RESEARCH ARTICLE

[View Article Online](#)
[View Journal](#) | [View Issue](#) Cite this: *Med. Chem. Commun.*,
2018, 9, 1961The emerging chemical patterns applied in
predicting human toll-like receptor 8 agonists†Shuheng Huang,^{ab} Hu Mei,^{ID} *^{ab} Duo Zhang,^b Yubin Ren,^b
MuliadiYeremia Kevin^b and Xianchao Pan^{*bc}

Toll-like receptors (TLRs) are important pattern recognition receptors to human innate immunity, which can recognize pathogen-associated molecular patterns and initiate innate immune responses. As the receptor of single stranded RNA (ssRNA), toll-like receptor 8 (TLR8) has potential in the treatment of tumors, microbial infection, and inflammatory diseases. Herein, an emerging chemical pattern (ECP) method was utilized to predict the key chemical patterns of TLR8 agonists. Based on the ECPs discovered, a robust and predictive ECP model was derived with prediction accuracies of 83.3%, 81.0%, and 80.0% for 132 training samples, 79 validation samples, and 75 test samples, respectively. When the ECP model was applied with a molecular docking method, the hit rate of TLR8 agonists was greatly enhanced. The results of ECP-based hierarchical cluster analysis and Connolly surface analysis of the TLR8 receptor showed that the H-bonding, hydrophilic and hydrophobic potentials as well as the unbalanced degree of property distributions are very important for distinguishing the TLR8 agonists from non-agonists.

Received 31st May 2018,
Accepted 19th September 2018

DOI: 10.1039/c8md00276b

rsc.li/medchemcomm

1. Introduction

Toll-like receptors (TLRs) belong to a class of critical proteins related to innate and adaptive immune responses.¹ In 1996, the first TLR was found by Jules A. Hoffmann² in the *Drosophila* immune system, which is activated upon the recognition of pathogen associated molecular patterns (PAMPs).^{3,4} In the past decades, many TLRs have been found in mammalian dendritic cells, monocytes, and mononuclear macrophages. These TLRs can recognize specific components among microorganisms and initiate innate immune responses.^{5–8} In mammals, there are 13 members in the TLR family: 10 have been found in humans (TLR1–9, TLR11) and others have been merely found in mice (TLR10, TLR12, and TLR13).⁹

Recently, there has been growing interest in human TLR7, TLR8 and TLR9, which belong to the same subfamily of leucocyte pattern-recognition receptors. TLR8 is expressed on the surface of endosomes and can activate the downstream signaling pathways by recognizing single-stranded viral RNA and various small molecules.^{10,11} Human TLR8 consists of an extracellular leucine-rich repeat (LRR) domain, a transmem-

brane (TM) domain, and an intracellular toll/IL-1R homology (TIR) domain. The extracellular domain can bind directly to TLR8 agonists and initiate a multimerization process. The intracellular signaling domain shares a high sequence similarity with the interleukin-1 receptor and is thus termed the TIR domain.^{12,13} In the signaling pathway downstream of the TIR domain, the adaptor protein MyD88 is required for the induction of inflammatory cytokines such as TNF and IL-12.

Fig. 1 shows the extracellular LRR domains of human TLR8 dimers in apo (PDB ID: 3W3G) and ligand-bound (PDB ID: 4R07) states, respectively. Each monomer contains 26 LRRs and is characterized by 6 short loop regions (LRR2, LRR5, LRR8, LRR11, LRR18, and LRR20) and a long Z-loop region (~40 amino acids) between LRR14 and LRR15 (Fig. 1a).^{11,14}

Two distinct binding sites for recognizing the degradation products of ssRNA^{15–17} can be observed in the ligand-bound state (Fig. 1b). One bound by a uridine mononucleotide is located in the dimerization interface, and this site can also be bound by a series of small molecules; the other bound by a short oligonucleotide UG is on the concave surface of the TLR8 horseshoe structure. The binding site of uridine involves LRR11–14 and LRR16*–18* (or LRR11*–14* and LRR16–18) and is mainly composed of Lys350, Gly351, Tyr353, Phe405 and Arg429 of chain A (blue) and Asp543*, Asp545*, and Gly572* of chain B (green). Uridine can form H-bonding interactions with Asp543*, Arg429 and Asp545* (Fig. 1c). The binding site of UG involves LRR10–13 and the

^a Key Laboratory of Biorheological Science and Technology (Ministry of Education), Chongqing University, Chongqing 400044, China.
E-mail: meihu@cqu.edu.cn; Tel: +86 23 65112677

^b College of Bioengineering, Chongqing University, Chongqing 400044, China

^c Department of Medicinal Chemistry, College of Pharmacy, Southwest Medical University, Luzhou, Sichuan, 646000, China. E-mail: panxc@swmu.edu.cn

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8md00276b

As notable pattern-recognition receptors, TLR7 and TLR8 have been suggested as important therapeutic targets of inflammatory disorders, autoimmune diseases, cancer, and viral infections. The first small-molecule TLR7 agonist, namely, imiquimod (R837), is clinically approved for the treatments of basal cell carcinoma, papilloma-induced genital warts and human papillomavirus. Subsequently, TLR7/8 dual agonist

