Supplementary Information

Contents:

Supplementary Methods

Specific modifications to the ROSETTA energy function

Phosphorous Desolvation

The LK_DGFREE term in the atom_properties.txt file of the database is changed from -24 to -4.1 and the LK_VOLUME term is changed from to 34.8 to 14.7.

Electrostatics

The standard.wts file for the Standard ROSETTA energy function was modified by the replacement of fa_pair term (weight of 0.49) with the hack_elec term (weight of 0.5).

LK_Ball

The Electrostatics energy function was modified by the addition of several energy terms corresponding to the orientation-dependent desolvation method (these terms replace fa_sol 0.65):

lk_ball 0.325, lk_polar 0.325, lk_polar_nw 0.65, lk_nonpolar 0.65, lk_charged 0.325,
lk_ball_xd 0.325, lk_polar_xd 0.325, lk_polar_nw_xd 0.65, lk_nonpolar_xd 0.65,
lk_ball_dd 0.5, lk_polar_nw_dd 0.5, lk_nonpolar_dd 0.65

Multi-Desolvation

The LK_DGFREE term was modified for the following atom types in the atom_properties.txt file:

NH20 from -10 to -7.8 (modifying Gln and Asn)
ONH2 from -10 to -5.85 (modifying Gln and Asn)
Nlys from -20 to -16 (modifying Lys)
Narg from -11 to -10 (modifying Arg)

Corresponding reference energy changes:

Asp from -0.67 to -0.75
Glu from -0.81 to -0.71
Lys from -0.65 to -1.2
Asn from -0.89 to -0.8
Gln from -0.97 to -0.78

Attractive

The fa_atr term was increased from 0.8 to 0.95. Corresponding reference energy changes:

Phe from 0.63 to 1.63
Trp from 0.91 to 2.21
Tyr from 0.51 to 0.91

Lysine Charge

The positive charge on the hydrogens on the terminal nitrogen of lysine was increased from 0.33 to 0.48.

Reference Energies

The following reference energies were modified:

Ala from 0.16 to 0.26
Cys from 1.7 to 0.5
Phe from 1.63 to 1.7
His from 0.56 to 0.8
Ile from 0.24 to -0.1
Met from -0.34 to -0.1
Arg from -0.98 to -0.65
Ser from -0.37 to -0.57
Thr from -0.27 to -0.8
Val from 0.29 to -0.1

Trp from 2.21 to 2.3
Tyr from 0.91 to 1.0

Additional changes, essentially negligible for sequence recovery, for this “Optimized” energy function include the change of hbond_sc from 1.1 to 1.17 and the addition of the two command line options “-*local_bb_sc_downweight* 0.2” and “-*apply_proton_chi_potential*”.

Code modification for the HighTemp-Packer calculation

The lower temperature of the simulated annealing algorithm was changed from 0.3 to 1.3 because it provided a similar level of diversity to the diversity derived from the DNA-Rebuild method. The final quenching step of the simulated annealing algorithm is also removed. The two lines of code for these changes are in the /mini/src/core/annealer/SimAnnealerBase.cc.

Original Line 44: const PackerEnergy SimAnnealerBase::lowtemp = 0.3;
New Line 44: const PackerEnergy SimAnnealerBase::lowtemp = 1.3;
Original Line 272: void SimAnnealerBase::set_to_quench(){ quench_ = true;}
New Line 272: void SimAnnealerBase::set_to_quench(){ quench_ = false;}

Available command-line options for motif Protocols

Used only with dna_motif_collector for collection of the motif library.
'keep_motif_xtal_location', 'Boolean', default = 'false', desc= 'controls whether motifs are moved away from original PDB location (comparison between motifs is easier if they are moved, so that's default).'
'pack_score_cutoff', 'Real', default = '-0.5', desc = 'fa_atr (attractive) + fa_rep (repulsive) energy threshold for a two-residue interaction to determine if it is a motif.'
'hb_score_cutoff', 'Real', default = '-0.3', desc = 'hbond_sc (sidechain-sidechain hydrogen bonding) energy threshold for a two-residue interaction to determine if it is a motif.'
'water_score_cutoff', 'Real', default = '-0.3', desc = 'h2o_hbond (water hydrogen bonding) energy threshold for a two-residue interaction to determine if it is a motif.'
'motif_output_directory', 'String', desc = 'path for the directory where all the motifs are collected as 2-residue pdbs.'
'eliminate_weak_motifs', 'Boolean', default = 'true', desc= 'controls whether only the top 1-2 motifs (instead of all possible interactions) are counted for every protein position in a protein-DNA interface.'
'duplicate_motif_cutoff', 'Real', default = '0.2', desc = 'RMSD cutoff for an identical, canonical base residue placed via a motif to see if that motif already exists in a motif library.'
'preminimize_motif_pdbs', 'Boolean', default = 'false', desc= 'controls whether the input PDB structure sidechains and backbone are minimized before motifs are collected.'
'preminimize_motif_pdbs_sconly', 'Boolean', default = 'false', desc= 'controls whether the input PDB structure sidechains are minimized before motifs are collected.'
'place_adduct_waters', 'Boolean', default = 'true', desc= 'whether or not adduct waters are placed before motifs are collected, there will be no water interaction energy calculated if this option is false.'

Example for dna_motif_collector:

```
/rosetta/bin/dna_motifs_collector.linuxgccrelease -motif_output_directory Motif_Dir_August2011/
-ignore_unrecognized_res -adducts dna_major_groove_water -database /minirosetta_database/ -l
```

list_PDBsAugust2011

> **'ignore_unrecognized_res'* and *'database'* are standard ROSETTA options that should be included in all commandlines and are not specific to motif protocols
> **'adducts'* is an option that can be used with non-motif protocols, but it is specific to DNA. The specification of 'dna_major_groove_water' results in the potential placement of adduct waters at canonical DNA major groove positions. The motif option *'place_adduct_waters'* will not function without this option.

Used with both motif_dna_packer_design and dna_fragment_rebuild_with_motifs.
'list_motifs', 'FileVector', desc= 'File(s) containing list(s) of two-residue, motif PDB files to process and use as the motif library.'
'motif_filename', 'String', desc= 'File containing motifs – can be used to build the motif library as an alternative to the -list_motifs that takes PDB files.'
'BPData', 'String', desc= 'File containing BuildPosition (designed protein position) specific motifs and/or rotamers.'
'list_dnaconformers', 'FileVector', desc= 'File(s) containing list(s) of PDB files to process. This option is included to allow the use of DNA residues collected from the PDB instead of a canonical DNA residue to determine if a motif passes required *z2* and *r2* cutoffs during the motif search. This option was not used during the work described in this paper.'
'target_dna_defs', 'StringVector', default = '', desc = 'Can be used in conjunction with the standard way of making DNA mutations/designable areas (*dna::design::dna_defs*) to specifically target the motif searching to a subset of the mutated bases or to do motif searches with multiple allowed DNA base types (for DNA target site prediction). The form is the same as for the standard *dna_def* – chain.position.type or X.409.ADE.'
'motif_build_defs', 'StringVector', default = '', desc = 'The protein positions that can be searched for motifs. This option is most useful to limit the amino acid types that can be included in the motif search. The format is same as for the dna_def, except that one letter codes are used – Z.33.SR would allow motifs of serine or arginine at position Z33.'
'r1', 'Real', default = '4.5', lower = '0', desc = 'RMSD cutoff between motif anchor position and motif target position for allowing a particular motif rotamer to continue on to expand with DNA conformers or just to pass with the canonical base. The RMSD is calculated between a canonical base and the nearest crystal structure base of the same type. Six atoms in the plane of the nucleobase portion of the DNA residues are used for the RMSD calculation.'
'r2', 'Real', default = '1.1', lower = '0', desc = 'RMSD cutoff between motif anchor position and motif target position for accepting the motif. The RMSD is calculated between a canonical base and the nearest crystal structure base of the same type. Six atoms in the plane of the nucleobase portion of the DNA residues are used for the RMSD calculation.'
'z1', 'Real', default = '0.75', lower = '0', desc = 'DNA motif specific: cutoff between motif target DNA position and canonical base for allowing a particular motif to continue on to expand with DNA conformers. This cutoff is a test for how parallel the canonical, placed base is to the nearest crystal structure base (dot product of two vectors from the base plane).'
'z2', 'Real', default = '0.95', lower = '0', desc = 'DNA motif specific: cutoff between motif target DNA position and DNA conformer, or repeated with a canonical base and equivalent to *z1*, placed according to motif for accepting the pair of residues. This cutoff is a test for how parallel the canonical, placed base is to the nearest crystal structure base (dot product of two vectors from the base plane).'
'dtest', 'Real', default = '5.5', lower = '0', desc = 'DNA motif specific: cutoff between motif target

DNA position and DNA conformer or canonical DNA base placed according to motif for accepting the pair of residues. This cutoff is based on the C1* atom that is found in all DNA residues.'

