Data S7

DNA shape features improve transcription factor binding site prediction *in vivo*

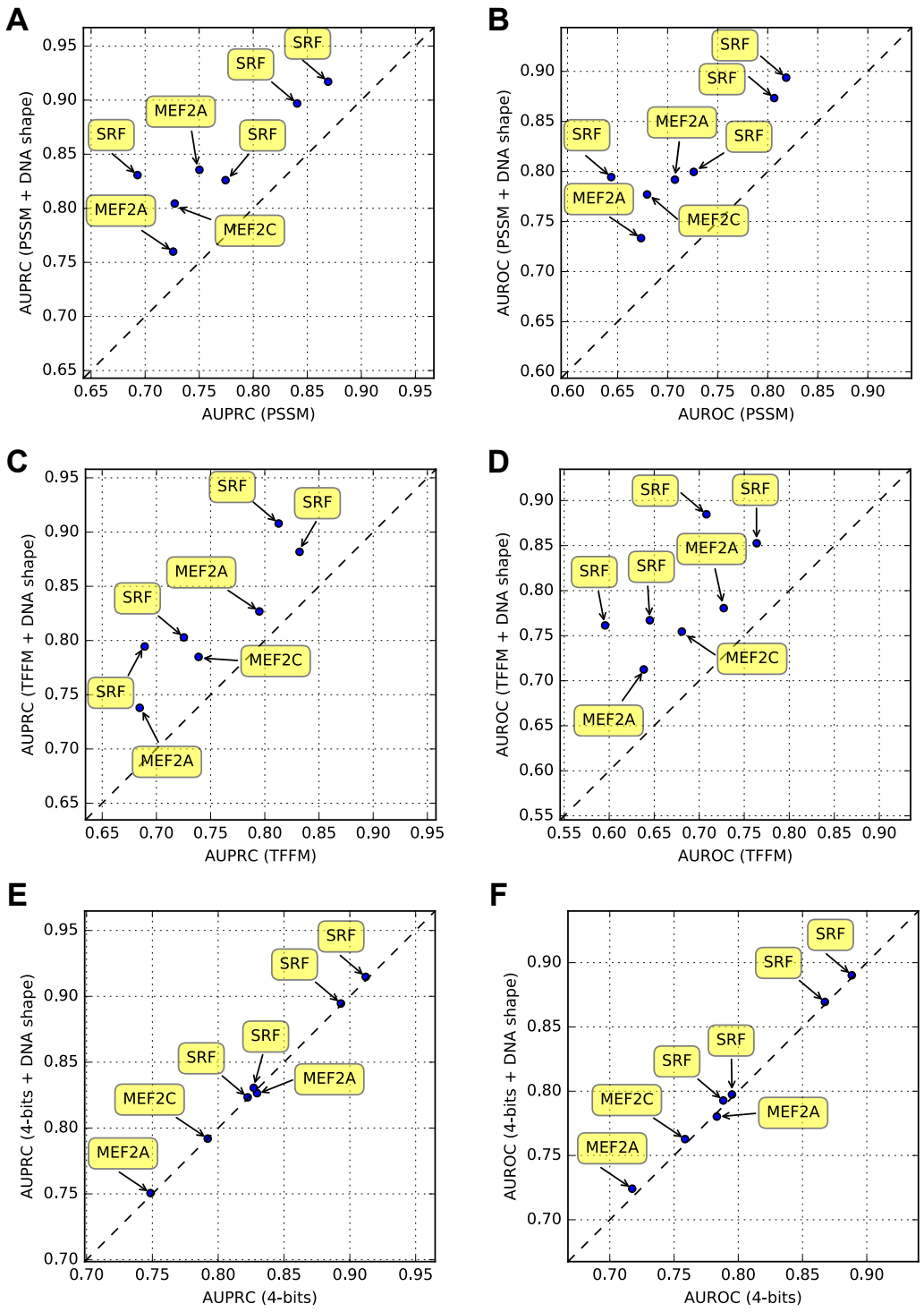
Anthony Mathelier[1,2], Beibei Xin[3], Tsu-Pei Chiu[3],
Lin Yang[3], Remo Rohs[3,*], and Wyeth W. Wasserman[1,*]