Based on 13 TLR7 and TLR8 agonists with heterocyclic scaffolds of oxazolo-zoquinoline, imidazo-quinoline, thiazolo-quinoline, triazolo-quinazoline and pyrazolo-quinolines *etc.*, Yoo *et al.*³¹ predicted TLR8-active and TLR7/8 dual active compounds by using linear discriminant analysis of Mulliken charges. All the 13 training samples were classified correctly. The results showed that the partial charges of atoms play an

Fig. 2 The structures of hTLR7 and hTLR8 agonists.

important role in the determination of the biological activities of TLR7/8 agonists. It should be noted that the predictive power of the linear discriminant model needs further validation on a large test dataset. By employing comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) methods, Deng *et al.*³² derived partial least squares (PLS) models from 55 TLR8 agonists. The R^2 , Q^2 and predictive Q^2 values were 0.98, 0.666, and 0.826 for the optimal CoMFA model, and 0.935, 0.627, and 0.872 for the optimal CoMSIA model, respectively. However, the significant differences between R^2 and Q^2 (*i.e.* $R^2 \gg Q^2$) indicate that the two PLS models are over-fitting in some degree although the prediction performances are quite high.

In spite of these exploratory studies, the development of an efficient prediction method of TLR8 agonists is still on the way. In this paper, a total of 286 samples consisting of 149 TLR8 agonists and 137 non-agonists were collected from the available literature. Herein, an emerging chemical pattern (ECP)³³ was employed to establish prediction models. The results showed that the ECP is an extremely efficient and promising method for predicting TLR8 agonists. The prediction accuracies of the optimal ECP model are 83.3%, 81.0% and 80.0% for 132 training samples, 79 validation samples and 75 test samples, respectively. By combining the ECP model with a molecular docking method, the hit rate of TLR8 agonists was greatly enhanced. Besides, the key structural features distinguishing TLR8 agonists from non-agonists were also investigated by ECP-based hierarchical cluster analysis and Connolly surface analysis of the TLR8 binding pocket.

2. Materials and methods

2.1 Datasets

A dataset of 211 molecules (97 TLR8 agonists/114 non-agonists) was derived from recent studies.^{9,20,21,24–27,31,34–36} For most of the compounds, the agonist activities were detected

by human TLR8-specific reporter gene assays, which were quantified by the induction of NF- κ B in human embryonic kidney 293 (HEK 293) cells. Herein, the compounds with an EC_{50} value larger than 100 μ M were assigned as “non-agonists”.³⁰

Before building prediction models, the TLR8 agonists were labeled 1 (positive) and non-agonists –1 (negative). Then, the 211 samples were randomly divided into 132 training samples (61 agonists/71 non-agonists) and 79 validation samples (36 agonists/43 non-agonists). In addition, 75 test samples (52 agonists/23 non-agonists) were collected from the most recent studies^{23,30} and used to test the predictive power of the resulting ECP models. The detailed sample information is shown in Table S1.†

2.2 Structural description and feature selection

VolSurf³⁷ is a grid-based structural description method based on 3-D molecular fields of interaction energies. VolSurf can quantify molecular steric, hydrophilic, electronic and H-bonding properties by using specific probes,³⁸ and it has been successfully employed to predict pharmacokinetic properties in the last decade.^{39–44}

Before calculating VolSurf descriptors, all the molecules were charged by the MMFF94 method and then optimized by the Tripos force field with a conjugate gradient minimizer (Sybyl 8.1). The maximum iteration steps and energy gradient were set to 5000 times and 0.05 kcal mol^{–1} Å^{–1}, respectively. Based on the optimal conformation obtained for each molecule, 118 VolSurf descriptors were calculated by using the 5 most common probes, *i.e.*, a water probe (H₂O), a hydrophobic probe (DRY), an amphipathic probe (BOTH), H-bonding carbonyl (O), and an sp² carboxy oxygen atom (O:). Besides, 102 classic physico-chemical descriptors were also calculated, which consist of atom, bond and ring counts, pharmacophore properties, potential energies, and surface

Table 1 A hypothetical dataset for illustration of the ECP concept

Class	Compound ID	Descriptor 1			Descriptor 2			Descriptor 3		
		I1	I2	I3	I1	I2	I3	I1	I2	I3
Positive	1	•				•			•	
	2	•				•			•	
	3			•	•			•		
Negative	4		•				•			•
	5	•				•			•	
	6			•		•				•

properties *etc.* Thus, a total of 220 molecular descriptors were generated for each molecule. Before ECP modeling, stepwise linear discriminant analysis (SLDA) was used for feature selection, of which the entry and removal probabilities were set to 0.1 and 0.5, respectively. Then, the candidate subsets of descriptors were employed for the following ECP modeling.

2.3 Emerging chemical pattern modeling

2.3.1. The concept of the emerging chemical pattern. The emerging chemical pattern (ECP) was derived from the concept of an emerging pattern firstly applied in gene expression studies and introduced into cheminformatics in 2007.^{33,45} As a powerful classification method, the ECP can construct accurate classifiers on a few positive samples with different skeletons.^{46,47}

Table 1 shows a hypothetical dataset comprising 3 positive samples and 3 negative samples. Each sample is characterized by 3 continuous variables. Herein, each variable is

mapped into 3 intervals with equal lengths, *i.e.*, I1, I2 and I3. For example, compound 1 can be characterized by a descriptor–interval set {(D1, I1), (D2, I2), (D3, I2)}. Any subset of the above descriptor–interval set is called a chemical pattern, *e.g.* {(D1, I1), (D2, I2)}. The occurrence frequency of a chemical pattern *x* in a given class is called ‘support’, abbreviated as $\text{supp}_C(x)$ (eqn (1)).

