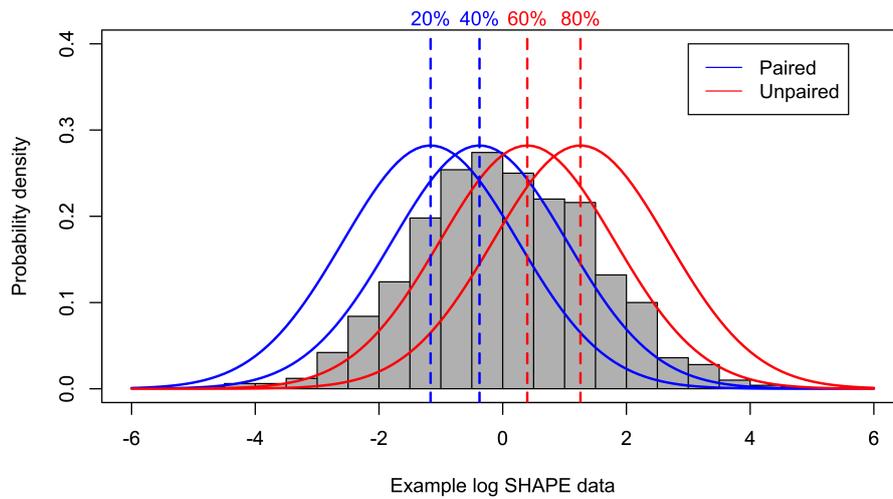


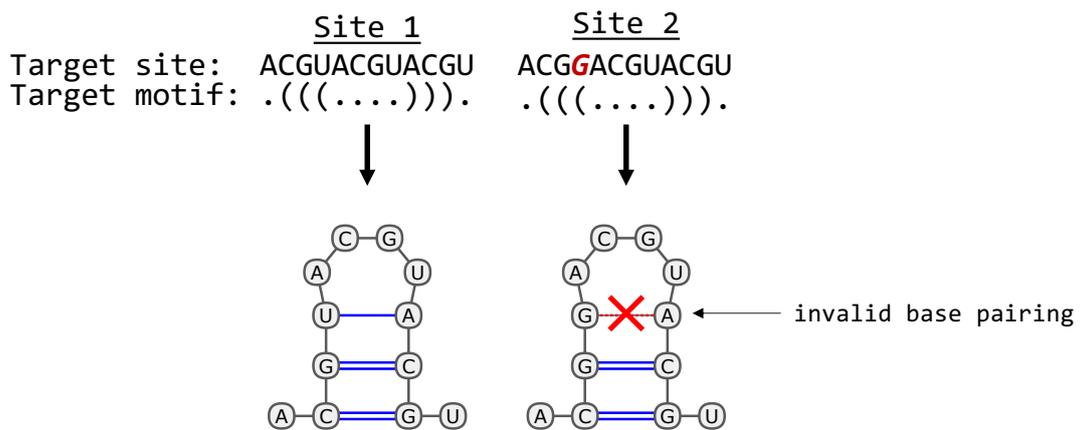
# Supplementary Materials: Automated recognition of RNA structure motifs by their SHAPE data signatures

