Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2025

Supplementary Information: Predicting the excited-state properties of crystalline organic semiconductors using GW+BSE and machine learning

Siyu Gao[£,†] Yiqun Luo[£,‡] Xingyu Liu,[†] and Noa Marom[*,†,‡,¶]

†*Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213*

‡*Department of Physics, Carnegie Mellon University, Pittsburgh, PA, 15213*

¶*Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, 15213*

E-mail: nmarom@andrew.cmu.edu

£ These authors contributed equally to this work.

Cost and Accuracy of SISSO models

Supplementary Table 1: Cost and accuracy of SISSO models. The cost is given in multiples of the computer time required for a PBE calculation of a single molecule in the ground state. The cost of each model is a sum over the costs of all the primary features it contains.

Property	Model	Relative cost	Train RMSE / eV	Test RMSE / eV	Test 2 RMSE / eV
Singlet excitation energy	$M_{1,1}$	150	0.201538	0.301245	0.458352
	$M_{2,1}$	150	0.171414	0.234759	0.424820
	$M_{3,1}$	151	0.157605	0.284217	0.180598
	$M_{4,1}$	154	0.151685	0.269321	0.337210
	$M_{1,2}$	150	0.173075	0.230781	0.408663
	$M_{2,2}$	187	0.155588	0.255873	0.392779
	$M_{3,2}$	187	0.134945	0.254811	0.390204
	$M_{4,2}$	190	0.118918	0.222241	0.356472
	$M_{1,3}$	154	0.160548	0.24449	0.304954
	$M_{2,3}$	192	0.132806	0.222161	0.334729
	$M_{3,3}$	192	0.111429	0.225212	0.324774
	$M_{4,3}$	285	0.095924	0.215078	0.366502
Triplet excitation energy	$M_{1,1}$	5	0.159575	0.228299	0.096798
	$M_{2,1}$	153	0.125877	0.224746	0.107431
	$M_{3,1}$	5	0.110092	0.220599	0.127470
	$M_{4,1}$	189	0.104579	0.204671	0.128824
	$M_{1,2}$	5	0.132829	0.223955	0.095949
	$M_{2,2}$	135	0.109783	0.206172	0.149406
	$M_{3,2}$	190	0.093099	0.169798	0.264473
	$M_{4,2}$	283	0.08229	0.185558	0.339817
	$M_{1,3}$	186	0.110615	0.189752	0.179680
	$M_{2,3}$	192	0.085168	0.193157	0.214821
	$M_{3,3}$	192	0.071477	0.221533	0.356647
	$M_{4,3}$	325	0.062224	0.143991	0.230104
Singlet-triplet gap	$M_{1,1}$	184	0.169191	0.16692	0.180974
	$M_{2,1}$	6	0.141445	0.140467	0.264898
	$M_{3,1}$	41	0.120577	0.12479	0.279805
	$M_{4,1}$	155	0.11456	0.140745	0.394192
	$M_{1,2}$	6	0.142564	0.147129	0.261600
	$M_{2,2}$	40	0.115737	0.139387	0.237051
	$M_{3,2}$	224	0.107635	0.137664	0.236437
	$M_{4,2}$	465	0.10035	0.13648	0.349764
	$M_{1,3}$	42	0.119323	0.125497	0.274649
	$M_{2,3}$	223	0.098058	0.126694	0.230956
	$M_{3,3}$	371	0.083525	0.123683	0.286843
	$M_{4,3}$	465	0.072515	0.303889	6.584734
Siglet binding energy	$M_{1,1}$	2	0.172525	0.277835	0.228964
	$M_{2,1}$	36	0.158016	0.2825	0.263096
	$M_{3,1}$	186	0.146498	0.284399	0.256042
	$M_{4,1}$	38	0.140104	0.273232	0.2389262
	$M_{1,2}$	35	0.161635	0.257793	0.202655
	$M_{2,2}$	38	0.142992	0.270665	0.204340
	$M_{3,2}$	173	0.128592	0.236831	0.201275
	$M_{4,2}$	186	0.118792	0.270636	0.364399
	$M_{1,3}$	186	0.149494	0.327204	0.364399
	$M_{2,3}$	231	0.121746	0.2212	0.310614
	$M_{3,3}$	324	0.101331	0.249754	0.318029
	$M_{4,3}$	324	0.091123	0.265422	1.083569

SISSO Models for the Optical Gap

$$M_{1,1} = 7.87 \times E_T^C / IP^S + 0.66$$

$$M_{2,1} = 7.54 \times E_T^C / IP^S - 1.58 \times \ln(\rho^C) + 0.38$$

$$M_{3,1} = 7.02 \times Gap^S / IP^S - 2.15 \times \rho^C + 0.000370 \times E_T^C \times AtomNum^C + 2.20$$

$$M_{4,1} = 7.13 \times Gap^S / IP^S - 0.833 \times e^{\rho^C} + 0.000328 \times E_T^S \times AtomNum^C - 0.818 \times E_T^S / E_T^C + 3.16$$

$$M_{1,2} = 16.1 \times \frac{E_T^C}{IP^S \times e^{\rho^C}} + 0.81$$

$$M_{2,2} = 17.0 \times \frac{E_T^C}{IP^S \times e^{\rho^C}} + 1.80 \times \frac{Gap^C}{AtomNum^C \times (E_T^S - Gap^S)} + 0.82$$

$$M_{3,2} = -0.611 \times \frac{E_T^C}{\rho^C \times (CB_{disp}^C - IP^S)} + 1.85 \times \frac{E_T^C}{AtomNum^C \times (E_T^S - Gap^S)} - 0.00555 \times \frac{Gap^C \times CB_{disp}^C}{|E_T^S - Gap^C|} + 0.79$$

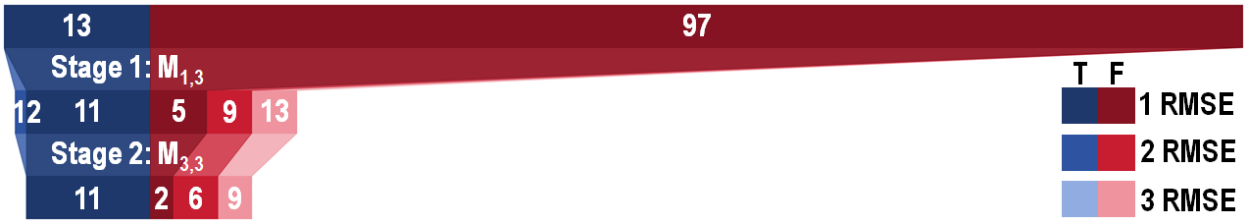
$$M_{4,2} = -5.53 \times \frac{E_T^C}{\rho^C \times (CB_{disp}^C - IP^S)} + 4.73 \times \frac{\rho^C}{AtomNum^C \times (E_T^S - Gap^S)} - 0.00525 \times \frac{Gap^C \times CB_{disp}^C}{|E_T^S - Gap^C|} - 0.108 \times \left|EA^S / Gap^S - VB_{disp}^C / EA^S\right| + 1.09$$

$$M_{1,3} = -7.32 \times \frac{E_T^C \times \ln(Gap^S / \rho^C)}{(E_T^S - IP^S) \times E_T^S} + 1.38$$

