**Supporting Information**

**Artificial intelligence-based multi-objective optimization protocol for protein structure refinement**

Di Wang, Ling Geng, Yu-Jun Zhao, Yang Yang, Yan Huang, Yang Zhang, and Hong-Bin Shen

Three energy functions are used as the objectives of the multi-objective PSO optimization algorithm:

*Objective 1*: The Rosetta scoring function is based on a Bayesian separation, which describes the likelihood of a particular structure for the given sequence. Since there are many different versions of the Rosetta energy functions, we use Rosetta Score12 (13), which is the default setting in the Rosetta software.

 (S1)

Here, each term is a kind of energy, and  is the corresponding weight factors for different energy terms:  is a backbone energy term for torsional angels  (phi) and  (psi),  is a statistical *ω*-angle potential,  is a side chain rotamer energy term derived from known protein structure, and  are relevant to the Lennard-Jones interactions,  represents the solvation energy, and  is the representation of the electrostatic interaction. Four kinds of the hydrogen bonding energies are termed as  (local backbone-backbone),  (non-local backbone),  (side chain–backbone), and  (side chain–side chain). The  is a proline ring closure energy.

*Objective 2*: The RWplus potential function is shown in Eq.3, which uses an ideal random-walk chain as the reference state (20). This function can be divided into two parts: a pair-wise distance-dependent term and an orientation dependent term:

(S2)

Here, *r* is the distance between atoms of  and , A and B are two vector pairs, which are defined for each residue based on three most representative side-chain atoms and are used to represent the orientation of the side-chain atoms. is the relative orientation between vector types A and B,  is the potential for the atom pair derived from the inverse of Boltzmann’s law, is the weighting factor to balance the energy terms, is 1 when vector pairs A and B are in contact and 0 otherwise,  and is the packing energy for two vector pairs A and B in relative orientation space.

*Objective 3*: The Chemistry at Harvard Macromolecular Mechanics **(**CHARMM) force field is shown in Eq.4. In this study, we use CHARMM22 (20), which is an all-atom empirical force field. It contains several terms that can be grouped into internal and external interactions. These interactions were derived from quantum chemical calculations.

 (S3)

Here, all *K* are force constants, ,,, are equilibrium constants, *b* is the bond length, *S* isUrey−Bradley 1,3-distance, *θ* is the bond angle, *χ* is the dihedral angle,  is the multiplicity of the function,  is the phase shift, and *φ* is the improper torsion angle. In the non-bonded terms,  is the Lennard-Jones well depth,  is the minimum interaction radius,  is the distance between atoms *i* and *j*, is the partial atomic charge, and *e*is the effective dielectric constant.

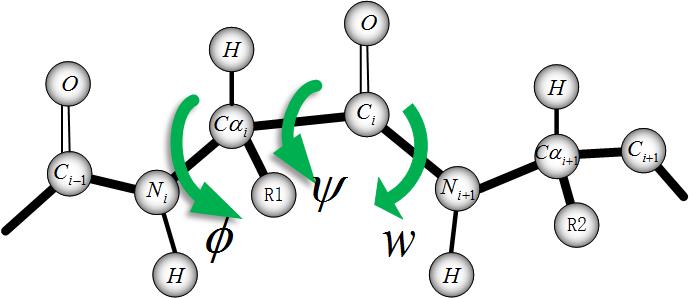
**Table S1**. Overview of the refinement results for each target (*N*=50)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test targets | Length | Initial modelc | Refined model (the best model) | Refined model ( model 1) |
| TM-score /  GDT-TS / RSMD | TM-score /  GDT-TS / RSMD | TM-score /  GDT-TS / RSMD |
| TR217a | 224 | 0.911/80.61/1.967 | 0.918/81.89/1.802 | 0.917/81.89/1.814 |
| TR228a | 84 | 0.692/71.73/3.918 | 0.700/73.81/3.083 | 0.700/73.81/3.083 |
| TR274a | 183 | 0.584/45.44/6.833 | 0.599/46.82/6.771 | 0.599/46.82/6.771 |
| TR280a | 96 | 0.780/78.19/4.065 | 0.789/78.46/3.888 | 0.789/78.46/3.895 |
| TR283a | 168 | 0.704/62.82/3.927 | 0.731/64.74/3.586 | 0.728/64.58/3.621 |
| TR520b | 321 | 0.930/79.13/1.974 | 0.941/79.70/1.834 | 0.941/79.70/1.834 |
| TR594b | 89 | 0.566/55.34/4.361 | 0.611/59.55/3.721 | 0.605/59.55/3.721 |
| TR759a | 61 | 0.565/62.70/4.239 | 0.596/66.80/3.732 | 0.593/66.39/3.732 |
| TR762a | 257 | 0.910/85.69/3.073 | 0.920/85.98/2.06 | 0.920/85.98/2.06 |
| TR765a | 76 | 0.757/77.67/2.512 | 0.816/82.67/2.416 | 0.816/82.67/2.42 |
| TR768a | 143 | 0.870/81.96/2.355 | 0.898/87.86/2.293 | 0.898/87.86/2.293 |
| TR769a | 97 | 0.814/77.32/1.879 | 0.803/76.21/2.078 | 0.803/75.98/2.078 |
| TR772a | 158 | 0.787/70.83/4.893 | 0.792/71.37/4.716 | 0.792/71.37/4.716 |
| TR774a | 155 | 0.671/58.11/4.998 | 0.675/58.45/4.768 | 0.672/58.45/4.782 |
| TR776a | 219 | 0.901/82.04/2.836 | 0.922/82.75/2.78 | 0.922/82.75/2.78 |
| TR780a | 95 | 0.780/75.00/2.62 | 0.788/76.86/2.299 | 0.783/76.33/2.348 |
| TR782a | 110 | 0.860/83.18/1.934 | 0.861/83.41/1.915 | 0.861/83.41/1.915 |
| TR783a | 243 | 0.890/79.15/3.073 | 0.897/80.53/2.89 | 0.897/80.21/2.912 |
| TR786a | 217 | 0.827/68.93/3.63 | 0.829/70.56/3.413 | 0.828/70.33/3.413 |
| TR792a | 80 | 0.730/74.33/2.329 | 0.769/78.67/2.15 | 0.766/78.67/2.202 |
| TR795a | 136 | 0.831/77.39/2.384 | 0.838/77.94/2.312 | 0.835/77.94/2.341 |
| TR803a | 134 | 0.602/51.87/5.988 | 0.607/53.73/5.872 | 0.607/53.73/5.872 |
| TR810a | 225 | 0.790/71.54/12.474 | 0.792/71.76/11.244 | 0.792/71.76/11.244 |
| TR816a | 68 | 0.661/70.53/2.57 | 0.678/71.92/2.256 | 0.673/71.15/2.278 |
| TR817a | 265 | 0.934/84.87/1.811 | 0.933/84.39/1.799 | 0.933/84.39/1.801 |
| TR821a | 255 | 0.861/70.34/2.456 | 0.891/74.29/2.231 | 0.891/74.29/2.231 |
| TR822a | 117 | 0.596/52.78/4.257 | 0.601/53.07/4.675 | 0.601/52.78/4.675 |
| TR829a | 67 | 0.600/66.54/6.216 | 0.750/79.23/2 | 0.750/79.23/2 |
| TR833a | 108 | 0.780/75.71/4.756 | 0.800/77.83/2.95 | 0.791/77.12/2.95 |
| TR848a | 138 | 0.798/75.18/3.785 | 0.801/76.29/3.764 | 0.801/75.92/3.764 |
| TR854a | 70 | 0.727/77.69/2.315 | 0.738/79.23/2.165 | 0.738/78.84/2.216 |
| TR856a | 159 | 0.869/79.87/2.681 | 0.871/79.87/2.251 | 0.871/79.42/2.251 |
| TR857a | 96 | 0.599/55.47/4.061 | 0.611/55.99/3.837 | 0.605/55.73/3.904 |
| TR862b | 101 | 0.520/54.84/5.964 | 0.537/56.18/5.484 | 0.537/55.91/5.484 |
| TR866b | 115 | 0.761/74.55/4.708 | 0.761/74.55/4.536 | 0.761/74.55/4.584 |
| TR868b | 116 | 0.803/76.72/3.01 | 0.815/79.09/2.796 | 0.815/78.66/2.81 |
| TR869b | 104 | 0.359/38.94/12.061 | 0.365/39.42/11.65 | 0.3650/39.18/11.724 |
| TR870b | 123 | 0.434/37.60/9.476 | 0.444/39.05/9.21 | 0.443/38.43/9.21 |
| TR872b | 88 | 0.737/73.84/5.853 | 0.761/75.00/4.192 | 0.761/75.00/4.21 |
| TR877b | 142 | 0.784/70.07/3.008 | 0.785/70.07/2.981 | 0.784/70.07/2.981 |
| TR879b | 220 | 0.847/79.20/5.502 | 0.855/79.77/5.375 | 0.855/79.77/5.375 |
| TR882b | 79 | 0.835/86.39/2.078 | 0.837/87.12/2.15 | 0.837/86.39/2.15 |
| TR884b | 71 | 0.593/64.29/3.773 | 0.608/65.00/3.481 | 0.608/65.00/3.561 |
| TR885b | 114 | 0.899/88.29/2.148 | 0.890/87.79/2.517 | 0.889/87.79/2.517 |
| TR891b | 112 | 0.920/91.20/1.599 | 0.922/91.52/1.472 | 0.922/91.52/1.472 |
| TR893b | 169 | 0.911/87.28/2.274 | 0.918/87.57/2.015 | 0.916/87.47/2.015 |
| TR894b | 54 | 0.640/74.54/2.324 | 0.671/78.30/1.726 | 0.664/78.30/1.726 |
| TR895b | 120 | 0.732/70.21/4.328 | 0.750/73.93/4.089 | 0.750/73.93/4.115 |
| TR909b | 340 | 0.773/60.21/6.473 | 0.775/61.25/6.431 | 0.773/60.06/6.445 |
| TR920b | 219 | 0.892/79.68/2.711 | 0.900/80.70/2.684 | 0.892/79.91/2.695 |
| TR921b | 138 | 0.759/69.02/3.508 | 0.771/70.83/3.412 | 0.762/70.04/3.412 |
| TR922b | 74 | 0.766/79.73/3.145 | 0.794/79.73/2.769 | 0.794/79.73/2.769 |
| TR928b | 341 | 0.788/63.27/5.965 | 0.785/62.43/6 | 0.785/62.43/6.014 |
| TR944b | 253 | 0.841/74.11/3.896 | 0.845/74.6/3.854 | 0.843/74.21/3.854 |
| TR945b | 396 | 0.764/58.67/8.408 | 0.766/59.87/8.284 | 0.765/59.47/8.374 |
| TR948b | 161 | 0.756/69.88/6.74 | 0.767/70.50/5.495 | 0.766/70.50/5.495 |
| Average | 152.93 | 0.755/71.12/4.145 | 0.769/72.62/3.788 | 0.768/72.43/3.802 |

a Targets from CASP11;

b Targets from CASP12;

c The initial structures released for each target in CASP refinement category. These are the best models, selected by the organization committee, from the submitted structures.



**Figure S1**. An illustration of the torsional angles.

****

**Figure S2**. An illustration of the definition of standard Denavit and Hartenberg link parameters 38 used in our protein refinement project. A covalent bond is considered a rigid body that defines the spatial relationship between two neighboring atom axes. A covalent bond can be specified by two parameters, its length and its twist . Atoms are also described by two parameters. The atom offset is the distance from one atom coordinate frame to the next along the axis of the atom. The atom angle is the rotation of one atom with respect to the next about the atom axis. In our paper, each is a constant, defined in the initial structure of the protein. However, each is a variable, which will change during the refinement procedure, effecting final refined structure of the protein.

**AIR algorithm (*****,*** *Maxit, N***)**

**Input**: – the initial models, *Maxit* – the max-number of the iteration, *N*– the number of initial particles

**Output**: *Cf* – the final refined structures set

 /\* the number of iterations\*/



**Initialization**(); /\*generate part of initial models, extract the initial coordinates, set the initial velocity and select initial non-dominated particles into\*/

**While** (*k < Maxit*)

**While** (*i < N*)

; /\* Randomly select a conformation from Pareto set \*/

;/\*select the best conformation particle *i* has had\*/

/\*update the conformation by the following two updating equations\*/



 /\* the newly updated conformation\*/

;/\*conformation evaluation by the three energy functions\*/

update (); /\*if  is a new non-dominated solution, add it into Pareto set\*/

*i :=i+1*;

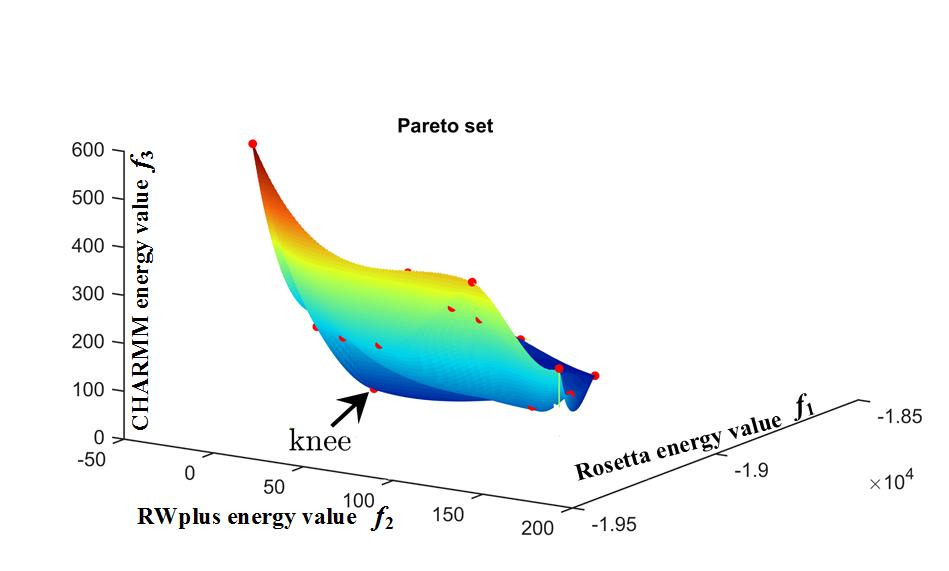
**End while**

*k := k+1*;

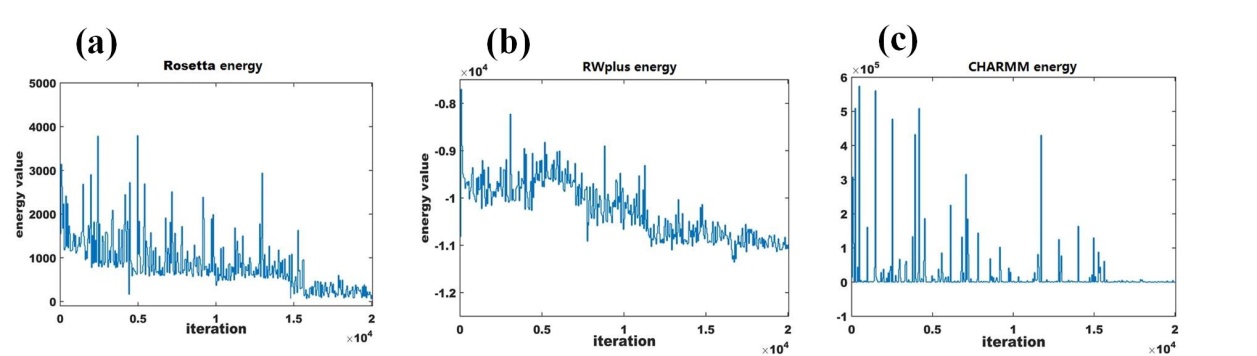
**End while**

*Cf* := rank\_and\_select()./\*rank the conformation in the Pareto set and select the top 5 as the final structure \*/

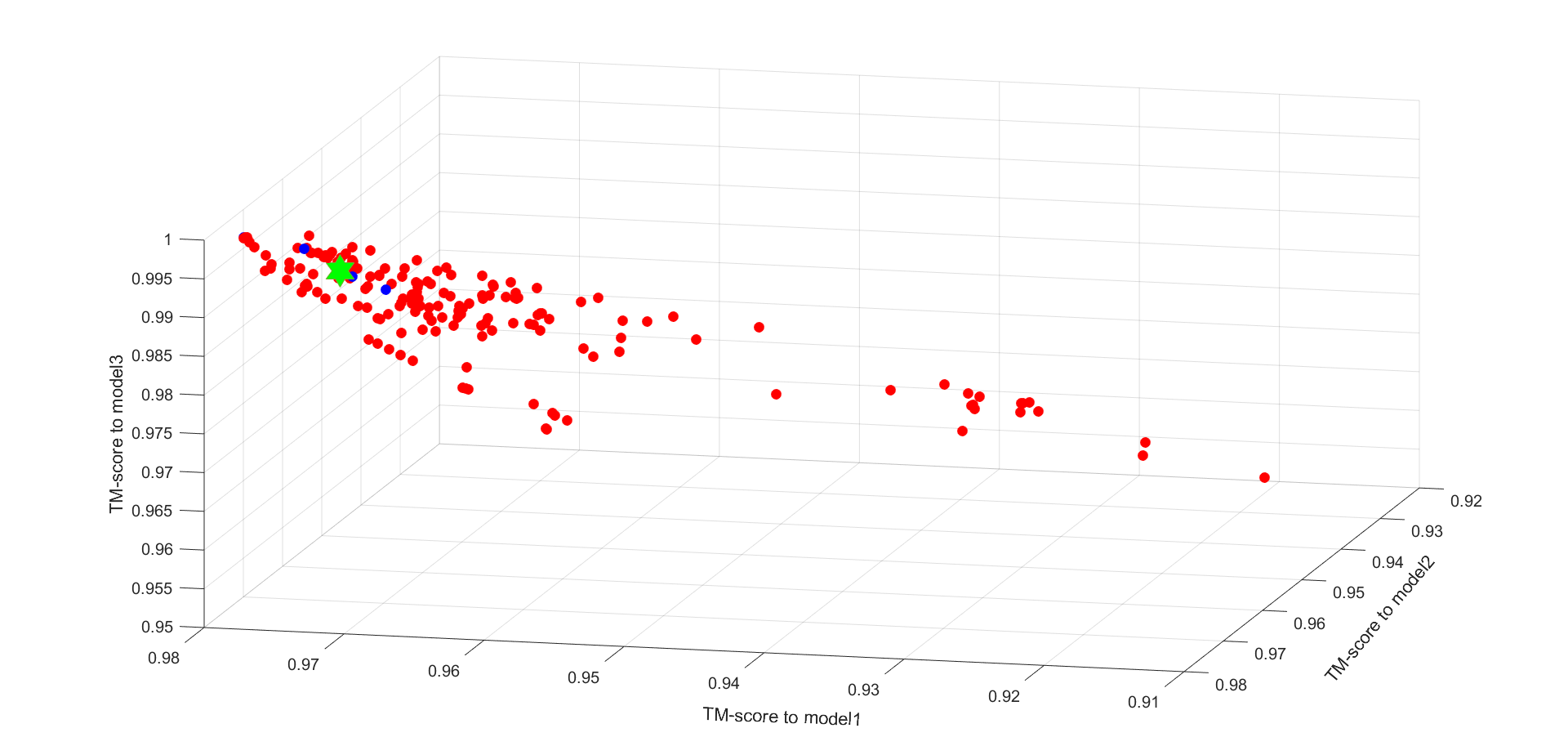
**Figure S3**. The pseudo code for AIR program.



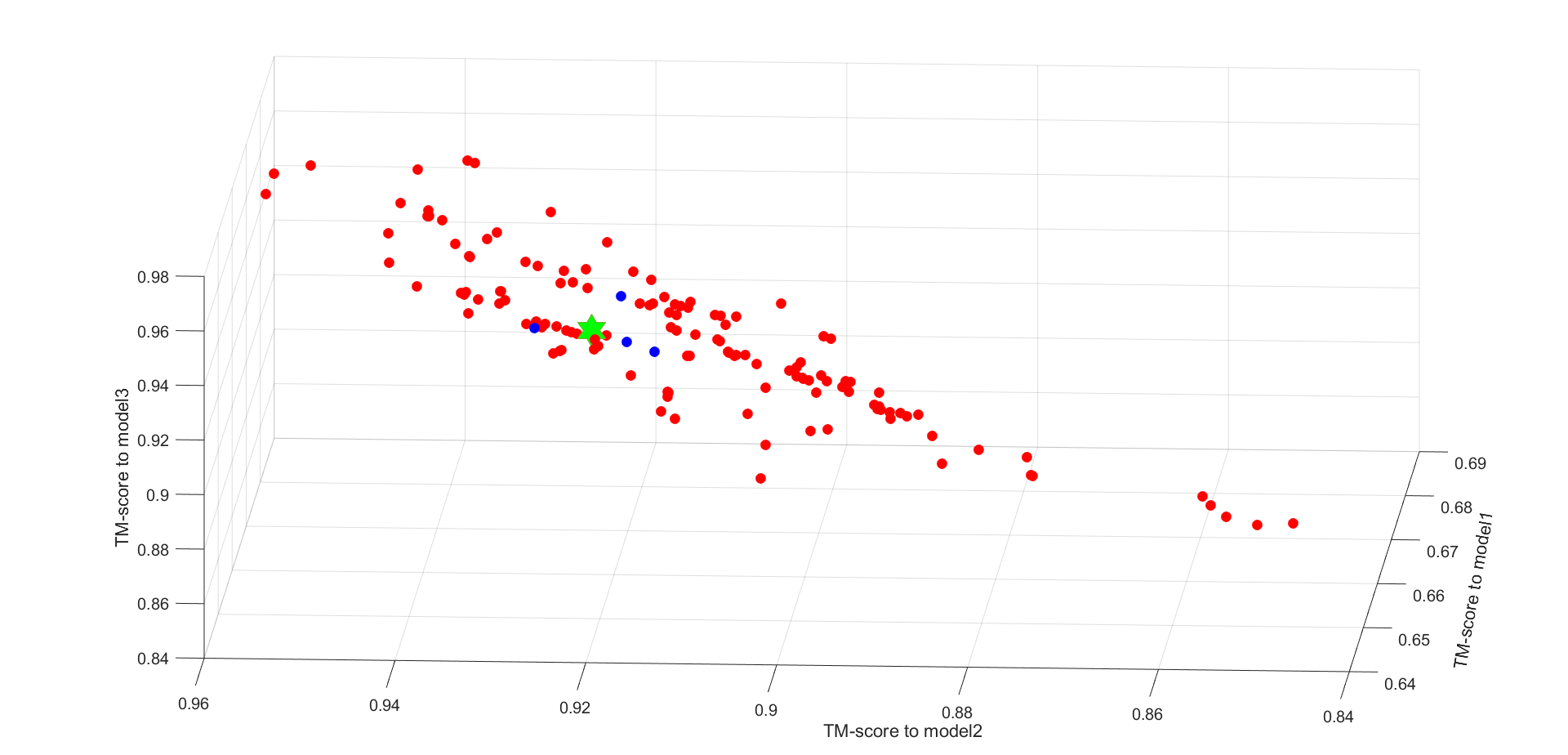
**Figure S4**. An intuitive figure showing Pareto-optimal front with a knee for target TR869.



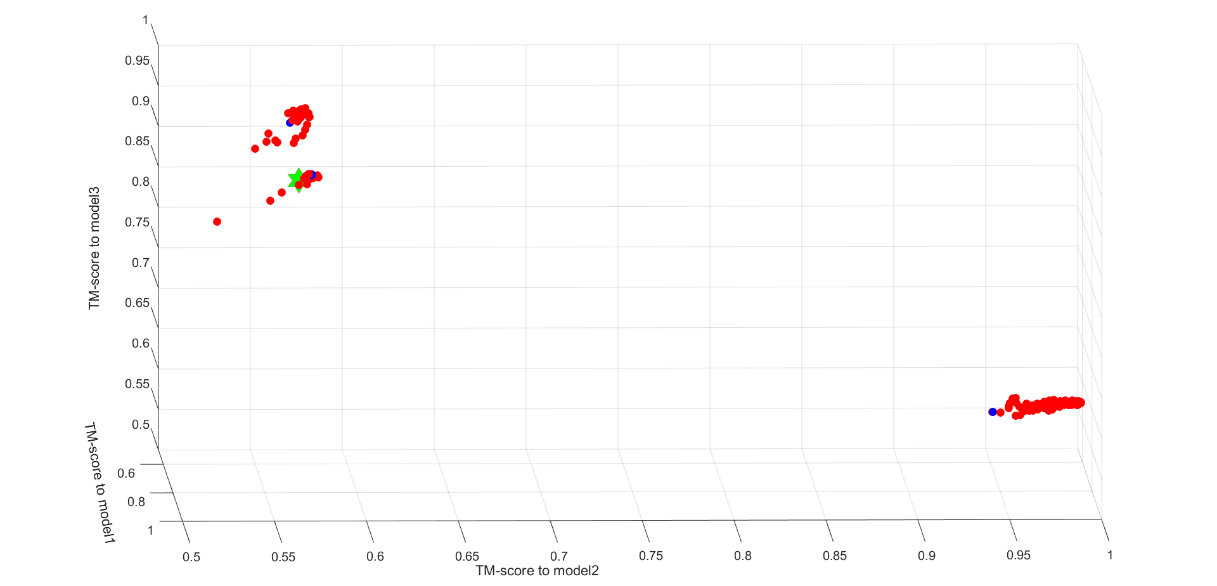
**Figure S5**. Energy function track with the PSO searching iterations: the Rosetta enenrgy value, the RWplus energy value and the CHARMM energy value in each iteration are shown in (a), (b) and (c), respectively.

