Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals Electronic Supplementary Information

Felix Musil,[*a] Sandip De,[*a] Jack Yang,[b‡] Joshua E. Campbell,[b‡] Graeme M. Day[b‡] and Michele Ceriotti[a]

1 Supplementary Data

The input data that we used to build the sketch-maps and perform the property predictions is provided in the Electronic Supplementary materials. Each systems has its structural information in extended xyz format (.xyz file) and its corresponding properties (.prop file). Interactive sketch-map visualizers for all the systems considered are available on http://interactive.sketchmap.org/ and as offline versions in an included zip file (see the README file for more details).

Most of our analysis where performed with the glosim python package[1] which relies on the QUIP code[2] to compute the SOAP vectors. To compute a SOAP-REMatch kernel, a typical command line is:

```
 ~/git/glosim/glosim.py traj-pentacene.xyz
-n 9 -l 9 -g 0.3 -c 3 --kernel rematch
--gamma 2 --periodic --nonorm
```

Note that for the cutoff radius of 3Å and 5Å, the parameters for the radial and angular expansion are respectively chosen as `-n 9 -l 9` and `-n 12 -l 12`.

2 Crystal Structure Prediction

CSP were performed with Global Lattice Energy Explorer (GLEE)[3] for possible crystal packings of a given molecules in the 23 most commonly adopted space groups for organic molecules in $Z' = 1$, and 12 common space groups for molecules that crystallise in $Z' = 2$.[4] This led to a total of 212,000 trial crystal structures, which were subsequently energy minimised in DMACRYS[5] using the W99 atom–atom intermolecular potentials[6–9], and multipolar electrostatics described by the distributed multipole model[10]. Duplicated crystal structures were removed using COMPACK[11] to consolidate a final list of structures for subsequent analysis.

3 Density Functional Calculation

Single point energy calculations for the discussed set of molecular crystals have been carried out within Density functional theory (DFT) with quantum espresso code[12]. Plane wave basis set with wavefuction cutoff of 100Ry and charge density cutoff of 400Ry has been used, together with projector augmented wave (PAW) type pseudo potentials (non-linear core correction and scalar relativistic) and Perdew-Burke-Ernzerhof (PBE)[13] exchange correlation functional. To account for vanderwals interaction, Grimme's vanderwals dispersion correction[14] has been used with a cutoff radius of 80 bohr. The energy has been converged within an accuracy of 10^{-6} Hartree. The correlation between W99 and DFT energies are shown in Fig. 4

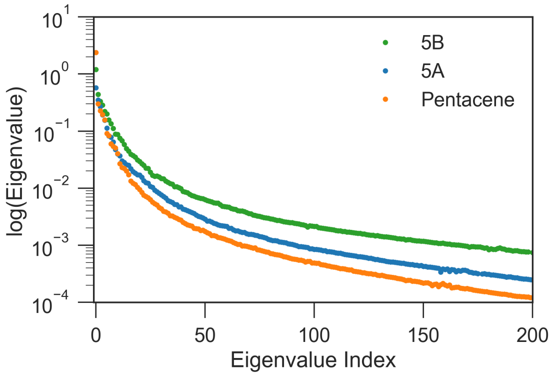


Figure 1 First 200 largest eigenvalues corresponding to the kernel matrices of pentacene, 5A and 5B datasets with cutoff radius of 5Å, gaussian width of 0.3Å and $\gamma = 2$.

4 Kernels and Sketch-maps

Fig. 1 shows the first 200 eigenvalues sorted by deacreasing order of the centered kernel matrices (see Ref. 15 for more details on the kernel matrix centering) of the pentacene, 5A and 5B datasets. As described in the main text, the 5B dataset has larger eigenvalues compared to the other two datasets which indicates that the 5B dataset is sparser in terms of structural diversity.

Figs. 2 and 3 show the sketch-map representation of the kernel matrices of the 5A and 5B dataset used for the lattice energy prediction (cutoff radius of 5Å). The visual correlation between the sketch-maps and the lattice energy seems quite good. Moreover, cluster found with HDBSCAN* match the sheet and γ heuristic class. However, these cluster do not appear to define clear cut structural motifs.

