SUPPLEMENTARY MATERIALS

REDESIGNING HUMAN 2'-DEOXYCYTIDINE KINASE ENANTIOSELECTIVITY FOR L-NUCLEOSIDE ANALOG AS REPORTERS IN POSITRON EMISSION TOMOGRAPHY

Pravin Muthu, Hannah X. Chen, Stefan Lutz *

Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, GA, 30322. USA,

*** To whom correspondence should be addressed: sal2@emory.edu**

Library A Score Function (A1 Score)

We assembled a linear model using experimental data associated with a given PDB entry (see manuscript and page SM3). More specifically, an energy minimized structure of each PDB structure was generated using the standard score function of the Rosetta program. The un-weighted score terms were extracted and associated to the natural log of the Michaels constant (ln K_M). A new score function was parameterized using non-negative least squares regression. The table below shows weights comparing the standard Rosetta score function and the re-weighted score function (A1 Score) used for Library A design.

Table S1: Reweighted score function

Score Term	Rosetta	A1 Score
Lennard Jones Attractive	0.800	0.000
Lennard Jones Repulsive	0.440	0.238
Larzardis-Karplus Solvation	0.650	0.182
Lennard Jones Intermolecular Repulsive	0.004	0.043
Proline Ring Closure	1.000	0.000
Salt Bridge Interactions	0.490	0.000
Hydrogen Bond (Short Range)	0.585	0.158
Hydrogen Bond (Long Range)	1.170	0.035
Hydrogen Bond (Backbone-Sidechain)	1.170	1.357
Hydrogen Bond (Sidechain-Sidechain)	1.100	0.209
Disulfide Bond Energy	1.000	0.000
Ramachandran Statistical Energy	0.200	0.000
Omega Torsion Statistical Energy	0.500	0.000
Dunbrack Statistical Energy	0.560	0.000
Amino Acid (Phi,Psi) Probability	0.320	0.426
Reference Energy	1.000	0.000

Library A Training Data

The re-weighted linear score function (A1 score) used to design Library A was based on crystallographic data of deoxycytidine kinases bound with various substrates listed in the table below. The structures were imported into the Rosetta program and energy minimized. While the substrate was modeled enumerating various torsional combinations, the minimized structure was always observed to a substrate conformation similar to that of the crystal structure. This observation is consistent with crystallographic data, as near identical torsions are observed for all bound structures.

Table S2: Training data for Library A

PDB	Substrate	Mutations	K_M (µM)	A1 Score
2NO1	L-Deoxycytidine	C4S	3	1.62
2NO7	D-Deoxycytidine	C4S	3	1.65
2NOA	Lamivudine	C4S	3	7.72
2NO6	Emtricitabine	C4S	4.9	1.25
3KFX	5-methyl-D-Deoxycytidine	WT	7.8	1.56
1P5Z	Cytarabine	WT	13.1	2.54
2NO9	Troxacitabine	WT	13.2	1.93
1P62	Gemcitabine	WT	16.1	3.34
3HP1	L-Thymidine	R104M/D133A	138	5.08
2ZI4	L-Deoxyadenosine	C4S	190	4.67

In order to assess the accuracy of the new score function (A1 score), we applied the function to an independent kinetic test set, from published data from Iyidogan and Lutz (1) (natural log of the Michaelis constant K_M). The initial structures were based on 2NO1, 2ZI7, 3KFX and 2ZI9 for D-deoxycytidine, D-deoxyguanosine, D-thymidine and D-deoxyadenosine. The relevant mutations were made using fixed backbone design, and the structure was energy minimized using the standard Rosetta score function. Only the crystalized substrate conformer was used. The resulting structure was scored using the A1 score, and is depicted in the table and figure below. The A1 score had moderate statistical correlation to the test data: $R^2 = 0.34$ and Pearson's r = 0.59.

