Supplementary Information

Rational Design of a Transferrin-Binding Peptide Sequence Tailored to Targeted Nanoparticle Internalization

Melissa Santi, [a,1,2]* Giuseppe Maccari,[a, 1] Paolo Mereghetti,[1] Valerio Voliani,[1] Silvia Rocchiccioli,[3] Nadia Ucciferri,[3] Stefano Luin,[2] Giovanni Signore[1]*

[1]Center for Nanotechnology Innovation@NEST, Istituto Italiano di Tecnologia, Piazza San Silvestro 12, Pisa 56127, Italy. [2] NEST, Scuola Normale Superiore and Istituto Nanoscienze-CNR, Piazza San Silvestro 12, Pisa 56127, Italy. [3]Institute of Clinical Physiology-CNR, Via Giuseppe Moruzzi 1, Pisa 56124, Italy.

[a] Equal contribution

Corresponding author: Giovanni Signore email: giovanni.signore@iit.it

Melissa Santi email: Melissa.santi@sns.it

LIST OF CONTENTS

SUPPLEMENTARY INFORMATIONS

Peptide encoding and QSAR

In order to present the simulated peptides to the learning algorithm, two types of descriptor encoding were employed (see **Table S1**): global descriptors and topological descriptors. Global descriptors describe the whole molecule, while topological descriptors represent the interaction of different residues along the amino acidic sequence. Charge and hydrophobicity related characteristics are among the most important properties in peptide binding.[1] Hydrophobicity determines folding, binding to receptors, and interactions of proteins and peptides with biological membranes. Therefore, both charge and hydrophobicity will be taken into account in global and topological descriptors. Topological description of the peptide sequence was accounted for by encoding QSAR descriptors into auto- and cross covariance (ACC) values. Classical ACC transformation was first introduced by Wold *et al* [2], resulting in two kinds of variables: auto covariance (AC) of the same descriptor and cross covariance (CC) between two different descriptors. Briefly, for a given protein sequence, ACC variables describe the average interactions between residues distributed a certain *lag* apart throughout the whole sequence. Besides describing the sequence order, ACC has the ability to transform each amino acid (AA) sequence of variable length into uniform equal-length vectors. This feature is very important in data mining methods, where a fixed-length vector describing each instance is required. However, averaging along the entire sequence may cause loss of information about strong and weak correlations. To cope with these limitations, the *Minimum and maximum of auto- and cross-covariances* (mMACC) algorithm was introduced[3], where positive and negative descriptor values are considered separately and only the minimum and maximum value of each lag is used. This allows accounting for both weak and strong correlations along the peptide sequence.

$$AC_{\min d} = MIN\left[Z_i^k * Z_i^{k+d}\right]; AC_{\max d} = MAX\left[Z_i^k * Z_i^{k+d}\right](k = 1,2,3..L-d)$$
$$CC_{\min d} = MIN\left[Z_i^k * Z_i^{k+d}\right]; CC_{\max d} = MAX\left[Z_i^k * Z_i^{k+d}\right](k = 1,2,3..L-d)$$

Equation 1. Minimum and Maximum of auto- and cross-covariances

Where Z^k_i is the *i*-th descriptor of residue *k* in the sequence, *d* is the lag, *L* is the length of the amino acid sequence. As in the MACC algorithm, the maximum value of each interaction is taken into account. However, in the mMACC each z-scale descriptor is shifted by the absolute minimal value in order to have only positive interactions. This reduces the number of combinations, while maintaining both information of strong and weak interactions. In this work, 'z-scale' variables [4] are encoded into topological descriptors. These are highly condensed variables derived from a principal component analysis (PCA) of several experimental or theoretical physicochemical properties for the 87 amino acids. In detail, these z-scale descriptors correspond to the first five principal components explaining the variance in the set: z_1, z_2, and z_3

represent the AA hydrophobicity, steric properties, and polarity, respectively. Finally, z_4 and z_5 describe the electronic effect of the residues [5].