Pierce Radecki, Mirko Ledda, and Sharon Aviran

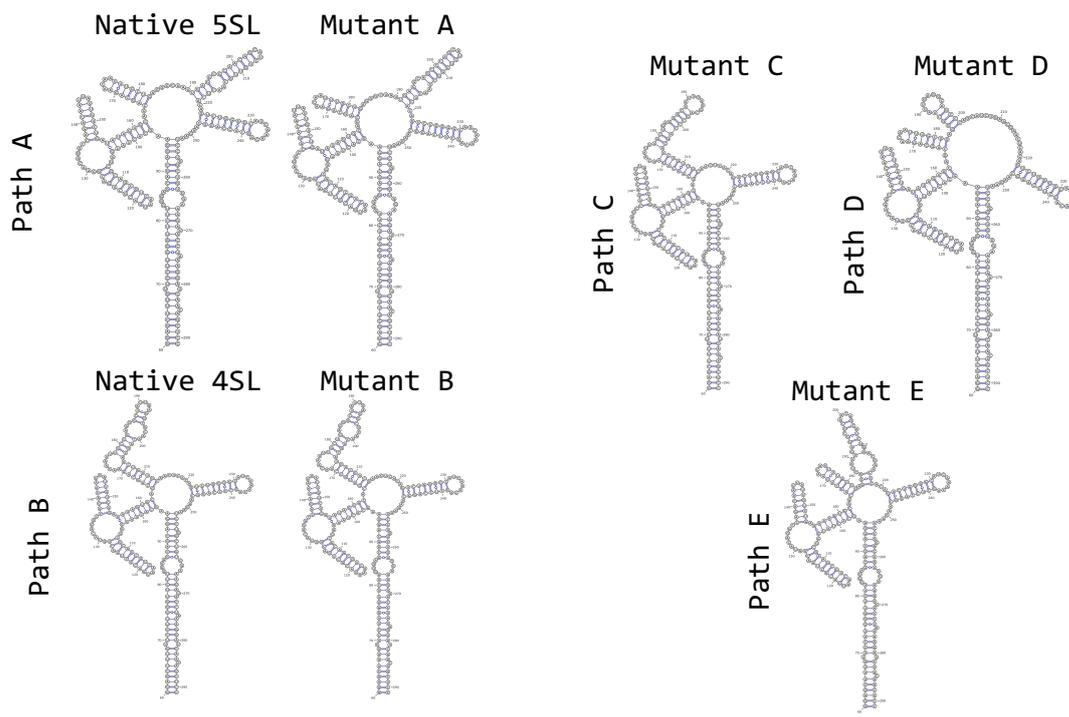
- Figure S1: Initialization of four Gaussian components using data percentiles.
- Figure S2: Illustration of sequence constraints.
- Figure S3: Secondary structures of the *in vitro* RREs.
- Figure S4: Sequences and pairing state paths of the SL III/SL IV region for RRE variants in the Sherpa set.
- Figure S5: *patteRNA* scores on the Sherpa set of RRE SHAPE profiles when searching full-length RRE paths.
- Figure S6: *patteRNA* scores for RRE motifs across four whole-genome HIV-1 structure profiles.
- Figure S7: Survival functions of *c*-scores for the 5SL and 4SL native structure of RRE across human transcriptome-wide PARS and HIV-1 SHAPE datasets.
- Figure S8: *patteRNA* score ratios (5SL/4SL) for mixtures of the 5SL and 4SL native isomers of the RRE.
- Figure S9: Comparison of trained models using an entire dataset and a reduced training subset.
- Table S1: *patteRNA* scoring of the SL III/SL IV region (nt 7409-7467) of RRE in genomic SHAPE data against the candidate paths described in the Sherpa set.



**Figure S1.** Initialization of four Gaussian components using data percentiles. Grey histograms represent the distribution of example data. In this case, the parameter  $K = 2$  (i.e., two components per pairing state) and each Gaussian component is represented by a solid line with blue indicating the two components used to model paired nucleotides, and red, unpaired ones. Gaussian means are spaced at regular percentile intervals, in this case at 20%, 40%, 60% and 80% of the data distribution density, respectively.