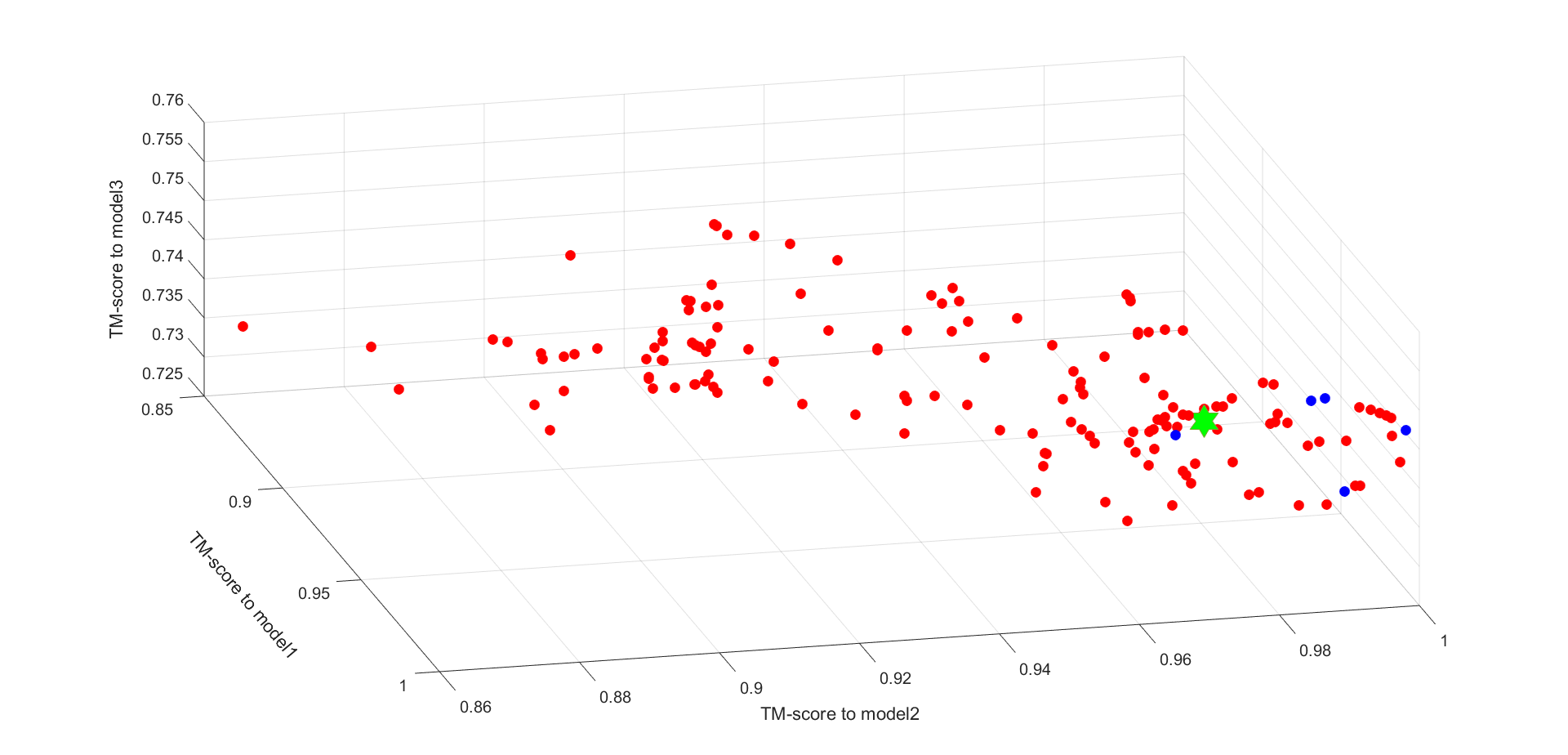


The TM-score space of TR217

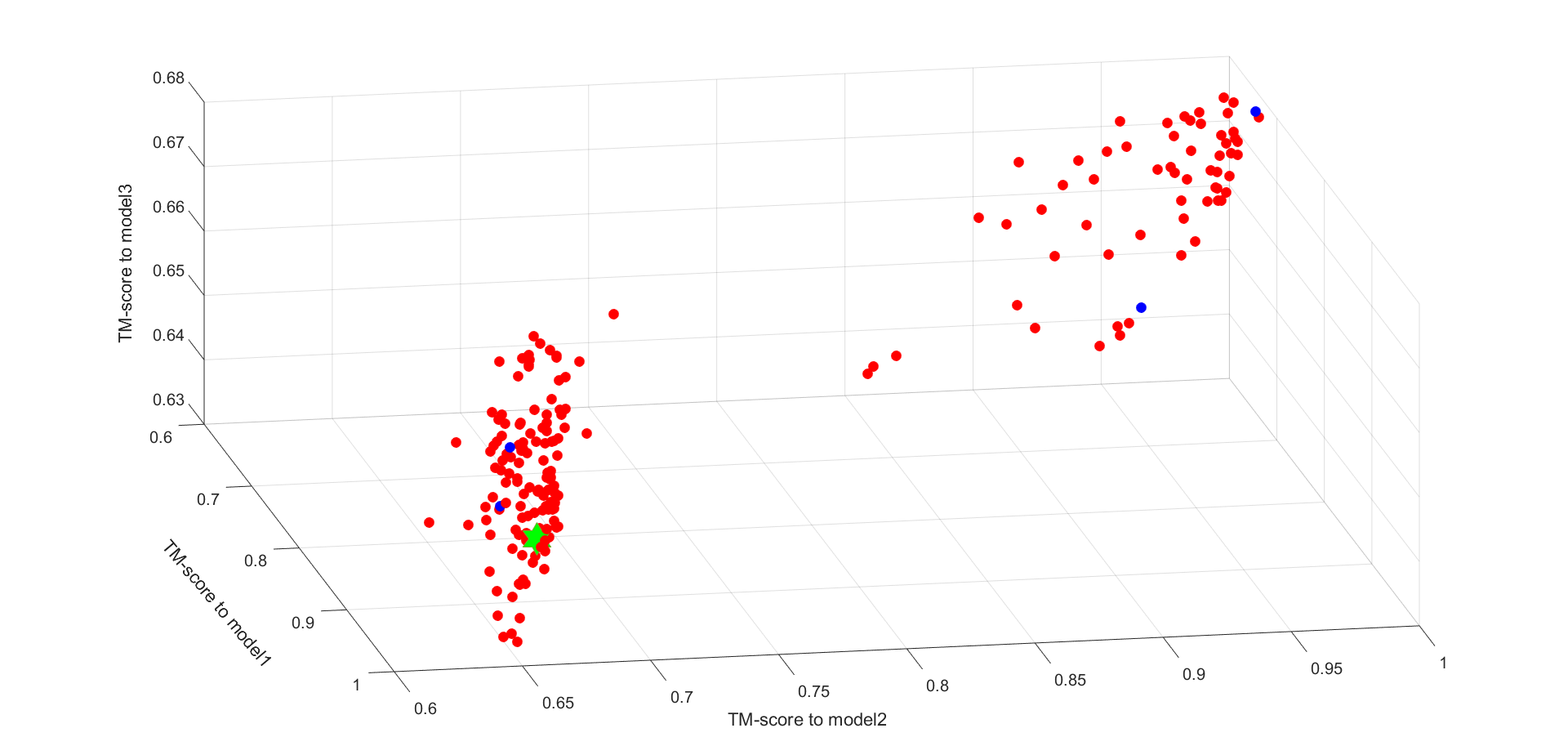


The TM-score space of TR228

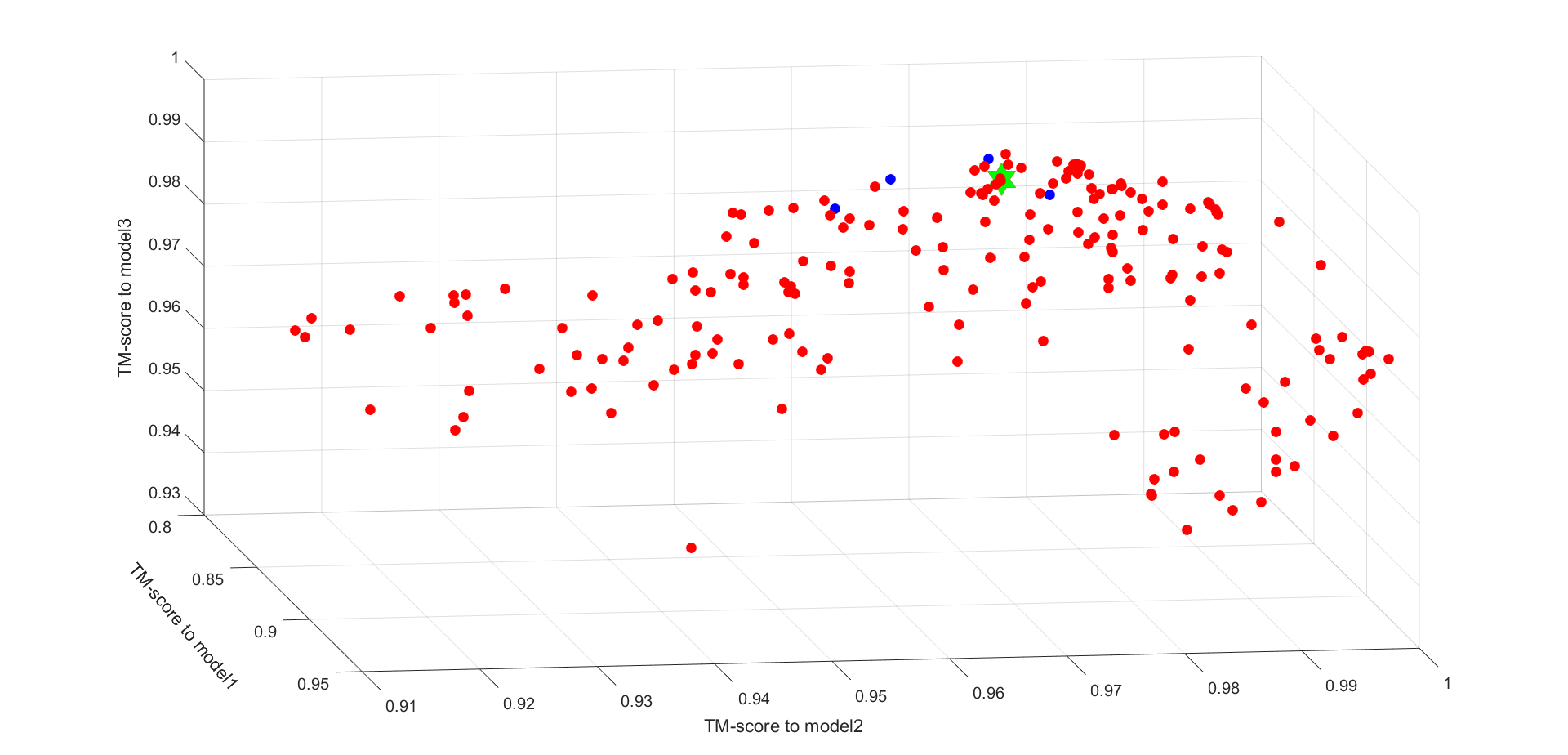
  
The TM-score space of TR274



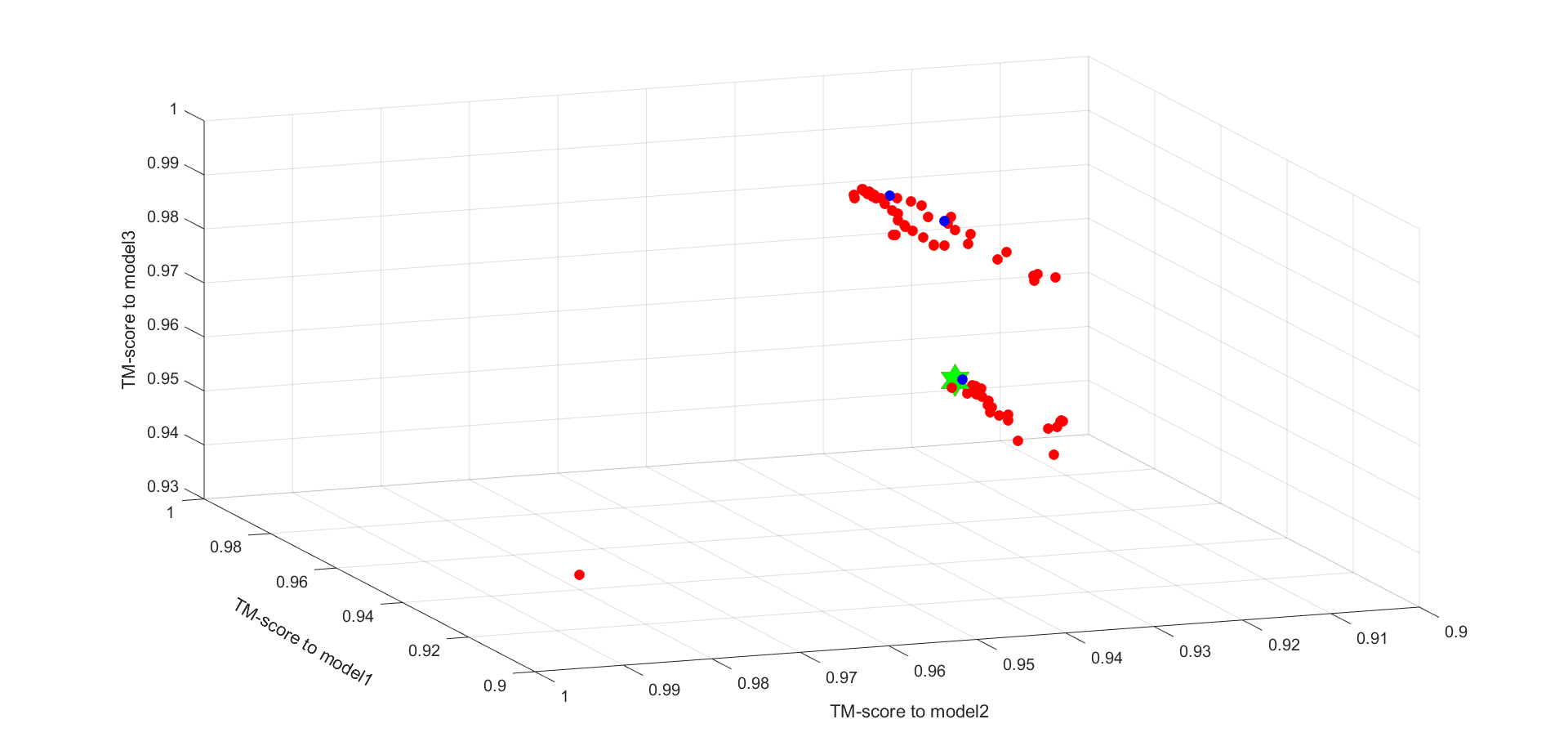
The TM-score space of TR280



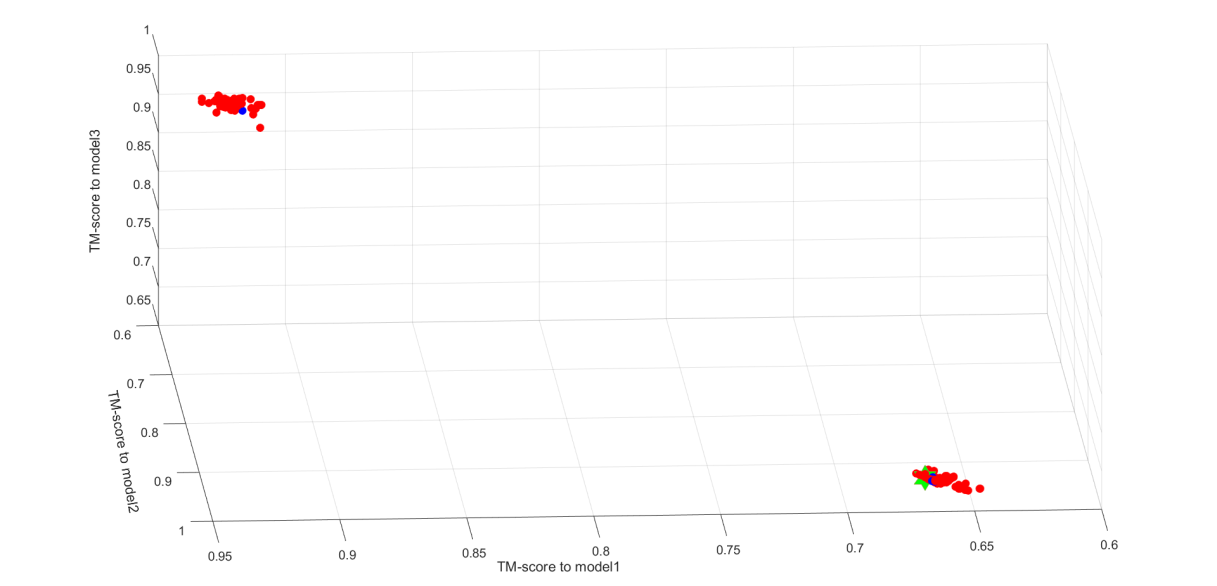
The TM-score space of TR283



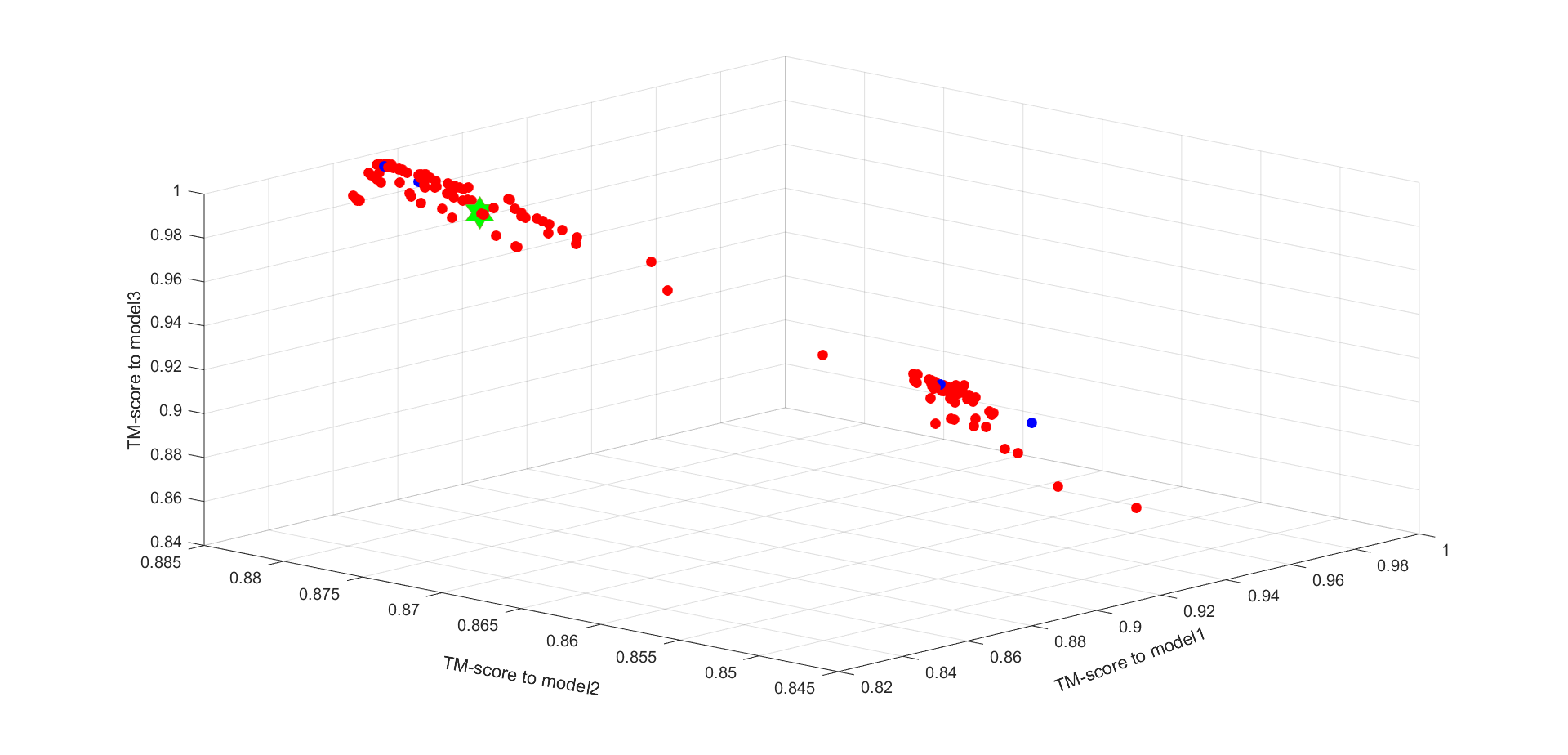
The TM-score space of TR759



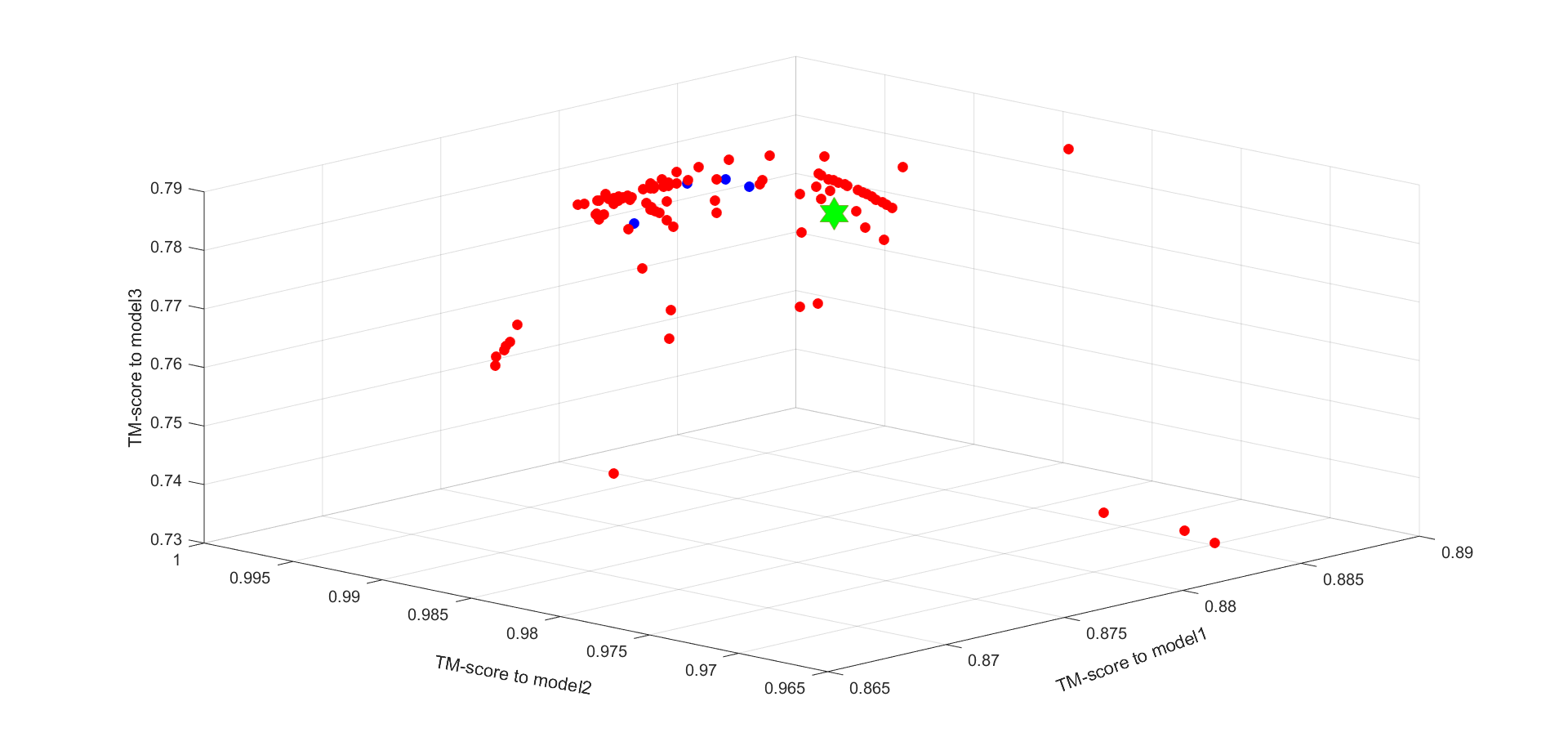
The TM-score space of TR762



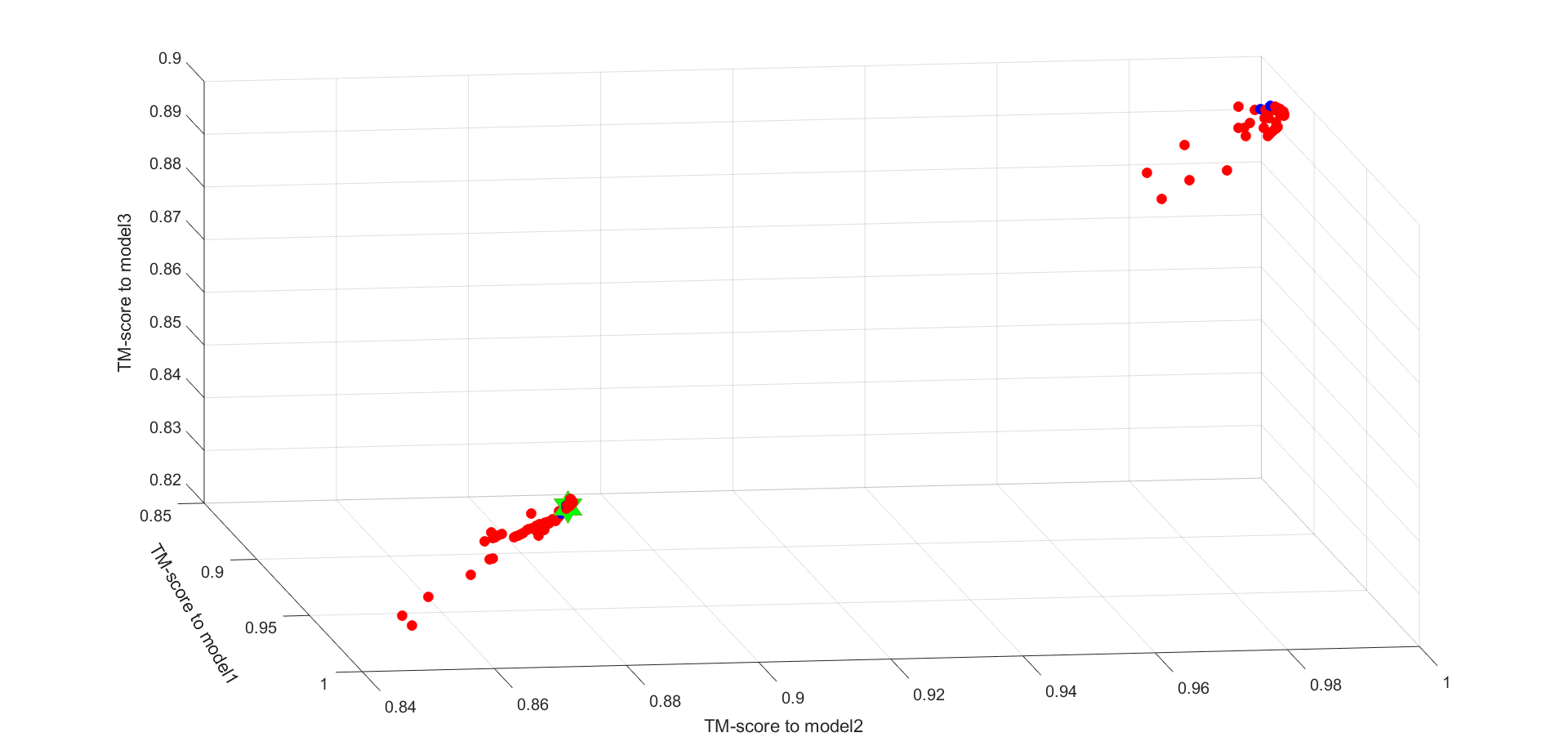
The TM-score space of TR765



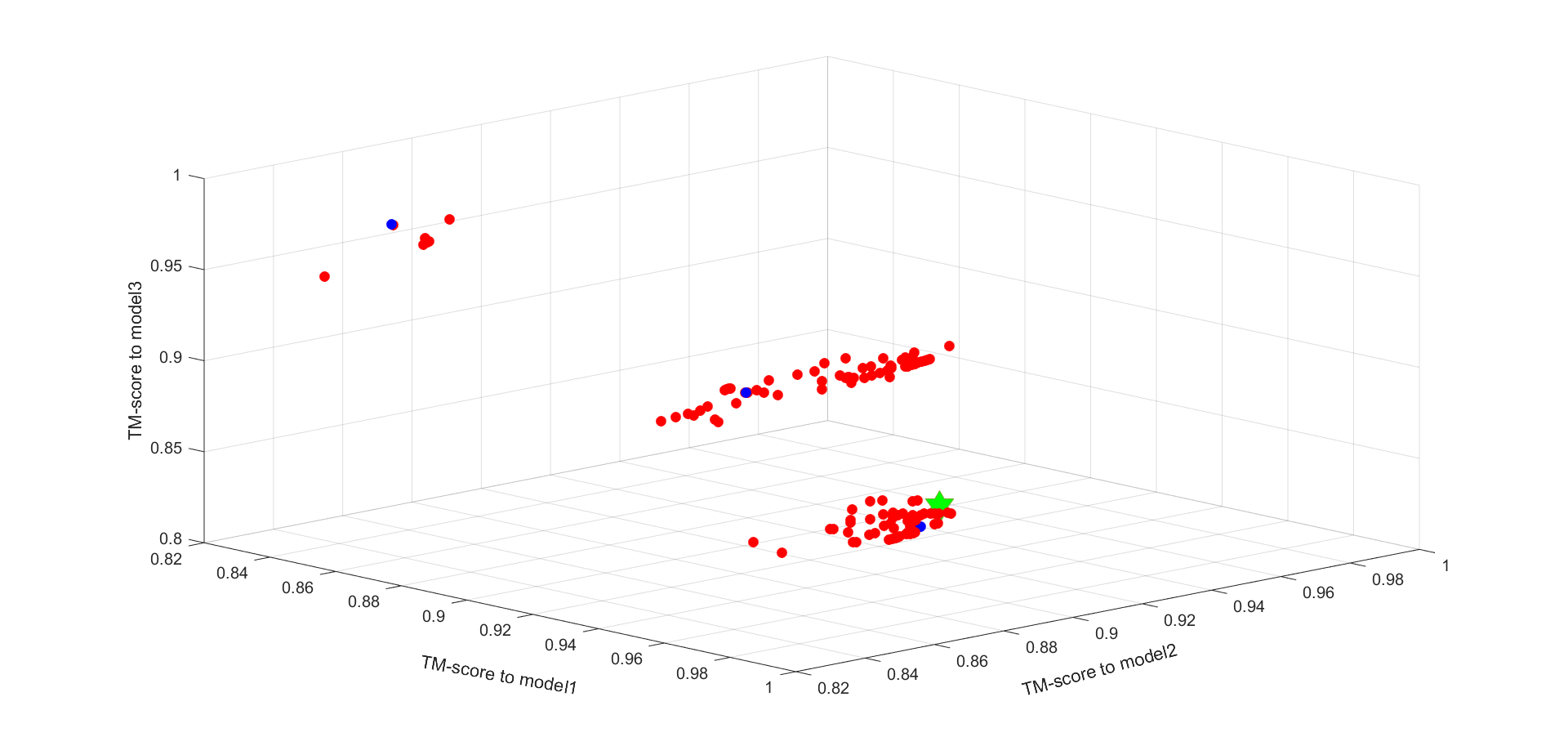
The TM-score space of TR768

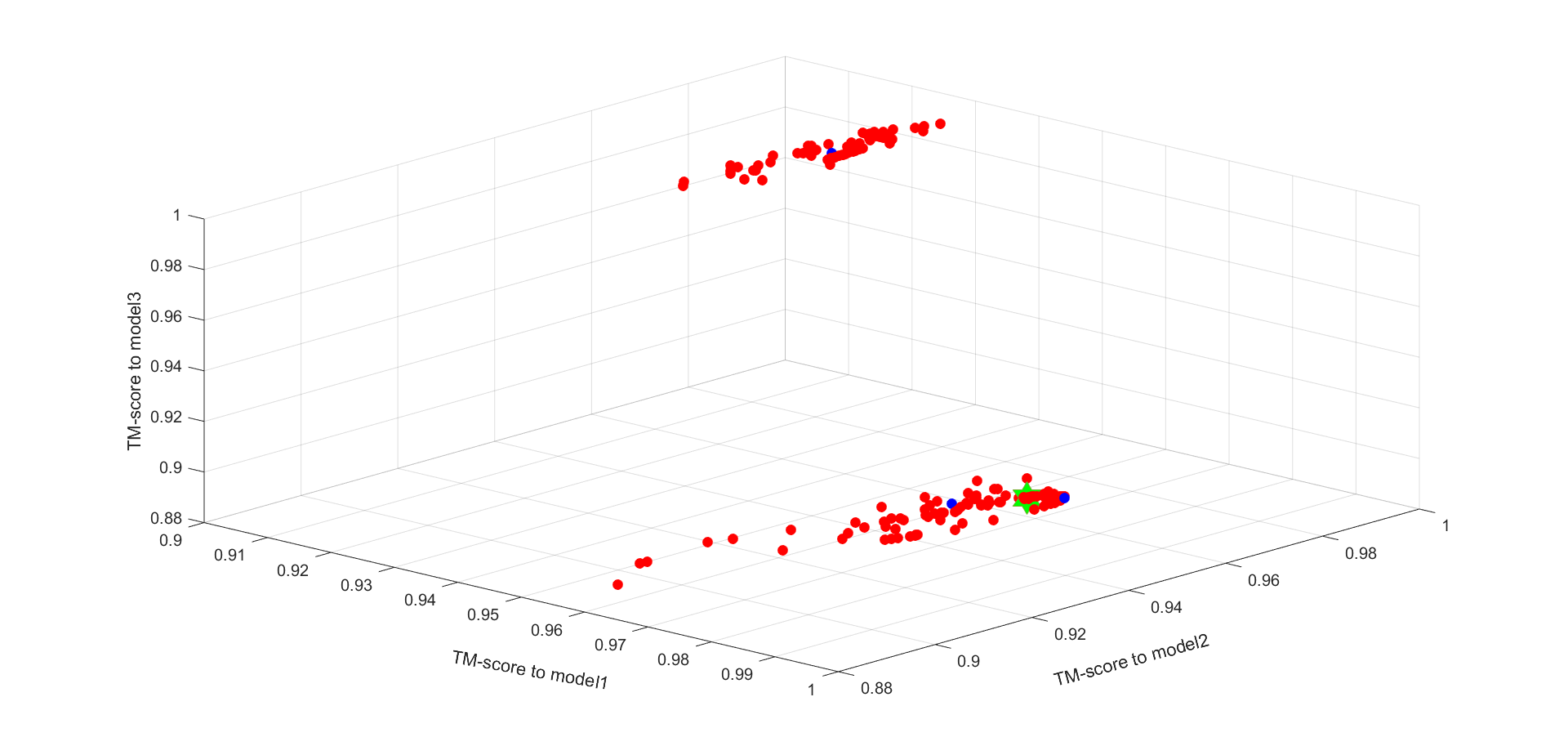


