**Supporting Information for:**

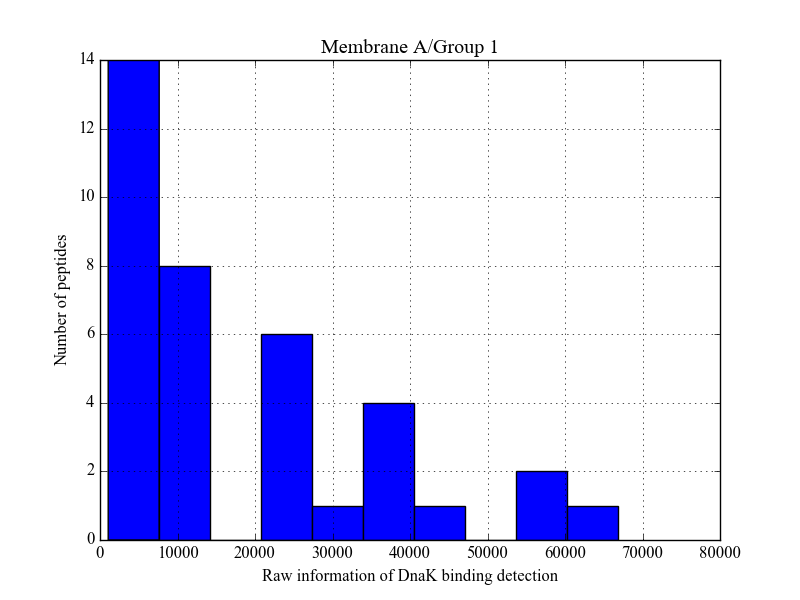
**ChaperISM: improved chaperone binding prediction using position-independent scoring matrices**

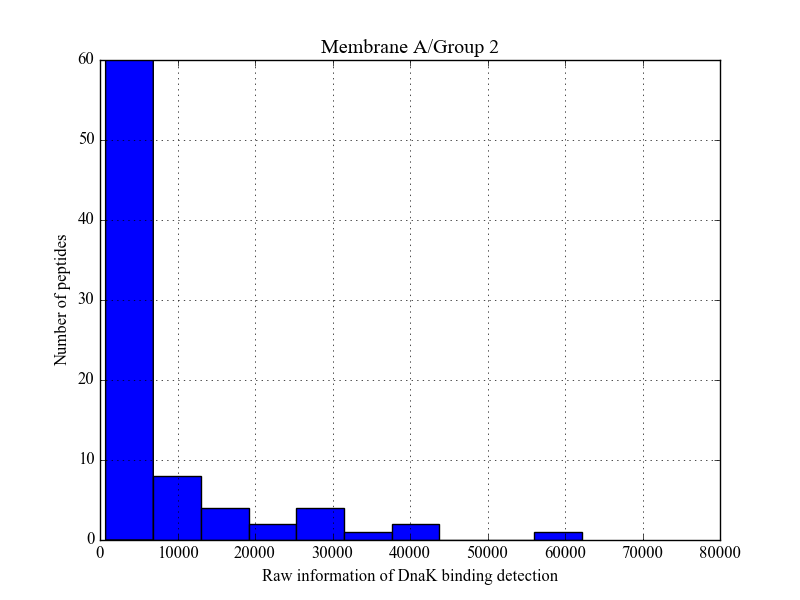
*M.B.B. Gutierres¹, C.B.C. Bonorino2,3, M. M. Rigo4,\*,*

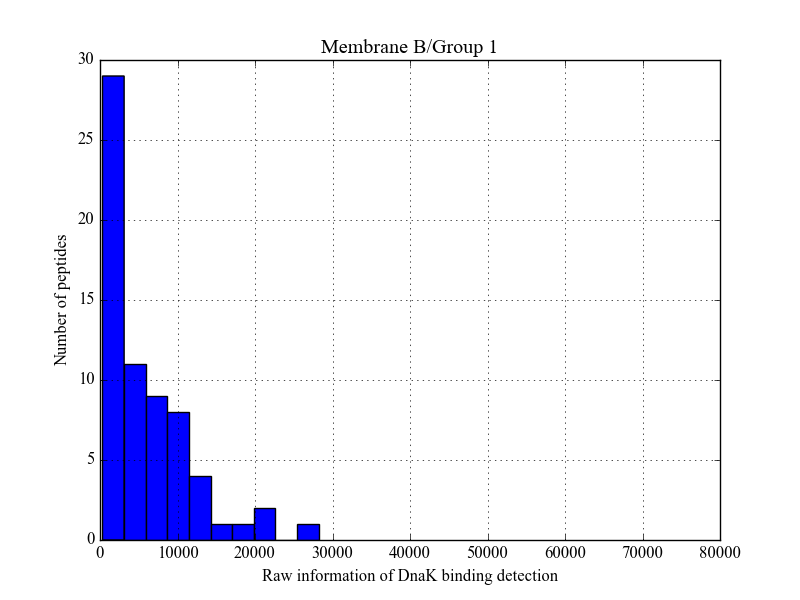
1 Escola de Ciências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil, 2 Laboratório de Imunoterapia, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre, Brazil, 3School of Medicine, University of Califórnia San Diego, La Jolla CA 92037, 4 Escola de Medicina, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil

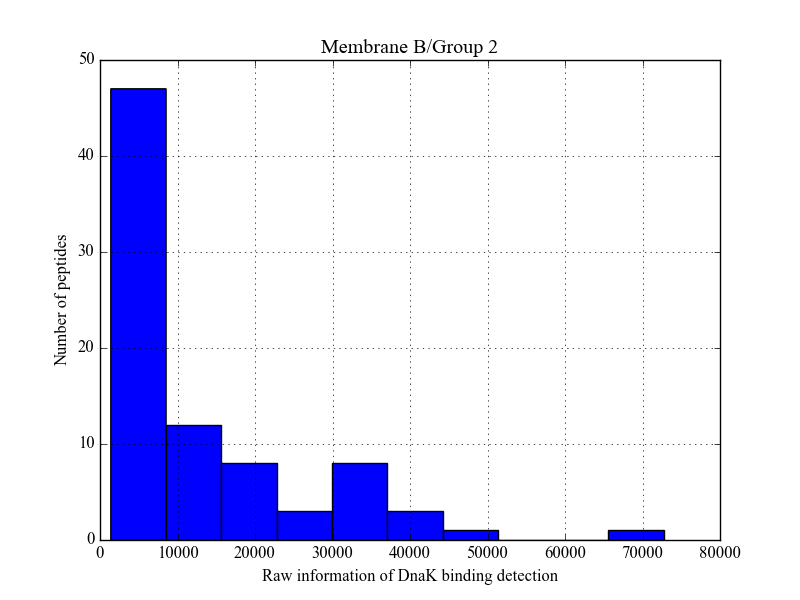
\*To whom correspondence should be addressed.

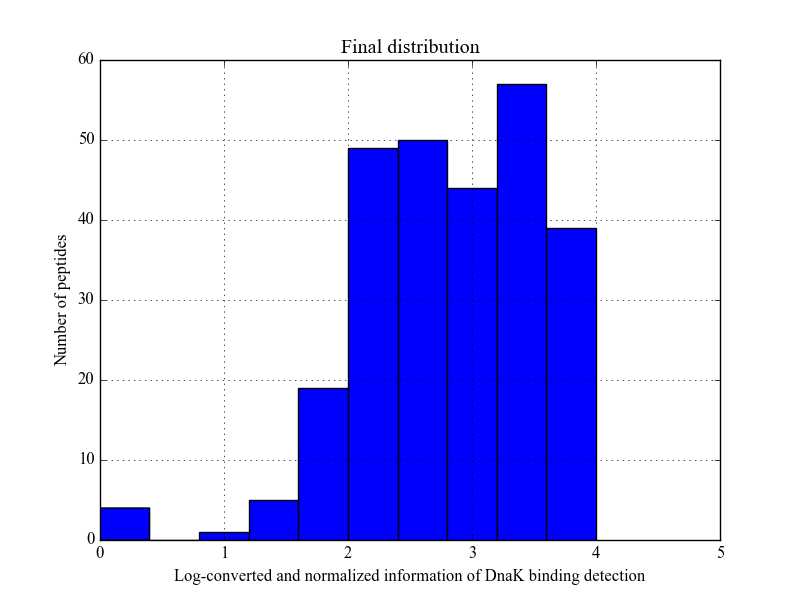
Email: mauricio.rigo@pucrs.br

**Supplementary Figure S1.** Distribution of Membrane A/Group 1 peptides according to raw DnaK detection data.

**Supplementary Figure S2.** Distribution of Membrane A/Group 2 peptides according to raw DnaK detection data.

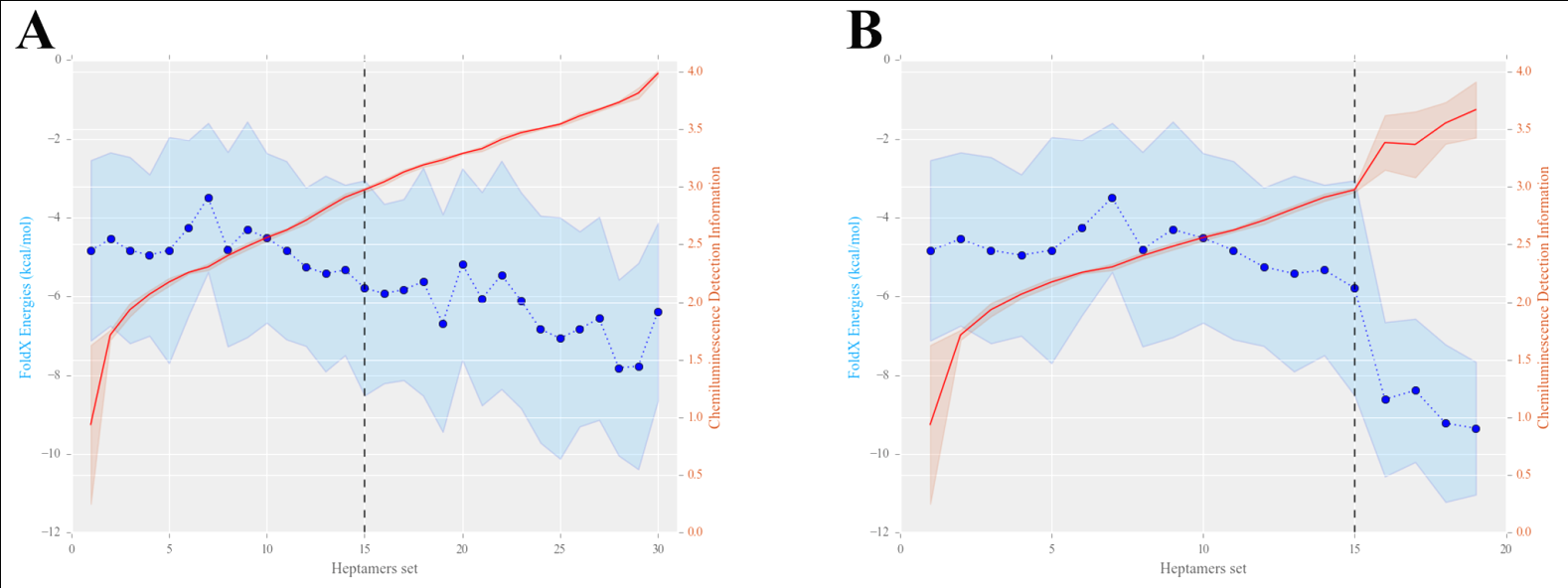
**Supplementary Figure S3.** Distribution of Membrane B/Group 1 peptides according to raw DnaK detection data.

**Supplementary Figure S4.** Distribution of Membrane B/Group 2 peptides according to raw DnaK detection data.

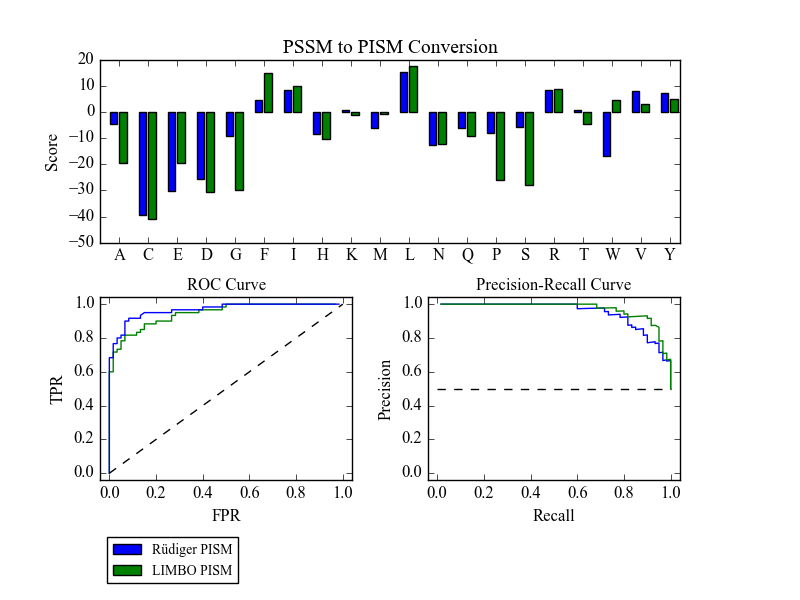
**Supplementary Figures S5.** Distribution for all 268 peptides from figures S1-S4 after preprocessing.