[1] Centre for Molecular Medicine at the Child and Family Research Institute,
Department of Medical Genetics, University of British Columbia,
980 West 28th Avenue, V5Z 4H4, Vancouver, BC, Canada

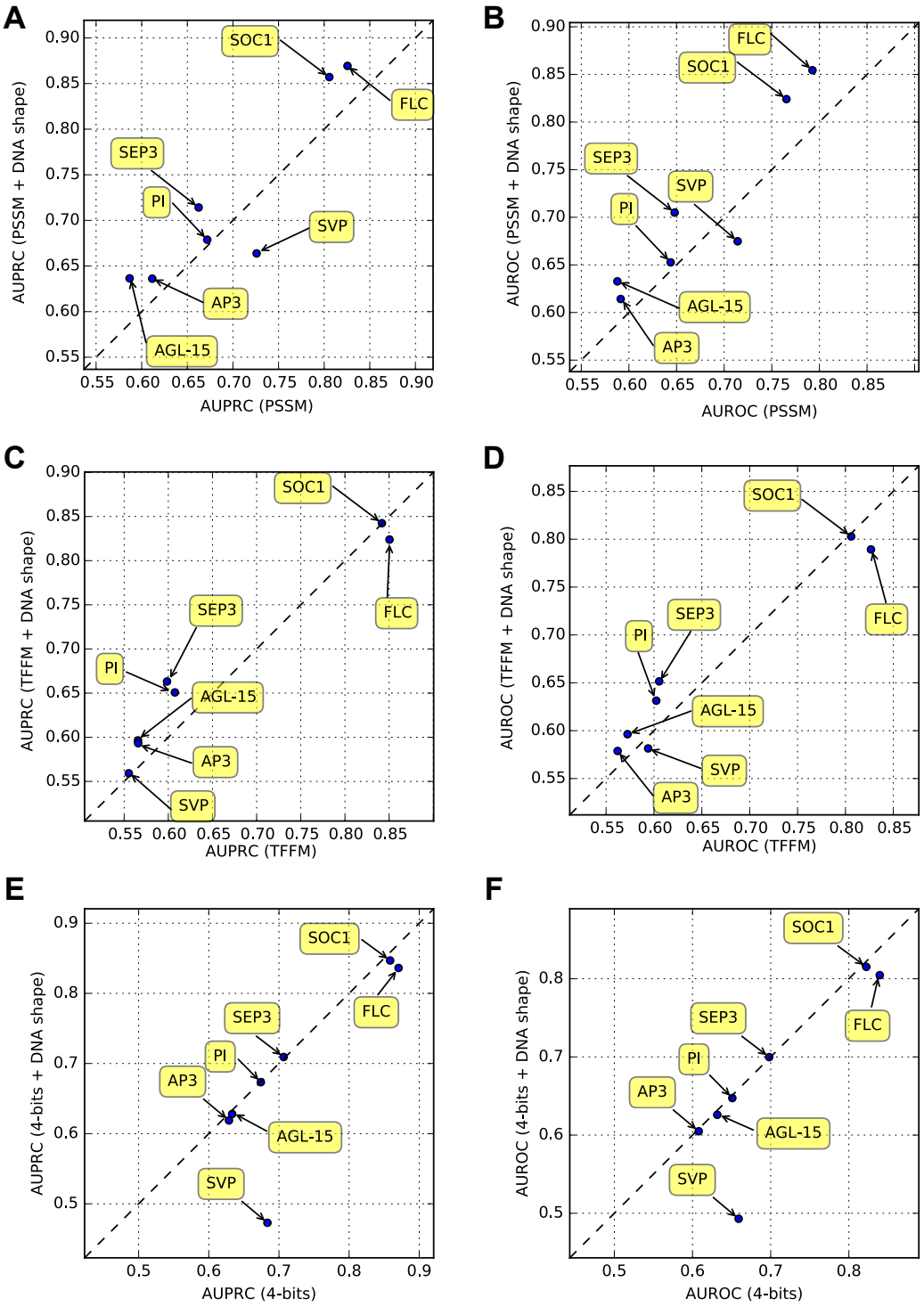
[2] Centre for Molecular Medicine Norway (NCMM), Nordic EMBL partnership,
University of Oslo and Oslo University Hospital, Norway