$$\text{supp}_C(x) = \frac{\text{count}_C(x)}{|C|} \quad (1)$$

where, $\text{count}_C(x)$ is the number of samples with pattern *x* in class *C*, and $|C|$ is the number of samples in class *C*. From Table 1, it is obvious that the support of pattern {(D1, I1), (D2, I2)} in the positive (C_P) and negative compounds (C_N) is 0.667 and 0.333, respectively. The ratio of the support in C_P to that in C_N is called ‘growth’, abbreviated as $\text{growth}_{C_P/C_N}(x)$ (eqn (2)).

$$\text{growth}_{C_P/C_N}(x) = \frac{\text{supp}_{C_P}(x)}{\text{supp}_{C_N}(x)} \quad (2)$$

Obviously, the $\text{growth}_{C_P/C_N}(x)$ of the pattern {(D1, I1), (D2, I2)} is 2. Growth can reflect the discriminant power of a given pattern, and a higher value indicates a better discriminatory capability.

2.3.2. Classification by aggregating emerging patterns. In most cases, a single pattern is unable to achieve satisfying prediction performance. Thus, it's necessary for a classifier to combine the strength of a set of chemical patterns to

Table 2 The performance of the ECP models

Model	No. of descriptors	No. of ECPs	Training set (%)			Validation set (%)		
			Acc	Sen	Spe	Acc	Sen	Spe
1	1	2	72.0	75.4	69.0	68.4	63.9	72.1
2	2	5	75.0	75.0	75.0	68.4	52.8	81.4
3	3	8	75.0	73.8	76.1	63.3	52.8	72.1
4	4	17	81.8	82.3	81.4	68.3	56.8	77.8
5	5	22	86.4	90.2	83.1	63.3	47.2	76.7
6 ^a	6 ^b	27	83.3	93.4	74.6	81.0	80.6	81.4
7	7	36	79.5	83.6	76.1	77.2	75.0	79.1
8	8	42	83.3	86.9	80.3	82.3	75.0	88.4
9	9	42	81.1	88.5	74.6	78.5	77.8	79.1
10	10	50	81.1	86.9	76.1	83.5	75.0	90.7
11	11	61	84.8	88.5	81.7	84.8	75.0	93.0
12	12	62	84.8	88.5	81.7	79.7	72.2	86.0
13	13	67	86.4	88.5	84.5	77.2	69.4	83.7
14	14	73	80.3	78.7	81.7	74.7	66.7	81.4
15	15	82	83.3	82.0	84.5	74.7	63.9	83.7
16	16	97	83.3	82.0	84.5	77.2	63.9	88.4
17	17	99	82.6	83.6	81.7	75.9	69.4	81.4
18	190	1358	86.4	91.8	81.7	78.5	72.2	83.7

^a The best ECP model with 6 descriptors. ^b Variables involved. nN: the number of nitrogen atoms; fHpbVSA: fractional hydrophobic van der Waals surface area; BV12-DRY: the best hydrophobic volumes generated by the hydrophobic probe at an energy level of $-1.0 \text{ kcal mol}^{-1}$; BV12-OH2: the best hydrophilic volumes generated by the OH2 probe at an energy level of $-3.0 \text{ kcal mol}^{-1}$; ID6-DRY: the distances between the molecular center of mass and the barycenter of the hydrophobic interaction region at an energy level of $-1.2 \text{ kcal mol}^{-1}$; mlogpUB: the Moriguchi octanol–water partition coefficient.

a **b**

Fig. 3 Predictive performances of the ECP models. (a) The performance on the training set; (b) the performance on the validation set.

improve the prediction accuracy. Classification by aggregating emerging patterns (CAEP), proposed by Dong *et al.*,⁴⁵ is the first ECP-based classifier. The principle of CAEP is to assign a sample label according to the summation of all the ECP contributions. The contribution of a pattern is defined as eqn (3).

$$\text{contribution}(x) = \frac{\text{growth}_{C_p C_N}(x)}{\text{growth}_{C_p C_N}(x) + 1} \times \text{supp}_C(x) \quad (3)$$

For a sample *s*, the aggregation score for class *C*, denoted as $\text{score}(s, C)$, is defined according to eqn (4)

$$\text{score}(s, C) = \sum_{x \subseteq s, s \in E(C)} \text{contribution}(x) \quad (4)$$

where $E(C)$ is a set of ECPs discovered in class *C*.

Of course, the number of ECPs discovered may not be balanced among different classes. Thus, $\text{score}(s, C)$ must be normalized (eqn (5)) when compared with different classes.

$$\text{norm_score}(s, C) = \frac{\text{score}(s, C)}{\text{base_score}(s, C)} \quad (5)$$

where $\text{base_score}(s, C)$ is the median score of the samples in class *C*. By comparing the norm_score obtained for each class, a sample is classified into the class with the highest norm_score . Please see ref. 46 for more detailed information.