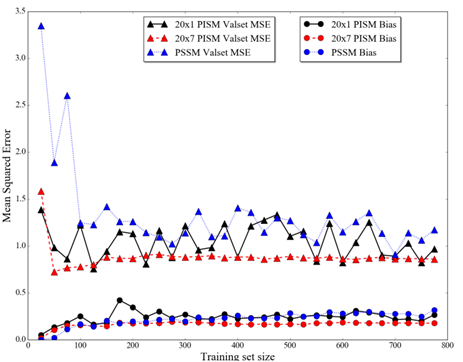


**Supplementary Figure S6.** Workflow of normalization, preprocessing and force-field based filtering steps from raw data. The 268 peptides depicted in Supplementary Figures S1-S4 were normalized as explained in *Material & Methods* section, split, and if the peptide source was considered as a binder (above 3.0) subjected to an energy filter. Energy and DnaK binding score inverse correlation shown in Figures 2A and 2B are also indicated.



**Supplementary Figure S7.** The normalized DnaK binding score and the interaction energies with DnaK SBD are shown for 1,488 (A) and 892 (B) heptamers. Data were grouped for every 50 heptamers and the mean values were plotted as a red solid line (DnaK binding score) and as blue points (interaction energy computed with FoldX force field). The shadowed area represents standard deviation (+-1 SD).

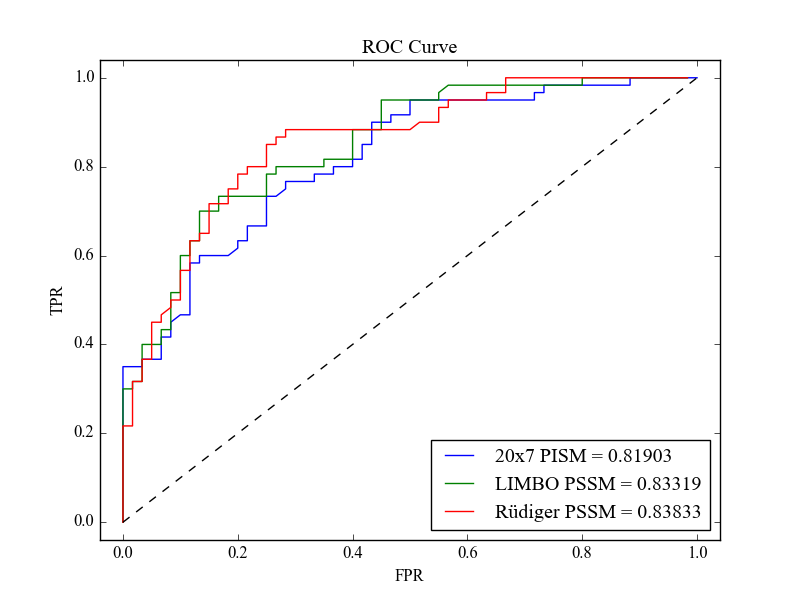
**Supplementary Figure S8.** Matrices from Rüdiger and LIMBO were converted from PSSM to PISM. Despite the simplification, this modification does not causes a decrease in performance evaluated through ROC and PR analysis in the validation set defined in manuscript main text. In addiction, both PISMs have a similar structure. Rüdiger and LIMBO PISMs are shown in blue and green, respectively. ROC and PR analysis solid line describes the PISM-converted predictive performance.



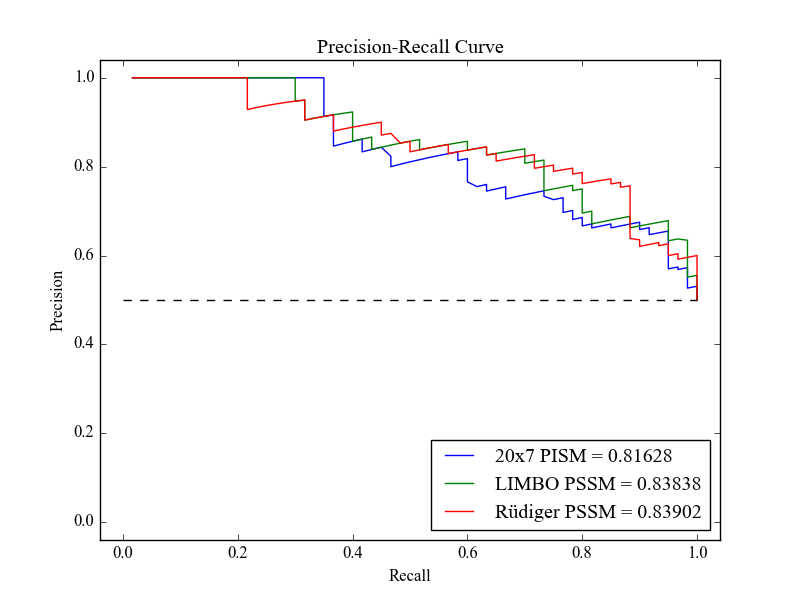
**Supplementary Figure S9.** Estimation of bias (MSE from training set, circles) and validation set MSE (triangles). Each point represents a classifier trained on a progressively increasing training set. In each step, twenty-five randomly selected heptamers are added to the training set.

**Performance of models trained without normalization step in data preprocessing (Figure S10—S11)**

***Contextualization:*** One step in the preprocessing of the 4 experiments consists in normalizing different experiments to the range of 1-10,000. This range was chosen after analysis of Supplementary Figures S1-S4. As a result, from this range distribution, the conversion to log10 leads to a new range of 0-4. Normalization, however, could change how true binders and true non-binders are distributed among the other experiments. For instance, the best interactor in Membrane B/Group 1 possess raw detection data of 28,208.86, while in Membrane A/Group 2 is 72,729.74. Normalizing to 1-10,000 independently for each experiment causes both peptides to possess an equal DnaK binding score of 10,000. Thus, we performed the same steps from workflow in Supplementary Figure S6 only removing normalization to evaluate this effect. Raw information from the 268 peptides was directly log10 converted. Minimum and maximum DnaK detection data after this were, respectively, 2.5118 and 4.8617. Selected threshold for energy filtering was chosen as 3.7, to similarly separate the same amount of binders and non-binders as the original threshold of 3.0. An independent validation set of 120 examples was extracted similarly as described in manuscript main text. Herein, we show ROC and PR analysis to one model derived from 20x7 PISM algorithm.



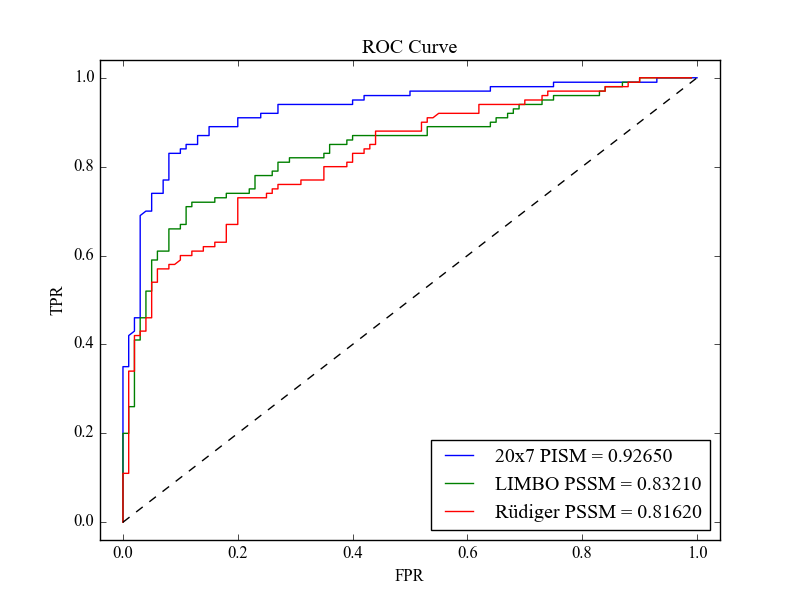
**Supplementary Figure S10.** ROC curve built upon validation set extracted without performing normalization. Area under the curves for each predictor is shown. Note both PSSMs from Rüdiger (red) and LIMBO (green) shown a decrease in performance in comparison to evaluation with validation set defined in manuscript main text (with normalization). Despite a reasonable performance, the 20x7 PISM (blue) does not outperform these matrices. Dashed line describes the behavior of a random predictor.

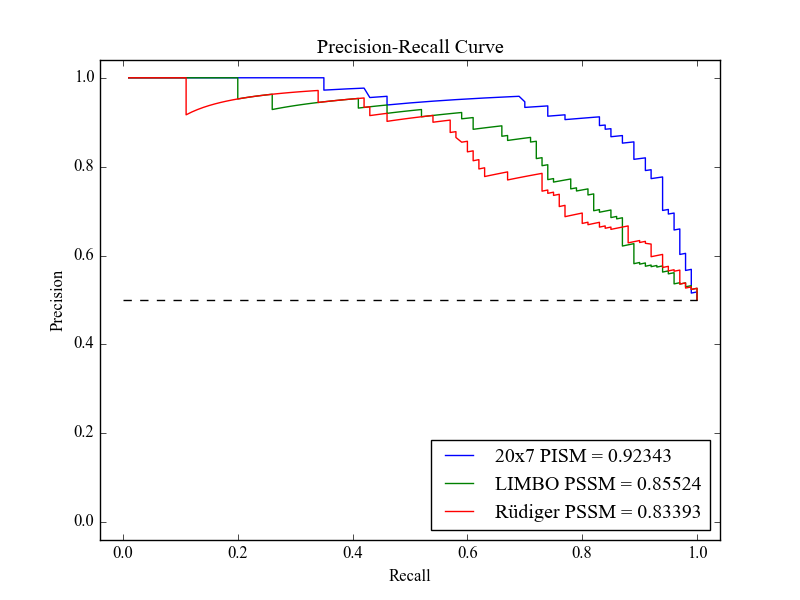
**Supplementary Figure S11.** PR curve built upon validation set extracted without performing normalization. Area under the curves for each predictor is shown. Dashed line describes the behavior of a random predictor. Area under curve is shown for PSSMs from Rüdiger (red), LIMBO (green) and 20x7 PISM (blue).

**Performance of models trained without energy filtering step in data preprocessing (S12-S15)**

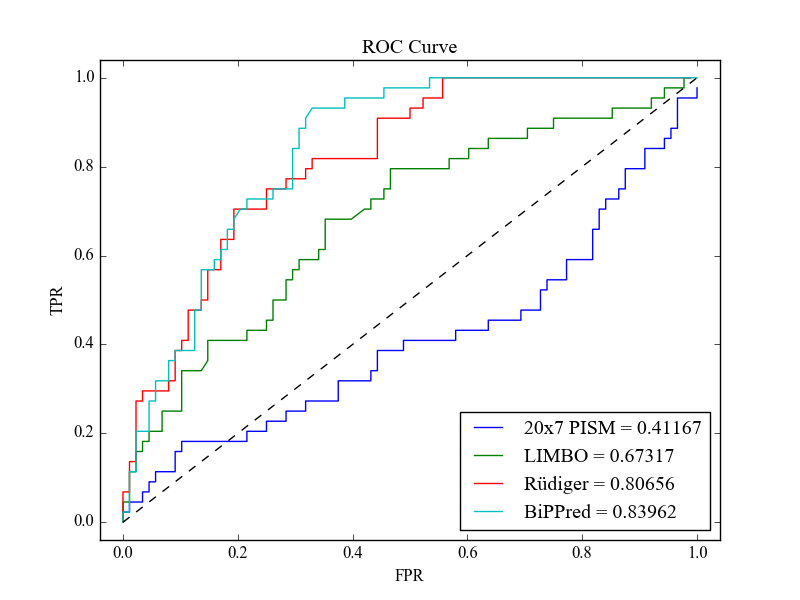
***Contextualization:*** Another preprocessing step we evaluated was the energy filtering (See *Material and Methods* in main text). This step removed 574 heptamers from the 1,488 total. After merging duplicates, the final available data to compose both training and validation steps consisted of 892 heptamers. Without performing this step, there is more available data to compose larger sets of training and validation. Merging duplicates leads to 1,364 heptamers (from the 1,488 total). As more data was available, we extracted a balanced validation set of 200 examples. 100 binders were randomly selected from the 200 top scored, while 100 non-binders were randomly extracted from the 200 least scored heptamers. Training set was composed of 1,164 examples. Herein, we show the performance of a model trained employing the 20x7 PISM algorithm in the validation set defined here, and for CD dataset obtained from Schneider *et al*.

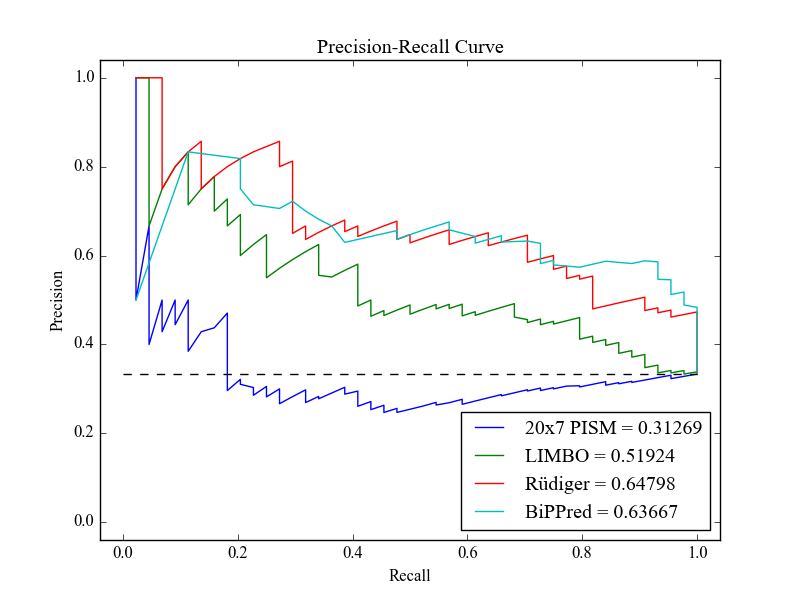
**A 20x7 PISM outperforms Rüdiger and LIMBO PSSMs on the validation set defined in this section**

**Supplementary Figure S12.** ROC curve built upon validation set extracted without performing energy filtering. Area under the curves for each predictor is shown. Dashed line describes the behavior of a random predictor. Area under curve is shown for PSSMs from Rüdiger (red), LIMBO (green) and 20x7 PISM (blue).