5 Error Analysis for Energy Calculations

We built a regression model to predict the energies of different polymorphs of pentacene, 5A and 5B molecular crystals, using as inputs the geometries optimized using the W99 empirical potential. Using this same set of inputs, and the kernels discussed in the main text, we performed regressions for (1) W99 energies (an insightful but rather academic benchmark exercise for the model); (2) dispersion-corrected DFT energies computed on the W99 geometries; (3) DFT+D energies computed by taking W99 energies as the baseline, that is energies predicted as $E_{\Delta} = E_{\mathrm{W99}} + ML(E_{\mathrm{DFT}} - E_{\mathrm{W99}})$. Fig. 4 shows the correlation be-

[0a] *National Center for Computational Design and Discovery of Novel Materials (MARVEL), Laboratory of Computational Science and Modelling, Institute of Materials, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, E-mail:sandip.de@epfl.ch*

[0b] *School of Chemistry, University of Southampton, Highfield, Southampton, UK.*

[0*] These authors contributed equally to this work.

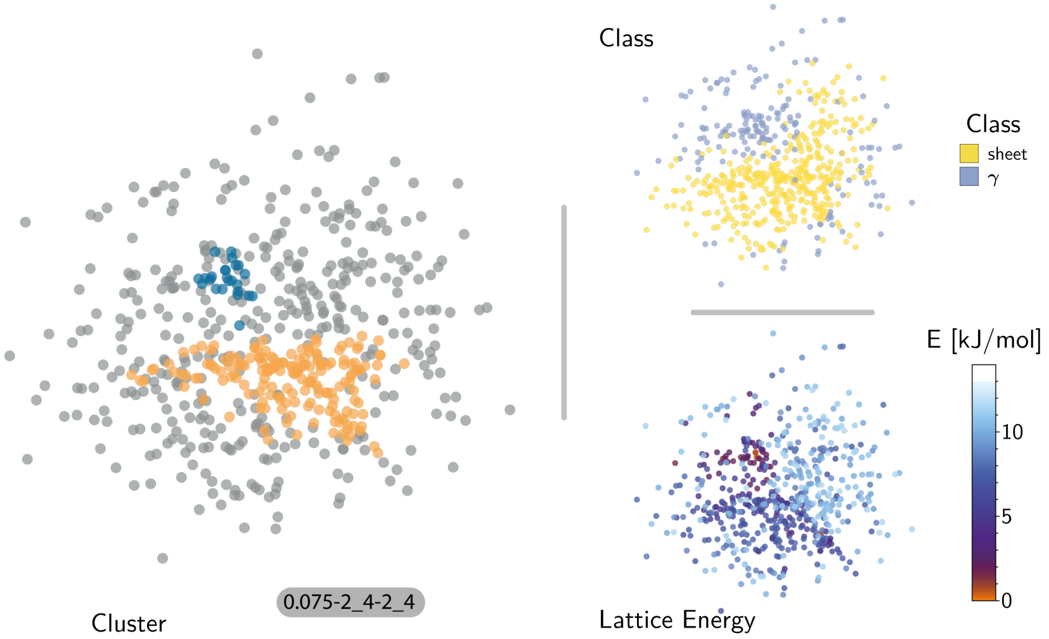


Figure 2 Representation of the 5A's similarity matrix with a cutoff radius of 5Å (projection parameters shown follow the scheme $\sigma_{map}\text{-}A\_B\text{-}a\_b$). The atomic configurations, i.e. disks, on the three sketchamps are color-coded according to their lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is represented using the front view of the pentacene molecules.

