TESS Phase light curves of binaries and search for a close match in a pre-compiled database

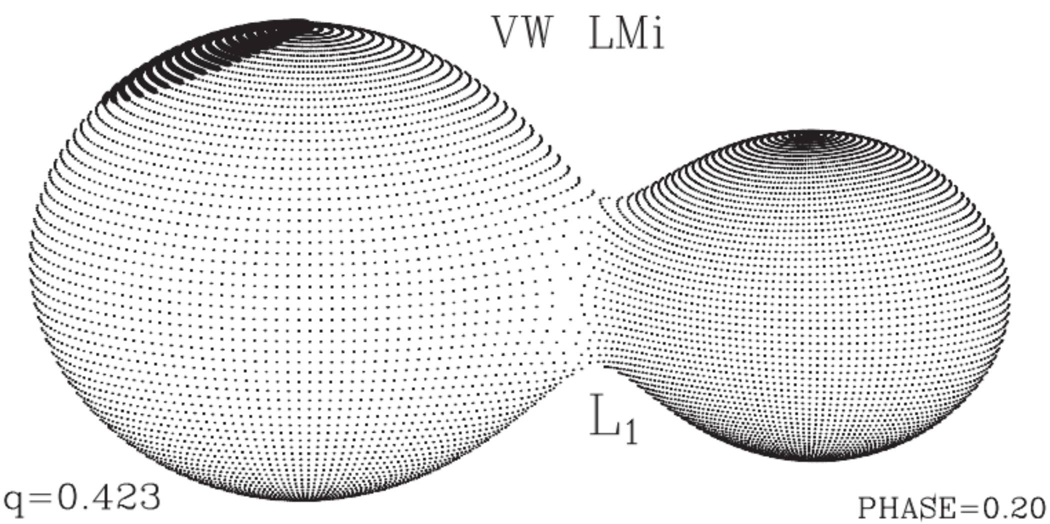
SKALNATÉ PLESO

Ľubomír Hambálek
with: Andrii Maliuk

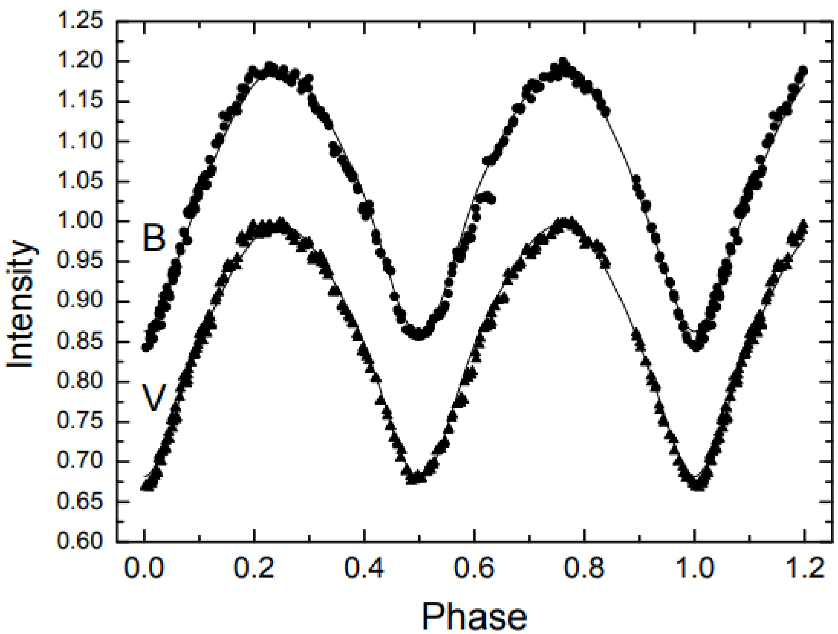
June 7, 2025

Contact binaries

- Binary stars with „small“ separation of components
- Shape dictated by surface equipotential Ω and mass ratio q
- Common evolution
- Circularized orbits with synchronized rotation
- Various fillings of Roche lobes, possible overflows (RLOs)
- If in contact (same Ω) – similar (~5%) surface temperature T



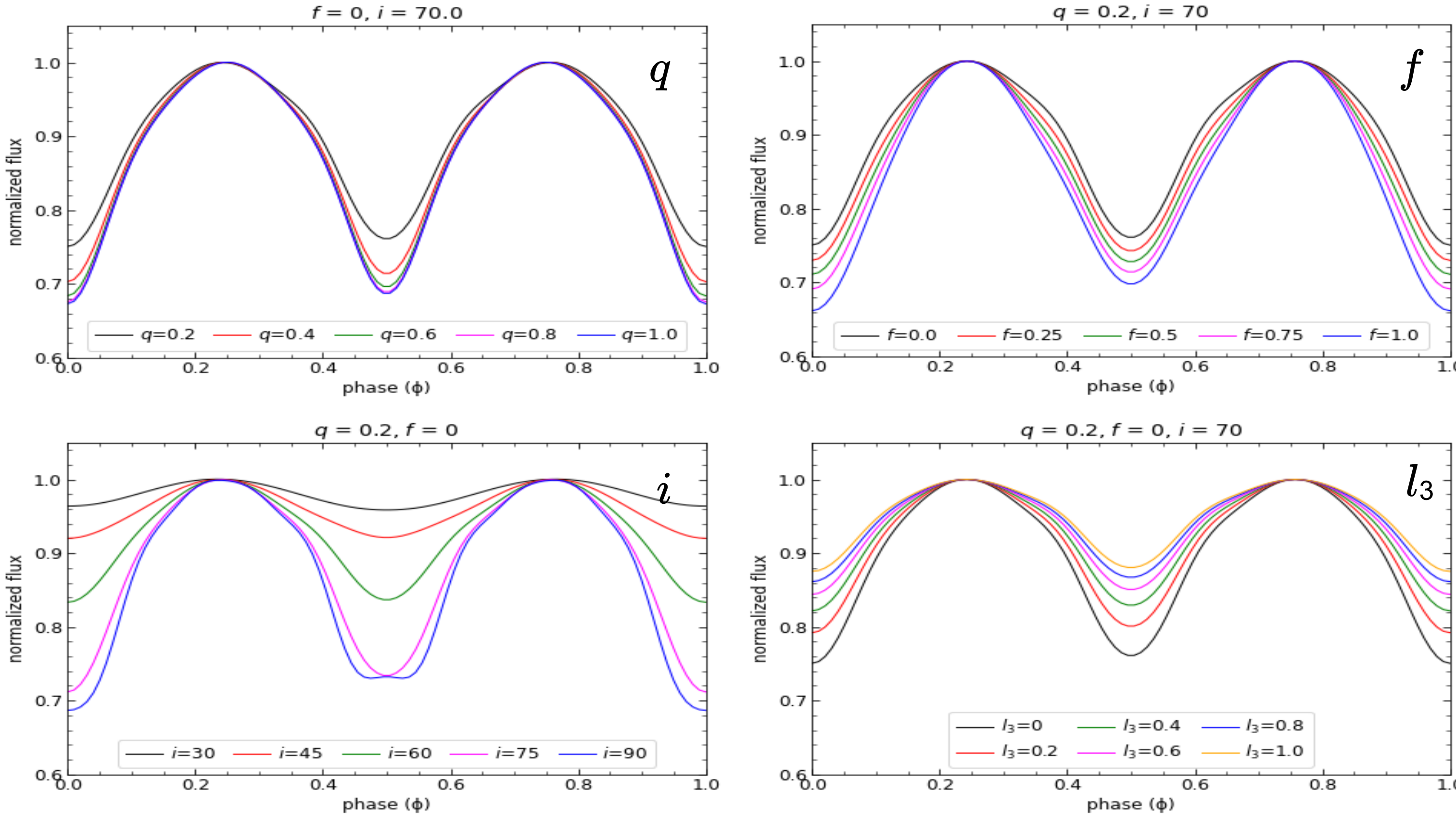
(Djurašević+, 2013)



(Pribulla+, 2008)

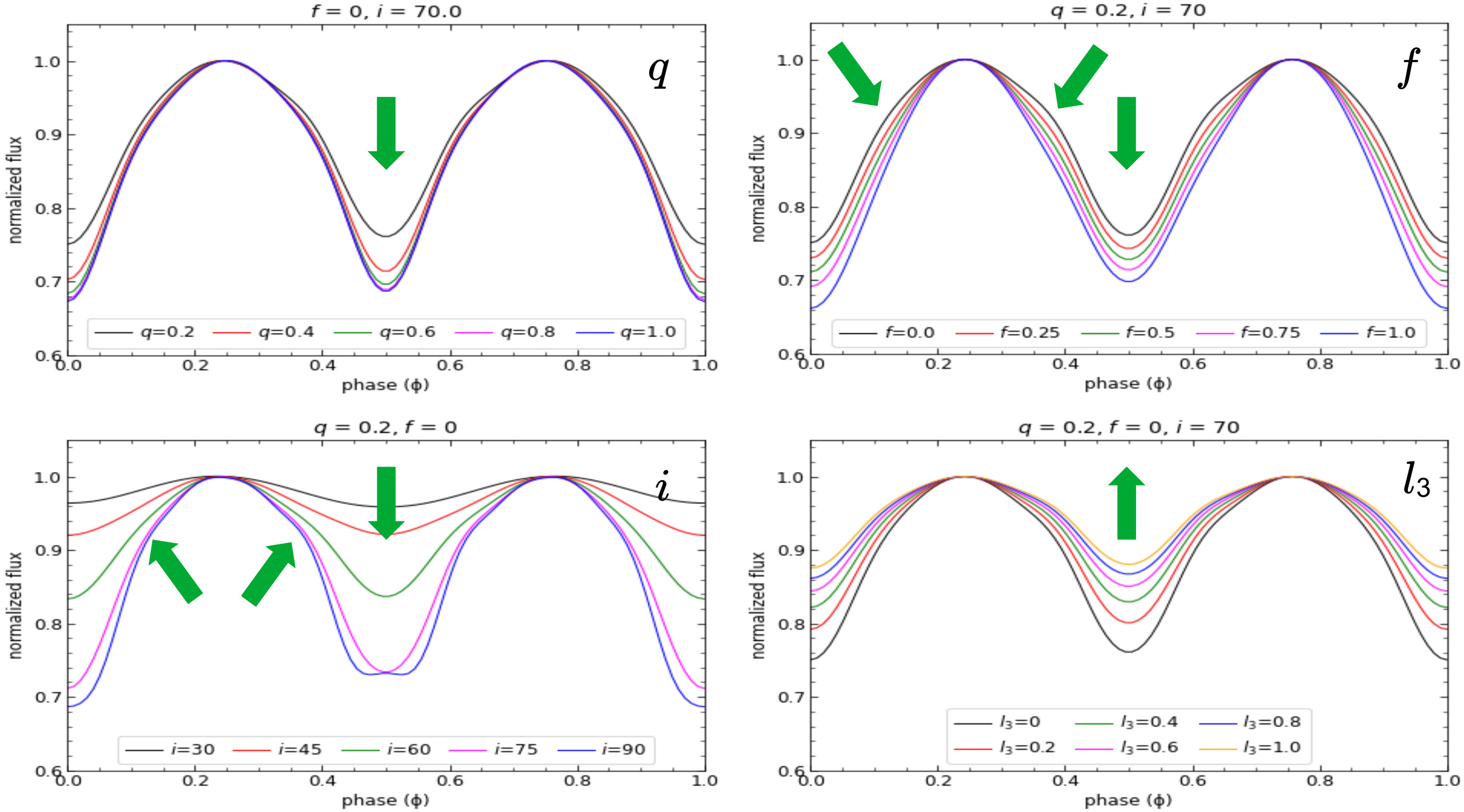
The „problem“ of photometric mass ratio

- Defined as $q_{\mathrm{ph}} = M_2/M_1$
- Correlates with orbital inclination $i(\Omega)$, fill-out $f(\Omega)$
- Close eclipsing binaries often part of multiple systems →light contamination (l_3 anticorrelates with i)



