Predicting lipid abundance in a murine brain section from spatial gene expression

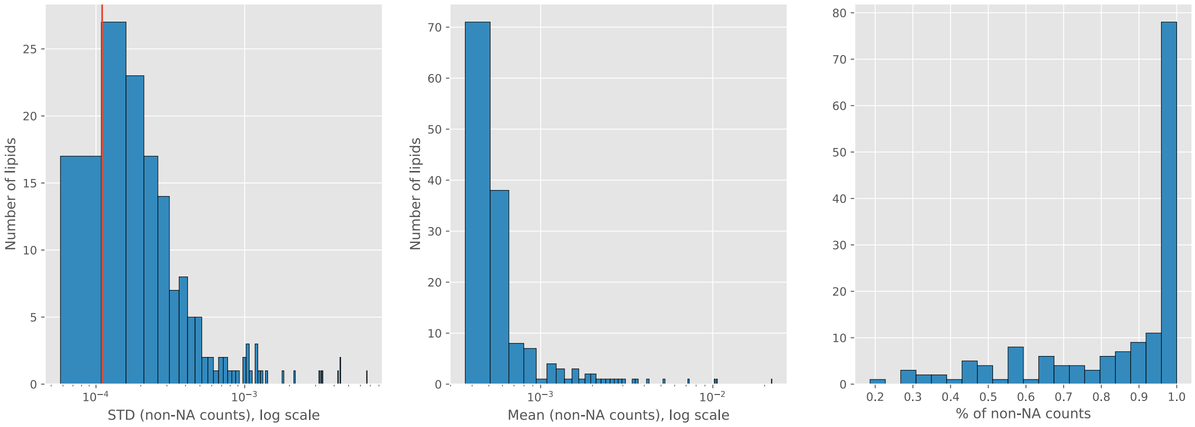
Lusine Khachatryan, Jules Perrin, Viola Renne

Tutor: Luca Fusar Bassini, Halima Hannah Schede
Supervisor: Gioele La Manno

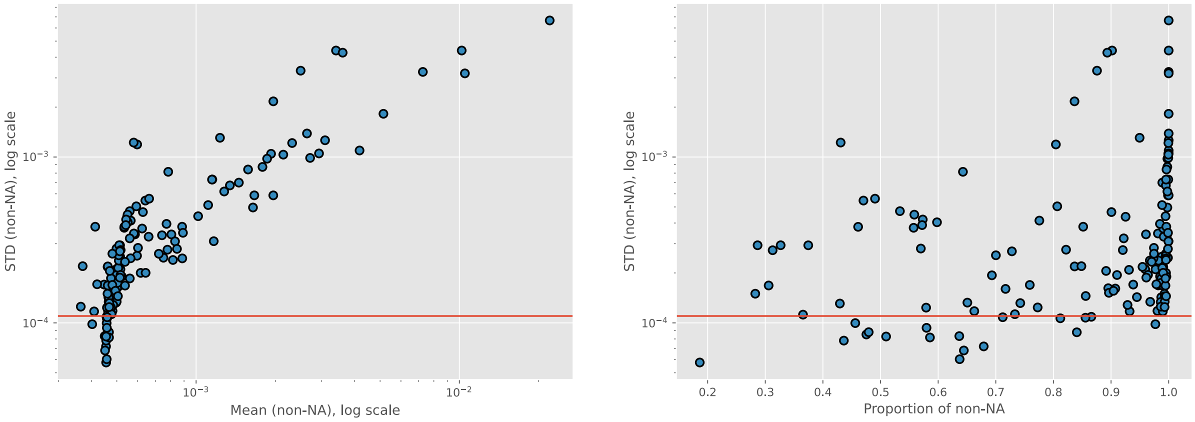
Laboratory of Brain Development and Biological Data Science, EPFL, Switzerland

Supplementary Materials

I. DATA EXPLORATION AND CLEANING



Supplementary Figure 1. Distribution of Standard deviations (STD), mean values and percentages of non-NA counts across all analyzed lipids. STD and mean values are calculated using non-NA values. Red vertical lire represents the STD cut-off (0.00011) selected for noisy lipids filtering.