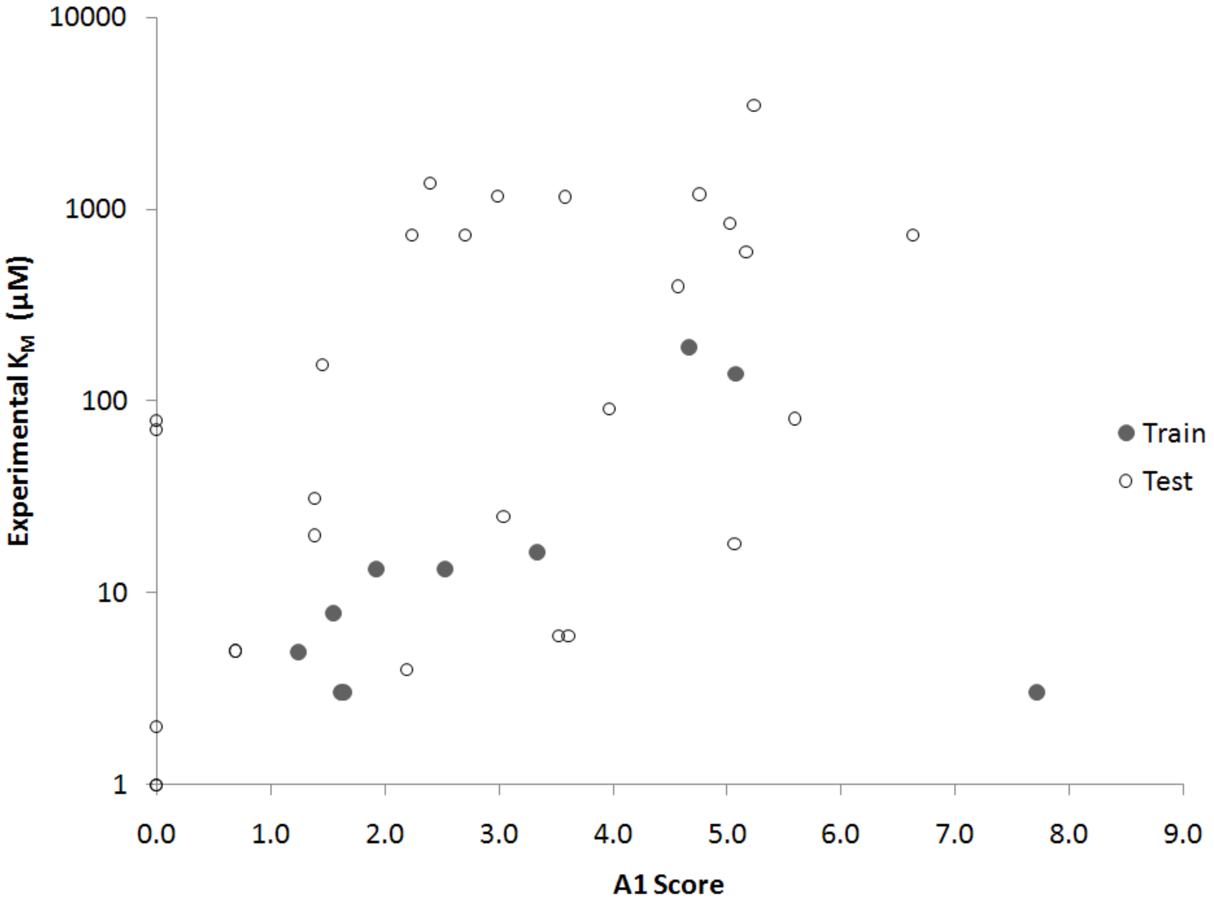

Figure S1: Comparison of test and training data using reweighted core function: A1 Score.

Table S3: Test Data for Library A (experimental data from ref. 1)

Mutations	Substrate	K_M (µM)	A1 Score
WT	D-Deoxycytidine	1	0.00
WT	D-Deoxyguanosine	155	1.46
WT	D-Thymidine	3480	5.24
WT	D-Deoxyadenosine	81	5.60
R104M,D133T	D-Deoxycytidine	20	1.39
R104M,D133T	D-Thymidine	6	3.61
R104M,D133T	D-Deoxyguanosine	1203	4.76
R104M,D133T	D-Deoxyadenosine	739	6.63
R104M,D133S	D-Deoxycytidine	5	0.69
R104M,D133S	D-Deoxyguanosine	1174	3.00
R104M,D133S	D-Deoxyadenosine	398	4.57
R104M,D133S	D-Thymidine	18	5.07
D47E,R104Q,D133G,N163I,F242L	D-Deoxycytidine	1	0.00
D47E,R104Q,D133G,N163I,F242L	D-Deoxyguanosine	79	0.00
D47E,R104Q,D133G,N163I,F242L	D-Thymidine	25	3.04
D47E,R104Q,D133G,N163I,F242L	D-Deoxyadenosine	91	3.97
A100V,R104M,D133T	D-Deoxycytidine	71	0.00
A100V,R104M,D133T	D-Thymidine	4	2.20
A100V,R104M,D133T	D-Deoxyadenosine	739	2.24
A100V,R104M,D133T	D-Deoxyguanosine	1164	3.58
A100V,R104M,D133S	D-Deoxycytidine	5	0.69
A100V,R104M,D133S	D-Deoxyguanosine	739	2.71
A100V,R104M,D133S	D-Thymidine	6	3.53
A100V,R104M,D133S	D-Deoxyadenosine	843	5.03
A100V,R104M,D133A	D-Deoxycytidine	2	0.00
A100V,R104M,D133A	D-Thymidine	31	1.39
A100V,R104M,D133A	D-Deoxyguanosine	1364	2.40
A100V,R104M,D133A	D-Deoxyadenosine	598	5.17