'rotlevel', 'Integer', default = '5', lower = '1', desc = 'level of rotamer sampling for motif search. The recommended sampling is 6, but 8 (slower and longer motif search runs) was used for the data collected in this paper.'

'num_repacks', 'Integer', default = '5', lower = '0', desc = 'number of cycles of dropping special_rot weight and design.'

'minimize', 'Boolean', default = 'true', desc = 'whether or not to minimize the motif rotamers that pass the cutoffs toward the xtal structure DNA (via the constraint of the motif interaction) prior to adding them to the packer.'

'run_motifs', 'Boolean', default = 'true', desc = 'whether or not to collect and include motif rotamers in design.'

'expand_motifs', 'Boolean', default = 'true', desc = 'whether or not to expand motifs by generating sets of designs for each type of motif (by amino acid type) identified in the search step. This option is used to generate the motif-based sequence constraint data.'

'aromatic_motifs', 'Boolean', default = 'false', desc = 'whether or not to use expand motifs, using aromatic motifs only.'

'dump_motifs', 'Boolean', default = 'false', desc = 'whether or not to output pdbs with the best rotamer/conformer for each motifs.'

'quick_and_dirty', 'Boolean', default = 'false', desc = 'quick motif run to get a list of all possible motifs before doing a real run. This type of run stops after the *dtest*, *z1*, and *r1* tests.'

'special_rotweight', 'Real', default = '-40.0', desc = 'starting weight for the weight on motif rotamers.'

'output_file', 'String', desc = 'name of output file for all the best motifs and rotamers or for the dna_motif_collector it is the file where all the motifs are dumped.'

'data_file', 'String', desc = 'name of output file for any data about how many rotamers and motifs pass what tests, etc.'

'clear_bprots', 'Boolean', default = 'false', desc = 'whether or not to clear the rotamers that were read in from a previous run and restart with only the motifs that were read in and the specified rotlevel.'

'rots2add', 'Integer', default = '100', lower = '1', desc = 'number of rotamers to add to the packer from the motif search for each amino acid type. Too many or all passing rotamers will is not allowable due to computational memory constraints.'

'restrict_to_wt', 'Boolean', default = 'true', desc = 'restrict the motif search to finding motifs of the same amino acid as the starting pose, such as for homology modeling applications.'

'rerun_motifsearch', 'Boolean', default = 'true', desc = 'setting the motif search process to run again, using the rotamers in the build position, most likely to change stringency or allowed amino acid type on a second run.'

Example for motif_dna_packer_design:

```
/rosetta/bin/motif_dna_packer_design.linuxgccrelease -run_motifs -dtest 2.0 -z1 0.97 -z2 0.97 -r1
1.0 -r2 1.0 -dna::design::z_cutoff 6.0 -motifs::rotlevel 8 -motifs::list_motifs
../list_August2011Motifs_2QOJ -motifs::output_file ./2QOJ.motifs -s ../2QOJ.pdb -score::weights
/work/sthyme/weights/dna_march2011_sr.wts
-ignore_unrecognized_res -database
/work/sthyme/sparse_databases/minirosetta_database_sparse_JA_march2011/
/minirosetta_database/ -ex1 -ex2 -ex1aro::level 6 -ex2aro::level 6 -extrachi_cutoff 0 -
dna::design::dna_defs X.409.CYT -special_rotweight -20.0 -num_repacks 4
```

*The motif search specific options shown here match the options used for all design calculations described in this work.
*'*ignore_unrecognized_res*' and '*database*' are standard ROSETTA options that should be included in all commandlines and are not specific to motif protocols
*'*score::weights*' is a standard ROSETTA option that allows the energy function to be externally chosen.
*'*ex1*', '*ex2*', '*ex1aro::level*', '*ex2aro::level*', and '*extrachi_cutoff*' are standard ROSETTA options that define the rotamer set used by the standard packer. The '*level*' specification for '*ex1aro*' and '*ex2aro*' refers to extra standard deviation sampling for $\chi1$ and $\chi2$ of aromatic residues. This level of rotamer sampling was used for all design calculations described in this work. The extra sampling for the aromatic residues was chosen because these residues are likely to be mis-designed if there is no available rotamer that is close to the native aromatic due to repulsive forces.
*'*dna::design::z_cutoff*' and '*dna::design::dna_defs*' are specific to DNA design methods, where the '*z_cutoff*' determines the designable residues and the '*dna_defs*' provide a way to mutate DNA positions and/or identify a designable area. The larger the input value to the '*z_cutoff*', the more residues surrounding the input bases to '*dna_defs*' will be considered designable (6.0 was used for all the work described here).

Currently only available in /workspace/blab version of ROSETTA.
'*minimize_dna*', 'Boolean', default = 'true', desc = 'whether or not to minimize DNA after every round of design with special_rot weight dropping.'
'*flex_sugar*', 'Boolean', default = 'false', desc = 'whether or not to add flexibility to the DNA sugar pucker.'
'*dna_geometry*', 'Boolean', default = 'false', desc = 'show DNA geometry calculations, only relevant for the dna_fragment_rebuild_with_motifs protocol.'
'*apply_proton_chi_potential*' #Used with "Optimized" energy function, but has negligible effect on sequence recovery.
'*local_bb_sc_downweight*' '0.2' #Used with "Optimized" energy function, but has negligible effect on sequence recovery.

Example for dna_fragment_rebuild_with_motifs:
Same as for the motif_dna_packer_design with the following additions:
-*minimize_dna*
-*double_frag_vall* ../double_fragment_vall_v2.lib
*'*dna::design::dna_defs*' should always be included for rebuilding with input DNA positions. The rebuilding will occur with the input dna_def position and the two surrounding bases, provided that none of the bases to be rebuilt are at termini or breaks in the DNA chain.

Table S1. Percent recoveries for iterations of ROSETTA energy function optimization.