The „problem“ of photometric mass ratio

- Defined as $q_{\mathrm{ph}} = M_2/M_1$
- Correlates with orbital inclination $i(\Omega)$, fill-out $f(\Omega)$
- Close eclipsing binaries often part of multiple systems →light contamination (l_3 anticorrelates with i)



Previously done

- Physical model of stars with ROCHE code (Pribulla, 2012)
- Parameter space:
    - $q \in <0.05, 1.00>$; $\Delta q = 0.025$
    - $f \in <0.0, 1.0>$; $\Delta f = 0.25$
    - $i \in <30, 90>$ deg; $\Delta i = 1$ deg
    - $l_3 \in <0.0, 1.0>$; $\Delta l_3 = 0.2$

Previously done

- Physical model of stars with `ROCHE` code (Pribulla, 2012)
- Parameter space:
  - $q \in <0.05, 1.00>$; $\Delta q = 0.025$ 39
  - $f \in <0.0, 1.0>$; $\Delta f = 0.25$ 5
  - $i \in <30, 90>$ deg; $\Delta i = 1$ deg 61
  - $l_3 \in <0.0, 1.0>$; $\Delta l_3 = 0.2$ 6

71 370

Previously done

- Physical model of stars with `ROCHE` code (Pribulla, 2012)
- Parameter space:
  - $q \in <\ 0.05,\ 1.00 >;\quad \Delta q = 0.025$
  - $f \in <\ 0.0,\ 1.0 >;\quad \Delta f = 0.25$
  - $i \in <\ 30,\ 90 >$ deg; $\Delta i = 1$ deg
  - $l_3 \in <\ 0.0,\ 1.0 >;\quad \Delta l_3 = 0.2$

71 370

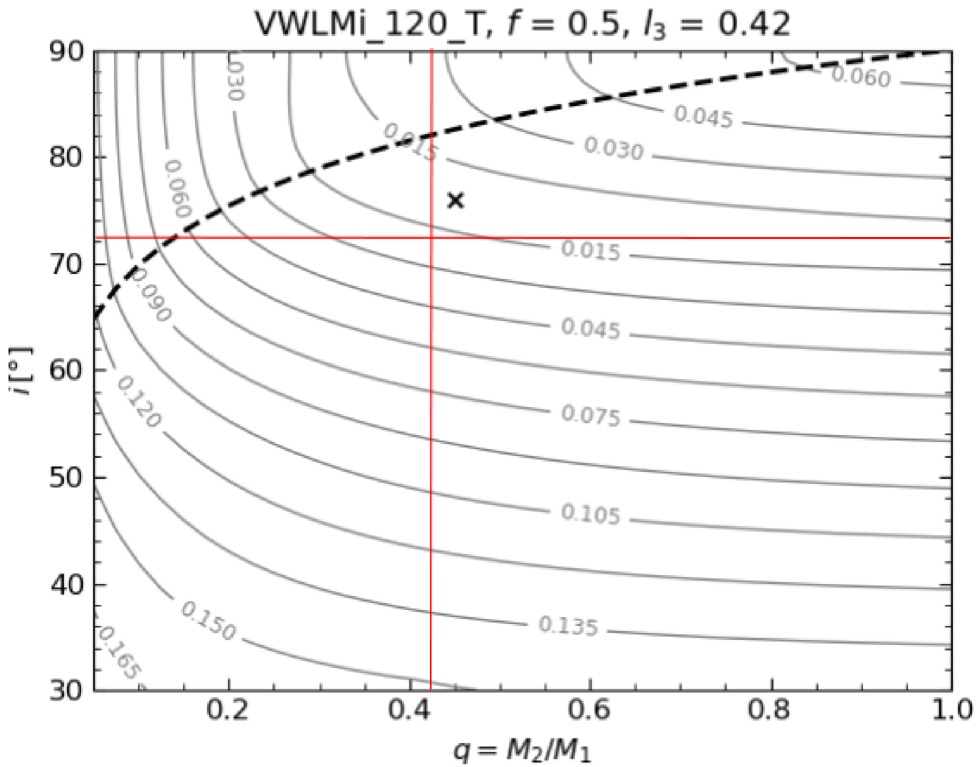
- Represent the LC with trigonometric polynomial:

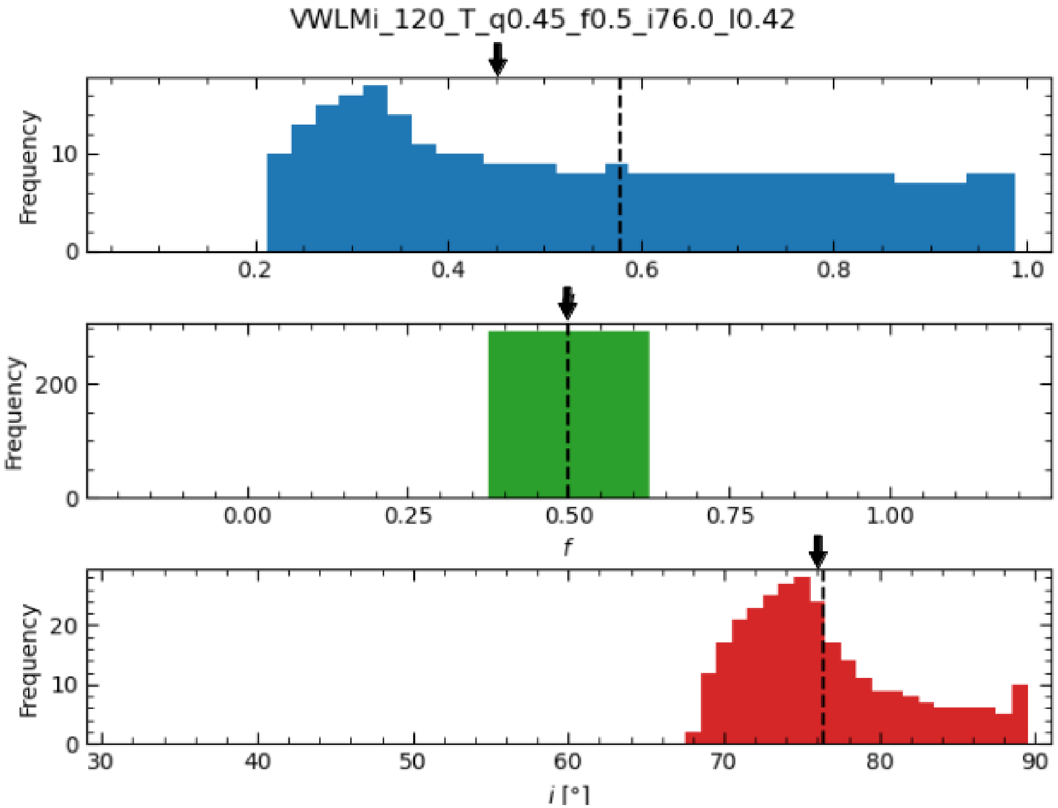
$$I(\varphi) = a_0 + \sum_{k=1}^{n} \cos(a_k) + \sum_{k=1}^{n} \sin(b_k) \tag{1}$$

- Consider only symmetrical LCs around $\varphi = 0.5$ ($\Rightarrow b_k = 0$)
- Sufficient up to $n = 10$ (Hambálek & Pribulla, 2013)

Grid search with real *TESS* data

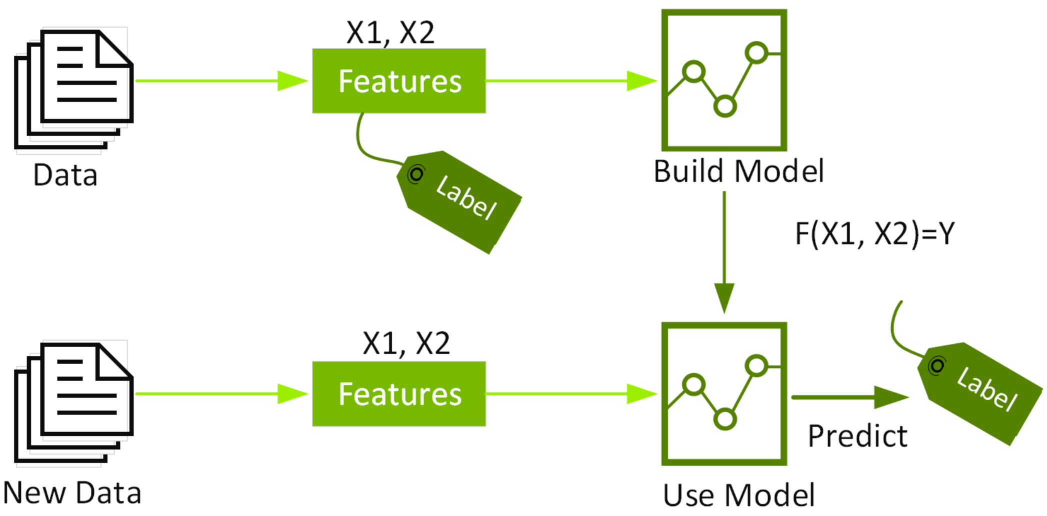
- Smoothed LC: Lest-square fit to **(1)** $\rightarrow a_k$
- Finding best (×,↓) LCs minimizing D
- Comparison with literature values

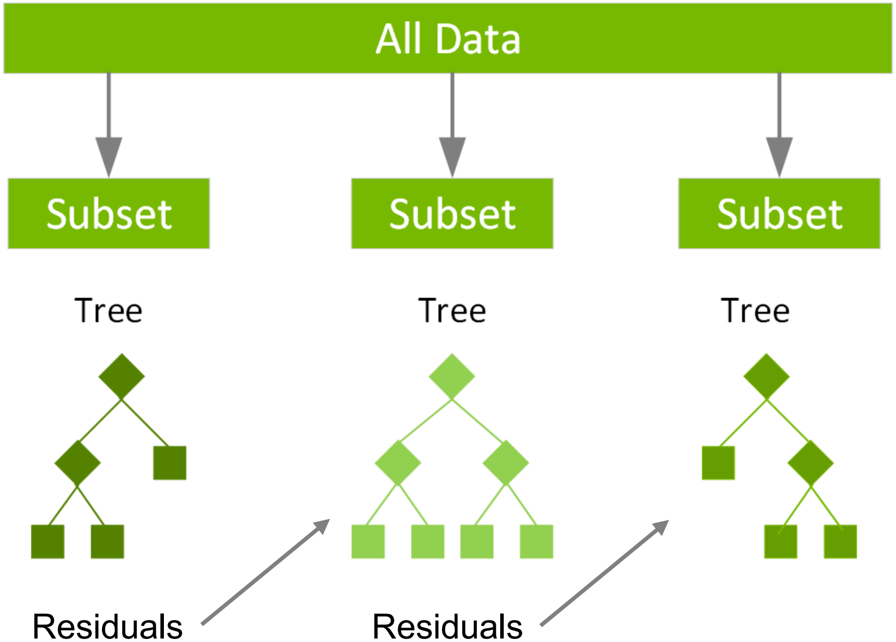




Can we try better?

