GLN	2504	Pos_rel	0,1510	Vol5	0,0980	Vol2	0,0839	Vol-3	0,0838	Vol-1	0,0829	Vol4	0,0651	4,4296	60,6124	7,7854
PHE	2928	Vol-3	0,2148	Vol-2	0,1099	Vol3	0,1066	Vol6	0,0888	Vol5	0,0705	Vol-1	0,0659	3,5674	41,6460	6,4534
TYR	1275	Vol-4	0,2747	Vol-3	0,1355	Vol-6	0,1111	Vol2	0,0936	T_Q2	0,0561	Vol3	0,0471	4,0024	60,1347	7,7547
TRP	947	Vol-1	0,3685	Vol3	0,2009	Vol5	0,0521	Vol2	0,0520	Vol6	0,0421	Vol-5	0,0382	4,3107	84,9141	9,2149
LYS	8821	Vol3	0,0928	Vol6	0,0855	Vol-5	0,0792	Vol2	0,0765	Vol-6	0,0742	Vol-4	0,0725	8,5886	166,4297	12,9008
HIS	5258	Vol-1	0,1669	Vol3	0,1296	Vol2	0,0886	Vol-2	0,0781	Vol-4	0,0714	Vol-3	0,0655	3,1346	31,0062	5,5683
ARG	1572	Vol-6	0,3594	Fun4	0,1299	Vol-4	0,0748	Vol-3	0,0704	Vol-2	0,0539	Vol1	0,0416	10,1136	304,4501	17,4485
ASP	3293	Pos_rel	0,2639	Vol-2	0,1132	Vol3	0,1116	Vol5	0,0906	Vol1	0,0645	Vol-4	0,0384	2,5156	17,3905	4,1702
GLU	6404	Vol-2	0,1956	Vol5	0,0665	Pos_rel	0,0658	Vol1	0,0601	Vol-5	0,0542	Vol-1	0,0508	5,1794	73,8164	8,5916

Table S4. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 6 amino acids preceding and following AA_0 in the primary sequence of the protein.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Vol4	0,8673	Vol-1	0,4215	Vol-4	0,4098	Vol-3	0,3903	Vol2	0,3813	Vol6	0,3737	0,5085	0,7594	0,8714
ALA	7694	Vol1	1,4419	Vol-4	0,5580	Vol-5	0,4331	Pos_rel	0,3939	Vol5	0,3824	Vol4	0,3394	0,7305	1,4633	1,2097
VAL	3572	Vol4	1,1559	Vol-6	0,4463	Vol1	0,4160	Vol-3	0,3949	Vol-4	0,3714	Vol5	0,3635	1,6341	6,8890	2,6247
LEU	8229	Vol-1	0,4251	Vol4	0,4047	Vol6	0,3624	Vol-4	0,3512	Vol2	0,3358	Vol-5	0,2403	3,2581	25,3667	5,0365
ILE	4101	Vol-6	0,5187	Vol-4	0,4597	Vol4	0,4170	Vol2	0,2938	Vol-5	0,2915	Vol1	0,2804	2,8417	19,6795	4,4362
MET	1105	Vol-1	0,5654	Vol2	0,4929	Vol-2	0,4103	Vol-4	0,1614	Vol1	0,1040	Vol5	0,0931	3,7151	45,7790	6,7660
PRO	1879	T_Q1	0,3476	Vol1	0,3094	Vol-4	0,2590	Vol-6	0,2550	Vol-2	0,2497	Vol3	0,2369	1,6850	6,8611	2,6194
SER	2724	T_Q3	1,4744	Vol3	1,4588	Vol5	1,4509	Vol-3	1,4495	Vol2	1,3373	Vol-5	1,2768	2,1998	4,4646	2,8613
THR	2575	Vol-1	0,4314	Vol-4	0,4114	Vol6	0,2878	Vol3	0,2839	Vol4	0,2494	T_Q1	0,2378	1,6942	6,7493	2,5979
CYS	12	Vol-5	0,1259	Vol-6	0,1185	Vol-1	0,1120	Vol3	0,1030	T_Q1	0,0631	Vol-2	0,0405	1,2641	1,7082	1,3070
ASN	850	Vol2	0,8380	Vol-1	0,3149	Vol-4	0,2613	Vol3	0,2600	Vol6	0,2263	Pos_rel	0,1752	2,3500	11,8889	3,4480
GLN	2504	Vol-4	0,4817	Vol2	0,3209	Vol5	0,3043	Vol3	0,2857	Vol1	0,2547	Vol-1	0,2423	5,1583	72,9790	8,5428
PHE	2928	Vol-3	0,5594	Vol6	0,4829	Vol-2	0,4828	Vol4	0,4686	Vol1	0,4298	Vol-4	0,3567	4,2988	50,0656	7,0757
TYR	1275	Vol5	0,9281	Vol-4	0,2171	Vol4	0,1906	Vol2	0,1505	Vol1	0,1477	Vol-2	0,1361	6,2633	125,2519	11,1916
TRP	947	Vol5	0,5520	Vol-1	0,1844	Vol-2	0,1755	Vol3	0,1442	Vol2	0,1135	Vol1	0,1106	5,1088	117,5901	10,8439
LYS	8821	Vol1	0,3911	Vol4	0,3503	Vol-5	0,3327	Vol-1	0,3124	Vol-3	0,2793	Vol2	0,2666	9,9423	205,4248	14,3326
HIS	5258	Vol3	0,4523	Vol-5	0,3744	Vol6	0,2888	Vol2	0,2843	Vol-2	0,2445	T_Q2	0,2402	3,4216	31,2673	5,5917
ARG	1572	Vol-4	0,1843	Vol6	0,1735	Pos_rel	0,1557	Vol-5	0,1504	Vol-3	0,1346	T_Q2	0,1289	10,2748	276,1318	16,6172
ASP	3293	Vol5	0,5271	Vol-4	0,4231	Vol-1	0,3460	Vol-2	0,3097	Vol2	0,3093	Vol-5	0,2822	2,8240	19,1003	4,3704
GLU	6404	Vol1	0,3760	Vol3	0,2407	Vol-2	0,2318	Vol-5	0,2241	Vol-3	0,2139	Vol6	0,2020	5,8611	84,9153	9,2149