Library A Mutations

Using a similar to protocol to obtain mutant structures and A1 scores, we attempted to find single mutations predicted to favor L-thymidine over D-thymidine within 3-shells of the substrate. A shell is defined to be all atoms within 4.0 angstroms from a given set of atoms. The PDB structures of 3KFX and 3HP1 were used to model D- and L- thymidine respectively. For each investigated position, all 20 amino acids were modeled in addition to the three base ssTK1A mutations. The difference of the A1 scores of L-thymidine and D-thymidine were used to evaluate the contribution of each mutation. Specifically, a negative difference suggests a favorable Michaelis constant for L-thymidine over D-thymidine. The top 3 single mutations were selected for experimental evaluation. Additionally, a secondary scoring mutation selected at each lead position.

Table S4: Top-predicted mutations for Library A

Variant	Mutants	A1 Score	A1 Score	Difference
		(D-Thymidine)	(L-Thymidine)	
A2	A100V,R104M,D133S,W58V	4.80	1.30	-3.5
A4	A100V,R104M,D133S,W58E	1.80	0.50	-1.3
A6	A100V,R104M,D133S,F96D	4.90	0.00	-4.9
A3	A100V,R104M,D133S,F96Y	5.60	1.20	-4.5
A5	A100V,R104M,D133S,E196L	3.70	0.60	-3.2
A1	A100V,R104M,D133S,E196A	3.00	1.10	-2.0

Library B Score Function (B1 and B2)

Variant structures were modeled similar to Library A, using fixed backbone design used to create initial mutant structures, followed by subsequent energy minimization to the standard Rosetta score function. However instead of a single resulting structure, an ensemble of five distinct structures was generated by independent trajectories. In addition the standard score terms, calculations specific to the bound D- or L-thymidine were additionally added. Using the corresponding experimental data (ln K_M), we attempted to find statistical correlation of un-weighted score terms for each structure. Correlation was evaluated using rank correlation, to an arbitrary significance of p-values less than 0.15. Based on this criteria, six score terms were selected as the feature set (highlighted in black).

Score Term	p-value
Lennard Jones Attractive	0.21
Lennard Jones Repulsive	0.09
Larzardis-Karplus Solvation	0.95
Lennard Jones Intermolecular Repulsive	0.40
Proline Ring Closure	0.54
Salt Bridge Interactions	0.09
Hydrogen Bond (Short Range)	0.92
Hydrogen Bond (Long Range)	0.45
Hydrogen Bond (Backbone-Sidechain)	0.37
Hydrogen Bond (Sidechain-Sidechain)	0.70
Dunbrack Statistical Energy	0.08
Amino Acid (Phi,Psi) Probability	0.74
Reference Energy	0.01
Lennard Jones Attractive (Substrate)	0.61
Lennard Jones Repulsive (Substrate)	0.15
Larzardis-Karplus Solvation (Substrate)	0.37
Hydrogen Bond (Substrate)	0.08

Figure S2: Statistical correlation of score terms to training data

Library B Training Data

The data used for Library B, was based on the experimental data from Library A. Using the feature set of six statistically correlated score terms, two separate score functions were created: B1 Score and B2 Score to model Michaels constant (ln K_M) and catalytic efficiency (ln k_{cat}/K_M). The performance of each score function was evaluated using leave on out validation, and the resulting values are tabulated below. After testing various machine-learning methods, the k-nearest neighbor algorithm had the best predictive performance. The non-parametric Library B score functions had slightly better correlations and are shown in the figure below. B1 Score: R^2 = 0.62, and Pearson's r = 0.72. B2 Score: R^2 = 0.49, and Pearson's r = 0.70.