Table S1: List of descriptors. Two classes of descriptors were used to describe a single hexapeptide sequence: global descriptors and topological descriptors. Column "Number" indicates the number of components for a given descriptor.

Type	**Abbreviation**	**Description**	**Number**
Global	NetCharge@5	Net charge at pH = 5	1
	NetCharge@7	Net charge at pH = 7	1
	NetCharge@9	Net charge at pH = 9	1
	IP	Isoelectric Point	1
	MW	Molecular weight	1
Topological	AC	Min and max auto covariance values between same descriptors	25
	CC	Min and max cross covariance between two descriptors	100

Figure S1: QCM-D measurements of Δf over time. Shift in frequency were detected at the binding of peptides (blue arrows) or the binding of increasing concentration (1, 10, 20, 50 μM) of hTransferrin (red arrows,upper panels) or BSA (red arrows, lower panels).

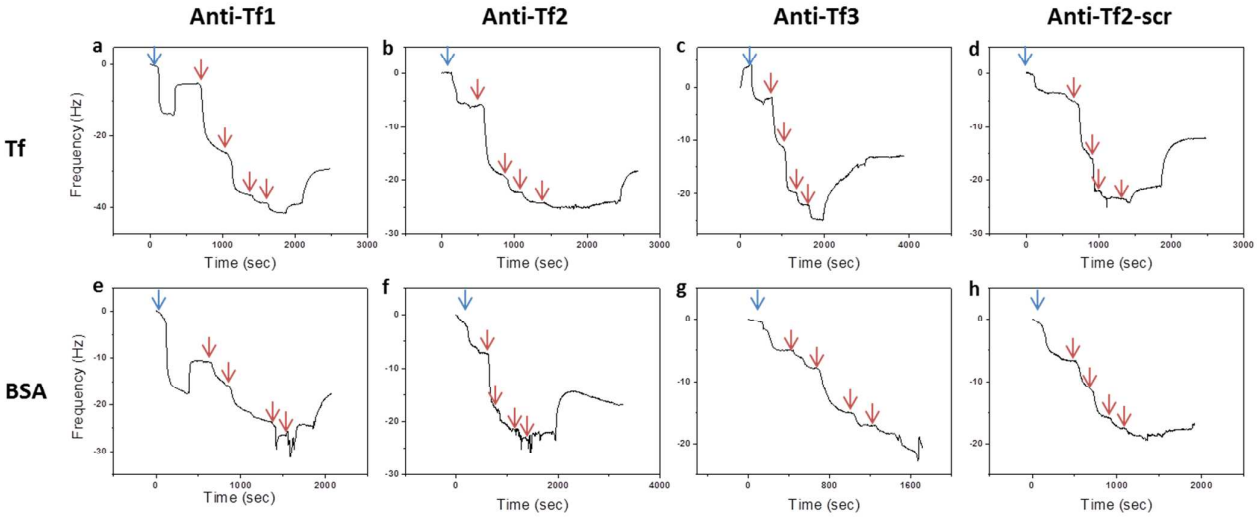

Figure S2: Internalization assay in adenocarcinoma pancreatic cells (Mia PaCa-2). Each peptide labelled with Atto633 (yellow channel) was incubated in cultured cells at concentration of 1µM for 30 minutes at 37°C and 5% CO_2 and then internalization was monitored by confocal fluorescence microscopy. (a) Evaluation of aspecific internalization of Tf2 using medium without serum and transferrin. (b) Evaluation of Tf2-scr internalization using medium with serum and adding unlabeled transferrin to a final concentration of 35µM. Scale bars 10µM.