Supplementary Figure 2. Scatter plots showing STD mean values and percentages of non-NA counts across all analyzed lipids. For STD and mean values log scale is used. Red horisontal lire represents the STD cut-off (0.00011) selected for "noisy" lipids filtering.

II. INITIAL FITTING USING PYCARET

List of PyCaret Algorithms tested during the initial fit:

- lr - Linear Regression
- lasso - Lasso Regression
- ridge - Ridge Regression
- en - Elastic Net
- lar - Least Angle Regression
- llar - Lasso Least Angle Regression
- omp - Orthogonal Matching Pursuit
- br - Bayesian Ridge
- ard - Automatic Relevance Determination
- par - Passive Aggressive Regressor
- ransac - Random Sample Consensus
- tr - TheilSen Regressor
- huber - Huber Regressor
- kr - Kernel Ridge
- svm - Support Vector Regression
- knn - K Neighbors Regressor
- dt - Decision Tree Regressor
- rf - Random Forest Regressor
- et - Extra Trees Regressor
- ada - AdaBoost Regressor
- gbr - Gradient Boosting Regressor
- mlp - MLP Regressor
- xgboost - Extreme Gradient Boosting
- lightgbm - Light Gradient Boosting Machine
- catboost - CatBoost Regressor

Supplementary Table 1. Mean R2 and R2 STD across 5 folds for 3 Pycaret models. Data used for training: n=1000, aggregation algorithm Negative Logarith decay

Lipid	CatBoost R2	CatBoost R2 std	KNN R2	KNN R2 std	XgBoost R2	XgBoost R2 std
LPC O- 18:3	0.6158	0.0015	0.6217	0.0022	0.5793	0.0405
LPC 15:1	0.8325	0.0017	0.8354	0.0021	0.8209	0.0254
LPC 20:4	0.7008	0.0034	0.6945	0.0019	0.6695	0.0476
LPC 22:6	0.6908	0.0035	0.6702	0.0020	0.6369	0.0292
Cer 40:2	0.5651	0.0031	0.5374	0.0050	0.539	0.0373
Cer 42:2	0.2915	0.004	0.2224	0.0029	0.2541	0.0118
HexCer(d32:2)	0.2803	0.0012	0.1704	0.0040	0.2648	0.0613
PE(32:1)	0.7099	0.0072	0.7065	0.0037	0.6728	0.0352
PA(34:1)	0.6771	0.0015	0.651	0.0014	0.6299	0.0592
PG(30:1)	0.2244	0.0037	0.131	0.0041	0.1766	0.0263
PA 36:4	0.6903	0.0025	0.6892	0.0047	0.6409	0.0212
PA(36:3)	0.5679	0.0076	0.5344	0.0058	0.5315	0.0493

