**Supplementary materials**

**DeepEfflux: a 2D Convolutional Neural Network Model for Identifying Families of Efflux Proteins in Transporters**

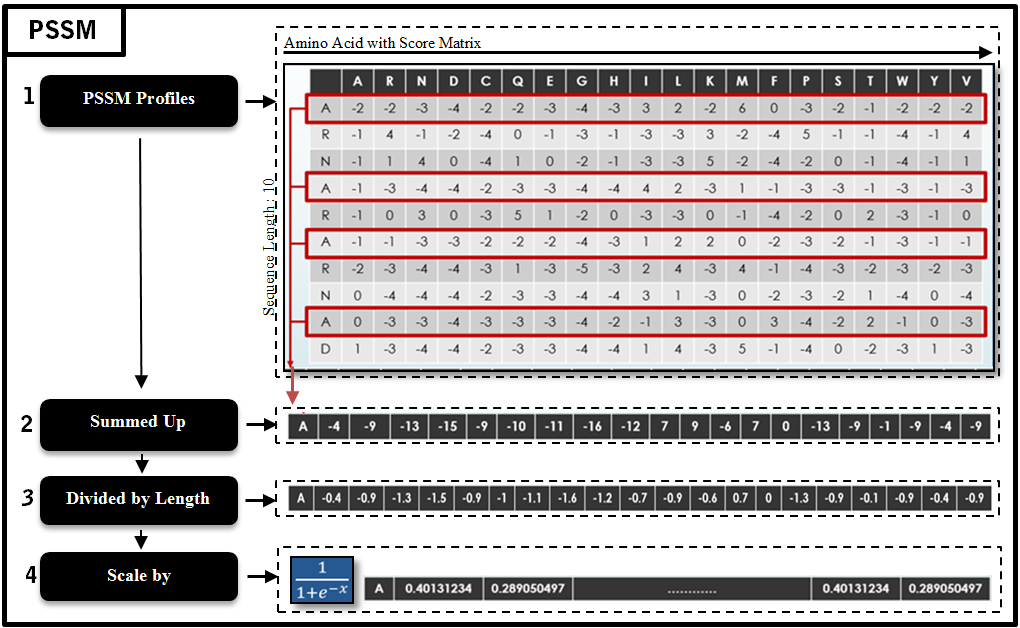
**F-Score Selection Feature**F-score or fisher score is a method to evaluate and measure the discrimination of two sets of real numbers consists of positive and negative datasets.

**Formula S1.** F-score Formula

Based on the F-score formula above, where *n+* is the number of positive instance, *n*- is the number of negative instance, *i* is each feature of the dataset and *k* is instance of the *i*th feature.  is the average of the *i*th feature of dataset, withis the average of the *i*th feature of positive datasets,  is the average of the *i*th feature of negative datasets, is the *i*th feature of the *k*th positive instance of feature and  is the *i*th feature of the *k*th negative instance of feature. Selection feature focus on the raking of *i*th features.

## Position Specific Scoring Matrix (PSSM)

A PSSM profile represents the vector {x(i,j) has *N* x 20 matrix or element (residue substitution) with *N* vectors {x(i,j), i=1,2,.....N and j=1,2,....n}, where *N* is the length of the target protein sequence and *n* is 20 amino acids. In the Figure S1 shows PSSM method to calculate 400 dimension input vectors from position component of PSSM profiles:

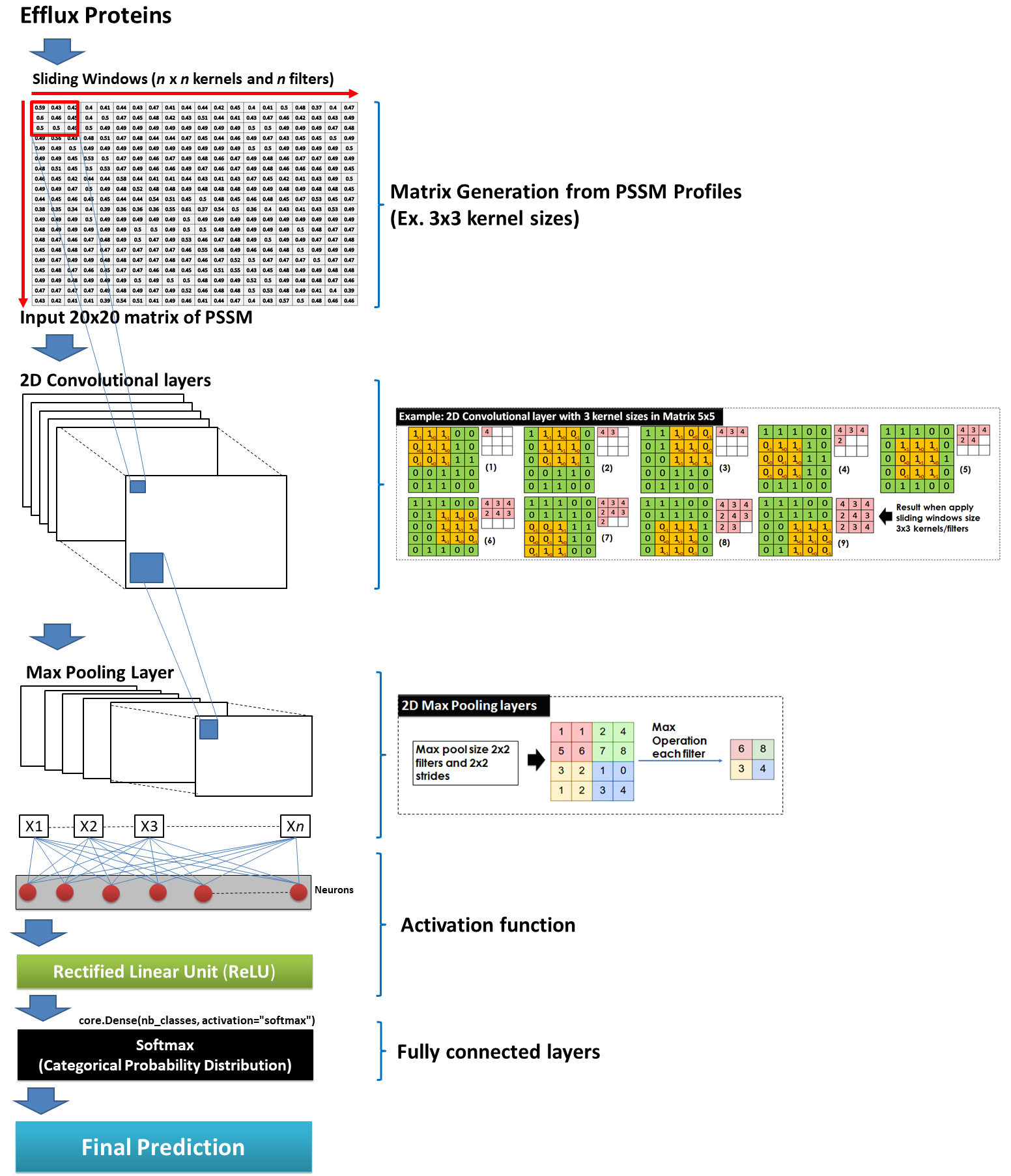


**Figure S1.** PSSM Method

**Assessment of predictive ability**

These classification performance indicators are often used by the following symbols (TP, FP, TN, FN are true positive, false positive, true negative, and false negative respectively). The relations between these symbols and the symbols in Eqs. (1,2,3 and 4) are given by:

where, for example, represents the number of class A proteins incorrectly predicted to be of non-A class, represents the number of non-A class proteins incorrectly predicted to be of class A proteins, represents the number of class A proteins investigated and represents the number of non-A proteins investigated. This convention is also applied to the classification of class B proteins with non-B class proteins, and class C proteins with non-C class proteins.



**Figure S2.** The procedure from inputting a PSSM profile into a CNN model, then passed through a series of convolutional, nonlinearity, pooling (down-sampling), and fully connected layers, and finally to an output.