- Simple model using `scikit-learn`, `XGBoost`, and `tensorflow`
- `XGBoost` for high-performance, sequence of decision trees
- Model predicts by evaluating a tree of if-then-else true/false questions (trees)
- Each tree corrects errors of previous one - possible non-linear relationships between input and target variables





Training

- Training data: 70% random of full set of 71 370 LCs as a_k
- Test data: the rest 30%
- By trial/error ⇒ `max_depth = 7`
- $i \in <30, 90>$ deg → sin(i) $\in <0.5, 1>$ since $q, f, l_3 \in <0.0, 1.0>$

	a0	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	q	f	i	l3
0	0.988591	0.003461	-0.010768	-0.001055	0.000191	-0.000014	0.000013	0.000010	0.000003	0.000002	0.000003	0.05	0.0	30.0	0.0
1	0.987924	0.003387	-0.011394	-0.001163	0.000223	-0.000019	0.000026	0.000007	-0.000001	0.000006	0.000004	0.05	0.0	31.0	0.0
2	0.987232	0.003301	-0.012032	-0.001269	0.000262	0.000002	0.000039	0.000000	0.000001	-0.000008	0.000007	0.05	0.0	32.0	0.0
3	0.986539	0.003180	-0.012673	-0.001383	0.000315	0.000017	0.000057	0.000006	0.000002	0.000002	0.000000	0.05	0.0	33.0	0.0
4	0.985849	0.003076	-0.013328	-0.001490	0.000373	0.000032	0.000057	0.000013	0.000001	0.000012	-0.000007	0.05	0.0	34.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
71369	0.856721	-0.003366	-0.155882	-0.001025	-0.019322	-0.000706	-0.010762	-0.000431	-0.006175	-0.000255	-0.003522	1.00	1.0	90.0	1.0

71370 rows × 15 columns

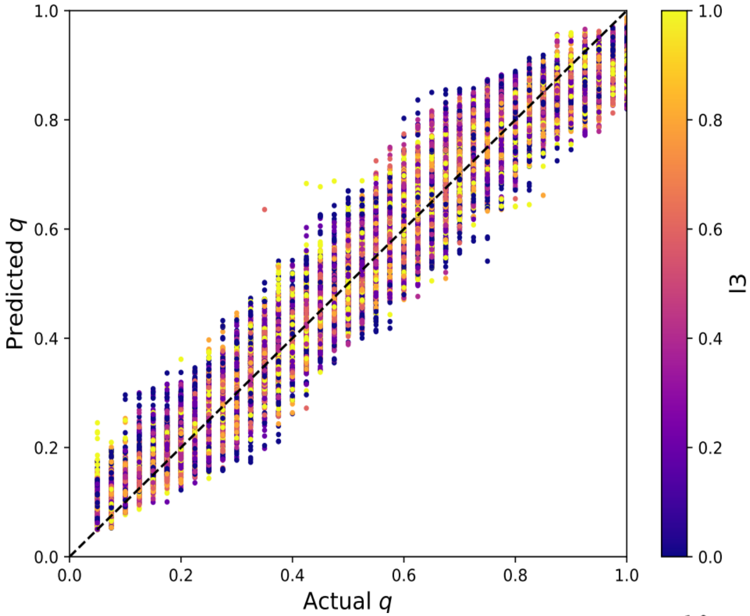
Training

- Training data: 70% random of full set of 71 370 LCs as a_k
- Test data: the rest 30%
- By trial/error ⇒ `max_depth = 7`
- $i \in <30, 90>$ deg → sin(i) $\in <0.5, 1>$ since $q, f, l_3 \in <0.0, 1.0>$

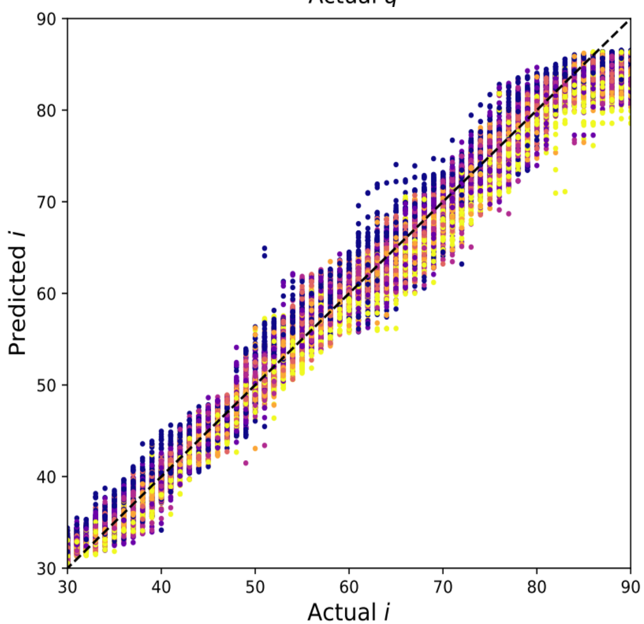
	a0	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	q	f	i	l3
0	0.988591	0.003461	-0.010768	-0.001055	0.000191	-0.000014	0.000013	0.000010	0.000003	0.000002	0.000003	0.05	0.0	30.0	0.0
1	0.987924	0.003387	-0.011394	-0.001163	0.000223	-0.000019	0.000026	0.000007	-0.000001	0.000006	0.000004	0.05	0.0	31.0	0.0
2	0.987232	0.003301	-0.012032	-0.001269	0.000262	0.000002	0.000039	0.000000	0.000001	-0.000008	0.000007	0.05	0.0	32.0	0.0
3	0.986539	0.003180	-0.012673	-0.001383	0.000315	0.000017	0.000057	0.000006	0.000002	0.000002	0.000000	0.05	0.0	33.0	0.0
4	0.985849	0.003076	-0.013328	-0.001490	0.000373	0.000032	0.000057	0.000013	0.000001	0.000012	-0.000007	0.05	0.0	34.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
71369	0.856721	-0.003366	-0.155882	-0.001025	-0.019322	-0.000706	-0.010762	-0.000431	-0.006175	-0.000255	-0.003522	1.00	1.0	90.0	1.0

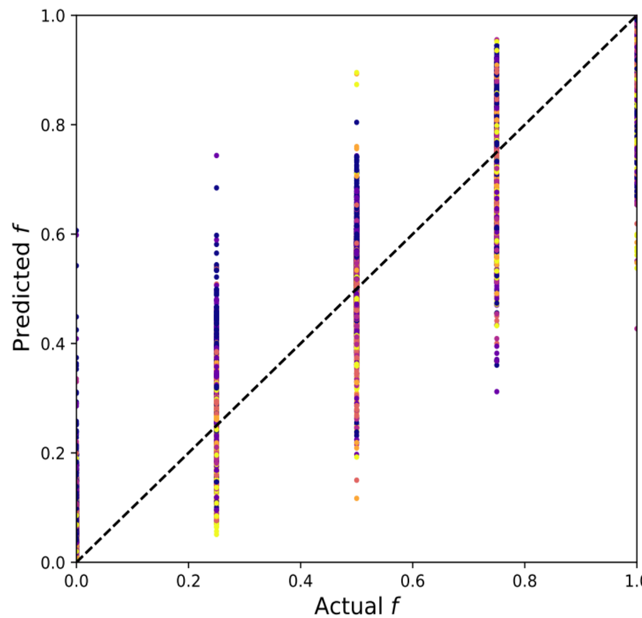
71370 rows × 15 columns

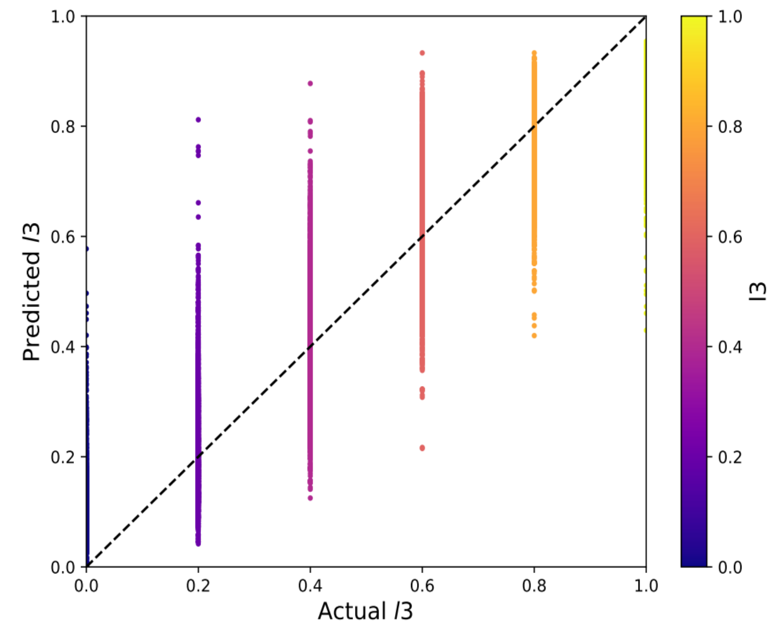
Model performance – random forest



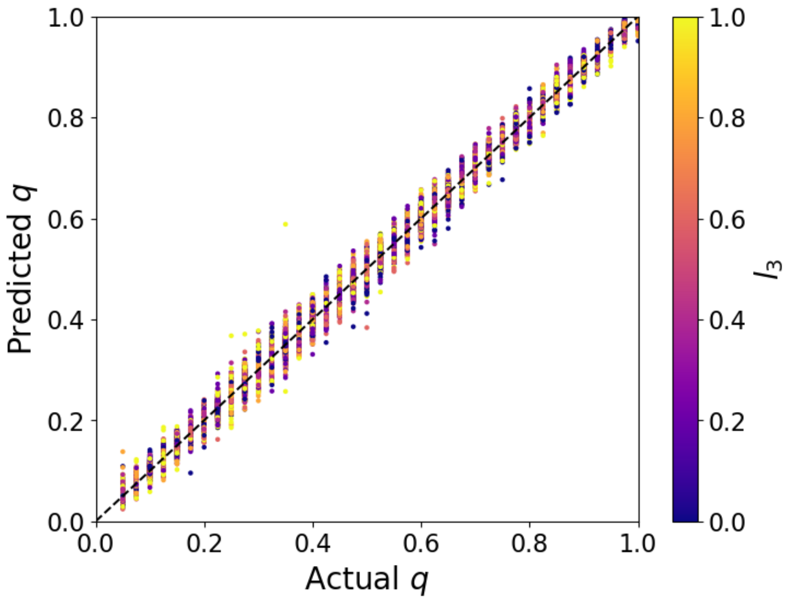
RMS	Training set	Test set
q	0.0546	0.0563
f	0.0578	0.0639
$\sin(i)$	0.0315	0.0324
l_3	0.0934	0.1051