PA 36:2	0.4141	0.004	0.345	0.0027	0.3725	0.0188
PE O-34:2	0.5083	0.0028	0.4526	0.0033	0.4517	0.0544
PA 36:1	0.6288	0.0023	0.6107	0.0046	0.5935	0.0153
SM 34:1	0.6431	0.0012	0.6229	0.0026	0.6011	0.0643
PE(O-34:1)	0.5703	0.0138	0.5328	0.0041	0.5335	0.0437
PC(30:0)	0.3081	0.0116	0.2231	0.0054	0.2729	0.0489
PG 32:0	0.6617	0.0039	0.6449	0.0072	0.6498	0.0434
PC 33:1	0.6307	0.0064	0.598	0.0031	0.6044	0.0532
PC 33:0	0.4359	0.0061	0.3794	0.0167	0.3675	0.0247
SM(d36:1)	0.7232	0.0031	0.7229	0.0045	0.6953	0.0428
PC(32:1)	0.6762	0.0018	0.6521	0.0024	0.6174	0.0206
PE-Cer(d38:1)	0.417	0.0026	0.3606	0.0049	0.3549	0.0782
PE 34:1	0.4034	0.0094	0.3461	0.0040	0.3351	0.0299
PC 34:2	0.3975	0.0029	0.3444	0.0032	0.3456	0.0750
PA 38:3	0.4698	0.0053	0.4351	0.0032	0.3963	0.0182
PI-Cer(t32:2)	0.7131	0.0028	0.7103	0.0061	0.6737	0.0284
PE(36:2)	0.5046	0.0081	0.4431	0.0046	0.4156	0.0286
PS(O-34:0(OH))	0.7206	0.0065	0.7149	0.0034	0.6882	0.0257
SM(d36:2)	0.6158	0.0012	0.578	0.0042	0.5827	0.0137
PA 39:1	0.6271	0.0011	0.6039	0.0047	0.5571	0.0890
PE O-36:2	0.384	0.0045	0.3524	0.0087	0.3226	0.0327
HexCer 36:0	0.3839	0.0015	0.3211	0.0019	0.3346	0.0516
PA(38:1)	0.599	0.002	0.5869	0.0086	0.5878	0.0668
PE 36:0	0.3269	0.0029	0.2642	0.0026	0.2314	0.0550
LBPA(34:1)	0.7188	0.0018	0.7001	0.0046	0.6687	0.0123
PG(34:0)	0.2466	0.0053	0.1529	0.0038	0.1709	0.0427
PA 40:4	0.2729	0.0014	0.206	0.0034	0.2306	0.0475
PC 35:0	0.7692	0.0018	0.7653	0.0031	0.7385	0.0283
PE 36:2	0.8281	0.0008	0.828	0.0018	0.8192	0.0360
HexCer 36:1	0.612	0.002	0.5952	0.0040	0.5714	0.0208
HexCer 40:2	0.7246	0.0024	0.711	0.0139	0.6848	0.0817
PE O-39:7	0.403	0.0099	0.3602	0.0051	0.3235	0.0420
CerP(t42:1)	0.256	0.0088	0.1529	0.0035	0.213	0.0424
PE(36:0)	0.5227	0.0027	0.4762	0.0056	0.435	0.0228
PC 36:2	0.3597	0.0026	0.3194	0.0035	0.3119	0.0347
PA(40:6(OH))	0.3667	0.0016	0.2874	0.0059	0.3251	0.0500
PA(40:5)	0.3344	0.0028	0.2689	0.0043	0.2845	0.0119
PE 37:2	0.5902	0.0016	0.5611	0.0044	0.5441	0.0516
PE 40:4	0.7934	0.0018	0.7961	0.0025	0.7844	0.0750
SM 38:1	0.1159	0.003	-0.0206	0.0029	0.1028	0.0325
PC 34.1	0.6495	0.0017	0.6204	0.0039	0.6173	0.0343
PG(36:1)	0.4845	0.001	0.4432	0.0112	0.4223	0.0788
PE O-40:6	0.4617	0.0022	0.4084	0.0016	0.3927	0.0131