**Figure S3.** Word cloud technique used to visualize *n*-gram with highest frequencies in Class A, B and C.

**Table S1.** 10 most frequent motifs in class A, B and C. The top 10 motif features represented as an *n*-gram contiguous sequence of amino acid residues. An *n*-gram of amino acid length 1 is referred to as a unigram; those of length 2, length 3, length 4 and length 5 are called bigram, trigram, four-gram and five-gram respectively. We also counted the number of time (Freg) these *n*-grams appeared in the protein sequences as well.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | Unigram | | Bigram | | Trigram | | Four-gram | | Five-gram | |
| Pair(s) | Freg | Pair(s) | Freg | Pair(s) | Freg | Pair(s) | Freg | Pair(s) | Freg |
| Class A | L | 10136 | LL | 1370 | LLL | 184 | LLLL | 35 | PESPR | 11 |
| A | 7675 | AL | 1043 | ALL | 164 | ALLL | 27 | LLLLL | 8 |
| G | 7022 | LG | 1025 | GLL | 144 | LGLA | 24 | ESPRW | 8 |
| S | 6755 | LA | 960 | LLA | 143 | LLAG | 23 | LPESP | 7 |
| V | 6302 | GL | 910 | LLG | 136 | LLGL | 23 | SPRWL | 7 |
| I | 5845 | LS | 858 | LGL | 133 | LLAA | 20 | GALLL | 7 |
| F | 5071 | AA | 819 | LAL | 126 | ALAL | 20 | GVVVL | 6 |
| T | 4630 | SL | 812 | ALG | 122 | AAAA | 19 | ALLPL | 6 |
| P | 3579 | VL | 794 | LLS | 122 | LLLG | 19 | LLIGL | 6 |
| R | 3390 | LV | 783 | AAA | 117 | GALL | 19 | ALLGS | 6 |
| Class B | L | 2804 | AA | 335 | AAA | 55 | AAAA | 12 | RIKEE | 5 |
| A | 2583 | LL | 332 | AAL | 44 | LLAL | 11 | LAVAL | 5 |
| V | 1928 | LA | 327 | LAA | 44 | AALL | 10 | AAGDP | 4 |
| S | 1771 | AL | 315 | LAL | 43 | LAVA | 10 | AARAA | 4 |
| G | 1659 | LV | 244 | ALL | 41 | LALL | 9 | ARAAG | 4 |
| T | 1477 | VL | 242 | AVA | 41 | LSLG | 9 | ATVVR | 4 |
| I | 1364 | VA | 223 | LAV | 41 | LLAA | 9 | DATVV | 4 |
| R | 1305 | AV | 222 | LLL | 40 | AVAL | 9 | GALSA | 4 |
| P | 1283 | GL | 206 | LLA | 39 | LALV | 9 | LGKHV | 4 |
| E | 1169 | LG | 205 | LVL | 36 | ALSA | 8 | TPALE | 4 |
| Class C | L | 21667 | LL | 2811 | LLL | 376 | LSGG | 91 | NNNNN | 60 |
| A | 16053 | AL | 2052 | ALL | 278 | NNNN | 76 | LSGGQ | 38 |
| V | 13584 | LA | 2003 | LLA | 269 | GKST | 65 | LDEPT | 37 |
| G | 13184 | AA | 1720 | LAL | 255 | LLLL | 62 | GSGKS | 36 |
| S | 12961 | VL | 1686 | ALA | 249 | LDEP | 55 | GKSTL | 32 |
| I | 12157 | LS | 1646 | AAL | 242 | GSGK | 53 | NGAGK | 32 |
| T | 10557 | LV | 1600 | LAA | 230 | GAGK | 50 | SGKST | 30 |
| F | 9021 | SL | 1596 | AAA | 228 | DEPT | 48 | GAGKT | 29 |
| E | 8992 | GL | 1569 | VLL | 228 | SGGQ | 48 | SGGQK | 27 |
| K | 8680 | LG | 1458 | SLL | 221 | ALLL | 47 | AGKTT | 24 |