Table S5. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 9 amino acids preceding and following AA_0 in the primary sequence of the protein.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Vol1	0,1773	Vol9	0,0836	Vol2	0,0790	Vol-1	0,0720	Vol4	0,0644	Vol-9	0,0576	0,4468	0,7666	0,8755
ALA	7694	Vol1	0,1093	Vol9	0,0824	Vol-2	0,0739	Vol6	0,0693	Vol5	0,0661	Vol-4	0,0500	0,6009	1,3414	1,1582
VAL	3572	Vol-8	0,1201	Vol7	0,1055	Vol5	0,0895	Vol4	0,0804	Vol-6	0,0788	Vol-4	0,0704	1,5162	6,9590	2,6380
LEU	8229	Vol-5	0,0818	Vol3	0,0744	Vol8	0,0691	Vol1	0,0611	Vol-3	0,0592	Vol-7	0,0571	2,4643	16,8462	4,1044
ILE	4101	Fun2	0,1432	Vol7	0,1140	Vol1	0,0785	Vol4	0,0666	Vol-7	0,0584	Vol9	0,0559	2,2002	12,6165	3,5520
MET	1105	Vol-7	0,3166	Vol2	0,2078	Vol-4	0,0628	Vol6	0,0494	Vol-6	0,0475	Vol-2	0,0383	3,0405	32,3352	5,6864
PRO	1879	Vol9	0,2120	Vol-8	0,1307	Vol1	0,1075	Vol-6	0,0789	Vol-7	0,0608	Vol-1	0,0586	1,3758	5,7758	2,4033
SER	2723	Vol-7	0,1271	Vol-5	0,1178	Vol8	0,0838	Vol1	0,0703	Vol6	0,0679	Vol-3	0,0631	1,0885	3,4037	1,8449
THR	2575	Vol2	0,1954	Vol-9	0,1434	Vol7	0,0720	Vol-2	0,0679	Vol5	0,0644	Vol-1	0,0590	1,4648	5,6602	2,3791
CYS	12	Pos_rel	0,0356	Vol-8	0,0329	Fun-8	0,0312	Vol-1	0,0306	Chem-2	0,0268	Vol-3	0,0255	1,6509	4,8096	2,1931
ASN	850	Vol6	0,2940	Vol1	0,1147	Vol-9	0,0661	Vol4	0,0529	Vol9	0,0463	Vol2	0,0430	1,8991	8,5390	2,9222
GLN	2504	Vol1	0,1914	Vol5	0,1320	Vol-3	0,1207	Vol2	0,1103	Vol4	0,0644	Vol3	0,0332	5,1124	70,1302	8,3744
PHE	2928	Vol-3	0,1923	Vol-7	0,1044	Vol8	0,1000	Vol6	0,0802	Vol3	0,0740	Vol-5	0,0703	3,4181	37,1660	6,0964
TYR	1275	Vol-3	0,2488	Vol-9	0,1096	Vol-6	0,0832	Fun-5	0,0808	Vol-2	0,0794	Vol8	0,0621	4,6987	71,3403	8,4463
TRP	947	Vol3	0,2127	Vol-3	0,1332	Vol-7	0,1058	Vol9	0,1005	Vol-4	0,0544	Vol-2	0,0444	4,2929	58,8710	7,6727
LYS	8821	Vol-4	0,0738	Vol-5	0,0725	Vol2	0,0649	Vol8	0,0589	Vol3	0,0531	Vol1	0,0497	8,5366	165,8054	12,8765
HIS	5258	Vol8	0,1570	Vol-1	0,0777	Vol-4	0,0707	Vol3	0,0657	Vol9	0,0602	T_Q1	0,0479	2,8392	24,4970	4,9494
ARG	1572	Vol-6	0,3469	Fun4	0,1020	Vol-3	0,0605	Vol9	0,0574	Vol-4	0,0557	Vol-2	0,0495	10,1957	303,4559	17,4200
ASP	3293	Pos_rel	0,2579	Vol-2	0,1090	Vol3	0,0912	Vol9	0,0591	Vol-7	0,0557	Vol1	0,0467	2,5252	16,4076	4,0506
GLU	6404	Vol-2	0,2157	Vol9	0,0850	Vol1	0,0468	Vol8	0,0449	Vol-8	0,0410	Pos_rel	0,0407	4,9968	60,2162	7,7599