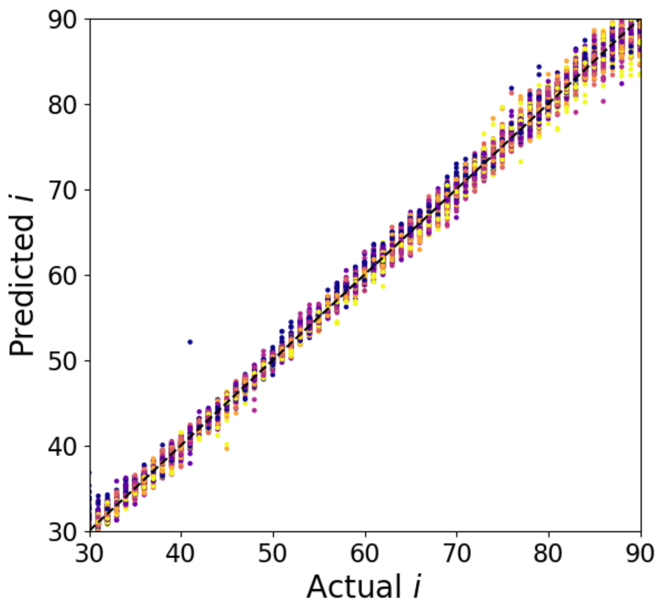


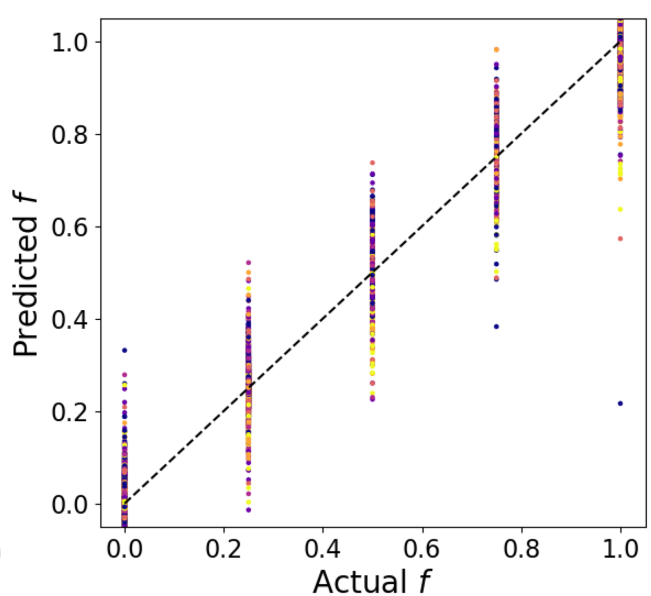


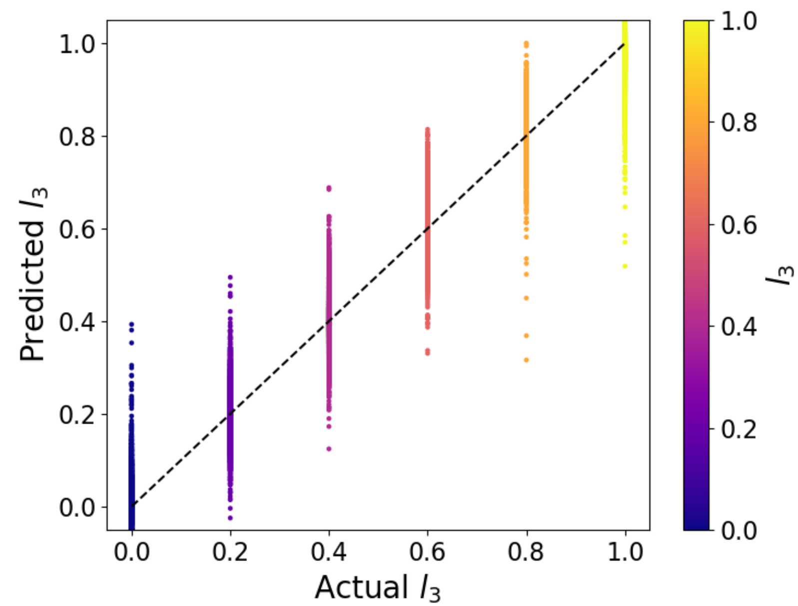
Model performance – XGBoost



RMS	Training set	Test set
q	0.0108	0.0151
f	0.0245	0.0408
$\sin(i)$	0.0037	0.0057
l_3	0.0388	0.0562

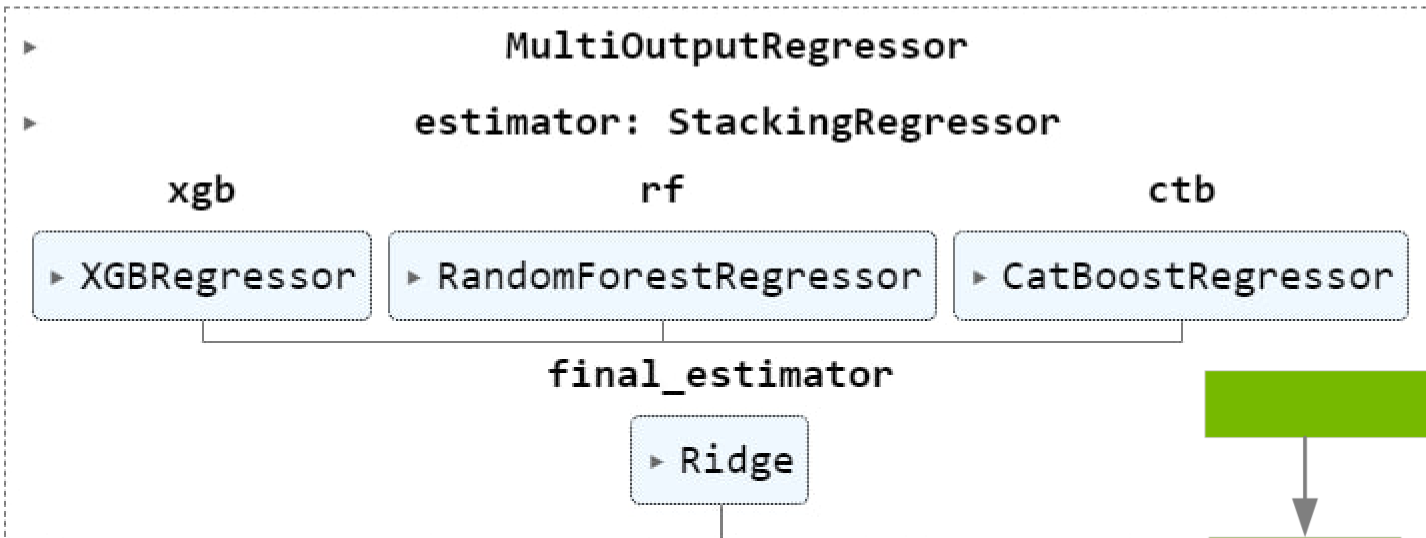


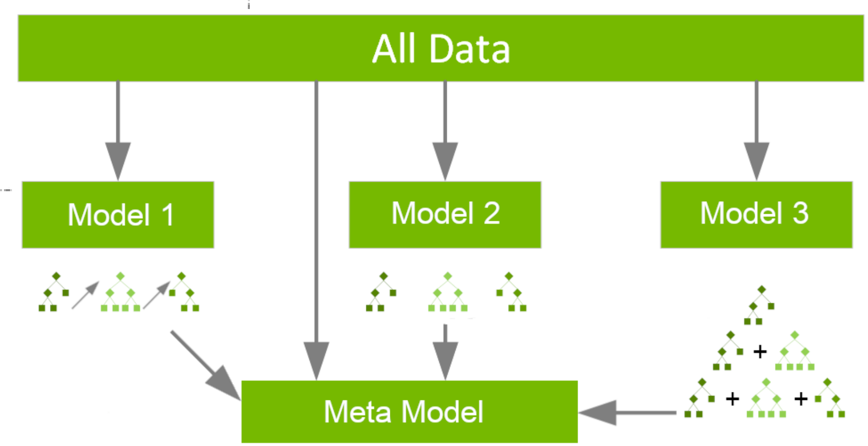




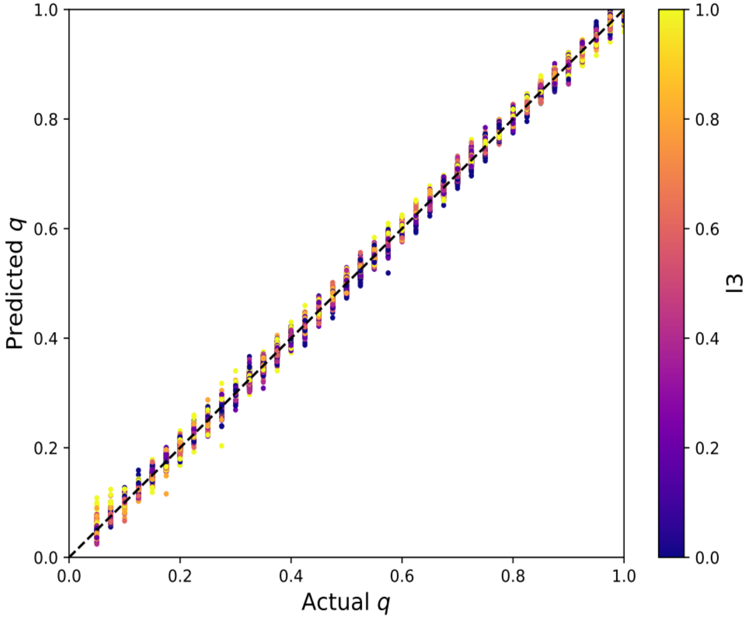
Next try - stacked models

- Using `MultiOutputRegressor`
- Combine previous Random forest with XGBoost add CatBoost
- Ridge regression used best for multicollinear data or if number of predictor variables > number of observations.

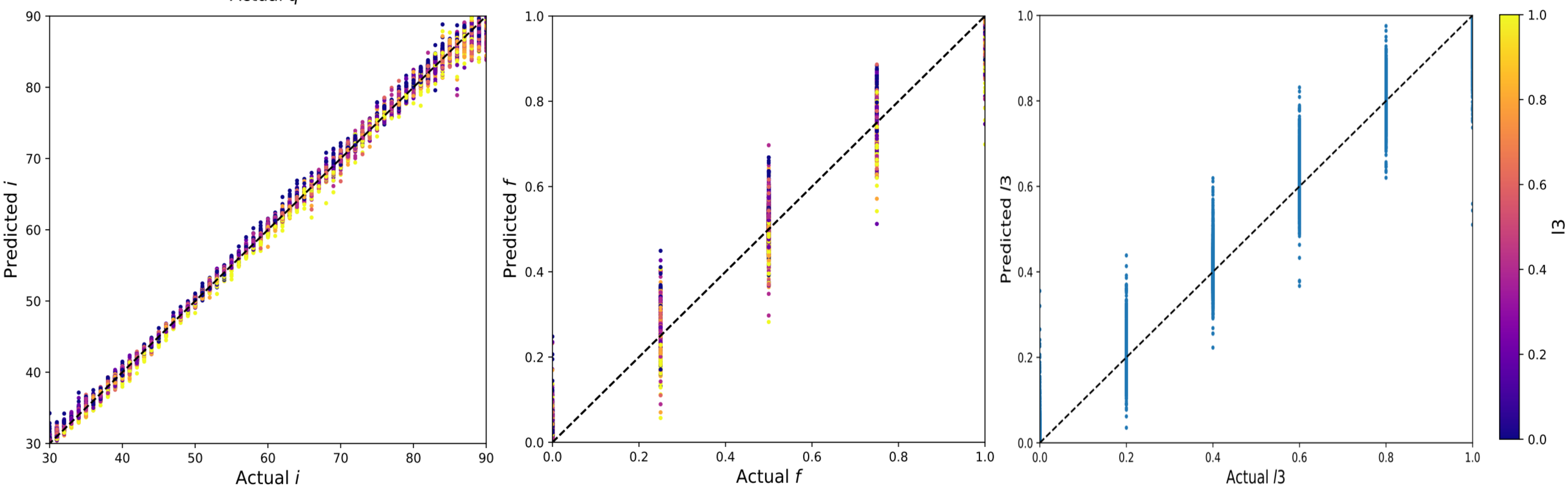




Model performance – Stacked Regressor



RMS	Training set	Test set
q	0.0068	0.0082
f	0.0210	0.0273
$\sin(i)$	0.0061	0.0081
l_3	0.0344	0.0431