PC 37:1	0.6863	0.0049	0.6721	0.0053	0.6526	0.0810
PG(34:1(OH))	0.4339	0.0007	0.3795	0.0064	0.3767	0.0332
PC 36.2	0.4365	0.001	0.3732	0.0025	0.3744	0.0266
HexCer 40:0	0.6712	0.0018	0.6581	0.0016	0.6621	0.1565
PC 35:2	0.3064	0.0012	0.2265	0.0049	0.2709	0.0342
HexCer 38:1	0.8244	0.0011	0.8237	0.0017	0.8181	0.0292
PA(42:8)	0.3787	0.002	0.3148	0.0049	0.2943	0.0645
PE(40:7)	0.3535	0.0038	0.2698	0.0073	0.3055	0.0159
PA 42:7	0.7586	0.0032	0.7436	0.0051	0.7195	0.0349
PE(P-40:6)	0.085	0.0057	0.053	0.0040	0.0797	0.0405
PE O-41:11	0.7775	0.0014	0.774	0.0065	0.7559	0.0297
PS(38:0)	0.7263	0.0052	0.7181	0.0052	0.6876	0.0329
PG(38:4)	0.6501	0.0078	0.6124	0.0065	0.6206	0.0373
PC O-39:9	0.3664	0.0018	0.2818	0.0074	0.3388	0.0385
PC 36:3	0.7197	0.0094	0.714	0.0022	0.6815	0.0385
HexCer 40:1	0.6407	0.001	0.6503	0.0085	0.5942	0.0433
PG(38:3)	0.4822	0.0015	0.424	0.0021	0.4373	0.0454
PE(40:9)	0.7473	0.0038	0.7416	0.0049	0.7127	0.0414
PE O-42:8	0.5474	0.0091	0.4934	0.0019	0.5144	0.0456
PG(38:2)	0.7904	0.0034	0.7879	0.0060	0.7739	0.0370
SM 40:1	0.6941	0.0031	0.6633	0.0051	0.6479	0.0614
PG(38:1)	0.6191	0.0039	0.6222	0.0047	0.5487	0.0447
PE 40:7	0.5306	0.007	0.4881	0.0037	0.4833	0.0137
PS 36:1	0.7902	0.0102	0.7888	0.0068	0.7603	0.0254
PC 38:5	0.8334	0.0032	0.7907	0.0041	0.7818	0.0866
PE(40:4)	0.8188	0.0035	0.8192	0.0031	0.8038	0.0624
Hex2Cer 32:1	0.4309	0.0009	0.4097	0.0043	0.3982	0.0339
HexCer 40:2;O3	0.3627	0.0015	0.2979	0.0034	0.3229	0.0380
PI-Cer(d38:0)	0.5689	0.0074	0.5507	0.0086	0.5154	0.0385
PC(P-40:6)	0.549	0.0073	0.5215	0.0079	0.5339	0.0501
HexCer(t40:0)	0.6897	0.0021	0.6754	0.0031	0.6427	0.0595
SM(t40:1)	0.4934	0.0079	0.4473	0.0048	0.4782	0.0577
PG(40:6)	0.1156	0.0011	0.0029	0.0052	0.1075	0.0958
PC(38:5)	0.759	0.008	0.7628	0.0024	0.7354	0.0305
PS(40:1)	0.7931	0.0028	0.7917	0.0042	0.7846	0.0626
PE-Cer(d46:3)	0.8094	0.0022	0.8072	0.0031	0.804	0.0232
HexCer(t42:2)	0.3545	0.0029	0.2914	0.0141	0.2806	0.0429
PG(40:4)	0.7643	0.002	0.7435	0.0037	0.7066	0.0499
PS 38:4	0.7196	0.0014	0.6943	0.0070	0.6831	0.0334
PC 38:3	0.6087	0.0017	0.5693	0.0061	0.5932	0.0274
HexCer 42:1	0.7546	0.0049	0.7574	0.0024	0.7456	0.0896
PG(40:3)	0.7407	0.0093	0.7344	0.0027	0.6768	0.0541
PC 38:2	0.7788	0.0017	0.7833	0.0038	0.7713	0.0384