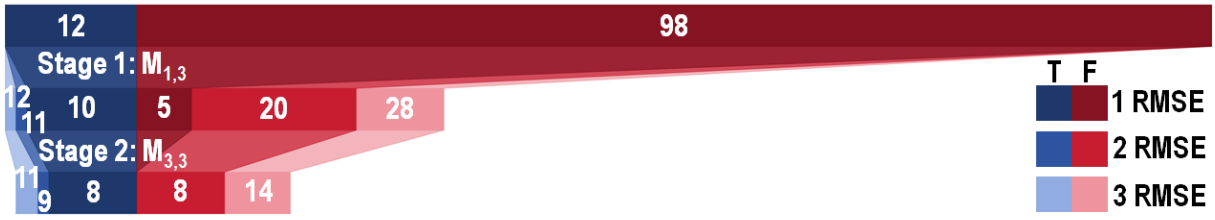
$$M_{2,3} = -7.32 \times \frac{E_T^C \times \ln(Gap^S / \rho^C)}{(E_T^S - IP^S) \times E_T^S} - 0.986 \times \frac{(EA^S + CB_{disp}^C) \times AtomNum^C}{PolarTensor^S \times (CB_{disp}^C - EA^S - Gap^S + Gap^C)} + 1.22$$

$$M_{3,3} = -7.33 \times \frac{E_T^C \times \ln(Gap^S / \rho^C)}{(E_T^S - IP^S) \times E_T^S} - 0.939 \times \frac{(EA^S + CB_{disp}^C) \times AtomNum^C}{PolarTensor^S \times (CB_{disp}^C - EA^S - Gap^S + Gap^C)} - 0.0641 \times \frac{1}{MolWt^S \times |E_T^S - Gap^C| \times |(E_T^C)^2 - Gap^C \times Gap^S|}$$

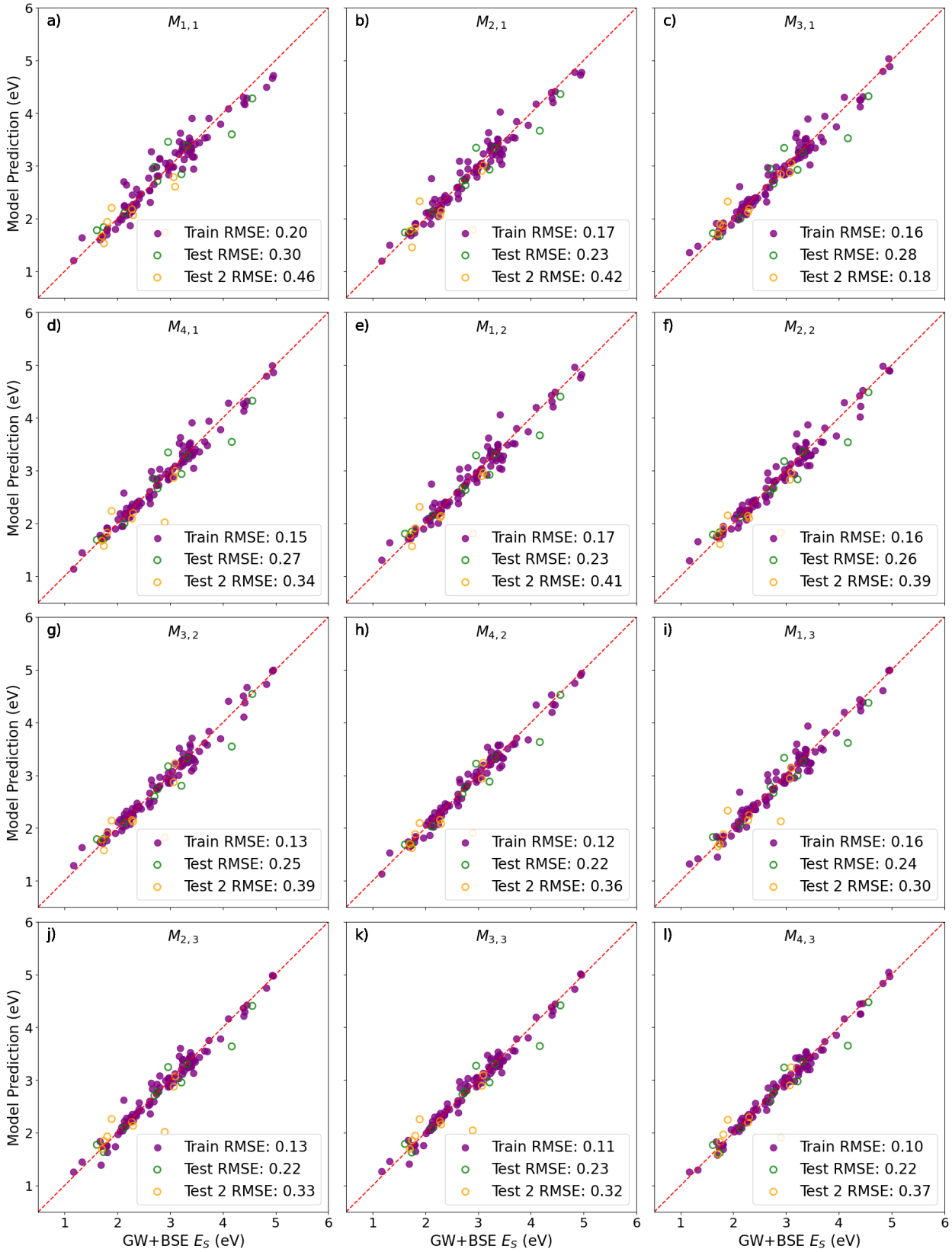
$$
\begin{aligned}
&+ 1.25 \\
M_{4,3} = &-7.57 \times \frac{E_T^C \times \ln(Gap^S/\rho^C)}{(E_T^C - IP^S) \times E_T^S} \\
&- 0.000409 \times \frac{H_{ab} + CB_{disp}^C}{|E_T^S - Gap^C| \times |E_T^C/Gap^C - Gap^S/E_T^C|} \\
&- 0.959 \times \frac{(EA^S + CB_{disp}^C) \times AtomNum^C}{PolarTensor^S \times (CB_{disp}^C - EA^S - Gap^S + Gap^C)} \\
&- 0.0000398 \times \frac{\left|E_T^C - Gap^C\right| \times MolWt^S}{E_T^S \times \left|E_T^C + CB_{disp}^C - Gap^S + VB_{disp}^C\right|} + 1.22
\end{aligned}
$$



Supplementary Figure 1: A two-stage screening workflow for materials with an optical gap in the red range of 1.6-2.0 eV. The first stage is $M_{1,3}$ and the second stage is $M_{3,3}$. The number of true positives (shades of blue) and the number of false positives (shades od red/pink) that pass each stage of screening is shown when the thresholds are set to one, two, and three times the training set RMSE of each model. In each case, $n/2\times$ RMSE is applied on either end of the target energy range.