**Supplementary Figure S13.** PR curve built upon validation set extracted without performing energy filtering. Area under the curves for each predictor is shown. Dashed line describes the behavior of a random predictor. Area under curve is shown for PSSMs from Rüdiger (red), LIMBO (green) and 20x7 PISM (blue).

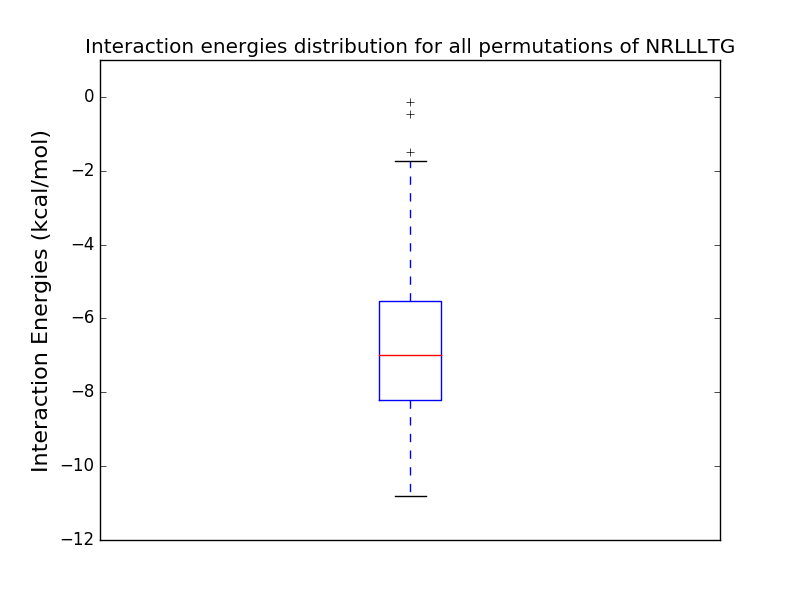
**Both Rüdiger and LIMBO PSSMs outperform a 20x7 PISM on the BiP related CD dataset**

**Supplementary Figure S14.** ROC curve built upon CD dataset. Area under the curves for each predictor is shown. Dashed line describes the behavior of a random predictor. Area under curve is shown for PSSMs from Rüdiger (red), LIMBO (green), 20x7 PISM trained without energy filtering step in preprocessing (blue) and BiPPred’s maximum predicted score (light green).

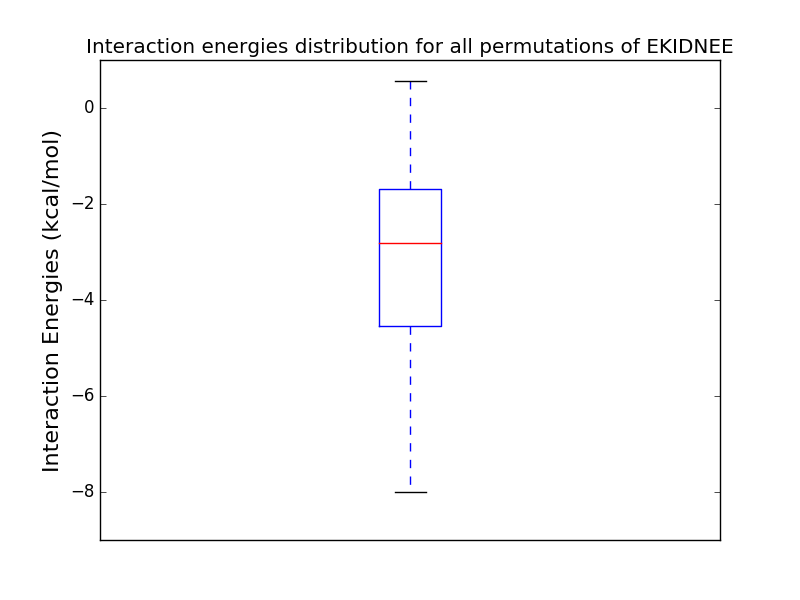


**Supplementary Figure S15.** PR curve built upon CD dataset. Area under the curves for each predictor is shown. Dashed line describes the behavior of a random predictor. Area under curve is shown for PSSMs from Rüdiger (red), LIMBO (green), 20x7 PISM trained without energy filtering step in preprocessing (blue) and BiPPred maximum predicted score (light green).

**Analysis of interaction energies for all possible permutations of a true binder and a true non-binder**

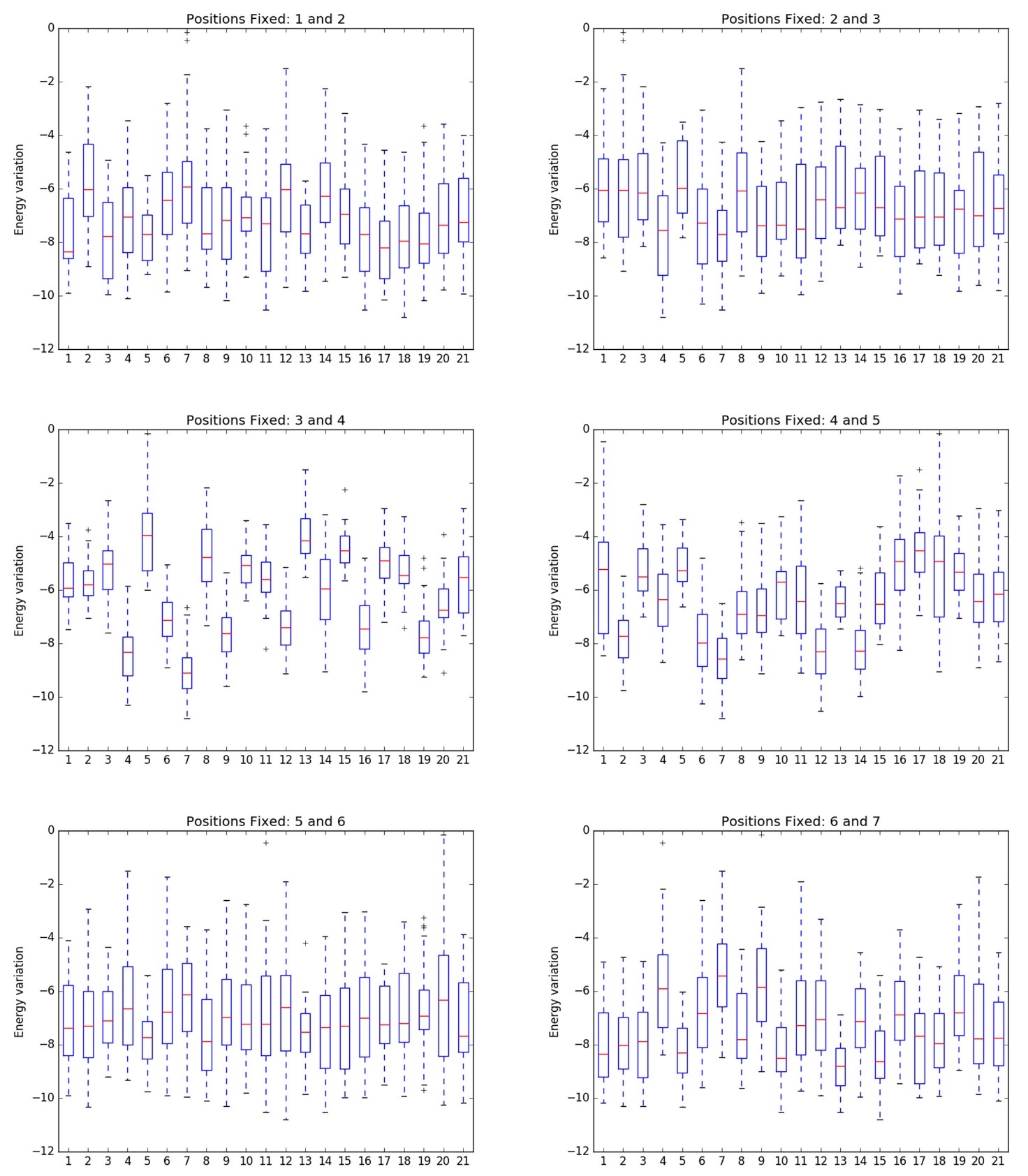
***Contextualization***: As a PISM does not account different scores for the different positions in the heptamer, it is expected that the interaction energies of all heptamers must be similar in order to explain the PISM superior performance over a regular PSSM. Therefore, we selected a true binder (NRLLLTG) and a true non-binder (EKIDNEE), performed its mutations in the crystal structure of DnaK bound to a heptamer to each possible permutation and computed the energy with FoldX. Figures below highlight that the variation is larger than expected.

**Supplementary Figure S1****6.** Interaction energies of all permutations of NRLLLTG computed with FoldX force field based on mutations of the crystal structure of DnaK SBD bound to NRLLLTG heptamer (PDB ID.: 1DKX).

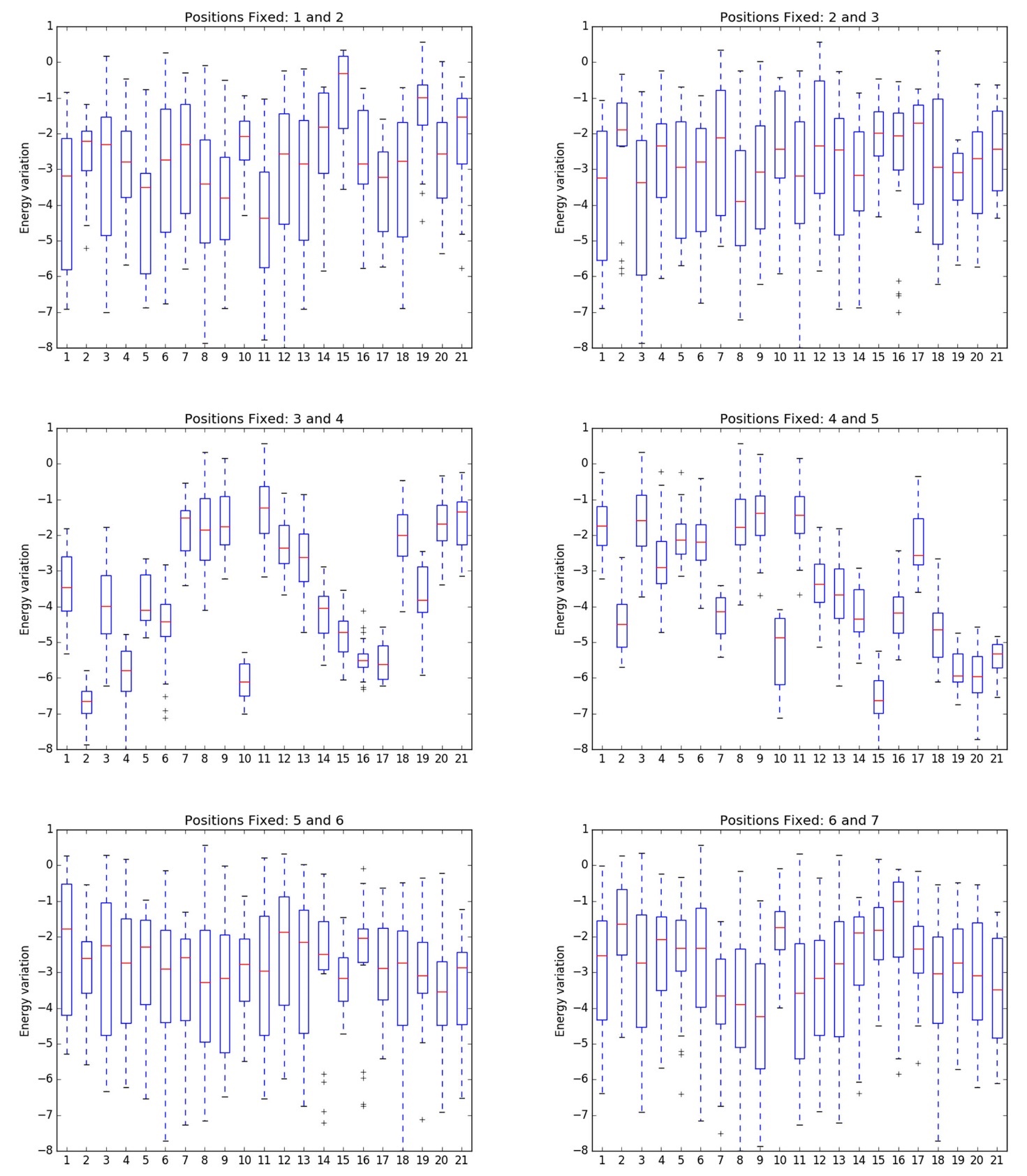
****

**Supplementary Figure S17.** Interaction energies of all permutations of EKIDNEE computed with FoldX force field based on mutations of the crystal structure of DnaK SBD bound to a heptamer (PDB ID.: 1DKX).