The TM-score space of TR769

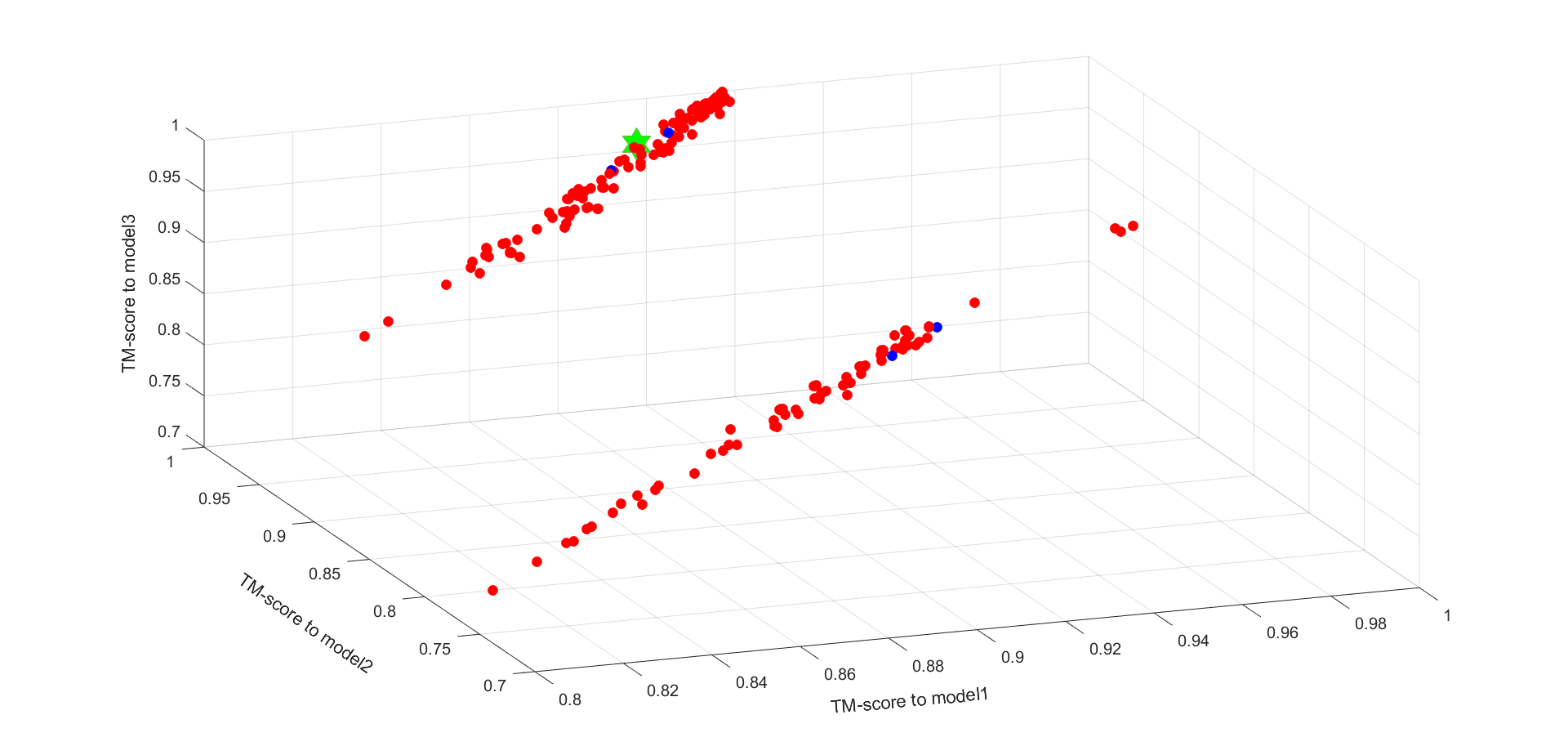


The TM-score space of TR772

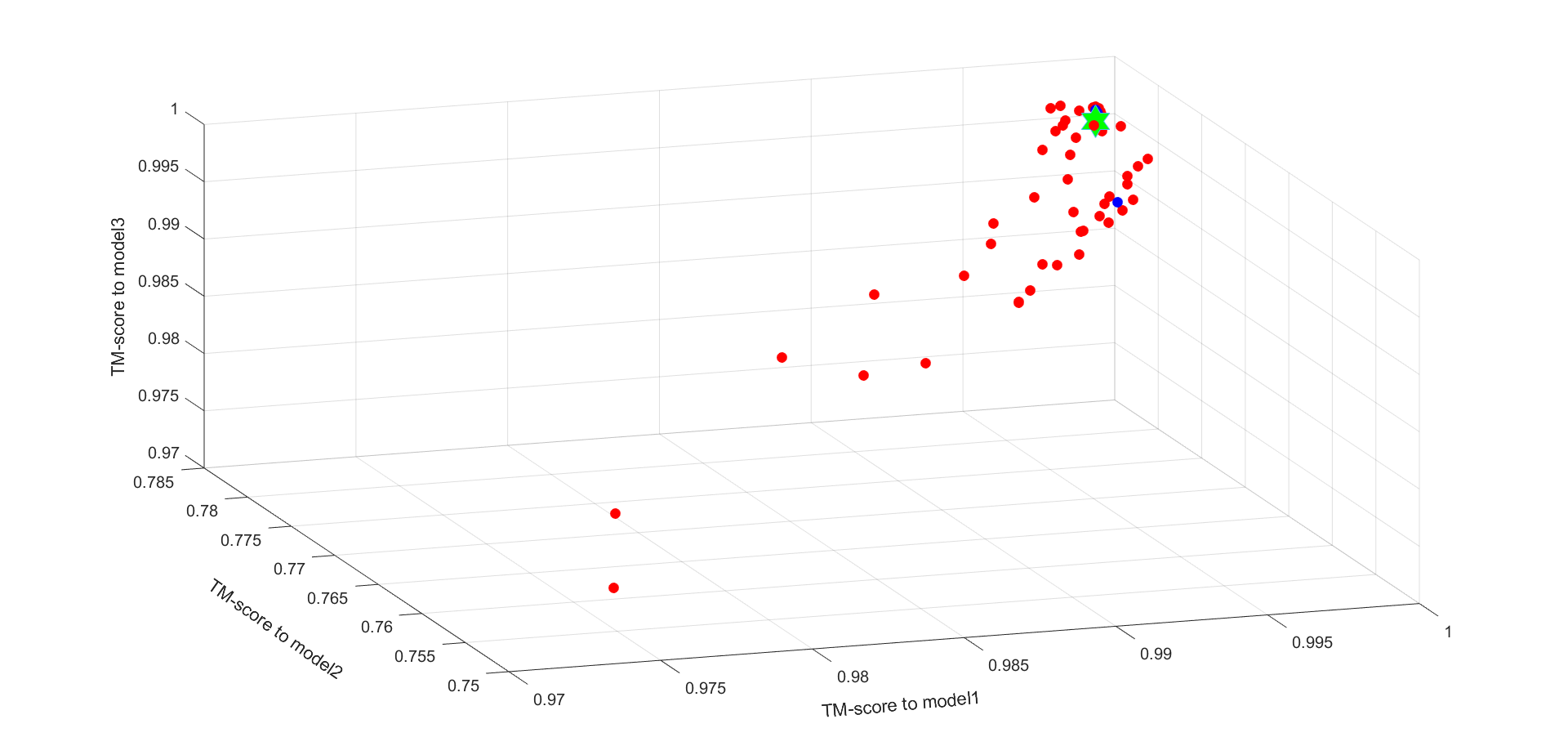
  
The TM-score space of TR774



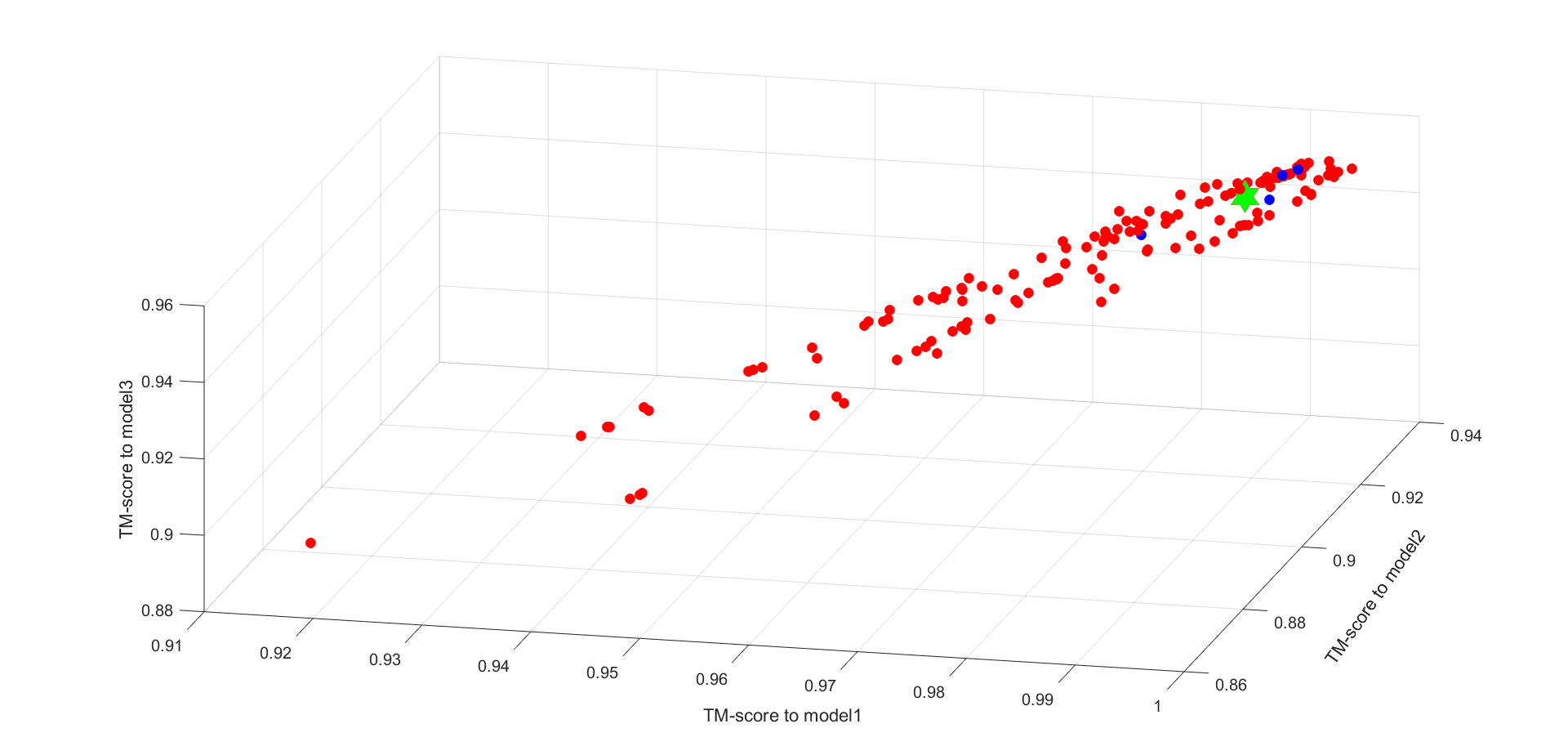
The TM-score space of TR776



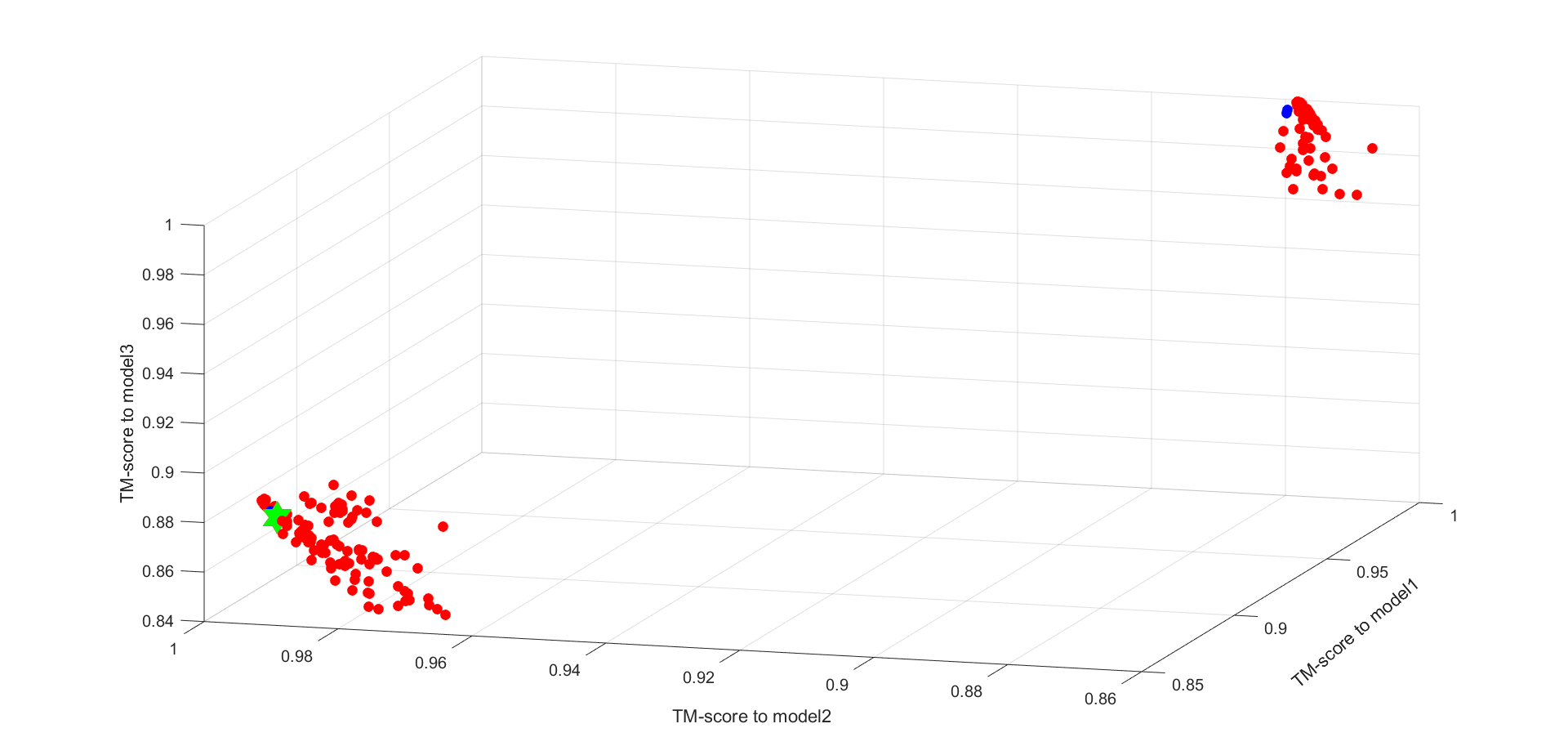
The TM-score space of TR780



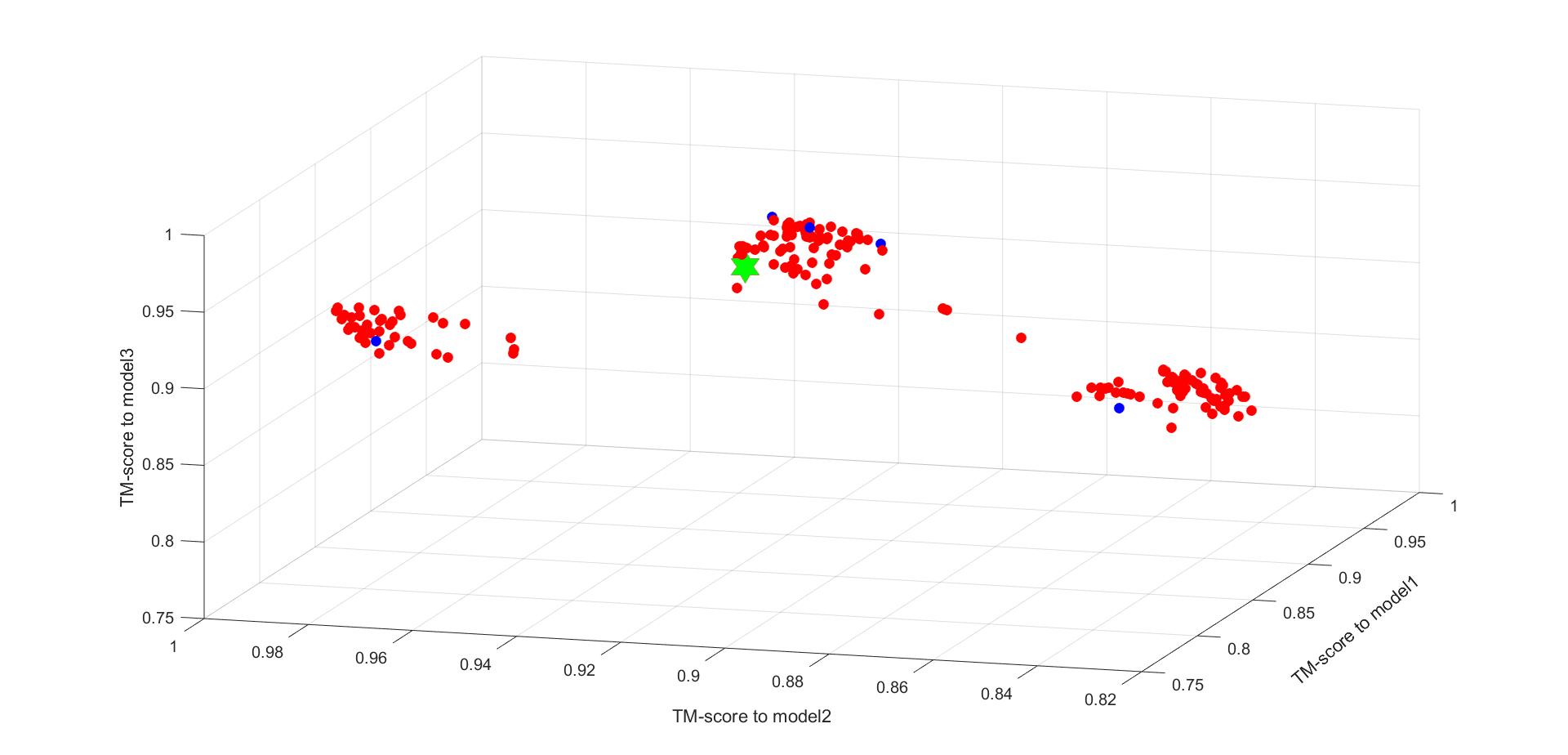
The TM-score space of TR782



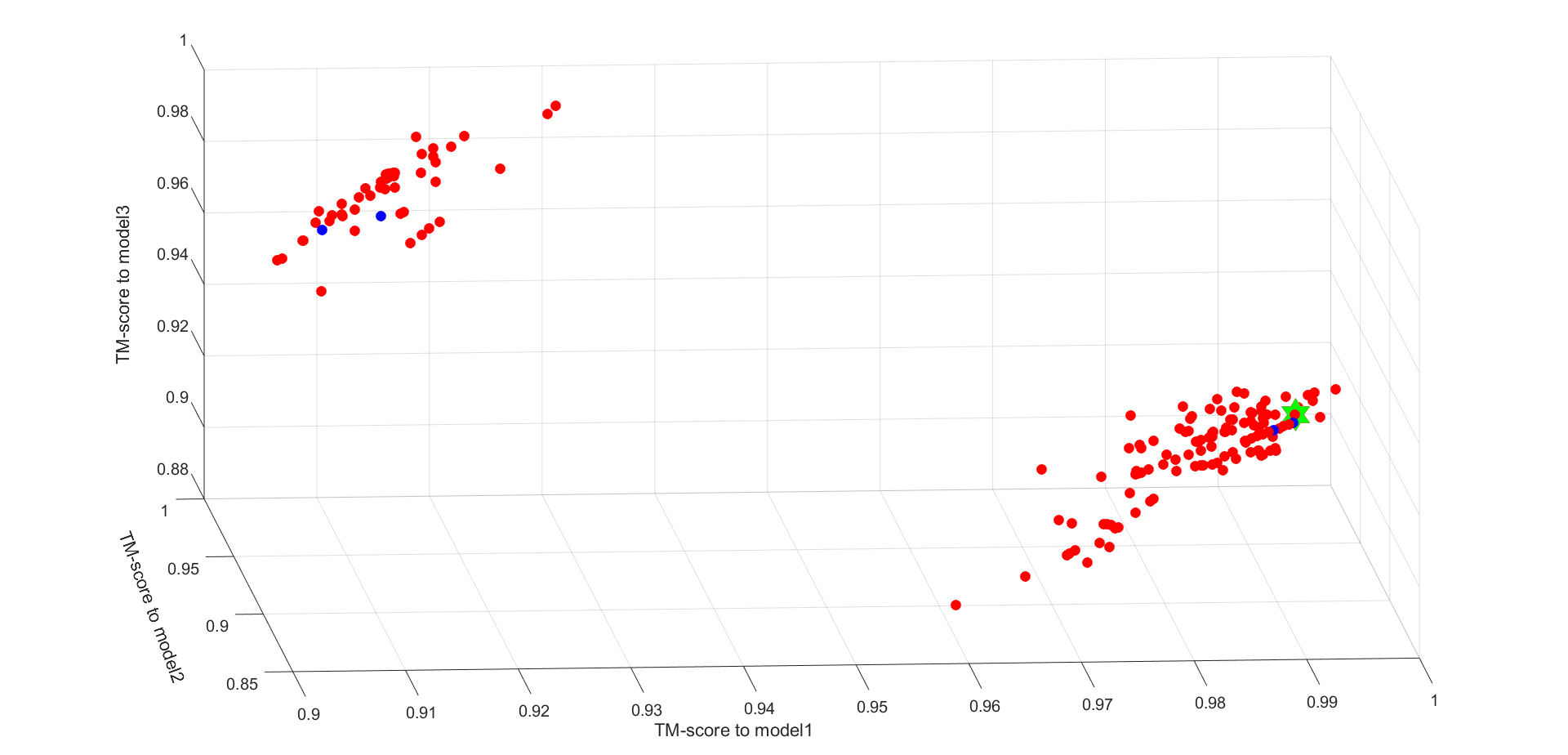
The TM-score space of TR783



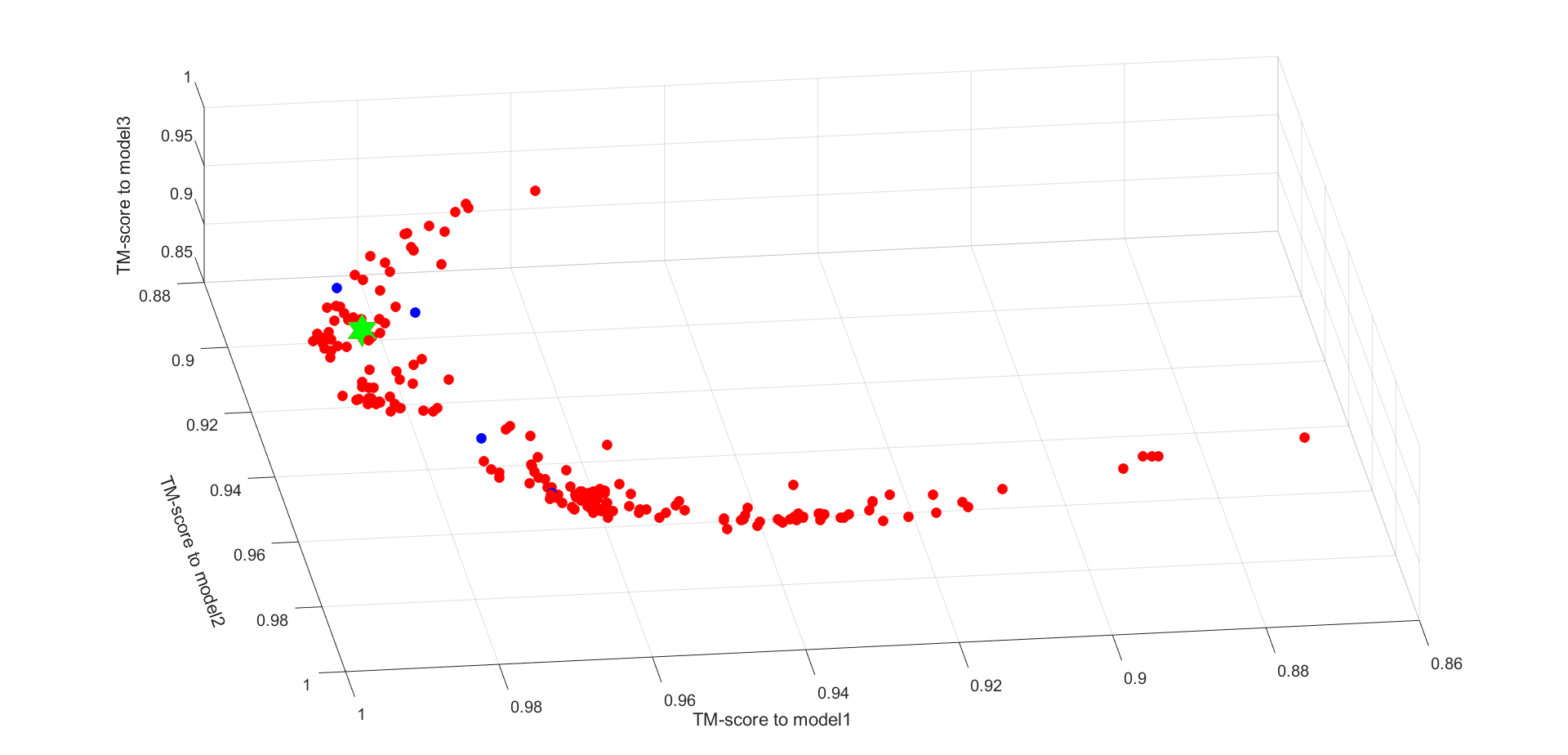
The TM-score space of TR786



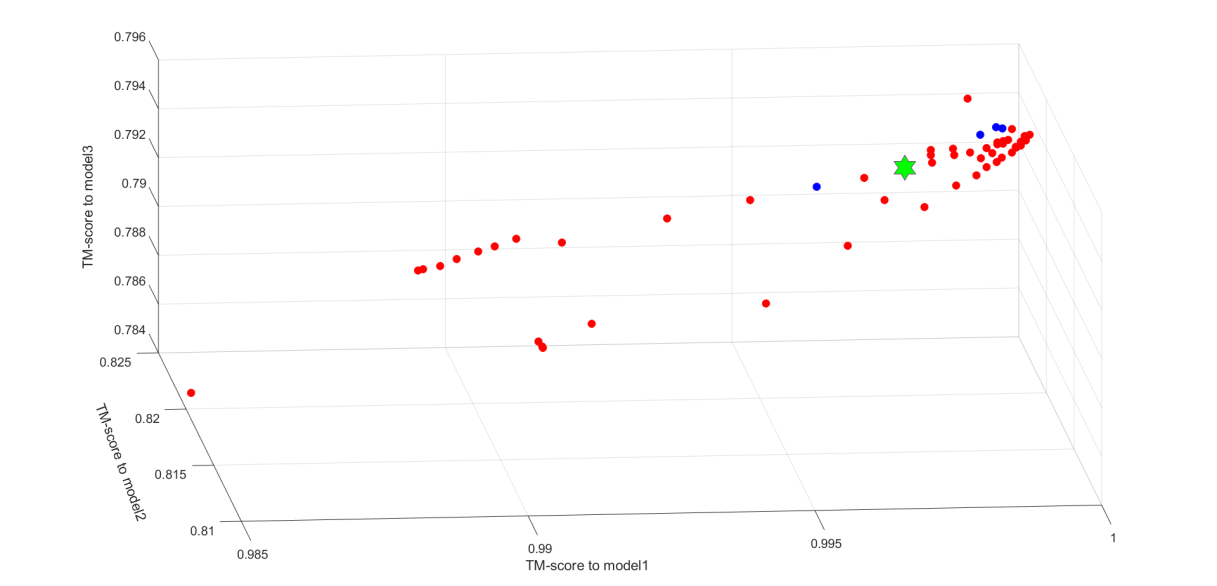
The TM-score space of TR792



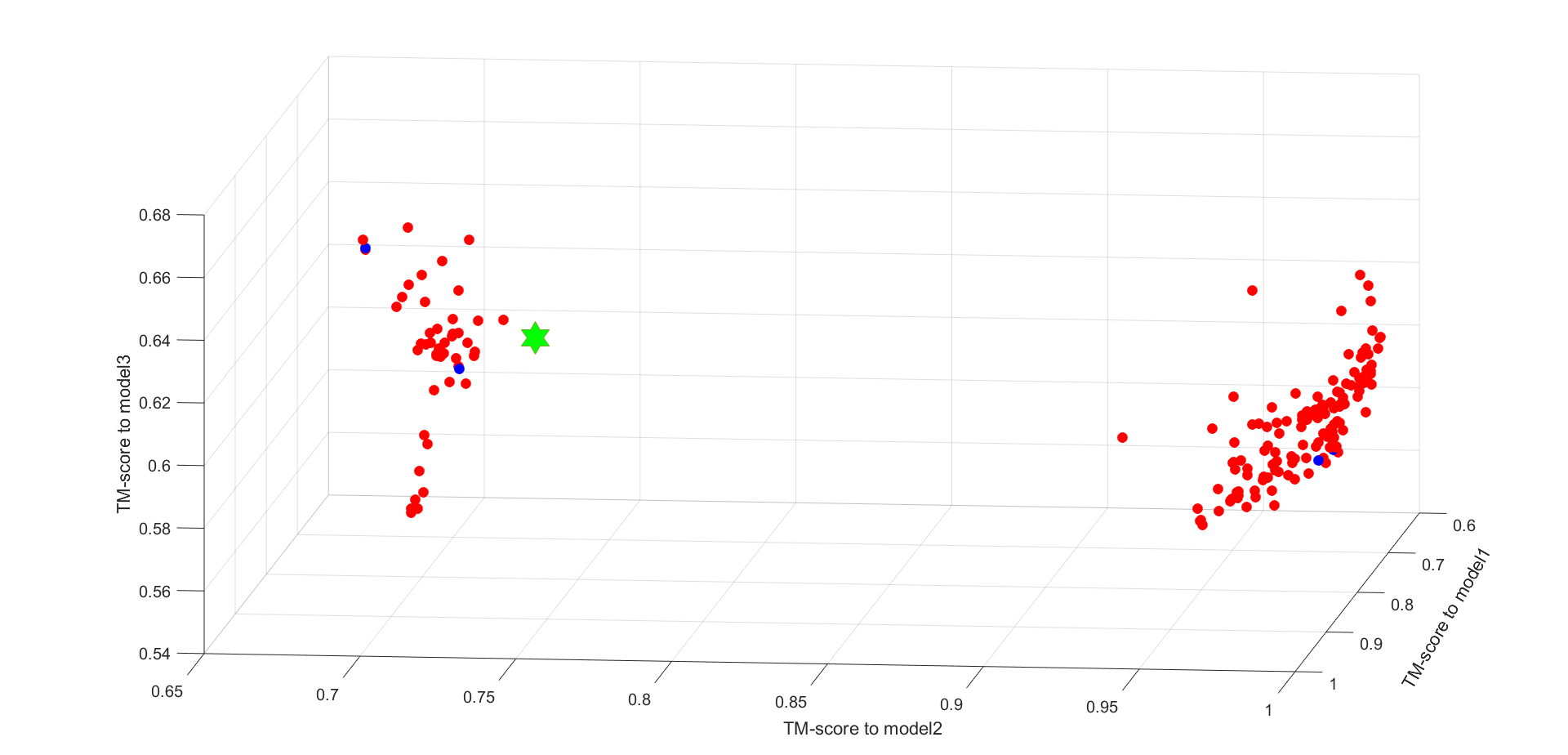