In this paper, the KT data analysis (KTDA) program⁴⁸ was employed for ECP modeling. Two different ECP discovering algorithms, *i.e.* a maximal frequent item set (MFI) algorithm and a decision tree (DT) algorithm were implemented. Herein, each structural descriptor was firstly mapped into 10

Table 4 The 14 ECPs discovered for TLR8 agonists

No.	Pattern information	Support (%)	Growth rate	Contribution (%)
P1	BV12-OH2 > 17.8 BV12-OH2 ≤ 21.3	36.1	2.8	26.7
P2	BV12-OH2 > 17.8 nN ≤ 3.5 fHpbVSA ≤ 0.4	62.3	4.9	51.8
P3	BV12-OH2 > 34.1 ID6-DRY ≤ 87.8	44.3	3.5	34.4
P4	BV12-OH2 > 35.2 BV12-DRY > 28.7 fHpbVSA ≤ 0.4	24.6	8.7	22.1
P5	BV12-OH2 > 43.4 BV12-DRY > 22 mlogpUB ≤ 7.5	27.9	19.8	26.5
P6	BV12-OH2 > 43.4 BV12-DRY > 28.7	23.0	8.1	20.4
P7	BV12-OH2 > 48 BV12-DRY ≤ 28.7 fHpbVSA ≤ 0.4	8.2	Infinity	8.2
P8	BV12-OH2 ≤ 20.1 BV12-DRY > 22 mlogpUB ≤ 7.5	26.23	9.3	23.7
P9	BV12-OH2 ≤ 20.1 BV12-DRY > 28.7 fHpbVSA ≤ 0.4	24.6	17.5	23.3
P10	BV12-OH2 ≤ 22.4 BV12-DRY > 28.7	37.7	2.4	26.7
P11	BV12-OH2 ≤ 22.4 BV12-DRY > 28.7 ID6-DRY > 34.7	32.8	4.7	27.0
P12	BV12-OH2 ≤ 43.4 BV12-DRY > 22 mlogpUB ≤ 7.5	44.3	2.0	29.3
P13	ID6-DRY > 101.1 ID6-DRY ≤ 121	8.2	Infinity	8.2
P14	fHpbVSA ≤ 0.4	91.8	1.4	53.8

Table 3 The prediction performances of the optimal ECP model and molecular docking

	Training set (%)			Validation set (%)			External test set (%)		
	Acc	Sen	Spe	Acc	Sen	Spe	Acc	Sen	Spe
ECP	83.3	93.4	74.6	81.0	80.6	81.4	80.0	76.9	87
Total score	69.7	73.8	66.2	72.2	72.2	72.1	81.3	86.5	69.6
ECP-Surflex	86.9	93.6	80.8	86.2	85.2	87.1	94.2	97.2	87.5

Fig. 4 Hierarchical cluster analysis of hTLR8 agonists (colored green and blue), and 2 distinct patterns are denoted by white-dashed lines.

bins with equal lengths. Then, the decision tree algorithm was used for ECP mining. The significance level of the ECP was set to 0.02, and the significance of splits during decision tree construction was set to 0.05. The accuracy (Acc), sensitivity (Sen), and specificity (Spe) (eqn (6)–(8)) were used for model evaluations.

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (6)$$

$$\text{Sen} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\% \quad (7)$$

$$\text{Spe} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\% \quad (8)$$

TP: true positive; TN: true negative; FP: false positive and FN: false negative.

2.4 Molecular docking

Surflex-dock (Sybyl 8.1, Tripos Inc.), a flexible molecular docking method, was employed in this paper.⁴⁹ Herein, two crystal structures of TLR8 were used for molecular docking. One is in the apo state (3W3G), and the other is in the bound

state (3W3J) with a co-crystallized agonist CL097. CL097 is an imidazoquinoline derivative similar to the better-known agonist imiquimod. Surflex-dock employs a protocol, an idealized ensemble of CH₄, NH and CO probes, to guide the generation of docking conformers of the ligands.

In this paper, the protocol was generated based on the residues within the 5 Å distance from CL097. The search grid and the number of additional starting conformations were set to 2 Å and 3, respectively. The self-scoring, ring flexibility, soft grid, and pre- and post-dock minimizations were considered in the docking processes. Five scoring functions, *i.e.*, total score, G score, PMF score, D score, and Chem score, were used for ligand ranking, respectively. Before docking, all ligands were optimized by the Tripos force field with a conjugate gradient minimizer (Sybyl 8.1). The maximum iteration steps and energy gradient were set to 5000 times and 0.05 kcal mol⁻¹ Å⁻¹, respectively.

3. Results and discussion

3.1 Feature selection and ECP modeling

A total of 17 variables comprising 10 VolSurf descriptors and 7 physico-chemical descriptors were obtained from the original 220 descriptors by SLDA. The variable selection process and performances of SLDA in 132 training sets are shown in Table S2.†

Based on the results of SLDA, ECP modeling was performed and the performances are shown in Table 2 and

Fig. 5 The ROC curves of the 5 scoring functions (a) 3W3G (apo state); (b) 3W3J (ligand-bound state).

Fig. 3. It can be seen that the prediction performances on the training and validation sets are fairly good for most of the ECP models. With the decreased number of descriptors, the accuracies are maintained at about 80.0% for both the training and validation sets. However, the prediction performances on the validation set begin to decrease drastically when the number of descriptors is less than 6. In comparison with other models, model 6 is simpler and more efficient, of which the number of ECPs is only 27. Thus, model 6 with 6 descriptors is chosen as the best ECP model, of which Acc, Sen and Spe are 83.3%, 93.4%, and 74.6% for the training set and 81.0%, 80.6%, and 81.4% for the validation set, respectively.