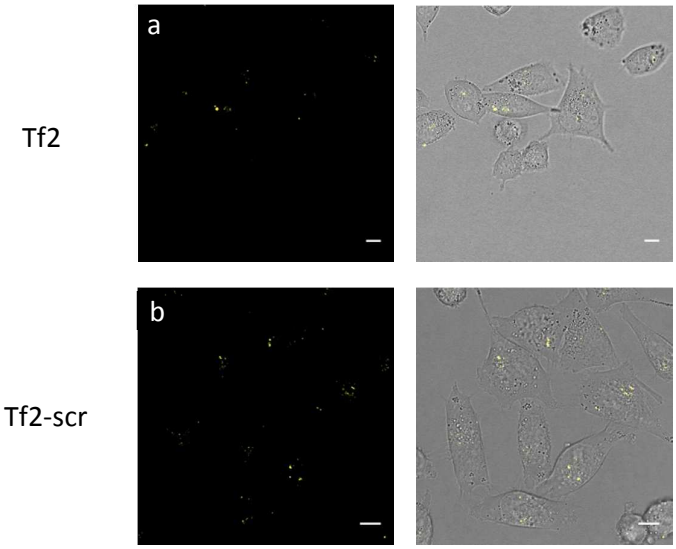

Figure S3: Competitive inhibition for binding of Tf2 to Tf. Tf2 (1 µM) internalizes in Mia PaCa-2 cells in the presence of Tf (35 µM, labeled/unlabeled 1/7), with extensive colocalization at membrane and lysosomal level (Fig. S3a). Internalization of labeled Tf2 (1 µM) is completely inhibited when saturating amounts of unlabeled Tf2 (1 mM) are added in the presence of Tf (35 µM, labeled/unlabeled 1/7) (Fig. S3b). Internalization of labeled Tf2 (1 µM) and of labeled Tf (5 µM) is almost completely inhibited by the addition of 600 µM of unlabeled Tf (Fig. S3c).

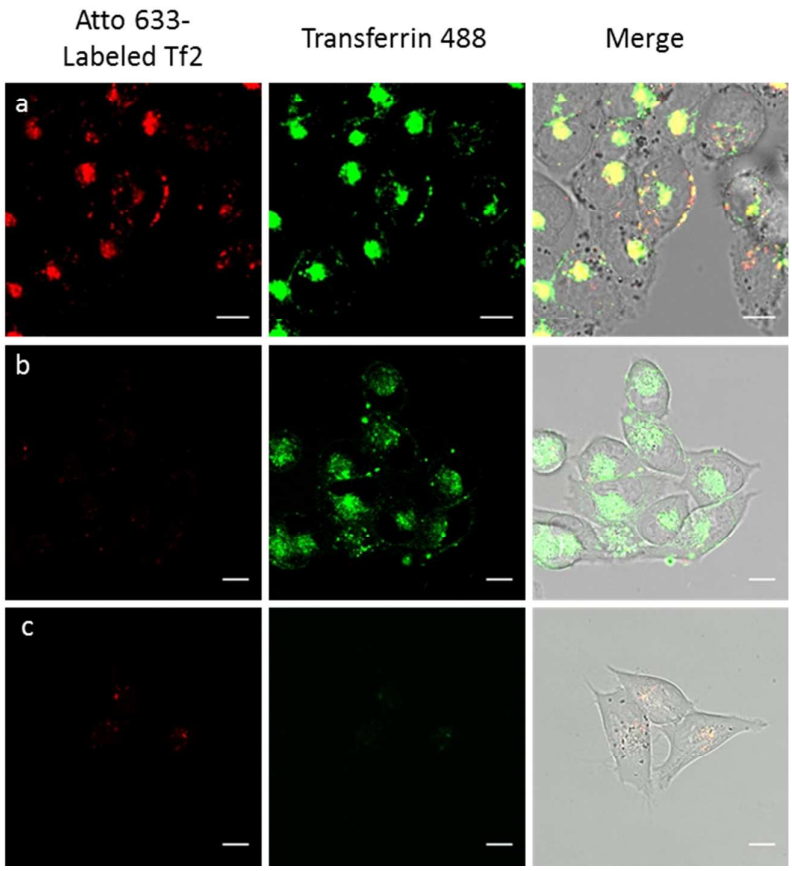

Table S2: Total identified proteins

The table contains all identified protein with at least 2 identified peptides (at 95% of confidence). %Cov: Coverage percentage at 95% confidence level.