Next, to investigate why it is possible that a PISM could outperform other PSSMs, we focused our analysis to sub-permutations. Thus, we grouped pairs of adjacent residues (e.g. NR-LTGLL, NR-LLGTL, NR-GLLLT …), and plotted the interaction energy variation in this context. The figures below indicate that variation is extremely reduced when residues fixed are in positions 4 and 5. This reveals that the PISM superior performance is most likely due to the fact that pockets 1, 2, 3, 6 and 7 show a character more ‘position-independent’ than the central pockets.

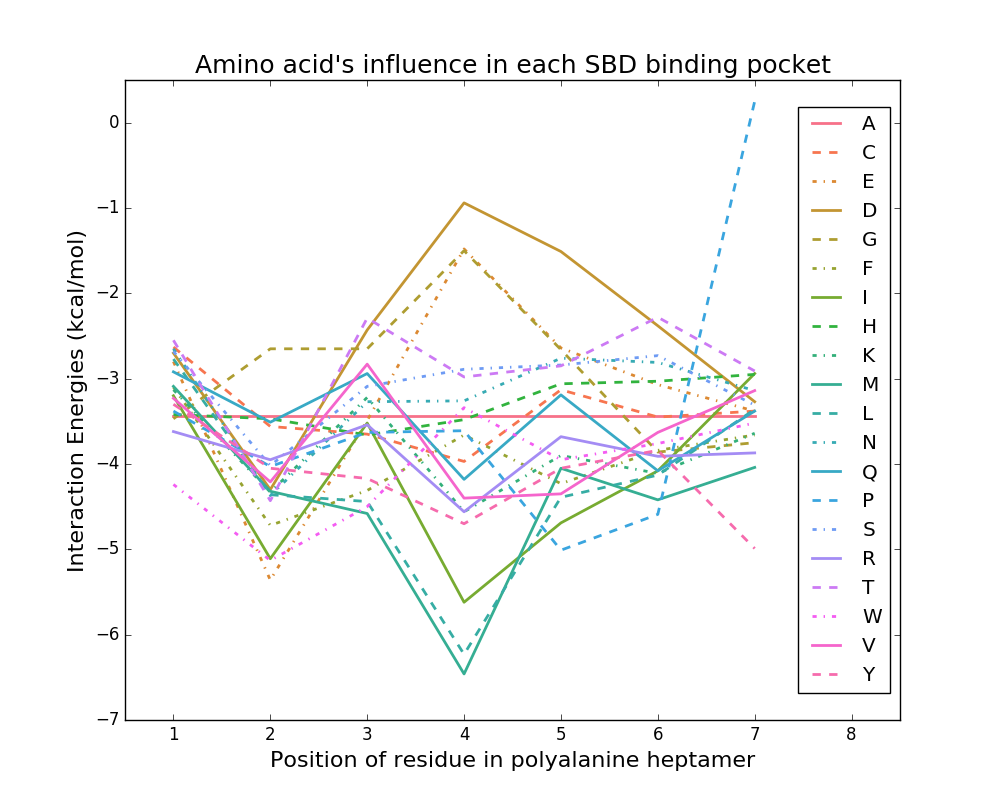


**Supplementary Figure S18.** Pairs of adjacent amino acids were grouped for the pool of NRLLLTG permutations, and the DnaK interaction energies were plotted. Positions 1 to 21 refers to residues RT, LG, GT, NL, GR, LN, LL, TR, TL, NG, TN, LT, RG, LR, TG, RL, GN, RN, GL, NT, NR.

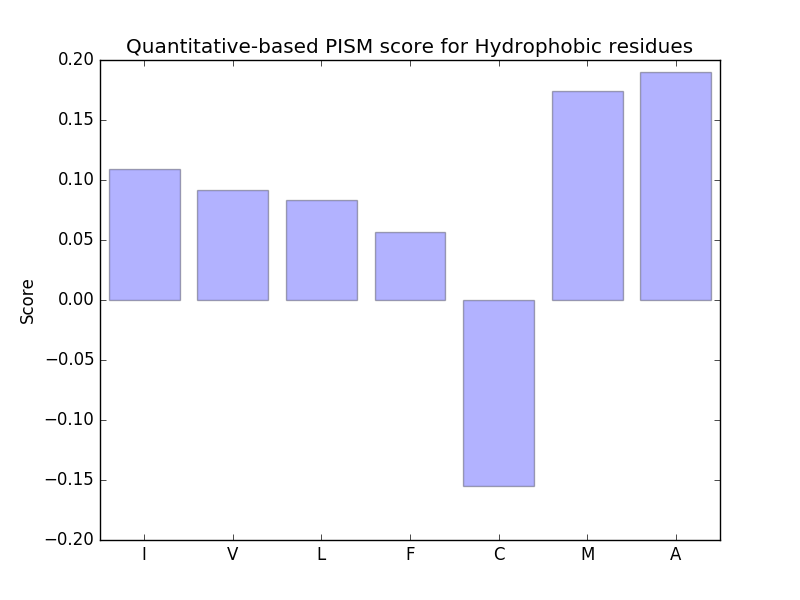


**Supplementary Figure S19.** Pairs of adjacent amino acids were grouped for the pool of EKIDNEE permutations, and the DnaK interaction energies were plotted. Positions 1 to 21 refers to residues DN, NI, EN, EI, DK, EK, KD, EE, ED, KI, DE, ND, NE, KN, IK, NK, DI, KE, IN, IE, ID.

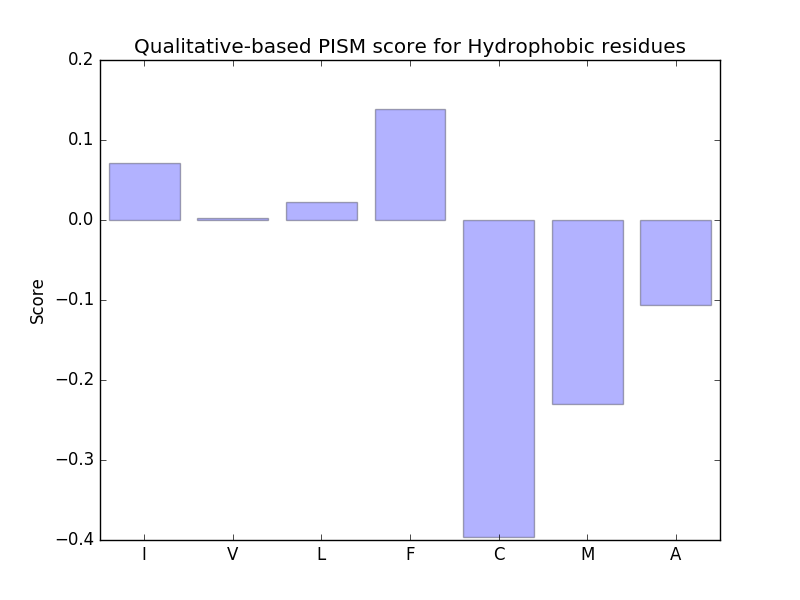
Finally, in agreement to pockets 4 and 5 being more relevant to the interaction energy of the whole heptamer, we performed a mutation scanning of a polyalanine heptamer in DnaK SBD. We mutated each position to the remaining amino acids and plotted the interaction energies for all positions.



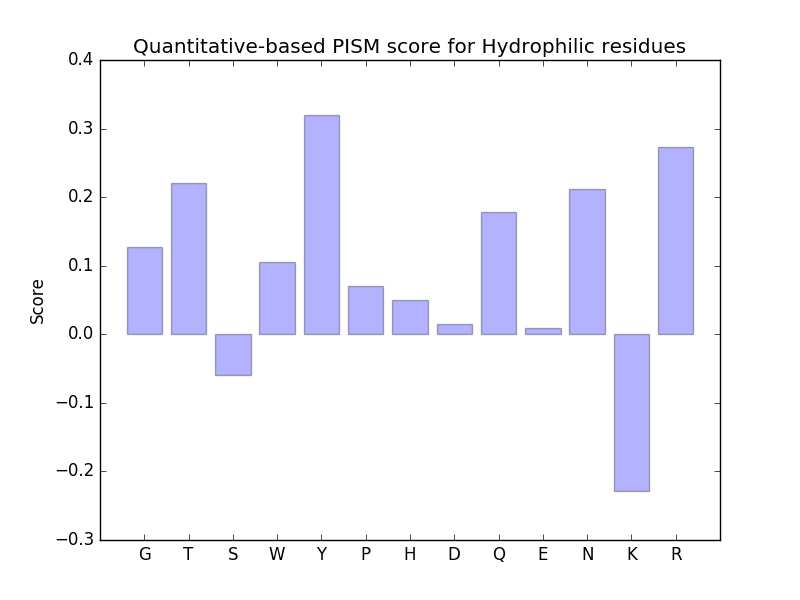
**Supplementary Figure S20.** The crystal structure of NRLLTG peptide bound to DnaK SBD (PDB ID.: 1DKX) was mutated to a polyalanine heptamer. Next, all positions were sequentially mutated to the other 19 amino acids. The interaction energies were computed with FoldX force field.

****

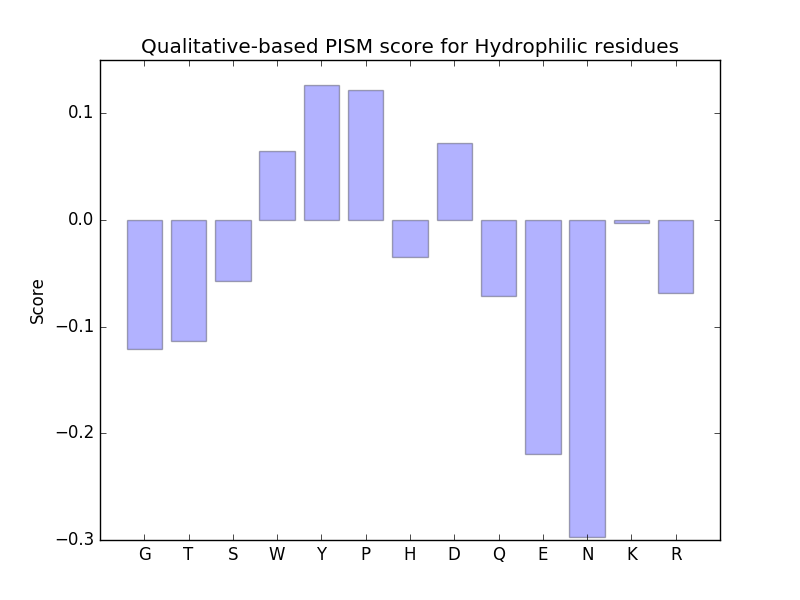
**Supplementary Figure S21.** The score of hydrophobic amino acids sorted by Kyte and Doolitle hydrophobicity scale according to quantitative-based 20x7 PISM.

****

**Supplementary Figure S22.** The core of hydrophobic amino acids sorted by Kyte and Doolitle hydrophobicity scale according to qualitative-based 20x7 PISM.

****

**Supplementary Figure S23.** The score of hydrophilic amino acids sorted by Kyte and Doolitle hydrophobicity scale according to quantitative-based 20x7 PISM.

****

**Supplementary Figure S24.** The score of hydrophilic amino acids sorted by Kyte and Doolitle hydrophobicity scale according to qualitative-based 20x7 PISM.