In comparison with the support vector machine (SVM) and SLDA models, the ECP model demonstrates strong discriminatory capabilities on both training and validation sets (Tables S2 and S3[†]). To test the robustness of model 6, ECP modeling was repeated 30 times based on the training and validation sets randomly divided. The means of Acc, Sen, Spe are $85.1 \pm 3.2\%$, $86.8 \pm 3.5\%$, and $83.5 \pm 6.0\%$ for the training set, and $76.2 \pm 4.0\%$, $77.3 \pm 7.1\%$, and $75.7 \pm 6.9\%$ for the validation set. The detailed information is shown in Table S4[†].

In addition, a total of 75 samples including 52 TLR8 agonists and 23 non-agonists were collected from the latest studies^{23,30} and used to further test the predictive power of the optimal ECP model. Acc, Sen and Spe on this external test set are 80.0%, 76.9% and 87.0%, respectively (Table 3). It can be deduced that the optimal ECP model has excellent predictive capability.

The variables included in model 6 are nN, fHpbVSA, BV12-DRY, BV12-OH2, ID6-DRY and mlogpUB. The detailed information of these 6 variables is shown in Table 2. It can be deduced that the activities of TLR8 agonists are closely related to hydrophobic, H-bonding, and polar interactions as well as the unbalanced degree of distribution of hydrophobic properties. It is worth noting that nN is the most discriminating variable among the 6 descriptors. The prediction performance of model 1 with only an nN descriptor is fairly good, of which the overall accuracies are around 70.0% for both the training and validation sets.

3.2 Two-dimensional hierarchical cluster analysis of TLR8 agonists and non-agonists

In the optimal ECP model, a total of 27 ECPs were discovered, of which 14 were for agonists and 13 were for non-agonists (Table 4). In consideration of both contributions and growth rates, P2, P3 and P12 are the most important for predicting TLR8 agonists. P2 is the sole ECP, which comprises the most discriminating nN descriptor. Among the 13 ECPs for non-agonists, P27, P25 and P21 have the largest contributions to the prediction of TLR8 non-agonists (Table S5[†]).

Two-dimensional hierarchical cluster analysis (HAC) was then applied to delineate the ECP profiles of TLR8 agonists and non-agonists, respectively. Herein, each molecule is characterized by a binary string of 27 bits, which denotes the presence or absence of the 27 ECPs, respectively. From Fig. 4, it can be seen that the TLR8 agonists and non-agonists happened to be dominated by ECPG1 and ECPG2, respectively. hTLR8 agonists were mainly divided into 2 classes, colored green and blue, respectively. The first class (green) was mainly characterized by P12, P2, P3 *etc.*, and the second (blue) by P12, P2, P10 *etc.*

3.3 Molecular docking

To validate the protocol of Surflex-dock, the co-crystallized CL097 was re-docked into the binding site of TLR8 (PDB ID: 3W3J). The total score of the optimal conformer is 9.43, and the root mean square deviation (RMSD) of heavy atoms is 0.67 Å (Fig. 7a). It can be inferred that Surflex-dock can reproduce the native conformation of CL097 very well.

Then, the TLR8 in bound (PDB ID: 3W3J) and apo (PDB ID: 3W3G) states were used for molecular docking, respectively. Fig. 5a shows the receiver operating characteristic curves (ROCs) in the case of 3W3G. It is clear that the predictive capabilities of the 5 scoring functions are very poor. The areas under the receiver operating characteristic curves (AUCs) are 0.54 (total score), 0.62 (G score), 0.57 (PMF score),

Fig. 6 The lipophilic (upper panel), electrostatic (central panel), and H-bonding (lower panel) potentials mapped into the Connolly surface of the hTLR8 binding site.

0.57 (D score) and 0.51 (Chem score), respectively. The reason for this may be that TLR8 in the apo state is not suitable for scoring ligands, and that the conformational change of TLR8 upon ligand binding should be considered in the docking process.

In the case of 3W3J, the results indicate that the total score has a better discriminatory capability than the other 4 scoring functions (Fig. 5b). The AUC, threshold and Youden index of the total score are 0.73, 6.65 and 0.40, respectively. Although the prediction performances on both the training and validation datasets are somewhat lower than those of the ECP model (Table 3), excellent results are achieved for the external test dataset. Acc, Sen and Spe for 75 test samples are

81.3%, 86.5% and 69.6%, respectively. For the molecular docking method, the prediction performances are, in general, quite satisfactory.

Furthermore, the predictive power caused by combining the ECP model with Surflex-dock (threshold = 6.65) was investigated. The results showed that about 75.0% (99/132) of the training samples, 73.4% (58/79) of the validation samples and 69.3% (52/75) of the test samples gave consistent prediction results. For the samples with consistent results, Acc, Sen, and Spe are 86.9%, 93.6%, and 80.8% for the training set, 86.2%, 85.2%, and 87.1% for the validation set, and 94.2%, 97.2%, and 87.5% for the test set, respectively (Table 3). Thus, it can be deduced that, in the case of large-

Fig. 7 The binding modes of hTLR8 agonists. (a) CL097; (b) 70 023; (c) 80 026; (d) 90 076. Chain a and chain b of hTLR8 are colored blue and green, respectively, and H-bonds are denoted by dashed lines.

scale screening, the hit rate of TLR8 agonists can be greatly enhanced when the ECP model is combined with the Surflex-dock method.