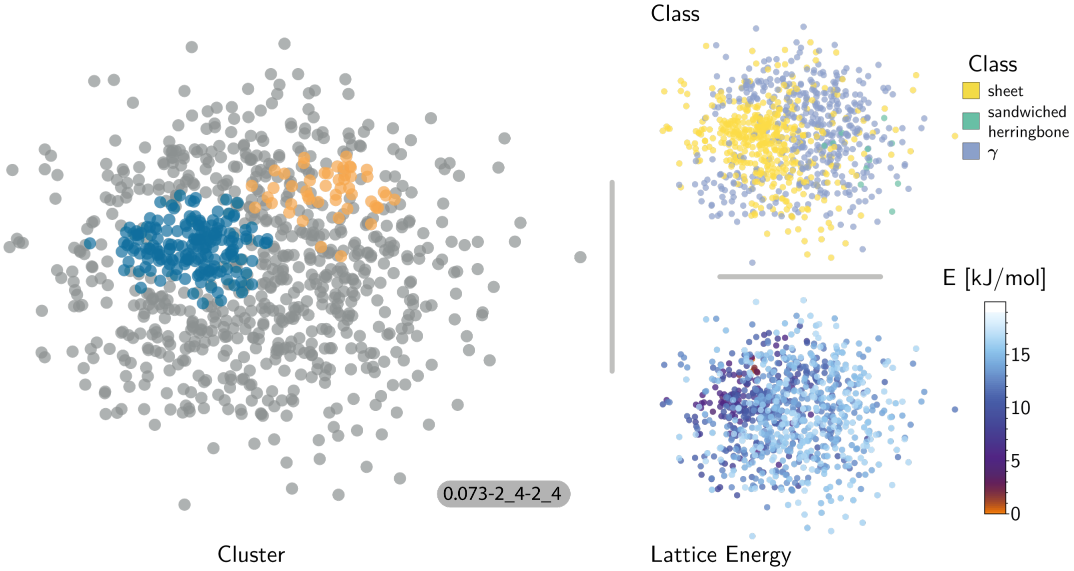


Figure 3 Representation of the 5B's similarity matrix with a cutoff radius of 5Å (projection parameters shown follow the scheme $\sigma_{map}\text{-}A\_B\text{-}a\_b$). The atomic configurations, i.e. disks, on the three sketchamps are color-coded according to their lattice energy (bottom right), class following the heuristic classification (top right) and cluster index (gray structure do not belong to a cluster) found using HDBSCAN* on the similarity matrix (left). The structural pattern of each cluster is represented using the front view of the pentacene molecules.

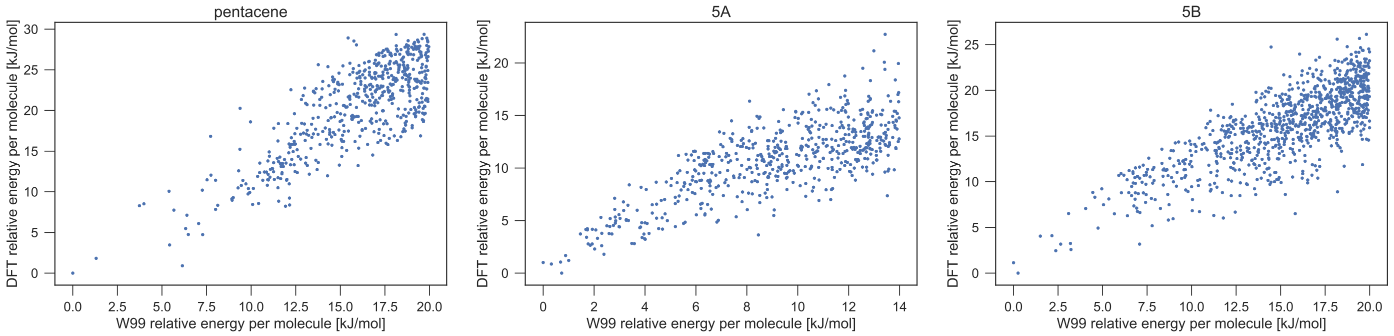


Figure 4 The correlation between the W99 and DFT relative lattice energy of pentacene, 5A and 5B crystals, for W99-optimized geometries.

Table 1 Summary of the lattice energy prediction scores for pentacene, 5A and 5B (respectively 564, 594 and 936 structures). Our best accuracies on these datasets are estimated from average scores from a 4-fold cross validation (75% of the dataset is used for training). Δ-learning refers to the learning of the difference between W99 and DFT energies.