Paragon Score	%Cov (95)	Accession	Entry name	Protein Name	Peptides (95%)
121.26	81.12	P02768	ALBU_HUMAN	Serum albumin	174
68.73	33.72	P02671	FIBA_HUMAN	Fibrinogen alpha chain	72
59.44	16.21	P21333	FLNA_HUMAN	Filamin-A	35
40.54	31.83	P04264	K2C1_HUMAN	Keratin, type II cytoskeletal 1	43
39.6	51.93	P02675	FIBB_HUMAN	Fibrinogen beta chain	34
35	9.745	P35579	MYH9_HUMAN	Myosin-9	23
32.69	25.64	P02787	TRFE_HUMAN	Serotransferrin	15
29.85	56.78	P02649	APOE_HUMAN	Apolipoprotein E	24
27.38	47.19	P02647	APOA1_HUMAN	Apolipoprotein A-I	24
26.42	34.73	P01871	IGHM_HUMAN	Ig mu chain C region	28
26.27	26.54	P13645	K1C10_HUMAN	Keratin, type I cytoskeletal 10	25
25.72	44.15	P02679	FIBG_HUMAN	Fibrinogen gamma chain	19
24.57	8.049	P12259	FA5_HUMAN	Coagulation factor V	21
24.04	50	P01857	IGHG1_HUMAN	Ig gamma-1 chain C region	20
22.77	15.53	P01042	KNG1_HUMAN	Kininogen-1	20
20.94	58.47	P67936	TPM4_HUMAN	Tropomyosin alpha-4 chain	18
20.29	26.28	P10909	CLUS_HUMAN	Clusterin	12
19.98	38.42	P00738	HPT_HUMAN	Haptoglobin	16
18.5	24.27	P04004	VTNC_HUMAN	Vitronectin	19
17.11	21.67	P35527	K1C9_HUMAN	Keratin, type I cytoskeletal 9	15
15.98	38.46	O75636	FCN3_HUMAN	Ficolin-3	17
15.28	5.706	Q9Y490	TLN1_HUMAN	Talin-1	10
15.2	35.13	P01876	IGHA1_HUMAN	Ig alpha-1 chain C region	10
14.77	18.18	P06727	APOA4_HUMAN	Apolipoprotein A-IV	9
13.64	11.64	P02730	B3AT_HUMAN	Band 3 anion transport protein	8
12.44	12.95	P04196	HRG_HUMAN	Histidine-rich glycoprotein	8
12	82.08	P01834	IGKC_HUMAN	Ig kappa chain C region	12

