ADVANCED FUNCTIONAL MATERIALS

Supporting Information

Machine-Learning Aided First-Principles Prediction of Earth-Abundant Pnictogen Chalcohalide Solid Solutions for Solar-Cell Devices

Cibrán López, Ivan Caño, David Rovira, Pol Benítez, José Miguel Asensi, Zacharie Jehl, Josep-Lluís Tamarit, Edgardo Saucedo and Claudio Cazorla*

Supporting Information for "Machine-Learning Aided First-Principles Prediction of Earth-Abundant Pnictogen Chalcohalide Solid Solutions for Solar-Cell Devices"

Cibrán López, Pol Benítez, Josep-Lluís Tamarit, and Claudio Cazorla

Group of Characterization of Materials, Departament de Física,

Universitat Politècnica de Catalunya, Campus Diagonal-Besòs,

Av. Eduard Maristany 10–14, 08019 Barcelona, Spain and

Research Center in Multiscale Science and Engineering,

Universitat Politècnica de Catalunya, Campus Diagonal-Besòs,

Av. Eduard Maristany 10–14, 08019 Barcelona, Spain

Ivan Caño, David Rovira, Zacharie Jehl, and Edgardo Saucedo

Research Center in Multiscale Science and Engineering,

Universitat Politècnica de Catalunya, Campus Diagonal-Besòs,

Av. Eduard Maristany 10–14, 08019 Barcelona, Spain and

Department of Electronic Engineering,

Universitat Politècnica de Catalunya, 08034 Barcelona, Spain

José Miguel Asensi

Departament de Física Aplicada, Universitat de Barcelona, 08007 Barcelona, Spain
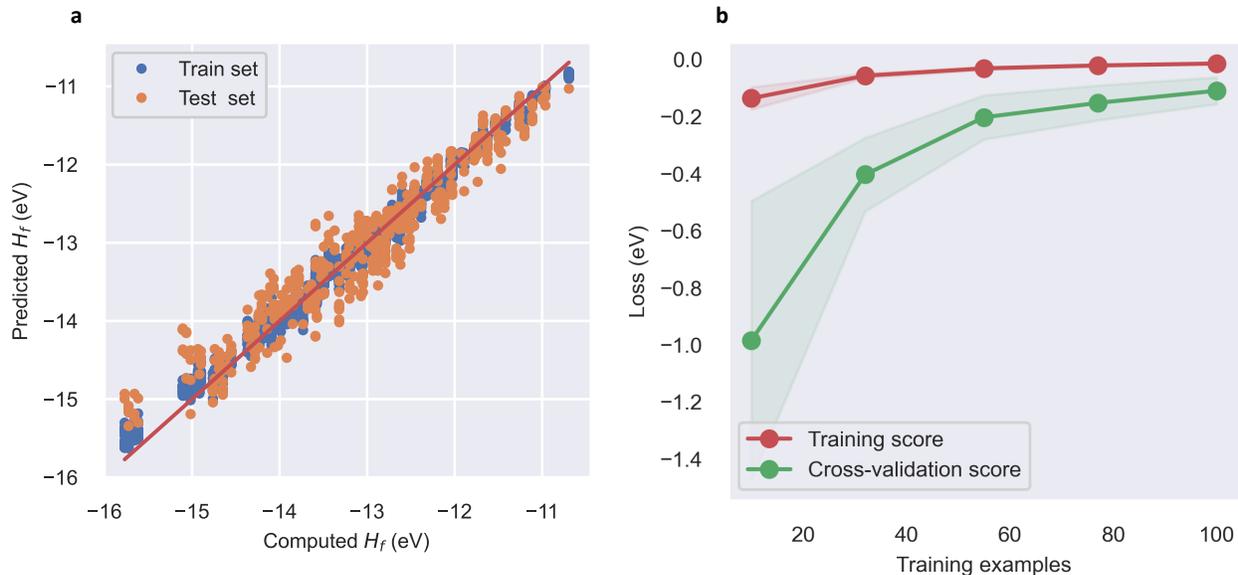
Figure S1: (a) Performance of the random forest regressor model trained for the prediction of the enthalpy of formation. The predicted values are presented against the DFT-computed data, representing the solid red line the ideal performance. (b) Performance of the model during its training for ground state energy prediction. For both training and test sets, the negative mean squared error as loss functions is represented versus the training epoch. A cross-validation strategy with 20 folds is implemented, obtaining mean absolute errors of 0.049 and 0.064 eV for the training and test datasets, respectively.
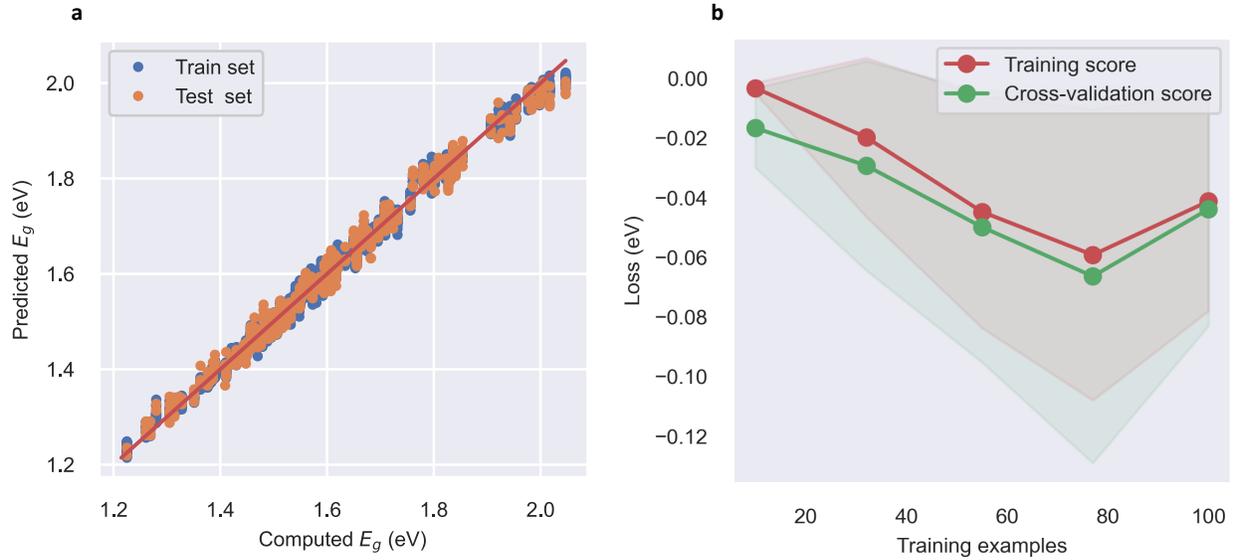