Supplementary Figure 2: A two-stage screening workflow for materials with an optical gap in the blue range of 2.5-2.8 eV. The first stage is $M_{1,3}$ and the second stage is $M_{3,3}$. The number of true positives (shades of blue) and the number of false positives (shades od red/pink) that pass each stage of screening is shown when the thresholds are set to one, two, and three times the training set RMSE of each model. In each case, $n/2\times$ RMSE is applied on either end of the target energy range.



Supplementary Figure 3: SISSO model predictions as a function of the GW+BSE reference data for the optical gap. The filled purple circles represent the training set and the open green circles represent the test set.

SISSO Models for the Triplet Excitation Energy

$$M_{1,1} = 7.9 \times (E_T^S/IP^S) - 0.21$$

$$M_{2,1} = 7.3 \times (E_T^S/IP^S) + 0.00047 \times (E_T^C \times AtomNum^C) - 0.19$$

$$M_{3,1} = 6.5 \times (E_T^S/IP^S) + 0.00051 \times (E_T^S \times AtomNum^C) + 0.0091 \times exp(E_T^S) - 0.052$$

$$M_{4,1} = 6.3 \times (E_T^S/IP^S) + 0.00052 \times (E_T^S \times AtomNum^C) + 0.0092 \times (E_T^C)^3) + 0.078(VB_{disp}^C/EA^S) - 0.014$$

$$M_{1,2} = 0.16 \times (E_T^S \times \ln(PolarTensor^S)) + 0.040$$

$$M_{2,2} = 0.081 \times ((E_T^S + Gap^S) \times \ln(PolarTensor^S)) - 15 \times \frac{VB_{disp}^C/AtomNum^C}{|EA^S - H_{ab}|} - 0.13$$

$$M_{3,2} = 0.079 \times ((E_T^C + E_T^S) \times \ln(PolarTensor^S)) + 55 \times \frac{(CB_{disp}^C)^3}{EA^S \times (AtomNum^C)} - 0.0000053 \times ((Gap^C - Gap^S) \times (AtomNum^C)^2) - 0.035$$

$$M_{4,2} = 1.2 \times ((EA^S + E_T^C) \times (Gap^S/IP^S)) + 0.089 \times ((E_T^C + VB_{disp}^C) \times \ln(AtomNum^C)) + 1.8 \times \frac{(CB_{disp}^C)^3}{EA^S \times E_T^S} + 0.21 \times \frac{H_{ab} - VB_{disp}^C}{EA^S + CB_{disp}^C} - 0.019$$

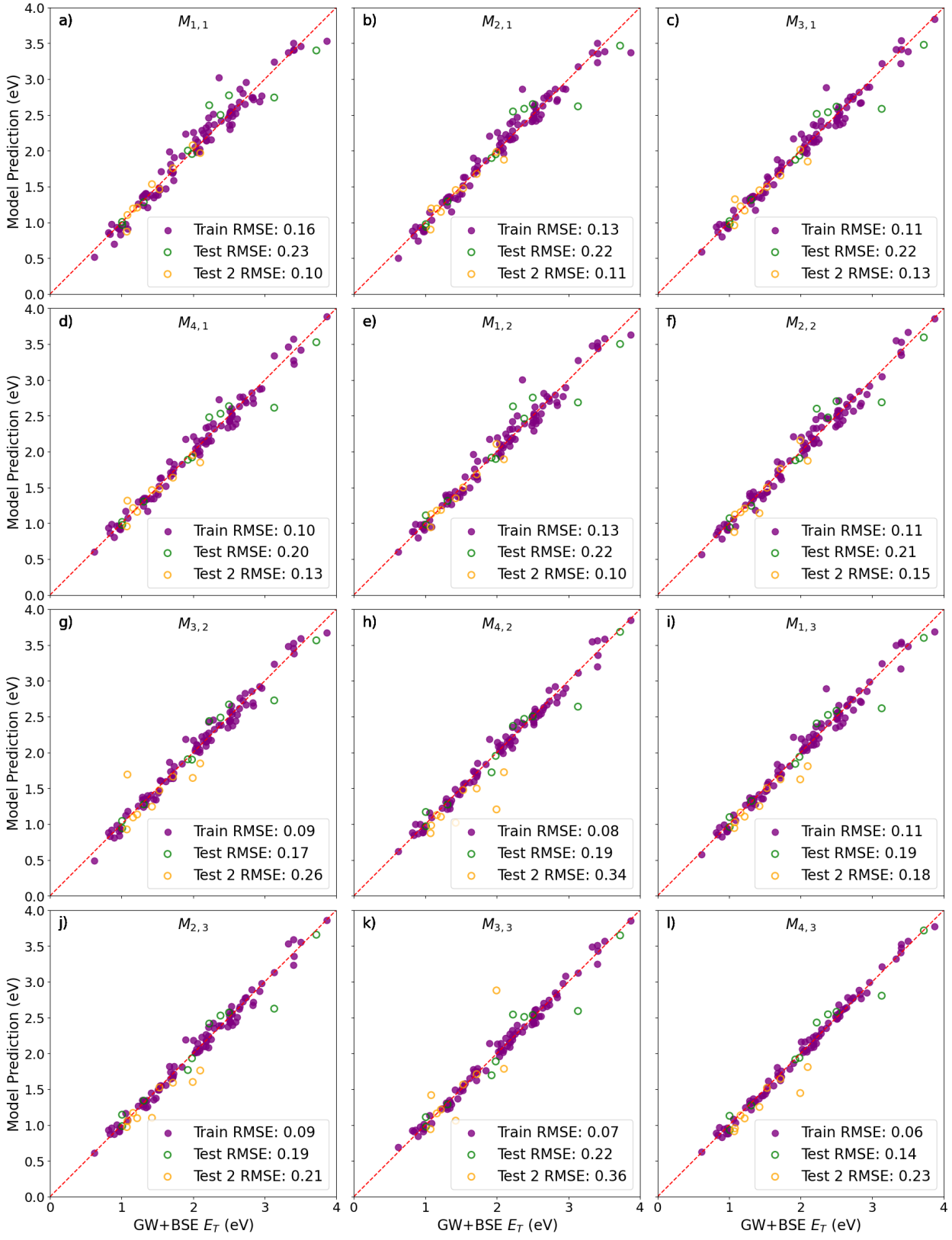
$$M_{1,3} = 0.089 \times ((E_T^C - CB_{disp}^C) \times \ln(AtomNum^C) + (E_T^S + CB_{disp}^C) \times \ln(PolarTensor^S)) + 0.029$$