3.4 The relationships between the surface properties of the ligand-binding pocket and the ECP features of TLR8 agonists

According to the results of hierarchical cluster analysis, 3 typical TLR8 agonists, *i.e.*, 70 023, 80 026 and 90 076, were selected for exploring the relationships between the surface properties of the ligand-binding pocket and the ECP profiles of TLR8 agonists. The 3 agonists have different scaffolds and are all characterized by P2, P3 and P12, the major ECPs of TLR8 agonists.

Fig. 6 shows the lipophilic, electrostatic and H-bonding donor/receptor potentials mapped into the Connolly surface of the ligand-binding pocket. It can be observed that the lipophilic potential is asymmetrically distributed in the ligand-binding pocket formed by two chains. The lipophilic region is mainly constituted by Ile349*, Gly351*, Gly376*, Val378*, Ile403* and Phe405* of chain B, and the hydrophilic region

is mainly formed by Thr574, His575, His576, Asp543, Asp545, and Tyr567 of chain A. The asymmetric distributions of the hydrophilic/hydrophobic properties of the two chains determine the specific properties of the TLR8 agonists.

Similarly, the electrostatic potential is also asymmetrically distributed. A significant feature is the negative charge center formed by Asp543 and Asp545 of chain A. The asymmetry in the property distribution can also be observed in the H-bonding potentials of the Connolly surface. In Fig. 6, it is obvious that H-bond donors/acceptors are mainly distributed in the hydrophilic region of chain A. From Fig. 7, it can be observed that H-bonds are important for the TLR8 agonists to interact with the binding-site residues of chain A. It can be seen that molecule 70023 forms five H-bonds with Asp543, Thr574 and Gly572, 80026 forms three H-bonds with Thr574 and Asp543, and 90076 forms one H-bond with Thr574.

Due to the asymmetry in the property distribution of the ligand-binding pocket, the ECPs of the TLR8 agonists show some unique features. As mentioned above, P2, P3 and P12 are the most important ECPs for the TLR8 agonists. P2 ($BV12-OH2 > 17.8$, $nN \leq 3.5$, $fHpbVSA \leq 0.4$) is the sole ECP

which comprises the descriptor nN. This pattern represents a delicate balance among the hydrophilic, hydrophobic and H-bonding properties characterized by BV12-OH2, fHpbVSA and nN.

As for P12 ($BV12-OH2 \leq 43.4$, $BV12-DRY > 22$, $mlogpUB \leq 7.5$), it directly characterizes the amphipathic properties of the TLR8 agonists. The 3 variables included present a subtle balance between the hydrophilic and hydrophobic properties of the TLR8 agonists. P3 ($BV12-OH2 > 34.1$, $ID6-DRY \leq 87.8$) describes not only the hydrophilic properties but also the unbalanced degree of distribution of the hydrophobic properties. Herein, ID6-DRY describes the distance between the molecular center of mass and the barycenter of the hydrophobic interaction region, which is extremely useful for depicting the asymmetry in the hydrophobic properties of the TLR8 agonists.

In general, the asymmetrical distributions of the lipophilic, electrostatic and H-bonding donor/receptor potentials of the binding pocket of TLR8 determine some specific and subtle characteristics of TLR8 agonists, which can be quantitatively characterized by P2, P3 and P12. In this context, the ECP has an overwhelming advantage over other structural description methods.

4. Conclusions

In this study, an ECP was introduced to establish a predictive model of TLR8 agonists. By using only 6 descriptors, an optimal ECP model was derived with excellent prediction performances on the training, validation and test datasets. The results demonstrate that the 27 ECPs discovered from the 6 descriptors are discriminatory and can capture the subtle structural differences between the TLR8 agonists and non-agonists. By combining the optimal ECP model with the Surflex-dock method, about 73.1% of the samples (209/286) gave consistent prediction results. The hit rate of TLR8 agonists was greatly enhanced with sensitivities of 93.6%, 85.2% and 97.2% for the training, validation and test datasets, respectively.

Besides, the latent relationships between the ECP profiles of the TLR8 agonists and the surface properties of the ligand-binding pocket of TLR8 were investigated by ECP mining and Connolly surface analysis. The results showed that P2, P3 and P12 are major prerequisites for TLR8 agonists, which can characterize their H-bonding, hydrophilic, and hydrophobic properties as well as the unbalanced degree of property distributions.

In summary, the ECP is a very efficient way to delineate the subtle structural properties of TLR8 agonists, and the strategy of combining the ECP with Surflex-dock can be widely used in molecular design and pattern recognition of TLR8 agonists. It should be noted that in the revision process of this manuscript, the ECP model was further blind-validated using 46 TLR8 agonists collected from the most recent studies.^{50–52} The prediction performances were as excellent as before (Tables S6 and S7[†]). However, the efficacy and

application domain of the ECP model still need to be tested by *in vitro* experiments.

Abbreviations

TLRs	Toll-like receptors
ECP	Emerging chemical pattern
ECPG	Emerging chemical pattern group
LRR	Leucine-rich repeat
ssRNA	Single stranded RNA
SLDA	Linear discriminant analysis
KTDA	KT data analysis
MFI	Maximal frequent itemsets
DT	Decision trees
Acc	Accuracy
Sen	Sensitivity
Spe	Specificity
TP	True positive
TN	True negative
FP	False positive
FN	False negative
HAC	Hierarchical cluster analysis

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (Project No. 106112017CDJQJ238816) and the Collaborative Fund of the Science and Technology Agency of Sichuan Province (Project No. 2016LZXNYD-J19). We are thankful to Prof. Roman Podraza (Warsaw University of Technology, Poland) for providing the KTDA system for ECP mining and modeling.