Training set	Dataset	SD[kJ/mol]	MAE [kJ/mol]	RMSE [kJ/mol]	R^2	Spearman Coefficient
75%	Pentacene(W99)	3.38	0.29 ± 0.03	0.49 ± 0.08	0.98	0.99
	Pentacene(DFT)	5.49	0.48 ± 0.04	0.68 ± 0.04	0.98	0.99
	Pentacene(Δ)	3.42	0.51 ± 0.04	0.70 ± 0.06	0.96	0.98
	5A(W99)	3.31	0.41 ± 0.02	0.59 ± 0.04	0.97	0.98
	5A(DFT)	3.56	0.64 ± 0.03	0.91 ± 0.07	0.93	0.95
	5A(Δ)	2.37	0.59 ± 0.03	0.85 ± 0.06	0.85	0.94
	5B(W99)	3.86	0.98 ± 0.03	1.31 ± 0.03	0.88	0.93
	5B(DFT)	4.23	1.09 ± 0.03	1.44 ± 0.04	0.87	0.93
	5B(Δ)	2.66	0.74 ± 0.04	1.00 ± 0.05	0.83	0.92
10%	Pentacene(W99)	3.38	0.63 ± 0.04	0.89 ± 0.08	0.92	0.94
	Pentacene(DFT)	5.49	1.04 ± 0.01	1.42 ± 0.04	0.93	0.95
	Pentacene(Δ)	3.42	0.89 ± 0.05	1.20 ± 0.10	0.85	0.94
	5A(W99)	3.31	1.10 ± 0.02	1.46 ± 0.04	0.76	0.88
	5A(DFT)	3.56	1.44 ± 0.04	1.88 ± 0.06	0.62	0.81
	5A(Δ)	2.37	1.02 ± 0.04	1.36 ± 0.05	0.54	0.84
	5B(W99)	3.86	1.90 ± 0.10	2.40 ± 0.10	0.36	0.72
	5B(DFT)	4.23	2.15 ± 0.06	2.77 ± 0.09	0.29	0.74
	5B(Δ)	2.66	1.25 ± 0.02	1.67 ± 0.05	0.36	0.78

tween W99 and DFT+D energies.

Table 1 reports the test errors using 75% and 10% train points for the different systems and learning strategies. As well as the MAE and RMSE error we report the R^2 and Spearman rank correlation coefficients, that characterize the learning efficiency for the correlations, and how well predictions preserve the ranking of different phases.

Figure 5 reports a thorough analysis of test and train error learning curves, and error distributions at different levels of training. It should be stressed that the automatic determination of the regularization parameter leads to train error curves that are not linear on a log-log scale, as a tighter fit is beneficial when the train set becomes denser. Judging from the very regular behavior of the test set errors, this does not lead to overfitting artifacts. The violin plots (Figs. 5 (c-e)) are consistent with a near-Gaussian distribution of errors for all but the smallest train set sizes, with a few outliers but no siginificant skewness or multi-modality apparent in the distributions.

6 Mobility Calculation

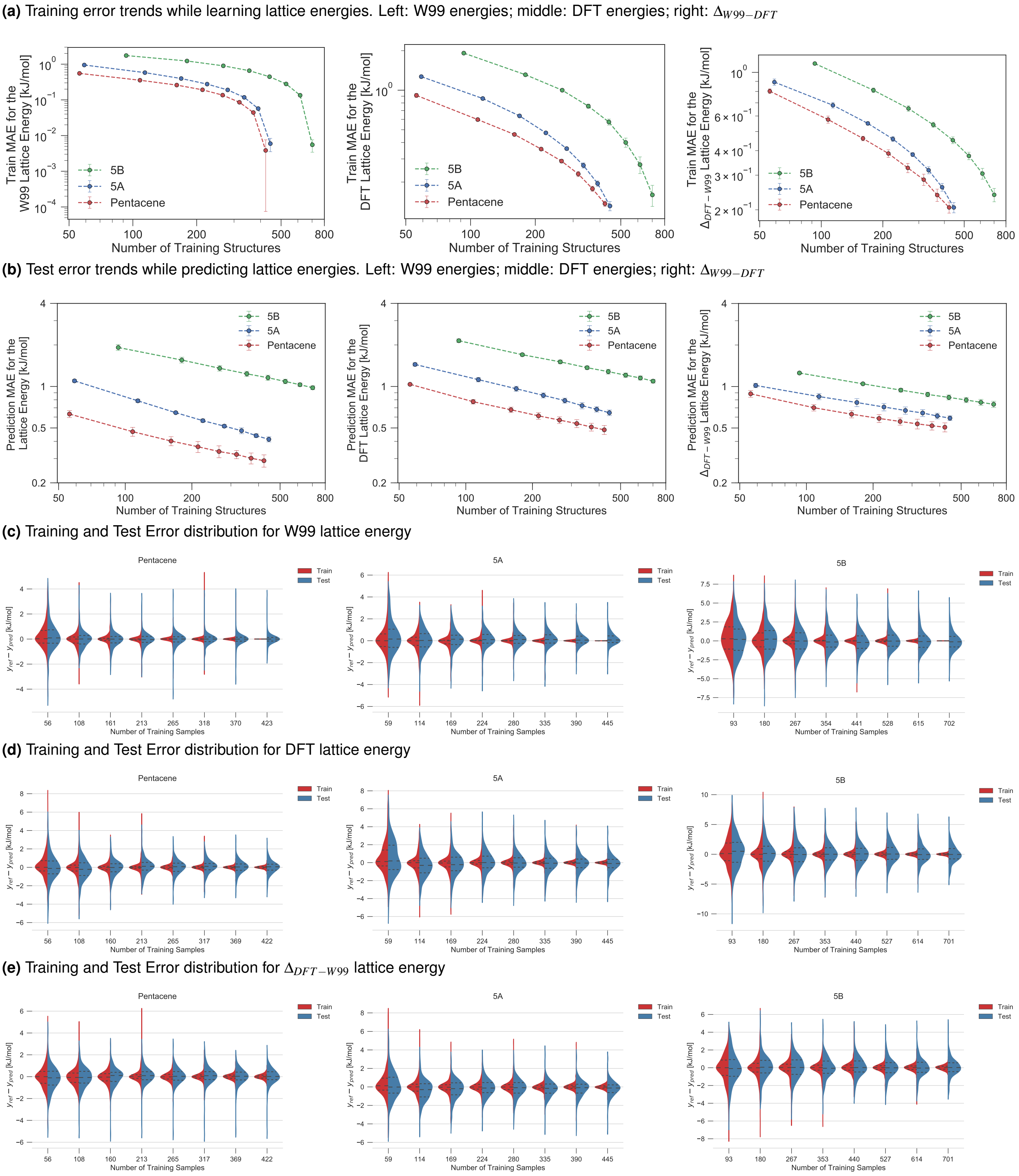
We outline the theories that underpin the calculations of charge mobilities in our work. To gather enough data for machine learning purposes, an extended list of predicted crystal structures (up to 15 kJ/mol above the predicted global minimum for 5A) were subjected to mobility calculations. For each crystal structure, the charge mobility was estimated using Einstein relationship:

$$\mu = \frac{e}{k_B T} D, \tag{1}$$

where e is the charge of electrons, k_B is the Boltzmann constant, T is the temperature and was set to 300 K. The electron diffusivity (D) was evaluated as:

$$D = \frac{1}{2nM} \sum_{n=1}^{M} \sum_{j=1}^{N_i} r_{ij}^2 k_{ij} P_{ij}, \tag{2}$$

in which M is the total number of symmetrically independent molecules in a crystal that can be related to the Z' number for a crystal. For the i–th symmetrically independent molecule, N_i number of nearest–neighbouring molecules will be extracted, which



(a) Training error trends while learning lattice energies. Left: W99 energies; middle: DFT energies; right: $\Delta_{W99-DFT}$

(b) Test error trends while predicting lattice energies. Left: W99 energies; middle: DFT energies; right: $\Delta_{W99-DFT}$

(c) Training and Test Error distribution for W99 lattice energy

(d) Training and Test Error distribution for DFT lattice energy

(e) Training and Test Error distribution for $\Delta_{DFT-W99}$ lattice energy

Train MAE for the W99 Lattice Energy [kJ/mol]
Train MAE for the DFT Lattice Energy [kJ/mol]
Train MAE for the $\Delta_{DFT-W99}$ Lattice Energy [kJ/mol]
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
6×10^{-1}
4×10^{-1}
3×10^{-1}
2×10^{-1}
5B
5A
Pentacene
50
100
200
500
800
Number of Training Structures
Prediction MAE for the Lattice Energy [kJ/mol]
Prediction MAE for the DFT Lattice Energy [kJ/mol]
Prediction MAE for the $\Delta_{DFT-W99}$ Lattice Energy [kJ/mol]
4
1
0.5
0.2
Pentacene
5A
5B
Train
Test
$y_{ref} - y_{pred}$ [kJ/mol]
Number of Training Samples
56
108
161
213
265
318
370
423
59
114
169
224
280
335
390
445
93
180
267
354
441
528
615
702
160
317
369
422
353
440
527
614
701

Figure 5 The figures report extensive analytics for the prediction of lattice energies of pentacene, 5A and 5B polymorphs.