Comparison with real *TESS* data

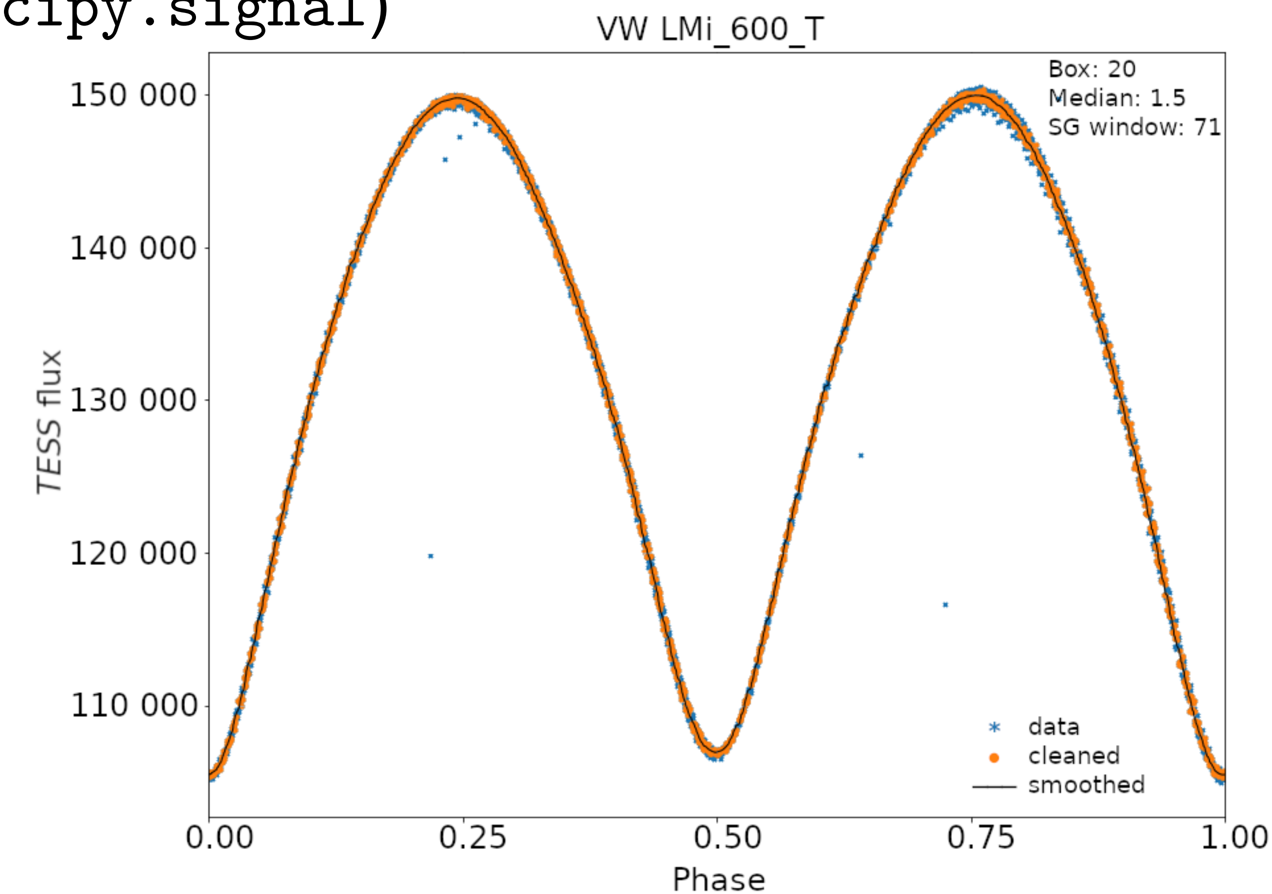
- Selected 14 stars with full range of $q_{\rm sp} \in < 0.066, 0.984 >$

star	$q_{\rm L}$	$f_{\rm L}$	$i_{\rm L}$ [deg]	$l_{3,\rm L}$	type
AG Vir	0.341 [a]	0.17 [b]	84 [b]	0.05 [a]	EW A
AW UMa	0.108 [c]	0.30 [c]	78 [d]	0.00 [c]	EW
DU Boo	0.206 [b]	0.56 [b]	81 [b]	0.00 [b]	EW A
EL Boo	0.248 [d]	0.00 [e]	74 [e]	1.00 [f]	EW
EQ Tau	0.442 [g]	0.09 [e]	82 [e]	0.00 [g]	EW A
FI Boo	0.372 [h]	0.50 [i]	38 [i]	0.30 [h]	EW W
FT UMa	0.984 [f]	N/A	60(3) [j]	1.01 [f]	EB
SW Lac	0.776 [k]	?	?	<0.05 [k]	EW W
SX Crv	0.066 [g]	?	65(5) [g]	0.00 [g]	EW A
V1191 Cyg	0.107 [l]	0.30 [m]	83(2) [m]	0.00 [l]	EW W
V523 Cas	0.516 [n]	0.00 [o]	84(1) [o]	0.00 [n]	EW W
V753 Mon	0.970 [p]	N/A	75 [q]	0.00 [p]	EB
VW LMi	0.423 [a]	0.47 [r]	79 [s]	0.42 [a]	EW W
W UMa	0.484 [t]	0.10 [u]	86 [u]	0.00 [t]	EW

Source: [a]Pribulla et al. (2006), [b]Pribulla et al. (2011), [c]Pribulla & Rucinski (2008), [d]Pribulla & Rucinski (2006), [e]Deb & Singh (2011), [f]Pribulla et al. (2009), [g]Rucinski et al. (2001), [h]Lu et al. (2001), [i]Christopoulou & Papageorgiou (2013), [j]Yuan (2011), [k]Rucinski et al. (2005), [l]Rucinski et al. (2008), [m]Ekmekçi et al. (2012), [n]Rucinski et al. (2003), [o]Mohammadi et al. (2016), [p]Rucinski et al. (2000), [q]Qian et al. (2013), [r]Sánchez-Bajo et al. (2007), [s]Pribulla et al. (2008), [t]Pribulla et al. (2007), [u]Linnell (1991).

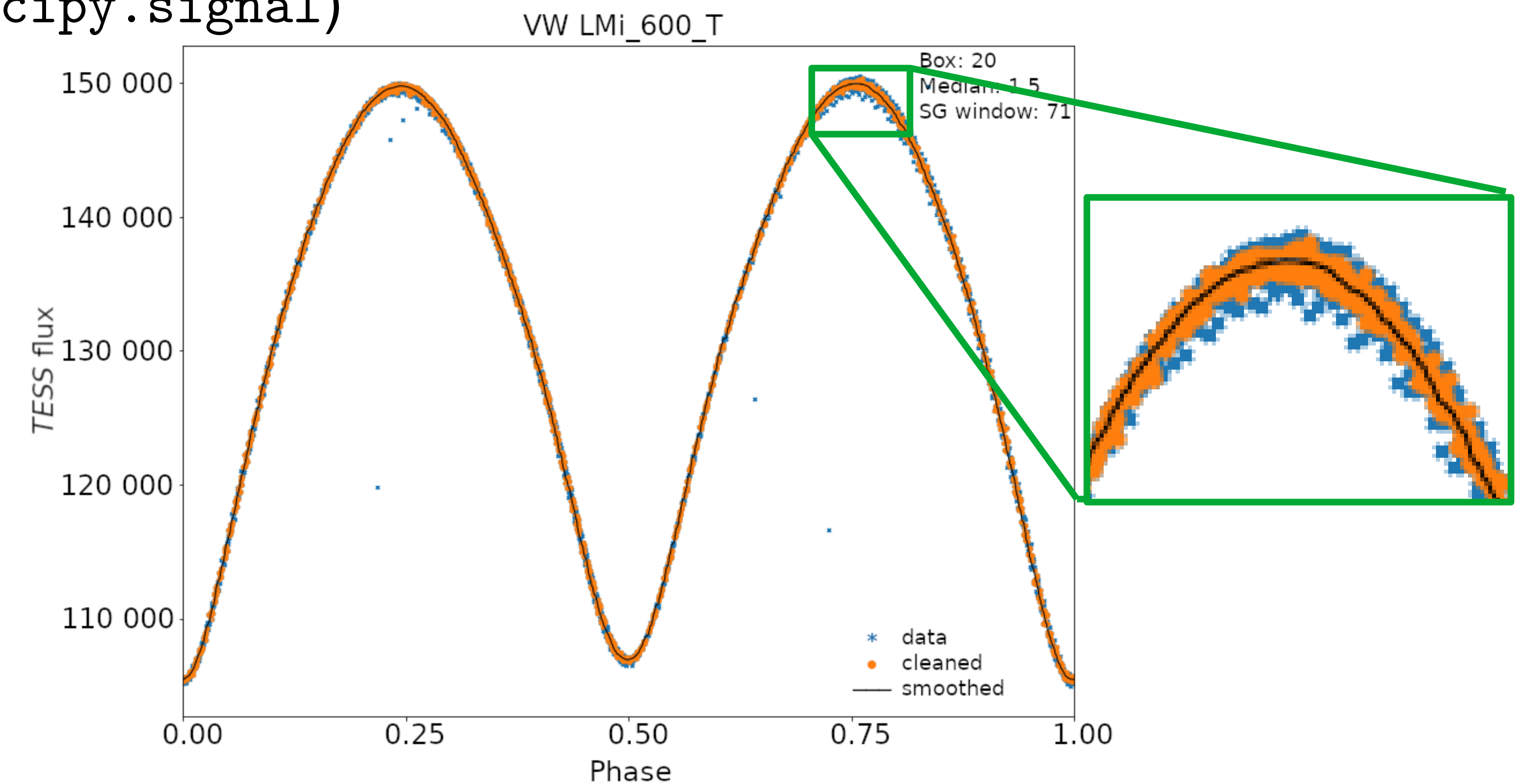
Comparison with real *TESS* data

- Selected 14 stars with full range of $q_{\mathrm{sp}} \in <0.066, 0.984>$
- LCs obtained by `lightkurve` (SPOC flux) → phase LC by period
- Running box outlier removal, smoothed by Savitzky-Golay filter (from `scipy.signal`)



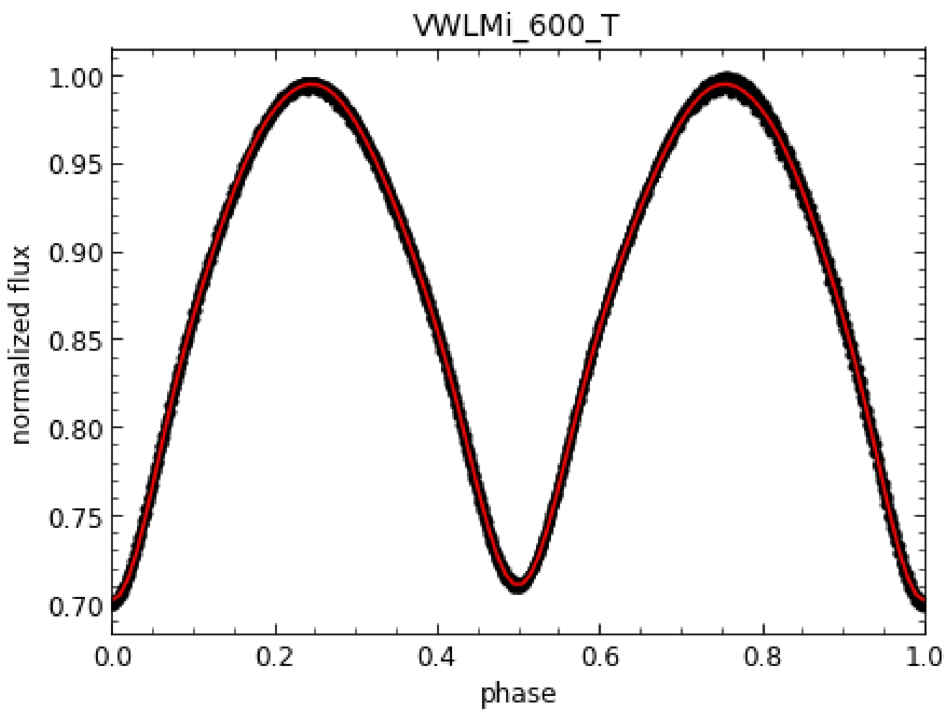
Comparison with real *TESS* data

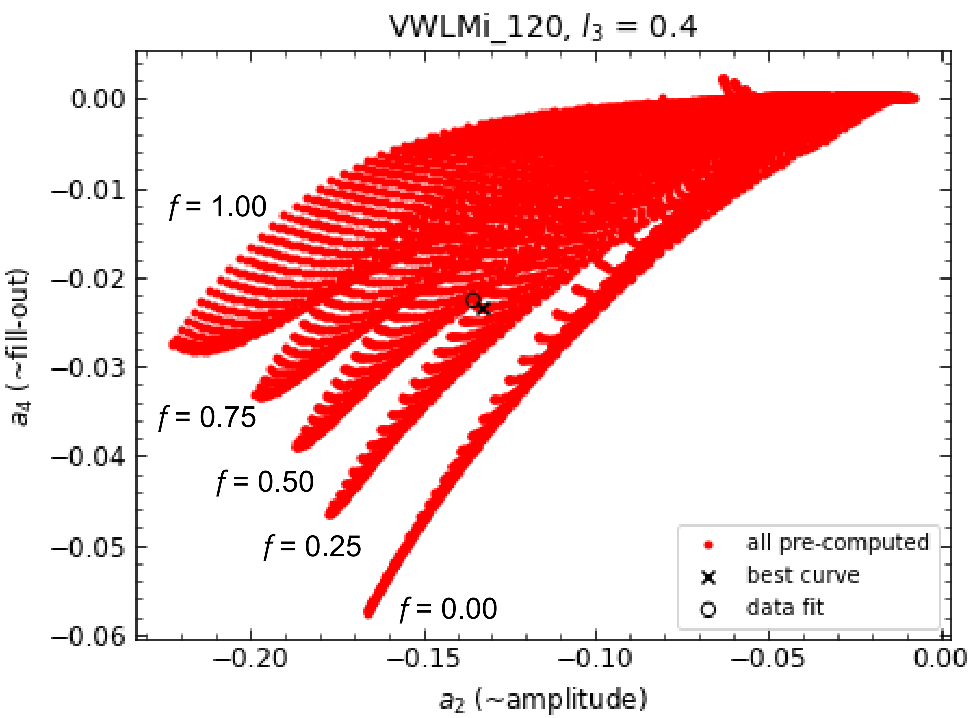
- Selected 14 stars with full range of $q_{\mathrm{sp}} \in < 0.066, 0.984 >$
- LCs obtained by `lightkurve` (SPOC flux) → phase LC by period
- Running box outlier removal, smoothed by Savitzky-Golay filter (from `scipy.signal`)