$$M_{2,3} = 0.090 \times ((E_T^C - CB_{disp}^C) \times \ln(AtomNum^C) + (E_T^S + CB_{disp}^C) \times \ln(PolarTensor^S)) - 0.28 \times \frac{\left|IP^S - Gap^S - |EA^S - Gap^C|\right|}{(EA^S \times IP^S) + (E_T^S \times CB_{disp}^C)} + 0.15$$

$$M_{3,3} = 0.090 \times ((E_T^C - CB_{disp}^C) \times \ln(AtomNum^C) + (E_T^S + CB_{disp}^C) \times \ln(PolarTensor^S)) + 0.098 \times \frac{(E_T^S - I_P^S) + |EA^S - Gap^C|}{|(E_T^S - EA^S) - (EA^S + Gap^S)|}$$

$$+ 1.6 \times \left| \frac{|E_T^C - Gap^C|}{exp(Gap^C)} - \left| |E_T^C - E_T^S| - |E_T^C - Gap^C| \right| \right| + 0.10$$

$$M_{4,3} = 0.093 \times ((E_T^C + CB_{disp}^C) \times \ln(PolarTensor^S) - (CB_{disp}^C - E_T^S) \times \ln(AtomNum^C))$$
$$+ 83 \times \frac{(CB_{disp}^C \times \epsilon^C)/(AtomNum^C)^2}{(EA^S)^2 - (H_{ab} \times Gap^C)}$$
$$+ 0.0000039 \times \frac{(EA^S - Gap^S) \times (VB_{disp}^C \times PolarTensor^S)}{|(E_T^C + CB_{disp}^C) - (Gap^S - VB_{disp}^C)|}$$
$$+ 0.080 \times \left| \frac{E_T^S \times CB_{disp}^C}{Gap^S - E_T^C} - ((H_{ab} + CB_{disp}^C) \times \ln(PolarTensor^S)) \right| - 0.082$$



Supplementary Figure 4: SISSO model predictions as a function of the GW+BSE reference data for the triplet exciton energy. The filled purple circles represent the training set and the open green circles represent the test set.

SISSO Models for the Singlet Triplet Gap

$$M_{1,1} = 1.0 \times (\Delta E_{ST}^{C} + \Delta E_{ST}^{S}) + 0.64$$

$$M_{2,1} = -3.6 \times (\sqrt{\rho^{C}}) - 0.26 \times (EA^{S} + E_{T}^{S})) + 4.83$$

$$M_{3,1} = -2.9 \times (\sqrt{\rho^{C}}) - 0.44 \times (EA^{S} + E_{T}^{S}) - 0.054 \times \frac{IP^{S}}{Gap^{C}} + 5.0$$

$$M_{4,1} = -0.28 \times (EA^{S} + Gap^{S}) + 2.7 \times \frac{Gap^{S}}{E_{T}^{S}} - 0.62 \times (Gap^{S} \times \rho^{C}) + 1.3 \times \ln(E_{T}^{C}) - 1.05$$

$$M_{1,2} = -0.41 \times ((\rho^{C})^{2} \times (EA^{S} + E_{T}^{S})) + 1.6$$

$$M_{2,2} = -0.52 \times (\rho^{C} \times (EA^{S} + Gap^{S})) - 0.036 \times \frac{Gap^{S}/\Delta E_{ST}^{S}}{(Gap^{C})^{2}} + 2.3$$

$$M_{3,2} = -0.52 \times (\rho^{C} \times (EA^{S} + Gap^{S})) - 0.037 \times \frac{Gap^{S}/\Delta E_{ST}^{S}}{(Gap^{C})^{2}} - 0.00043 \times \frac{\Delta E_{ST}^{C}/VB_{disp}^{C}}{E_{T}^{S} - Gap^{C}} + 2.3$$

$$M_{4,2} = -0.48 \times \frac{(\rho^{C})^{2}}{\sqrt[3]{\Delta E_{ST}^{S}}} + 3.6 \times \frac{E_{T}^{C}/Gap^{S}}{EA^{S} + E_{T}^{S}} + 0.0077 \times \frac{\left|H_{ab} - \Delta E_{ST}^{C}\right|}{E_{T}^{S} - Gap^{C}} - 0.0000046 \times \frac{AtomNum^{C}/EA^{S}}{\Delta E_{ST}^{C} + \Delta E_{ST}^{S}} + 0.29$$