Table 2 Prediction accuracy of TI while training on different fraction of the full dataset of 6697 pentacene dimers, selected with Farthest point sampling method on similarity kernel data as well as randomly. All the MAE, RMSE and SUP values noted in the table are in eV and averaged over values obtained by randomly choosing 10 different training and test set.

n_{train}		FPS	Selection			Random	Selection	
($\sigma = 0.025$)	MAE	RMSE	SUP	R^2	MAE	RMSE	SUP	R^2
670	0.0101	0.0190	0.1720	0.45	0.0095	0.0202	0.2608	0.51
1674	0.0060	0.0121	0.1009	0.77	0.0071	0.0168	0.2563	0.66
3348	0.0030	0.0059	0.0489	0.94	0.0054	0.0130	0.2102	0.78
5023	0.0023	0.0048	0.0319	0.96	0.0049	0.0121	0.1577	0.80
5358	0.0021	0.0045	0.0287	0.96	0.0045	0.0105	0.1110	0.85

Table 3 Prediction accuracy of TI while training on different fraction of the full dataset of 7305 5A dimers, selected with Farthest point sampling method on similarity kernel data as well as randomly. All the MAE, RMSE and SUP values noted in the table are in eV and averaged over values obtained by randomly choosing 10 different training and test set.

n_{train}		FPS	Selection		Random	Selection		
($\sigma = 0.052$)	MAE	RMSE	SUP	R^2	MAE	RMSE	SUP	R^2
730	0.0134	0.0294	0.20	0.69	0.0053	0.0118	0.17	0.95
1826	0.0028	0.0050	0.06	0.99	0.0035	0.0081	0.10	0.97
3652	0.0017	0.0033	0.03	0.99	0.0027	0.0064	0.09	0.98
5479	0.0008	0.0020	0.02	0.99	0.0023	0.0054	0.07	0.99
5844	0.0005	0.0017	0.02	0.99	0.0022	0.0053	0.07	0.99

gives rise to a total of MN_i dimer pairs in the crystal structure. Symmetrically equivalent dimers based on an RMSD$<$ 0.1 criteria was filtered out from explicit transfer integral calculations with DFT to descrease the overall computational cost. For each dimer, r_{ij} denotes its inter–centroid distance, k_{ij} is the corresponding charge hopping rate, derived from Marcus theory:

$$k_{ij} = \frac{t_{ij}^2}{\hbar}\sqrt{\frac{\pi}{\lambda k_B T}}\exp\left[-\frac{\lambda}{4k_B T}\right], \tag{3}$$

where t_{ij}, the transfer integral, describes the intermolecular electronic coupling which depends on the relative positions and orientations of the molecules in the crystal structure and λ is the intramolecular reorganisation energy, and was calculated here using the conventional four–point models at B3LYP/6-311G** level of theory with GAUSSIAN09. P_{ij} is the probability for charge to hop between molecule i and j and it is related to the transfer integral as:

$$P_{ij} = \frac{k_{ij}}{\sum_{j=1}^{N_i} k_{ij}} = \frac{t_{ij}^2}{\sum_{j=1}^{N_i} t_{ij}^2}. \tag{4}$$

It should be clear from the above discussions that the key quantity that varies across crystal structures is t_{ij}, which is explicitly calculated with frozen–density embedding (FDE) DFT scheme. The calculations were performed at PW91/DZ level of theory with the non–additive kinetic energy modelled with PW91k functional. A threshold of $S < 10^{-2}$, below which the Penrose pesudoinverse was applied in the final calculations of TI, was applied globally for all dimers considered, in order to avoid numerical instabilities when the orbital overlap between two monomers, S, is less than 10^{-2}. Hence our key effort here in accelerating mobility calculations will be focusing on direct prediction of t_{ij}'s for all dimers extracted from predicted crystal structures. The origins of non pairwise additivities in the t_{ij} values can be partially understood from the theory of FDE applied to calculate t_{ij}'s which will be briefly discussed as following.

FDE was built on the basis that the total electron densities of two interacting systems can be exactly partitioned into the sum of electron densities of two interacting systems as $\rho(\mathbf{r}) = \rho_I(\mathbf{r}) + \rho_{II}(\mathbf{r})$. In a Kohn–Sham scheme, where the total energy of the system is a functional of the total charge densities $E[\rho(\mathbf{r})]$, the same partition scheme for density does not apply for the total energy, in which a interacting non–additive component must be included as

$$E[\rho(\mathbf{r})] = E_I[\rho_I(\mathbf{r})] + E_{II}[\rho_{II}(\mathbf{r})] + E_{int}[\rho_I(\mathbf{r}), \rho_{II}(\mathbf{r})]. \tag{5}$$