References

- 1 S. Akira, K. Takeda and T. Kaisho, *Nat. Immunol.*, 2001, 2, 675–680.
- 2 B. Lemaitre, E. Nicolas, L. Michaut, J. M. Reichhart and J. A. Hoffmann, *Cell*, 1996, 86, 973–983.
- 3 C. A. Janeway, *Cold Spring Harbor Symp. Quant. Biol.*, 1989, 54, 1–13.
- 4 T. Kawai and S. Akira, *Nat. Immunol.*, 2010, 11, 373–384.
- 5 S. Akira and K. Takeda, *Nat. Rev. Immunol.*, 2004, 4, 499–511.
- 6 A. Aderem and R. J. Ulevitch, *Nature*, 2000, 406, 782–787.
- 7 H. Hemmi, O. Takeuchi, T. Kawai, T. Kaisho, S. Sato, H. Sanjo, M. Matsumoto, K. Hoshino, H. Wagner, K. Takeda and S. Akira, *Nature*, 2000, 408, 740–745.
- 8 K. Takeda, T. Kaisho and S. Akira, *Annu. Rev. Immunol.*, 2003, 21, 335–376.
- 9 M. Czarniecki, *J. Med. Chem.*, 2008, 51, 6621–6626.

- 10 S. S. Diebold, T. Kaisho, H. Hemmi, S. Akira and C. R. E. Sousa, *Science*, 2004, **303**, 1529–1531.
- 11 F. Heil, H. Hemmi, H. Hochrein, F. Ampenberger, C. Kirschning, S. Akira, G. Lipford, H. Wagner and S. Bauer, *Science*, 2004, **303**, 1526–1529.
- 12 J. Y. Kang and J. O. Lee, *Annu. Rev. Biochem.*, 2011, **80**, 917–941.
- 13 U. Ohto, H. Tanji and T. Shimizu, *Microbes Infect.*, 2014, **16**, 273–282.
- 14 M. Jurk, F. Heil, J. Vollmer, C. Schetter, A. M. Krieg, H. Wagner, G. Lipford and S. Bauer, *Nat. Immunol.*, 2002, **3**, 499.
- 15 H. Tanji, U. Ohto, T. Shibata, K. Miyake and T. Shimizu, *Science*, 2013, **339**, 1426–1429.
- 16 H. Tanji, U. Ohto, T. Shibata, M. Taoka, Y. Yamauchi, T. Isobe, K. Miyake and T. Shimizu, *Nat. Struct. Mol. Biol.*, 2015, **22**, 109–115.
- 17 M. Geyer, K. Pelka and E. Latz, *Nat. Struct. Mol. Biol.*, 2015, **22**, 99–101.
- 18 U. Wille-Reece, B. J. Flynn, K. Lore, R. A. Koup, R. M. Kedl, J. J. Mattapallil, W. R. Weiss, M. Roederer and R. A. Seder, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 15190–15194.
- 19 H. Kanzler, F. J. Barrat, E. M. Hessel and R. L. Coffman, *Nat. Med.*, 2007, **13**, 552–559.
- 20 H. P. Kokatla, D. Sil, S. S. Malladi, R. Balakrishna, A. R. Hermanson, L. M. Fox, X. Wang, A. Dixit and S. A. David, *J. Med. Chem.*, 2013, **56**, 6871–6885.
- 21 D. B. Salunke, E. Yoo, N. M. Shukla, R. Balakrishna, S. S. Malladi, K. J. Serafin, V. W. Day, X. K. Wang and S. A. David, *J. Med. Chem.*, 2012, **55**, 8137–8151.
- 22 H. P. Kokatla, E. Yoo, D. B. Salunke, D. Sil, C. F. Ng, R. Balakrishna, S. S. Malladi, L. M. Fox and S. A. David, *Org. Biomol. Chem.*, 2013, **11**, 1179–1198.
- 23 M. Beesu, G. Caruso, A. C. D. Salyer, N. M. Shukla, K. K. Khetani, L. J. Smith, L. M. Fox, H. Tanji, U. Ohto, T. Shimizu and S. A. David, *J. Med. Chem.*, 2016, **59**, 3311–3330.
- 24 M. Beesu, G. Caruso, A. C. D. Salyer, K. K. Khetani, D. Sil, M. Weerasinghe, H. Tanji, U. Ohto, T. Shimizu and S. A. David, *J. Med. Chem.*, 2015, **58**, 7833–7849.
- 25 M. Beesu, S. S. Malladi, L. M. Fox, C. D. Jones, A. Dixit and S. A. David, *J. Med. Chem.*, 2014, **57**, 7325–7341.
- 26 C. E. Schiaffo, C. Shi, Z. Xiong, M. Olin, J. R. Ohlfest, C. C. Aldrich and D. M. Ferguson, *J. Med. Chem.*, 2014, **57**, 339–347.
- 27 C. Shi, Z. Xiong, P. Chittep, C. C. Aldrich, J. R. Ohlfest and D. M. Ferguson, *ACS Med. Chem. Lett.*, 2012, **3**, 501–504.
- 28 B. Pope, J. Sigindere, E. Chourmouzis, P. MacIntyre and M. Goodman, *Cancer Immunol. Immunother.*, 1994, **38**, 83–91.