Table S6. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 9 amino acids preceding and following AA_0 in the primary sequence of the protein.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Vol4	0,6262	Vol-3	0,4771	Vol8	0,4374	Vol-1	0,2824	Vol2	0,2592	Vol7	0,2531	0,4813	0,6144	0,7839
ALA	7694	Vol1	0,4943	Vol5	0,4866	Vol4	0,3666	Vol-9	0,2977	Vol-4	0,2831	Vol-8	0,2807	0,6569	1,1581	1,0762
VAL	3572	Vol4	0,4700	Vol-7	0,3716	Vol1	0,3591	Vol-3	0,3255	Vol-8	0,3251	Vol-6	0,3206	1,7685	6,9803	2,6420
LEU	8229	Vol8	0,2856	Vol4	0,2808	Vol-4	0,2532	Vol-9	0,2408	Vol2	0,2404	Vol-1	0,2148	3,2014	22,8169	4,7767
ILE	4101	Vol7	0,6416	Vol-4	0,5097	Vol4	0,3186	Vol-6	0,2919	Vol-7	0,2615	Vol9	0,1442	2,6705	17,1133	4,1368
MET	1105	Vol-1	0,4665	Vol2	0,4102	Vol-2	0,1858	Vol8	0,1009	Vol-7	0,0743	Vol-5	0,0571	4,1665	63,7989	7,9874
PRO	1879	Vol1	0,3496	Vol-8	0,3342	T_Q1	0,2380	Vol-6	0,2231	Vol3	0,1837	Vol2	0,1473	1,5146	6,7319	2,5946
SER	2724	Vol5	1,3331	Vol-9	1,2909	Vol3	1,2698	Vol2	1,2647	T_Q3	1,2569	Vol-3	1,2529	2,0998	4,1364	2,7710
THR	2575	Vol7	0,4619	Vol-4	0,2654	Vol-1	0,2355	Vol1	0,2344	Vol6	0,2228	Vol2	0,2097	1,7111	7,2330	2,6894
CYS	12	Vol5	0,0155	Vol-7	0,0065	Pos_rel	0,0046	Vol-3	0,0028	Vol-1	0,0012	Vol1	0,0010	1,2494	2,6384	1,6243
ASN	850	Vol2	0,8157	Vol-8	0,2706	Vol-4	0,1940	Vol1	0,1858	Vol6	0,1492	Vol-7	0,1091	2,1876	10,1797	3,1906
GLN	2504	Vol-4	0,4470	Vol1	0,4387	Vol2	0,2434	Vol3	0,1984	Vol4	0,1841	Vol5	0,1513	5,6628	87,1403	9,3349
PHE	2928	Vol8	0,5194	Vol4	0,3965	Vol-3	0,3691	Vol-8	0,3168	Vol5	0,2785	Vol-4	0,2507	3,8403	38,6336	6,2156
TYR	1275	Vol-3	0,2027	Vol-4	0,1286	Vol5	0,1003	Vol9	0,0983	Vol-6	0,0565	Vol2	0,0494	7,1454	144,7457	12,0310
TRP	947	Vol-8	0,2164	Vol-3	0,2085	Vol-7	0,1919	Vol5	0,1302	Vol4	0,0777	Vol9	0,0571	5,5169	107,7148	10,3786
LYS	8821	Vol-4	0,2317	Vol8	0,2213	Vol1	0,2192	Vol-3	0,2125	Vol-5	0,2035	Vol-1	0,2006	9,4384	187,7494	13,7022
HIS	5258	Vol-5	0,2910	Vol3	0,2669	Vol8	0,2639	T_Q2	0,2584	Vol6	0,2439	Vol2	0,2262	3,3458	28,8594	5,3721
ARG	1572	Vol-3	0,1355	Vol-5	0,1349	T_Q1	0,1202	Vol7	0,1089	Vol-6	0,1033	Vol9	0,0999	10,2835	274,4339	16,5660
ASP	3293	Vol5	0,3949	Vol-4	0,3304	Vol-8	0,2699	Vol-2	0,2640	Vol-1	0,2595	Vol-5	0,2462	2,7899	17,8939	4,2301
GLU	6404	Vol1	0,2157	Vol-2	0,2125	T_Q2	0,1968	Vol9	0,1760	Vol-3	0,1721	Vol-1	0,1657	5,4416	65,0220	8,0636