	Unweighted Recovery						Weighted Recovery					
Motif Weight	**NA***	**0.00**	**-1.25**	**-2.50**	**-5.00**	**-10.0**	**NA***	**0.00**	**-1.25**	**-2.50**	**-5.00**	**-10.0**
Standard Training	28.25 ±0.02	28.92 ±0.06	34.16 ±0.14	36.17 ±0.23	35.40 ±0.08	33.75 ±0.19	29.64 ±0.06	29.96 ±0.00	32.28 ±0.02	33.33 ±0.12	31.11 ±0.13	28.83 ±0.10
Standard Test	29.54 ±0.13	30.93 ±0.22	35.09 ±0.20	36.24 ±0.29	35.71 ±0.22	33.55 ±0.13	31.55 ±0.26	32.24 ±0.22	33.25 ±0.25	33.27 ±0.19	30.81 ±0.00	27.93 ±0.43
Stringent HBonds Training	32.0	34.3	38.4	39.8	39.6	36.9	32.7	33.7	35.1	35.3	33.4	29.8
Stringent HBonds Test	33.2	35.8	39.5	40.4	40.0	37.0	32.9	35.3	36.0	35.8	33.8	29.7
Phos. Desolvation Training	35.2	37.2	42.1	42.6	41.8	38.7	34.4	35.9	38.2	37.7	34.9	31.0
Phos. Desolvation Test	38.0	40.3	43.5	44.2	42.4	39.3	39.2	40.5	41.4	39.9	35.9	32.2
Electrostatics Training	40.0	41.8	46.3	46.0	44.1	41.0	37.4	38.4	41.4	39.7	35.8	31.7
Electrostatics Test	42.6	45.1	46.8	47.2	44.7	40.9	41.5	43.1	42.3	41.8	37.2	32.8
LK_Ball Training	39.1	41.4	44.1	45.1	44.0	40.8	39.0	39.5	40.8	40.0	36.7	32.4
LK_Ball Test	40.1	42.5	45.1	45.9	44.0	40.2	40.0	42.3	43.3	42.6	38.2	33.0
Multi-Desolvation Training	41.9	44.2	46.7	47.4	45.4	42.3	39.3	40.8	41.5	41.5	37.3	33.0
Multi-Desolvation Test	43.2	45.3	47.6	47.3	45.2	41.1	41.0	42.4	45.0	42.0	38.2	32.8
Attractive (fa_atr) Training	43.5	46.0	47.3	46.2	44.7	42.4	39.2	40.7	40.6	38.5	35.7	32.8
Attractive (fa_atr) Test	45.0	46.2	47.8	47.6	45.6	41.4	42.3	41.8	43.9	41.5	38.4	33.1
Lysine Charge Training	44.0	46.0	47.8	47.1	45.6	42.9	38.7	40.7	40.6	39.2	37.1	33.4
Lysine Charge Test	45.2	47.9	48.4	48.2	46.0	41.8	41.3	43.4	43.6	41.2	37.7	32.5
Reference Energies Training	45.2	49.2	50.0	48.5	46.8	43.9	41.0	43.3	42.9	39.6	37.3	33.7
Reference Energiest Test	46.8	49.3	50.7	50.0	46.4	42.3	43.2	44.8	45.1	43.0	38.5	33.6

*NA = No motif rotamers added

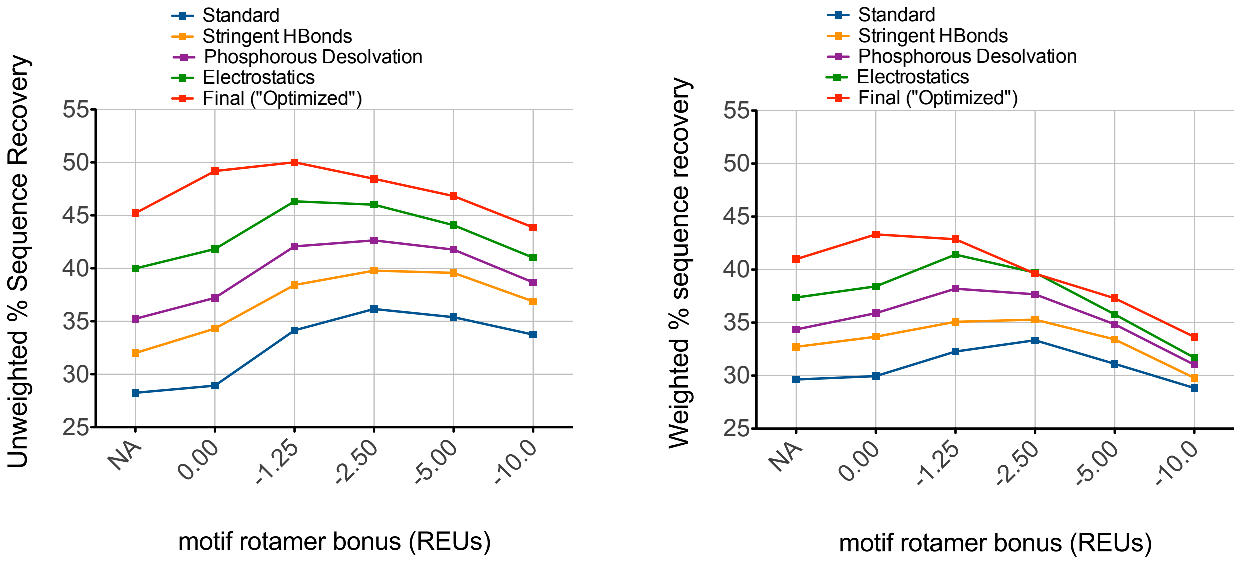


Figure S1. Sequence recovery metrics calculated for training set. The full set of 112 proteins was divided into a training set of 48 complexes and a test set of 64 complexes. The sequence recovery for different iterations of energy function optimization and varying motif weight is shown here for the training set. The data from the test set is in the main text.