**Supplementary Tables**

**Supplementary Table 1.** Description of available DnaK peptide binding data from Van Durme *et al.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Membrane/ Group** | **Number of Peptides** | **Min. Length** | **Max. Length** | **Min. Raw Data** | **Max. Raw Data** | **Avg Raw Data** | **Min. Converted Data** | **Max. Converted Data** | **Avg Converted Data** |
| A/G1 | 37 | 10 | 15 | 1,056.56 | 66,870.97 | 18,839.39 | 1 | 10,000 | 2,702.69 |
| A/G2 | 82 | 10 | 20 | 724.37 | 62,212.7 | 7,674.09 | 1 | 10,000 | 1,131.14 |
| B/G1 | 66 | 7 | 14 | 325.01 | 28,208.86 | 5,828.84 | 1 | 10,000 | 1,974.65 |
| B/G2 | 83 | 10 | 10 | 1,393.81 | 72,729.74 | 12,623.25 | 1 | 10,000 | 1,575.01 |

**Supplementary Table 2.** Performance in the independent validation set defined in manuscript main text for models generated from five independents rounds of training with all six algorithms.

|  |  |  |
| --- | --- | --- |
| **PSSM - Qualitative information** | **auROC** | **auPR** |
| Model 1 | 0.870 | 0.894 |
| Model 2 | 0.829 | 0.856 |
| Model 3 | 0.824 | 0.830 |
| Model 4 | 0.696 | 0.688 |
| Model 5 | 0.791 | 0.810 |
| **PSSM - Quantitative information** | **auROC** | **auPR** |
| Model 1 | 0.818 | 0.825 |
| Model 2 | 0.718 | 0.763 |
| Model 3 | 0.768 | 0.735 |
| Model 4 | 0.783 | 0.778 |
| Model 5 | 0.758 | 0.747 |
| **PISM - Qualitative information** | **auROC** | **auPR** |
| Model 1 | 0.664 | 0.674 |
| Model 2 | 0.898 | 0.912 |
| Model 3 | 0.926 | 0.933 |
| Model 4 | 0.946 | 0.956 |
| Model 5 | 0.940 | 0.939 |
| **PISM - Quantitative information** | **auROC** | **auPR** |
| Model 1 | 0.801 | 0.793 |
| Model 2 | 0.934 | 0.925 |
| Model 3 | 0.784 | 0.806 |
| Model 4 | 0.925 | 0.913 |
| Model 5 | 0.924 | 0.923 |
| **PISM (20x7) - Qualitative information** | **auROC** | **auPR** |
| Model 1 | 0.973 | 0.974 |
| Model 2 | 0.972 | 0.974 |
| Model 3 | 0.972 | 0.974 |
| Model 4 | 0.972 | 0.974 |
| Model 5 | 0.972 | 0.974 |
| **PISM (20x7) - Quantitative information** | **auROC** | **auPR** |
| Model 1 | 0.982 | 0.967 |
| Model 2 | 0.982 | 0.968 |
| Model 3 | 0.983 | 0.970 |
| Model 4 | 0.982 | 0.969 |
| Model 5 | 0.982 | 0.969 |

***Datasets and Associated codes***

**Supplementary Table 3.** Internal evaluation set.

|  |  |  |  |
| --- | --- | --- | --- |
| **True Positives** | | **True Negatives** | |
| **Heptamer** | **Normalized Score** | **Heptamer** | **Normalized Score** |
| HILLFTR | 3,511519714 | IMQKEQC | 1,751243833 |
| DLLIVPF | 3,654485032 | VAGTLTN | 0 |
| TVHNPII | 3,425429496 | EKIDNEE | 0 |
| AALLSPY | 3,567143371 | MAGAAAA | 1,675798196 |
| KLVFFAE | 3,661326413 | NTLVDVT | 1,775518858 |
| IRAIRAK | 3,610918029 | LDDATPE | 0 |
| EKYLYII | 3,506734632 | KIDNEEV | 0 |
| ALLAVSL | 3,592400989 | LENGHVV | 1,668333133 |
| LYWLGRK | 3,709111333 | VLIMSES | 1,165210652 |
| LFIINLH | 3,518556963 | EEVLIMS | 1,165210652 |
| QSLLLAL | 3,731889472 | TLVDVTA | 1,775518858 |
| LYFVLDF | 3,676453226 | LVDVTAD | 1,775518858 |
| LLKNKSD | 3,394746561 | GIPEMIP | 1,833501684 |
| SITLFII | 3,518556963 | GNTLVDV | 1,775518858 |
| HILLFIR | 3,660630544 | MKIEEIS | 1,763702146 |
| WPKTKLR | 3,640307756 | IDNEEVL | 0 |
| GNRLLTG | 3,376461844 | MFRVELE | 1,311808169 |
| ERRLARG | 3,773782387 | VGGLGAA | 1,675798196 |
| YRGMINK | 3,521151914 | STVIIEA | 1,689732383 |
| HILLFSR | 3,447478986 | LQGDSDA | 1,687157027 |
| PLVIVTA | 3,350586045 | TLPNTMF | 1,650779866 |
| AFFTIAK | 3,938085754 | PEKLENE | 1,477440775 |
| SLLLLLG | 3,801363908 | IPDIQVE | 1,693750851 |
| IINLHRS | 3,518556963 | LETLPNT | 1,775770109 |
| SLLLLFA | 3,563869677 | RSELLLE | 1,354751509 |
| HILLFVR | 3,497134301 | AGAAAAG | 1,675798196 |
| HILLFRR | 3,649090441 | LENEIEV | 1,477440775 |
| WIVMRQN | 3,725304385 | VDVTADH | 1,775518858 |
| AALLAVS | 3,592400989 | IMEGARD | 1,678356642 |
| DEIYYIS | 3,452509343 | KHMAGAA | 1,675798196 |
| HILLFFR | 3,684748511 | VLETLPN | 1,775770109 |
| NLYLVLH | 3,439526679 | GAAAAGA | 1,675798196 |
| KGRIVFR | 3,834998851 | NMKHMAG | 1,675798196 |
| RLLLLLG | 3,851552662 | GHVVTAH | 1,668333133 |
| DKLLLFT | 3,610676292 | NTMFRVE | 1,650779866 |
| LLLLFAL | 3,563869677 | GTVLETL | 1,775770109 |
| LKNKSDN | 3,394746561 | ELQGDSD | 1,687157027 |
| AFFSIGV | 3,935443438 | EGNTLVD | 1,775518858 |
| GRIVFRS | 3,834998851 | ENGHVVT | 1,668333133 |
| LYWLNPR | 3,547588346 | DVTADHA | 1,775518858 |
| NFALEIL | 3,43227955 | AGTLTNK | 0 |
| SLLLALR | 3,731889472 | FIMEGAR | 1,678356642 |
| LILSHLR | 4 | KLENEIE | 1,477440775 |
| DRVIVKR | 3,673321884 | VLEPTGP | 1,450937305 |
| IIELGQR | 3,585774357 | RVELENG | 1,311808169 |
| **True Positives** | | **True Negatives** | |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| QKLVFFA | 3,579513835 | PDIQVEA | 1,693750851 |
| YKLTKNI | 3,604380787 | NIEMQGT | 1,244947169 |
| TQHLTPS | 3,397645565 | DATPEKL | 0 |
| EKLLRNF | 3,782072752 | MKHMAGA | 1,675798196 |
| HSFINRL | 3,661197618 | MEGARDG | 1,678356642 |
| QKLIVLG | 3,731817171 | DNEEVLI | 0 |
| HILLFKR | 3,802235048 | AAAGAVV | 1,675798196 |
| HILLFAR | 3,439874078 | SDMIVAG | 0 |
| IIKIFFI | 3,749989277 | EHMKIEE | 1,763702146 |
| GKLAKKH | 3,508689765 | MIVAGTL | 0 |
| TLVIVRA | 3,492958501 | AFIMEGA | 1,678356642 |
| FLLIMRE | 3,539354667 | IELQGDS | 1,687157027 |
| KLLLFTA | 3,789982713 | AAAAGAV | 1,675798196 |
| PLMVKVL | 3,896366999 | EMQGTVL | 1,244947169 |
| RLARGLK | 3,773782387 | AAGAVVG | 1,675798196 |