PE(40:2(OH))	0.7798	0.0089	0.7802	0.0108	0.7449	0.0763
PC 38:1	0.7557	0.0129	0.7495	0.0031	0.696	0.0593
PG(40:1)	0.4643	0.0052	0.4276	0.0028	0.3835	0.0487
HexCer 42:2;O3	0.7971	0.0022	0.8011	0.0042	0.7872	0.0128
HexCer 42:1;O3	0.3952	0.001	0.3639	0.0020	0.349	0.1029
HexCer(t42:0)	0.2859	0.0111	0.2185	0.0038	0.2533	0.0561
SM(t42:1)	0.8059	0.0044	0.7943	0.0032	0.7754	0.0303
PC(40:7)	0.6214	0.0013	0.5976	0.0117	0.5682	0.0396
PG(42:6)	0.629	0.0057	0.6106	0.0055	0.6034	0.0538
Hex2Cer 32:0	0.7602	0.0022	0.7624	0.0045	0.7494	0.0289
SHexCer 38:1;3	0.5836	0.0019	0.5608	0.0044	0.5456	0.0473
PC(40:4)	0.6669	0.0021	0.6525	0.0042	0.6567	0.0463
PI-Cer(t32:1)	0.2247	0.0037	0.1259	0.0052	0.1598	0.0270
PC 34:0	0.5879	0.0075	0.5478	0.0036	0.5288	0.0201
PG(36:1(OH))	0.7944	0.0036	0.7924	0.0019	0.7805	0.0756
PC 36:1	0.6163	0.0012	0.5879	0.0087	0.5615	0.0467
SM 42:2	0.5397	0.0015	0.5312	0.0024	0.475	0.0190
PE 40:6	0.7758	0.0017	0.781	0.0029	0.7478	0.0566
PC 38:6	0.3623	0.0088	0.2745	0.0040	0.3278	0.0733
PE 38:6	0.6703	0.0027	0.6356	0.0021	0.6203	0.0441
PC 40:6	0.5978	0.0071	0.5868	0.0050	0.5869	0.0485
PE 38:4	0.1835	0.0021	0.098	0.0021	0.1464	0.0535
PC 34:1	0.4228	0.008	0.3501	0.0013	0.3967	0.0371
HexCer 42:2	0.5292	0.0067	0.4908	0.0037	0.4491	0.0439
PC 35:1	0.5377	0.0047	0.5041	0.0015	0.5086	0.0241
PE O-38:5	0.7903	0.0086	0.7903	0.0020	0.7675	0.0559
PI-Cer(t30:1)	0.349	0.0053	0.2669	0.0014	0.2662	0.0283
PC 32:0	0.4736	0.006	0.4258	0.0051	0.4097	0.0507
PC 32:1	0.4264	0.0021	0.39	0.0028	0.3472	0.0126
SM 36:1	0.2855	0.0088	0.1997	0.0079	0.2385	0.0472
PC 31:0	0.6734	0.0036	0.6558	0.0037	0.6025	0.0182
PC 38:4	0.4335	0.0027	0.3775	0.0022	0.3778	0.0201
PE 36:4	0.1392	0.0089	0.0436	0.0034	0.0732	0.0385
PI-Cer(t28:0)	0.5934	0.0067	0.5647	0.0019	0.5701	0.0625
LPC 18:1	0.8136	0.0034	0.8112	0.0045	0.7925	0.0672
LPC O-16:2	0.7231	0.0082	0.702	0.0089	0.6782	0.0279
LPC 16:0	0.2758	0.0013	0.1977	0.0026	0.1993	0.0448
LPC O-18:2	0.7379	0.0024	0.7343	0.0021	0.6992	0.0329
LPC 18:0	0.6821	0.0107	0.6705	0.0032	0.6508	0.0174
PC 36:4	0.6909	0.006	0.6889	0.0017	0.6397	0.0806

III. MLP

Supplementary Table 2. Loss (MSE) and R2 score across 5 folds with default values for activation functions, kernel initializer (GlorotUniform) and Adam optimizer with 0.001 as learning rate. Sigmoid is used as output layer activation function.