Table S5: Training data for Library B

Variant	Mutant	Substrate	B1 Score	K_M (µM)	B2 Score	k_{cat}/K_M ($s^{-1}mM^{-1}$)
WT		L-Thymidine	7.6	nd	-5.4	nd
WT		D-Thymidine	6.5	nd	-5.8	nd
ssTK1A	A100V,R104M,D133S	L-Thymidine	4.7	20.7	-5.6	130
ssTK1A	A100V,R104M,D133S	D-Thymidine	4.1	11.7	-4.7	240
A1	A100V,R104M,D133S,F96D	L-Thymidine	4.1	25.5	-5.4	10
A1	A100V,R104M,D133S,F96D	D-Thymidine	4.3	76.6	-4.1	40
A2	A100V,R104M,D133S,W58E	L-Thymidine	4.1	569	-5.4	1
A2	A100V,R104M,D133S,W58E	D-Thymidine	4.9	1164	-6.0	1
A3	A100V,R104M,D133S,E196L	L-Thymidine	4.9	351	-6.7	1
A3	A100V,R104M,D133S,E196L	D-Thymidine	5.7	1112	-6.2	1
A4	A100V,R104M,D133S,F96Y	L-Thymidine	6.7	1652	-5.7	1
A4	A100V,R104M,D133S,F96Y	D-Thymidine	5.1	1263	-6.0	1
A5	A100V,R104M,D133S,W58V	L-Thymidine	6.7	nd	-5.7	nd
A5	A100V,R104M,D133S,W58V	D-Thymidine	6.5	nd	-5.9	nd
A6	A100V,R104M,D133S,E196A	L-Thymidine	6.9	nd	-5.6	nd
A6	A100V,R104M,D133S,E196A	D-Thymidine	6.5	nd	-5.8	nd
ssTK3	R104M,D133N	L-Thymidine	4.1	24.6	-4.6	100
ssTK3	R104M,D133N	D-Thymidine	4.9	27.9	-5.3	90

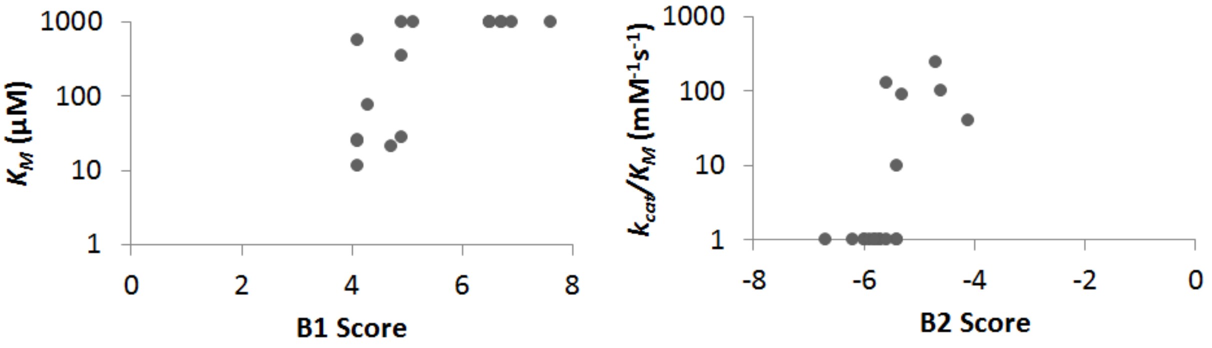

Figure S3: Leave one out validation of B1 and B2 scores to experimental K_M and k_{cat}/K_M respectively

Library B Mutations

Using the modified ensemble approach, B1 and B2 scores were calculated for D- and L-thymidine interactions for all 20 amino acid single mutations within 3-shells of the substrate, using the two ssTK3 base mutations. The difference in B1 and B2 scores were used to evaluate each mutation. For B1 a negative difference would suggest a favorable Michaelis constant for L-thymidine, while for B2 a positive difference would suggest overall improved catalytic performance for L-thymidine. The top 4 mutations using the B1 and B2 scores were selected for experimental validation.