Table S7. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 3 amino acids preceding and following AA_0 in the primary sequence of the protein without considering their volumes.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Pos_rel	0,1666	T_Q2	0,1434	T_Q1	0,1395	C2_Q1	0,1222	C1_Q1	0,1086	C1_Q2	0,1054	1,2735	4,1107	2,0275
ALA	7694	T_Q1	0,1639	T_Q2	0,1559	C1_Q1	0,1237	C1_Q2	0,1166	C2_Q2	0,1149	Pos_rel	0,1133	1,6812	9,4946	3,0813
VAL	3572	T_Q1	0,1732	T_Q2	0,1491	C1_Q1	0,1157	C2_Q1	0,1030	C1_Q2	0,0988	Pos_rel	0,0984	3,7327	37,3959	6,1152
LEU	8229	Pos_rel	0,1599	T_Q1	0,1497	T_Q2	0,1329	C1_Q1	0,0996	C2_Q1	0,0992	C2_Q2	0,0963	5,2988	73,3236	8,5629
ILE	4101	Pos_rel	0,1668	Fun2	0,1480	T_Q2	0,1280	T_Q1	0,1206	C1_Q2	0,1073	C2_Q1	0,0984	5,0065	68,3621	8,2681
MET	1105	T_Q1	0,1356	T_Q2	0,1259	Pos_rel	0,1022	Chem-1	0,0967	C2_Q2	0,0852	C1_Q2	0,0837	6,7622	114,7796	10,7135
PRO	1879	T_Q1	0,1947	Pos_rel	0,1419	C2_Q1	0,1417	T_Q2	0,1304	C1_Q1	0,1204	C2_Q2	0,0959	3,3439	42,6728	6,5324
SER	2723	T_Q2	0,2709	T_Q1	0,1701	Pos_rel	0,1085	C2_Q2	0,1066	C1_Q2	0,0982	C1_Q1	0,0908	2,2181	13,3151	3,6490
THR	2575	T_Q1	0,2715	Pos_rel	0,1375	T_Q2	0,1369	C2_Q2	0,0972	C2_Q1	0,0936	C1_Q2	0,0884	3,0784	25,8705	5,0863
CYS	12	Hydro1	0,1050	Fun-1	0,1019	Hydro2	0,1017	Fun2	0,0984	Pos_rel	0,0824	Chem-1	0,0789	4,5023	28,4631	5,3351
ASN	850	T_Q1	0,1688	Pos_rel	0,1315	T_Q2	0,1165	C2_Q2	0,0847	C1_Q1	0,0835	C1_Q2	0,0810	3,2641	25,6902	5,0685
GLN	2504	Pos_rel	0,2385	T_Q1	0,1373	T_Q2	0,1072	C2_Q1	0,1061	C1_Q1	0,0893	C2_Q2	0,0863	10,0373	183,7821	13,5566
PHE	2928	Pos_rel	0,1625	T_Q2	0,1425	T_Q1	0,1194	C1_Q1	0,1015	C2_Q2	0,0977	C1_Q2	0,0966	9,1226	200,4889	14,1594
TYR	1275	T_Q2	0,1987	Pos_rel	0,1862	T_Q1	0,1728	C2_Q2	0,1064	C2_Q1	0,0961	C1_Q1	0,0881	9,7599	243,6182	15,6083
TRP	947	T_Q1	0,1722	T_Q2	0,1505	Chem2	0,1348	C2_Q2	0,1091	C1_Q2	0,1030	Pos_rel	0,1022	12,4209	443,2285	21,0530
LYS	8821	T_Q1	0,1570	T_Q2	0,1504	C2_Q2	0,1343	C1_Q1	0,1166	Pos_rel	0,1110	C1_Q2	0,1096	13,5857	383,1098	19,5732
HIS	5258	T_Q2	0,1687	T_Q1	0,1517	Pos_rel	0,1193	C2_Q2	0,1048	C1_Q2	0,1014	C1_Q1	0,0959	6,5153	106,9160	10,3400
ARG	1572	Pos_rel	0,3813	T_Q1	0,1822	T_Q2	0,1118	C2_Q2	0,0650	C1_Q1	0,0648	C1_Q2	0,0564	15,4856	611,0910	24,7203
ASP	3293	Chem-2	0,2085	T_Q2	0,1335	C1_Q1	0,1292	T_Q1	0,1203	Pos_rel	0,0911	C1_Q2	0,0878	4,4294	40,9467	6,3990
GLU	6404	Pos_rel	0,1612	T_Q2	0,1155	T_Q1	0,1048	Chem3	0,1001	C2_Q1	0,0909	C1_Q1	0,0861	8,0992	137,4181	11,7225

Table S8. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 3 amino acids preceding and following AA_0 in the protein without their volumes.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Pos_rel	0,1729	T_Q2	0,0865	T_Q1	0,0821	Hydro3	0,0335	Chem3	0,0285	Chem1	0,0213	1,5946	4,9402	2,2227
ALA	7694	T_Q1	0,0535	T_Q2	0,0527	Pos_rel	0,0477	Chem-1	0,0287	Chem-3	0,0093	Chem2	0,0086	2,1218	11,1592	3,3405
VAL	3572	T_Q1	0,1570	Pos_rel	0,0884	T_Q2	0,0300	Hydro-2	0,0277	Hydro2	0,0218	Hydro-3	0,0123	4,8140	50,3098	7,0929
LEU	8229	Pos_rel	0,2112	Chem1	0,1053	T_Q2	0,0804	Hydro-2	0,0654	T_Q1	0,0506	Hydro1	0,0419	6,8586	93,2199	9,6550
ILE	4101	Pos_rel	0,1116	T_Q1	0,0796	Chem-2	0,0504	Hydro1	0,0425	Fun2	0,0346	T_Q2	0,0324	6,8473	93,7131	9,6806
MET	1105	Hydro-2	0,1481	T_Q2	0,0340	Chem-3	0,0244	Chem-1	0,0204	T_Q1	0,0198	Chem1	0,0197	7,8958	118,5317	10,8872
PRO	1879	T_Q1	0,0575	Pos_rel	0,0415	T_Q2	0,0137	Hydro3	0,0097	Fun-3	0,0065	Chem-2	0,0059	3,9080	47,1117	6,8638
SER	2724	T_Q2	1,2521	Pos_rel	1,2485	T_Q3	1,1391	Hydro3	1,0566	Chem3	1,0508	Fun-3	1,0216	3,7085	16,6581	4,9570
THR	2575	T_Q1	0,4352	T_Q2	0,1135	Chem-1	0,0173	Chem-2	0,0169	Hydro-3	0,0144	Pos_rel	0,0143	3,8131	31,5964	5,6211
CYS	12	T_Q1	0,3208	T_Q2	0,2766	Pos_rel	0,2482	Chem-3	0,0471	Hydro-3	0,0407	Chem1	0,0187	3,4752	18,2250	4,2691
ASN	850	Pos_rel	0,4484	T_Q2	0,4362	Hydro1	0,1821	T_Q1	0,1371	Chem2	0,0957	Hydro-3	0,0290	4,1435	33,6089	5,7973
GLN	2504	Pos_rel	0,2281	T_Q1	0,0966	T_Q2	0,0839	Hydro1	0,0575	Hydro-2	0,0311	Chem-2	0,0249	11,1820	205,7520	14,3441
PHE	2928	Chem-1	0,1171	Fun-1	0,0411	T_Q1	0,0379	T_Q2	0,0373	Hydro-1	0,0211	Hydro1	0,0135	11,4726	248,3436	15,7589
TYR	1275	Pos_rel	0,1112	T_Q2	0,1107	T_Q1	0,0723	Hydro1	0,0288	Chem1	0,0203	Fun1	0,0092	12,9676	318,9281	17,8586
TRP	947	Hydro2	0,0578	Hydro-3	0,0547	Chem2	0,0380	Fun2	0,0327	T_Q1	0,0273	Chem-1	0,0124	15,7977	541,6705	23,2738
LYS	8821	T_Q2	0,1458	T_Q1	0,0895	Pos_rel	0,0781	Chem-2	0,0773	Chem3	0,0558	Chem-1	0,0473	16,5806	479,7815	21,9039
HIS	5258	T_Q2	0,2151	T_Q1	0,0870	Pos_rel	0,0783	Chem3	0,0314	Hydro-1	0,0102	Hydro1	0,0097	8,5671	140,7214	11,8626
ARG	1572	Pos_rel	0,4912	T_Q1	0,2465	T_Q2	0,1879	Chem1	0,0126	Hydro-1	0,0070	Fun-2	0,0053	17,4146	674,7626	25,9762
ASP	3293	T_Q1	0,1578	T_Q2	0,1140	Pos_rel	0,1000	Hydro2	0,0300	Fun2	0,0254	Chem1	0,0243	5,6960	55,6428	7,4594
GLU	6404	Pos_rel	0,2161	T_Q1	0,1922	T_Q2	0,1526	Hydro-1	0,0531	Hydro-2	0,0420	Chem-2	0,0385	10,4980	184,3700	13,5783