Comparison with real *TESS* data

- Smoothed LC: Lest-square fit to **(1)** $\rightarrow a_k$
- Finding best (×,↓) LCs minimizing D



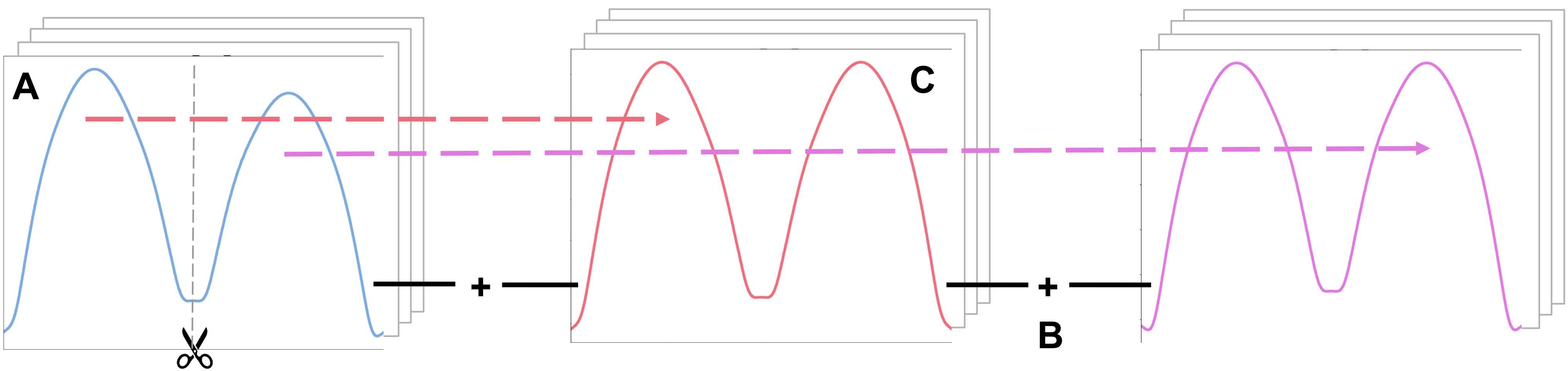


Subsets of real data

- **U** = Previous predictions from matching with pre-computed library (code `UNIQUE`)
- **M** = New predictions from `XGBoost` (**M**achine learning)

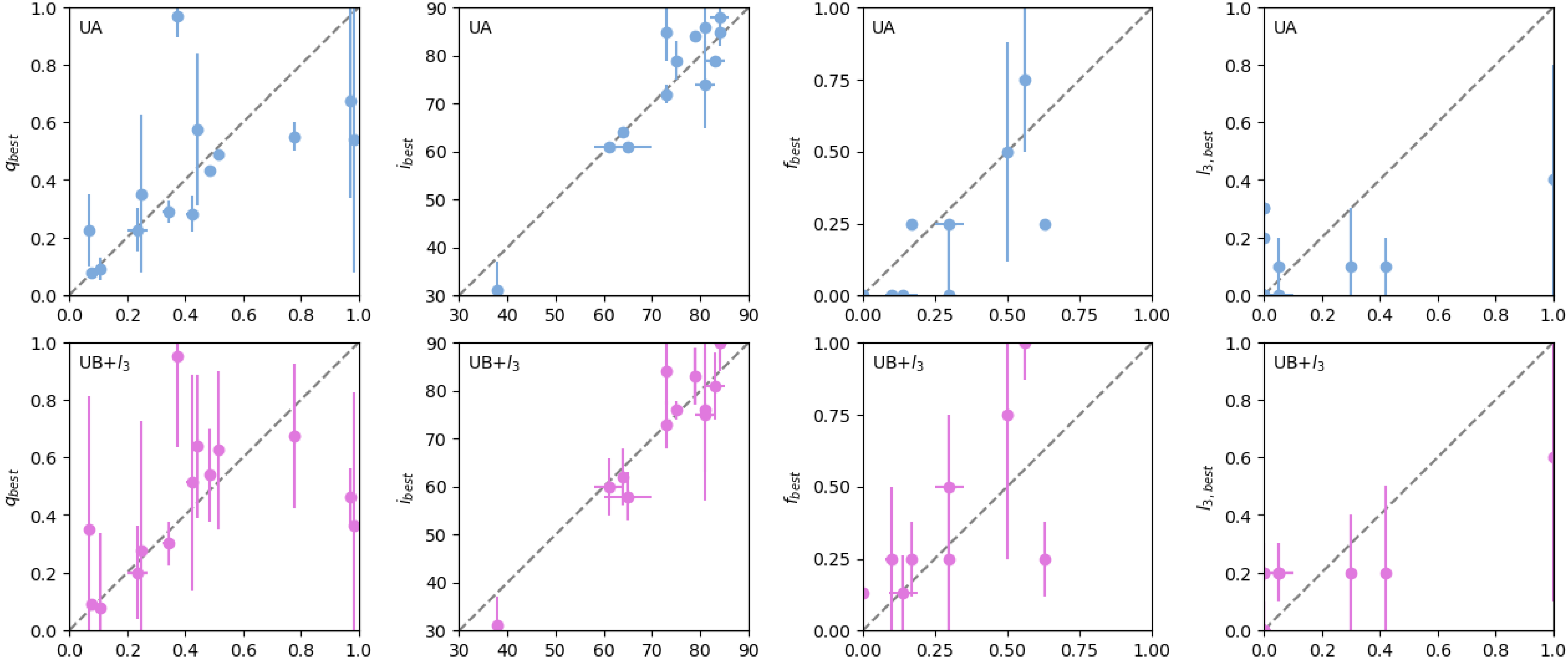
Subsets of real data

- **U** = Previous predictions from matching with pre-computed library (code `UNIQUE`)
- **M** = New predictions from `XGBoost` (**M**achine learning)
- **A** = Initial set from *TESS*
- **B** = A + artificially symmetric LCs by mirroring at φ = 0.5
- **C** = A + subset of B with only LCs with higher peak mirrored („no spot“)

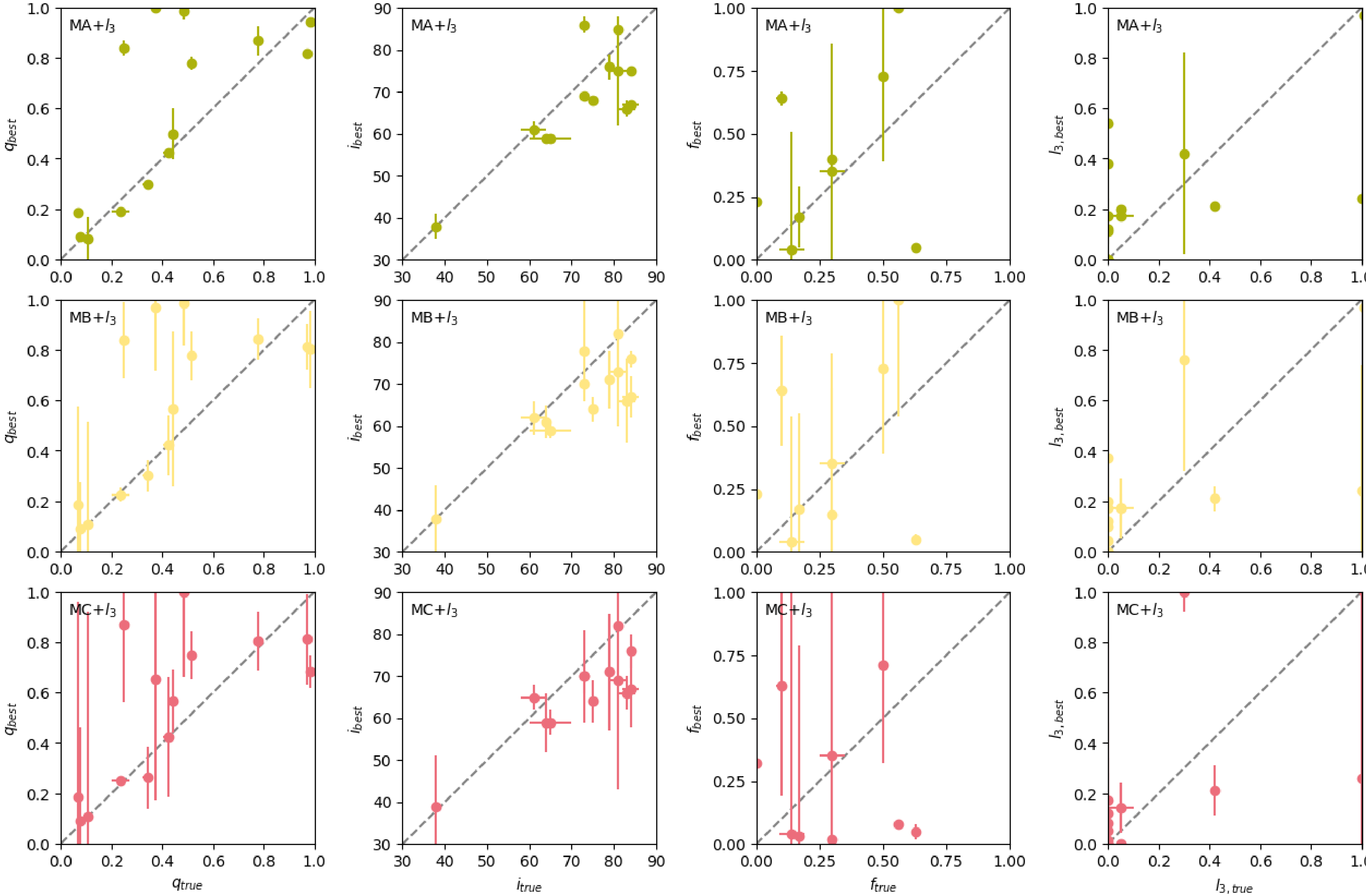


Results - correspondence

- predicted values (y-axis) vs actual values (x-axis)