- 29 D. C. McGowan, F. Herschke, F. Pauwels, B. Stoops, S. Last, S. Pieters, A. Scholliers, T. Thoné, B. V. Schoubroeck and D. D. Pooter, *J. Med. Chem.*, 2016, **59**, 7936.
- 30 M. Beesu, A. C. D. Salyer, K. L. Trautman, J. K. Hill and S. A. David, *J. Med. Chem.*, 2016, **59**, 8082–8093.
- 31 E. Yoo, D. B. Salunke, D. Sil, X. Guo, A. C. D. Salyer, A. R. Hermanson, M. Kumar, S. S. Malladi, R. Balakrishna, W. H. Thompson, H. Tanji, U. Ohto, T. Shimizu and S. A. David, *J. Med. Chem.*, 2014, **57**, 7955–7970.
- 32 F. F. Deng, S. Y. Ma, M. H. Xie, X. Y. Zhang, P. Z. Li and H. L. Zhai, *Mol. Biosyst.*, 2014, **10**, 2202–2214.
- 33 J. Auer and J. Bajorath, *J. Chem. Inf. Model.*, 2006, **46**, 2502–2514.
- 34 J. E. Neve, H. P. Wijesekera, S. Duffy, I. D. Jenkins, J. A. Ripper, S. J. Teague, M. Campitelli, A. Garavelas, G. Nikolakopoulos, P. V. Le, P. D. Leone, N. B. Pham, P. Shelton, N. Fraser, A. R. Carroll, V. M. Avery, C. McCrae, N. Williams and R. J. Quinn, *J. Med. Chem.*, 2014, **57**, 1252–1275.
- 35 P. A. Roethle, R. M. McFadden, H. Yang, P. Hrvatin, H. Hui, M. Graupe, B. Gallagher, J. Chao, J. Hesselgesser, P. Duatschek, J. Zheng, B. Lu, D. B. Tumas, J. Perry and R. L. Halcomb, *J. Med. Chem.*, 2013, **56**, 7324–7333.
- 36 A. Kurimoto, K. Hashimoto, T. Nakamura, K. Norimura, H. Ogita, H. Takaku, R. Bonnert, T. McNally, H. Wada and Y. Isobe, *J. Med. Chem.*, 2010, **53**, 2964–2972.
- 37 C. Cruciani, P. Crivori, P. A. Carrupt and B. Testa, *J. Mol. Struct.: THEOCHEM*, 2000, **503**, 17–30.
- 38 I. Zamora, T. Oprea, G. Cruciani, M. Pastor and A. L. Ungell, *J. Med. Chem.*, 2003, **46**, 25–33.
- 39 H. van de Waterbeemd and E. Gifford, *Nat. Rev. Drug Discovery*, 2003, **2**, 192–204.
- 40 P. Crivori, G. Cruciani, P. A. Carrupt and B. Testa, *J. Med. Chem.*, 2000, **43**, 2204–2216.
- 41 G. Cruciani, M. Pastor and W. Guba, *Eur. J. Pharm. Sci.*, 2000, **11**, S29–S39.
- 42 C. Tintori, M. Magnani, S. Schenone and M. Botta, *Eur. J. Med. Chem.*, 2009, **44**, 990–1000.
- 43 G. Ermondi, G. Caron, I. G. Pintos, M. Gerbaldo, M. Perez, D. I. Perez, Z. Gandara, A. Martinez, G. Gomez and Y. Fall, *Eur. J. Med. Chem.*, 2011, **46**, 860–869.
- 44 S. J. Qu, S. H. Huang, X. C. Pang, L. Yang and H. Mei, *J. Chem. Inf. Model.*, 2016, **56**, 2061–2068.
- 45 G. Z. Dong, X. Z. Zhang, L. S. Wong and J. Y. Li, *Discovery Science, Proceedings*, 1999, vol. 1721, pp. 30–42.
- 46 X. C. Pan, L. Chao, W. Tan, L. Yang, R. Podraza and H. Mei, *Chemom. Intell. Lab. Syst.*, 2014, **137**, 140–145.
- 47 X. C. Pan, H. Mei, S. J. Qu, S. H. Huang, J. Y. Sun, L. Yang and H. Chen, *Int. J. Pharm.*, 2016, **502**, 61–69.
- 48 R. Podraza and K. Tomaszewski, *Annales UMCS, Informatica*, 2006, **4**, 279–290.
- 49 A. N. Jain, *J. Med. Chem.*, 2003, **46**, 499–511.
- 50 M. Beesu, H. P. Kokatla and S. A. David, *Methods Mol. Biol.*, 2017, **1494**, 29–44.
- 51 P. Larson, T. A. Kucaba, Z. Xiong, M. Olin, T. S. Griffith and D. M. Ferguson, *ACS Med. Chem. Lett.*, 2017, **8**, 1148–1152.
- 52 S. Pieters, D. McGowan, F. Herschke, F. Pauwels, B. Stoops, S. Last, W. Embrechts, A. Scholliers, W. Mostmans, K. Van Dijck, B. Van Schoubroeck, T. Thoné, D. De Pooter, G. Fanning, M. L. Rosauro, M. D. Khamlichi, I. Houppis, E. Arnoult, T. H. M. Jonckers and P. Raboisson, *Bioorg. Med. Chem. Lett.*, 2018, **28**, 711–719.