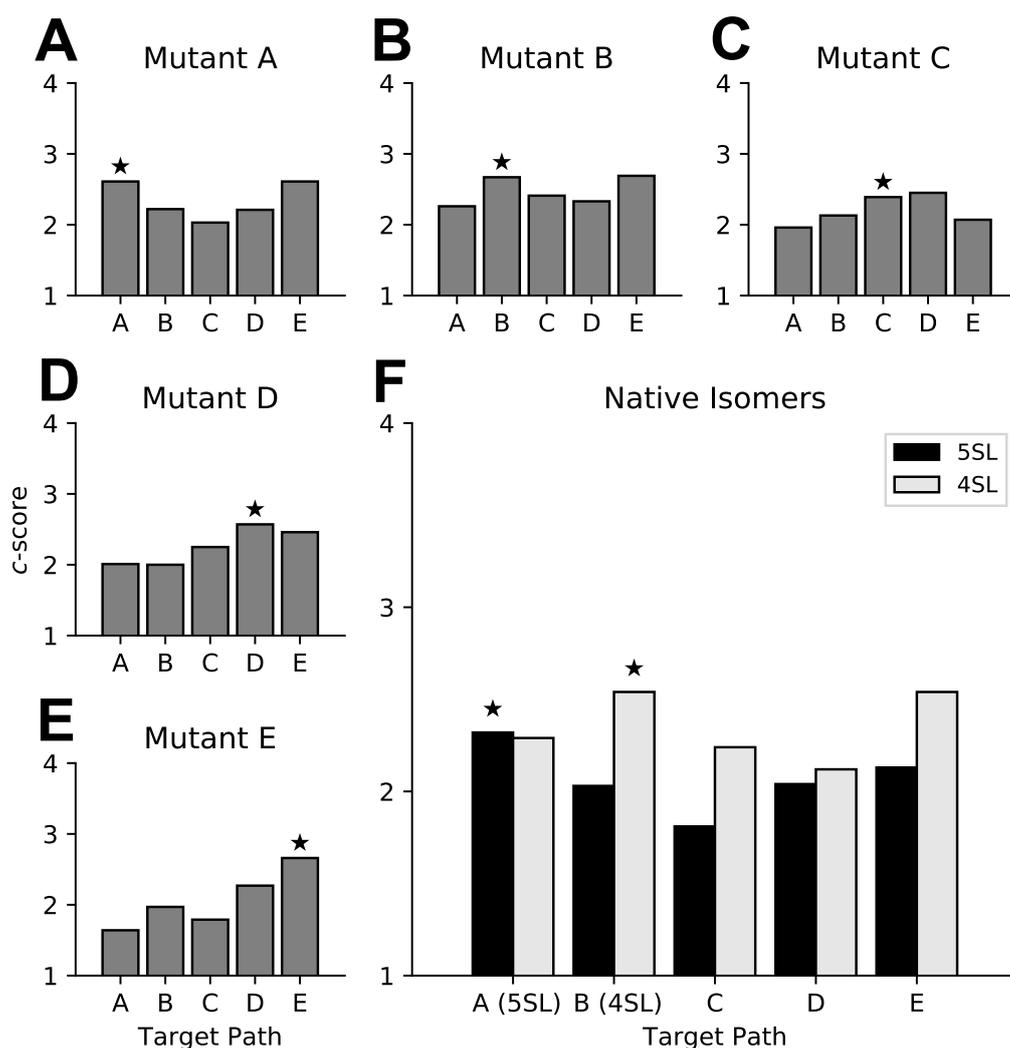


**Figure S2.** Illustration of sequence constraints. When comparing the target motif to the nucleotide sequence in Site 1, all base pairings follow the canonical rules (G–C, A–U, G–U allowed). This site consequently “passes” sequence constraints. On the contrary, the nucleotide sequence in Site 2 gives rise to non-canonical base pairings. Specifically, a G–A pairing is deemed invalid. As such, this site violates sequence constraints.