The TM-score space of TR795



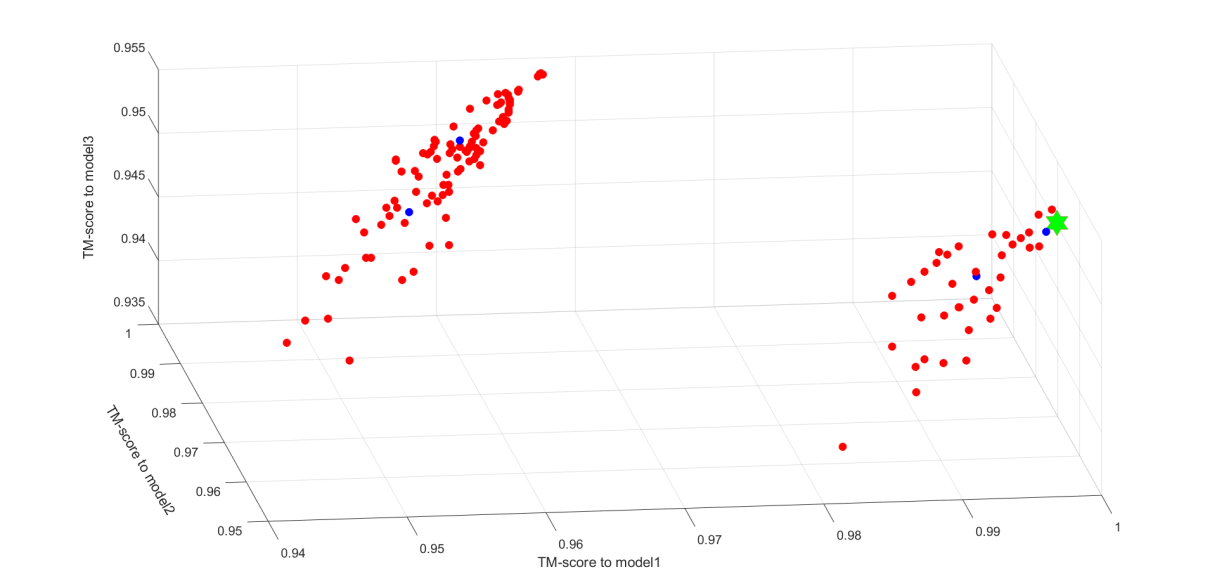
The TM-score space of TR803



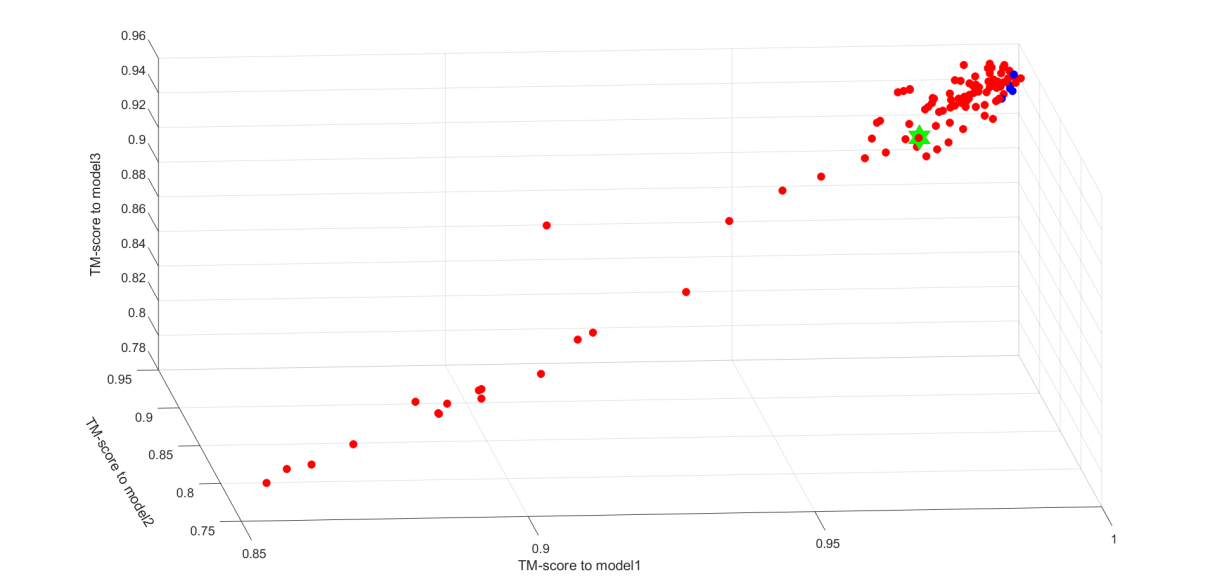
The TM-score space of TR810



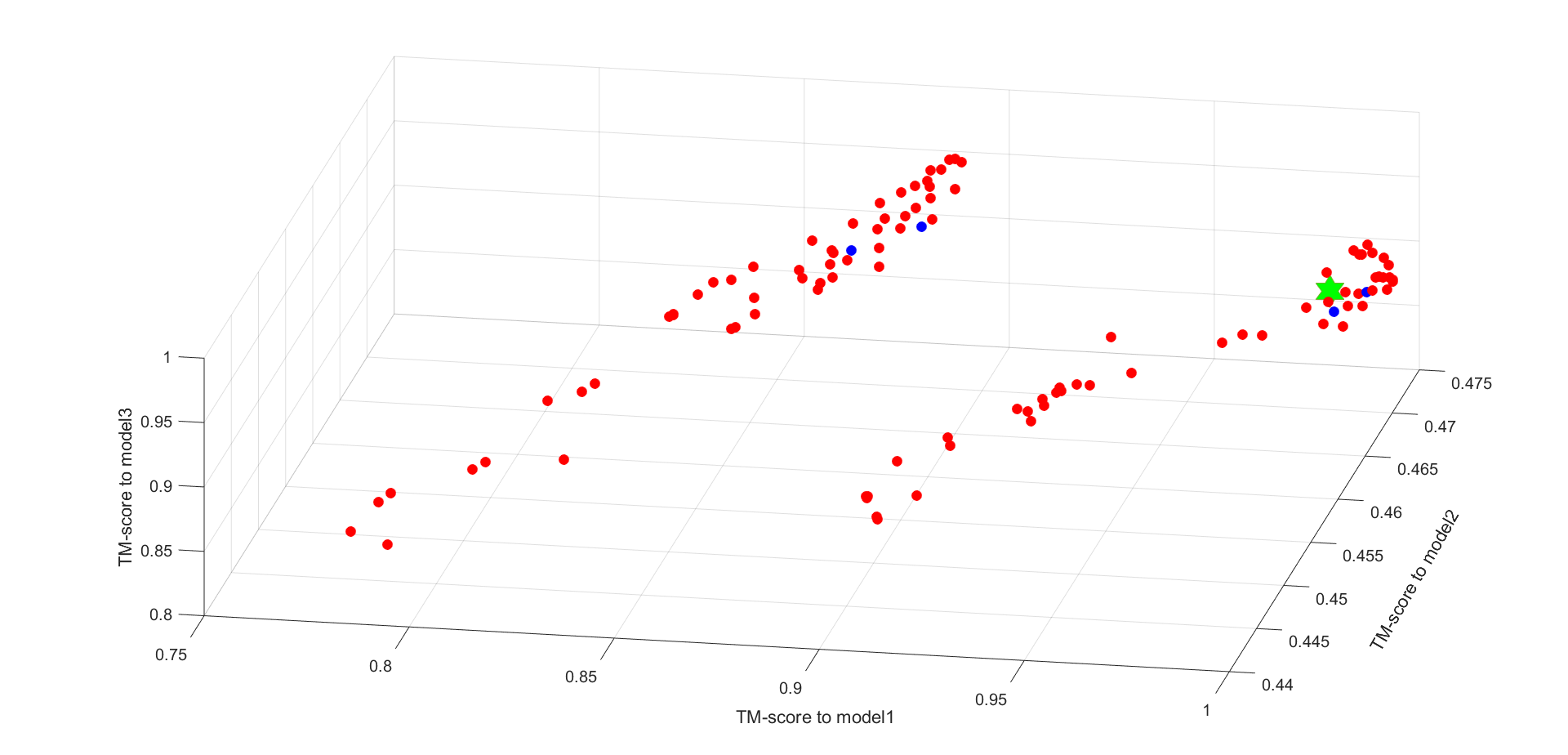
The TM-score space of TR816



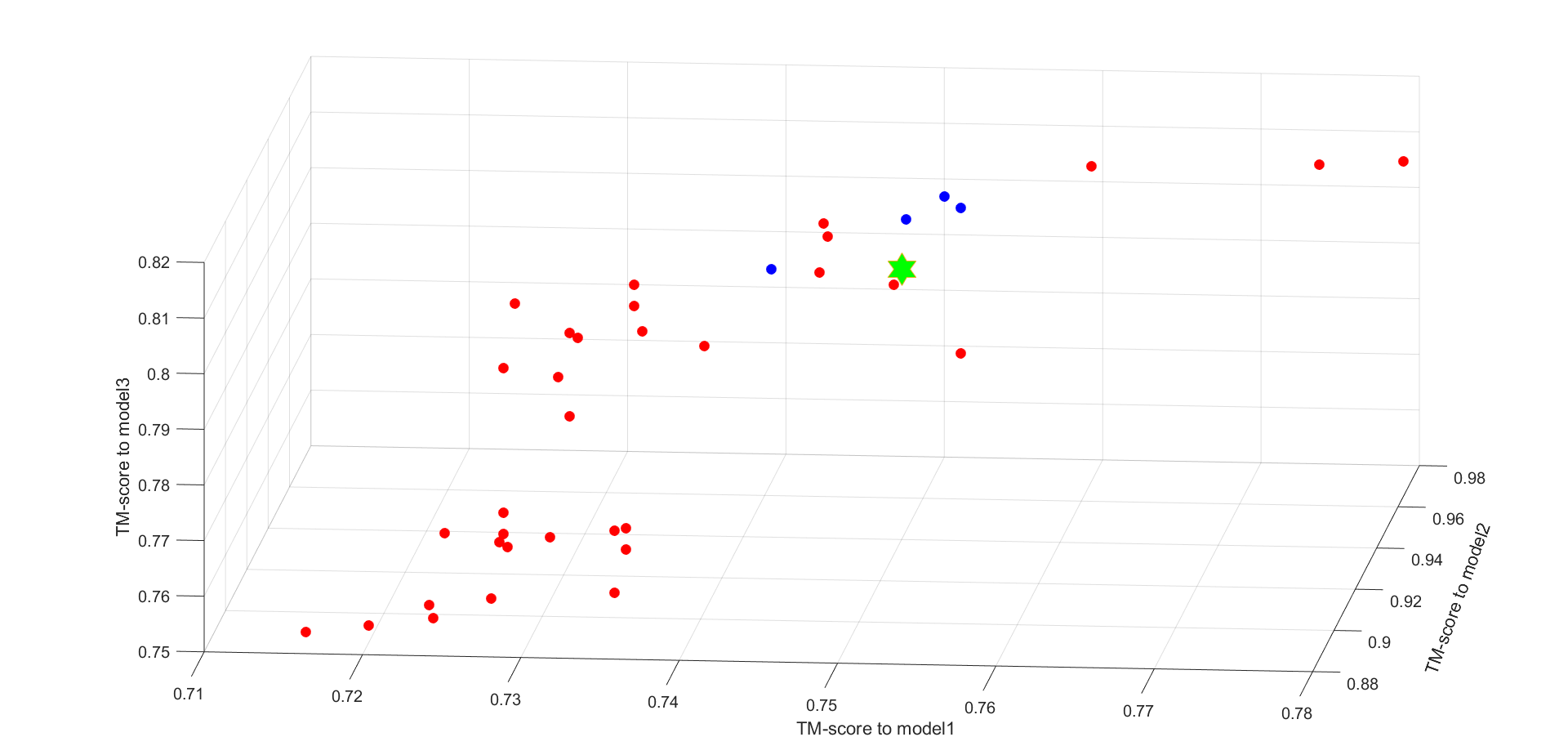
The TM-score space of TR817



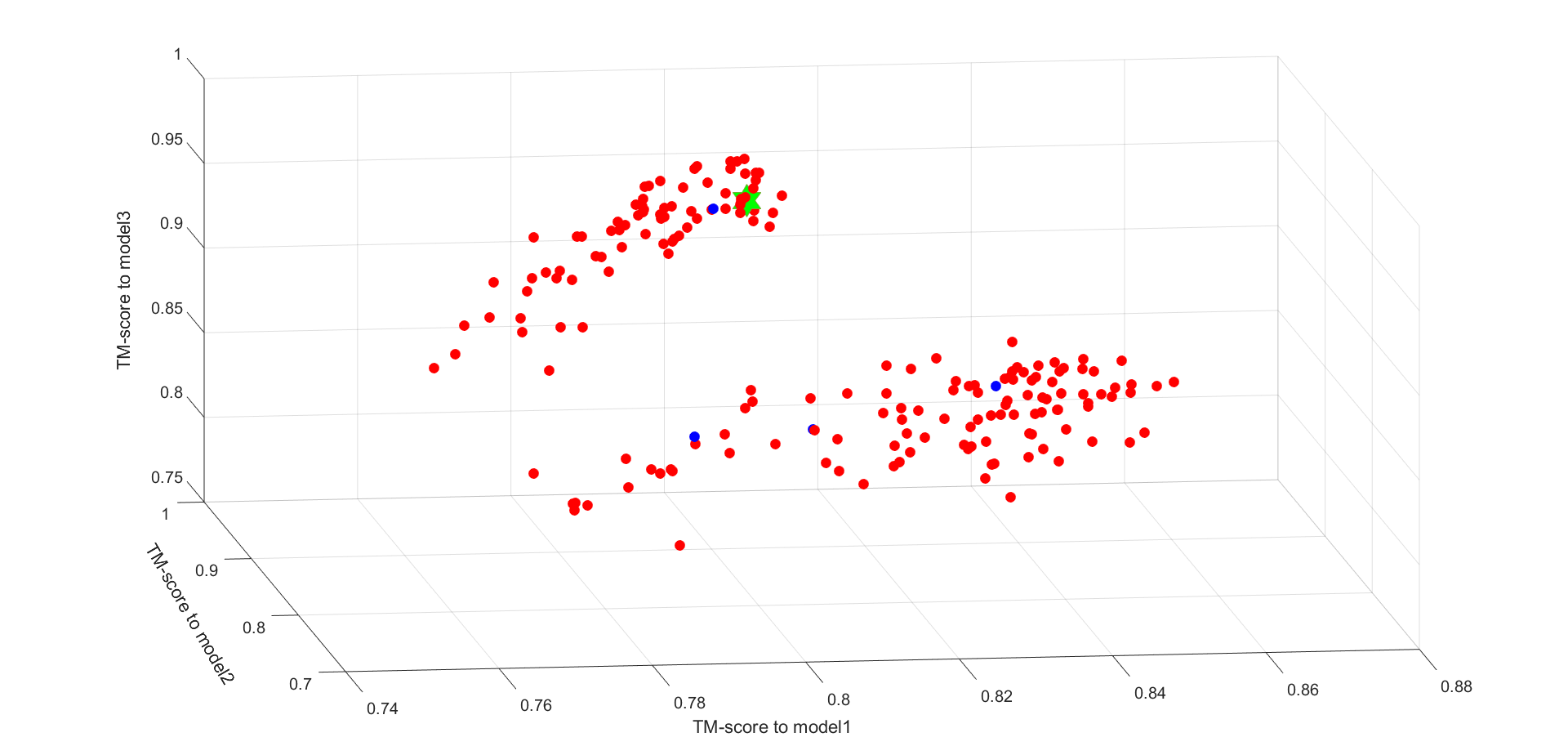
The TM-score space of TR821



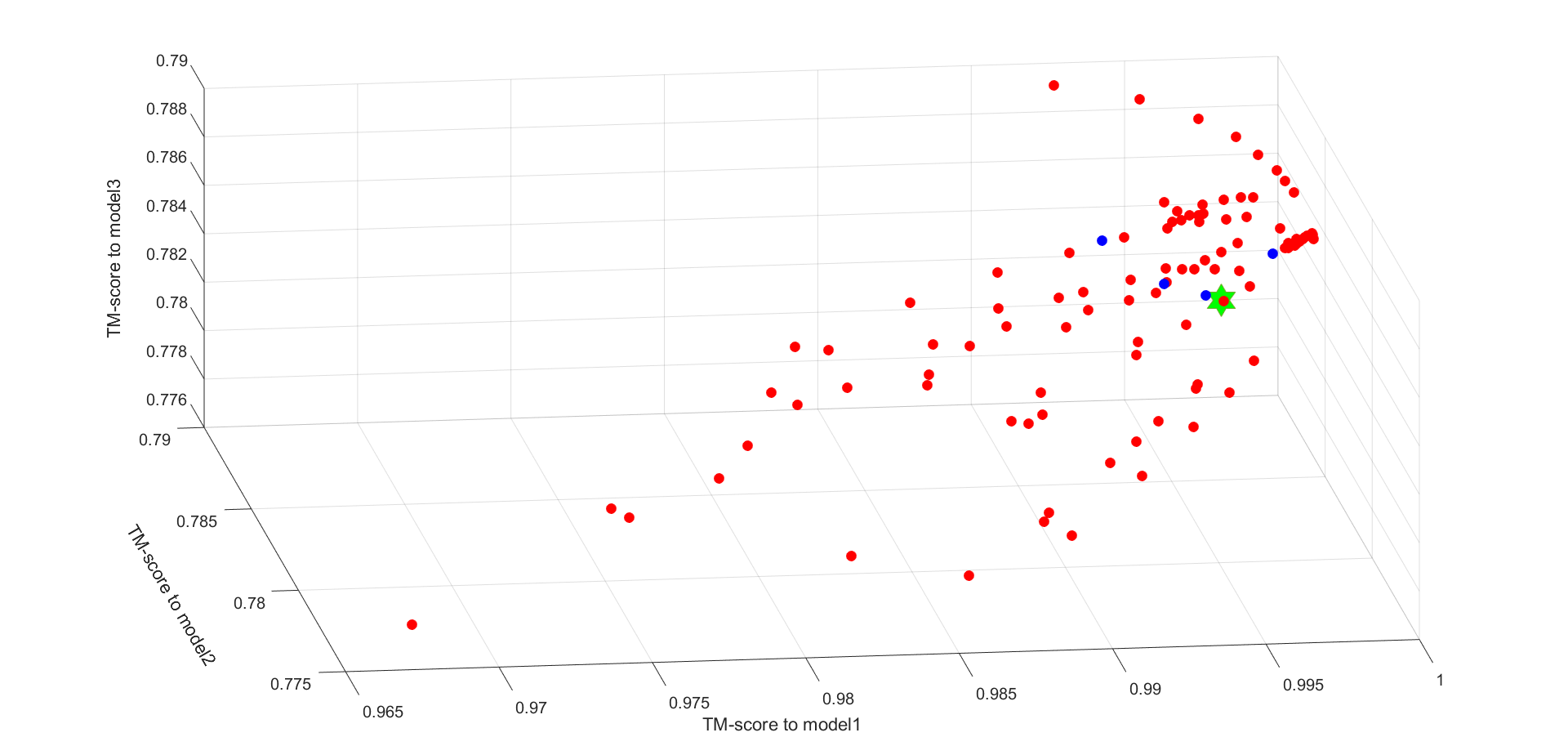
The TM-score space of TR822



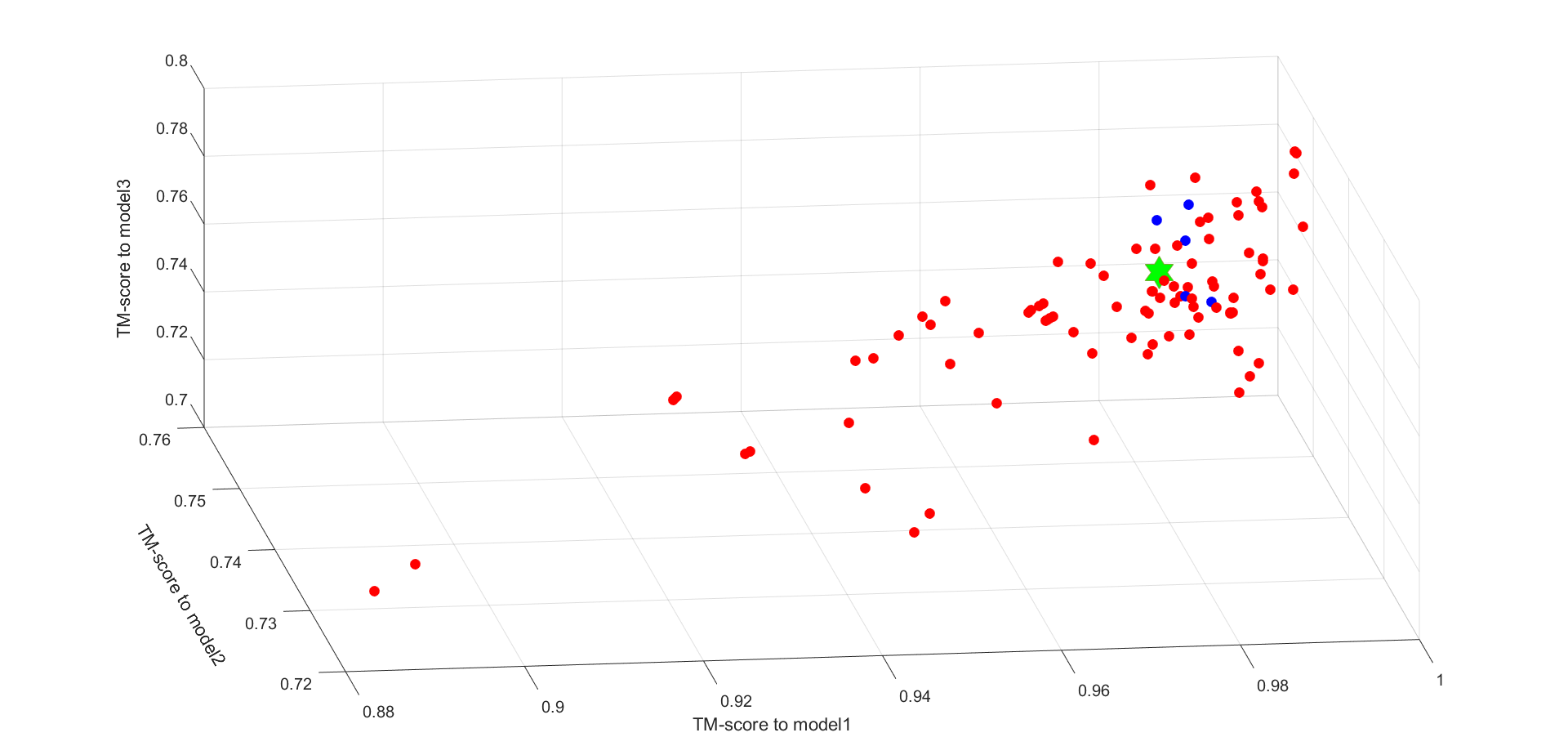
The TM-score space of TR829



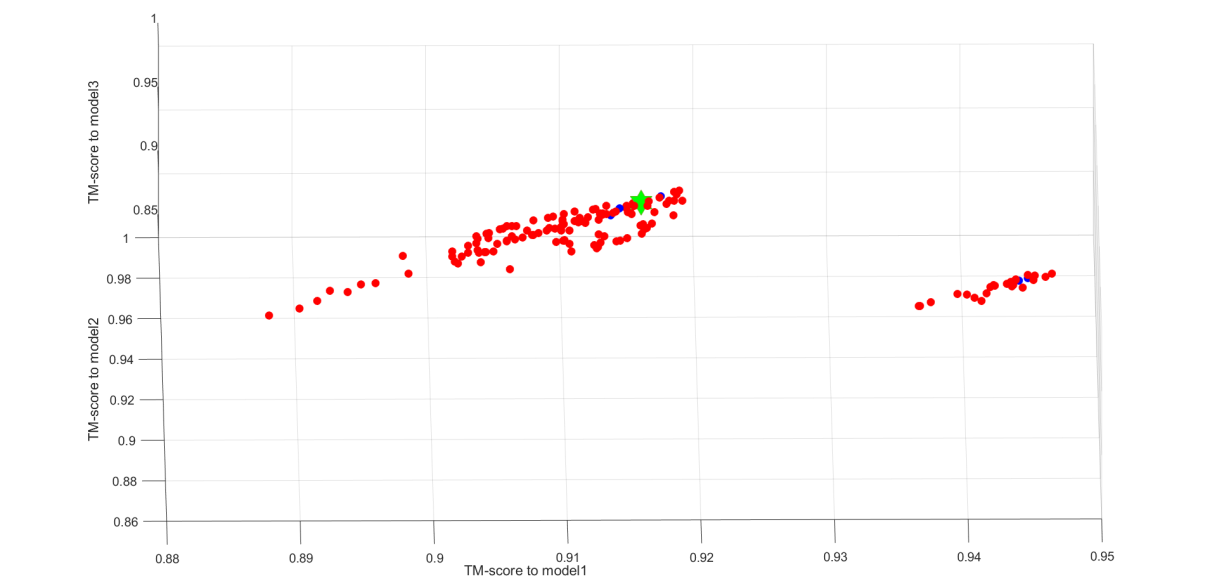
The TM-score space of TR833



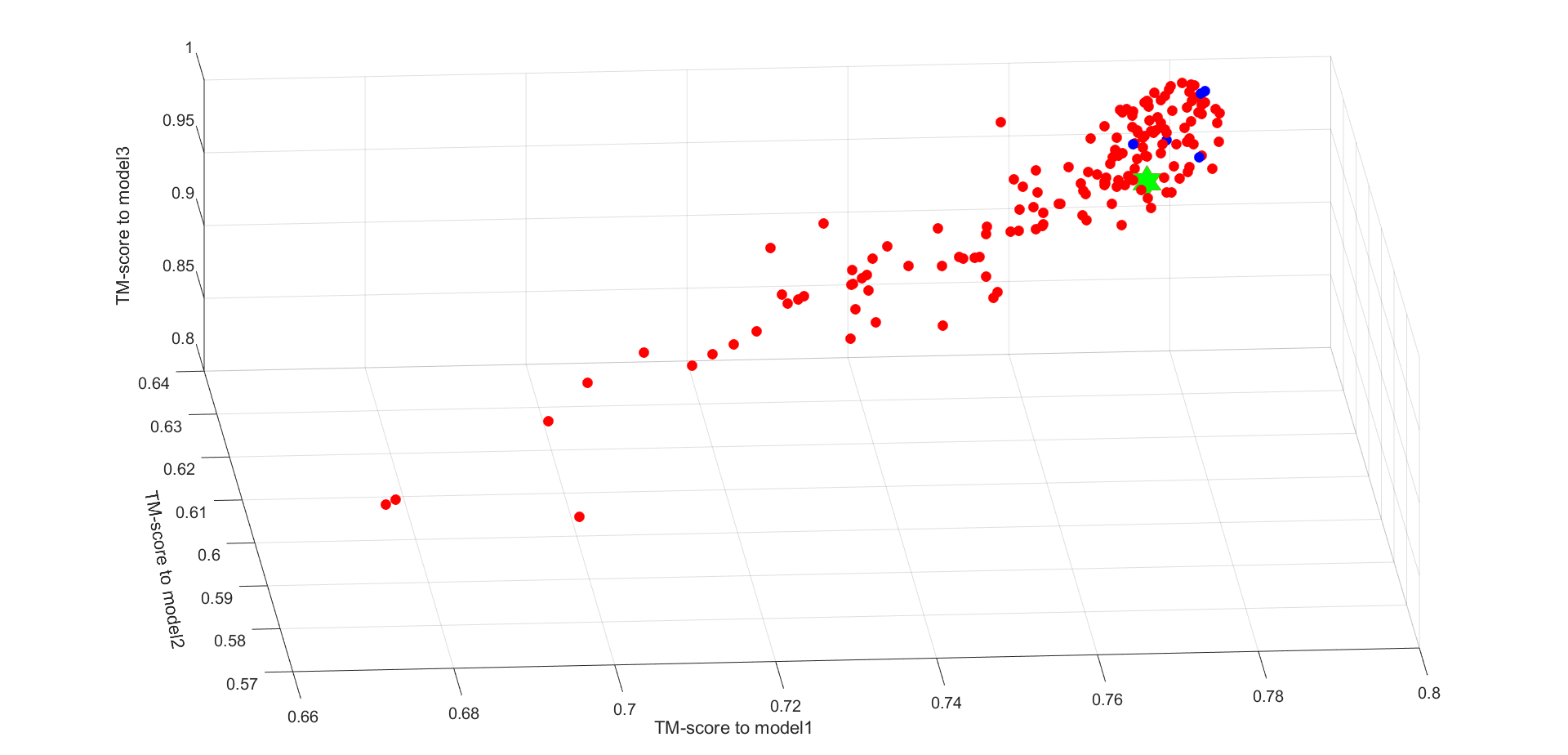
The TM-score space of TR848