In FDE, this is achieved by including a embedding potential $v_{emb}(\mathbf{r})$ in the Kohn–Sham equation, which takes into account contributions from non–additive kinetic and exchange–correlation energies. Furthermore, the embedding potential $v_{emb}^{I(II)}(\mathbf{r})$ acting on subsystem I (II) contains a Coulomb interaction between $\rho^I(\mathbf{r})$ and $\rho^{II}(\mathbf{r})$, and this was solved iteratively via 'freeze–and–thaw' cycles by updating the electron densities of one subsystem while keeping the other one frozen. For the evaluation of t_{ij}, one needs to introduce an additional electron/hole into the charge densities of the subsystems, thus $E_{int}[\rho_I(\mathbf{r}), \rho_{II}(\mathbf{r})]$ in Eq. (5) would also involve energetic contributions from polarised electron densities, which is also non–pairwise additive.

7 TI prediction protocol

The protocol we propose to evaluate the TI for dimer configurations found in the low-lying crystalline polymorphs is the following:

1. Extract all possible dimers from the molecular crystals whose charge carrier mobilities are to be screened, in the same way as one would have done for the mobility calculations.

2. Select a training subset from all the dimers such that maximum structural diversity is captured. Farthest point sampling based on the same kernel used for predictions would be one of the recommended paths to achieve robust, systematically-improvable accuracy.

3. Calculate the TI with an appropriate method capable of providing the desired accuracy, only for the training set dimers.

4. Combine the weight vector for the training set with the kernel computed between training configurations and the remainder of the dimers to predict their TI values.

Table 4 Prediction accuracy of TI while training on different fraction of the full dataset of 11581 5B dimers, selected with Farthest point sampling method on similarity kernel data as well as randomly. All the MAE, RMSE and SUP values noted in the table are in eV and averaged over values obtained by randomly choosing 10 different training and test set.

n_{train}		FPS	Selection			Random	Selection	
$(\sigma = 0.046)$	MAE	RMSE	SUP	R^2	MAE	RMSE	SUP	R^2
1158	0.0118	0.0260	0.2080	0.70	0.0061	0.0114	0.1419	0.94
2895	0.0059	0.0133	0.1458	0.93	0.0041	0.0082	0.1271	0.97
5790	0.0021	0.0046	0.0533	0.99	0.0031	0.0065	0.0911	0.98
8686	0.0011	0.0029	0.0282	0.99	0.0027	0.0055	0.0704	0.98
9265	0.0009	0.0026	0.0252	0.99	0.0026	0.0055	0.0655	0.98

5. Combine the predicted TI values following the standard method described in supplementary information to obtain the charge carrier mobility in each crystal structure.

The TI prediction accuracy and performance data for pentacene, 5A and 5B dimers, using the kernel parameters discussed in the main text, are reported for reference in Tables 2 to 4.

References

[1] M. Cerrioti, S. De and F. Musil, *Glosim package*, https://github.com/cosmo-epfl/glosim.

[2] G. Csanyi, J. Kermode and N. Bernstein, *QUIP and quippy documentation*, http://libatoms.github.io/QUIP/index.htmlhttps://libatoms.github.io/QUIP/.

[3] D. H. Case, J. E. Campbell, P. J. Bygrave and G. M. Day, *J. Chem. Theory Comput.*, 2015.

[4] C. P. Brock and J. D. Dunitz, *Chem. Mater.*, 1994, **6**, 1118–1127.

[5] S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.

[6] D. E. Williams, *J. Mol. Struct.*, 1999, **485-486**, 321–347.

[7] D. E. Williams, *J. Comp. Chem.*, 2001, **22**, 1–20.

[8] D. E. Williams, *J. Comp. Chem.*, 2001, **22**, 1154–1166.

[9] E. O. Pyzer-Knapp, H. P. Thompson and G. M. Day, *Acta Cryst. B*, 2016, **72**, 477.

[10] A. Stone and M. Alderton, *Molecular Physics*, 2002, **100**, 221–233.

[11] J. A. Chisholm and S. Motherwell, *J. Appl. Cryst*, 2005, **38**, 228–231.

[12] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari and R. M. Wentzcovitch, *Journal of Physics: Condensed Matter*, 2009, **21**, 395502.

[13] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

[14] S. Grimme, *Journal of Computational Chemistry*, 2006, **27**, 1787–1799.

[15] B. Schölkopf, A. Smola and K.-R. Müller, *Neural Computation*, 1998, **10**, 1299–1319.