**Table S2** - Properties with 20 Topmost F-scores of AAIndex

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **AAindex** | **The description of each index** | **F-score** |
| **Class A** | **VASM830101** | Relative population of conformational state A (Vasquez et al., 1983) | 0.342460827 |
| **AURR980111** | Normalized positional residue frequency at helix termini C5 (Aurora-Rose, 1998) | 0.22854235 |
| **WERD780104** | Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978) | 0.219960606 |
| **AURR980106** | Normalized positional residue frequency at helix termini N1 (Aurora-Rose, 1998) | 0.219906999 |
| **FUKS010104** | Surface composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001) | 0.213159864 |
| **AURR980119** | Normalized positional residue frequency at helix termini C"' (Aurora-Rose, 1998) | 0.211508336 |
| **PRAM900101** | Hydrophobicity (Prabhakaran, 1990) | 0.203872806 |
| **JANJ780101** | Average accessible surface area (Janin et al., 1978) | 0.197629038 |
| **FUKS010101** | Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001) | 0.196095873 |
| **JANJ780103** | Percentage of exposed residues (Janin et al., 1978) | 0.195883979 |
| **JACR890101** | Weights from the IFH scale (Jacobs-White, 1989) | 0.195326358 |
| **KRIW790102** | Fraction of site occupied by water (Krigbaum-Komoriya, 1979) | 0.194565114 |
| **FUKS010102** | Surface composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001) | 0.192749547 |
| **EISD860102** | Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986) | 0.191356573 |
| **ZIMJ680103** | Polarity (Zimmerman et al., 1968) | 0.189555749 |
| **RICJ880116** | Relative preference value at C' (Richardson-Richardson, 1988) | 0.189099037 |
| **DESM900102** | Average membrane preference: AMP07 (Degli Esposti et al., 1990) | 0.186099746 |
| **WOLS870103** | Principal property value z3 (Wold et al., 1987) | 0.180035419 |
| **PARS000102** | p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000) | 0.178836994 |
| **KRIW790101** | Side chain interaction parameter (Krigbaum-Komoriya, 1979) | 0.174803732 |
| **Class B** | **GARJ730101** | Partition coefficient (Garel et al., 1973) | 0.296978111 |
| **WIMW960101** | Free energies of transfer of AcWl-X-LL peptides from bilayer interface to water (Wimley-White, 1996) | 0.274329375 |
| **OOBM850102** | Optimized propensity to form reverse turn (Oobatake et al., 1985) | 0.243505547 |
| **NOZY710101** | Transfer energy, organic solvent/water (Nozaki-Tanford, 1971) | 0.227153082 |
| **MEEJ800101** | Retention coefficient in HPLC, pH7.4 (Meek, 1980) | 0.220987127 |
| **ZASB820101** | Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982) | 0.220362644 |
| **RICJ880113** | Relative preference value at C2 (Richardson-Richardson, 1988) | 0.217285247 |
| **MEEJ800102** | Retention coefficient in HPLC, pH2.1 (Meek, 1980) | 0.210276678 |
| **MEEJ810102** | Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981) | 0.206647342 |
| **OOBM770103** | Long range non-bonded energy per atom (Oobatake-Ooi, 1977) | 0.204857174 |
| **ZHOH040101** | The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004) | 0.193541368 |
| **LEVM760107** | van der Waals parameter epsilon (Levitt, 1976) | 0.192850155 |
| **NADH010105** | Hydropathy scale based on self-information values in the two-state model (25% accessibility) (Naderi-Manesh et al., 2001) | 0.19279994 |
| **MEEJ810101** | Retention coefficient in NaClO4 (Meek-Rossetti, 1981) | 0.191013455 |
| **NAKH900104** | Normalized composition of mt-proteins (Nakashima et al., 1990) | 0.188007988 |
| **MEIH800103** | Average side chain orientation angle (Meirovitch et al., 1980) | 0.187925774 |
| **VASM830101** | Relative population of conformational state A (Vasquez et al., 1983) | 0.187675516 |
| **RADA880102** | Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988) | 0.18691035 |
| **AURR980107** | Normalized positional residue frequency at helix termini N2 (Aurora-Rose, 1998) | 0.184972275 |
| **ROBB790101** | Hydration free energy (Robson-Osguthorpe, 1979) | 0.184374374 |
| **Class C** | **VASM830101** | Relative population of conformational state A (Vasquez et al., 1983) | 0.150822796 |
| **WOLS870103** | Principal property value z3 (Wold et al., 1987) | 0.105607418 |
| **WERD780104** | Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978) | 0.099871764 |
| **RACS820112** | Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982) | 0.099512295 |
| **EISD860102** | Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986) | 0.094805553 |
| **AURR980119** | Normalized positional residue frequency at helix termini C"' (Aurora-Rose, 1998) | 0.093718211 |
| **AURR980106** | Normalized positional residue frequency at helix termini N1 (Aurora-Rose, 1998) | 0.088546859 |
| **JANJ780101** | Average accessible surface area (Janin et al., 1978) | 0.087616886 |
| **AURR980111** | Normalized positional residue frequency at helix termini C5 (Aurora-Rose, 1998) | 0.086799187 |
| **JANJ780103** | Percentage of exposed residues (Janin et al., 1978) | 0.086797855 |
| **FUKS010101** | Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001) | 0.083930301 |
| **AURR980101** | Normalized positional residue frequency at helix termini N4'(Aurora-Rose, 1998) | 0.082480464 |
| **PRAM900101** | Hydrophobicity (Prabhakaran, 1990) | 0.082064889 |
| **FUKS010104** | Surface composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001) | 0.08154551 |
| **RACS820103** | Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982) | 0.081386685 |
| **ZIMJ680103** | Polarity (Zimmerman et al., 1968) | 0.081298788 |
| **PARS000102** | p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000) | 0.07919487 |
| **HOPA770101** | Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982) | 0.075894642 |
| **DESM900102** | Average membrane preference: AMP07 (Degli Esposti et al., 1990) | 0.0757959 |
| **JACR890101** | Weights from the IFH scale (Jacobs-White, 1989) | 0.075440664 |