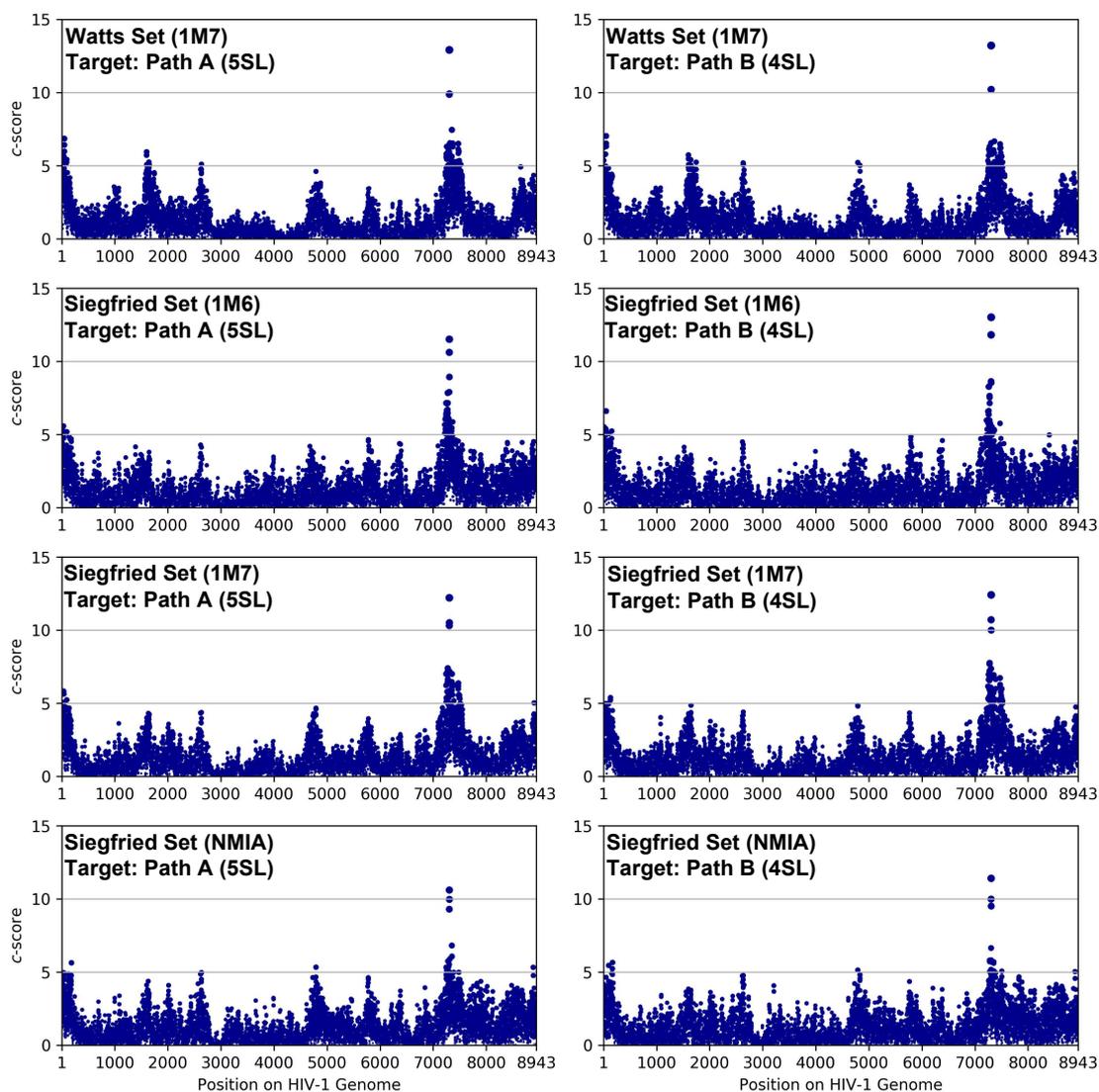


**Figure S3.** Secondary structures of the *in vitro* RREs (nt 60-291), as predicted by Sherpa *et al.*

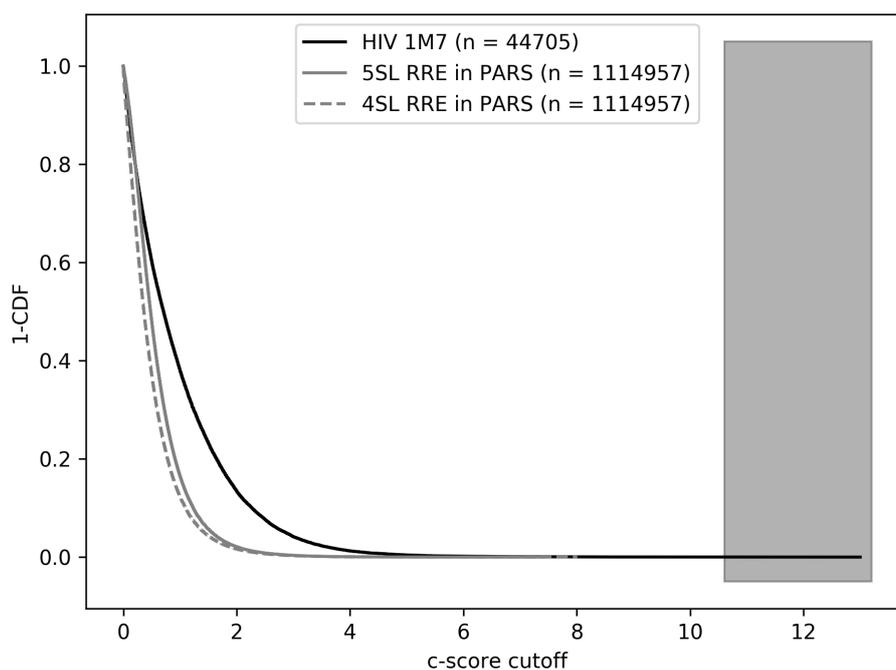




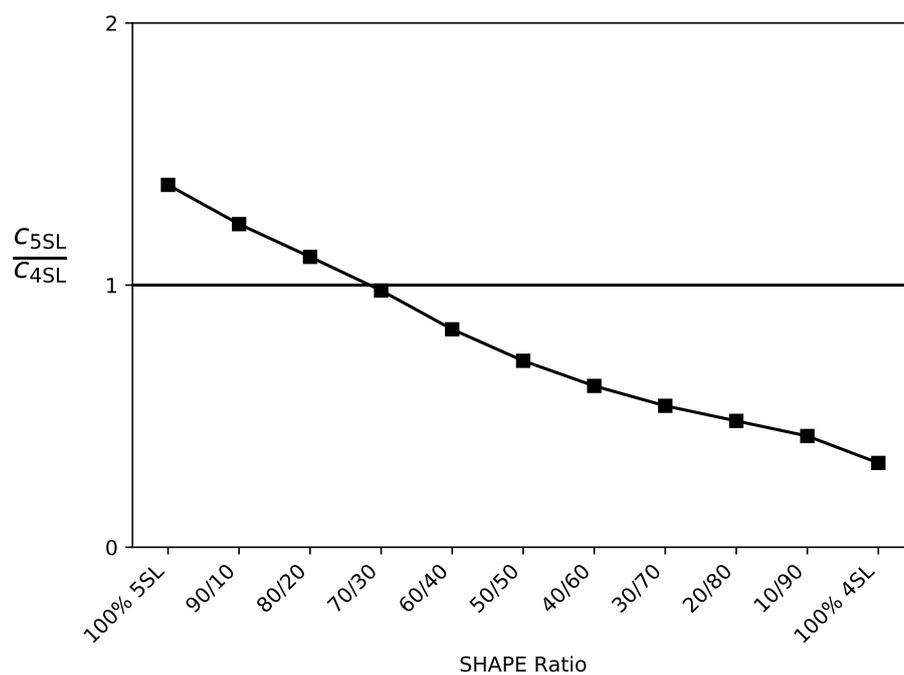
**Figure S5.** *patteRNA* scores on the Sherpa set of RRE SHAPE profiles when searching full-length RRE paths. **(A-E)** Each panel corresponds to a SHAPE profile for an RRE mutant. Grey bars indicate *patteRNA*'s *c*-scores for the five paths A-E. Highlighted with a star is the score for the predicted path in the tested profile. **(F)** *c*-scores for the two native 5SL and 4SL isomers. Bars correspond to scores for paths A-E on the 5SL (black) and 4SL (grey) profiles. Similar to the other panels, stars highlight scores for the predicted path in each profile, namely