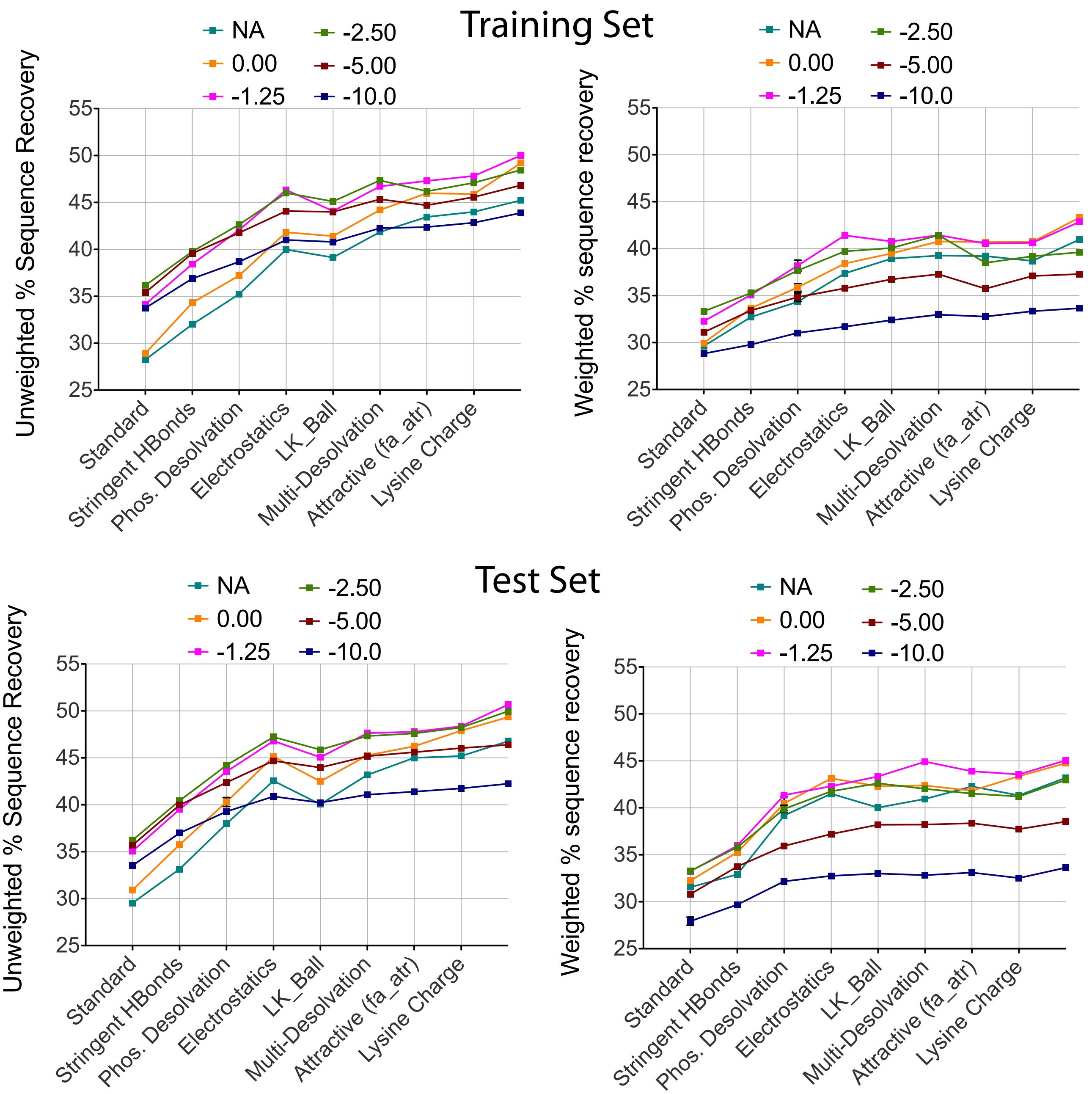
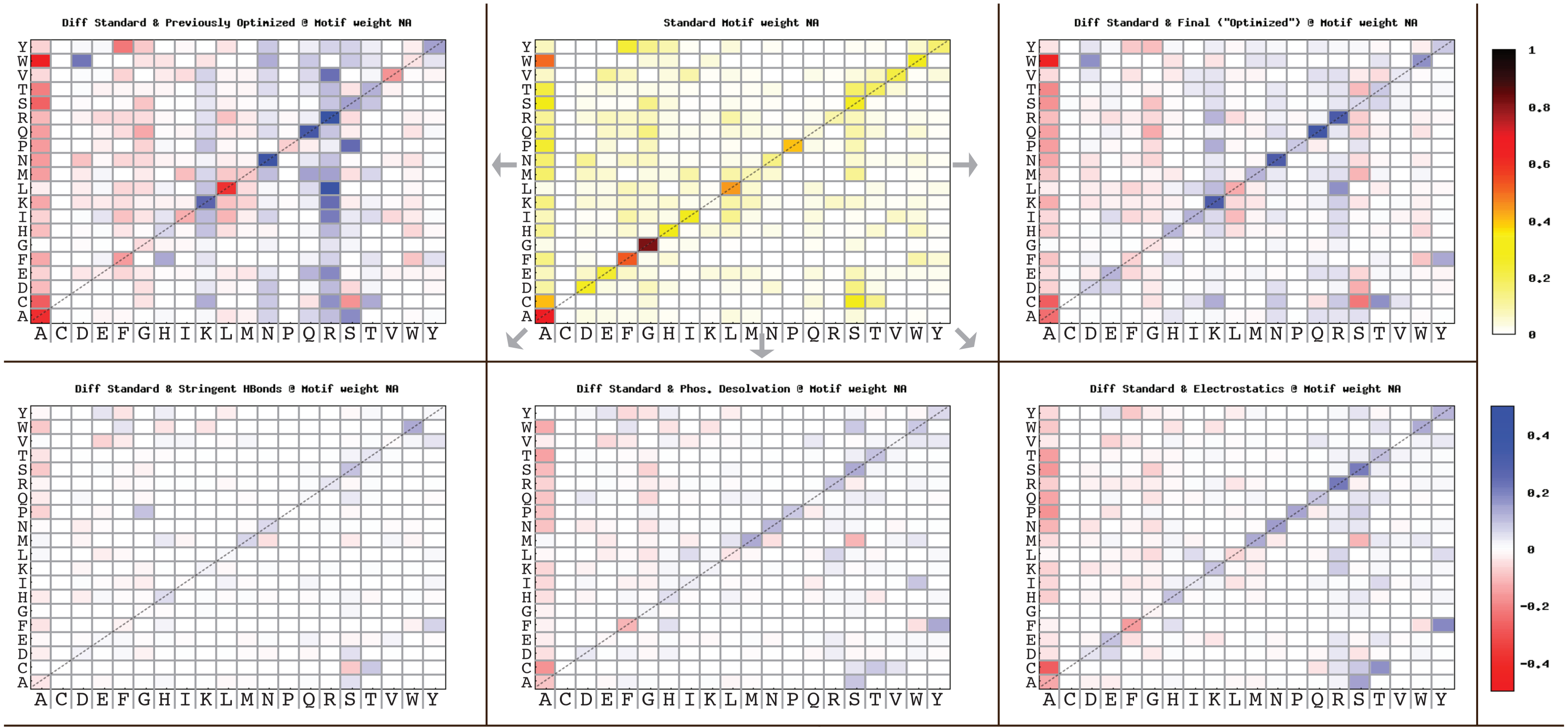


Figure S2. Percent recoveries for iterations of ROSETTA energy function optimization. The detailed sequence recovery values for both the test and training sets over all the iterations of energy function optimization and tested motif weights. Each line represents the recovery with a different weight on the added motif rotamers, and NA is the recovery with no motif rotamers added.



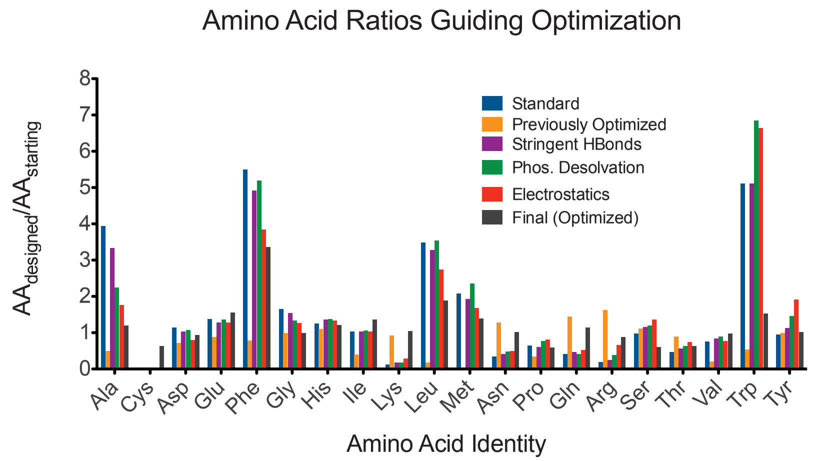
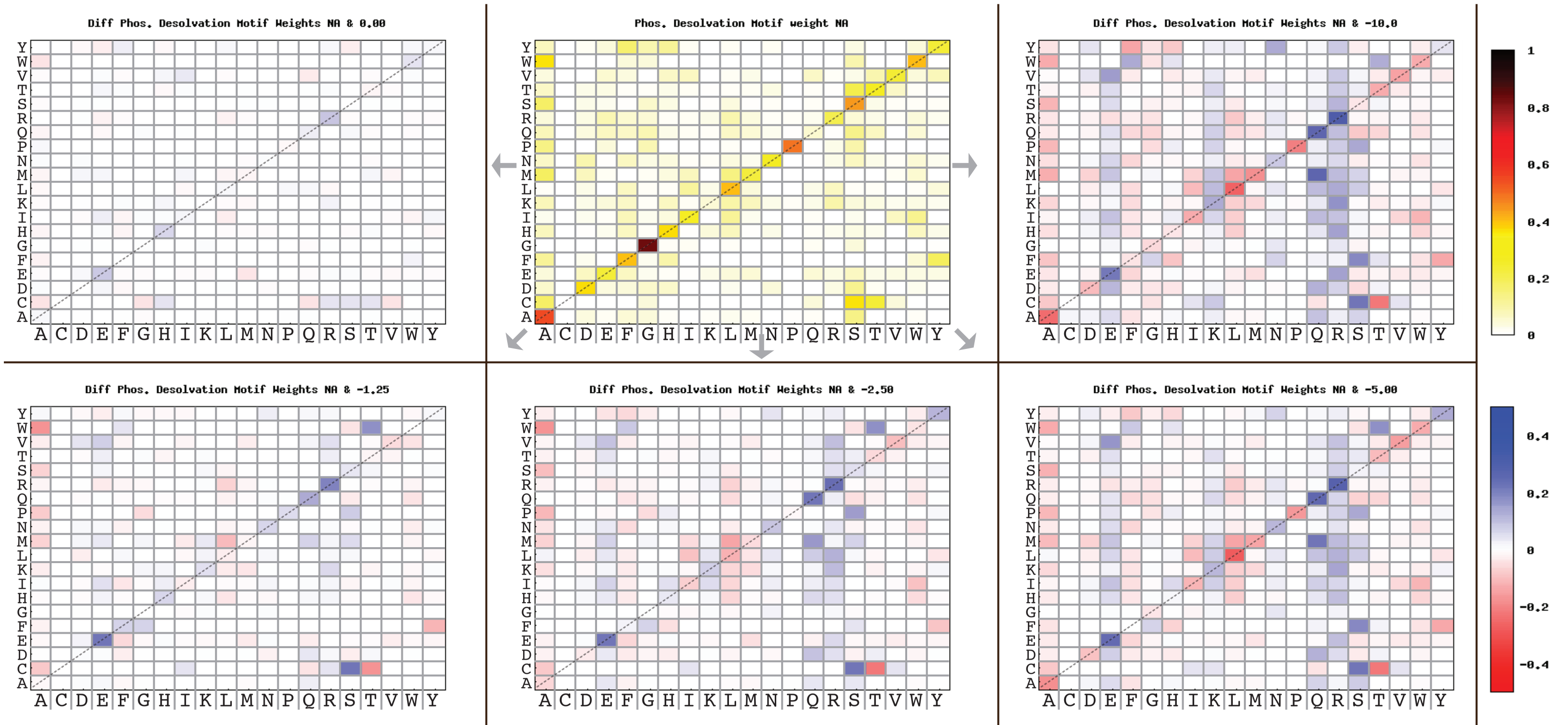