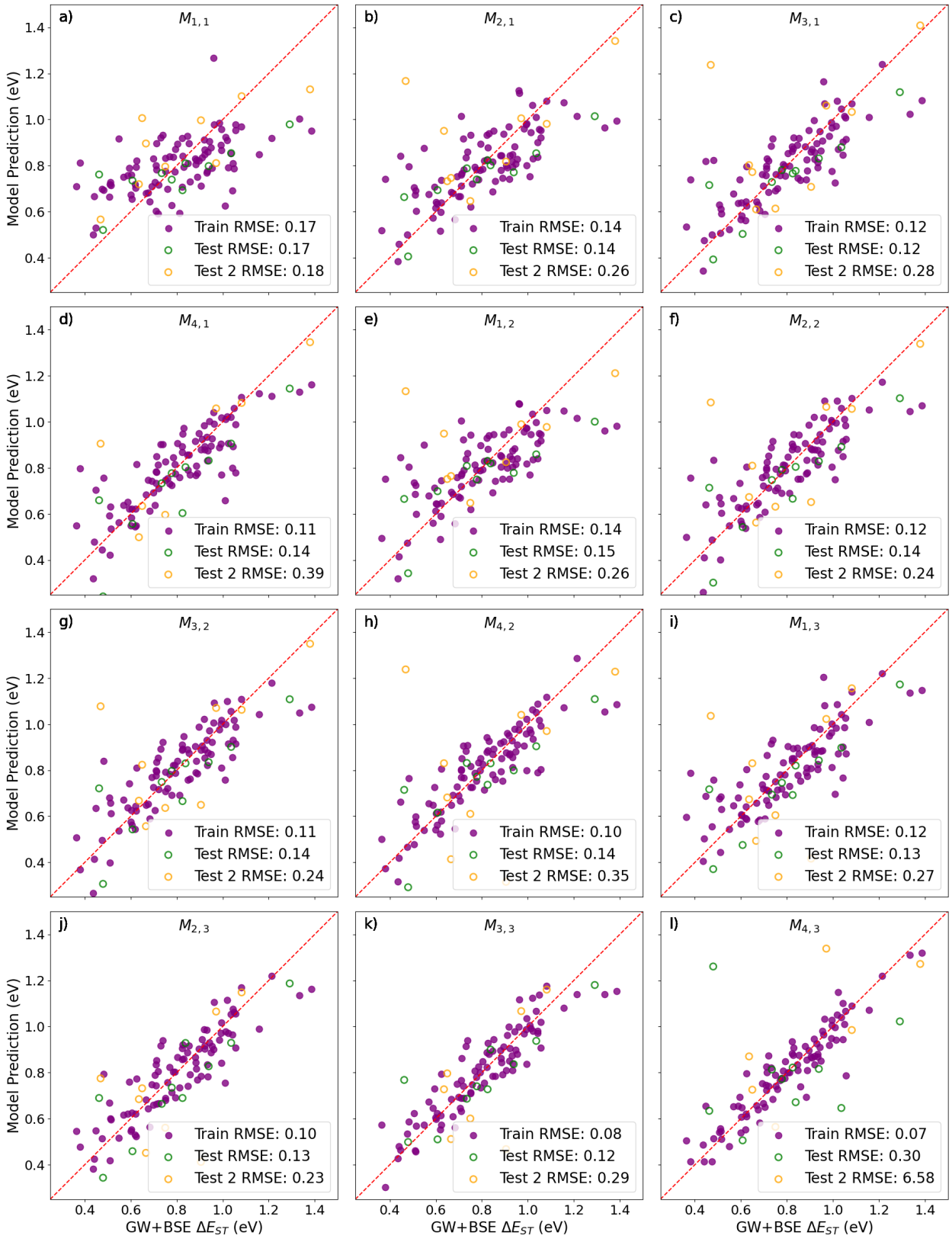
$$M_{1,3} = -5.3 \times \frac{\sqrt{Gap^{C}/\rho^{C}}}{(\Delta E_{ST}^{S} - E_{T}^{S}) - (EA^{S} + E_{T}^{S})} - 0.88$$

$$M_{2,3} = -130 \times \frac{\sqrt{Gap^{C}}/exp(\rho^{C})}{(\Delta E_{ST}^{S} - E_{T}^{S}) - (EA^{S} + E_{T}^{S})} + 0.12 \times \frac{\Delta E_{ST}^{C} \times CB_{disp}^{C}/(E_{T}^{S} - Gap^{C})}{(VB_{disp}^{C}/\Delta E_{ST}^{C}) + VB_{disp}^{C}/CB_{disp}^{C}} - 0.86$$

$$M_{3,3} = -121 \times \frac{\sqrt{Gap^{C}}/exp(\rho^{C})}{(\Delta E_{ST}^{S} - E_{T}^{S}) - (EA^{S} + E_{T}^{S})} - 0.0021 \frac{(EA^{S} - Gap^{C})/(\Delta E_{ST}^{C} + \Delta E_{ST}^{S})}{(E_{T}^{S} - Gap^{C})/(E_{T}^{C} - E_{T}^{S})} + 0.071 \frac{(EA^{S} \times \Delta E_{ST}^{C})/(\Delta E_{ST}^{C} + CB_{disp}^{C})}{(E_{T}^{C}/EA^{S}) + (EA^{S}/\Delta E_{ST}^{C})} - 0.73$$

$$M_{4,3} = 0.45 \times ((\Delta E_{ST}^{S} - EA^{S}) - (E_{T}^{S} - H_{ab}) \times (\rho^{C}/Gap^{C}) \times (Gap^{S} + \Delta E_{ST}^{S}))$$

$$+ 0.000019 \times \frac{(E_T^S)^3/|E_T^S - Gap^C|}{(CB_{disp}^C + VB_{disp}^C) - |EA^S - H_{ab}|}$$
$$+ 0.034 \times \frac{(CB_{disp}^C)^2/(\Delta E_{ST}^C + CB_{disp}^C)}{\big||EA^S - E_T^S| - (EA^S + CB_{disp}^C)\big|}$$
$$+ 0.011 \times (\frac{|CB_{disp}^C - VB_{disp}^C|/|EA^S - CB_{disp}^C|}{\big||EA^S - E_T^C| - (CB_{disp}^C + VB_{disp}^C)\big|} + 1.95)$$



Supplementary Figure 5: SISSO model predictions as a function of the GW+BSE reference data for the singlet-triplet gap. The filled purple circles represent the training set and the open green circles represent the test set.

SISSO Models for the Singlet Exciton Binding Energy

$$M_{1,1} = 0.24 \times (IP^S/\rho^C) - 1.3$$

$$M_{2,1} = 0.037 \times (IP^S \times Gap^S) - 0.0049 \times (CB^C_{disp} \times AtomNum^C) + 0.19$$

$$M_{3,1} = 0.074 \times (E^S_T \times IP^S) - 0.42 \times (E^C_T \times \rho^C)$$
$$- 0.0041 \times (CB^C_{disp} \times AtomNum^C) + 0.40$$

$$M_{4,1} = 0.072 \times (E^S_T \times IP^S) - 0.41 \times (Gap^C \times Rho^C)$$
$$- 0.0048 \times (CB^C_{disp} \times AtomNum^C) - 1.6 \times (VB^C_{disp})^3 + 0.40$$

$$M_{1,2} = -0.34 \times \frac{(CB^C_{disp} - IP^S)}{\sqrt{\rho^C}} - 1.8$$

$$M_{2,2} = 0.012 \times (E^S_T + IP^S) \times (IP^S/\rho^C)$$
$$+ 0.020 \times ((E^S_T - Gap^S) \times (CB^C_{disp} \times AtomNum^C)) - 0.17$$

$$M_{3,2} = -0.00061 \times ((IP^S)^3 \times (CB^C_{disp} - Gap^C))$$
$$- 0.17 \times \frac{(CB^C_{disp} \times \epsilon^C)}{(H_{ab} + CB^C_{disp})} - 3.0 \times \frac{(AtomNum^C)^{-1}}{(E^S_T - Gap^S)} + 0.63$$