Figure S2: (a) Performance of the random forest regressor model trained for the prediction of the enery band gap. The predicted values are presented against the DFT-computed data, representing the solid red line the ideal performance. (b) Performance of the model during its training for the energy band gap prediction. For both training and test sets, the negative mean squared error as loss functions is represented versus the training epoch. A cross-validation strategy with 20 folds is implemented, obtaining mean absolute errors of 0.0063 and 0.0043 eV for the training and test datasets, respectively.
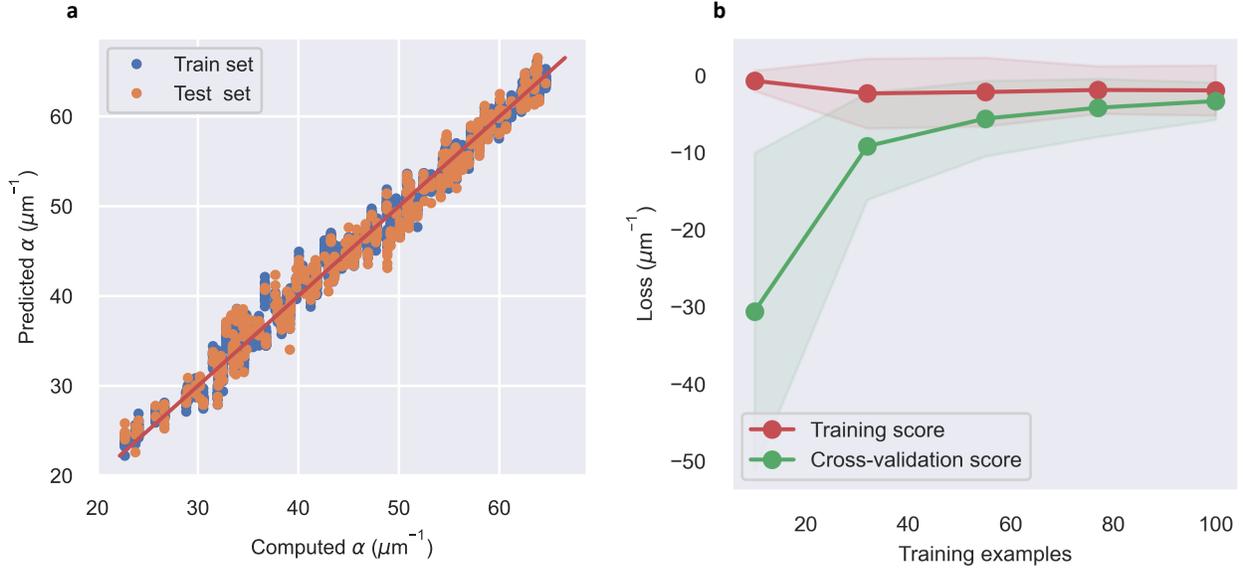
Figure S3: (a) Performance of the random forest regressor model trained for the prediction of absorption coefficient just above the band gap. The predicted values are presented against the DFT-computed data, representing the solid red line the ideal performance. (b) Performance of the model during its training for the absorption coefficient prediction. For both the training and test datasets, the negative mean squared error as loss