**Supplementary Table 4.** Training set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Heptamer** | **Normalized Score** | **Heptamer** | **Normalized Score** |
| VKRKEVE | 3,322304665 | KVTVELT | 2,38993065 |
| LKLNYPE | 3,345576219 | TVELTPY | 2,38993065 |
| RGLKLNY | 3,345576219 | VTVELTP | 2,38993065 |
| GNTLVIV | 3,346048239 | EVLAVGN | 2,3950255 |
| SFYLLYY | 3,365748416 | GEVLAVG | 2,3950255 |
| KLVTVHN | 3,425429496 | RGEVLAV | 2,3950255 |
| VVFILYD | 3,432981805 | TRGEVLA | 2,3950255 |
| RLVIVRA | 3,434746113 | GITLENE | 2,39532425 |
| LYLVLHN | 3,439526679 | ITLENEH | 2,39532425 |
| KNYIRIL | 3,457753029 | LENEHMK | 2,39532425 |
| HILLFYR | 3,475520463 | TLENEHM | 2,39532425 |
| GIVKTFD | 3,485106354 | PSRSQGL | 2,41275532 |
| KLVIVKA | 3,493533229 | RSQGLEA | 2,41275532 |
| HILLFQR | 3,502227789 | SRSQGLE | 2,41275532 |
| TLFIINL | 3,518556963 | TPSRSQG | 2,41275532 |
| GLFSFSV | 3,522003044 | CDLLVCG | 2,41597753 |
| SDGLFSF | 3,522003044 | EQCDLLV | 2,41597753 |
| TLVIVTA | 3,539741074 | KEQCDLL | 2,41597753 |
| VYLLYWL | 3,547588346 | QCDLLVC | 2,41597753 |
| FILYDQM | 3,630446145 | ATGIAVQ | 2,42728643 |
| QWPKTKL | 3,640307756 | GIAVQIL | 2,42728643 |
| NRLMPAY | 3,641712898 | MATGIAV | 2,42728643 |
| HIIKVLY | 3,644052696 | TGIAVQI | 2,42728643 |
| RVIVKRK | 3,673321884 | HNQGITL | 2,43438929 |
| KLRIERG | 3,694843612 | LHNQGIT | 2,43438929 |
| KTKLRIE | 3,694843612 | NQGITLE | 2,43438929 |
| QVHISAF | 3,696799841 | VLHNQGI | 2,43438929 |
| VWIVMRQ | 3,725304385 | ALVNLSA | 2,43656533 |
| KLYKLTK | 3,728745878 | LVNLSAV | 2,43656533 |
| HILLFLR | 3,743046707 | MALVNLS | 2,43656533 |
| EIIIKIF | 3,749989277 | VNLSAVA | 2,43656533 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| IIIKIFF | 3,749989277 | DADHAAA | 2,43772501 |
| VMVIPET | 3,759188377 | DLVIVDA | 2,43772501 |
| KEKLLRN | 3,782072752 | EGNDLVI | 2,43772501 |
| FAFFSIG | 3,935443438 | IVDADHA | 2,43772501 |
| FFSIGVQ | 3,935443438 | LVIVDAD | 2,43772501 |
| KGLIAVV | 4 | NDLVIVD | 2,43772501 |
| KTLILSH | 4 | VDADHAA | 2,43772501 |
| SILLMFS | 4 | VIVDADH | 2,43772501 |
| TLIVLAA | 4 | DGYGVKS | 2,43787702 |
| ATPEKLE | 0 | FNDGYGV | 2,43787702 |
| DDATPEK | 0 | IFNDGYG | 2,43787702 |
| DFNKFAA | 0 | NDGYGVK | 2,43787702 |
| DMIVAGT | 0 | EAMIRAI | 2,46031149 |
| GTLTNKM | 0 | GLEAMIR | 2,46031149 |
| IVAGTLT | 0 | LEAMIRA | 2,46031149 |
| TLTNKMA | 0 | QGLEAMI | 2,46031149 |
| EVLIMSE | 1,165210652 | DRSPQNS | 2,46441134 |
| LIMSESD | 1,165210652 | PQNSIQG | 2,46441134 |
| IEMQGTV | 1,244947169 | RSPQNSI | 2,46441134 |
| MQGTVLE | 1,244947169 | SPQNSIQ | 2,46441134 |
| FRVELEN | 1,311808169 | ADVISAF | 2,47020364 |
| VELENGH | 1,311808169 | AFGSVLS | 2,47020364 |
| MSSRSEL | 1,354751509 | DVISAFG | 2,47020364 |
| SRSELLL | 1,354751509 | FGSVLSD | 2,47020364 |
| SSRSELL | 1,354751509 | GSVLSDP | 2,47020364 |
| EPTGPLH | 1,450937305 | ISAFGSV | 2,47020364 |
| LEPTGPL | 1,450937305 | SAFGSVL | 2,47020364 |
| PTGPLHT | 1,450937305 | VISAFGS | 2,47020364 |
| EKLENEI | 1,477440775 | ETKSAGG | 2,49199925 |
| LPNTMFR | 1,650779866 | EVETKSA | 2,49199925 |
| PNTMFRV | 1,650779866 | TKSAGGI | 2,49199925 |
| NGHVVTA | 1,668333133 | VETKSAG | 2,49199925 |
| AGAVVGG | 1,675798196 | HIDDGLS | 2,51169936 |
| AVVGGLG | 1,675798196 | IDDGLSE | 2,51169936 |
| GAVVGGL | 1,675798196 | IHIDDGL | 2,51169936 |
| HMAGAAA | 1,675798196 | LIHIDDG | 2,51169936 |
| KTNMKHM | 1,675798196 | EILGVVI | 2,51223086 |
| TNMKHMA | 1,675798196 | GEILGVV | 2,51223086 |
| VVGGLGA | 1,675798196 | GVVIVES | 2,51223086 |
| IIELQGD | 1,687157027 | ILGVVIV | 2,51223086 |
| DIQVEAT | 1,693750851 | KQGEILG | 2,51223086 |
| MIPDIQV | 1,693750851 | LGVVIVE | 2,51223086 |
| EIMQKEQ | 1,751243833 | QGEILGV | 2,51223086 |
| LEIMQKE | 1,751243833 | VVIVESG | 2,51223086 |
| MQKEQCD | 1,751243833 | ARHNDAH | 2,51370242 |
| HMKIEEI | 1,763702146 | HNDAHLT | 2,51370242 |
| NEHMKIE | 1,763702146 | LARHNDA | 2,51370242 |
| TVLETLP | 1,775770109 | RHNDAHL | 2,51370242 |
| EGIPEMI | 1,833501684 | DEYMMIY | 2,51572986 |
| IPEMIPD | 1,833501684 | EYMMIYG | 2,51572986 |
| PEMIPDI | 1,833501684 | MMIYGVC | 2,51572986 |
| ANDEYMM | 1,859445901 | YMMIYGV | 2,51572986 |
| DANDEYM | 1,859445901 | GSAAAKS | 2,51917508 |
| LSDANDE | 1,859445901 | LTGSAAA | 2,51917508 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| SDANDEY | 1,859445901 | TGSAAAK | 2,51917508 |
| ATFPDGS | 1,873269249 | VLTGSAA | 2,51917508 |
| EATFPDG | 1,873269249 | AVTKGRG | 2,520623 |
| QVEATFP | 1,873269249 | DVIAVTK | 2,520623 |
| VEATFPD | 1,873269249 | EIDVIAV | 2,520623 |
| GNNQQNY | 1,891561098 | IAVTKGR | 2,520623 |
| NNQQNYA | 1,891561098 | IDVIAVT | 2,520623 |
| NQQNYAA | 1,891561098 | SEIDVIA | 2,520623 |
| TVIIEAA | 1,895956269 | VIAVTKG | 2,520623 |
| CPVLEPT | 1,931132413 | VTKGRGV | 2,520623 |
| FSCPVLE | 1,931132413 | INKLSAD | 2,52721767 |
| SCPVLEP | 1,931132413 | KLSADLL | 2,52721767 |
| VFSCPVL | 1,931132413 | MINKLSA | 2,52721767 |
| AMSRPII | 1,966053077 | NKLSADL | 2,52721767 |
| FGSDYED | 1,966053077 | EKMALTQ | 2,54588842 |
| GSAMSRP | 1,966053077 | FEKMALT | 2,54588842 |
| GSDYEDA | 1,966053077 | KMALTQH | 2,54588842 |
| HFGSDYE | 1,966053077 | MALTQHL | 2,54588842 |
| IHFGSDY | 1,966053077 | HVTLSQA | 2,54615728 |
| IIHFGSD | 1,966053077 | NHVTLSQ | 2,54615728 |
| LGSAMSR | 1,966053077 | VTLSQAA | 2,54615728 |
| MLGSAMS | 1,966053077 | DAAIVKG | 2,56950722 |
| MSRPIIH | 1,966053077 | DSDAAIV | 2,56950722 |
| PIIHFGS | 1,966053077 | GDSDAAI | 2,56950722 |
| RPIIHFG | 1,966053077 | SDAAIVK | 2,56950722 |
| SAMSRPI | 1,966053077 | IIPSDGR | 2,57571664 |
| SDYEDAA | 1,966053077 | IPSDGRK | 2,57571664 |
| SRPIIHF | 1,966053077 | PSDGRKE | 2,57571664 |
| YMLGSAM | 1,966053077 | SDGRKEV | 2,57571664 |
| AEVLIPG | 1,969772595 | AKTAAAL | 2,57599607 |
| DAEVLIP | 1,969772595 | KTAAALH | 2,57599607 |
| EVLIPGL | 1,969772595 | MAKTAAA | 2,57599607 |
| VLIPGLR | 1,969772595 | TAAALHI | 2,57599607 |
| DYGILQI | 1,994944121 | DGFITIT | 2,58506629 |
| GILQINS | 1,994944121 | FITITGG | 2,58506629 |
| ILQINSR | 1,994944121 | GFITITG | 2,58506629 |
| INSRWWC | 1,994944121 | GLDGFIT | 2,58506629 |
| LQINSRW | 1,994944121 | ITGGKLM | 2,58506629 |
| NSRWWCA | 1,994944121 | ITITGGK | 2,58506629 |
| QINSRWW | 1,994944121 | LDGFITI | 2,58506629 |
| SRWWCAA | 1,994944121 | TITGGKL | 2,58506629 |
| STDYGIL | 1,994944121 | EVKPLDV | 2,58631507 |
| TDYGILQ | 1,994944121 | GEVKPLD | 2,58631507 |
| YGILQIN | 1,994944121 | NGEVKPL | 2,58631507 |
| AKEDNIE | 2,016147514 | VKPLDVK | 2,58631507 |
| EDNIEMQ | 2,016147514 | FPTDNSN | 2,58707049 |
| KEDNIEM | 2,016147514 | GKFPTDN | 2,58707049 |
| MAKEDNI | 2,016147514 | KFPTDNS | 2,58707049 |
| NNQQNYQ | 2,028249555 | PTDNSNF | 2,58707049 |
| NQQNYQA | 2,028249555 | NFLVHAA | 2,59416033 |
| QQNYQAA | 2,028249555 | AGAQSLL | 2,61312839 |
| GVKSEKI | 2,034989728 | EAGAQSL | 2,61312839 |
| KSEKIDN | 2,034989728 | GAQSLLL | 2,61312839 |
| VKSEKID | 2,034989728 | YEAGAQS | 2,61312839 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| YGVKSEK | 2,034989728 | AVGGGSR | 2,61358494 |
| AALMEEG | 2,048076005 | AVVMAVG | 2,61358494 |
| KSVAALM | 2,048076005 | EFQAVVM | 2,61358494 |
| SVAALME | 2,048076005 | FQAVVMA | 2,61358494 |
| VAALMEE | 2,048076005 | MAVGGGS | 2,61358494 |
| EEISSSD | 2,054104511 | QAVVMAV | 2,61358494 |
| EISSSDN | 2,054104511 | VMAVGGG | 2,61358494 |
| IEEISSS | 2,054104511 | VVMAVGG | 2,61358494 |
| ISSSDNK | 2,054104511 | GLRGPTA | 2,61433407 |
| AQGIIEL | 2,059681035 | LRGPTAA | 2,61433407 |
| NAQGIIE | 2,059681035 | NGLRGPT | 2,61433407 |
| QGIIELQ | 2,059681035 | VNGLRGP | 2,61433407 |
| QNAQGII | 2,059681035 | ADFGKLA | 2,6144971 |
| GQMVPAF | 2,074819432 | GADFGKL | 2,6144971 |
| MVPAFDK | 2,074819432 | KNGADFG | 2,6144971 |
| QMVPAFD | 2,074819432 | NGADFGK | 2,6144971 |
| VPAFDKV | 2,074819432 | DNKHYYA | 2,61608275 |
| ETLLEIM | 2,076685948 | NKHYYAG | 2,61608275 |
| MPETLLE | 2,076685948 | SDNKHYY | 2,61608275 |
| PETLLEI | 2,076685948 | SSDNKHY | 2,61608275 |
| TLLEIMQ | 2,076685948 | GRILENG | 2,61656308 |
| ELTPREK | 2,087115652 | ILENGEV | 2,61656308 |
| LTPREKD | 2,087115652 | LENGEVK | 2,61656308 |
| MELTPRE | 2,087115652 | RILENGE | 2,61656308 |
| TPREKDK | 2,087115652 | AALHILV | 2,618235 |
| STVIFEA | 2,098955672 | ALHILVK | 2,618235 |
| TVIFEAA | 2,098955672 | HILVKEE | 2,618235 |
| KTVIIEA | 2,102180154 | LHILVKE | 2,618235 |
| AMSAAVG | 2,107403357 | IQGCQSQ | 2,62769673 |
| IAMSAAV | 2,107403357 | NSIQGCQ | 2,62769673 |
| MSAAVGL | 2,107403357 | QGCQSQV | 2,62769673 |
| VIAMSAA | 2,107403357 | SIQGCQS | 2,62769673 |
| DVRPWFE | 2,11789237 | ESVALIS | 2,6317037 |
| PWFEKMA | 2,11789237 | NYPESVA | 2,6317037 |
| RPWFEKM | 2,11789237 | PESVALI | 2,6317037 |
| VRPWFEK | 2,11789237 | YPESVAL | 2,6317037 |
| EQVMEGI | 2,122784541 | GPYLCYE | 2,63619353 |
| QVMEGIP | 2,122784541 | LCYEAGA | 2,63619353 |
| REQVMEG | 2,122784541 | PYLCYEA | 2,63619353 |
| VMEGIPE | 2,122784541 | YLCYEAG | 2,63619353 |
| EEKLALD | 2,123794752 | EPHILLF | 2,64252883 |
| KEEKLAL | 2,123794752 | PEPHILL | 2,64252883 |
| LVKEEKL | 2,123794752 | ATEDILQ | 2,65305296 |
| VKEEKLA | 2,123794752 | FPATEDI | 2,65305296 |
| EGNTLVE | 2,137914622 | PATEDIL | 2,65305296 |
| EVTADHA | 2,137914622 | YFPATED | 2,65305296 |
| GNTLVEV | 2,137914622 | CVNITIK | 2,66477221 |
| LVEVTAD | 2,137914622 | DCVNITI | 2,66477221 |
| NTLVEVT | 2,137914622 | HTVTTTT | 2,66477221 |
| TLVEVTA | 2,137914622 | IKQHTVT | 2,66477221 |
| VEVTADH | 2,137914622 | ITIKQHT | 2,66477221 |
| EFRQGQM | 2,16073412 | KQHTVTT | 2,66477221 |
| FRQGQMV | 2,16073412 | NITIKQH | 2,66477221 |
| GEFRQGQ | 2,16073412 | QHTVTTT | 2,66477221 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| RQGQMVP | 2,16073412 | TIKQHTV | 2,66477221 |
| AFTPRDA | 2,172693856 | TVTTTTA | 2,66477221 |
| FTPRDAE | 2,172693856 | VNITIKQ | 2,66477221 |
| PRDAEVL | 2,172693856 | VTTTTAA | 2,66477221 |
| TPRDAEV | 2,172693856 | KSMENLY | 2,68007344 |
| SVTSTFA | 2,178492525 | SMENLYL | 2,68007344 |
| VTSTFAA | 2,178492525 | VKSMENL | 2,68007344 |
| CREKDAA | 2,181363726 | VVKSMEN | 2,68007344 |
| FCREKDA | 2,181363726 | ELTPYDL | 2,68418191 |
| HILLFCR | 2,181363726 | LTPYDLS | 2,68418191 |
| ILLFCRE | 2,181363726 | PYDLSKG | 2,68418191 |
| LFCREKD | 2,181363726 | TPYDLSK | 2,68418191 |
| LLFCREK | 2,181363726 | ALLPDKE | 2,69510208 |
| PHILLFC | 2,181363726 | LLPDKEK | 2,69510208 |
| DVKVGDI | 2,188287941 | LPDKEKL | 2,69510208 |
| LDVKVGD | 2,188287941 | MALLPDK | 2,69510208 |
| PLDVKVG | 2,188287941 | EFCRVNG | 2,70487552 |
| VKVGDIV | 2,188287941 | FCRVNGL | 2,70487552 |
| DREKDAA | 2,220999021 | RVEFCRV | 2,70487552 |
| FDREKDA | 2,220999021 | VEFCRVN | 2,70487552 |
| HILLFDR | 2,220999021 | GNDLVIV | 2,72952038 |
| ILLFDRE | 2,220999021 | DVVNFDV | 2,72973524 |
| LFDREKD | 2,220999021 | NFDVRPW | 2,72973524 |
| LLFDREK | 2,220999021 | VNFDVRP | 2,72973524 |
| PHILLFD | 2,220999021 | VVNFDVR | 2,72973524 |
| AANVYLS | 2,227294045 | CPSGKRG | 2,73365107 |
| GPTAANV | 2,227294045 | GKRGGDL | 2,73365107 |
| PTAANVY | 2,227294045 | PSGKRGG | 2,73365107 |
| TAANVYL | 2,227294045 | SGKRGGD | 2,73365107 |
| ARDGKSV | 2,235353661 | AAIVALE | 2,73896535 |
| DGKSVAA | 2,235353661 | AIVALEA | 2,73896535 |
| GARDGKS | 2,235353661 | ELMAAIV | 2,73896535 |
| RDGKSVA | 2,235353661 | IVALEAL | 2,73896535 |
| AISGNEE | 2,242879295 | LMAAIVA | 2,73896535 |
| GVAISGN | 2,242879295 | MAAIVAL | 2,73896535 |
| IGVAISG | 2,242879295 | MELMAAI | 2,73896535 |
| VAISGNE | 2,242879295 | RMELMAA | 2,73896535 |
| EGNTLVP | 2,251544281 | AACFLET | 2,75043429 |
| GNTLVPV | 2,251544281 | ACFLETL | 2,75043429 |
| LVPVTAD | 2,251544281 | FAACFLE | 2,75043429 |
| NTLVPVT | 2,251544281 | KLQQFAA | 2,75043429 |
| PVTADHA | 2,251544281 | LKLQQFA | 2,75043429 |
| TLVPVTA | 2,251544281 | LQQFAAC | 2,75043429 |
| VPVTADH | 2,251544281 | QFAACFL | 2,75043429 |
| EEDALLV | 2,26092081 | QQFAACF | 2,75043429 |
| GNEEDAL | 2,26092081 | GQRLPEL | 2,75835714 |
| NEEDALL | 2,26092081 | LPELRDE | 2,75835714 |
| SGNEEDA | 2,26092081 | QRLPELR | 2,75835714 |
| AILSSAA | 2,26334659 | RLPELRD | 2,75835714 |
| FGAILSS | 2,26334659 | ALLVNKA | 2,78172912 |
| GAILSSA | 2,26334659 | DALLVNK | 2,78172912 |
| NFGAILS | 2,26334659 | LLVNKAL | 2,78172912 |
| NNFGAIL | 2,26334659 | LVNKALE | 2,78172912 |
| SNNFGAI | 2,26334659 | EADHAAA | 2,78611146 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| CYVSFHP | 2,265819382 | EGNELVI | 2,78611146 |
| DIEVDLL | 2,265819382 | ELVIVEA | 2,78611146 |
| EVDLLKA | 2,265819382 | IVEADHA | 2,78611146 |
| FHPSDIE | 2,265819382 | LVIVEAD | 2,78611146 |
| FLNCYVS | 2,265819382 | NELVIVE | 2,78611146 |
| HPSDIEV | 2,265819382 | VEADHAA | 2,78611146 |
| IEVDLLK | 2,265819382 | VIVEADH | 2,78611146 |
| LNCYVSF | 2,265819382 | ILTGDKV | 2,80038712 |
| NCYVSFH | 2,265819382 | LTGDKVT | 2,80038712 |
| NFLNCYV | 2,265819382 | RILTGDK | 2,80038712 |
| PSDIEVD | 2,265819382 | TGDKVTV | 2,80038712 |
| SDIEVDL | 2,265819382 | KMTGIVK | 2,81237392 |
| SFHPSDI | 2,265819382 | MSRKMTG | 2,81237392 |
| SNFLNCY | 2,265819382 | RKMTGIV | 2,81237392 |
| VDLLKAA | 2,265819382 | SRKMTGI | 2,81237392 |
| VSFHPSD | 2,265819382 | EGNTLVK | 2,82052236 |
| YVSFHPS | 2,265819382 | GNTLVKV | 2,82052236 |
| EQIKNGA | 2,266847724 | KVTADHA | 2,82052236 |
| LEQIKNG | 2,266847724 | LVKVTAD | 2,82052236 |
| LLEQIKN | 2,266847724 | NTLVKVT | 2,82052236 |
| QIKNGAD | 2,266847724 | TLVKVTA | 2,82052236 |
| MTPQDVV | 2,267421536 | VKVTADH | 2,82052236 |
| PQDVVNF | 2,267421536 | KMMTSYV | 2,83717422 |
| QMTPQDV | 2,267421536 | LTKMMTS | 2,83717422 |
| TPQDVVN | 2,267421536 | MMTSYVI | 2,83717422 |
| ERGEMPE | 2,268905705 | MTSYVIG | 2,83717422 |
| GEMPETL | 2,268905705 | SYVIGQA | 2,83717422 |
| IERGEMP | 2,268905705 | TKMMTSY | 2,83717422 |
| RGEMPET | 2,268905705 | TSYVIGQ | 2,83717422 |
| AAAKSTR | 2,27201631 | YVIGQAM | 2,83717422 |
| AAKSTRG | 2,27201631 | AEKIGIG | 2,84910992 |
| AKSTRGE | 2,27201631 | EKIGIGS | 2,84910992 |
| KSTRGEV | 2,27201631 | FAEKIGI | 2,84910992 |
| AVGNGRI | 2,282817426 | KFAEKIG | 2,84910992 |
| GNGRILE | 2,282817426 | ALDLLEQ | 2,85070309 |
| LAVGNGR | 2,282817426 | KLALDLL | 2,85070309 |
| VGNGRIL | 2,282817426 | LALDLLE | 2,85070309 |
| EIEVVVK | 2,287283648 | LDLLEQI | 2,85070309 |
| EVVVKSM | 2,287283648 | AHISGKM | 2,85840591 |
| IEVVVKS | 2,287283648 | TAHISGK | 2,85840591 |
| NEIEVVV | 2,287283648 | VTAHISG | 2,85840591 |
| EEGRHVL | 2,299377752 | VVTAHIS | 2,85840591 |
| EGRHVLS | 2,299377752 | FHREKDA | 2,86378018 |
| LMEEGRH | 2,299377752 | HILLFHR | 2,86378018 |
| MEEGRHV | 2,299377752 | HREKDAA | 2,86378018 |
| DEDRSPQ | 2,303809329 | ILLFHRE | 2,86378018 |
| ELRDEDR | 2,303809329 | LFHREKD | 2,86378018 |
| LRDEDRS | 2,303809329 | LLFHREK | 2,86378018 |
| RDEDRSP | 2,303809329 | PHILLFH | 2,86378018 |
| DGSKLVT | 2,309263764 | GKMRKNY | 2,89175403 |
| FPDGSKL | 2,309263764 | ISGKMRK | 2,89175403 |
| GSKLVTV | 2,309263764 | KMRKNYI | 2,89175403 |
| PDGSKLV | 2,309263764 | SGKMRKN | 2,89175403 |
| FNENRLC | 2,3235216 | AYKHIGV | 2,90630513 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| ISFNENR | 2,3235216 | KHIGVAI | 2,90630513 |
| NENRLCS | 2,3235216 | MAYKHIG | 2,90630513 |
| SFNENRL | 2,3235216 | YKHIGVA | 2,90630513 |
| FGSVQFV | 2,324629805 | AALLSMP | 2,92168964 |
| GSVQFVA | 2,324629805 | AVGAALL | 2,92168964 |
| NFGSVQF | 2,324629805 | GAALLSM | 2,92168964 |
| SVQFVAA | 2,324629805 | LQRVAVG | 2,92168964 |
| AENGGPY | 2,341325716 | QRVAVGA | 2,92168964 |
| ENGGPYL | 2,341325716 | RVAVGAA | 2,92168964 |
| FAENGGP | 2,341325716 | VAVGAAL | 2,92168964 |
| NGGPYLC | 2,341325716 | VGAALLS | 2,92168964 |
| ESDILAI | 2,345031392 | ALISAFI | 2,92719723 |
| MSESDIL | 2,345031392 | ISAFIME | 2,92719723 |
| SDILAIV | 2,345031392 | LISAFIM | 2,92719723 |
| SESDILA | 2,345031392 | VALISAF | 2,92719723 |
| ANWEEKY | 2,348908948 | AQKTVEG | 2,938638 |
| CANWEEK | 2,348908948 | AVAQKTV | 2,938638 |
| NWEEKYL | 2,348908948 | AVVTGVT | 2,938638 |
| RCANWEE | 2,348908948 | GAVVTGV | 2,938638 |
| DLGEFRQ | 2,358326604 | GVTAVAQ | 2,938638 |
| GDLGEFR | 2,358326604 | KTVEGAA | 2,938638 |
| GGDLGEF | 2,358326604 | QKTVEGA | 2,938638 |
| RGGDLGE | 2,358326604 | TAVAQKT | 2,938638 |
| AGGIVLT | 2,365913198 | TGVTAVA | 2,938638 |
| GGIVLTG | 2,365913198 | TVEGAAA | 2,938638 |
| GIVLTGS | 2,365913198 | VAQKTVE | 2,938638 |
| SAGGIVL | 2,365913198 | VTAVAQK | 2,938638 |
| GFIIPSD | 2,373450242 | VTGVTAV | 2,938638 |
| GKGFIIP | 2,373450242 | VVTGVTA | 2,938638 |
| KGFIIPS | 2,373450242 | EGNTLVR | 2,94631143 |
| SGKGFII | 2,373450242 | GNTLVRV | 2,94631143 |
| TADHAAA | 2,386362309 | LVRVTAD | 2,94631143 |
| VTADHAA | 2,386362309 | NTLVRVT | 2,94631143 |
| DKVTVEL | 2,389930651 | RVTADHA | 2,94631143 |
| PHILLFG | 2,988107989 | TLVRVTA | 2,94631143 |
| GNELVIV | 2,989034466 | VRVTADH | 2,94631143 |
| ALRFPLD | 2,990098958 | HSICPSG | 2,94977 |
| FPLDDAT | 2,990098958 | KHSICPS | 2,94977 |
| LRFPLDD | 2,990098958 | KKHSICP | 2,94977 |
| RFPLDDA | 2,990098958 | SICPSGK | 2,94977 |
| ELLLEKF | 2,990464812 | EREKDAA | 2,95839323 |
| LEKFAEK | 2,990464812 | FEREKDA | 2,95839323 |
| LLEKFAE | 2,990464812 | HILLFER | 2,95839323 |
| LLLEKFA | 2,990464812 | ILLFERE | 2,95839323 |
| GLLTTLN | 2,995859596 | LFEREKD | 2,95839323 |
| HVGLLTT | 2,995859596 | LLFEREK | 2,95839323 |
| LLTTLNF | 2,995859596 | PHILLFE | 2,95839323 |
| LTTLNFG | 2,995859596 | DKVVFSC | 2,96571015 |
| SHVGLLT | 2,995859596 | FDKVVFS | 2,96571015 |
| TLNFGDG | 2,995859596 | KVVFSCP | 2,96571015 |
| TTLNFGD | 2,995859596 | VVFSCPV | 2,96571015 |
| VGLLTTL | 2,995859596 | FGREKDA | 2,98810799 |
| GLRVEFC | 2,999080324 | GREKDAA | 2,98810799 |
| IPGLRVE | 2,999080324 | HILLFGR | 2,98810799 |
| **Normalized Score** | **Heptamer** | **Heptamer** | **Normalized Score** |
| LRVEFCR | 2,999080324 | ILLFGRE | 2,98810799 |
| PGLRVEF | 2,999080324 | LFGREKD | 2,98810799 |
| VCGKFPT | 3,005469211 | LLFGREK | 2,98810799 |
| DLVIVTA | 3,021315746 | RNFLRCA | 3,17738367 |
| HLTLIHI | 3,027755686 | GIYFPAT | 3,18815474 |
| LTLIHID | 3,027755686 | PGIYFPA | 3,18815474 |
| RHVLSRE | 3,032731728 | ELVIVTA | 3,19195747 |
| HILLFNR | 3,050698636 | GPLHTQF | 3,19639177 |
| IRPLHDR | 3,062532886 | GNKLVIV | 3,2095174 |
| MNIRPLH | 3,062532886 | KLVIVTA | 3,2095174 |
| ALEILNA | 3,064671901 | SLLMKLF | 3,21117998 |
| RLCSFAI | 3,0673361 | TLVIVDA | 3,21210143 |
| LGVLGIR | 3,082889793 | QYLLAVD | 3,22032032 |
| QEVISLG | 3,082889793 | TQYLLAV | 3,22032032 |
| VLAVYAE | 3,101603515 | RKEVQVH | 3,22651438 |
| TLVIVPA | 3,113378376 | IYYISLS | 3,24275138 |
| FDRKSGK | 3,136026965 | LSELYPG | 3,25205027 |
| HILLFWR | 3,141946561 | LQLLKNK | 3,26453923 |
| VEKVVFW | 3,145072335 | AALLAER | 3,27149824 |
| VFWLHDS | 3,145072335 | ANLWFAE | 3,27575539 |
| VVFWLHD | 3,145072335 | NANLWFA | 3,27575539 |
| LLVCGHH | 3,145973852 | HIARNYA | 3,28037596 |
| DIVIFND | 3,14788463 | VVHIARN | 3,28037596 |
| ALAVLAV | 3,161492051 | TLVIVEA | 3,28630221 |
| HILLFMR | 3,167157782 | AFPLFAL | 3,28781133 |
| HILLFPR | 3,308676875 | AYLVYNA | 3,2950841 |
| TLVIVKA | 3,319167049 | QVWIVMR | 3,2960842 |
| GNRLVIV | 3,320153495 | SQVWIVM | 3,2960842 |
| ALELARH | 3,320316204 | PLVIVPA | 3,29960681 |
| NKALELA | 3,320316204 | GIGSISF | 3,30332329 |
| IVKRKEV | 3,322304665 | RKEVETK | 3,32230467 |