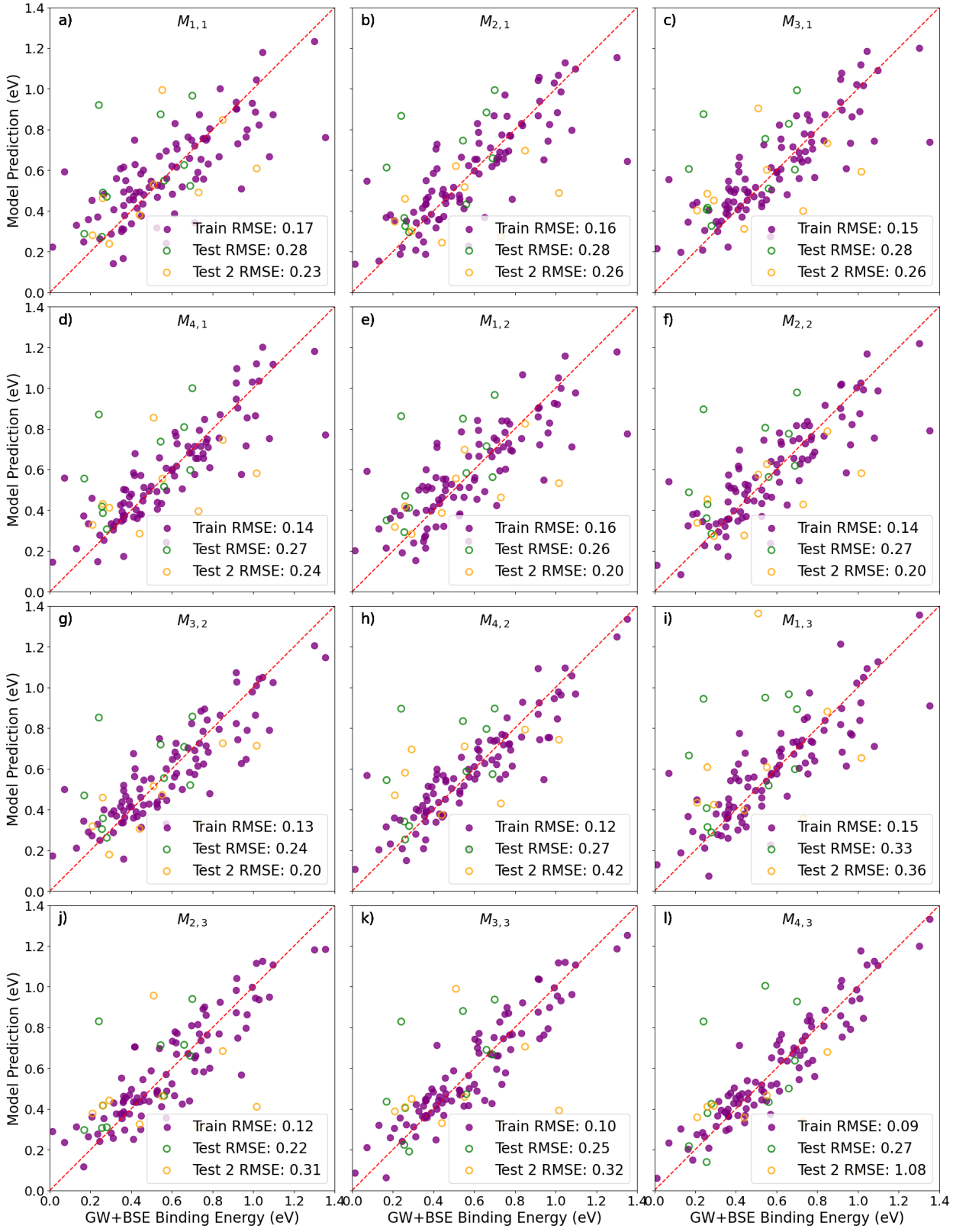
$$M_{4,2} = 0.011 \times (IP^S + Gap^C) \times (IP^S/\rho^C) - 5.4 \times \frac{(\rho^C/AtomNum^C)}{(E^S_T - Gap^S)}$$
$$- 0.014 \times \frac{|E^C_T - Gap^C|}{(E^S_T - Gap^C)} - 1.0 \times \frac{(CB^C_{disp} - Gap^S)}{E^C_T + VB^C_{disp}} - 1.3$$

$$M_{1,3} = 0.80 \times \left[\frac{(E^S_T + IP^S)}{\ln(AtomNum^C)} + |E^C_T - E^S_T| - (CB^C_{disp} + VB^C_{disp})\right]$$
$$- 0.54$$

$$M_{2,3} = 5.2 \times \frac{|(Gap^S)^2 - (IP^S \times CB^C_{disp})|}{\sqrt[3]{MolWt^S} \times (Gap^S \times \epsilon^C)}$$
$$+ 0.033 \times \frac{(E^C_T - E^S_T)/(E^C_T - Gap^S)}{\ln(|EA^S - CB^C_{disp}|)} + 0.24$$

$$M_{3,3} = 5.2 \times \frac{|(Gap^S)^2 - (IP^S \times CB^C_{disp})|}{\sqrt[3]{MolWt^S} \times (Gap^S \times \epsilon^C)}$$
$$+ 0.035 \times \frac{(E^C_T - E^S_T)/(E^C_T - Gap^S)}{\ln(|EA^S - CB^C_{disp}|)}$$

$$
\begin{aligned}
&+ 0.49 \times \frac{(H_{ab} + VB^{C}_{disp}) - |CB^{C}_{disp} - VB^{C}_{disp}|}{(E^{S}_{T}/VB^{C}_{disp}) - (VB^{C}_{disp}/H_{ab})} + 0.25 \\
M_{4,3} = 5.1 &\times \frac{|(Gap^{S})^{2} - (IP^{S} \times CB^{C}_{disp})|}{\sqrt[3]{MolWt^{S}} \times (Gap^{S} \times \epsilon^{C})} \\
&- 0.50 \times \frac{(E^{C}_{T} - E^{S}_{T})/(E^{C}_{T} - Gap^{S})}{\ln(|EA^{S} - CB^{C}_{disp}|)} \\
&+ 6.5 \times \frac{(H^{2}_{ab}/|E^{S}_{T} - Gap^{C}|)}{(E^{S}_{T}/VB^{C}_{disp}) - (VB^{C}_{disp}/H_{ab})} \\
&- 0.15 \times \frac{(H_{ab} \times VB^{C}_{disp})/\ln(Gap^{S})}{((E^{C}_{T} - E^{S}_{T}) + |H_{ab} - CB^{C}_{disp}|)} + 0.25
\end{aligned}
$$