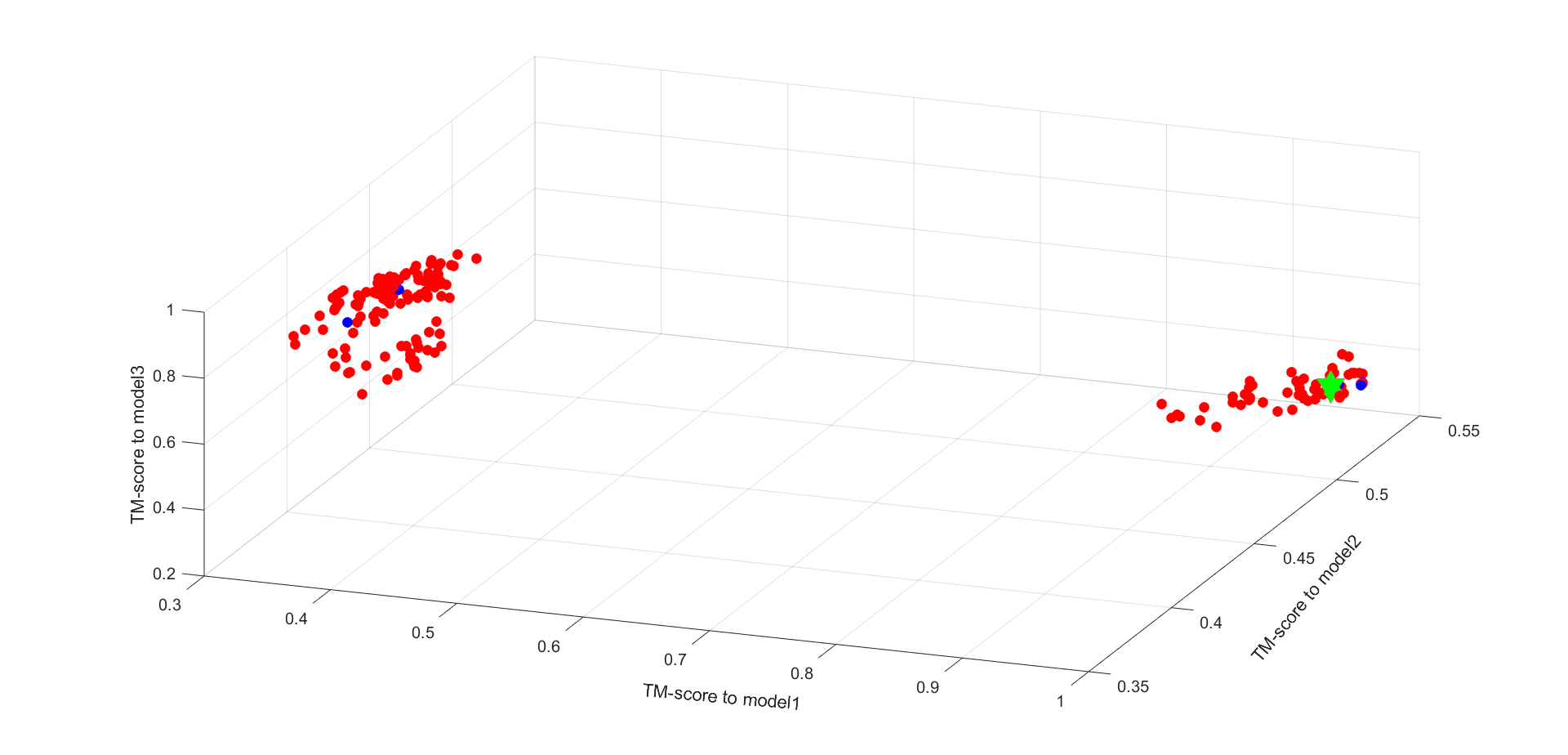
The TM-score space of TR854



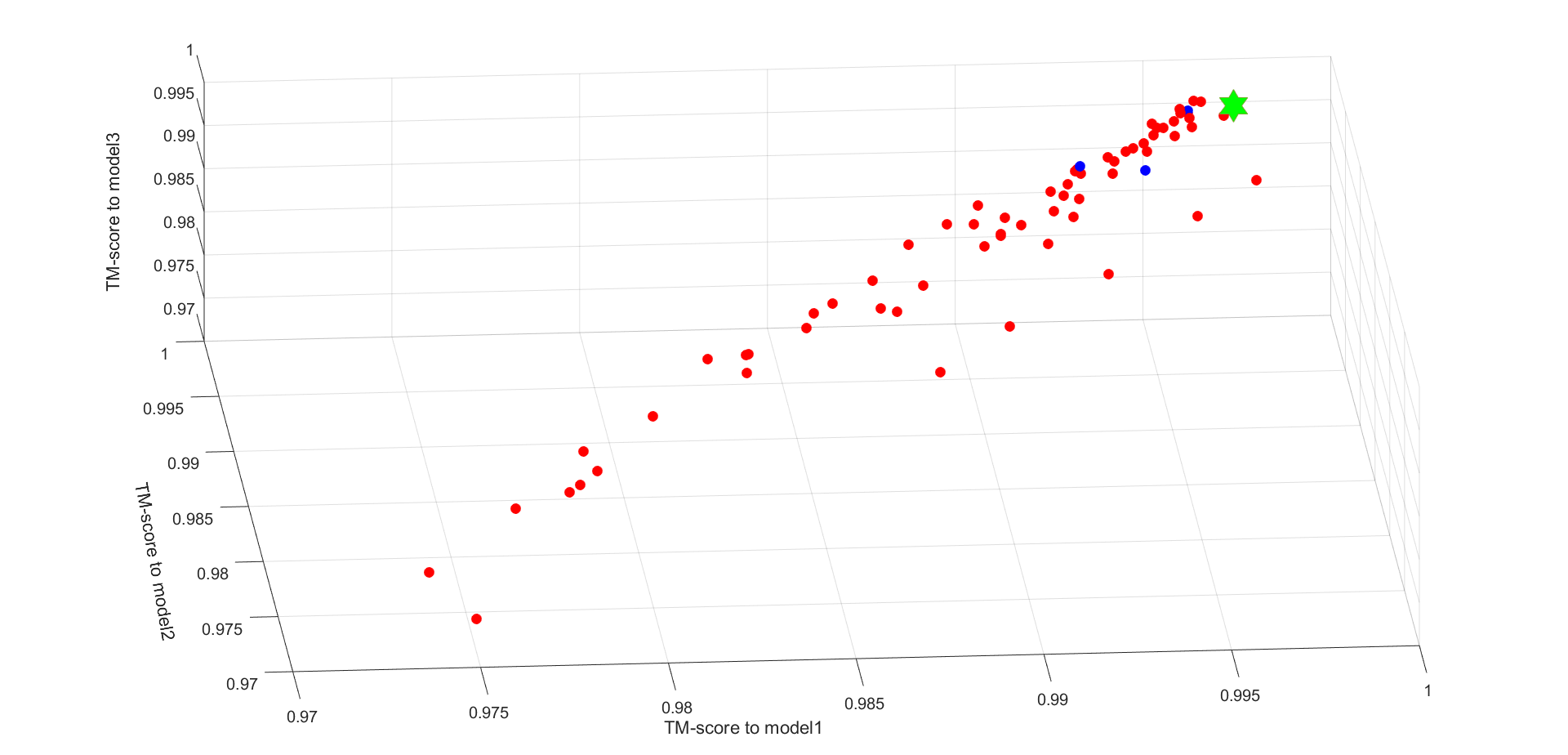
The TM-score space of TR856



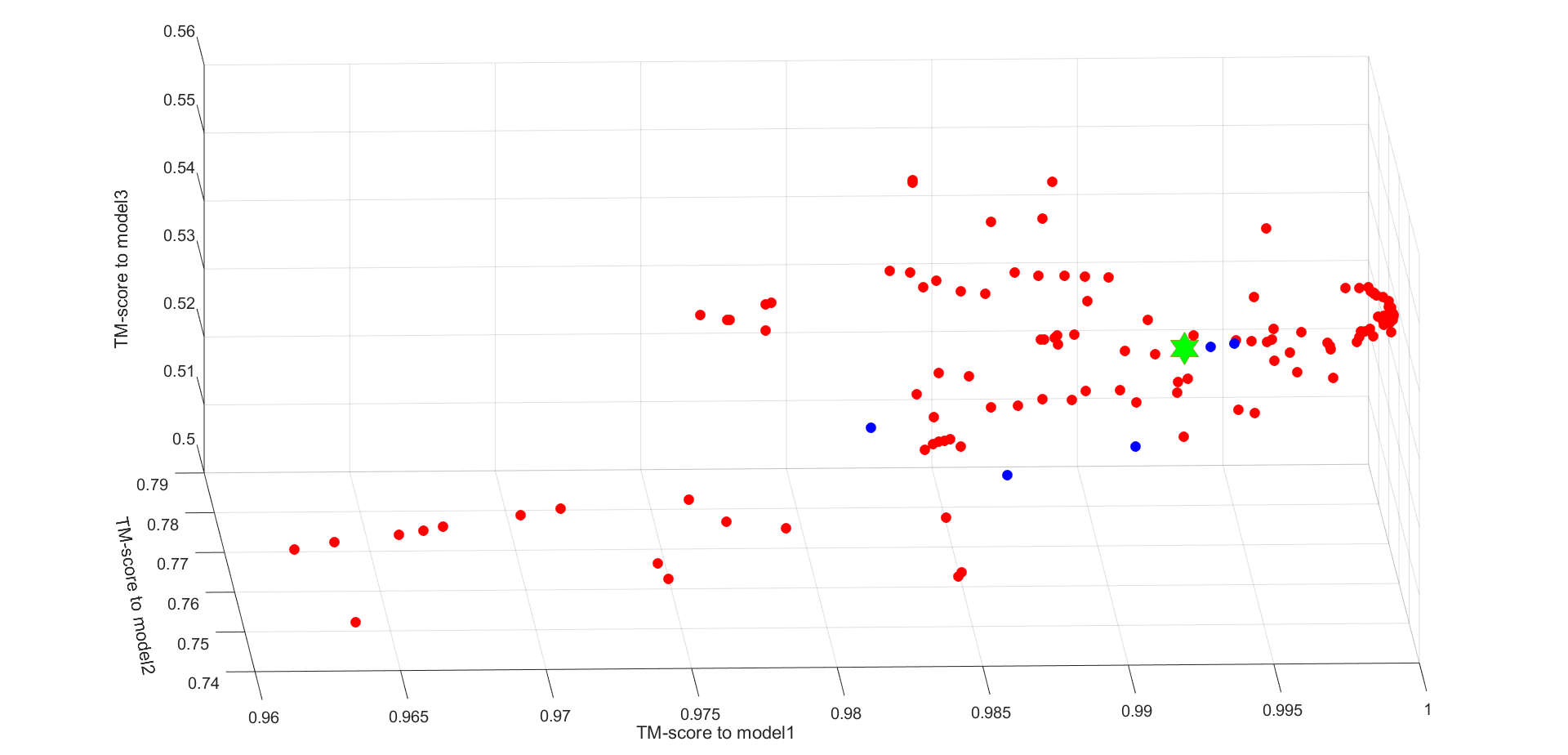
The TM-score space of TR857



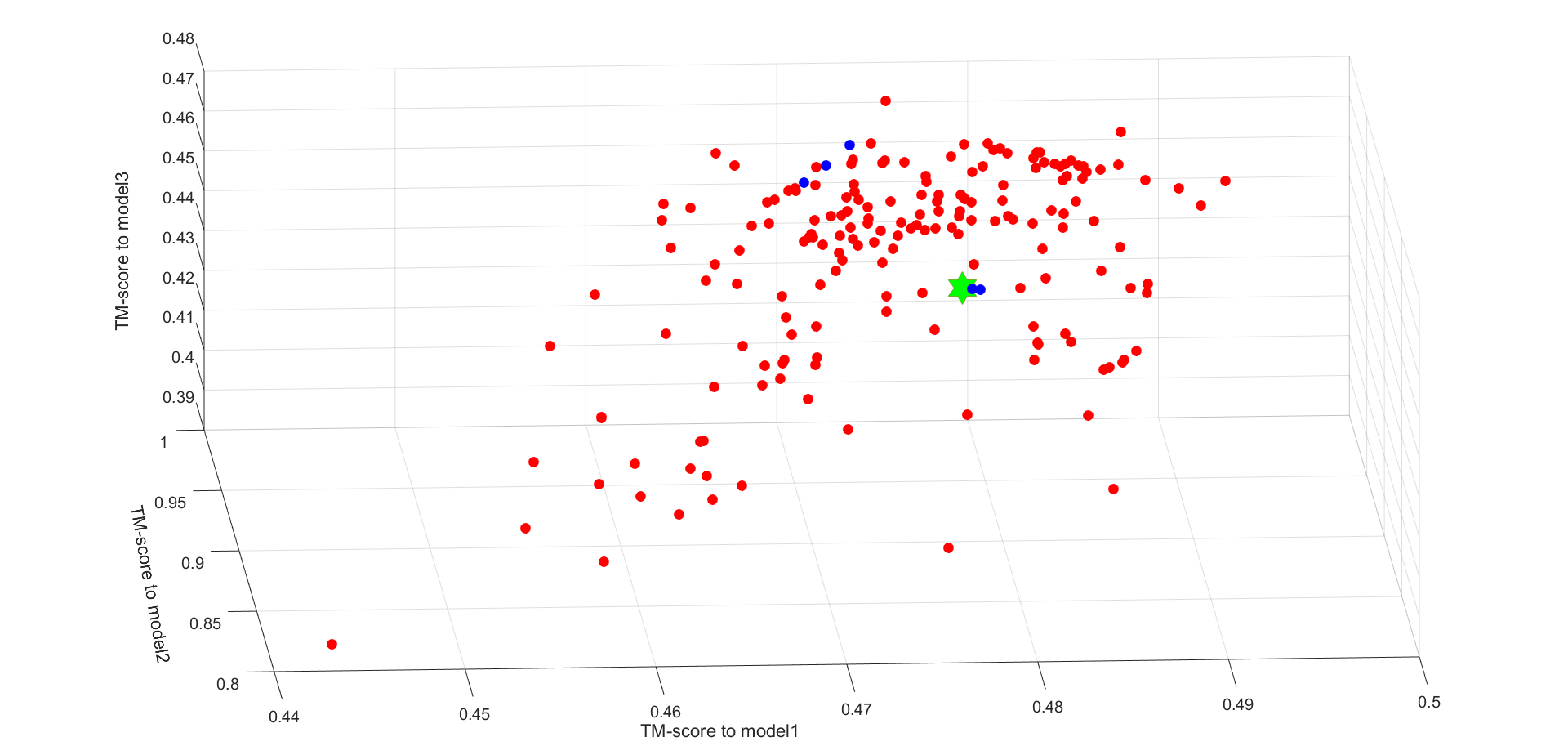
The TM-score space of TR862



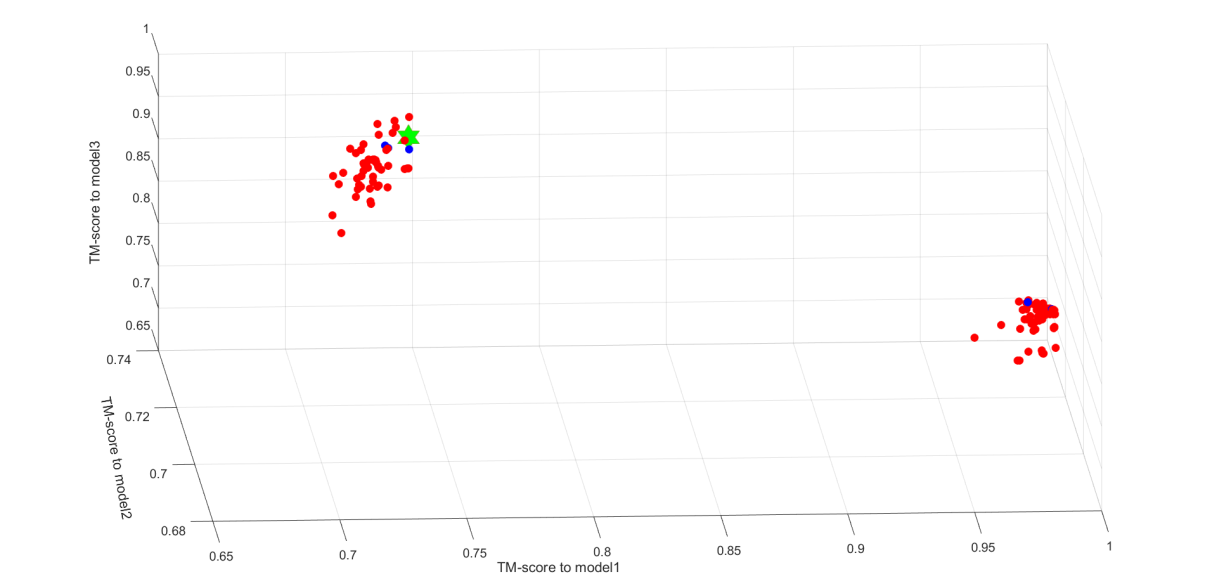
The TM-score space of TR866



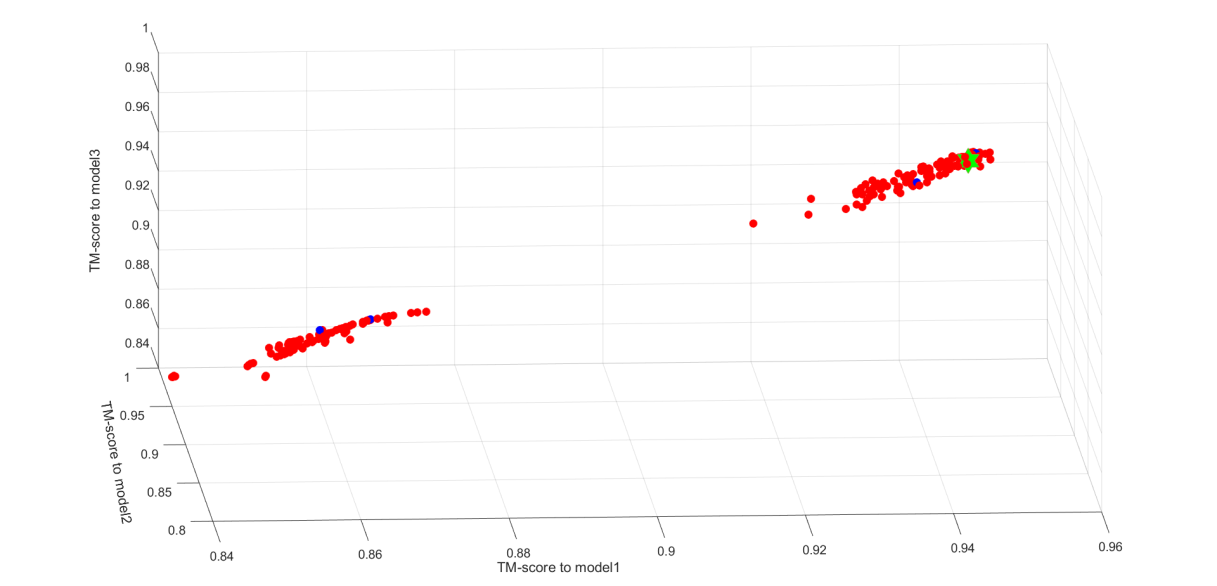
The TM-score space of TR868



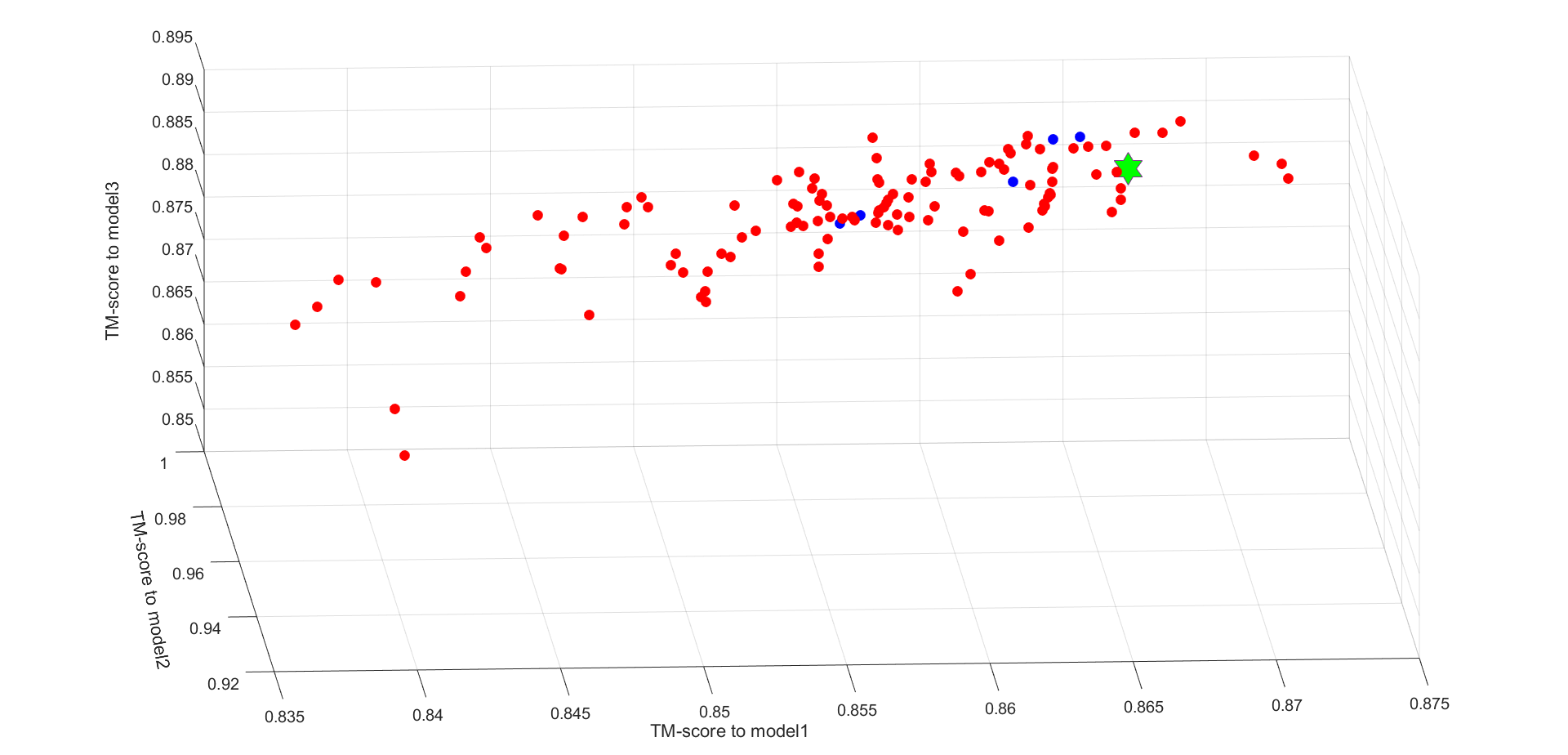
The TM-score space of TR870



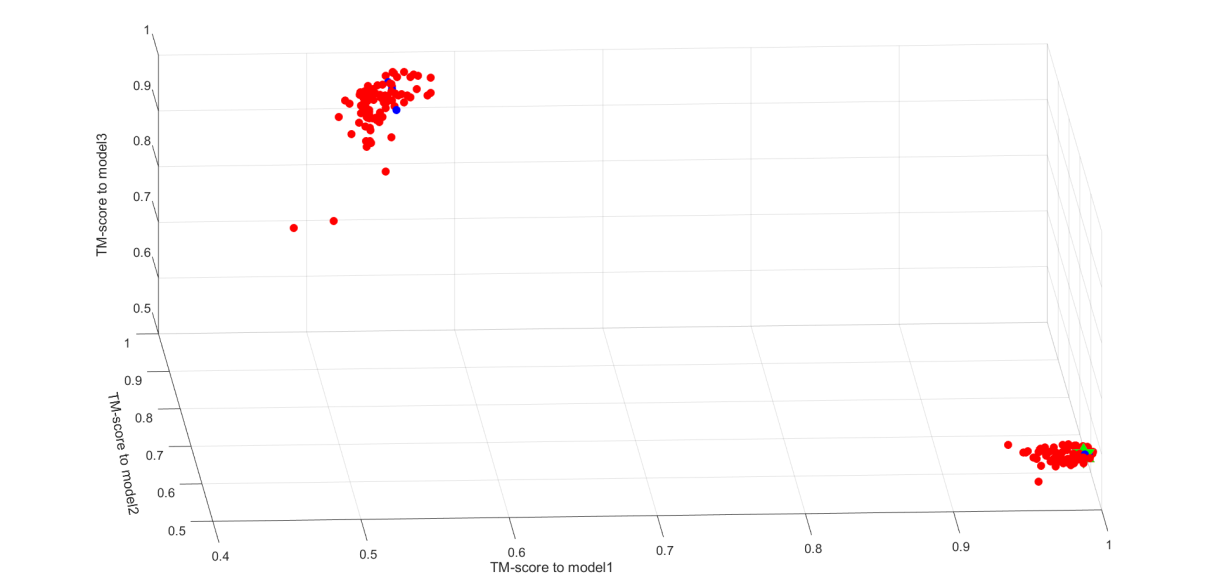
The TM-score space of TR872



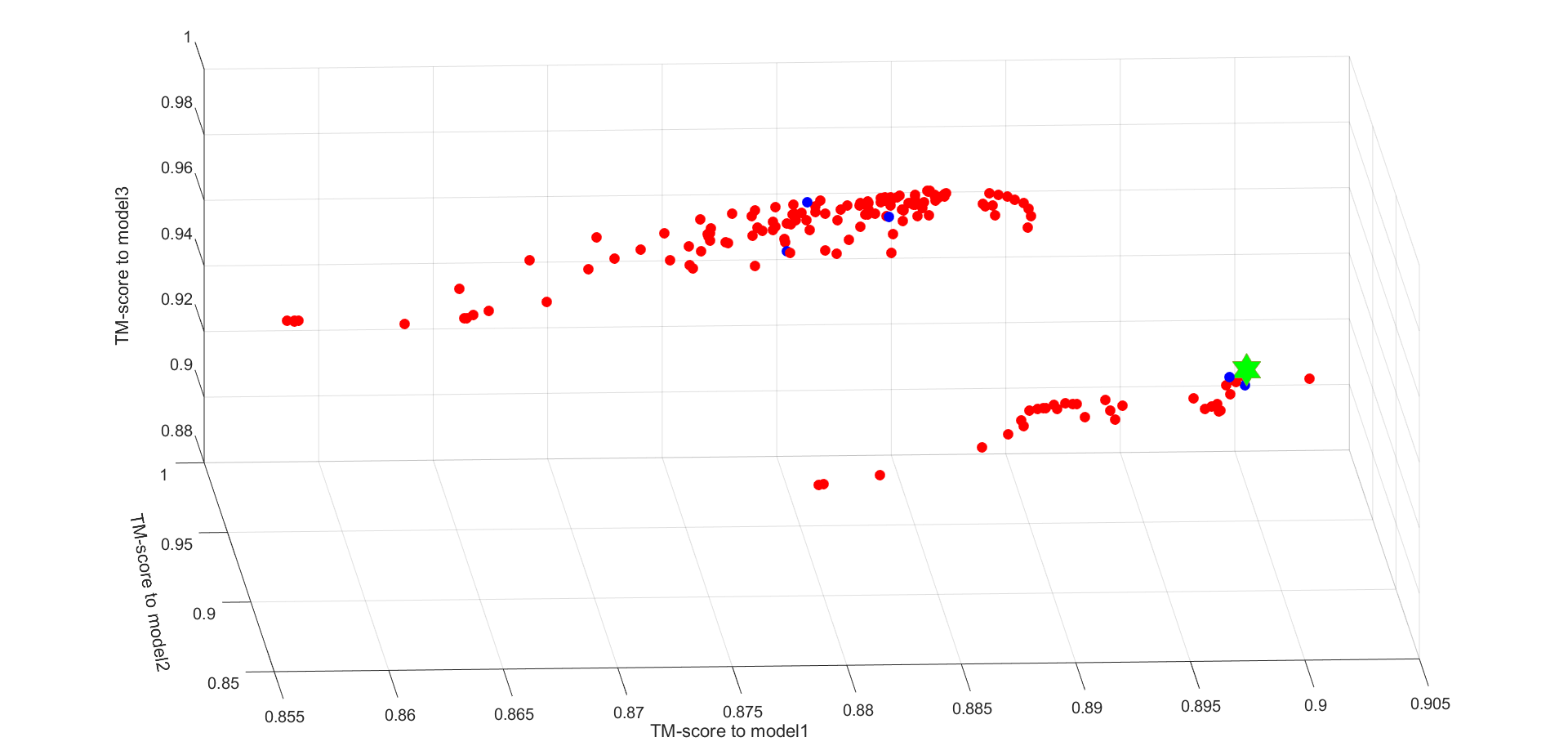
The TM-score space of TR877



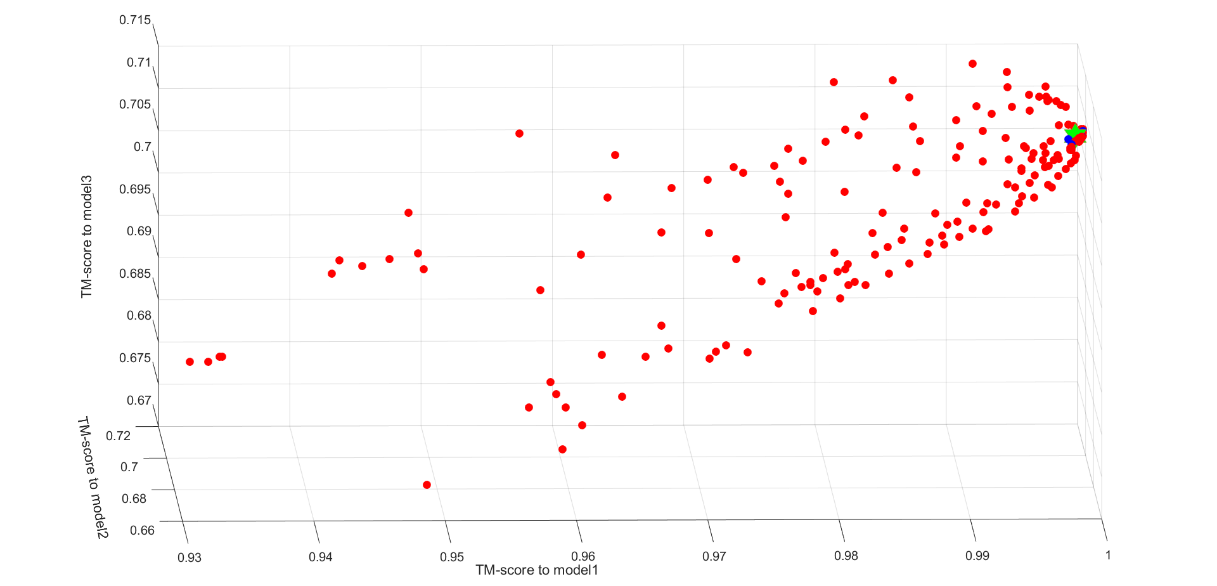