Figure S3. Comparison of the Standard energy function with steps in the energy function optimization and comparison to a previously optimized energy function. The heat maps from "Standard Motif Weight NA" heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the "Standard" energy function as the baseline for the differences. All plots in this figure are for design calculations with no motif rotamers added. The difference heat maps are (counterclockwise) 1) a previously optimized energy function that shows a significant bias toward designing of K, N, Q, and especially R residue types. 2) The addition of "Stringent HBonds". 3) The addition of improved "Phosphate Desolvation" in conjunction with the "Stringent HBonds". 4) The addition of the "Electrostatics" model in conjunction with the "Phosphate Desolvation" and "Stringent HBonds". 5) The "Final (”Optimized”)" Energy function.



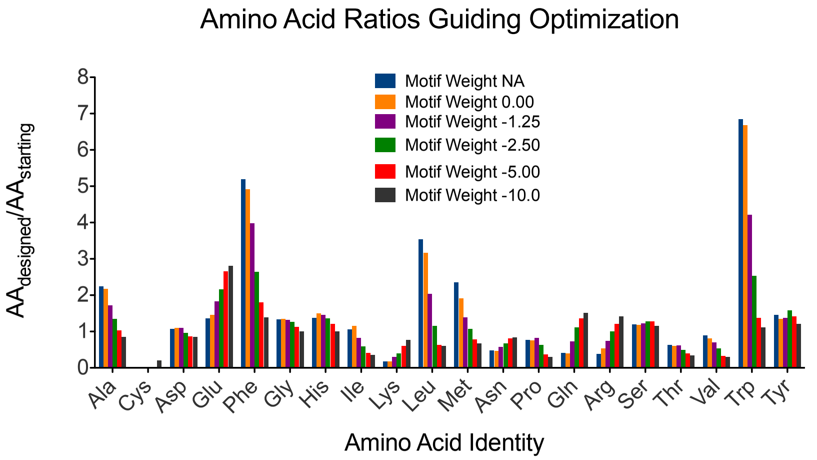
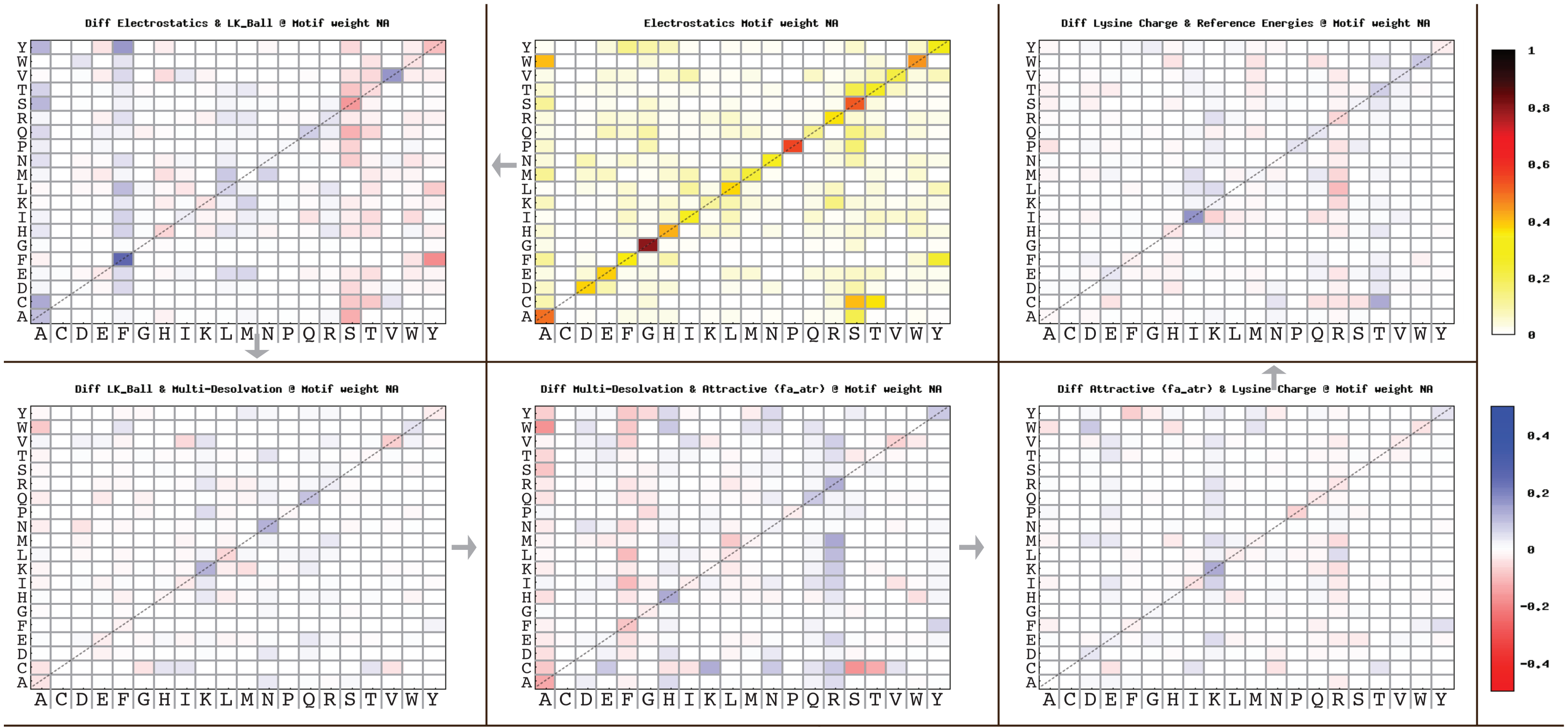


Figure S4. Comparison of the Standard energy function with “Stringent HBonds” and “Phosphate Desolvation” corrections to the same energy function with different motif weights. The heat maps from “Phos. Desolvation Motif Weight NA” heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the “Phos. Desolvation” energy function as the baseline for the differences. The difference heat maps are increasing weights on the motif rotamers in ROSETTA Energy Units or REUs (weight increases in counterclockwise direction). 1) Motif rotamers added with no weight. 2) A weight of -1.25 REUs. 3) A weight of -2.50 REUs. 4) A weight of -5.00 REUs. 5) A weight of -10.0 REUs.