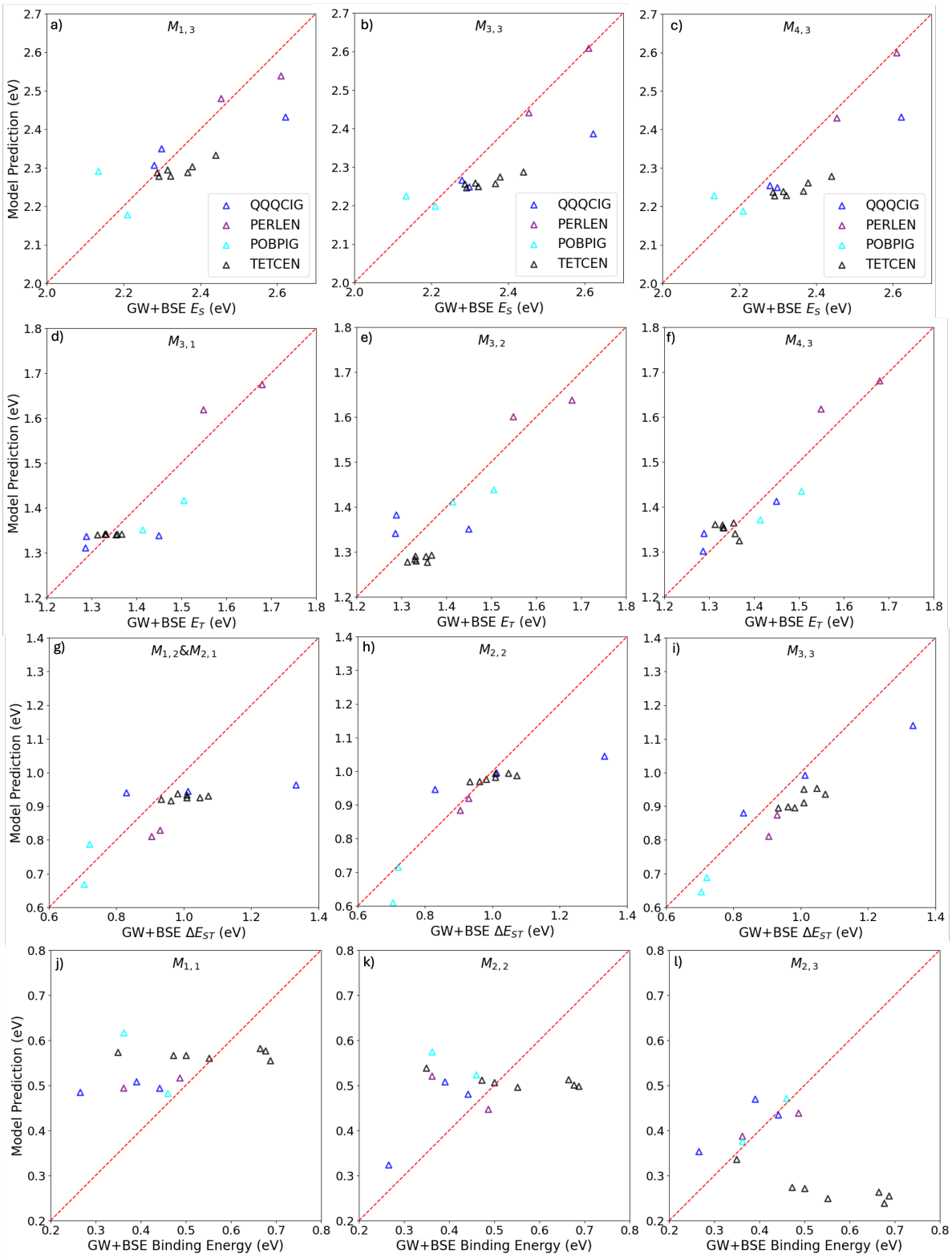
Supplementary Figure 6: SISSO model predictions as a function of the GW+BSE reference data for the singlet exciton binding energy. The filled purple circles represent the training set, the open green circles represent the test set, and the open red circle is terphenyl.

Model Performance for Polymorphs

Supplementary Figure 7 shows the correlation between the predictions of selected SISSO models and the GW+BSE reference data for the optical gap, triplet exciton energy, singlet-triplet gap, and singlet exciton binding energy for polymorphic systems. The PAH101 set contains only three materials with polymorphs: rubrene (QQQCIG), perylene (PERLEN), and diindeno[1,2,3-cd:1',2',3'-lm]perylene (POBPIG). The putative polymorphs of teracene from Ref. 1. are also shown. Overall, the model performance for polymorphs is variable. The models for the optical gap and triplet exciton energy perform better than the models for the singlet-triplet gap and singlet exciton binding energy. The performance for rubrene, perylene, and POBPIG tends to be better than the putative tetracene polymorphs. This may be attributed to the small differences in the target property values between the tetracene polymorphs combined with overfitting to the training set. We also note that models based only on single molecule features can distinguish between polymorphs only through the effect of crystal packing on the molecular conformation, which hardly changes for tetracene because it is a very rigid molecule. As previously discussed in Ref. 2, there is a very small number of polymorphic systems in the PAH101 set. Achieving better performance for polymorphs would require additional data acquisition for observed and/or putative polymorphs.

1. Tom R, Gao S, Yang Y, Zhao K, Bier I, Buchanan EA, Zaykov A, Havlas Z, Michl J, Marom N. Inverse design of tetracene polymorphs with enhanced singlet fission performance by property-based genetic algorithm optimization. Chemistry of Materials. 2023 Jan 21;35(3):1373-86

2. Wang X, Gao S, Luo Y, Liu X, Tom R, Zhao K, Chang V, Marom N. Computational Discovery of Intermolecular Singlet Fission Materials Using Many-Body Perturbation Theory. The Journal of Physical Chemistry C. 2024 May 1;128(19):7841-64



Supplementary Figure 7: Predictions of selected SISSO models for the polymorphs of rubrene (QQQCIG), perylene (PERLEN), and diindeno[1,2,3-cd:1',2',3'-lm]perylene (POBPIG) from PAH101 and the putative polymorphs of tetracene (TETCEN) as a function of the GW+BSE reference data for a-c) singlet excitation Energy, d-f) triplet excitation energy, g-i) singlet-triplet gap, and j-l)singlet exciton binding energy.