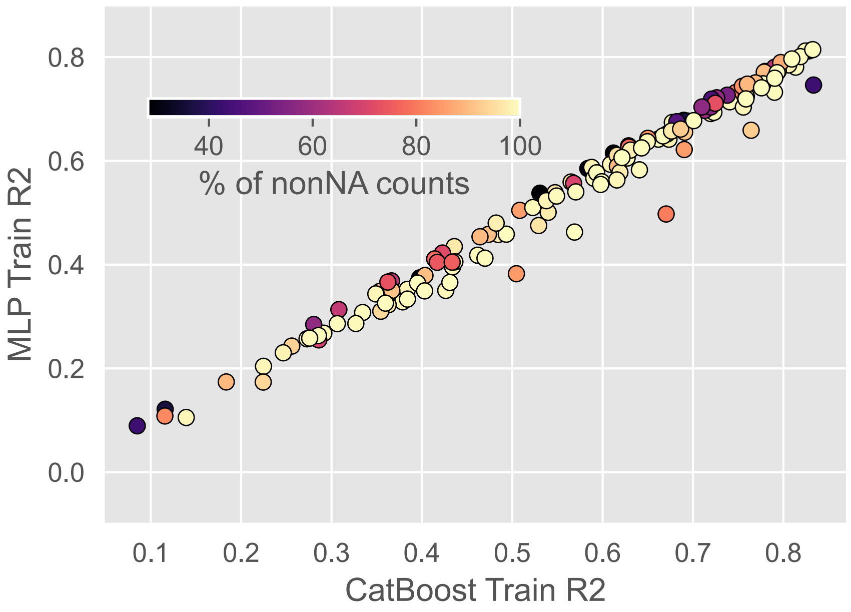
Number of hidden layers	Number of parameters	Loss (MSE)	R2
2	72,810	0.005381 ± 9.607e-5	0.4042 ± 3.465e-3
2	**38,698**	**0.005356 ± 1.196e-4**	**0.4070 ± 6.004e-3**
2	18,906	0.005394 ± 1.007e-4	0.4017 ± 2.081e-3
3	**79,018**	**0.005375 ± 0.6074e-5**	**0.4053 ± 3.831e-3**
3	37,018	0.005404 ± 9.660e-5	0.4008 ± 2.651e-3
3	17,938	0.005510 ± 1.099e-4	0.3863 ± 3.022e-3
4	77,338	0.005409 ± 1.086e-4	0.4004 ± 3.937e-3
4	36,050	0.005516 ± 1.128e-4	0.3858 ± 2.844e-3
4	17,422	0.005773 ± 1.232e-4	0.3541 ± 4.573e-3

Supplementary Table 3. Loss (MSE) and R2 score across 5 folds with different activation functions and default values for kernel initializer (GlorotUniform) and Adam optimizer with 0.001 as learning rate. Sigmoid is used as output layer activation function.

Number of hidden layers	Number of parameters	Activation (hidden layers)	Loss	R2
2	38,698	Sigmoid	0.004058 ± 1.069e-4	0.5359 ± 5.075e-3
2	38,698	Tanh	0.004154 ± 1.043e-4	0.5243 ± 4.886e-3
2	38,698	ReLU	0.004318 ± 1.371e-4	0.5054 ± 9.103e-3
2	38,698	GELU	0.004151 ± 1.4345e-4	0.5240 ± 9.560e-3
3	**79,018**	**Sigmoid**	**0.003991 ± 7.274e-5**	**0.5434 ± 7.768e-3**
3	79,018	Tanh	0.004144 ± 1.237e-4	0.5257 ± 5.693e-3
3	79,018	ReLU	0.004155 ± 1.566e-4	0.5251 ± 1.067e-2
3	**79,018**	**GELU**	**0.004045 ± 9.400e-5**	**0.5372 ± 5.124e-3**