10.27	20.8	P63261	ACTG_HUMAN	Actin, cytoplasmic 2	8
10.27	20.8	P60709	ACTB_HUMAN	Actin, cytoplasmic 1	8
10	32.65	P68871	HBB_HUMAN	Hemoglobin subunit beta	8
9.89	5.726	P07996	TSP1_HUMAN	Thrombospondin-1	6
16.22	22.69	P35908	K22E_HUMAN	Keratin, type II cytoskeletal 2 epidermal	18
9.74	28.91	P02775	CXCL7_HUMAN	Platelet basic protein	7
7.49	13.28	P04040	CATA_HUMAN	Catalase	5
7.46	2.34	P11277	SPTB1_HUMAN	Spectrin beta chain, erythrocytic	4
7.31	4.524	P08514	ITA2B_HUMAN	Integrin alpha-IIb	5
17.15	40.49	P01859	IGHG2_HUMAN	Ig gamma-2 chain C region	14
6.52	15.8	P02765	FETUA_HUMAN	Alpha-2-HS-glycoprotein	5
6.4	25.83	P60660	MYL6_HUMAN	Myosin light polypeptide 6	4
6	23.27	P01591	IGJ_HUMAN	Immunoglobulin J chain	3
5.97	6.472	P05106	ITB3_HUMAN	Integrin beta-3	4
5.9	56.36	P81605	DCD_HUMAN	Dermcidin	8
5.82	2.02	P16157	ANK1_HUMAN	Ankyrin-1	3
5.36	2.108	P02549	SPTA1_HUMAN	Spectrin alpha chain, erythrocytic 1	4
5.14	30.3	P02656	APOC3_HUMAN	Apolipoprotein C-III	5
4.94	28.17	P69905	HBA_HUMAN	Hemoglobin subunit alpha	5
4.55	43.56	P02655	APOC2_HUMAN	Apolipoprotein C-II	4
4.04	1.323	P01024	CO3_HUMAN	Complement C3	2
4.04	28.3	P0CG06	LAC3_HUMAN	Ig lambda-3 chain C regions	9
4.04	28.3	P0CG05	LAC2_HUMAN	Ig lambda-2 chain C regions	9
4.04	28.3	P0CG04	LAC1_HUMAN	Ig lambda-1 chain C regions	9
4.04	14.02	B9A064	IGLL5_HUMAN	Immunoglobulin lambda-like polypeptide 5	9
4	28.3	P0CF74	LAC6_HUMAN	Ig lambda-6 chain C region	9
4.02	5.24	P15169	CBPN_HUMAN	Carboxypeptidase N catalytic chain	3
4.02	2.662	P11171	41_HUMAN	Protein 4.1	2
20.07	31.71	P04220	MUCB_HUMAN	Ig mu heavy chain disease protein	19
4	3.333	Q14624	ITIH4_HUMAN	Inter-alpha-trypsin inhibitor heavy chain H4	2

3.7	12.14	Q15485	FCN2_HUMAN	Ficolin-2	4
3.6	2.279	Q92954	PRG4_HUMAN	Proteoglycan 4	3
3.43	4.762	P02790	HEMO_HUMAN	Hemopexin	2
3.28	6.982	P07437	TBB5_HUMAN	Tubulin beta chain	2
3.18	10.19	P13224	GP1BB_HUMAN	Platelet glycoprotein Ib beta chain	3
2.75	7.536	P02749	APOH_HUMAN	Beta-2-glycoprotein 1	2
2.67	10.95	O43866	CD5L_HUMAN	CD5 antigen-like	2
2.18	24.56	P06702	S10A9_HUMAN	Protein S100-A9	3
2.03	4.049	P07477	TRY1_HUMAN	Trypsin-1	2
2	10	P02652	APOA2_HUMAN	Apolipoprotein A-II	2
1.78	4.43	P02774	VTDB_HUMAN	Vitamin D-binding protein	2

References

(1) Hu, L., Ai, Z., Liu, P., Xiong, Q., Min, M., Lan, C., Wang, J., Fan, L., and Chen, D. (2010) Predicting the Binding Affinity of Epitope-Peptides with HLA-A*0201 by Encoding Atom-Pair Non-covalent Interaction Information between Receptor and Ligands. *Chem. Biol. Drug Des. 75*, 597–606.
(2) Wold, S., Jonsson, J., Sjörström, M., Sandberg, M., and Rännar, S. (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta 277*, 239–253.
(3) Maccari, G., Di Luca, M., Nifosí, R., Cardarelli, F., Signore, G., Boccardi, C., and Bifone, A. (2013) Antimicrobial peptides design by evolutionary multiobjective optimization. *PLoS Comput. Biol. 9*, e1003212.
(4) Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S. (1987) Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem. 30*, 1126–35.
(5) Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S. (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem. 41*, 2481–91.