The TM-score space of TR882



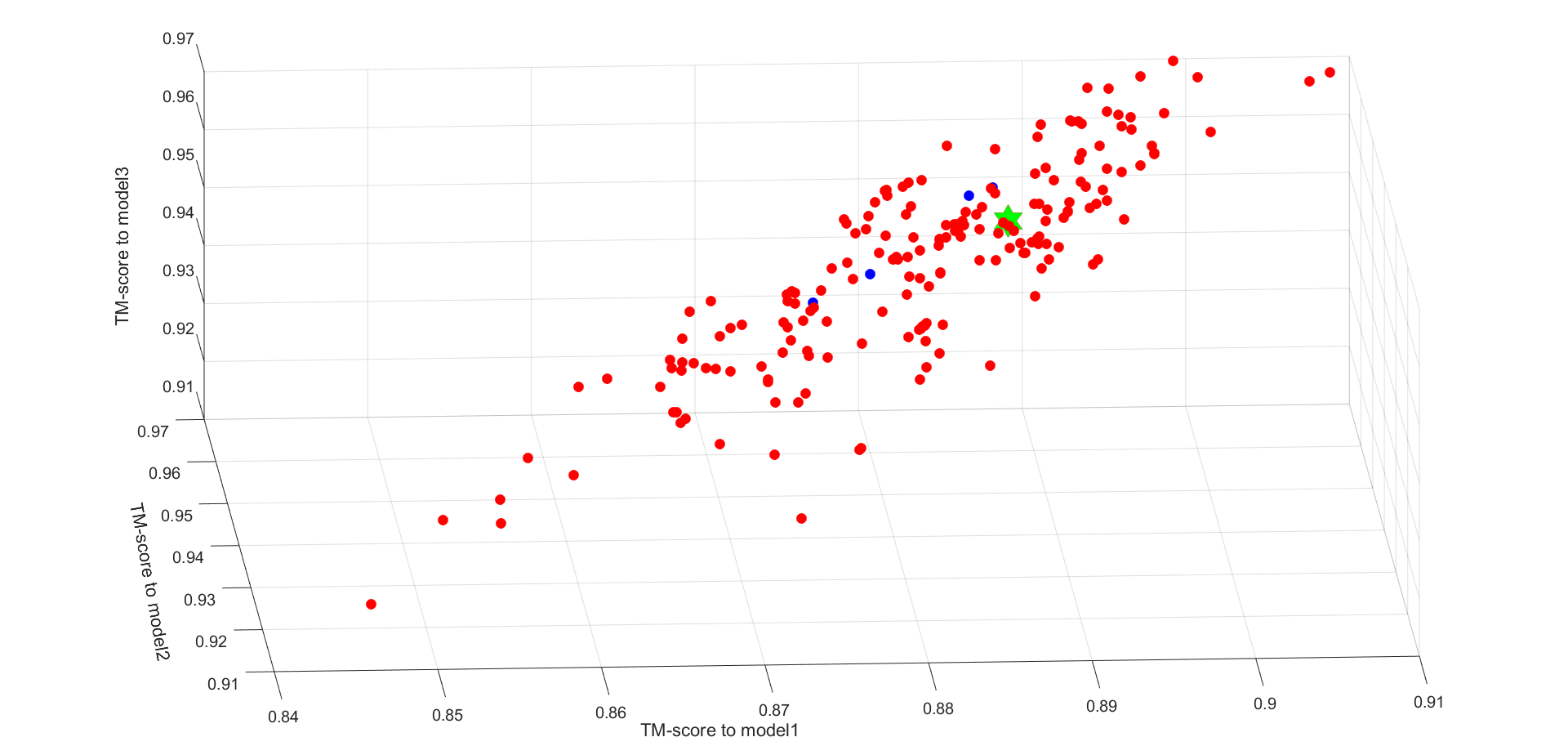
The TM-score space of TR884



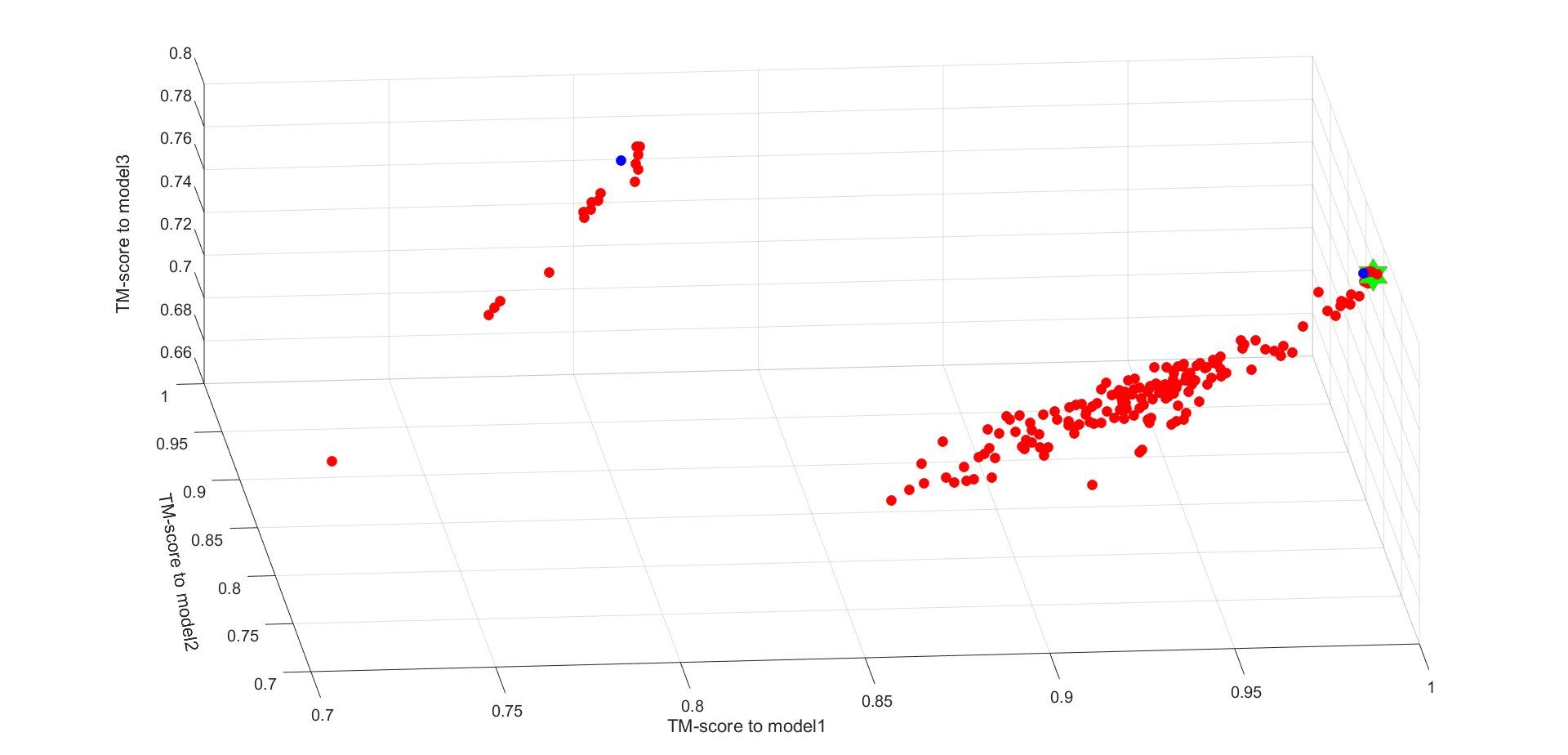
The TM-score space of TR885



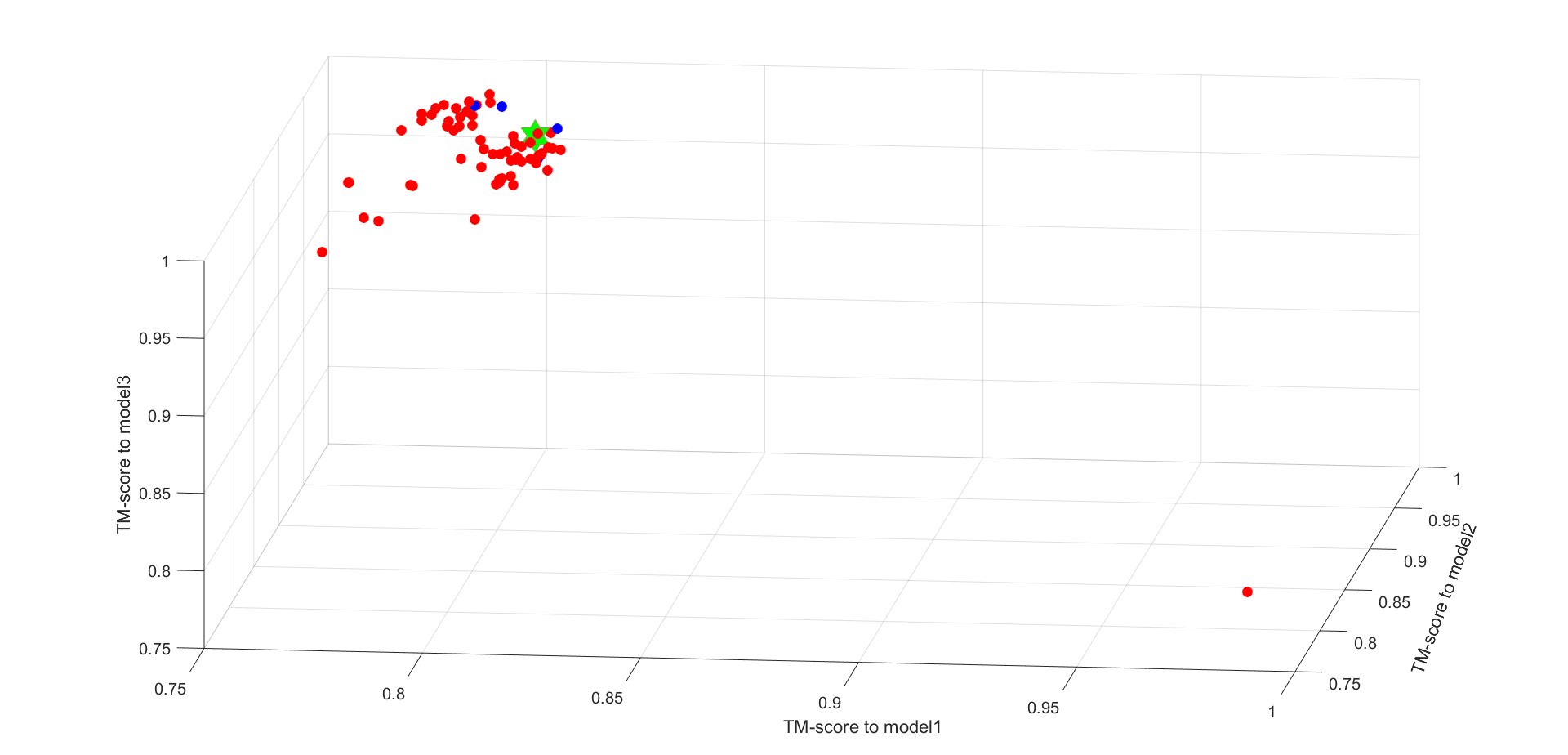
The TM-score space of TR909



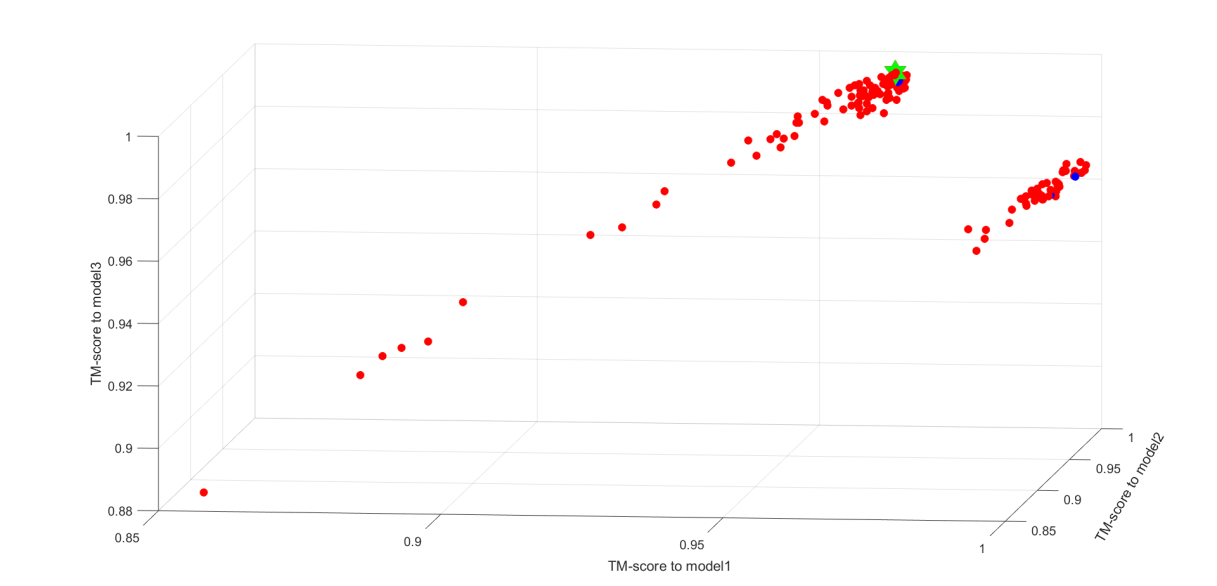
The TM-score space of TR922



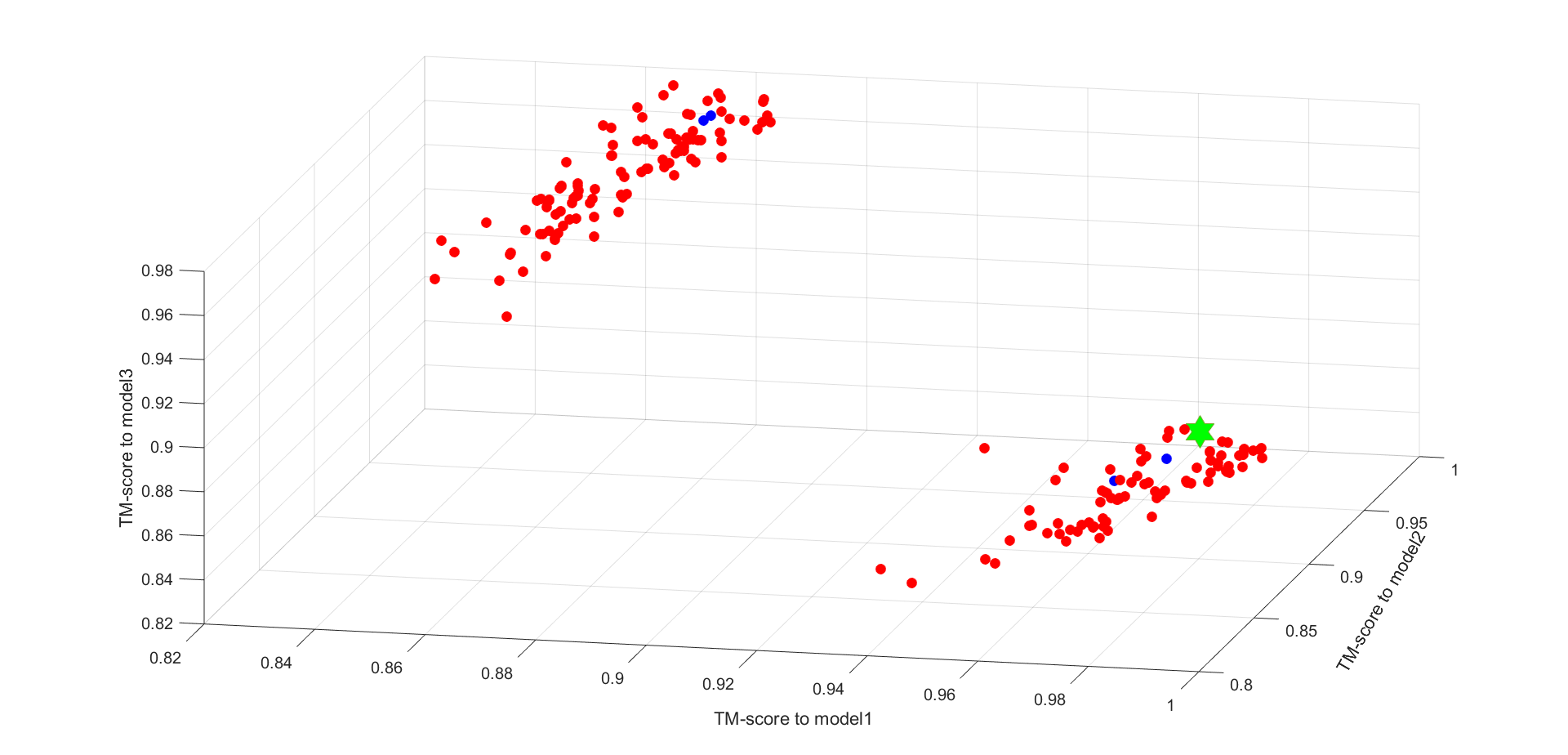
The TM-score space of TR945



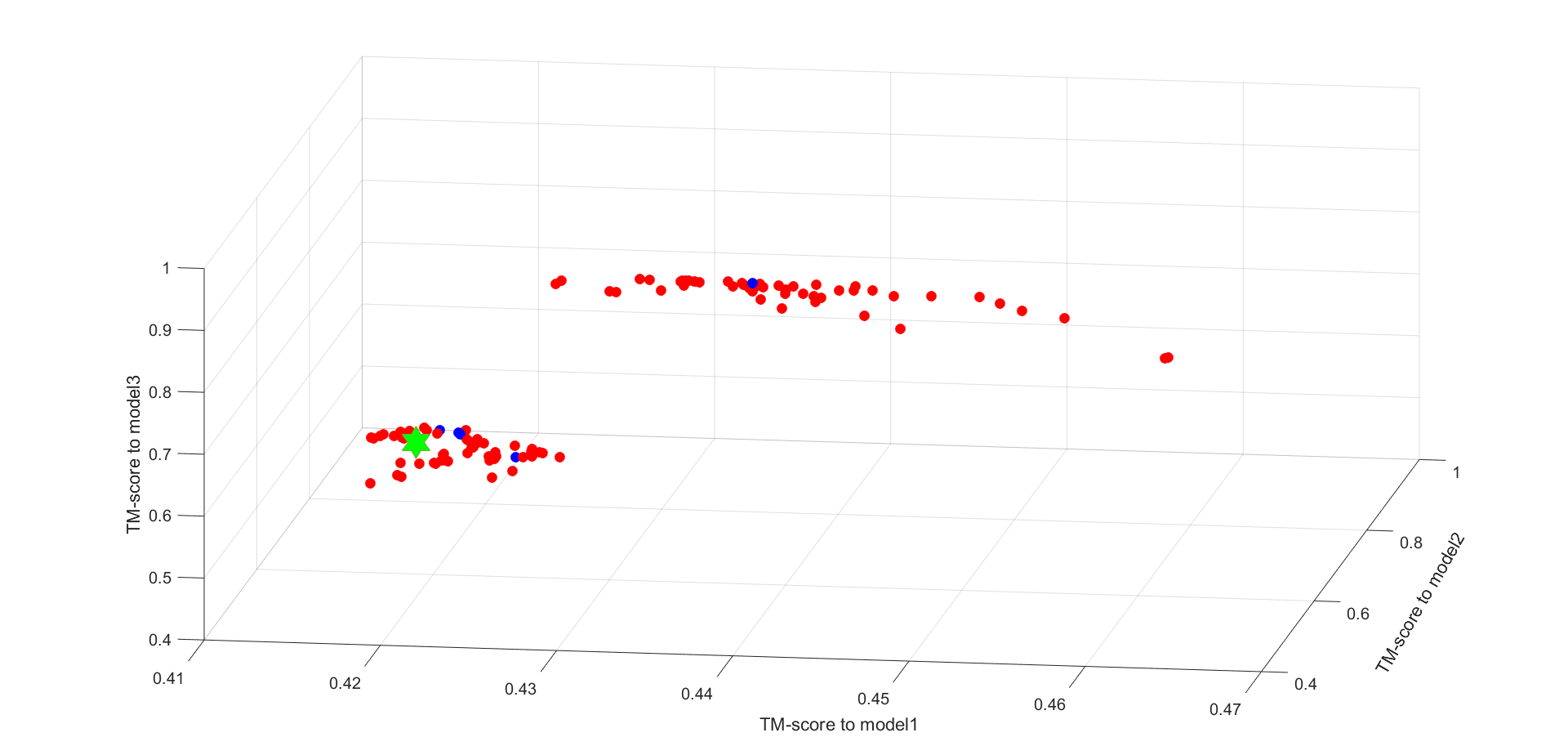
The TM-score space of TR948



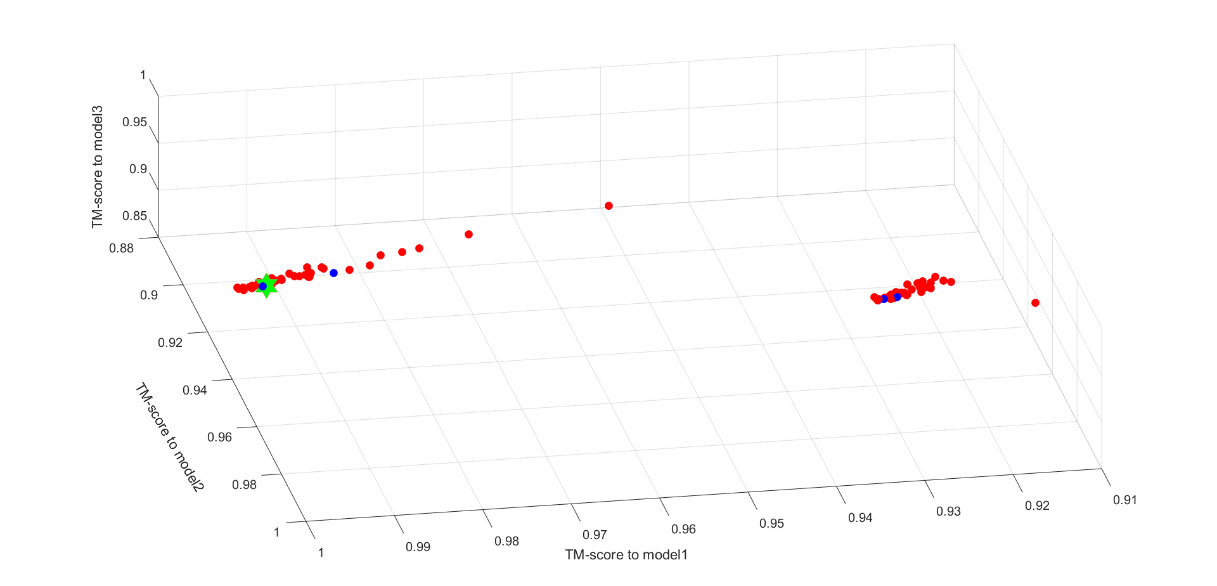
The TM-score space of TR520



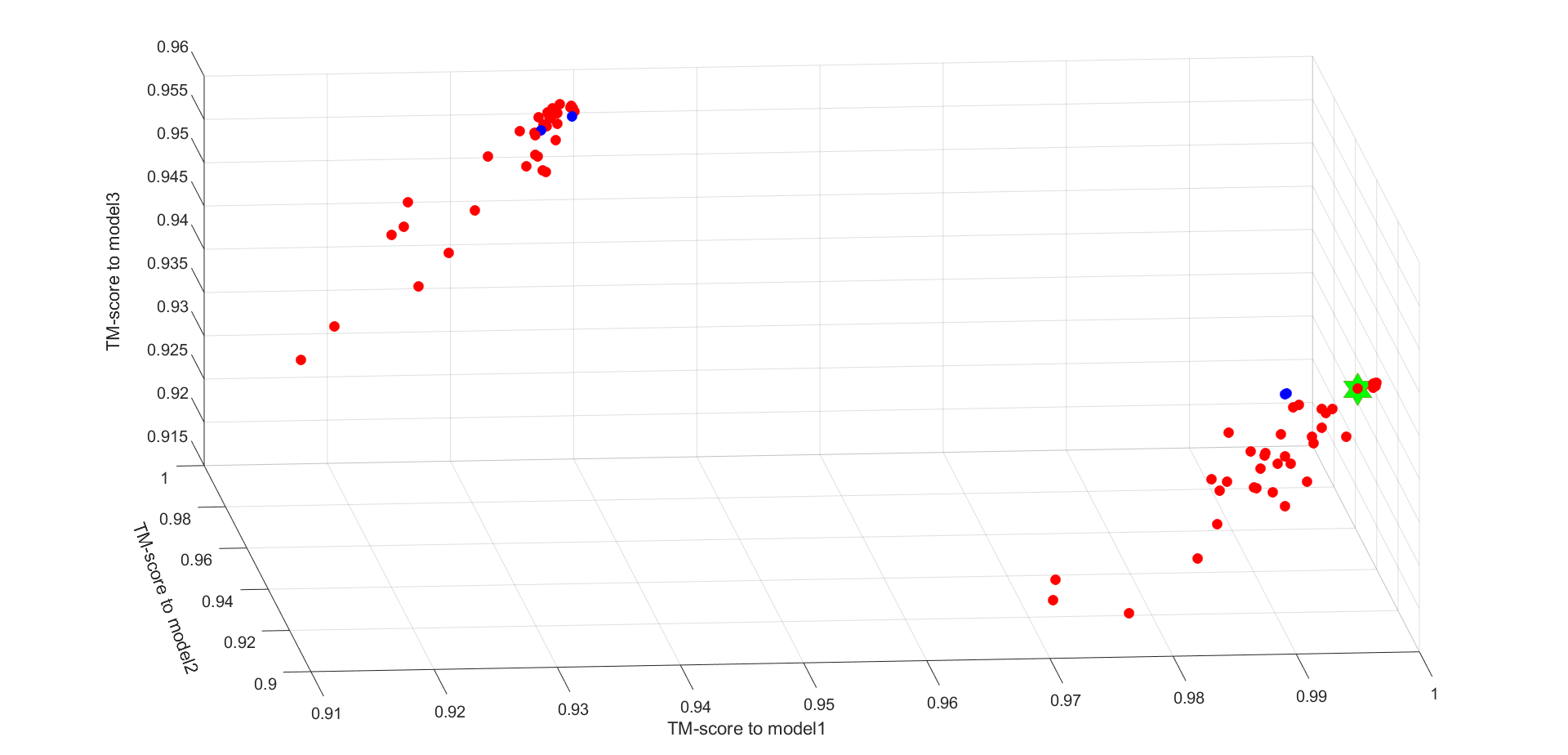
The TM-score space of TR594



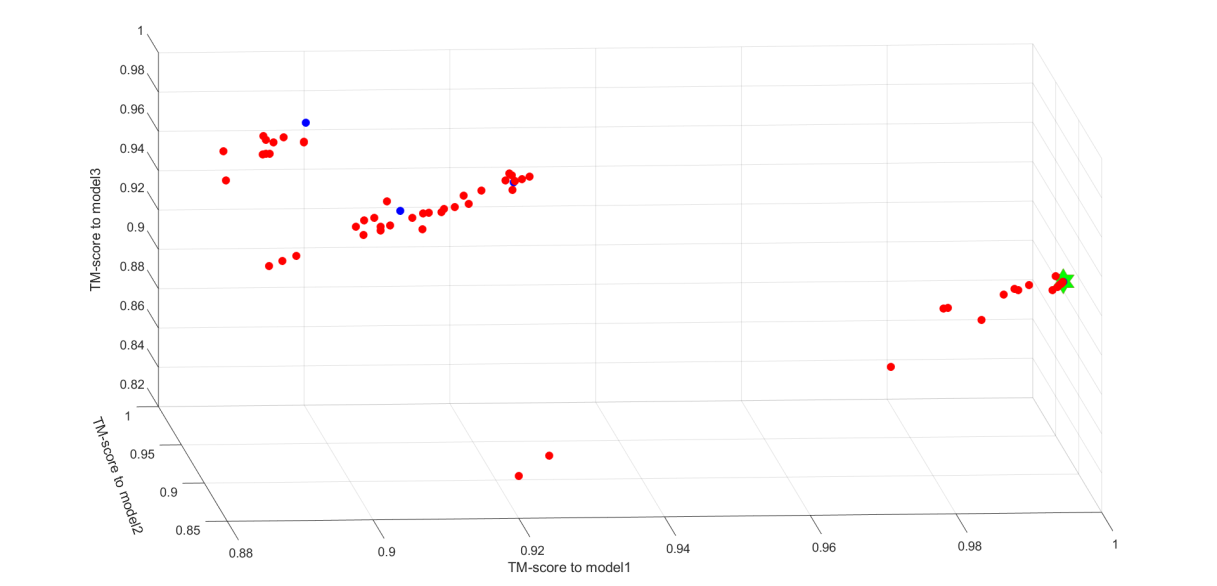
The TM-score space of TR869