[3] Molecular and Computational Biology Program, Departments of Biological Sciences,
Chemistry, Physics, and Computer Science,
University of Southern California, Los Angeles, CA 90089, USA

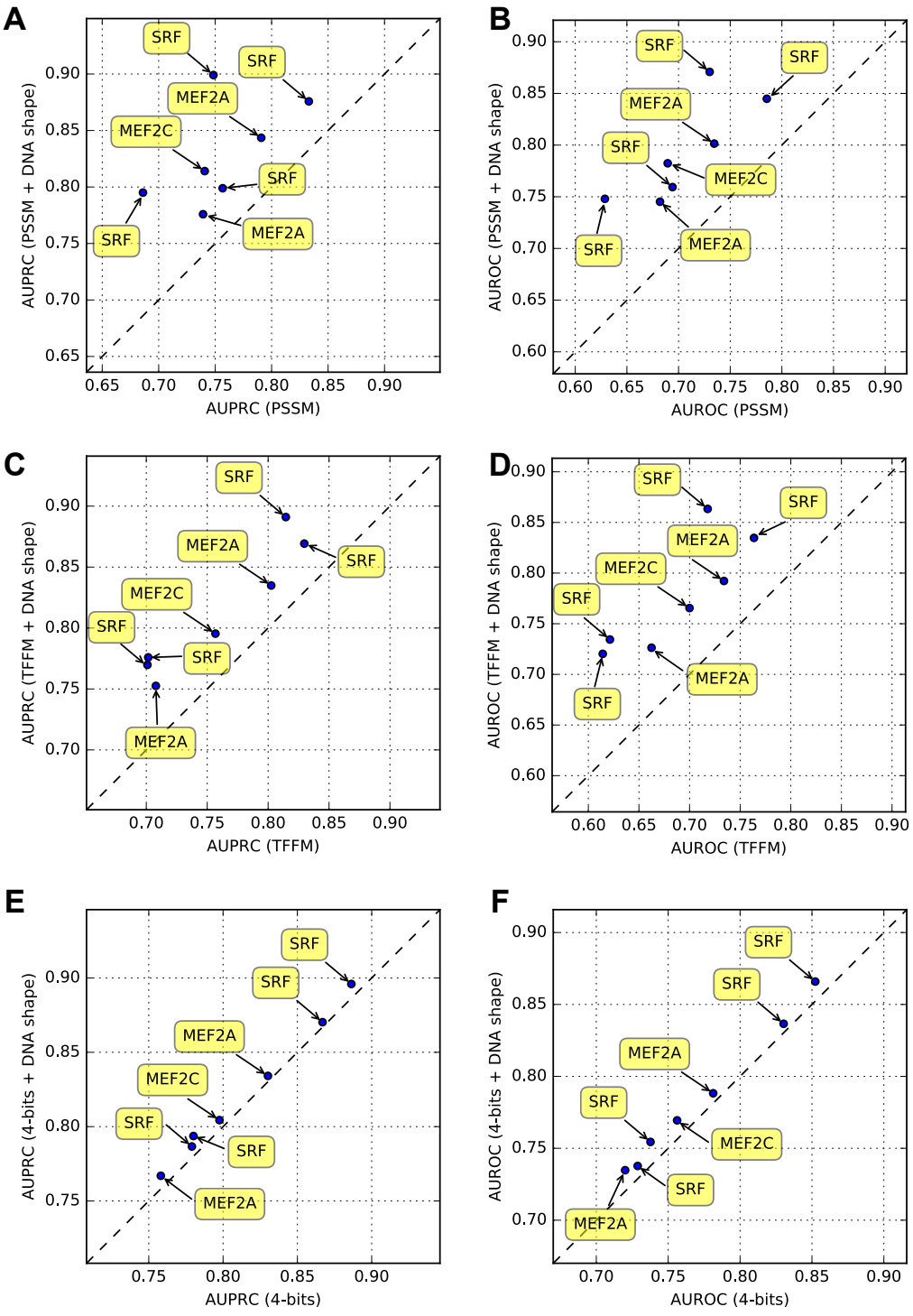
[*] Co-corresponding authors



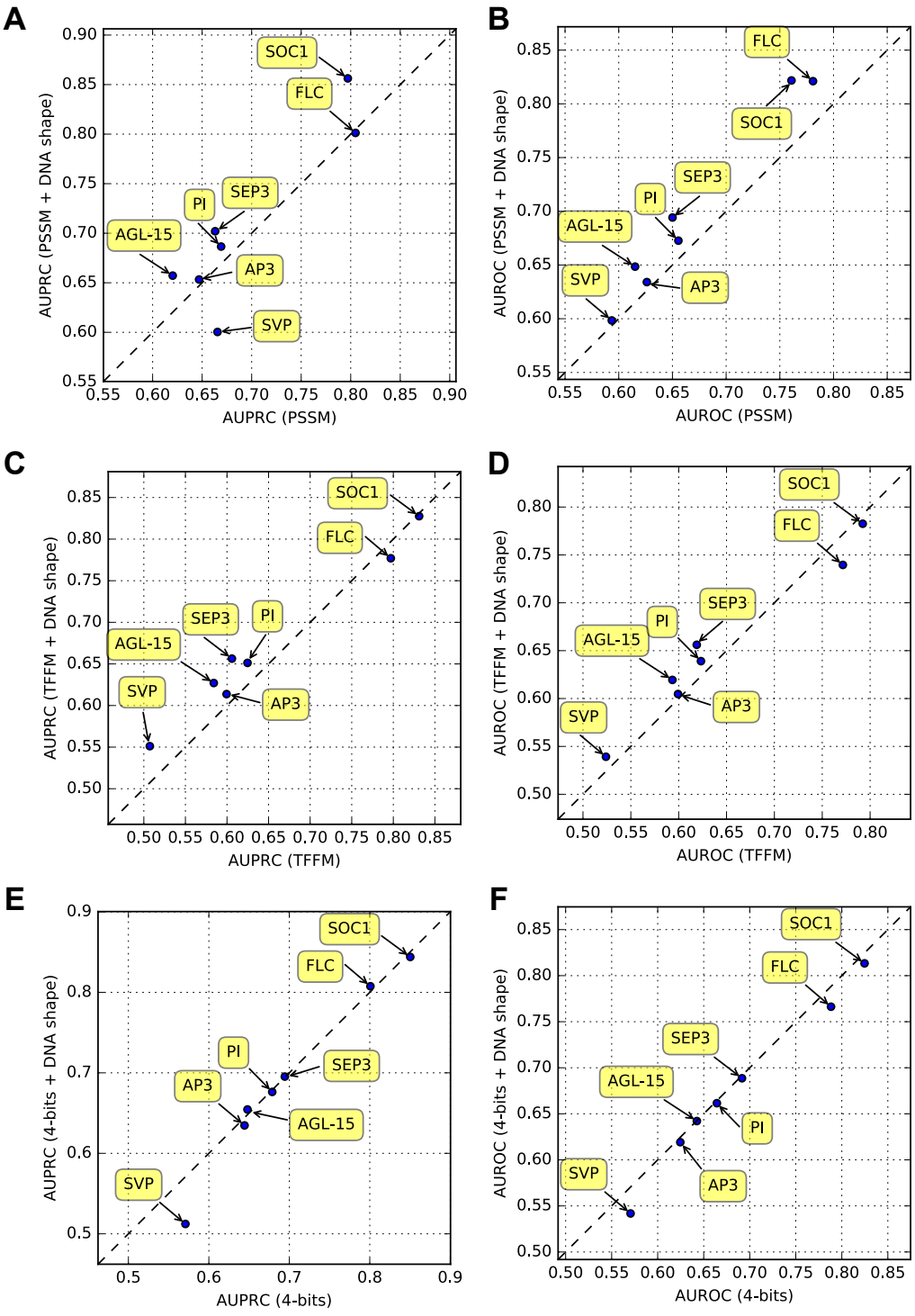
*Related to Figure 2. Comparison of the AUPRC (**A**, **C**, and **E**) or AUROC (**B**, **D**, and **F**) obtained for the 7 human ENCODE MADS TF ChIP-seq data sets when using the PSSM scores (x-axis; **A**-**B**), the TFFM scores (x-axis; **C**-**D**), or the 4-bits classifier (x-axis; **E**-**F**; see (Zhou* et al.*, 2015) for the 4-bits encoding) versus the PSSM + DNA shape (y-axis; **A**-**B**), TFFM + DNA shape (y-axis; **C**-**D**), or 4-bits + DNA shape (y-axis; **E**-**F**) classifiers.*