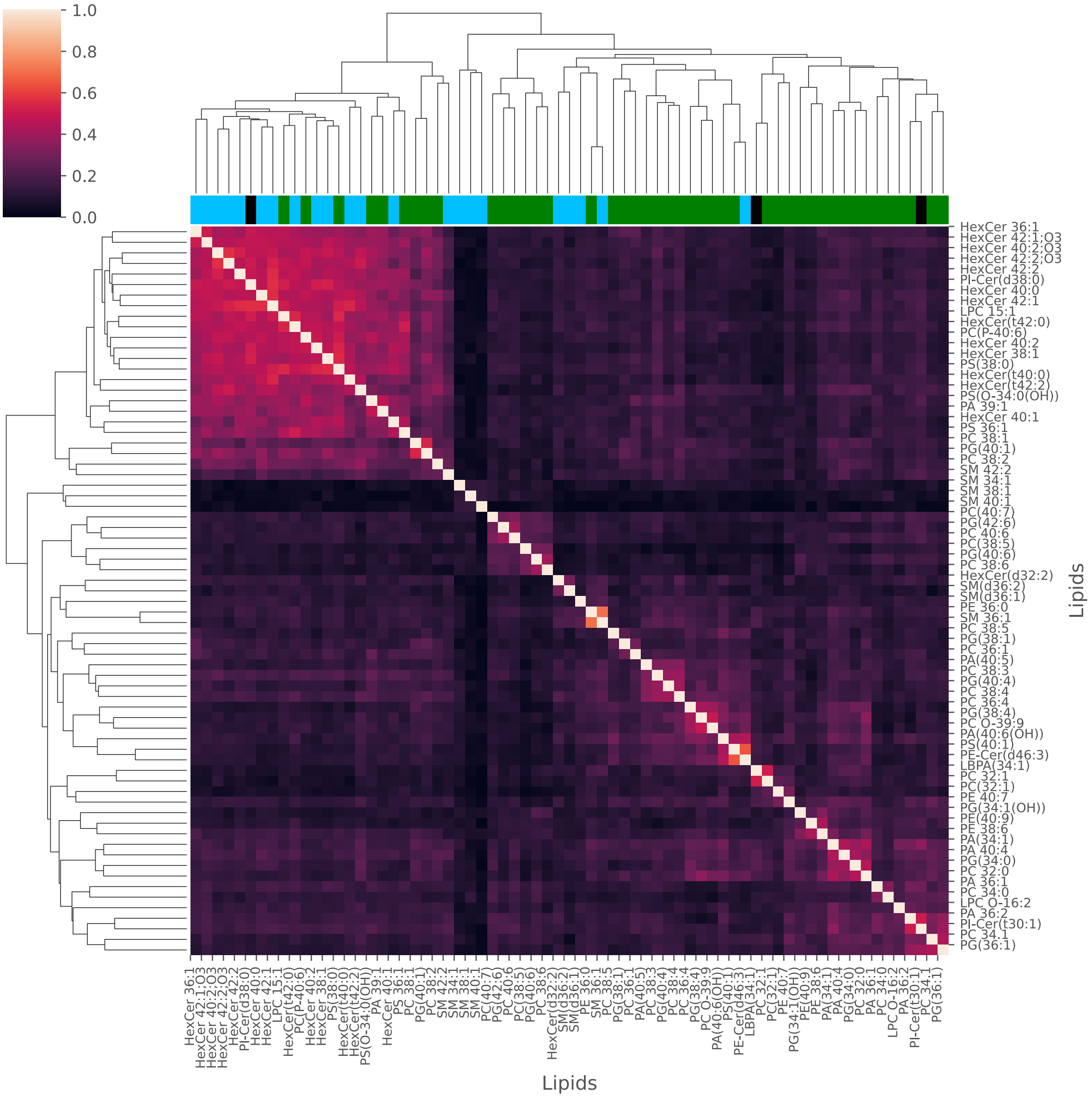
Supplementary Table 4. Loss (MSE) and R2 score across 5 folds. Adam optimizer with 0.001 as learning rate. Sigmoid is used as output layer activation function.

Number of hidden layers	Number of parameters	Activation (hidden layers)	Kernel Initializer	Loss	R2
3	**79,018**	**Sigmoid**	**GlorotUniform**	**0.003991 ± 7.274e-5**	**0.5434 ± 7.768e-3**
3	**79,018**	**Sigmoid**	**HeNormal**	**0.004015 ± 7.305e-5**	**0.5411 ± 5.040e-3**
3	79,018	GELU	GlorotUniform	0.004045 ± 9.400e-5	0.5372 ± 5.124e-3
3	79,018	GELU	HeNormal	0.004046 ± 9.765e-5	0.5373 ± 4.564e-3