path A for 5SL and path B for 4SL. Note that *y*-axes start at 1 to better highlight differences in *c*-scores between paths, which relate primarily to differences in 59 out of 232 nucleotides when searching the full-length path.



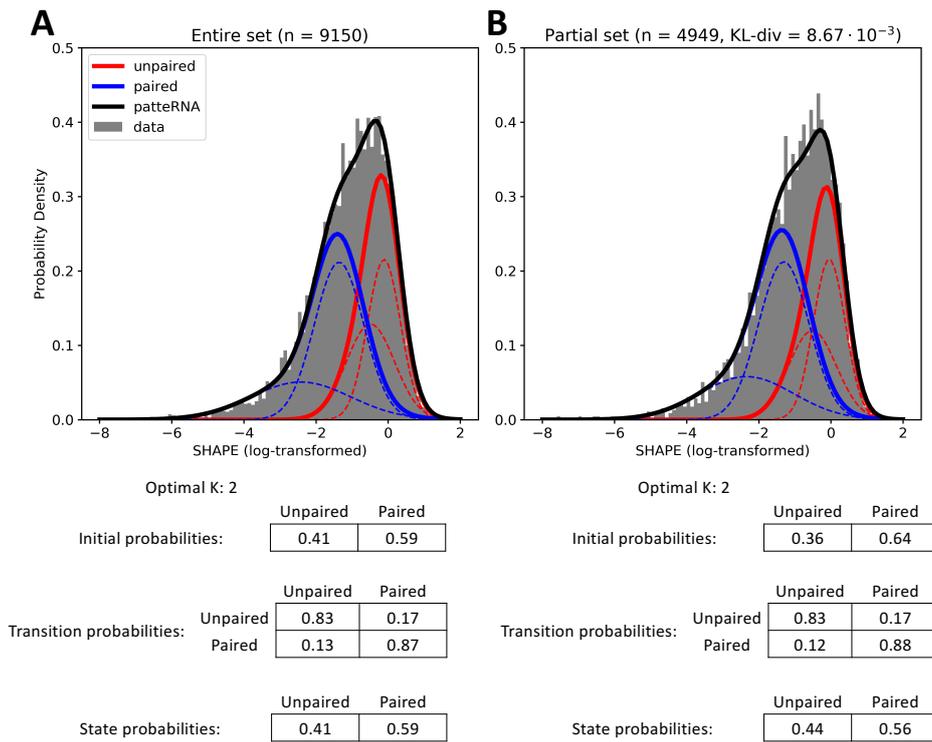
**Figure S6.** *patteRNA* scores for RRE motifs across four whole-genome HIV-1 structure profiles. *c*-scores for full-length paths A (5SL structure, left panels) and B (4SL structure, right panels) across all sites in the HIV-1 genome. Dataset and modifying reagents used are indicated in each panel and include the Watts set (SHAPE assayed with 1M7) and three profiles from the Siegfried set (SHAPE-MaP assayed with 1M6, 1M7, and NMIA, respectively). Peaks at nucleotide 7306 correspond to the known start location of the RRE.