functions is represented versus the training epoch. A cross-validation strategy with 20 folds is implemented, obtaining mean absolute errors of 0.56 and 0.36 μ m$_1$ for the training and test datasets, respectively.
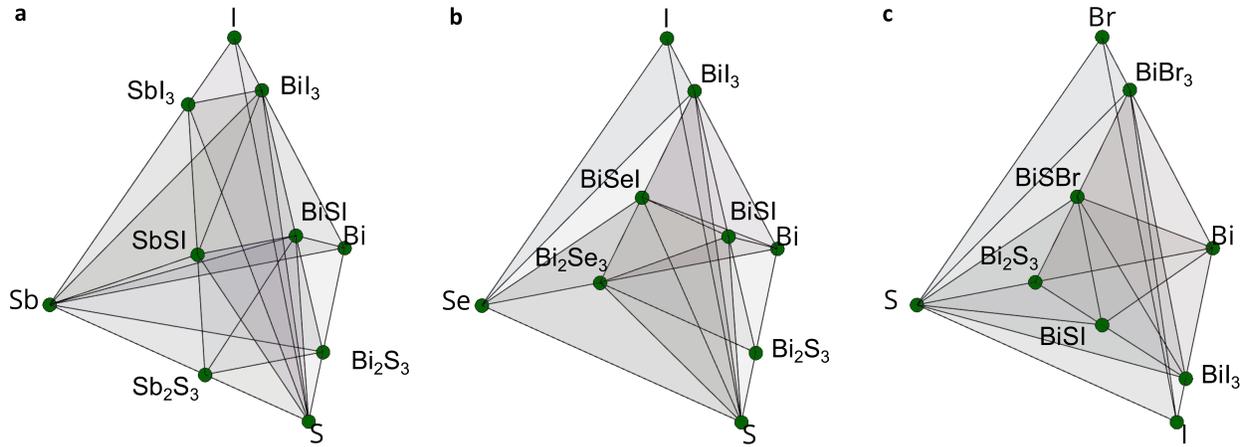
Figure S4: Convex-hull representations for the MChX-based solid solutions (a) $Bi_xSb_{1-x}SI$, (b) BiS_ySe_{1-y} and (c) $BiSI_zBr_{1-z}$. The plots include chemical species, secondary phases, and parent pnictogen-chalcohalides. Each line represents possible decomposition paths for each compound. A new compound located in the figure can decompose into the secondary phases at the vertices of the parallelepiped that contains it. If the energy of this compound is lower than the expected energy from that decomposition, the compound is thermodynamically stable and becomes a new point of the convex-hull diagram. Calculations were performed at the HSEsol+SOC level of theory.
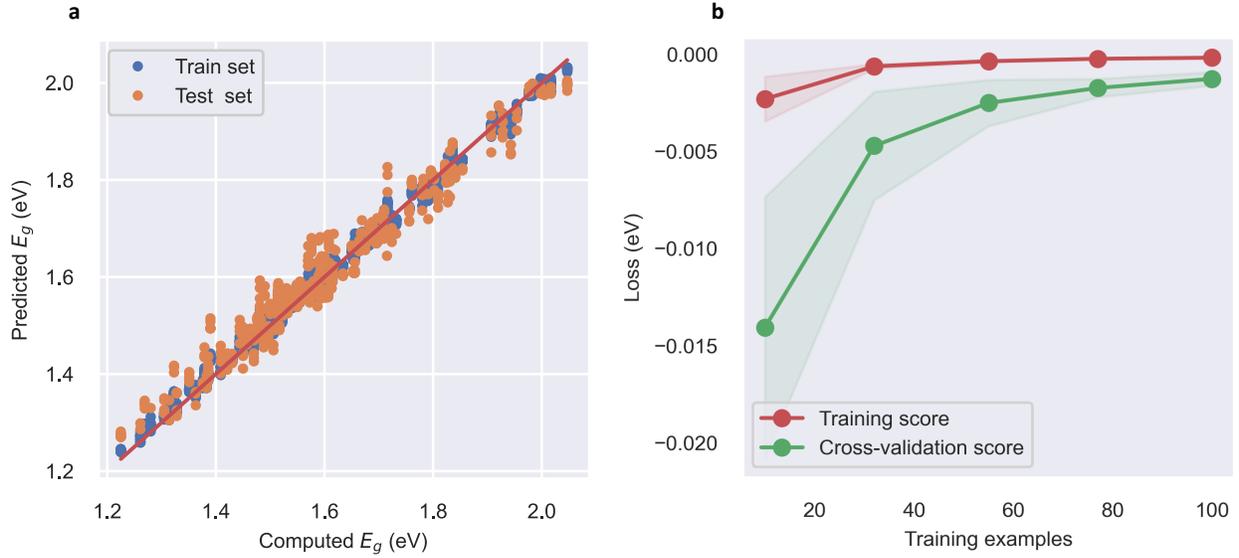