Table S6: Top-predicted mutations for Library B based on K_M

Variant	Mutations	B1 Score (L-Thymidine)	B1 Score (D-Thymidine)	Difference
B5	R104M,D133N,V55E	3.1	4.9	-1.8
B6	R104M,D133N,L191A	1.2	2.5	-1.2
B8	R104M,D133N,V55F	5.7	6.5	-0.8
B3	R104M,D133N,L102Y	3.3	3.7	-0.4

Table S7: Top-predicted mutations for Library B based on k_{cat}/K_M

Variants	Mutations	B2 Score (L-Thymidine)	B2 Score (D-Thymidine)	Difference
B4	R104M,D133N,M85Y	-4.6	-5.4	0.8
B7	R104M,D133N,V130T	-6.0	-6.7	0.7
B1	R104M,D133N,P89F	-4.5	-5.0	0.5
B2	R104M,D133N,A138I	-3.9	-4.4	0.5

Overall Predictive Performance (B3 Score)

In a summative capacity, the B3 score was used to evaluate predictive performance using the final set of kinetic data. Using a similar methodology, six score terms were non-parametrically fit to catalytic efficiency (k_{cat}/K_M) using the k-nearest neighbor algorithm. The standard Rosetta score function has little/no correlation to the experimental data (R^2 = 0.02, and Pearson's r = 0.08). The B3 Score has moderate statistical correlation (R^2 = 0.62, and Pearson's r = 0.82), and has a slight improvement to the predecessor function B2 Score (compiled using less data points). For plotting purposes, both the Rosetta and B3 Score have been normalized.

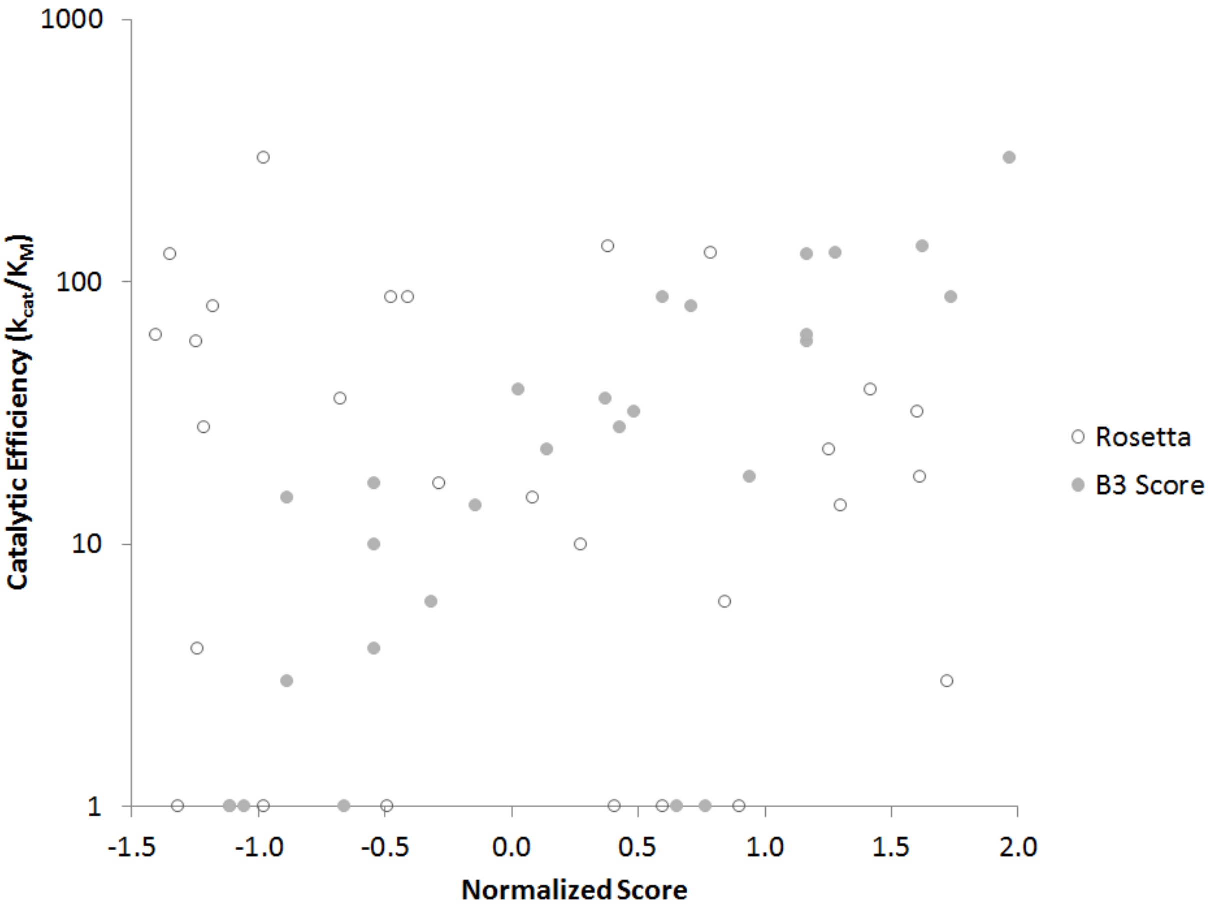

Figure S4: Comparison of Predictive Performance of the Standard Rosetta and B3 Score Function