*Related to Figure 2. Comparison of the AUPRC (**A**, **C**, and **E**) or AUROC (**B**, **D**, and **F**) obtained for the 7 plant MADS TF ChIP-seq data sets when using the PSSM scores (x-axis; **A**-**B**), the TFFM scores (x-axis; **C**-**D**), or the 4-bits classifier (x-axis; **E**-**F**; see (Zhou* et al., *2015) for the 4-bits encoding) versus the PSSM + DNA shape (y-axis; **A**-**B**), TFFM + DNA shape (y-axis; **C**-**D**), or 4-bits + DNA shape (y-axis; **E**-**F**) classifiers.*



*Related to Figure 2. Impact of DNA shape on predicting human MADS-box TFBSs with background sequences matching the dinucleotide composition of ChIP-seq regions. Comparison of the AUPRC (**A**, **C**, and **E**) or AUROC (**B**, **D**, and **F**) obtained for the 7 human ENCODE MADS TF ChIP-seq data sets when using the PSSM scores (x-axis; **A**-**B**), the TFFM scores (x-axis; **C**-**D**), or the 4-bits classifier (x-axis; **E**-**F**; see (Zhou* et al.*, 2015) for the 4-bits encoding) versus the PSSM + DNA shape (y-axis; **A**-**B**), TFFM + DNA shape (y-axis; **C**-**D**), or 4-bits + DNA shape (y-axis; **E**-**F**) classifiers.*



*Related to Figure 2. Impact of DNA shape on predicting plant MADS-box TFBSs with background sequences matching the dinucleotide composition of ChIP-seq regions.Comparison of the AUPRC (**A**, **C**, and **E**) or AUROC (**B**, **D**, and **F**) obtained for the 7 plant MADS TF ChIP-seq data sets when using the PSSM scores (x-axis; **A**-**B**), the TFFM scores (x-axis; **C**-**D**), or the 4-bits classifier (x-axis; **E**-**F**; see (Zhou* et al., *2015) for the 4-bits encoding) versus the PSSM + DNA shape (y-axis; **A**-**B**), TFFM + DNA shape (y-axis; **C**-**D**), or 4-bits + DNA shape (y-axis; **E**-**F**) classifiers.*