Figure S5: (a) Performance of the convolutional neural network model trained for the prediction of the energy band gap. The predicted values are presented against the DFT-computed data, representing the solid red line the ideal performance. (b) Performance of the model during its training for the energy band gap prediction. For both the training and test datasets, the negative mean squared error as loss functions is represented versus the training epoch. A cross-validation strategy with 20 folds is implemented, obtaining mean absolute errors of 0.0054 and 0.0071 eV for the training and test datasets, respectively.
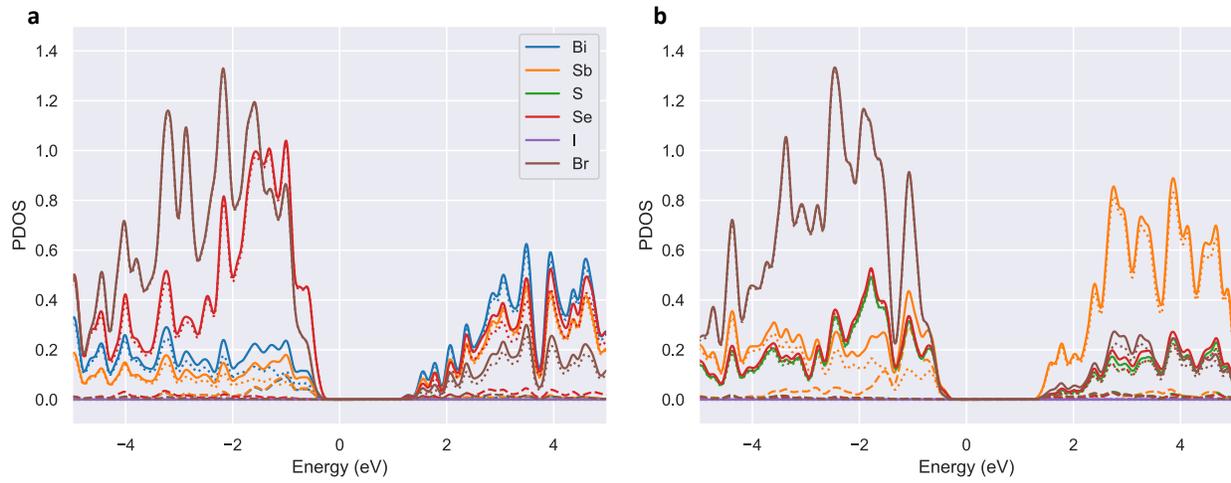
Figure S6: Partial density of states for each atomic species in the MChX solid solutions (a) $Bi_{0.5}Sb_{0.5}SeBr$ and (b) $SbS_{0.5}Se_{0.5}Br$. Dashed and doted lines correspond to s and p orbital contributions, respectively, while solid lines represent the total DOS.
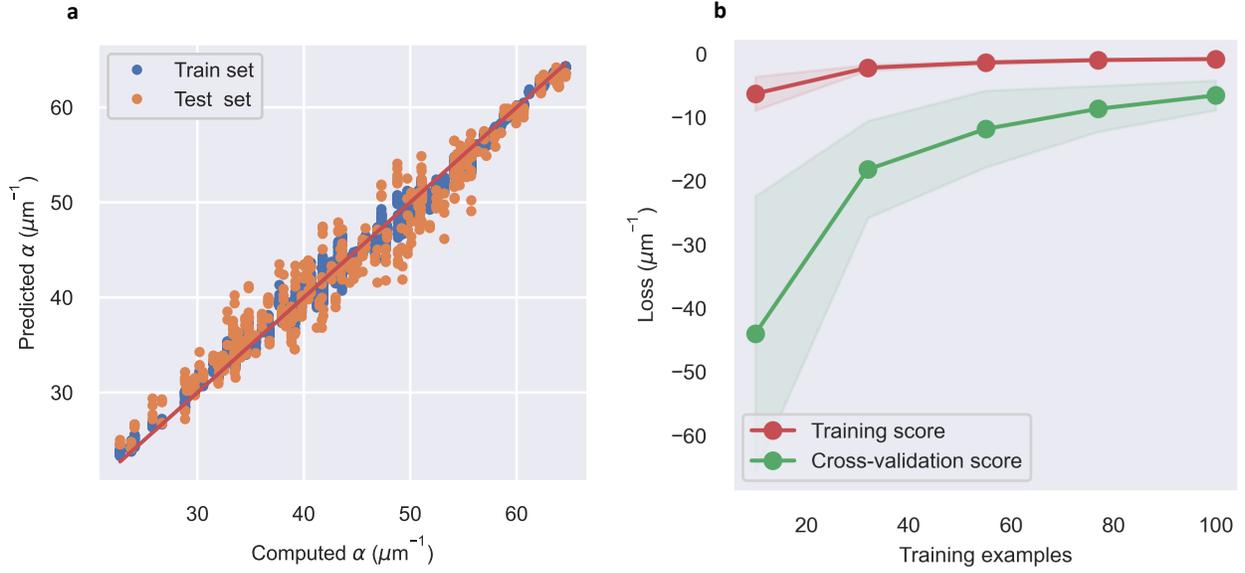