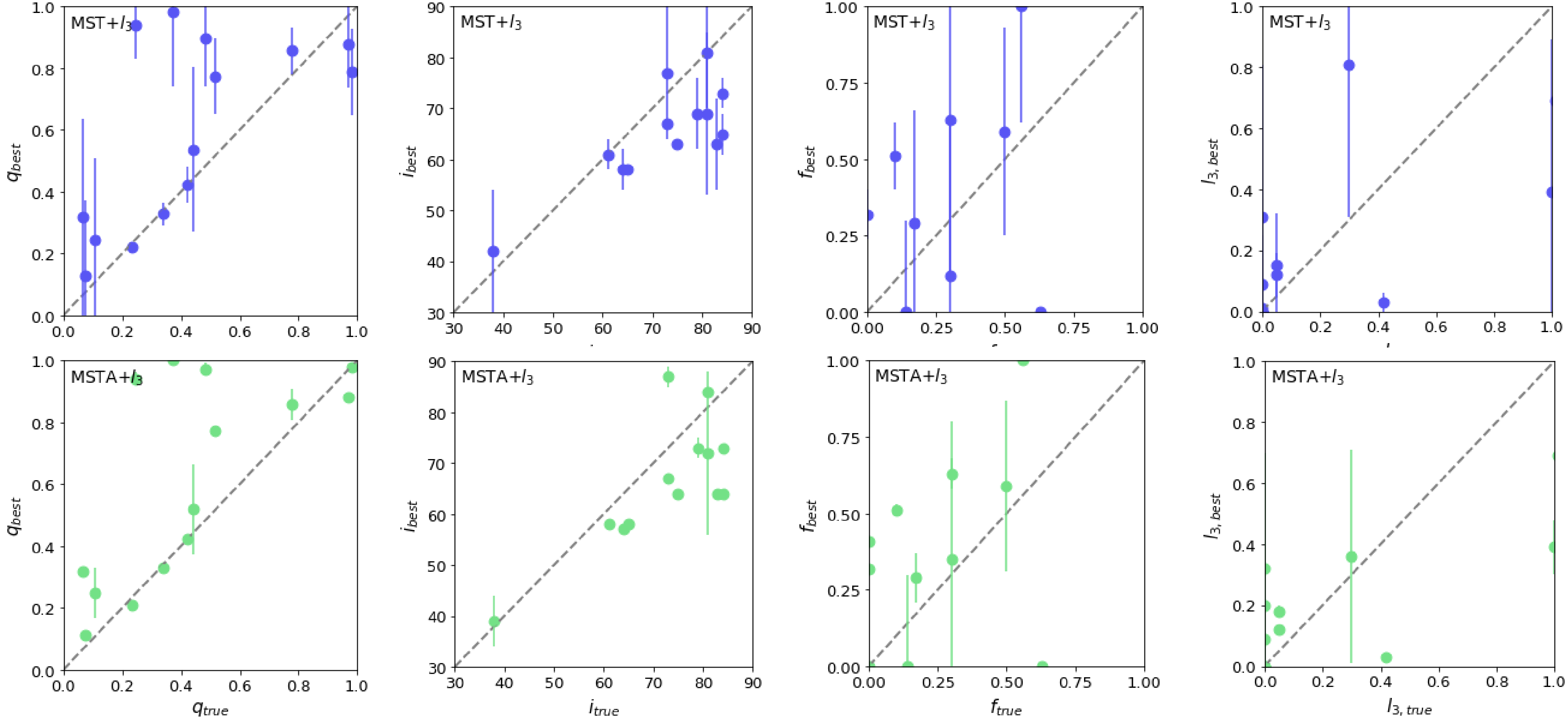


Results - correspondence



Results - correspondence

- **ST** = stacked models
- MST – stacked models, **C** dataset
- MSTA – stacked models, **A** dataset



Results – weighted correlation

- **U** = matching with pre-computed library
- **M** = New predictions from `XGBoost`
  - **A** = Initial set from *TESS*
  - **B** = **A** + artificially symmetric LCs by mirroring at φ = 0.5
  - **C** = **A** + only LCs with higher peak mirrored („no spot“)
- MST – stacked models, **C** dataset
- MSTA – stacked models, **A** dataset

model	q	i	f	l_3
UA	0.978	0.955	0.708	0.083
UB+l_3	0.857	0.973	0.573	0.547
MA+l_3	0.787	0.952	-0.548	0.096
MB+l_3	0.839	0.896	-0.618	0.759
MC+l_3	0.897	0.853	-0.702	0.999
MST+l_3	0.806	0.998	-0.846	0.458
MSTA+l_3	0.748	0.864	0.342	-0.089

Results – predictions of mass ratio

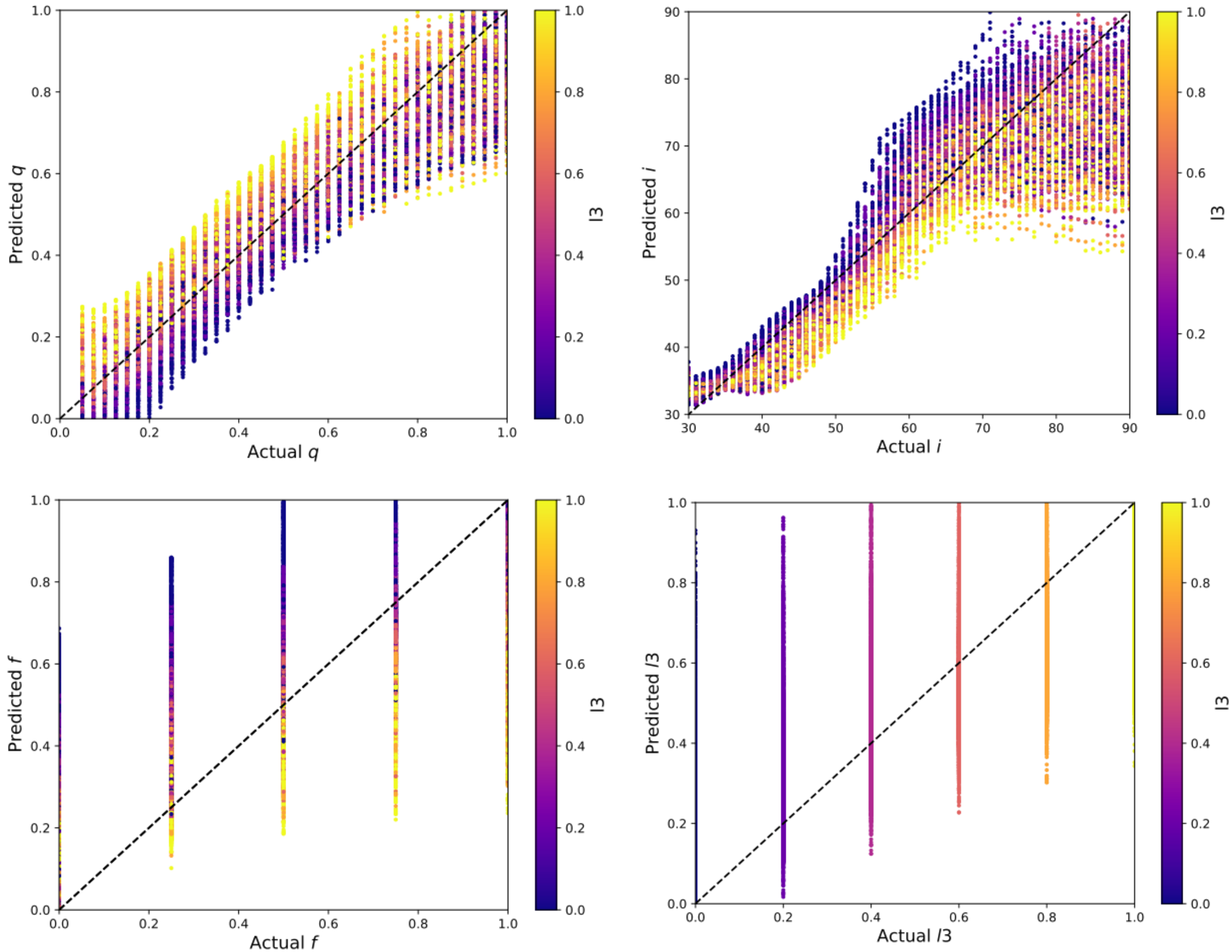
- best predictions of q in different models with [(max - min)/2]

Object	q_{true}	UA	UB+l_3	MA+l_3	MB+l_3	MC+l_3	MST+l_3	MSTA+l_3
AG Vir	0.341(21)	0.288[38]	**0.300[75]**	0.296[8]	**0.300[62]**	0.261[123]	**0.328[36]**	**0.328[5]**
AW UMa	0.075(5)	**0.075[0]**	0.088[13]	0.089[10]	0.089[187]	0.089[374]	0.126[245]	0.111[4]
DU Boo	0.234(35)	**0.225[75]**	0.200[163]	0.190[0]	**0.226[30]**	**0.250[0]**	**0.219[19]**	0.209[0]
EL Boo	0.248(7)	0.350[275]	**0.275[450]**	0.839[30]	0.839[153]	0.869[306]	0.941[109]	0.941[20]
EQ Tau	0.442(10)	0.575[263]	0.638[250]	**0.498[103]**	0.566[308]	0.566[126]	0.537[265]	0.518[114]
FI Boo	0.327(9)	0.970[75]	0.950[313]	1.000[5]	0.970[252]	0.651[481]	0.981[240]	1.000[9]
FT UMa	0.984(19)	0.538[463]	0.363[463]	**0.941[14]**	0.802[153]	0.682[65]	0.788[139]	**0.979[16]**
SW Lac	0.776(14)	0.550[50]	0.675[250]	0.867[57]	0.843[82]	**0.804[115]**	0.856[74]	0.858[50]
SX Crv	0.066(3)	0.225[125]	0.350[463]	0.186[0]	0.186[388]	0.186[776]	0.318[319]	0.318[0]
V1191 Cyg	0.107(5)	0.089[38]	0.075[263]	0.083[85]	**0.105[408]**	**0.105[814]**	0.245[264]	0.248[81]
V523 Cas	0.516(8)	**0.488[13]**	0.625[275]	0.777[25]	0.777[98]	0.747[94]	0.774[121]	0.774[8]
V753 Mon	0.970(11)	0.675[338]	0.463[100]	0.817[9]	0.812[90]	0.812[180]	0.878[142]	0.882[3]
VW LMi	0.423(21)	0.281[63]	0.513[375]	**0.422[0]**	**0.422[119]**	**0.422[238]**	**0.424[58]**	**0.424[0]**
W UMa	0.484(3)	**0.433[13]**	0.538[163]	0.985[32]	0.985[170]	1.000[339]	0.896[154]	0.972[20]

Done so far...

- Based on training – stacked model looks as best approach
- Very small sample to proof concept
- Need for larger ensemble of sectors, individual LCs of the same object
- Better predictions for systems with total eclipses
- Further analysis of O'Connell effect, shapes of minima of Lcs
- Represent LCs in phases rather than trigonometric polynomials
- *TESS* vs. V-band, small bins of f, l_3 – needs new training sample
- $\vdots$
- Use neural network as meta-model

Done so far… Initial neural network prediction „power“



New grid generation for training

Physical model of stars with `PHOEBE` code (Prša, 2011)

Parameter space:

$q \in < 0.05, 1.00 >$;	$\Delta q = 0.025$	39
$f \in < 0.05, 0.95 >$;	$\Delta f = 0.1$	10
$i \in < 30, 90 >$ deg;	$\Delta i = 1.5$ deg	41
$T_1 \in < 4\,100, 9\,600 >$ K;	$\Delta T_1 = 250$ K	23
$T_2/T_1 \in < 0.5, 1.0 >$;	$\Delta T_2/T_1 = 0.05$	11

New grid generation for training

Physical model of stars with `PHOEBE` code (Prša, 2011)

Parameter space:

$q \in < 0.05, 1.00 >$;	$\Delta q = 0.025$	39
$f \in < 0.05, 0.95 >$;	$\Delta f = 0.1$	10
$i \in < 30, 90 >$ deg;	$\Delta i = 1.5$ deg	41
$T_1 \in < 4\,100, 9\,600 >$ K;	$\Delta T_1 = 250$ K	23
$T_2/T_1 \in < 0.5, 1.0 >$;	$\Delta T_2/T_1 = 0.05$	11

4 045 470

Efficiency

Other considerations:

- Still symmetrical LCs around φ = 0.5 ($\Rightarrow b_k$ = 0) – generate ½ LC
- **NO** LC fit with **(1)** – each represented by 65 phase points
- Eclipse regions 2x higher bin density

Add steps via interpolation between i parameters while all others fixed $\rightarrow$ final library almost 2-times bigger