Table S9. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 6 amino acids preceding and following AA_0 in the protein without their volumes.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Pos_rel	0,1737	T_Q2	0,1409	T_Q1	0,1263	C2_Q1	0,1029	C1_Q2	0,0970	C1_Q1	0,0968	1,2706	3,8456	1,9610
ALA	7694	T_Q1	0,1575	T_Q2	0,1518	C1_Q2	0,1178	C2_Q2	0,1166	C1_Q1	0,1087	Pos_rel	0,1041	1,5962	7,8409	2,8002
VAL	3572	T_Q1	0,1591	T_Q2	0,1532	C1_Q1	0,1052	Pos_rel	0,0975	C2_Q1	0,0973	C2_Q2	0,0918	3,1969	28,3987	5,3290
LEU	8229	Pos_rel	0,1549	T_Q1	0,1325	T_Q2	0,1289	C2_Q1	0,0908	C1_Q1	0,0888	C1_Q2	0,0878	5,3895	77,3227	8,7933
ILE	4101	Pos_rel	0,1539	Fun2	0,1475	T_Q2	0,1271	T_Q1	0,1162	C1_Q2	0,1039	C2_Q1	0,0932	4,8841	67,0290	8,1871
MET	1105	T_Q1	0,1285	Pos_rel	0,1231	C2_Q1	0,1077	T_Q2	0,1043	C1_Q1	0,0986	C1_Q2	0,0712	5,7727	108,9736	10,4390
PRO	1879	T_Q1	0,1882	C2_Q1	0,1395	Pos_rel	0,1315	T_Q2	0,1239	C1_Q1	0,1065	C2_Q2	0,0948	3,1011	40,2511	6,3444
SER	2723	T_Q2	0,2409	T_Q1	0,1423	Pos_rel	0,1019	C2_Q2	0,0996	C1_Q1	0,0939	C1_Q2	0,0912	2,3348	14,2305	3,7723
THR	2575	T_Q1	0,2560	T_Q2	0,1224	Pos_rel	0,1223	C2_Q2	0,0901	C2_Q1	0,0880	C1_Q2	0,0862	2,9837	25,3166	5,0316
CYS	12	Hydro2	0,1049	Pos_rel	0,0857	Fun-1	0,0722	T_Q1	0,0695	Chem-6	0,0659	Fun2	0,0654	3,9005	26,0431	5,1032
ASN	850	T_Q1	0,1482	Pos_rel	0,1058	T_Q2	0,1048	C2_Q2	0,0839	C1_Q2	0,0787	C1_Q1	0,0773	3,2059	24,9948	4,9995
GLN	2504	Pos_rel	0,2200	T_Q1	0,1448	T_Q2	0,1039	C2_Q1	0,0960	C1_Q1	0,0907	C1_Q2	0,0784	8,1344	156,4942	12,5098
PHE	2928	Pos_rel	0,1509	T_Q2	0,1356	T_Q1	0,1200	C1_Q2	0,1034	C2_Q2	0,1011	C1_Q1	0,0996	8,0773	155,0028	12,4500
TYR	1275	T_Q2	0,1987	T_Q1	0,1769	Pos_rel	0,1427	C1_Q2	0,1032	C2_Q1	0,0938	C1_Q1	0,0911	7,3757	161,3654	12,7030
TRP	947	T_Q2	0,1372	T_Q1	0,1360	Chem2	0,1356	Pos_rel	0,1024	C2_Q1	0,1018	C1_Q2	0,0920	10,4333	337,8107	18,3796
LYS	8821	T_Q1	0,1490	T_Q2	0,1473	C2_Q2	0,1278	C1_Q1	0,1075	Pos_rel	0,1013	C1_Q2	0,1012	13,3227	375,6712	19,3822
HIS	5258	T_Q2	0,1816	T_Q1	0,1358	Pos_rel	0,1281	C2_Q2	0,0951	C1_Q2	0,0926	C1_Q1	0,0857	6,0474	98,6265	9,9311
ARG	1572	Pos_rel	0,2748	T_Q1	0,1308	Chem4	0,1230	T_Q2	0,0879	C1_Q1	0,0628	C2_Q2	0,0569	14,9165	584,1808	24,1698
ASP	3293	Pos_rel	0,3150	T_Q2	0,1058	T_Q1	0,1038	C1_Q2	0,0840	C1_Q1	0,0773	C2_Q1	0,0760	4,1312	34,4588	5,8702
GLU	6404	Pos_rel	0,1288	Chem3	0,1064	T_Q1	0,1040	T_Q2	0,0999	C2_Q2	0,0884	Chem-2	0,0873	7,7240	127,5978	11,2959

Table S10. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 6 amino acids preceding and following AA_0 in the protein without their volumes.