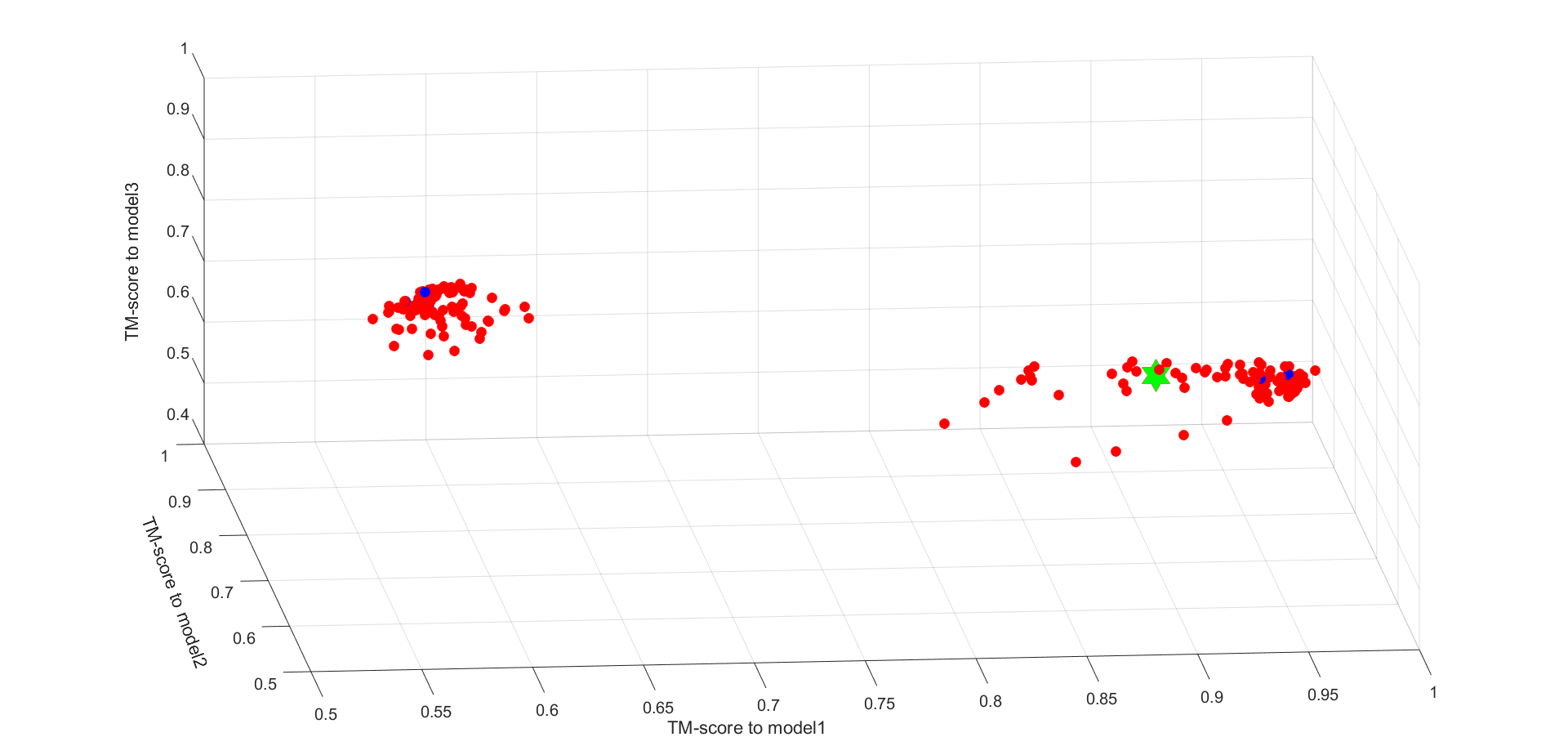
The TM-score space of TR879



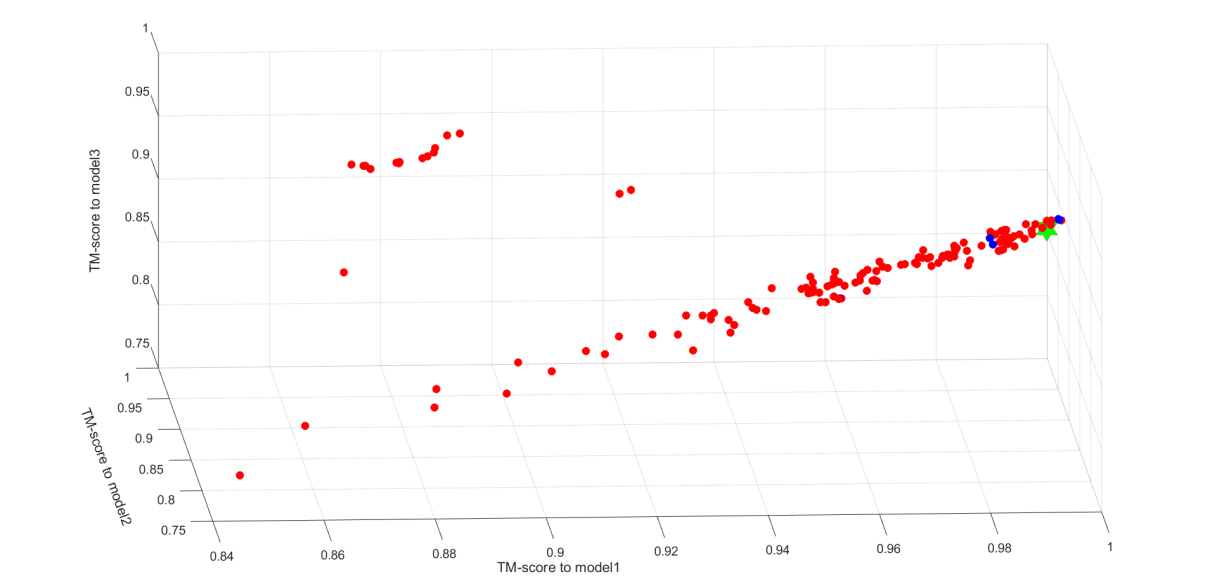
The TM-score space of TR891



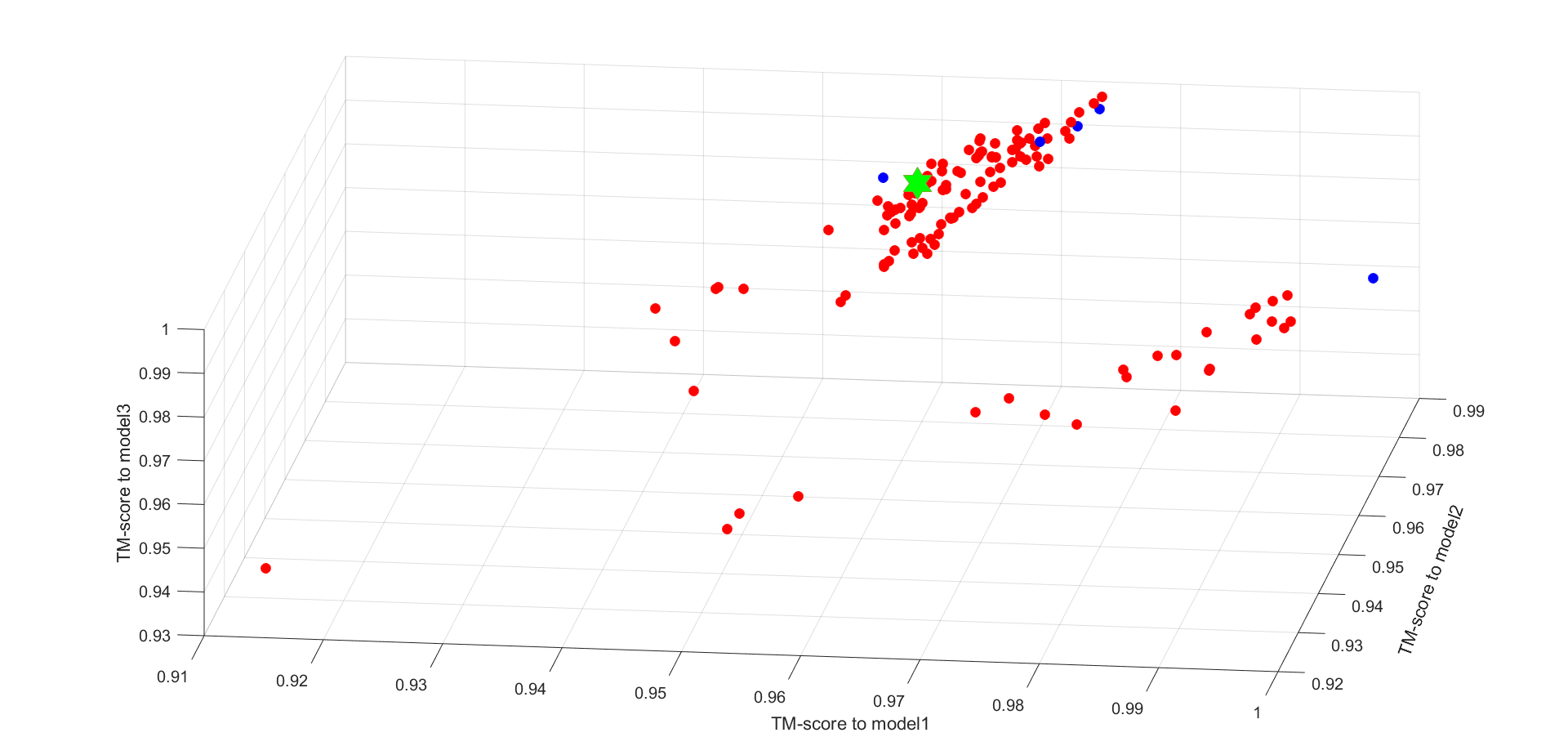
The TM-score space of TR893



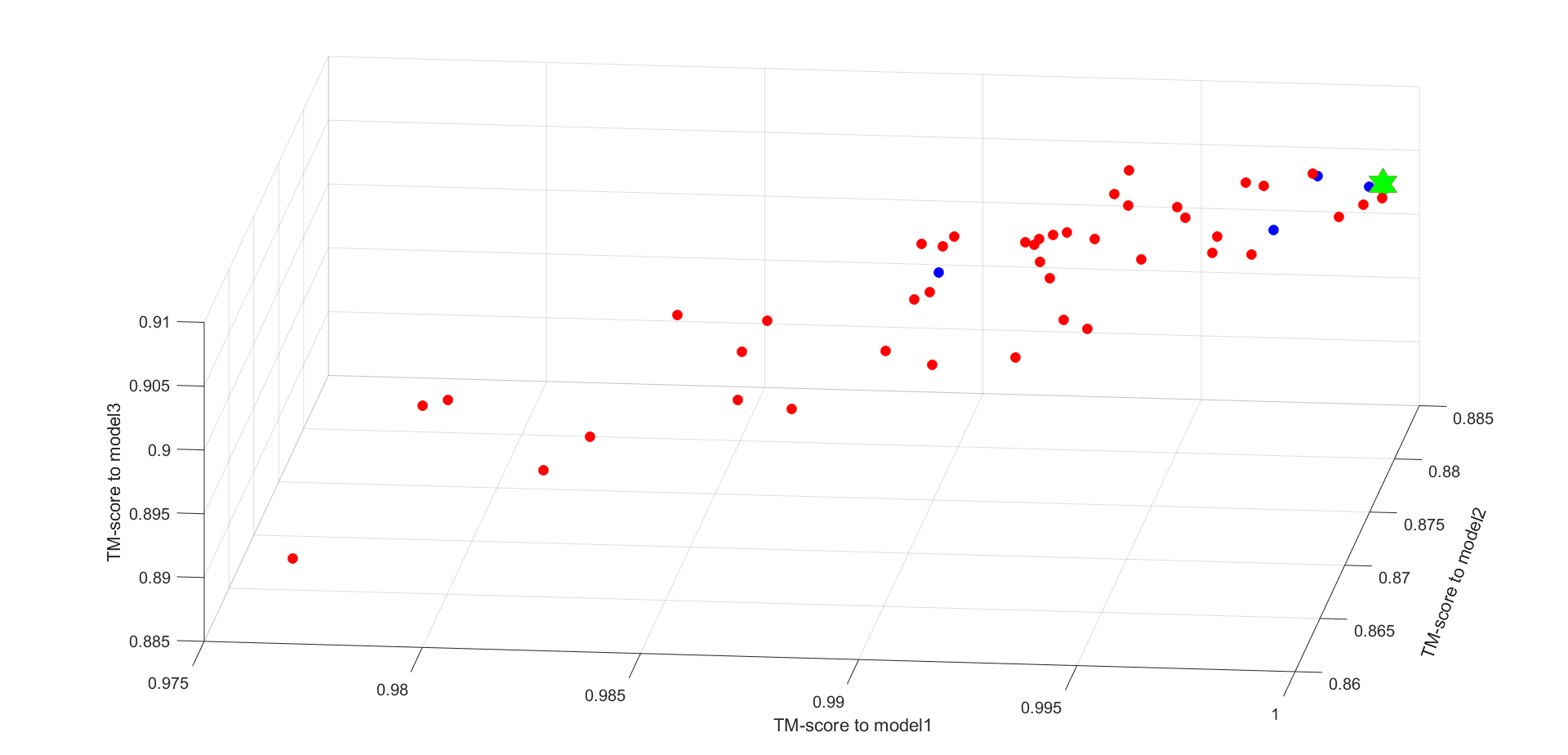
The TM-score space of TR894



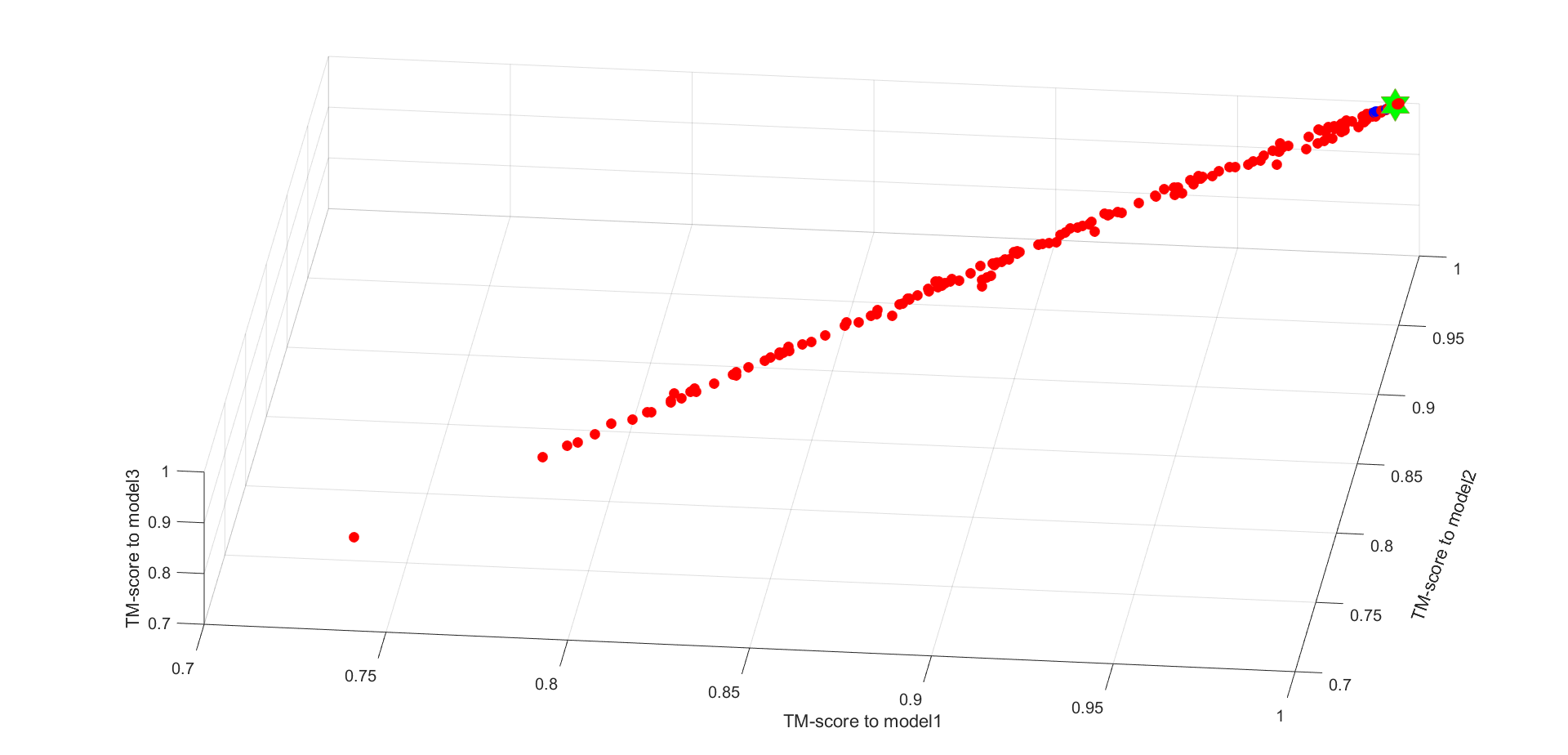
The TM-score space of TR895



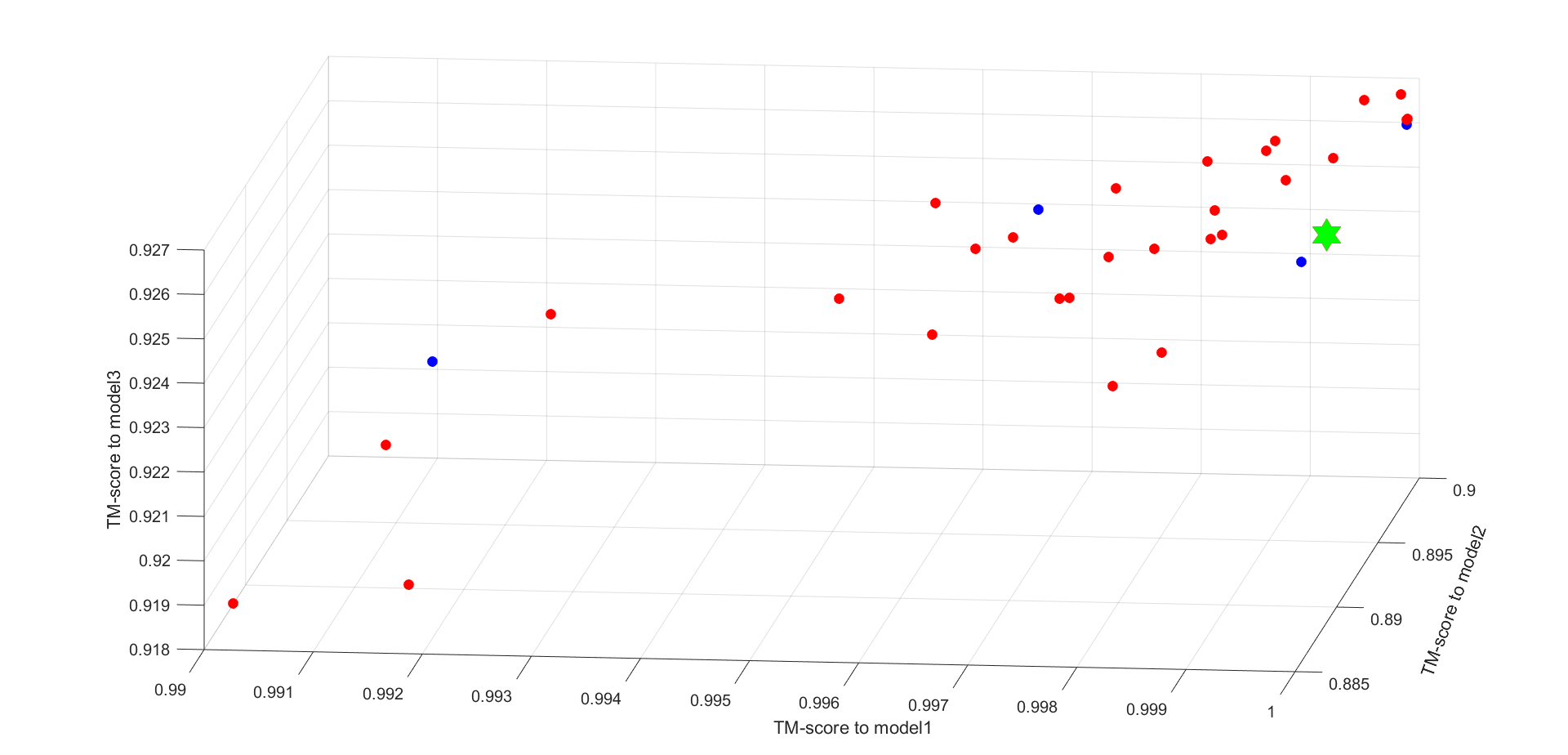
The TM-score space of TR920



The TM-score space of TR921



The TM-score space of TR928



The TM-score space of TR944

**Figure S6**. The top five final selected models, the best output models and the candidate models of each target in the TM-score space.