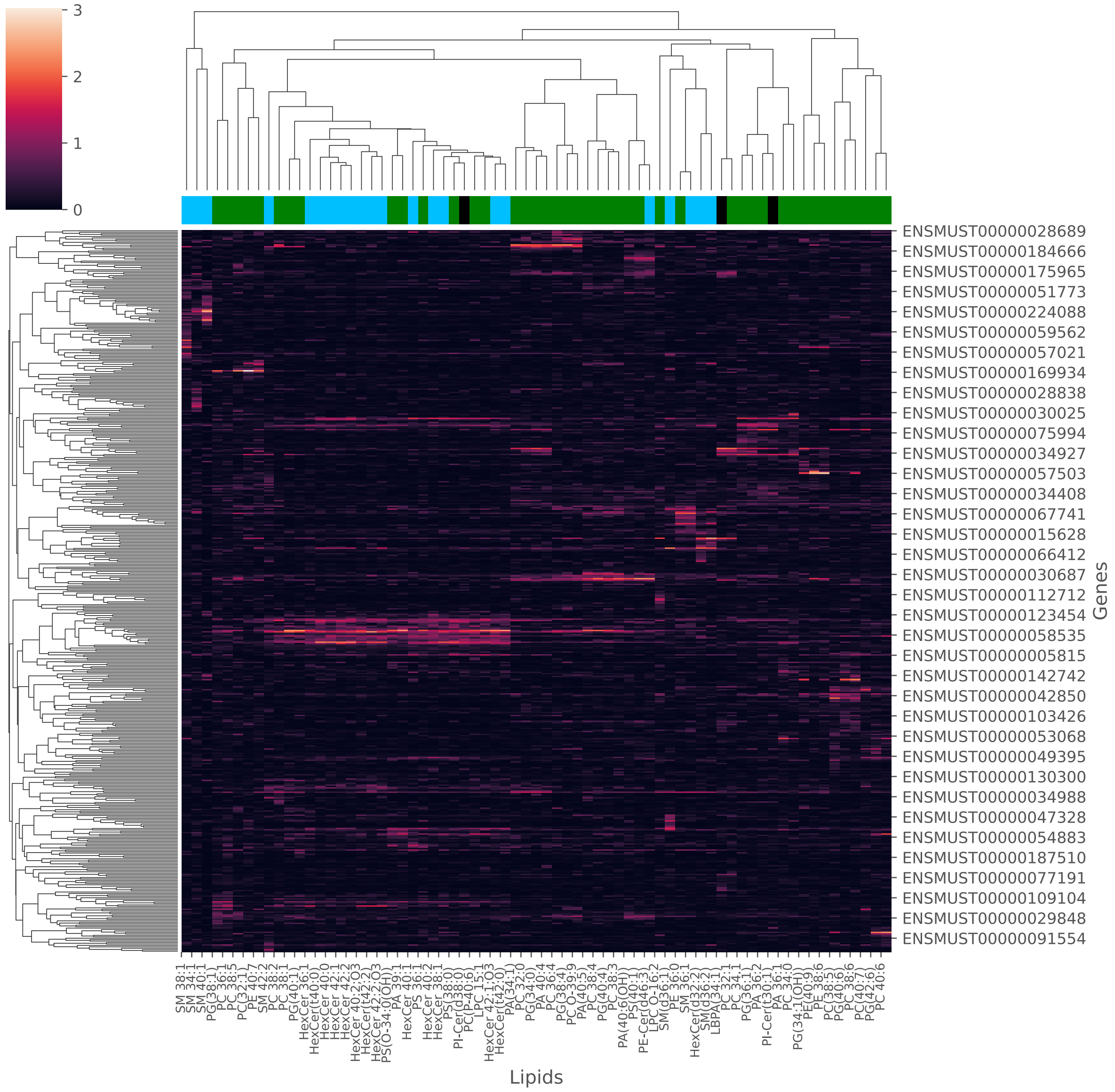


Supplementary Figure 3. Train R^2 performance of MLP (y axis) and CatBoost (x axis) colored by the percentage of non-NA counts observed for that lipid.

IV. CATBOOST MODELS INTERPRETATION



Supplementary Figure 4. Pairwise Jaccard index for lipids calculated using the list of most important genes (50% cumulative importance) reported by the CatBoost model. Lipids color-coding is based on the lipid category (Shingolipids – blue, Glycerophospholipids – green, other – black). Only results for best lipids with high CatBoost prediction quality ($R^2 > 0.6$, 69 lipids in total) are shown.



Supplementary Figure 5. Correlation of genes importances obtained from CatBoost models. Clustermap represents lipids (columns) and genes (rows). Lipids color-coding is based on the lipid category (Shingolipids – blue, Glycerophospholipids – green, other – black). Each cell represents the importance (from 0 to 100) of the specific gene for the Catoost model of the specific lipid. Only results for best lipids with high CatBoost prediction quality ($R^2 > 0.6$, 69 lipids in total) are shown.