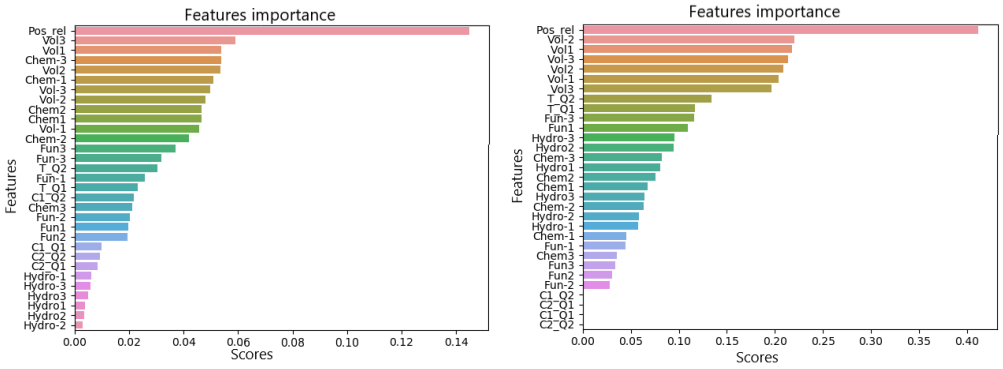
Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	T_Q1	0,1249	Pos_rel	0,1086	T_Q2	0,0923	Hydro-6	0,0844	Hydro3	0,0747	Chem3	0,0500	1,6256	5,0569	2,2487
ALA	7694	T_Q2	0,0637	Pos_rel	0,0534	T_Q1	0,0364	Hydro-4	0,0361	Hydro-5	0,0182	Chem-5	0,0167	1,9387	8,9031	2,9838
VAL	3572	Pos_rel	0,1220	T_Q1	0,1141	T_Q2	0,1071	Hydro-4	0,0398	Chem-1	0,0374	Hydro-6	0,0309	4,4088	41,7691	6,4629
LEU	8229	T_Q1	0,1168	Hydro5	0,1007	T_Q2	0,0932	Chem-6	0,0816	Chem5	0,0650	Hydro-2	0,0439	6,7841	93,2647	9,6574
ILE	4101	Pos_rel	0,0385	T_Q1	0,0262	Hydro5	0,0221	Chem-2	0,0194	Hydro-6	0,0123	T_Q2	0,0119	6,9205	94,5820	9,7253
MET	1105	Pos_rel	0,0405	Chem-1	0,0200	Chem-2	0,0177	Hydro-2	0,0164	Chem-3	0,0128	Hydro-1	0,0117	7,9650	138,5073	11,7689
PRO	1879	T_Q1	0,0390	Pos_rel	0,0068	T_Q2	0,0055	Hydro3	0,0051	Fun6	0,0022	Chem-2	0,0021	3,9733	48,7966	6,9855
SER	2724	Pos_rel	1,2113	T_Q3	1,1511	T_Q2	1,1143	Chem6	1,0559	Fun-6	1,0233	Chem3	1,0190	3,8249	17,6683	5,0827
THR	2575	T_Q1	0,1904	T_Q2	0,1206	Chem-1	0,0226	Hydro1	0,0214	Fun1	0,0202	Chem-5	0,0158	3,6381	30,3610	5,5101
CYS	12	T_Q2	0,4996	T_Q1	0,4818	Pos_rel	0,2295	Chem5	0,0575	Hydro-3	0,0419	Fun3	0,0196	2,6678	10,4382	3,2308
ASN	850	Pos_rel	0,6548	T_Q2	0,3224	T_Q1	0,0952	Chem2	0,0843	Chem-6	0,0335	Hydro6	0,0251	4,1670	32,4074	5,6927
GLN	2504	Pos_rel	0,1424	T_Q1	0,0579	Hydro1	0,0312	T_Q2	0,0264	Chem-2	0,0225	Chem5	0,0179	10,7260	213,1497	14,5996
PHE	2928	Chem-1	0,0894	T_Q1	0,0506	Fun-1	0,0180	Hydro6	0,0114	Pos_rel	0,0113	Fun4	0,0112	11,2789	223,3444	14,9447
TYR	1275	T_Q1	0,1453	T_Q2	0,1368	Pos_rel	0,0825	Chem-4	0,0151	Hydro4	0,0079	Chem5	0,0068	11,5654	286,7572	16,9339
TRP	947	Hydro2	0,1120	Fun2	0,0490	Chem2	0,0235	Hydro4	0,0166	Chem4	0,0137	Chem5	0,0128	15,5721	519,3814	22,7899
LYS	8821	T_Q2	0,2599	Pos_rel	0,1450	T_Q1	0,1299	Chem4	0,0751	Hydro4	0,0405	Chem-1	0,0295	16,0659	464,5117	21,5525
HIS	5258	T_Q2	0,3072	T_Q1	0,2028	Pos_rel	0,1126	Chem3	0,0635	Chem6	0,0373	Hydro2	0,0226	7,8663	128,2602	11,3252
ARG	1572	Pos_rel	0,2158	T_Q1	0,1498	T_Q2	0,1356	Chem4	0,1172	Chem1	0,0109	Hydro-1	0,0065	16,1652	601,9164	24,5340
ASP	3293	Pos_rel	0,1867	T_Q2	0,1818	Chem6	0,1342	T_Q1	0,1048	Chem-4	0,0674	Chem-6	0,0303	5,2004	48,9104	6,9936
GLU	6404	T_Q1	0,3068	Pos_rel	0,0952	T_Q2	0,0941	Hydro-5	0,0464	Chem-4	0,0399	Chem-2	0,0330	9,9959	175,3662	13,2426

Table S11. Features Importance RF: the features and their scores, that have most influenced lysine volume prediction, training RF on data set containing information about the 9 amino acids preceding and following AA_0 in the protein without their volumes.

Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	T_Q2	0,1342	Pos_rel	0,1243	T_Q1	0,1187	C2_Q1	0,1059	C1_Q2	0,0924	C1_Q1	0,0868	1,2317	3,9861	1,9965
ALA	7694	T_Q1	0,1518	T_Q2	0,1449	C1_Q1	0,1117	C2_Q2	0,1085	C1_Q2	0,1081	C2_Q1	0,1057	1,5688	7,0250	2,6505
VAL	3572	T_Q2	0,1418	T_Q1	0,1393	C1_Q1	0,1088	C1_Q2	0,0895	C2_Q1	0,0872	C2_Q2	0,0832	3,4376	31,9331	5,6509
LEU	8229	Pos_rel	0,1303	T_Q2	0,1270	T_Q1	0,1224	C1_Q1	0,0928	C1_Q2	0,0862	C2_Q2	0,0848	5,4283	77,6611	8,8126
ILE	4101	Pos_rel	0,1476	Fun2	0,1469	T_Q2	0,1221	T_Q1	0,1119	C1_Q2	0,0994	C2_Q1	0,0901	4,8464	64,8402	8,0523
MET	1105	T_Q1	0,1394	T_Q2	0,1075	Pos_rel	0,0975	C2_Q2	0,0907	C1_Q2	0,0848	C2_Q1	0,0824	6,7815	128,8413	11,3508
PRO	1879	T_Q1	0,1811	C2_Q1	0,1420	Pos_rel	0,1304	T_Q2	0,1225	C1_Q1	0,0931	C2_Q2	0,0846	3,0363	40,1024	6,3326
SER	2723	T_Q2	0,1983	Pos_rel	0,1660	T_Q1	0,1421	C2_Q2	0,1014	C1_Q2	0,0947	C2_Q1	0,0831	2,2340	13,5833	3,6855
THR	2575	T_Q1	0,2509	T_Q2	0,1206	Pos_rel	0,1080	C2_Q2	0,0893	C2_Q1	0,0879	C1_Q2	0,0852	2,9687	25,0911	5,0091
CYS	12	Fun-8	0,0659	Chem-1	0,0390	Fun9	0,0370	Pos_rel	0,0362	Chem-7	0,0319	Chem-4	0,0274	1,6636	4,8609	2,2047
ASN	850	T_Q1	0,1267	T_Q2	0,1155	C2_Q2	0,1120	C2_Q1	0,0922	C1_Q1	0,0890	Fun9	0,0858	3,1648	20,6769	4,5472
GLN	2504	T_Q1	0,1481	C1_Q1	0,1115	T_Q2	0,1071	C1_Q2	0,0917	C2_Q1	0,0867	C2_Q2	0,0860	9,4416	179,4646	13,3964
PHE	2928	Pos_rel	0,1603	T_Q2	0,1237	T_Q1	0,1130	C1_Q2	0,0866	C1_Q1	0,0851	C2_Q1	0,0814	8,8567	180,9724	13,4526
TYR	1275	Pos_rel	0,1397	C2_Q1	0,1279	T_Q1	0,1233	T_Q2	0,1170	C1_Q1	0,1087	C1_Q2	0,0800	11,2591	273,9231	16,5506
TRP	947	T_Q2	0,1862	T_Q1	0,1788	C1_Q2	0,0974	C2_Q2	0,0894	C2_Q1	0,0854	C1_Q1	0,0818	7,7724	197,0886	14,0388
LYS	8821	T_Q2	0,1428	T_Q1	0,1386	C2_Q2	0,1131	Pos_rel	0,1099	C1_Q2	0,1015	C2_Q1	0,1008	12,9321	364,3683	19,0884
HIS	5258	T_Q2	0,1385	T_Q1	0,1260	C1_Q2	0,1028	Pos_rel	0,0961	C2_Q2	0,0942	C1_Q1	0,0861	6,0178	96,9934	9,8485
ARG	1572	Pos_rel	0,2599	T_Q1	0,1304	Chem4	0,1207	T_Q2	0,0849	C1_Q1	0,0598	C2_Q2	0,0560	14,8545	582,3117	24,1311
ASP	3293	Pos_rel	0,2869	T_Q1	0,1025	T_Q2	0,1008	C1_Q1	0,0839	C2_Q1	0,0800	C1_Q2	0,0793	4,2591	35,5586	5,9631
GLU	6404	Pos_rel	0,1467	T_Q2	0,1184	T_Q1	0,0973	Chem3	0,0885	C2_Q1	0,0773	C1_Q1	0,0773	7,4725	107,1938	10,3534

Table S12. Features Importance MLP: the features and their scores, that have most influenced lysine volume prediction, training MLP on data set containing information about the 9 amino acids preceding and following AA_0 in the protein without their volumes.