Table S8: Tabulated data for normalized Rosetta and R3 Scores

Variant	Mutations	Substrate	k_{cat}/K_M ($mM^{-1}s^{-1}$)	Rosetta	B3 Score
WT		D-Thymidine	0	-0.5	-0.6
WT		L-Thymidine	0	0.9	-0.9
ssTK1A	A100V,R104M,D133S	D-Thymidine	298	-1.0	2.0
ssTK1A	A100V,R104M,D133S	L-Thymidine	137	0.4	1.6
A1	A100V,R104M,D133S,F96D	D-Thymidine	0	0.1	-1.1
A1	A100V,R104M,D133S,F96D	L-Thymidine	0	-0.2	-0.8
A2	A100V,R104M,D133S,W58E	D-Thymidine	0	-0.9	-1.1
A2	A100V,R104M,D133S,W58E	L-Thymidine	0	-1.1	-1.1
A3	A100V,R104M,D133S,E196L	D-Thymidine	0	-0.1	-0.6
A3	A100V,R104M,D133S,E196L	L-Thymidine	0	0.6	-1.0
A4	A100V,R104M,D133S,F96Y	D-Thymidine	1	-1.0	-1.1
A4	A100V,R104M,D133S,F96Y	L-Thymidine	1	0.6	-0.7
A5	A100V,R104M,D133S,W58V	D-Thymidine	1	0.4	-1.1
A5	A100V,R104M,D133S,W58V	L-Thymidine	1	0.9	-1.1
A6	A100V,R104M,D133S,E196A	D-Thymidine	59	-1.2	1.2
A6	A100V,R104M,D133S,E196A	L-Thymidine	87	-0.5	1.7
ssTK3	R104M,D133N	D-Thymidine	81	-1.2	0.7
ssTK3	R104M,D133N	L-Thymidine	87	-0.4	0.6
B1	R104M,D133N,P89F	D-Thymidine	0	-0.6	-0.9
B1	R104M,D133N,P89F	L-Thymidine	0	-1.1	-0.3
B2	R104M,D133N,A138I	D-Thymidine	0	1.4	-1.1
B2	R104M,D133N,A138I	L-Thymidine	0	0.7	-0.5
B3	R104M,D133N,L102Y	D-Thymidine	127	-1.3	1.2
B3	R104M,D133N,L102Y	L-Thymidine	129	0.8	1.3
B4	R104M,D133N,M85Y	D-Thymidine	28	-1.2	0.4
B4	R104M,D133N,M85Y	L-Thymidine	32	1.6	0.5
B5	R104M,D133N,V55E	D-Thymidine	1	-0.5	0.8
B5	R104M,D133N,V55E	L-Thymidine	1	-1.3	0.7
B6	R104M,D133N,L191A	D-Thymidine	23	1.3	0.1
B6	R104M,D133N,L191A	L-Thymidine	39	1.4	0.0
B7	R104M,D133N,V130T	D-Thymidine	17	-0.3	-0.5
B7	R104M,D133N,V130T	L-Thymidine	36	-0.7	0.4
B8	R104M,D133N,V55F	D-Thymidine	6	0.8	-0.3
B8	R104M,D133N,V55F	L-Thymidine	14	1.3	-0.1
B6-II	R104M, D133N, V130T, L191A	D-Thymidine	18	1.6	0.9
B6-II	R104M, D133N, V130T, L191A	L-Thymidine	63	-1.4	1.2
B8-II	R104M, D133N, V55F, V130T	D-Thymidine	4	-1.2	-0.5
B8-II	R104M, D133N, V55F, V130T	L-Thymidine	15	0.1	-0.9
B6-III	R104M, D133N, V55F, V130T, L191A	D-Thymidine	3	1.7	-0.9
B6-III	R104M, D133N, V55F, V130T, L191A	L-Thymidine	10	0.3	-0.5

References

(1) Iyidogan, P., and Lutz, S. (2008) Systematic exploration of active site mutations on human deoxycytidine kinase substrate specificity, Biochemistry **47**, 4711-4720.