Comparison with Baseline Models

To evaluate the performance of SISSO compared to baseline models, we have trained linear regression (LR) and Gaussian process regression (GPR) models. Models for each property were trained based on the SISSO primary features and the many-body tensor representation (MBTR)[3]. The LR models were trained using the Python package scikit-learn[4] and the GPR models were trained using GPyTorch[5]. For MBTR, we used a combination of molecular and crystal representations, each containing 300 features. We set the geometry function to be angle, the number of grids to be 20, axes with minimum 0 and maximum 180, and the grid Gaussian broadening standard deviation to be 0.1. For other parameters we used the default values suggested in the MBTR documentation. For the singlet excitation energy, the triplet excitation energy, and the singlet-triplet gap, we also provide a comparison with a linear fit based only on the corresponding DFT-level approximations, namely the single molecule HOMO-LUMO gap and crystal band gap (Gap^S, Gap^C) for the singlet excitation energy, the single molecule and crystal triplet formation energy (E_T^S, E_T^C) for the triplet excitation energy, and the DFT estimate for the single molecule and crystal singlet-triplet gap (ΔE_{ST}^S, ΔE_{ST}^C), evaluated by calculating the difference between the aforementioned features. For the singlet exciton binding energy there is no corresponding DFT feature. The results are presented in Table 2.

Because the LR and GPR models based on the SISSO primary features use all the available primary features, their cost is roughly equivalent to the most complex models generated by SISSO (the LR model equations are provided below). In all cases the train and test RMSE of these baseline models are higher than SISSO models of similar cost. In

3. Huo H, Rupp M. Unified representation of molecules and crystals for machine learning. Machine Learning: Science and Technology. 2022 Nov 21;3(4):045017.

4. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

5. Gardner J, Pleiss G, Weinberger KQ, Bindel D, Wilson AG. Gpytorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration. Advances in neural information processing systems. 2018;31.

addition, their performance deteriorates, in some cases very significantly, for the test set, indicating over-fitting. When switching from the SISSO primary features to the MBTR structural representation, the LR models perform very poorly because they are extremely over-fitted to the training set. The performance of the GPR models based on MBTR is similar to the models based on the SISSO primary features, and worse than the SISSO-generated models.

Supplementary Table 2: Performance of linear regression (LR) and Gaussian process regression (GPR) models based on the SISSO primary features used in this work and the many-body tensor representation (MBTR), as well as LR models based on a single DFT feature that can be considered as a surrogate model for the target property.

Task	Descriptor	Method	Train RMSE (eV)	Test RMSE (eV)	Relative Cost
Singlet excitation energy	SISSO primary features	LR	0.170	0.242	326
		GPR	0.328	0.888	
	MBTR	LR	9.20×10^{-14}	1.169	N/A
		GPR	0.178	0.543	
	Gap^S only	LR	0.265	0.377	1
	Gap^C only		0.224	0.266	33
Triplet excitation energy	SISSO primary features	LR	0.128	0.228	326
		GPR	0.201	0.810	
	MBTR	LR	9.99×10^{-14}	1.203	N/A
		GPR	0.191	0.602	
	E_T^S only	LR	0.216	0.216	1
	E_T^C only		0.229	0.173	148
Singlet binding energy	SISSO primary features	LR	0.150	0.256	326
		GPR	0.130	0.263	
	MBTR	LR	7.58×10^{-14}	0.412	N/A
		GPR	0.189	0.288	
Singlet-triplet gap	SISSO primary features	LR	0.120	0.167	326
		GPR	0.131	0.238	
	MBTR	LR	1.57×10^{-14}	0.295	N/A
		GPR	0.163	0.168	
	ΔE_{ST}^S only	LR	0.180	0.210	3
	ΔE_{ST}^C only		0.194	0.197	181

For the singlet excitation energy, the LR models based on Gap^S and Gap^C have higher train and test RMSE values, but their computational cost is lower than the best performing

SISSO models. For the triplet excitation energy, the LR model based on E_T^S only has a similar cost to the $M_{3,1}$ SISSO model but a higher RMSE. The LR model based on E_T^C only has a comparable cost, but a higher RMSE than the $M_{3,2}$ SISSO model. For the singlet-triplet gap, the LR model based on ΔE_{ST}^S only has a similar cost to the $M_{1,2}$ and $M_{2,1}$ SISSO models, but higher train and test RMSE values. The LR model based on ΔE_{ST}^C only has a similar cost, but higher RMSE than the $M_{3,3}$ SISSO model.

Equations of the LR models trained on the SISSO primary features:

Singlet excitation energy :

$$0.215 \times Gap^C + 0.667 \times E_T^C + 0.18 \times VB_{disp}^C + 0.284 \times CB_{disp}^C - 0.263 \times H_{ab} -$$
$$-0.106 \times Gap^S + 0.398 \times E_T^S - 0.468 \times IP^S - 0.0151 \times EA^S - 0.000201 \times PolarTensor^S +$$
$$+0.000969 \times AtomNum^C - 2.24 \times \rho^C + 0.0502 \times \epsilon^C - 2.32 \times 10^{-5} \times MolWt^S + 5.35$$

Triplet excitation energy :

$$-0.324 \times Gap^C + 0.433 \times E_T^C - 0.093 \times VB_{disp}^C + 0.106 \times CB_{disp}^C + 1.02 \times H_{ab} +$$
$$+0.112 \times Gap^S + 1.14 \times E_T^S - 0.332 \times IP^S + 0.294 \times EA^S - 1.33 \times 10^{-5} \times PolarTensor^S +$$
$$+0.000722 \times AtomNumC - 0.712 \times \rho^C + 0.0651 \times \epsilon^C - 5.43 \times 10^{-5} \times MolWt^S + 1.49$$

Singlet binding energy :

$$0.244 \times Gap^C - 0.325 \times E_T^C - 0.344 \times VB_{disp}^C - 0.345 \times CB_{disp}^C + 1.24 \times H_{ab} -$$
$$-0.559 \times Gap^S + 0.553 \times E_T^S + 0.479 \times IP^S - 0.219 \times EA^S + 0.000272 \times PolarTensor^S -$$
$$-0.000501 \times AtomNum^C - 0.71 \times \rho^C - 0.0422 \times \epsilon^C - 6.92 \times 10^{-5} \times MolWt^S - 1.36$$

Singlet $-$ triplet gap :

$$0.437 \times Gap^C + 0.336 \times E_T^C + 0.273 \times VB_{disp}^C + 0.178 \times CB_{disp}^C - 1.29 \times H_{ab} -$$
$$-0.391 \times Gap^S - 0.564 \times E_T^S - 0.136 \times IP^S - 0.309 \times EA^S + 0.102 \times \Delta E_{ST}^C +$$
$$+0.173 \times \Delta E_{ST}^S - 0.000188 \times PolarTensor^S + 0.000246 * AtomNum^C - 1.53 \times \rho^C -$$
$$-0.0149 \times \epsilon^C + 3.11 \times 10^{-5} \times MolWt^S + 3.87$$