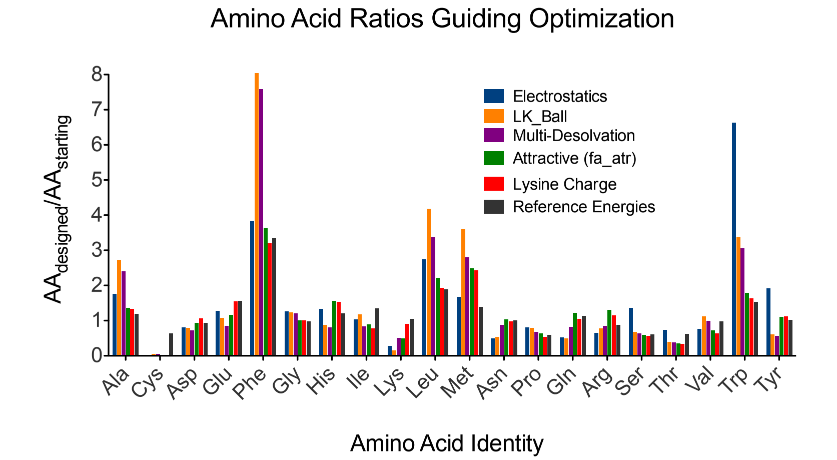


Figure S5. Changes to the energy function over the optimization process, starting at the Electrostatics energy function. The heat maps from "Electrostatics Motif Weight NA" heatmap shows the identity of the wild-type amino acid type on the y-axis and the designed amino acid type on the x-axis. Each of the additional five plots are different heatmaps, where red indicates a loss in frequency and blue is a gain. The arrows indicate that each different plot is calculated using the preceeding energy function as the baseline for the differences. All plots in this figure are for design calculations with no motif rotamers added. The difference heat maps are (counterclockwise) 1) The difference between the "Electrostatics" energy function and the "Electrostatics" energy function with "LK_Ball" added. 2) The addition of "Multi-Desolvation" changes to the #1 energy function. 3) The increased "Attractive (fa_atr)" added to the #2 energy function. 4) The increased "Lysine Charge" added to the #3 energy function. 5) The "Final ("Optimized")" Energy function or the addition of modified "Reference Energies" to the #4 energy function.

Table S2. Summary of I-AniI interface randomization sequencing data.

Pos/AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Tot.
18	2	1														1	2		12	2	20
20	3					13										1	1	3			21
22	2	1				11	1							2		3	1				21
24							12		9												21
25	4		1	1		2				1				1	3	3	2	1			19
26	1					4	1	1	1				2		3	1	4	1	1	1	21
27					1		7			2									4	7	21
29	10					3				2						3	3				21
31	2	2		3				2			2		1				8	1			21
33	4					11										5					20
35	2	1	1	7						1				2		6		1			21
37	5	1		1		6									3	2		3			21
55	1	1		1	1	2		2			2		2		1	1	4	2			20
57	1	2				1					2			1	1	5	4	2		1	20
59									1						20						21
61							1		1						18			1			21
64	1	1			1	1		1		1		1		3	2	4	3	1		1	21
66	3	1								4		1		2	1	1	4	2		2	21
68	2					3			2	1	1				5	2	3	3			22
70										1					20						21
72															21						21
148				19																1	20
150		17														5					22
152	3	1				1		1				1	1		1	4	5			1	19
154		2	1				3									1				14	21
156	2					2			1	1			2	1	3		7	1			20
157	3				1	1	1			3	1	1			5	3	2				21
160	3	1	1			1	2			2		1	2		1	2	1	2	2		21
162	1	1	1		7													1		7	18
164	4	3						1		1	2	1		1		1	3	3			20
166	4		2			1						5				8					20
168		6	8		1	1										5					21
170	6	2				7			1							5					21
172					1	3	1			3		2		1	4	4	1	1			21
189	3			1		4				2			2			4	4	1			21
192	3	3				7	1			2	1				2	1			1		21
194	3	4	2			3	1					1	3			5	1				23
196	4			1		4	5		2	2					5	1	3		1	1	29
198								1		6	1		2				10	6			26
200					1		3		1			2			12			1		2	22
202									20												20
204	1			1		7						1	2			3	1	4			20
205	3	1			1	7	1		1			3	1	1	1	1					21
243															21						21