**Supplementary Table 5.** Final quantitative-based 20x7 PISM.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Position** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **A** | 0,14567794 | 0,04818767 | 0,00665627 | 0,03211395 | -0,0343707 | 0,01161968 | 0,04473559 |
| **C** | -0,2715942 | -0,0557392 | 0,06694989 | 0,08702948 | 0,17277077 | 0,02670373 | 0,11699927 |
| **E** | 0,02432079 | -0,1914373 | 0,09043395 | -0,0054715 | 0,08613284 | 0,08931813 | -0,015254 |
| **D** | 0,09261986 | -0,2061135 | 0,18241395 | 0,23323361 | 0,0322968 | -0,1602488 | -0,0773423 |
| **G** | 0,18166801 | 0,0677326 | 0,08549923 | -0,1114399 | 0,02749219 | -0,0182826 | -0,0549524 |
| **F** | -0,0847618 | 0,01291043 | -0,144703 | 0,03835858 | 0,19123185 | 0,25543387 | 0,1412575 |
| **I** | 0,09966293 | 0,04018111 | 0,08091449 | 0,18619379 | -0,0613654 | -0,0156231 | 0,00960857 |
| **H** | 0,00316637 | 0,00750568 | 0,19636926 | 0,31979451 | -0,2408857 | -0,0387696 | 0,04646512 |
| **K** | -0,2337895 | 0,2867628 | 0,18598707 | 0,33846578 | -0,1075575 | -0,0788761 | 0,00537582 |
| **M** | -0,0194343 | -0,3557637 | 0,01043346 | 0,26776317 | 0,01824333 | -0,0335918 | 0,19372862 |
| **L** | 0,01698865 | -0,0024529 | 0,07489881 | -0,0249672 | 0,09667284 | 0,15280091 | 0,06608233 |
| **N** | 0,0287688 | -0,070507 | -0,0358937 | 0,06284256 | 0,29931243 | -0,3519158 | 0,18299936 |
| **Q** | 0,2429028 | 0,05286953 | -0,0763199 | 0,38173389 | -0,2521255 | -0,0676313 | -0,0646074 |
| **P** | -0,0598986 | -0,0213685 | 0,35385948 | 0,05360025 | -0,4068352 | 0,07202069 | 0,13091775 |
| **S** | 0,07414246 | 0,1579762 | 0,1366677 | -0,0164196 | -0,0157308 | -0,021856 | -0,1328965 |
| **R** | -0,2389082 | 0,01013346 | 0,18307768 | -0,1126362 | 0,07725322 | 0,00154875 | 0,51172313 |
| **T** | 0,14647671 | 0,15445086 | -0,1768573 | 0,13509232 | 0,04916194 | -0,1691293 | 0,07435063 |
| **W** | -0,0654971 | -0,020803 | -0,0776952 | 0,19539848 | -0,0520748 | 0,13122955 | 0,17026478 |
| **V** | -0,06889 | -0,0009964 | 0,17202196 | 0,12998881 | -0,004906 | -0,0391157 | 0,16087556 |
| **Y** | 0,05127709 | -0,145112 | 0,22904897 | 0,06417733 | 0,30099858 | -0,2997115 | 0,26846789 |