Figure S7: (a) Performance of the convolutional neural network model trained for the prediction of absorption coefficient just above the energy band gap. The predicted values are presented against the DFT-computed data, representing the solid red line the ideal performance. (b) Performance of the model during its training for the absorption coefficient prediction. For both the training and test datasets, the negative mean squared error as loss functions is represented versus the training epoch. A cross-validation strategy with 20 folds is implemented, obtaining mean absolute errors of 0.38 and 0.48 μ m$_1$ for both the training and test datasets, respectively.
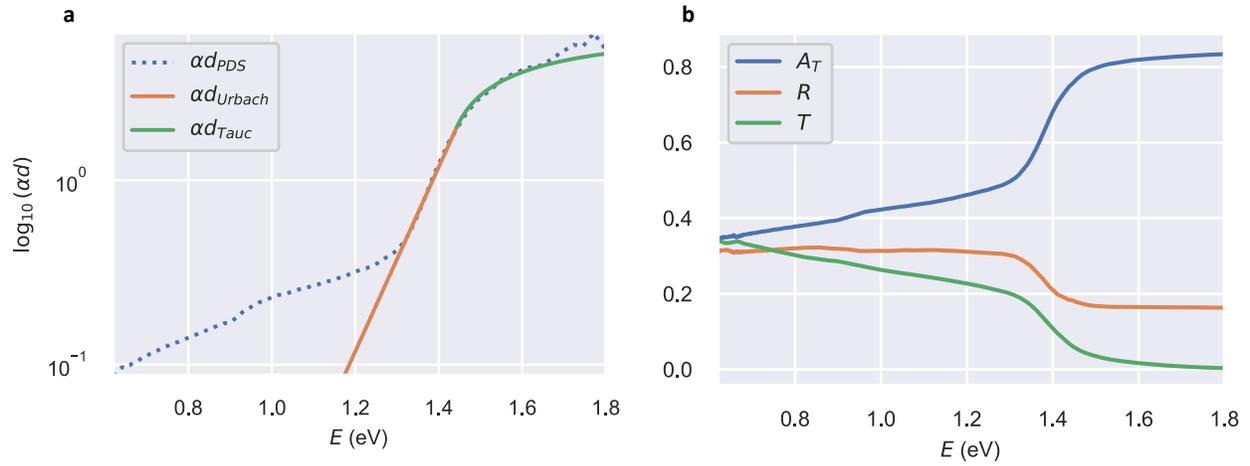
Figure S8: Experimental measurements performed for the synthesized MChX solid solution. (a) Energy band gap, with the PDS technique. (b) Optical properties (transmissivity, reflectivity and absorptivity).
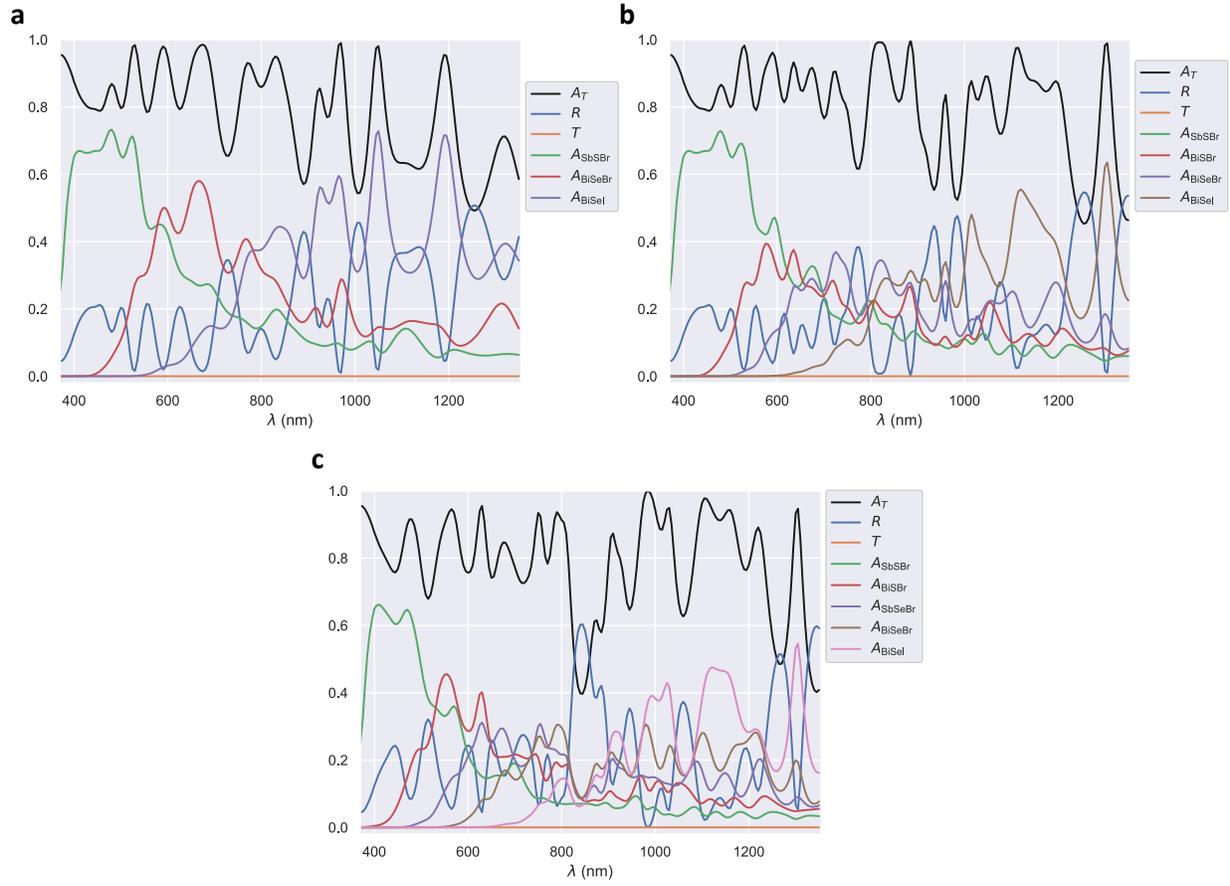
Figure S9: Absorptivities, reflectivities and transmissivities of the optimal multi-junction MChX solar cell devices shown in Table 3 of the main text. Results are obtained considering (a) 3, (b) 4 and (c) 5 MChX absorber layers.