**Figure S7.** Survival functions of  $c$ -scores for the 5SL and 4SL native structure of RRE across human transcriptome-wide PARS and HIV1 SHAPE datasets. We report  $c$ -scores for searches conducted across 649 transcripts in the PARS set with data density above 75% (i.e.  $\leq 25\%$  missing data), as well as  $c$ -scores from the entire HIV-1 RNA genome as probed with 1M7 by Siegfried *et al.*. The  $y$ -axis represents the proportion of data points with  $c$ -scores above the cutoff reported on the  $x$ -axis, i.e. the survival function defined as  $1 - \text{CDF}(c)$ , where  $\text{CDF}(c)$  is the cumulative distribution function. The grey rectangle highlights the dynamic range of  $c$ -scores (10.6 to 13.2) obtained at the location of the RRE for all considered RRE paths and HIV-1 SHAPE profiles (see Table 1 for details).



**Figure S8.** *patterNA* score ratios (5SL/4SL) for mixtures of the 5SL and 4SL native isomers of the RRE. The *x*-axis corresponds to SHAPE profiles emulating various mixtures of the 5SL/4SL conformations. The *y*-axis corresponds to *c*-score ratios between the 5SL and the 4SL paths ( $c_{5SL}/c_{4SL}$ ). Results indicate a stable progression of *c*-score ratios initially favoring the 5SL structure until the SHAPE data is comprised by 30% 4SL, at which point the 4SL structure receives higher scores.



**Figure S9.** Comparison of trained models using an entire dataset and a reduced training subset. The input data are based on the HIV-1 genome probed with 1M7 from the Siegfried set and partitioned into 100 bp fragments to mimic multiple transcripts. Gaussian Mixture Models (black lines) learned by *patteRNA* as well as Hidden Markov Model parameters for **(A)** the entire dataset and **(B)** a training subset determined using KL-divergence. Grey histograms represent the distribution of the SHAPE data. Distributions associated with paired and unpaired nucleotides are shown in blue and red solid lines, respectively (solid colored lines). Individual Gaussian components are highlighted by dashed colored lines (two for each pairing state as the optimal  $K = 2$  for this dataset).

**Table S1.** *patteRNA* scoring of the SL III/SL IV RRE region (nt. 7409-7467) in genomic SHAPE data against the candidate paths A–E described in the Sherpa set. *c*-scores with and without sequence constraints are included. Paths C and D violate sequence constraints and are not reported. *c*-score percentiles are included to gauge the significance of the reported values. Percentiles correspond to the percentage of scored sites falling below the reported *c*-score.

Dataset	Reagent	Search Target	<i>c</i> -score (Percentile)	
			<i>no sequence constraints</i>	<i>sequence constraints</i>
Siegfried Set	NMIA	5SL/Path A	0.75 (81st)	0.75
		4SL/Path B	1.58 (97th)	1.52
		Path C	1.18 (93rd)	<i>invalid</i>
		Path D	0.23 (41st)	<i>invalid</i>
		Path E	0.83 (84th)	0.78
	1M6	5SL/Path A	1.39 (95th)	1.39
		4SL/Path B	1.22 (94th)	1.22
		Path C	0.55 (71st)	<i>invalid</i>
		Path D	0.50 (67th)	<i>invalid</i>
		Path E	0.65 (77th)	0.65
	1M7	5SL/Path A	1.42 (96th)	1.48
		4SL/Path B	2.77 (99th)	2.75
		Path C	2.02 (99th)	<i>invalid</i>
		Path D	0.81 (84th)	<i>invalid</i>
		Path E	0.98 (89th)	0.99
Watts Set	1M7	5SL/Path A	2.60 (99th)	2.60
		4SL/Path B	2.50 (99th)	2.50
		Path C	2.18 (99th)	<i>invalid</i>
		Path D	0.16 (31st)	<i>invalid</i>
		Path E	0.45 (63rd)	0.45