**Table S2**. Student’s *t*-test the difference in the performance of different methods

|  |  |  |  |
| --- | --- | --- | --- |
| Pairs of different methods | AIR & Baker | AIR & Seok | AIR & Zhang |
| *p*-value | 0.8504 | 0.0498 | 0.0139 |

**Table S3**. Overview of the average refinement results for different energy function combinations

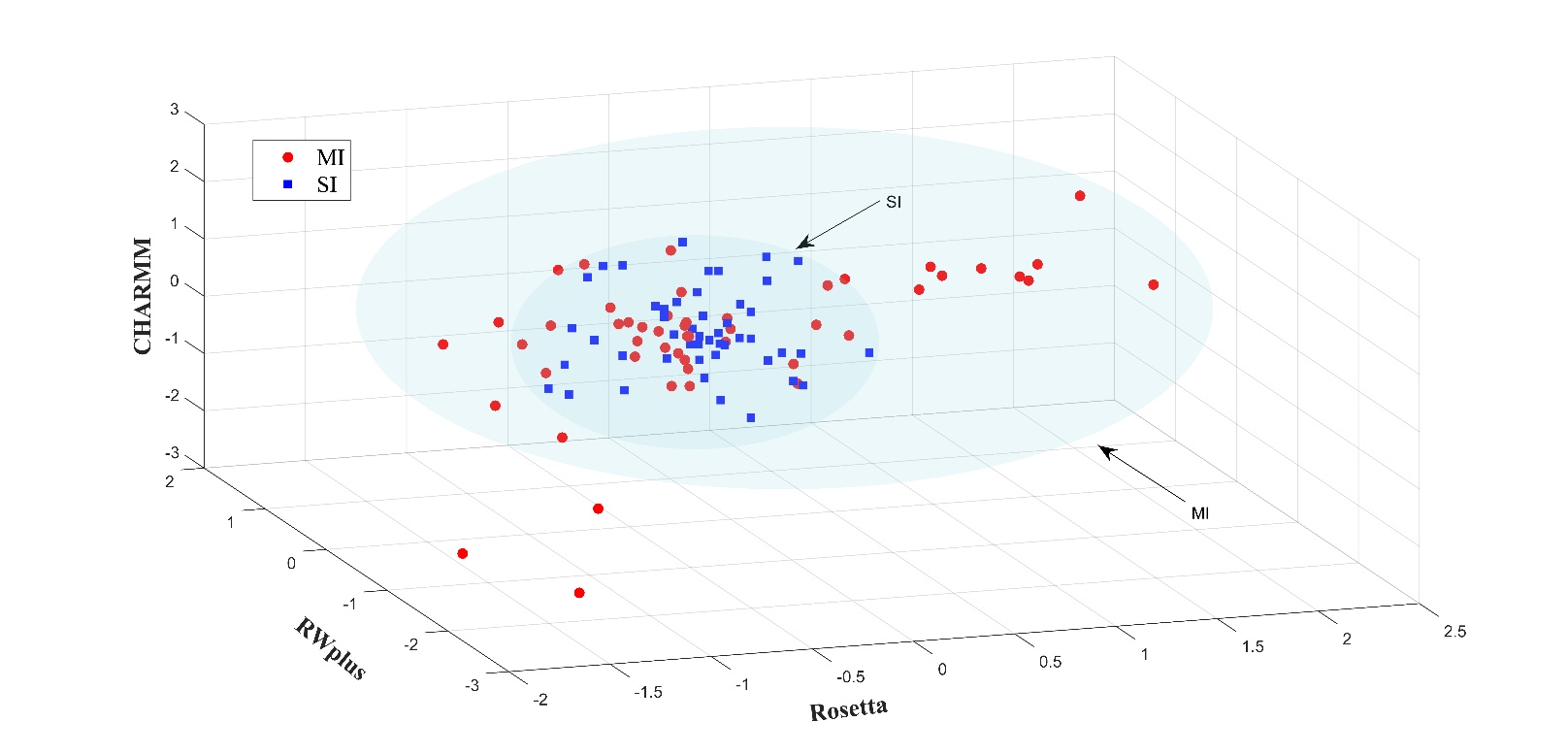
|  |  |  |
| --- | --- | --- |
| Combinations  of energy functions | Refined model (the best model) | Refined model ( model 1) |
| Increase of TM-score/ GDT-TS | Increase of TM-score/ GDT-TS |
| Rosetta & RWplus | 0.012/1.45 | 0.012/1.24 |
| CHRAMM & RWplus | 0.010/1.32 | 0.009/1.19 |
| CHARMM & Rosetta | **0.013/1.50** | **0.012/1.31** |

**Table S4**. Performance comparison between AIR with 3Drefine on targets of CASP13

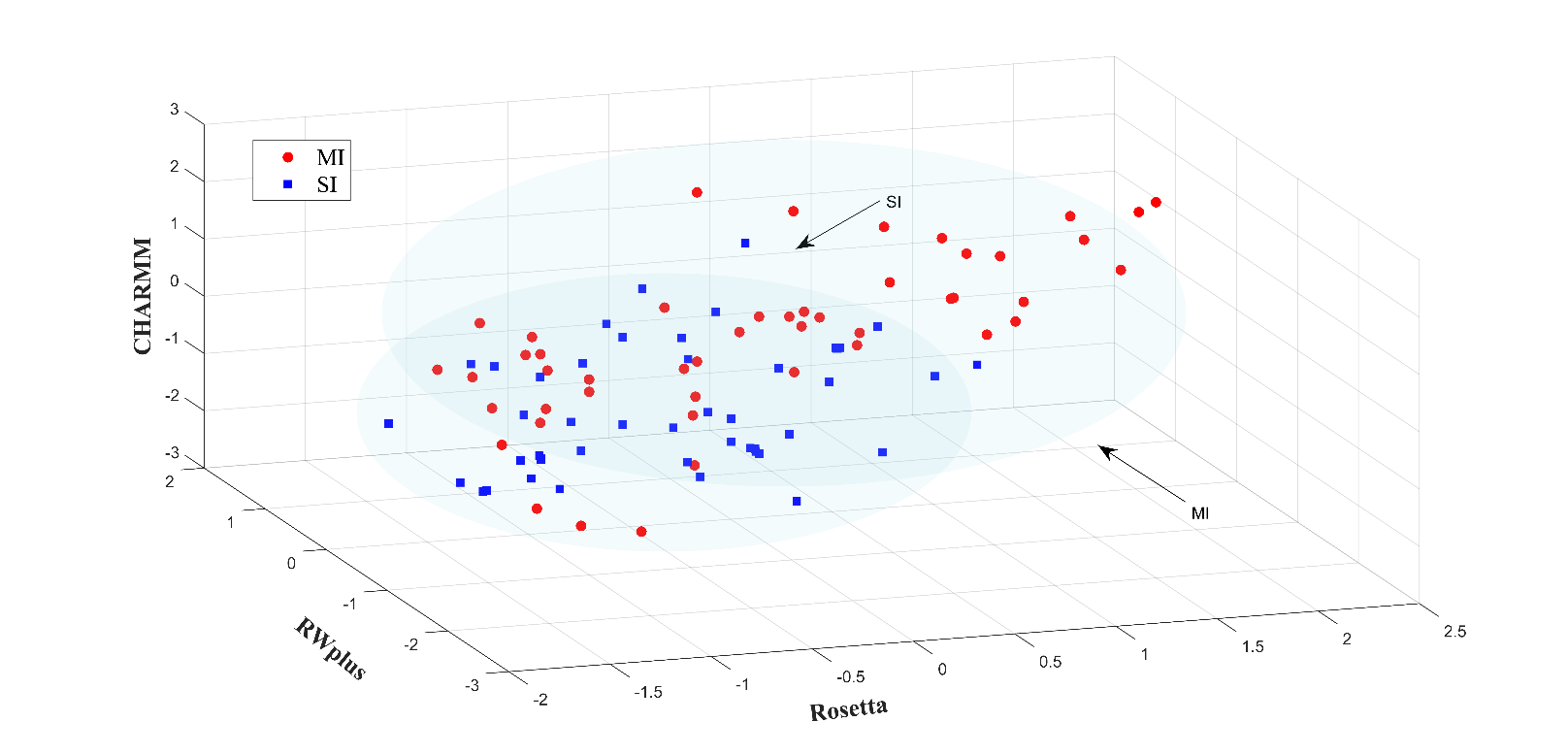
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test targets | Length | Initial model | Model2 a | Model3 a | AIR (one template) b | AIR (three templates) | 3Drefine b |
| TM-score/ GDT-TS | TM-score/ GDT-TS | TM-score/ GDT-TS | TM-score/ GDT-TS | TM-score/ GDT-TS | TM-score/  GDT-TS |
| R0968s1 | 118 | 0.74/66.74 | 0.72/65.80 | 0.70/62.71 | 0.74/67.16 | **0.75/67.80** | 0.71/65.04 |
| R0968s2 | 116 | 0.77/71.30 | 0.70/62.92 | 0.71/63.48 | **0.80/75.43** | 0.77/70.87 | 0.77/71.74 |
| R1016 | 203 | 0.89/81.06 | 0.89/79.83 | 0.88/80.07 | 0.88/78.22 | 0.89/79.70 | **0.89/80.45** |

a GDT-TS and TM-score for the other two templates used in AIR for the 3-template refinement protocol.

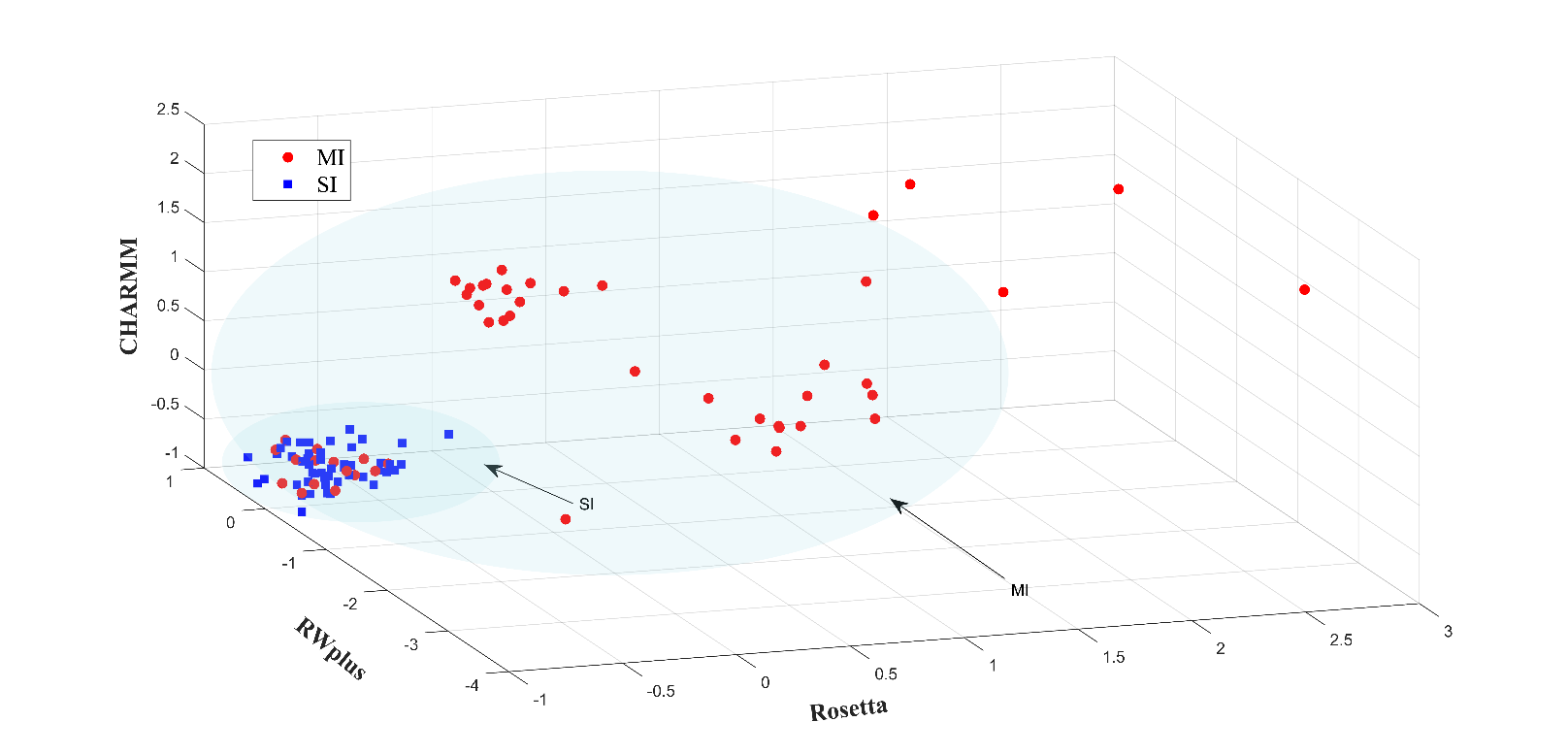
b Only the initial model is used as the input.

****

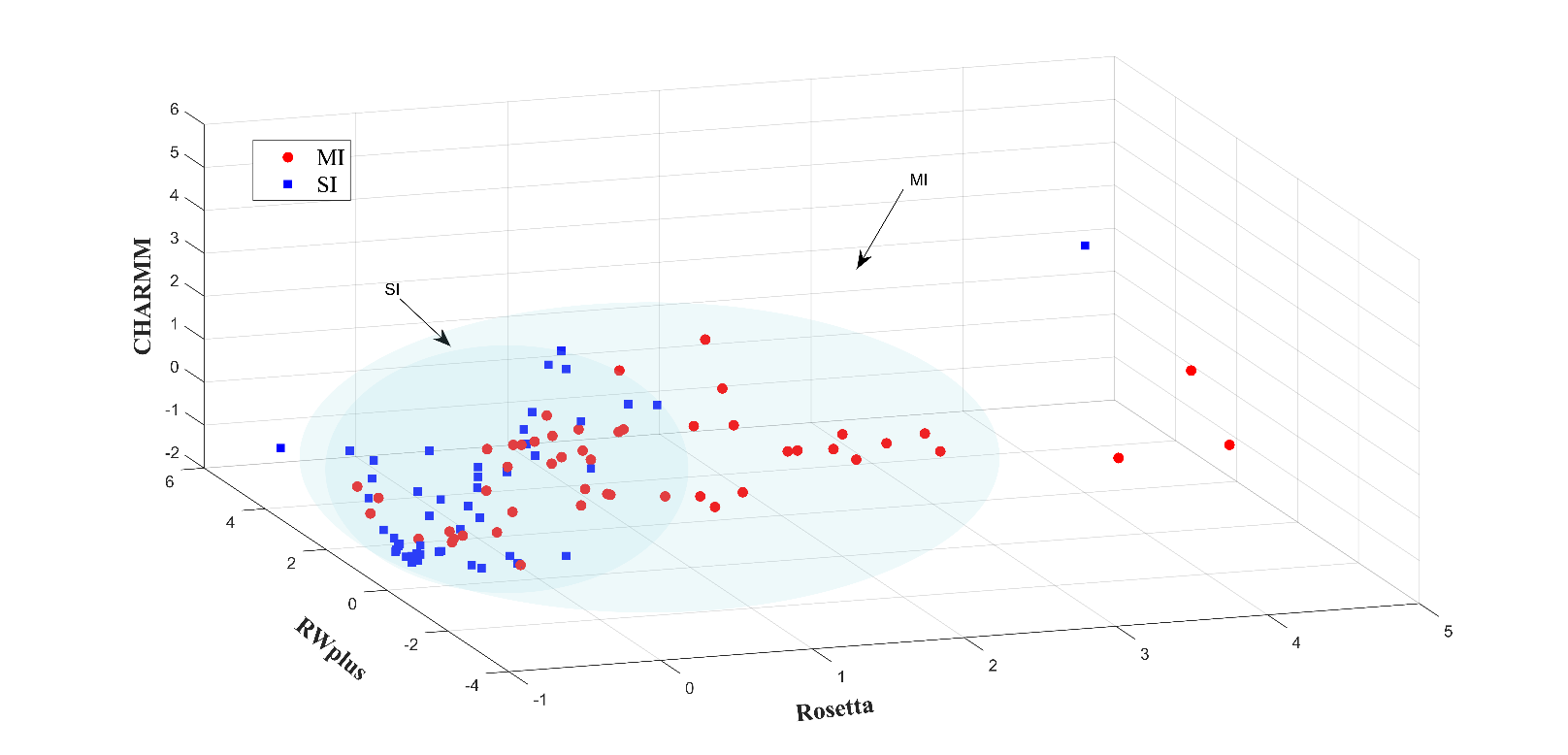
1. Energy space of R0968s1 between Single-template Input (SI) and Multi-template Input (MI) on the 1st iteration.

****

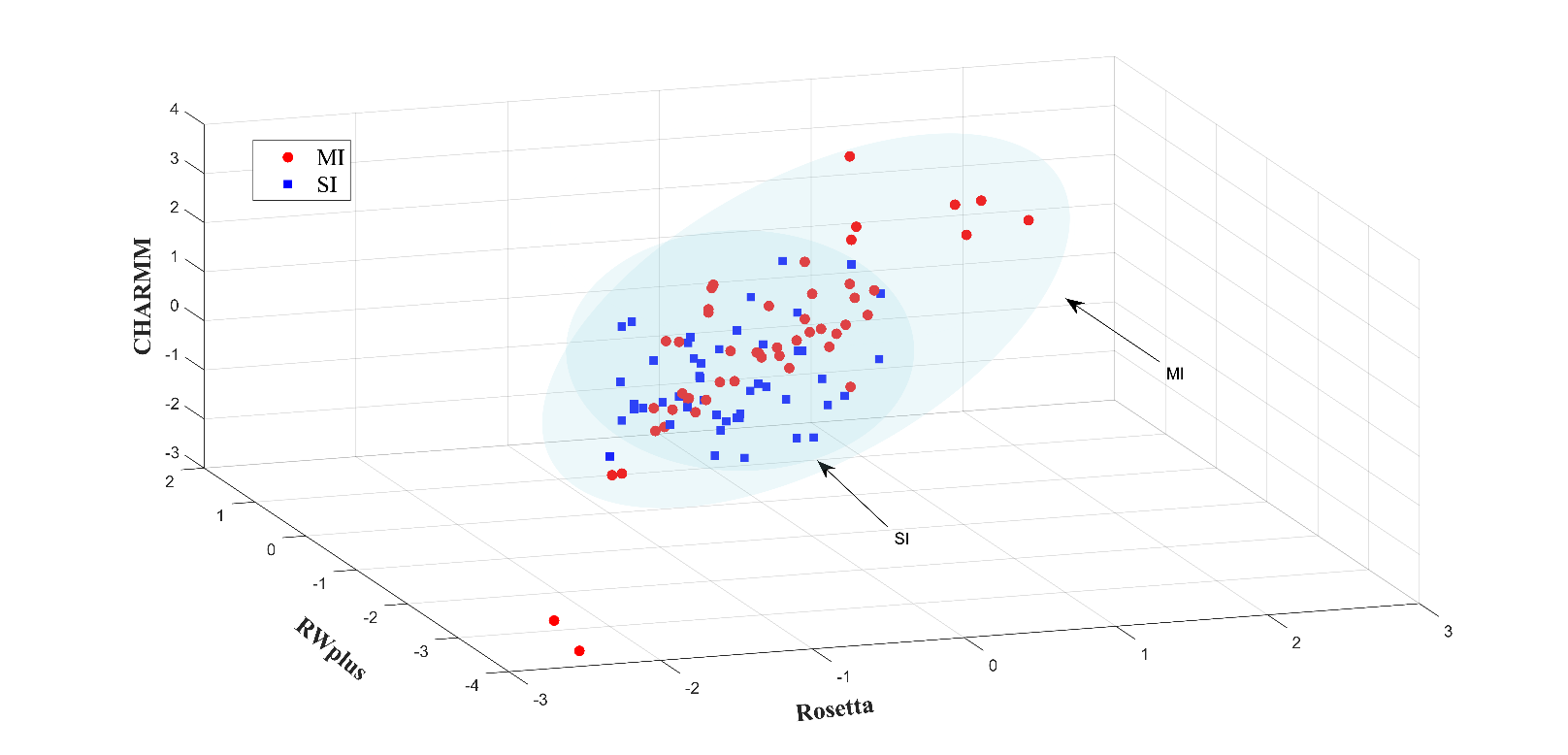
1. Energy space of R0968s1 between Single-template Input (SI) and Multi-template Input (MI) on the 3000th iterations.

****

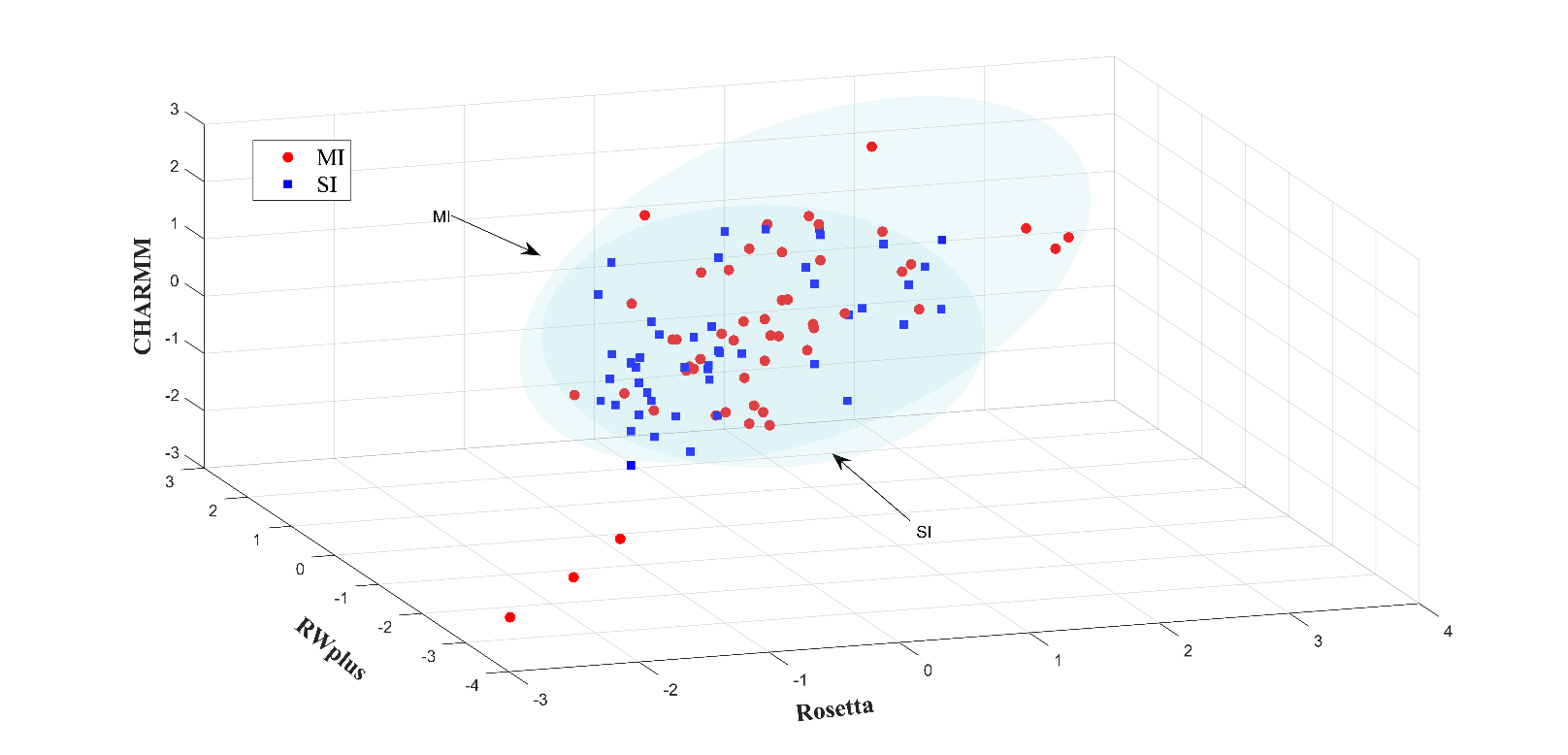
1. Energy space of R0968s2 between Single-template Input (SI) and Multi-template Input (MI) on the 1st iteration.

****

1. Energy space of R0968s2 between Single-template Input (SI) and Multi-template Input (MI) on the 3000th iterations.

****

1. Energy space of R1016 between Single-template Input (SI) and Multi-template Input (MI) on the 1st iteration.

****

1. Energy space of R1016 between Single-template Input (SI) and Multi-template Input (MI) on the 3000th iterations.

**Figure S7**.The comparison of particle distribution in energy space between Single-template Input (SI) and Multi-template Input (MI).