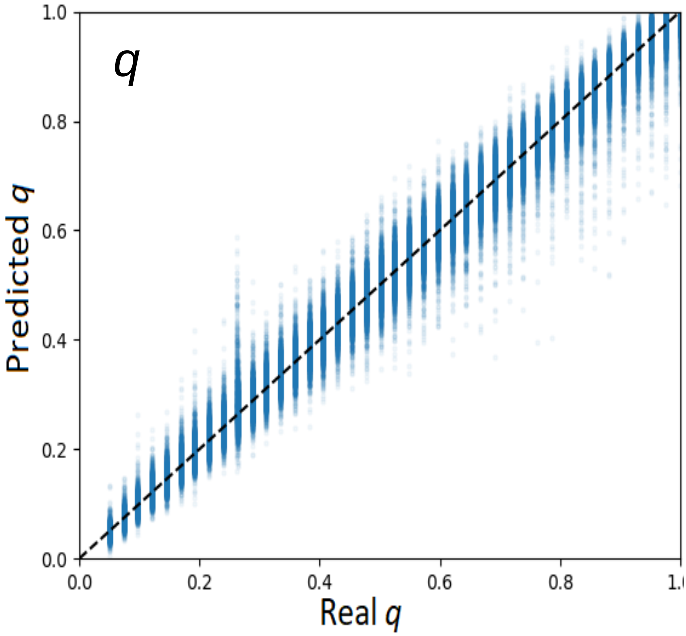
- PCA applied to reduce dimensionality and retain the most significant variations in the data – 20 principal components with 99.9% of the total variance.
- Reduced complexity, noise, redundancy and higher efficiency

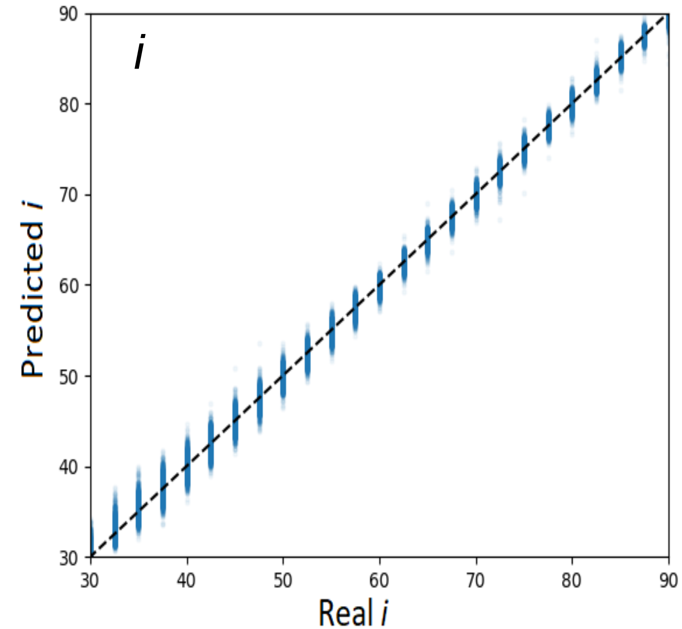
New regressor

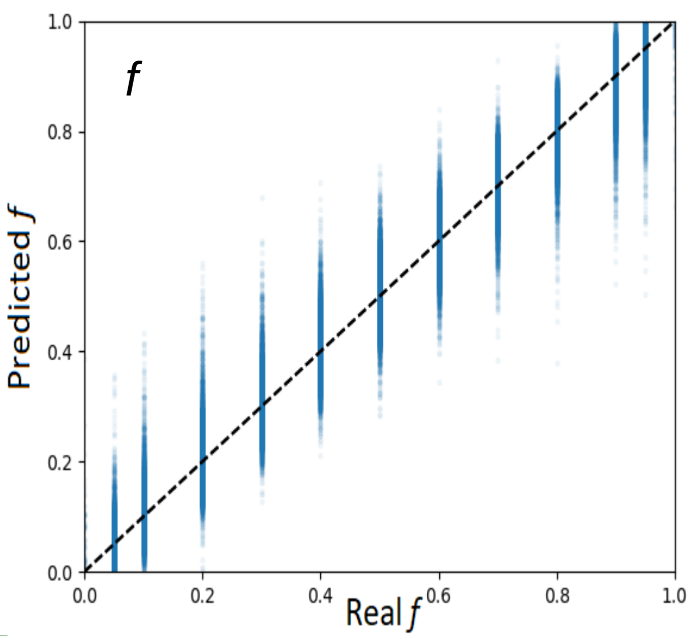
- Using `MultiOutputRegressor`
- Combine previous Random forest with XGBoost add CatBoost
- Ridge regression used best for multicollinear data or if number of predictor variables > number of observations.
- XGBoost configured with `max_depth=16`, `learning_rate=0.4`, and strong `reg_lambda=500`, enabling it to capture complex patterns while controlling overfitting.
- CatBoost used with the `MultiRMSE` loss function, `tree_depth=12`, and a low learning rate of 0.015, offering high accuracy and robustness to feature noise.
- Random forest regularized via `ccp_alpha=0.1` and parallelized with `n_jobs=8` to optimize training speed and generalization.

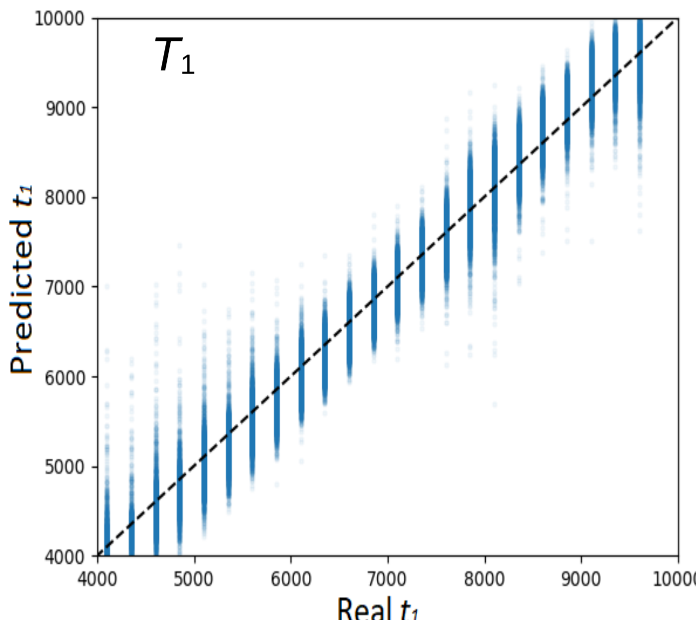
New regressor

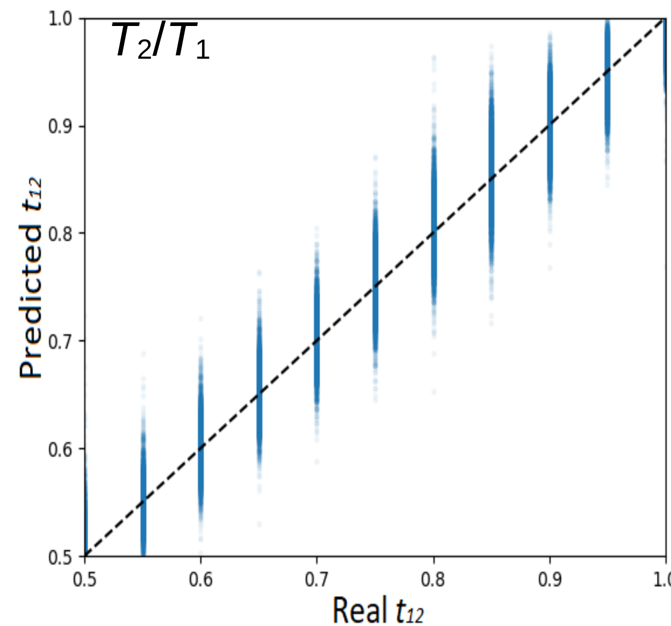
RMS	Training set	Test set
q	0.0250	0.0251
f	0.0352	0.0352
i [deg]	0.5679	0.5666
T_1 [K]	190.88	191.07
T_2/T_1	0.0171	0.170



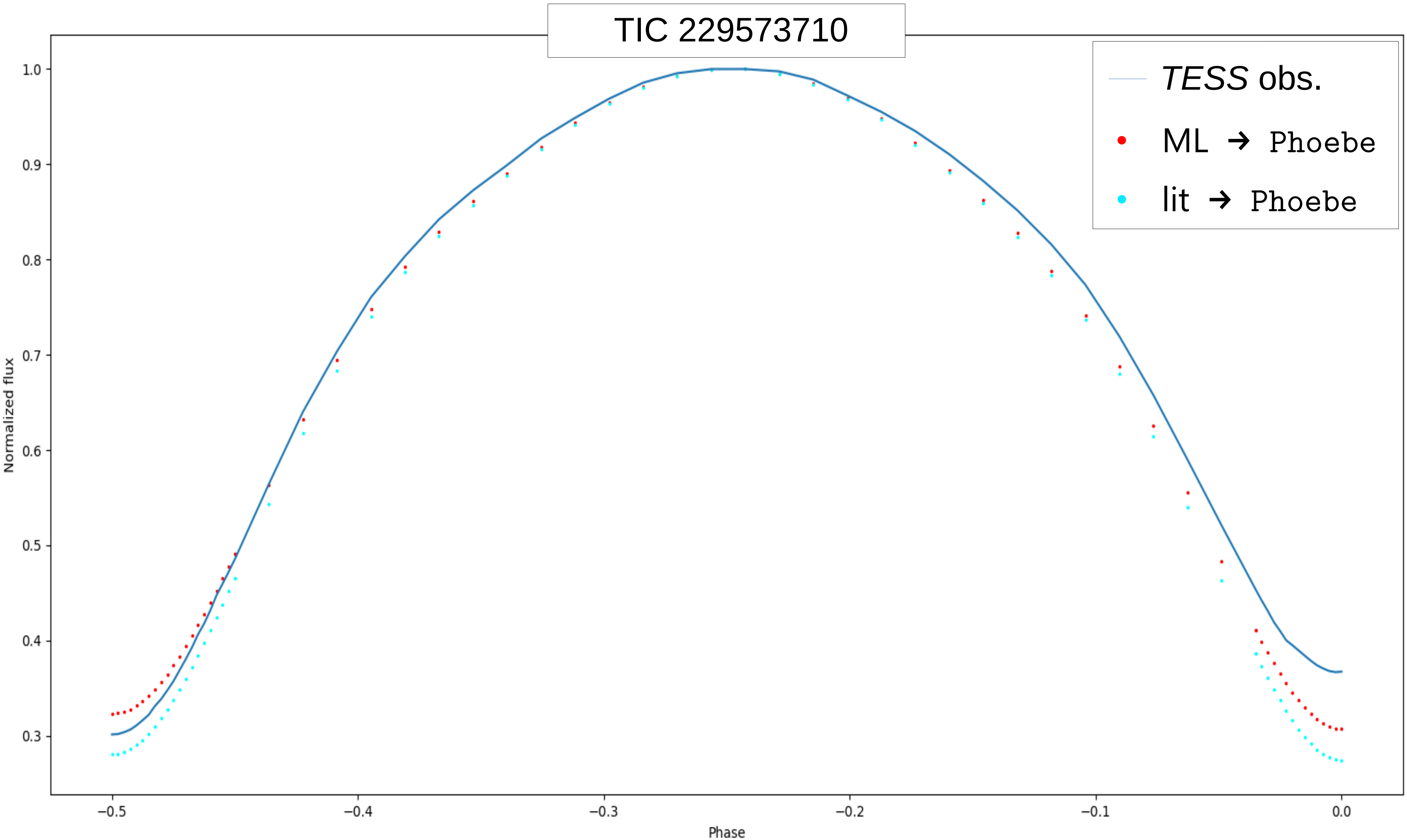








Results so far...



TESS Phase light curves of binaries and search for a close match in a pre-compiled database

SKALNATÉ PLESO

Thank you !

This work was supported by grants:
APVV-20-0148 and VEGA 2/0031/22

AGENTÚRA NA PODPORU VÝSKUMU A VÝVOJA

MINISTERSTVO ŠKOLSTVA, VÝSKUMU, VÝVOJA A MLÁDEŽE SLOVENSKEJ REPUBLIKY