Offset term: 0.718675619345

**Supplementary Table 6.** Final qualitative-based 20x7 PISM.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Position** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **A** | 0,02772892 | 0,03356412 | -0,0274479 | 0,0080969 | 0,02334192 | 0,01645188 | -0,1345558 |
| **C** | -0,2423557 | -0,0190308 | -0,0902136 | 0,04371711 | -0,0402091 | 0,31587864 | -0,153475 |
| **E** | -0,3546581 | 0,0237687 | -0,0043667 | 0,01545498 | -0,076613 | 0,13557834 | 0,13511838 |
| **D** | 0,0767015 | -0,0619147 | -0,0055943 | -0,0617203 | -0,064502 | -0,0140991 | -0,004613 |
| **G** | -0,1504576 | 0,24528891 | 0,07361844 | -0,0087505 | -0,0562485 | -0,2506829 | 0,0293025 |
| **F** | 0,12160976 | 0,00636235 | 0,04975532 | 0,14426455 | 0,02620517 | -0,311801 | 0,01742275 |
| **I** | 0,07188344 | 0,20029367 | -0,1223247 | -0,0539893 | -0,0511216 | 0,00709132 | -0,000248 |
| **H** | -0,0026876 | 0,19977328 | -0,1245267 | 0,13262339 | 0,02067887 | -0,2147967 | -0,0323243 |
| **K** | -0,0571738 | -0,0699915 | -0,0732875 | -0,0829186 | 0,29446825 | -0,0488725 | 0,05478155 |
| **M** | -0,2181167 | 0,01953522 | 0,10066954 | 0,02882645 | 0,00856765 | -0,0218294 | -0,0124394 |
| **L** | -0,1406314 | -0,1510123 | 0,07858373 | -0,1653882 | 0,01463639 | 0,24639886 | 0,16340288 |
| **N** | 0,04451703 | -0,0036594 | 0,0359383 | 0,08069534 | 0,07544078 | 0,03416336 | -0,3420097 |
| **Q** | -0,1512499 | 0,19934364 | -0,2552993 | 0,22905532 | -0,013511 | -0,1447299 | 0,08011651 |
| **P** | -0,0463946 | 0,07141876 | -0,0750862 | -0,1527967 | -0,0228232 | -0,0004052 | 0,16872326 |
| **S** | 0,06522657 | -0,0322716 | 0,11654486 | 0,04817205 | 0,01227283 | -0,1620065 | -0,1222289 |
| **R** | 0,08174114 | 0,17448505 | 0,13144211 | 0,08108816 | 0,09652098 | -0,348292 | -0,1500995 |
| **T** | -0,0262457 | 0,03703667 | -0,234901 | 0,25541162 | -0,0215262 | 0,00726724 | -0,0873706 |
| **W** | -0,0055601 | -0,0265819 | 0,10839476 | 0,043646 | -0,0339119 | -0,0136019 | 0,07059761 |
| **V** | 0,20890139 | 0,17162778 | -0,0839869 | -0,0590273 | -0,0153962 | 0,00423243 | -0,206164 |
| **Y** | -0,093552 | 0,01138308 | -0,0933564 | 0,08862024 | -0,0353932 | 0,0206736 | 0,22008913 |

Offset term: 0.338948385187

**Code 1:** Python 2.7 program for training a PSSM based on quantitative information.

Input:

*‘Training.txt’* is a text file containing two columns, the first being the sequence of the heptamer, and the second the normalized detection score.

*‘True\_Positives.txt’* is a text file in the same format as *‘Training.txt* containing the information for true binders.

*‘True\_Negatives.txt’ is a text file in the same format as ‘Training.txt* containing the information for true non binders.

Output:

A vector containing the PSSM information saved with pickle library. To read and use the PSSM, use the funcion *convert*.

def readfiles(fname):

ret = {}

with open(fname,'r') as F:

lines = F.readlines()

for line in lines:

linesplit = line.split()

score = linesplit[1]

score = score.replace(']','')

score = score.replace('[','')

ret.update({linesplit[0]:float(score)})

return ret

def runMat(pep,mat,offset):

ret = 0.0

for x in xrange(len(pep)):

ret += mat[pep[x]][x]

return ret+offset

def phi(dat, dict\_mat,offset):

ret\_sum = 0.0

for key in dat:

score = runMat(key,dict\_mat,offset)

ret\_sum+=(score-dat[key])\*\*2

return ret\_sum

def compute\_Bias(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm/len(data)

def compute\_Variance(vector):

dataP = readfiles('True\_Negatives.txt')

dataN = readfiles('True\_Positives.txt')

mat, offset = convert(vector)

firstTermP = phi(dataP,mat,offset)

firstTermN = phi(dataN,mat,offset)

firstTerm = firstTermP + firstTermN

n = len(dataP) + len(dataN)

return firstTerm/n

def convert(vec):

ret = {}

aas = ('A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y')

lines = []

y = 0

offset = vec[0]

for x in xrange(1,141,7):

ret.update({aas[y]:vec[x:x+7]})

y+=1

return ret,offset

def psi(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm

def gen\_mat():

import random

import numpy

range1 = -0.2

range2 = 0.2

rangeoffset1 = -0.5

rangeoffset2 = 0.5

mat = []

mat.append(random.uniform(rangeoffset1,rangeoffset2))

for x in xrange(140):

mat.append(random.uniform(range1,range2))

return numpy.array(mat)

def callback\_func(x):

ret,offset=convert(x)

print "\*"

print compute\_Bias(x),',',compute\_Variance(x),',',psi(x)

def minimize\_all\_methods(function, initialGuess):

from scipy.optimize import minimize

ret = {}

methods = ['nelder-mead']#,'cg','bfgs','l-bfgs-b','tnc','slsqp')

for m in methods:

print "Method:",m

res = minimize(psi, matVec,method=m,

options={'xtol': 1e-9,'disp': False},

callback=callback\_func)

print "\_"

ret.update({m:res})

return ret

# Main:

import numpy

matVec=gen\_mat()

dic = minimize\_all\_methods(psi,matVec)

import pickle

for key in dic:

with open('%s\_Gutierres\_2019.pickle'%key, 'wb') as handle:

pickle.dump(dic[key].x, handle, protocol=pickle.HIGHEST\_PROTOCOL)

**Code 2:** Python 2.7 program for training a PISM based on quantitative information.

Input:

*‘Training.txt’* is a text file containing two columns, the first being the sequence of the heptamer, and the second the normalized detection score.

*‘True\_Positives.txt’* is a text file in the same format as *‘Training.txt* containing the information for true binders.

*‘True\_Negatives.txt’* is a text file in the same format as *‘Training.txt* containing the information for true non binders.

Output:

A vector containing the PISM information saved with pickle library. To read and use the PISM, use the funcion *convert*.

def readfiles(fname):

ret = {}

with open(fname,'r') as F:

lines = F.readlines()

for line in lines:

linesplit = line.split()

score = linesplit[1]

score = score.replace(']','')

score = score.replace('[','')

ret.update({linesplit[0]:float(score)})

return ret

def runMat(pep,mat,offset):

ret = 0.0

for aa in pep:

ret += mat[aa]

return ret+offset

def phi(dat, dict\_mat,offset):

ret\_sum = 0.0

for key in dat:

score = runMat(key,dict\_mat,offset)

ret\_sum+=(score-dat[key])\*\*2

return ret\_sum

def compute\_Bias(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm/len(data)

def compute\_Variance(vector):

dataP = readfiles('True\_Negatives.txt')

dataN = readfiles('True\_PositIves.txt')

mat, offset = convert(vector)

firstTermP = phi(dataP,mat,offset)

firstTermN = phi(dataN,mat,offset)

firstTerm = firstTermP + firstTermN

n = len(dataP) + len(dataN)

return firstTerm/n

def convert(vec):

ret = {}

aas = ('A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y')

lines = []

y = 0

offset = vec[0]

for x in xrange(1,21,1):

ret.update({aas[y]:vec[x]})

y+=1

return ret,offset

def psi(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm

def gen\_mat():

import random

import numpy

range1 = -0.2

range2 = 0.2

rangeoffset1 = -0.5

rangeoffset2 = 0.5

mat = []

mat.append(random.uniform(rangeoffset1,rangeoffset2))

for x in xrange(20):

mat.append(random.uniform(range1,range2))

return numpy.array(mat)

def callback\_func(x):

ret,offset=convert(x)

print "\*"

print compute\_Bias(x),',',compute\_Variance(x),',',psi(x)

def minimize\_all\_methods(function, initialGuess):

from scipy.optimize import minimize

ret = {}

methods = ['nelder-mead']#,'cg','bfgs','l-bfgs-b','tnc','slsqp')

for m in methods:

print "Method:",m

res = minimize(psi, matVec,method=m,

options={'xtol': 1e-10,'disp': False},

callback=callback\_func)

print "\_"

ret.update({m:res})

return ret

# Main:

import numpy

matVec=gen\_mat()

dic = minimize\_all\_methods(psi,matVec)

import pickle

for key in dic:

with open('%s\_Gutierres\_2019.pickle'%key, 'wb') as handle:

pickle.dump(dic[key].x, handle, protocol=pickle.HIGHEST\_PROTOCOL)

**Code 3:** Python 2.7 program for training a 20x7 PISM based on quantitative information.

Input:

*‘Training.txt’* is a text file containing two columns, the first being the sequence of the heptamer, and the second the normalized detection score.

*‘True\_Positives.txt’* is a text file in the same format as *‘Training.txt* containing the information for true binders.

*‘True\_Negatives.txt’ is a text file in the same format as ‘Training.txt* containing the information for true non binders.

Output:

A vector containing the 20x7 PISM information saved with pickle library. To read and use the 20x7 PISM, use the funcion *convert*.

def readfiles(fname):

ret = {}

with open(fname,'r') as F:

lines = F.readlines()

for line in lines:

linesplit = line.split()

score = linesplit[1]

score = score.replace(']','')

score = score.replace('[','')

ret.update({linesplit[0]:float(score)})

return ret

def runMat(pep,mat,offset):

ret = 0.0

for aa in pep:

ret += mat[aa]

return ret+offset

def phi(dat, dict\_mat,offset):

ret\_sum = 0.0

for key in dat:

score = runMat(key,dict\_mat,offset)

ret\_sum+=(score-dat[key])\*\*2

return ret\_sum

def compute\_Bias(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm/len(data)

def compute\_Variance(vector):

dataP = readfiles('True\_Negatives.txt')

dataN = readfiles('True\_Positives.txt')

mat, offset = convert(vector)

firstTermP = phi(dataP,mat,offset)

firstTermN = phi(dataN,mat,offset)

firstTerm = firstTermP + firstTermN

n = len(dataP) + len(dataN)

return firstTerm/n

def convert(vec):

ret = {}

aas = ('A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y')

lines = []

y = 0

offset = vec[0]

for x in xrange(1,141,7):

ret.update({aas[y]:vec[x:x+7]})

y+=1

return ret,offset

def psi(vector):

data = readfiles('Training.txt')

mat, offset = convert(vector)

firstTerm = phi(data,mat,offset)

return firstTerm

def gen\_mat():

import random

import numpy

range1 = -0.2

range2 = 0.2

rangeoffset1 = -0.5

rangeoffset2 = 0.5

mat = []

mat.append(random.uniform(rangeoffset1,rangeoffset2))

for x in xrange(20):

mat.append(random.uniform(range1,range2))

return numpy.array(mat)

def callback\_func(x):

ret,offset=convert(x)

print "\*"

print compute\_Bias(x),',',compute\_Variance(x),',',psi(x)

def minimize\_all\_methods(function, initialGuess):

from scipy.optimize import minimize

ret = {}

methods = ['nelder-mead']#,'cg','bfgs','l-bfgs-b','tnc','slsqp')

for m in methods:

print "Method:",m

res = minimize(psi, matVec,method=m,

options={'xtol': 1e-10,'disp': False},

callback=callback\_func)

print "\_"

ret.update({m:res})

return ret

# Main:

import numpy

matVec=gen\_mat()

dic = minimize\_all\_methods(psi,matVec)

import pickle

for key in dic:

with open('%s\_Gutierres\_2019.pickle'%key, 'wb') as handle:

pickle.dump(dic[key].x, handle, protocol=pickle.HIGHEST\_PROTOCOL)

*# For training a PISM, PSSM or 20x7 PISM based on qualitative information use the function readfiles below.*

def readfiles(fname):

ret = {}

with open(fname,'r') as F:

lines = F.readlines()

for line in lines:

linesplit = line.split()

score = linesplit[1]

score = score.replace(']','')

score = score.replace('[','')

if float(score) <= 3.0: # threshold to consider positive

ret.update({linesplit[0]:0})

else:

ret.update({linesplit[0]:1})

return ret