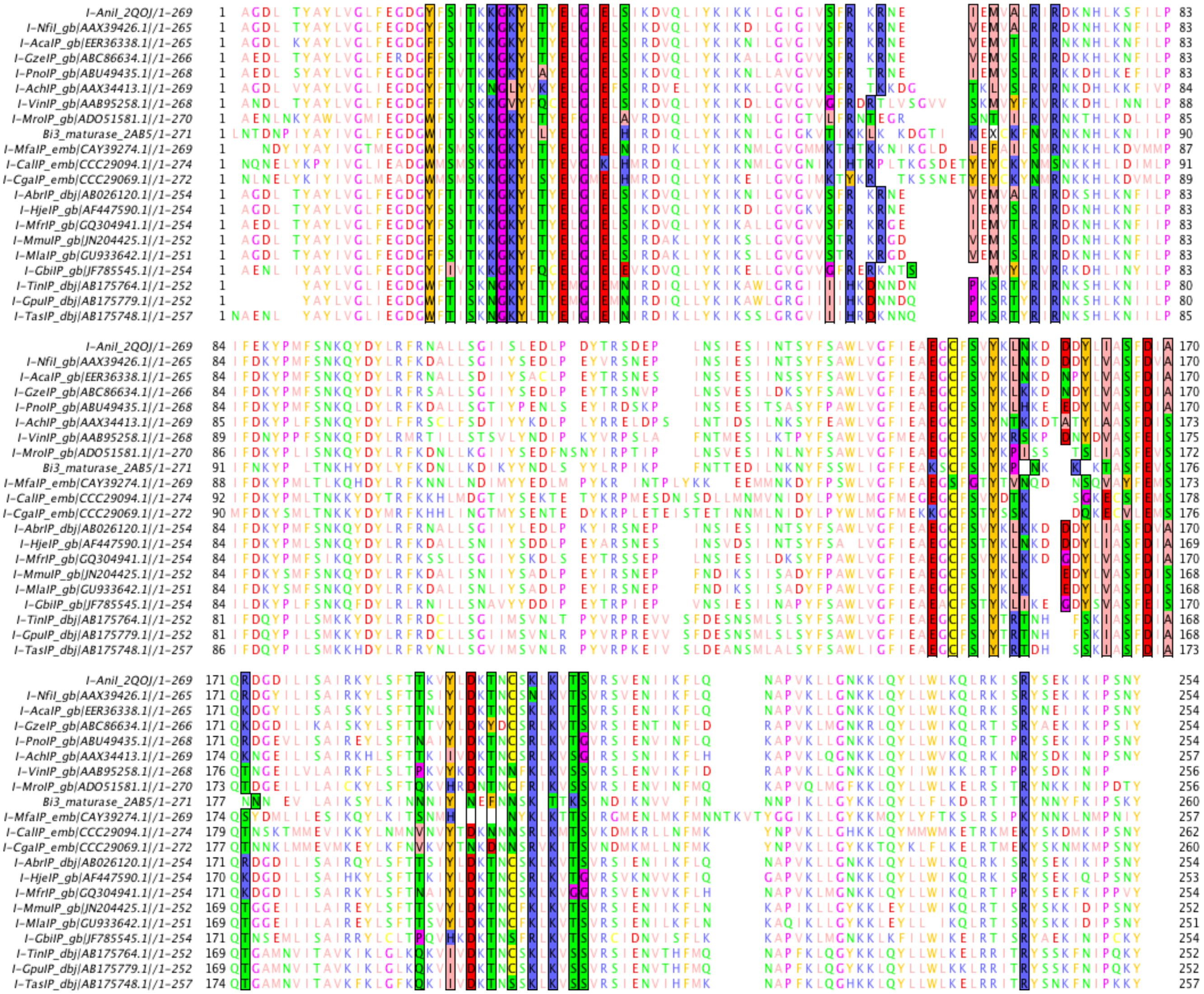

I-AniI_2QOJ/1-269
I-NfiI_gb|AAX39426.1|/1-265
I-AcaIP_gb|EER36338.1|/1-265
I-GzeIP_gb|ABC86634.1|/1-266
I-PnoIP_gb|ABU49435.1|/1-268
I-AchIP_gb|AAX34413.1|/1-269
I-VinIP_gb|AAB95258.1|/1-268
I-MroIP_gb|ADO51581.1|/1-270
Bi3_maturase_2AB5/1-271
I-MfaIP_emb|CAY39274.1|/1-269
I-CalIP_emb|CCC29094.1|/1-274
I-CgaIP_emb|CCC29069.1|/1-272
I-AbrIP_dbj|AB026120.1|/1-254
I-HjeIP_gb|AF447590.1|/1-254
I-MfrIP_gb|GQ304941.1|/1-254
I-MmuIP_gb|JN204425.1|/1-252
I-MlaIP_gb|GU933642.1|/1-251
I-GbiIP_gb|JF785545.1|/1-254
I-TinIP_dbj|AB175764.1|/1-252
I-GpuIP_dbj|AB175779.1|/1-252
I-TasIP_dbj|AB175748.1|/1-257

Figure S6. Alignment of proteins homologous to I-AniI.

Multiple sequence alignment of proteins homologous to I-AniI found using NCBI blastp and tblastn. The most distant homologue in this alignment, I-MfaIP, has 47% sequence identity and 70% similarity. Alignments with less than 40% identity were predicted to have significantly divergent putative target site sequences (data not shown) and were not chosen for examination. The 44 positions that were randomized in the bacterial selection experiment described in this paper are boxed and highlighted in this multiple sequence alignment. Many of these positions show less variability in the alignment compared to the selection results, most likely due to the differences in conditions and selection pressures between the engineering experiment and natural evolution. It is important to point out as well that, while these enzymes most likely have some activity on the native I-AniI target site, it is probable that enzymes with positions in this alignment that have a high abundance of an amino acid type not observed in the selection experiment are likely targeting substrates with single or multiple nucleotide substitutions.

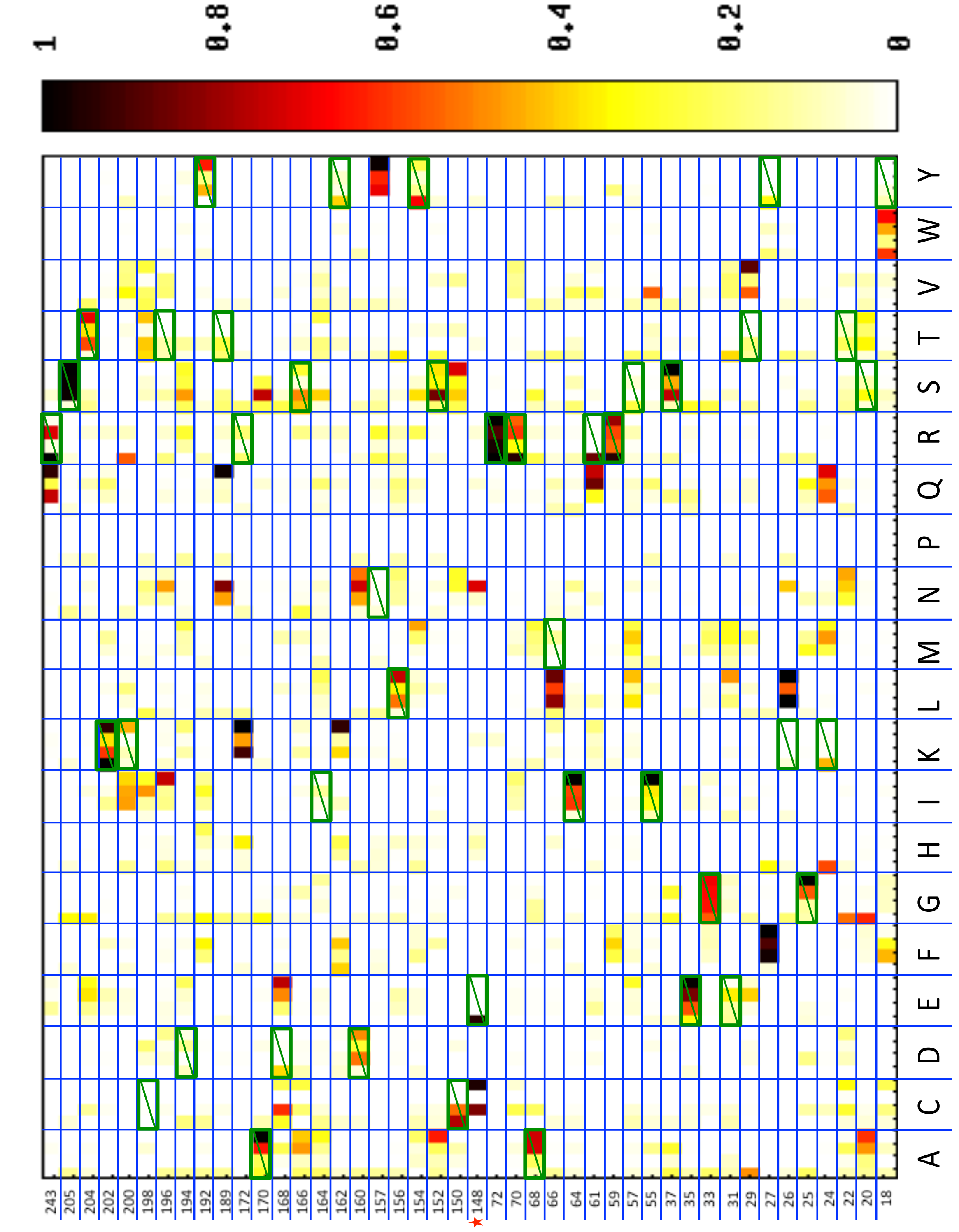


Figure S7. Heatmap comparison of experimental data with predictions from three computational methods. Frequencies of amino acid occurrence for each I-AniI interface position in 1) Experimental data 2) HighTemp-Packer 3) DNA-Rebuild 4) Standard with "Optimized" energy function. Wild-type is boxed in green and the order in the boxes is 1-4 from left to right. Each computational protocol was run 56 times.