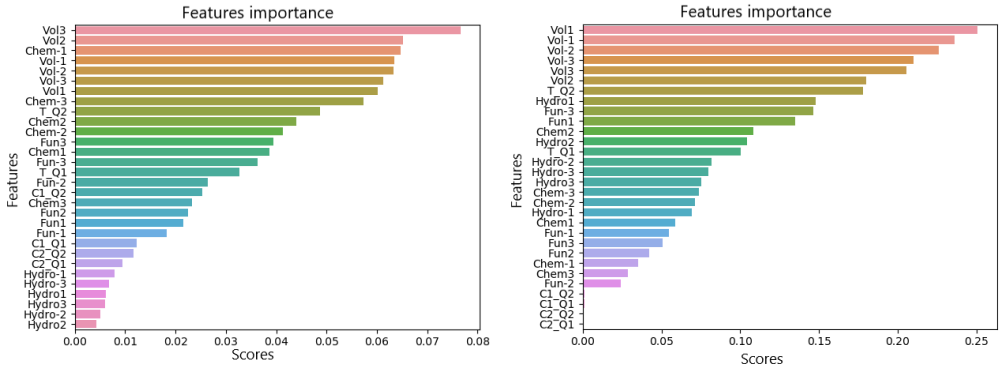
Amino	FREQ	F1	Val	F2	Val	F3	Val	F4	Val	F5	Val	F6	Val	MAE	MSE	RMS
GLY	5672	Pos_rel	0,1315	T_Q1	0,1271	T_Q2	0,1196	Hydro-6	0,0853	Chem4	0,0770	Hydro3	0,0692	1,6121	4,9393	2,2225
ALA	7694	T_Q2	0,0918	T_Q1	0,0826	Pos_rel	0,0532	Hydro-4	0,0403	Hydro-5	0,0315	Chem-5	0,0152	1,8956	8,2349	2,8697
VAL	3572	Chem8	0,1403	Hydro-8	0,0864	Chem-1	0,0454	Hydro-4	0,0360	T_Q1	0,0357	Chem-4	0,0257	4,1212	36,6524	6,0541
LEU	8229	T_Q2	0,1240	T_Q1	0,1189	Chem-6	0,0663	Hydro5	0,0639	Hydro-2	0,0391	Chem5	0,0372	6,7312	93,1337	9,6506
ILE	4101	T_Q1	0,0211	Pos_rel	0,0182	Hydro5	0,0157	Hydro9	0,0128	T_Q2	0,0105	Hydro-8	0,0103	6,8182	93,1517	9,6515
MET	1105	Chem9	0,0771	Hydro-2	0,0678	Chem-1	0,0470	Pos_rel	0,0342	Chem-3	0,0321	T_Q2	0,0298	7,6145	124,5784	11,1615
PRO	1879	T_Q1	0,0614	T_Q2	0,0240	Pos_rel	0,0237	Chem-4	0,0175	Hydro3	0,0119	Chem-7	0,0105	3,7529	48,0048	6,9286
SER	2724	Pos_rel	1,2545	T_Q3	1,1299	T_Q2	1,0989	Chem6	1,0587	Chem10	1,0378	Fun-6	1,0267	3,8237	17,9648	5,1188
THR	2575	T_Q1	0,1608	T_Q2	0,1483	Chem7	0,0364	Fun-5	0,0134	Hydro-4	0,0133	Hydro1	0,0132	3,7133	30,9326	5,5617
CYS	12	T_Q1	0,0461	Pos_rel	0,0084	Chem-9	0,0022	Chem3	0,0013	Fun-6	0,0012	Hydro-6	0,0011	2,0410	5,0489	2,2470
ASN	850	T_Q2	0,1876	Pos_rel	0,1855	Chem-9	0,0669	Hydro-3	0,0644	Hydro-9	0,0525	Chem2	0,0451	3,9476	25,3953	5,0394
GLN	2504	Pos_rel	0,1150	T_Q1	0,0434	Hydro-9	0,0407	Chem-9	0,0287	Fun-9	0,0286	Chem-6	0,0224	10,7624	211,8817	14,5562
PHE	2928	Chem-1	0,1004	Hydro7	0,0528	Chem-7	0,0439	T_Q1	0,0433	T_Q2	0,0312	Chem7	0,0219	12,1458	261,9531	16,1850
TYR	1275	T_Q1	0,0296	Pos_rel	0,0202	Fun4	0,0162	Hydro4	0,0140	Chem4	0,0061	Chem6	0,0044	14,0738	328,5074	18,1248
TRP	947	Hydro1	0,0123	Fun1	0,0079	Chem-2	0,0025	Chem-6	0,0007	Chem7	0,0004	Hydro8	0,0001	8,6417	204,3007	14,2934
LYS	8821	T_Q2	0,2676	T_Q1	0,1934	Pos_rel	0,1222	Hydro-9	0,0480	Chem-5	0,0471	Chem4	0,0396	15,6154	451,5171	21,2489
HIS	5258	T_Q2	0,3339	T_Q1	0,2945	Hydro8	0,0734	Chem6	0,0733	Chem3	0,0508	Pos_rel	0,0467	7,5599	122,2233	11,0555
ARG	1572	Pos_rel	0,1728	T_Q2	0,1509	T_Q1	0,1340	Chem4	0,1005	Chem1	0,0110	Hydro3	0,0099	15,9363	591,0136	24,3108
ASP	3293	T_Q2	0,1385	Pos_rel	0,0930	T_Q1	0,0785	Chem-4	0,0656	Chem6	0,0423	Chem-6	0,0224	5,1691	46,9522	6,8522
GLU	6404	T_Q1	0,2719	T_Q2	0,1062	Hydro-5	0,0477	Hydro-1	0,0395	Chem-1	0,0308	Pos_rel	0,0306	9,7335	160,3813	12,6642



(a) Analysis with RF

(b) Analysis with MLP

Fig. S4. Feature importance obtained for volume prediction of all amino acids, trained on data about the 3 amino acids preceding and following the AA_0 in the primary sequence of the protein.



(a) Analysis with RF

(b) Analysis with MLP

Fig. S5. Feature importance obtained for volume prediction of all amino acids, trained on data about the 3 amino acids preceding and following the AA_0 in the primary sequence of the protein, excluding its relative position in the chain.