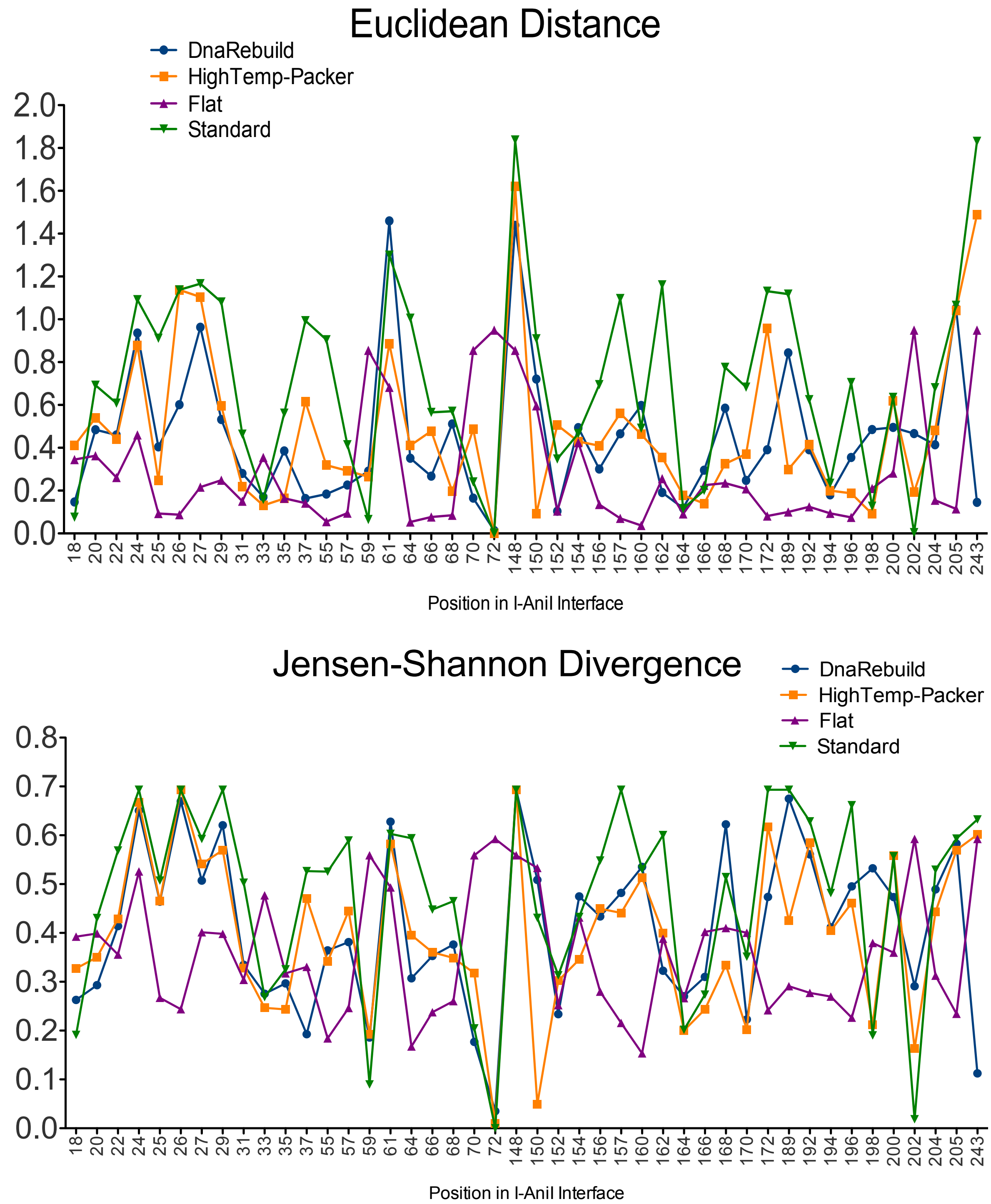


Figure S8. Comparison of experimental data with predictions from three computational methods using Euclidean distance and Jensen-Shannon divergence. Positions in the I-AniI interface that were randomized and screened for sequence tolerances are shown on the X-axis. The divergence values for comparisons between the predicted distributions and the experimentally derived distributions were calculated for each randomized position. Lower values of divergence indicate a better match between the calculated and experimental distribution. The "Flat" distribution is a distribution with 0.05 at every position, showing that the computational prediction at many positions is still not displaying as high of sequence diversity as the experimental data.

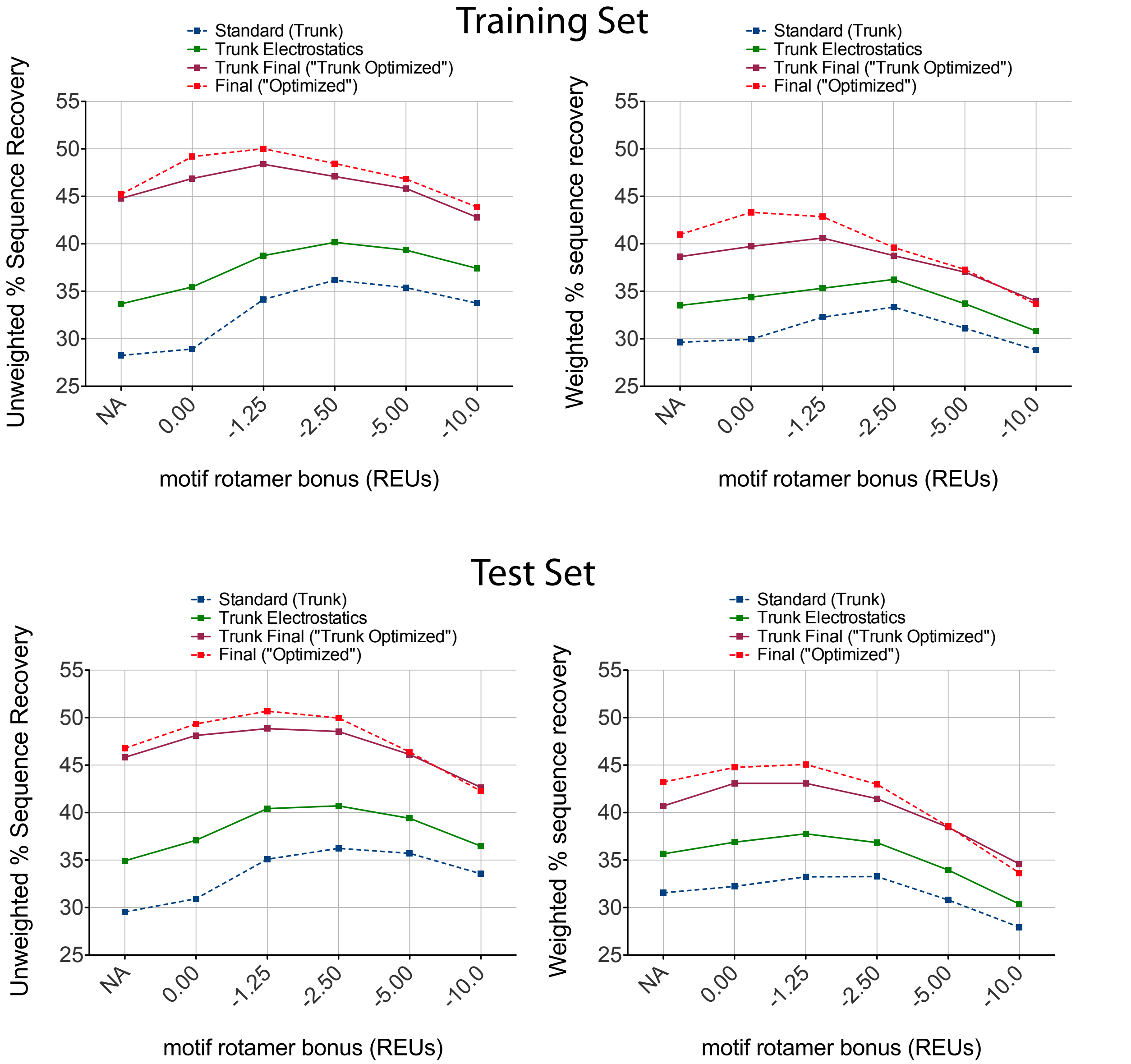


Figure S9. Corresponding energy function optimizations in the Trunk version of ROSETTA. The optimizations to the ROSETTA energy function discussed in this work were completed in a branch of ROSETTA with a focus on protein-DNA interactions. The ROSETTA energy function was additionally optimized in the context of the "Trunk" or main version of ROSETTA. The stringent hydrogen bonds and orientation-dependent desolvation model were not available in the "Trunk" version when these calculations were completed. The differences between the "Trunk Optimized" and "Final ("Optimized")" energy functions are the orientation-dependent desolvation, the stringent hydrogen bonds, and differences in the amino acid specific references energies to account for the missing terms. The dashed lines indicate energy functions discussed in the main text of the paper, included for comparison to the trunk optimizations.